Opportunities and Challenges in Big Data Neuroscience

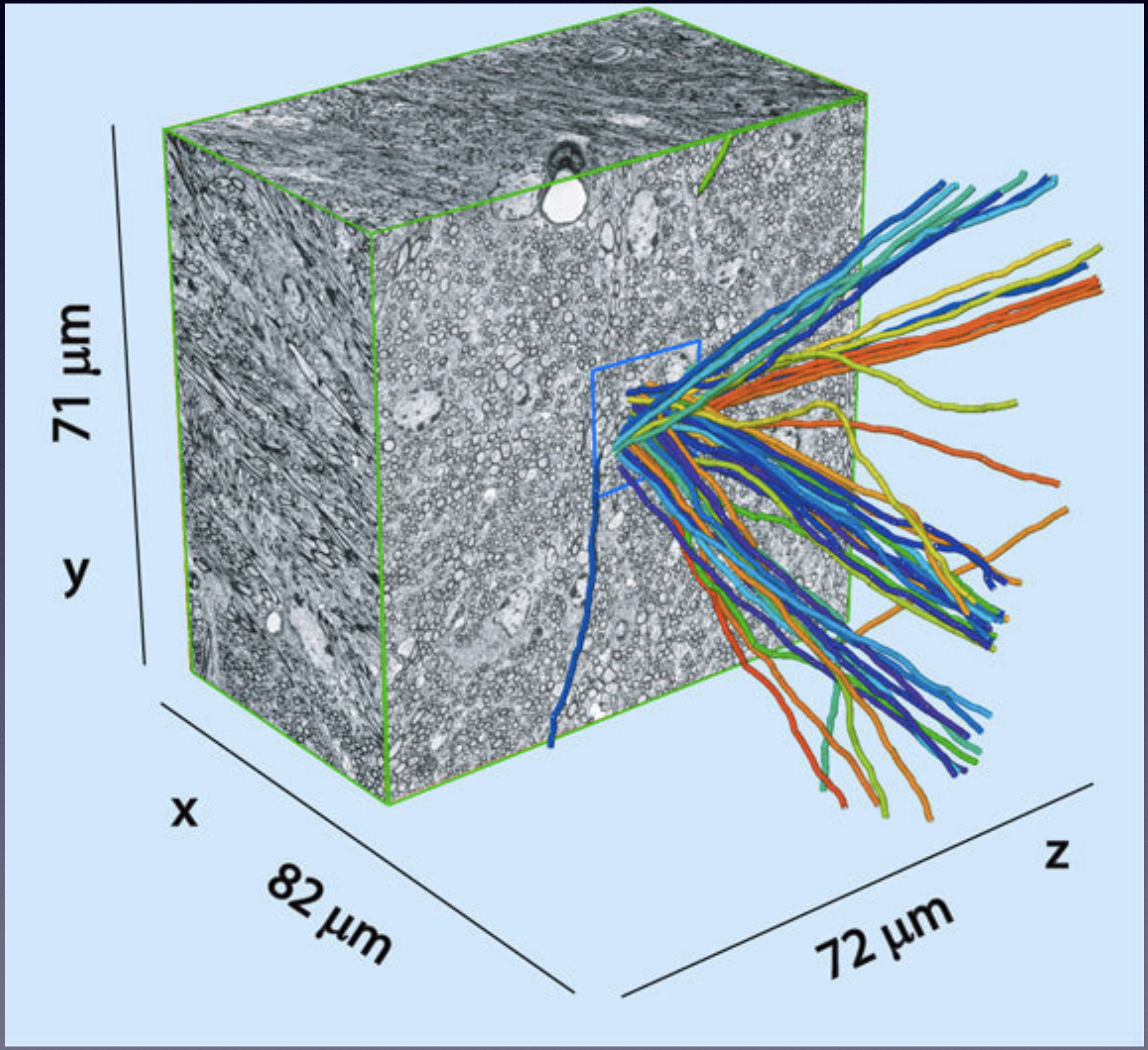
Joshua T. Vogelstein
{BME, ICM, CIS, IDIES}@JHU
Co-founder and Director of the **Open Connectome Project**
e: jovo@jhu.edu, w: http://ocp.me

Why is it hard?

- **volume**: individual datasets larger than RAM, or local storage
- **variety**: each modality requires domain specific expertise & code
- **velocity**: some technologies can generate terabytes a day per lab
- **veracity**: data are noisy, missing, etc., analysis must be robust
- **parallelism**: many want to both read & **write** to the data in parallel
- **complexity**: raw data are “images”, useful data are semantically tagged, data —> knowledge at scale is a serious challenge
- **logistics**: who pays for & prioritizes storage & computation

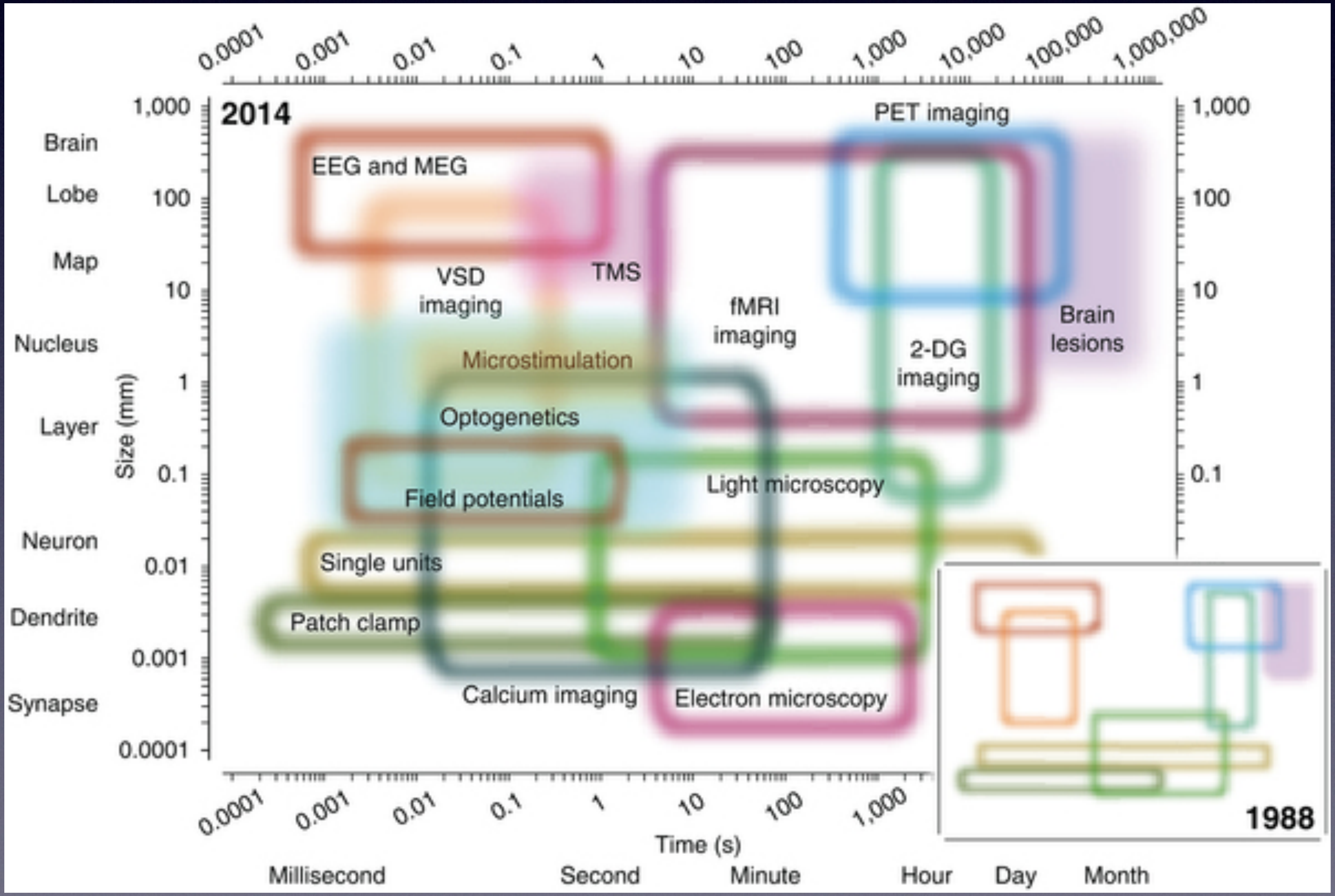
Why is it hard?

- **volume**: individual datasets larger than RAM, or local storage



Why is it hard?

- **variety**: each modality requires domain specific expertise & code

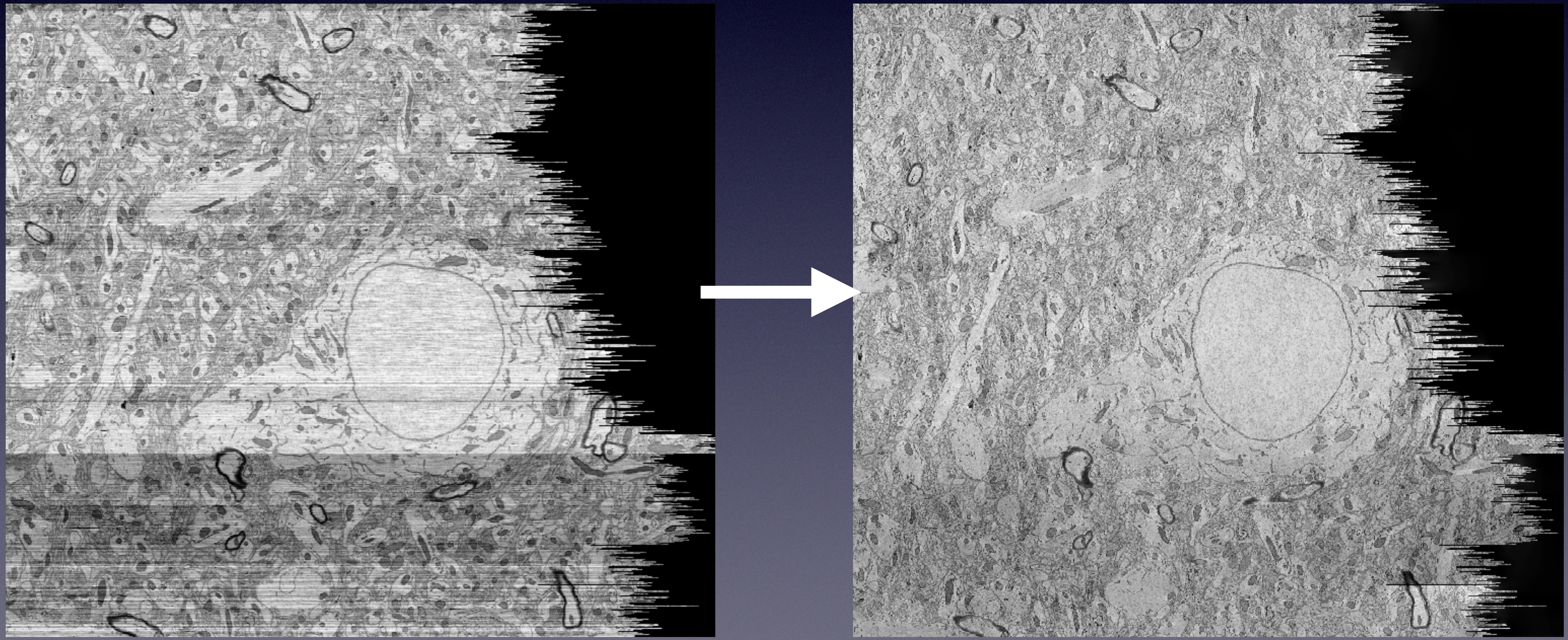


Why is it hard?

- **velocity**: some technologies can generate TB/hr per lab

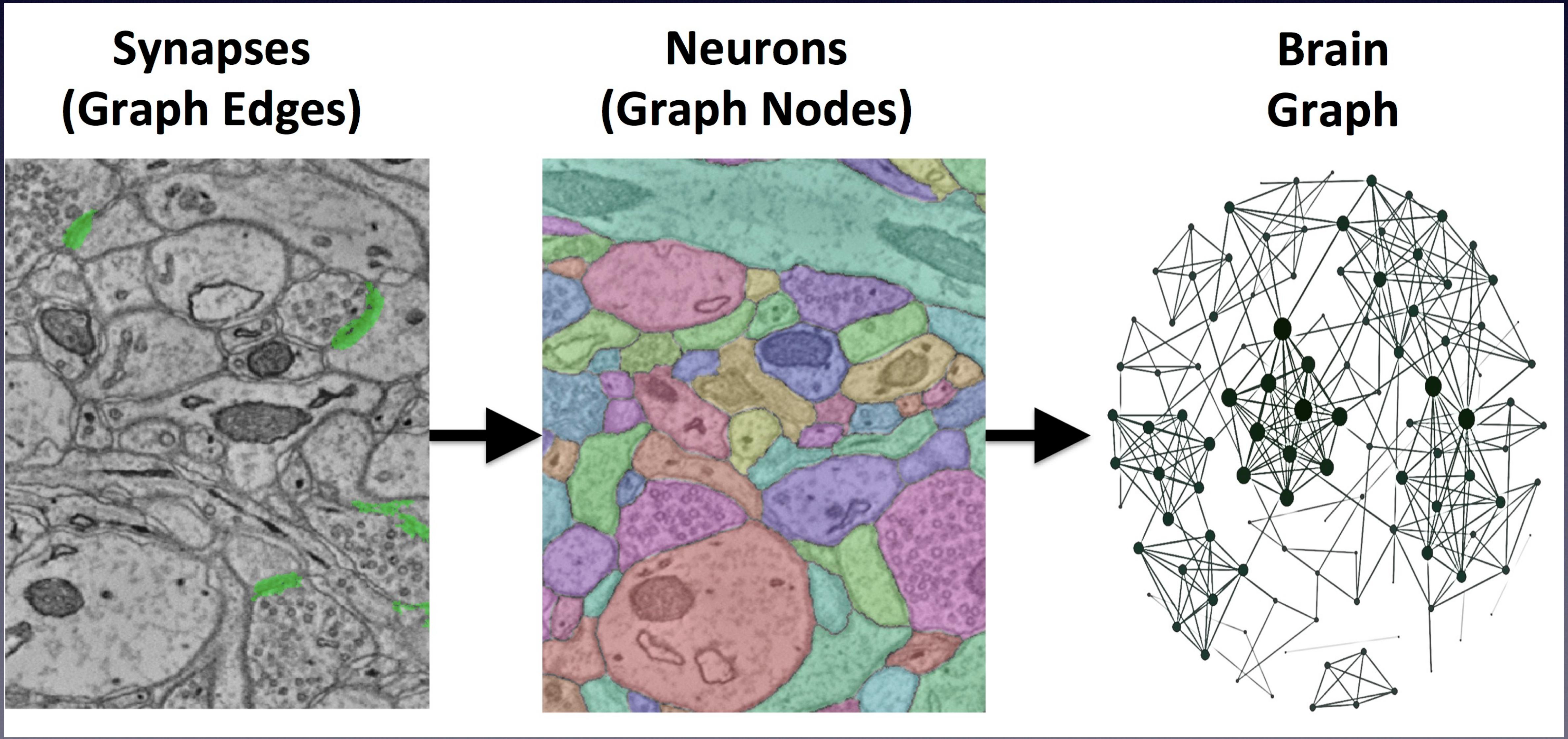
Why is it hard?

- **veracity**: data are noisy, missing, etc., analysis must be robust



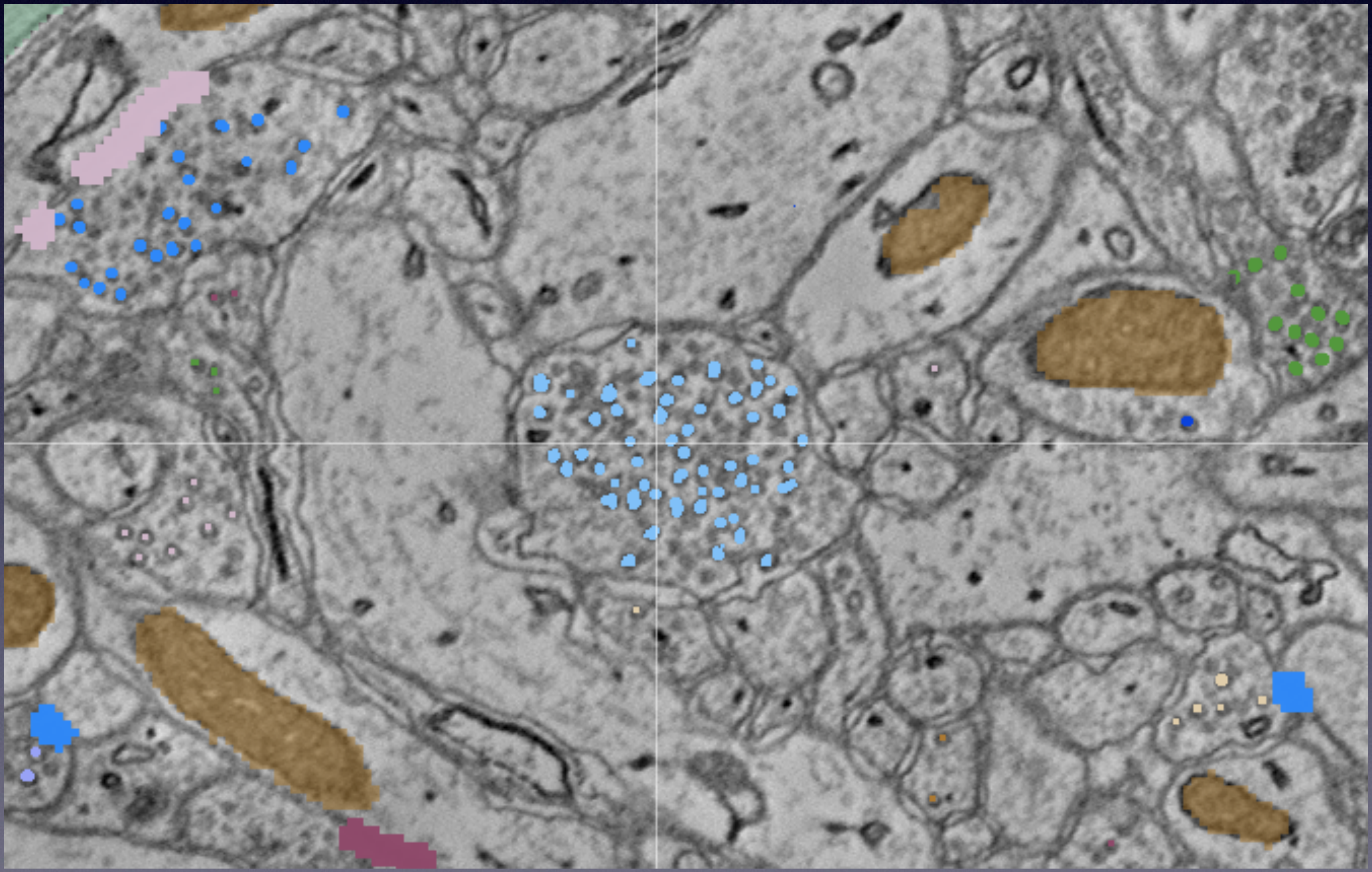
Why is it hard?

- **complexity**: raw data are “images”, useful data are semantically tagged, data —> knowledge at scale is a serious challenge



Why is it hard?

- **parallelism**: many want to both read & **write** to the data in parallel



A Potential Solution

BRAINOMIC

DOE ~~GENOMIC~~ SCIENCE

SYSTEMS BIOLOGY
FOR ENERGY AND
ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

The Brainomic Science Program uses brain and behavioral data, high-throughput analytical technologies, and modeling and simulation to develop a predictive understanding of neural systems behavior relevant to solving energy and environmental challenges including computational energy savings.

What Should it Do?

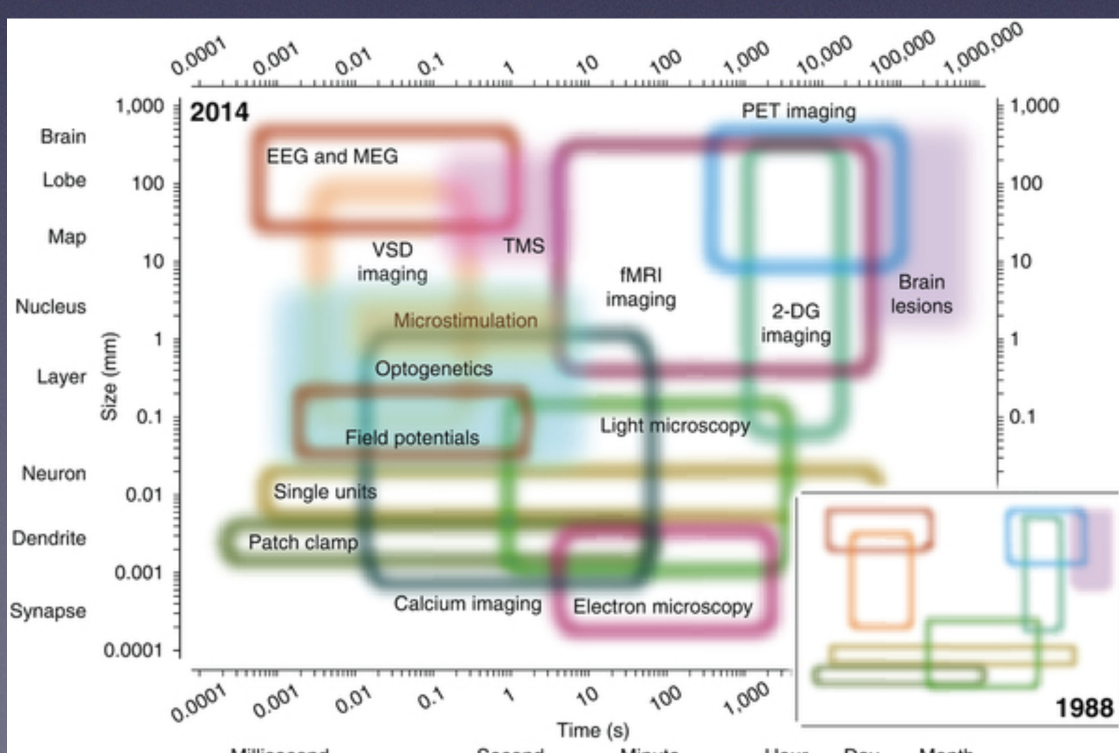
1. “anybody” can upload “anything”
2. scalable computer vision that works
3. enable efficient semantic+spatial queries
4. incorporate/interoperate with UI’s for analysis & visualization
5. statistical machine learning for big icky data

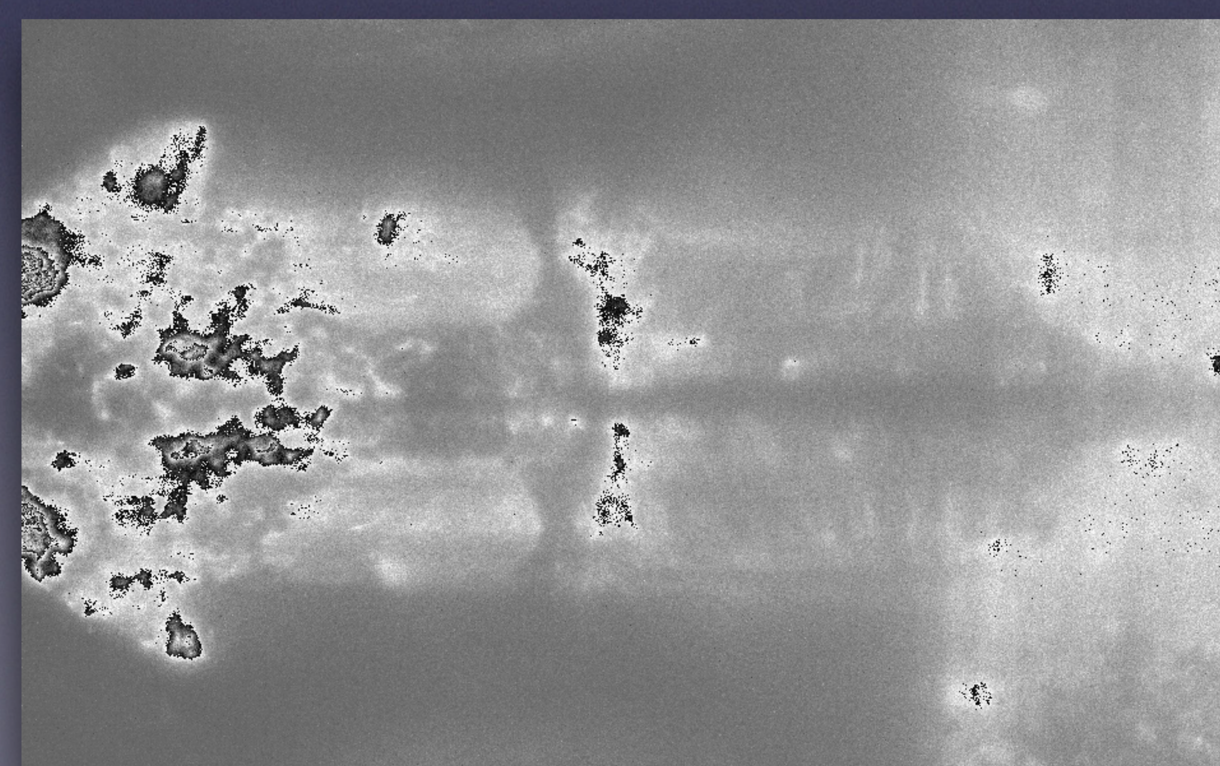
Any Example Workflow

1. upload 100 new human DTI brains
2. Estimate connectome from each
3. find AAL ROIs in 1-hop neighborhood from hippocampus in any of the new brains
4. download the shape of all such ROIs and load into R
5. build a classifier operating on these shapes

1. “anybody” can upload “anything”

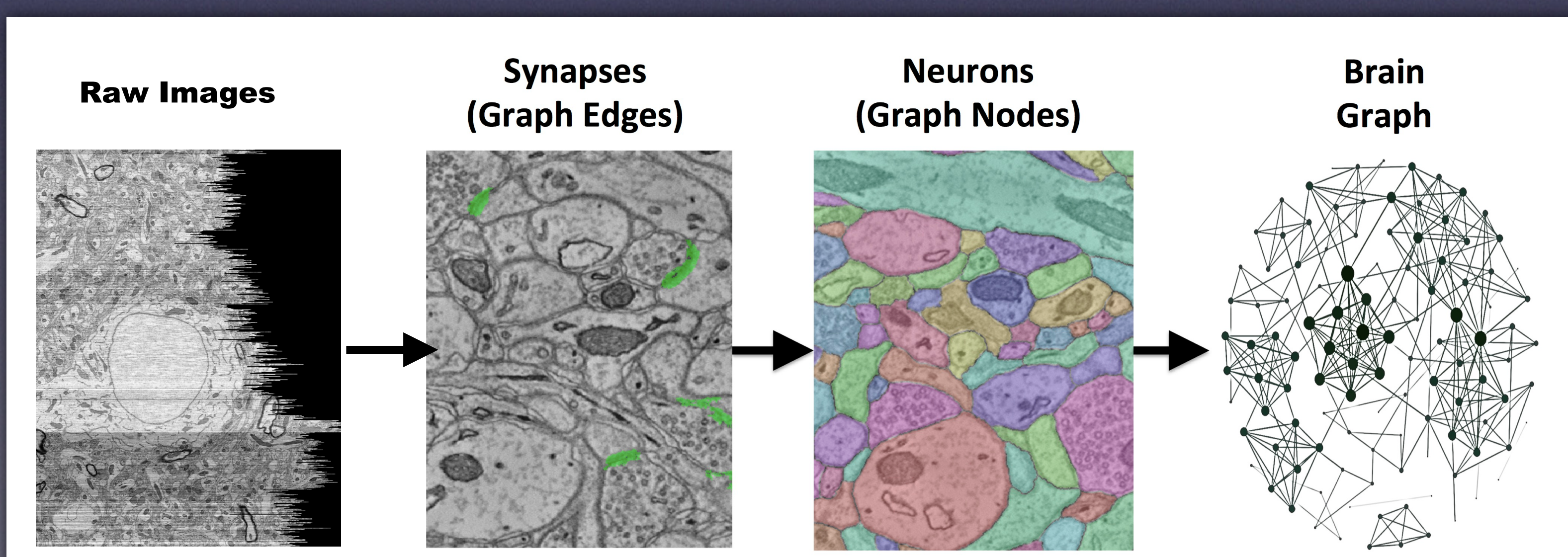
	challenge	solution
variety	each modality requires different ingest specs, database types	domain specific ingest code
velocity	data rates > 1 TB / hr / machine	streaming algorithms
privacy	some data not ready to be public, EHR	flexibility, security, local anonymization

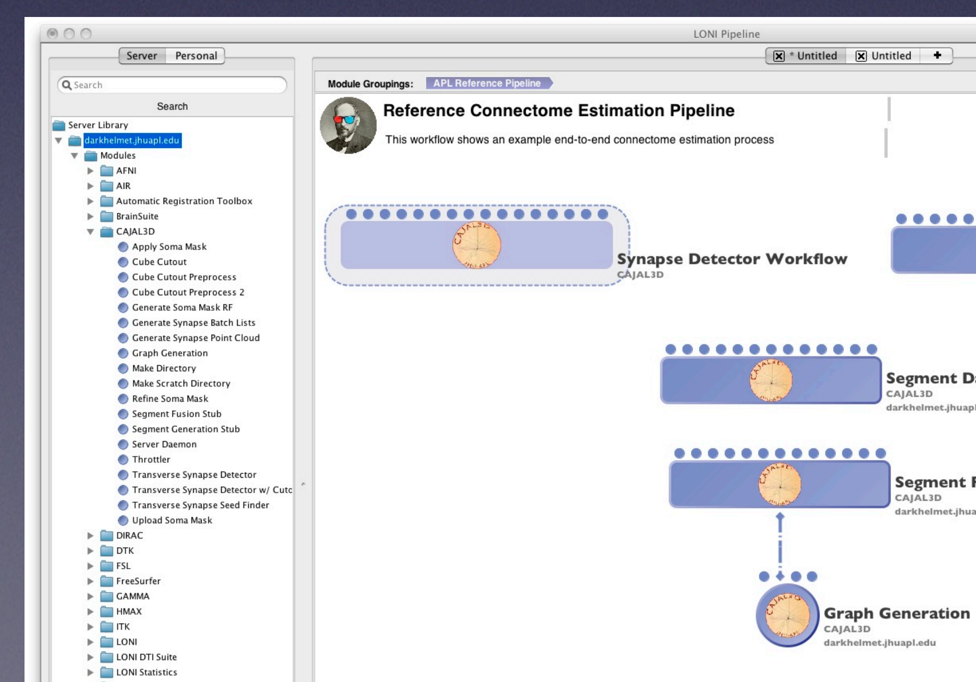




2. scalable computer vision that works

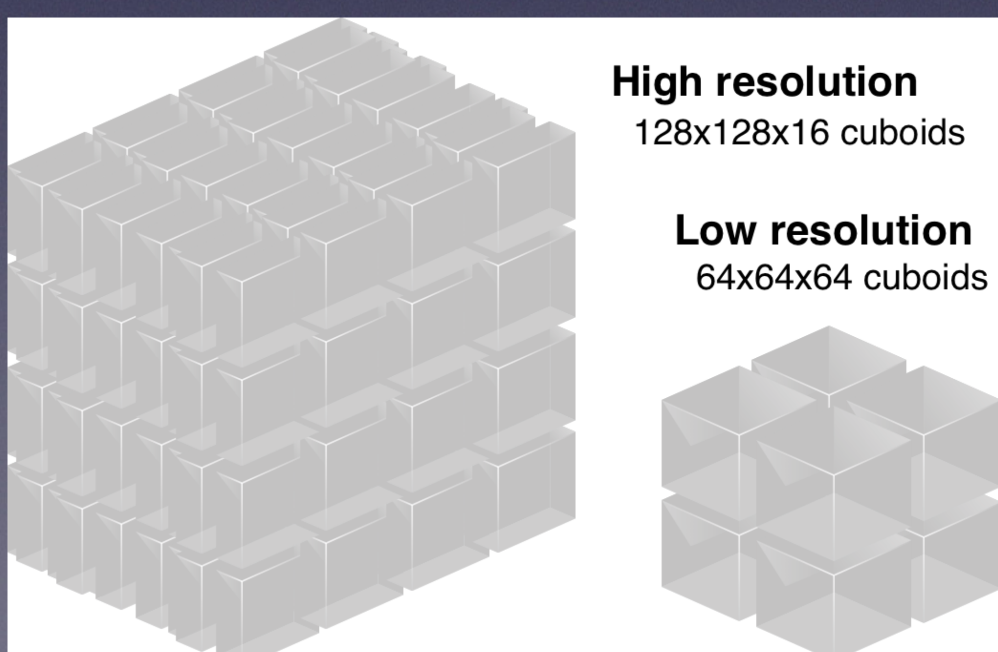
	challenge	solution
veracity	trade-offs to enable big data introduce noise & corruptions	fully automatic robust data munging
complexity	requires deep domain specific code	GALA, Rhoana

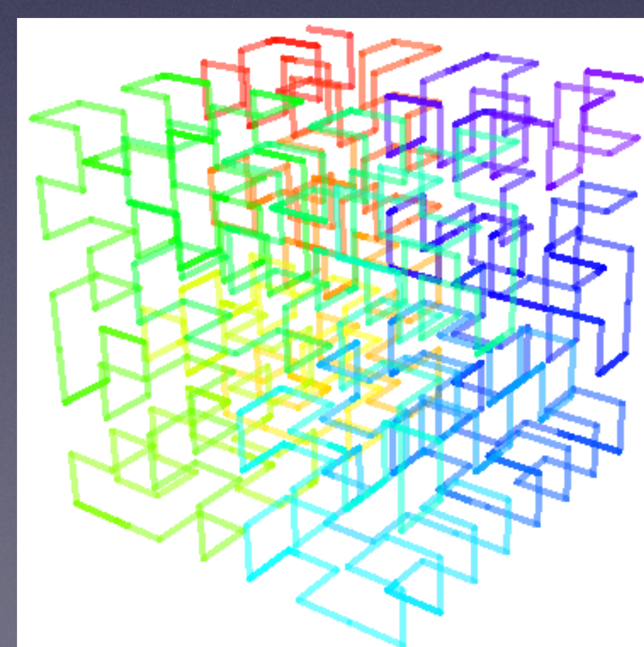


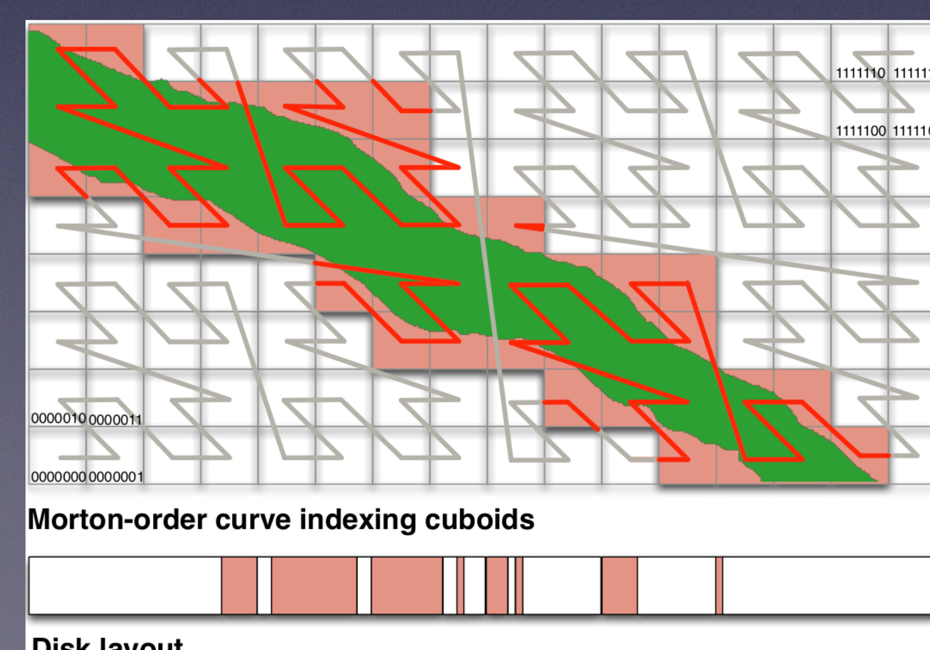


3. enable efficient semantic+spatial queries

	challenge	solution
volume	individual datasets size >> RAM / local storage, currently > 100 TB	scale-out database
cutouts	desirable to extract local subvolumes for analysis/viz	space-filling curves
locality	individual semantic objects may span the entire bounding box	sparse cutouts

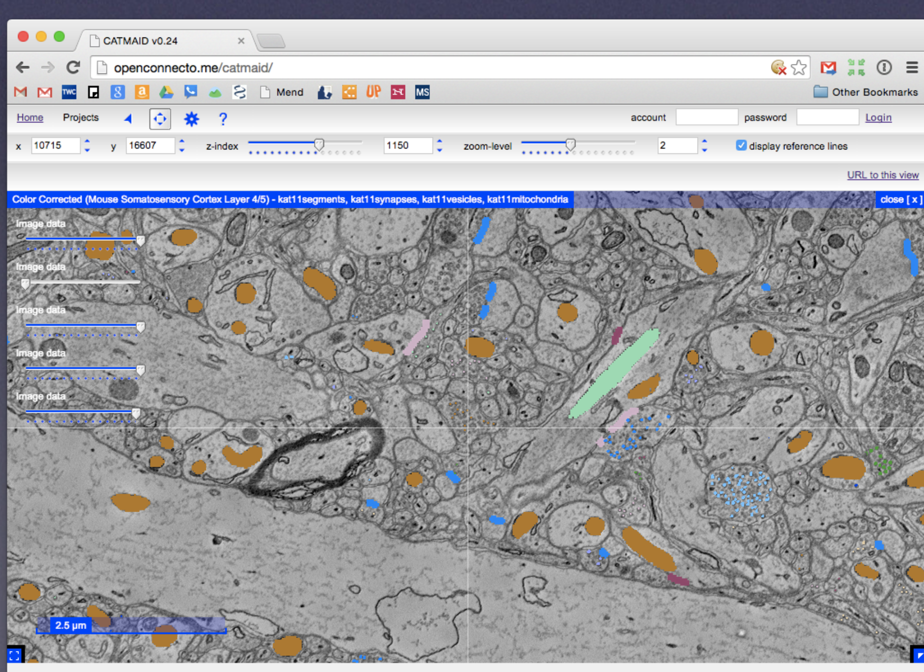


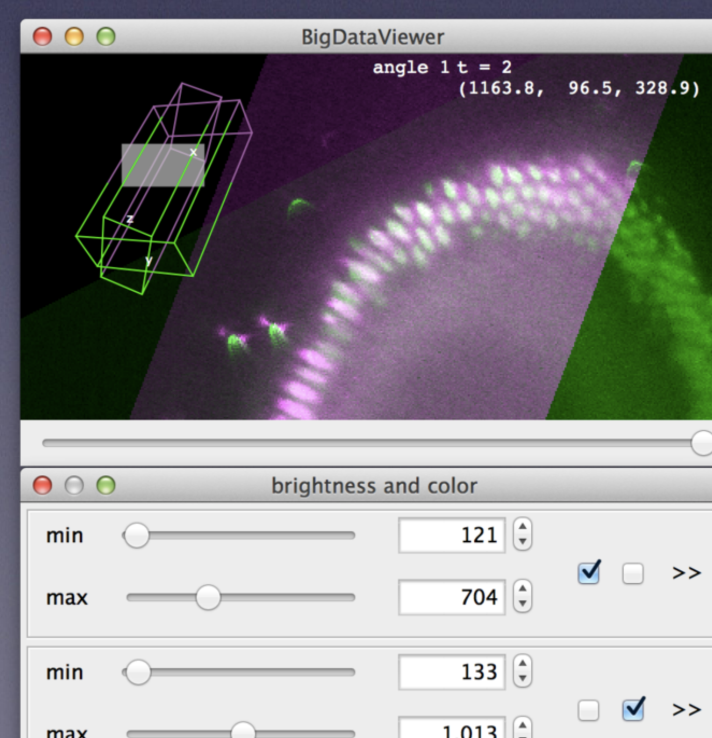


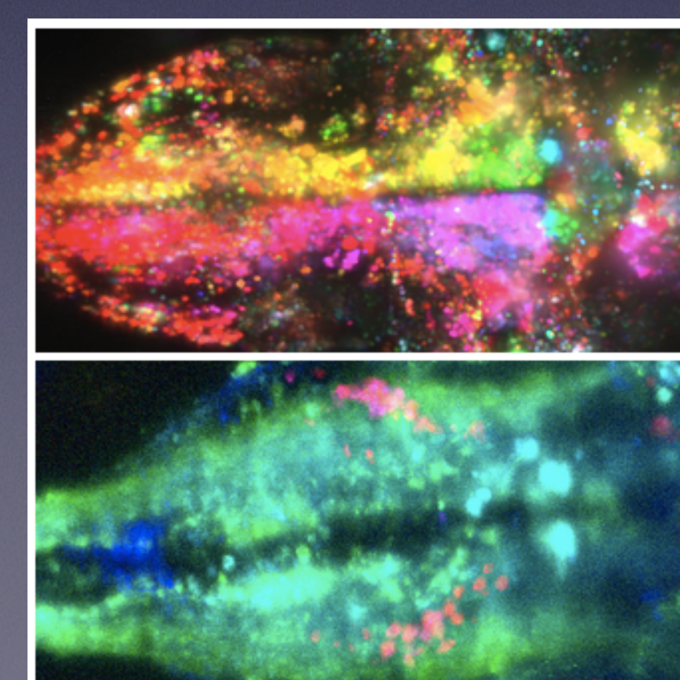


4. incorporate/interoperate with UI's for analysis & visualization

	challenge	solution
volume	interactive speed from remote databases	smart fetching/caching
user interface	simple, flexible, powerful, open	plug-ins for Fiji / ITK-SNAP
variety	different users will want to use different tools (eg, R, ITK-SNAP)	common API

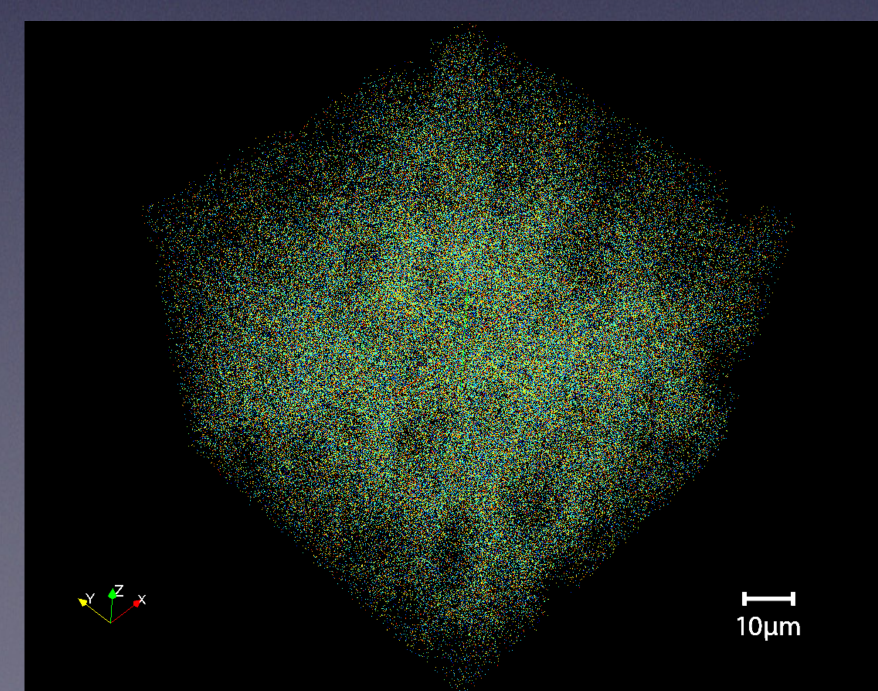






5. statistical machine learning for big icky data

	challenge	solution
complexity	objects are not vectors, they are shapes, nodes, networks, etc.	(supervised) manifold learning
scale	datasets include >10^6 objects	semi-external memory, out-of-sample embed



Thank you. Questions?

Stats	C. Priebe, M Maggioni, D Dunson, G Sapiro, B Caffo, M Miller
Code	Randal Burns, OPEN CONNECTOME PROJECT — COLLECTIVELY REVERSE-ENGINEERING THE BRAIN ONE SYNAPSE AT A TIME.
Data	C Reid, M Milham, K Deisseroth, J Lichtman, S Smith, M Ahrens
Funds	TRA (NIH), XDATA & GRAPHS (DARPA), BIGDATA & CRCNS (NIH/NSF)
Love	yummy, family, friends, earth, universe, multiverse!?

e: jovo@jhu.edu, c:443.858.9911
w: jovo.me, openconnecto.me

How to Fail

1. Work in isolation
2. Only put one person on it
3. Work on some technology you are not an expert on
4. Write code that nobody else can run
5. Work on problems that you know are publishable inside your community

How to Succeed

1. Work in close **collaboration** with a develop/user of your tool
2. Devote **>50%** of your group to solving one of these problems
3. Solve a problem for which you are already a world **expert** in a different application
4. Make sure somebody external can **reproduce** your results
5. Generate **useful** (rather than publishable) solutions

extra slides

Necessary Constituents of Ideal System

Domain Agnostic	Domain Specific
scale-out spatial database	ingest specs
UI for 2D & 3D+ viz	ingest code
UI for 2D & 3D+ annotation	robust data munging code
push/pull annotation support	robust registration code
co-registered annotation DB	robust scene parsing code
unified semantics for access	quality control capabilities
massively parallel writes	computational statistics
just queuing policy	

Necessary Constituents of Ideal System

Domain Agnostic	Domain Specific
scale-out spatial database	ingest specs
UI for 2D & 3D+ viz	ingest code
UI for 2D & 3D+ annotation	**robust data munging code**
push/pull annotation support	robust registration code
co-registered annotation DB	**robust scene parsing code**
unified semantics for access	quality control capabilities
massively parallel writes	**computational statistics**
just queuing policy	

Functionality of an Ideal

Open-Science Platforms for Heterogeneous **Brain** Data

1. eats multiple d
2. enables efficie
3. enables efficie
4. integrated with