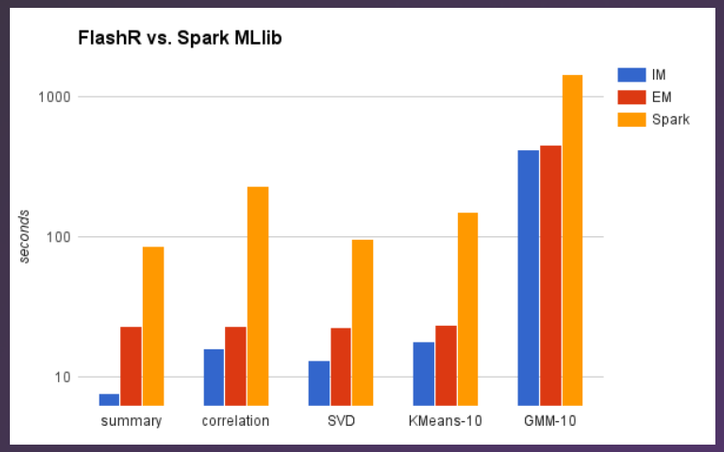
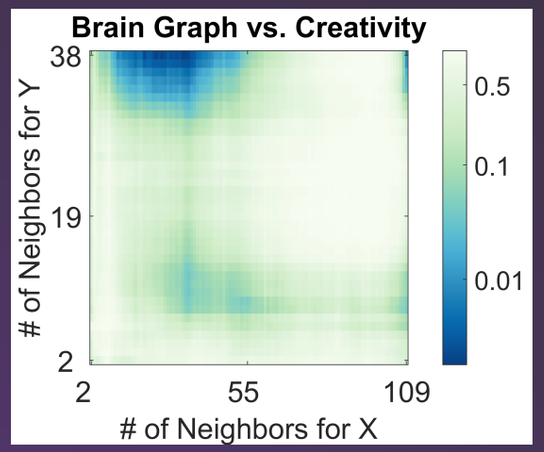
JHU SIMPLEX: Enabling Terascale Neuroscience for Everyone

(1) FlashX for Data Science

- Added a variety of generalized matrix operations to FlashX
- This enables many basic data science routines to be written in native R code but scale to arbitrarily big data on a single machine

http://flashx.io/

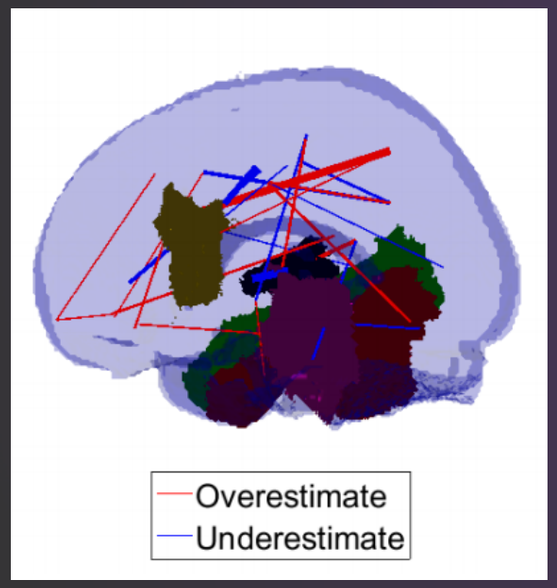


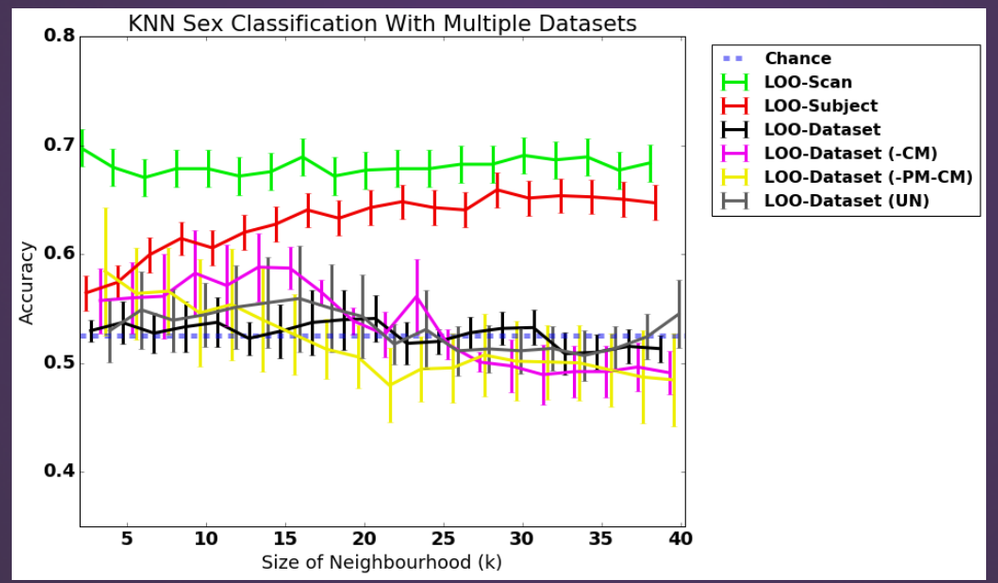


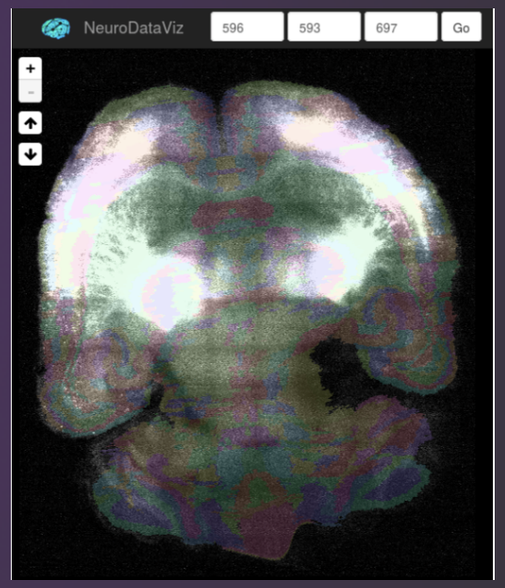
(2) MGC Dependence Testing

- Extended previous dependence testing (dcorr) by sparsifying graphs
- We now have theory and methods (in MATLAB and R) that statistically dominates the previous best method
- Use in several novel neuroscience applications to discover the scales of dependency

https://github.com/jovo/MGC







(3) Law of Large Graphs

- Proved spectral regularization is more efficient than naive estimate of of average graph
- Proved robust variant is even more efficient in the presence of outliers
- Demonstrated on real data to discover false positives and negatives in previous estimates of mean connectome

https://github.com/jhu-graphstat/LLG

(4) MR Batch Effect

- Using data data and pipeline from GRAPHS, discovered the existence of batch effects
- Simple linear approaches to removing batch effects failed
- Extending now to nonlinear models

http://m2g.io

(5) CLARITY

- Multimodal LDDMM for registering CLARITY to Atlas and other CLARITY brains
- Distributed ROI histogram Web-service
- Ingested 12 ~1TB CLARITY brains and registered them each to Allen Atlas

https://github.com/neurodata/ndreg

In summary, we provide a suite of tools and techniques which extend the boundaries of scientific discovery

FlashX for Data Science

- Existing big data analytics—for both graph traversal & machine learning—leverage cluster computing, and therefore must deal with a **communication bottleneck**, because transferring data across computers within a cluster is orders of magnitude slower than other data transfer costs

- As one ascends the memory hierarchy, the speed increases by about 10x, eg, across computers > spinning hard disks > solid state disks (SSD) > RAM > Cache

- The trick is to keep as much data in fast storage

- FlashX uses a **semi-external memory** model, we store the dense data matrix or sparse graph on a disk array in a single machine, and only keep whatever is necessary for the specific algorithm in RAM

- This **eliminates the communication bottleneck** required for cluster computing.

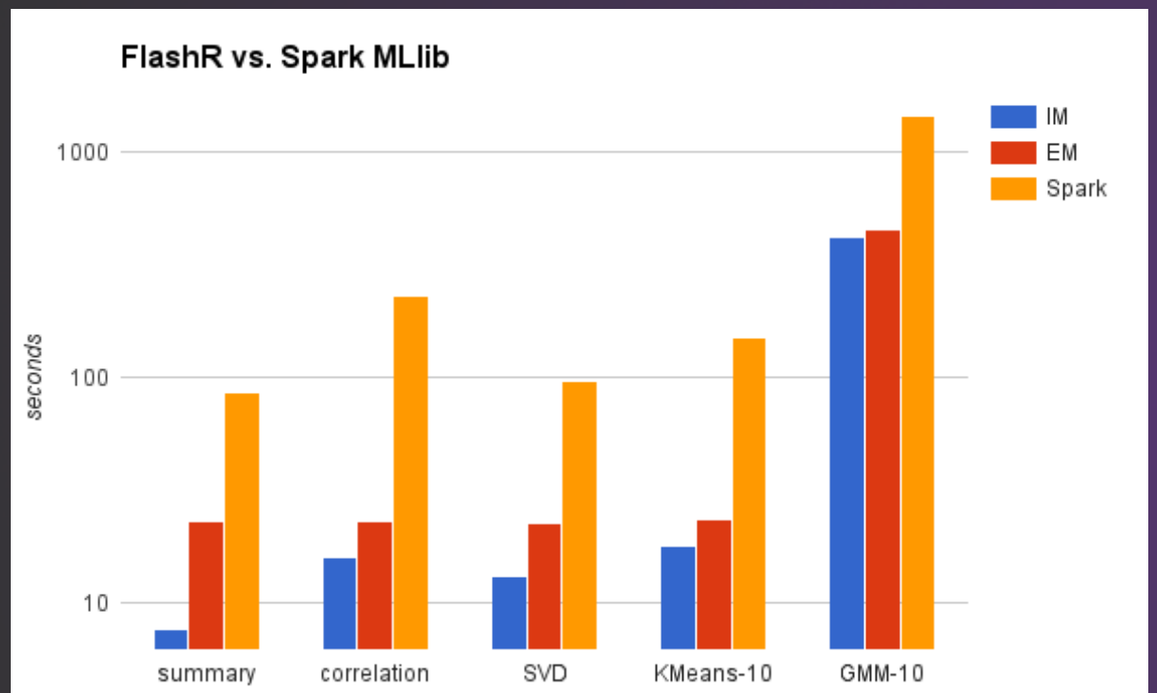
- 1 big-memory node can running FlashX outperforms

    - MLlib for machine learning on same hardware by 10x

    - 300 node cluster running Google's Pregel for graph traversal

- Machine learning algorithms can now utilize FlashX's parallel data science primitives using native R code, which will substantially accelerate the pace of making more algorithms parallel on single machines.

FlashX is ~10x Faster than MLlib



	Compuation	I/O
Summary	n*p	n*p
Correlation	n*p^2	n*p
svd	n*p^2	n*p
k-means	n*p*k	n*p
gmm	n*p^2*k + p^3*k	n*p + n*k

For each experiment, n=1B, p=32,
summary computes: min, max, mean, L1 norm, L2 norm, nnz

MGC for Dependence Testing

- Dependence testing is the first step in many data science problems
- Multiscale Generalized Correlation (MGC) is the first approach that works well in **high-dimensions, low sample size, arbitrary objects** (shapes, graphs, time-series), and provides insight into **scales of dependency.**
- Combined ideas from correlation testing and manifold learning
- First theory to prove **statistical dominance**, meaning that we are always better than (or at least as good as) previous methods.

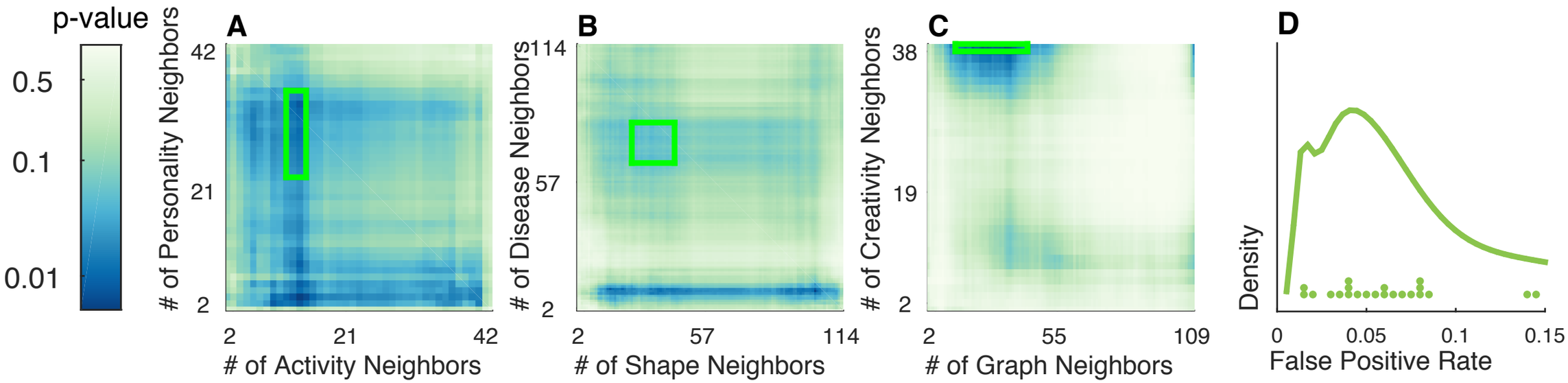
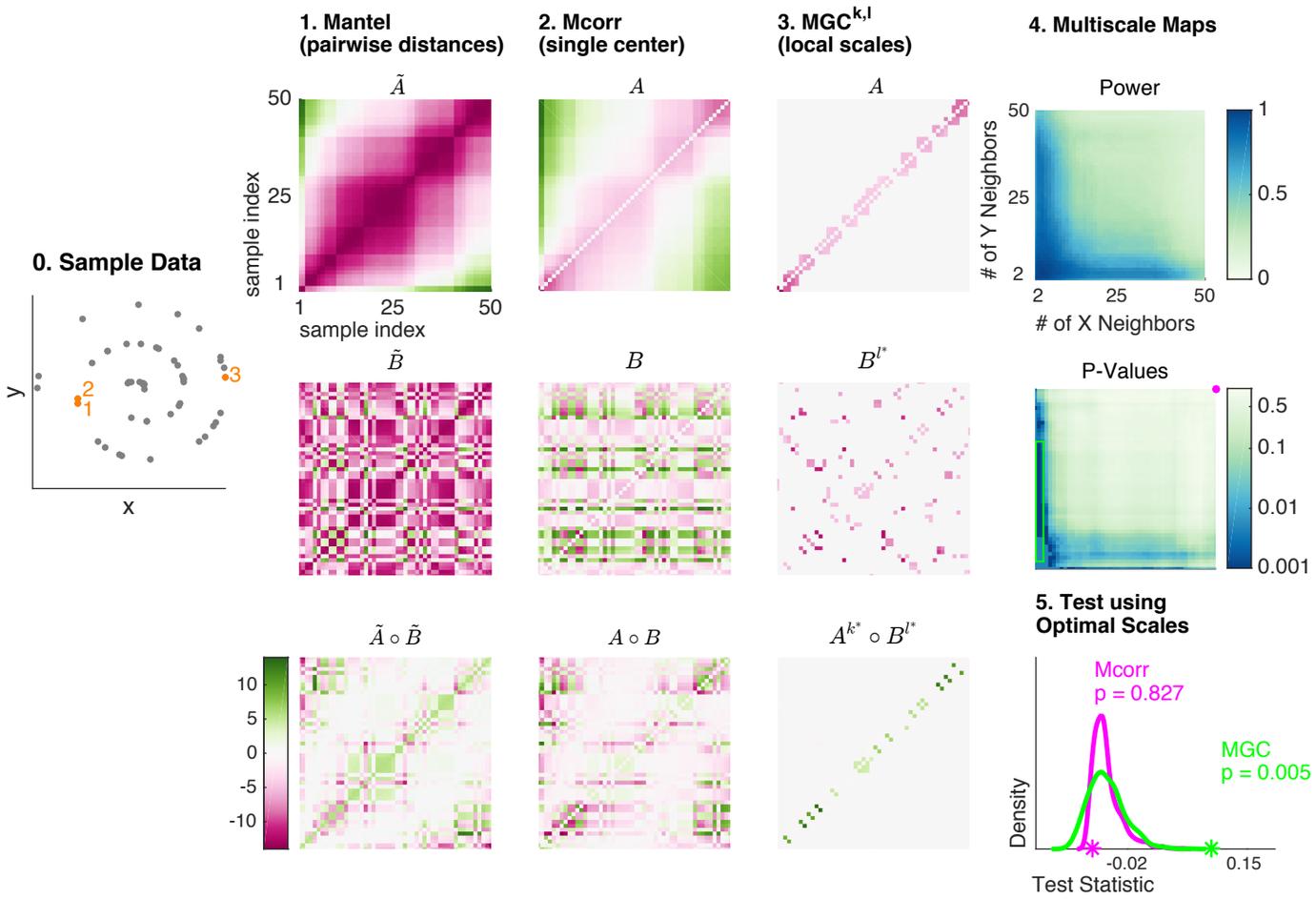


Figure 5: MGC discovers the scales of dependence between various brain and mental properties when they exist, and does not detect dependence when it does not exist. The left three panels show multiscale p-value maps and their corresponding estimated optimal scales for three different settings: **A** brain activity vs. five-factor personality model, **B** brain shape vs depressive disease, and **C** brain networks vs. creativity. Sample size is 42, 114, and 109, respectively, though the ordinate of these panels only goes as high as the largest possible neighborhood size due to repeated entries. For all three, MGC yields as low or lower p-value as the global test and reveals the optimal scales of dependence (green rectangles). (D) Density estimate for the false positive rates of sample MGC on the brain vs synthetic independent noise experiments, with the actual rate of each experiment shown as dots above the x-axis. The mean $\pm$ standard deviation is 0.0576 ± 0.0452 respectively, demonstrating that for these real data, MGC is a valid test.



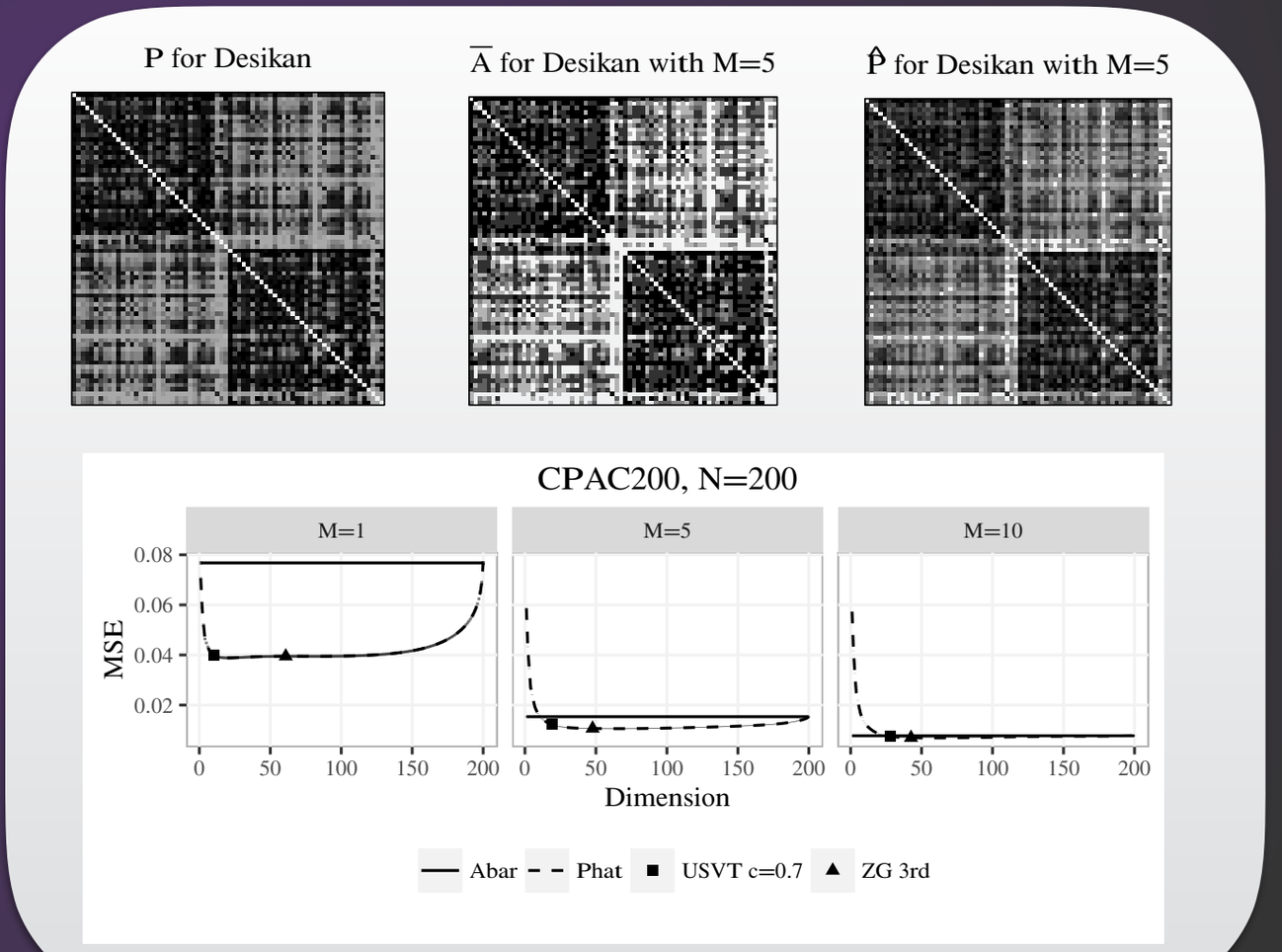
δ_x(1,2)	-2.42	-5.21	-5.07
δ_y(1,2)	-1.58	-0.91	-0.12
$\delta_x \times \delta_y$	3.82	4.74	0.61
δ_x(2,3)	0.70	0.61	0.14
δ_y(2,3)	-0.91	-0.28	0.12
$\delta_x \times \delta_y$	-0.63	-0.17	0.02
$\sum \delta_x \times \delta_y$	-162.14	-93.04	116.41
test statistic	-0.02	-0.02	0.16

Figure 1: (caption on next page.)

Figure 1: Flowchart schematizing Multiscale Generalized Correlation (MGC). Columns listed from left to right. **Column 1:** 50 pairs of observations (x_i, y_i) are nonlinearly (spirally) dependent on one another. **Column 2:** Compute all pairwise distances for x and y yielding interpoint comparison matrices $\tilde{A}$ (top) and $\tilde{B}$ (middle), and their element-wise product $\tilde{C}$ (bottom), whose sum is the MANTEL statistic [7]. Note that implementing it requires choosing appropriate distances for both x and y. **Column 3:** Double centering—subtracting the row-sums and column-sums to eliminate bias due to individual samples—yields $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$, which we use to compute C, whose sum is the un-normalized MCORR statistic [10]. **Column 4:** Rank truncating yields A^k, B^l, and $c^{k,l}$ at $k = l = 4$. k and l can be chosen using either the multiscale p-value or power map, MGC is the sum of the elements of $c^{k,l}$, which detects dependence only when the sum is large. **Column 5:** (Top) The empirical null distribution for MCORR, as well as our MGC, and the corresponding observed test statistics for each. Multiscale maps are used to determine the optimal scales, using p-values (middle) in the absence of the true distribution or training data, and simulated power (bottom) when the true distribution or training data are available. Whereas MCORR, the global test, has very low power and therefore yields a non-significant p-value (0.827), there are many local scales that achieve nearly perfect power, resulting in highly significant p-values (≈ 0.005), as well as revealing the scales of dependency. Table illustrating how MGC is able to detect dependence even in highly nonlinear and low-sample size settings. The three colored points in the scatter plot of Figure 1 indicate the three points considered in this table. MGC detects local dependence across x and y, whereas the global methods get confused by many nonlinearly related pairs.

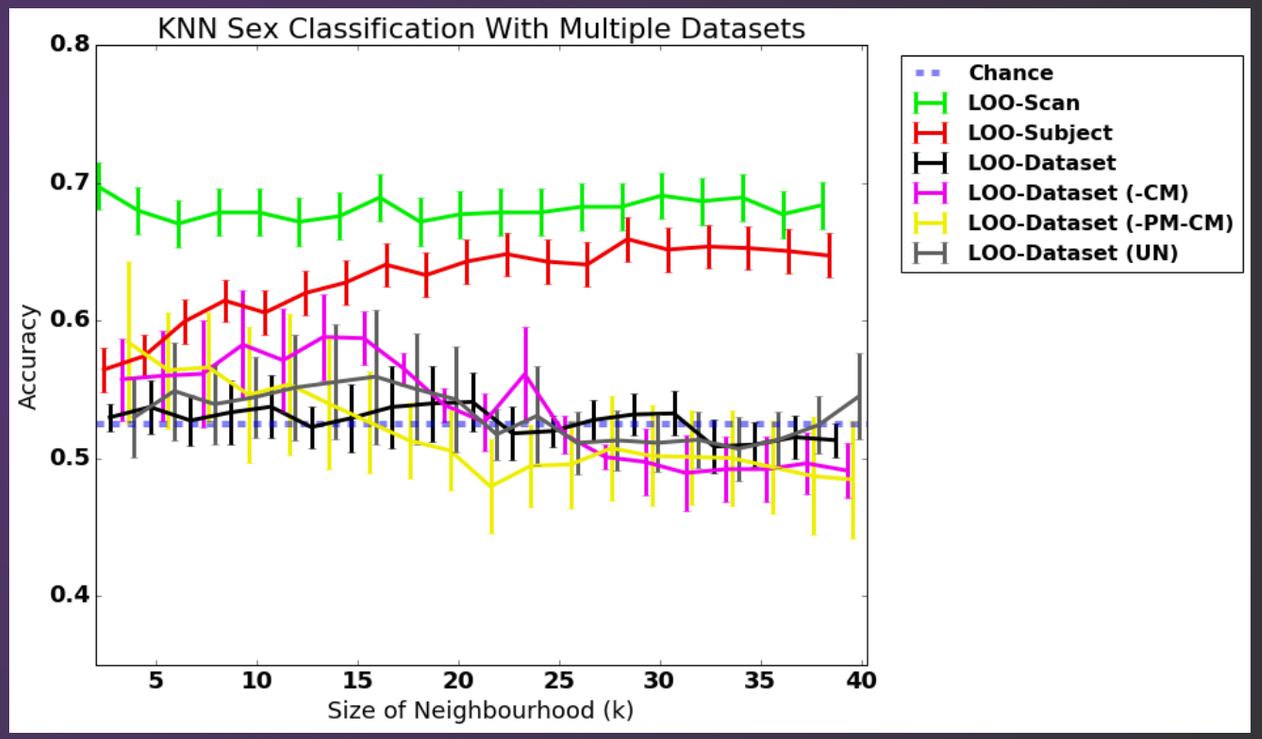
Law of Large Graphs

- Average connectome is a key result of the $40M Human Connectome Project
- For nearly all studies, sample size is small, and for all clinical studies, cohort size is even smaller (~5)
- Proved our **low-rank estimator** is asymptotically more efficient (smaller errorbars) than the naive estimator
- In practice, our low-rank estimator for connectomes yields **better estimates whenever sample size is < 10** (which is typical)

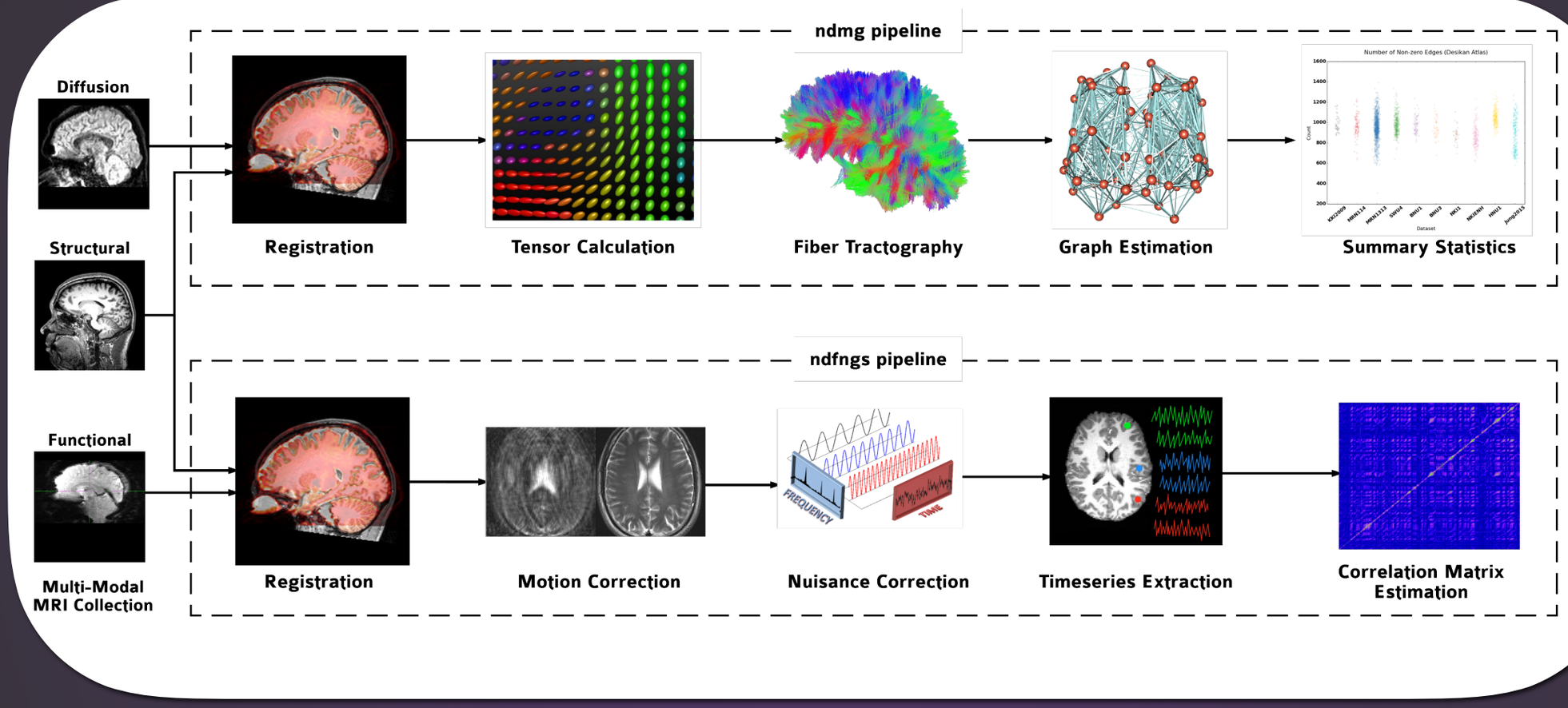


MRI Batch effect

- Batch effects are sources of variance due to experimental conditions
- For clinical utility, batch effects must be mitigated
- Removing batch effects is the key statistical step to move from research to clinic
- We demonstrate using 15 different datasets processed the exact same way that **batch effects completely eliminate signal**
- Standard methods for batch removal failed



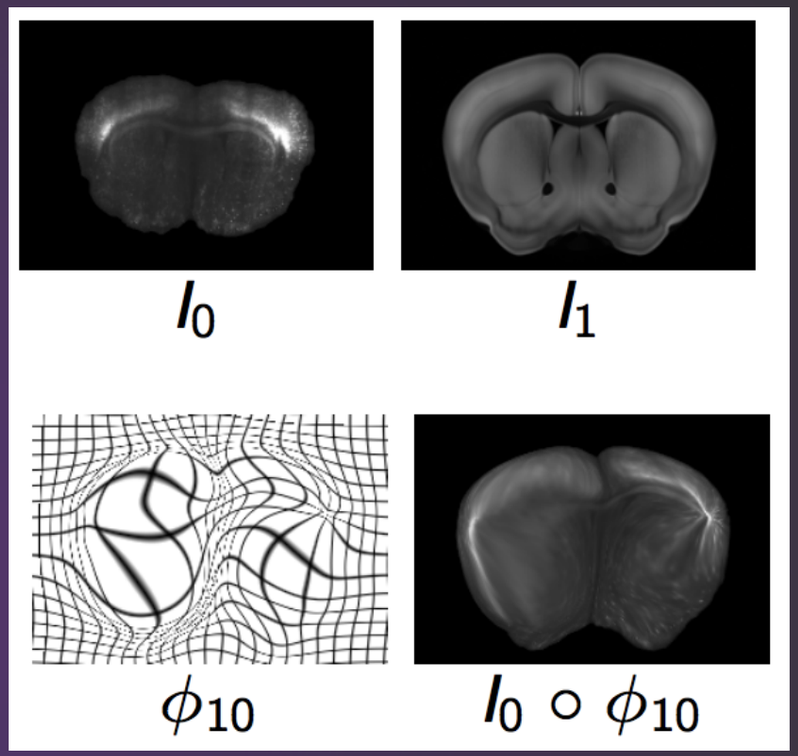
Multimodal MRI Pipeline

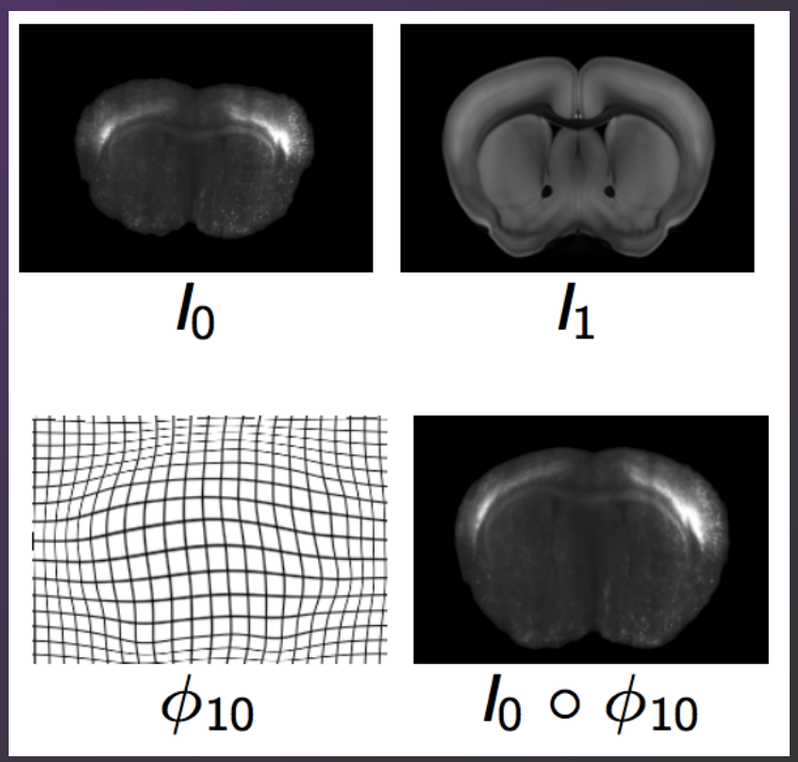


- Pipeline is open source, fully reproducible and distributed: https://github.com/neurodata/ndmg
- Can be 1-click launched via Docker or Singularity: https://github.com/neurodata/ndmg/blob/master/Dockerfiles/README.md
- Optimized to maximize subject discriminability & all subsequent inferences
- Requires about 1 hr per subject per core to run.

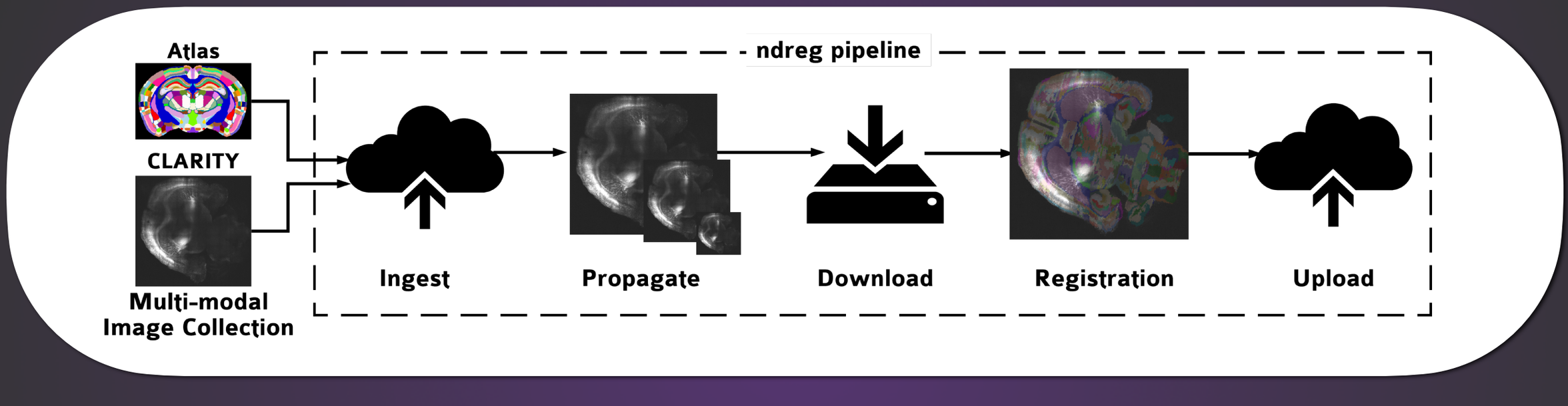
Multimodal Nonlinear Registration

- CLARITY brains are increasingly important
- Registering to one another and histology atlases is key to understanding
- Multimodal data are bright in different places
- Squared Error Nonlinear registration (LDDMM) maps bright spots to bright spots, which fails (top figure)
- **Mutual information LDDMM** solves this problem (bottom figure)





Multimodal Image Pipeline



- CLARITY datasets are 1-10 terabytes (TB), so we store data in spatial database in cloud
- Atlases are much smaller, and different modalities, so registration require multimodal techniques and resampling
- Requires about 3 hr per per brain per core to run.