Statistical Data Integration for Large-Scale Multi-Modal Medical Studies

Genevera I. Allen

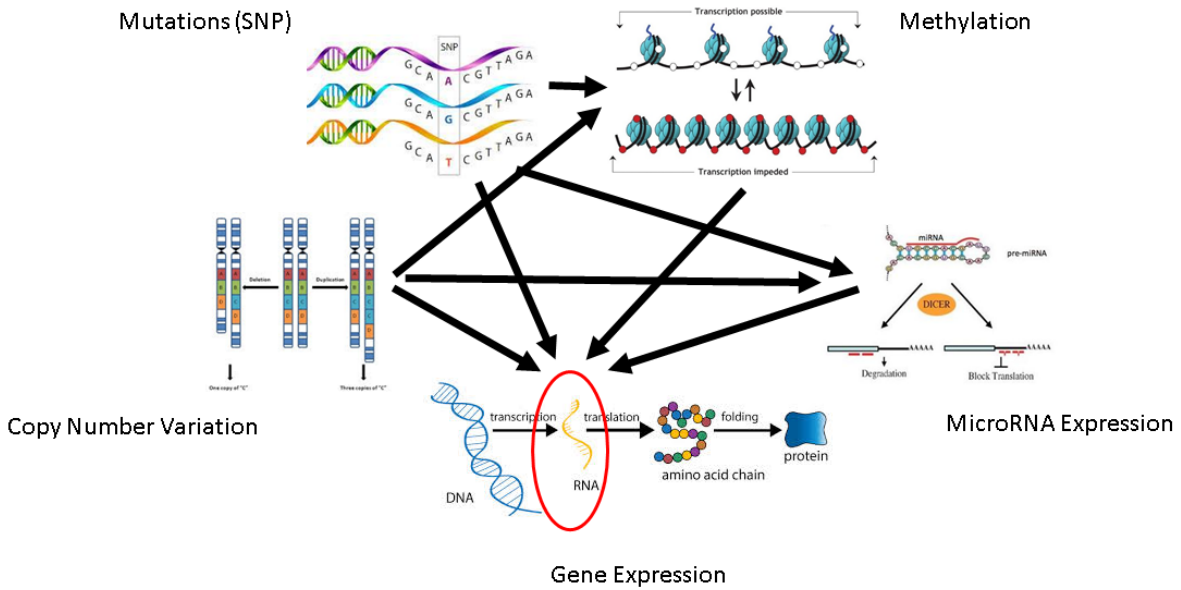
Dobelman Family Junior Chair,
Department of Statistics and Electrical and Computer Engineering, Rice University,
Department of Pediatrics-Neurology, Baylor College of Medicine,
Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital.

June 8, 2016

Data Integration for Medical Studies

Combine Different Types of Medical Data ...

- Clinical.
- EHRs.
- Social.
- Behavioral.
- Environmental.
- Genetics.
- Proteomics.
- Neuroimaging.
- Metabolomics.
- Microbiome.

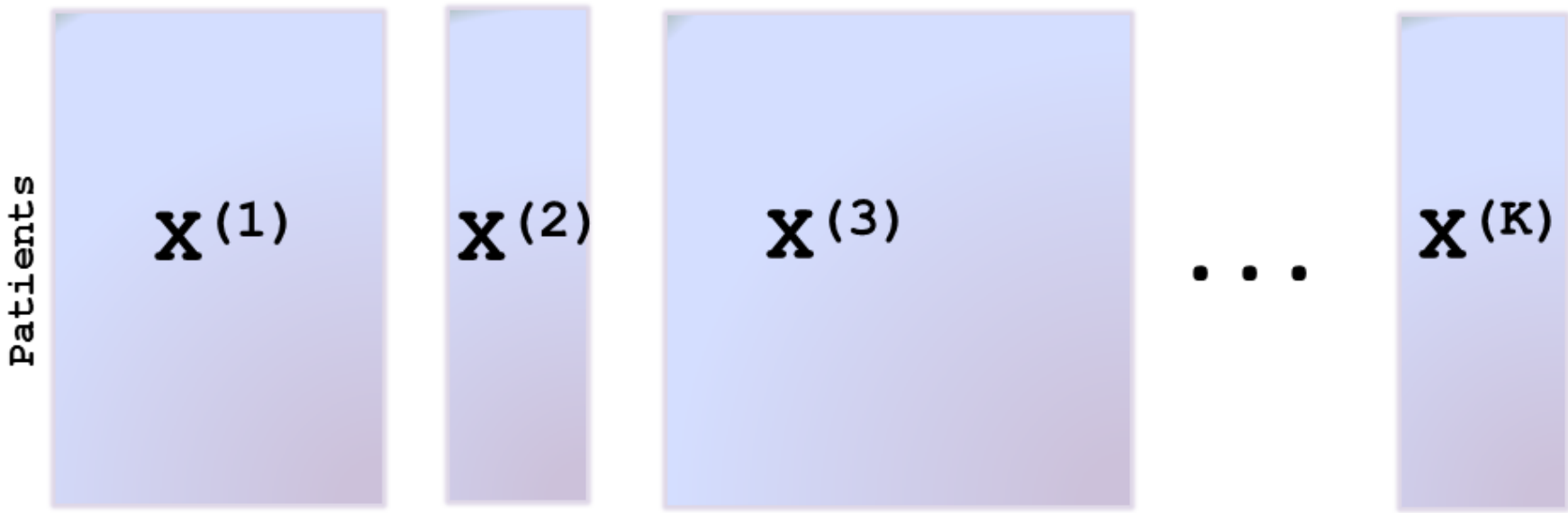


... To Better Understand Complex Diseases.

Multi-Modal Data

Multi-Modal Data

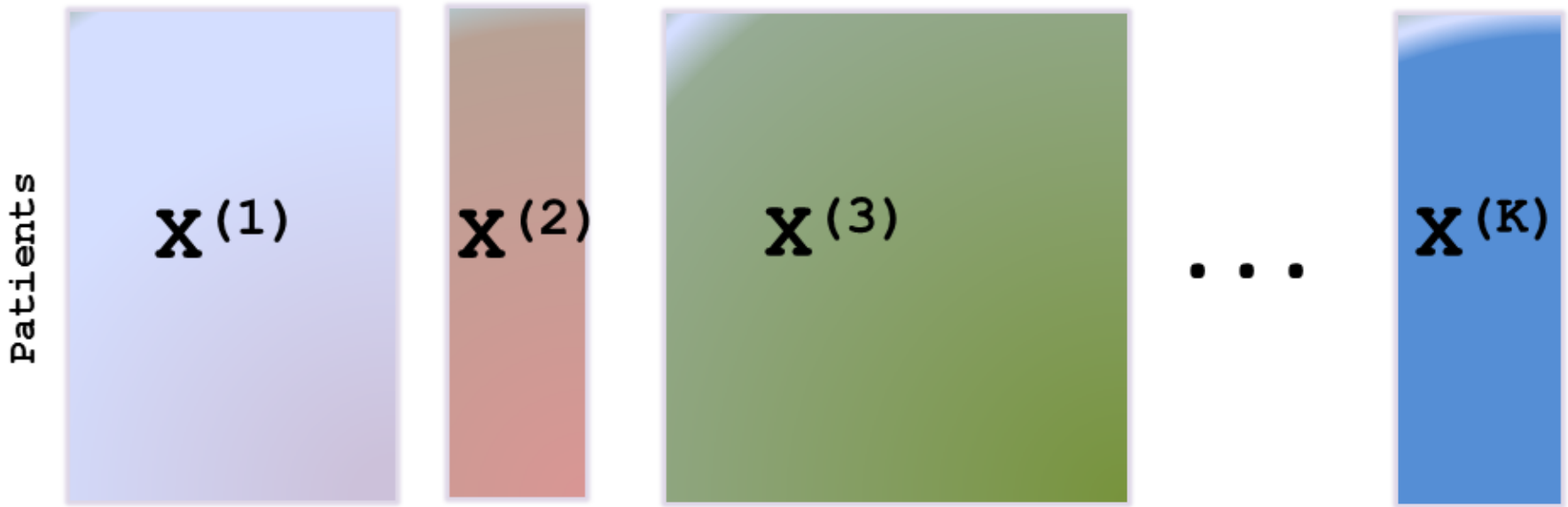
Multiple sources of data (sets of features) measured for the same set of subjects or observations.



Multi-Modal Data

Mixed, Multi-Modal Data

Mixed or heterogeneous types of data-modalities measured for the same set of subjects.



Multi-Modal Data

Opposite: Meta-Analysis.

- We know how to aggregate multiple sets of subjects (n) to conduct inference for features (p).
  - Example: Aggregating patients from multiple GWAS studies to determine associations of rare variants.

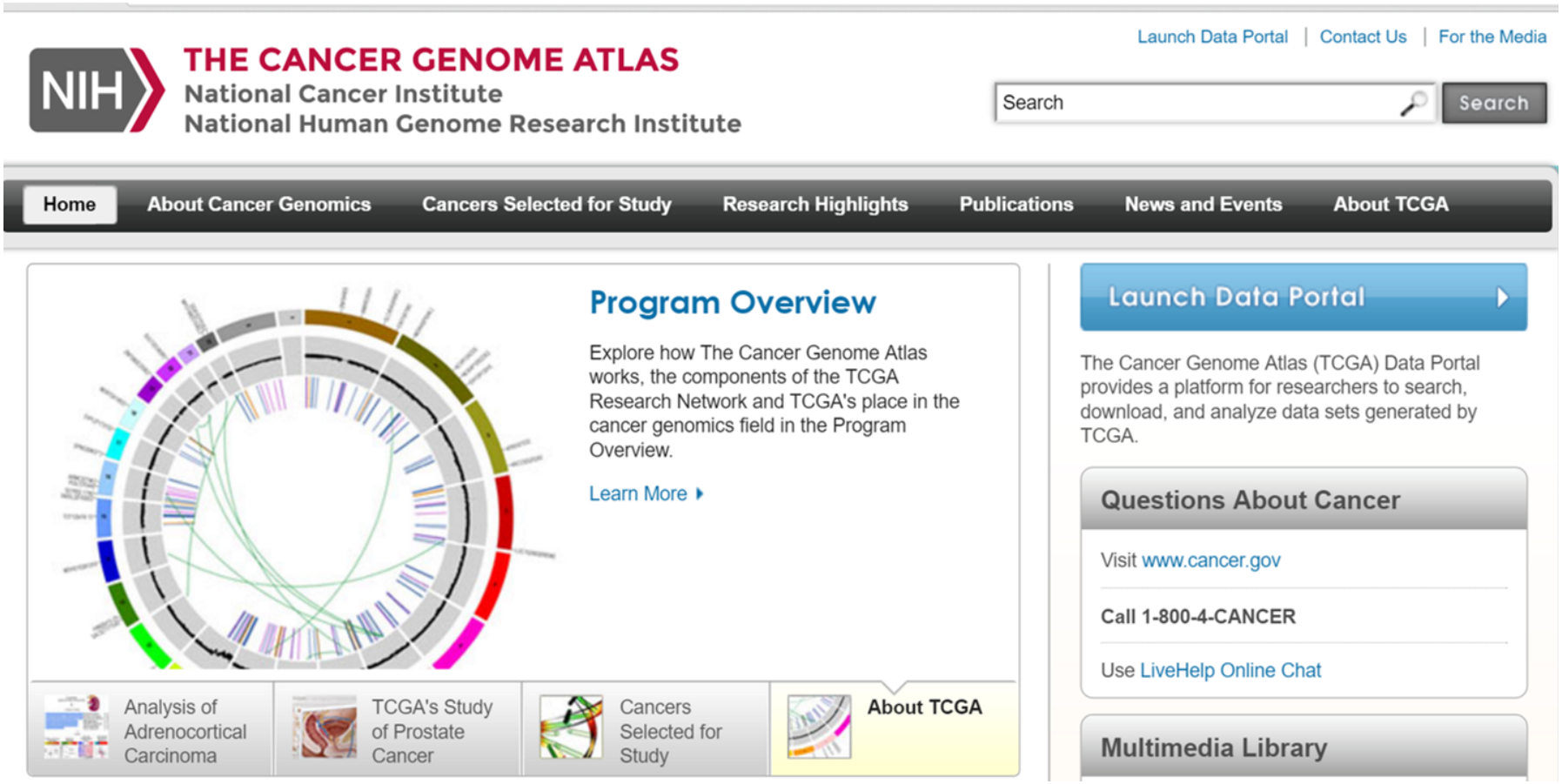
Our Focus

How do we aggregate multiple (mixed) sets of features (p) to conduct inference on subjects (n)?

Multi-Modal Statistical Data Integration.

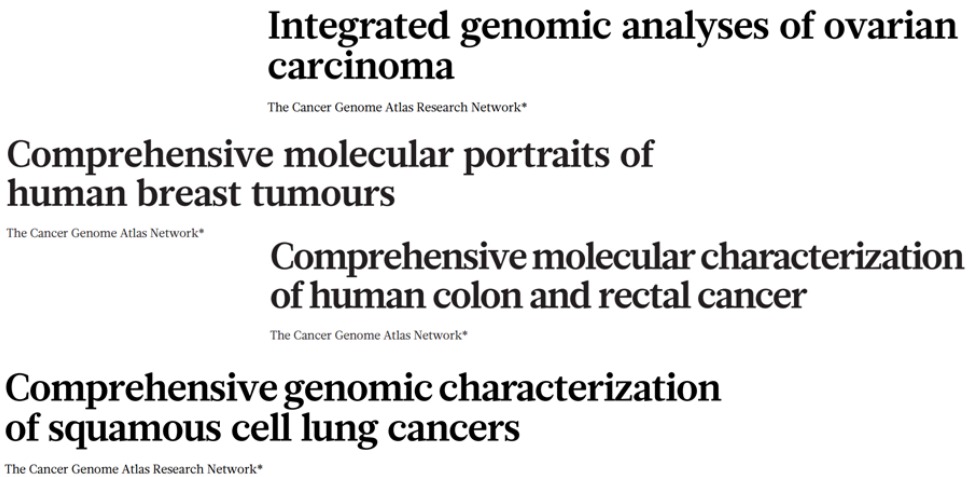
1. Motivating Case Study: The Cancer Genome Atlas
2. Motivating Case Study: The ROS & MAP Studies
3. Challenges & Open Problems
   - Data Challenges
   - Statistical Challenges
     - Multivariate Modeling Challenges: Integrative Networks
     - Integrative Data Mining Challenges
4. The Big(ger) Picture

The Cancer Genome Atlas (TCGA)



The Cancer Genome Atlas (TCGA)

- 33 different cancer types.
- Over 11,000 patients!
- 7 different types of "omics" data (e.g. gene expression, microRNA expression, mutations, copy number aberrations and variation, methylation).
- 2.5 Petabytes worth of data.

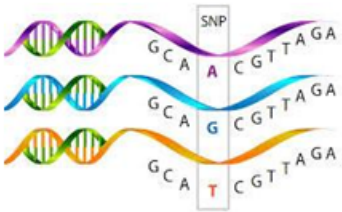


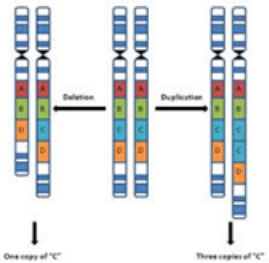
TCGA Data-Modalities

The Cancer Genome Atlas — *Understanding genomics to improve cancer care*

Sequence Changes

Mutations (SNP)

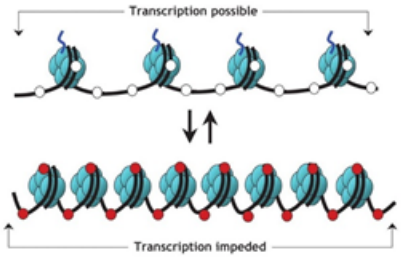


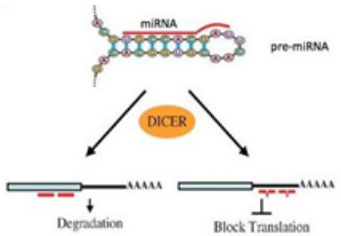


Copy Number Variation

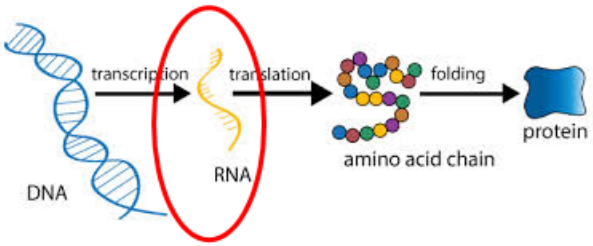
Epigenetics

Methylation





MicroRNA Expression



Gene Expression

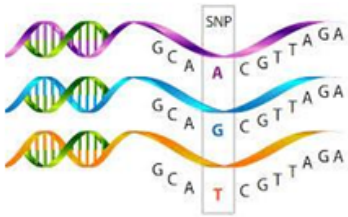
Functional Genetics

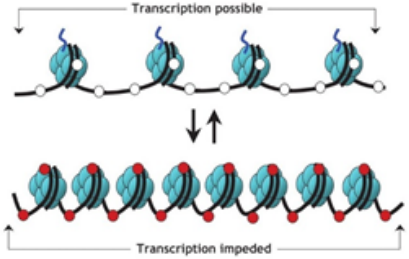
TCGA Data-Modalities

The Cancer Genome Atlas — *Understanding genomics to improve cancer care*

Mutations (SNP)

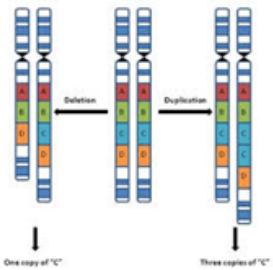
- ~ 100K – 20 Million
- **Binary / Categorical**





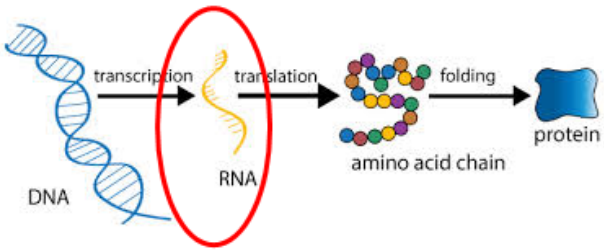
Methylation

- ~ 30K – 450K
- **Bounded Continuous**



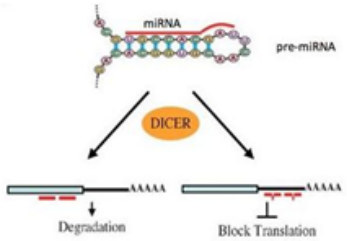
Copy Number Variation

- ~ 20K – 200K
- **Continuous**



Gene Expression

- ~ 1K – 10K
- **Continuous** (array)
- **Counts** (Sequencing)



MicroRNA Expression

- ~ 1K – 10K
- **Continuous** (array)
- **Counts** (Seq)

TCGA Data-Modalities

TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R.

TCGA Data-Modalities

Appendix A: Summary of cancer types and omics-profiles.

Cancer name	Acronym	RNASeq V2	RNASeq	miRNASeq	CNA_SNP	CNV_SNP	CNA_CGH	Methylation (27K)	Methylation (450K)	Mutation	mRNA_Array	miRNA_Array
Adrenocortical carcinoma	**ACC**	Y		Y	Y	Y			Y	Y		
Bladder urothelial carcinoma	**BLCA**	Y	Y	Y	Y	Y			Y	Y		
Breast invasive carcinoma	**BRCA**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Cervical and endocervical cancers	**CESC**	Y		Y	Y	Y			Y	Y		
Cholangiocarcinoma	**CHOL**	Y		Y	Y	Y			Y			
Colon adenocarcinoma	**COAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Colorectal adenocarcinoma	**COADREAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	**DLBC**	Y		Y	Y	Y			Y			
Esophageal carcinoma	**ESCA**		Y	Y	Y	Y			Y			
FFPE Pilot Phase II	**FPPP**			Y								
Glioblastoma multiforme	**GBM**	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Glioma	**GBMLGG**	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y
Head and Neck squamous cell carcinoma	**HNSC**	Y	Y	Y	Y	Y			Y	Y		
Kidney Chromophobe	**KICH**	Y		Y	Y	Y			Y	Y		
Pan-kidney cohort (KICH+KIRC+KIRP)	**KIPAN**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal clear cell carcinoma	**KIRC**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Kidney renal papillary cell carcinoma	**KIRP**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Acute Myeloid Leukemia	**LAML**	Y	Y	Y	Y	Y		Y	Y	Y		
Brain Lower Grade Glioma	**LGG**	Y		Y	Y	Y			Y	Y	Y	
Liver hepatocellular carcinoma	**LIHC**	Y	Y	Y	Y	Y			Y	Y		
Lung adenocarcinoma	**LUAD**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Lung squamous cell carcinoma	**LUSC**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
Mesothelioma	**MESO**	Y		Y	Y	Y			Y			
Ovarian serous cystadenocarcinoma	**OV**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Pancreatic adenocarcinoma	**PAAD**	Y		Y	Y	Y			Y	Y		
Pheochromocytoma and Paraganglioma	**PCPG**	Y		Y	Y	Y			Y	Y		
Prostate adenocarcinoma	**PRAD**	Y		Y	Y	Y			Y	Y		
Rectum adenocarcinoma	**READ**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Sarcoma	**SARC**	Y		Y	Y	Y			Y			
Skin Cutaneous Melanoma	**SKCM**	Y		Y	Y	Y			Y	Y		
Stomach adenocarcinoma	**STAD**		Y	Y	Y	Y		Y	Y	Y		
Testicular Germ Cell Tumors	**TGCT**	Y		Y	Y	Y			Y	Y		
Thyroid carcinoma	**THCA**	Y	Y	Y	Y	Y			Y	Y		
Thymoma	**THYM**	Y		Y	Y	Y			Y			
Uterine Corpus Endometrial Carcinoma	**UCEC**	Y	Y	Y	Y	Y		Y	Y	Y	Y	
Uterine Carcinosarcoma	**UCS**	Y		Y	Y	Y			Y	Y		
Uveal Melanoma	**UVM**	Y		Y	Y	Y			Y	Y		

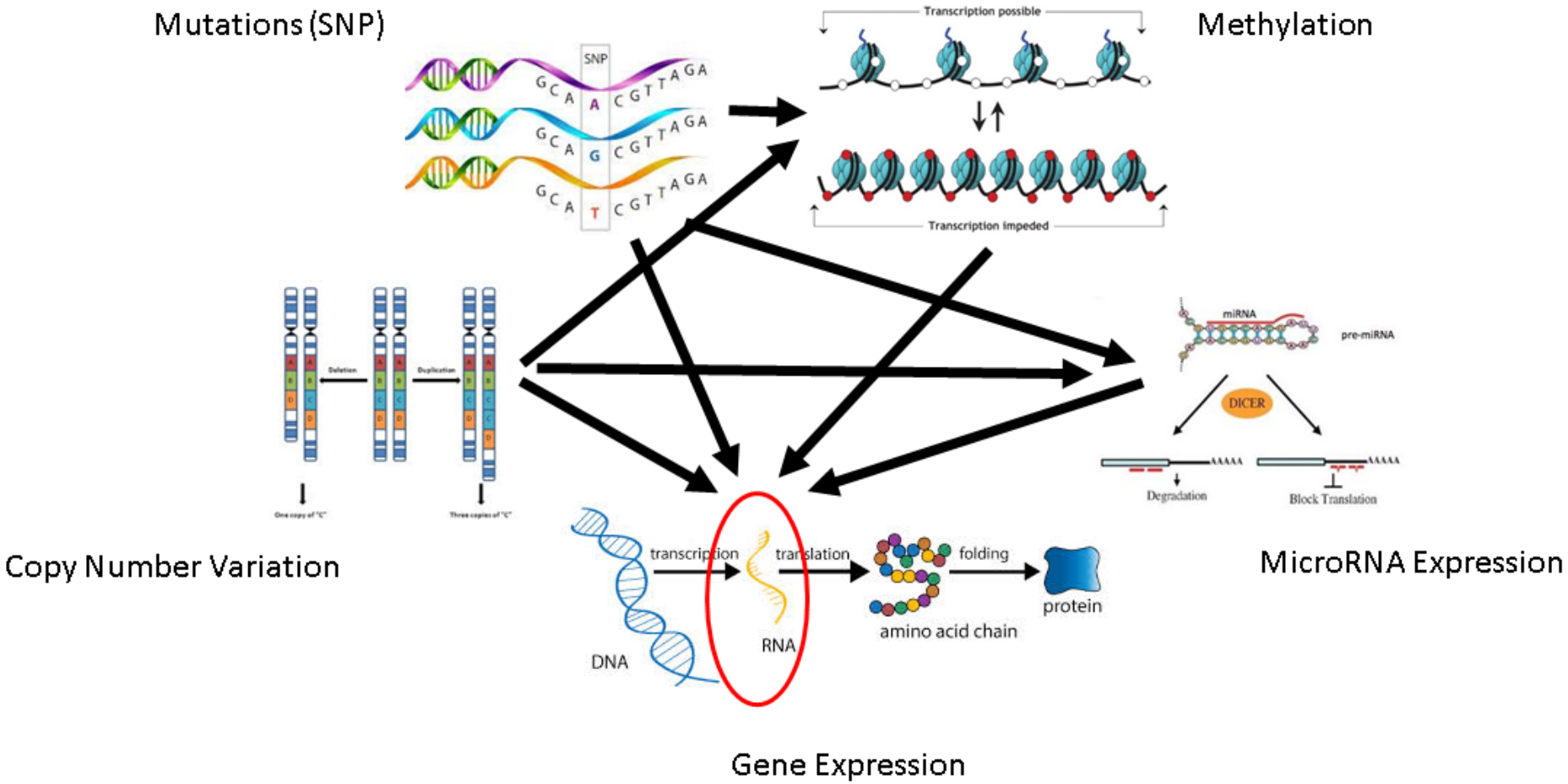
TCGA Objectives & Data Integration

- Cancer is complex & heterogeneous.
- TCGA targeted cancer types where etiology unknown, prognosis is poor and/or few therapies exist.

Objectives:

- Discover sets of mutations and aberrations, gene expression changes, and epigenetic changes that cause tumor cells to grow and proliferate.
- Discover cancer subtypes.
  - Groups of patients with similar molecular tumor characteristics.
  - Similar outcomes & Respond similarly to therapies.
- Discover new personalized therapies.
  - Many genes are more easily targeted by manipulating miRNAs or methylation levels.
  - Example: Ovarian cancer tumors with BRCA1/2 mutations sensitive to PARP inhibitors.

TCGA Objectives & Data Integration



1. Motivating Case Study: The Cancer Genome Atlas
2. Motivating Case Study: The ROS & MAP Studies
3. Challenges & Open Problems
   - Data Challenges
   - Statistical Challenges
     - Multivariate Modeling Challenges: Integrative Networks
     - Integrative Data Mining Challenges
4. The Big(ger) Picture

Alzheimer's Disease & Dementia

Alzheimer's Disease:

- 6^{th} leading cause of death in the US.
- Only top ten cause of death that cannot be prevented, cured, or slowed.
- **35.6 million** people worldwide are currently living with Alzheimer's Disease, with an estimated **115 million** people by 2050.
- Characterized by progressive declines in memory & cognition (dementia).

Other Causes of Dementias:

- Lewy Bodies.
- Parkinson's Disease.
- Infarcts.

ROS/MAP Studies

ROS: Religious Orders Study (1994 - present).

- Catholic priests, nuns & brothers.

MAP: Rush Memory and Aging Project (1997 - present).

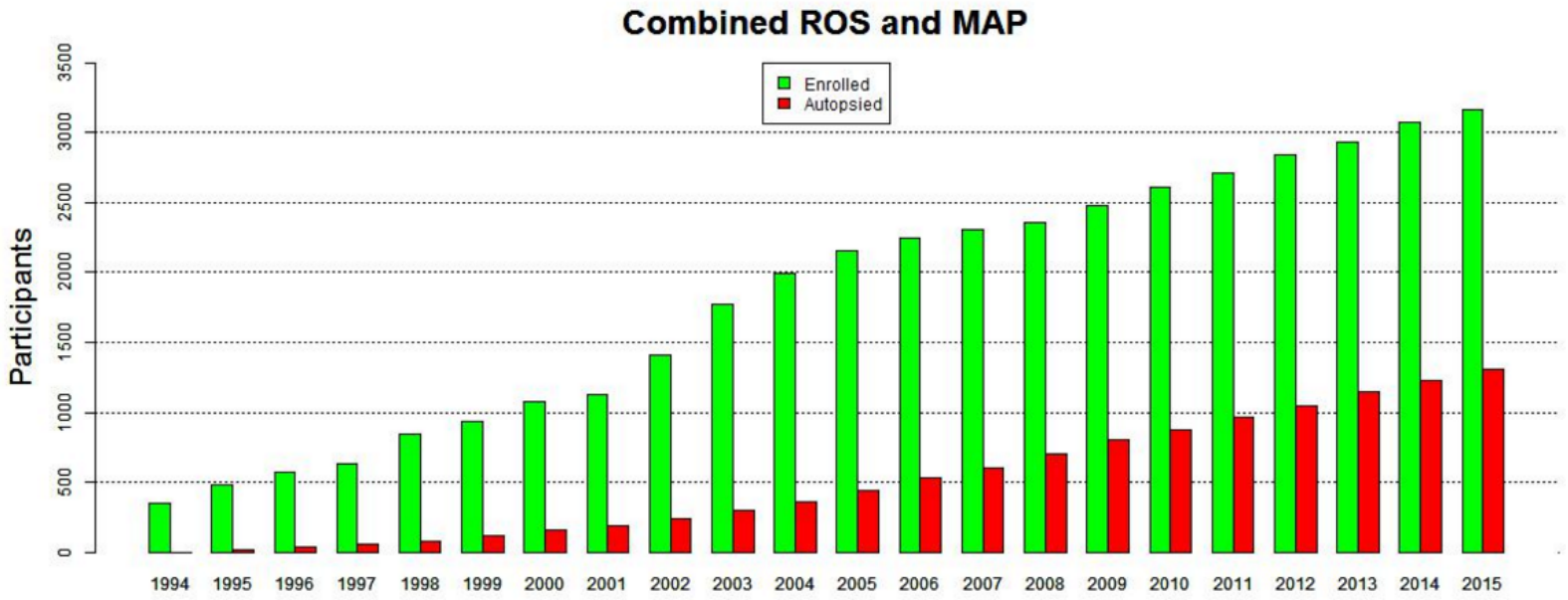
- Older adults living in assisted living communities.

Prospective, longitudinal studies of aging & dementia:

- $>$2800 subjects enrolled.
- Free of dementia at time of enrollment.
- Follow-up rate among survivors: $> 90\%$

David Bennett, PI.

ROS/MAP Studies

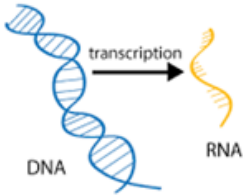


- Brains donated at death for autopsy (Anatomical Gift Act).
- Autopsy rate among deceased: $> 90\%$
- $>$1000 brains autopsied to date.

ROS/MAP Studies

	ROS	MAP
n	1156	1642
Age at BL	75.6 (7.5)	79.9 (7.6)
Years of FU, median [min – max]	9 [0 – 19]	4 [0 – 15]
Male sex	355 (31%)	435 (26%)
Education, yrs	18 (3.3)	14.5 (3.3)
Clinical dx of AD	377 (22%)	361 (33%)
Self-reported EA	1020 (88%)	1443 (88%)
Dead	635 (55%)	649 (40%)
Age at death	87.2 (7.0)	88.8 (6.2)
Pathologic dx of AD	355 (62%)	306 (62%)

ROS/MAP Data-Modalities Collected



Baseline: GWAS

Annual: Independent Clinical Evaluation & Cognitive Testing

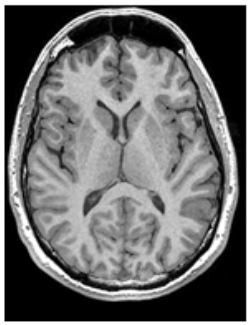
Post-Mortem: Transcriptomics & Epigenetics

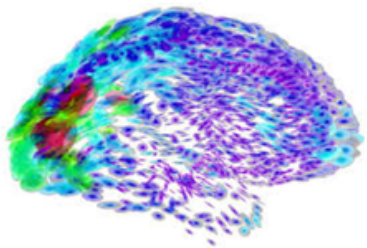
Post-Mortem: Neuroimaging

Biannual: Structural Neuroimaging (2009-present)

Baseline: Demographics, Behavioral, Environmental, Social, and Lifestyle Factors

Post-Mortem: Comprehensive Neuropathology

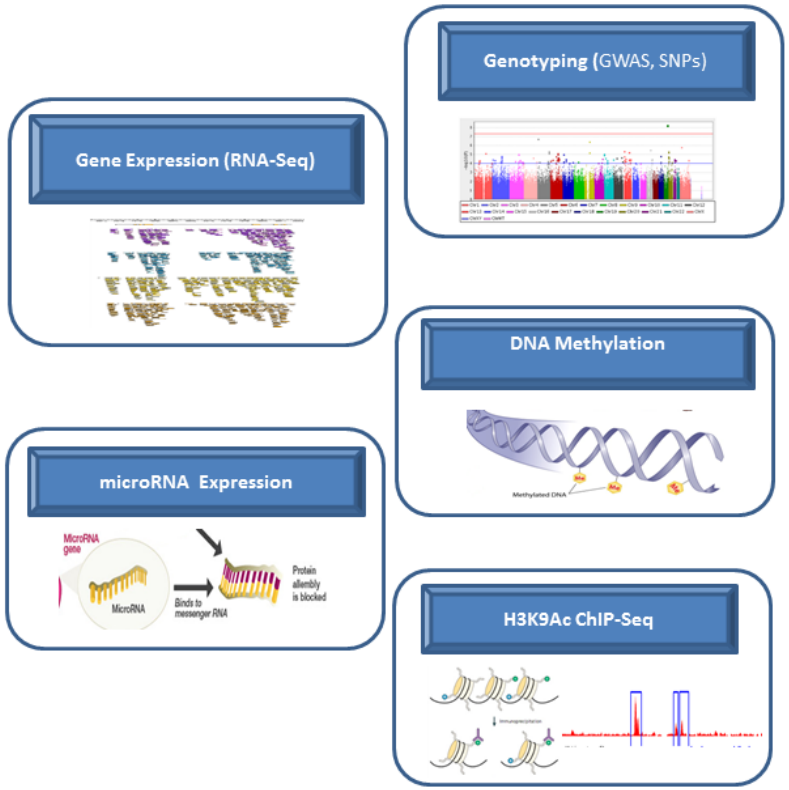




ROS/MAP Data-Modalities Collected

Genetics Data.

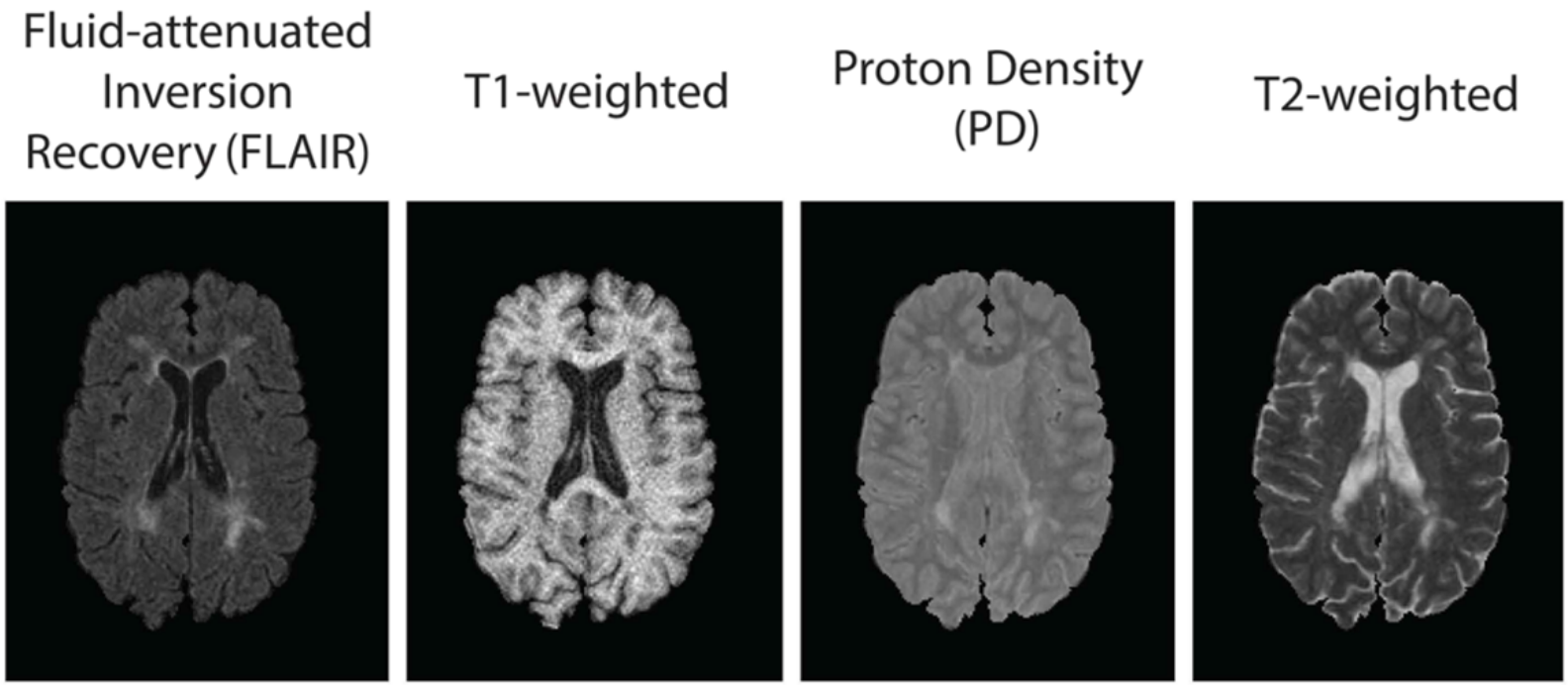
- Genotyping.
  - Single Nucleotide Polymorphisms ($n = 2295$).
  - Whole Exome Sequencing ($n = 783$).
- Gene Expression.
  - Next generation RNASequencing ($n = 636$).
- Epigenetics.
  - MicroRNA Expression ($n = 702$).
  - Histone acetylation ($n = 714$).
  - DNA Methylation ($n = 748$).
- Proteomics & Metabolomics. (In Progress)



ROS/MAP Data-Modalities Collected

Neuroimaging Data (Ante & Post-Mortem).

- Multi-Parametric Structural MRI (FLAIR, T1-Weighted, Quantitative T2, SWI); 1601 scans on $n = 854$ patients.
- Diffusion Tensor Imaging (DTI).
- Functional MRI (fMRI; resting-state).



ROS/MAP Data-Modalities Collected

Other Data.

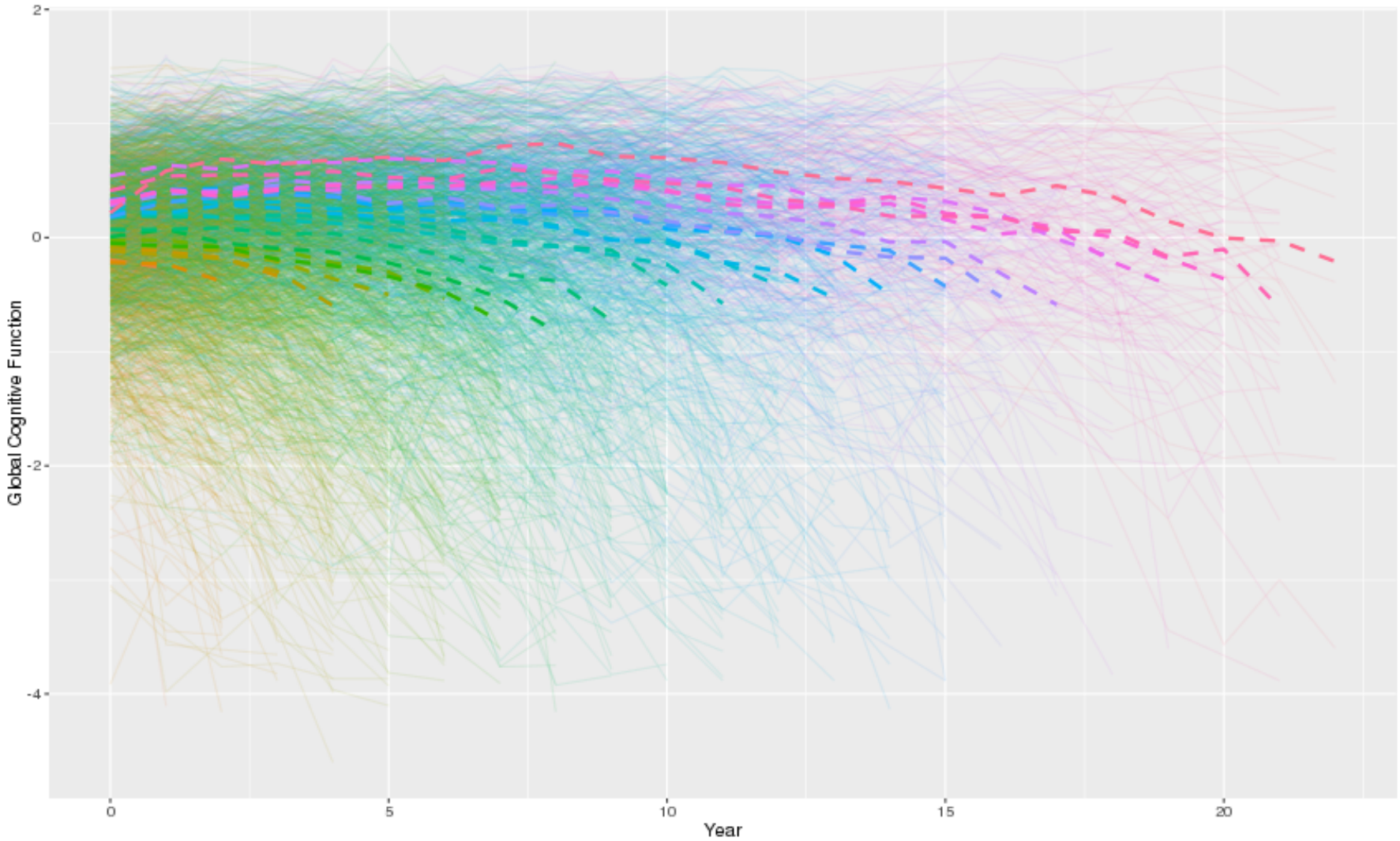
- Demographics (age, education, gender, race, etc.)
- Clinical Diagnosis & Medical Conditions.

- Cognitive Testing.
  - 19 different tests! (Episodic memory, working memory, semantic memory, perceptual orientation, perceptual speed).

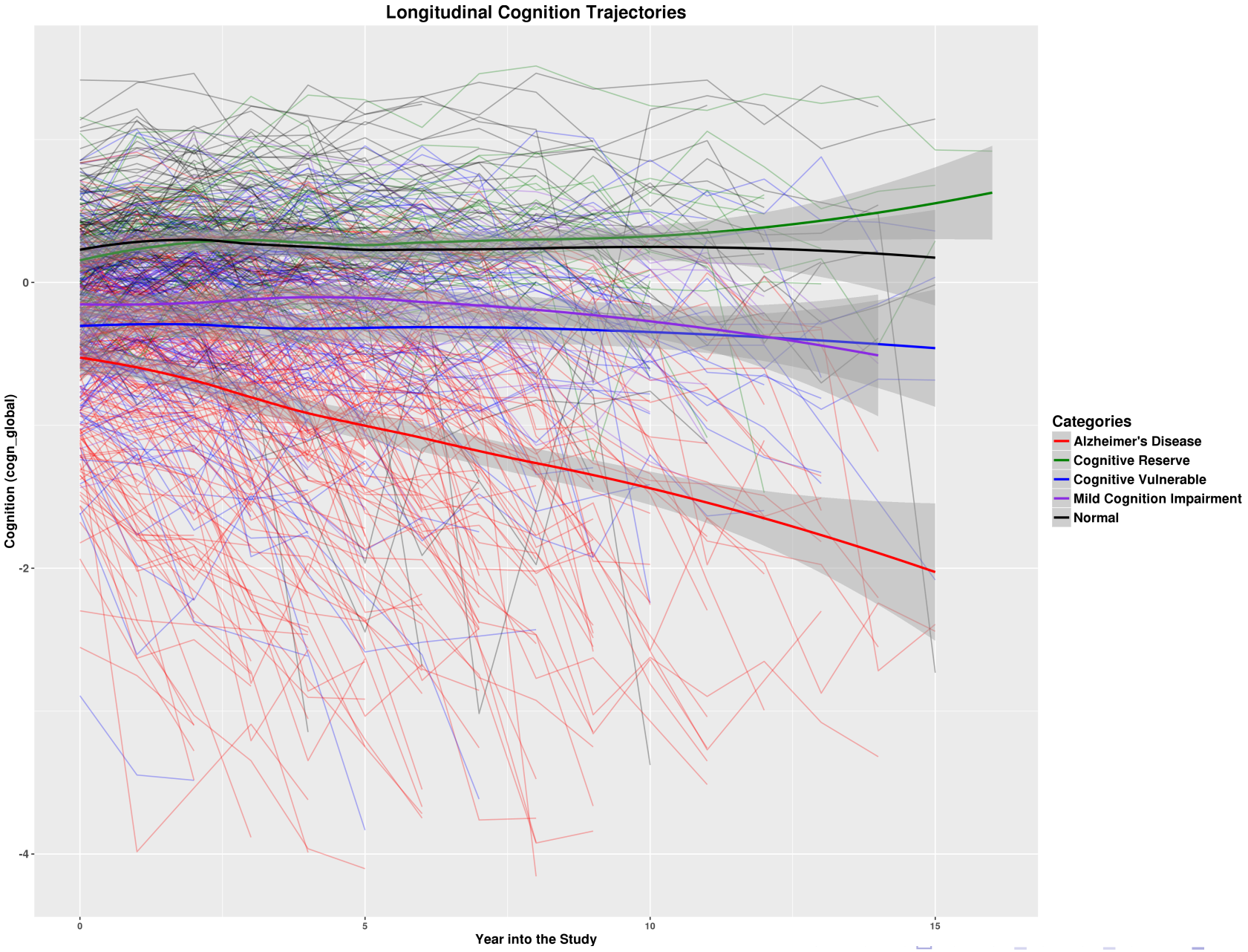
- Neuro-Pathology.
  - Immunohistology of Beta-Amyloid & Tau-Tangle Density for 8 brain regions, Lewy Bodies, Infarcts, TDP-43, etc.

- Life Style & Personality (physical activities, social interactions, cognitive activities, purpose in life, etc.)
- Motor & Gait Measures.

ROS/MAP Data-Modalities Collected



ROS/MAP Data-Modalities Collected



ROS/MAP Objectives & Data Integration

Alzheimer's Disease (Late Onset):

- Etiology unknown.
- Few known risk factors.
  - APOE e4 (10-15% in population) associated with higher risk of AD, but 75% with APOE e4 don't develop AD and only 50% of AD patients have APOE e4.
  - Small associations with educational attainment, social interactions, cognitive activities, and lifestyle.
- Pathologically characterized by Amyloid β plagues and τ-tangles.
  - Causative or a by-product of another disease-causing process?
  - Cognitive Reserve: $\approx$ 44% of patients exhibit AD-like neuro-pathology but show no signs of cognitive decline.

ROS/MAP Objectives & Data Integration

- Alzheimer's Disease is very complex.
- Individual Data-Modalities (e.g. just genetics, just cognition, lifestyle, just neuroimaging, just neuropatholgy) have been extensively studied ... with few successes.
- Data Integration of all possible Data-Modalities to gain a more complete picture into the etiology & possible therapies for AD!
  - Examples: Find a genetic basis for amyloid plagues and for when these cause cognitive decline.
  - Find an earlier neuroimaging marker for plagues that are likely to cause cognitive decline.
  - Find epigenetic markers that are possible drug targets modifying the expansion of amyloid plagues.

1 Motivating Case Study: The Cancer Genome Atlas

2 Motivating Case Study: The ROS & MAP Studies

3 Challenges & Open Problems
- Data Challenges
- Statistical Challenges
  - Multivariate Modeling Challenges: Integrative Networks
  - Integrative Data Mining Challenges

4 The Big(ger) Picture

Data Challenge: Batch Effects

Problem:

- Data acquired in groups or in different labs / clinics.
- Over time, technology can change.
- Results in differences in way data is produced and processed.

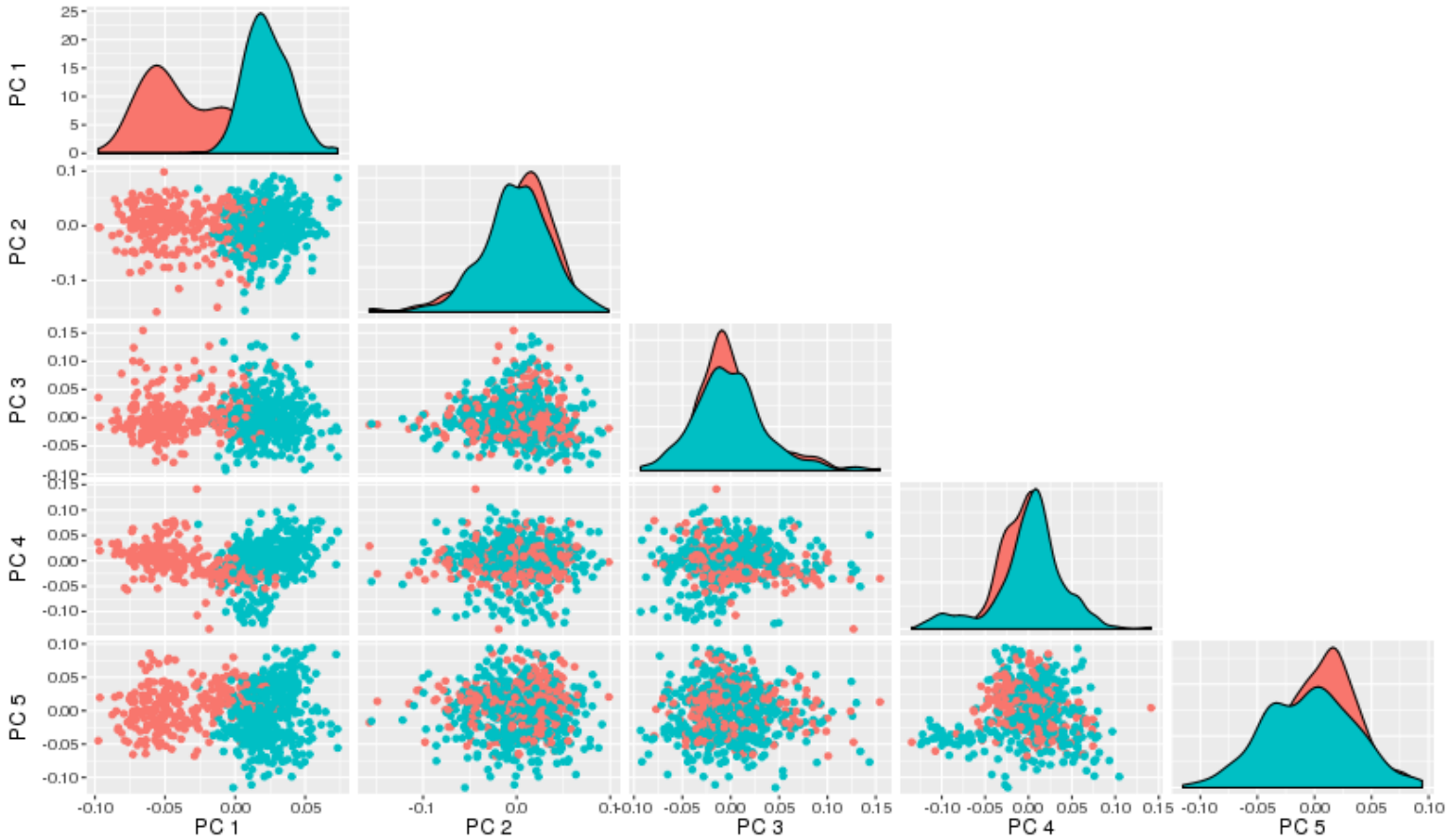
Batch Effects!

Major Challenges:

- Batch Effects can be confounded over time.
- Batch Effects can be confounded across Data-Modalities.
- Technologies change or are replaced over time.

Batch Effects in ROS/MAP

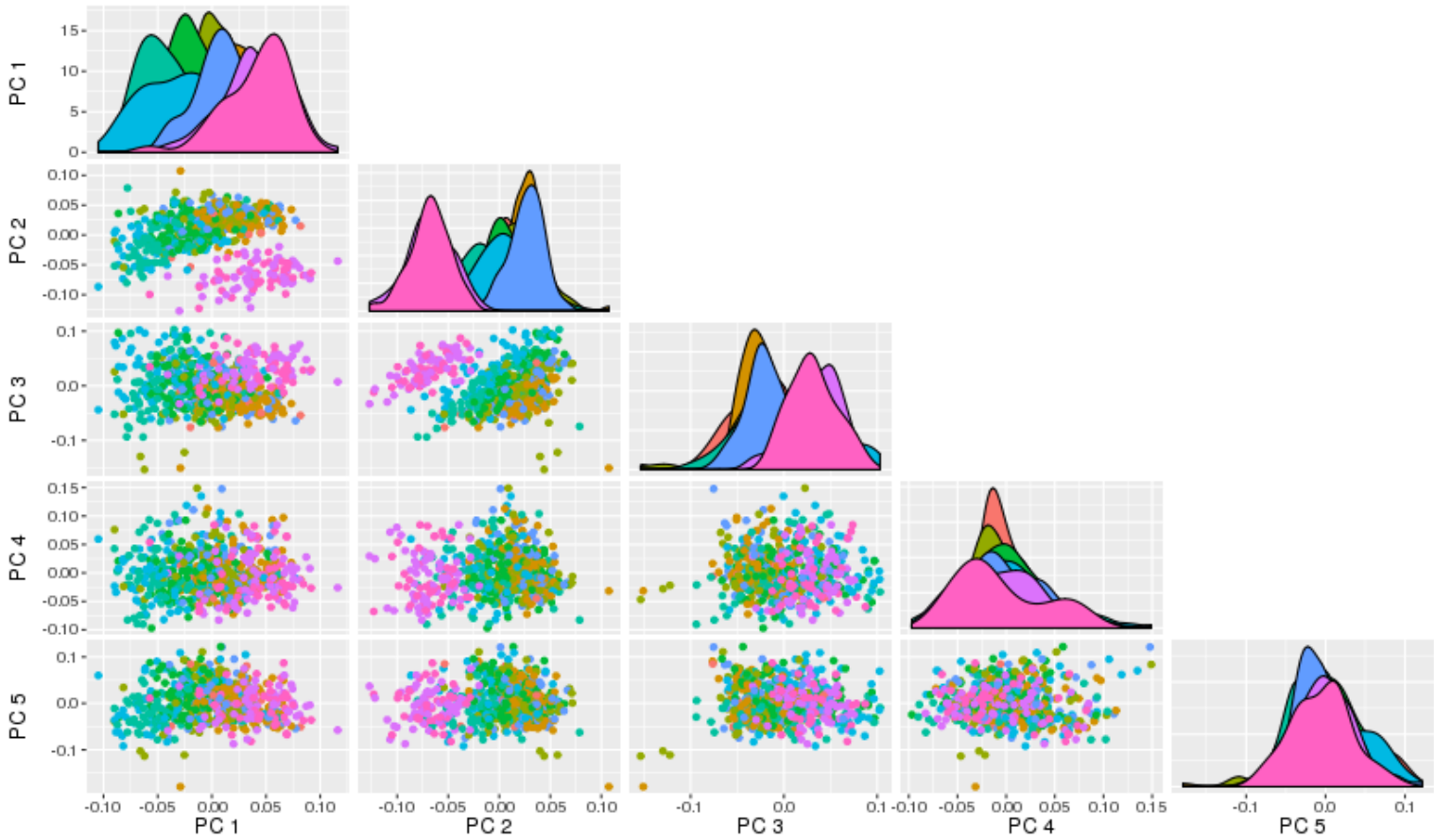
DNA Methylation Data.



- Thermocycler replaced.

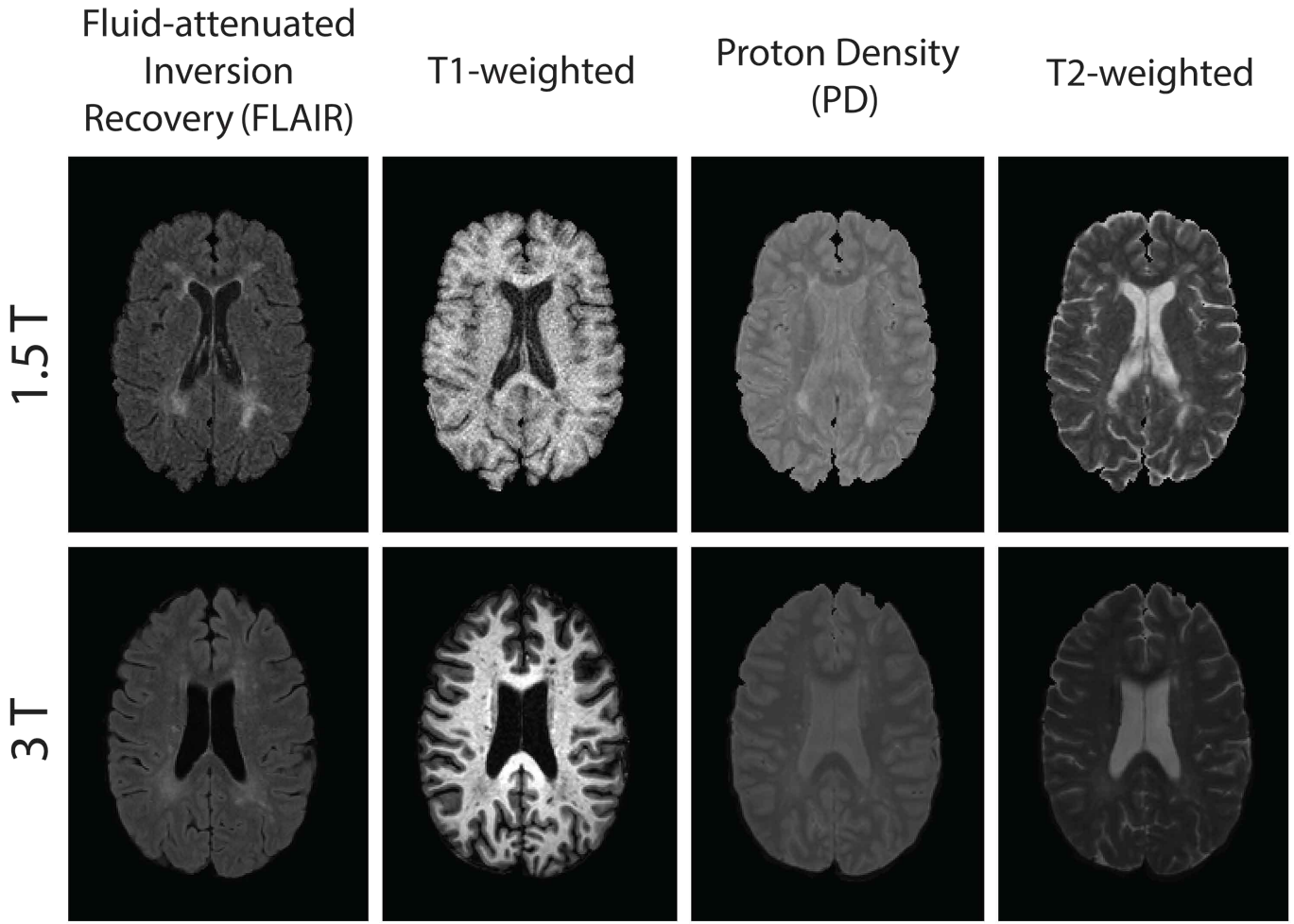
Batch Effects in ROS/MAP

RNA-Sequencing Gene Expression Data.



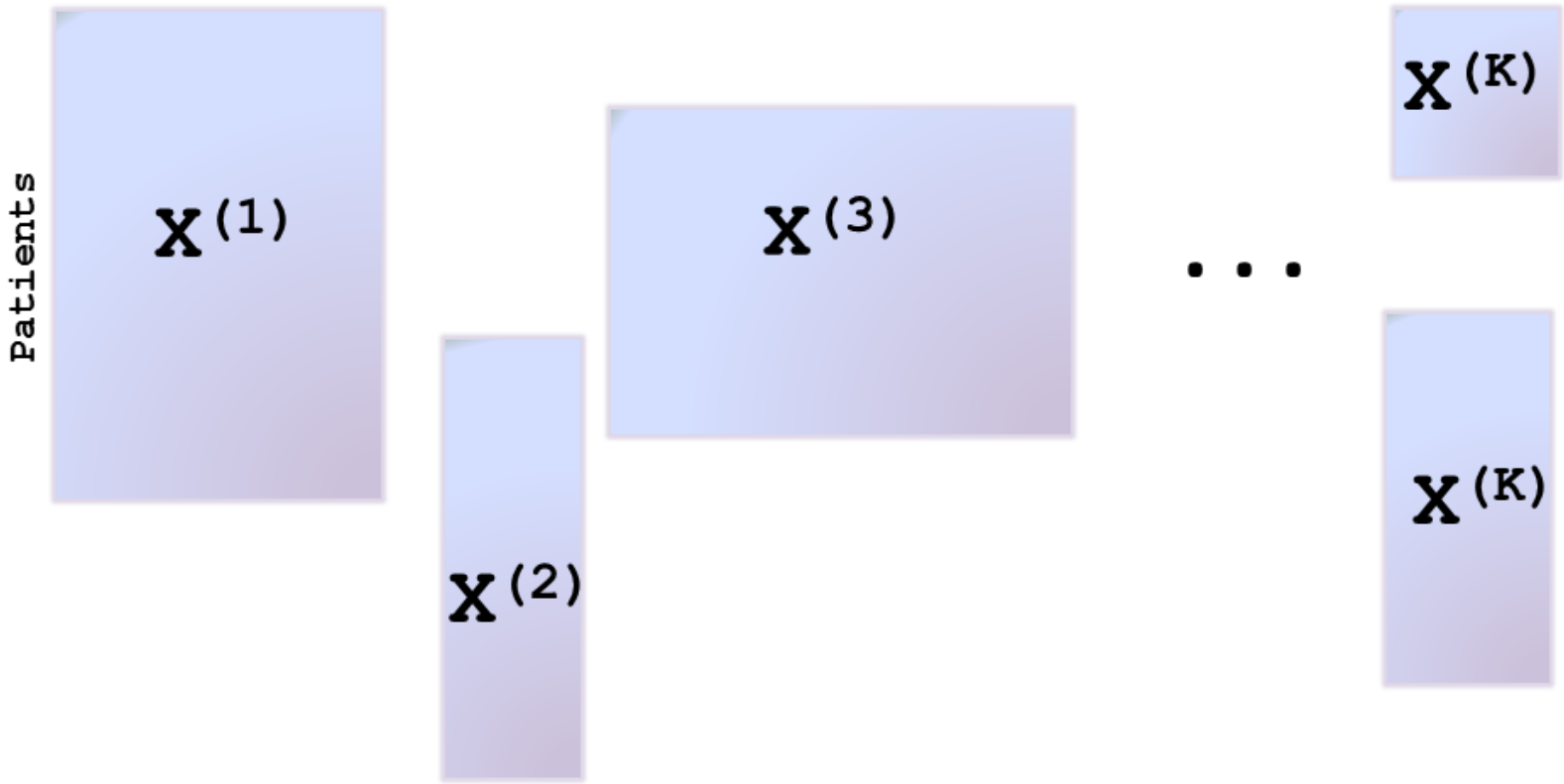
Batch Effects in ROS/MAP

Structural Neuroimaging Data.



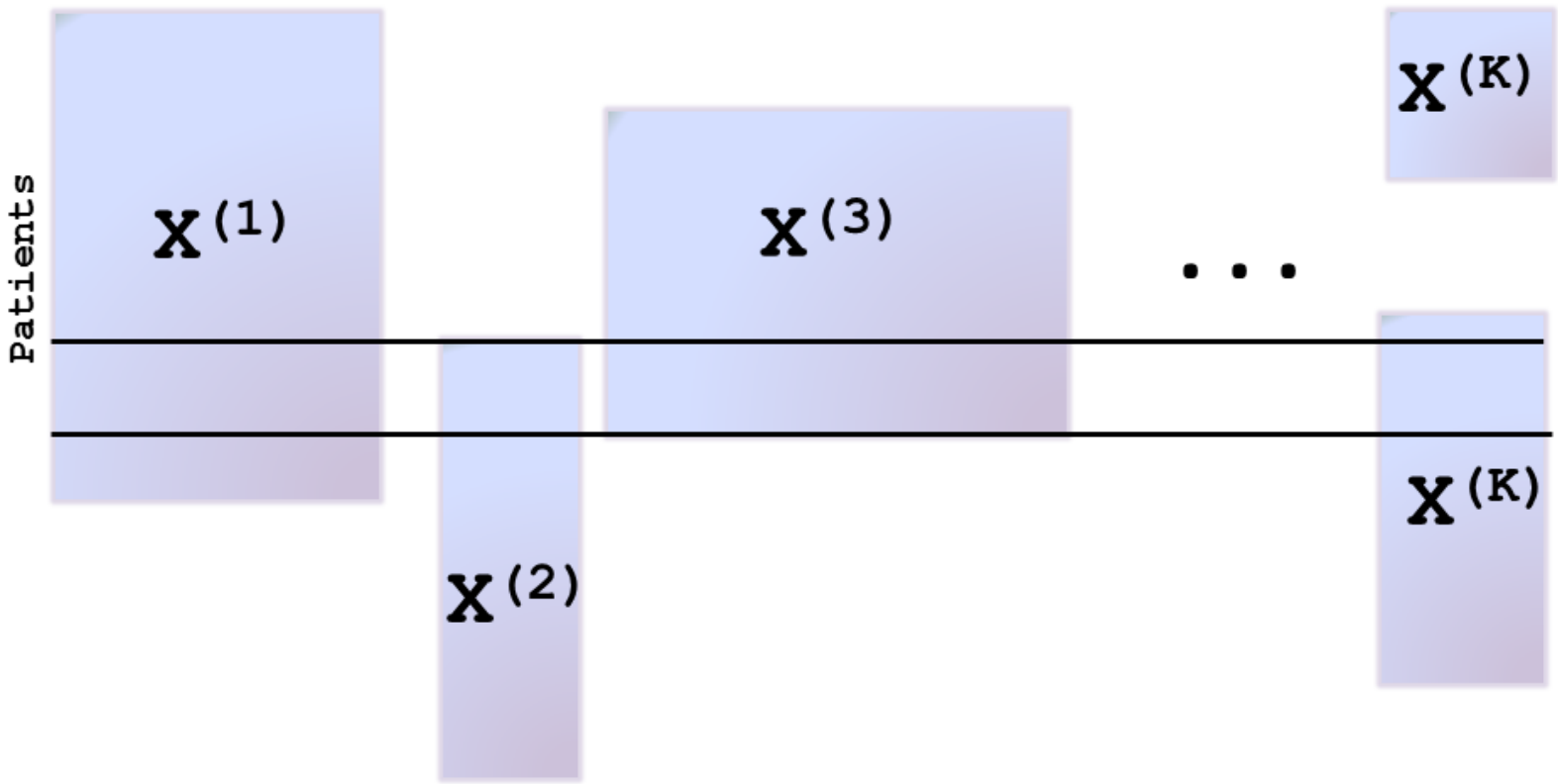
- 1.5 Tesla magnet used before 2012, 3 Tesla after.

Data Challenge: Missing / Unaligned Data-Modalities



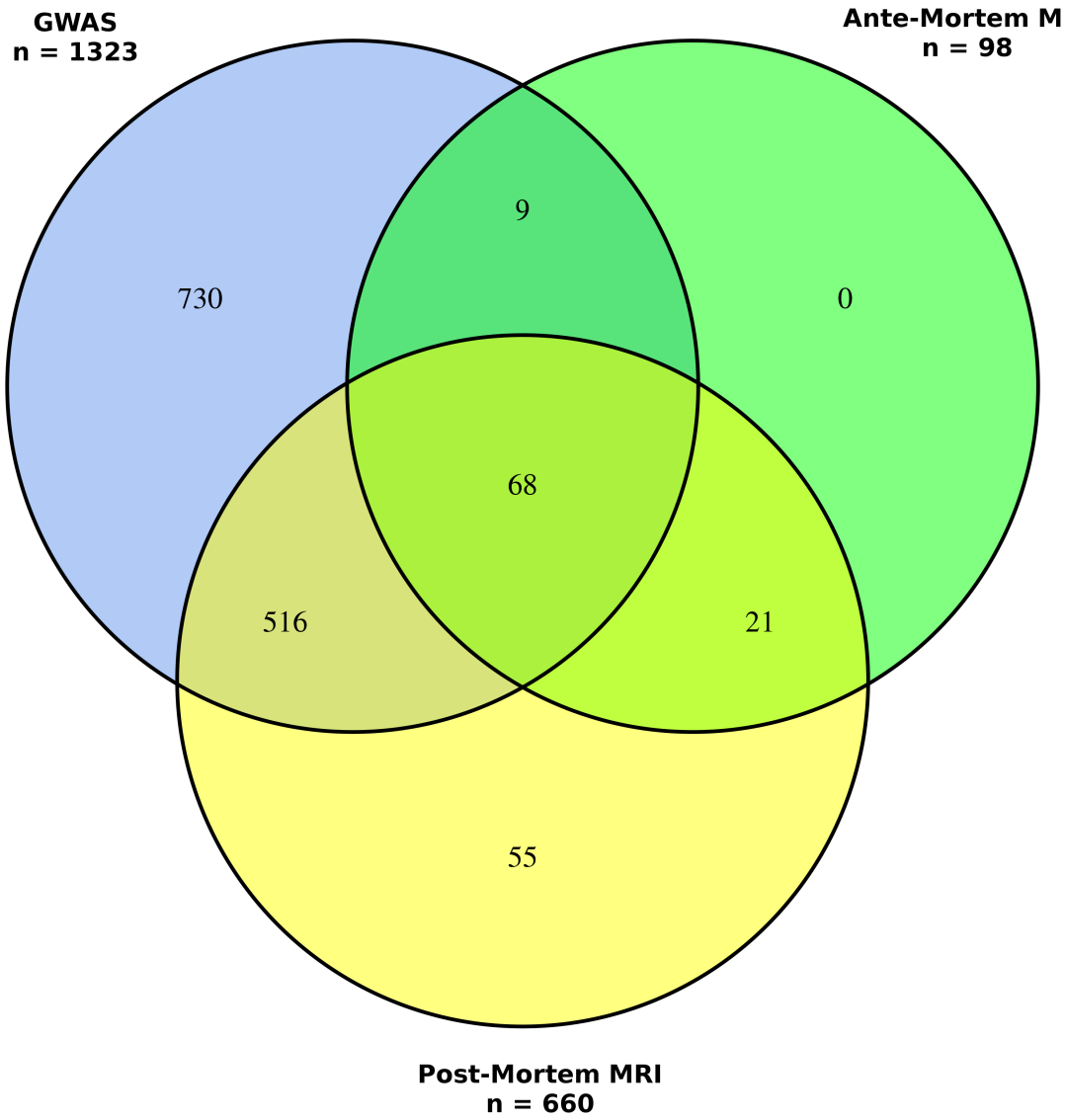
- Not all Data-Modalities measured for all subjects.

Data Challenge: Missing / Unaligned Data-Modalities

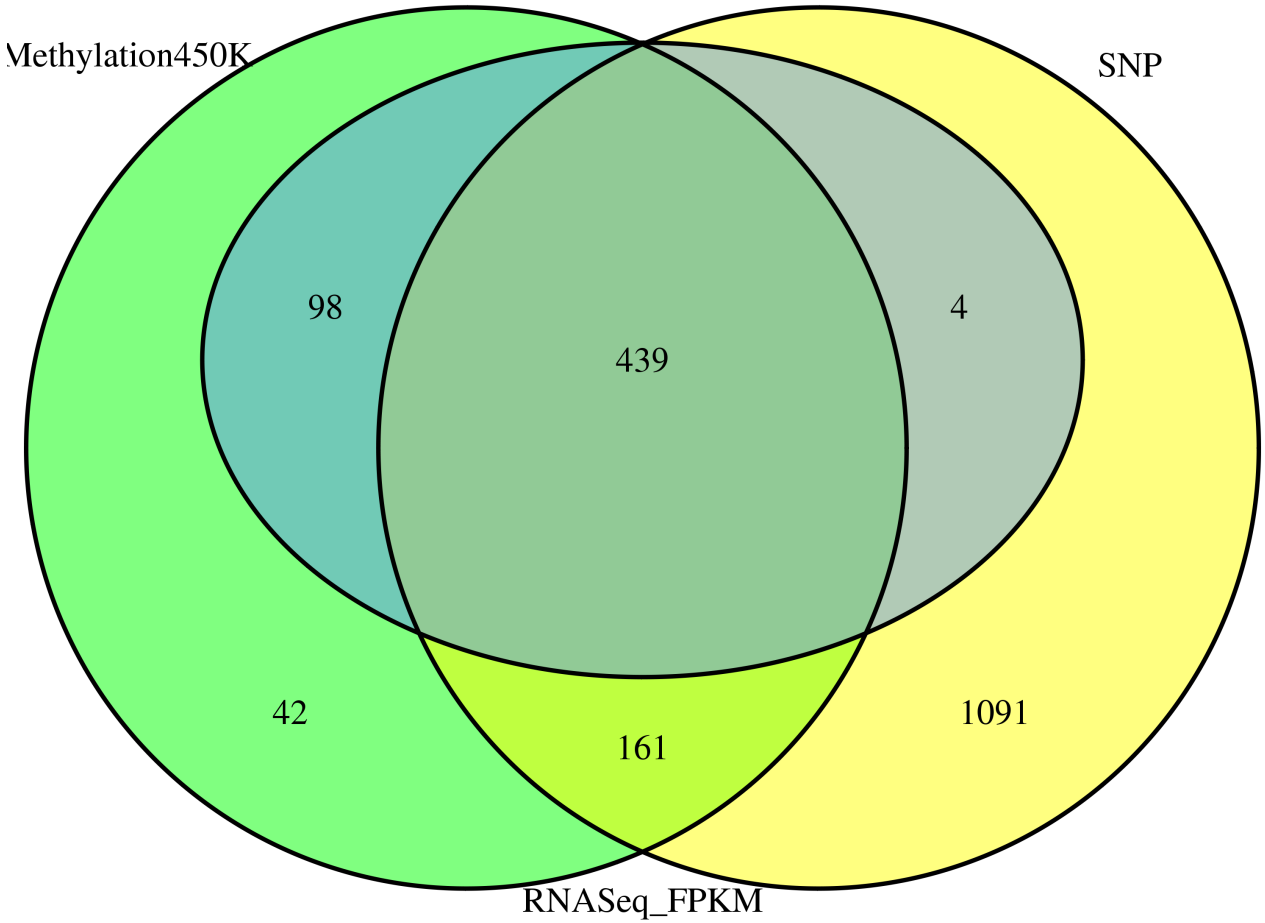


- Limited sample size if use complete cases.

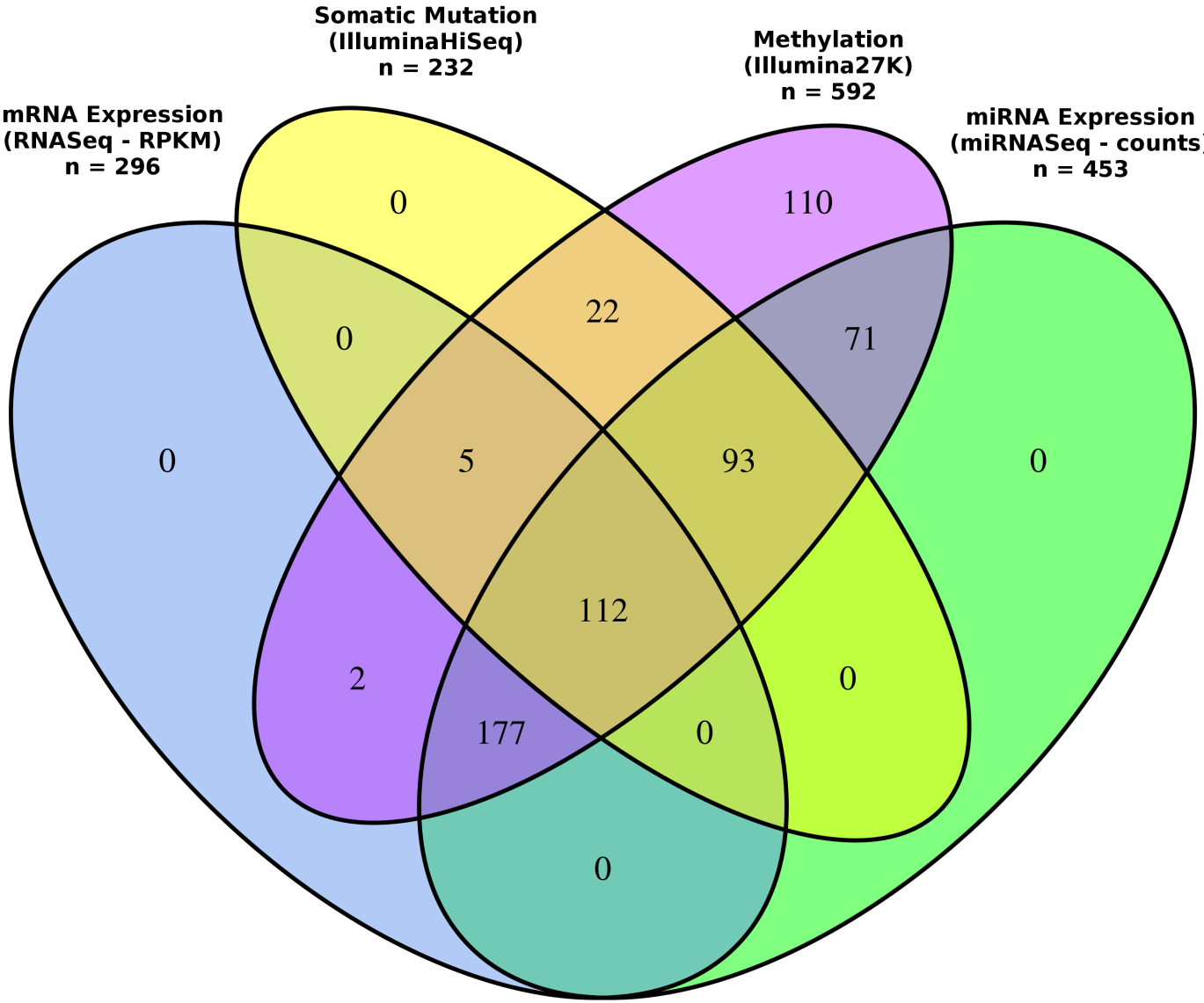
Missing Data Chunks: ROS/MAP Data



Missing Data Chunks: ROS/MAP Data



Missing Data Chunks: TCGA Ovarian Cancer



1. Motivating Case Study: The Cancer Genome Atlas
2. Motivating Case Study: The ROS & MAP Studies
3. Challenges & Open Problems
   - Data Challenges
   - Statistical Challenges
     - Multivariate Modeling Challenges: Integrative Networks
     - Integrative Data Mining Challenges
4. The Big(ger) Picture

Prediction vs. Data-Driven Discoveries

Prediction is (relatively) Easy ...

- Black-Box methods that can handle mixed data.
  - Example: Random Forests, RBMs, Deep Learning.
- Ensemble Learning.
  - Fit different model to each Data-Modality & Ensemble them together.
- Feature Learning on each Data-Modality.
  - Feature learning (e.g. PCA, PLS, RBM, etc.) on each Data-Modality.
  - Fit supervised model to learned features from all Data-Modalities.

Prediction vs. Data-Driven Discoveries

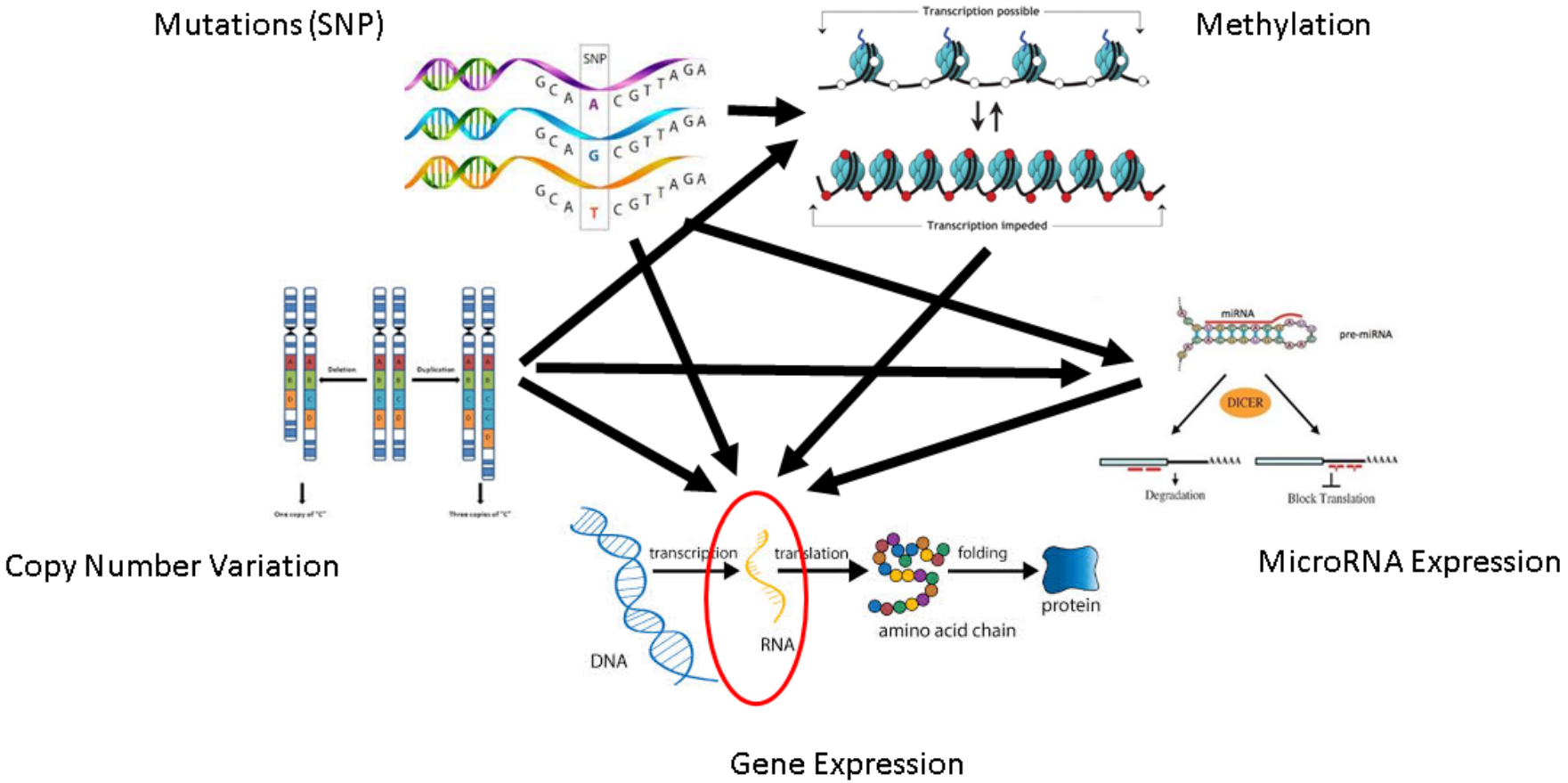
Prediction is (relatively) Easy ...

- Black-Box methods that can handle mixed data.
  - Example: Random Forests, RBMs, Deep Learning.
- Ensemble Learning.
  - Fit different model to each Data-Modality & Ensemble them together.
- Feature Learning on each Data-Modality.
  - Feature learning (e.g. PCA, PLS, RBM, etc.) on each Data-Modality.
  - Fit supervised model to learned features from all Data-Modalities.

... Discoveries from Mixed, Multi-Modal Data are Hard.

Data-Driven Discoveries

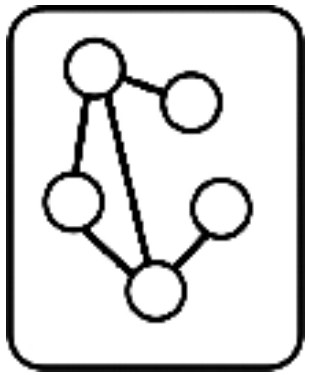
Integrative Genetic Networks

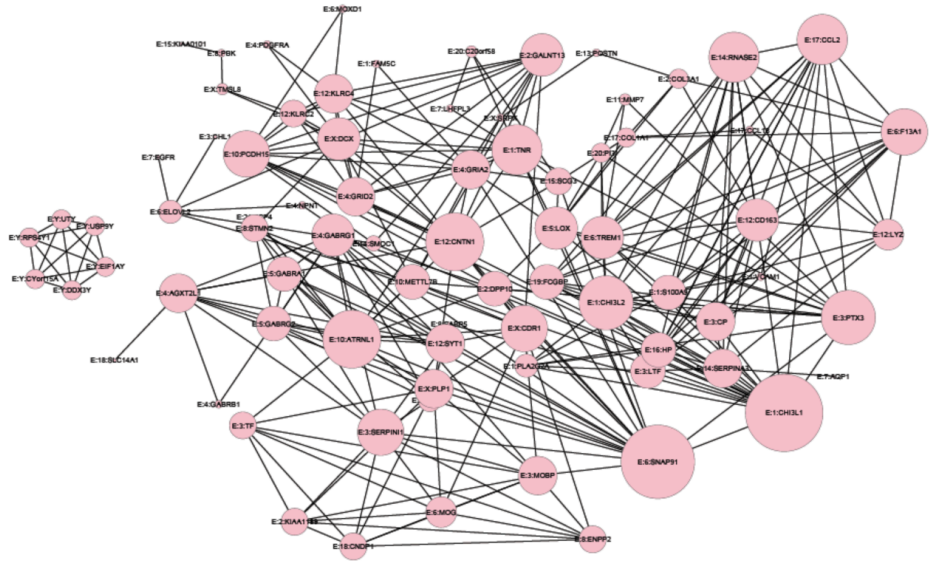


Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).





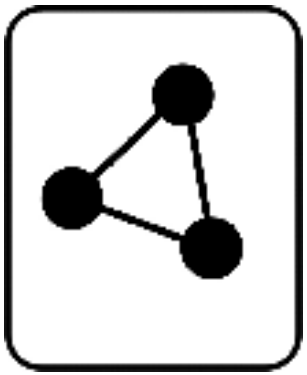
Glioblastoma gene expression network (microarray).

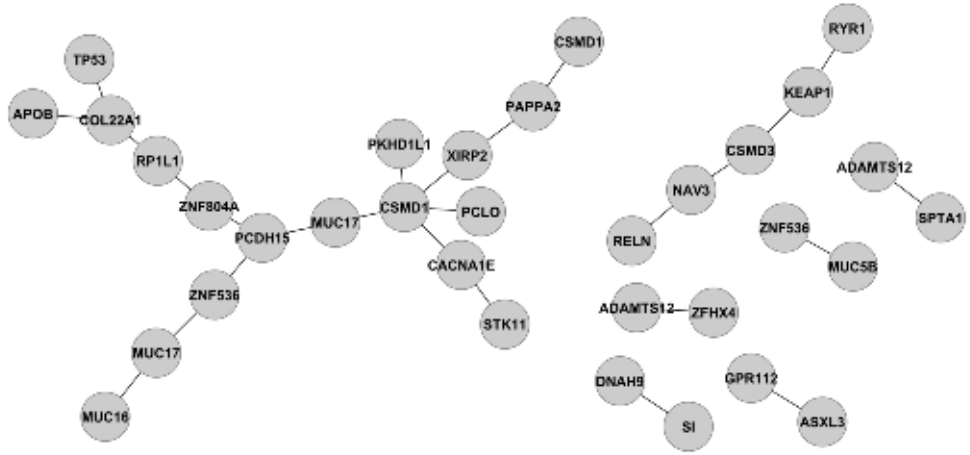
2. Ising Models (Binary-Valued).
3. Gaussian-Ising Models.

Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).





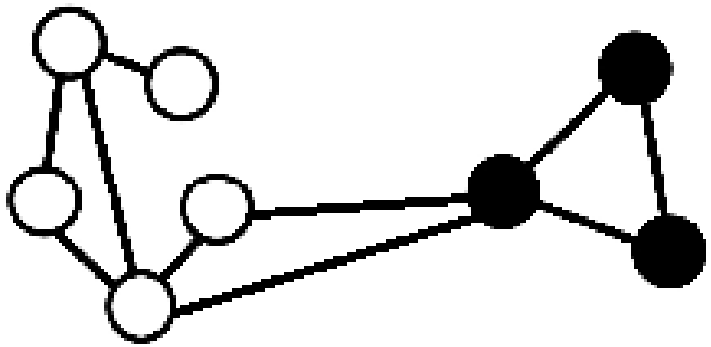
Modules from lung cancer somatic mutation network.

3. Gaussian-Ising Models.

Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).
3. Gaussian-Ising Models.

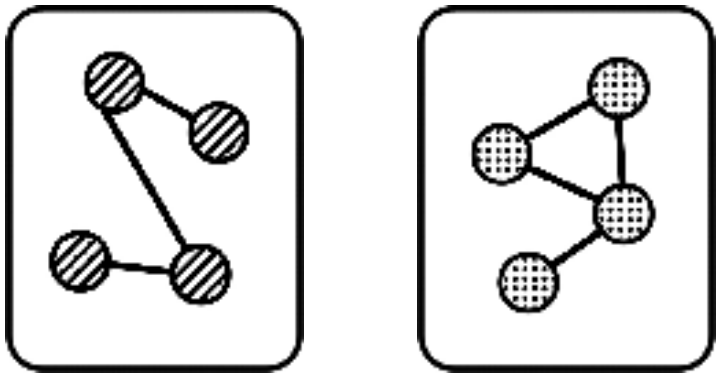


Networks for Different Data Types

Existing (Markov) Network Types:

1. Gaussian Graphical Models (Continuous-Valued).
2. Ising Models (Binary-Valued).
3. Gaussian-Ising Models.

What about count-valued data? Others?



RNA-sequencing data? Methylation data?

Graphical Models via Exponential Families

- Key Assumption: Conditional distributions are Exponential Families.
  - Ex: Gaussian, Bernoulli, Poisson, Exponential, Negative Binomial, etc.

Review: Exponential Family Distributions.

$$P(X) = \exp\left(\theta\, B(X) + C(X) - D(\theta)\right)$$

- θ is the canonical parameter.
- $B(X)$ is the sufficient statistic.
- $C(X)$ is the base measure.
- $D(\theta)$ is the log-partition function.

Graphical Models via Exponential Families

Theorem

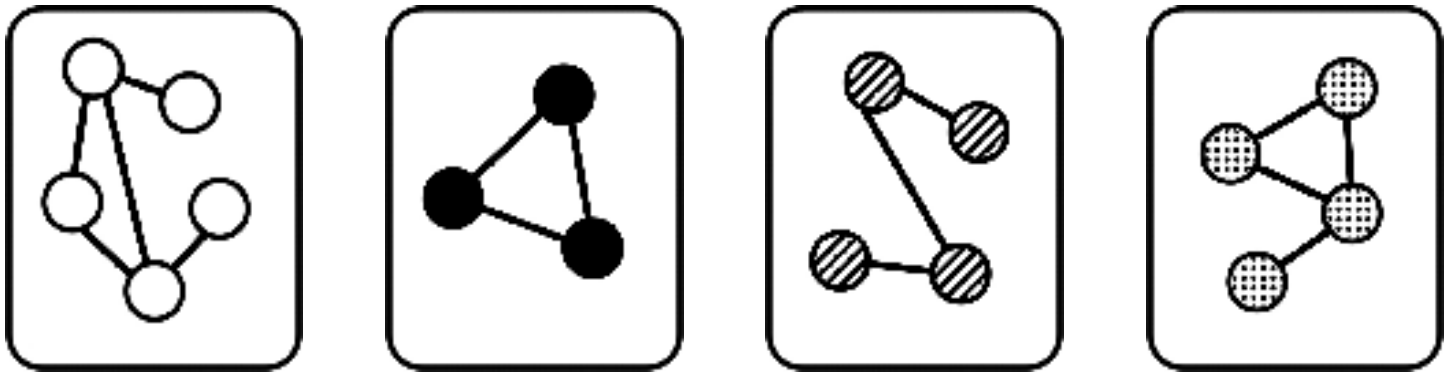
Joint Density necessarily has the form:

$$P(X) = \exp\Bigg\{ \sum_s \theta_s B(X_s) + \sum_{s \in V} \sum_{t \in N(s)} \theta_{st}\, B(X_s) B(X_t)$$

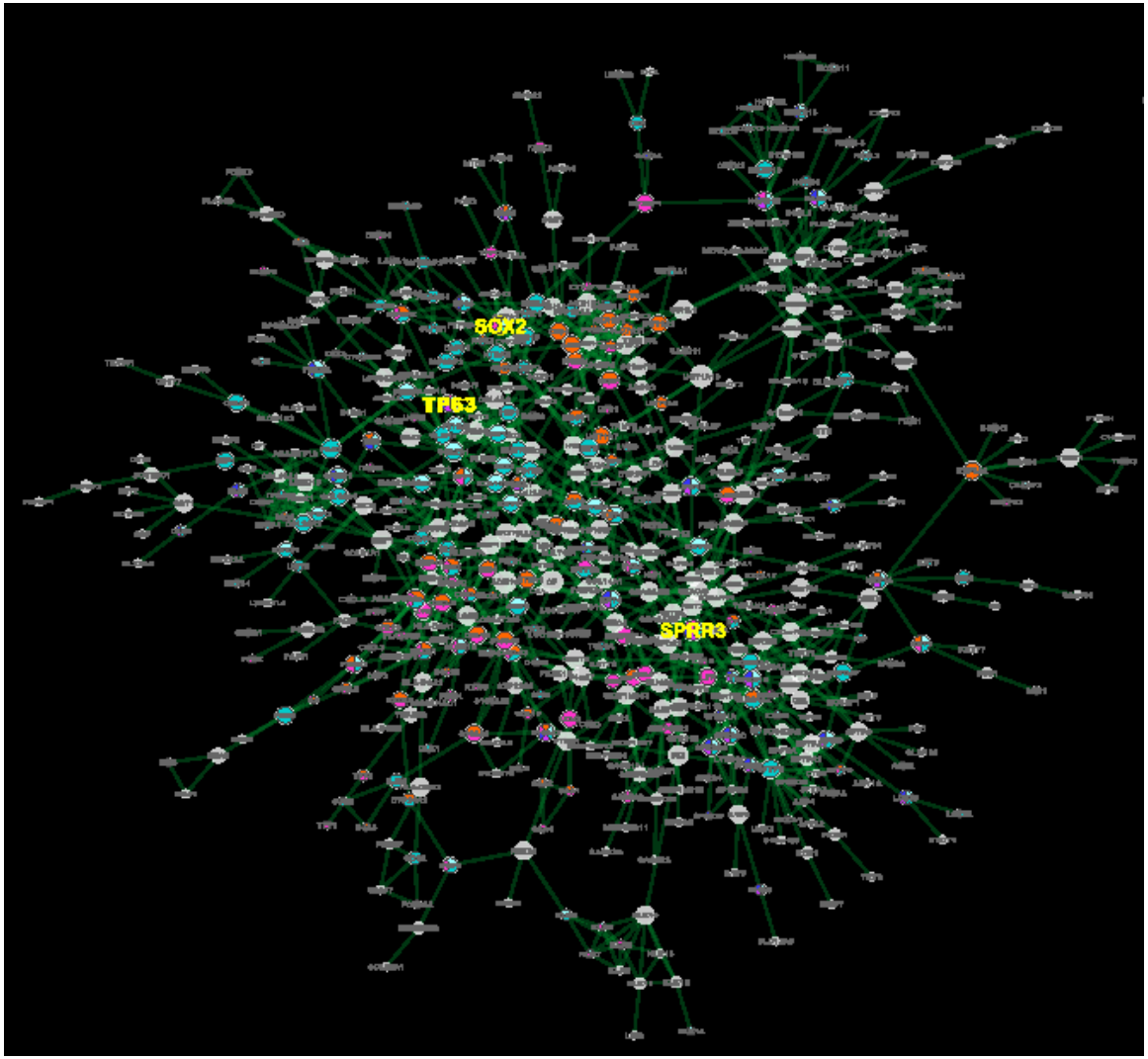
$$+ \sum_{s \in V} \sum_{t_2, \ldots, t_k \in N(s)} \theta_{s \ldots t_k}\, B(X_s) \prod_{j=2}^{k} B(X_{t_j}) + \sum_s C(X_s) - A(\theta) \Bigg\}$$

- Network inference via penalized conditional maximum likelihood estimation (neighborhood selection).
- Penalized (and possibly constrained) GLMs!
- Strong theoretical guarantees for parameter estimation and network recovery.

Graphical Models via Exponential Families

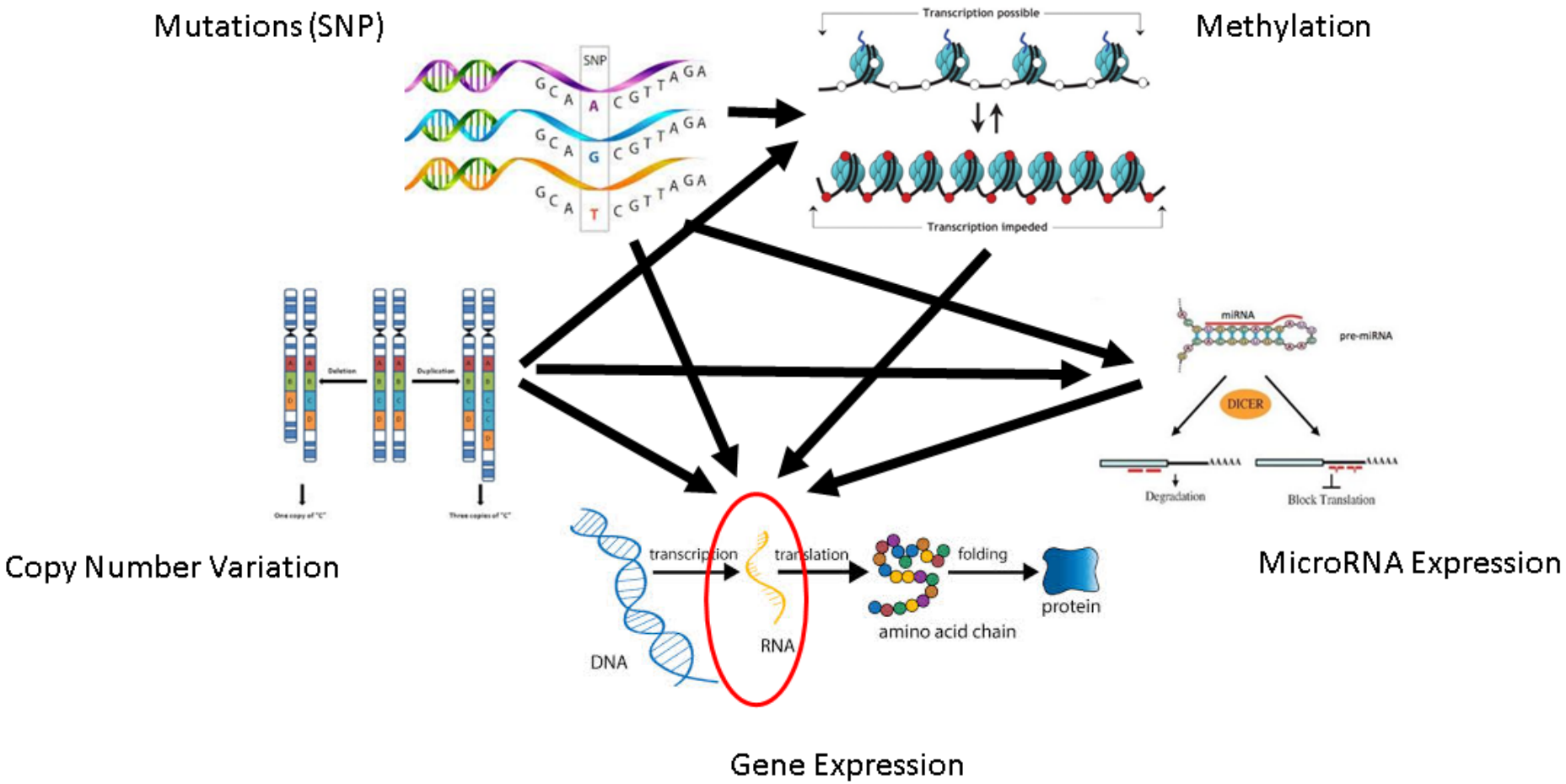


Graphical Models via Exponential Families



Lung Cancer Gene Expression Network (RNA-Seq)
Inferred via Poisson Graphical Model.

Integrated Network Models



Integrated Network Models

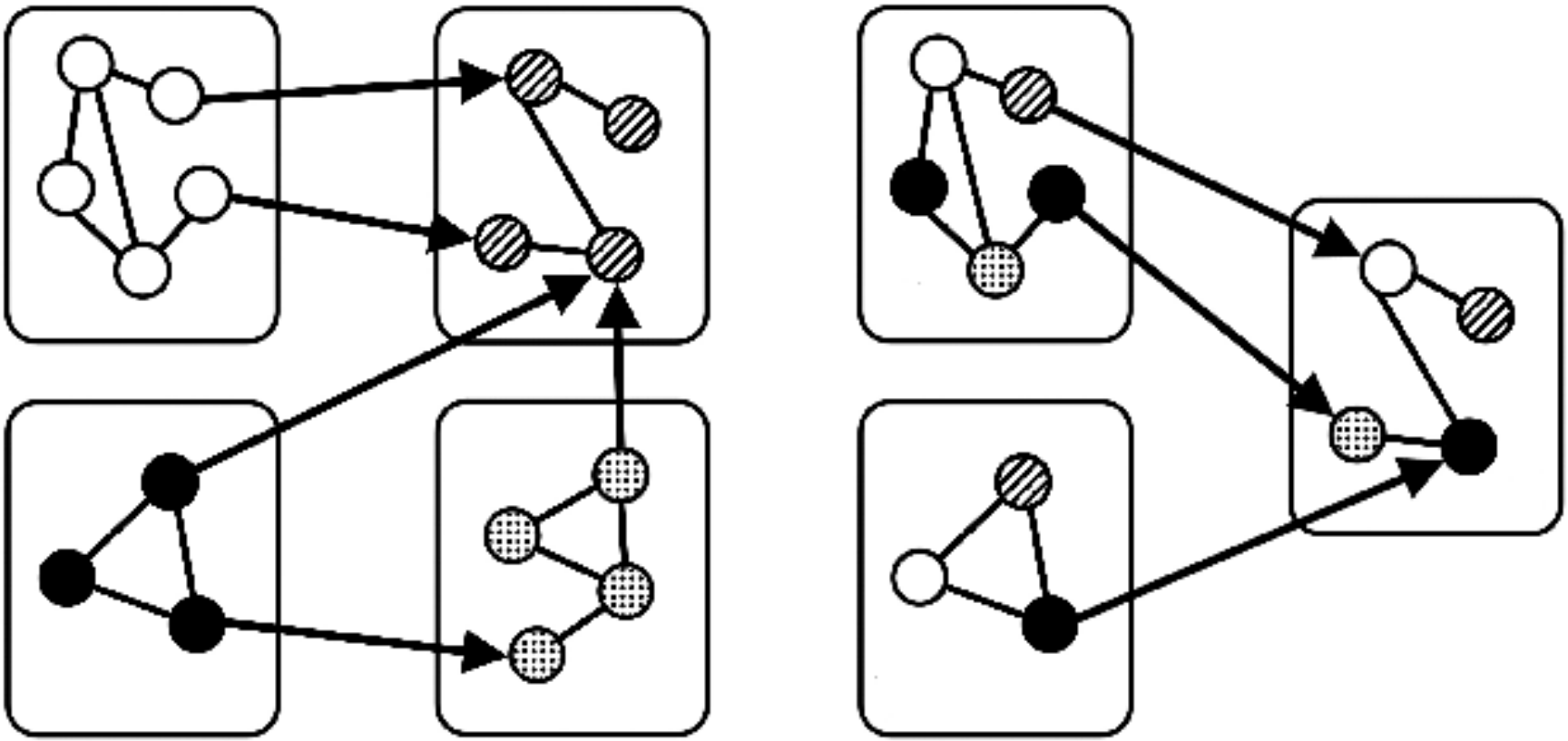
Mixed Chain Graphical Models via Exponential Families.

- Key Assumptions:
  - Conditional distributions are (different) Exponential Families.
  - Variables belong to known groups and the directionality of dependencies between groups is known.

- Theorem: Joint integrated distribution exists and has a closed form!
  - Dependencies parameterized by products of sufficient statistics from different distributions.
  - Strong statistical guarantees for network inference (penalized MLEs).
  - Permits wide-range of dependence structures.

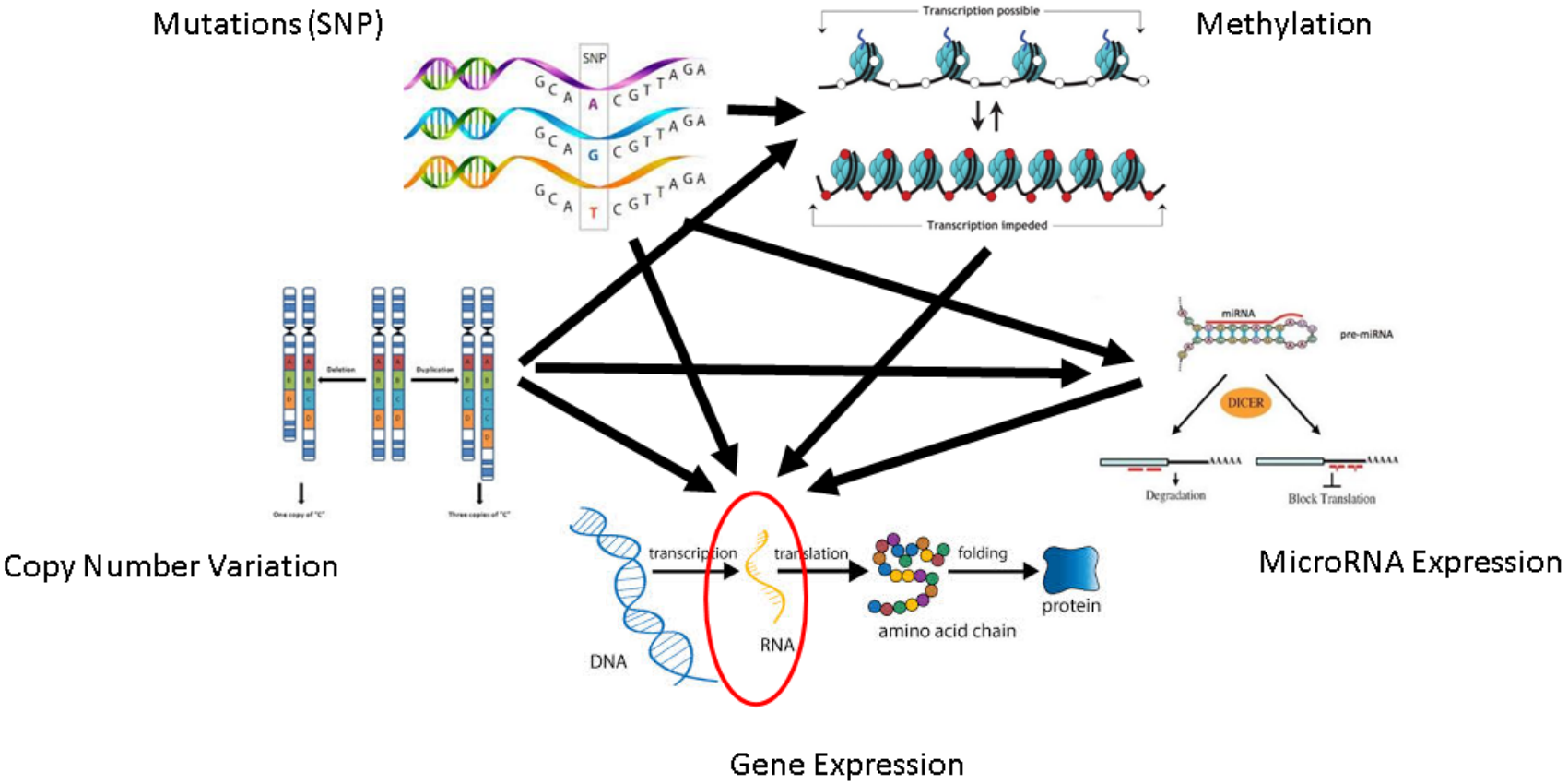
Integrated Network Models

Mixed Chain Graphical Models via Exponential Families.



Integrated Network Models

Mixed Chain Graphical Models via Exponential Families.

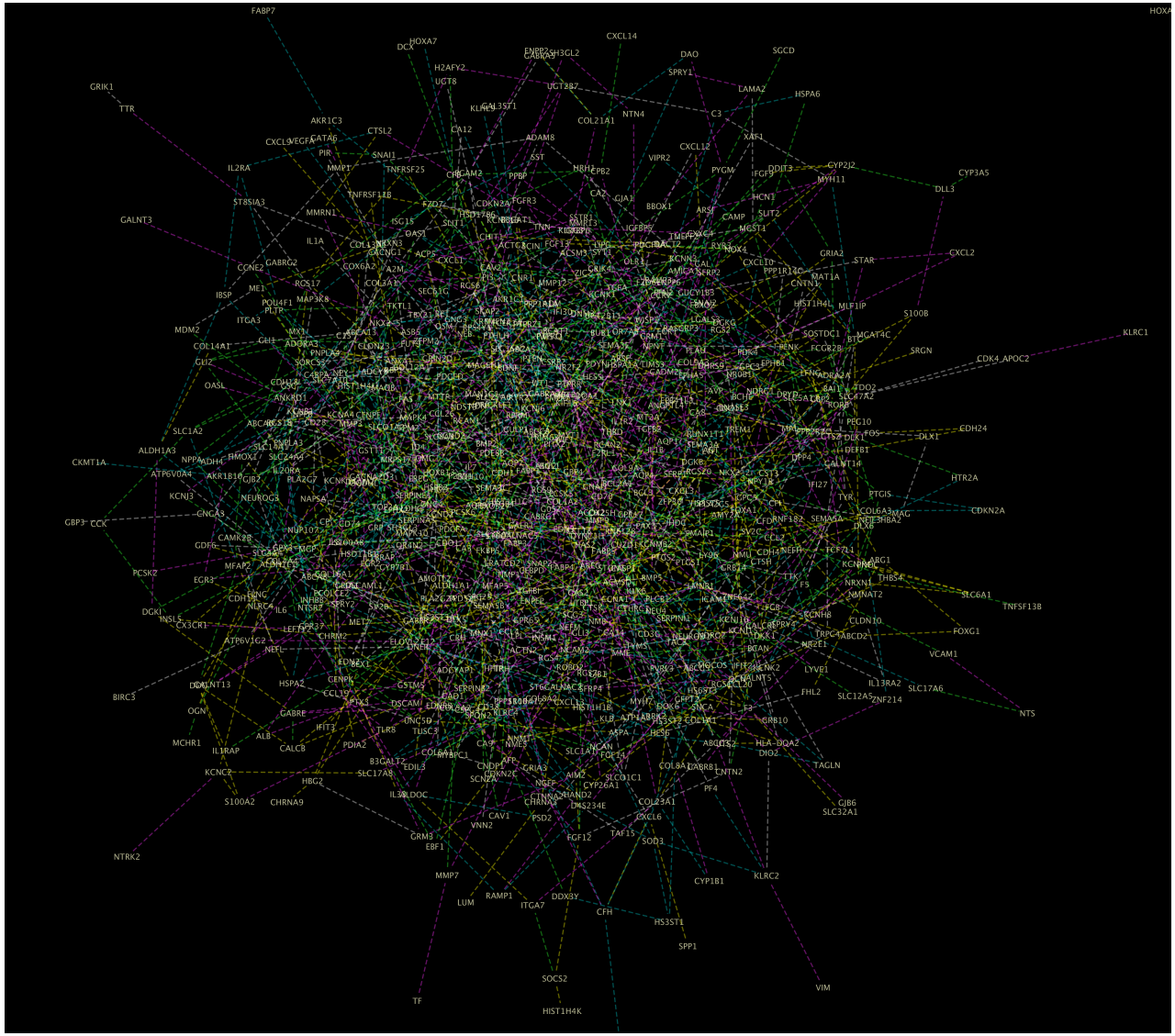


Applications & Implications

Implication

First multivariate distribution that can directly parameterize dependencies for mixed data types.

Applications & Implications



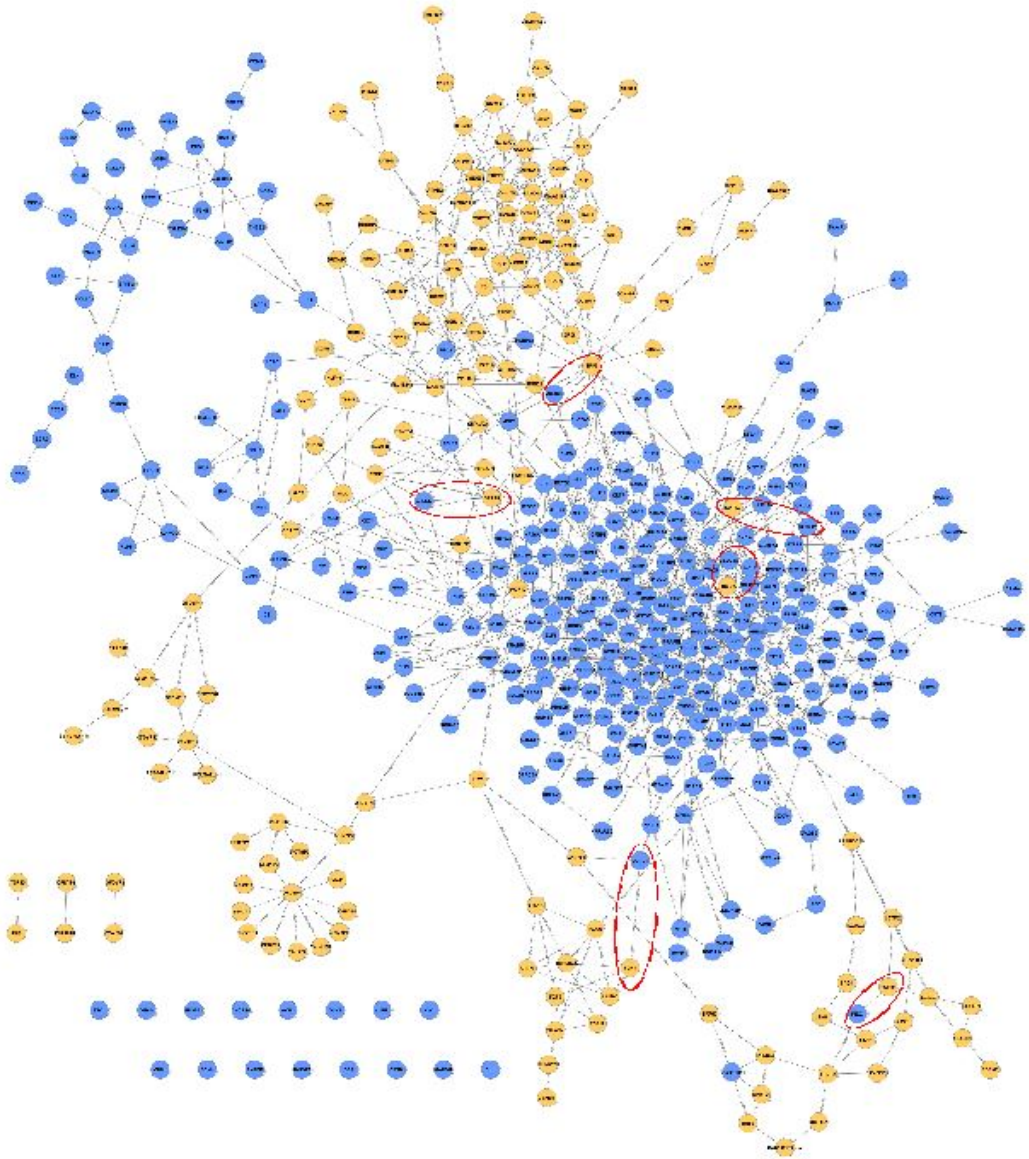
Glioblastoma Integrated Network.

Applications & Implications

Breast Cancer Integrated Mutation-Gene Expression Network.

Blue nodes: Genes (RNA-Sequencing - counts)
Yellow nodes: Mutations & Aberrations (aggregated at the gene level - binary)

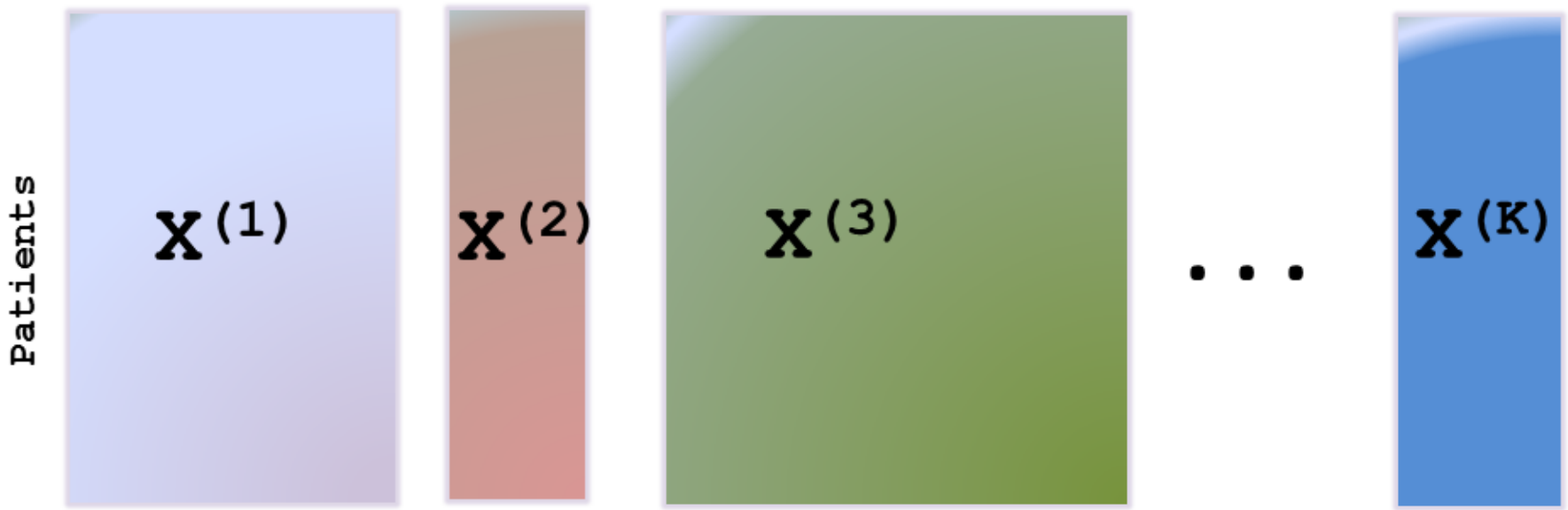
Inferred via Poisson-Ising Graphical Model.



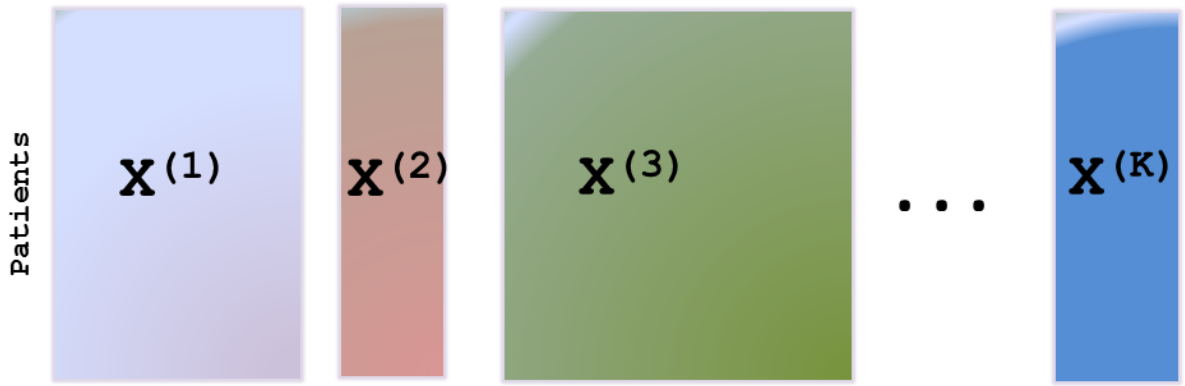
Challenge: Feature Selection from Mixed Data

Problem

Inferring Integrative Graphical Models via Exponential Families requires performing feature selection for GLMs from Mixed Multi-Modal Data.



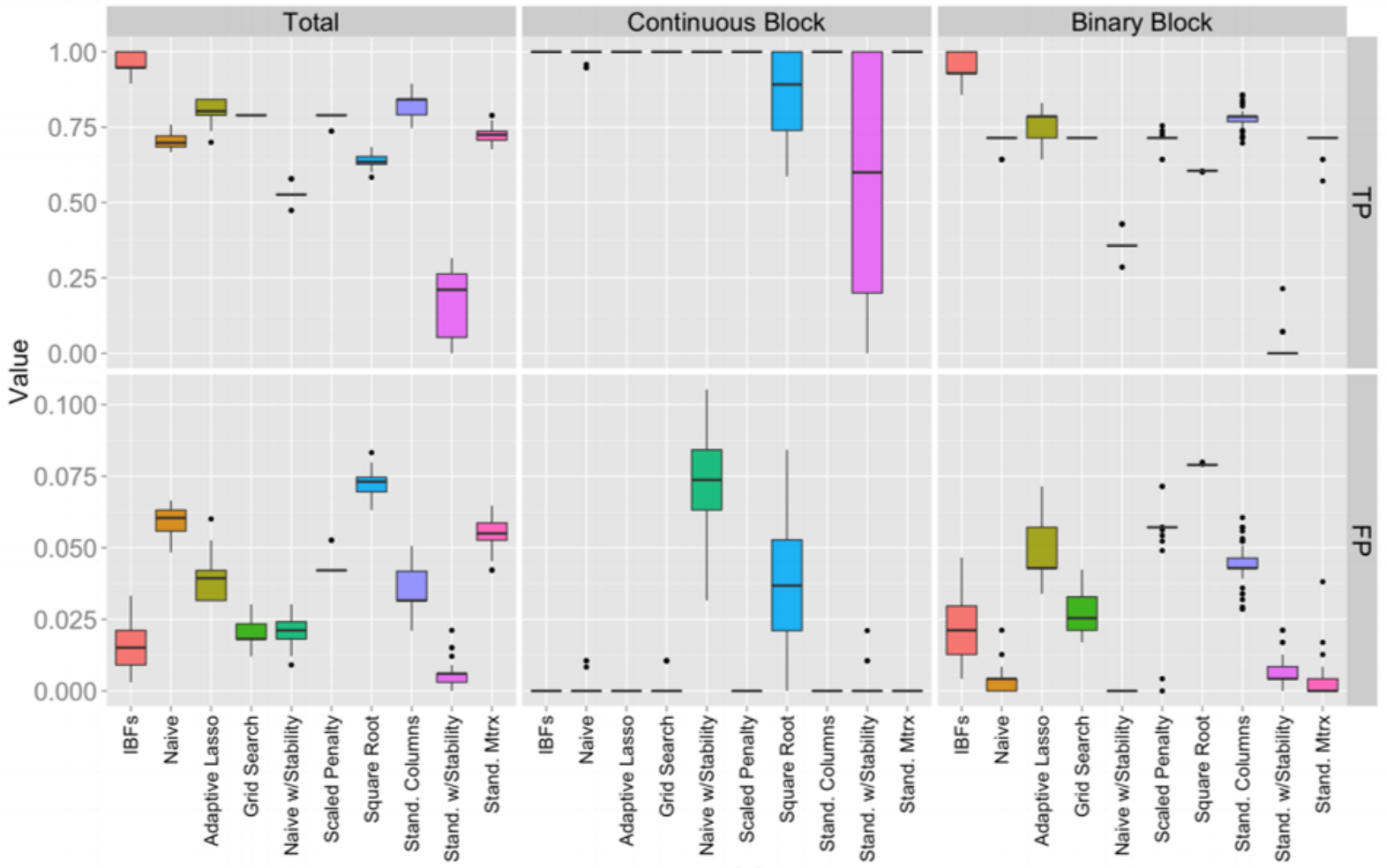
Challenge: Feature Selection from Mixed Data



Feature Selection Challenges:

1. Data-Modalities on Different **Scales**.
   - Requires different regularization levels.
   - Standardizing binary, count-valued, skewed variable etc. can make things worse!
2. Signal Interference across Data-Modalities.
   - Correlation within and between Data-Modalities can obscure weaker signals from other Data-Modalities.

Challenge: Feature Selection from Mixed Data

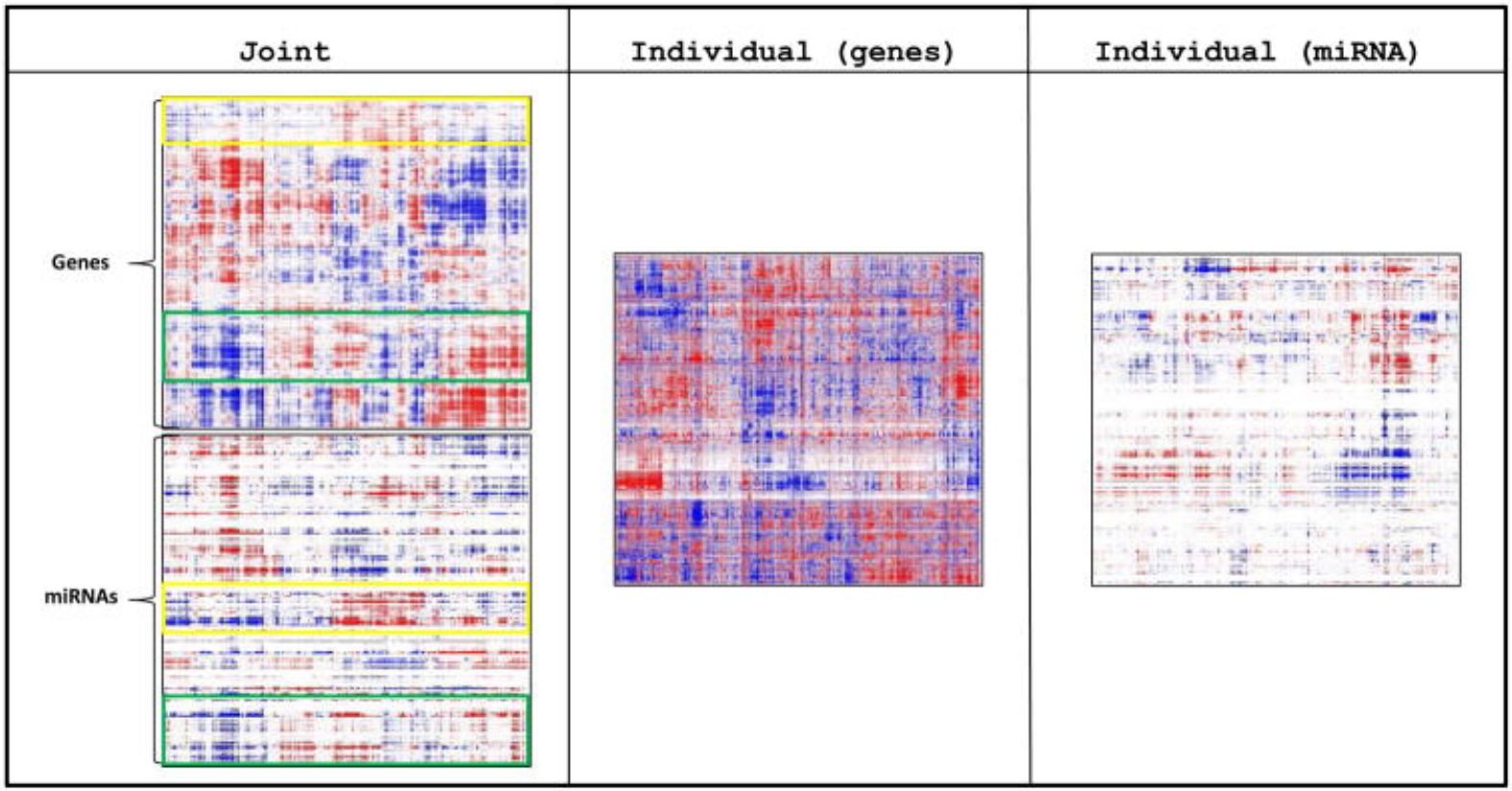


Simulation: Lasso with $n = 100$, $p_1 = 250$ & $p_2 = 250$.

1. Motivating Case Study: The Cancer Genome Atlas
2. Motivating Case Study: The ROS & MAP Studies
3. Challenges & Open Problems
   - Data Challenges
   - Statistical Challenges
     - Multivariate Modeling Challenges: Integrative Networks
     - Integrative Data Mining Challenges
4. The Big(ger) Picture

Integrative Dimension Reduction / Pattern Recognition

Existing Approaches:



Joint & Individual Variation Explained (JIVE; Lock et al., 2013).

Integrative Dimension Reduction / Pattern Recognition

Existing Approaches:

- Modern Canonical Correlations Analysis (same types of data).
- Coupled PCA / Coupled Matrix Decompositions (same types of data).
- PCA for Exponential Families (single data sets).

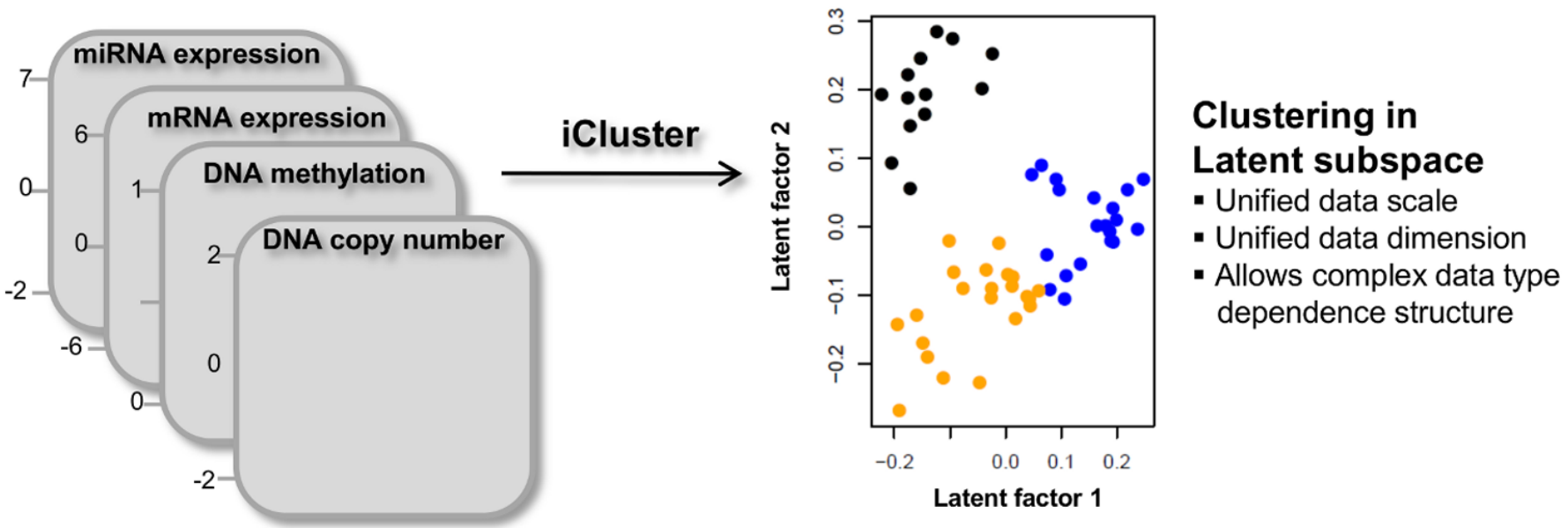
Open Statistical Problem: Dimension Reduction / Pattern Recognition for Mixed Multi-Modal Data.

Major Challenges:

- Scaling, regularization, signal interference across mixed data-modalities.
- Missing / unaligned data chunks.
- Reproducible discoveries & inference in high-dimensions.

Integrative Clustering

Existing Approaches:



Integrative Clustering (iCluster; Shen et al., 2009).

Integrative Clustering

Existing Approaches:

- Integrative Clustering (via latent variables and coupled matrix factorizations; same type of data).
- Exponential Family Clustering (single data set).

Open Statistical Problem: Clustering for Mixed Multi-Modal Data.

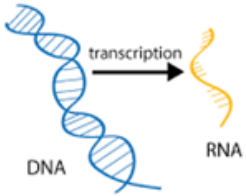
Major Challenges:

- Scaling, regularization, signal interference across mixed data-modalities.
- Reproducible discoveries & inference in high-dimensions.
- Missing / unaligned data chunks.

Longitudinal Modeling & Imaging-Genetics

Baseline: GWAS

Annual: Independent Clinical Evaluation & Cognitive Testing

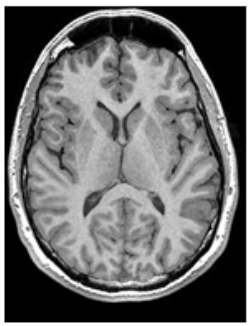


Post-Mortem: Transcriptomics & Epigenetics

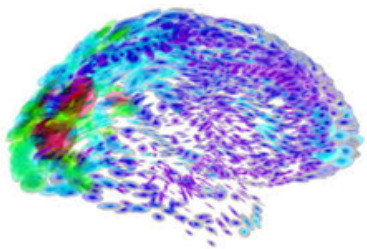
Post-Mortem: Neuroimaging

Baseline: Demographics, Behavioral, Environmental, Social, and Lifestyle Factors

Biannual: Structural Neuroimaging (2009-present)



Post-Mortem: Comprehensive Neuropathology



Longitudinal Modeling & Imaging-Genetics

Existing Approaches:

- Multi-modal neuroimaging (e.g. multi-parametric structural MRI, MRI-DTI, fMRI-EEG, fMRI-DTI).
- Imaging-genetics (mostly GWAS; mostly univariate).
- Sparse, uneven longitudinal mixed effects models.
- Variable selection for longitudinal mixed effects models.

Open Statistical Problems:

- Integrate: Integrative Genetics + Integrative Neuroimaging.
- Data Integration for Longitudinal Data.
- Modern / High-Dimensional Longitudinal Mixed Effects Models.
- Missing and Longitudinally Unaligned Data-Modalities.

1. Motivating Case Study: The Cancer Genome Atlas
2. Motivating Case Study: The ROS & MAP Studies
3. Challenges & Open Problems
   - Data Challenges
   - Statistical Challenges
     - Multivariate Modeling Challenges: Integrative Networks
     - Integrative Data Mining Challenges
4. The Big(ger) Picture

Integration of Multiple Studies & Data Modalities

AMP-AD: Accelerating Medicine Partnership - Alzheimer's Disease

Accelerating Alzheimer's Research and Drug Development

AMP-AD is an initiative of the Accelerating Medicines Partnership (AMP), a bold new venture among the NIH, 10 biopharmaceutical companies, and several nonprofit organizations aiming to transform the current model for developing new diagnostics and treatments for chronic diseases.

- $92.5 M partnerships between government (NIH-NIA) & industry.

Integration of Multiple Studies & Data Modalities

AMP-AD Data Available:

- **Disease**
- *Alzheimers Disease*
- Progressive Supranuclear Palsy
- Tauopathy, Pathological Aging
- Parkinsons Disease
- Mild Cognitive Impairment
- Amyotrophic Lateral Sclerosis
- Corticobasal Degeneration
- Autosomal Dominant Parkinsons Disease
- Frontotemporal Dementia

- **Tissue type**
- Temporal Cortex
- Frontal Pole
- Occipital Visual Cortex
- Inferior Temporal Gyrus
- Middle Temporal Gyrus
- Superior Temporal Gyrus
- Posterior Cingulate Cortex
- .....

- **Data type**
- RNA-seq
- Array Genotype
- Imputed Genotype
- Clinical
- miRNA nanostring
- H3K9Ac ChIP-Seq
- DNA Methylation
- Mass Spectrometry
- Gene Expression
- Array Genotype
- Array Expression
- Genotype
- eSNP Results
- Nanostring Expression
- Exome sequencing
- Coexpression Networks
- TLR Genotype
- Confocal imaging REST

- **Study**
- ROSMAP
- Emory
- ACT
- BLSA
- HBTRC
- MSBB
- TAUAPPms
- MayoEGWAS
- MayoLOADGWAS
- MayoPilot
- MayoRNAseq
- MayoBB
- MayoLOADGWAS
- IL10
- BroadiPSC
- Upenn
- TAUmicroglial
- MSMM
- MSDM

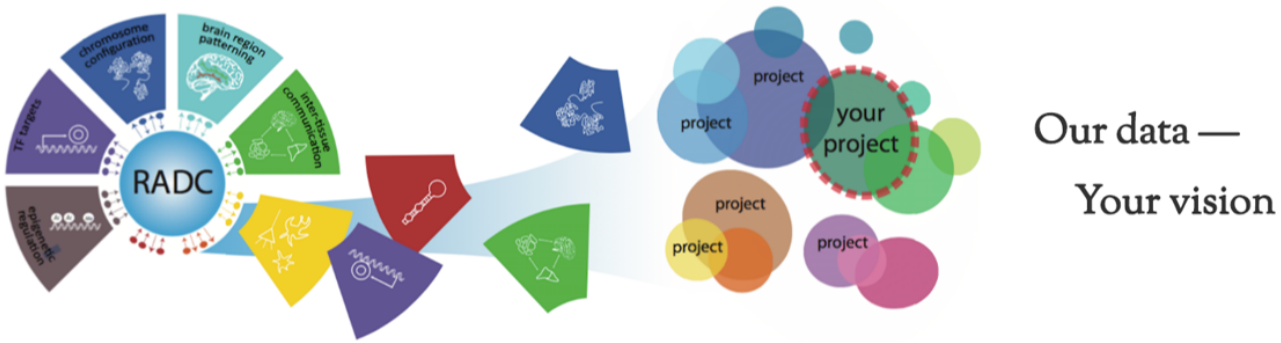
- **Center**
- Broad-Rush
- Emory
- Mount Sinai
- UFL-Mayo-ISB
- Myers-NIAGADS
- Harvard-MIT

NIH National Institute on Aging
Turning Discovery Into Health

Ongoing Challenges

1. Data Access, Data Quality & Data Privacy.
2. Reproducible Research.
3. Reproducible Data-Driven Discoveries & Inference.

RADC Research Resource Sharing Hub



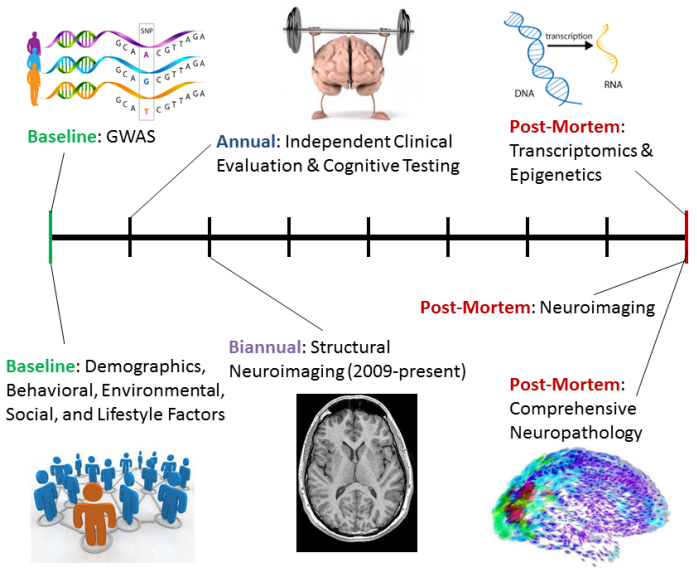
The Rush Alzheimer's Disease Center (RADC), one of 29 Alzheimer's disease (AD) Research Centers across the country designated and funded by the National Institute on Aging (NIA), is dedicated to supporting research about the cause, treatment, and prevention of AD, other dementias, and a range of other common chronic conditions of aging. The many RADC studies generate an enormous variety of unique data and biospecimens to support this effort. RADC faculty and staff are committed to sharing these resources with the wider aging and AD research community to accelerate the pace at which new knowledge is created for the treatment and prevention of dementia and other age-related chronic neurologic conditions.

The **RADC Research Resource Sharing Hub** was specifically designed to help you, the non-RADC investigator, navigate the complex data and biospecimens available for sharing, and to assist you in identifying data and biospecimens that you can use to support your own projects. We invite you to explore the site, see what is available, and submit your data and/or biospecimen request.

—David A. Bennett, MD, Director of RADC

Ongoing Challenges

1. Data Access, Data Quality & Data Privacy.
2. Reproducible Research.
3. Reproducible Data-Driven Discoveries & Inference.



- Different data-modalities go through different processing and quality control procedures, often performed by different people using different software, and etc.

Ongoing Challenges

1. Data Access, Data Quality & Data Privacy.
2. Reproducible Research.
3. Reproducible Data-Driven Discoveries & Inference.

- Multi-Step Analysis Pipeline: Large-Scale Multi-Modal Data goes through different (stochastic) processing, feature extraction, and feature learning pipelines.
- Inference typically conducted on the last step.
- For inference to be valid, must account for the stochastic nature of the *entire analysis pipeline*, not just the last step.
- Challenge: Most processing, feature extraction, feature learning, and machine learning produce estimators whose distributions are unknown (PSI).

Summary

Summary

- Analysis of individual Data-Modalities has yielded limited progress for several complex diseases.
- Data Integration is needed across Data-Modalities to better understand disease etiology, prognosis, and discover new therapies.
- Statistical Challenges:
  - Data-Driven Discoveries with mixed data, unaligned data, & longitudinally unaligned data.
  - Inference with mixed, multi-modal data & multi-step pipeline analyses.

Summary

Summary

- Analysis of individual Data-Modalities has yielded limited progress for several complex diseases.
- Data Integration is needed across Data-Modalities to better understand disease etiology, prognosis, and discover new therapies.
- Statistical Challenges:
  - Data-Driven Discoveries with mixed data, unaligned data, & longitudinally unaligned data.
  - Inference with mixed, multi-modal data & multi-step pipeline analyses.

What does it take?

1. A Team!
2. A willingness to get dirty!
3. Long term objectives, planning, and resources.

Acknowledgments

My Group:

- John Nagorski.
- Yulia Baker.
- Tinayi Yao.
- Elizabeth Sweeney.
- Michael Weylandt.
- Frederick Campbell.

Collaborators:

- Zhandong Liu, BCM.
- Ying-Wooi Wan, BCM.
- Joshua Shulman, BCM.
- Matthew Anderson, BCM.

ROS/MAP Collaborators:

- David Bennett, Rush.
- Sue Leurgans, Rush.
- Konstantinos Arfanakis, Rush.

Thank You!