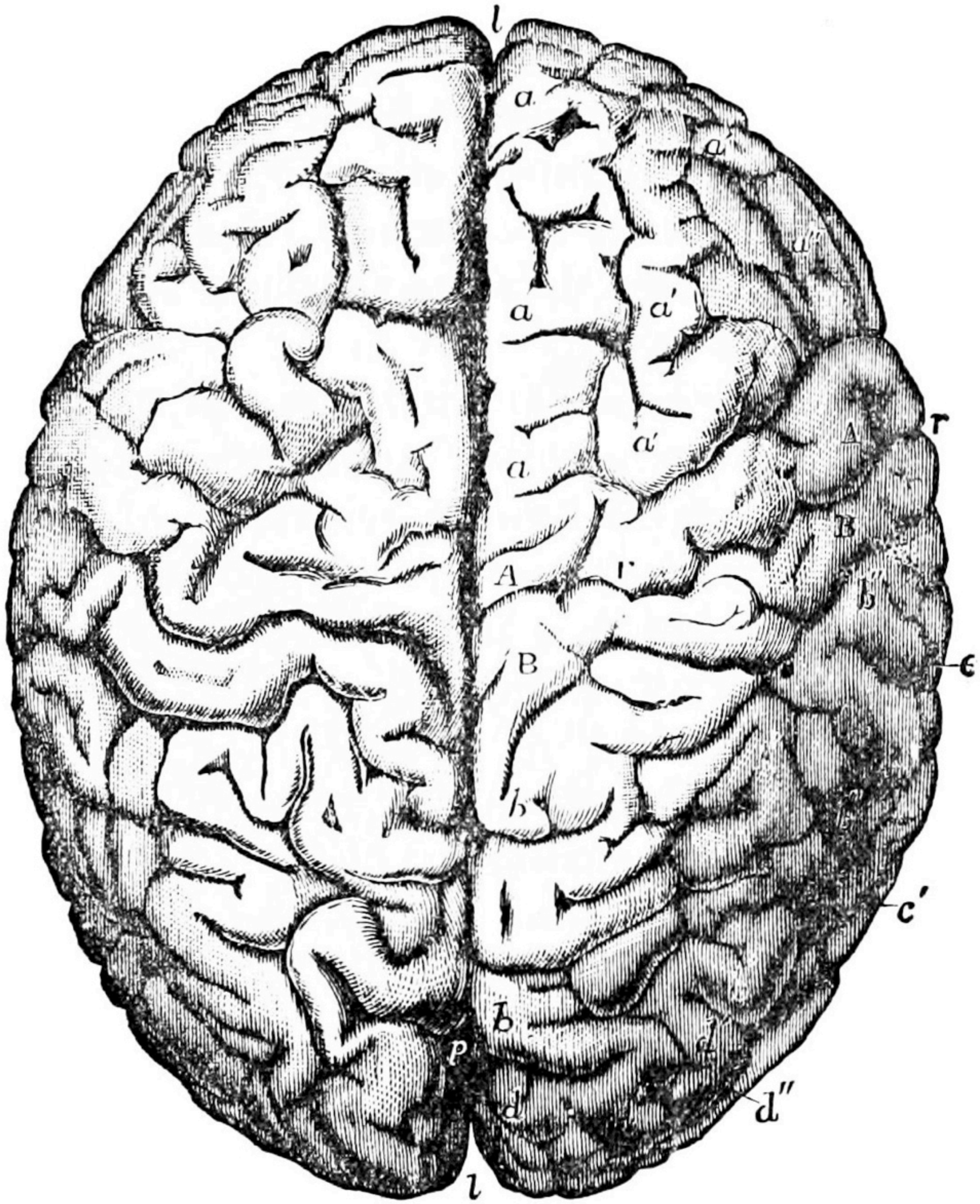
# Designing scalable architectures for neural data acquisition

Adam Marblestone
Church Lab
Harvard Biophysics

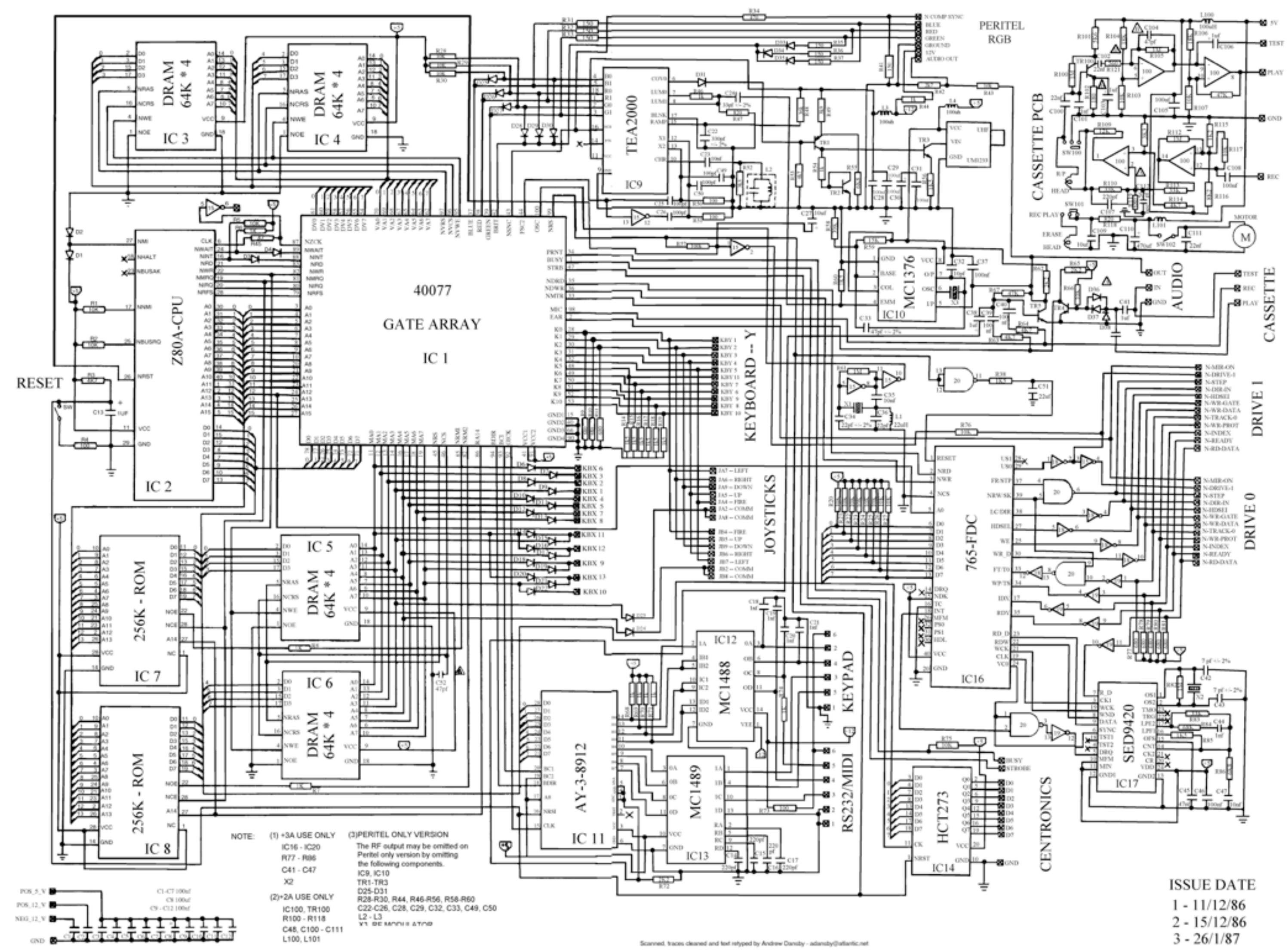
PhD Thesis Defense
April 2014

neuro

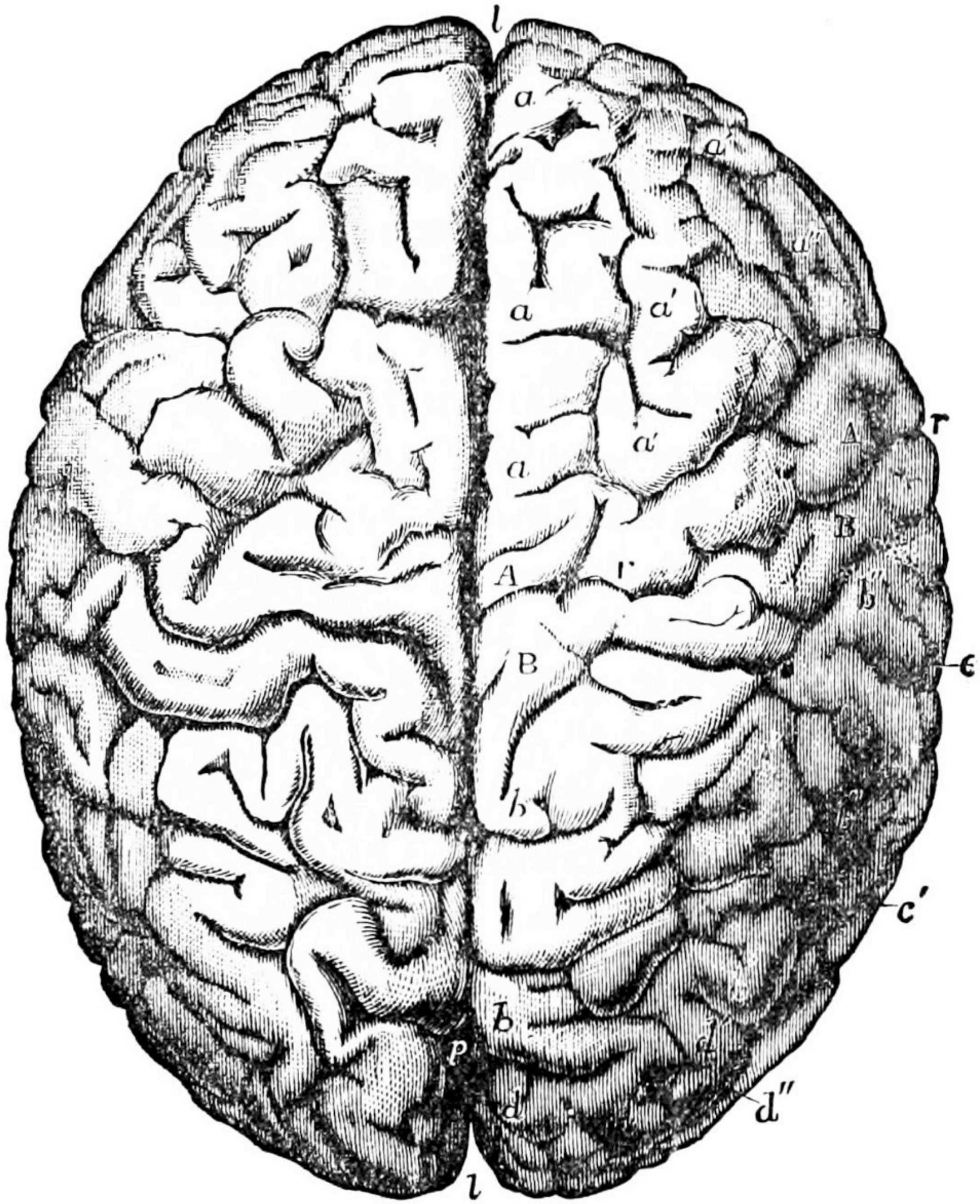
info



? =


40077
GATE ARRAY
IC 1
Z80A-CPU
IC 2
RESET
DRAM 64K * 4
IC 3
IC 4
IC 5
IC 6
256K - ROM
IC 7
IC 8
TEA2000
IC9
MC1376
IC10
AY-3-8912
IC 11
MC1488
IC12
MC1489
IC13
HCT273
IC14
765-FDC
IC16
SED9420
IC17
PERITEL RGB
CASSETTE PCB
AUDIO
CASSETTE
KEYBOARD -- Y
JOYSTICKS
KEYPAD
RS232/MIDI
CENTRONICS
DRIVE 1
DRIVE 0
ISSUE DATE
1 - 11/12/86
2 - 15/12/86
3 - 26/1/87

neuro




?
=

info

$$y_t(i) = P(^{-}e_t | c_i(t)) \propto \prod_{j=1}^{M} \lambda_t^{m_j}(r_i^{m_j}) \quad (2)$$

where coincidence pattern $c_i$ is the co-occurrence of $r_i^{m_1}$'th Markov chain from child 1, $r_i^{m_2}$'th Markov chain from child 2, $\cdots$, and $r_i^{m_M}$'th Markov chain from child $M$.

$$\lambda_t(g_r) = P(^{-}e_0^t | g_r(t)) \propto \sum_{c_i(t) \in C^k} \alpha_t(c_i, g_r) \quad (3)$$

$$\alpha_t(c_i, g_r) = P(^{-}e_t | c_i(t)) \sum_{c_j(t-1) \in C^k} P(c_i(t) | c_j(t-1), g_r) \alpha_{t-1}(c_j, g_r) \quad (4)$$

$$\alpha_0(c_i, g_r) = P(^{-}e_0 | c_i(t=0)) P(c_i(t=0) | g_r) \quad (5)$$

$$Bel_t(c_i) \propto \sum_{g_r \in G^k} \beta_t(c_i, g_r) \quad (6)$$

$$\beta_t(c_i, g_r) = P(^{-}e_t | c_i(t)) \sum_{c_j(t-1) \in C^k} P(c_i(t) | c_j(t-1), g_r) \beta_{t-1}(c_j, g_r) \quad (7)$$

$$\beta_0(c_i, g_r) = P(^{-}e_0 | c_i(t=0)) P(c_i | g_r) \pi_0(g_r) \quad (8)$$

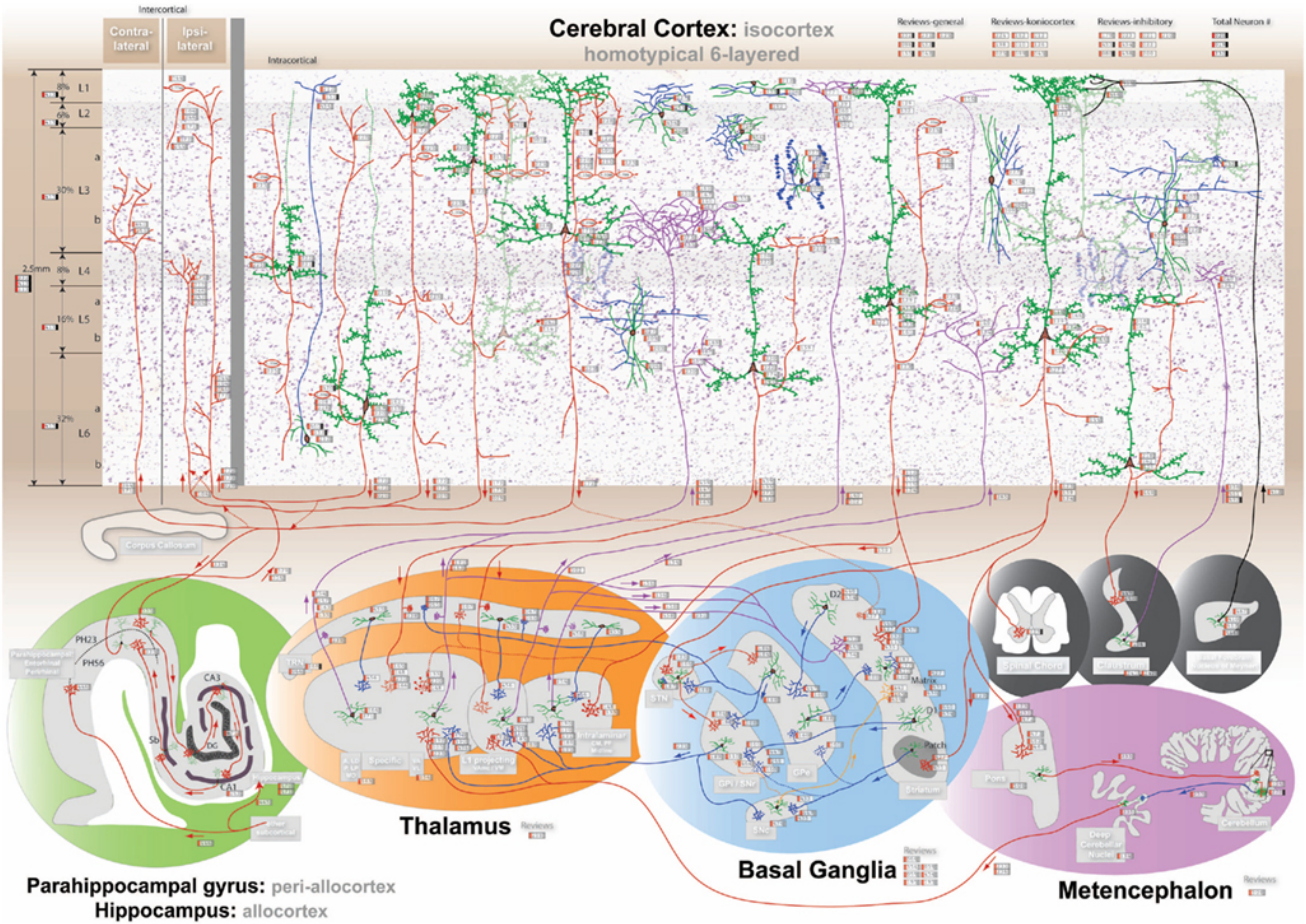
$$\pi^{m_i}(g_r) \propto \sum_i I(c_i) Bel(c_i) \quad (9)$$

where

$$I(c_i) = \begin{cases} 1, \text{ if } g_r^{m_i} \text{ is a component of } c_i \\ 0, \text{ otherwise} \end{cases} \quad (10)$$

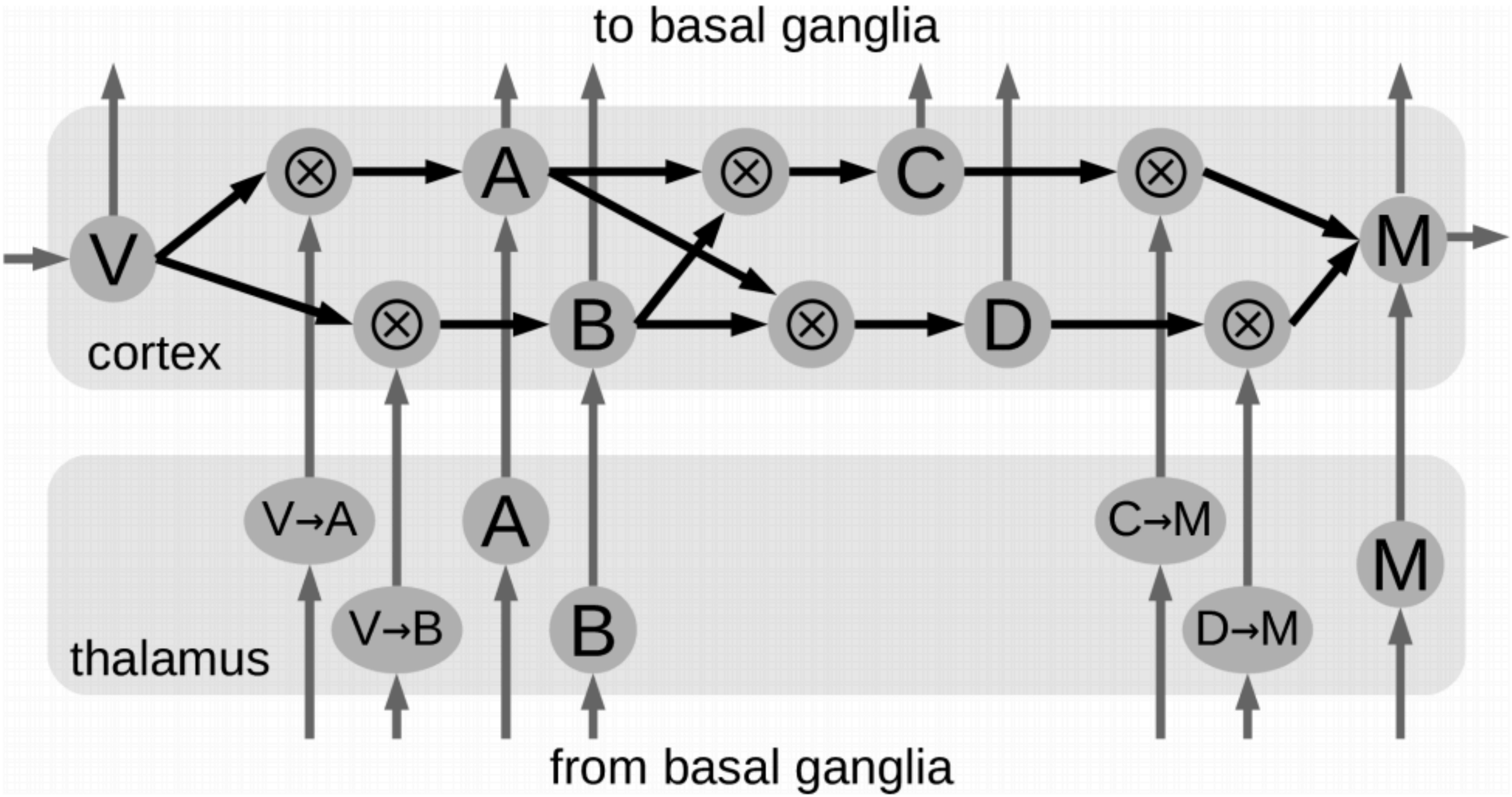
George & Hawkins, 2009

# **Coarse-grained brain** *maps* **give un-testable** *models*




Solari and Stoner 2011

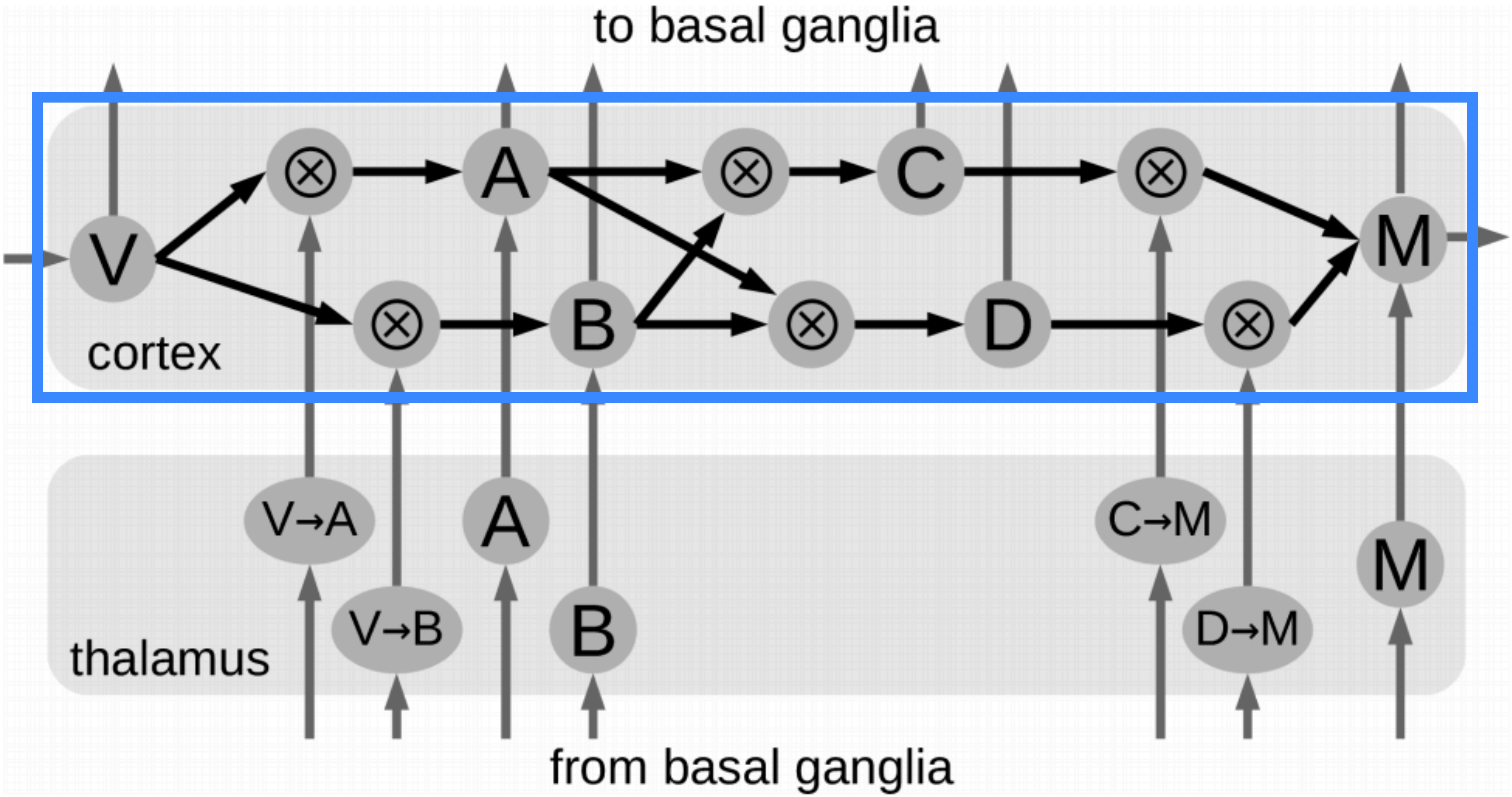
# **Coarse-grained brain** *maps* **give un-testable** *models*




thalamic gating of "copy and paste" operations between cortical working memory buffers, executing a sequence of steps controlled by the basal ganglia

Stewart, Eliasmith et al 2010

# **Coarse-grained brain** *maps* **give un-testable** *models*

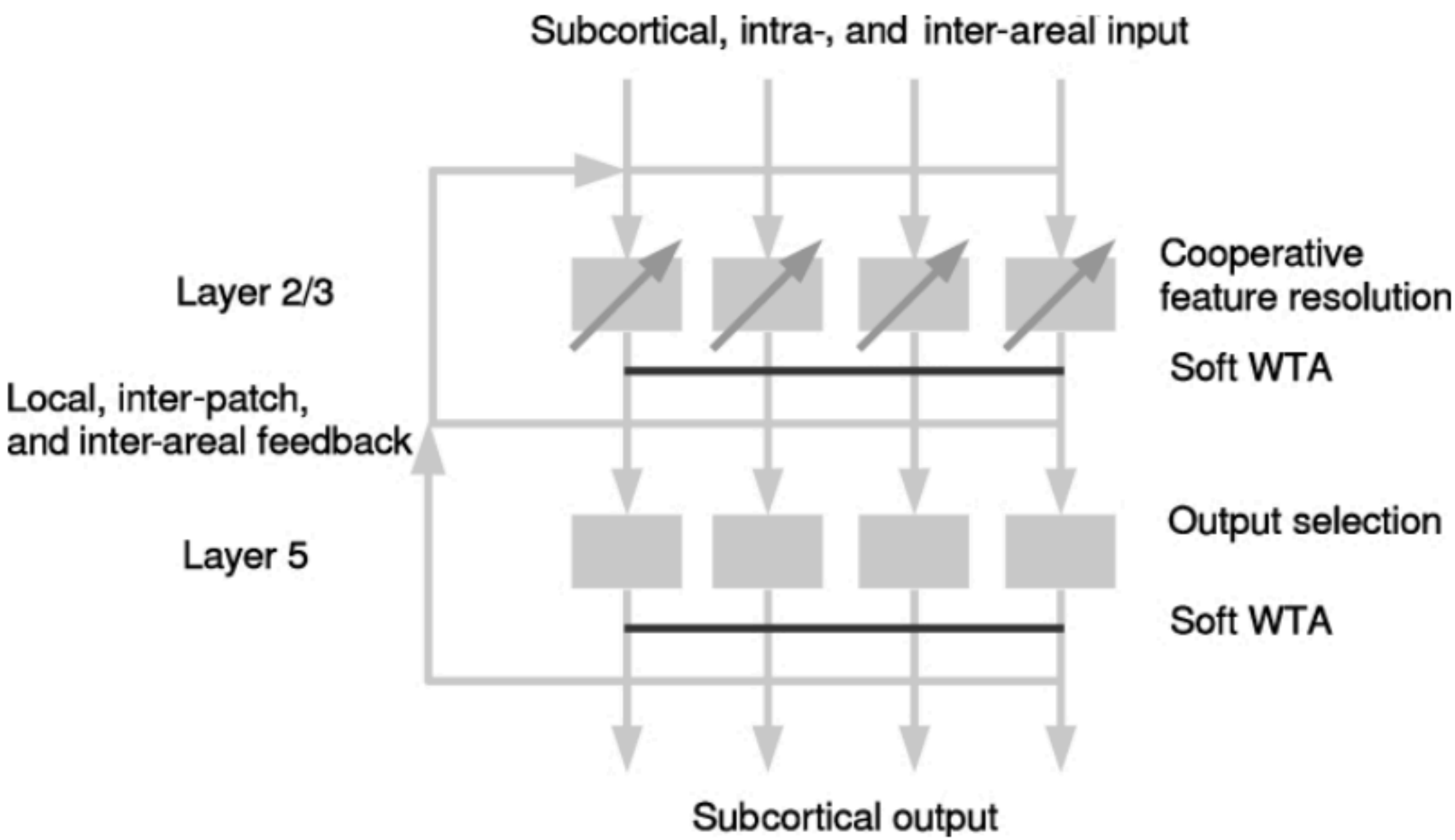



thalamic gating of "copy and paste" operations between cortical working memory buffers, executing a sequence of steps controlled by the basal ganglia

Stewart, Eliasmith et al 2010

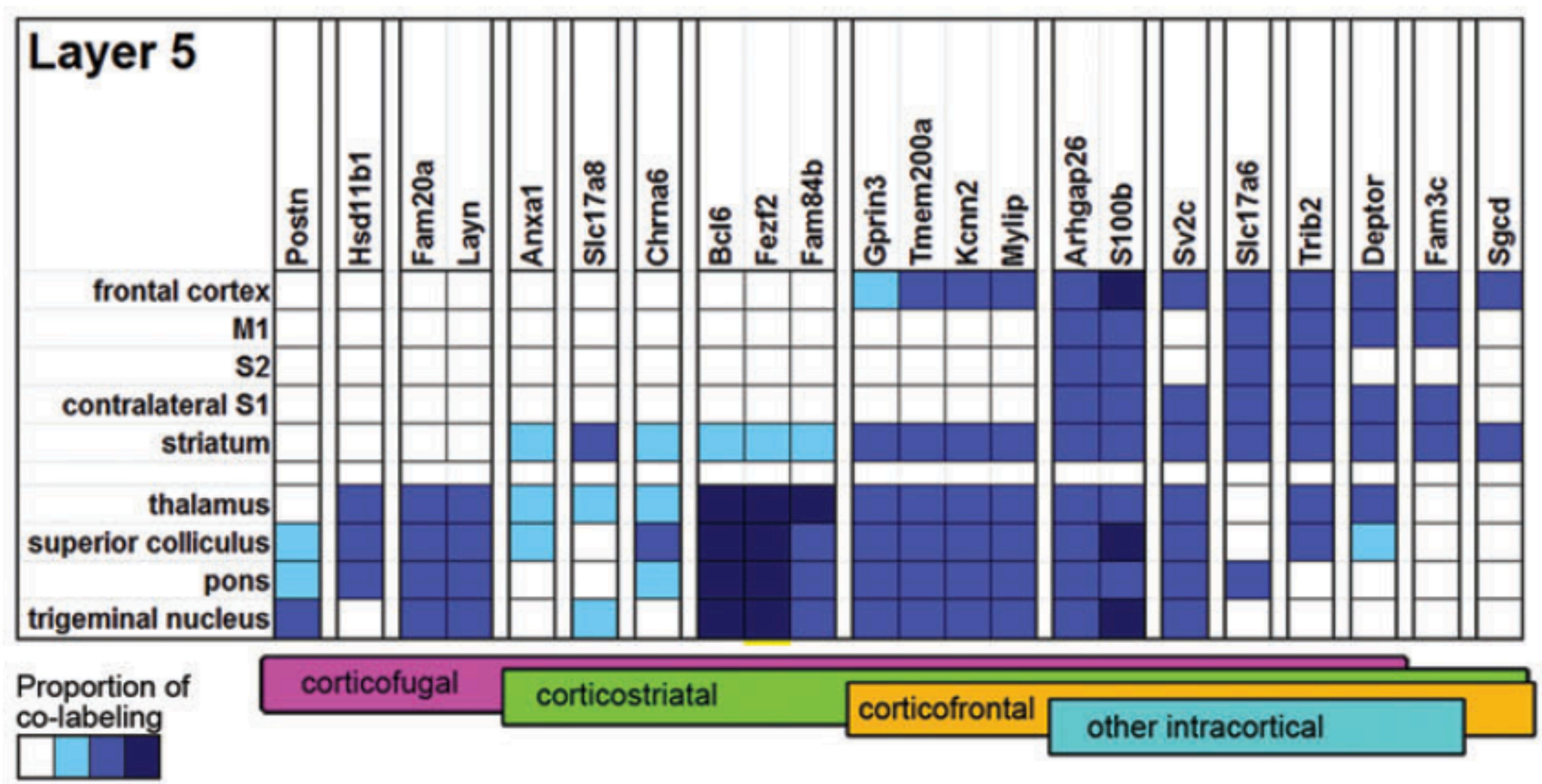
# The dream of a *single* “canonical” cortical circuit

## Likely a gross over-simplification of reality




Douglas & Martin 2004

# Molecularly-defined cortical connectivity




(retrograde labeling + FISH)

Sorensen et al, Cerebral Cortex, 2013

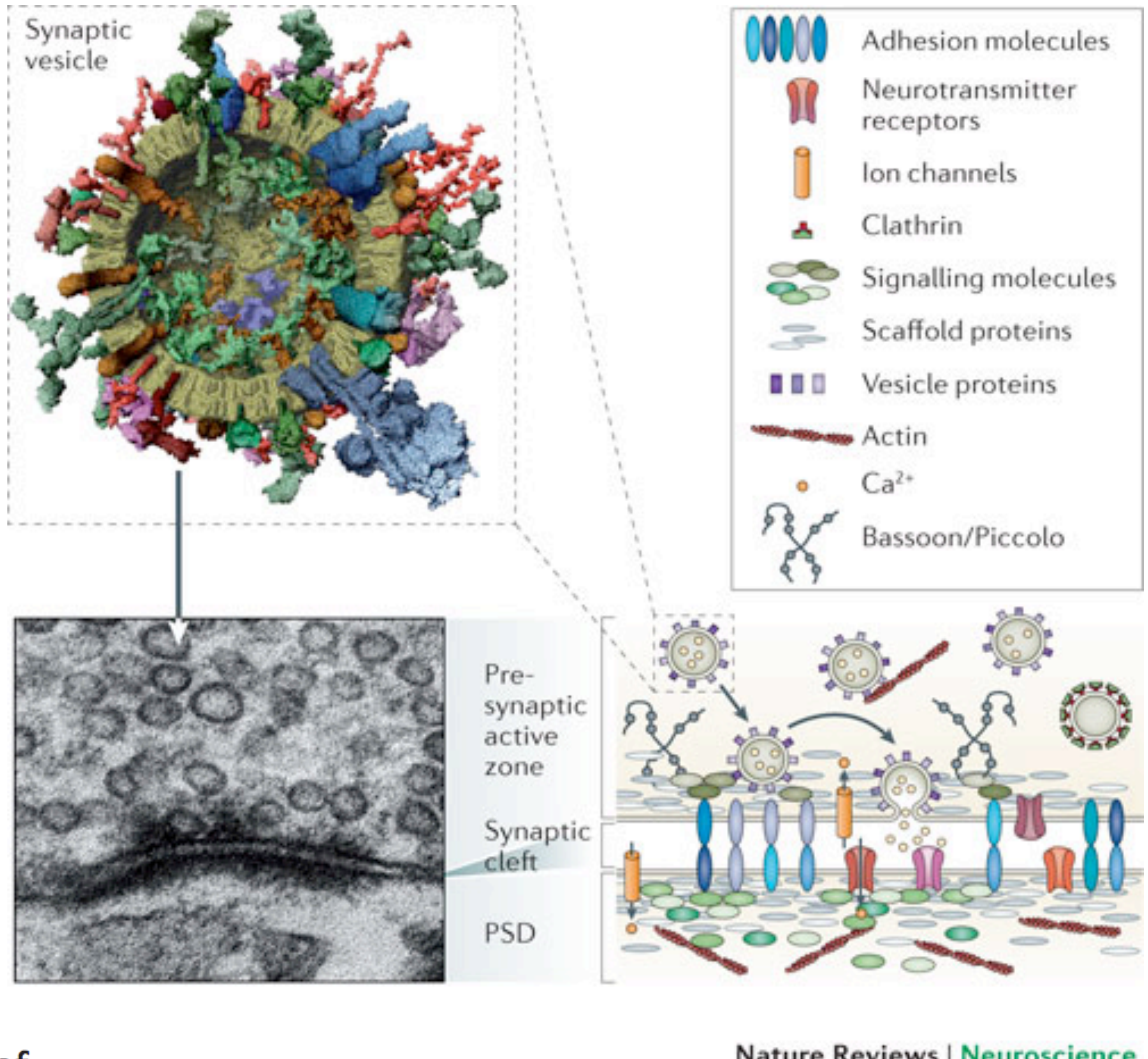
# Emerging picture: *many* computations, not one

| Computation | Algorithmic/ representational realization | Neural implementation(s) | Brain location(s) |
|---|---|---|---|
| Rapid perceptual classification | Receptive fields, pooling and local contrast normalization [58,61] | Hierarchies of simple and complex cells [62] | Visual system |
| Complex spatiotemporal pattern recognition | Bayesian belief propagation [19,63] | Feedforward and feedback pathways in cortical hierarchy [19] | Sensory hierarchies |
| Learning efficient coding of inputs | Sparse coding [64] | Thresholding and local competition [65] | Sensory and other systems |
| Working memory | Continuous or discrete attractor states in networks [66,67] | Persistent activity in recurrent networks [68] | Prefrontal cortex |
| Decision making | Temporal-difference reinforcement learning algorithms [69,70]; actor-critic models [71] | Cortically implemented Bayesian inference networks combined with temporal difference reinforcement learning via the dopamine system and action selection systems in the basal ganglia [72] | Prefrontal cortex |
| | Winner-take-all networks [73] | Recurrent networks coupled via lateral inhibition [73] | Prefrontal cortex |
| Gating of information flow | Context-dependent tuning of activity in recurrent network dynamics [74] | Recurrent neural networks implementing line attractors and selection vectors [74] | Prefrontal cortex |
| | Shifter circuits [64,75] | Divergent excitatory relays and input-selective shunting inhibition in dendrites [76] | Visual system |
| Gain control | Divisive normalization [35] | Shunting inhibition in networks or balanced background synaptic excitation and inhibition [77] | Common across many cortical areas |
| Sequencing of events over time [78] | Feed-forward cascades; Serial working memories [79] | Synfire chains [80-82]; Thalamo-cortico-striatal loops [83,84] | Common across many cortical areas |
| Representation and transformation of variables | Population coding [85] | Time-varying firing rates of cosine-tuned neurons representing dot products with encoding vectors | Motor cortex |
| Variable binding | Holographic reduced representations [56,86] | Circular convolution of vectors represented by neural population codes | Cortical areas involved in sequential or symbolic processing |
| | Dynamic binding [87,88] | Neural synchronization [89] | |

Marcus, Marblestone, Dean, *in prep*.

# Multi-level complexity

*Example: hundreds of proteins in each synaptic vesicle*




Deep molecular diversity of mammalian synapses: why it matters and how to measure it

*Nancy A. O'Rourke, Nicholas C. Weiler, Kristina D. Micheva and Stephen J. Smith*

# Multi-level complexity

*Example: the physical basis of long-term memory is still uncertain*

**Very long-term memories may be stored in the pattern of holes in the perineuronal net** (i.e., in the extracellular matrix)

**Roger Y. Tsien**[1]

Department of Pharmacology, Department of Chemistry and Biochemistry, and Howard Hughes Medical Institute, University of California at San Diego, La Jolla, CA 92093-0647

Contributed by Roger Y. Tsien, June 3, 2013 (sent for review May 16, 2013)

(vs. bistable auto-phosphorylating kinases at the synapse)

# How to cope with the brain's complexity?

**A**ctivity history
**B**ehavior
**C**onnectivity (circuit diagram)
**D**evelopment (cell lineage tree)
**E**xpression (epigenetic cell types,
+ single-synapse proteomes)

**Need:** a technology to cheaply / rapidly measure all these variables in a single brain.

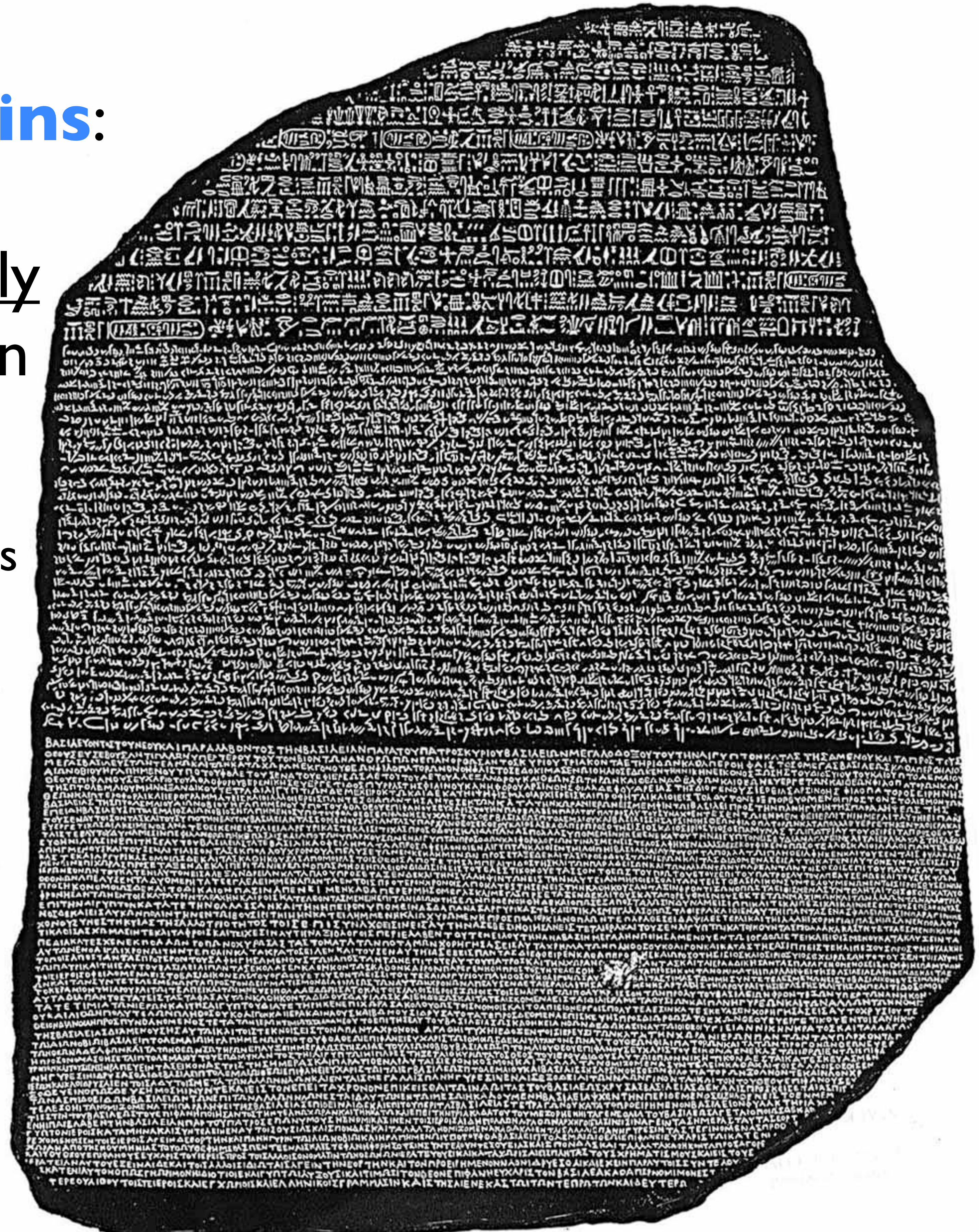
**Need:** (sub)cellular resolution + whole brain scope.

see upcoming essay by Church, Marblestone & Kalhor

# How to cope with the brain's complexity?

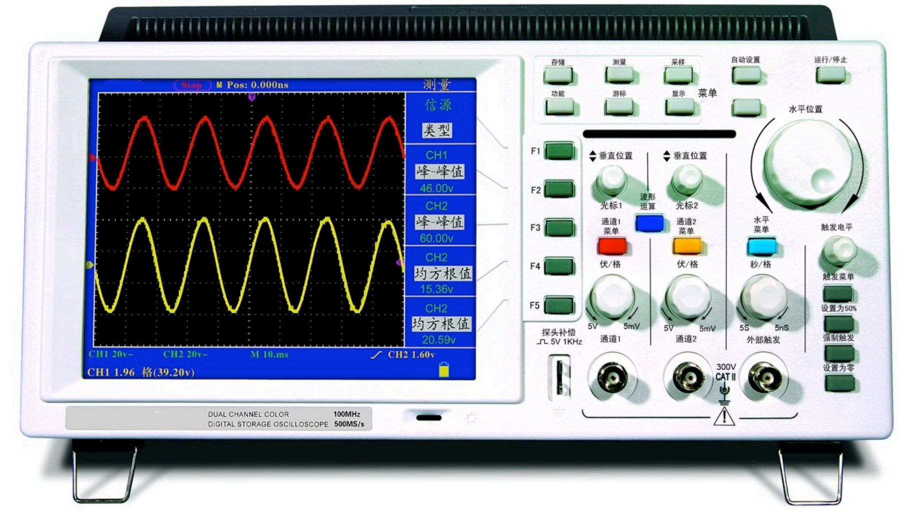
**Rosetta Brains**:
A, B, C, D & E
<u>comprehensively</u>
on a single brain

allows correlations
across variables/levels



**Q for thesis:**

What would such a technology conceivably look like?



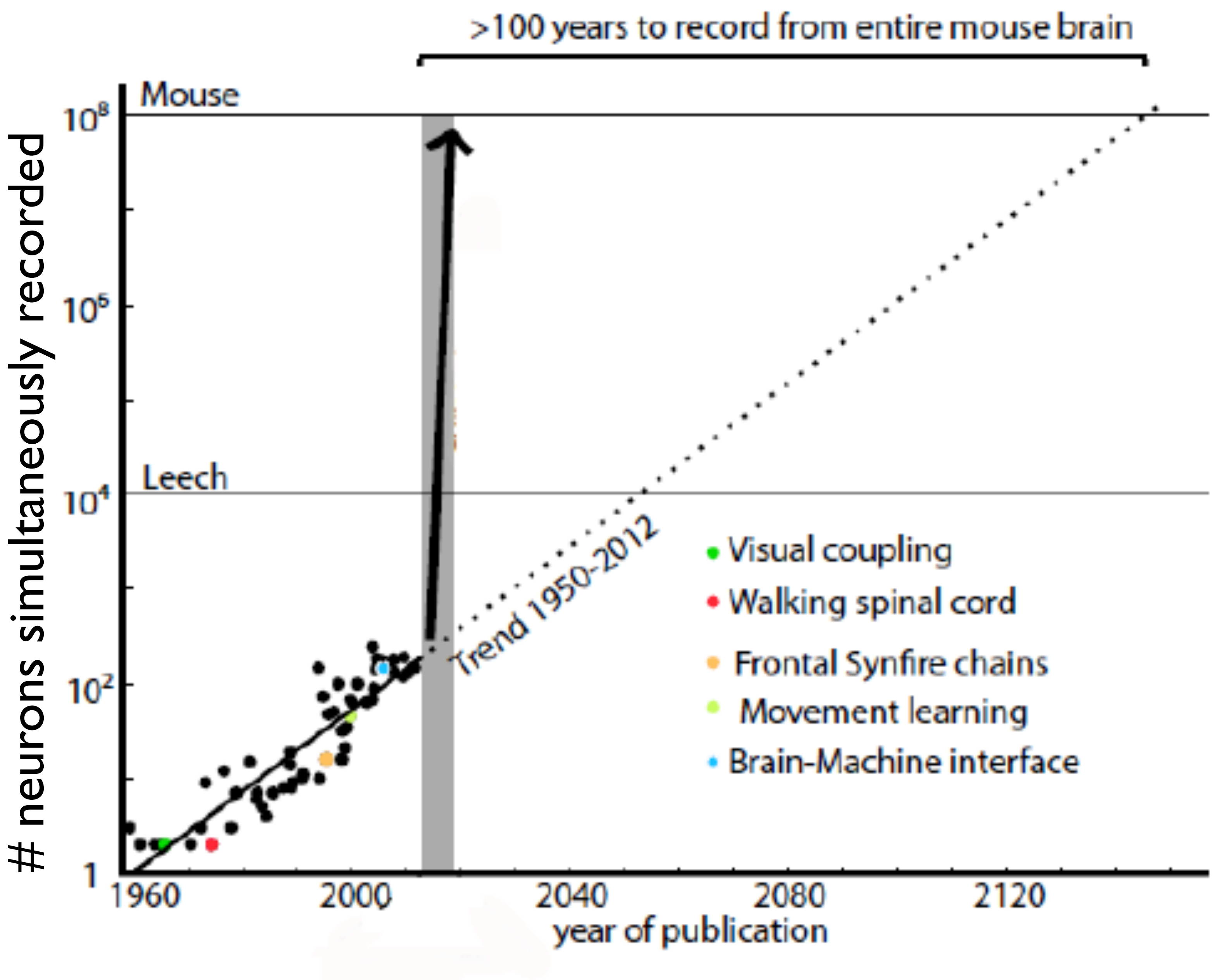
**Activity history**
**B**ehavior
**C**onnectivity (circuit diagram)
**D**evelopment (cell lineage tree)
**E**xpression (epigenetic cell types,
+ single-synapse proteomes)

We want a technology which can cheaply / easily measure all these variables in a single brain.

Need (sub)cellular resolution + whole brain scope.

see upcoming essay by Church, Marblestone & Kalhor

# “every spike from every neuron” in a mammalian brain


Mouse brain:

- 100M neurons
- >100B connections
- 1-1000 Hz firing rates
- 0.5 cc volume


Konrad Kording 2013; Alivasatos et al 2012

**nobody has <u>written down a design</u> that clearly**

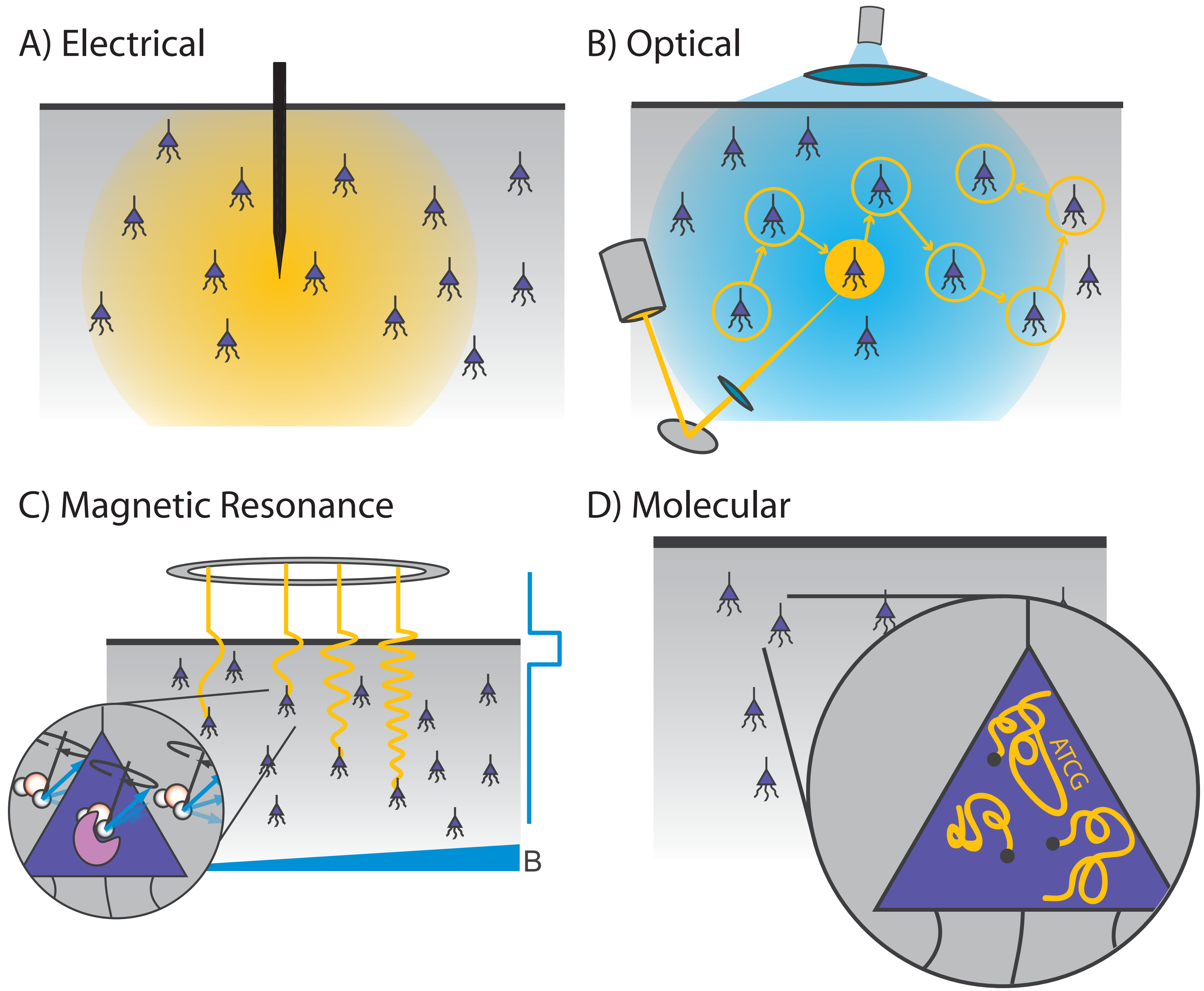
**solves the problem**
**does not violate any laws of physics**
**does not severely damage the brain**

# **Physics constraints** on brain activity mapping

< 2C temperature change:

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

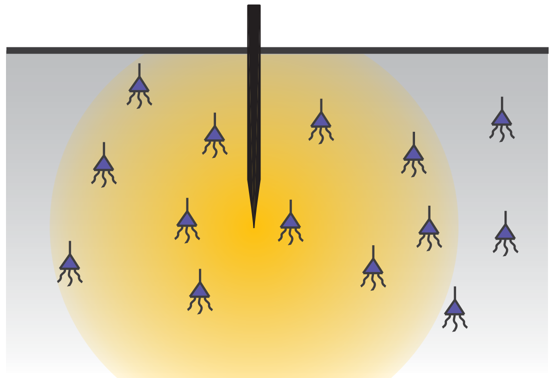
< 1% tissue volume displacement



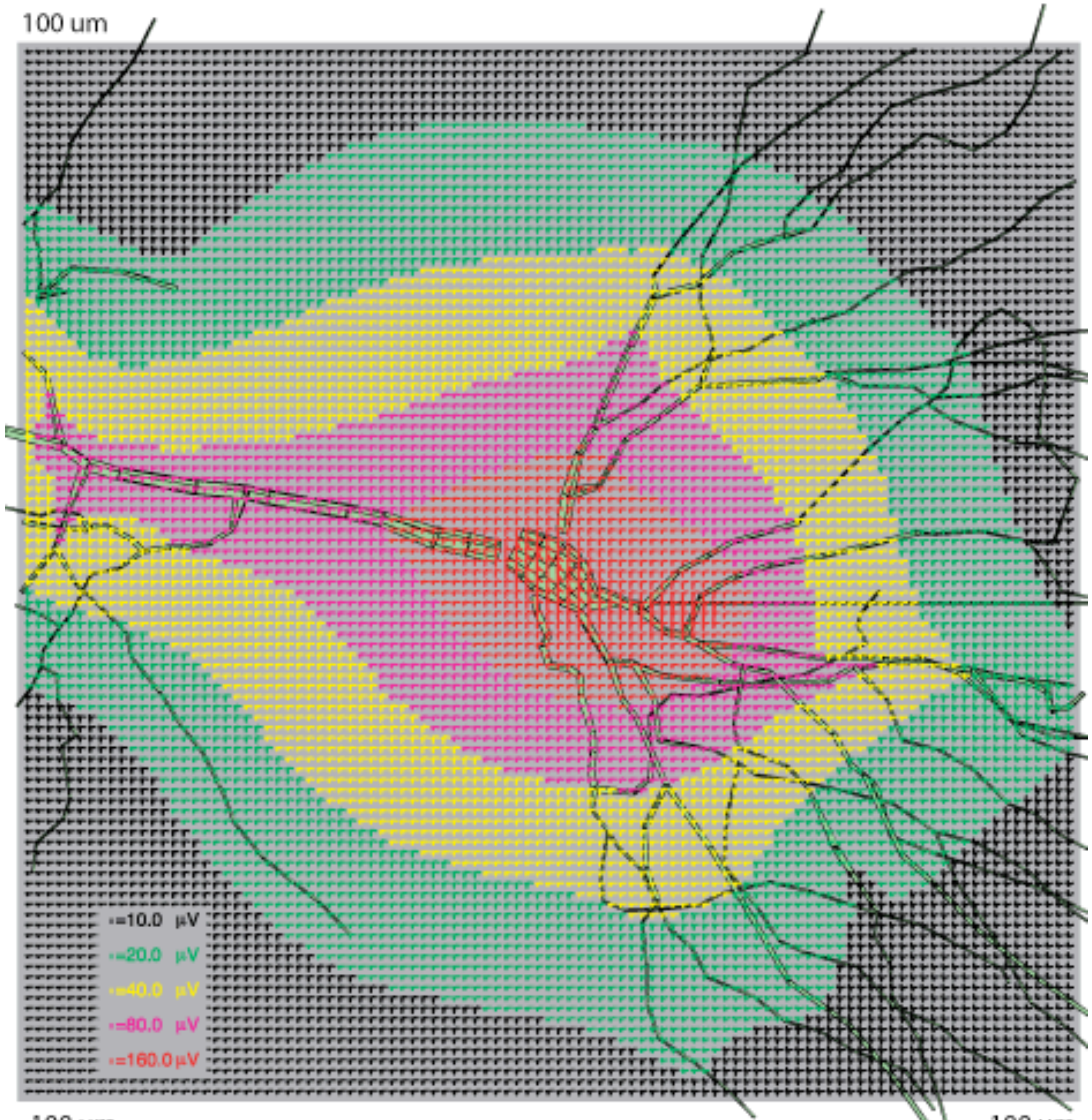

Marblestone et al, Front. Comp. Neurosci (2013): **with >17 co-authors**

http://arxiv.org/abs/1306.5709

# **Physics constraints** on brain activity mapping



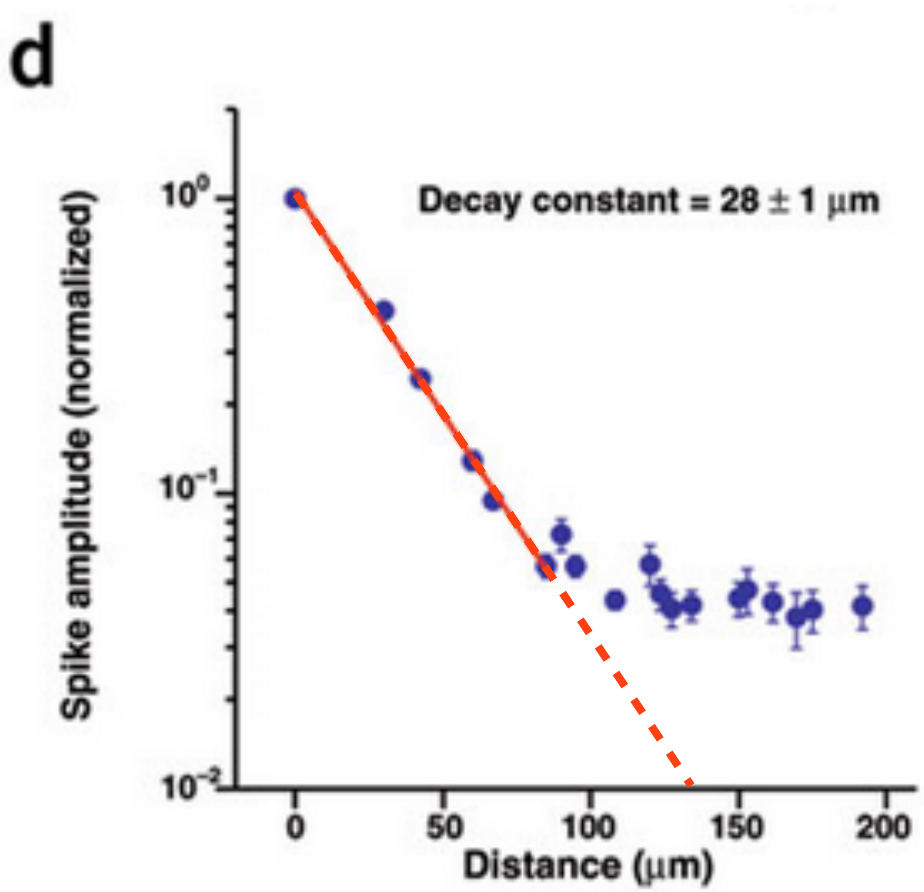
**100-200 um max recording radius**




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

[Neuron/Matlab simulation based on Gold and Koch]

# **Physics constraints** on brain activity mapping




$r_{max}$ **~ 130 um**

**signal size ~ 100-1000 uV**

**noise floor ~ 10-20 uV**

**SNR <100**

$V_{johnson} = [4k_bTR(BW)]^{1/2}$ ~ 9 uV

amplifier noise ~ 1 uV

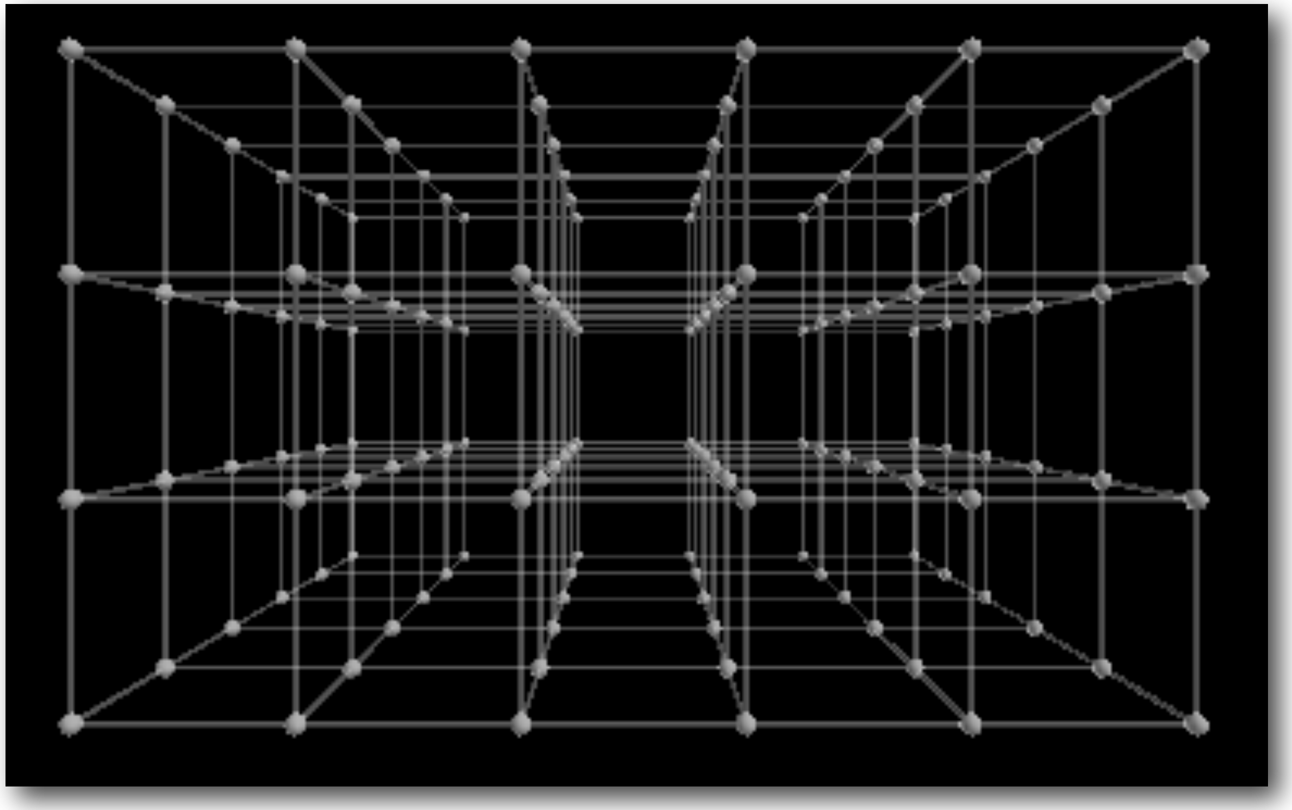
noise from other neurons

data: **Berry lab**, Princeton

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

# **Physics constraints** on brain activity mapping

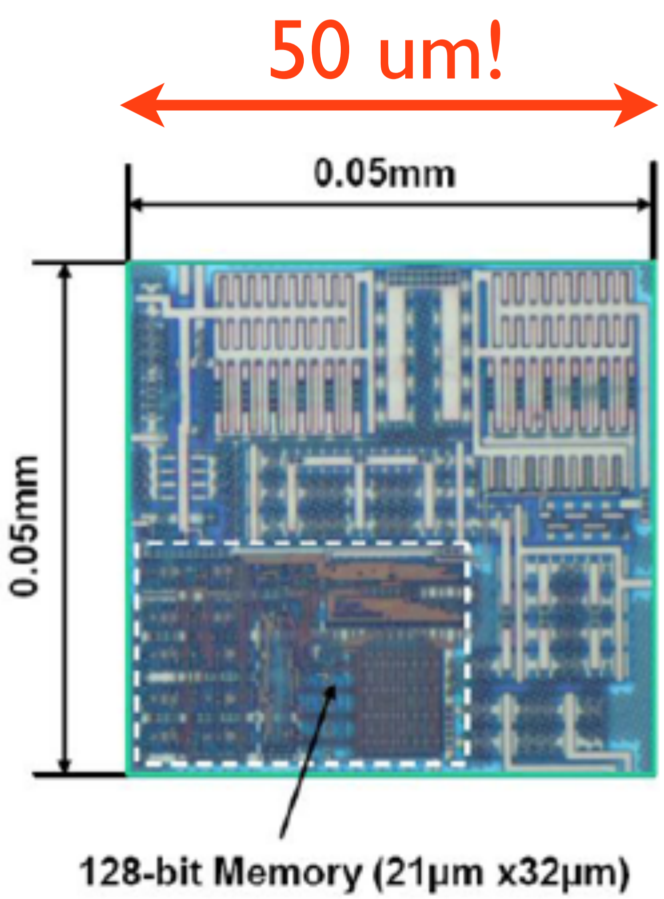
**750k -7.5M electrodes for 75M mouse neurons**



current spike sorting algorithms: < 10 neurons / electrode
information theory estimate: < 100 neurons / electrode
estimate from max recording radius: < 1000 neurons / electrode

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

# **Physics constraints** on brain activity mapping




**wirelessly-powered RFID chip**

> 10 pJ per transmitted bit

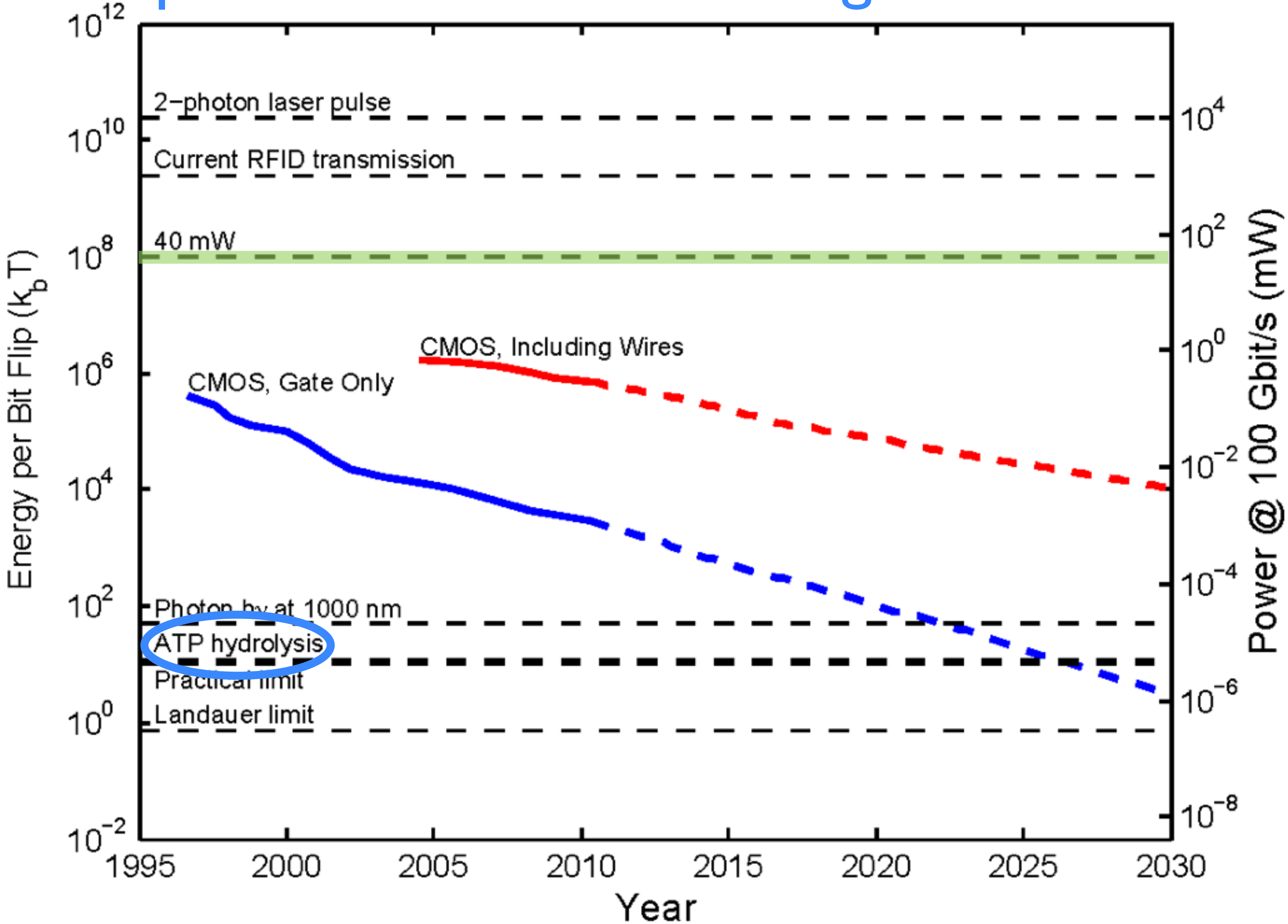
x 100 Gbit/sec = **1W**

need at **2-3 orders of magnitude** improvement
in power efficiency of electronics
to use embedded chips for whole-brain recording

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

# **Physics constraints** on brain activity mapping

## power limits on switching elements

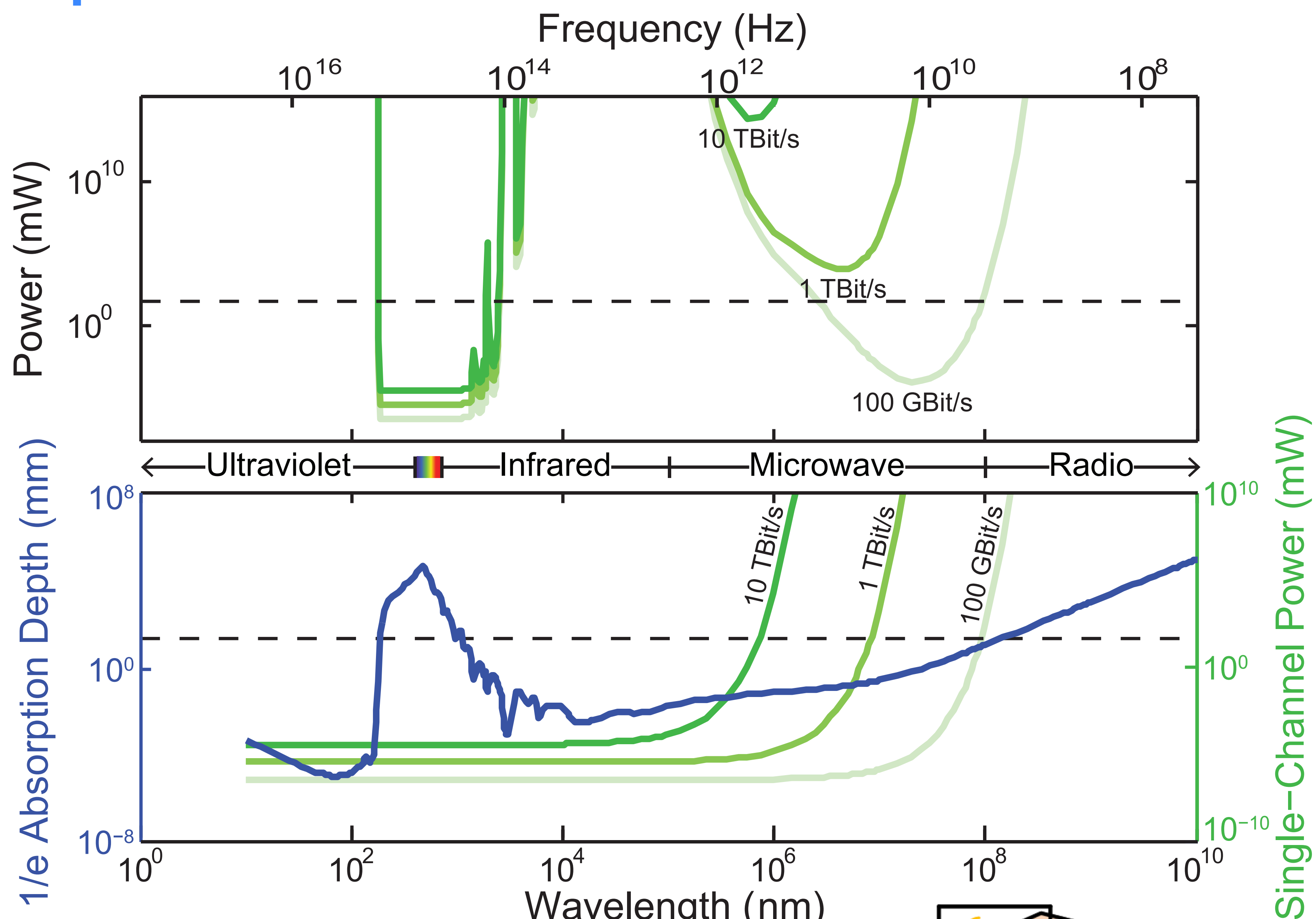


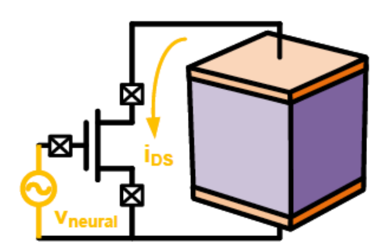

[CMOS data from Rod Tucker]

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

# **Physics constraints** on brain activity mapping

## power limits on data transmission




Seo et al

<u>Workarounds</u>:

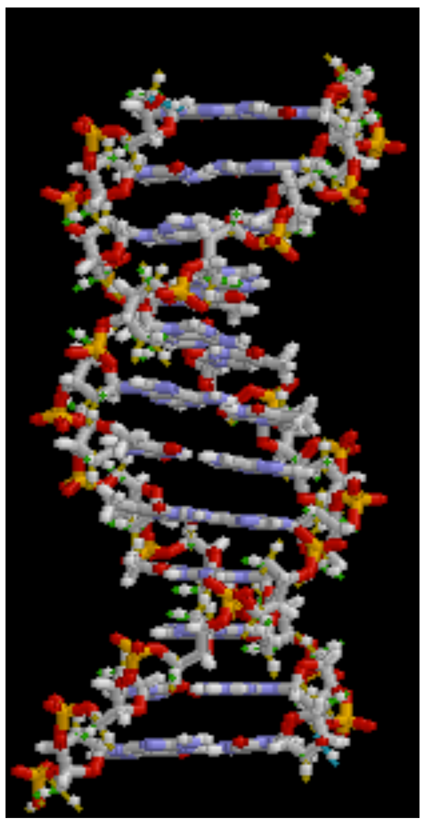
store data locally and read out slowly

use ultrasound or infrared/visible light rather than RF / microwave transmission

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

# Back to the drawing board

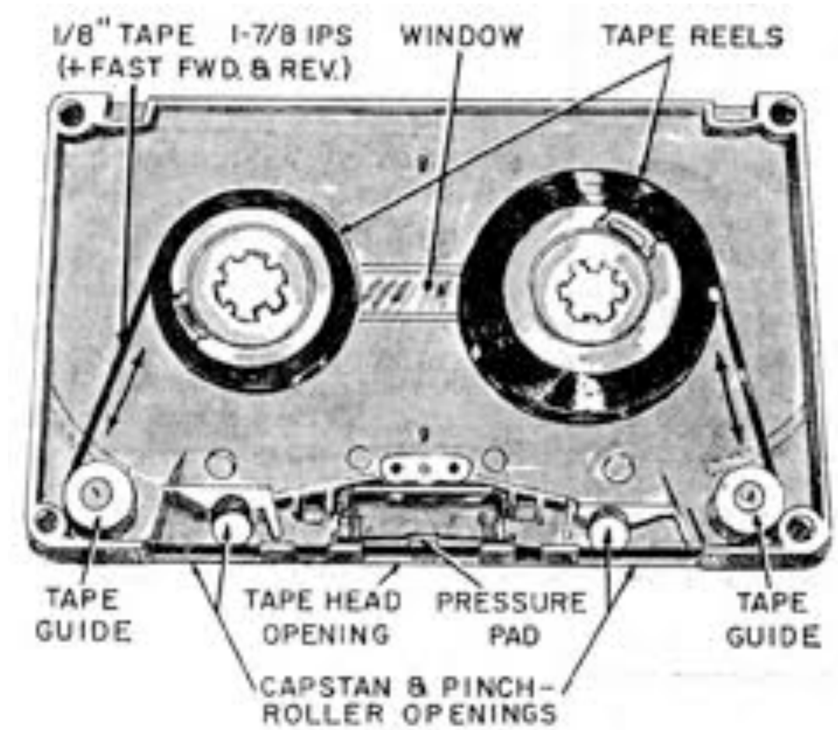
# What if each cell could record its own activity?



=

0101001010101010010101
0101010101001010110111
0010110101010111010011

=

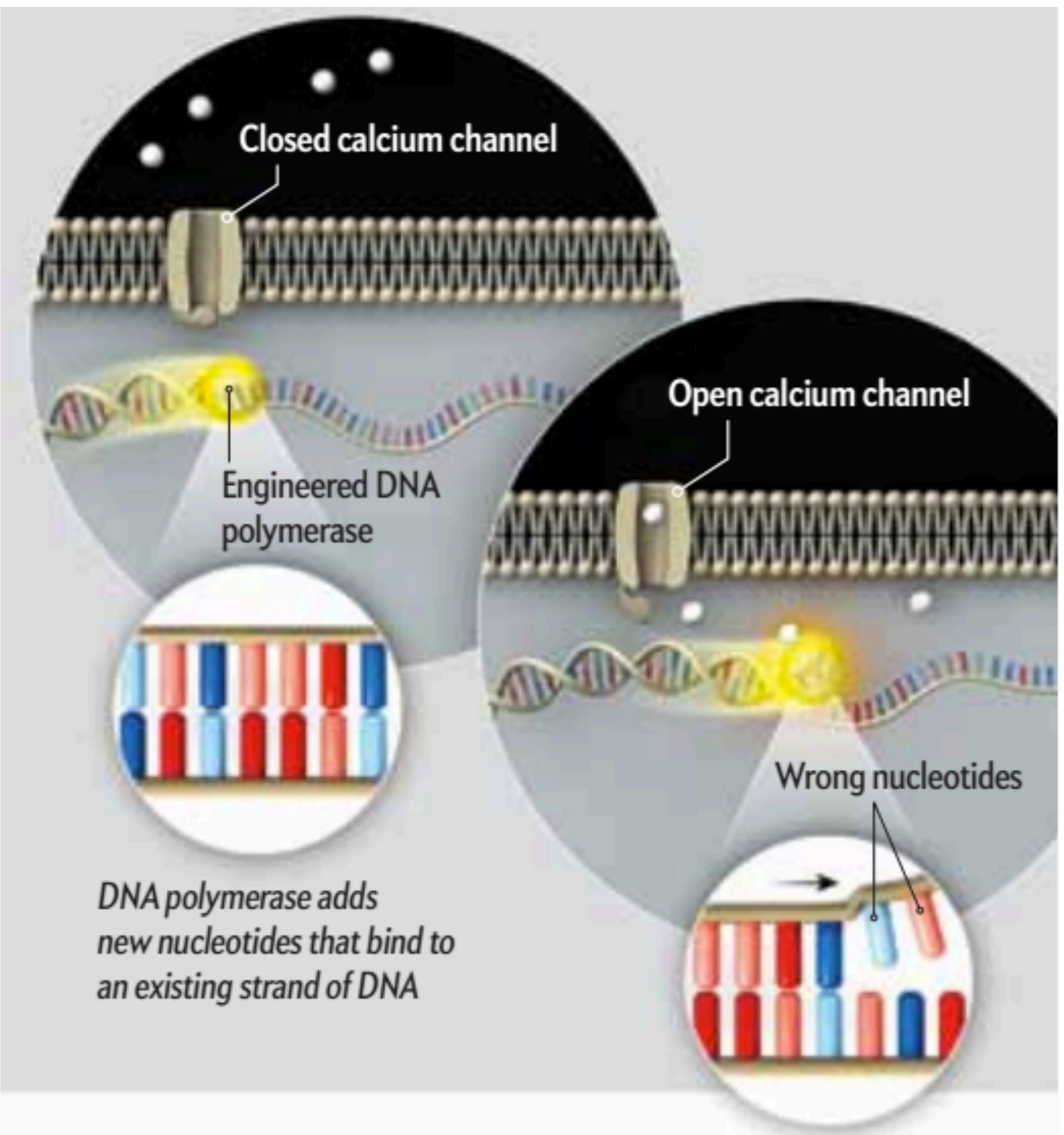


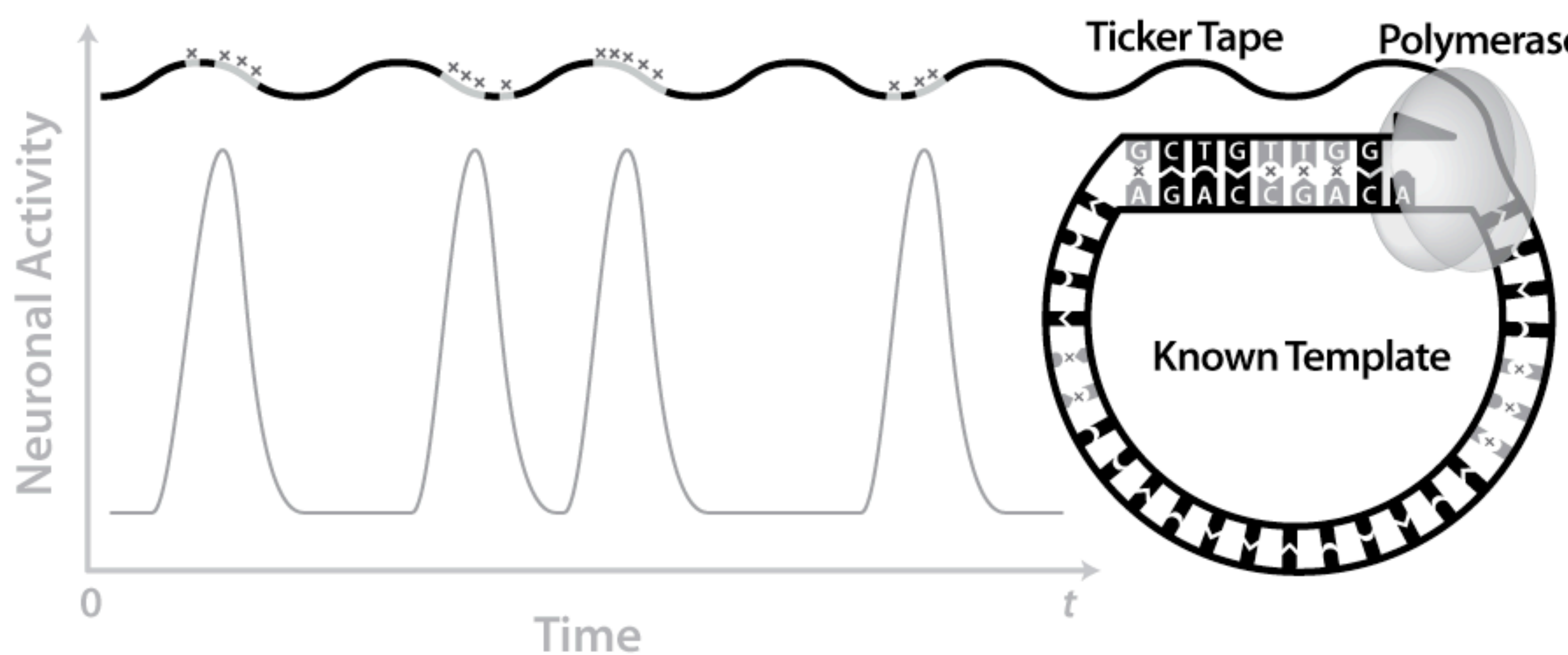

got $9M NIH grant to pursue this

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

# a molecular recording device in each neuron

**Encode information via control of polymerase error rate**

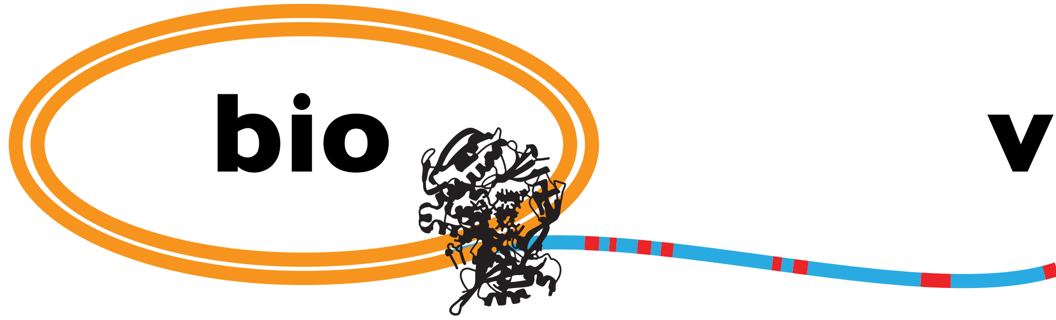
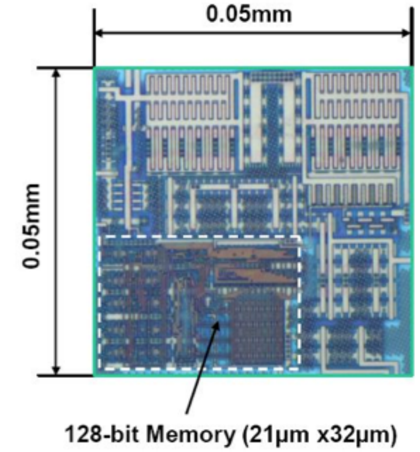



Kording, 2011

Yuste and Church, Scientific American, 2014

Figure by Reza Kalhor

**bio**  **vs.** 

**electronic**

**Bandwidth:** (1 bit / 1 ms / neuron) x (1e8 neurons) = **100 gigabit / s**
to read out a <u>mouse</u> brain at 1 bit per ms per neuron in real time

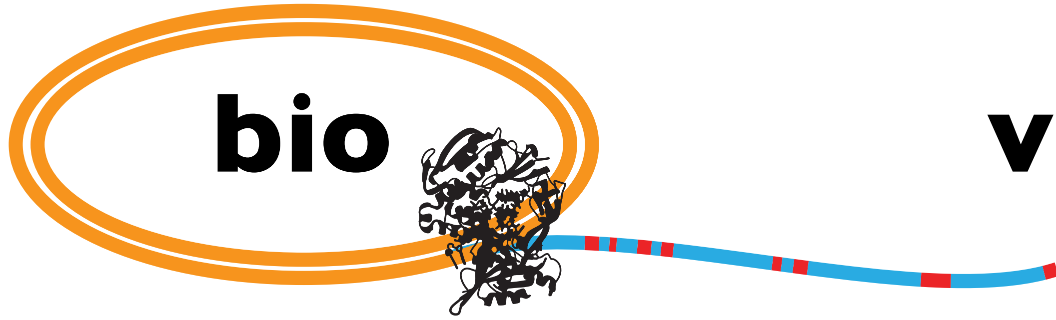
**readout of information <u>stored</u> in biomolecules can be <u>arbitrarily slow</u>**

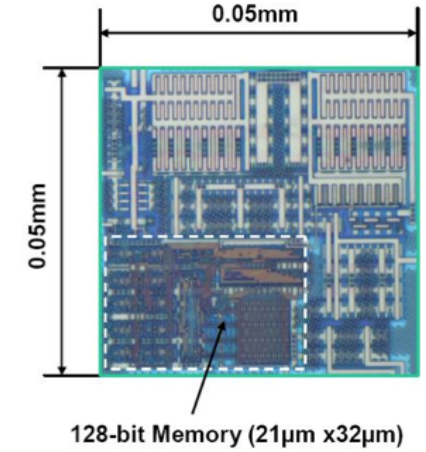
**Power:** CMOS bit switching at 1000 Hz consumes more energy than a neuron

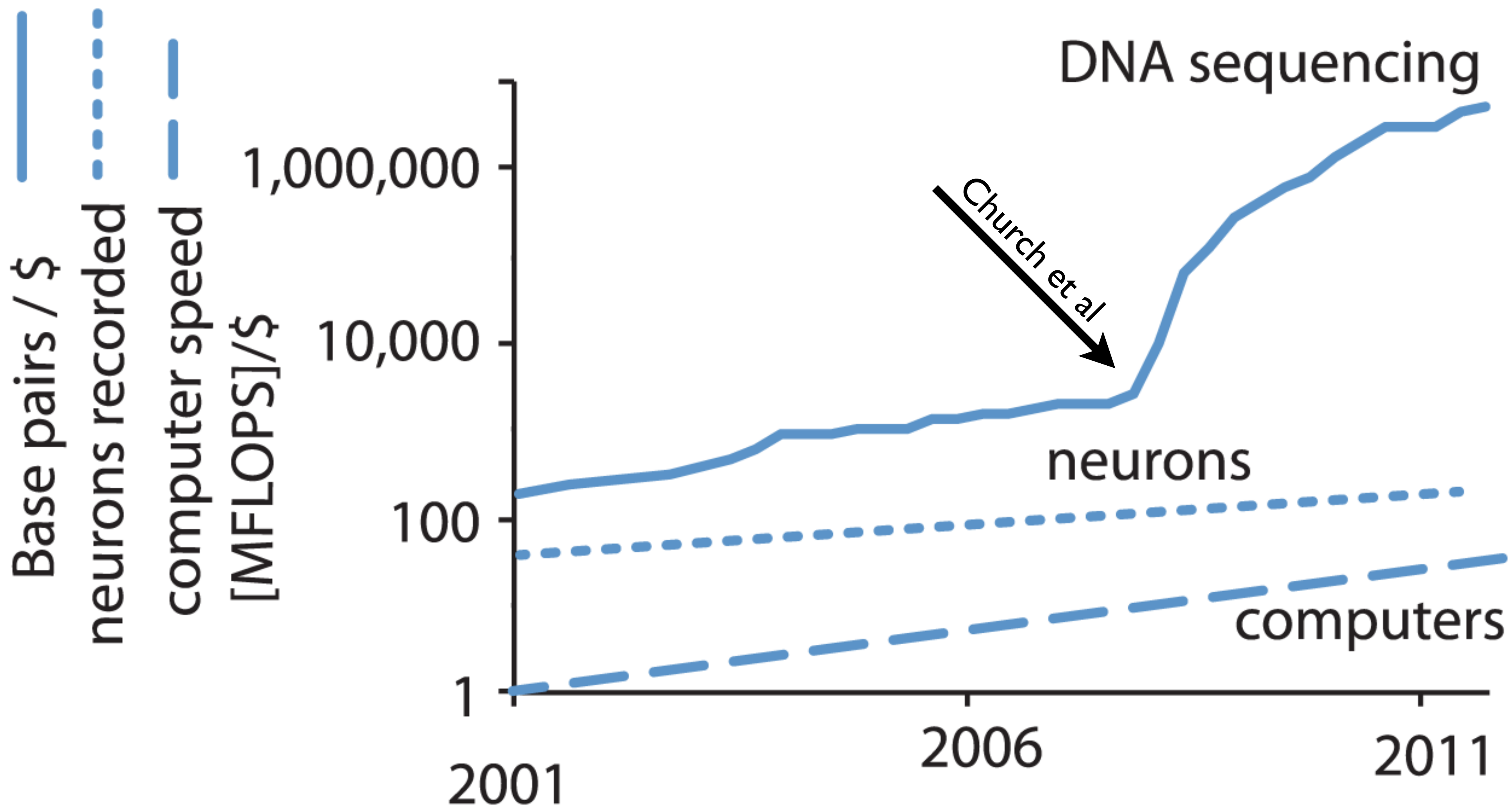
**molecular recorders are 5-6 orders of magnitude more efficient**

| | | |
|---|---|---|
| <u>ultimate</u> limit: | kT x log(2) per bit x 1000 bps | **3e-18 W** |
| <u>practical</u> limit for electronics: | 40 kT per bit x 1000 bps | **1e-16 W** |
| <u>current</u> CMOS: | 1e6 worse than the 40 kT per bit limit | **1e-10 W** |
| | 1fF x (1V)^2 x 1000 Hz | **1e-12 W** |
| 1 <u>neuron</u>: | (1e8 ATP / spike) x 100 Hz | **5e-10 W** |
| | (25W / human brain) x (1 human brain / 1e11 neurons) | **2.5e-10 W** |
| <u>ATP</u> consumption at 2000 Hz: | 2000 Hz x (5e-20J / ATP) | **1e-16 W** |

# bio vs. electronic

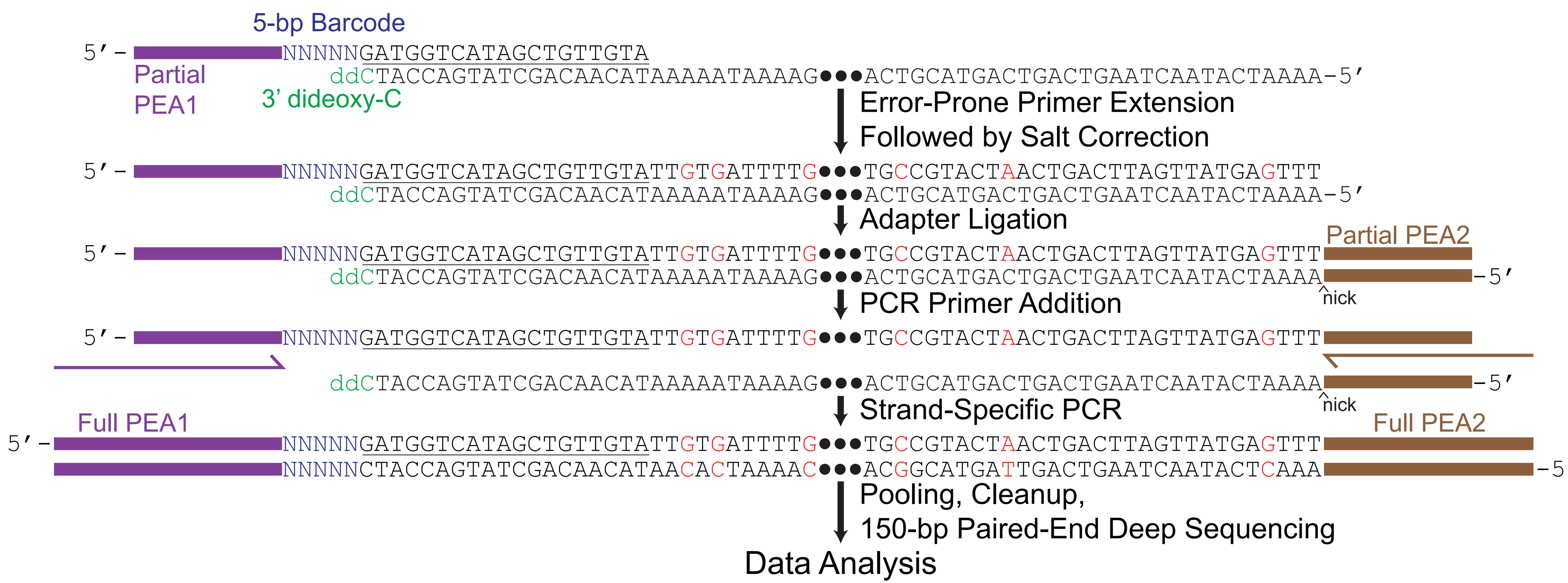





Konrad Kording

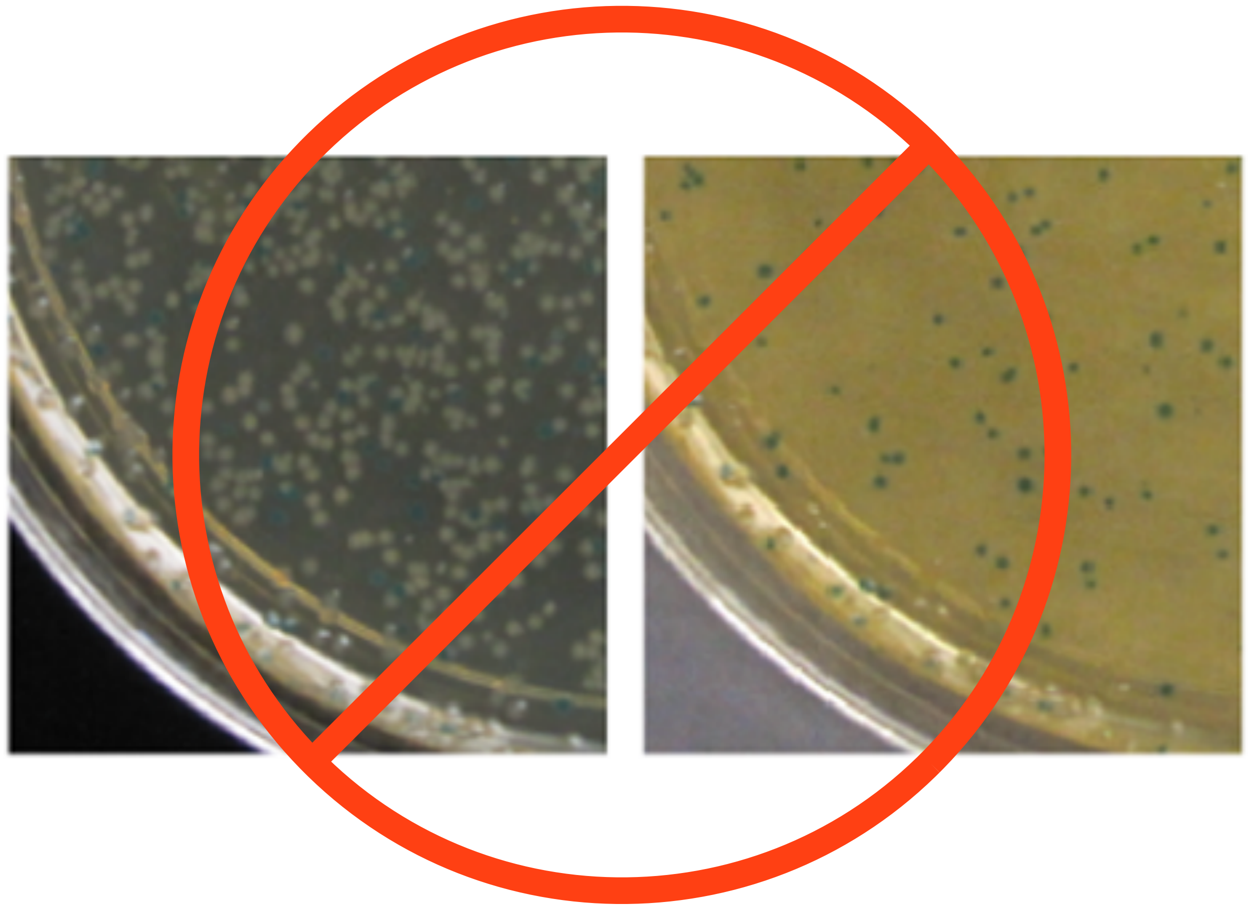
# Sequencing-based screening for ion-dependent DNA pol error rates




biochemistry via sequencing

[Zamft*, Marblestone*, et al. PLoS ONE (2012)]

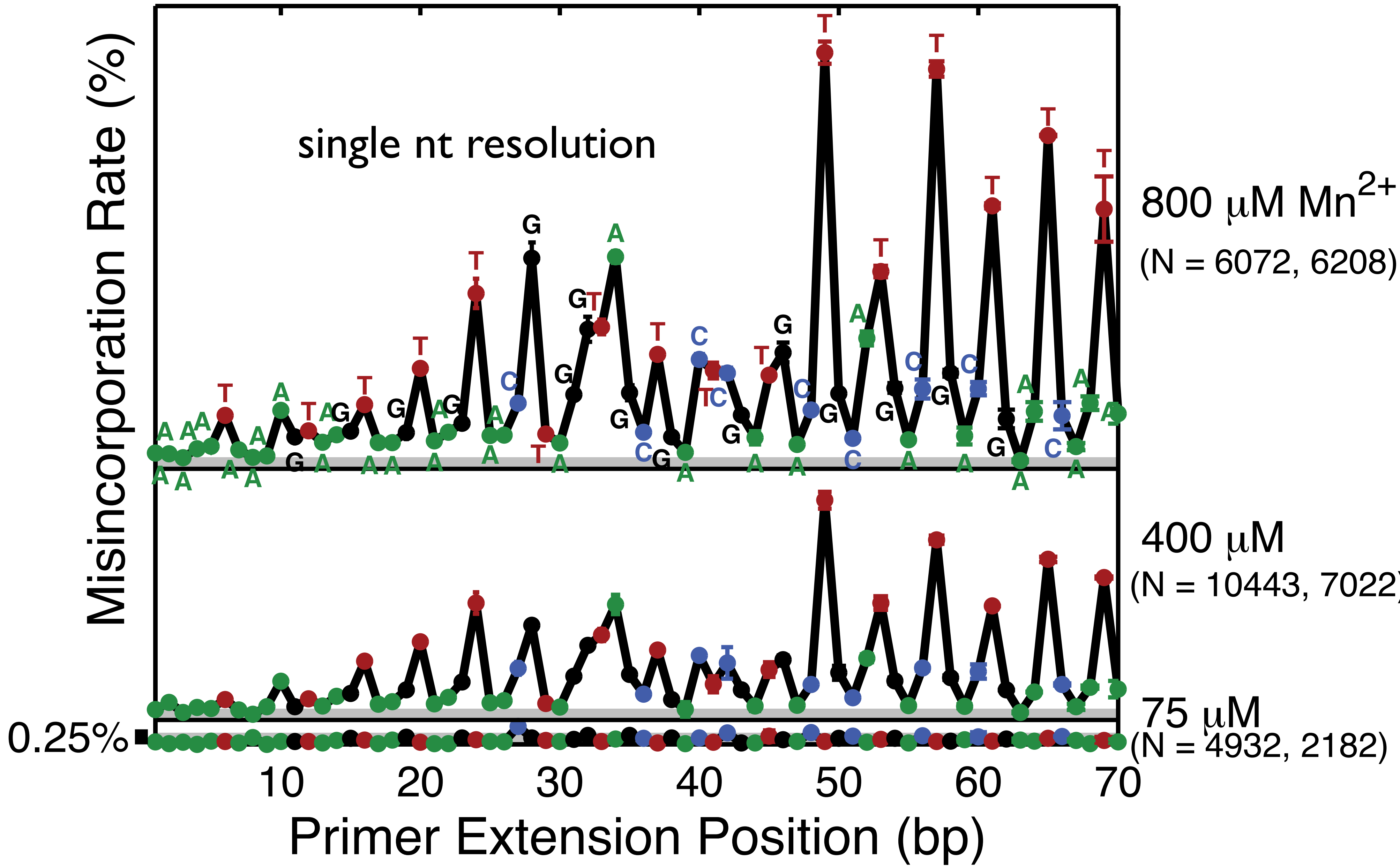
# Sequencing-based screening for ion-dependent DNA pol error rates



biochemistry via sequencing

[Zamft*, Marblestone*, et al. PLoS ONE (2012)]

# Sequencing-based screening for ion-dependent DNA pol error rates




[Zamft*, Marblestone*, et al. PLoS ONE (2012)]

# Iota pol: gets “A” wrong 70% of the time

Example of a good read aligned with revcomp(template):

forward primer
binding site

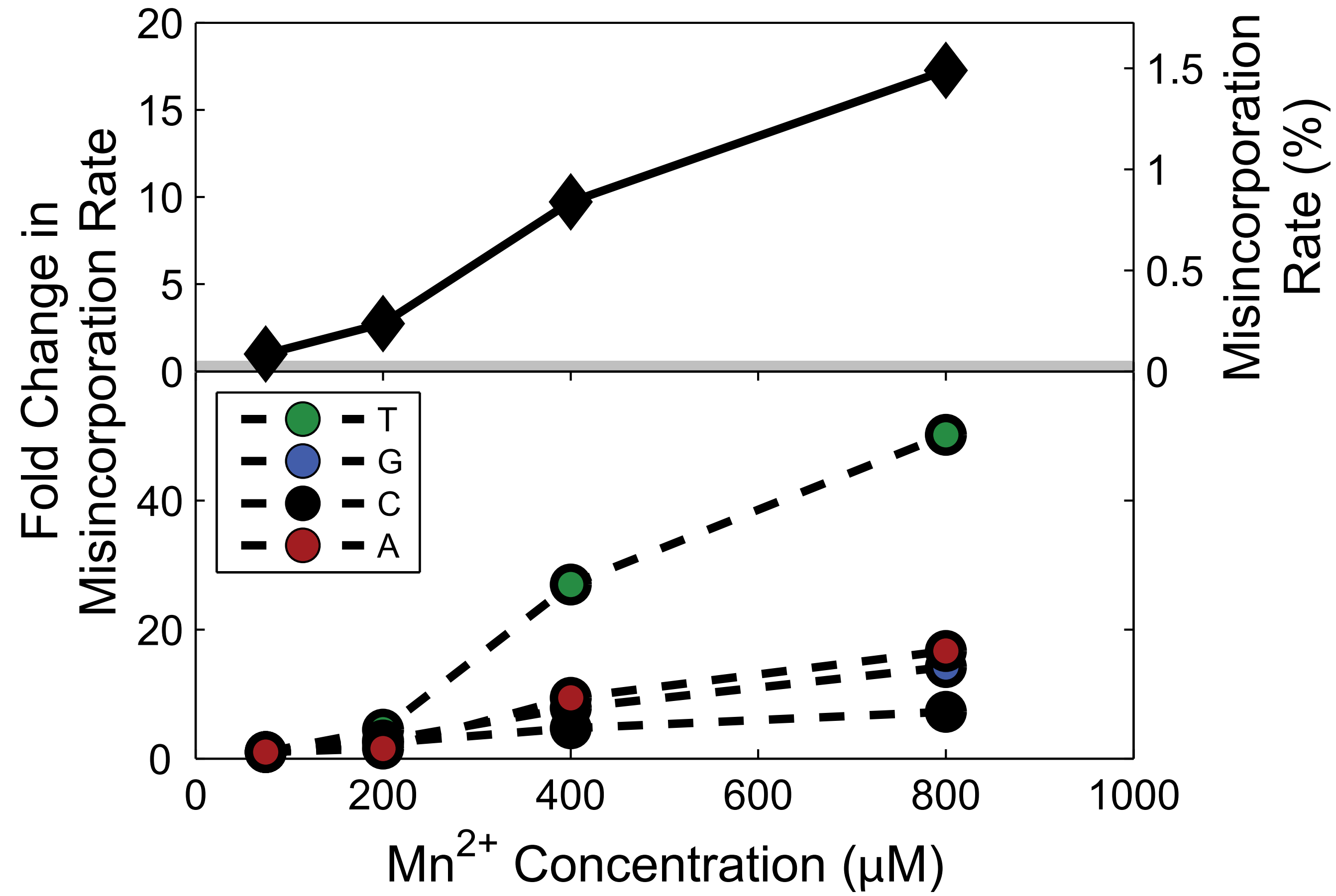
deletion

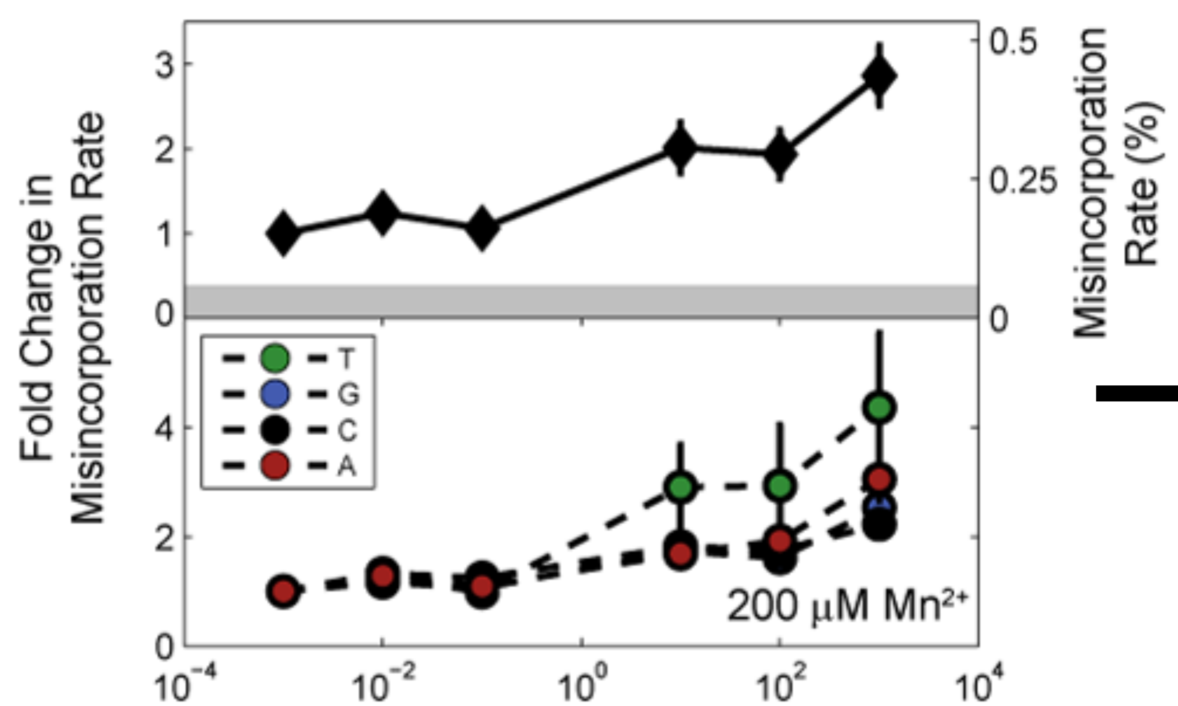
TTTTTGATGGTCATAGCTGTTGTATTTTTGTTTTCATTCTTTCGTTTGTTGCTTCCTTCGGCTGTGCTGCTGGCATGCTGATTGGCTT GTTTTGATTT

GATGGTCATAGCTGTTGTATTTTTA TTTTCATTCATTCATTCA TTGCATCCATCGACTGAGCT ACTGACGTA CTGACTGACTTAGTTATGATTT

with Brad Zamft

# DNA pol offerings from Nature

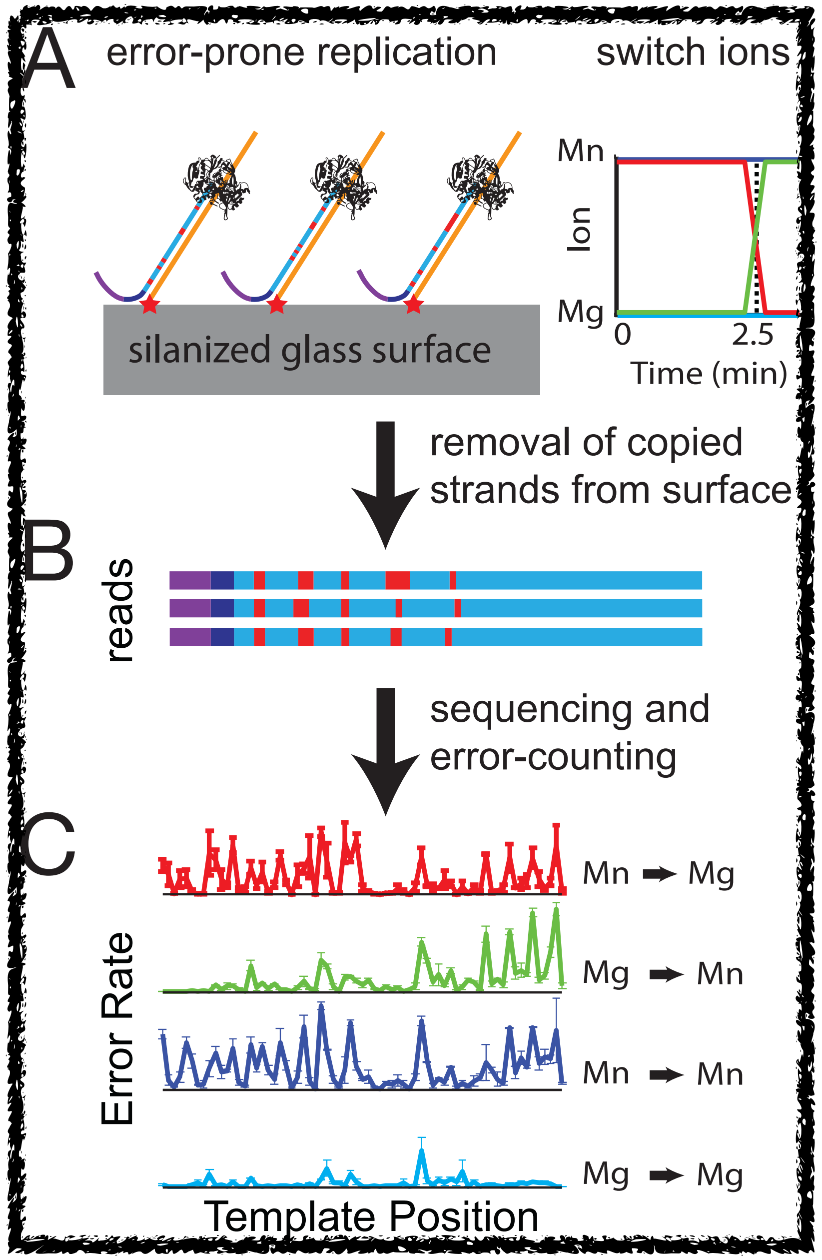


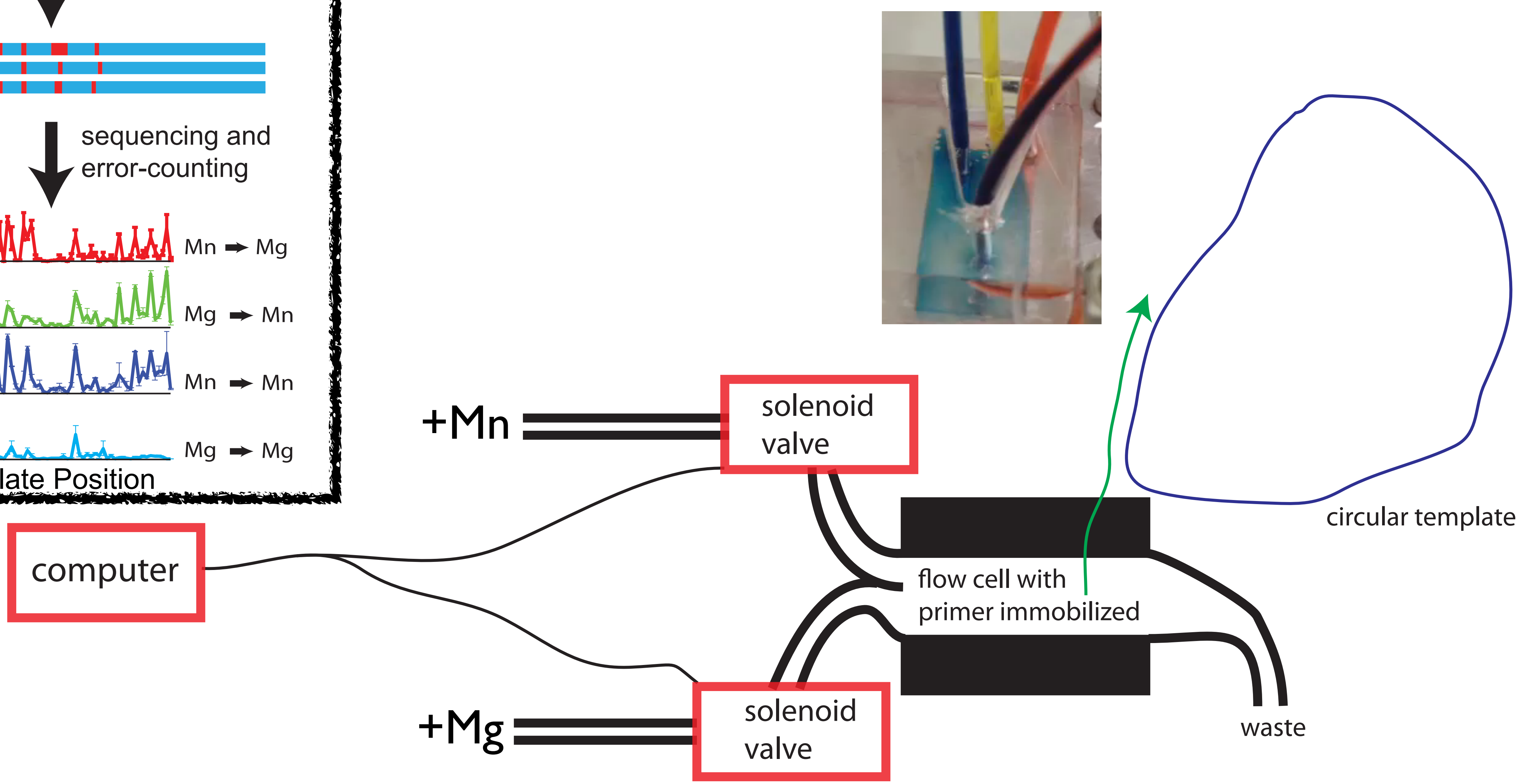




needs custom pol engineering

[Zamft*, Marblestone*, et al. PLoS ONE (2012)]

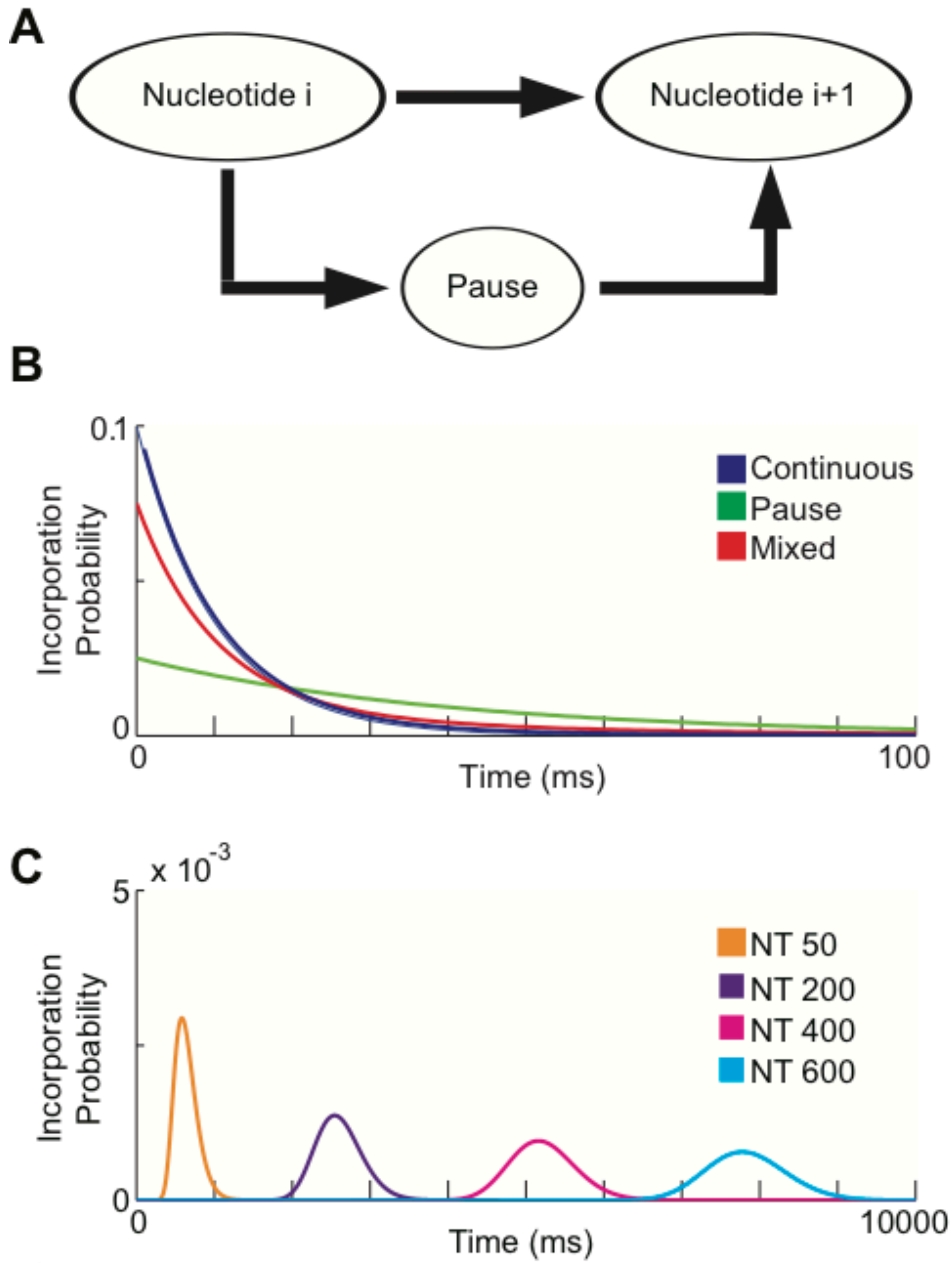
# Encoding time-series into DNA




with Brad Zamft and Noah Donoghue; time-coding results are **preliminary**

# Theoretical temporal resolution of a DNAP ensemble




$$p(\Delta t)_{2-paths} = P\frac{e^{-\Delta t/\tau_p}}{\tau_p} + (1-P)\frac{e^{-\Delta t/\tau_C}}{\tau_C}$$

even with no "paused state", dwell times are stochastic

nucleotide ⟷ time mapping smears over time

[Glaser, Zamft*, Marblestone*, et al. PLoS Computational Biology (2013)]

# Theoretical temporal resolution of a DNAP ensemble

**In the absence of a synchronization / clock mechanism:**

**<10 ms** temporal resolution does not persist for > seconds, even with >1000 simultaneously-replicated templates per cell

unless a faster-than-natural polymerase can be engineered (while maintaining other favorable properties)

**~100 ms** temporal resolution is feasible w/ optimal combination of individual naturally-occurring polymerase parameters

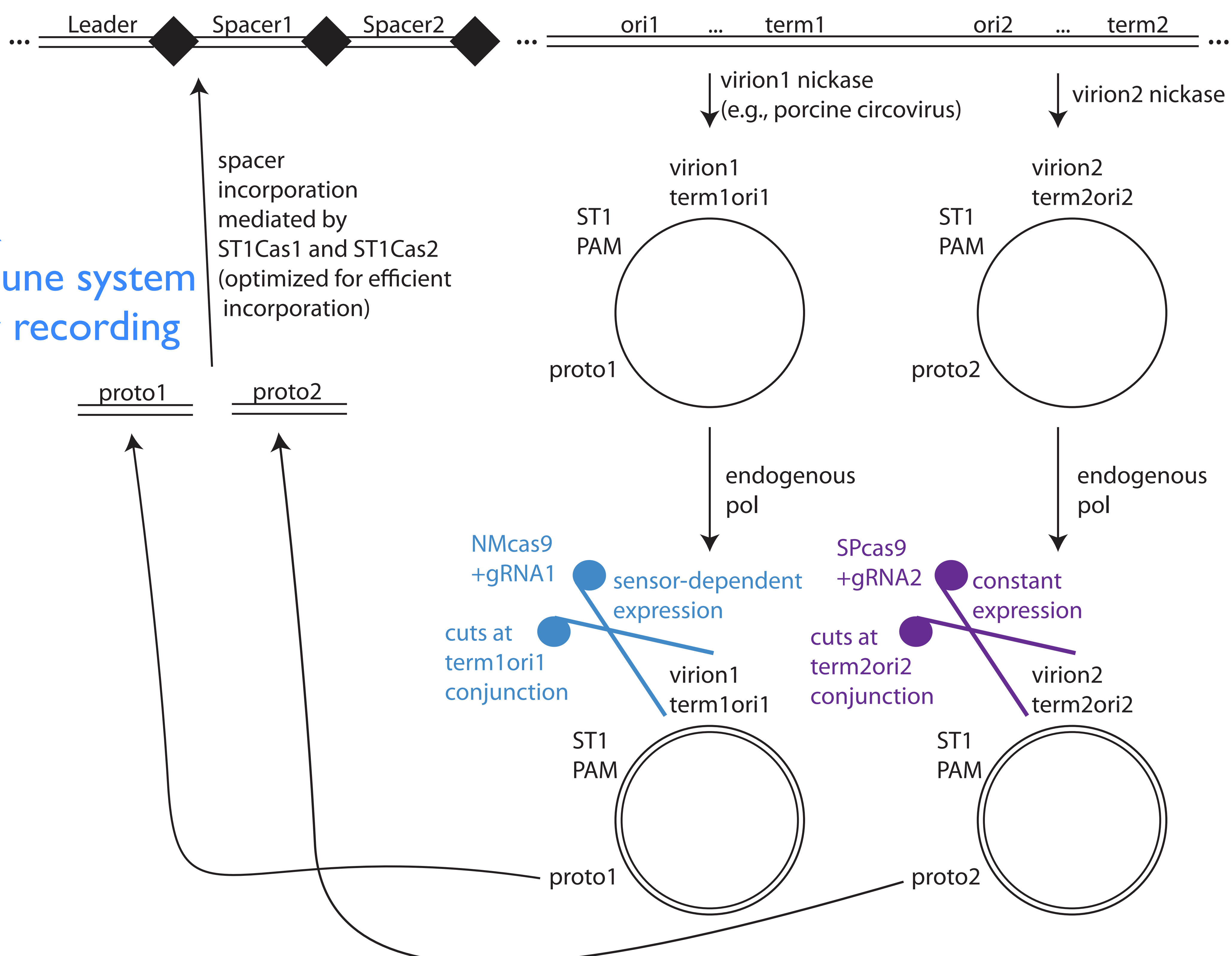
**~1000 ms** temporal resolution feasible with realistic near-term polymerase parameters

**need ensemble synchronization to reach single-spike resolution**

[Glaser, Zamft*, Marblestone*, et al. PLoS Computational Biology (2013)]

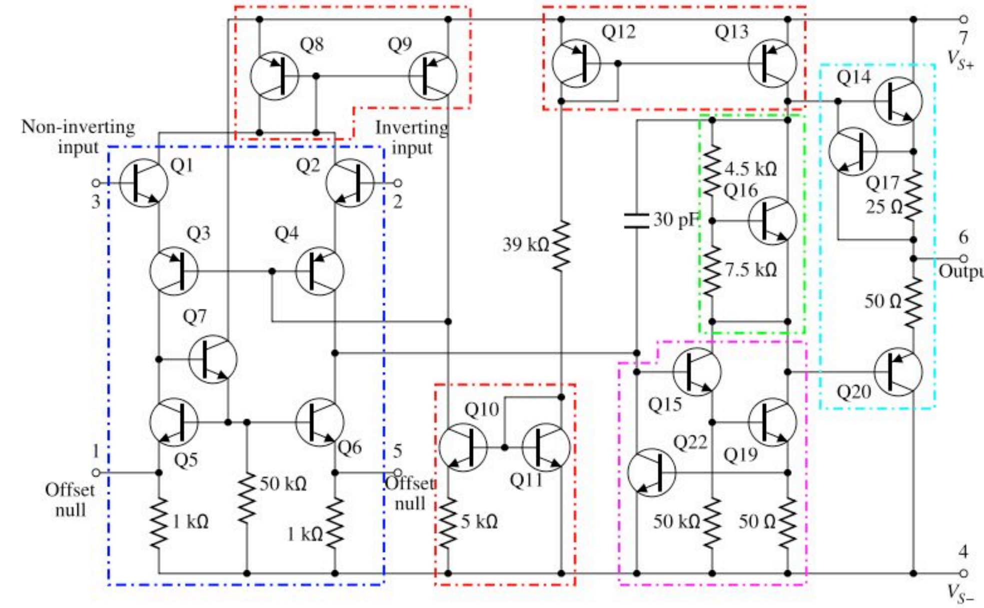
# slow-timescale *gene expression* tickertape

Hack CRISPR
bacterial immune system
for molecular recording




control expression of NMcas9 relative to SPcas9
or use the same cas9 and control gRNA expression
(which could allow generalization to many sensors)

with Esvelt, Chavez, Church, Gootenberg



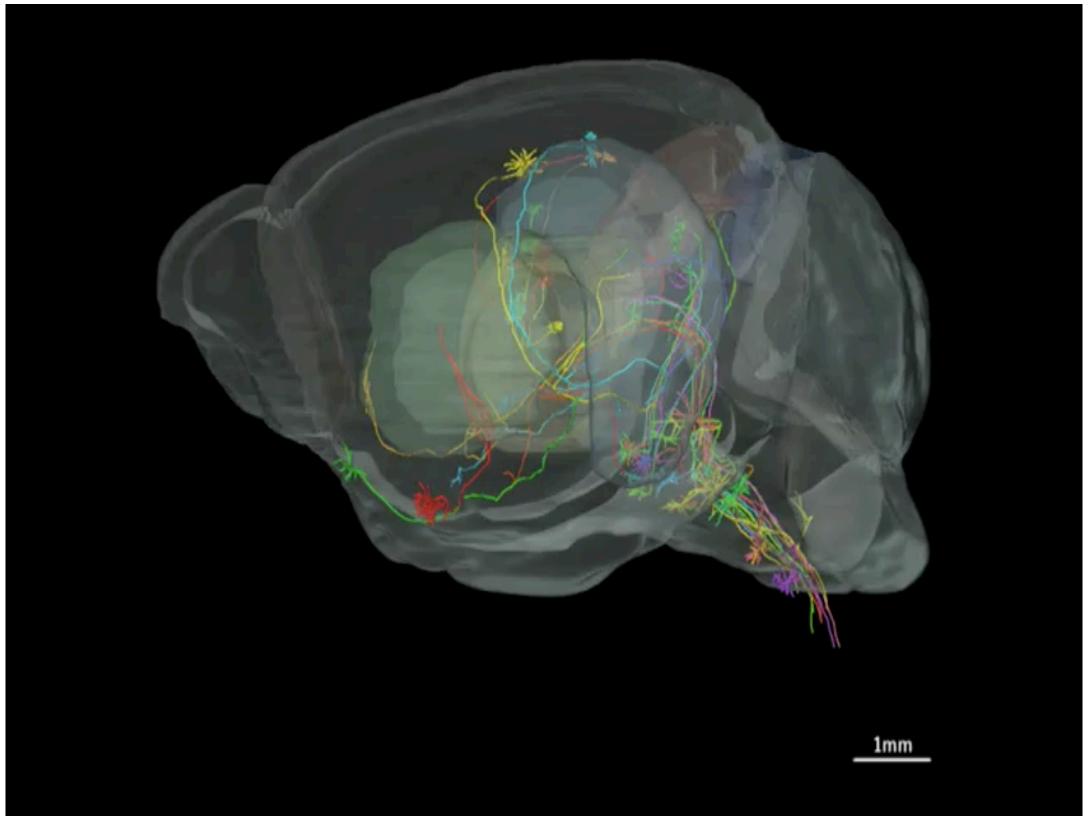

**A**ctivity history
**B**ehavior
**Connectivity** (circuit diagram)
**D**evelopment (cell lineage tree)
**E**xpression (epigenetic cell types,
+ single-synapse proteomes)

**Need:** a technology to cheaply / rapidly
measure all these variables in a single brain.

**Need:** (sub)cellular resolution + whole brain scope.

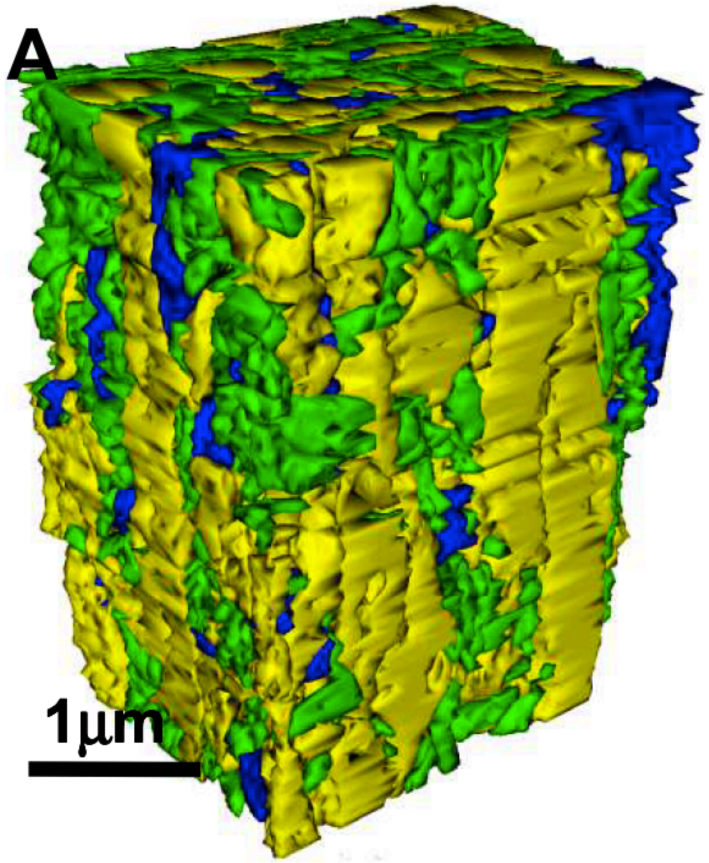
see upcoming essay by Church, Marblestone & Kalhor

long-range connections

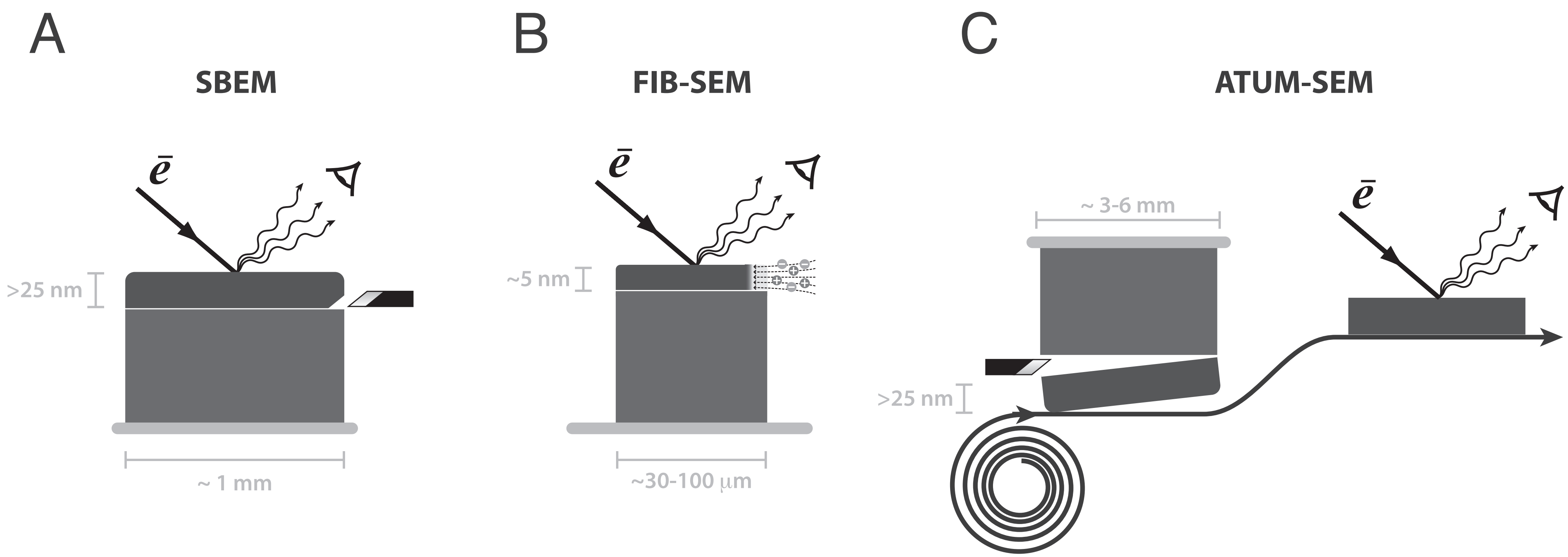



[Gong et al, 2013]

dense 3D circuitry:
>1 synapse (connection) per $um^3$

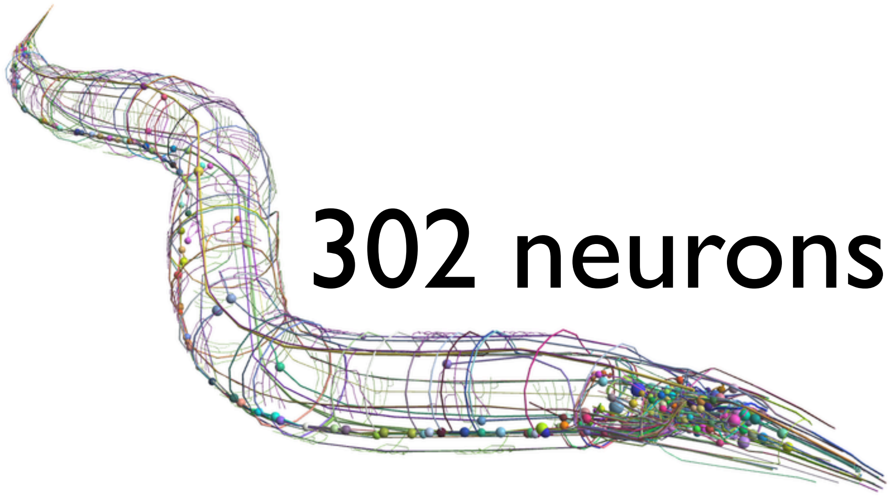



[Mischenko, 2010]




*Assumption*: widely believed that electron microscopy (EM) is the <u>only viable method</u> for cellular-resolution connectomics

Marblestone et al 2013



302 neurons: 50 person-years for C. elegans

analysis

**$1B** for mouse w/ **no molecular info**

Can’t run a <u>simulation</u>

Marblestone et al 2013

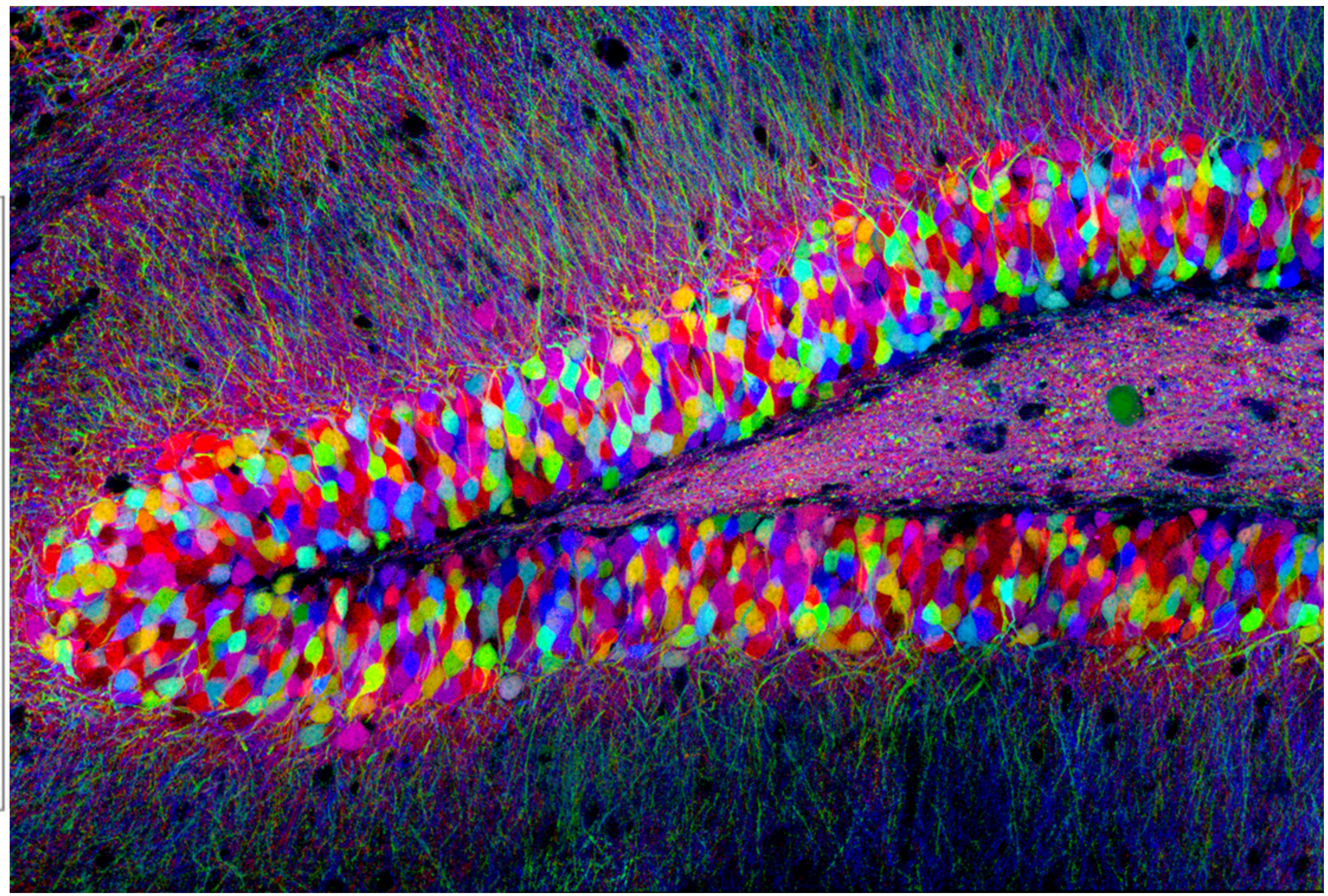
# BrainBow: random color for each neuron

... *analog* implementation of a profound *digital* idea

**Problem**: only ~20 distinguishable colors in practice...

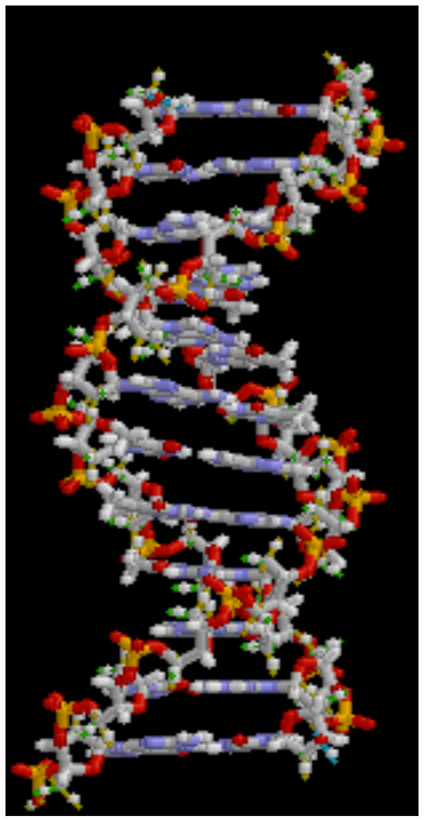
**a** XFP combinations

| Outcome for each copy 1 | 2 | 3 | Resulting colour |
|---|---|---|---|
| C | C | C | Blue |
| C | C | Y | Light blue |
| C | Y | Y | Blue-green |
| Y | Y | Y | Green |
| Y | Y | R | Light green |
| Y | R | R | Orange |
| R | R | R | Red |
| R | R | C | Magenta |
| R | C | C | Purple |
| R | C | Y | Grey |



Lichtman, Sanes, Livet et al

# $4^N$ possible DNA sequences of length N “letters”

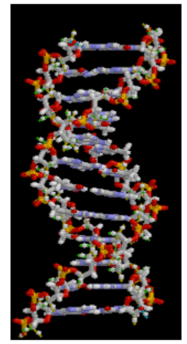


=

**0101001010101010010101**
**0101010101001010110111**
**0010110101010111010011**

=

w/ Zador, Church et al



=

**Strategy #1**: deliver random DNA strings by cloning into a virus

<u>Duplicate barcode probability per mouse brain</u>: negligible for 30 nt barcodes

$$P(j,n) = n! * \text{Binomial}(4^j, n)/(4^j)^n$$

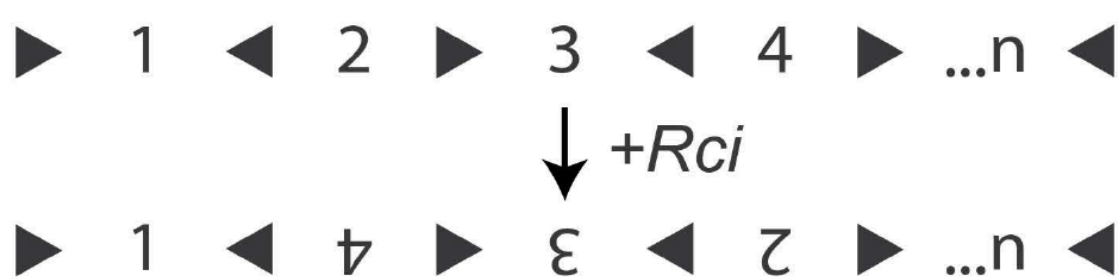
where $n$ is the size of the cell population and $j$ is the DNA barcode length in nucleotides. $P(j,n)$ can be approximated by

$$P(j,n) \approx \exp(-\frac{n^2/2}{4^j})$$

such that for $n = 7.5 * 10^7$ neurons and $j = 31$ base-long barcodes, the probability of a duplication $1 - P(j,n) < 0.001$. This corresponds to a total barcode population size of $4^{31} \approx 5 * 10^{18}$.

**Strategy #2**: have each cell *make* its own barcode in-vivo

<u>Example: invertase</u>



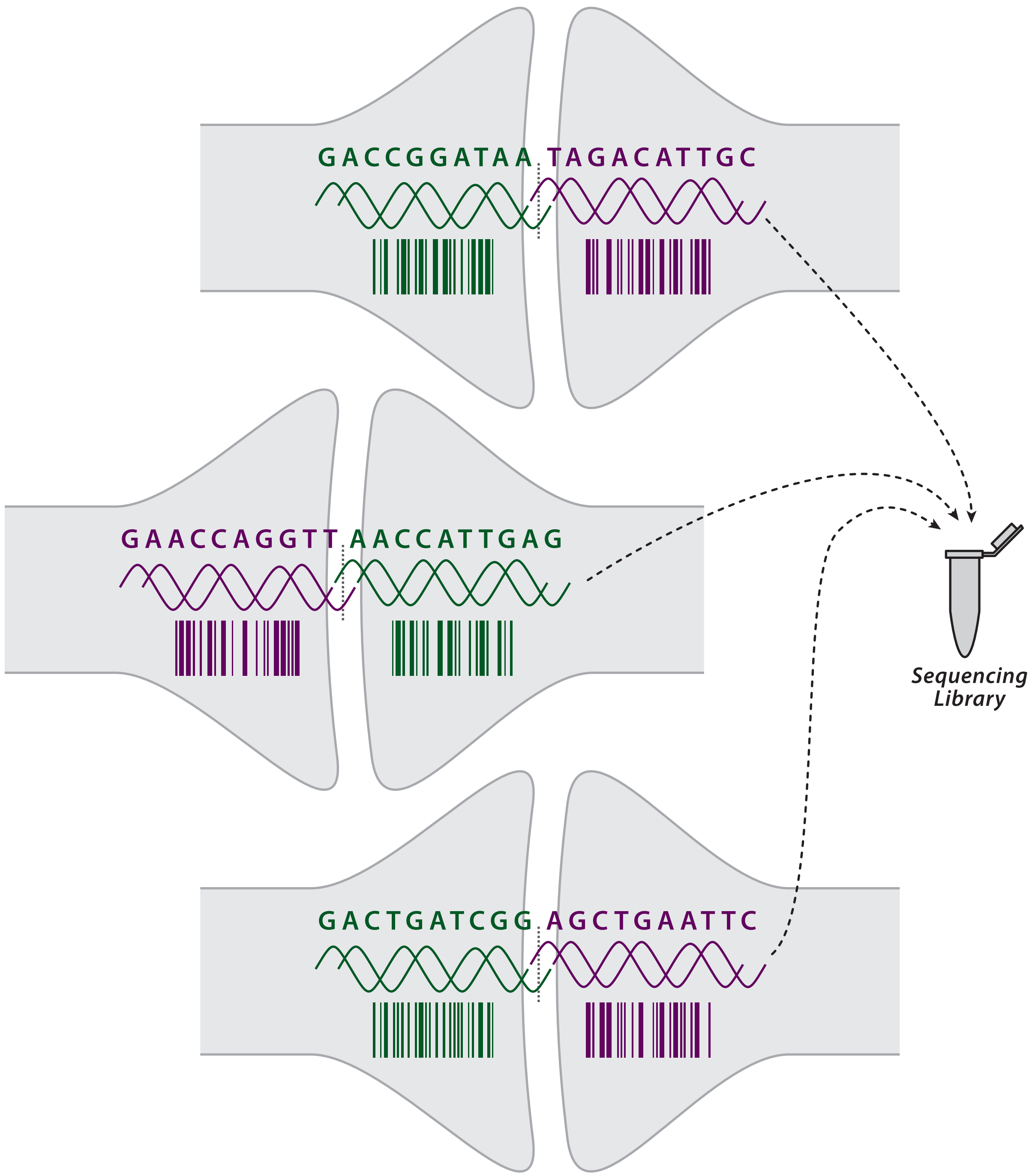

<u>Theoretical diversity</u>

$$k! \times 2^k$$

*Other in-vivo barcoding strategies are also under way in the Church lab*

Peikon, Gizatullina & Zador, 2014

# **Zador:** Pair barcodes of connected cells, then sequence




**extracts connectivity**

**<u>but</u>**

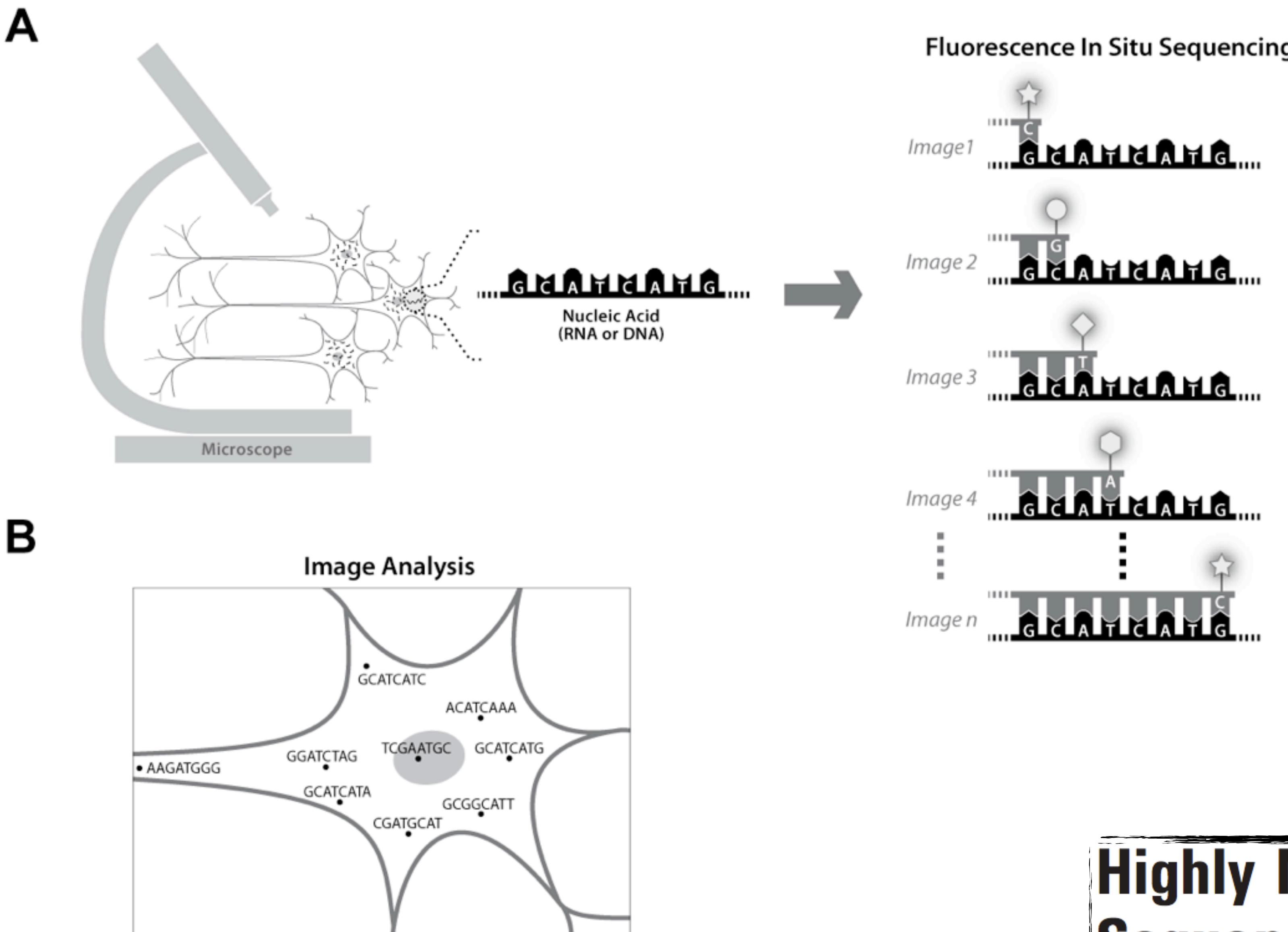
**1) scrambles cell positions**

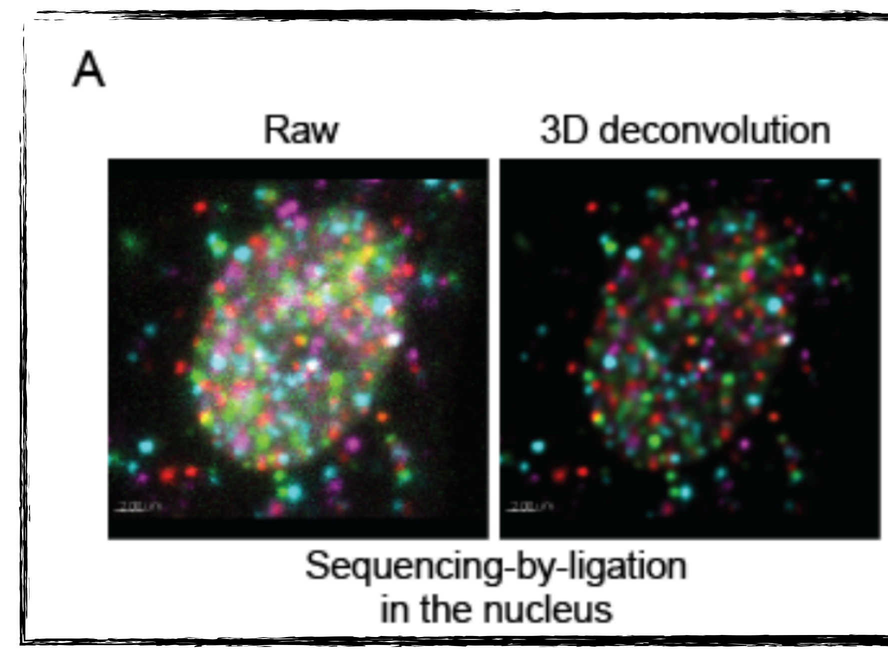
**2) hard to integrate w/ molecular "annotations" (e.g., gene expression)**

Marblestone, Daugharthy, Kalhor et al, 2014

# digital $4^N$ **color** microscopy

Fluorescent In-Situ DNA Sequencing (FISSEQ):





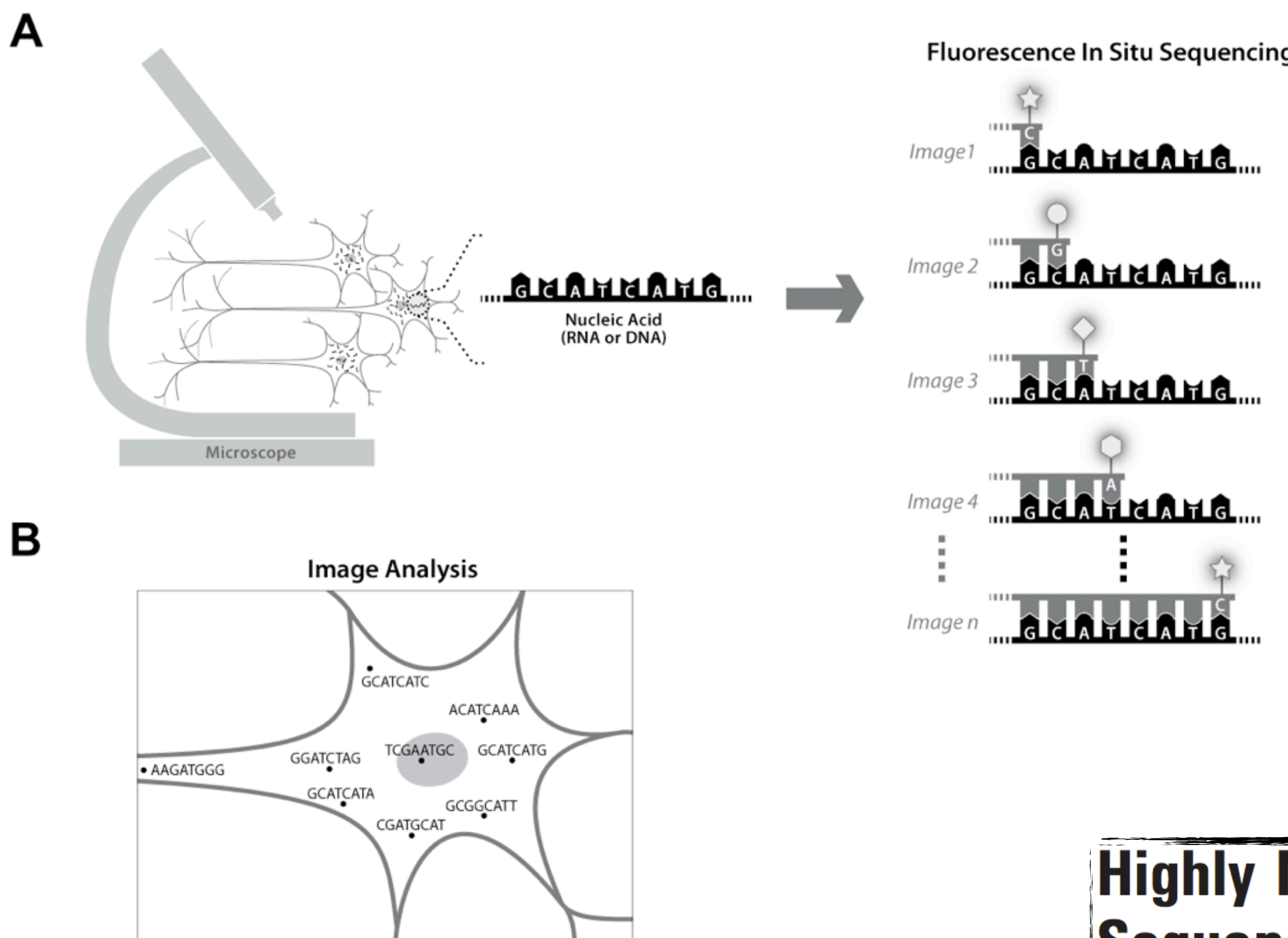
**Highly Multiplexed Subcellular RNA Sequencing in Situ**

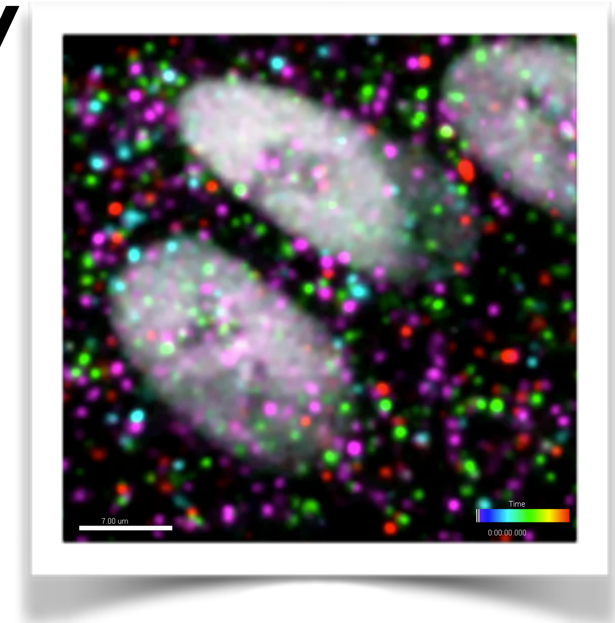
Je Hyuk Lee,[1,2*†] Evan R. Daugharthy,[1,2,4*] Jonathan Scheiman,[1,2] Reza Kalhor,[2] Joyce L. Yang,[2] Thomas C. Ferrante,[1] Richard Terry,[1] Sauveur S. F. Jeanty,[1] Chao Li,[1] Ryoji Amamoto,[3] Derek T. Peters,[3] Brian M. Turczyk,[1] Adam H. Marblestone,[1,2] Samuel A. Inverso,[1] Amy Bernard,[5] Prashant Mali,[2] Xavier Rios,[2] John Aach,[2] George M. Church[1,2†]

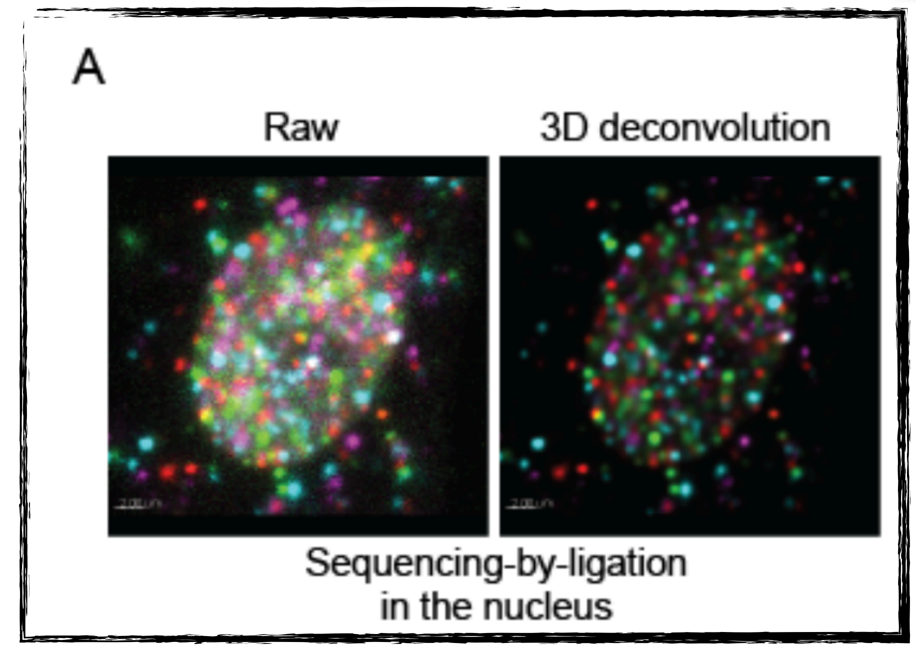
Figure by Reza Kalhor

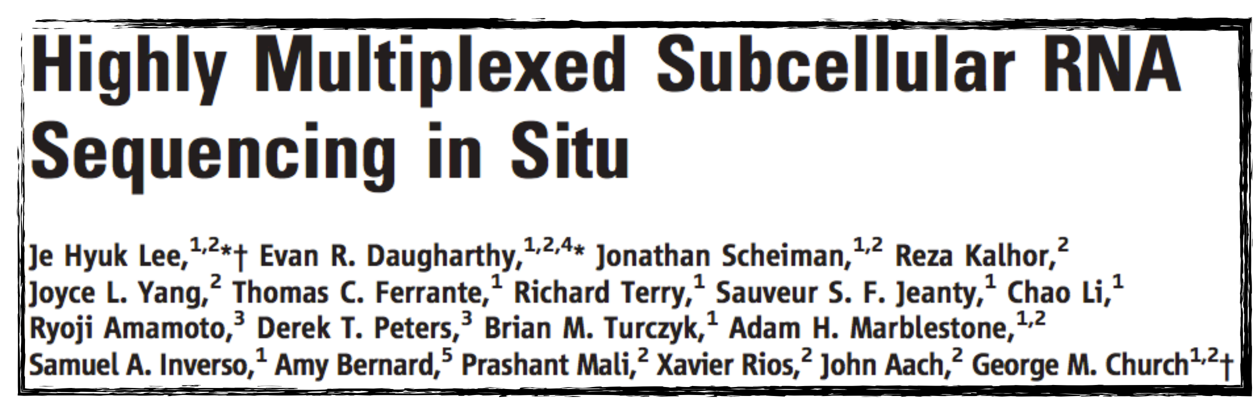
# digital $\mathbf{4^N}$ **color** microscopy

Fluorescent In-Situ DNA Sequencing (FISSEQ):





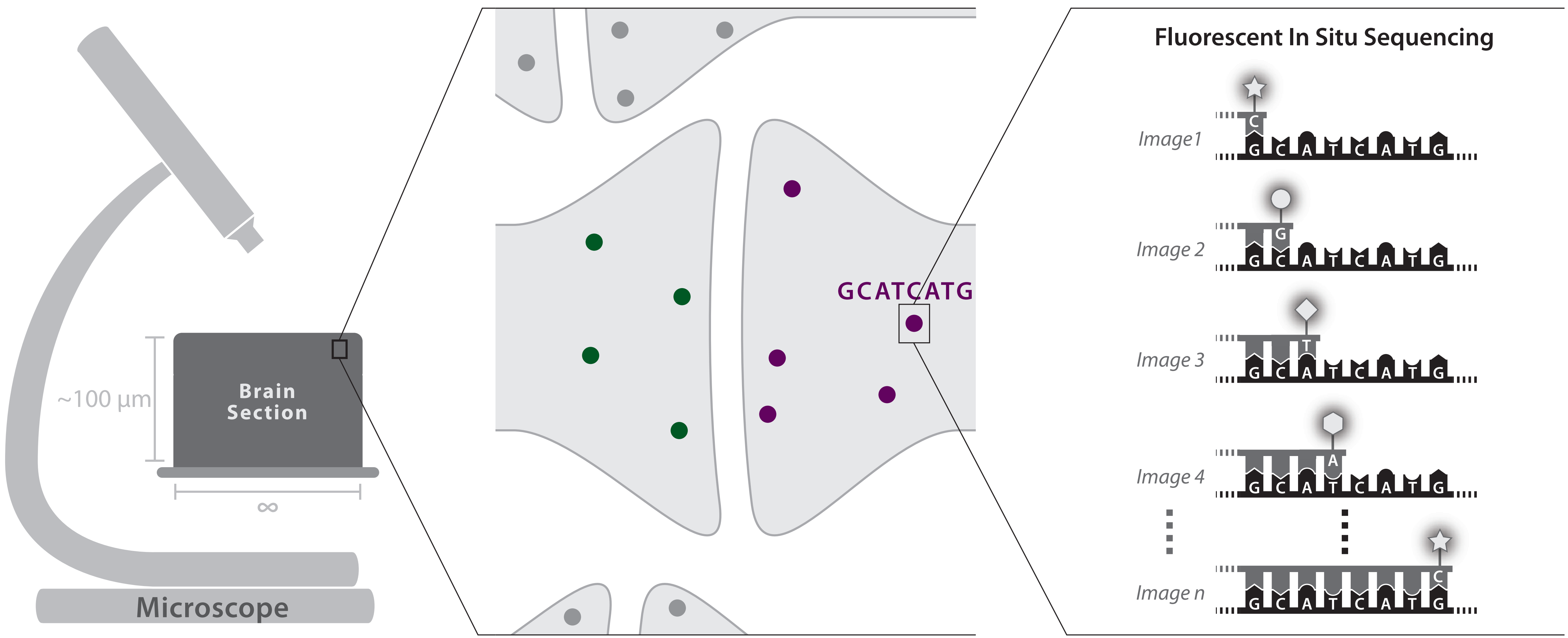



**Highly Multiplexed Subcellular RNA Sequencing in Situ**

Je Hyuk Lee,[1,2*†] Evan R. Daugharthy,[1,2,4*] Jonathan Scheiman,[1,2] Reza Kalhor,[2] Joyce L. Yang,[2] Thomas C. Ferrante,[1] Richard Terry,[1] Sauveur S. F. Jeanty,[1] Chao Li,[1] Ryoji Amamoto,[3] Derek T. Peters,[3] Brian M. Turczyk,[1] Adam H. Marblestone,[1,2] Samuel A. Inverso,[1] Amy Bernard,[5] Prashant Mali,[2] Xavier Rios,[2] John Aach,[2] George M. Church[1,2†]

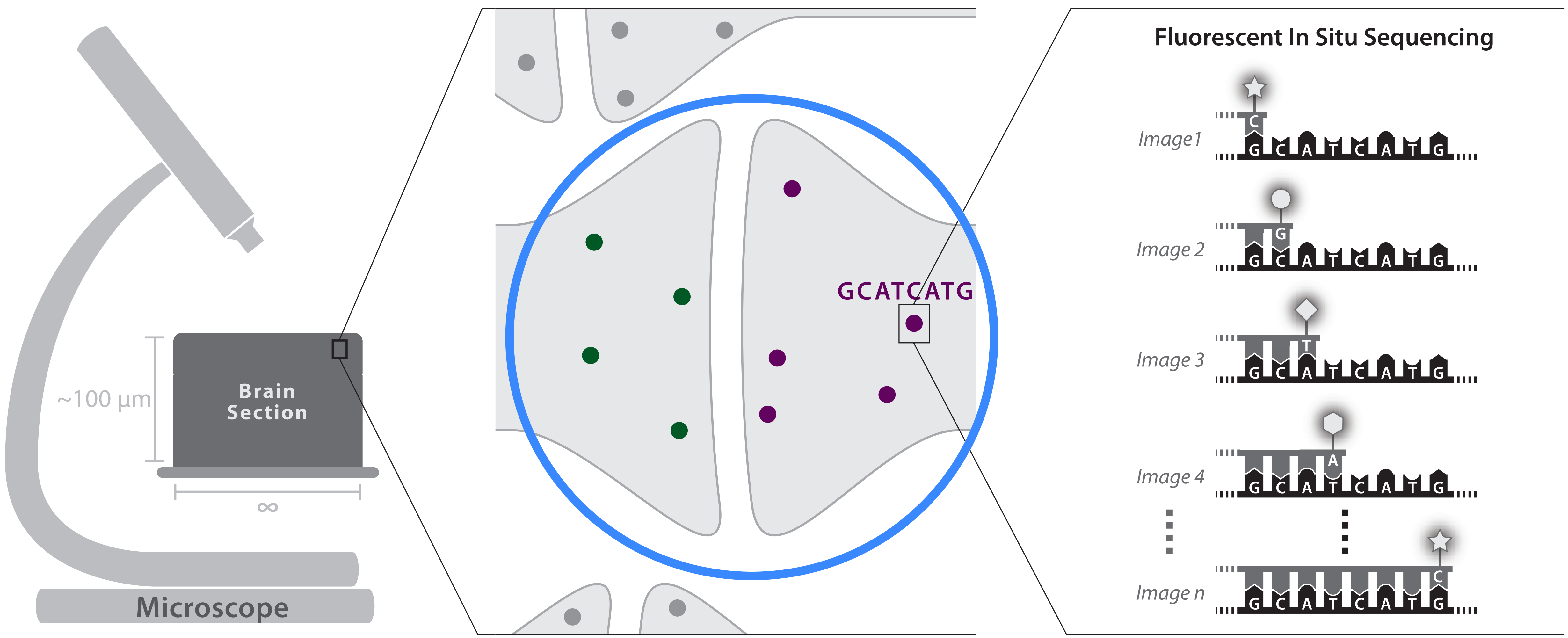
Figure by Reza Kalhor

# digital $4^N$ color microscopy



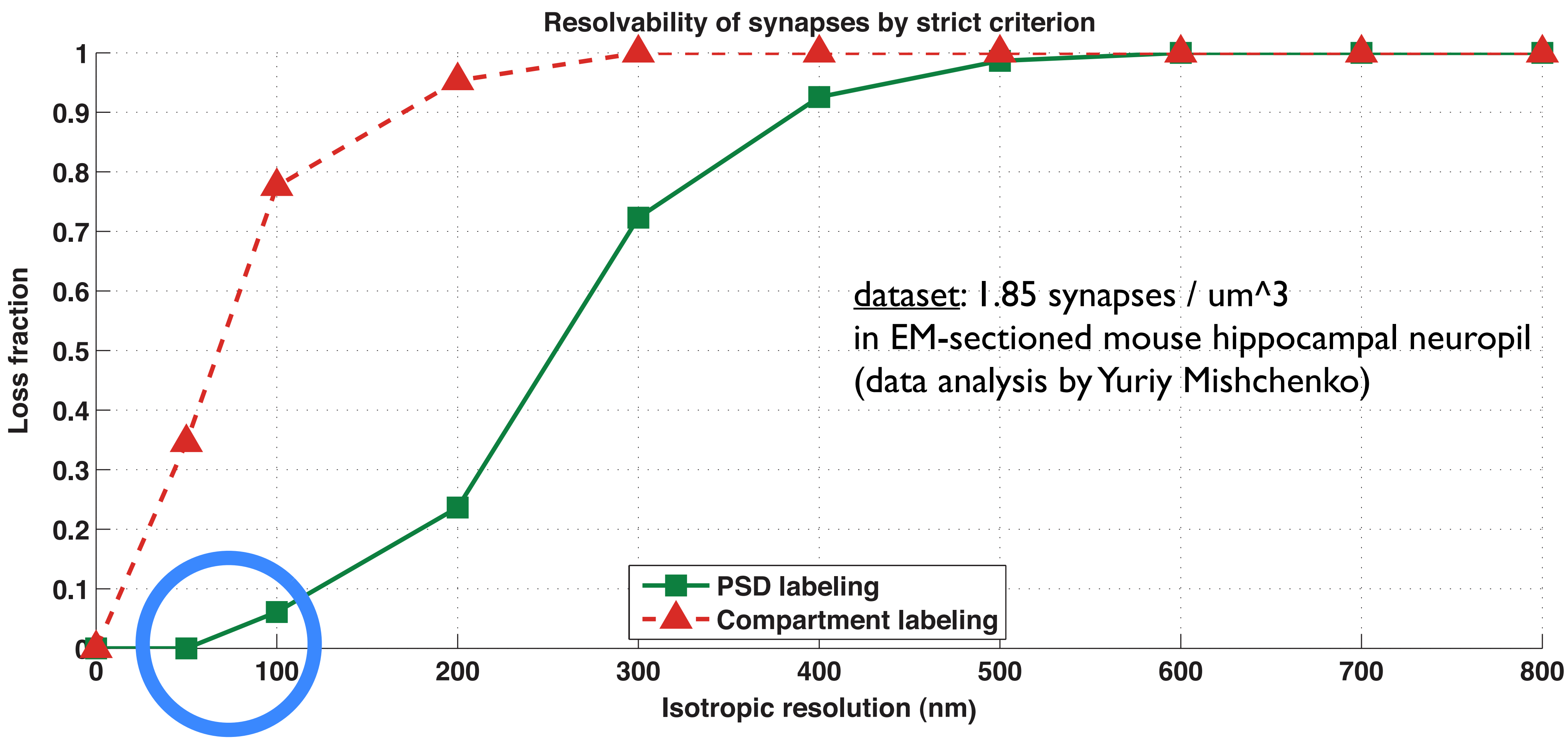

Marblestone, Daugharthy, Kalhor et al, 2014

# digital $4^N$ color microscopy




Marblestone, Daugharthy, Kalhor et al, 2014

# digital $4^N$ **color** microscopy

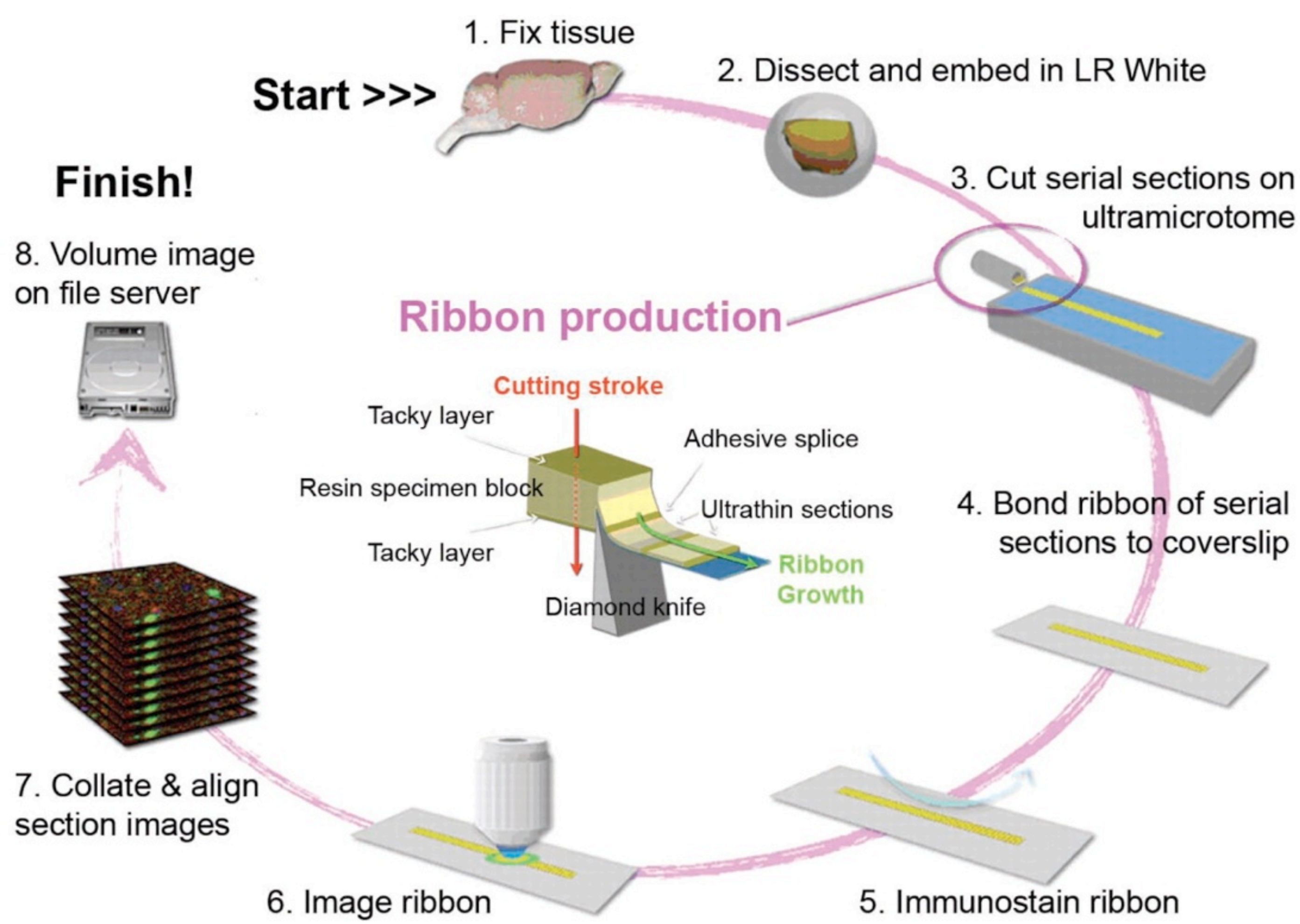



**A <u>much easier problem</u> than whole-brain electron microscopy!**

Marblestone, Daugharthy, Kalhor et al, 2014

# digital $4^N$ color microscopy

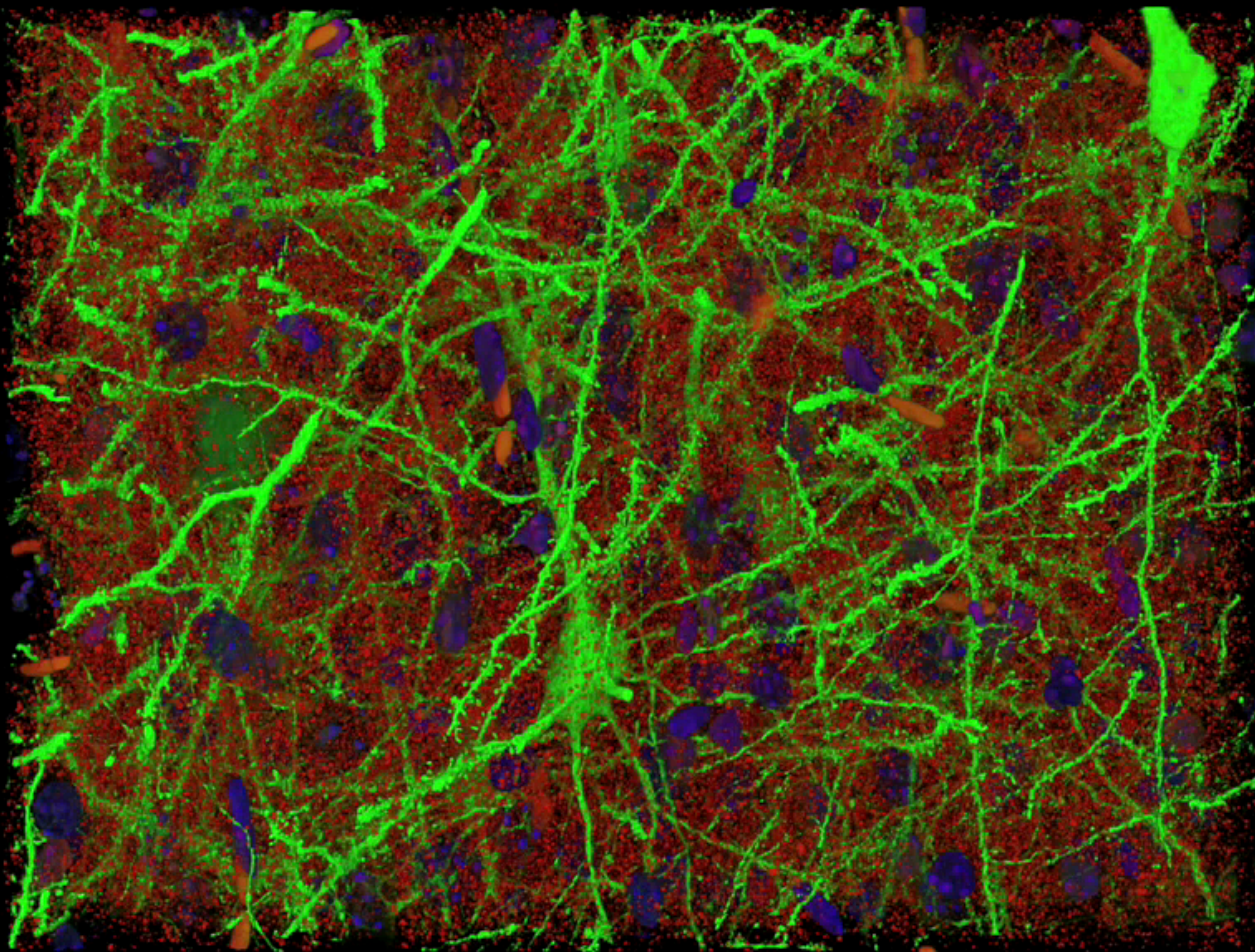
**Example: “array tomography” thin sectioning + fluorescence microscopy**




**Array Tomography: A New Tool for Imaging the Molecular Architecture and Ultrastructure of Neural Circuits**

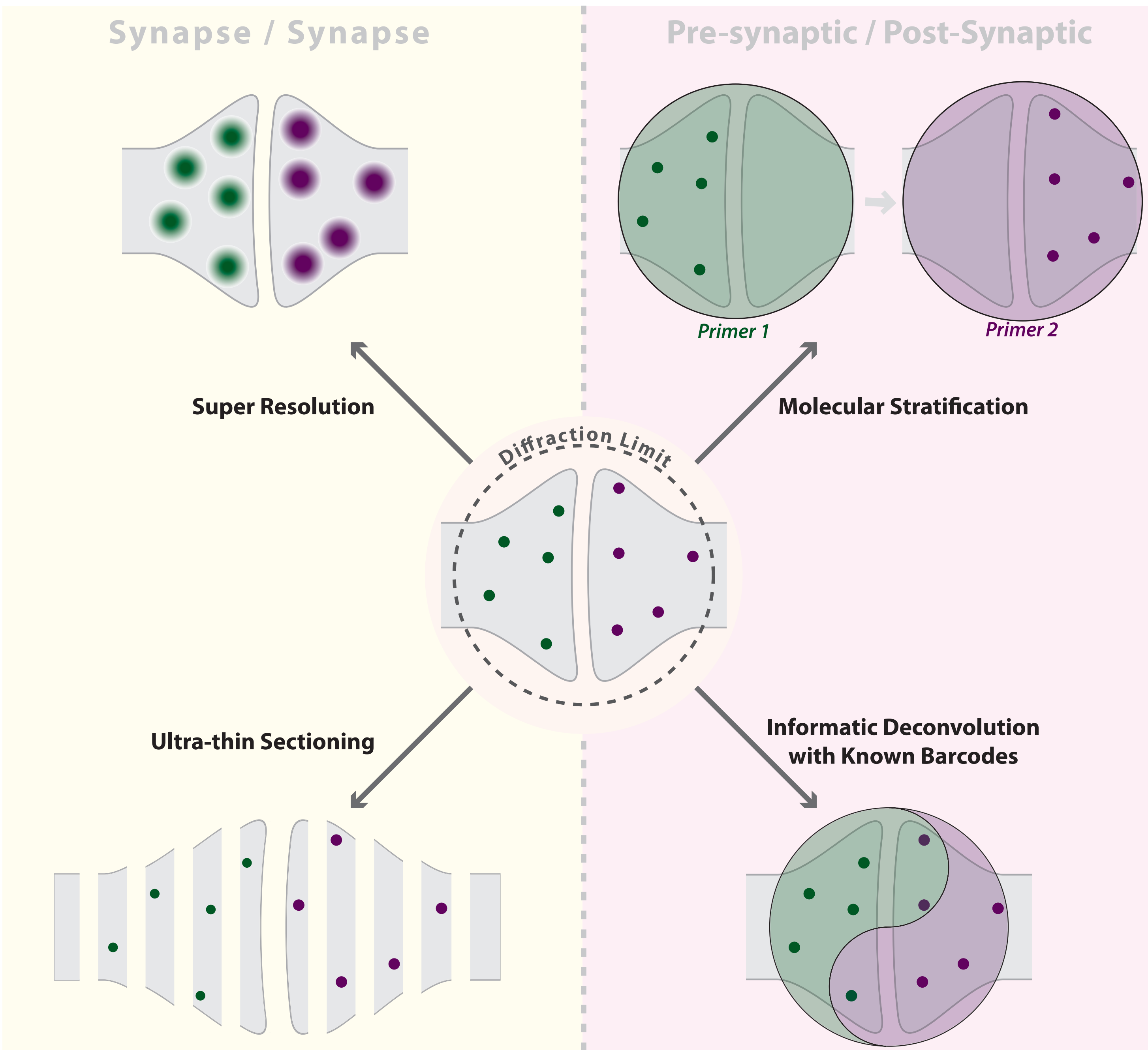
Kristina D. Micheva[1,*] and Stephen J Smith[1,*]

## *Red dots are synapses, resolved w/ diff. limited optics via thin-sectioning*



**Smith lab, Stanford**

# digital $4^N$ **color** microscopy




Enhance effective resolution by "stratifying" molecules into different image frames:

S^(1/3) xyz improvement where S is slowdown factor

Use prior information about entire pool of barcodes to <u>de-mix multiple signals from a single resolution voxel</u>:

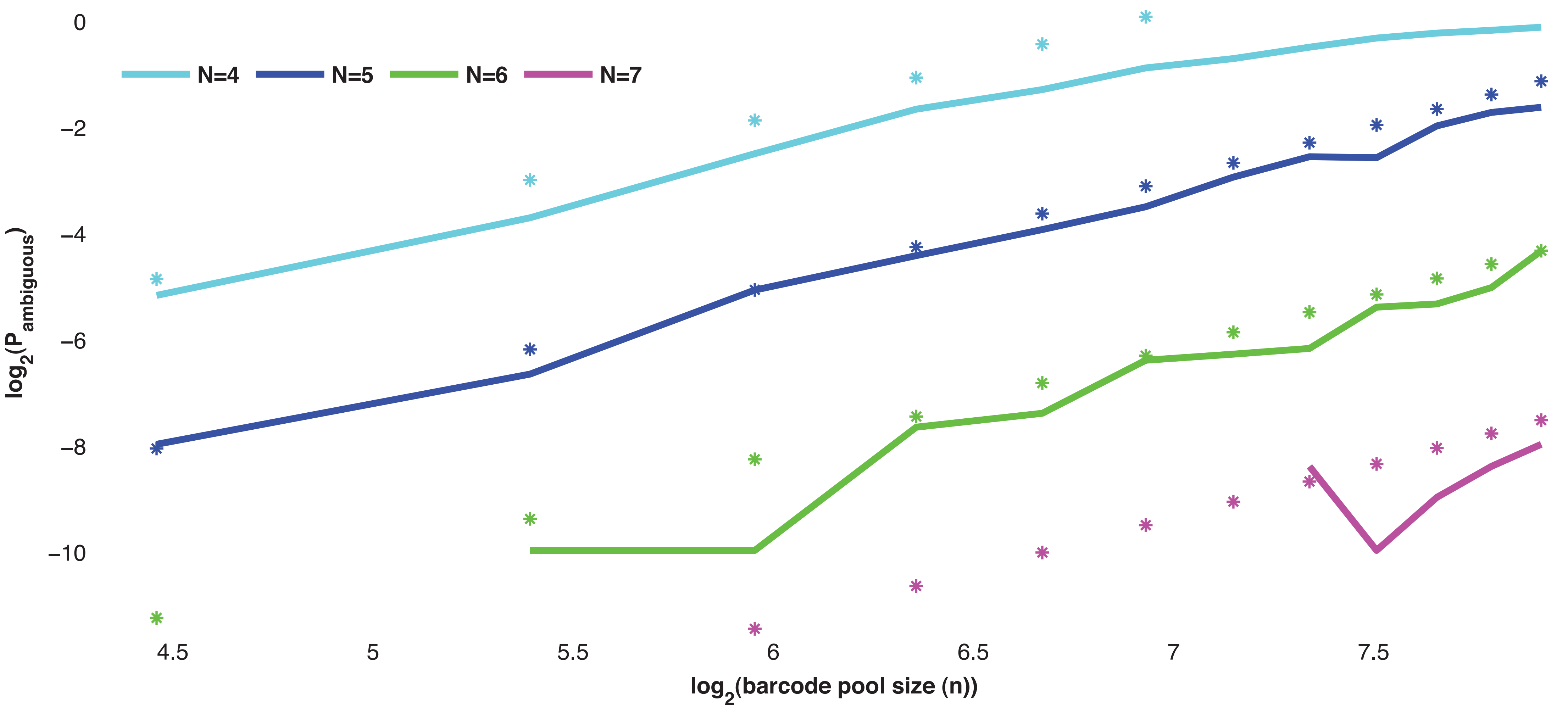
35-base barcodes suffice for no deconvolution ambiguities in mouse

Marblestone, Daugharthy et al 2014

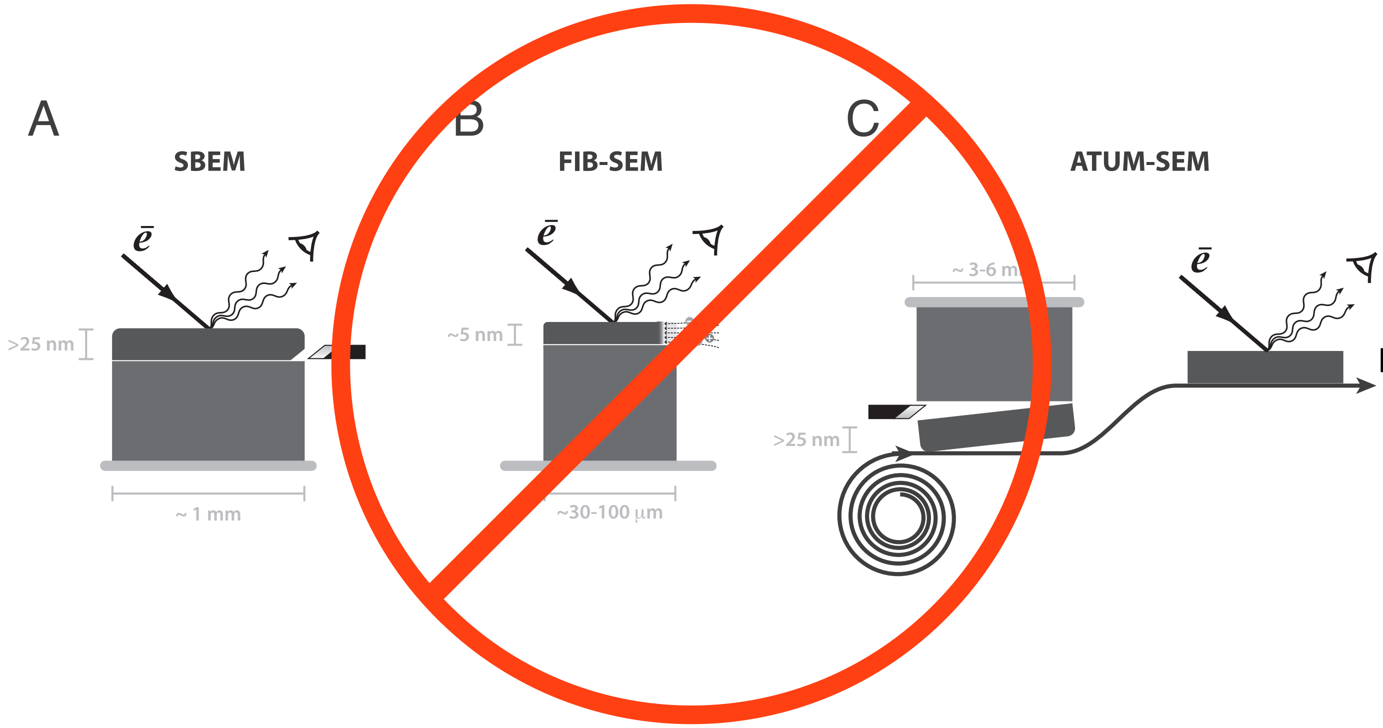
# digital $4^N$ color microscopy

$$P(\text{ambiguity}) < \frac{n^2 \cdot 2^N}{4^{2N}} = n^2 \cdot 2^{-3N}$$

*Simulations of informatic deconvolution verify estimated scaling behavior*




Marblestone, Daugharthy et al 2014




**>$1B, without molecular info**

DNA barcoding

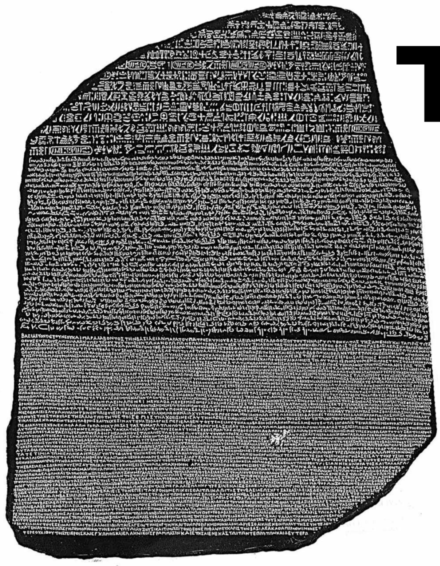
intracellular protein trafficking

fluorescent in-situ RNA sequencing

fast 100 nm microscopy

**<$10M, <u>with</u> molecular info**

Marblestone et al 2014; with Church, Zador, Boyden, Daugharthy, Kalhor, Lee, Peikon, Chen, Tillberg et al



# The Rosetta Brain integration project

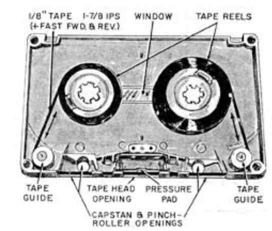
**Activity**

**B**ehavior

**Connectivity**

**Development**

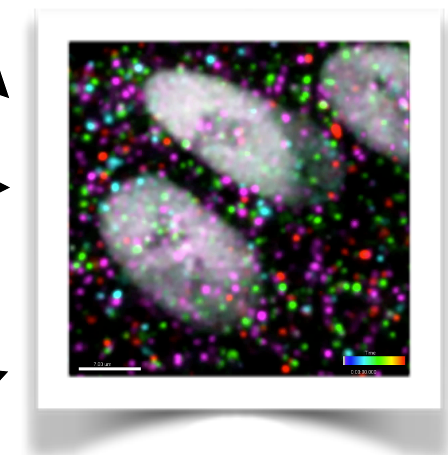
**Expression**




Molecular recordings

In-vivo-generated
cell barcodes
(update 1x per division)

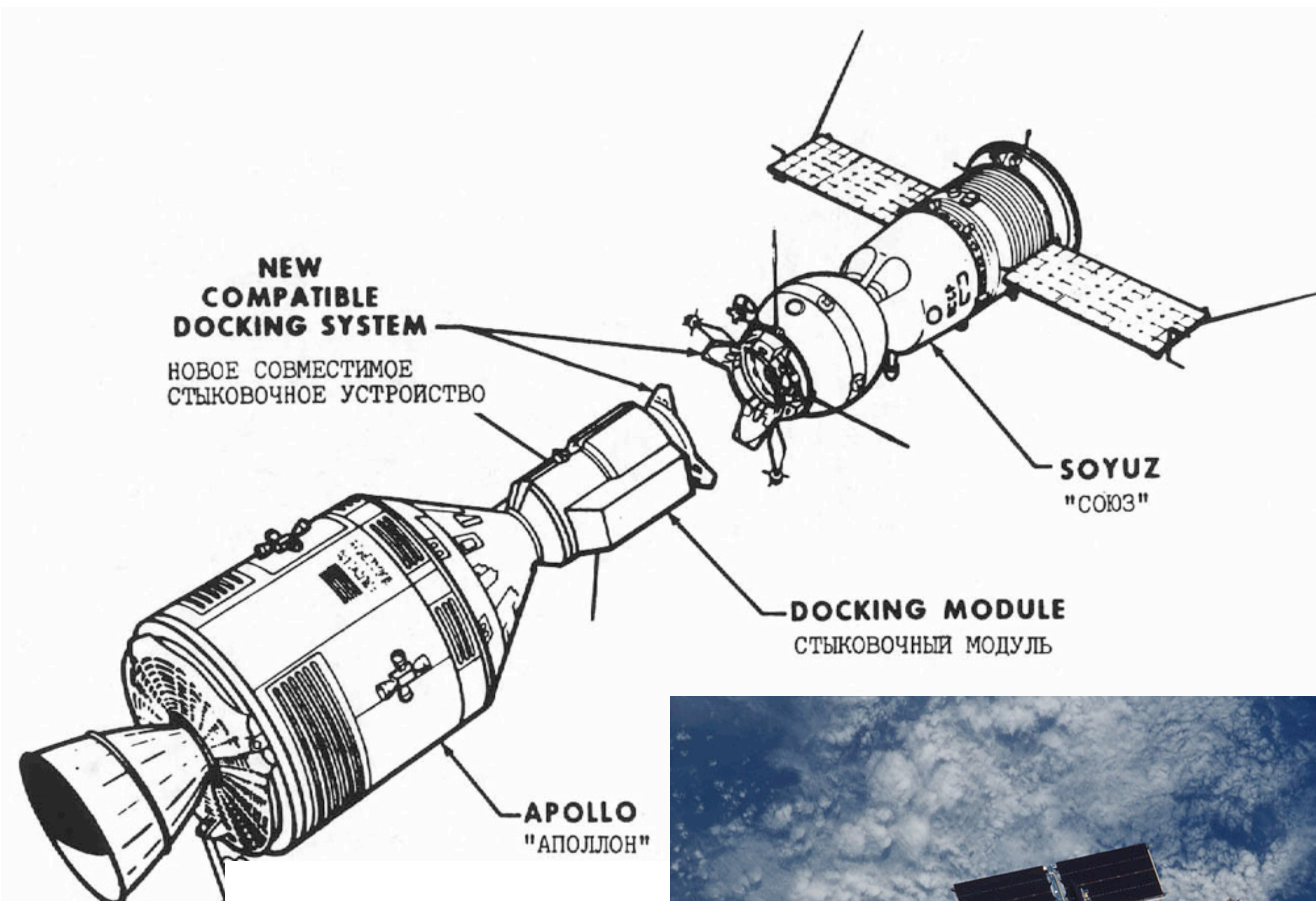
Transcripts
DNA-barcoded antibodies

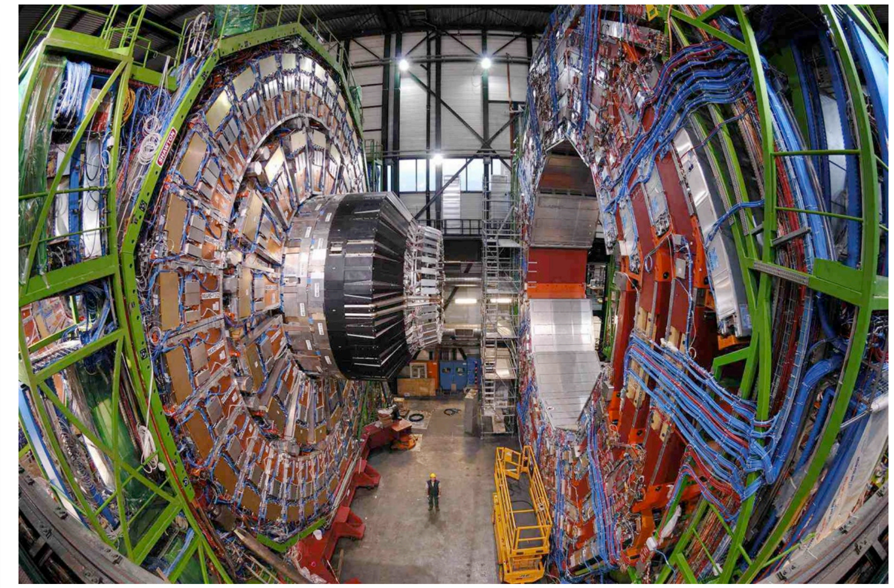


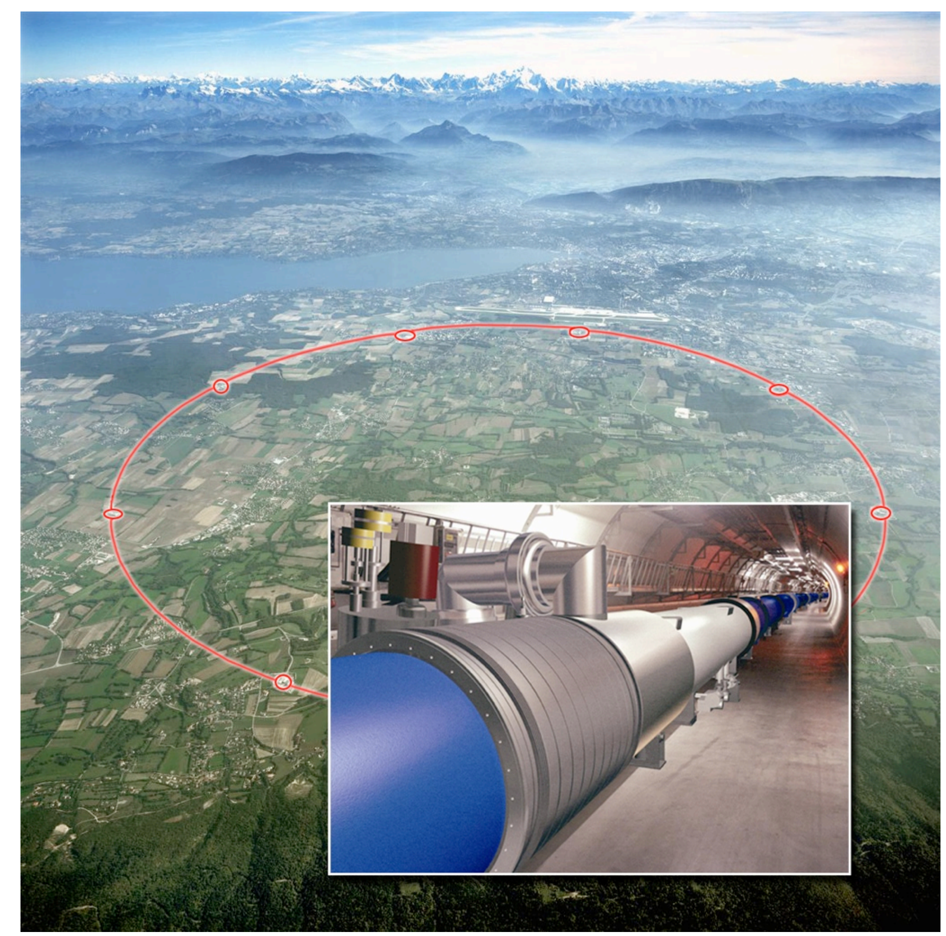
FISSEQ
(Lee et al 2014)

see upcoming essay by Church, Marblestone & Kalhor

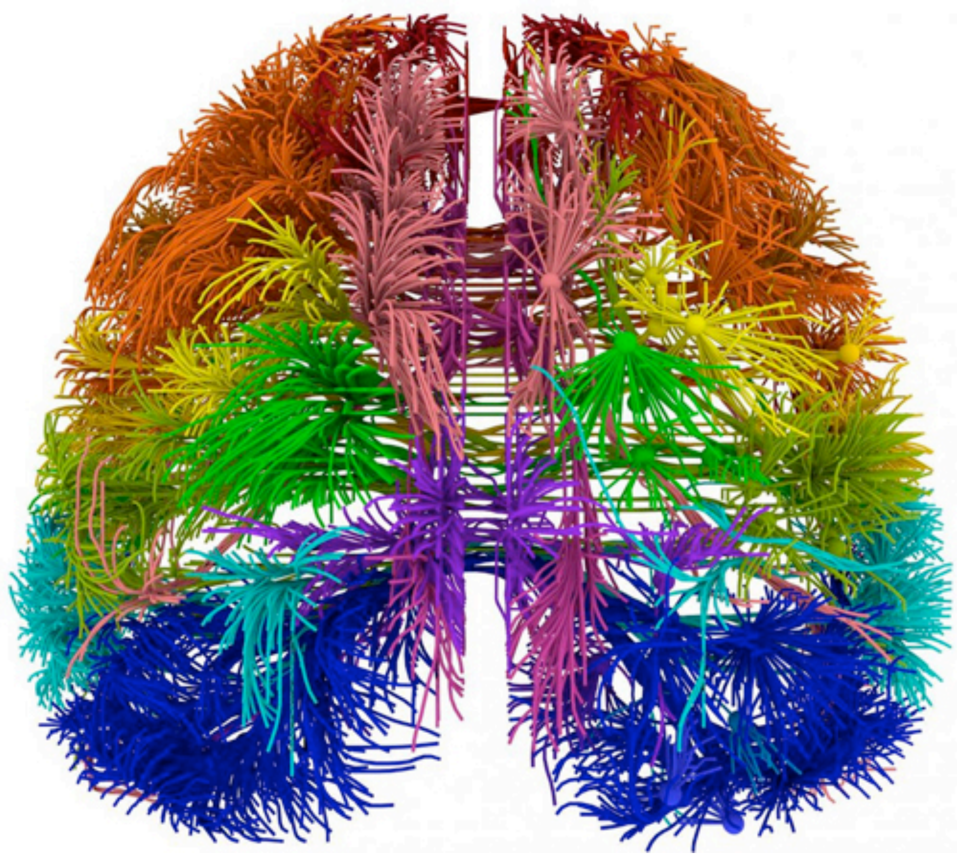
# Well-known "systems engineering" methods by-&-for the sciences of *simplicity*...

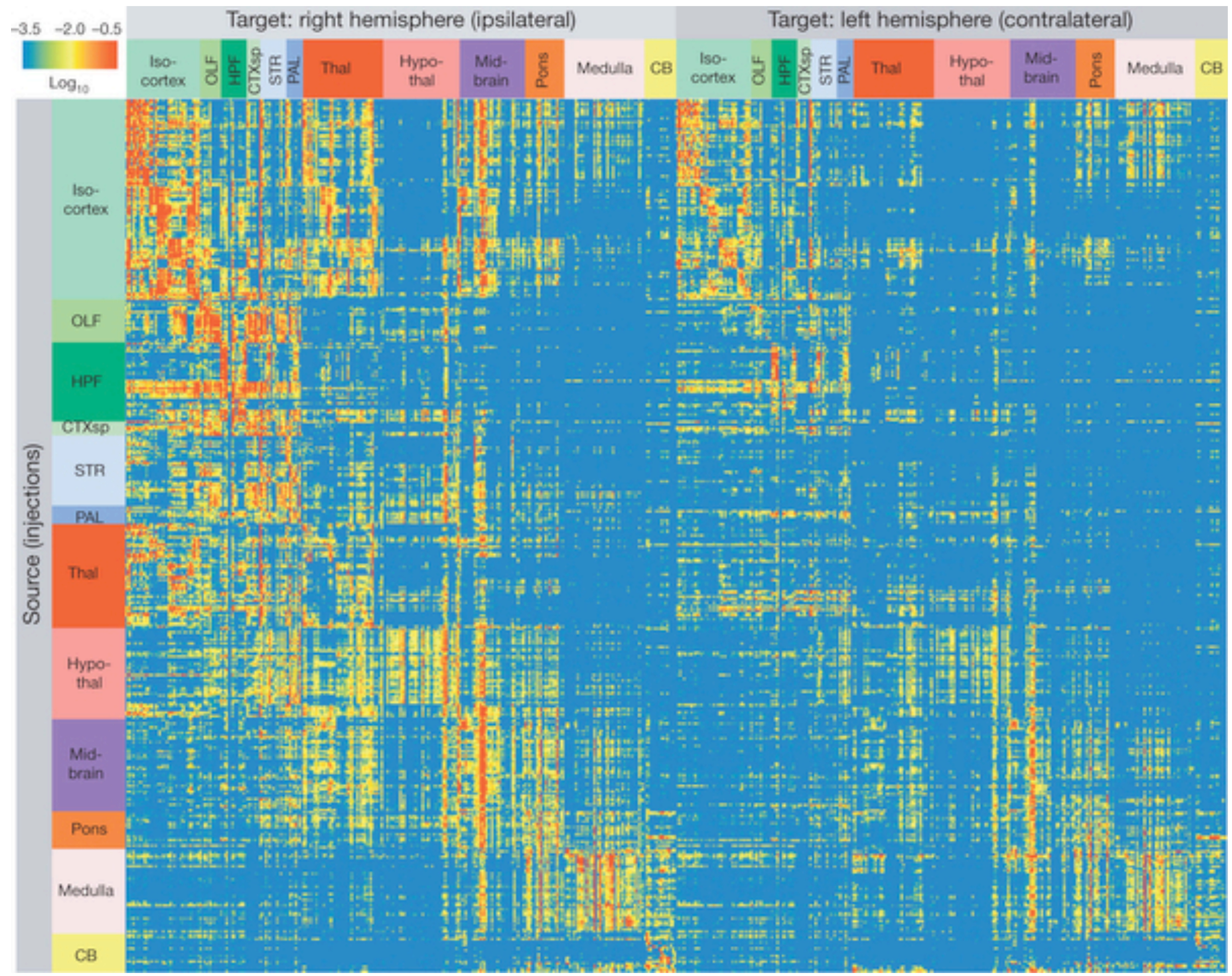




# How to extend to the sciences of *complexity*?






Allen Institute mouse brain connectivity atlas (2014)

# Further reading

frontiers in COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 21 October 2013
doi: 10.3389/fncom.2013.00137

## Physical principles for scalable neural recording

*Adam H. Marblestone[1,2*†], Bradley M. Zamft[3†], Yael G. Maguire[3,4], Mikhail G. Shapiro[5], Thaddeus R. Cybulski[6], Joshua I. Glaser[6], Dario Amodei[7], P. Benjamin Stranges[3], Reza Kalhor[3], David A. Dalrymple[1,8,9], Dongjin Seo[10], Elad Alon[10], Michel M. Maharbiz[10], Jose M. Carmena[10,11], Jan M. Rabaey[10], Edward S. Boyden[9,12‡], George M. Church[1,2,3‡] and Konrad P. Kording[13,14‡]*

## Rosetta Brains: A Strategy for Molecularly-Annotated Connectomics

Adam H Marblestone, Evan R Daugharthy, Reza Kalhor, Ian D Peikon, Justus M Kebschull, Seth L Shipman, Yuriy Mishchenko, Je Hyuk Lee, Konrad P Kording, Edward S Boyden, Anthony M Zador, George M Church

OPEN ACCESS Freely available online

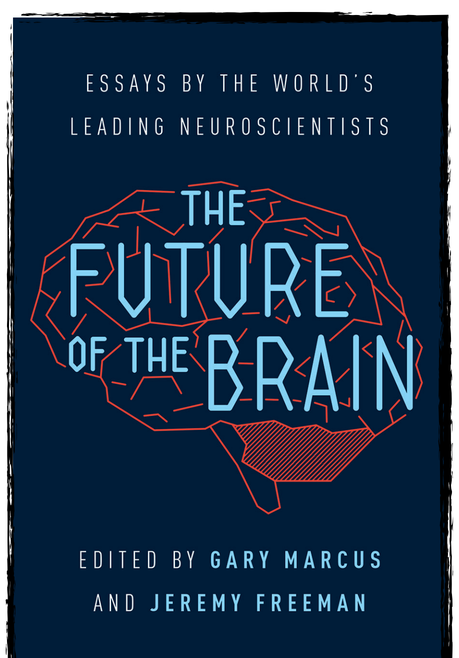
PLOS | ONE

## Measuring Cation Dependent DNA Polymerase Fidelity Landscapes by Deep Sequencing

**Bradley Michael Zamft[1], Adam H. Marblestone[2,9], Konrad Kording[3,8], Daniel Schmidt[4,5,7], Daniel Martin-Alarcon[4,5,7], Keith Tyo[6], Edward S. Boyden[4,5,7], George Church[1,2*]**

## Highly Multiplexed Subcellular RNA Sequencing in Situ

Je Hyuk Lee,[1,2*†] Evan R. Daugharthy,[1,2,4*] Jonathan Scheiman,[1,2] Reza Kalhor,[2] Joyce L. Yang,[2] Thomas C. Ferrante,[1] Richard Terry,[1] Sauveur S. F. Jeanty,[1] Chao Li,[1] Ryoji Amamoto,[3] Derek T. Peters,[3] Brian M. Turczyk,[1] Adam H. Marblestone,[1,2] Samuel A. Inverso,[1] Amy Bernard,[5] Prashant Mali,[2] Xavier Rios,[2] John Aach,[2] George M. Church[1,2†]




OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

## Statistical Analysis of Molecular Signal Recording

**Joshua I. Glaser[1*], Bradley M. Zamft[2], Adam H. Marblestone[3,4], Jeffrey R. Moffitt[5], Keith Tyo[6], Edward S. Boyden[7,8,9], George Church[2,3,4], Konrad P. Kording[1,10,11]**

## Conneconomics: The Economics of Dense, Large-Scale, High-Resolution Neural Connectomics

Adam H Marblestone, Evan R Daugharthy, Reza Kalhor, Ian D Peikon, Justus M Kebschull, Seth L Shipman, Yuriy Mishchenko, Je Hyuk Lee, David A Dalrymple, Bradley M Zamft, Konrad P Kording, Edward S Boyden, Anthony M Zador, George M Church

**doi:** 10.1101/001214