MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

UNIVERSITY OF MICHIGAN

Data integration with diverse datasets

Alfred Hero

Co-director, Michigan Institute for Data Science

Dept. of EECS, Dept. of BME, Dept. of Statistics

University of Michigan

May 20, 2016

CATS Workshop on Refining the Concept of Scientific Inference When Working With Big Data

midas.umich.edu

Acknowledgements

- Bala Rajaratnam - Stanford University
- Geoff Ginsburg and the Duke team – Duke University
- Yongsheng Huang – Merck Research
- Pin Yu Chen – Univ of Michigan

- ARO MURI Non-commutative Information Structures
- ARO MURI Value of Information
- ARO Social Analytics
- ARO Nonlinear Dynamics of Reaction Networks
- AFRL ATR Center Program
- DARPA Biochronicity Program
- DOE Consortium for Verification Technology Program
- NSF: Theoretical Foundations Program
- NIH Biomedical Imaging Institute

Outline

1. Integration of diverse HD data
2. Illustrative example
3. Network inference
4. Concluding remarks

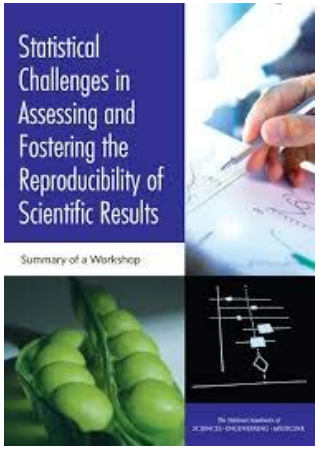
Outline

1. Integration of diverse HD data
2. Illustrative example
3. Network inference
4. Concluding remarks

The Value of Big Data for Inference

- Big data: demonstrated value in private and public sector [FTC 2016]
- Caveat: bigger data does not necessarily lead to better inferences
- Accuracy of predictions using big data is severely affected by biases
  - Lack of replicability [CATS 2016]: under-sampling, variability across studies
  - Confusion of correlation vs causation: observational vs interventional data
- It is important to quantify and unmask hidden biases
  - Strength of evidence for predictor accuracy: confidence intervals, p-values
  - Necessary conditions for replicability: sample size, noise, missing values

Federal Trade Commission, Jan 2016



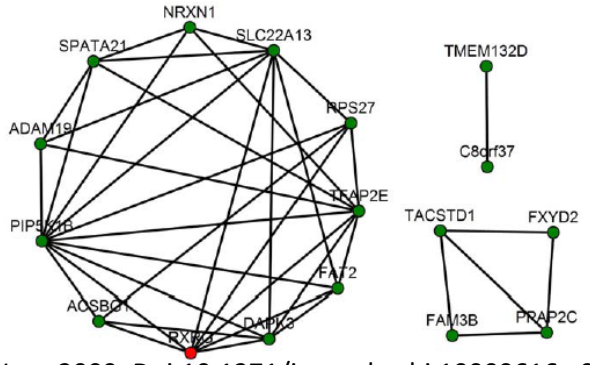
CATS, Natl. Acad. 2016

Data diversity and integration

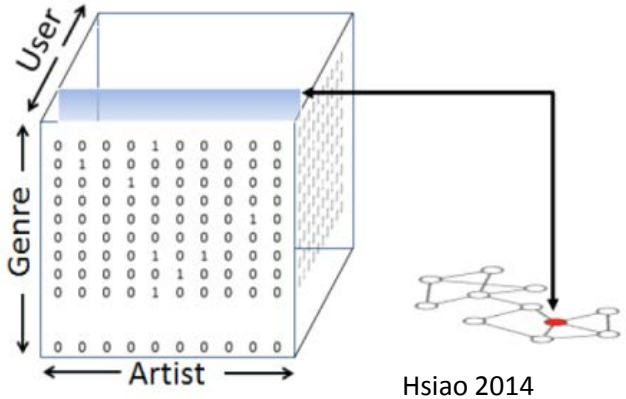
- Most Big Data applications make use of diverse data sources for
  - Better predictors: targeted advertising, credit worthiness, learning outcome
  - Better descriptive models: 3D Nucleome, Connectome, ImageNet
- But good integration requires more sophisticated methods
  - Integration with weighting or normalization of the data sources
  - Assessment of bias and replicability is more difficult, esp in high dimension
- Data integration falls into several categories
  1. Integration of diverse data within a single study
  2. Integration of primary data across several studies
  3. Integration of meta-data across several studies

Integration of diverse data within a single study

- Experimenter controls data heterogeneity and experiment design
  - Technical noise diversity: batch effects, misregistration, multi-modalities
  - Biological diversity: mixed population, longitudinal study, inter-species
- Examples
  - Integrated Personal Omics Profile (IPOP) [Chen 2012]
  - Inferring inter-species co-expression networks [Wang 2009]
  - Socio-collaborative database retrieval systems [Hsiao 2015]



Wang 2009: Doi:10.1371/journal.pcbi.10000616.g006

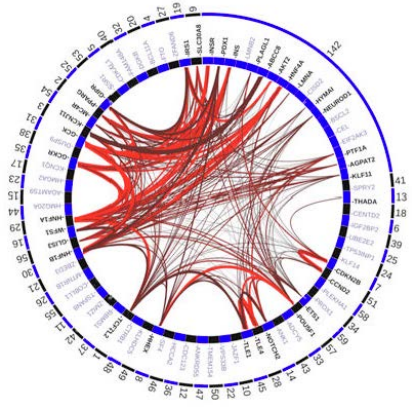


Hsiao 2014

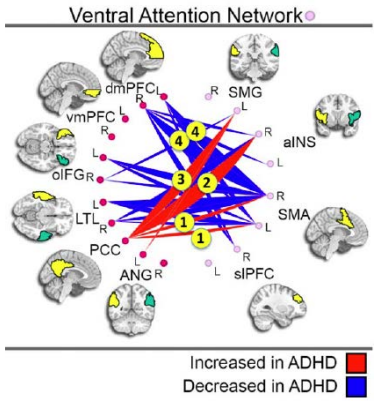
*Chen et al, **Personal omics profiling reveals dynamic molecular and medical phenotypes**, *Cell* 2012.
*Wang et al, **Meta-analysis of Inter-species Liver Co-expression Networks Elucidates Traits Associated with Common Human Diseases**, *PLoS Computational Biology*, 2008
*Hsiao, Kulesza, Hero, **Social collaborative retrieval**, *IEEE J Selected Topics in Signal Processing*, 2014.

Integration of primary data across studies

- Federated experiments: data heterogeneity is uncontrolled
  - Benefit of increased sample size can often lead to improved power
  - Must account for technical/biological variability, diverse sample coverage
- Examples
  - Combining genotype information across T2D GWAS studies [Morris 2013]
  - Combining ADHD-200 multisite brain connectomics studies [Sripada 2014]
  - Combining expertly and crowdsourced annotated databases [Deng 2009]



Morris 2013, PPI network GRAIL connectivity plot

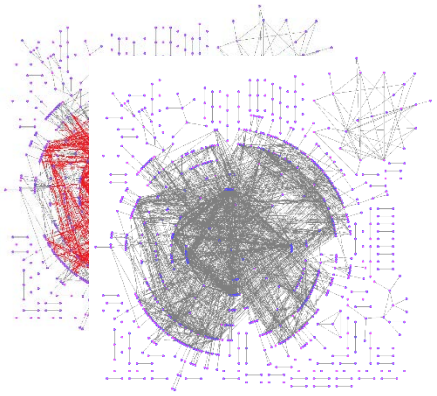


Sripada 2014: attention network

*Morris et al, **Large-scale association analysis provides insights into genetic architecture and pathophysiology of T2 diabetes**, *Nat Gen* 2013.
*Sripada et al, **Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder**, *Human brain mapping*, 2014.
*Deng et al, **Imagenet: A large-scale hierarchical image database**, *Computer Vision and Pattern Recognition, 2009*

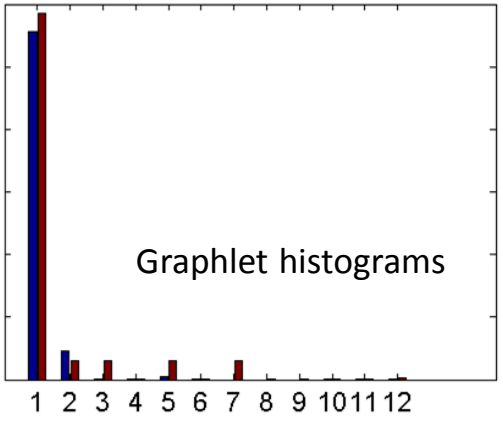
Integration of meta-data across studies

- Meta-analysis of meta-data from many related published studies
  - Combining aggregated effect sizes, computed p-values, imputed relations
  - Combining lists of variables, e.g. rankings of influential genes or pathways
- Examples
  - Combining p-values from multiple RNAseq studies [Rau 2014]
  - Aligning co-expression networks from multiple PPI studies [Singh 2008]
  - Identifying hub nodes in integrated correlation networks [Langfelder 2013]



Two interaction networks

Graphlet motifs

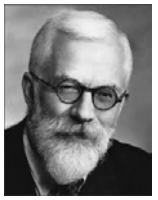


*Rau, Marot, Jaffrezik, **Differential meta-analysis of RNA-seq data from multiple studies**, *BMC Bioinformatics* 2014
*Singh et al, **Global alignment of multiple protein interaction networks with application to functional orthology detection**. *PNAS* 2008
*Langfelder et al, **When is hub gene selection better than standard meta-analysis?** *PLoS One*, 2013.

Statistical Principles for Data Integration

- How to best design integration function z such that $z(X,Y)$ integrates datasets X, Y?
- Ideally, the integration function should depend on **inference task**
- Assume model $f(X,Y|\theta) =$ joint density given a parameter θ
- Fisher's sufficiency principle: $z(X,Y) = T(X,Y)$ = the **minimal sufficient statistic** satisfying Fisher-Neyman Factorization,

$$f(X,Y|\theta) = g_\theta\big(T(X,Y)\big)h(X,Y)$$

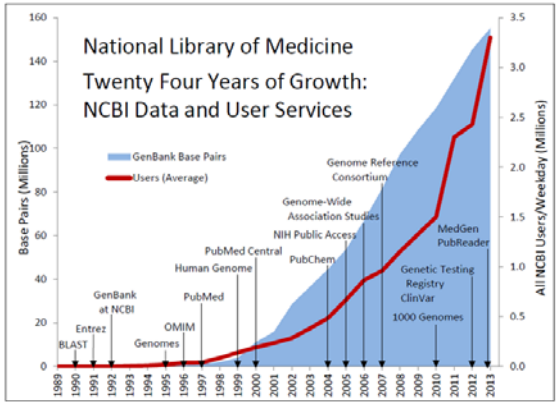



- Bayes imputation principle: if $\theta = Z$ is latent variable w/ prior $f(Z)$

$$z(X,Y) = \begin{cases} \text{amax}_Z f(X,Y|Z)f(Z), & \text{maximum mode imputation} \\ E[Z|X,Y], & \text{minimum MSE imputation} \end{cases}$$

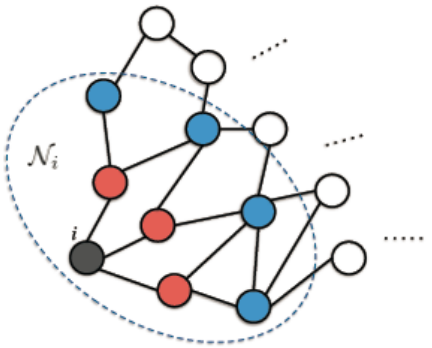
- "All statistical models are wrong but some are useful" (G. Box)

Challenges for Integrating Ultra HD Data



https://www.nlm.nih.gov/about/2015CJ.html

http://projectredcap.org/



Meng et al, 2014

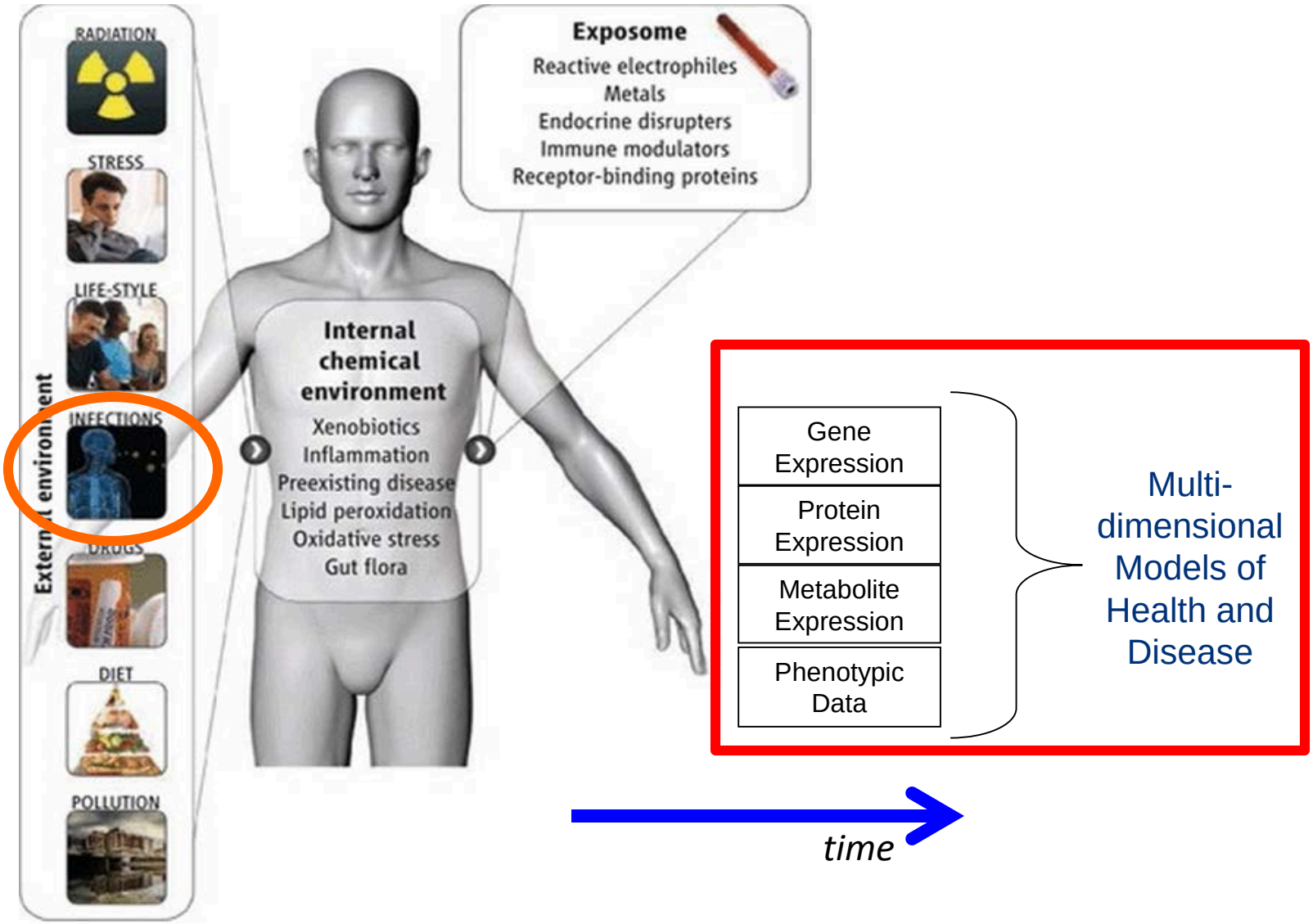
- Cloud data archiving is evolving into cloud computing capacity
- Full integration of ultra high dimensional data sets remains impractical
  - Distributed learning: local info sharing w/ loopy BP for GMs [Wainright 2008]
  - Sometimes local info sharing is **sufficient**, e.g., GGMs [Meng 2014]
- Privacy issues may allow only partial access to datasets
  - Privacy heterogeneity: site-specific levels of privacy protection [Song 2015]
  - Inference is complicated by challenging missing data problem [Duchi 2014]

Wainright and Jordan, M. Wainwright and M. Jordan**, Graphical models, exponential families, and variational inference**, *Foundations&Trends in ML* 2008.
Meng et al, **Distributed Learning of Gaussian Graphical Models via Marginal Likelihoods**, *IEEE Trans on Signal Processing*, 2014.
Song et al, **Learning from Data with Heterogeneous Noise using SGD,** *AISTATS* 2015
Duchi et al, **Privacy Aware Learning**, *Journal of the Association for Computing Machinery*, 2014

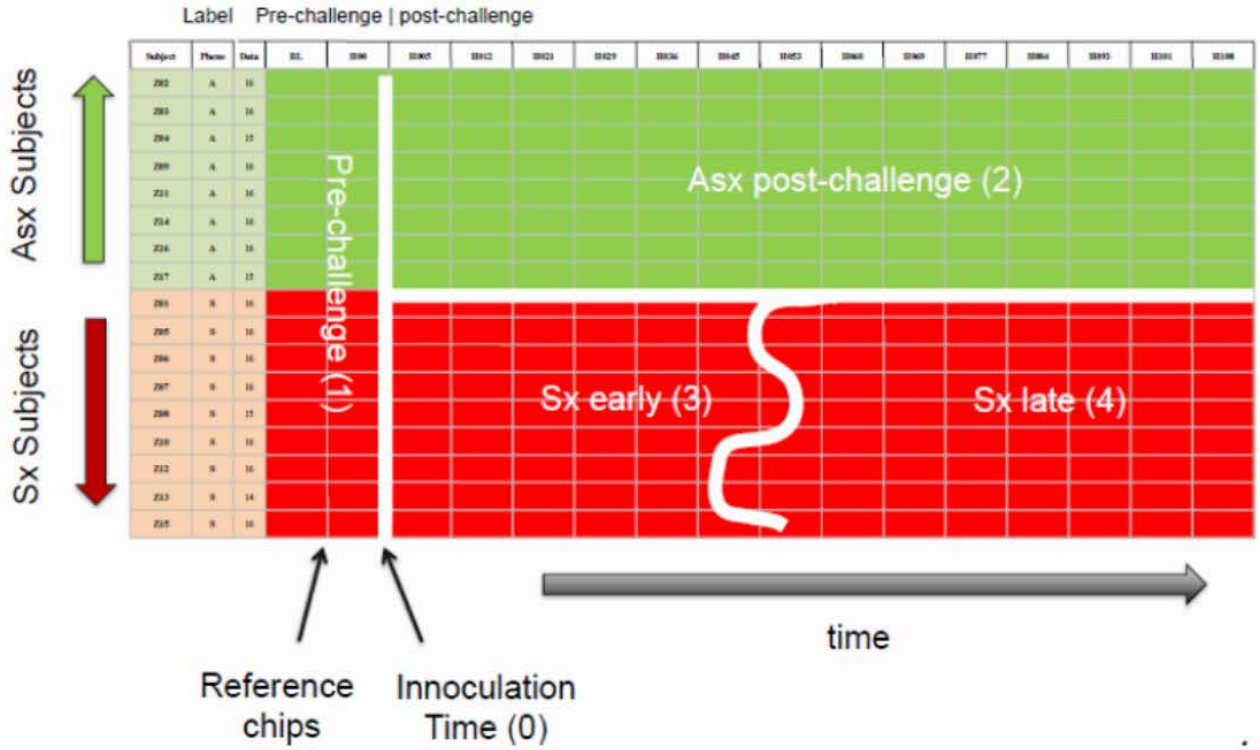
Outline

1. Integration of diverse HD data
2. Illustrative example
3. Network inference
4. Concluding remarks

Exposures, the Exposome, Exposomics



2006-2015 DARPA Predicting Health and Disease



Zaas *et al*, Cell, Host and Microbe, 2009

Chen *et al*, IEEE Trans. Biomedical Eng, 2010

Chen *et al* BMC Bioinformatics, 2011

Puig *et al* IEEE Trans. Signal Processing, 2011

Huang *et al*, PLoS Genetics, 2011

Woods *et al*, PLoS One, 2012

Bazot *et al*, BMC Bioinformatics, 2013

Zaas *et al*, Science Translation Medicine, 2014

Liu et al, BMC Bioinformatics, 2016

Mclain et al, J. Infectious Diseases, 2016

2006-2015 DARPA Predicting Health and Disease

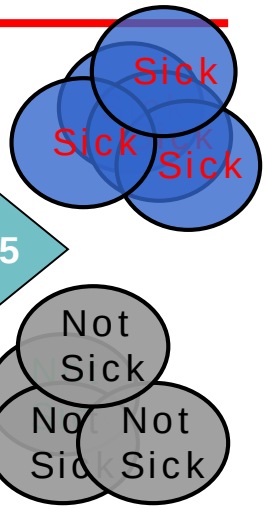
2 Collections

Day-1

HRV/RSV/INF Challenge (Day 0)

3X Daily Collections

Day 1 | Day 2 | Day 3 | Day 4 | Day 5



- Blood RNA Expression
- Blood proteins/metabolites
- Nasal/Breath/Urine cytkns
- Self Reported Symptoms

- Blood RNA Expression
- Blood proteins/metabolites
- Nasal/Breath/Urine cytkns
- Self Reported Symptoms

Inference tasks

Predict ARVI at 0.8T with 0.9 sensitivity at 0.1 specificity

Identify molecular pathways associated with acute Sx infection

Analytical Modeling

Analytical Modeling

DARPA PHD Team

PI: G. Ginsburg

Duke University

University of Michigan

University of Virginia

Children's Mercy Hospital

Retroscreen

ARVI – Acute Respiratory Viral Infection

PHD - 7 challenge studies performed*

- 121 human subjects in 7 cohorts quarantined for 5+ days
- Samples collected 3 times per day from each subject
- Over 16,600 samples assayed
- Attack rate was approximately 50%
- Factors complicating statistically accurate inference
  - High dimension, cohort variability, assay variability, missing samples

Challenge	Virus	Year	Location	IRB protocol	Duration (hrs)	# Subjects	# Time Points
DEE1	RSV	2008	Retroscreen	Pro00002796	166	20	21
DEE2	H3N2	2009	Retroscreen	Pro00006750	166	17	21
DEE3	H1N1	2009	Retroscreen	Pro00018132	166	24	20
DEE4	H1N1	2010	Retroscreen	Pro00019238	166	19	21
DEE5	H3N2	2011	Retroscreen	Pro00029521	680	21	23
HRV UVA	HRV	2008	Univ. of Virginia	Pro00003477	120	20	15
HRV Duke	HRV	2010	Duke Univ.	Pro00022448	136	30	19

*Data available on GEO, accession number GSE73072

Selected findings

- Novel factor analysis method isolates pan-viral Sx/Asx ARVI signature [1,2]
- Sparse ENet mRNA predictor developed and validated on all studies [3]
- ARVI signature appears as a strong *sentinel imprint* at baseline [4]
- Use of a personalized reference sample can improve predictor [5]
- Whole blood mRNA is overall best modality for ARVI prediction

Questions

- Could more baseline samples further improve prediction accuracy?
- Could some combination of non-invasive modalities do as well as blood?
- How well do findings generalize to a wildtype (unquarantined) sample?

New DARPA Biochronicity study designed to answer these questions

[1] Huang et al, **Temporal Dynamics of Host Molecular Responses Differentiate Sx and Asx Influenza …,** *PLoS Genetics*, 2011

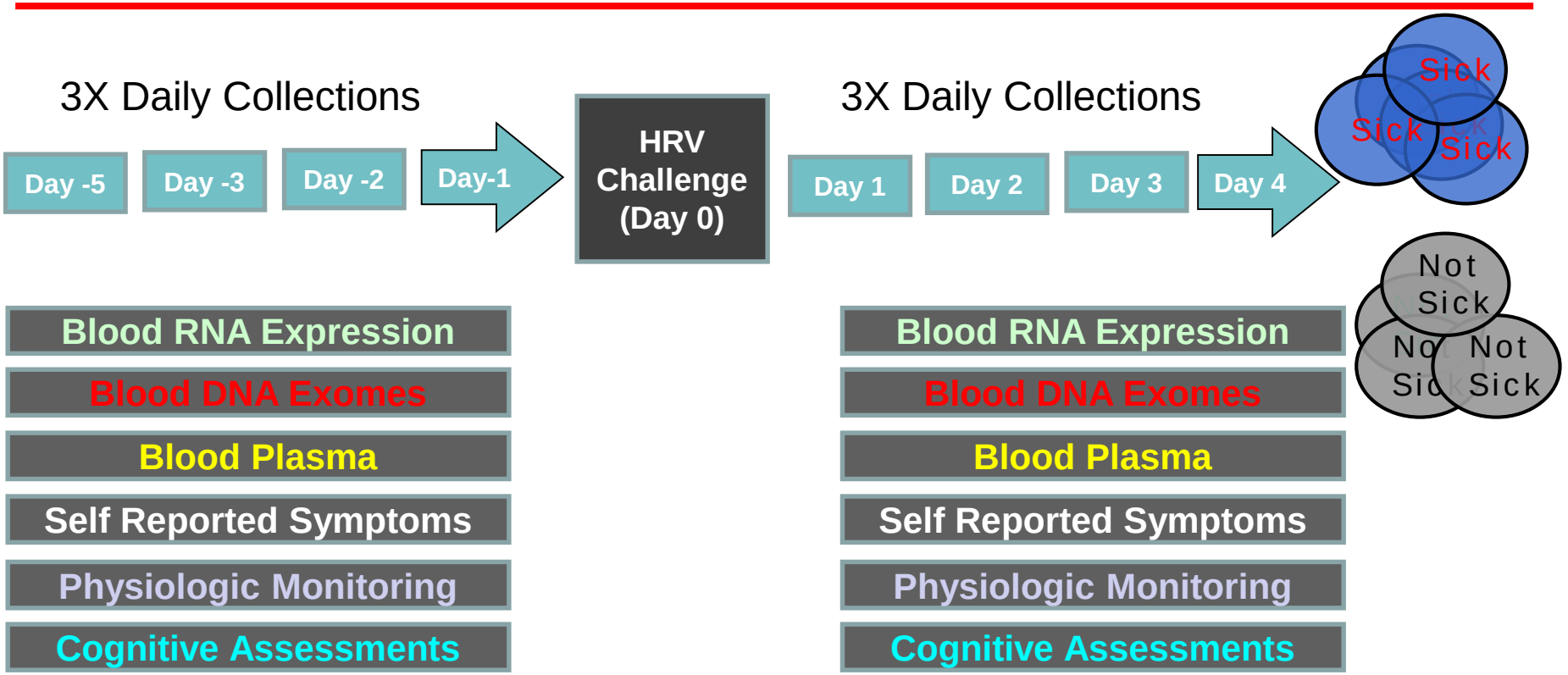
[2] Woods et al, **A Host Transcriptional Signature for Pre-symptomatic Detection of Infection…,** *PLoS One*, 2013

[3] Woods et al, **A Host-Based RT-PCR Gene Expression Signature to Identify ARVI,** *Sci Transl Med, 2013*

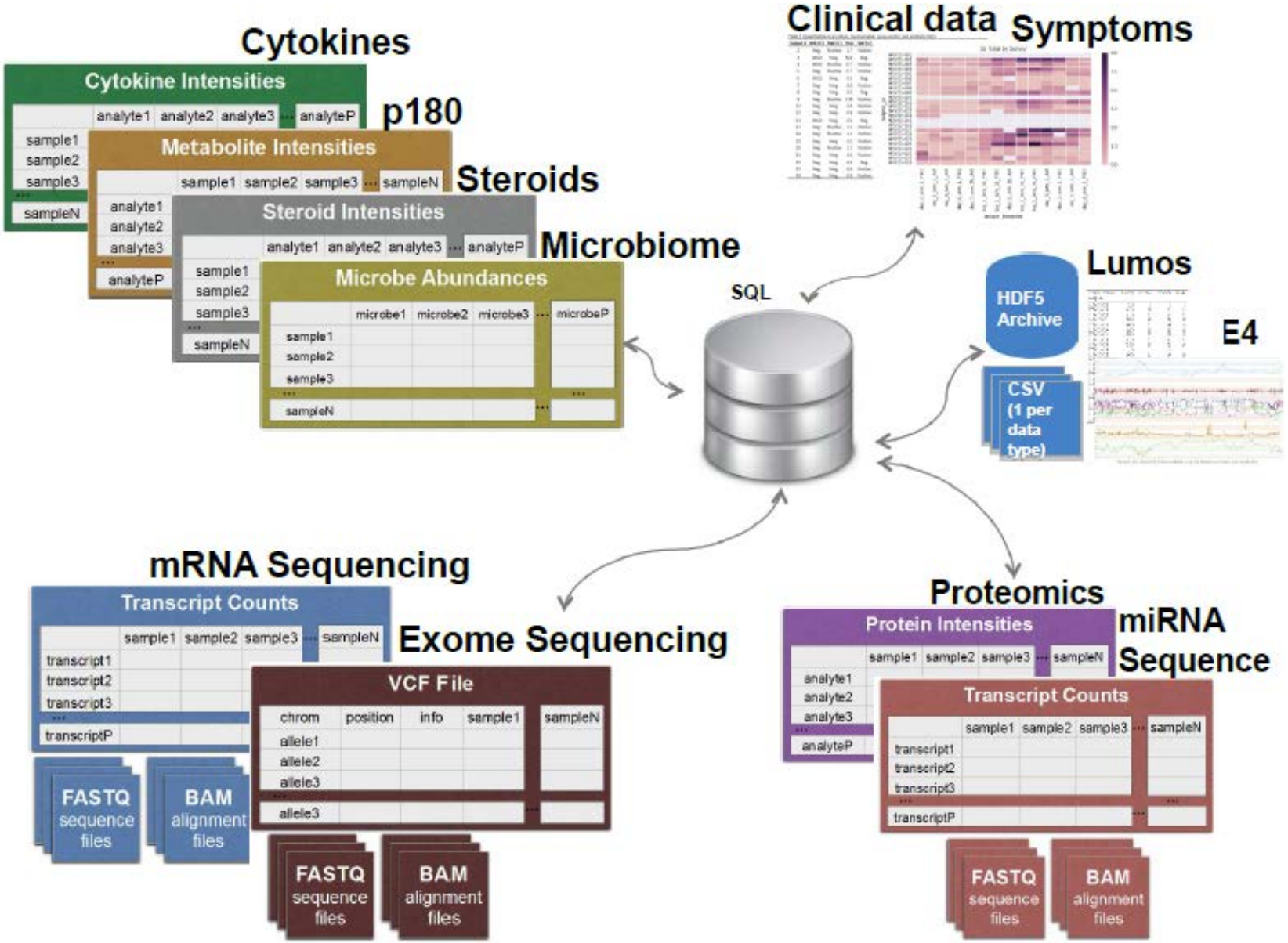
[4] Hero and Rajaratnam, **Large scale correlation mining for biomolecular network discovery**, *in Big Data Over Networks 2015*

[5] Liu et al, **An individualized predictor of health and disease using paired reference and target,** *BMC Bioinformatics* 2016.

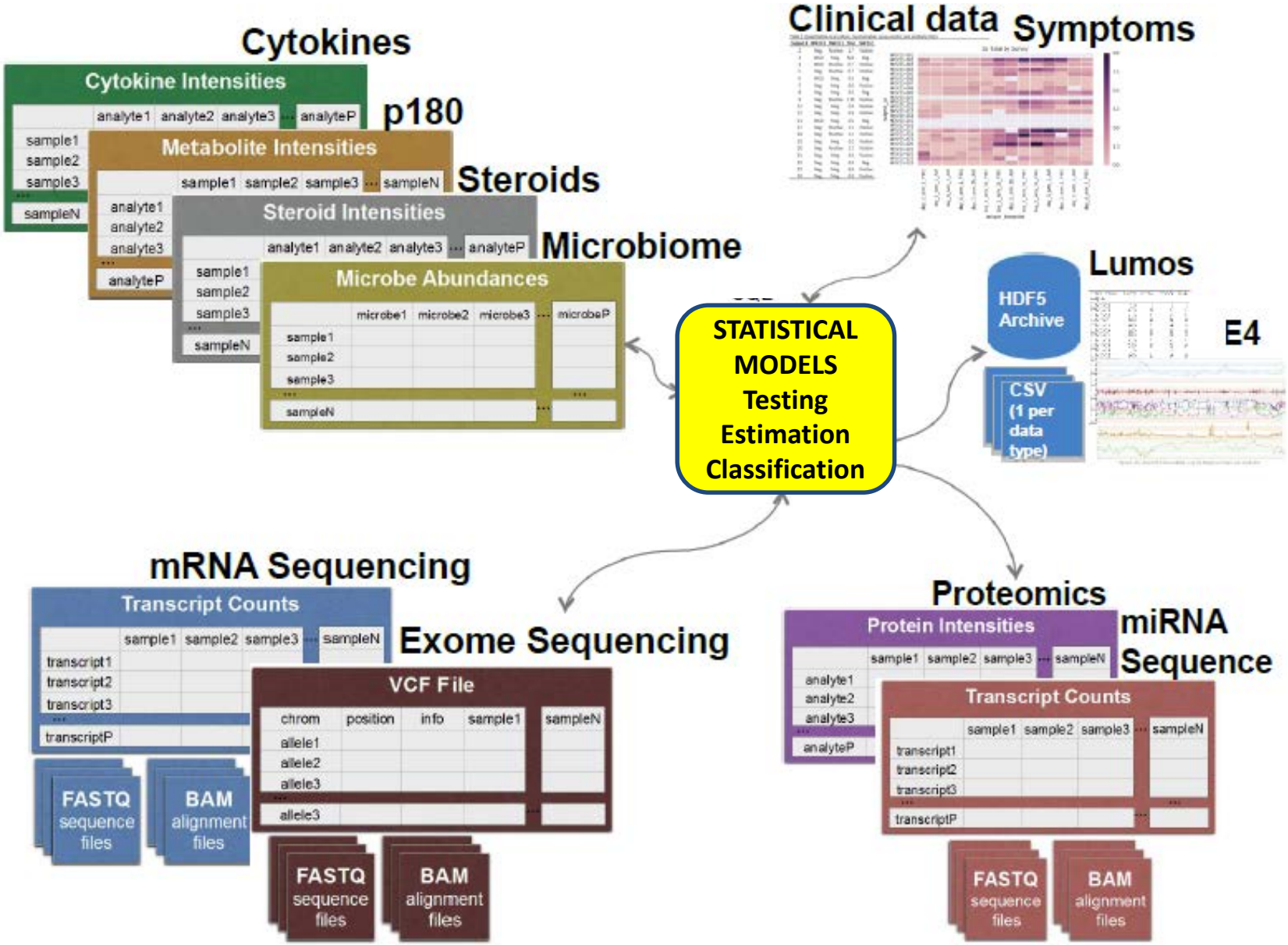
2015-2016 DARPA Biochronicity



Data Integration: Database Point of View



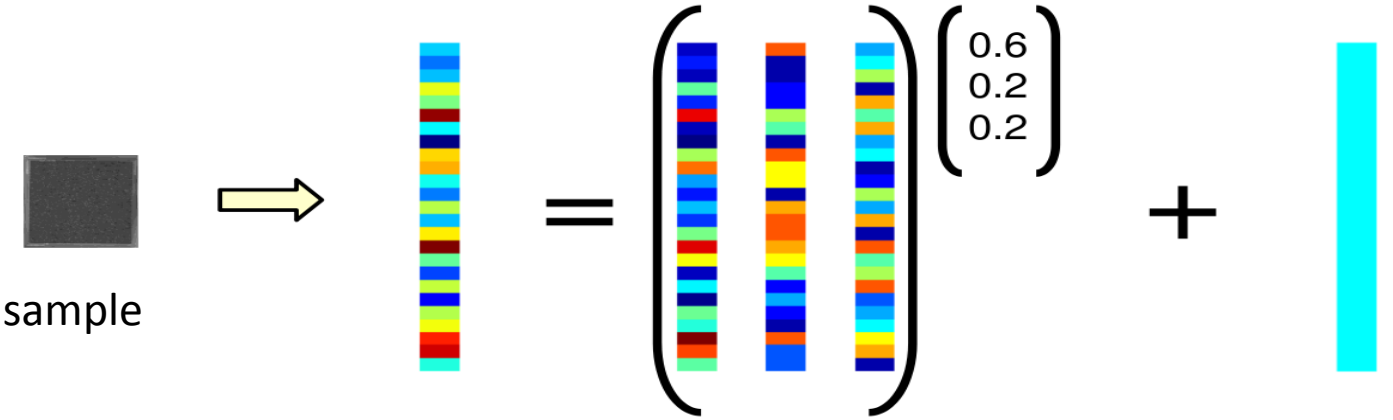
Data Integration: Inference Point of View



Biomarker Discovery: Bayesian Linear Unmixing (BLU)

Inference task: estimate small number of explanatory variables: M, A

Each column of matrix **X** vector follows linear mixing model $\boldsymbol{x} = \boldsymbol{Ma} + \boldsymbol{n}$



(3 factor model)

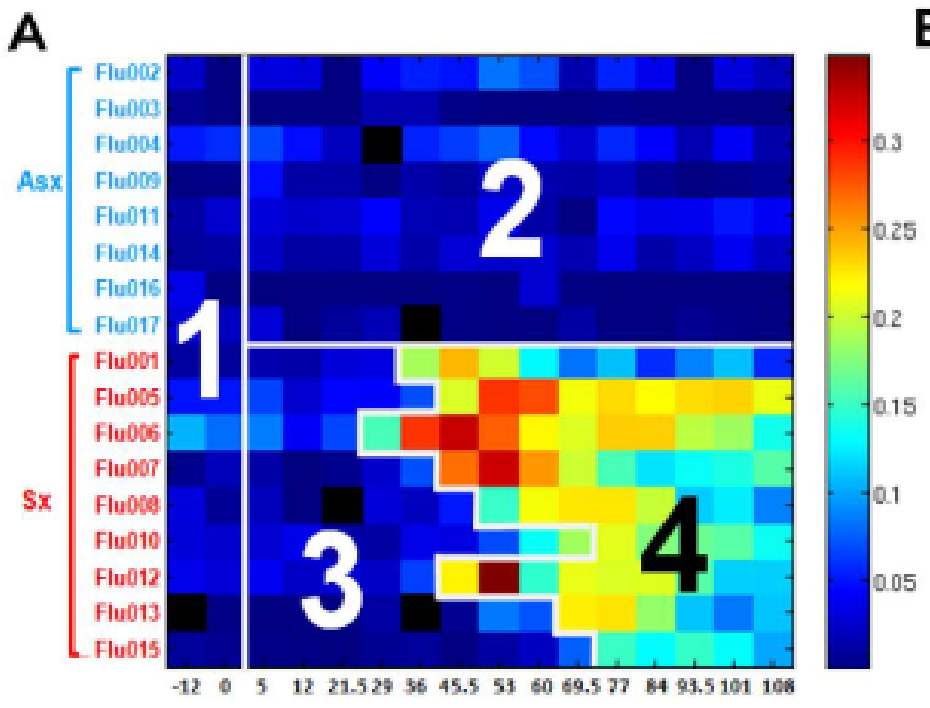
- Full matrix model: $[x_1,\ldots, x_n]=X=MA+N$
- BLU is a factor analysis method using specially adapted priors on **M,A**
  - Non-negative factor loadings (cols ***M***)
  - Factor scores (cols ***A***) are proportions (sum-to-one)
  - Unsupervised BLU (uBLU) uses reversible jump model to estimate #factors=rank(**M**)
- Full posterior distribution f(**A,M**|**X**) obtained by Gibbs sampling.

Huang et al, **Temporal Dynamics of Host Molecular Responses Differentiate Sx and Asy Influenza A Infection,** PLoS Genetics, 2011

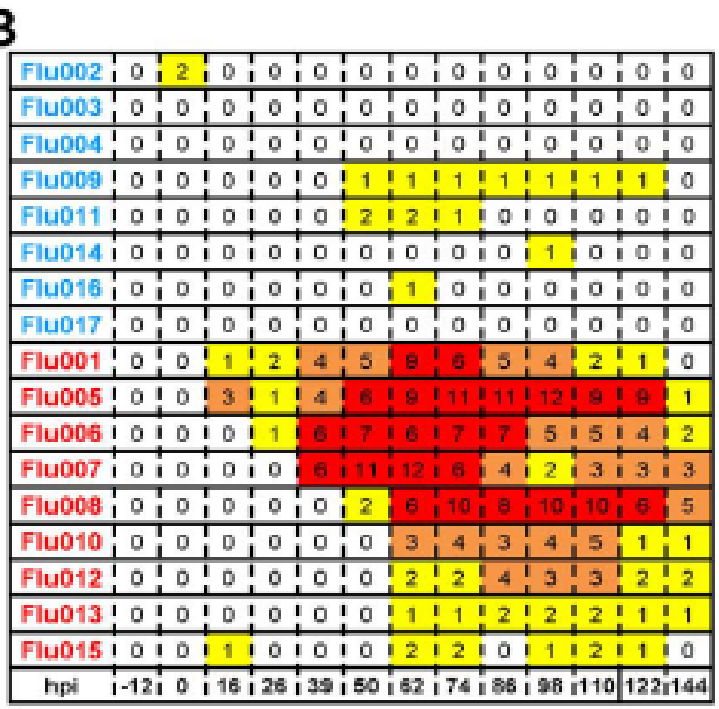
Bazot et al, **Unsupervised Bayesian linear unmixing of gene expression microarrays**, BMC Bioinformatics, 2013

mRNA uBLU vs Symptom Scores for DEE2

BLU/PNMF Factor 1 scores



Symptom scores

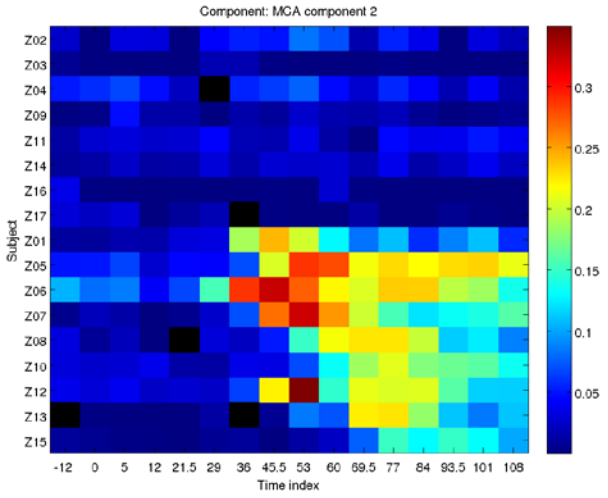


- Factor 1 scores: are in strong concordance with symptoms
- Factor 1 loadings: TRIM22, IFI144, IFIT1, IFIT3, LY6E, LAMP3, OAS1, OAS2,..
- Factor 1 variables constitute “ARVI signature” that has been validated

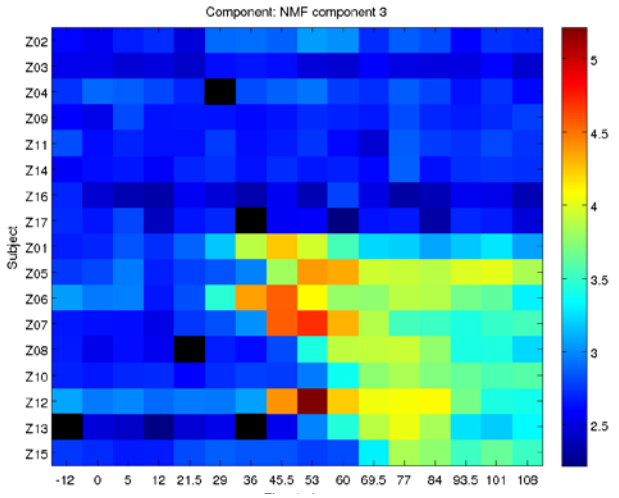
Huang et al, **Temporal Dynamics of Host Molecular Responses Differentiate Sx and Asy Influenza A Infection,** PLoS Genetics, 2011
Bazot et al, **Unsupervised Bayesian linear unmixing of gene expression microarrays**, BMC Bioinformatics, 2013

uBLU compared to other FA methods

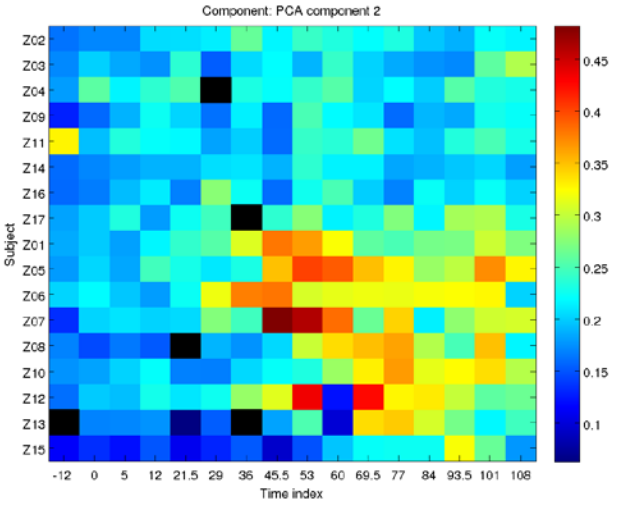
BLU



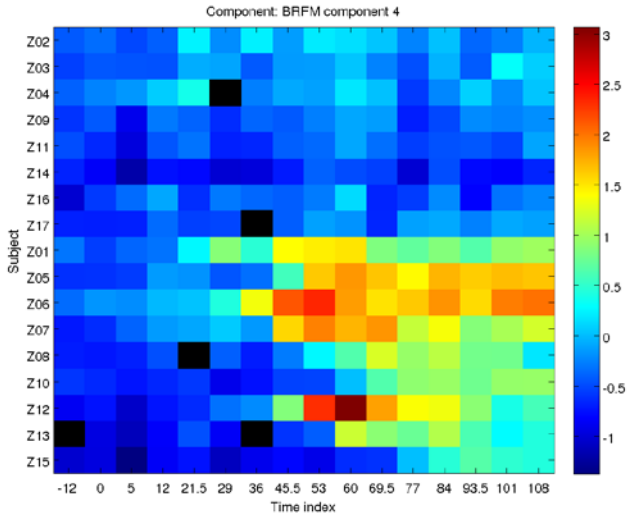
NMF



PCA



BRFM



Bazot et al, **Unsupervised Bayesian linear unmixing of gene expression microarrays**, BMC Bioinformatics, 2013

Validation with NCI Pathway Interaction Database

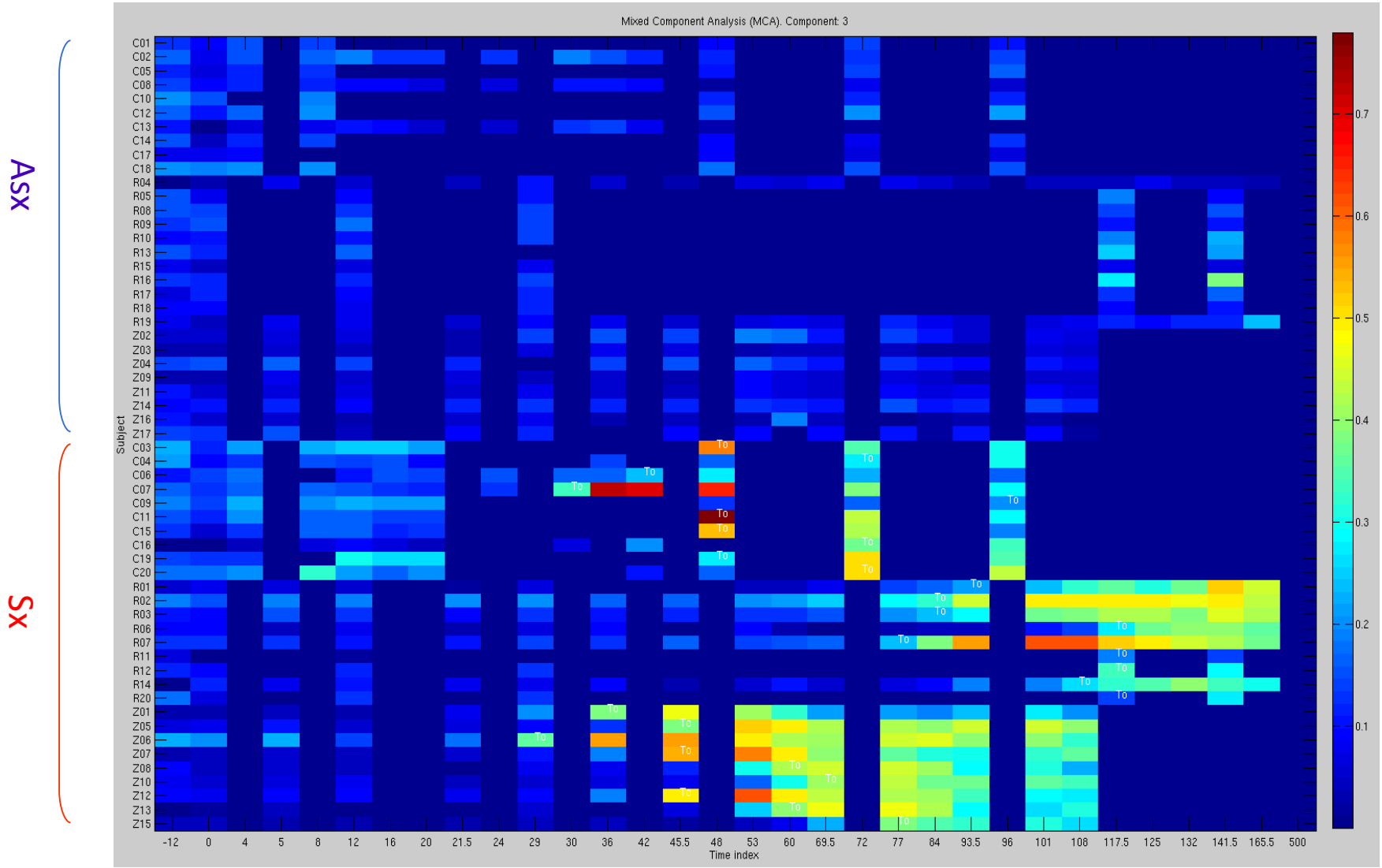
Table 6 NCI-curated pathway associations of group of genes contributing to uBLU inflammatory component

Pathway name	Genes	P-value
IFN-gamma pathway	CASP1, CEBPB, IL1B, IRF1, IRF9, PRKCD, SOCS1, STAT1, STAT3	1.34e-09
PDGFR-beta signaling pathway	DOCK4, EIF2AK2, FYN, HCK, LYN, PRKCD, SLA, SRC, STAT1, STAT3, STAT5A, STAT5B	3.26e-08
IL23-mediated signaling events	CCL2, CXCL1, CXCL9, IL1B, STAT1, STAT3, STAT5A	2.18e-07
Signaling events mediated by TCPTP	EIF2AK2, SRC, STAT1, STAT3, STAT5A, STAT5B, STAT6	6.38e-07
Signaling events mediated by PTP1B	FYN, HCK, LYN, SRC, STAT3, STAT5A, STAT5B	2.40e-06
GMCSF-mediated signaling events	CCL2, LYN, STAT1, STAT3, STAT5A, STAT5B	3.70e-06
IL12-mediated signaling events	HLA-A, IL1B, SOCS1, STAT1, STAT3, STAT5A, STAT6	1.32e-05
IL6-mediated signaling events	CEBPB, HCK, IRF1, PRKCD, STAT1, STAT3	1.80e-05

NB: P-value is 2 orders of magnitude better than the corresponding pathway association table for NMF

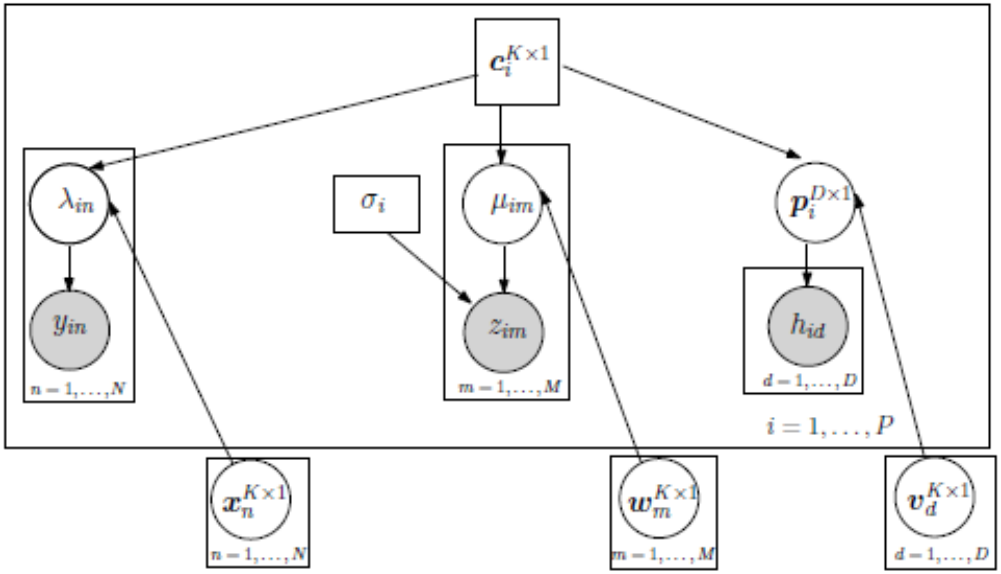
Bazot et al, **Unsupervised Bayesian linear unmixing of gene expression microarrays**, BMC Bioinformatics, 2013

Multi-study HRV/RSV/H3N2 Application of uBLU



Factor 1 scores

BLU can easily be extended to diverse datatypes



$$y_{in}|\mathbf{x}_n \sim \text{Pois}(e^{\mathbf{c}_i^T \mathbf{x}_n}), \qquad z_{im}|\mathbf{w}_m \sim \mathcal{N}(\mathbf{c}_i^T \mathbf{w}_m, \sigma_i^2),$$

$$\mathbf{x}_n \sim \mathcal{N}(\zeta, \mathbf{R}) \qquad \mathbf{w}_m \sim \mathcal{N}(\alpha, \mathbf{S})$$

$$h_{id}|\{\mathbf{v}_d\} \sim \text{Mult}\left(L_i; \frac{e^{\mathbf{c}_i^T \mathbf{v}_1}}{\sum_{d=1}^D e^{\mathbf{c}_i^T \mathbf{v}_d}}, \ldots, \frac{e^{\mathbf{c}_i^T \mathbf{v}_D}}{\sum_{d=1}^D e^{\mathbf{c}_i^T \mathbf{v}_d}}\right),$$

$$\mathbf{v}_d \sim \mathcal{N}(\beta, \mathbf{Q}),$$

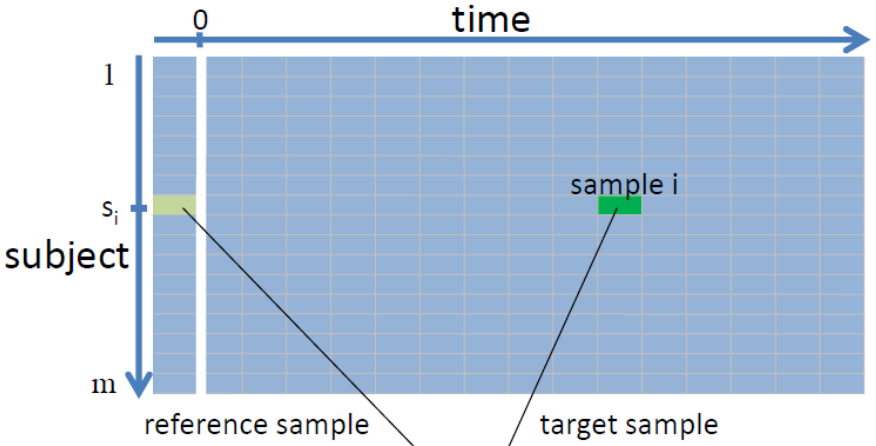
$$\theta = \{\mathbf{c}_i, \zeta, \mathbf{R}, \sigma_i^2, \alpha, \mathbf{S}, \beta, \mathbf{Q}\} : \text{ parameters}$$

Y. Yilmaz and A. Hero, **Multimodal Factor Analysis**, *IEEE Workshop on Machine Learning for Signal Processing*, 2015
Y. Yilmaz and A. Hero, **Multimodal Event Detection in Twitter Hashtag Networks**, to appear *J. Sig. Proc Systems*, 2016.

Paired Reference Predictor

Inference task: state classification

1. Each target sample paired w/ ref.

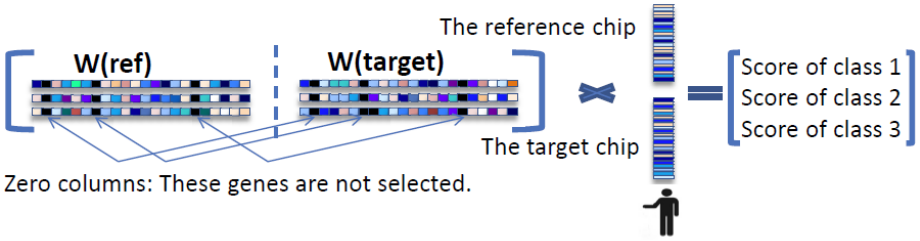


$$\mathbf{x}_i = [\ \ ]$$

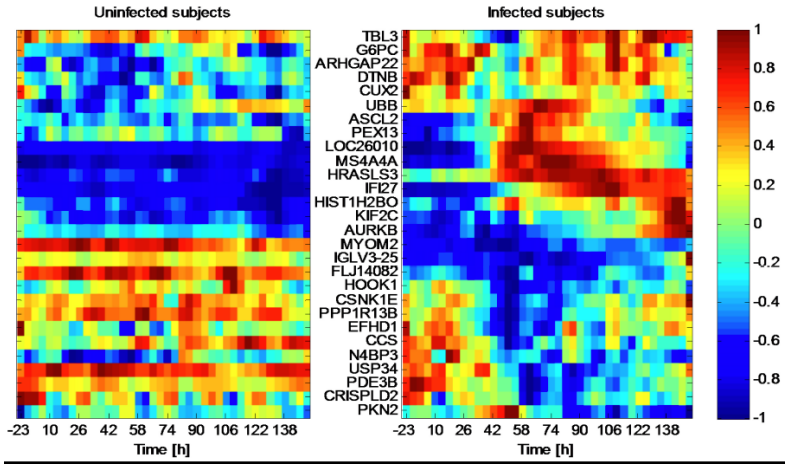
$$y_i \in \{1, 2, ..., K\}$$

$$s_i \in \{1, 2, ..., m\}$$

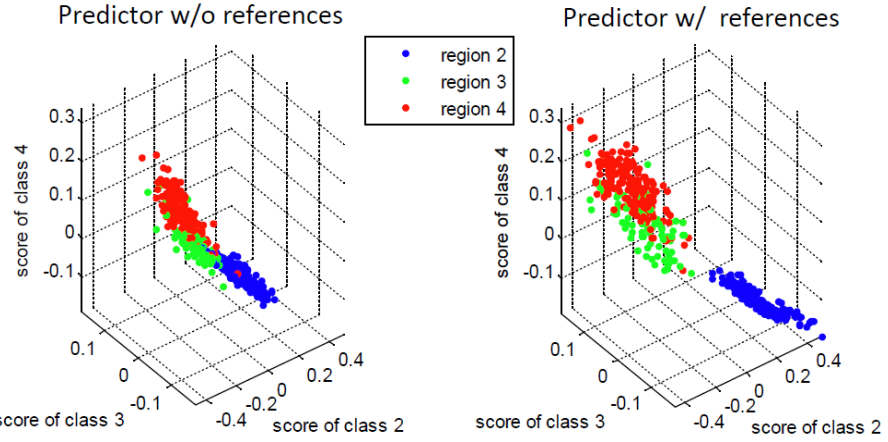
2. Structured-sparse learning model



3. Personalized predictor variables



4. Leads to improved state classification



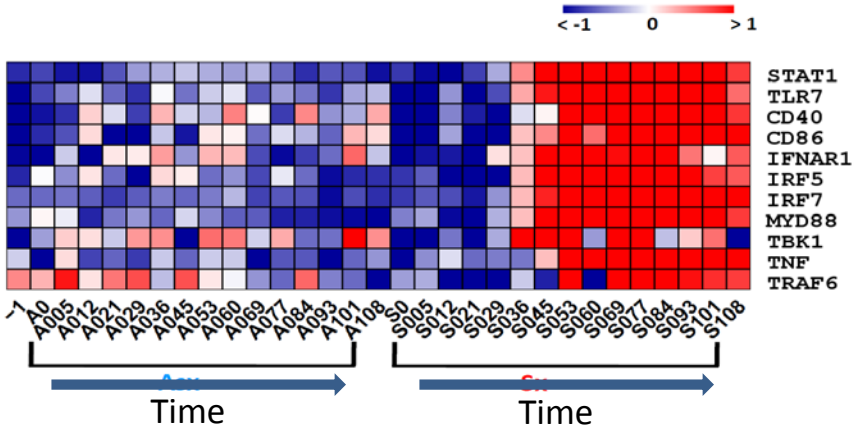
Liu et al, **An individualized predictor of health and disease using paired reference and target samples,** BMC Bioinformatics, 2016

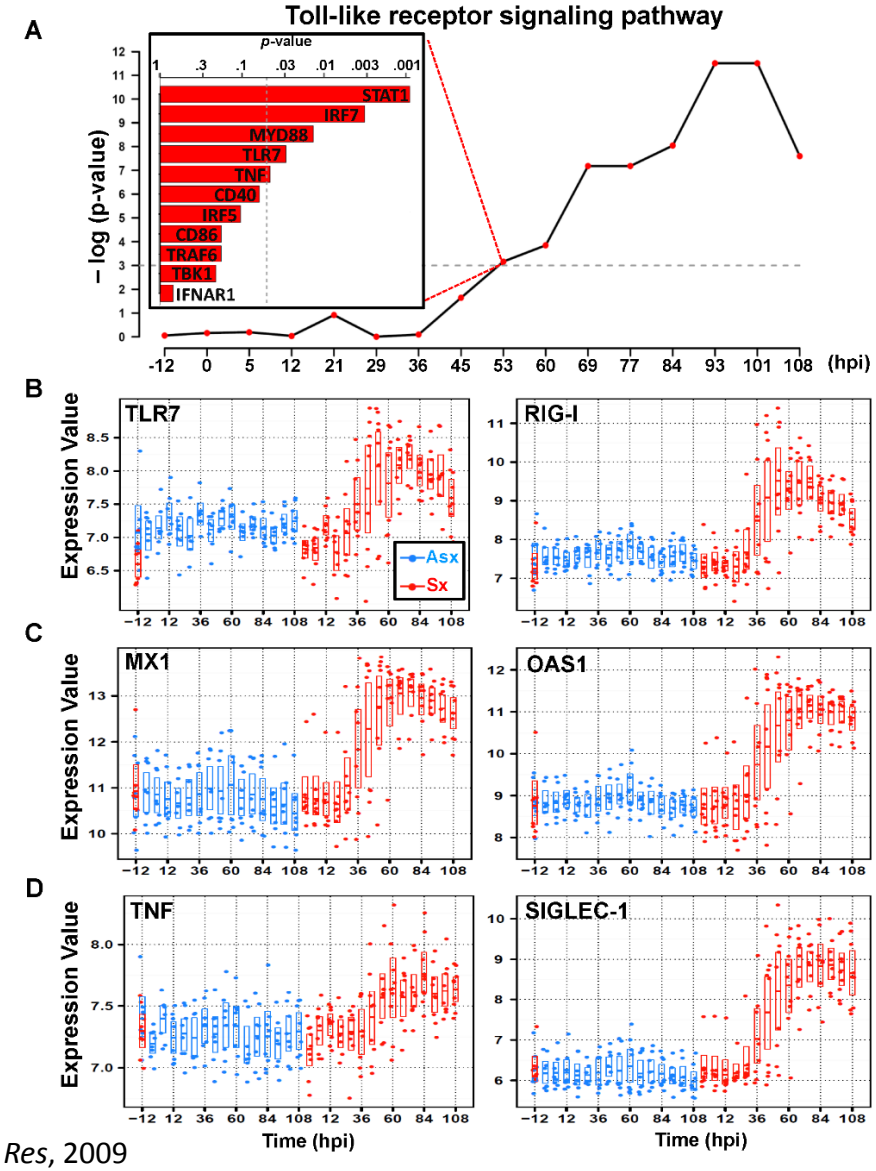
Testing Gene Pathways: Gene Set Enrichment Analysis

Inference task: test Sx/Asx differential expression of pathways at each time

GSEA integrates variables into known molecular pathways: sets of genes with similar function [Irrizary 2009]

GSEA p-value time profile refects statistical significance of Sx vs Asx differential expression at each time point.





Irrizary et al, **Gene set enrichment analysis made simple,** *Stat. Meth. Med Res*, 2009
Huang et al**, Temporal Dynamics of Host Molecular Responses Differentiate Sx and Asx Influenza A Infection**, PLoS Genetics, 2011

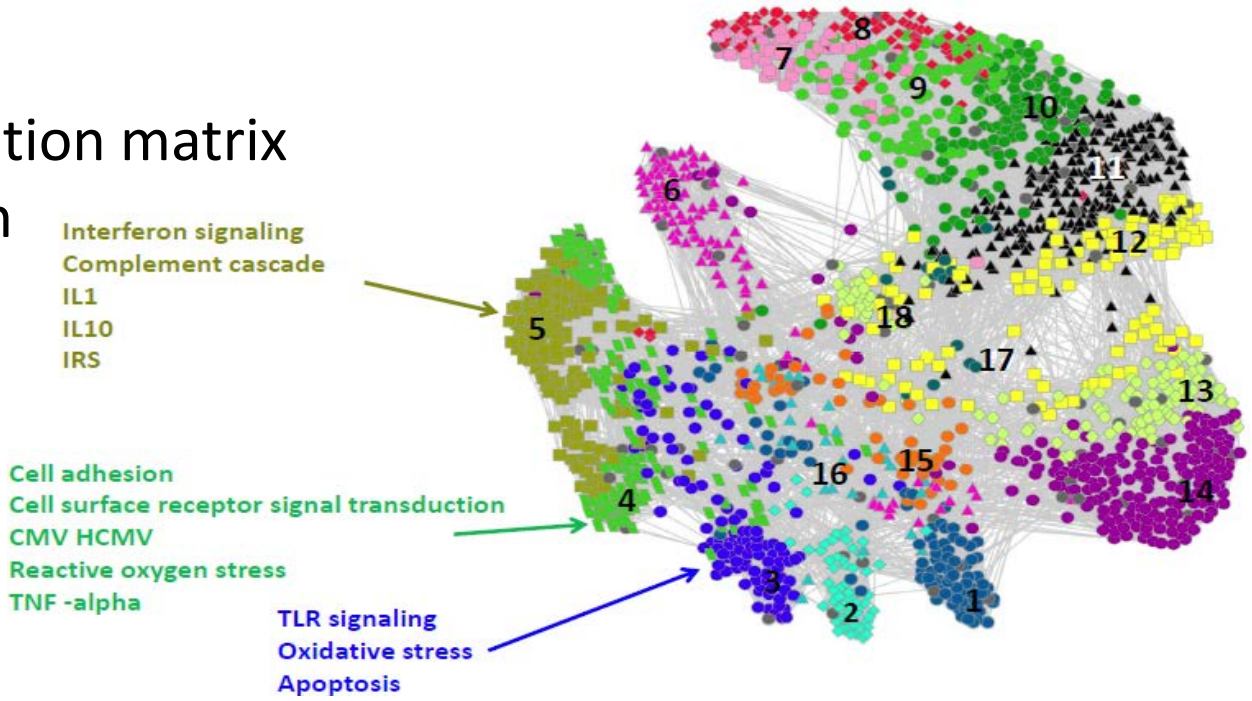
Integrated Network of Temporally Similar Pathways

Inference task: cluster pathways into groups w/ similar pv profiles

Correlation between GSEA p-value time-courses yields a pathway correlation matrix.

Thresholding of correlation matrix yields correlation graph

Spectral clustering of p-value profiles classifies pathways having similar patterns of differential expression



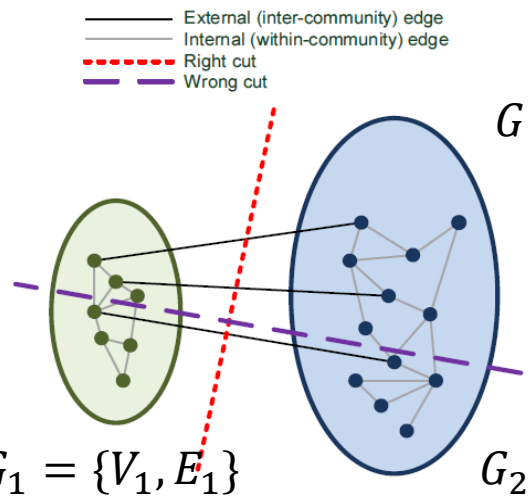
Huang et al, **Temporal Dynamics of Host Molecular Responses Differentiate Sx and Asx Influenza A Infection**, PLoS Genetics, 2011

Outline

1. Integration of diverse HD data
2. Illustrative example
3. Network inference
4. Concluding remarks

Goals of network inference

Discover communities/pathways



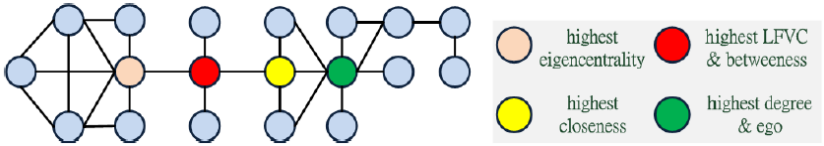
$G = \{V, E\}$

$G_1 = \{V_1, E_1\}$ $G_2 = \{V_2, E_2\}$

$$\mathbf{A} = \begin{array}{c} \\ n_1 \\ n_2 \end{array} \begin{array}{c} \begin{array}{cc} n_1 & n_2 \end{array} \\ \begin{pmatrix} \mathbf{A}_1 & \mathbf{C} \\ \mathbf{C}^T & \mathbf{A}_2 \end{pmatrix} \end{array}$$

G_1, G_2 obtained by partitioning adjacency matrix A (spectral clustering)

Discover nodes of high centrality…



Hub discovery table

Sub-graph	d	$N_{=d}$	$N_{\geq d}$
	1	29	39
	2	9	20
	3	0	1
	4	1	1
	5	0	0
…	…	…	…

$N_{\geq 1} = 2\, N_{edges}$

… while controlling class errors

$P(V_1 \,\Delta\, V_{1,true} \text{ not empty})$

…while controlling false discoveries

$P(N_{\geq d} > 0 | H_0)$

Multimodal network inference

Combine information from several observed networks

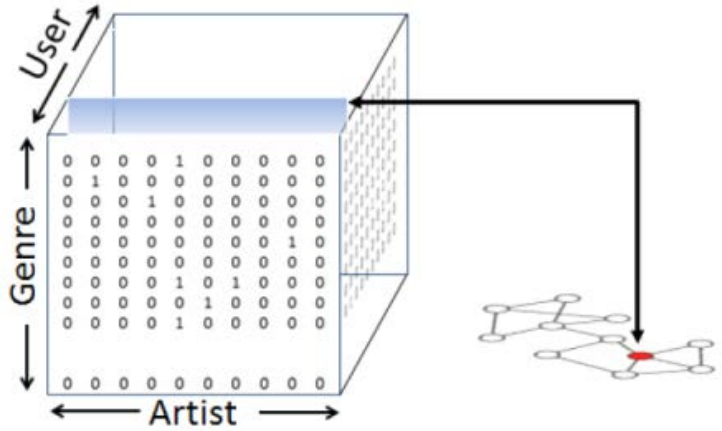
Networks may have different provenances

- Edges derived from correlations of node attributes, e.g., behavior
- Edges directly from node pair relations

Network integration methods

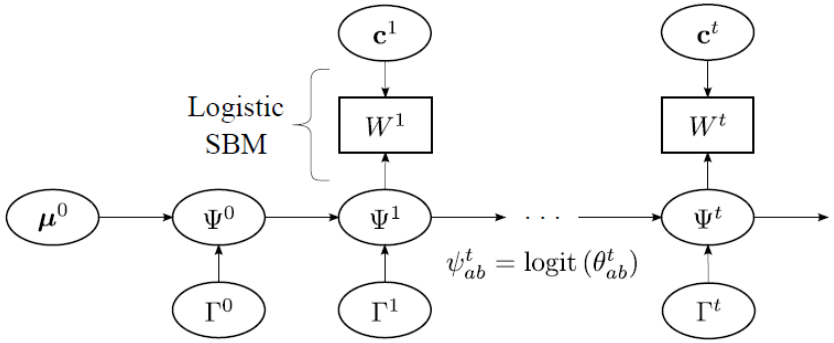
- Edge averaging [1]: $aA_1 + (1-a)A_2$
- Multicriteria optim [2,3]:Pareto fronts
- Centrality[4,5]: p-value aggregation
- Latent variable methods[6,7]: LSBM

Socio-collaborative retrieval [6]



User behavior matrix User friendship network

Dynamic SBM for time evolving Social Nets [7]



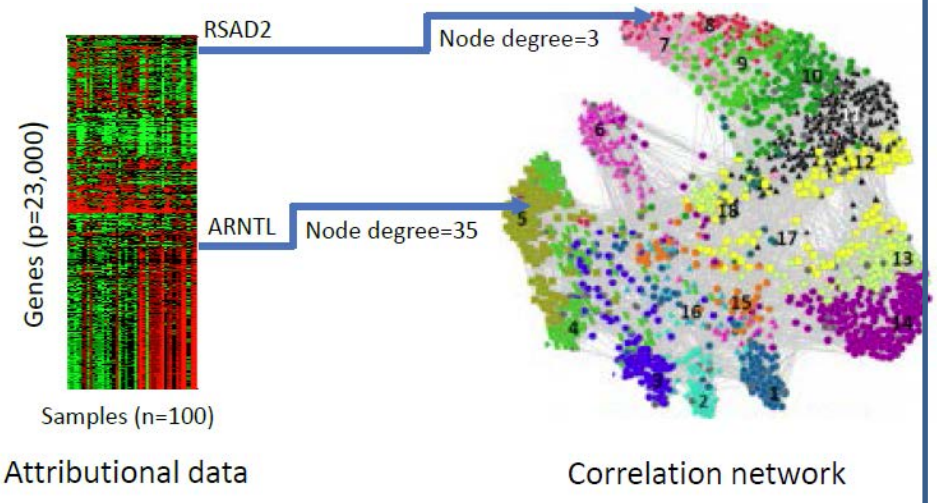
[1] Taylor, Shai, Stanley, Mucha, arXiv:1511.05271, May 2016
[2] Hsiao, Xu, Calder, Hero, NIPS 2012. arXiv:1110.3741
[3] Oselio, Kulesza, Hero, IEEE JSTSP 2014. arXiv:1309.5124
[4] Hero and Rajaratnam, JASA 2011. arXiv:1102.1204
[5] Chen, Wei, Newstadt, Simmons, H, IEEE ICIP, 2015. arxiv 1502.07432
[6] Hsiao, Kulesza, Hero, IEEE JSTSP, 2014. arxiv:1404.2342
[7] Xu, Hero, IEEE Journ Sel. Topic Sig Proc (JSTSP) 2014. arXiv:1403.0921

Often Network is High Dimensional

- Large number *p* of nodes: e.g.,
  - *p*=1.4 billion for Twitter
  - *p=12-30* thousand for genome
- Small number *n* of samples: e.g.,
  - *n=3* tweets per day per user
  - *n=17* human subjects assayed

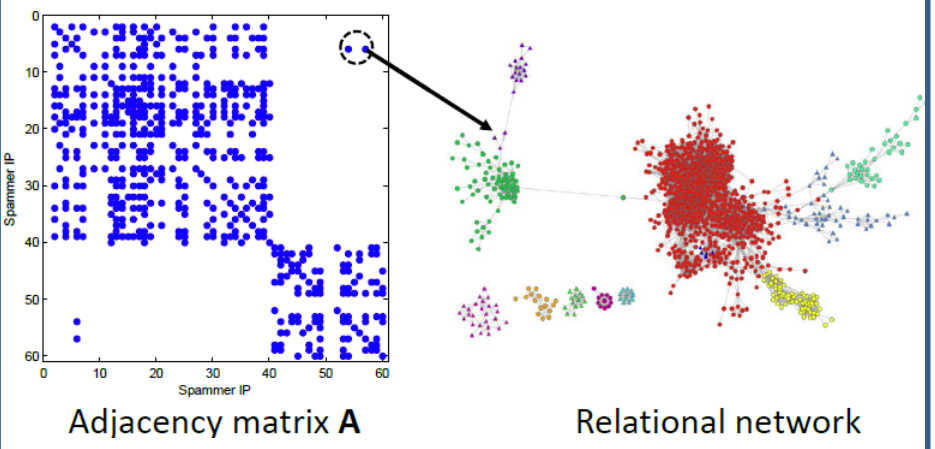
Q: conditions on *n,p* ensuring high quality of network inferences?

Attributional (Correlation) network



- How reliable are node degrees in the recovered network?

Relational network



- What is resiliency of clusters to spurious inter-cluster edges?

High Dimensional Error Analysis

- Standard asymptotic regimes:
  - Classical (CLT) regime: p fixed, n→ ∞. Not useful for large networks
  - Mixed high dimensional setting: p, n → ∞. Not useful for small samples
- New asymptotic regime (Hero and Rajaratnam 2011, 2012, 2016):
  - Purely high dimensional setting: n fixed, p→ ∞. Useful for large networks

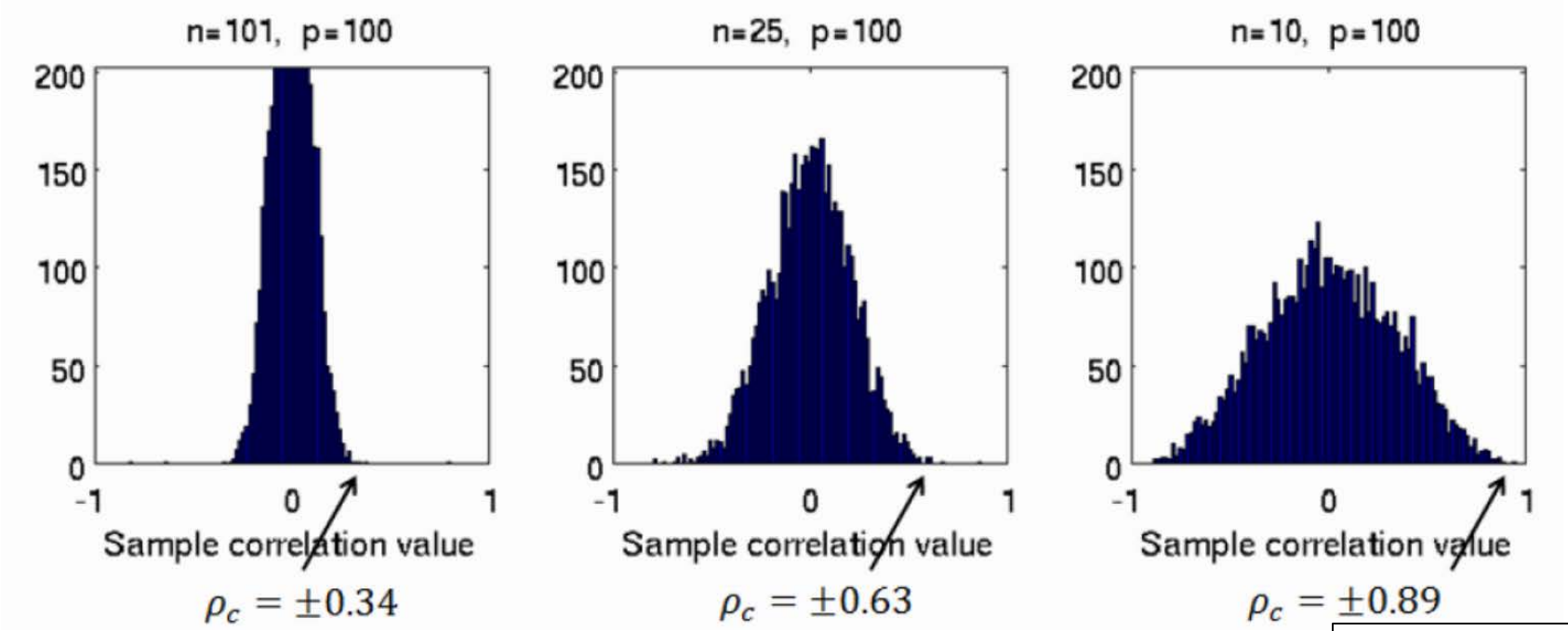
Asymptotic framework	Terminology	Sample size	Dimension	Application setting	References
		n	p		
Classical (or sample increasing)	small dimensional	$\longrightarrow \infty$	fixed	"small data"	Fisher [28, 29], Rao [68, 69], Neyman and Pearson [61], Wilks [84], Wald [79, 80, 81, 82], Cramér [16, 15], Le Cam [51, 52], Chernoff [13], Kiefer and Wolfowitz[46], Bahadur [3], Efron [22]
Mixed asymptotics	high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$	"medium sized" data (mega or giga scales)	Donoho [20], Zhao and Yu [87], Meinshausen and Bühlmann [58], Candès and Tao [10], Bickel, Ritov, and Tsybakov[6], Peng, Wang, Zhou, and Zhu [64], Wainwright [77, 78], Khare, Oh, and Rajaratnam, [44]
	very high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$		
	ultra high dimensional	$\longrightarrow \infty$	$\longrightarrow \infty$		
Purely high dimensional	purely high dimensional	fixed	$\longrightarrow \infty$	"Big Data" (tera, peta and exascales)	Hero and Rajaratnam [35] Hero and Rajaratnam [36] Firouzi, Hero and Rajaratnam [25]

Hero and Rajaratnam, **Large scale correlation mining**, *Journ. Am. Stat. Assoc., 2011*
Hero and Rajaratnam, **Hub discovery in partial correlation graphs**, *IEEE Trans Iinformation Theory,* 2012
Hero and Rajaratnam, **Foundational principles for large-scale inference**, *IEEE Proceedings,* 2016

Phase transitions for correlation networks

- When n << p there can be many false edges (FE) in the graph
- There is a critical phase transition below which FE's dominate

Multivariate Gaussian sample with diagonal 100 x 100 covariance matrix



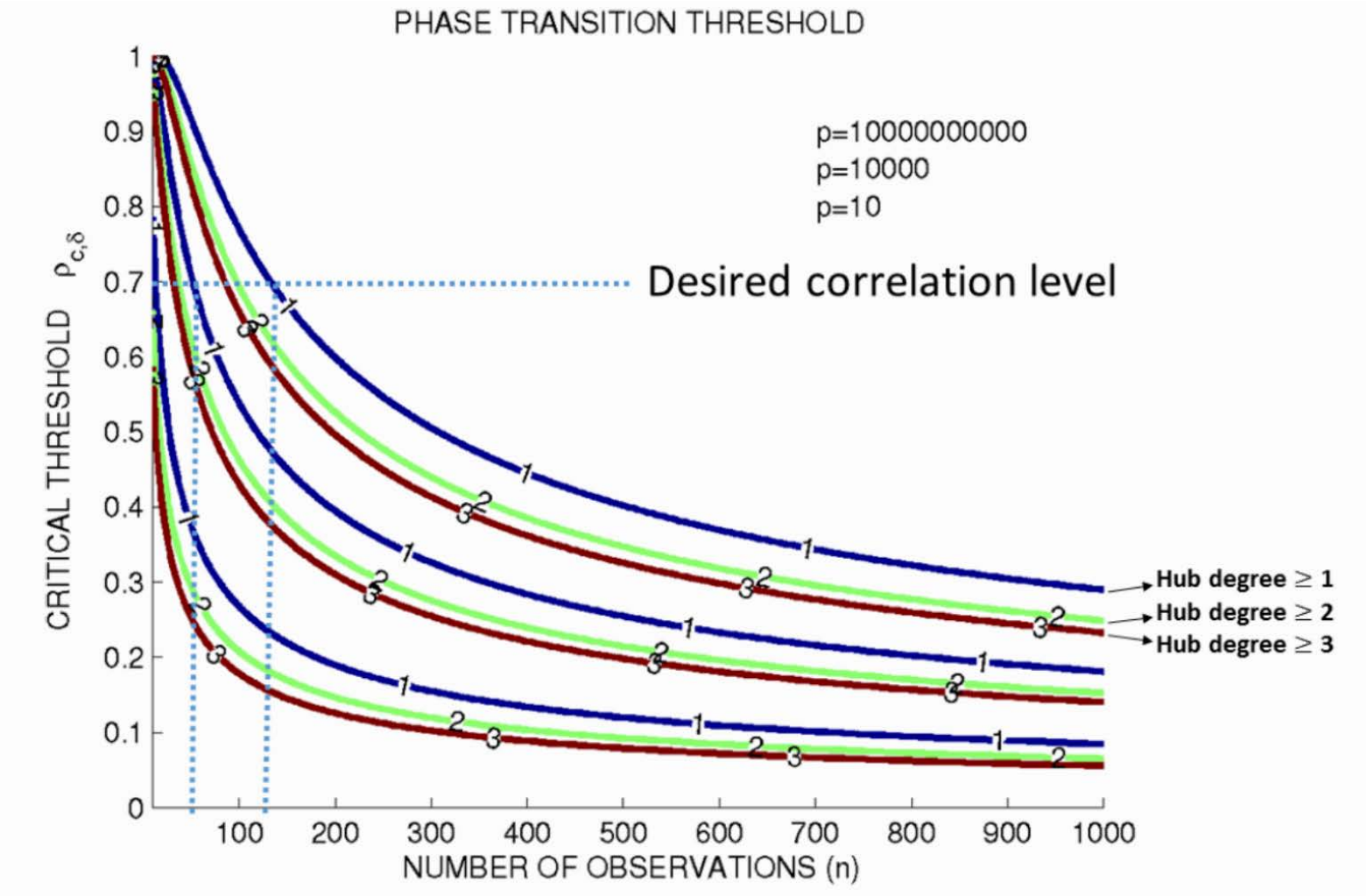
$$\rho_c = \sqrt{1 - c_n (p-1)^{-2/(n-4)}}$$

Follows from purely high D analysis

- There is a similar critical phase transition threshold for false hubs

Hero and Rajaratnam, **Large scale correlation mining**, *Journ. Am. Stat. Assoc., 2011*

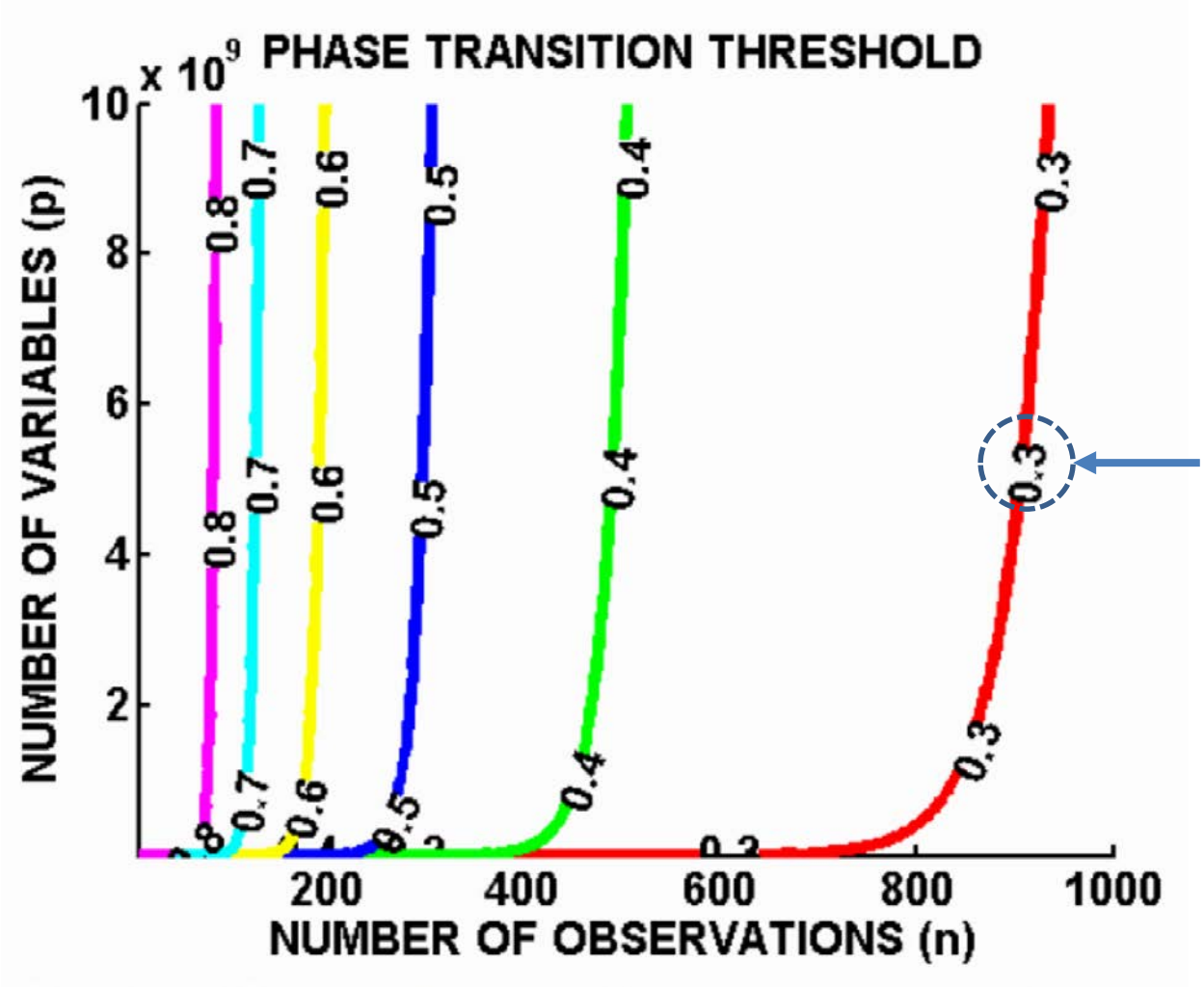
Critical threshold scaling law: n=O(log(p)): the "blessing of high dimensionality"



Hero and Rajaratnam, **Hub discovery in partial correlation graphs**, *IEEE Trans IT* 2012

Hero and Rajaratnam, **Foundational principles for large-scale inference**, *IEEE Proceedings* 2016

The "blessing of high dimensionality"



Hero and Rajaratnam, **Hub discovery in partial correlation graphs**, *IEEE Trans IT* 2012
Hero and Rajaratnam, **Foundational principles for large-scale inference**, *IEEE Proceedings* 2016

From phase transitions to false discovery probability

Asymptotics of hub screening[1]: (H and Rajaratnam 2012): Assume that rows of $\mathbb{X}$ are i.i.d. with bounded elliptically contoured density and row sparse covariance $\mathbf{\Sigma}$.

Theorem

Let p and $\rho = \rho_p$ satisfy $\lim_{p\to\infty} p^{1/\delta}(p-1)(1-\rho_p^2)^{(n-2)/2} = e_{n,\delta}$. Then

$$P(N_{\delta,\rho} > 0) \to \begin{cases} 1 - \exp(-\lambda_{\delta,\rho,n}/2), & \delta = 1 \\ 1 - \exp(-\lambda_{\delta,\rho,n}), & \delta > 1 \end{cases}.$$

$$\lambda_{\delta,\rho,n} = p\binom{p-1}{\delta}(P_0(\rho,n))^{\delta}\ J(\mathbf{\Sigma})$$

$$P_0(\rho,n) = 2B((n-2)/2,1/2)\int_{\rho}^{1}(1-u^2)^{\frac{n-4}{2}}\,du$$

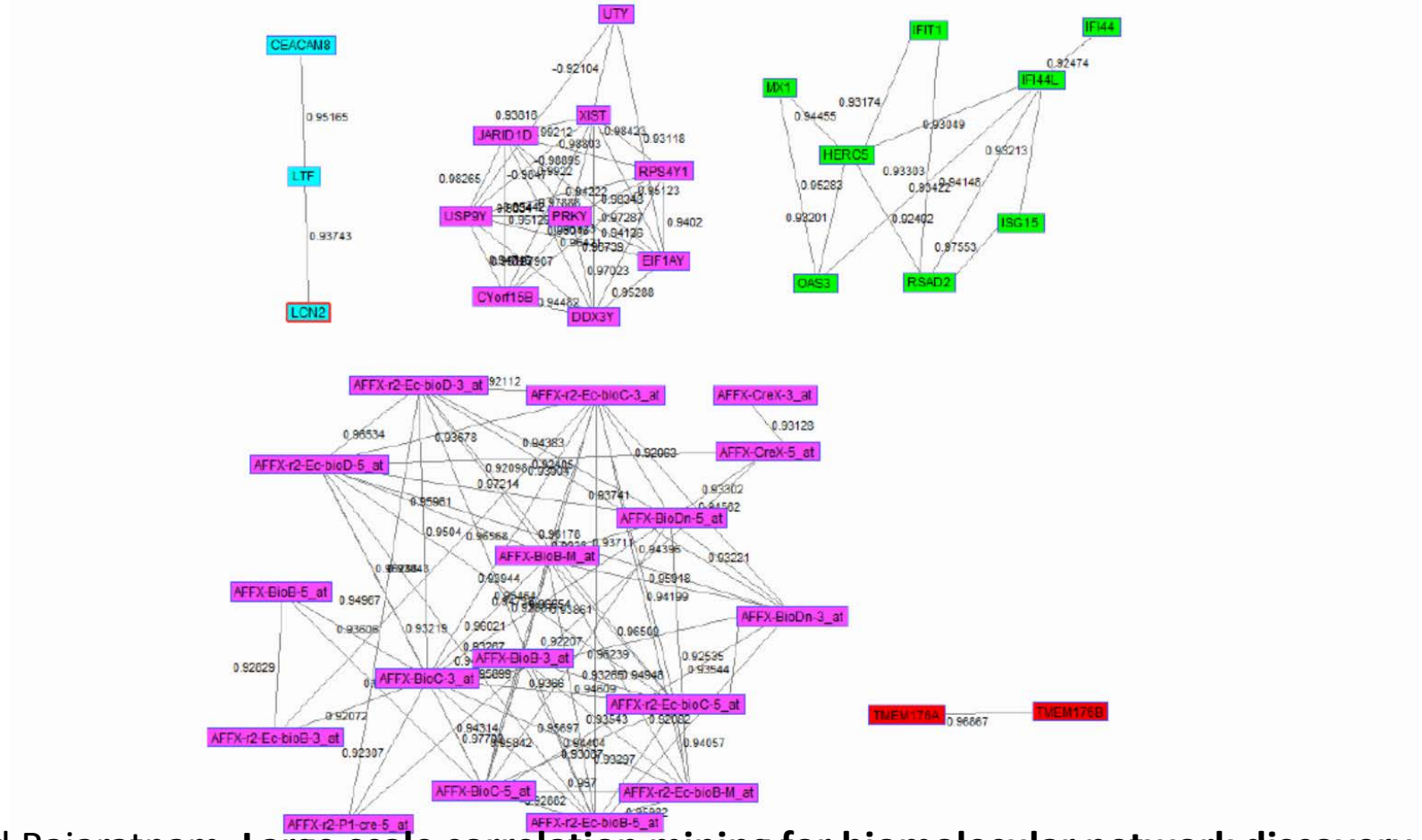
Hero and Rajaratnam, **Hub discovery in partial correlation graphs**, *IEEE Trans IT* 2012

Application to H3N2 DEE2 Study

Hero and Rajaratnam, **Large scale correlation mining for biomolecular network discovery**, in *Big Data Over Networks, 2015*

Sentinel Pathway Analysis: Pre-inoc Samples

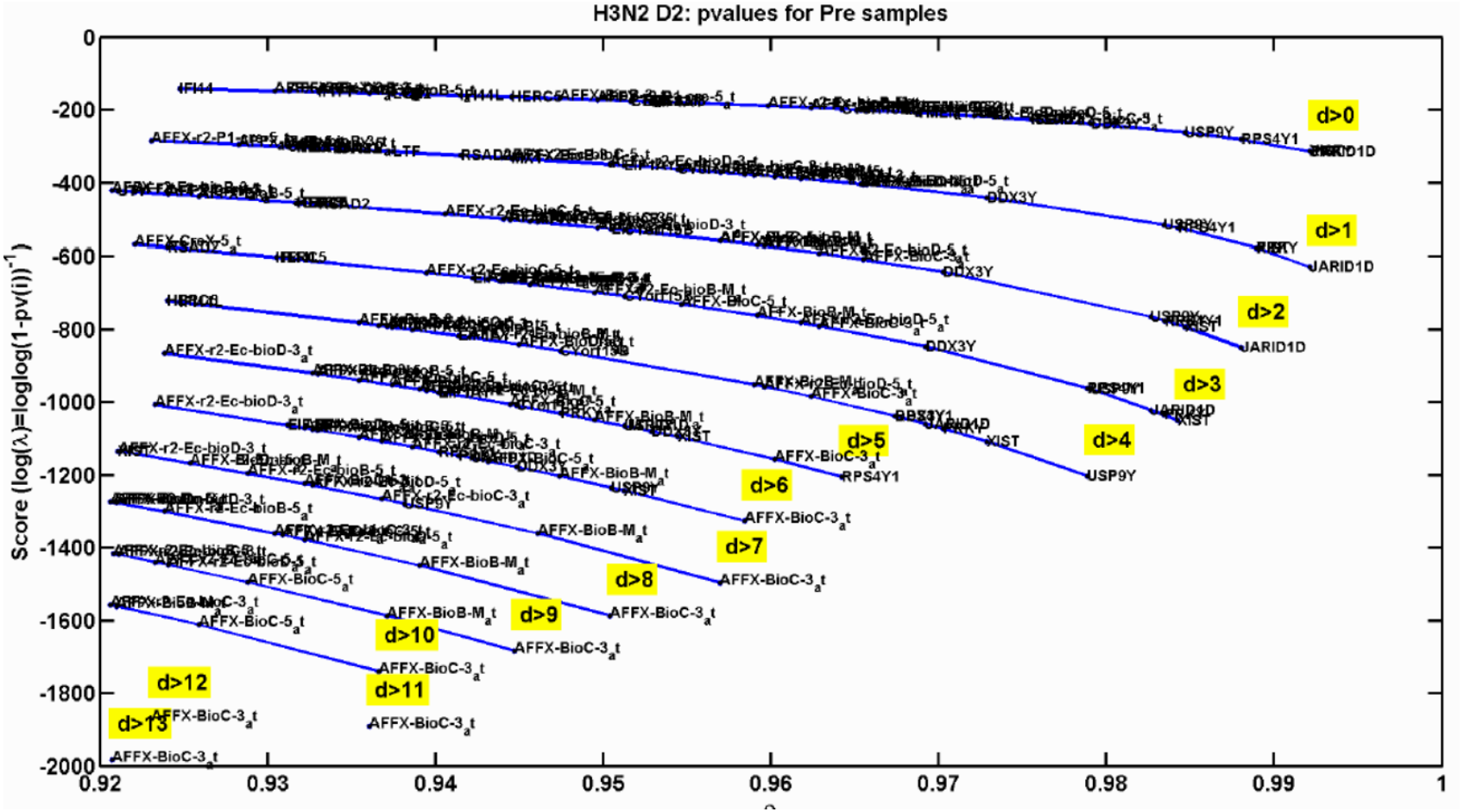
- Screen correlation at FWER 10^{-6}: 1658 genes, 8718 edges
- Screen partial correlation at FWER 10^{-6}: 39 genes, 111 edges



Hero and Rajaratnam, **Large scale correlation mining for biomolecular network discovery**, in *Big Data Over Networks, 2015*

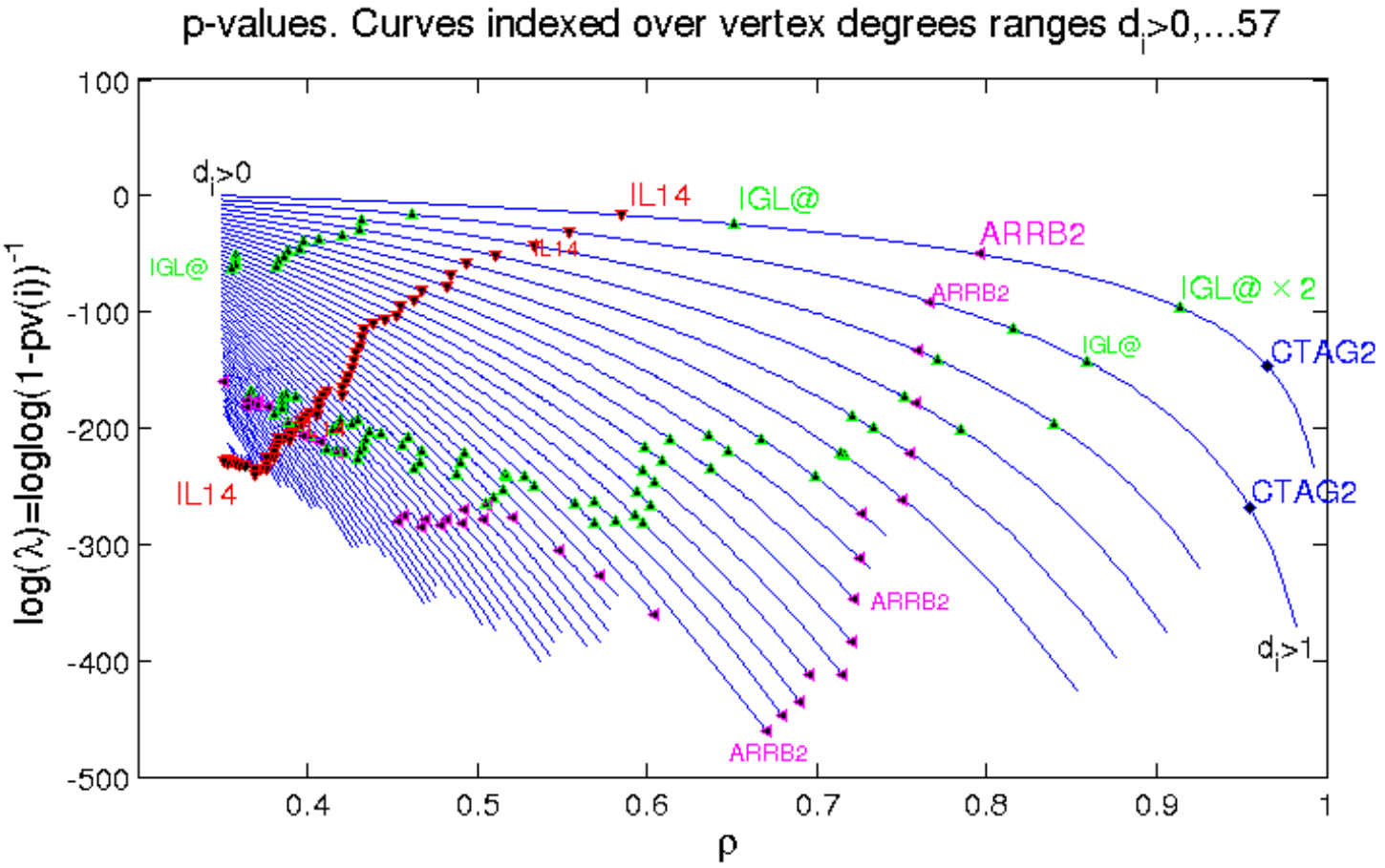
P-value Waterfall Plot: Pre-inoc Samples

- A more explicit way to visualize the hubs in previous graph
- This plot summarizes statistical significance and minimal correlation of all 37 hubs of varying degrees $d \in \{1, \ldots, 13\}$



P-value Waterfall Plot: Post-inoc Sx Samples

- This plot shows all of the 12,000+ nodes (Affymetrix genes) - statistical significance and partial correlations for Sx samples.



Hero and Rajaratnam, **Hub discovery in partial correlation graphs**, *IEEE Trans IT* 2012

Extensions

- Theory provides universal phase transitions and scaling laws for significance testing of hubs in sparse correlation networks.
- Applies to mixed-type data under elliptical multivariate joint dsn

- Worthwhile extensions
  - Heterogeneous populations having multi-modal distribution
  - Missing entries of the correlation matrix
  - Significance testing of more general network motifs than hubs (stars)
  - Higher order correlation, e.g., mutual information

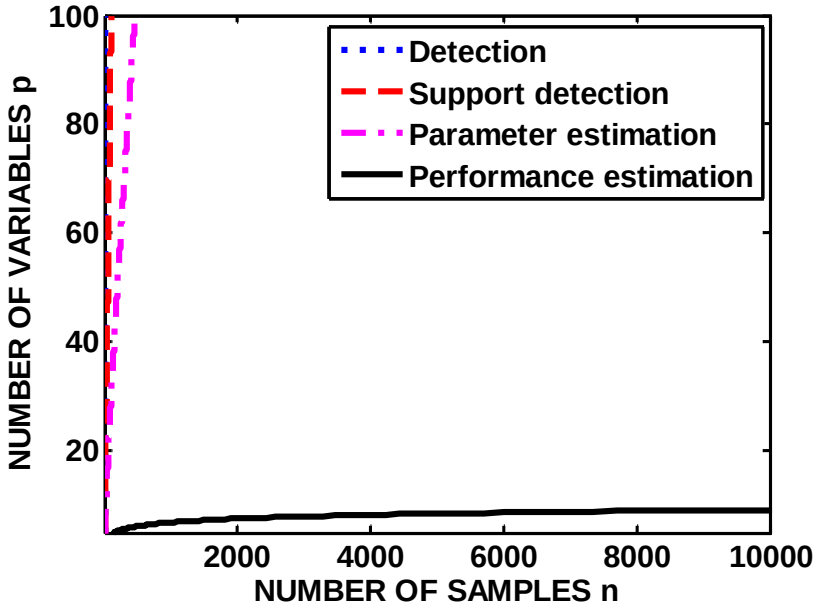
- Note: purely high dimensional regime of small *n* and large *p* may not be useful for other inference tasks [Hero and Rajaratnam 2016]

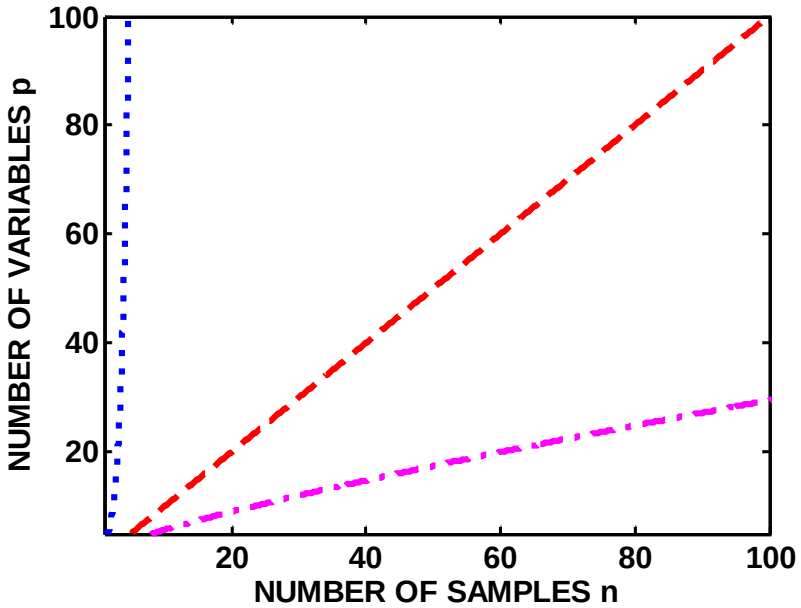
Hero and Rajaratnam, **Foundational principles for large-scale inference**, *IEEE Proceedings* 2016

Scaling laws across inference tasks

- Correlation graph inference tasks can be rank ordered in sample complexity from logarithmically easy to exponentially hard in p

Detection	Support recovery	Param. estimation	Perform. estimation
$P(N_e > 0)$	$P(\lvert\{\mathcal{S}\Delta\hat{\mathcal{S}}\}\rvert = \phi)$	$E[\lVert\Omega - \hat{\Omega}\rVert_F^2]$	$\int E[(f_\Omega(\mathbf{x}) - \hat{f}(\mathbf{x}))^2]d\mathbf{x}$
$pe^{-n\beta}$	$2^{p^D}e^{-n\beta}$	$\frac{p\log p}{n}\beta$	$n^{-2/(1+p)}\beta$
$\frac{\log p}{n} \to \alpha$	$\frac{p^D}{n} \to \alpha$	$\frac{p\log p}{n} \to \alpha$	$\frac{p}{\log n} \to \alpha$
$\rho_C \to \rho^*$	$\rho_C \to 0$	$\rho_C \to 0$	$\rho_C \to 0$

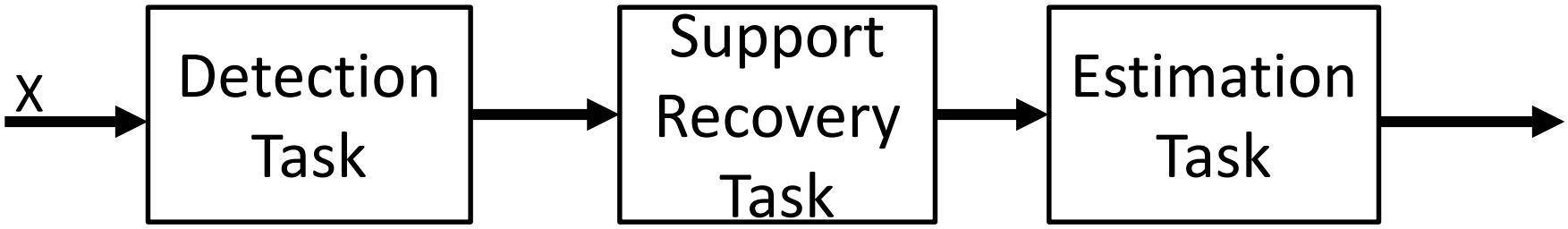


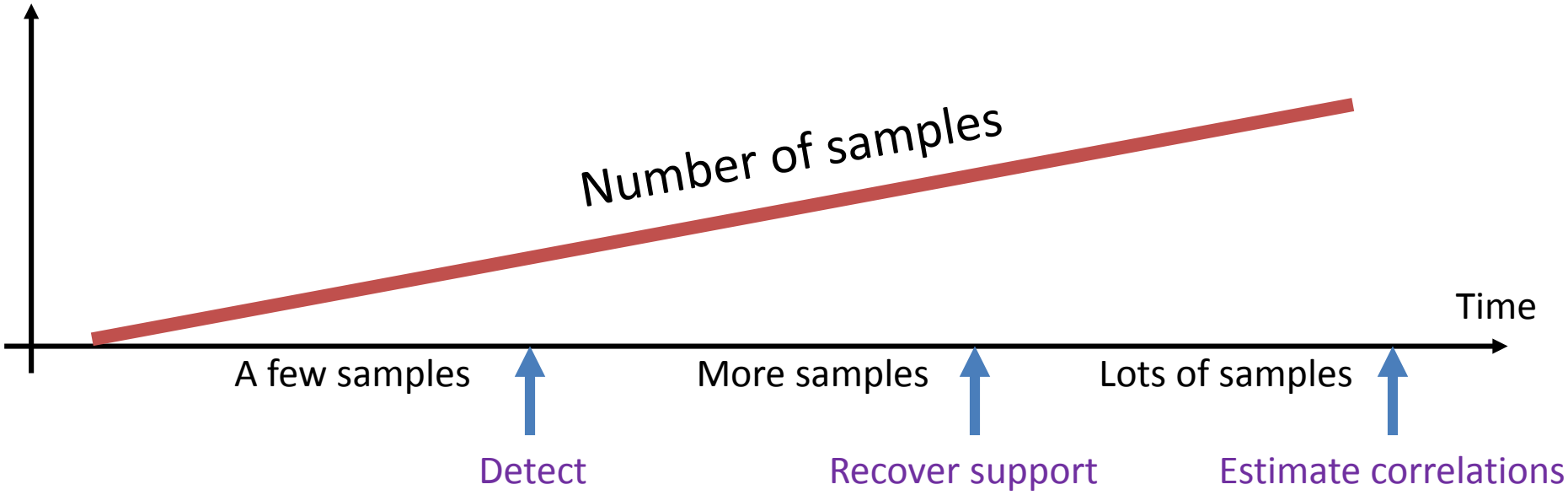


Hero and Rajaratnam, **Foundational principles for large-scale inference**, *IEEE Proceedings* 2016

Implication: right size the task to sample size

Sequential data collection: adapt inference task to achievable accuracy





[2] Firouzi, Rajaratnam and Hero, **Two-stage Sampling, Prediction and Adaptive Regression via Correlation Screening (SPARCS),** arxiv 1502:06189.

Outline

1. Integration of diverse HD data
2. Illustrative example
3. Network inference
4. Concluding remarks

Concluding Remarks

- New principles for reliable multimodality data integration are needed
  - Methods should be closely tied to inference task
  - Study of phase transitions and error behavior is fundamentally important
  - High dimensional analysis can be practically useful

- Many problems are open, including:
  - Integrating highly dichotomous data
    - Sparse graphical models for heterogenous data [Zhou 2014]
    - Posterior Pareto analysis [Hero, Fleury, 2003], [Hsiao, Xu, Calder, H, NIPS 2012]
  - Integrating time varying dynamic network data
    - Spectral correlation networks [Firouzi, Wei, H 2013]
    - Kalman tracking nets and DBN's [Xu, Kliger, H 2014]
  - Phase transitions for aggregated data
    - Aggregated homogeneous SBM's [Dane et al, 2016]
  - Integration of directed and undirected graphical models [Wainright&Jordan 2008], [Rao, States, Engel, H 2007], [Gleich 2016]