Angle-based Joint and Individual Variation Explained (AJIVE)

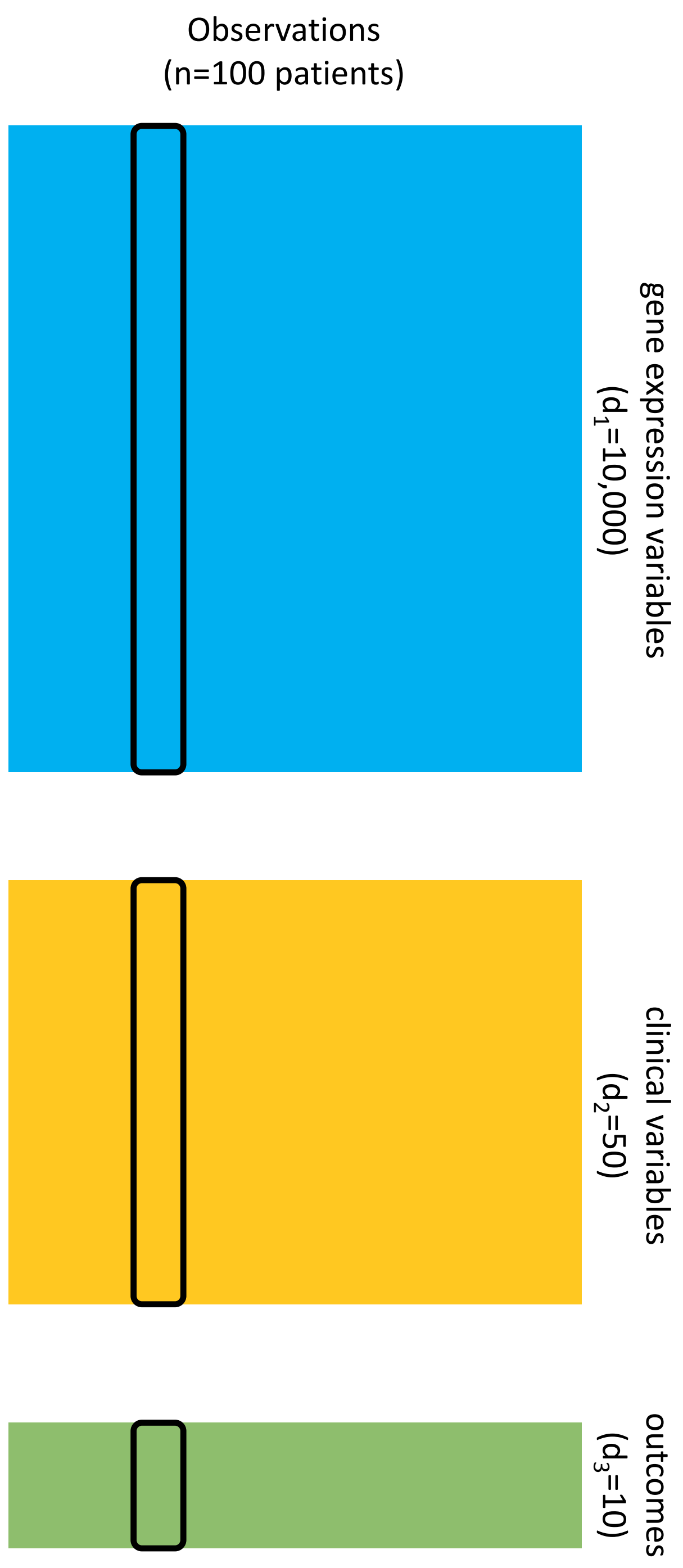
NSF Big Data Grant 1633074 Statistical Approaches to Big Data Analytics
Investigators: J. S. Marron, Jan Hannig, Iain Carmichael
University of North Carolina, Chapel Hill, Dept. of Statistics & Operations Research

Project Description and Impacts

Data integration, **feature extraction** and **compare/contrast** capabilities for multiple, heterogeneous data blocks

Multi-block data: fixed set of observations, multiple sets of variables

- Gene expression, mutation, copy number, proteins, …
- Tumor H&E images and genetic data
- fMRI and behavioral scores
- Text and image data
- Citation network and text documents



Applications in

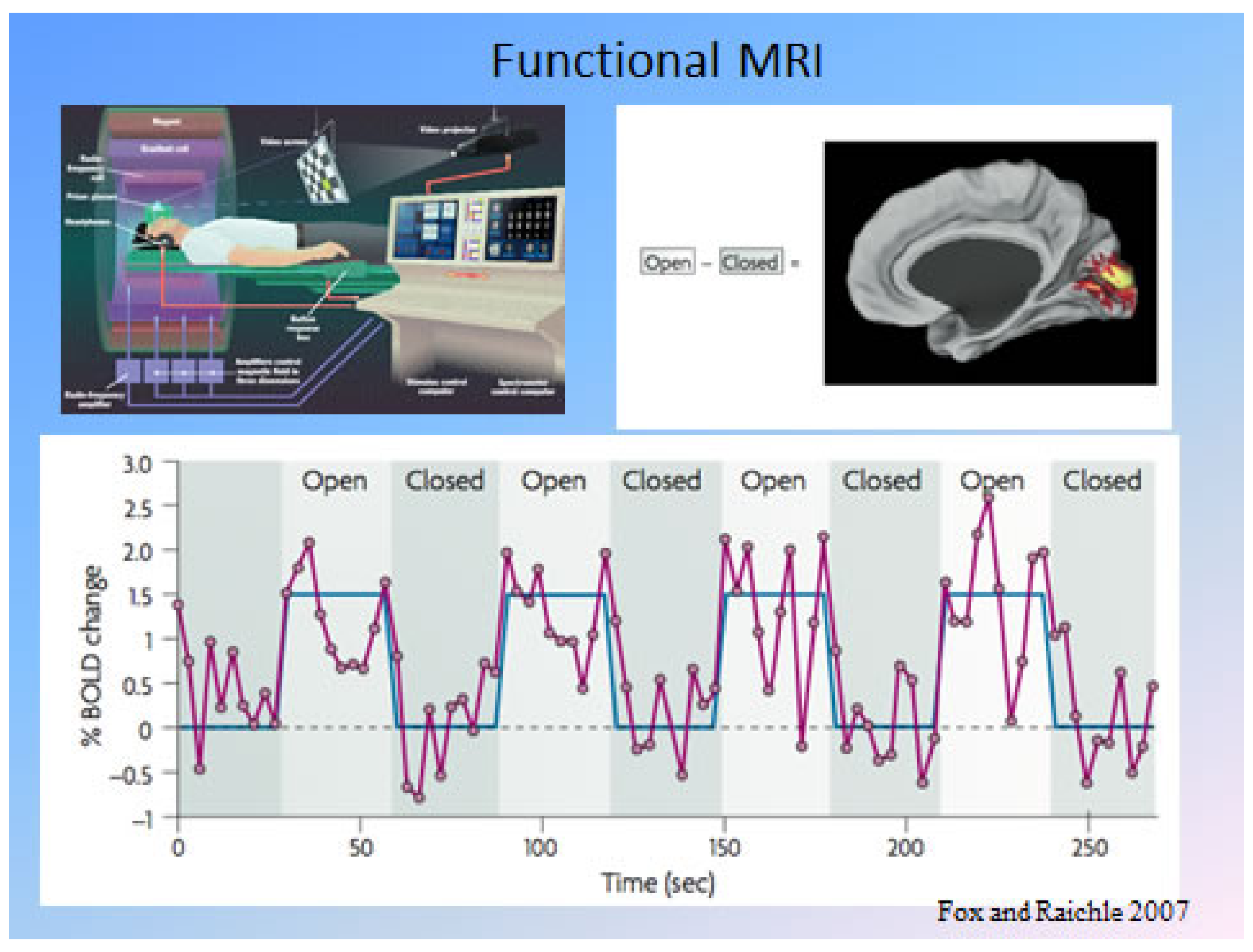
- Cancer genetics
- Neuroscience
- Natural language processing
- Medical image analysis
- Multi-modal machine learning

Unique challenges for multi-block setting

- Statistical inference
- Non-orthogonal decomposition
- Wildly different block scaling and dimensionality
- Heterogeneous signals (block individual, block joint, partially shared)
- Multiple, non-standard rank estimation problems
- Batch effects

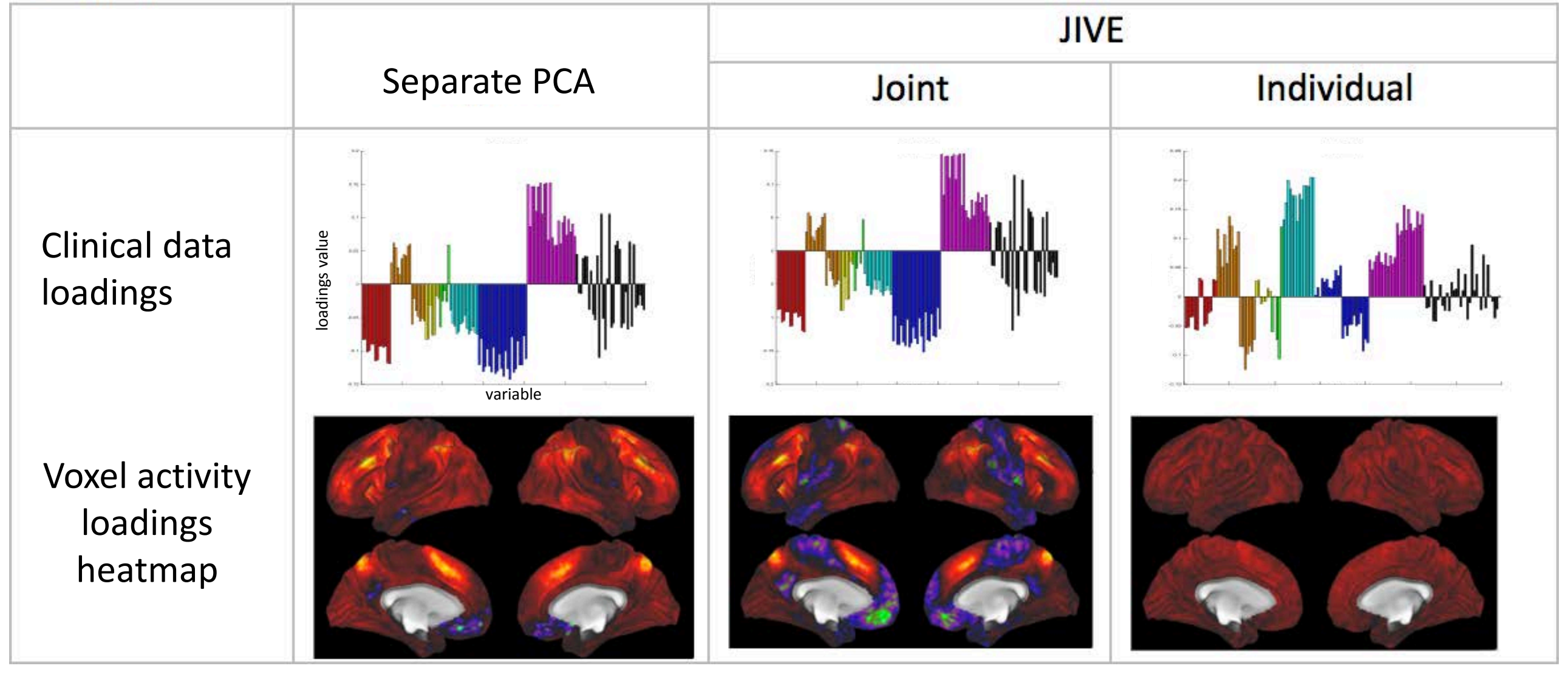
Functional MR Imaging

Disentangle important modes of variation



Fox and Raichle 2007

Visualize Using PC Loadings



Yu et al, 2017

Approach

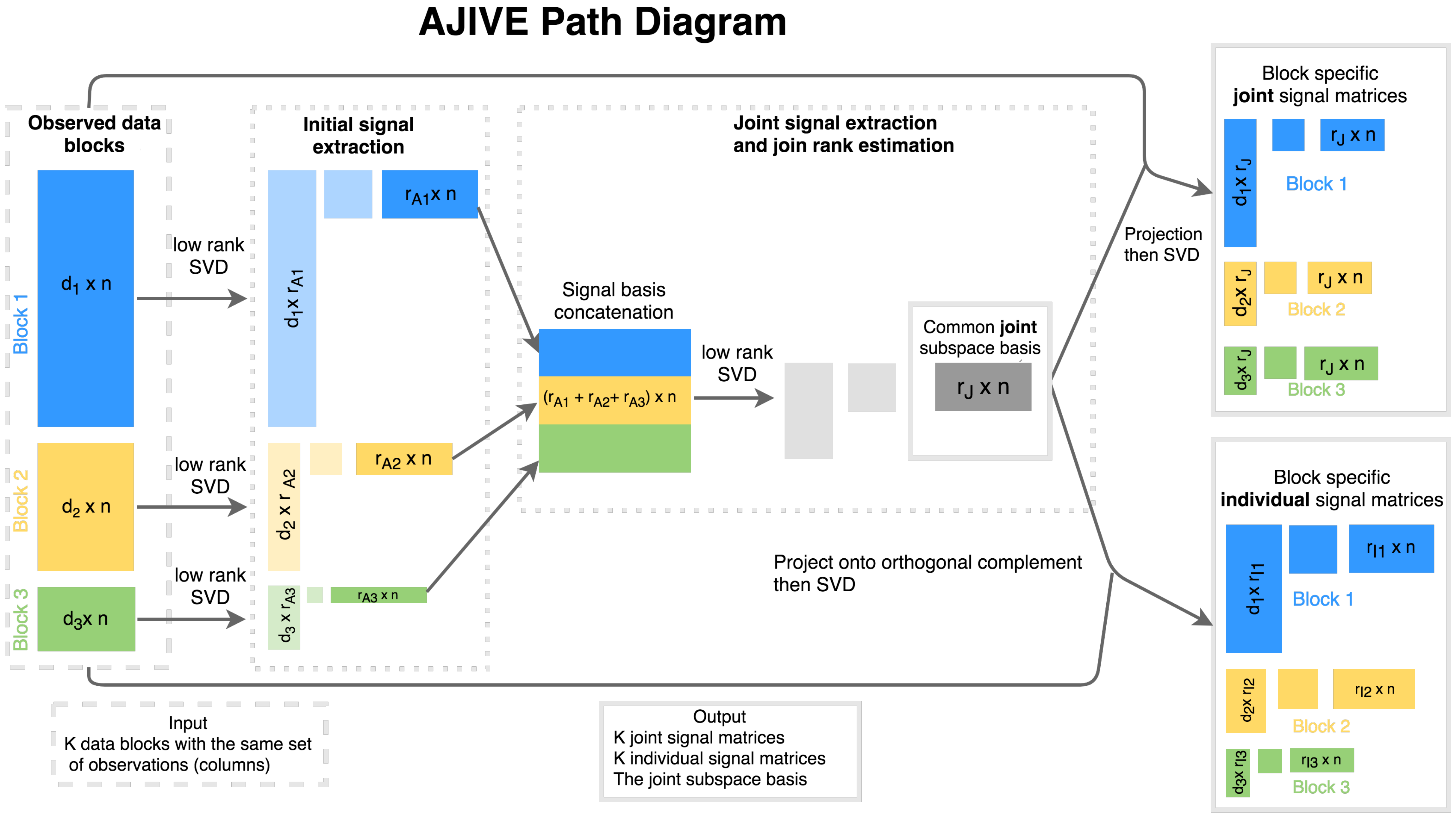
Statistical inference to identify **multiple kinds of shared signals**

- Shared by all blocks
- Partially shared (e.g. blocks 1, 3, 5)
- Individual (e.g. present in block 2 only)

Mathematical/statistical tools

- Principal angle analysis
- Singular Value Decomposition
- Second order cone programming
- Perturbation analysis
- Bootstrap

AJIVE Path Diagram



Progress and Future Directions

Partially shared block analysis via **new perturbation framework**

- New direction-based approach
- Bootstrap based improvement to Wedin bound
- Major improvement for non-square matrices

Improved **methodology and computation**

- Manifold optimization
- Difference of convex functions (DC) programming

Future applications

- Supervised JIVE for cancer genetics
    - Incorporate subtype information
    - Clinical outcomes/survival
- Deep learning integration with JIVE
    - Medical image analysis and genetic applications
    - Search for genetic drivers
    - Interpretation of extracted features
    - True integration
- Bayesian JIVE (Zhao et al, 2016)
    - Novel priors
    - Non-orthogonal partially shared blocks
    - Subspace based sampling
- Continuous time data blocks
    - Language evolution with word embeddings
    - High-frequency financial data

References

Feng, Q., Jiang, M., Hannig, J., & Marron, J. S. (2018). Angle-based joint and individual variation explained. Journal of Multivariate Analysis, 166, 241–265.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics*, *7*(1), 523.

Yu, Q., Risk, B. B., Zhang, K., & Marron, J. S. (2017). JIVE integration of imaging and behavioral data. *NeuroImage*, *152*, 38-49.

Zhou, G., Cichocki, A., Zhang, Y., & Mandic, D. P. (2016). Group component analysis for multiblock data: Common and individual feature extraction. IEEE transactions on neural networks and learning systems, 27(11), 2426-2439.

Zhao, S., Gao, C., Mukherjee, S., & Engelhardt, B. (2016). Bayesian group factor analysis with structured sparsity. Journal of Machine Learning Research, 17, 1–47.

NSF

UNC Statistics & OR