Road-mapping Biology

How the organization of biological knowledge impacts revolution strategy

Adam Marblestone
Revolutionary Ventures 2015

Agenda

- What is a scientific roadmap?
- Structural inefficiencies in biology:
  - Why does biology need roadmaps?
  - Why are there so many hidden gems?
- Examples of roadmaps, and of hidden gems
- Improved software for mapping science

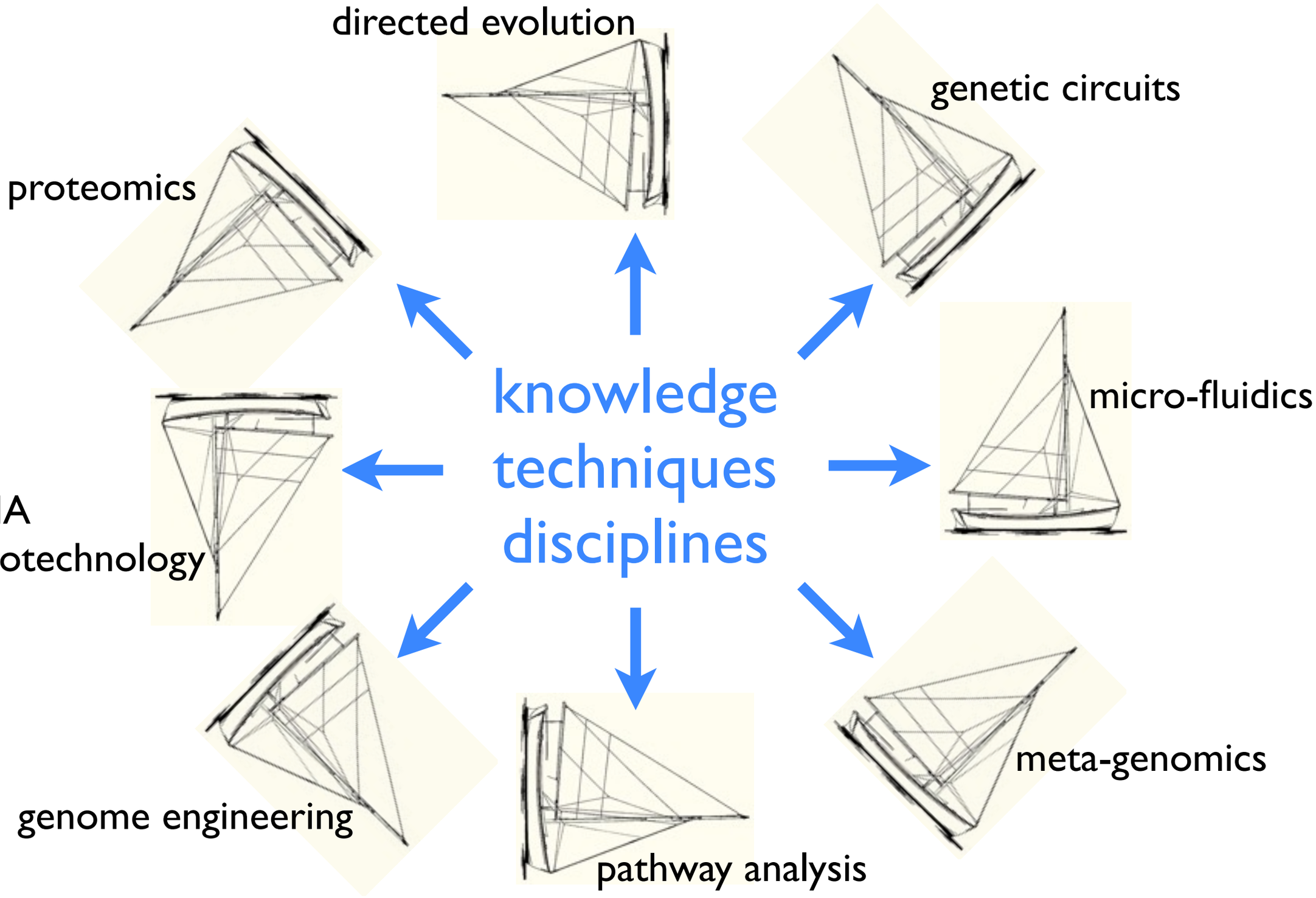
- **Roadmap**: A map of constraints, on the way towards a goal, and of potential workarounds for those constraints

- **Engineered Serendipity**: Biology breakthroughs depend on serendipity. We can make serendipity more likely by systematically surveying for hidden gems.

Why does biology need roadmaps?

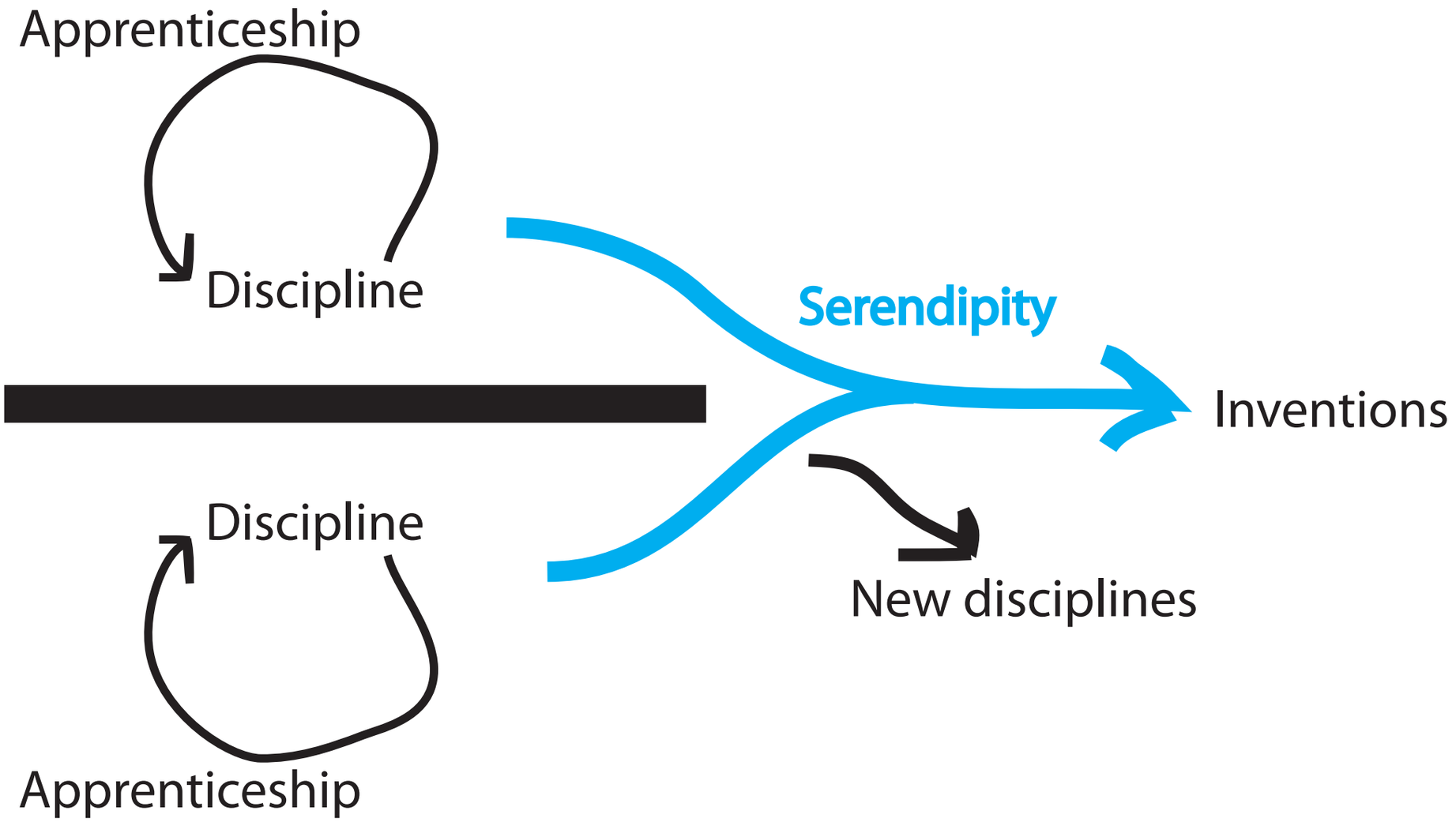
Why are there so many hidden gems?

“Science” <u>does not have a plan</u> for solving biotech grand challenges



A “sociological big bang”:

does not scale well with problem complexity


Apprenticeship
Discipline
Serendipity
Inventions
Discipline
New disciplines
Apprenticeship

The elephant in the room:

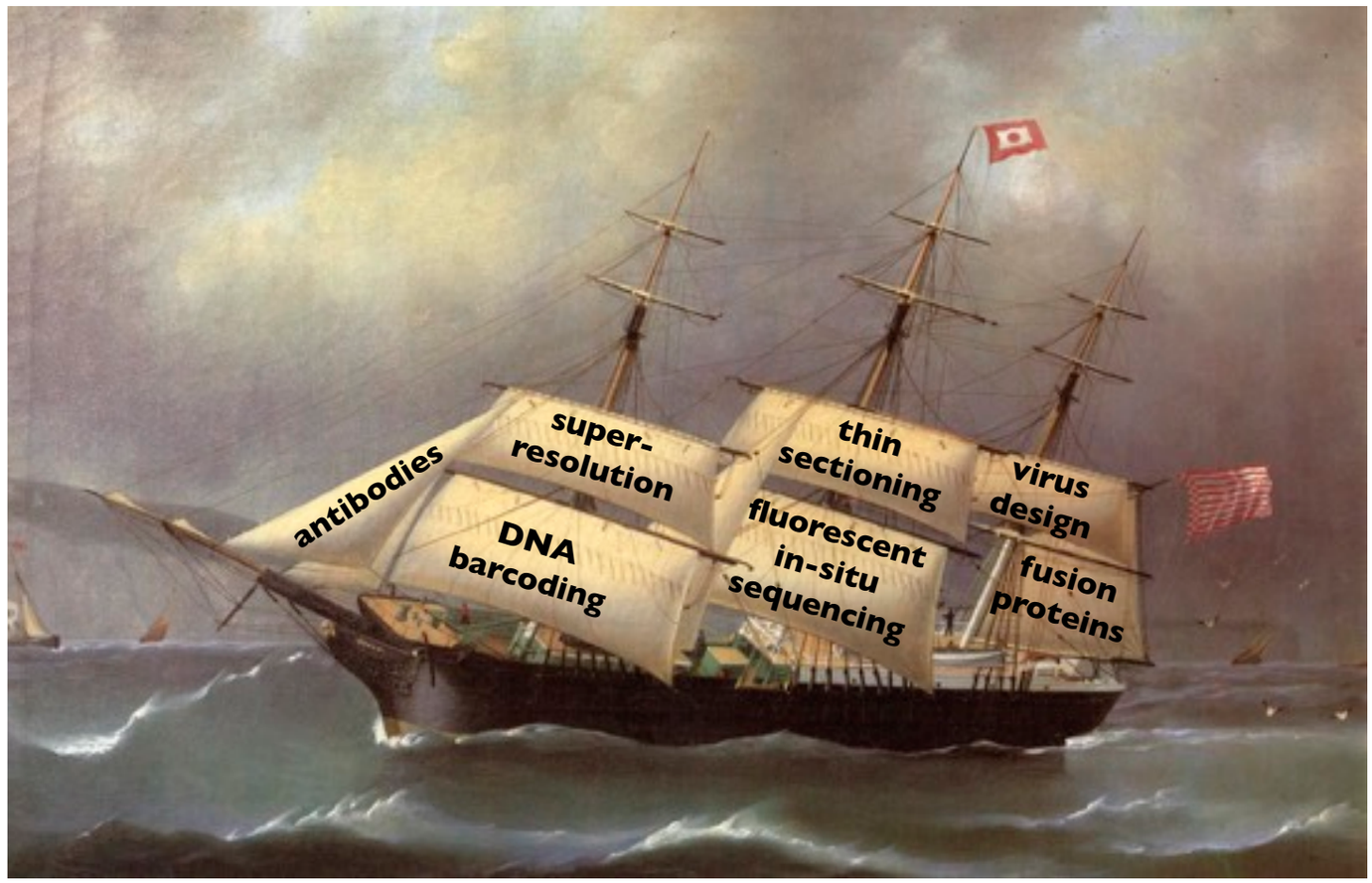
to make bio-technological quantum leaps
we must change how biology is done

revolutions by chance → **revolutions by design**

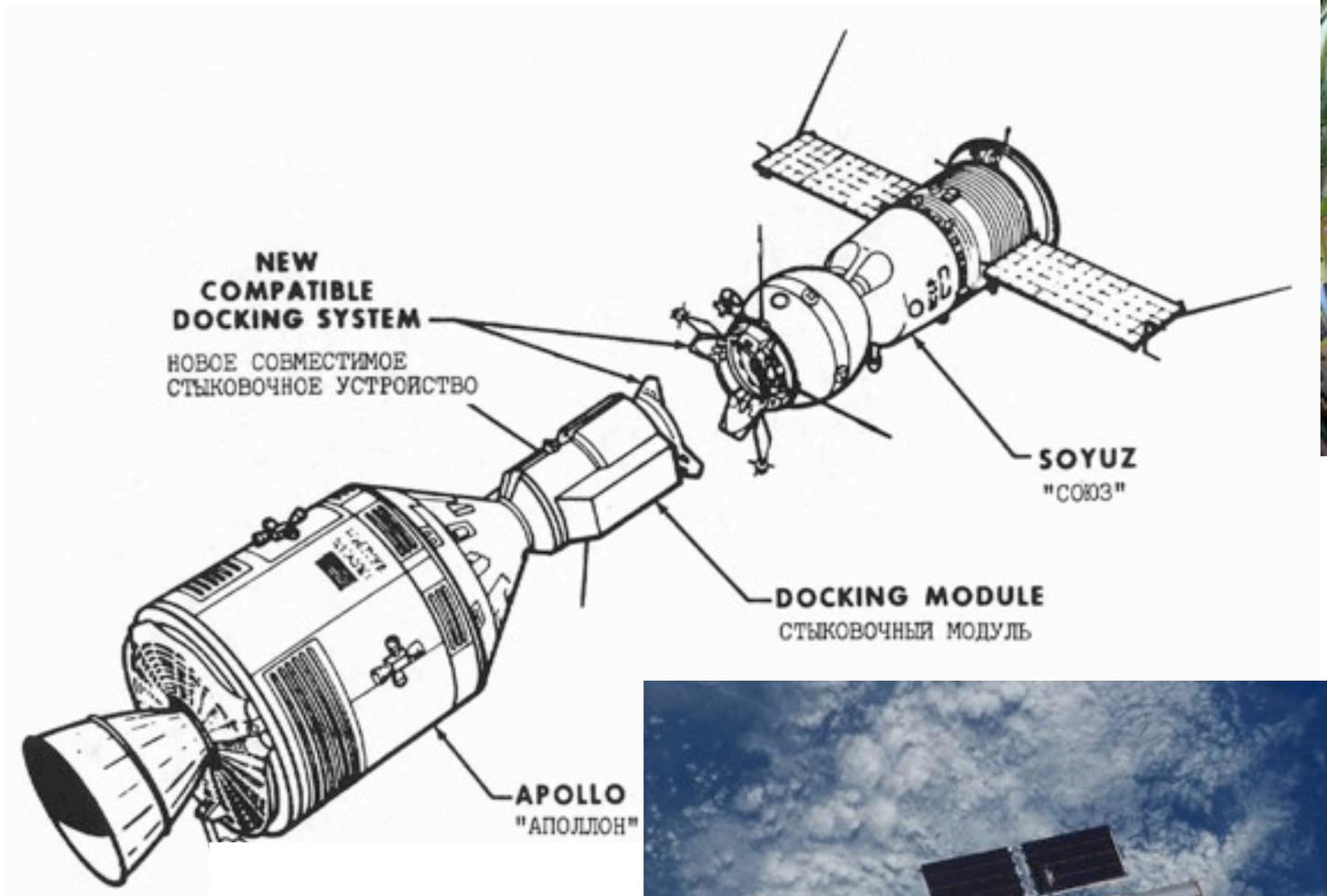
Two conspicuously missing elements in modern biology

maps of entire domains

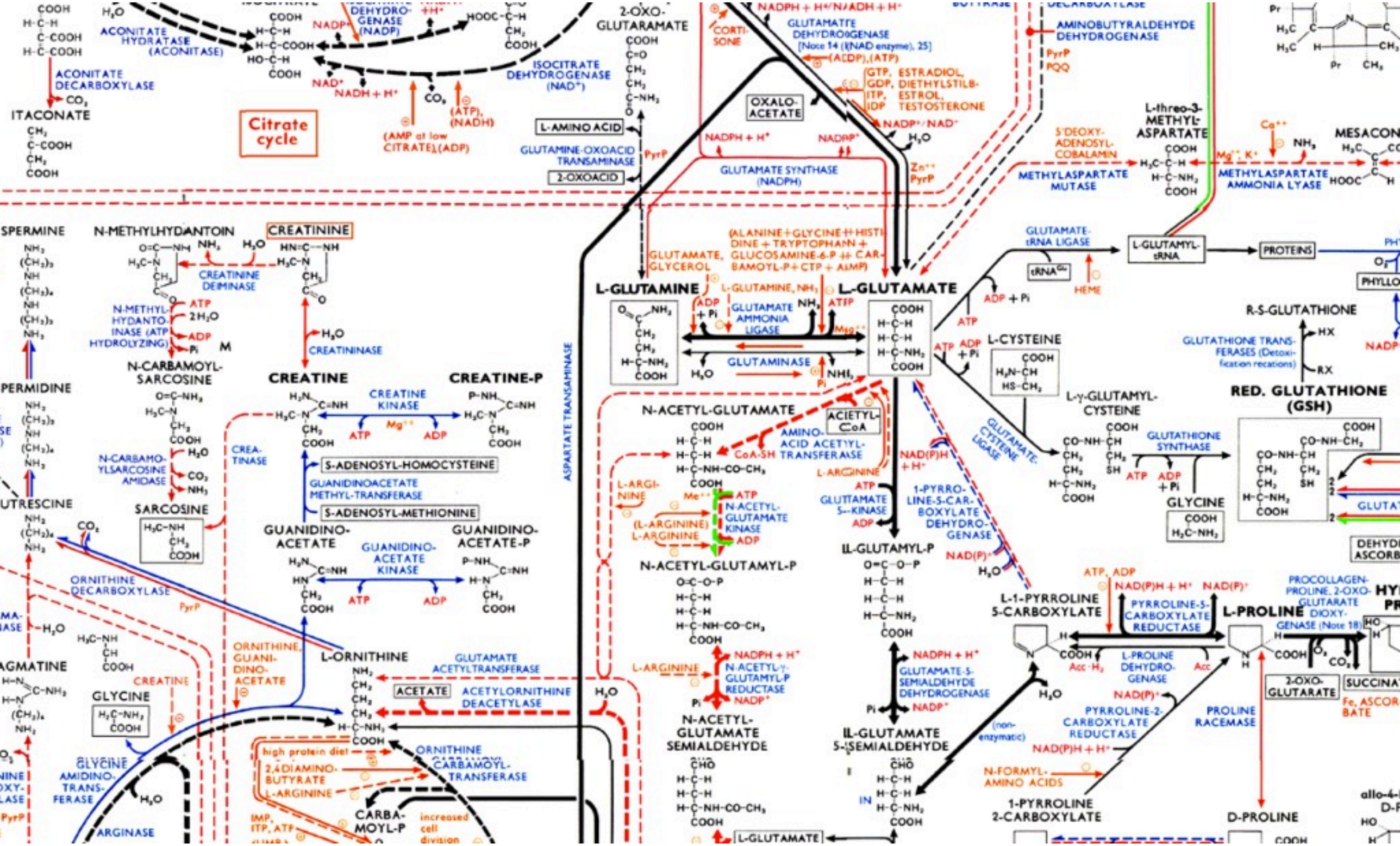
tech architectures / strategies



Well-known “systems engineering” methods in the sciences of *simplicity*...



How to extend to the sciences of complexity?



innovation “algorithm” & “business model” in bio is ripe for change...

How to create a bio-methods revolution

(only partially joking)

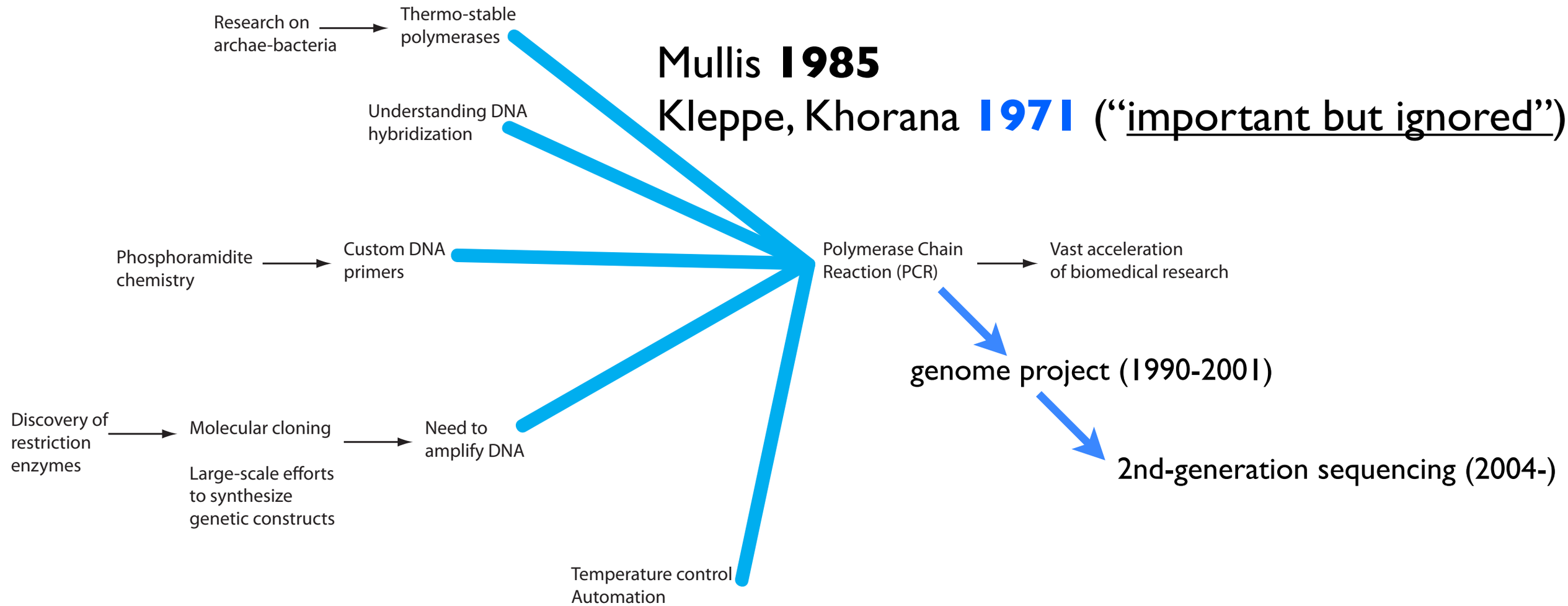
1) Understand landscape of *constraints* (physics, design robustness)
2) Find a *conceptual* assumption-violating workaround
3) Find a *hidden gem* (e.g., in nature) that implements this concept

#3 can be done via:

a) *searching* existing knowledge and transplanting to new domain
b) *screening* libraries of elements from nature (mining evolution)

PCR: a gem "hiding in plain sight"

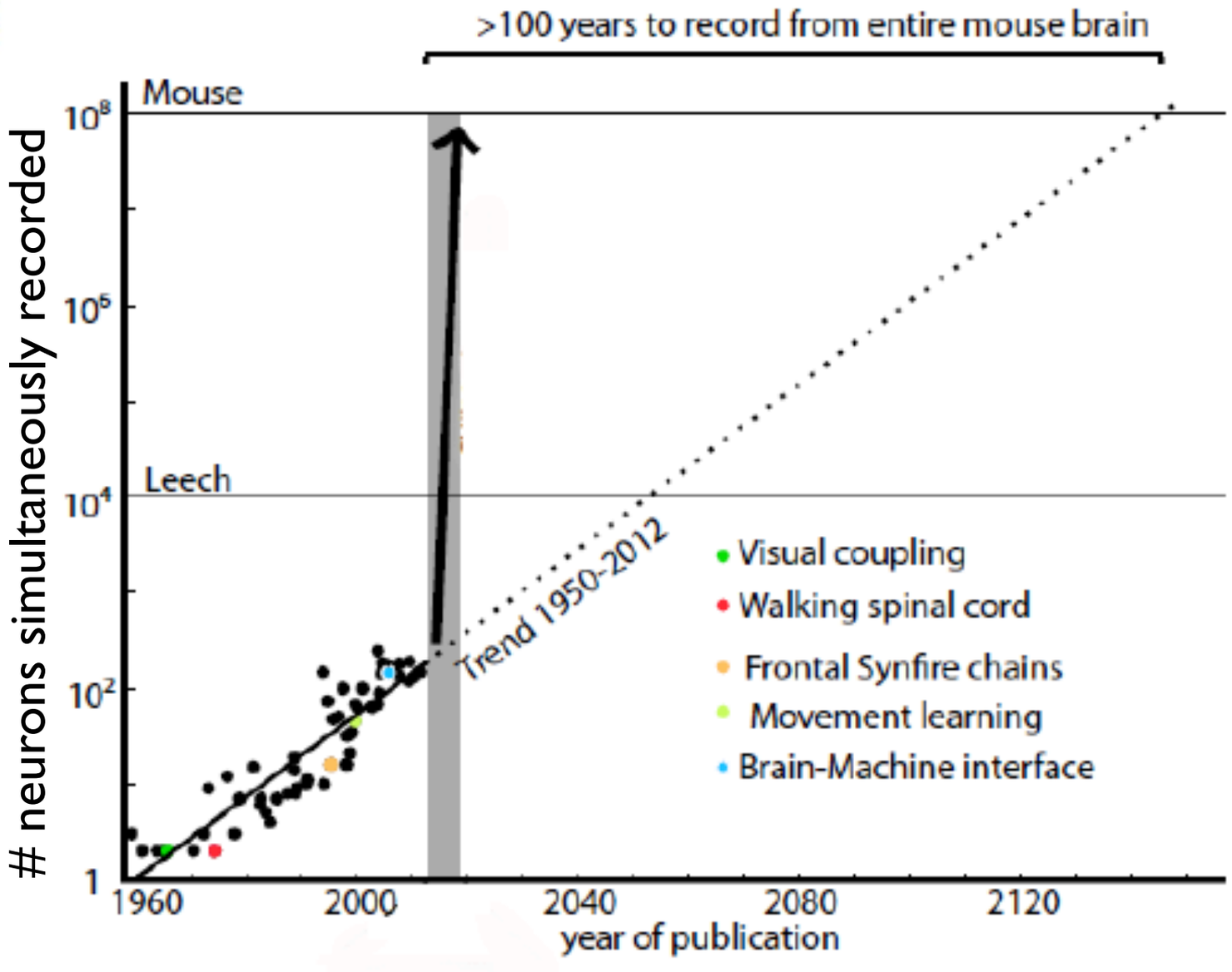
Goal: make lots of copies of an arbitrary DNA sequence



Mullis **1985**

Kleppe, Khorana **1971** ("important but ignored")

Goal: record every “spike” from every neuron in a mammalian brain



Konrad Kording 2013; Alivasatos et al 2012

nobody has written down a design that clearly

solves the problem

does not violate any laws of physics

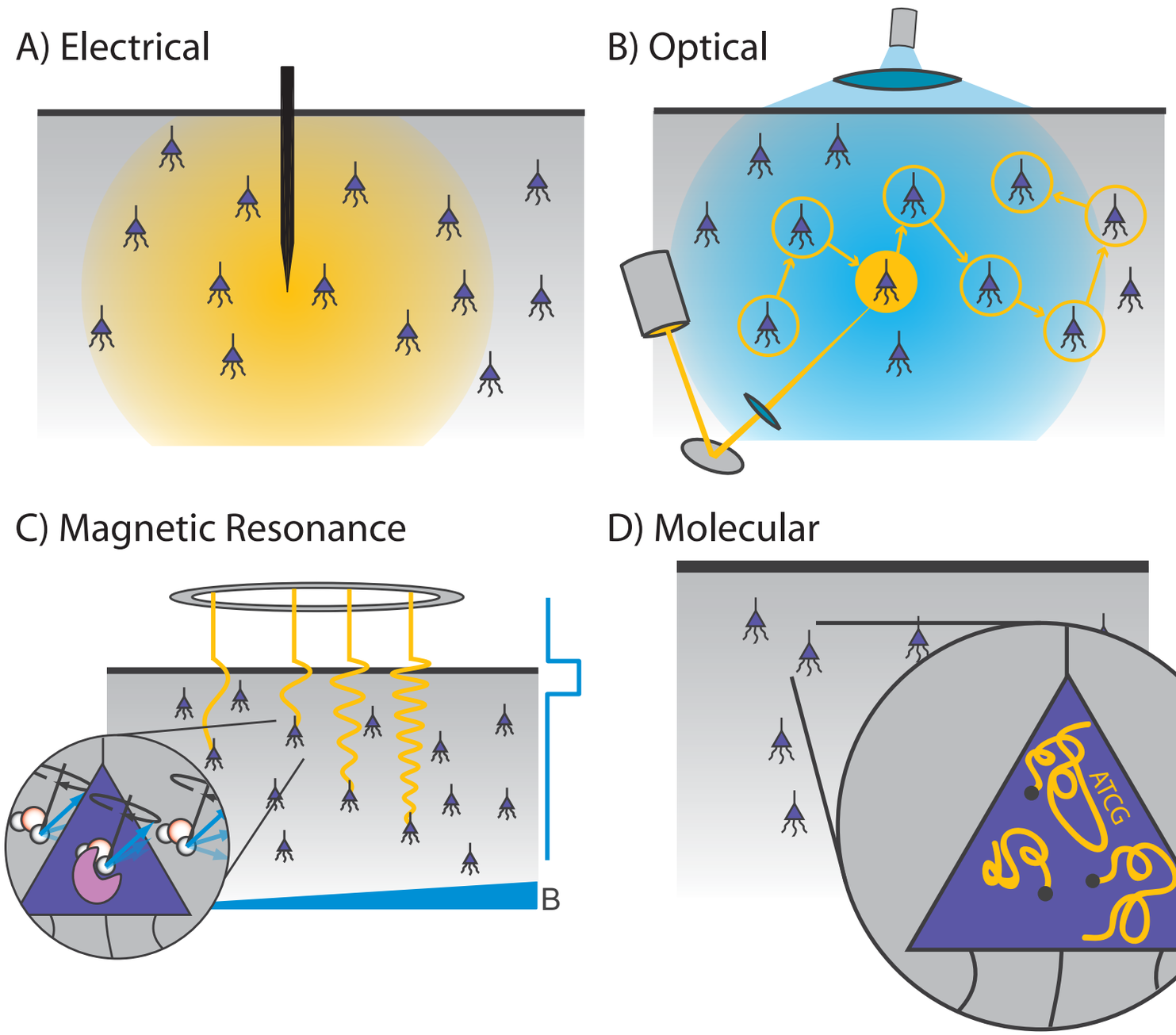
does not severely damage the brain

Landscape of constraints on brain activity mapping

< 2C temperature change:

→ **< 50 mW** steady-state power dissipation

< 1% tissue volume displacement



Marblestone et al, Front. Comp. Neurosci (2013): **with >17 co-authors**

http://arxiv.org/abs/1306.5709

Modality	Analysis Strategy	Assumptions	Conclusions
Extracellular electrical recording	Compute minimal number of recorders based on max distance from recorder to recorded neuron Compute channel capacity limits to spike sorting	Decay profile of extracellular voltage Approximate noise levels at recording site	Maximum recording distance $r_{max} \approx 100$–$200\,\mu m$ from electrode to neuron measured $\sim 10^5$ recording sites are required per mouse brain at current noise levels assuming perfect spike sorting $\sim 10^6$ recording sites are required at current noise levels at the physical limits of spike sorting $\sim 10^7$ recording sites are required using current spike sorting algorithms
Implanted electrical recorders	Compute power dissipation of electronic devices that digitally sample neuronal activity	Physical limit: $k_B T \ln(2)$/ bit erased Practical limit: $\sim 10 k_B T$/ bit processed Current CMOS digital circuits: $> 10^5 k_B T$/ bit processed	Requires 2–3 orders of magnitude increase in the power efficiency of electronics relative to current devices to scale to whole-brain simultaneous recordings Minimalist architectures could be developed to reduce local data processing overhead
Wireless data transmission	Compute tradeoff between power dissipation and channel bandwidth using information theory	Transmitter must supply enough power to overcome noise and path loss	Transmission at optical or near-optical frequencies is needed to achieve sufficient single-channel data rates using electromagnetic radiation. Radio-frequency (RF) electromagnetic transmission of whole-brain activity data draws excessive power due to bandwidth constraints Bandwidth cannot be split over multiple independent RF channels, but IR light or ultrasound may allow spatial multiplexing
Optical imaging	Relate the scattering and absorption lengths of optical wavelengths in brain tissue to signal-to-noise ratios for optical imaging	Approximate values of scattering and absorption lengths as a function of wavelength	Light scattering imposes severe constraints, but strategies exist which could negate the effects of scattering, such as implantable optics, infrared indicators, signal modulation, and online inversion of the scattering matrix
Multi-photon optical imaging	Compute minimum total excitation light power to excite multi-photon transitions from indicators within each neuron in every imaging frame	Approximate values of multi-photon cross-sections Pulse durations similar to those currently used in multi-photon imaging	Whole-brain multi-photon excitation will over-heat the brain except in very short experiments, unless ultra-high-cross-section indicators are used
Beam scanning microscopies	Calculate device and indicator parameters necessary for fast beam repositioning and signal detection	Fast optical phase modulators could reposition beams at ~1 GHz switching rates Fluorescence lifetimes in the 0.1–1.0 ns range	Beam repositioning time limits the speed of current systems but these are far from the physical limits Fluorescence lifetimes of indicators constrain design of ultra-fast scanning microscopies
Magnetic resonance imaging	Calculate spatial and temporal resolution of MRI based on spin relaxation times and spin diffusion	Proton MRI using tissue water Approximate T_1 and T_2 relaxation times and self-diffusion times for tissue water	Proton MRI is limited by the T_1 relaxation time of water to ~ 100 ms temporal resolution and by the self-diffusion of water to spatial resolutions of $\sim 40\,\mu m$. T_1 pre-mapping could allow T_2 contrast on a ~10 ms timescale. Achieving these limits for functional imaging requires going beyond BOLD contrast
Ultrasound	Calculate spatial resolution, signal strength and bandwidth limits on ultrasound imaging	Speed of sound in brain Attenuation length of ultrasound in brain	Attenuation of ultrasound by brain tissue and bone may be prohibitive at the ~ 100 MHz frequencies needed for single-cell resolution ultrasound imaging Ultrasound may be viable for spatially multiplexed data transmission from embedded devices [70]
Molecular recording	Compute metabolic load and volume constraint for rapid synthesis of large nucleic acid polymers Evaluate temporal resolution in simulated experiments using kinetic models [6]	Polymerase biochemical parameter ranges Metabolic requirements of genome replication	Molecular recording devices appear to fall within physical limits but their development poses multiple major challenges in synthetic biology Synchronization or time-stamping mechanisms are required for temporal resolution to approach the millisecond scale

many electrodes are needed

embedded electronics are too power-hungry, at present

use IR or ultrasound, not RF, for data-transmission

light scattering can potentially be overcome in several ways

multi-photon optics is too dissipative

requires many parallel scanned beams

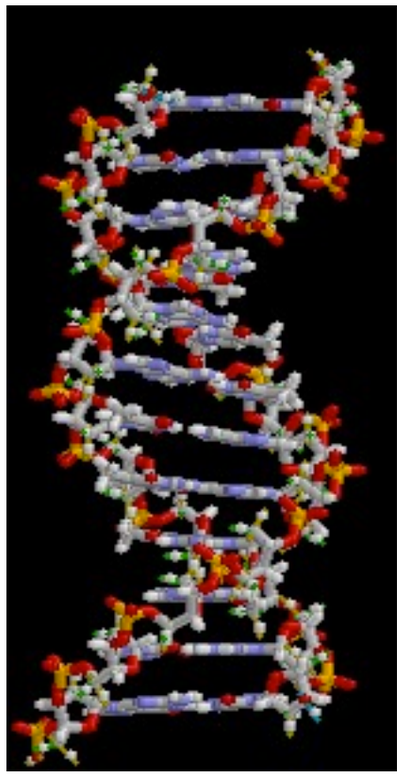
MRI needs new contrast mechanisms

ultrasound is potentially powerful

molecular recording is possible but hard

Need for a conceptual work-around

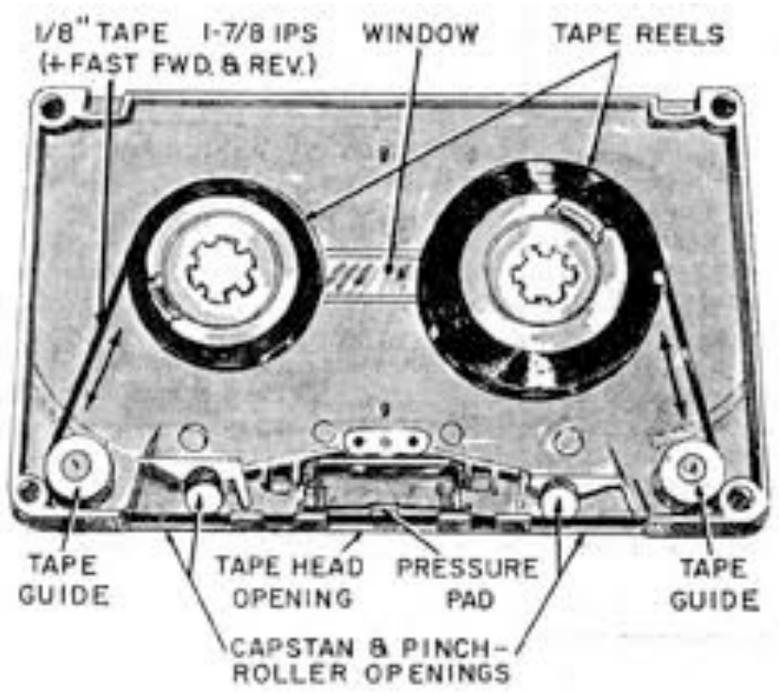
Assumption-violating <u>concept</u>:
what if each cell could record its own activity?



=

0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1
0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 1 0 1 1 1
0 0 1 0 1 1 0 1 0 1 0 1 0 1 0 1 1 1 0 1 0 0 1 1

=

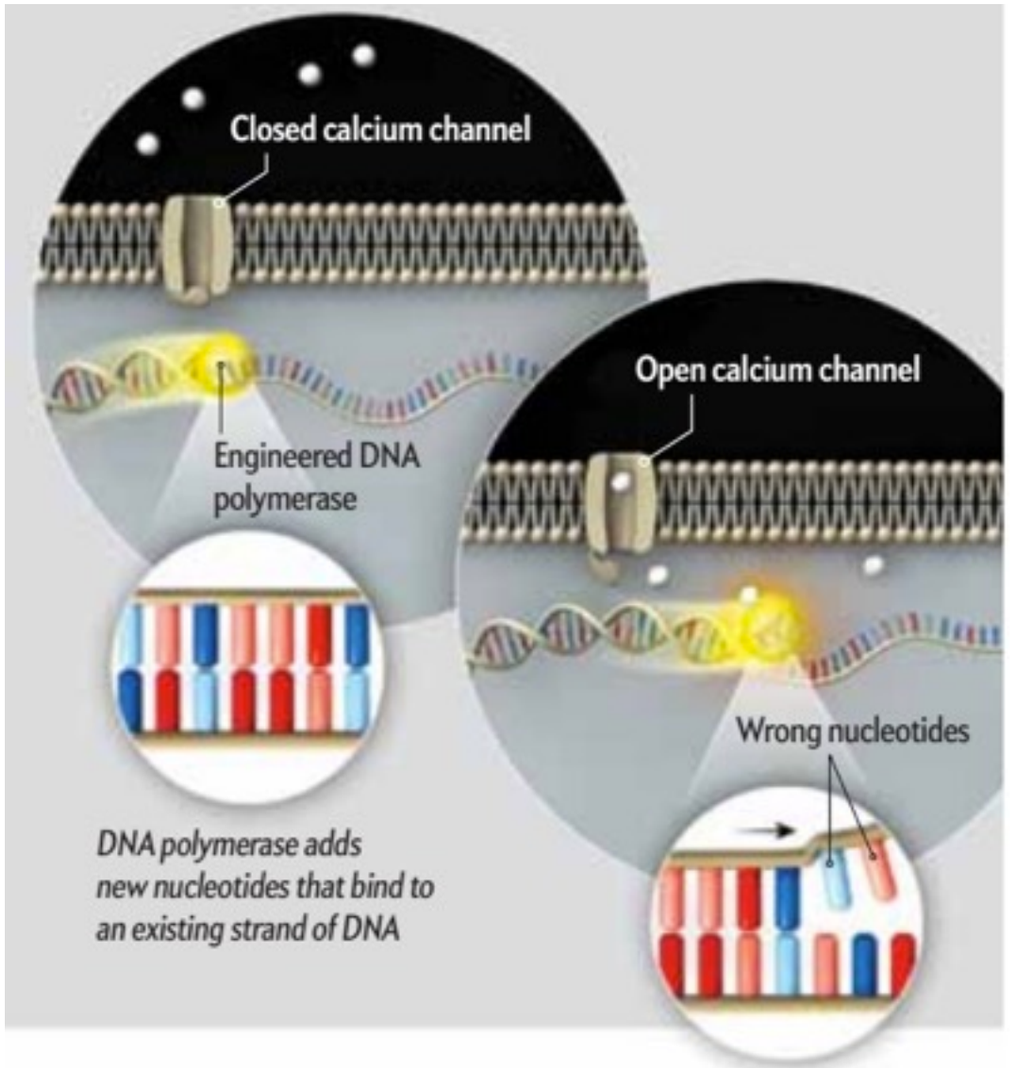


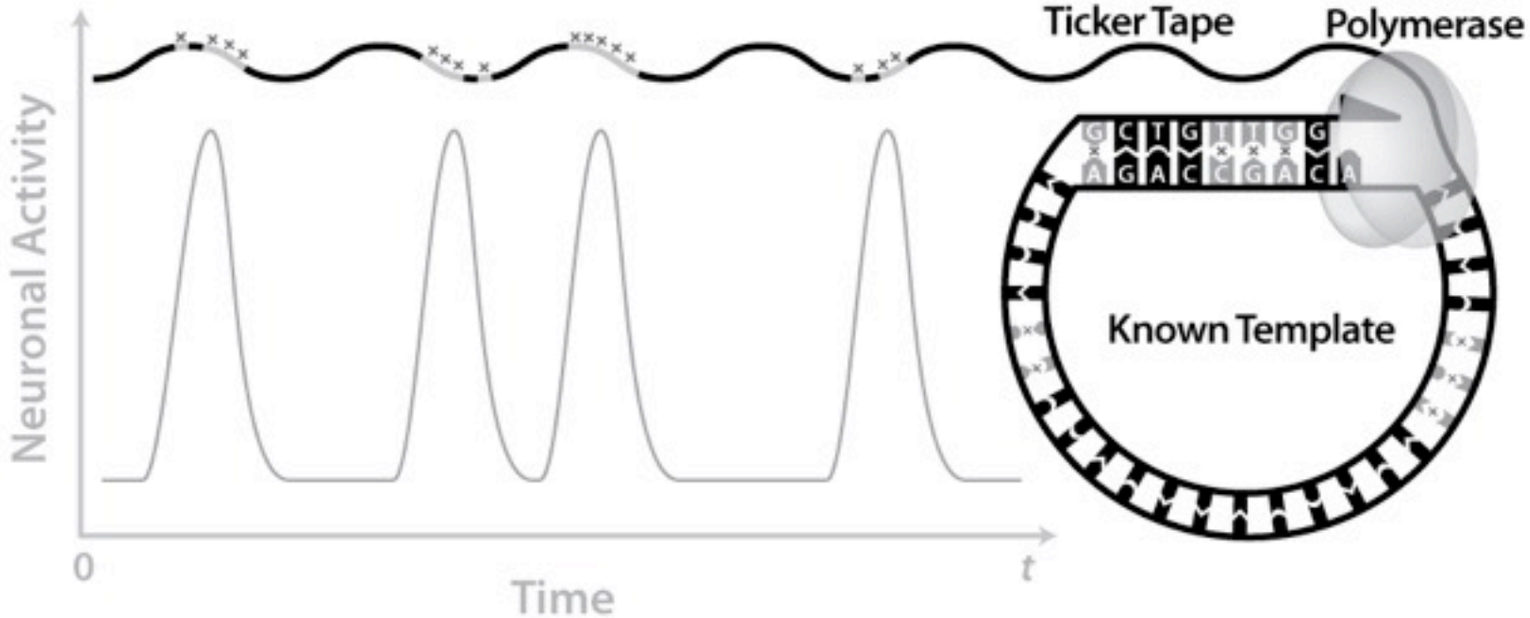
$9M NIH grant to pursue this

w/ **Kording**, **Zamft**, Church, Boyden et al

Molecular implementation of the concept:
a molecular recording device in each neuron

Encode information via control of DNA polymerase copying error rate





Yuste and Church, Scientific American, 2014

Figure by Reza Kalhor

Mapping out the space of *theories* about the brain

Computation	Algorithmic/ representational realization	Neural implementation(s)	Brain location(s)
Rapid perceptual classification	Receptive fields, pooling and local contrast normalization [51,55]	Hierarchies of simple and complex cells[56]	Visual system
Complex spatiotemporal pattern recognition	Bayesian belief propagation [19,57]	Feedforward and feedback pathways in cortical hierarchy[19]	Sensory hierarchies
Learning efficient coding of inputs	Sparse coding[58]	Thresholding and local competition[59]	Sensory and other systems
Working memory	Continuous or discrete attractor states in networks[60,61]	Persistent activity in recurrent networks[62]	Prefrontal cortex
Decision making	Temporal-difference reinforcement learning algorithms[63,64]; actor-critic models[65]	Cortically implemented Bayesian inference networks combined with temporal difference reinforcement learning via the dopamine system and action selection systems in the basal gangla[66]	Prefrontal cortex
	Winner-take-all networks[67]	Recurrent networks coupled via lateral inhibition[67]	Prefrontal cortex
Gating of information flow	Context-dependent tuning of activity in recurrent network dynamics[68]	Recurrent neural networks implementing line attractors and selection vectors[68]	Prefrontal cortex
	Shifter circuits[69]	Divergent excitatory relays and input-selective shunting inhibition in dendrites[69]	Visual system
Gain control	Divisive normalization[52]	Shunting inhibition in networks or balanced background synaptic excitation and inhibition [70]	Common across many cortical areas
Sequencing of events over time [71]	Feed-forward cascades; Serial working memories[72]	Synfire chains[73-75]; Thalamo-cortico-striatal loops[76,77]	Common across many cortical areas
Representation and transformation of variables	Population coding [78]	Time-varying firing rates of cosine-tuned neurons representing dot products with encoding vectors	Motor cortex
Variable binding	Holographic reduced representations [49,79]	Circular convolution of vectors represented by neural population codes	Cortical areas involved in sequential or symbolic processing
	Dynamic binding[80,81]	Neural synchronization[82]	

with Gary Marcus
(NYU + Allen Institute)
and Tom Dean (Google)

Can software tools accelerate the uptake of cross-disciplinary knowledge, helping us find the hidden gems?

Automatically learning the "meanings" of science words

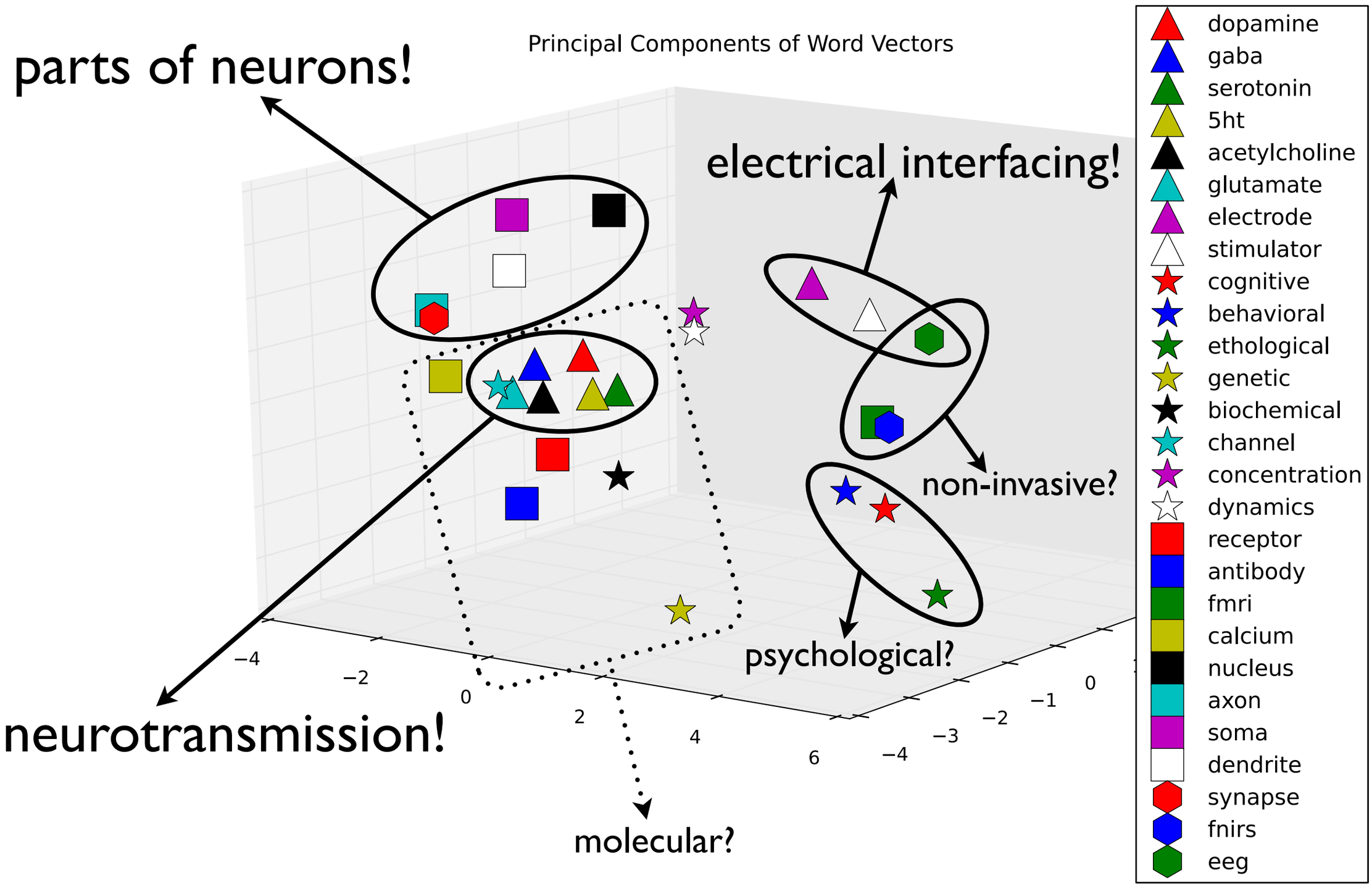
Word2Vec model trained on 150k PubMed abstracts

```
prompt> dopamin
dopamine          dopaminergic      dopaminergically  dopaminoceptive   dopaminomimetic   dopaminomimetics

prompt> dopamine

Similar words:
--------------
serotonergic
nsda
5ht
dopamine
noradrenergic
monoamines
nigro
striatal
dopaminergic
cholinergic
monoamine
hydroxytryptamine
nigrostriatal
reuptake
serotonin
nigral
norepinephrine
midbrain
msns
sert
striatum
noradrenalin
noradrenaline
daergic
mamph
```

statistical model was only fed raw text, and knows nothing about neuroscience (or anything), yet it "discovers" neurotransmitters

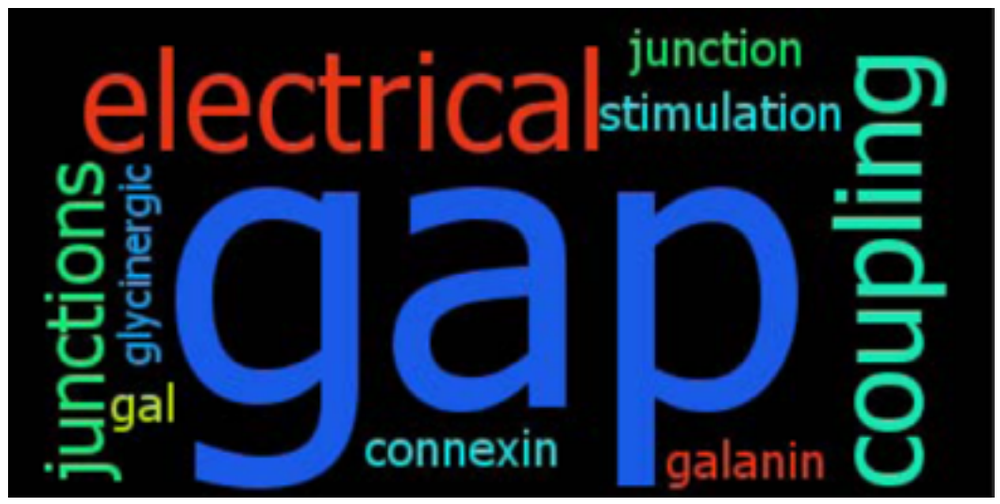
Automatically learning the "meanings" of science words

Word2Vec model trained on 150k PubMed abstracts



Identifying the sub-fields/sub-topics:
unsupervised document classification model

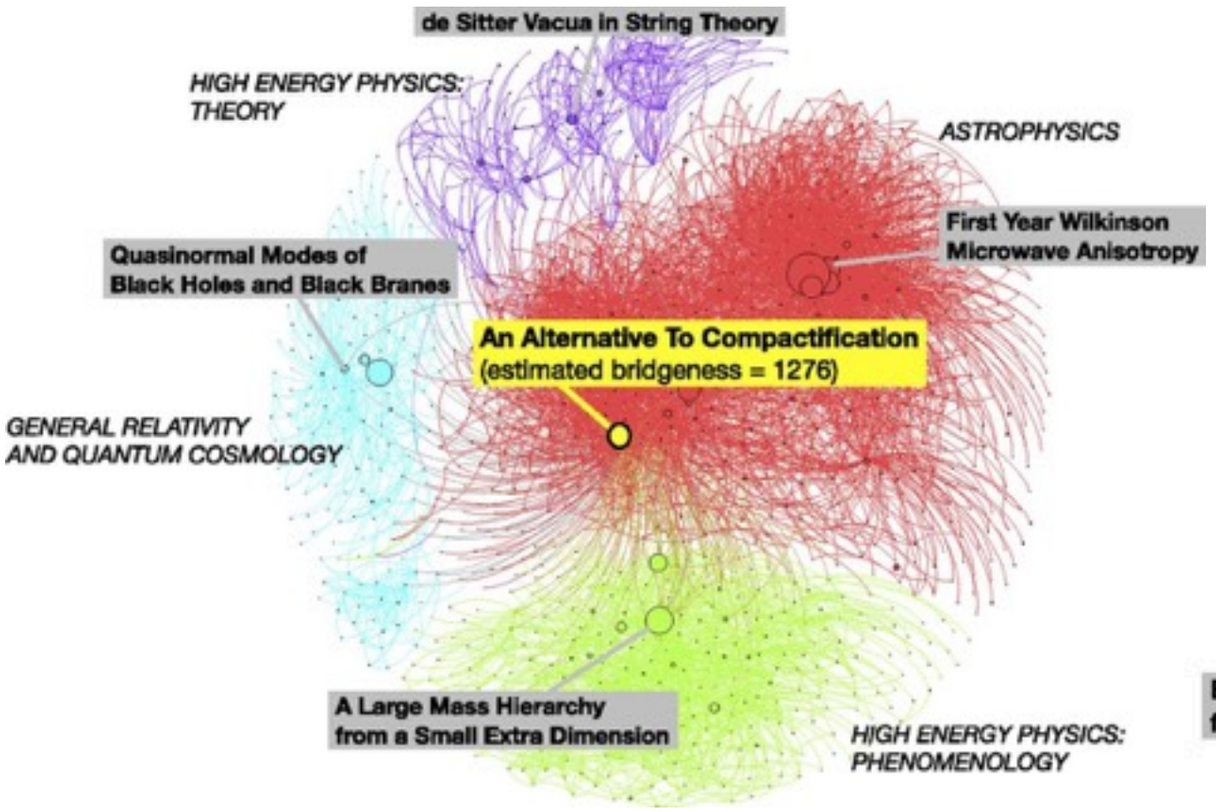
gaba is an inhibitory neurotransmitter

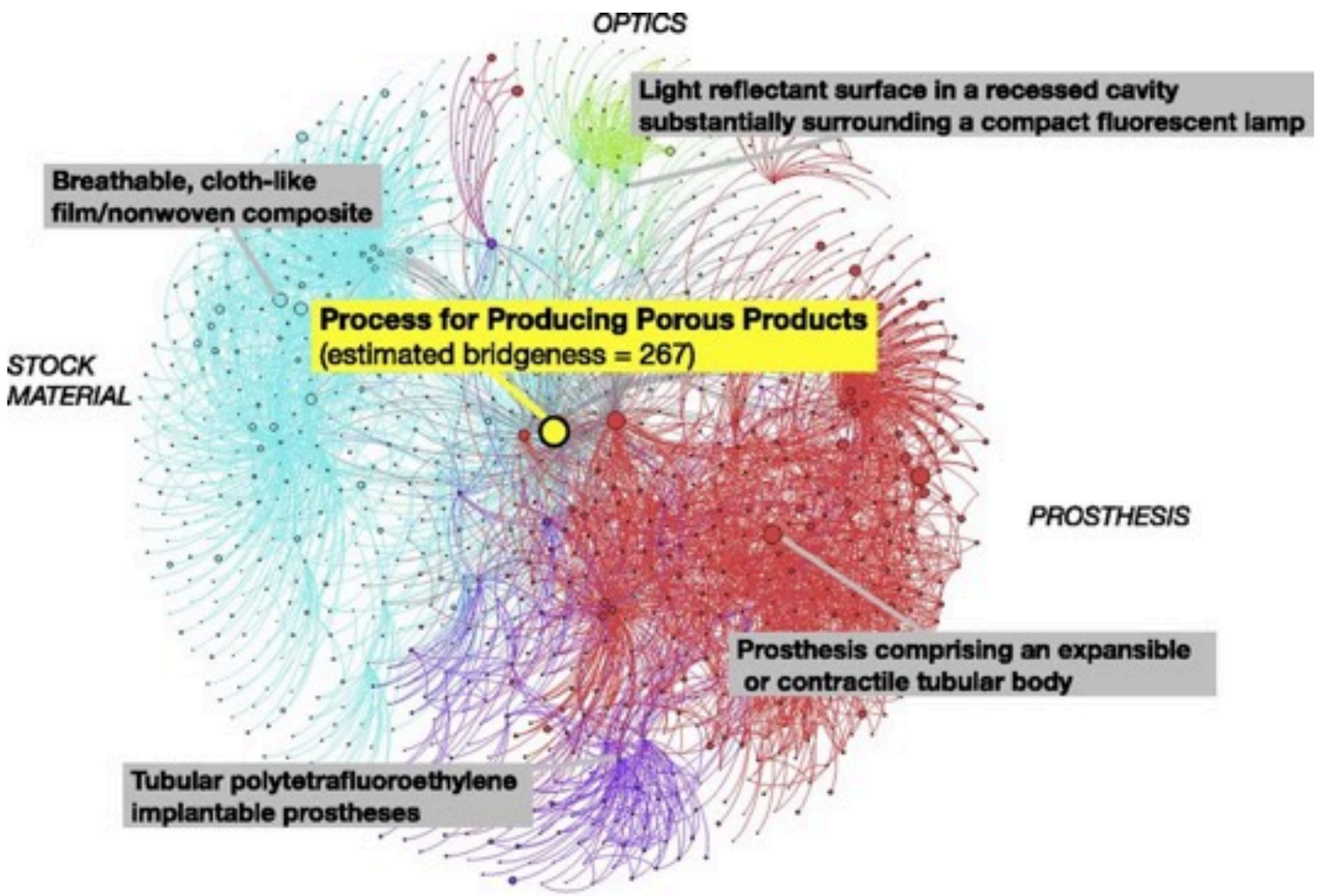


connexin is a gap-junction protein

the *dopamine* system lives in the *substantia nigra* region of the *striatum*

Identifying the sub-fields/sub-topics: community-detection on the citation graph





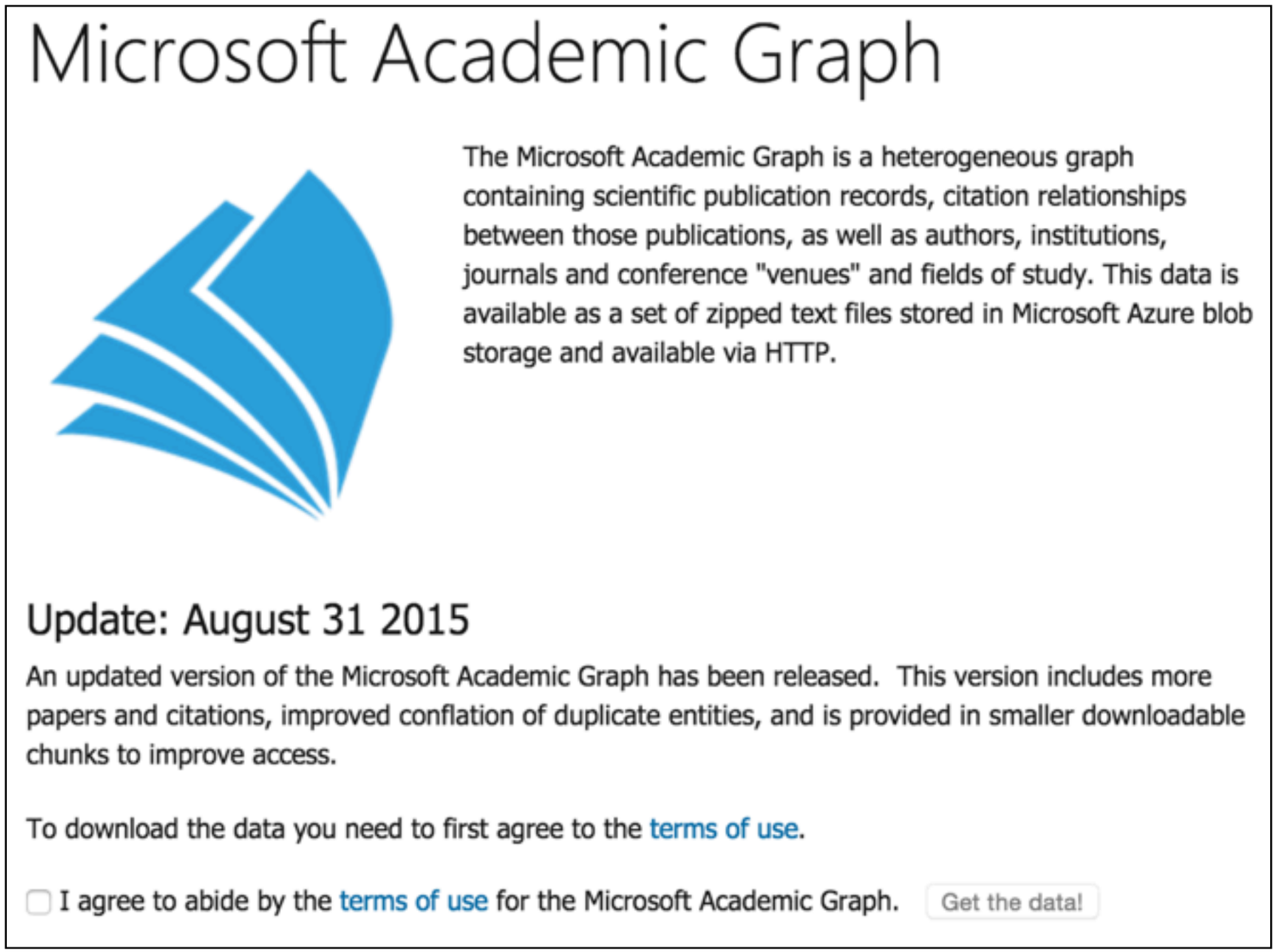
Efficient discovery of overlapping communities in massive networks

Prem K. Gopalan[1] and David M. Blei

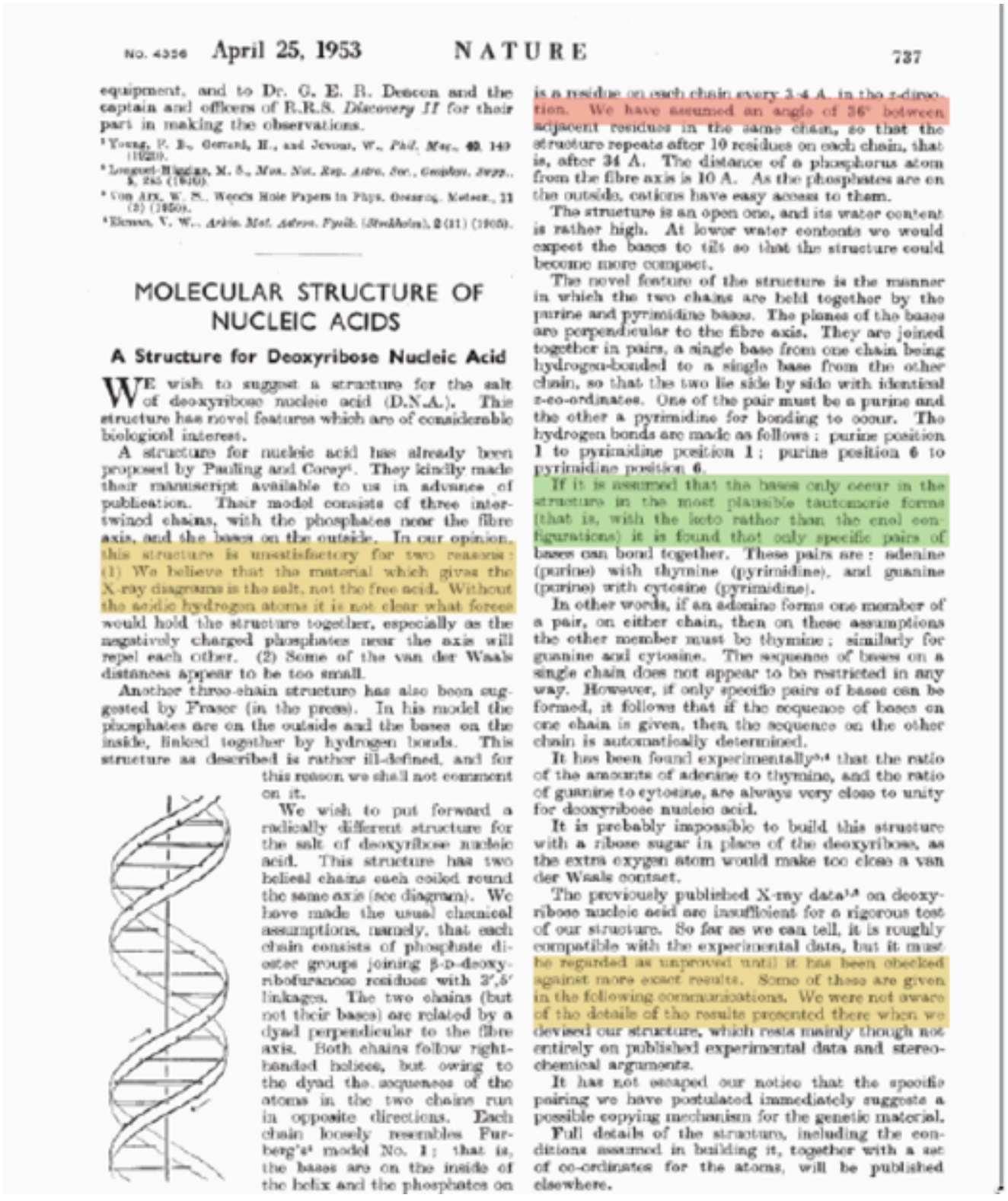
Department of Computer Science, Princeton University, Princeton, NJ 08540

Identifying the sub-fields/sub-topics:
community-detection on the citation graph

The first large-scale publicly available citation graph



Towards human-computer synergy for accelerated science



equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, P. E., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Van Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (2) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three inter-twined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

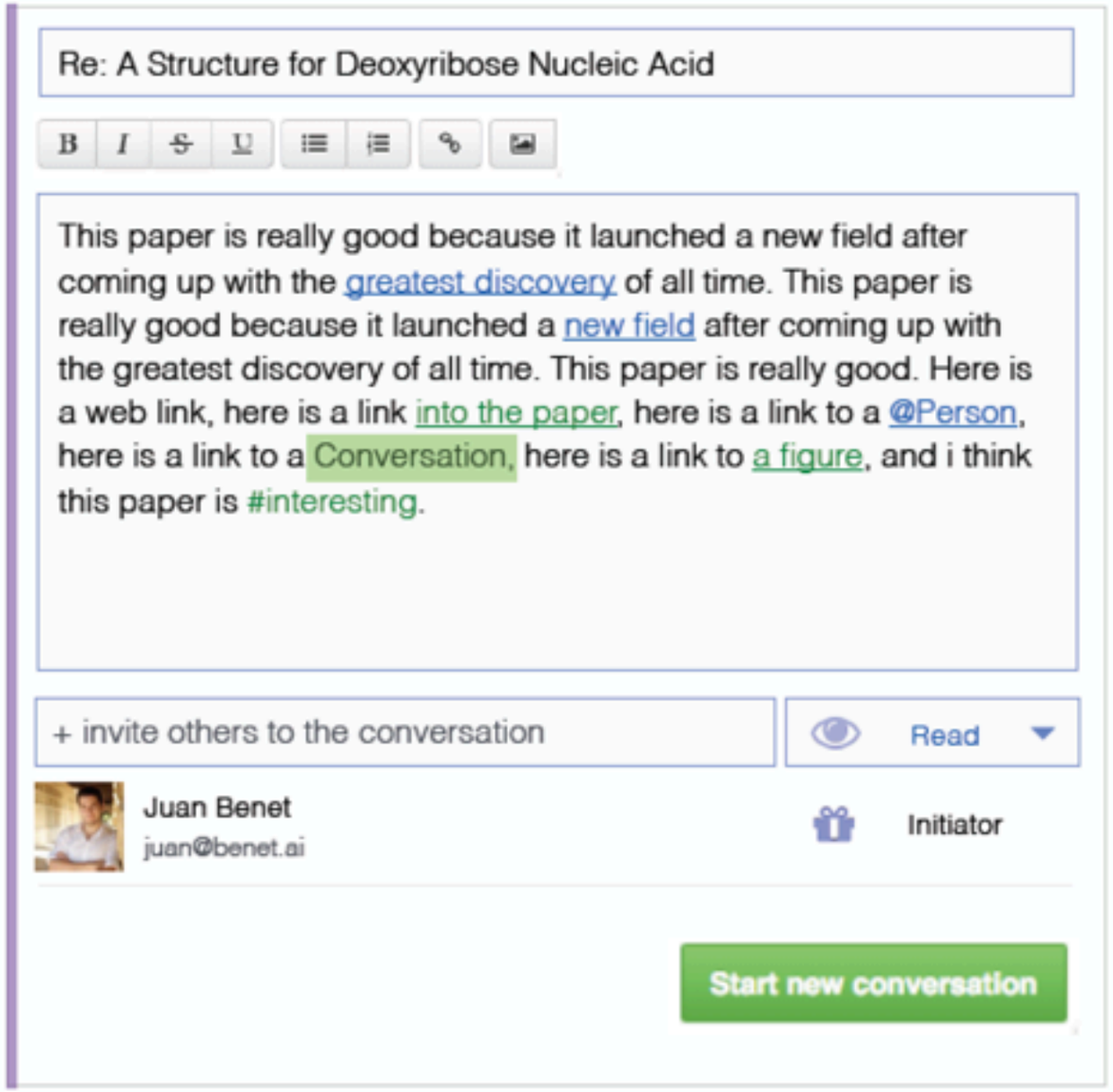
The previously published X-ray data[1,3] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

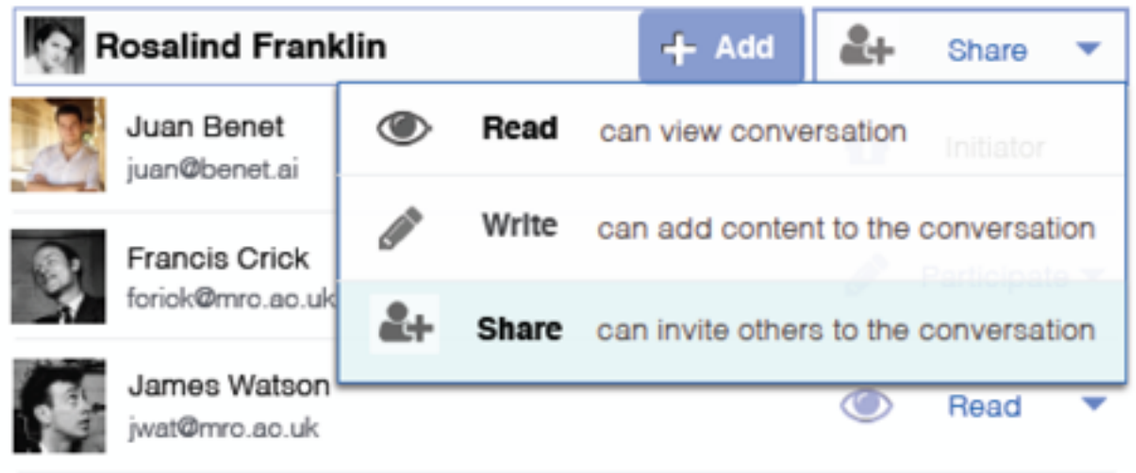
It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

With Juan Batiz-Benet, Richard Littauer, Ed Boyden

Towards human-computer synergy for accelerated science





With Juan Batiz-Benet, Richard Littauer, Ed Boyden