Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Working models of working memory

Omri Barak[1] and Misha Tsodyks[2]

Working memory is a system that maintains and manipulates information for several seconds during the planning and execution of many cognitive tasks. Traditionally, it was believed that the neuronal underpinning of working memory is stationary persistent firing of selective neuronal populations. Recent advances introduced new ideas regarding possible mechanisms of working memory, such as short-term synaptic facilitation, precise tuning of recurrent excitation and inhibition, and intrinsic network dynamics. These ideas are motivated by computational considerations and careful analysis of experimental data. Taken together, they may indicate the plethora of different processes underlying working memory in the brain.

Addresses
[1] Faculty of Medicine, Technion – Israel Institute of Technology, 1 Efron St., Haifa 31096, Israel
[2] Department of Neurobiology, Weizmann Institute of Science, Herzl St., Rehovot 76100, Israel

Corresponding authors: Tsodyks, Misha (misha@weizmann.ac.il)

Current Opinion in Neurobiology 2014, **25**:20–24

This review comes from a themed issue on **Theoretical and computational neuroscience**

Edited by **Adrienne Fairhall** and **Haim Sompolinsky**

For a complete overview see the Issue and the Editorial

Available online 4th December 2013

0959-4388/$ – see front matter,

http://dx.doi.org/10.1016/j.conb.2013.10.008

Introduction

Working memory is a crucial component in the execution of many cognitive tasks that require holding and manipulating information for short periods of time (see e.g., [1]). In this review, we will focus on the holding of information for a time period of several seconds. From the mechanistic point of view, working memory differs from long term memory in that no structural changes are hypothesized to be involved – it is a transient phenomenon. Models of working memory are presented with two types of challenges: data-driven and computational-driven (Figure 1, middle). The data-driven challenges arise from the analysis of behavior and neuronal recordings in animals performing working memory tasks. Animals were shown to be able to maintain several items simultaneously in memory, remember their order, and manipulate them (see e.g., [2] for a recent account). Among the common physiological observations, it was reported that neurons typically exhibit irregular firing activity at a low rate, the activity related to storing a fixed item is not stationary, and there is a large heterogeneity in the firing profiles of different neurons [3,4,5•,6]. From the computational side, the network activity representing a memorized item should exhibit a sufficient degree of stability to ensure memory retainment. This requirement is especially challenging for storing continuous variables, such as orientation or spatial position of a visual cue, because of an inevitable drift along the variable's representation. Furthermore, integrating the various data-driven challenges in a self-consistent manner is often a non-trivial computational problem.

To cope with these challenges, various models incorporate different amounts of biophysical detail – highlighting the contribution of model elements to the various challenges (Figure 1, right). In the current review, we will briefly present the classic models of working memory, and proceed to highlight several recent attempts at addressing the different challenges. The focus of this review is on network mechanisms of working memory. For alternative mechanisms based on single cells persistent activity see [7,8].

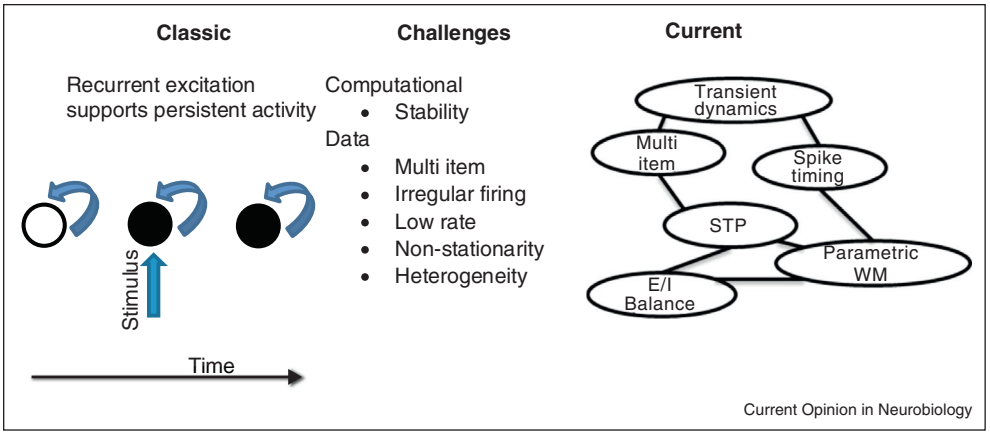
The classic models

The classic view is that items are embedded in long term memory via specific synaptic modifications, and presentation of these items leads to activation of stable activity patterns in the network ('attractors') [9,10]. Thus the information on which item is currently in working memory is stored in the persistent firing of these attractors. Supporting this theory, neurons exhibiting persistent activity after the removal of a stimulus were observed in the inferior temporal and prefrontal cortices of monkeys [11,12] (Figure 1, left).

Multi-item memory

The majority of foundational work on models of working memory were motivated by delayed memory experiments where only one item had to be retained in memory [11,13,14]. Working memory, of course, is not limited to a single item [15], and accordingly electrophysiological recordings were done on monkeys performing tasks requiring the maintenance of several items in working memory [12]. The mechanistic challenge of maintaining more than one item arises due to interference between the activations of the different items. Amit *et al.* [16] proposed that such interference is reduced when items are encoded by sparse patterns – every item is represented by a small fraction of the neuronal population. This approach was extended by [17] to account for the storage of both learned and novel items.

Figure 1



Concepts in working memory models. **Left**: the classic account of working memory is that a strong recurrent excitation enables the network to sustain persistent activity after removal of a transient stimulus. **Middle**: models of working memory face challenges on computational and data-driven fronts. **Right**: current models of working memory introduce various biophysical considerations to cope with the challenges, while attempting to remain simple enough to understand.

Both inhibition [18] and excitation [19] were shown to influence the capacity of multi item working memory. In both of these works, the authors showed how a network storing a continuous value can be dynamically partitioned to maintain several localized bumps of activity, each representing one memorized value of this variable. The balance of excitation and inhibition determines both the number of items that can be held, and their mode of failure (fade out or merge). Continuous attractors required tuned connectivity, but this tuning can be relaxed by incorporating more biological detail into the model. Specifically, Rolls *et al.* [20] showed that synaptic facilitation (detailed in the next section) increases the capacity of working memory. Moving beyond capacity considerations, Dempere-Marco *et al.* [21] showed that salient items (those presented with higher intensity) can be guaranteed a higher chance of maintenance at the expense of less salient items.

A conceptually different method of holding multiple items in working memory is to multiplex them in time rather than in space [22,23]. In this approach, the activated items are all oscillating at some frequency in different phases, and capacity is determined by the ratio of this frequency to the temporal width of each activation. In principle, this method can store information about the order of the items as well as their identity.

Effects of NMDA receptors on persistent activity

Early network models of persistent activity used a highly simplified description of neuronal and synaptic dynamics, resulting in certain difficulties in reproducing a realistic range of firing rates during working memory [24]. As first pointed out by [25], this issue can be resolved by considering networks with slow recurrent excitatory currents that are reminiscent of NMDA currents. Indeed, it was recently reported that blocking NMDA, but not AMPA, receptors during a working memory task abolishes persistent activity in prefrontal neurons [26•]. Moreover, the relative efficacy of NMDA currents is sensitive to Dopamine modulation, thus providing a possible mechanism of regulating working memory [27]. In particular, strengthening NMDA currents during the delay period of memory tasks can enhance the robustness of persistent activity to intervening stimuli. More intriguingly, NMDA currents can also affect the temporal aspects of neuronal activity, for example, by enhancing the burstiness of firing, thus potentially mediating the more complex forms of persistent activity compared to simple steady-state asynchronous states [27].

Short term synaptic plasticity

The model of [20] mentioned above relied on the slow timescale of synaptic facilitation to stabilize the persistent firing state (see also [28]). Synaptic facilitation, and other forms of short term synaptic plasticity, enable synapses to temporarily modify their efficacy in response to stimuli [29,30]. Recently, Itskov *et al.* [31] examined the effect of synaptic facilitation on a network storing a continuous variable via the 'line attractor' mechanism, that is, a continuous one-dimensional set of marginally stable activity states, and showed that facilitation reduces the inherent drift of the system, thus prolonging memory lifetime significantly.

A more dominant role for synaptic facilitation was suggested by Mongillo *et al.* [23], who proposed that a stimulus-selective pattern of synaptic facilitation can itself maintain working memory in the absence of

increased spiking activity. In this scenario, neuronal activity is only required when information is extracted from synaptic into spiking form at the end of the delay period. Thus, synaptic facilitation does not stabilize persistent firing activity, but replaces it. This property of the model is compatible with the analysis of the neuronal recordings from the Romo lab, showing that overall activity in the prefrontal cortex exhibits significant *reduction* over the course of delay period, slowly recovering to the pre-stimulus level towards the presentation of the second stimulus [3]. A recent model utilizing gating neurons instead of synapses has some functional similarity to this idea [32].

Finally, synaptic facilitation does not only bestow the network with slow timescales, but it also provides a nonlinear relation between the presynaptic firing rate and postsynaptic currents [33]. Hansel and Mato [34•] demonstrated that this nonlinearity is crucial for a network to display persistent activity with realistic spiking statistics. Specifically, it is known that neurons fire in a highly irregular manner, and this phenomenon was explained by a fluctuation driven regime where excitation and inhibition balance each other [35]. This balanced state, however, is characterized by a linear input–output transformation of firing rates that precludes the bistability necessary for many working memory models. By incorporating synaptic facilitation into a balanced network, Hansel and Mato [34•] showed that bistability is restored, and their model exhibits realistic spike firing in the persistent state.

Excitatory/inhibitory balance

Besides guaranteeing irregular spiking activity, the interplay of excitation and inhibition can stabilize working memory as demonstrated by several recent models. McDougal (PhD Thesis, Ohio State University, 2011) studied a model of excitatory and inhibitory populations, where an arbitrary subpopulation of the excitatory neurons can maintain elevated firing rates after a transient stimulus. Interestingly, there is no excitatory feedback, but rather these cells activate inhibitory neurons which in turn inhibit the excitatory population. Persistent firing is enabled due to a post inhibitory rebound current (I_h). This current is Calcium dependent, and hence only the previously active excitatory cells have an elevated calcium level, serving as an identifying tag and prolonging their firing. This E–I–E loop creates a gamma frequency signature during memory maintenance. A similar mechanism was demonstrated experimentally in LP neurons of pyloric network of the crab *Cancer borealis* in [36].

Two recent results demonstrate that fast inhibition followed by matched slower excitation can stabilize the memory of a continuous parameter. Boerlin *et al.* [37••] considered the implications of encoding abstract variables using a population of spiking neurons. They assumed that every spike is only emitted when it improves the decoding accuracy of an encoded variable. The resulting activity is highly irregular and yet the overall population can accurately represent the variable (a similar idea was explored in [38], but firing rate rather than precise timing of spikes was used as the information carrier). In order for the network to function, fast recurrent inhibition is needed to notify the entire network every time a neuron spikes, so that the same prediction error will not be corrected twice. The dynamics of the abstract variable itself are managed by slower, excitatory, connections that are matched in strength to the inhibitory ones. Lim and Goldman [39••] proposed a similar mechanism from a different perspective. The authors argued that in order to reduce the drift of a memorized variable, a friction-like term should be added to the dynamics. Thus, they suggested that negative derivative feedback could stabilize the memory. In order to implement this idea in the neural network, they noticed that fast inhibition followed by balanced slower excitation produces a signal that is proportional to the negative temporal derivative of the population activity.

Dynamic mechanisms of memory

The obvious candidate for storing a fixed item in memory is a fixed state of the network – in the simplest case the persistent activity of neurons [40]. A closer look at the data, however, reveals that the activity of typical cells rarely adheres to this concept of persistent activity. The information stored in populations of prefrontal neurons seems to decline and reappear during the delay period [3,4]. The tuning of neurons to stimuli changes from the stimulus to the delay periods [3,5], and in general the activity of neurons is more heterogeneous than predicted by most models [6]. These observations triggered new theoretical ideas regarding the mechanisms subserving working memory. One solution, mentioned above, is to rely on other biophysical processes as the state of the system [23] (McDougal, PhD Thesis, Ohio State University, 2011), but in those cases as well this state is a fixed point (or limit cycle) of the system.

An alternative view is that memory of an item could be maintained by highly non-stationary activity, as illustrated by the framework of reservoir computing [41,42]. In this framework a stimulus impinges upon a randomly connected network, eliciting some trajectory in state space. Recurrent connectivity enables this trajectory to last for substantial time before the network returns to baseline. During this time, the memory can be decoded from the activity of the network.

The plausibility of such a mechanism depends on the temporal capacity of the network – the duration in which the stimulus can be decoded. This capacity has most often been numerically and analytically studied by injecting white noise into the network and checking the amount of information present in the current state of

the network about the past values of the stimulus. Results from considering linear [43,44] and nonlinear networks [45,46•,47•,48], in discrete and also continuous [49] time, have shown that the memory of completely random networks only scales logarithmically with network size while a structured (generally more feedforward) network can have a memory that scales linearly with network size.

Given the acceptable performance level of a random network, and the substantial performance gain in a structured network, a family of working memory models can be obtained by training an initially random network to perform a memory task. Barak *et al.* [50••] used this approach to compare models of varying level of structure to data collected from monkeys performing delayed vibrotactile discrimination. They found that both a random reservoir-type model, and a structured fixed-point model [51] can match the behavior of the monkeys, but that their firing rate profiles are either too consistent or not enough consistent across time, compared to the data. An intermediate model, obtained by training an initially random network, provided a better match to the experimental findings.

A different route was taken by [52], who trained chaotic neural networks to do what they were already doing. Specifically, they chose an arbitrary existing trajectory of the network, and by making it the target of a training algorithm stabilized it. Thus, this seemingly random trajectory became an attracting trajectory and could be harnessed for functional uses such as measuring elapsed time. A similar idea was explored by Szatmáry and Izhikevich [53] with an emphasis on exact spike timing. The authors argued that a random network has, by chance, many short spatiotemporal patterns that are more likely to occur than others. By introducing an associative form of short term plasticity, they showed that these patterns can be stabilized and spontaneously reactivated, supporting working memory.

This line of work hinges upon training networks to perform a certain task without dictating exactly how the network should do it. It is probable that in some cases training will result in fixed point mechanisms of working memory, but other, unexpected, solutions are also possible. Recently, Sussillo and Barak [54•] developed a method to reverse engineer such trained networks, revealing the dynamical structures underlying their operation.

Conclusions

Working memory is vital to our everyday behaviors. At the same time the neuronal processes underlying working memory present intriguing computational problems. Thus it is tempting to find the one particular process responsible for working memory, and we have reviewed several noteworthy attempts of doing so. Biological systems, however, do not have to choose one mechanism. It is highly possible that many of the mechanisms mentioned above are utilized by the brain to sustain working memory, perhaps even affording some degree of robustness to the failure of one particular mechanism.

Acknowledgements

We thank Ron Meir for helpful comments on the manuscript. MT is supported by the Israeli Science Foundation and Foundation Adelis.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Dudai Y: *Memory from A to Z: Keywords, Concepts and Beyond*. New York: Oxford University Press; 2002.
2. Warden MR, Miller EK: **Task-dependent changes in short-term memory in the prefrontal cortex**. *J Neurosci* 2010, **30**:15801-15810.
3. Barak O, Tsodyks M, Romo R: **Neuronal population coding of parametric working memory**. *J Neurosci* 2010, **30**:9424.
4. Rainer G, Miller EK: **Timecourse of object-related neural activity in the primate prefrontal cortex during a short-term memory task**. *Eur J Neurosci* 2002, **15**:1244.
5. • Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J: **Dynamic coding for cognitive control in prefrontal cortex**. *Neuron* 2013, **78**:364-375.

Latest in the series of recent experimental papers emphasizing the highly dynamic nature of working memory representations during delayed memory tasks in monkeys.

6. Jun JK, Miller P, Hernández A, Zainos A, Lemus L, Brody CD, Romo R: **Heterogenous population coding of a short-term memory and decision task**. *J Neurosci Off J Soc Neurosci* 2010, **30**:916-929.
7. Egorov AV, Hamam BN, Fransén E, Hasselmo ME, Alonso AA: **Graded persistent activity in entorhinal cortex neurons**. *Nature* 2002, **420**:173-178.
8. Hasselmo ME, Stern CE: **Mechanisms underlying working memory for novel information**. *Trends Cogn Sci* 2006, **10**:487-493.
9. Hebb DO: *The Organization of Behavior: A Neuropsychological Theory*. Oxford, England: Wiley; 1949.
10. Hopfield JJ: **Neural networks and physical systems with emergent collective computational abilities**. *Proc Natl Acad Sci USA* 1982, **79**:2554.
11. Fuster JM: **Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory**. *J Neurophysiol* 1973, **36**:61-78.
12. Miller EK, Erickson CA, Desimone R: **Neural mechanisms of visual working memory in prefrontal cortex of the macaque**. *J Neurosci* 1996, **16**:5154-5167.
13. Funahashi S, Bruce CJ, Goldman-Rakic PS: **Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex**. *J Neurophysiol* 1989, **61**:331-349.
14. Miyashita Y, Chang HS: **Neuronal correlate of pictorial short-term memory in the primate temporal cortex**. *Nature* 1988, **331**:68-70.
15. Miller GA: **The magical number seven, plus or minus two: some limits on our capacity for processing information**. *Psychol Rev* 1956, **63**:81.
16. Amit DJ, Bernacchia A, Yakovlev V: **Multiple-object working memory—a model for behavioral performance**. *Cereb Cortex* 2003, **13**:435-443.
17. Amit Y, Yakovlev V, Hochstein S: **Modeling behavior in different delay match to sample tasks in one simple network**. *Front Hum Neurosci* 2013, **7**:408.

18. Edin F, Klingberg T, Johansson P, McNab F, Tegnér J, Compte A: **Mechanism for top-down control of working memory capacity**. *Proc Natl Acad Sci USA* 2009, **106**:6802-6807.

19. Wei Z, Wang X-J, Wang D-H: **From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization**. *J Neurosci* 2012, **32**:11228-11240.

20. Rolls ET, Dempere-Marco L, Deco G: **Holding multiple items in short term memory: a neural mechanism**. *PLoS ONE* 2013, **8**:e61078.

21. Dempere-Marco L, Melcher DP, Deco G: **Effective visual working memory capacity: an emergent effect from the neural dynamics in an attractor network**. *PLoS ONE* 2012, **7**:e42719.

22. Lisman JE, Idiart MA: **Storage of 7 ± 2 short-term memories in oscillatory subcycles**. *Science* 1995, **267**:1512-1515.

23. Mongillo G, Barak O, Tsodyks M: **Synaptic theory of working memory**. *Science* 2008, **319**:1543-1546.

24. Amit DJ, Brunel N: **Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex**. *Cereb Cortex* 1997, **7**:237-252.

25. Wang X-J: **Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory**. *J Neurosci* 1999, **19**:9587-9603.

26. • Wang M, Yang Y, Wang C-J, Gamo NJ, Jin LE, Mazer JA, Morrison JH, Wang X-J, Arnsten AF: **NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex**. *Neuron* 2013, **77**:736-749.
First direct experimental support for the importance of NMDA receptors for delay activity in working memory tasks.

27. Durstewitz D: **Implications of synaptic biophysics for recurrent network dynamics and active memory**. *Neural Netw* 2009, **22**:1189-1200.

28. Hempel CM, Hartman KH, Wang X-J, Turrigiano GG, Nelson SB: **Multiple forms of short-term plasticity at excitatory synapses in rat medial prefrontal cortex**. *J Neurophysiol* 2000, **83**:3031-3041.

29. Markram H, Wang Y, Tsodyks M: **Differential signaling via the same axon of neocortical pyramidal neurons**. *Proc Natl Acad Sci USA* 1998, **95**:5323-5328.

30. Zucker RS, Regehr WG: **Short-term synaptic plasticity**. *Annu Rev Physiol* 2002, **64**:355-405.

31. Itskov V, Hansel D, Tsodyks M: **Short-term facilitation may stabilize parametric working memory trace (Internet)**. *Front Comput Neurosci* 2011:5.

32. Conde-Sousa E, Aguiar P: **A working memory model for serial order that stores information in the intrinsic excitability properties of neurons**. *J Comput Neurosci* 2013 http://dx.doi.org/10.1007/s10827-013-0447-457.

33. Barak O, Tsodyks M: **Persistent activity in neural networks with dynamic synapses**. *PLoS Comput Biol* 2007, **3**:e35.

34. • Hansel D, Mato G: **Short-term plasticity explains irregular persistent activity in working memory tasks**. *J Neurosci* 2013, **33**:133-149.
The authors show that the nonlinearity associated with short term synaptic plasticity enables persistent delay activity in balanced neural networks.

35. Van Vreeswijk C, Sompolinsky H: **Chaos in neuronal networks with balanced excitatory and inhibitory activity**. *Science* 1996, **274**:1724-1726.

36. Goaillard J-M, Taylor AL, Pulver SR, Marder E: **Slow and persistent postinhibitory rebound acts as an intrinsic short-term memory mechanism**. *J Neurosci* 2010, **30**:4687-4692.

37. •• Boerlin M, Machens CK, Denève S: **Predictive Coding of Dynamical Variables in Balanced Spiking Networks**. *PLOS Computational Biology* 2013, **9**:e1003258.
The authors design a recurrent network of spiking neurons that can represent in its output the time-integrated external input. Fast inhibition and slow excitation are tightly balanced to each other. In the absence of the input, the network sustains its activity pattern.

38. Eliasmith C, Anderson CCH: *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*. Cambridge, Massachusetts: MIT Press; 2004.

39. •• Lim S, Goldman MS: **Balanced cortical microcircuitry for maintaining information in working memory**. *Nat Neurosci* 2013, **16**:1306-1314.
Another implementation of balanced fast excitation and slow inhibition to stabilize persistent activity, similar to Boerlin *et al.*'s work. The authors arrive at this solution, however, from a control theory perspective, utilizing the concept of negative derivative feedback.

40. Amit DJ, Fusi S, Yakovlev V: **Paradigmatic working memory (attractor) cell in IT cortex**. *Neural Comput* 1997, **9**:1071-1092.

41. Jaeger H: *The 'echo state' approach to analysing and training recurrent neural networks-with an erratum note'. Technical Report GMD Report 148, German National Research Center for Information Technology*. 2001.

42. Maass W, Natschläger T, Markram H:: **Real-time computing without stable states: a new framework for neural computation based on perturbations**. *Neural Comput* 2002, **14**:2531-2560.

43. White OL, Lee DD, Sompolinsky H: **Short-term memory in orthogonal neural networks**. *Phys Rev Lett* 2004, **92**:148102.

44. Ganguli S, Huh D, Sompolinsky H: **Memory traces in dynamical systems**. *Proc Natl Acad Sci USA* 2008, **105**:18970-18975.

45. Lim S, Goldman MS: **Noise tolerance of attractor and feedforward memory models**. *Neural Comput* 2011, **24**:332-390.

46. • Dambre J, Verstraeten D, Schrauwen B, Massar S: **Information processing capacity of dynamical systems**. *Sci Rep* 2012:2.
The authors show that almost any dynamical system has an ability to memorize past inputs, and perform non-linear computations on them. Furthermore, they show a tradeoff between these two properties.

47. • Toyoizumi T: **Nearly extensive sequential memory lifetime achieved by coupled nonlinear neurons**. *Neural Comput* 2012, **24**:2678-2699.
The author shows that a synfire-like feedforward architecture can enable a network of nonlinear neurons to represent the past state of a stimulus up to a time that scales linearly with network size. This favorable scaling is due to exploiting the nonlinearity of the neurons for error correcting.

48. Wallace E, Maei HR, Latham PE: **Randomly connected networks have short temporal memory**. *Neural Comput* 2013, **25**:1408-1439.

49. Hermans M, Schrauwen B: **Memory in linear recurrent neural networks in continuous time**. *Neural Netw* 2010, **23**:341-355.

50. •• Barak O, Sussillo D, Romo R, Tsodyks M, Abbott LF: **From fixed points to chaos: three models of delayed discrimination**. *Prog Neurobiol* 2013 http://dx.doi.org/10.1016/j.pneurobio.02.002.
The authors show that parametric working memory tasks can be implemented in networks with vastly different organizing principles, from fine-tuned line attractors to unstructured chaotic networks. By comparing the emerging properties of model neurons with experimental data, they conclude that initially random networks that undergo a moderate amount of training are most compatible with the data.

51. Machens CK, Romo R, Brody CD: **Flexible control of mutual inhibition: a neural model of two-interval discrimination**. *Science* 2005, **307**:1121-1124.

52. Laje R, Buonomano DV: **Robust timing and motor patterns by taming chaos in recurrent neural networks**. *Nat Neurosci* 2013, **16**:925-933.

53. Szatmáry B, Izhikevich EM: **Spike-timing theory of working memory**. *PLoS Comput Biol* 2010, **6**:e1000879.

54. • Sussillo D, Barak O: **Opening the Black Box: low-dimensional dynamics in high-dimensional recurrent neural networks**. *Neural Comput* 2013, **25**:626-649.
Reservoir computing techniques often specify what the network has to do, but not how it should do it. The authors present an algorithm that allows to discover dynamical objects of trained recurrent neural networks, and learn the 'how'.