1 Platonic model of mind as an approximation to neurodynamics

Włodzisław Duch

Department of Computer Methods, Nicholas Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland.
E-mail: duch@phys.uni.torun.pl

Abstract. One of the biggest challenges of science today is to outline connections between the subjective world of human experience, as studied by psychology, and the objective world of measurable brain events, as studied by neuroscience. In this paper a series of approximations to neural dynamics is outlined, leading to a phenomenological theory of mind based on concepts directly related to human cognition. Behaviorism is based on an engineering approach, treating the mind as a control system for the organism. This corresponds to an approximation of the recurrent neural dynamics (brain states) by finite state automata (behavioral states). Another approximations to neural dynamics is described, leading to a Platonic-like model of mind based on psychological spaces. Objects and events in these spaces correspond to quasi-stable states of brain dynamics and may be interpreted from psychological point of view. Platonic model bridges the gap between the neurophysiological brain events and higher cognitive functions realized by the mind. Categorization experiments with human subjects are presented as a challenge for mind-brain theories. Wider implications of this model as a basis for cognitive science are discussed and possible extensions outlined. [1]

1.1 Introduction

Two distinct approaches to understanding of human mind were developed in science. Artificial intelligence aims at building intelligent systems starting from the processing of symbols [1]. There are serious problems at the very foundation of such an approach, starting with the famous mind-body problem (how can the non-material mind interact with matter), the symbol grounding problem (how can the meaning be defined in a self-referential symbolic system) or the frame problem (catastrophic breakdowns of intelligent behavior for “obvious” tasks) [3]. There is no doubt that higher cognitive functions result from the activity of the brain and thus should be implementable by neural networks [4]. It is clear that the present shortcomings of neural networks are connected with the lack of modularity and low complexity of the models rather then with the inherent limitations of the neural modeling itself. Computational cognitive neuroscience [6], recent descendant of neurodynamics [5], tries to unravel the details of neural processes responsible for

[1] This is an extended version of a paper written for the “Brain-like Computing and Intelligent Information System” book (ed. by S-i. Amari and N. Kasabov, Springer 1997)

brain functions. Recently even developmental psychology has been influenced by neurodynamics [7].

Although large body of empirical facts have been accumulated in cognitive psychology [2] so far only very few attempts that aim at a unified theory of cognition have been made. They came mostly from the artificial intelligence perspective. John Anderson's series of ACT models [2], developed by him and his collaborators at the Carnegi-Mellon University and elsewhere in the past 20 years [9,10] were perhaps the first projects aimed at global theory of cognitive science. The model is based on a classical production system, using IF ... THEN rules, extended in more recent ACT-R version to deal with dynamical aspects of cognition by optimizing the structure of the system. The model has three types of memories, declarative, procedural and working memory and is quite successful in modeling a variety of high level cognitive phenomena, such as memory effects, priming, or even learning simple programming techniques and theorem proving. Another project, Model Human Processor (MHP), was started at Xerox PARC company as a model to design human-machine interfaces. Allen Newell, a co-author of the MHP project, wrote a book *Unified theories of cognition* [2] promoting the view that there is enough empirical evidence from cognitive psychology and neuropsychology to create many alternative unified theories of cognition. His own attempts are based on an expert system called SOAR[3], a sophisticated system using production rules and symbol processing. Both ACT and SOAR are well developed systems quite successful in modeling many psychological phenomena and aiming at architectures capable of a full range of cognitive tasks. Their success indicates that higher cognition may be to some degree approximated by production rules. However, direct evidence for independent production rules, with condition-action pairs stored in human memory, is missing. It is quite feasible that production systems are powerful enough to model any behavior. Thus it is not clear that human cognition may really be understood in this way. Third large-scale project in the logical tradition, named OSCAR, is pursued by John Pollock [11] at the University of Arizona, and is based on probabilistic reasoning [4]. The goal of this project is to create a fully functioning rational agent. OSCAR inference engine is used in some real-world applications, such as decision support systems for medicine. The project is slowly gaining momentum but even if it will be successful it may not tell us much about human cognition.

Computational neuroscience provides perhaps a better path for understanding of mind through models of brain functions, although nothing comparable to ACT or SOAR systems has been build so far using connectionist paradigm. One grand proposal for the theory of cognition based on specific brain models has been worked out by Callataÿ [12]. Unfortunately his book is too speculative and not supported by computational models to attract greater interest. An interesting attempt at simplification of modular cerebral cortex architecture has been presented by Burnod [13]. His book certainly deserves wider recognition and may serve as the basis of

[2] See the WWW homepage of ACT at: http://sands.psy.cmu.edu/

[3] WWW homepage: http://www.isi.edu/soar/soar-homepage.html

[4] WWW homepage: http://www.u.arizona.edu/˜pollock/

brain-like information processing systems, although it ignores cognitive psychology. Brooks [14] started an interesting project called Cog[5], aimed at development of the behavior-based intelligence of a humanoid robot. It remains to be seen what level of intelligence this approach will achieve.

Computational neuroscience may be our best approach to ultimate understanding of the brain and mind but chances that neural models are going to explain soon all aspect of cognition are small. Can we understand higher mental activity directly in terms of neural processes in the brain? It does not seem likely. Physicists claimed that chemistry has been reduced to physics since the birth of quantum mechanics. Quantum chemistry, basic theory of chemical systems, has found wider acceptation by chemists only very recently. Even in chemistry and physics phenomenological concepts that are not easily reducible to fundamental interactions, are still used, in fact experimental chemists have hardly been affected by the development of quantum chemistry [8]. Macroscopical theories are reducible only in principle to microscopical descriptions, in practice phenomenological approach to complex systems is most fruitful. Language of neuroscience and language of psychology are quite different. Understanding of the brain requires intermediate theories, between neural and mental, physical and symbolic. Great progress in revealing of biochemical and neurological mechanisms has not yet lead to the comparable progress in understanding of higher cognitive functions of the mind. Computational neuroscience has been born only recently but bearing in mind the complexity of the systems it has to deal with it may take many years before any interesting predictions would be possible. Perhaps all we can hope for is to have a general view on the problem, to look for specific theories on different levels of complexity and to search for the bridges between different levels.

There are two large branches associated with neural modeling paradigm. In the first branch connectionist systems are used for cognitive modeling, replacing production systems with networks of interacting pseudo-neurons [15]. Neuronic equations of Caianello [16] were among the first interesting qualitative models. A very interesting work using modular neural networks for categorization and learning has been published by Murre [17]. His CALM (Categorization And Learning Module) networks, inspired by the neocortical minicolumns, represent quite successful attempt at biologically and psychologically plausible computational models so far. Such models are a step closer to what the brain does, although rarely there is a connection with biological reality.

The second branch attempts to link computer simulations of brain structures with experimental approaches. Successes are still rare and somehow restricted to lower-level cognition (cf. [18,4]), although there are exceptions. Analysis of Miyashita experiments [19] on visual perception in monkeys by Griniasty *et.al*[20] using attractor neural network of the Hopfield type elucidated results of the single-neuron recordings and showed how temporal correlations in the sequence of pictures are changed into spatial correlations between attractors in the phase space. One area in which there is some theoretical and experimental interplay is in the development of

[5] WWW page: http://www.ai.mit.edu/projects/cog/

topographical feature maps, for example in the visual system (cf. Erwin *et.al*[21]). Several books on application of the theory of dynamical systems to human development process appeared recently (cf. [7,24]). Although very promising and useful as a metaphoric language dynamical systems theory has not yet produced quantitative explanations of the brain functions. A series of papers by Ingberg [22] on statistical mechanics of neocortical interactions (SMNI) written in over more than a decade formulate a mesoscopic theory of neural tissue. Of particular importance is his analysis of multiple scales in scalp EEG and explanation of the 7±2 and similar rules in psychology. Global properties of short term memory (STM) are easier to describe using statistical rather than microscopic theories.

Cognitive science seems to be a collection of loosely related subjects without central theory (cf. [25]). How should we approach the study of mind? In this paper I argue that a new language to speak of mind events, linking neurodynamics with cognitive psychology, is needed. Although we clearly are still at the beginning a plausible view of the mind is possible today and many threads in the tapestry of the theory of mind are already visible. Such a view, from brain processes to mind events or from computational neuroscience to cognitive psychology, is presented in this paper. It is speculative but it gives a badly needed framework to the cognitive sciences. In the next section models and approximations to neural functions at various levels, from subcellural to the large brain structures, are discussed. Relations between different levels of modeling and possibilities of reduction of these models to more fundamental level are of particular interest. In the third section the question how to model the mind is addressed directly and basic ideas of the Platonic model of mind are introduced. Feature Space Mapping, a specific neurofuzzy realization based on inspirations derived from Platonic model, is presented in the next section. In the fifth section categorization experiments, the simplest higher-cognition phenomena in psychology, are discussed. A very challenging problem in cognitive science is to understand the results of these experiments using different approximations to neural dynamics. A brief discussion on the future of the Platonic model closes this paper.

1.2 Hierarchical approximations to neurodynamics

Nervous system has distinct organization that may be investigated at many different temporal and spatial scales or levels [4], starting from single molecules (spatial scale of the order of 10^{-10} m), through synapses, neurons, small networks, topographical maps, brain systems, up to the whole brain and central nervous system level. Understanding the relations between different levels, approximations to reach from lower to higher levels, and trying to find the place for mind in this scheme is a very fruitful exercise. Each level of description has its specific theories, methods and goals. The very notion of “levels” is approximate since different levels cannot usually be decoupled.

Cognitive phenomena	Levels and Models	Theories/Methods
Reasoning, problem solving, thinking, creativity	Brain, rational mind; knowledge-based systems, symbolic processing	Artificial intelligence, psychology of thinking
Behavior, immediate reactions ($t \approx 1$ s), associations	Large brain structures; probabilistic and finite state automata	Cognitive and behavioristic psychology, machine learning
Intuition, categorization, recognition	Computational maps, transcortical neural assemblies; Platonic models, feature spaces	Psychology, biology, machine learning
Memory effects	Small networks; recurrent neural networks and dynamical systems	Computational neuroscience, EEG, MEG, fMRI
Learning, internal representations	Neural assemblies; spiking neurons	Neurophysiology, single and multielectrode recordings
Conditioning, basic forms of learning	Single neurons, electric compartmental models, LTP, LTD	Biophysics, neurophysiology
Moods, habits, addictions	Molecular and synaptic level, biophysics of ion channels, membrane properties	Neurochemistry, Genetics, Biophysics, Psychopharmacology
?	Quantum level, small molecules	Neurochemistry, Biophysics

Table 1. Levels of modeling, from neurochemistry to psychology of thinking.

1. Quantum level.

Processes at the genetic and molecular level have direct influence on the states of the brain and contents of mind. Ira Black may be right [26]: information processing is ultimately done on the molecular level, unique brain states are due to the real, physical concentration of different neurochemicals. Understanding the quality of mental experience, called by philosophers of mind “the qualia problem” [27], including consciousness itself, may depend directly on the molecular level. Consciousness may in fact be a special category refering to real, physical states of the brain endowed with structures of sufficient complexity and appropriate control and feedback architecture. In such a case attempts to understand the mind fully in terms of bioelectrical information processing, or even more simplified models, would ultimately fail. Without denying the importance of molecular processes I will present below several levels of description and various approximations to the information processing. In the worst case we should be able at least to simulate the behavior, if not the “real thing”.

Quantum mechanics has been very successful in description of normal matter giving detailed description of interactions of atoms and molecules. Some authors, such as Penrose [28], Stapp [29] or Eccles [30] argue that without quantum mechanics we cannot understand the unity of human experience. This line of reasoning has not been fruitful so far and it seems to be fundamentally wrong, trying to bridge many levels of approximation at once. Is quantum mechanics necessary to understand interactions of neurochemicals with membranes or can we understand the behavior of neurons using classical physics? It is true that in special, well understood conditions a single quantum event, like a photon falling on the retina, may become amplified, and may even influence global dynamics of the brain. On the other hand effects requiring quantum description are observable either in interactions of a few small molecules or in very low temperatures. Properties of large biomolecules are frequently investigated using molecular dynamics, a classical physics theory based on Newton's equations and electrostatic interaction potentials. Details of transition from quantum to classical world are still discussed by physicists but it is highly unlikely that this discussion has any relevance to cognitive science.

2. Molecular and synaptic level.

At molecular level [26,6] genetics and molecular biology provide information for neurochemistry. Psychopharmacology investigates (in a purely phenomenological way) direct influence of changes in neurochemistry on the working of mind, as well as the indirect influence on other neurochemicals, in particular on neurotransmitters. Influence of neurochemicals on the dynamics of ionic channels is most important for understanding the bioelectric properties of neurons. Current computer simulations of these processes give results comparable in many details with the results of neurophysiological measurements [31]. Theory is well prepared to accommodate new experimental findings, such as the role of new neurotransmitters or neuromodulators in growth and development processes. Diffusive neurotransmitters, such as nitric oxide (NO), acting in a less specific way than classical neurotransmitters, seem to contribute to formation of larger cortical structures such as topographic maps [32].

Processes at the synaptic level are crucial to the overall functioning of the neural systems. This is evident from the efficiency of the new generation of drugs that regulate the levels of such neurotransmitters as serotonin or dopamine. Unfortunately higher-level theories rarely take details of synaptic properties resulting from physics of ionic channels into account. Models approximating the flow of ionic currents are at the interface between bioelectric and molecular level. Dendritic spines, where most synapses are formed, probably help to isolate individual contributions of synapses to overall depolarization of membrane, perhaps allowing for realization of logical functions [13] or instantaneous learning [12], although the precise function of dendritic spines is not yet clear. Currents generated by several ionic channels are responsible for the dynamics of different kinds of synapses. Mechanisms of fast changes, leading to action potentials, seem to be better understood than those responsible for slow, permanent changes, although the mechanisms of Long Term Potentiation (LTP) and Long Term Depression (LTD) and their role in memory processes [6] are slowly uncovered (cf. recent articles in *Nature* and neurobiological journals). Permanent synaptic changes cannot be decoupled from the molecular and genetic level while understanding of action potentials is possible at the bioelectrical level. Neuromodulation acting at some distance from the origin of release of such substances as serotonin or acetylocholine creates additional complications. The fast messengers such as nitric oxide (NO) diffusing quickly through extra-synaptic space [32] allow for volume learning, i.e. synaptic changes in a diffusion defined region, and may play an important role in the organization of topographic maps.

3. Single neuron level.

Spatial and temporal integration of charges over the neural membrane is well described by the Hodgkin-Huxley equations. Quite detailed simulations of model cerebellar Purkinje cells, using 4550 dendritic compartments and 8021 ionic channels have already been published [31]. Simulations reproduce experimental results with sufficient accuracy to give us confidence that the bioelectrical description of single neuron is essentially correct. Simulation of the influence of psychoactive chemicals on postsynaptic potentials may be directly compared with experiments, for example barbiturates increase inhibitory postsynaptic potential (IPSP) time constants, bicuculline makes it smaller and diazepam makes it bigger. Such facts are very useful when comparison between experiments and simulations of populations of neurons is made.

Anderson and van Essen [59] argue that since the neurobiological systems deal with analog inputs and outputs, theory of such systems should be based on analog quantities as well. They suggest that “the firing rates of cortical neurons encode an irregular, but complete, sampling of PDF’s of multidimensional space”, where PDF is an abbreviation for Probability Density Functions. In particular they analyze multimodal topographical maps of the superior colliculus in terms of PDFs, maps integrating sensory and motoric components (saccadic eye movements) and discuss computation of PDFs by visual cortex. Indeed the idea that Turing models are not the best foundation for biological computing has been discussed for some time (cf. Siegelmann [60,61]). When real coefficients are allowed in neural network model super-Turing capabilities may appear and if exponential time of computation is allowed they have unbounded power.

Several ways of analysis of neuron responses to stimuli are used. Population coding is a well known mechanism described later in this section, but perhaps better and less known method is based on Bayesian analysis [63]. To compute the posterior probability $P(s|r) = P(stimulus|response)$ for responses of many neurons $r = (r_1, r_2, ...r_N)$, assuming that the variability of responses is statistically independent and that estimation of $P(r_i|s)$ has been computed directly from multi-electrode measurements, Bayes law is used:

$$P(s|r) = P(s|r_1, r_2, ...r_N) = \frac{P(s)\prod_{i=1}^{N} P(r_i|s)}{\sum_{s'} P(s')\prod_{i=1}^{N} P(r_i|s')} \tag{1.1}$$

where a uniform prior is usually assumed for $P(s)$. In experiments with the estimation of visual cortex neuron responses this method showed much faster convergence to correct stimuli than the population vector method [63]. Direct possibility to relate metric properties of trains of spikes to stimuli has been considered by several authors (cf. [62]). It seems that both the absolute time and the intervals between the spikes may carry information allowing for discrimination of different clusters of stimuli. Analysis of the temporal structure of spike trains may be based on metric spaces (spaces defined by a set of points with a metric function). The distance $d(A, B)$ is defined by the lower bound for the number of steps needed to convert spike train A into the spike train B, including insertion and deletion of single spikes and small-step shifts.

4. Neural assemblies.

The next step involves simulation of collective behavior of groups of neurons. To achieve this, simplifications of the Hodgkin-Huxley neuron models are necessary. Using simple models of biologically realistic neurons it is possible to simulate on a supercomputer behavior of thousands of interacting neurons. Results may be compared with single neuron measurements in brain slices. Detailed comparison of simulations with experiments at this level of complexity is unfortunately difficult due to the enormous amount of experimental details, such as synaptic connections and geometry of each cell. Comparison of general trends observed in live slices of neural tissue or in some brain structures (such as rat hippocampus) with computer simulations are quite successful. Although single neurons may send burst of hundreds of spikes per second groups of inhibitory neurons (such as interneurons) produce synchronized 40 Hz gamma rhythms [35] driven by fast (glutamate) receptors. Influence of certain drugs on inhibitory postsynaptic potentials (IPSP) is modeled by changing the time constants in model neurons. Simulations of influence of IPSP on the gamma rhythms agree quite well with experiments [36].

There is ample evidence that neurons use the timings of spikes to encode information [38]. Although the noisy “integrate-and-fire” neurons are the only well-justified approximate models of real neurons a common simplification to avoid the complexities of temporal behavior is based on an assumption that spikes are needed just to build the action potential on the axon hillock. Thus the activity of the neuron is described by a single parameter that should in principle be proportional to the number of spikes per second produced by the neurons. Relation of this potential, obtained by the integration of the synaptic inputs, to the output firing rate is usually assumed to be of sigmoidal type, i.e. once the potential exceeds a threshold value the output slowly grows in semi-linear fashion until saturation is reached (corresponding to the maximum firing rate). In neural network community this scenario is almost never questioned. In fact it is based on the slow potential theory of neuron [33] and all the derivations [18] or experimental measurements of the firing rates [34] make unrealistic assumptions about lack of correlations with simultaneous increase of all inputs. This type of neuron behavior may be observed when a current is transmitted through a large area of neural tissue. However, much weaker but correlated inputs from a few synapses may induce a burst of neural activity, showing that the simple, monotonic neuron transfer functions are a gross oversimplification. It is hard to justify the transition from spiking neurons to weighted threshold neurons. Neurons are able not only to recognize strong activation by adjusting their firing rates, but they also recognize specific combination of inputs that fall into the temporal integration time constant with proper timing.

Firing rate approximation is in contradiction with fast reaction times to visual stimuli. For example Rolls [37] estimates that a single cortical area completes processing of visual information in 20-30 msec. Only a few spikes are emitted in such a short time and certainly this is not sufficient to estimate firing rates. Using a small number of inputs and changing the phases and frequencies of the incoming spike trains quite complex output patterns may be observed (Duch and Ludwiczewski, in preparation). There is no reason to exclude non-monotonic transfer functions, especially that there is evidence that associative memories based on neurons with such functions has larger capacity and are able to escape from local minima [39].

In the most common models of simplified neurons with sigmoidal processing functions the activity is computed as a scalar product $I = \mathbf{W} \cdot \mathbf{X}$ or, for fixed norm of weights and input signals, as $I = I_{max} - d(\mathbf{W}, \mathbf{X})^{\mathbf{2}}$, i.e. activity is a function of distance $d()$ between inputs are weights. In this way one can justify neural models based on localized transfer functions, where weights play the role of prototypes. What does it mean in terms of spiking neuron models? Large activation is obtained when positive weights or excitatory synapses are matched with positive dendritic inputs, i.e. those with impulse frequencies above the average, while negative weights, or inhibitory synapses, are matched with negative inputs, i.e. those with below average frequency.

5. Small networks.

The concept of neural cell assemblies was introduced already in 1949 by Donald Hebb in his seminal book [40]. The cerebral cortex has indeed a very modular structure [65,13,42]. Macrocolumns, distinguishable using neuroanatomical techniques, contain between $10^4 - 10^5$ neurons in a $1 - 2$ mm high column spanning six layers of the neocortex, within the cortical area of a fraction of mm ^{2}. Axons of some NCA neurons spread horizontally on several millimeters enabling mutual excitation of different NCAs. Within a macrocolumn one may distinguish minicolumns, much smaller functional groups of neurons with inhibitory connections. They have a diameter of 30 μm and only 110 neurons, except in the primary visual cortex (V1 area), where they contain about twice as many neurons in orientation columns. These minicolumns behave as oscillators and recurrent excitations of such oscillators leads to entrainment and synchronization of neuronal pulses [43], called “synfire chains” by Abeles [44]. Vertical connections inside these minicolumns are largely excitatory and the density of these connections is of an order of magnitude higher than of the connections with neurons outside of the column. Development of the specific orientation columns in visual system may be quite successfully modeled by the self-organizing networks [45].

Every representational scheme has to face a problem of combinatorial explosion: concepts are combined in an infinite number of ways creating new concepts [64]. Activation of a few basic neural units representing primitive concepts should invoke representation of a more complex concept, but how are these activations remembered? An obvious solution is that “grandmother neurons” reacting to complex concepts should exist. This idea is at least partially wrong. Although integration of sensory processing is greater than it was believed for many years at least in the linguistic realm there are no neurons reacting selectively to complex sentences. It would be impossible to learn the language with such organization of memory. There are many other arguments (cf. [64]) in favor of neural cell assemblies (NCA), group of neurons that strongly activate each other. If the assemblies overlap one neuron in NCA may respond to a number of different stimuli, increasing the number of different patterns that neural network is able to recognize. If there are 1000 representational units and only one is active at a time there are 1000 states but if 5 of these units may be active simultaneously the number of possible states is 10^{85}, astronomically large. Hebb introduced cell assemblies as a bridge between neurophysiological nervous activity and psychological mind events. Connectivity within an assembly should be about an order of magnitude higher than between different assemblies. Several other mechanisms for binding complex information have been suggested [64]. Neurons in NCA activate strongly each other but may also activate neurons in other groups. Manipulating thresholds for intergroup activation one can maintain quasi-periodic activity of several NCAs in a network. Synchronization of the spiking activities of neurons belonging to NCA should create enough activity to excited linked NCAs. Such mutual excitation should be seen as correlation of the spiking activities in different NCAs.

There are still many controversies surrounding NCA, their role, size and structure. From the neuroanatomical point of view cortical minicolumns described by Szentagotai (1975) and Mountcastle [65] are natural candidates for NCAs. These vertically oriented cords of cells contain in almost all neocortical areas about 110 cells, except for the striate cortex of primates, where they contain about 260 neurons. The diameter of such a column is about 30 microns and the number of such minicolumns in the neocortex is about 500 millions. Hundreds of such minicolumns are gathered into *cortical columns* (sometimes called also maxicolumns) of much larger size, with 10-100 thousands of neurons within an area of about 0.2-1 mm ^{2}. The brain contains about 0.5 million of these cortical columns. Their shapes and structure varies in different neocortex areas and the connections between different columns are reciprocal. In addition the thin neocortex has a hierarchical, layered structure, with six principal layers.

Thus the neocortex has highly modular structure, something that must be taken into account in models of the brain functions [66]. Such modular neuroanatomical structure enables functional modules. To minimize the interference between independent tasks each system designed for parallel processing must contain functionally independent modules. Humans are capable of performing several tasks at the same time if these tasks are of a different type – for example there is little interference between talking, visual observation and walking. The type of learning we are capable of is restricted by the brain structures we have. Learning is restricted and guided by the available brain resources that evolved in the evolutionary adaptive process.

Although temporal coding in the neocortex has been considered of primary importance for some time [46] only quite recently in computer simulations and experiments with real neurons Traub *et.al*[47] showed how groups of small columns, composed of inhibitory interneurons connected via excitatory piramidal cells synchronize and communicate over larger distances. Studies of synchronization in thin slices of live brain tissue were done identifying the neuroreceptors (belonging to the class of metabotropic glutamate receptors) responsible for the 40 Hz oscillatory activity of the network. 40 Hz gamma oscillations sustained by inhibitory neurons provide a local clock (τ=25 ms) and spikes produced by excitatory cells appear about 5 ms after the main beats, forming a kind of “doublets” that allow to bind together activity of widely separated groups of neurons. Gamma oscillations seem to provide a temporal structure for synchronization of neuron activities and at least in principle allow to solve the binding problem [47].

Since temporal behavior seems to be so important one should consider the question: is it possible to find a mathematically sound approximations leading from models of “integrate-and-fire” spiking neurons, where information is coded in temporal correlations, to models based on graded response neurons, where information is coded in the patterns of activations. So far only the reverse has been shown [38]: spiking neuron models can compute everything graded response neurons can, and sometimes fewer spiking neurons suffice to do the same work.

Several authors, including Amari [48], Freeman [49], Cowan [77] and more recently Mallot and Giannakopoulos [50] take another approach, stressing not the single neuron activity but the whole populations of neurons. Instead of the activity of single neurons a global parameter, called neuroactivity [49], is used. In fact such continuos models have long history, starting with the book published in 1938 by Rashevsky [5]. Modeling brain structures should take into account lamination of cortical layers, columnar structure, geometry of connections, modulation of densities of different type of cells, formation of topographic maps. Continuos theory of cortical networks aims at comparison with neurophysiological data at the level of EEG, MEG, field potentials and optical recordings. For microscopical theories it may also provide an environment in which individual neurons are embedded, similarly as it is done in chemistry where solvation effects are frequently modeled using continuos media while local interactions are simulated using molecular dynamics.

6. Transcortical neural networks.

Attractor neural networks [18] and other neurodynamical models are useful to understand basic mental phenomena such as recognition and categorization. Approximations and simplifications of such models are necessary to understand higher-order cognition. The low level cognitive processes, realized mostly by various topographical maps, define features of internal representations (some of which are hidden from the external world) [4]. These features represent many types of data: analog sensory signals, linguistic variables, numbers, visual images. Real mind objects are composed primarily of preprocessed sensory data, iconic representations, perception-action multidimensional objects. Mental events are correlated with attractors dynamic of local and transcortical neural cell assemblies (TNCAs) [51].

Humans are very good at visual classification tasks but quite poor at classification of multidimensional numerical data. Proper preprocessing, extraction of features in the sensory signal, is crucial to correct classification. Low level sensory and motoric processing is mostly done using various topographical maps. Sensory information is relayed by subcortical structures (dorsal horn, thalamus) and enters layer 4 of the neocortex, sparsely connected with axons in cortical columns, in one of the brain areas specializing in processing information within selected modality. This processing is typically done by a network of neurons that work as computational maps, i.e. specific features of signals activate localized groups of neurons by increasing their discharge frequencies. Although several competing theories of formation of topographic maps exist [21] simple self-organized mappings seems to be quite satisfactory to explain many neurobiological details [45,94].

Computational maps are created in self-organized unsupervised way in the early stages of brain development. Their rough structure is genetically coded but the final development is due to the interactions with the environment and they retain some plasticity even in mature brains (for example, stimulation of fingers may change the areas devoted to representation of somatosensory information [4]). Brain areas where computational maps are located not only receive but also send back the information to the feeding areas creating recurrent network structures. Thus artificial stimulation of the neocortex areas may create strong hallucinations by actual arousal of the neurons connected directly to sensory cells. Recent comparison of models of development of orientation and ocular dominance columns in the visual cortex [21] showed that self-organized feature maps are able to explain most of their neuroanatomical features. Topographical maps are not restricted to the neocortex, there are well known topographical representations in the old cortex and in some subcortical nuclei [6]. Spatial orientation temporary topographical representation maps (coding absolute direction of sight line) were found recently in monkeys [37].

Another type of computational map, used for example in motoric activity, codes attributes (such as direction of movements) by the distribution of activity of populations of neurons in certain brain area [52], therefore it is called “population coding map”. Population coding seems to be an ubiquitous mechanism of the brain information processing. Initially discovered in the motoric areas of the brain, in recordings from the motor cortex of monkeys performing mental rotation tasks, later it was found in premotor, parietal, temporal, occipital and cerebellar cortex. Population vector $P = \sum f_i u_i$, where u_i is a unitary vector, associated with each cell in the population, oriented in the direction of movement corresponding to the maximum cell activity and f_i is the discharge frequency of neuron, is an implicit representation of the direction of movement and its norm determines when the actual movement will take place. Thus both information and significance of this information is present in the population coding mechanism (dual coding principle, [53]). A maxicolumn measuring 0.5 mm by 0.5 mm and containing 100-1000 minicolumns, or about $10^4 - 10^5$ neurons, may be identified with the population. Arbitrary vector field corresponding to a complex information encoded by a population of neurons is a linear combination of elementary basis fields. Force fields generated by the premotor circuits in a frog's spinal cord are a combination of only four basis fields. Motor behavior is obtained as a superposition of time-dependent basis fields, or pattern generators [54]. Dual population coding of more abstract multidimensional attributes should provide a model for representation of complex information facilitating also the use of this information by other mental processes.

A direct attempt to model sensorimotor integration in geometrical terms has been made by Pellionisz and Llinás [55] and is known as the tensor theory. Sensorimotor functions are described by non-euclidean coordinate transformations in frames of reference intrinsic to biological systems. Tensor theory has been used to analyze population responses of cerebellum Purkinje neurons, computing from multielectrode recordings covariant and contravariant tensors fully characterizing neural geometry inherent in cerebellar coordination, in complete agreement with results of skeletomuscular model [56].

Two fundamental questions arise at this level. First, what are the precise internal features of representation of the sensory data that the brain is using in cognitive tasks? Some of these features are already known and neuroscientists are working hard to discover others. The answer to this question is crucial to understanding of some cognitive phenomena. Prosopagnosia, or the inability to recognize faces [57] evidently must depend on specialized internal representation facilitating this complex recognition task. Visual system has been analyzed in some details and it is known that such attributes as the local form, color and texture are passed to the infero-temporal (IT) cortex which implements the memory of the visual object and is essential in recognition of objects.

Second, how and where is the information from computational maps integrated? For a long time it seemed that processing by computational maps is separated among physically distinct areas that do not communicate much. It seems now that sensory convergence is probably a fundamental characteristic of all animal nervous systems [58]. Sensory information from visual, auditory and somatosensory areas converges in multiple areas in such subcortical structures as superior colliculus. There are strong suggestions [58] that these areas integrate also motor responses, creating "multisensory multimotor map". Crick [93] proposed that clastrum, a small subcortical sheet of neurons, may be involved in integration of visual inputs. Neurons responding selectively to faces were found in amygdala and other hypothalamic structures [6]. Baars [98] focused attention on another subcortical brain structure, the nonspecific thalamus. The nonspecific nuclei of thalamus are densely interconnected and project not only to and from the neocortex but also to the reticular activating system involved in attention. The function of these structures is not quite clear but neuroscientists concluded that they are not involved in more complex brain functions, such as conscious processes. Therefore Newman and Baars [72] look for the integrative brain functions involved in higher cognitive functions in the rhythmic cortical processes.

With a few exceptions neurons responding to multimodal features are not really candidates for "grandmother cells" or cells that get activated when certain specific object or event is recognized. Their existence seem rather to indicate that there are higher order feature detectors in the brain. Internal features that are specific combinations of sounds and shapes may be useful in classification. Complex information has to be bound together in some way and it must happen at the neural network level. Neurons in the infero-temporal cortex in the visual area are sensitive (through computational maps based on population coding) only to basic patterns, such as geometrical shapes or simple natural objects. These patterns, together with information from other brain areas coding different sensory modalities, are activated synchronously and appear as mental events or objects of the mind.

Specific nature of attractor states carrying internal representations of categories and symbols is not known. Straightforward implementation of Hebb's suggestion that reverberations in neural circuits are responsible for working memory leads to problems with stability. A signal should be stable in a loop for $10^{2} - 10^{3}$ cycles and it is hard to sustain a signal in a loop with biological neurons. Single synapses are too weak to excite neurons they connect to so it must be a statistical phenomenon. A model introduced by Zipser [67] avoids these problems. A population of units in a recurrent net is capable of holding a signal. A gating unit controls the access to the recurrent net. If the gating unit is turned on the signal is held, if its off it is released from the network and another signal may be captured. Stability of the global neurodynamics has been considered in details by Amit and Brunel [68], who solved the problem of spontaneous activity and stability of the background dynamics of networks of spiking neurons. Solution of this basic problem requires modular structure of the network, including inhibitory interneurons within NCAs. Learning creates local attractors without destabilizing the background dynamics. Predictions from such models may in some cases be directly compared with neurophysiological experiments on monkeys.

Thus we may assume [51] that the original idea of local reverberations in groups of cortical neurons coding the internal representations of categories, put forth by psychologist Donald Hebb already in 1949, is correct. Local circuits seem to be involved in perception and in memory processes. Analysis of integration of information from the visual receptive fields in terms of modules composed of dense local cortical circuitry [79] allows for explanation of a broad range of experimental data on orientation, direction selectivity and supersaturation. It would be most surprising if the brain mechanisms operating at the perceptual level were not used at higher levels of information processing.

7. Large brain structures.

Mind depends not only on the neural cell assemblies in cortex but also on a large number of specific, subcortical structures that I will not discuss here. Short-term memory (STM) is a very complex phenomenon. Psychologist started to use this concept when a famous law of 7 was discovered by Miller [2]: approximately 7 chunks (plus or minus two) or items may be held in the short-term memory, with the half lifetime of about 7 seconds. There is evidence that also some animals have this type of restriction. Sensory buffers: visual, auditory, vocal, even motoric, work on even shorter time scale. STM seems to be a dynamic phenomena due to stable patterns of reverberatory excitations involving large parts of the neocortex and some subcortical structures. Proposal for a STM (Short-Term Memory) mechanism based on the modulation of 40 Hz oscillations refreshed by the relatively slow neurotransmitters, such as acetylocholine and serotonin, has been put forth quite recently by Lisman and Idiart [69]. Already at the end of 1950s it was found (for a review see [72]) that the theta EEG rhythm (2-8 Hz) is associated with the longitudinal currents flowing between the cell bodies of the pyramidal cells. Rather early neuroscientists proposed that these pervasive wave processes may integrate information in the brain. Although recent discoveries of Traub *et.al*[47] throw some light on the details of this process they cannot be easily used to explain such a high-level phenomena as STM.

Direct local stimulation of the cortex with an electrode may evoke specific hallucinations and memories [81]. Conscious perception or simple reception requires certain resonance between the incoming data and the inner representations. Creation of basic representations and categories is slow, while recognition must be fast. This is true for the brain, where the development of human mind takes many years but recognition processes are very fast. It is also true for many models of artificial neural networks, where training phases require many repetitions before the network learns, but subsequent recognition and classification is very fast. Statistical mechanics of neocortical interactions (SMNI) of Ingberg [22], a mesoscopic theory of neural tissue, averages neural activity at multiple spatial and time-scales and is able to explain the lifetimes and the number of STM memories as quasi-stable dynamical patters in the model brain with typical size and shape.

Associative memory models based on simple recurrent networks, such as Hopfield models, seem also to be useful in studying psychological responses to drugs and understanding of some psychiatric phenomena (cf. the review of computational psychiatry [73] and the book [74]), although such models are still used in a highly metaphoric way since direct comparison with neurophysiological experiments is not possible. So far associative models used in psychology were rather simple but a new breed of such models is forthcoming. Successful models of memory, such as the tracelink model of Murre [80], make good use of this modular structure, postulating that each episodic memory is coded in a number of memory traces that are simultaneously activated and their activity dominates the global dynamics of the brain, reinstating similar neural state as during the actual episode. Koerner *et.al*[78] describe a modular recurrent neural network based on the functional organization of cortical columns in which forward input description is combined with feedback generated hypothesis. The network has been used to model robust object recognition.

Further simplifications of neural models are necessary to relate psychological concepts to brain activity. There is good experimental evidence, coming from the recordings of the single-neuron activity in the infero-temporal cortex of monkeys performing delayed match-to-sample tasks (cf. [18,51]), showing that the activity of a neural cell assembly (presumably a microcolumn within a macrocolumn) has attractor dynamics. Several stable patterns of local reverberations may form, each coding a specific perceptual or cognitive representation. Via axon collaterals of pyramidal cells extending at distances of several millimeters, each NCAs excites other NCAs coding related representations. From the mathematical point of view the structure of local activations is determined by attractors in the dynamics of neural cell assemblies. Such networks should be properly described as a collection of mode-locking spiking neurons. Simple models of competitive networks with spiking neurons have been created to explain such psychological processes as attention (cf. [83]). Realistic simulations of the dynamics of microcolumns, giving results comparable with experiment, should soon be possible, although have not been done yet.

8. Symbols, reasoning and the rational mind.

Architecture of the human mind seen at the highest, cognitive level has been considered only from the symbolic artificial intelligence perspective [1], without relations to neural issues. Simple perception may be explained at a level of short-term global dynamics of the brain. Symbols corresponding to categories or object recognitions should correspond to quasi-stable attractor states of large-scale dynamics. Cognitive processes are based on memory, internal representations and categorization of the data provided by environment. The next step – rational mind – requires understanding of the long-term dynamics of transitions between the attractor states. These transitions, in reasoning processes, seem to be controlled by higher-order attractor dynamics. At the end of this hierarchical approach the most complex features of brains, such as collection of concepts representing self, are formed. Formulation of such dynamical models is at present beyond our capabilities. What is feasible and important is to see how logics and reasoning may be obtained as an approximation to the dynamical systems behavior.

Simple associations or categorizations realized by typical neural network models are not sufficient to explain cognitive competence of humans [85], especially linguistic competence. Fodor [86] and Fodor and Pylyshin [87] has made a valid criticism of simple connectionist approaches to cognition. What is needed and is still poorly understood are neurobiologically plausible mechanisms of going from simple associations to logical rules and to the first order logics. A drastic, although quite natural simplification of the neurodynamics, leads to a discreet finite automata (DFA), such as the hidden Markov models [88]. Such models are usually defined without any relation to neurodynamics, but it should be possible to derive them as an approximation describing transitions between attractors. Finite state models should help in understanding sequential reasoning processes. Goldfarb *et.al*[70] have criticized both symbolic (finite state) and connectionist (vector space) models as inadequate for inductive reasoning.

Elman [71] claims that cognition (and language in particular) are best understood as the behavior of a dynamical system. In his view representations are not abstract symbols but rather regions of the state space, and rules and logics are embedded in the dynamics of the system. In his experiments with language semantic and category of words are learned from the context of a corpus of 10.000 sentences. The network learns to predict the next most probable word in a sequence and a hierarchical clustering of activations of hidden units in a feedforward network shows that internal representations of similar concepts are close to each other (using simple Euclidean metric in the activation space of hidden units). In the dendrogram verbs and nouns are well separated, animate and inamimate objects are a bit closer etc. Internal representations may thus be identified with patterns of activities of neurons in recurrent networks. Grammatical constructions are represented by trajectories in the state space.

1.3 Platonic mind

In the previous section I have briefly discussed some approximations that are commonly done at different levels of neural modelling. Transitions between these levels and attempts to derive higher level approximations from lower-level information are particularly interesting, although rarely studied. The central question is: how to go from neurochemistry and biophysical phenomena to the description of single neurons, to small groups of spiking neurons, to larger groups of simplified neural units, to non-spiking recurrent networks and various other neural models, and finally to the finite state probabilistic or deterministic automata, semantic networks or rule-based systems modeling behavior. If it is possible to describe the behavior without the mind where does the mind comes from and how to approach it? A model describing the stream of mind events – recognition of objects, categorizations, trains of thoughts, intuitive decisions and logical reasoning – is needed to place cognitive science on solid grounds. In this section I will sketch such a model based on a more subtle approximation to neurodynamics than finite state models. This model is called here "the Platonic model" since it treats the space in which mind events take place seriously and it represents concepts as idealized objects in this space. However, in contrast to what Plato believed in, what we experience as content of our minds is just a shadow of neurodynamics, taking place in the brain rather than being a shadow of some ideal world.

The Platonic model, which is a development of my earlier model [23], is based on a few observations. Sometimes it is possible to simplify the large-scale neurodynamics responsible for behavior describing it in low-dimensional spaces. For example Kelso [24] has observed that although the dynamics of movements of fingers is controlled by millions of neurons there are simple phase transitions which are described in a low-dimensional (in fact one-dimensional) subspaces. He bases his analysis on the enslaving principle of synergetics developed by Haken [84], stating that in some circumstances all modes of a complex dynamical system may be controlled (enslaved) by only a few modes. Attractors of such systems lie on a low-dimensional hyperplane in the state space of a huge number of dimensions. Recently Edelman and Intrator [89] proposed in context of perceptual problems that learning should be viewed as extraction of low-dimensional representations. In the Platonic model all mind events, including learning, take place in relatively low dimensional spaces.

Why should the Platonic model make the theory of mind easier or make it better than psychological theories we have today? First, it may be (at least in principle, and sometimes in practice) derived as an approximation to neurodynamics. Second, psychologists are used to the concept of "psychological spaces", known also as feature spaces or conceptual spaces, and discuss some phenomena in such spaces [90]. It is easier to discuss mental phenomena using the language of feature spaces rather than to talk about neural states. Third, psychology lacks the concept of space understood as an arena of mind events. It was only after concepts of space and time were established that physics started to develop. Fourth, such point of view leads to models of neurofuzzy systems and generalization of the memory-based systems [91,92] useful for technical applications and cognitive modeling.

Platonic model of mind is based on the assumption that the objects in the feature spaces correspond directly to the attractors of the large-scale dynamics of the brain. To make a step towards psychology attractor states of neurodynamics should be identified, basins of attractors outlined and transition probabilities between different attractors found. In the olfactory system it was experimentally found [76] that the dynamics is chaotic and reaches an attractor only when external input is given as a cue. The same may be expected for the dynamics of NCAs. Specific external input provides a proper combination of features initiating activation of an object (concept, category) coded by one or a group of neural cell assemblies. From the neurodynamical point of view external input puts the system in a basin of one of the local attractors. Such neural networks map input vectors (cues) into multidimensional fuzzy prototypes that contain output actions.

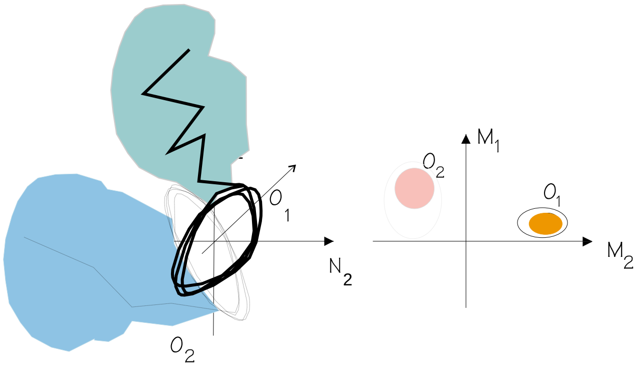

Fig. 1.1. Relation between attractors representing correlation of the spiking activity of a group of neurons (here just two) in the space of $N_i, i = 1..n$ and objects in the feature space $M_i, i = 1..m$, where $n \gg m$.

General description: a coordinate system based on the features of mental representations obtained from lower-level modules, such as topographical maps or computational maps based on population coding, defines a multidimensional space [23], called here “the mind space”, serving as an arena in which mind events take place. In this space a “mind function” is defined, describing fuzzy “mind objects” as regions of space where the mind function has non-zero values. The size and shape of these mind objects should correspond to basins of attractors of the underlying neurodynamical processes. High density in some area of the mind space means that if a specific combination of features that fall in this area is given as an input to the neurodynamical system, the time to reach an attractor will be short. The input variables (stimuli) and the output variables (reactions) together define the mind space while the internal variables of neurodynamics are not explicitly present in the model, although they have influence on the topography of the mind objects. The name “mind space” is replaced by more modest “feature space” for Platonic models using inputs of single modality, such as computational maps.

In simple situations one may try to construct neurodynamical model and the corresponding Platonic model as an approximation to neurodynamics (an example of such approach is given in the next section). To model a real mind corresponding to a real brain one should know all possible attractors of the brain’s dynamics, clearly an impossible tasks. The total number of required dimensions would also be quite large, one could even consider the possibility of continuous number of dimensions [61]. In complex situations Platonic models are constructed phenomenologically. Experimental techniques of cognitive psychology, such as probing the immediate associations and measuring the response time give sufficient information to place basic mind objects corresponding to some concepts or perceptions in the mind space. In the simplified version of the model mind objects are created and positioned using training data and unsupervised as well as supervised methods of learning, similar to the learning vector quantization [94] or other local learning techniques [95].

The model has three time scales. Very slow evolutionary processes impose constraints on the construction of mind spaces and their topography. The type of information mammal brains have at their disposal is fixed by the construction of sensory and motoric computational maps. At the mind space level differences between different type of brains are reflected in the number of dimensions, the character of the axis and the topology of the mind space. Long time scale is associated with learning or changes in the topography of mind spaces, creation of new objects, changing of their mutual relations, forgetting or decay of some objects. Such changes depend on the plasticity of the brain, i.e. on the real physical changes during the learning processes, therefore they are slow. Faster time scale is connected with the actual dynamics of activation of mental representations, trains of thoughts or perceptions, activating one mind object after another. At a given moment combination of features obtained from lower-level processing modules activates only one (or in large systems only a few) attractors in the neurodynamics. This corresponds to a certain “mind state” given usually by a fuzzy point (since the corresponding neurodynamics is always to some degree noisy) in the mind space. Mind objects in the region of the current mind state are “activated” or “recognized”. Evolution of the mind state is equivalent to a series of activations of objects in the mind space – a searchlight beam lighting successive objects is a good metaphor here [93]).

The idea of a “mind space” or “conceptual space” is not more metaphorical than such concept of physics as space-time or a state of the system described by a wavefunction. Mathematical description of such mind spaces should (hopefully) be easier than direct investigation of neurodynamics since low-dimensional spaces are used whenever possible. Psychological interpretation is granted in spaces based on input stimuli, while only very indirect interpretation of neurodynamical states is usually possible. Still the problem is quite difficult because even in rather simple cases feature spaces have more than three dimensions and complicated non-Euclidean metrics. Human conceptual space seems to be better approximated by a number of linked low-dimensional spaces rather than one large space admitting all possible features of internal representations. Local feature spaces model complex feature extraction at the level of topographical maps, providing even more complex features to higher-level “hypothesis” feature spaces that integrate multimodal representations, and at the highest level creating the space in which mind events take place (content of the mind is created). Edelman [96] uses the concept of first and second-order re-entrant maps, and value/category combinations which takes place at the intermediate “hypothesis” level in the Platonic model (Fig 1.2). Values provided by the emotional system may be included in the same way as other features in this model.

Qualitative description of basic concepts: Platonic model provides a useful language for description of mind events. Below some definitions and properties are summarized.

- Mind objects are defined as points, sets of points or fuzzy areas representing the probability density of combinations of features (or a subset of features) that fall into a single category. Several prototypes may create a complex shape corresponding to category that cannot be defined by simple enumeration of features [90]. Some objects may continuously change into other objects (think about color or emotions) in which case density will have maxima only in the subspace corresponding to labels (names) of the objects and change continuously in other dimensions.
- Mind objects are represented by a “mind function” with the value $M(X^{(i)})$ proportional to the confidence of assigning the combination of features at the point $X^{(i)}$ to certain object. In particular $M(X)$ may be a sum of all joint probability density functions for all objects contained in the mind space, with X corresponding to all relevant features, including symbolic names treated as one of the dimensions.
- Learning is equivalent to creation of new objects or to changing topographical relations among existing objects of the mind space, i.e. in the longer time frame the mind function changes with time. Topography may also be partially fixed by *a priori* design (long time-scale evolutionary processes) or knowledge in form of natural laws. Recognition or categorization of unknown combination of

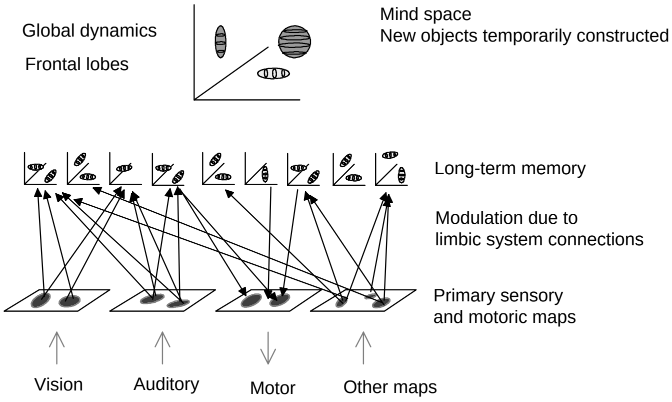

Fig. 1.2. Local feature spaces provide complex features for intermediate level, with competing local "hypothesis spaces" based on long-term memory, providing data for the highest-level "mind space".

features X_k is based on gradient dynamics, i.e. finding a local maximum of the mind function. Sometimes this local maximum is due to the name (symbolic label of the object) only.

- Learning and recognition processes form together the static part of the model which may be treated as a generalization of memory-based (exemplar or case-based) methods used in artificial intelligence. Objects in the feature spaces are equivalent to long-term memory traces and this part of the Platonic model works as an associative memory. Practical realization of the static model is done by neural network based on separable functions estimating joint probability densities of inputs and outputs. Such realization allows for the fuzzy logic interpretation. Static model should be sufficient to explain immediate ("intuitive") reactions in short-time frame experiments. "Intuition" is based on the topography of the mind space and is successful if this topography correctly reflects true relations between input structures.
- A collection of time dependent features $X(t)$ of internal representation is identified with the "mind state" in a given mind space. An object represented by the density $O(X^{(i)})$ is activated when the mind state $X(t) \in O(X^{(i)})$ points at or belongs to it. Simple recognition and learning processes activate only one object at a time but in the hierarchical model Fig. 1.2 each feature space has

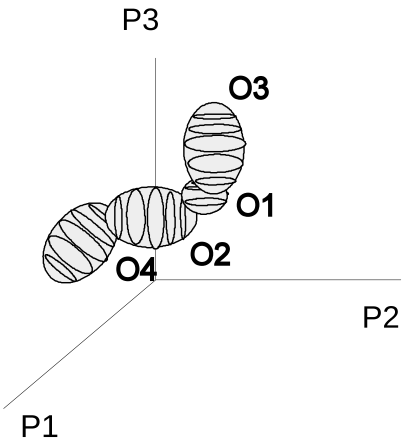

Fig. 1.3. Mind objects corresponding to psychological categories are not defined by simple enumeration of features or sets of exemplars.

separate evolution providing information based on local mind states to spaces higher in the hierarchy.

- Evolution of the mind state, including the dynamics of activation of mind objects, forms the dynamic part of the Platonic model. Mind states have inertia, the tendency to stay inside mind objects, i.e. areas of high $M(X)$. Momentum of the mind state is proportional to the derivative $\dot{X}(t) = \partial X(t)/\partial t$ of the state vector. One can treat $V(X) = -M(X)$ as a potential function and during the evolution of the mind state along the path from one object to another require energy proportional to the difference $\max M(X) - \min M(X)$. This energy is provided either by external inputs or by the internal noise (cf. the role of stochastic resonance in neural systems [101]).
- Transition probability $p(A \rightarrow B)$ from the mind object A to the mind object B is given by the conditional expectancy (cf. Sommerhoff [100]: "The brain's internal representations of the world consist of linked sets of conditional expectancies of the what-leads-to-what kind, which are in the main based on the past experiences". Knowledge is contained in the topography of mind spaces, while reasoning and logic are approximations to mind state dynamics.
- Short-term memory (STM) mechanisms are identified with the highest "mind space" level in the hierarchical Platonic model. Here new objects appear and decay after a short time. Primary objects are composed of preprocessed sensory data and are stored permanently after several repetitions in one or more feature spaces. Secondary objects (for example mathematical concepts) are defined as combinations of more abstract features and appear due to the internal dynamics of the system. Although objects in many local feature spaces may be active at the same time only a few will fit in the STM schemes based on expectations. At this level mind space plays similar role to frames in artificial intelligence. Once

a partial frame is build feature spaces provide more details and the competition among them is won by those that fit in the STM space (cf. [98]).

- Since STM capacity is limited creation of complex object relies on the “chunking” mechanism, which is the basic mechanism of creating higher levels of abstraction [2]. Symbols or linguistic labels are particularly effective (since they are less fuzzy than other features, facilitating faster identification) in identifying complex objects and whole subspaces, therefore chunking mechanism relies primarily on symbols. Inductive learning and problem solving require structured mind objects at the STM level. The dynamics at this level may use Evolving Transformation Systems approach [70], which has quite natural geometrical interpretation, including parametrized distance functions.

Full realization of the general model presented above is rather difficult, but partial realizations may be sufficient to model some cognitive phenomena and is useful in practical applications. Platonic model is an open system, with new subspaces constantly added and deleted and the topography changed in the course of time. All properties of this model result from approximations to neural dynamics, preserving more details than the finite state automata approaches. Low level feature spaces are identified with the primary sensory data processing areas and topographic maps, higher level feature spaces with higher level of processing (for example shape recognition in the inferotemporal (IT) cortex [103]) or transcortical neural cell assemblies coding complex spatio-temporal memory traces such as iconic representations or perception-action multidimensional objects. The forward cortical projections are accompanied by prominent projections back to the original sites creating attractor states involving top-down and bottom-up activations [104] which are modeled here as links between feature spaces at different levels of hierarchy. The state of mind is constantly changing due to the changing sensory stimuli (including the proprioceptive stimuli) and the internal noise in the system. In the absence of external stimuli dreaming, day-dreaming, hallucinations, granularity of experience and other such phenomena should be expected, depending on the level of noise and couplings between the subspaces. At the highest level inductive learning processes give it the capability to solve abstract problems.

Reduction of real microscopic neural dynamics to Platonic model should be possible in a few simple cases, but in general such models may be build using phenomenological data. It is interesting to note that common English language expressions: to put in mind, to have in mind, to keep in mind, to make up one’s mind, be of one mind ... (space) are quite natural in this model. Platonic model allows to discuss mental events in a language close to psychology but using concepts that are justified by neuroscience. As an illustration of use of the language presented here consider the masking phenomenon [27] in which a disk is briefly shown on the screen. The actual mind state is pushed by the external inputs in the direction of the object representing perception of such disk (Fig 1.4), but because of the inertia of mind state it takes about 500 ms to activate this object and send the information from the feature space dealing with object recognition to the higher level mind space. The effect of priming may shorten this time by placing the state of mind closer to the objects that will be activated in near future. Once the object is activated it appears as the content of mind or conscious experience. The duration of the perceptual input has to be sufficiently long to provide sufficient change of momentum (force) of the mind state to reach and activate the object invoking perception of a ring. If it is too short, or if after brief 30 ms exposure another object is briefly shown the mind state will be pushed in direction of another object, without activation of the first one. This is analogous to the scattering phenomena in physics and leads to masking of the first stimuli by the second. The same analysis may be done at the level of neurodynamics and attractors but connections with mental events are not so obvious as in the case of Platonic model.

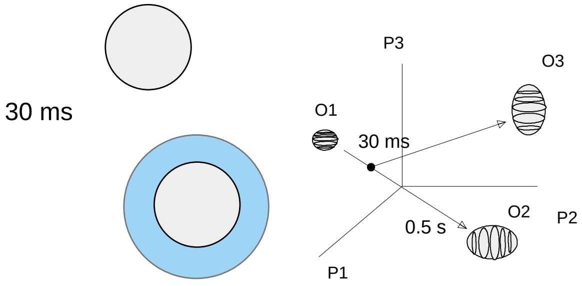

Fig. 1.4. Illustration of the masking phenomenon.

Mathematical realization:

General description of the Platonic model given above admits several mathematical conceptualizations. I will make here some remarks on preferred choices, but at this initial phase of development of the model one should keep many possibilities open. First, feature spaces have axis with meaningful labels, therefore instead of general linear space model dimensional spaces should be used. Function $M(X_i(t))$ describing topography of the feature spaces has natural realization in form of a modular neural network. In particular it may be mixture density network modeling joint probability distributions of the inputs. So far we have used the Feature Space Mapping (FSM) network [91] which is an ontogenic network (with variable number of neurons) based on separable transfer functions. Radial functions may also be used, although products $\sigma(x)\sigma(-x+\theta)$ or combinations $\sigma(x)-\sigma(x-\theta)$ of sigmoids give greater flexibility [105].

Since objects in the feature spaces are modeled by fuzzy areas of non-vanishing $M(X)$ values, and associations among mind objects, corresponding to the transition probabilities between different attractors, should be based on the distance between them, selection of metric is of particular importance. There is no reason why transition probabilities should be symmetric or transitive, and in fact psychological experiments with similarity judgments show that there is no such symmetry. This has been a problem for Shepard [106], who proposed that the likelihood of invoking the same response by two stimuli should be proportional to the proximity of these stimuli in a psychological representation space. There are two simple solutions to this problem. First, the use of local coordinate systems for each, or at least for some, objects in feature spaces. The distance from A to B is thus measured using local metric at A, and the distance from B to A using the local metric at B. From neurodynamical point of view this is reasonable because activation of object A, represented by local neural cell assembly, spreads to other cell assembly representing B, and the time it takes, proportional to the distance in the feature space, should be measured in the local coordinate frame of A. A mathematically well established approach is based on Finsler spaces [102]. Distances are measured here by the action integral and may directly be related to the transition times between different attractors. The idea of Finsler geometry is simple: if time is used as a measure of distance (as is frequently the case on the mountain paths) than the distance between two points connected via a curve $X(t)$ parametrized by t should depend not only on the intermediate positions $X(t+dt)$ but also on the derivative $\dot{X}(t)$.

$$s(A,B) = \int_A^B L(X(t), \dot{X}(t))dt \tag{1.2}$$

where $L()$ is the metric function (Lagrangian in physics). The distance between A and B may than be taken as a minimum of $s(A,B)$ over all curves connecting the two points, which leads to a variational problem (in physics this integral is called “action” and all fundamental laws of physics may be presented in such variational form). Finsler spaces seem to be sufficient for our purpose, with metric function defined by the mind function:

$$s(A,B) = \int_A^B exp\left[\alpha(\nabla M(X(t))^2 - \beta M(X(t))\right] dt \tag{1.3}$$

where α and β are constants and the gradient is calculated along the $X(t)$ curve. For flat densities (constant $M(X)$) $s(A,B)$ is proportional to Euclidean distance, for linearly increasing $M(X(t))$ it grows exponentially with t and for linearly decreasing $M(X(t))$ it grows slowly like $1 - e^{-Ct}$. Finsler metric may be obtained from psychophysical distance measures between different stimuli, which are derived from similarity matrices applying multidimensional scaling (MDS) in low-dimensional spaces, where the final dimension used should be sufficiently high to minimize stress coefficient. How to find a metric for feature space directly from neurobiology? One may either derive it directly from the comparison of trains of impulses [62], use population vector analysis [52], or use Bayesian analysis (Eq. 1.1). In the last case the matrix $P(s|r)$ may be subjected to MDS analysis in the same way as the matrix of psychophysical responses. In phenomenological approach it is the metric that determines the actual $M(X)$ function values, since transition probabilities should be proportional to distances rather than values of the mind function itself. From neurobiological point of view strong association between two objects is related to large transition probability between corresponding attractors, and this in turn should mean that the two representations use partially overlapping sets of neural cells [18].

For all stimuli (input states) $X \in O(A_1)$ belonging to the basin of attractor A_1 averaged responses Y of neurodynamical system are obtained (for example, using Bayesian analysis). Although the basin of attractor $O(A_1)$ is really defined in the very high-dimensional space of neuron activities it is represented here by the value (density) of $M(X,Y)$ function in the low-dimensional (X,Y) space. One way to do it is to measure the transients (the time that is needed to reach an attractor) from a given initial value of X and average them over different internal states. All points belonging to, or very close to, an attractor identified with the response Y are represented by an area in which $M(X,Y)$ has large constant value. The distance (in Finsler metric sense) between the point (X,Y) and the area representing attractor should be equal to the value of the transient. For simple dynamical systems we have defined the $M(X,Y)$ values in such a way that the number of steps in a simple gradient dynamics searching for maximum of $M(X,Y)$ by moving a fixed step size in the direction of largest increase (gradient) is equal to the number of iteration steps needed to reach an attractor. The gradient dynamics stops at the flat area of $M(X,Y)$ corresponding to the points that are sufficiently close to an attractor. In this case different objects in the feature space, corresponding to different attractors, are infinitely far from each other. The decision borders between different basins of attractors in the feature space are not so sharply defined as in the space of neuron activations. This is due to the averaging over internal neural variables and leaving only the input/output variables in the feature space. Adding some noise to the dynamics reduces distances since transition probabilities to go from one object to another increases.

How to find an approximation to neural dynamics and replace it by simpler dynamics in feature spaces? Symbolic approach to dynamics, although a drastic simplification, gives very interesting results even for chaotical systems [99]. Cell mapping method of Hsu [107] is also useful in characterization of dynamical systems. Approximations to neurodynamics introduced here may use simplified trajectories provided by cell mapping or symbolic dynamics. As a final simplification step that completely abandons the original dynamical approach transition probabilities between attractors or between objects in feature spaces (in presence of noise) should be computed, leading to the probabilistic finite state automata. Once transition probabilities $p(A \rightarrow B)$ are computed stochastic dynamics between finite states may be reintroduced. It is interesting to contemplate a quantum-like formalism in which such probabilities are computed as an integral between two objects in the feature space using a suitable operator representing momentum, i.e. $p(A \rightarrow B) =< M(A)|\hat{P}|M(B) >$.

Is this sufficient? It is useful to discriminate between the static and the dynamic cognitive functions. Static functions are related to the knowledge that is readily available, intuitive, memory-based, used in recognition and immediate evaluation. Dynamic functions of mind are used in reasoning and problem solving. The mind space approach is sufficient to describe the static aspects of human cognition. How well can the dynamical aspects of human thinking and problem solving be modeled using such systems? Many problem solving tasks, such as playing chess, seem to be based on a large memory (large number of mind objects) and on a memory-based reasoning [97] with a rather limited exploration of the search space. Memory-based reasoning is related to probabilistic neural networks and outperforms in many cases other learning methods, including neural networks [97]. It is not yet clear how much human thinking is dominated by learned skills; transfer of general thinking skills seems to be an illusion and some experts even ask if humans are rational at all [85]. In any case, adding hidden dimensions (corresponding to internal features that influence the dynamics but are not accessible through inputs or outputs of the system) allows to model arbitrary transition probabilities (associations of mind objects).

Related work and possible extensions Several ideas in the literature seem to point in the same direction as the Platonic model. As already mentioned Anderson and van Essen [59] regarded the firing rates as sampling of probability density functions, and objects in feature spaces are essentially such densities. In the perception-related similarity experiments discussed by Shepard [108] psychological representation spaces were always low-dimensional. In rotation-invariant object recognition experiments Edelman, Intrator and Poggio [109] obtained quite good results using nearest-neighbor similarity based computations, while a small number of principal component basis functions was sufficient for high accuracy of shape recognition. In vision problems although the measurement space (number of receptive fields) is huge internal representation of shapes seems to be low dimensional because many real world tasks have intrinsic low dimensional representations [89]. Processing of spatial information by the sensory and motoric systems is basically three-dimensional, although the number of internal dimensions (degrees of freedom) is very large – many muscles work synchroniously in simplest tasks, therefore it should be possible to describe neural activity controlling the movements using low-dimensional hypersurfaces in the huge state space. In philosophy the idea of conceptual spaces is pursued by Peter Gärdenfors and his collaborators in Lund [6] [112].

In linguistics the idea of mental spaces, called also conceptual spaces or feature spaces, is well established may be dated back to William James. The problem of finding meaning of words (semantics) and the related problem of word sense disambiguation requires taking context into account. A simple simulations (W. Duch and A. Naud, unpublished) shows that representation of words as objects in feature spaces allows to explain human similarity judgments. In an experiment by Ripps *et.al*[113] subjects rated relations between pairs of words that were names of 13 birds and 12 mammals. These authors proposed a feature comparison model of concept representation which would require rather complex neural processes acting on representations. We have used 30 verbal descriptions of the same animals creating their prototype representations in feature space and comparing the multidimensional scaling maps obtained from experimental similarity data and from computed distances in the feature space. The two maps show high degree of similarity, including higher order category formation (for example goose, duck and chicken are close together indicating a household bird category) Similarity ratings between

[6] see http://lucs.fil.lu.se/Projects/Conceptual.Spaces/

different categories of animals may therefore be explained on the basis of their distances increasing our confidence in models of categories based on feature space objects. Similar word-sense maps were obtained by Ritter and Kohonen [114] using the Self-Organizing Map, although no comparison with real psychological data was made.

The cost function that we use to measure topographical distortion of the low-dimensional representation of high-dimensional input data:

$$0 \leq D_n(r) = \sum_{i>j}^{N} (R_{ij} - r_{ij})^2 \Bigg/ \left(\sum_{i>j}^{N} r_{ij}^2 + \sum_{i>j}^{N} R_{ij}^2 \right) \leq 1 \tag{1.4}$$

is different than used in MDS (cf. [108]). Here r are distances in the target space (minimization parameters) and R are distances in the input space. Although we do not have good quantitative measure of the absolute amount of information stored in a particular pattern of memory traces this measure estimates the loss of information when low-dimensional representations are used. Another way of representing the proximity of concepts is by using the popular "mind maps" [118], or graphs connecting related concepts. Such concepts are represented in the mind space by objects overlapping in several dimensions.

The problem of deep semantics across heterogeneous sources has become especially acute in the Digital Library Initiative[7], a large-scale project aimed at building digital repositories of scientific and engineering literature. Searching for information in related fields the meaning of keywords is captured using concept spaces, based on co-ocurrence analysis of many terms. This approach is in fact a very simple version of the Platonic model in which objects are represented by single points (vectors) in Euclidean spaces, but for this particular application it is quite effective. In 1995 two symposia on semantic spaces were organized: "Developing Cognitive and Neural models of High-dimensional Semantic Space", in Montreal, Quebeck, and "Using High-dimensional Semantic Spaces Derived from Large Text Corpora", during Cognitive Science Conference in Pittsburg[8]. Semantic spaces are used for modeling of lexical ambiguity, typicality effects, synonym judgments, lexical/semantic access, priming effects, semantic constrained in parsing and many other linguistic phenomena. HAL, or the Hyperspace Analog to Language, is a model of semantics based on analysis of very large text corpus [110]. Conceptual spaces are used to solve the co-reference problems [111].

Platonic model may also include a subspace for emotions. In a paper by Yanaru *et.al*[115] an "engineering approach" to emotions has been presented. 8 primary emotions (joy, anger, expectation, hate, sadness, fear, surprise, acceptance) are used to define 68 linear combinations corresponding to "mixed emotions". The goal of this approach is to be able to predict the change of emotional state when prompted by external inputs (words, sentences, advertisements). However, linear approximation for mixture of emotions used in their approach is not justified (for example,

[7] See http://www.grainger.uiuc.edu/dli/

[8] see http://locutus.ucr.edu/hds.html

rage may quickly replace fear) and the method of assigning the weights to pure emotions was quite subjective. Despite these shortcomings the task of representing the emotional space and analysis of emotion dynamics by following the trajectories in such space may at least partially be feasible. Moreover, it seems possible to connect results of psychological experiments with EEG measurements allowing to differentiate between emotional states of subjects [116]. Two approaches for determination of the topography of emotional space from psychological observations are proposed here.

The first approach requires a series of stories, or maybe cartoons or videos, showing some typical situations in which people may feel a particular emotion. The subject is asked to identify him/herself with one of the characters of the story and express different degrees of pure emotions at the end; for example, telling a short story that describes a serene nature place, relaxing there, contemplating nature, lazy thought coming through our head, we may try to create an atmosphere of calmness and than ask people to express the degree of their pure feelings in such a situation on a scale: very much, much, some, little, not at all. After presenting many such stories, some of them trying to recreate in different wordings the same emotions one could analyze the resulting vectors to see if a hypothesis that they form different clusters is justified, or to create a fuzzy representation of mixed emotions. Such a representation is created by taking results collected for a single mixed emotion, taking each of the eight pure emotion component for such mixed emotion and fitting one-dimensional Gaussian function to it, recovering both the mean value and the dispersion of the pure emotion content. The total representation of mixed emotion will then be given by 8-dimensional Gaussian function, with the highest density around the most probable values (means) of the pure emotion components. After renormalization of such a function it could serve as a probabilistic description allowing to answer question like: given an observation (a vector of pure emotions), to what degree can it be classified as mixed emotion X? The total emotional space containing all 68 Gaussian representation of mixed emotions will then represent the posterior probability $P(X|V)$, i.e. given observation V, what is the probability that it is an emotion X ?

The second approach is quite different and relies on the similarities of emotions. Each subject is given a list of words designating mixed emotions and after some time of preparation to get familiar with all these descriptive words a question is asked: what emotions are most similar to X? Please list 10, placing the most similar at the top of your list. The same question is asked about all emotions. At the end of experiment similarity matrices $S(X|Y)$ are obtained, showing how similar are two emotions X and Y. This data is then used to create a list of clusters in the 8-dimensional space using multidimensional scaling, i.e. trying to preserve the relative distances between all concepts. A proper rotation and rescaling may be necessary to align the axis of such representation with the pure emotion axis coordinate system. In effect emotional space is obtained, with centers of clusters defined by the similarities. It may be converted to a probabilistic space by placing multidimensional Gaussians at these centers and defining the dispersions using the nearest neighbor criteria. It should be very interesting to compare results coming from these two approaches - how similar will be the final spaces ?

Van Loocke presented a connectionist approach to the dynamics of concepts [117]. It would be interesting to repeat some of his results using the Platonic mind model. All applications mentioned above are only special cases of the static version of the Platonic model using symmetric metric functions. Below I will show one particular realization of the static version of the model and then present an application of this model to psychology.

1.4 Feature Space Mapping network.

A natural practical realization of the static version of the Platonic model is obtained by a modular neural network, with nodes specializing in description of groups of objects in the mind space (coded in different attractor states of the same transcortical neural cell assembly). The function of each node of the network is an approximation to the activity of an attractor neural network, or a fragment of the neocortex that responds to stimulations by stable reverberations of persistent spiking activity. Nodes of such network do not represent single neurons but rather averaged activity of neural cell assemblies. Such network may be considered from several points of view: as a neural network based on estimation of joint probability densities, as a fuzzy expert system based on representation of knowledge by fuzzy sets or as a generalization of memory-based approaches in which exemplar and prototype theories of categorization [90] find natural extension.

Feature Space Mapping (FSM) network [91,92] has some unique properties, rather different from those of most artificial neural network models. The goal is to create a feature space representation of the incoming data, therefore flexible functions with small number of adaptive parameters should be used as node transfer functions. As a first step rough description of the topography of the mind space is created using clustering techniques based on dendrograms or decision trees [121]. To avoid large matrices of distances in the dendrogram method for large input datasets resolution of the data vectors is decreased using integer arithmetics until the number of distinct data vectors becomes manageable. Small number of important nodes are created in this phase using training data and additional symbolic knowledge. The FSM network receives input vectors X as well as desired outputs or classes Y (if known) and finds local maxima $M(X', Y')$, where the value of the function determines the confidence of classification (Fig. 1.5). In the learning phase new nodes are added and existing nodes are modified to account for the new data. Finally in the decision phase either the most probable mapping is derived from the local maxima of $M(X, Y)$, class and confidences values determined or logical rules (fuzzy or crisp) realized by nodes extracted from the network. There is no division between supervised or unsupervised learning since any partial input is used to improve the mind space topography in the subspace in which it is defined. Completion of partially known inputs or inverse problems are treated in the same way as simple classifications, i.e. by finding the local maxima of $M(X,Y)$ function first in the subspace of known inputs and then in the remaining dimensions.

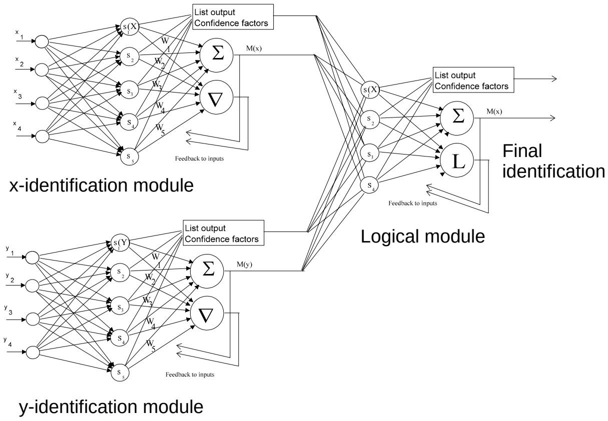

Fig. 1.5. Structure of the FSM network: two recognition and one logical module.

Preliminaries: given a set of training examples $\mathcal{D} = \{X^k, Y^k\}$ create $M(X,Y;P)$ in a network form (P are parameters), giving outputs approximating the landscape of the joint probability density $p(X,Y|\mathcal{D})$. Function $M(X,Y;P)$ should be neither equal to, or proportional to, this density; all that is required is that the maxima of conditional probabilities $Y_p(X) = \max_Y p(Y|X,\mathcal{D})$ agree with the corresponding maxima of $M(X,Y;\theta)$ obtained by calculation of $Y_M(X) = \max_Y M(X,Y;\theta)$. This task is simpler than the full estimation of joint or conditional probabilities.

Processing functions: FSM uses separable processing functions for description of fuzzy data in the mind space (in this section the mind space and the feature space are synonymous). In the special case when gaussian processing functions are used by the network nodes (gaussians are the only radial basis functions that are separable [119]) this model belongs to the family of the growing and shrinking Hyper Basis Function (HBF) networks, such as RAN networks [120]. Separable transfer functions $s(\mathbf{X};P) = \prod_i = 1^N s_i(X_i;P_i)$, where P signifies the set of adaptive parameters, are chosen because: 1) they allow for easy calculation of projections of N-dimensional objects on arbitrary subspaces, thus facilitating learning and recognition in lower-dimensional subspaces; 2) they allow for fuzzy logic interpretation, with $s_i(X_i;P_i)$ being local membership functions (the concept of local membership functions is context-dependent extension of the usual fuzzy logic membership functions); 3) separable functions are more flexible than radial basis functions [105]. In particular in FSM biradial transfer functions are used:

$$SBi((\mathbf{X};\mathbf{t},\mathbf{b},\mathbf{s}) = \prod_{\mathbf{i=1}}^{\mathbf{N}}(\alpha+\sigma(\mathbf{e}^{\mathbf{s_i}}\cdot(\mathbf{X_i}-\mathbf{t_i}+\mathbf{e}^{\mathbf{b_i}})))(\mathbf{1}-\beta\sigma(\mathbf{e}^{\mathbf{s_i}}\cdot(\mathbf{X_i}-\mathbf{t_i}-\mathbf{e}^{\mathbf{b_i}}))) \tag{1.5}$$

This function does not vanish for large $|X|$, for $\alpha = 0$, $\beta = 1$ it becomes localized while for $\alpha = \beta = 0$ each component under the product turns into the usual sigmoidal function. The first sigmoidal factor in the product is growing for increasing input X_i while the second is decreasing, localizing the function around t_i. Shape adaptation of the density provided by this function is possible by shifting centers $\mathbf{t}$, rescaling $\mathbf{b}$ and $\mathbf{s}$. Product form leads to well-localized convex densities of biradial functions. Exponentials e^{s_i} and e^{b_i} are used instead of s_i and b_i parameters to prevent oscillations during the learning procedure.

These functions are more flexible than Gaussian functions in description of multidimensional densities of arbitrary shapes. Each variable X_i defines a new dimension, the data vector $\mathbf{X}$ is a point in the feature space and the input data vector together with the associated uncertainties of the inputs defines a fuzzy region in the feature space, described by the combination of $s(\mathbf{X};P)$ functions. We have compared the convergence of various neural networks on classification and approximation problems and found biradial functions to be superior ([105] and work in progress). Moreover, increasing the slopes of these functions they may be changed into a window like rectangular functions $L(X_i;t_i,t'_i) = \sigma(X_i)(1-\sigma(X_i+B)) \rightarrow \Theta(X_i)(1-\Theta(X_i+B))$ (where Θ is a step function) and thus allow for a smooth change from fuzzy density contours to cuboidal, or from fuzzy logic rules to crisp logic.

Rotation of densities. An important drawback of RBF and other density networks is their inability to use rotated densities in high-dimensional spaces and thus provide simple description of skewed distributions. The $N \times N$ rotation matrix operating on the inputs $\mathbf{RX}$ is very hard to parametrize with $N-1$ independent angles (Euler's angles) and calculate the derivatives necessary for the gradient optimization procedures. In practice covariance matrices in Mahalanobis distance calculations are always diagonal (except in a few dimensional spaces). Using separable functions one can easily cut in N-dimensional space a “slice” of density perpendicular to the direction specified by a vector $\mathbf{K}$:

$$S(\mathbf{X};\mathbf{t},\mathbf{t}',\mathbf{K}) = \mathbf{L}(\mathbf{KX},\mathbf{t},\mathbf{t}')\prod_{\mathbf{i=1}}^{\mathbf{N-1}}\mathbf{L}(\mathbf{x_i},\mathbf{t_i},\mathbf{t'_i}) \tag{1.6}$$

We use rotations in the initialization phase as well as in the learning phase.

Learning: Since training of networks with fixed architecture is NP-hard [123], a robust constructive algorithm is used to build the FSM network. In the constructive algorithm performance is checked on the validation dataset and training is stopped when performance decreases, obtaining complexity of the network that is close to optimal.

Parameter t numbers the training epochs, parameter τ_k is a "local time", growing differently for each network node. One of the problems with RBF networks is their inability to select relevant input features. In FSM feature selection is performed by adding penalty term for small dispersions to the error function:

$$E(P) = E_0(P) + \lambda \sum_{i}^{N} 1/(1+\sigma_i) \tag{1.7}$$

where $E_0(P)$ is the standard quadratic error function and V represents all adaptive parameters, such as positions and dispersions $\sigma_i = |b_i - b'_i|$ of localized units. The sum runs over all active inputs for the node that is the most active upon presentation of a given training vector. The penalty term encourages dispersions to grow and if σ_i includes the whole range of input data the connection to X_i is deleted. Formally we could assign the weights $W_i = 1/(1+\sigma_i)$ to the X_i input connections and minimize the sum of weights. After the training each node has class label and has localized density in relevant dimensions of the input subspace (and constant density in irrelevant dimensions). An alternative approach to modification of the error function is to expand dispersions after several training epochs as much as possible without increasing the classification error. After initialization of the FSM architecture and parameters learning algorithm proceeds as follows:

1. Increase all dispersions until $M(X;P) > 0.5$ for all training vectors X; in this way approximation to the probability density $p(X|\mathcal{D})$ is more smooth and nodes covering outliers are not created.
2. Estimate dispersions σ_{ini} for each class. This is used to create new nodes. Estimated dispersions are slowly decreased during training increasing precision of classification borders.
3. Start of the learning epoch: read new data vector X; find the node N_k that is maximally active; find the node N'_k of the same class as X and closest to X (and $N_k \neq N'_k$).
4. If N'_k and N_k belong to the same class C and the activity of $G(N_k)$ is greater than a given threshold (experience shows that 0.6 is a good value) there is no need for further optimization and the program goes back to step 3; otherwise parameters of the N_k node are optimized:

$$m_k \leftarrow m_k + 1 \tag{1.8}$$

$$W_k \leftarrow W_k + \eta \cdot (C - M(\mathbf{X};\mathbf{D},\sigma)) \cdot \nabla_W M(\mathbf{X};\mathbf{D},\sigma) \tag{1.9}$$

$$D_{kj} \leftarrow D_{kj} + \gamma(t) \cdot (X_j - D_{kj})/m_k \tag{1.10}$$

$$\sigma_{kj} \leftarrow \sigma_{kj} + \alpha(t)(1 - G(\mathbf{X};\mathbf{D_k},\sigma_\mathbf{k}))|X_j - D_{kj}| \tag{1.11}$$

$$\alpha(t) \leftarrow \mathit{\Lambda}(1 + (t-\tau_k)/\epsilon)^{-2} \tag{1.12}$$

$$\gamma(t) \leftarrow \mathit{\Gamma}(1 + (t-\tau_k)/\epsilon)^{-1} \tag{1.13}$$

Here ϵ is about 100 (time in epochs) and is constant, $\mathit{\Lambda}$ and $\mathit{\Gamma}$ are also constants. The "mass" m_k of nodes is set to zero at the beginning of each epoch. Node that is selected frequently gains a large mass and its parameters are changed more slowly.

5. If N'_k and N_k belong to different classes check two conditions: is $||\mathbf{X} - \mathbf{D}(N_k)|| > \sigma(N_k)\ \sqrt{ln(10)}$, i.e. is the new vector $\mathbf{X}$ sufficiently far from the center of the nearest cluster? If yes, check $G(N_k) <$ min$_{act}$, i.e. does the activity of the node exceed certain minimum? min$_{act}$ is set to 0.2 at the beginning of training and after some time, when learning slows down, it is decreased by two. If both conditions are fulfilled create new node;
otherwise go back to 3.
6. Create new node: initial parameters are: $\mathbf{D_c} = \mathbf{X}$; $W_c = C - M(\mathbf{X}; \mathbf{D}, \sigma)$. Initial values of dispersion components σ_{ci} are set to σ_{ini} if $|D_{ci} - D_{ki}| > \sigma_{ki} + \sigma_{ini}$, i.e. if the nearest center is far enough to assume standard initial dispersion; otherwise $\sigma_{ci} = |D_{ci} - D_{ki}|$.
7. The second most excited node is also optimized if it belongs to a different class than the most excited node. Three cases are distinguished: if the vector $\mathbf{X}$ is within the range of this node, i.e. $||\mathbf{X} - \mathbf{D}||) < \sigma$, dispersion of the node is decreased:

$$\sigma_k = \sigma_k - \vartheta \frac{G(\mathbf{X}; \mathbf{D}, \sigma)(|X_k - D_k|) - \sigma_k)}{1 - (t - \tau_k)/\epsilon}; \quad \tau_k = \tau_k + (t - \tau_k)/2 \quad (1.14)$$

If the vector $\mathbf{X}$ is not within the range of the second most excited node, and the node's activity $G(\mathbf{X}; \mathbf{D}, \sigma)$ is greater than some threshold (we use 0.6 for this threshold):

$$\sigma_k = \sigma_k - \vartheta \frac{G(\mathbf{X}; \mathbf{D}, \sigma)\sigma_k}{1 - (t - \tau_k)/\epsilon}; \quad \tau_k = \tau_k + (t - \tau_k)/2 \quad (1.15)$$

Finally if the node is far and is not excited so strongly its dispersion is decreased by:

$$\sigma_k = \sigma_k - G(\mathbf{X}; \mathbf{D}, \sigma)\sigma_k; \quad \tau_k = \tau_k + (t - \tau_k)/2 \quad (1.16)$$

Remarks on the learning procedure:

Increasing the number of nodes leads to 100% classification accuracy on the training set, overfitting the data. The simplest way to avoid it is to assume lower goal for accuracy and check the performance on a test dataset. This requires several stops and checks while the networks adapts itself more and more closely to the data. For small datasets an alternative approach based on maximum certainty may be recommended, since it requires only training data.

After the training epoch is finished the quality Q_k of each node k is estimated by dividing the number of correctly classified vectors through the number of all vectors handled by this node (i.e. vectors for which this nodes was a winners and its activity exceeded certain threshold). Classification results may improve if dispersions are reduced after each epoch by $\sigma_k = \sigma_k - (1 - Q)\sigma_k$. Some units may have changed so much that they do not classify any vectors at all, or their quality Q is close to zero. Such nodes are removed from the network, allowing it to grow more "healthy" nodes in other areas of the input space. Sometimes restricting the number of nodes created in each epoch leads to smaller networks, but it may also lead to slower learning without any gains. Distance $||\mathbf{X} - \mathbf{D}(N_k)||$ does not have to be Euclidean. If two nodes show almost equal activity and one of them belongs to the wrong class it is selected as the winner to allow further adaptation.

Missing values in input are handled using linear search strategy, as described in [91]. To simplify searching in complex, real time classification tasks a hierarchical approach may be used. Dendrograms identify superclusters represented by nodes with large dispersion. Activation of a few such supernodes is checked first and only the nodes with densities contained in the winner node are searched further. Search time may always be reduced to the $\log$ of the number of network nodes.

FSM as fuzzy expert system. Representation of data by fuzzy regions of high density in the mind space make the FSM system equivalent to a fuzzy expert system. The rules of the fuzzy expert systems are of the following type:

$$\begin{aligned} \text{IF } & (x_1 \in X_1 \wedge x_2 \in X_2 \wedge ...x_N \in X_N) \\ & \text{THEN} \, (y_1 \in Y_1 \wedge y_2 \in Y_2 \wedge ...y_M \in Y_N) \end{aligned} \tag{1.17}$$

The rules in fuzzy expert systems are unique, i.e. the same IF part should not have a few different THEN parts. These rules may be directly programmed in the FSM network if many outputs from network nodes are allowed. More general rules: IF$\mathbf{X} \in C_X$THEN$Y \in C_Y$ may also be used in the FSM system. FSM system may contain a number of recognition, or identification, modules and some modules implementing logical operations (Fig. 1.5). This design corresponds to a number of separate mind spaces, each with its own coordinate system based on the unique input features. The results of identifications in these mind spaces are integrated in spaces that are higher in the hierarchy.

A number of logical problems have been solved using FSM density estimation as a heuristic in the search process. For example, representing discrete changes of variables for any additive $A = B + C$, multiplicative $A = B \cdot C$ or inverse additive $A^{-1} = B^{-1} + C^{-1}$ law as $\Delta A = +$ for increase, $\Delta A = -$ for decrease and $\Delta A = 0$ for no change 13 out of 27 facts in 3-dimensional feature space $(\Delta A, \Delta B, \Delta C)$ are true (for example $\Delta A = 0$ if both $\Delta B = \Delta C = 0$). If many such laws are applicable for N variables out of 3^N possible solutions only a few are in agreement with all laws. For example, if $A_1 = f(A_2, A_3); A_2 = f(A_3, A_4)...A_{N-2} = f(A_{N-1}, A_N)$ the number of possible solutions is $4N + 1$, for a large N being a negligible fraction of the 3^N possibilities. FSM can create the mind space corresponding to product of densities generated by all $A = f(B, C)$ laws and use it effectively in a task completion task. An application to the analysis of a simple electric circuit using network that knows Ohm and Kirchoff laws [91] shows how such symbolic knowledge helps to find a solution to a problem that for most neural networks is quite difficult.

In input-completion problems, where only some elements of the input vectors are defined, separable functions allow to find a projection of mind objects on a subspace of the known inputs first – unknown factors are temporarily dropped from the product 1.5. Once a point in this subspace is fixed unknown inputs are found by a one-dimensional line search procedure restricted to these nodes only that gave non–zero projections. For noisy inputs and large networks the complexity of search in high dimensional mind spaces is reduced using the dynamical scaling technique. If gradients of the M-function at point **X** are small, making the nearest mind object hard to find, fuzziness of all mind objects is temporarily increased at the beginning of the search, leaving only a rough representation of mind objects. This corresponds to the initial orientation step in human information processing, determining what the problem is about. After the local maximum of the *M*-function is found the FSM system focuses on the problem by changing the fuzziness of all objects to standard values and performing more detailed search using only a subset of all network nodes. In a typical classification or input completion problems most nodes use localized functions and identification of nodes that contribute to non-zero $M(X)$ value is easy. Several answers to the problem may be found by temporarily switching off the nodes representing mind objects found so far and repeating the search procedure. FSM may also answer questions of the type: find all objects similar to X and evaluate their similarity.

In addition local two-dimensional maps of the mind space objects around the solution found help to visualize the multidimensional relations among mind objects. These maps are obtained by minimization of the measure of topography preservation Eq. 1.4. FSM network has been applied to a number of classification problems, logical rule extraction, task completion problems and logical problems [92,122] with very good results. To summarize, the Feature Space Mapping model is an ontogenic density network realization of the static part of the Platonic model. It allows for a direct modeling of the mind (feature) spaces with crisp or fuzzy facts, using training data or laws constraining possible values of inputs. It enables symbolic interpretation of the objects stored in feature spaces. Generalization is controlled by the degree of fuzziness that may be changed globally or locally, associations are based either on the distance between mind objects (more precisely, between local maxima) or on the overlaps of the densities representing mind objects. Initialization is based on clusterization, learning is done by a combination of supervised and unsupervised techniques, adding more nodes or removing existing nodes of the network if necessary, with changes of network parameters restricted only to those nodes that have influence on the neighborhood of new data. Implementation of typical expert system production rules is straightforward. Reasoning and input completion tasks are solved using one-dimensional line searches, focusing on a single unknown variable each time although gradient techniques are also applicable. Formation of categories and metaconcepts for groups of objects is possible by investigating their projections on various subspaces. Complexity of this network scales at most linearly with the number of training data and parallel implementation is not difficult.

1.5 Categorization in psychology: a challenge

It would be very interesting to analyze a higher order cognitive task at different levels, starting with the brain dynamics and going up to psychology, comparing results at each level and trying to find links between different levels of explanation. Categorization, or creation of concepts, is one of the most important cognitive processes. Although current research on category learning and concept formation frequently ignores constraints coming from required neural plausibility of postulated mechanisms there is enough experimental data from psychology, brain imaging and recordings of neural activity [51] during category learning tasks in monkeys to make the understanding of category learning from several perspectives a great challenge for cognitive sciences in the next few years. In this section a sketch showing how to use Platonic model as an approximation to neurodynamics to explain results of category learning in psychology is presented.

Although the exemplar theory of categorization is usually presented as an alternative to the prototype theory [90] neurodynamics lies at the basis of both theories. Since neural dynamics in biological networks is noisy (due to spontaneous background cortex activity and other sources) several similar exemplars become so fuzzy that a single prototype is formed. To see it clearly a complementary description via feature spaces is helpful. Several models of categorization of perceptual tasks have been compared by Cohen and Massaro [124], including Fuzzy Logical Model of Perception (FLMP), Gaussian Multidimensional Scaling Model (GMM), Theory of Signal Detection (TSD), Feedforward Connectionist Model (FCM) and Interactive Activation and Competition Model (IAC). All these models predict probabilities of responses in a prototypical two and four-response situations in an almost equivalent way.

A classic category learning task experiment has been performed by Shepard *et.al*in 1961 [127] and replicated by Nosofsky *et.al*[125]. Subject were tested on six types of classification problems for which results were determined by logical rules. For example, categories of Type II problems had the XOR structure (i.e. XOR combination of two features determines which category to select) that may be described by the following dynamical system:

$$V(x,y,z) = 3xyz + \frac{1}{2}\left(x^2+y^2+z^2\right)^2 \tag{1.18}$$

$$\begin{aligned}\dot{x} &= -\frac{\partial V}{\partial x} = -3yz - \left(x^2+y^2+z^2\right)x \\ \dot{y} &= -\frac{\partial V}{\partial y} = -3xz - \left(x^2+y^2+z^2\right)y \\ \dot{z} &= -\frac{\partial V}{\partial z} = -3xy - \left(x^2+y^2+z^2\right)z\end{aligned} \tag{1.19}$$

Such equations may be found assuming that

$$\dot{W} = -\frac{\partial V}{\partial x_i} \tag{1.20}$$

$$V(x_1, x_2, x_3) = \sum_{ij} \lambda_{ij} x_i x_j + \sum_{ijk} \lambda_{ijk} x_i x_j x_k + \sum_{ijkl} \lambda_{ijkl} x_i x_j x_k x_l \tag{1.21}$$

and training this 117 parameter system using the XOR training data. We may write such dynamical equations for all category learning tasks. Although the dynamics of the brain during category learning is not so simple we may treat the equations given above as a canonical or prototype dynamics for all tasks where decision is based on the XOR rule. Although we do not know the exact equations governing brain dynamics in category learning these equations may be simplified to this prototype dynamics, with two inputs and one output. In this example, as well as in the remaining five types of classification problems of Shepard *et.al*[125], it is possible to follow the path from neural dynamics to the behavior of experimental subjects during classification task.

The system 1.18 has 5 attractors $(0, 0, 0)$, $(-1, -1, -1)$, $(1, 1, -1)$, $(-1, 1, 1)$, $(1, -1, 1)$; the first attractor is of the saddle point type and defines a separatrix for the basins of the other four. Such dynamical system may be realized by different neural networks. Starting from examples of patterns serving as point attractors it is always possible to construct a formal dynamics and realize it in the form of a set of frequency locking nonlinear oscillators [84].

It is convenient to describe this categorization problem in a feature space. In case of Shepard experiments [127] it contains axis for shape, color and size. Our goal is to illustrate neural dynamics as a process in the feature space. Is it possible to distinguish between categorization based on prototypes and exemplars? In the first case basins of attractors should be large and the corresponding objects in feature spaces should be large and fuzzy. A prototype is not simply a point with average features for a given set of examples, but a complex fuzzy object in the feature space. If categorization is based on exemplars there are point-like attractors corresponding to these exemplars and the feature space objects are crisp. Intermediate cases are also possible, going from set of points representing exemplars, to a fuzzy object containing all the exemplars. Although representation is different both theories may give similar behavioral results if processes acting on these representations are different. Noise in neural system will destroy weak local attractors, changing a set of localized objects representing exemplars to a fuzzy prototype with some internal structure.

Our goal is to show how neural dynamics is connected to processes in the feature spaces. Neural dynamics models physical processes at the level of brain events while feature spaces model mental processes providing precise language to speak about the mind events. Psychological models of categorization should be justified as approximations to real neural dynamics. Attractors activated by specific inputs X_{inp} divide the input space into areas corresponding to basins of different attractors. For example, a cortical microcolumn may learn to solve the A.XOR.B problem establishing attractors presented in Fig. 1.6. In the input space (feature space) the four vertices of the cube represent the shortest transients of the phase space trajectories and the basins of attractors belongs to the neighborhood of these vertices. Introducing the density of feature space objects $M(S)$ proportional to the length of transients of the neural dynamics (the time it takes to reach an attractor from a given initial conditions S_0 neural dynamics defined by the activity of a large number of neurons may be approximated by simple gradient dynamics in the feature space [126]. Equations Eq. 1.18 were solved for large number of points in a cube twice as large as the unit cube and for each initial point the number of iterations in the Runge-Kutta procedure needed for convergence to the point attractor were recorded in a three-dimensional matrix $T(x_i, y_j, z_k)$ (Duch and Kisielewski, in preparation). These values were fitted to several functions $M(x, y, z)$, with best results (accuracy within a few percent) obtained by using hyperbolic tangent basis functions. Original dynamics based on differential equations was than replaced by the gradient dynamics, with most trajectories looking very similar to the original ones. It should be quite interesting to model the development of these attractors (mind objects) during learning.

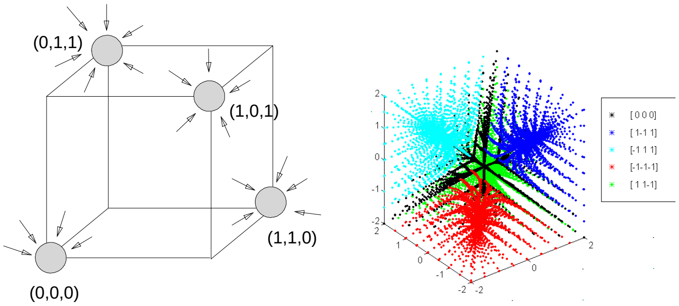

Fig. 1.6. Direct representation of the attractors in the XOR problems. The density of localized function at a given point depends on the time to reach an attractor from such initial conditions. Arrows show local gradients of this density.

Inverse base rate effects. People learn relative frequencies (base rates) of categories and use this knowledge for classification. This is known as the base rate effect. Frequently repeated stimuli create deep basins of attractors (large densities of feature space objects). The size of these basins depends on the inherent noise and variability of the stimuli. Such effects are relatively simple to model. The inverse base rate effect [128] shows that in some cases predictions contrary to the base rates are made. Names of two diseases, C (for Common) and R (for Rare), are presented to participants, the first linked to symptoms I and PC, and the second I and PR. Thus PC and PR are perfect predictors of the disease C and R. Associations (I,PC) → C are presented 3 times more often than (I,PR) → R. After a period of learning participants are asked to predict which disease corresponds to a novel combination of symptoms. For a single symptom I most (about 80%) predict C, in agreement with the base rates. For combination of symptoms PC+I+PR most (60%) choose C, again with agreement with the base rates. However, 60% participants associate the combination PR+PC with the disease R, contrary to the base rate expectations. For many years this effect has eluded explanation until Kruschke and Erickson [129] have introduced a model integrating six psychological principles of human category learning: error-driven association learning, rapid shift of attention, base rate learning, short term memory effects, strategic guessing and representations based on exemplars and their fragments. While strategic guessing in novel situations (assigning novel stimuli to still-to-be-learned categories) is certainly a higher order cognitive process all other principles may be absorbed in construction of representations of categories rather than in processes acting on these representations.

Neurodynamical point of view	Psychological point of view
Learning I+PC more frequent → stronger synaptic connections, larger and deeper basins of attractors.	Learning Symptoms I, PC typical for C because they appear more often.
To avoid attractor around I+PC leading to C deeper, localized attractor around PR is created.	Rare disease R - symptom I is misleading, attention shifted to PR associated with R.
Probing Activation by I leads to C because longer training on I+PC creates larger common basin than I+PR.	Probing I → C in agreement with base rates, more frequent stimuli I+PC are recalled more often.
Activation by I+PC+PR leads more frequently to C because I+PC puts the system in the middle of C basin.	I+PC+PR → C because all symptoms are present and C is more frequent (base rates again).
Activation by PR and PC leads more frequently to R because the basin of attractor for R is deeper.	PC+PR → R because R is distinct symptom, although PC is more common.

Table 2. Comparison of neurodynamical and psychological points of view in the inverse base rate problem.

The answers are determined by the sizes of the basins of attractors corresponding to shapes of objects in the feature space. The memory function describing these objects may be fitted to obtain observed probabilities of answers, as is usually done in psychological modeling [124]. Unfortunately they are defined in 4-dimensional space and are therefore hard to visualize. The C basin is larger, extends between I and PC+I vertices, forcing the R basin to be flatter and be closer to the PR+PC vertex than the C basin is, leading to the inverse base rate effect. Processes acting on representations in feature spaces define physics of mental events, with forces reflecting the underlying neural dynamics. In the absence of cues the state vector $S(t)$ moves randomly in the feature space. Base rate effects influence the size of the basins of attractors (size of the feature space objects). Specifying value of a feature that frequently appears in combination with other features gives momentum to the state vector in the direction parallel to the axis of this feature, initiating a search for a value of unspecified features. The neurodynamical point of view reduced to the feature space and the psychological point of view are compared in the Table here.

Using simulations of the inverse base rate effect tasks we can make one novel prediction: weak effects due to order and timing of presentation (PC, PR) and (PR, PC), due to trapping of the mind state by different attractors. Theoretical predictions and psychophysical experiments confirm the idea that object recognition is affected not only by their similarity but also by the order in which images are presented [130]. Such effects should be observable also in categorization experiments.

Psychological models of categorization have been developed in the past 40 years and are already quite sophisticated. To show that these models contain some truth one should try to justify them as approximations to neural dynamics. Therefore it is interesting to note that the FLMP, GMM and TSD categorization models [124] may be derived as static approximations to the dynamic feature space model described here. Linking neural dynamics with psychological models using feature spaces leads to a complementary description of brain processes and mental events. The laws governing these mental events result from approximations to neural dynamics. Modified feature space models should be useful in analysis of data from many psychological experiments. Learning how to link simplest neural dynamics with feature space representations is just one small step, but many more challenges remain. Hopefully this approach may offer not only good fits to the observations, but also interesting interpretation of mental events.

1.6 Summary

Understanding human mind and its relation to information processing by the brain is the greatest challenge that science has ever faced. A fruitful approach to this problem is to identify a hierarchy of models describing the brain processes at different levels of details and to look for approximations allowing to justify, at least in principle, simplified higher level theories as approximations to more detailed lower level models. Some of these models and theories were identified here. Neural dynamics of large groups of neurons is usually approximated by finite state automata leading to description of behavior. Between these two levels of modeling more detailed approximation to neurodynamics leads to identification of attractors with objects defined in feature or conceptual spaces. Internal representations used by the mind are placed in a space endowed with Finsler metrics. These objects are supported by quasi-stable (attractor) neurodynamics and they are described in relatively low dimensional feature spaces. Platonic model of mind, geometric model treating these feature or conceptual spaces as an arena for mental events, has a great potential to bridge the gap between neuroscience and psychology. It may integrate several trends in cognitive science, in particular linguistics, vision (object recognition) research and psychology (categorization, emotions), providing a useful language for analysis of mental phenomena. It may benefit from modern geometry (in particular Finsler spaces), theory of dynamical systems (especially symbolic dynamics), probability and estimation theory, neural networks, pattern recognition and inductive methods of artificial intelligence. It provides also an important inspiration for development of neurofuzzy systems. The static version of the Platonic model has been implemented in the Feature Space Mapping neurofuzzy system and applied in a number of classification, approximation, logical rule extraction and logical reasoning problems. In the last section categorization problems in psychology have been presented as an area of research where models of higher cognition should meet neuroscience and neurodynamical models on the middle ground of Platonic models. Although such models are in their initial phase of development in future they may play a central role in our understanding of the brain information processing capabilities.

Acknowledgments: Support by the Polish Committee for Scientific Research, grant 8T11F 00308, is gratefully acknowledged. It is a pleasure to thank prof. Jerzy Korczak for his hospitality at the Louise Pasteur Université in Strasbourg, France, where this paper was finished. My assistants, Rafał Adamczak and Norbert Jankowski were involved in development and applications of the FSM system.

References

1. Newell A, Simon H. A. (1976) *Computer Science as empirical inquiry: symbols and search*. Communic. of the ACM 19: 113-126;
2. Newell A, *Unified theories of cognition.* (Harvard Univ. Press, Cambridge, MA 1990)
3. Harnad, S. (1990) *The symbol grounding problem.* Physica D 42: 335-346; Harnad, S. (1993) *Problems, problems: the frame problem as a symptom of the symbol grounding problem.* PSYCOLOQUY 4 (34) frame-problem.11; Rakover, S.S. (1993). *Precise of Metapsychology: Missing Links in Behavior, Mind, and Science*. PSYCOLOQUY 4(55) metapsychology.1.rakover.
4. P.S. Churchland, T.J. Sejnowski, *The computational brain* (MIT, Bradford Book 1992)
5. N. Rashevsky, *Mathematical Biophysics* (Dover, NY 1960)
6. M. S. Gazzaniga, ed. *The Cognitive Neurosciences* (MIT, Bradford Book 1995)
7. E. Thelen, L.B. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action* (MIT Press 1994)
8. Primas H (1981) *Chemistry, quantum mechanics and reductionism* (Springer Verlag, Berlin)
9. Anderson JR (1993) *Rules of the Mind* (Lawrence Erlbaum Associates)
10. Anderson JR (1995) *Learning and Memory* (J. Wiley and Sons, NY)
11. Pollock J.L, *Cognitive Carpentery. A Blueprint for how to build a person*. (Bradford Book, 1995)
12. de Callataÿ AM (1992) *Natural and artificial intelligence. Misconceptions about brains and neural networks.* North Holland.
13. Burnod Y, *An Adaptive Neural Network. The Cerebral Cortex,* London: Prentice-Hall, 1990
14. Brooks, Rodney A., Lynn Andrea Stein. Building Brains for Bodies (MIT AI Lab Memo 1439), August 1993.
15. Levine DS (1991) *Introduction to neural and cognitive modeling* (L. Erlbaum, London)

16. Caianiello E.R., *Outline of a theory of thought processes and thinking machines.* Journal of Theor. Biology 2 (1961) 204-235; E.R. Caianiello, *A theory of neural networks.* In: Neural Computing Architectures, ed. I. Aleksander (MIT Press, MA 1989)
17. Murre J., *CALM, Categorization and Learning Modules* (Erlbaum 1992)
18. Amit D.J, Fusi S, Yakovlev V, Paradigmatic working memory (attractor) cell in IT cortex, Neuural Computations 9 (1997) 1101; Amit D. J, Brunel N, Tsodyks M.V, *Correlations of cortical Hebbian reverberations: experiment versus theory*, J. Neuroscience, 14 (1994) 6445; D.J. Amit, *Modeling brain function. The world of attractor neural networks* (Cambridge Univ. Press 1989)
19. Miyashita Y (1990) *Associative representation of the visual long-term memory in neurons of the primate temporal cortex*, in: Iwai E and Mishkin M, eds, *Vision, memory and the temporal lobe* (Elsevier, New York), pp. 75-87
20. Griniasty M., M. Tsodyks, D. Amit (1993) *Conversion of temporal correlations between stimuli to spatial correlations between attractors*. Neural Comput. **5** 1-17
21. Erwin E., K. Obermayer, K. Schulten, *Models of Orientation and Ocular Dominance Columns in the Visual Cortex: A Critical Comparison*. Neural Computation **7** (1995) 425-468
22. Ingber L, *Statistical mechanics of multiple scales of neocortical interactions.* in: Neocortical dynamics and Human EEG Rhythms, ed. Nunez PL (Oxford University Press 1995), p. 628-681; Ingber L, *Generic mesoscopic neural networks based on statistical mechanics of neocortical interactions.* Phys. Rev. A **45** (1992) R2183-2186
23. Duch W, A solution to the fundamental problems of cognitive sciences, UMK - KMK - TR 1/94 report (1994), available from ftp.phys.uni.torun.pl/pub/papers/kmk and from the International Philosophical Preprints Exchange.
24. Kelso J.A.S, *Dynamic Patterns*, Bradford Book, MIT Press 1995
25. Stillings N.A., Feinstein M.H, Garfield J.L, Rissland E.L, Rosenbaum D.A, Wiesler S.E, Baker-Ward L. Cognitive Science: An Introduction. (MIT Press 1987)
26. I. Black, *Information in the Brain A Molecular Perspective*, A Bradford Book 1994.
27. D.C. Dennett, Consciousness explained (Little Brown, Boston 1991)
28. Penrose R, *The Emperor's new mind* (Oxford Univ. Press 1989); *In the Shadow of the Mind* (Oxford Univ. Press 1994)
29. Stapp H.P (1993) *Mind, matter and quantum mechanics* (Springer Verlag, Heidelberg)
30. Eccles J.C. (1994) *How the self controls its brain* (Springer Verlag, Berlin)
31. J. M. Bower, D. Beeman, *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System* (Springer 1994); see also http://www.bbb.caltech.edu/GENESIS
32. Montague P.R, Dayan P, Sejnowski T.J, Volume learning: signaling covariance through neural tissue, in: Eeckman F.H, Bower J.M (Eds.), Computation and neural systems. Kluver 1993, pp. 377-381; Krekelberg B, Taylor J.G, *Nitric Oxide in Cortical Map Formation* Journal of Chemical Neuroanatomy, 10 (1996) 191-196
33. C. Stevens, *Neurophysiology: a Primer*. New York, Wiley 1996
34. J.A. Anderson, *An Introduction to Neural Networks*, A Bradford Book 1995
35. Whittington M.A, Traub R.D, Jefferys J.G.R, Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. Nature 373 (1995) 612-615
36. Traub R.D, Whittington M.A, Colling S.B, Buzsaki G, Jefferys J.G.R, Analysis of gamma rhythms in the rat hippocampus in vitro and in vivo. Journal of Physiology 493 (1996) 471-484
37. Rolls E.T, Brain mechanisms for invariant visual recognition and learning, Behavioral Processes 33 (1994) 113-138

38. Maass W, Fast sigmoidal networks via spiking neurons, Neural Computation 9 (1997) 279-304.

39. Yanai Hiro-Fumi, Amari Shun-ichi, Auto-associative memory with two-stage dynamics of non-monotonic neurons, IEEE Transactions on Neural Networks, vol. 7, pp. 803-815

40. Hebb D, *The Organization of Behavior* (J. Wiley, NY 1949)

41. Szentagothai, J. (1975). *The 'module-concept' in the cerebral cortex architecture.* Brain Research, 95, 475-496.

42. Calvin W.H, *Cortical columns, modules and Hebbian cell assemblies*, in: M. A. Arbib, Editor, *The Handbook of Brain Theory and Neural Networks* (MIT Press 1995), pp. 269-272

43. Singer W, *Synchronization of neuronal responses as a putative binding mechanism*, in: M. A. Arbib, Editor, *The Handbook of Brain Theory and Neural Networks* (MIT Press 1995), pp. 960-964

44. Abeles M, Corticotronics (New York, Cambridge University Press 1991)

45. Sirosh, J., Miikkulainen, R., and Choe, Y., editors, *Lateral Interactions in the Cortex: Structure and Function.* The UTCS Neural Networks Research Group, Austin, TX, 1996. Electronic book, http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96.

46. Engel A.K., P. König, A.K. Kreiter, T.B. Schillen, W. Singer (1992) *Temporal coding in the neocortex: new vistas on integration in the nervous system.* Trends in Neurosc. **15:** 218-226

47. Traub R.D, Whittington M.A, Stanford I.M, Jefferys J.G.R, A mechanism for generation of long-range synchronous fast oscillations in the cortex. Nature 382 (1996) 621-624; Jefferys J.G.R, Traub R.D, Whittington M.A, Neuronal networks for induced "40 Hz" rhythms. Trends in Neurosciences 19 (1996) 202-208

48. S-i. Amari, Field theory of self-organizing neural nets. IEEE Transations on Systems, Man and cybernetics 13 (1983) 741-748

49. W.J. Freeman, *Tutorial in Neurobiology: From Single Neurons to Brain Chaos.* International Journal of Bifurcation and Chaos 2 (1992) 451-482.

50. Mallot H.A, Giannakopoulos F, Population networks: a large scale framework for modeling cortical neural networks. Max-Planck-Institute of biological cybernetics, Technical Report 24 (1996)

51. Amit D.J, The Hebbian paradigm reintegrated: local reverberations as internal representations. Brain and Behavioral Science 18 (1995) 617-657

52. Georgopoulos AP, Taira M, Lukashin A, *Cognitive neurophysiology of the motor cortex*, Science 260 (1993) 47-52

53. Koechlin E, Burnod Y, *Dual population coding in the neocortex: a model of interaction between representation and attention in the visual cortex*. Tech. Report, Inst. des Neurosciences, Paris 1995.

54. Mussa-Ivaldi F.A, From basis functions to basis fields: using vector primitives to capture vector patterns. Biolg. Cybernetics 67 (1992) 479-489

55. Pellionisz A, Llinás R, Tensorial approach to the geometry of brain function: cerebellar coordination via metric tensor. Neuroscience 5 (1980) 1125-1136

56. Pellionisz A, Tomko D.L, Bloedel J.R, Neural geometry revealed by neurocomputer analysis of multi-unit recordings. In: Eeckman F.H, Bower J.M (Eds.), Computation and neural systems. Kluver 1993, pp. 67-71

57. Damasio, A.R, Damasio H, Van Hoesen G.W, *Prosopagnosia: anatomic basis and behavioral mechanisms.* Neurology 32 (1982) 331-341.

58. Stein B. E, Meredith M. A, *The merging of the senses*. (MIT Press, Cambridge, MA 1993)

59. Anderson C, van Essen D, Neurobiological computational systems, in: computational intelligence imitating life, ed. J.M. urada, R.J. Marks, C.J. Robinson, IEEE Press, NY 1994

60. Siegelmann H.T, Computation beyond the Turing limit, Science 268 (1995) 383-396; Siegelmann H.T. The simple dynamics of super Turing theories. Theoretical Computer Science, 168 (1996) 461-472

61. B. MacLennan, *Field computation in the brain*, CS-92-174 (Univ. of Tennessee, Knoxville, TN 37996)

62. Jeden z ostatnich numerw Neural Computations (1997)

63. Földíak P, The 'Ideal homunculus': statistical inferences from neural population responses. In: Eeckman F.H, Bower J.M (Eds.), Computation and neural systems. Kluver 1993, pp. 55-60

64. Palm G. (1990) *Cell assemblies as a guidline for brain research*, Concepts in Neuroscience, **1**: 133-147

65. Mountcastle V.B. (1978) *An organizing principle for cerebral function. The unit module and the distributed system.* In: The mindful brain, eds. Edelman GE and Mountcastle VB, MIT-Press, Cambridge, MA

66. Happel BLM and Murre JMJ (1994) *The Design and Evolution of Modular Neural Network Architectures.* Neural Networks 7: 985-1004.

67. Zipser D (1991) *Reccurent network model of the neural mechanism of short-term active memory.* Neural Computation **3**: 178-192

68. D.J. Amit, N. Brunel, Global spontaneous activity and local structured (learned) delay activity in cortex (preprint, Inst. of Physics, Univ. of Rome, 1995)

69. Lisman J.E. and Idiart M.A.P. *Storage of 7 ± 2 short-term memories in oscillatory subcycles*, Science 267 (1995) 1512-1515

70. Goldfarb L, Abela J, Bhavsar V.C, Kamat V.N, Can a vector space based learning algorithm discover inductive class generalization in symbolic environment? Pattern Recognition Letters 16 (1995) 719-726

71. Elman J.L, Language as a dynamical system, in: R.F. Port, T. van Gelder, Eds, Mind as motion: explorations in the dynamics of cognition (Cambridge, MA, MIT Press 1995), pp. 195-223

72. J. Newman and B.J. Baars, Neural Global Workspace Model, Concepts in Neuroscience 4 (1993) 255-290

73. Ruppin E, Neural modelling of psychiatric disorders, Network 6 (1995) 635-656

74. Ruppin E, Reggia J, Berndt R (Eds.), Neural modeling of brain and cognitive disorders. Singapore, World Scientific 1996

75. Freeman W.J., *Mass Action in the Nervous system* (Academic Press, NY 1975); Freeman W.J, Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. Biolog. Cybernetics 56 (1987) 139-150

76. Skarda C, W.J. Freeman, *How brains make chaos to make sense of the world.* The Behavioral and Brain Sci. **10** (1987) 161-195;

77. Cowan J.D., *A statistical mechanics of nervous activity.* Lectures on Math. in Life Sciences 2 (1970) 1-57, ed. by M. Gerstenhaber (Am. Math. Soc, Providence RI)

78. Koerner E, Tsujino H, Masutani T, A cortical-type modular neural network for hypothetical reasoning, Neural Networsk (in print)

79. Somers D. C, Todorov E.V., Siapas A.G, Sur M, Vector-space integration of local and long-range information in visual cortex. AI memo 1556, November 1995.

80. Murre J, TraceLink: A model of amnesia and consolidation of memory. Hippocampus 6 (1996) 675-684

81. Libet B. (1985) *Unconscious cerebral initiative and the role of conscious will in voluntary action.* The Behavioral and Brain Sciences 8: 529-566

82. Libet B. (1993) *Neurophysiology of Consciousness. Collected papers and new essays* (Birkhuser, Boston, Basel Berlin)

83. Taylor J.G, Alavi F.N, Mathematical analysis of a competitive network for attention. In: J.G. Taylor, ed. Mathematical Approaches to Neural Networks (Elsevier 1993), pp.341-382

84. Haken H, Synergetic Computers and Cognition. Springer 1991

85. A. Garnham and J. Oakhill, *Thinking and reasoning*. (Oxford, Blackwell 1994)

86. Fodor J. *Psychosemantics*. MIT Press, Cambridge, MA 1987)

87. Fodor J, Pylyshin Z, *Critical analysis of connectionism.* Cognition 28 (1988) 3-72

88. Casey M.P, *Computation in Discrete-Time Dynamical Systems* (PhD thesis, UCSD 1995, available in neuroprose).

89. Edelman S, Intrator N, Learning as extraction of low-dimensional representations. In: Medin D, Goldstone R, Schyns P (Eds.), Mechanism of Percpetual Learning (Academic Press, in print)

90. I. Roth, V. Bruce *Perception and Representation*, (Open University Press, 2n ed, 1995)

91. Duch W, Diercksen G.H.F, *Feature Space Mapping as a universal adaptive system*. Computer Physics Communications **87** (1995) 341-371; Duch W, *Floating Gaussian Mapping: a new model of adaptive systems*, Neural Network World 4 (1994) 645-654; Duch W, Adamczak R, Jankowski N, *New developments in the Feature Space Mapping model*, Third Conference on Neural Networks and Their Applications, Kule, October 1997 (in print)

92. Duch W, Adamczak R, Jankowski N, Naud A, *Feature Space Mapping: a neurofuzzy network for system identification*, Engineering Applications of Neural Networks, Helsinki 1995, pp. 221–224

93. Crick F, *The Astonishing hypothesis. The scientific search for the soul.* (Charles Scribner's Sons: New York 1994)

94. T. Kohonen, *An Introduction to Neural Computing.* Neural Networks 1 (1988) 3-16; T. Kohonen, *Self-organization and Associative Memory* (Springer-Verlag 1984, 3rd edition: 1989); T. Kohonen, *Self-organizing Maps* (Springer-Verlag 1995).

95. L. Bottou, V. Vapnik, *Local learning algorithms*, Neural Comput. 4 (1992) 888-901; V. Vapnik, L. Bottou, *Local Algorithms for Pattern Recognition and Dependencies Estimation*, Neural Comput, 1993, v.5, pp. 893-909

96. Edelman G, Bright Air, Brillant Fire. On the matter of mind. (Penguin 1992)

97. D.L. Waltz, *Memory-based reasoning*, in: M. A. Arbib, Editor, *The Handbook of Brain Theory and Neural Networks* (MIT Press 1995), pp. 568-570

98. Baars B.J. (1988) *A Cognitive Theory of Consciousness* (Cambridge University Press, Cambridge, MA)

99. T. Bedford, M. Keane and C. Series, *Ergodic theory, symbolic dynamics and hyperbolic spaces* (Oxford University Press 1991)

100. Sommerhoff, G. (1990) Life, brain and consciousness (North Holland: Amsterdam)

101. Parnas B.R, Stochastic resonance and noise in the neural coding and senosry signals. In: Bower J.M (Ed.), Computation neuroscience. Trends in research 1995. Academic Press 199, pp. 113-118

102. P.L. Antonelli, R.S. Ingarden, M. Matsumoto, The Theory of Sprays and Finsler Spaces with Applications in Physics and Biology (Kluver Academic, Dodrecht 1993)

103. Tanaka K, Inferotemporal cortex and object vision, Ann. Review of Neuroscience 19 (1996) 109-139

104. Ullman S, High-level vision. Object recognition and visual cognition. MIT Press 1996

105. Duch W, Jankowski N, New neural transfer functions. Applied Mathematics and Computer Science (in print, 1997)

106. Shepard R.N, Toward a universal law of generalization for psychological science. Science 237 (1987) 1317-1323

107. Hsu C.S, Global analysis by cell mapping, J. of Bifurcation and Chaos 2 (1994) 727-771

108. Shepard R.N, Multidimensional scaling, tree fitting and clustering. Science 210 (1980) 390-397

109. Edelman S, Intrator N, Poggio T, Complex Cells and Object Recognition (submitted to NIPS'97)

110. Lund K, Hyperspace Analog to Language: a General Model of Semantic Representation. TENNET VI, Sixth Annual Conference in Theoretical and Experimental Neuropsychology, Montreal, Quebec 1995

111. G. Fauconniere, *Mental Spaces* (Cambridge Univ. Press 1994)

112. Gärdenfors P, Holmqvist K, Concept formation in dimensional spaces, Lund University Cognitive Studies Report 26 (1994)

113. Ripps L.J, Shoben E.J, Smith E.E, Semantic distance and the verification of semantic relations. Journal of Verbal Learning and Verbal Behavior 12 (1973) 1-20

114. Ritter H, Kohonen T, Self-organizing semantic maps. Biolog. Cybernetics 61 (1989) 241-254

115. Yanaru T, Hirotja T, Kimura N, An emotion-processing system based on fuzzy inference and its subjective observations. Int. J. Approximate Reasoning 10 (1994) 99-122

116. Musha T, EEG - emotions, Proc. of 3rd confernce on Soft Computing, Iizuka 1996, pp.

117. Van Loocke P, The Dynamics of Concepts. A connectionist model. Lecture Notes in Artificial Intelligence, Vol. 766 (Springer Verlag 1994)

118. Buzan T, (1989) *Use your head*. (BBC Books: London)

119. Poggio T, Girosi F, *Networks for approximation and learning*. Proc. of the IEEE 78 (1990) 1481-1497

120. Platt J, *A resource-allocating network for function interpolation.* Neural Computation 3 (1991) 213-225; Kadirkamanathan V, Niranjan M, *A function estimation approach to sequential learning with neural networks.* Neural Computation 5 (1993) 954-975

121. Duch W, Adamczak R, Jankowski N, *Initialization of adaptive parameters in density networks*, Third Conference on Neural Networks and Their Applications, Kule, Poland (in print)

122. Duch W, Adamczak R, Grbczewski K, *Extraction of crisp logical rules using constrained backpropagation networks.* International Conference on Artificial Neural Networks (ICNN'97), Houston, TX, 9-12.6.1997, pp. 2384-2389

123. C. Bishop, Neural networks for pattern recognition (Clarendon Press, Oxford 1995)

124. Cohen M.M, Massaro D.W, On the similarity of categorization models, In: F.G. Ashby, ed. Multidimensional models of perception and cognition (LEA, Hillsdale, NJ 1992), chapter 15.

125. Nosofsky R.M, Gluck M.A, Palmeri T.J, McKinley S.C, Glauthier P, Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland and Jenkins (1961). Memory and Cognition 22 (1994) 352-369

126. Duch W. (1994) *Towards Artificial Minds*, Proc. of I National Conference on neural networks and applications, Kule, April 1994, pp. 17-28

127. R.N. Shepard, C.I. Hovland and H.M. Jenkins (1961) Learning and memorization of classifications. Psychological Monographs, issue 517

128. Medin D.L, Edelson S.M, Problem structure and the use of base-rate information from experience. Journ. of Exp. Psych: General 117 (1988) 68-85

129. Kruschke J. K, Erickson M.A, Five principles for models of category learning. In: Z. Dienes (ed.), Connectionism and Human Learning (Oxford, England: Oxford University Press 1996)

130. Wallis G, Presentation order affects human object recognition learning, Technical Report, Max-Planck Inst. of Biological Cybernetics, Aug. 1996