Machine Learning for Healthcare

HST.956, 6.S897

Lecture 19: Disease progression modeling & subtyping, Part 2

David Sontag

Recap of goals of disease progression modeling

- Predictive:
  - *What will this patient's future trajectory look like?*
- Descriptive:
  - *Find markers of disease stage and progression, statistics of what to expect when*
  - *Discover new disease subtypes*
- Key challenges we will tackle:
  - Seldom directly observe disease stage, but rather only indirect observations (e.g. symptoms)
  - Data is censored – don't observe beginning to end

Outline of today's lecture

1. Staging from cross-sectional data
   - Wang, Sontag, Wang, *KDD* 2014
   - Pseudo-time methods from computational biology
2. Simultaneous staging & subtyping
   - Young et al., *Nature Communications* 2018

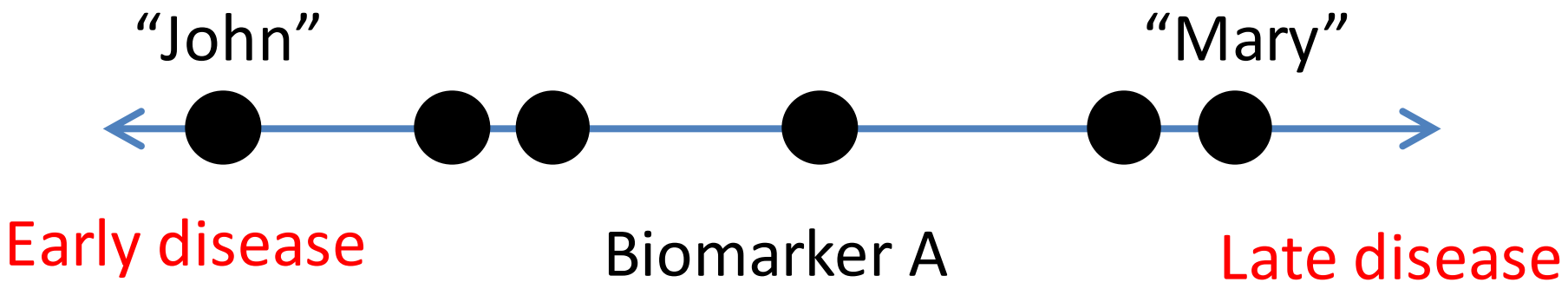
Outline of today's lecture

1. **Staging from cross-sectional data**
   - Wang, Sontag, Wang, *KDD* 2014
   - Pseudo-time methods from computational biology
2. Simultaneous staging & subtyping
   - Young et al., *Nature Communications* 2018

Stage vs. subtype

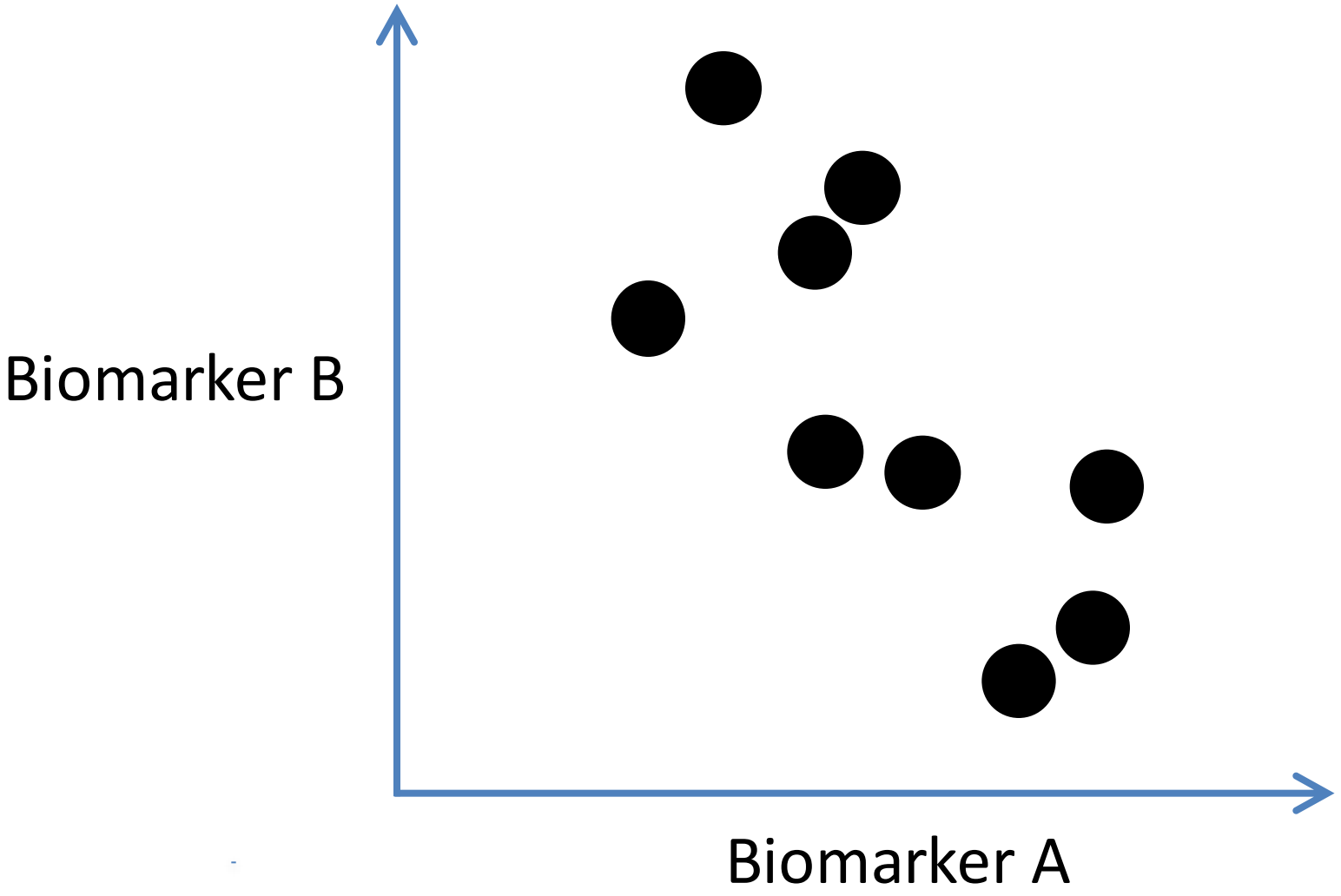
- Staging: sort patients into early-late disease or severity, i.e. discover the trajectory
- Cross-sectional data: only 1 time point observed per patient
  - More generally, censored to be a short window
- Naïve clustering can't differentiate between *stage* and *subtype*
  - Patients assumed to be aligned at baseline
- **Let's build some intuition around how staging from cross-sectional data might be possible…**

In 1-D, might assume that low values correspond to an early disease stage (or vice-versa)

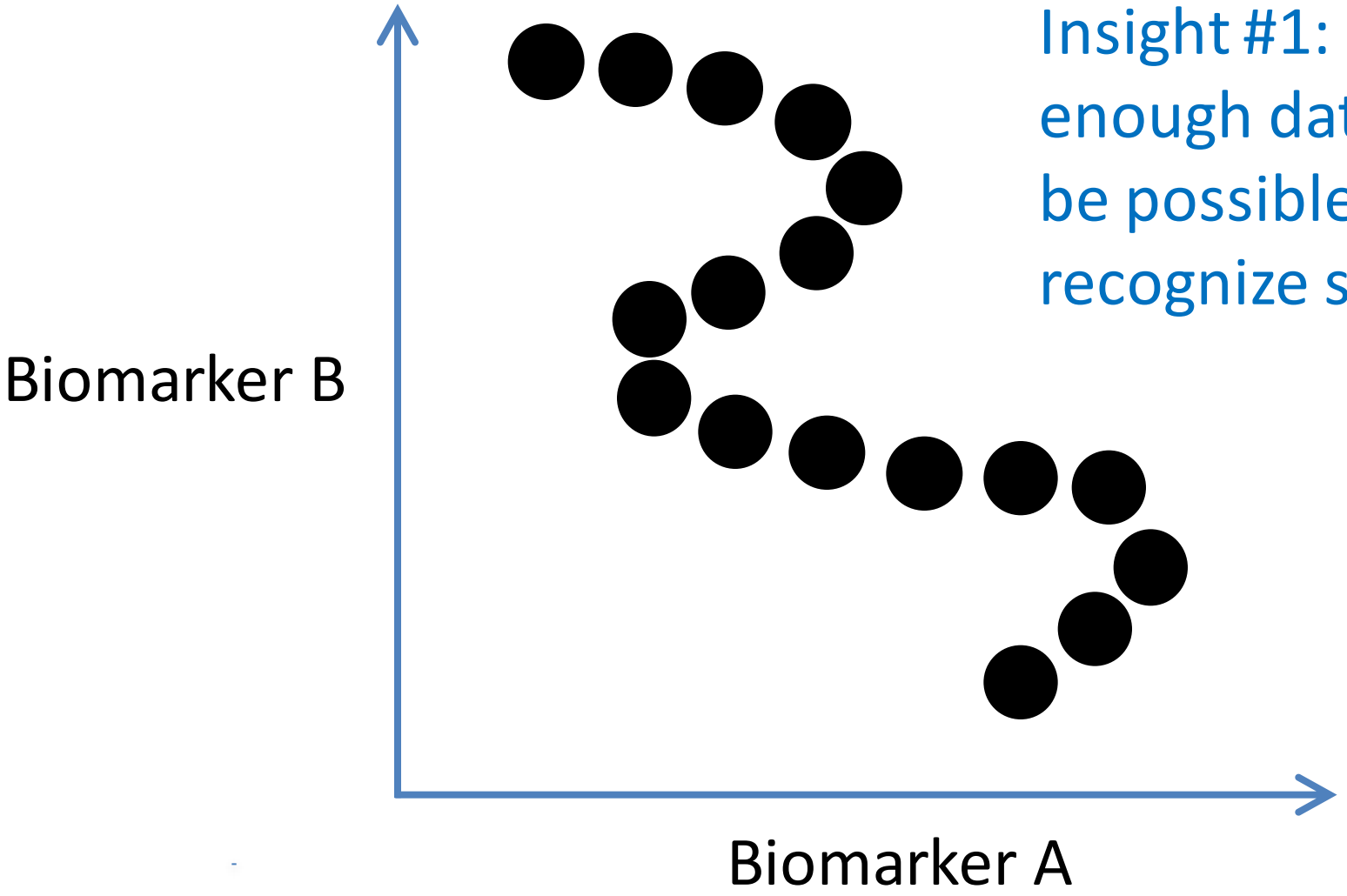


Assume samples were all taken today

What about in higher dimensions?



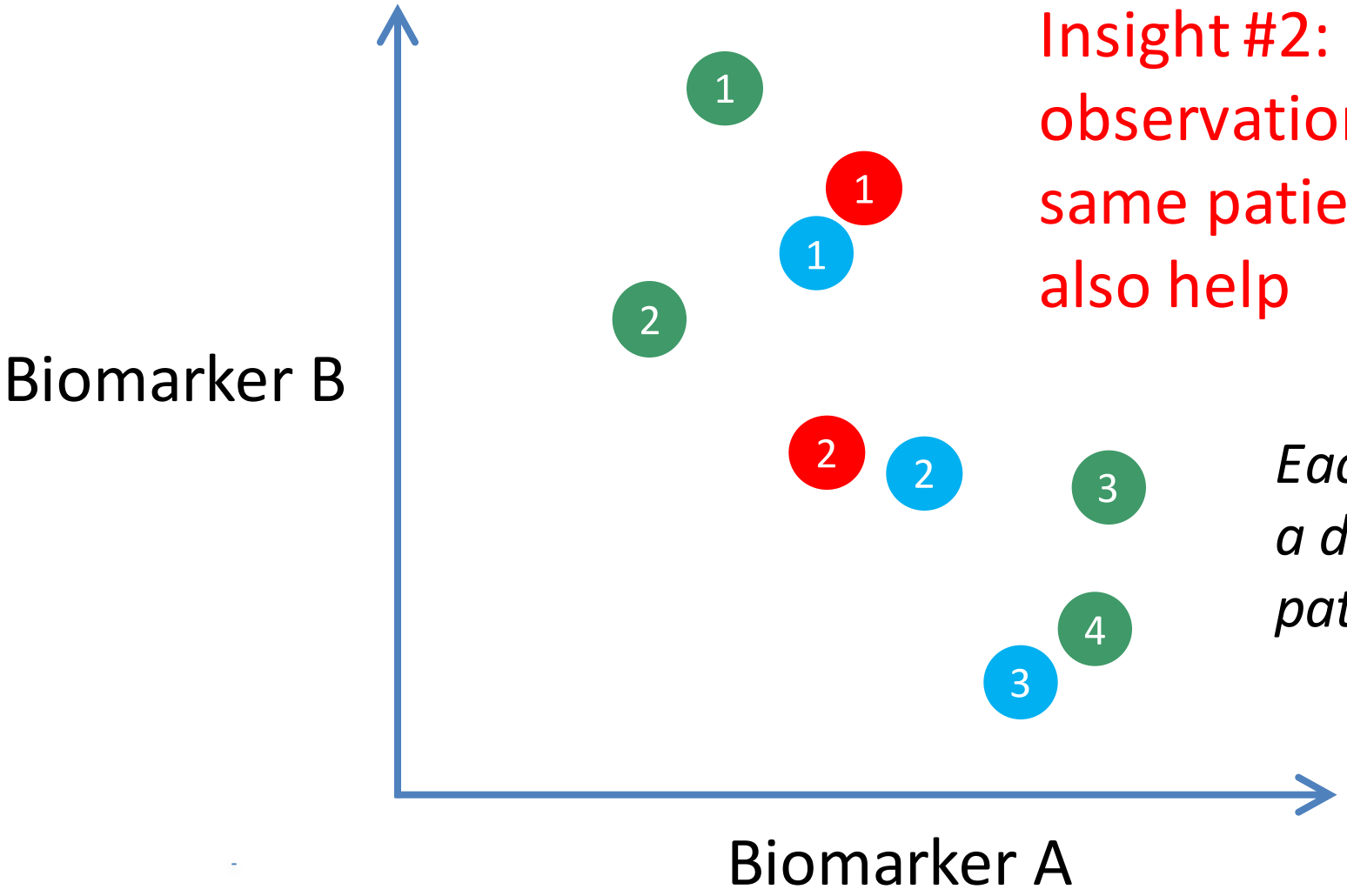
What about in higher dimensions?



Insight #1: with enough data, may be possible to recognize structure

[Bendall et al., Cell 2014 (human B cell development)]

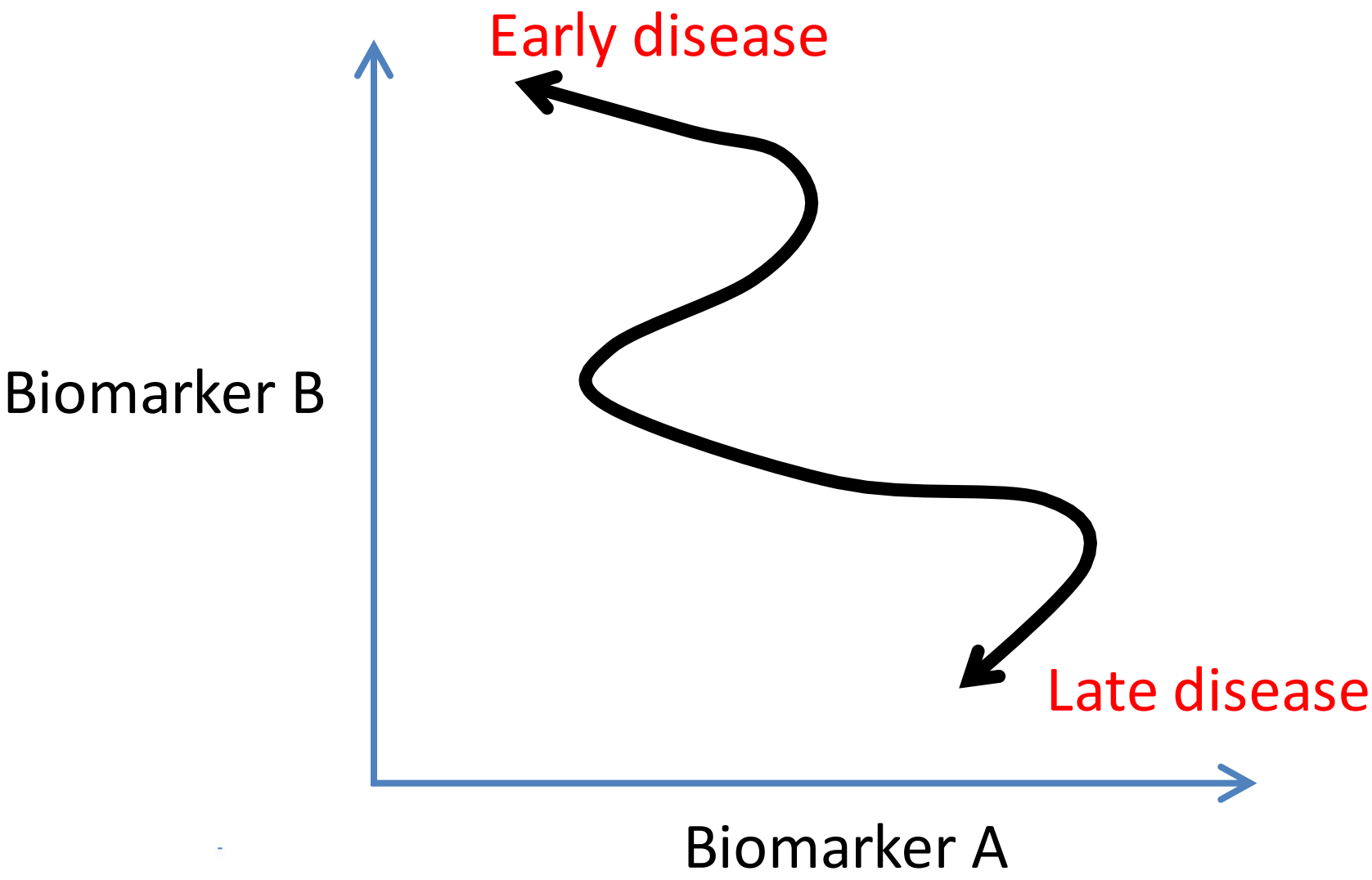
What about in higher dimensions?



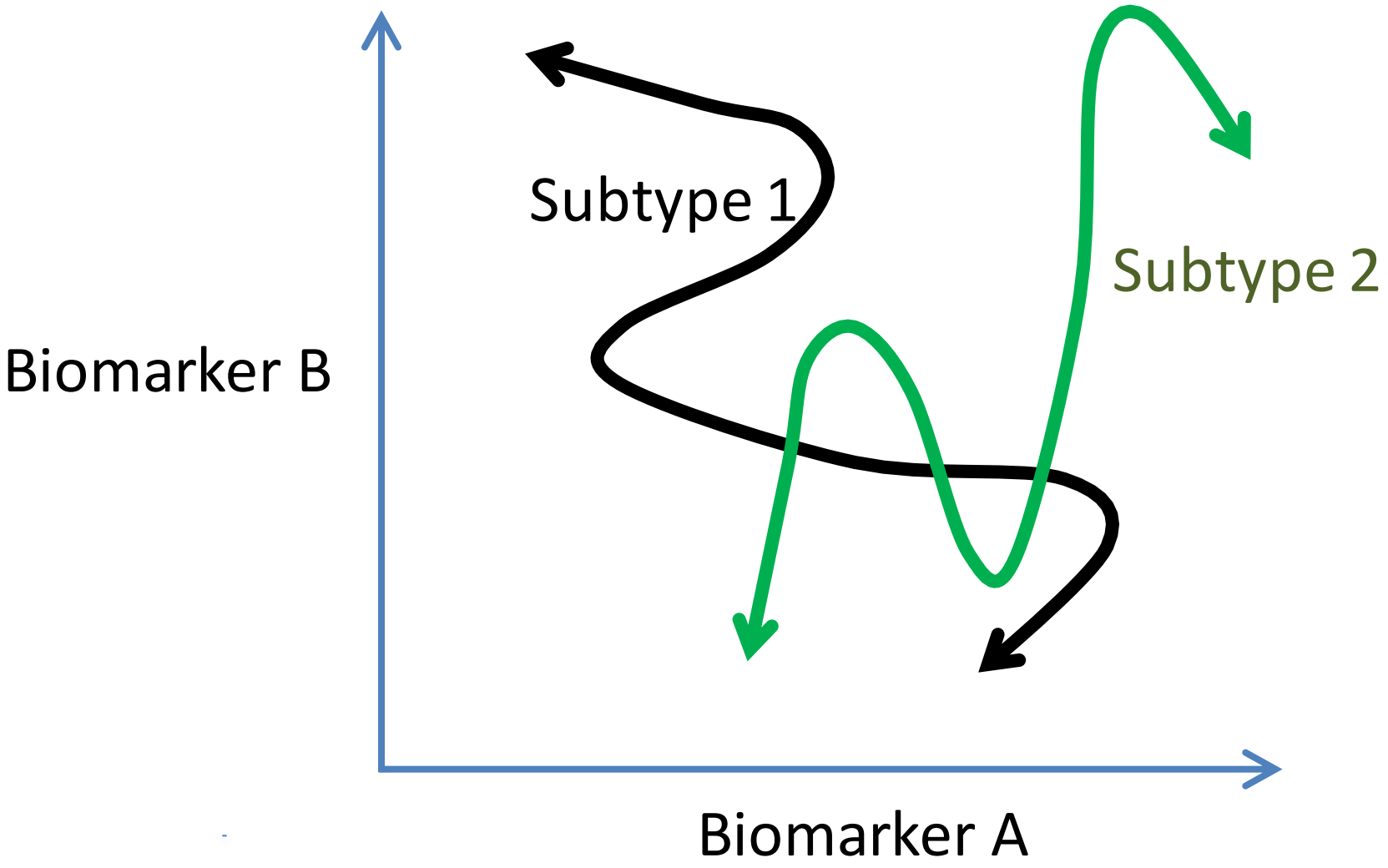
Insight #2: sequential observations from same patient can also help

Each color is a different patient

What about in higher dimensions?



May also seek to discover disease subtypes



Outline of today's lecture

1. Staging from cross-sectional data
   - **Wang, Sontag, Wang, *KDD* 2014**
   - Pseudo-time methods from computational biology
2. Simultaneous staging & subtyping
   - Young et al., *Nature Communications* 2018

COPD diagnosis & progression

- COPD diagnosis made using a breath test – fraction of air expelled in first second of exhalation < 70%
- Most doctors use GOLD criteria to stage the disease and measure its progression:

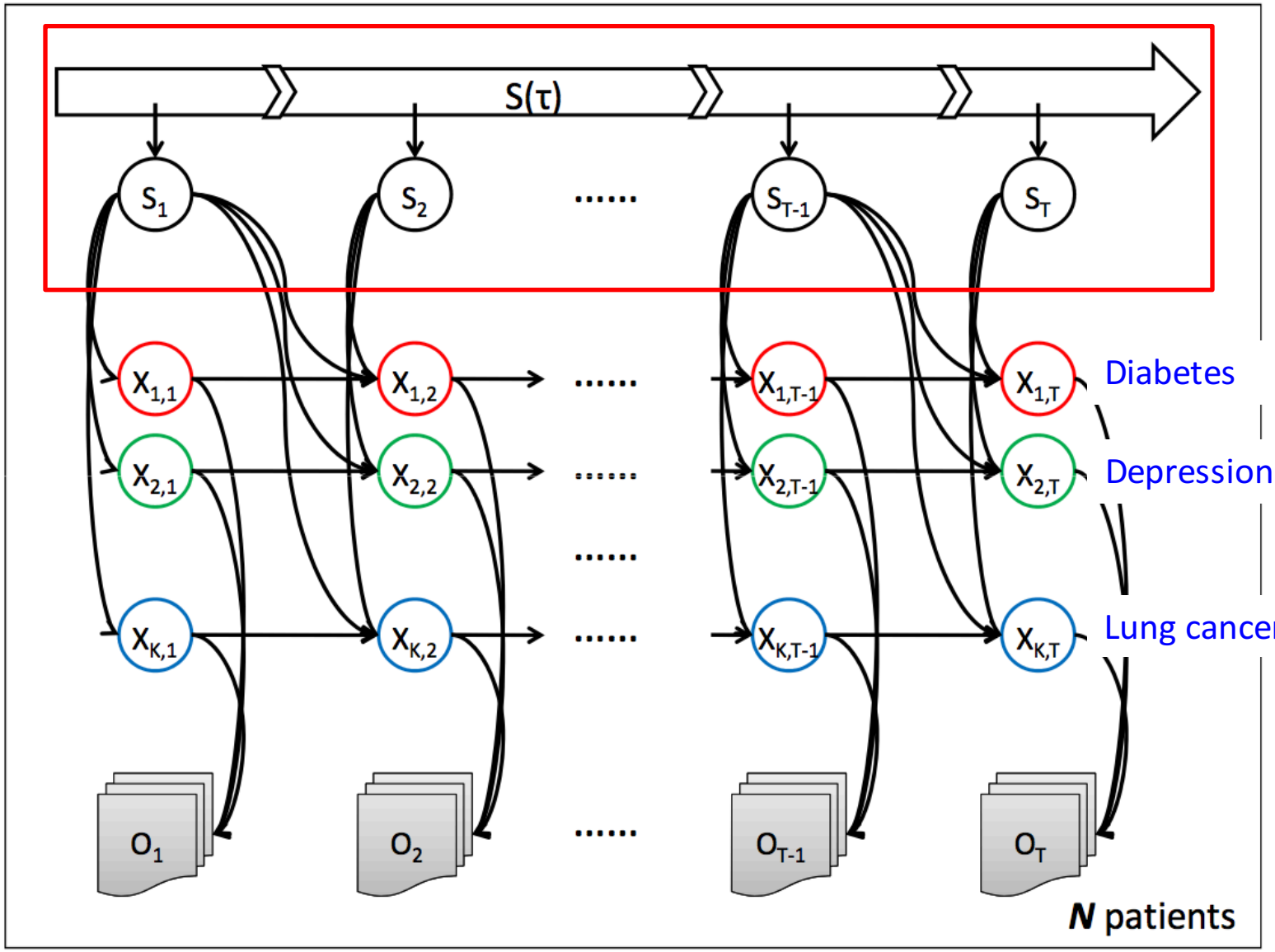
	1 (mild)	2 (moderate)	3 (severe)	4 (very severe)
FEV_1:FVC	<0·70	<0·70	<0·70	<0·70
FEV_1	≥80% of predicted	50–80% of predicted	30–50% of predicted	<30% of predicted or <50% of predicted plus chronic respiratory failure
Treatment	Influenza vaccination and short-acting bronchodilator* when needed	Influenza vaccination, short-acting and ≥1 long-acting bronchodilator* when needed; consider respiratory rehabilitation	Influenza vaccination and short-acting and ≥1 long-acting bronchodilator* when needed, inhaled glucocorticosteroid if repeated exacerbations; consider respiratory rehabilitation	Influenza vaccination and short-acting and ≥1 long-acting bronchodilator* when needed, inhaled glucocorticosteroid if repeated exacerbations, long-term oxygen if chronic respiratory failure occurs; consider respiratory rehabilitation and surgery

GOLD=Global Initiative on Obstructive Lung Disease. *β_2 agonists or anticholinergics.

Table: **Therapy at each stage of chronic obstructive pulmonary disease, by GOLD stage[1]**

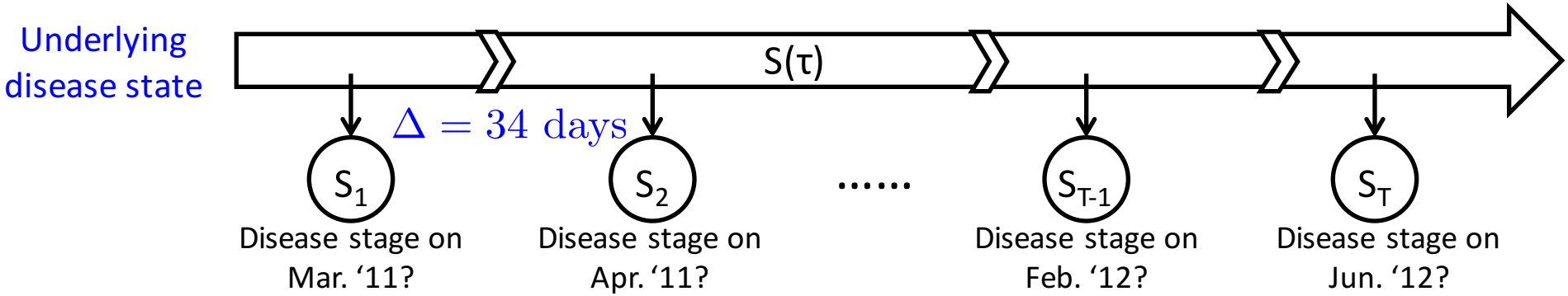
Chronic obstructive pulmonary disease. The Lancet, Volume 379, Issue 9823, Pages 1341 - 1351, 7 April 2012

The big picture: generative model for patient data



[Wang, Sontag, Wang, "Unsupervised learning of Disease Progression Models", KDD 2014]

Model for patient's disease progression across time

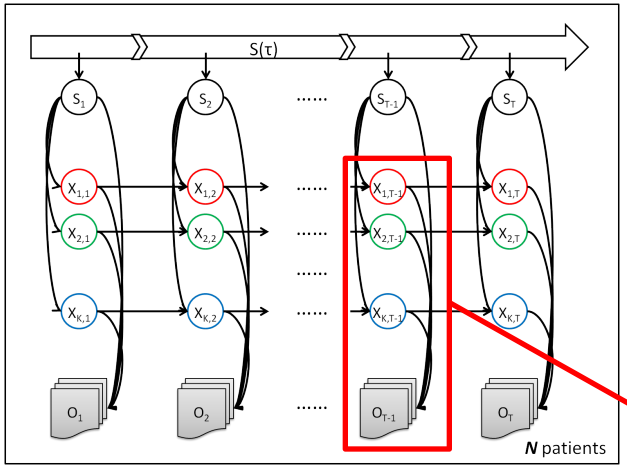


- A continuous-time Markov process with irregular discrete-time observations
- The transition probability is defined by an intensity matrix and the time interval:

$$A_{ij}(\Delta) \triangleq P(S_t = j | S_{t-1} = i, \tau_t - \tau_{t-1} = \Delta; Q)$$
$$= \text{expm}(\Delta Q)_{ij},$$

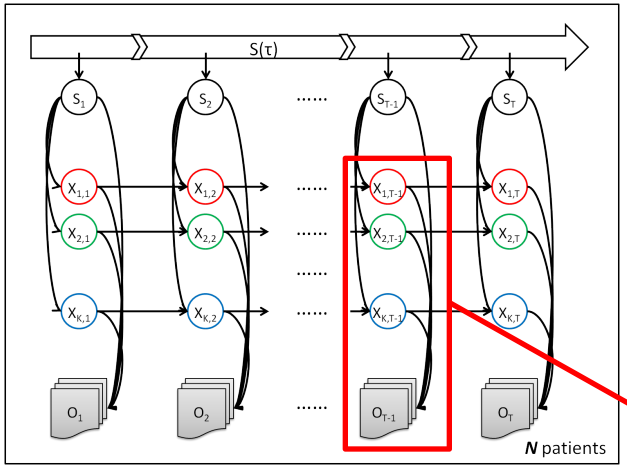
Matrix Q: Parameters to learn

Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. ’91)

Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. '91)

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

Diabetes

Depression

Lung cancer

X_K

L

All binary variables

Diagnosis codes, medications, etc.

205.02

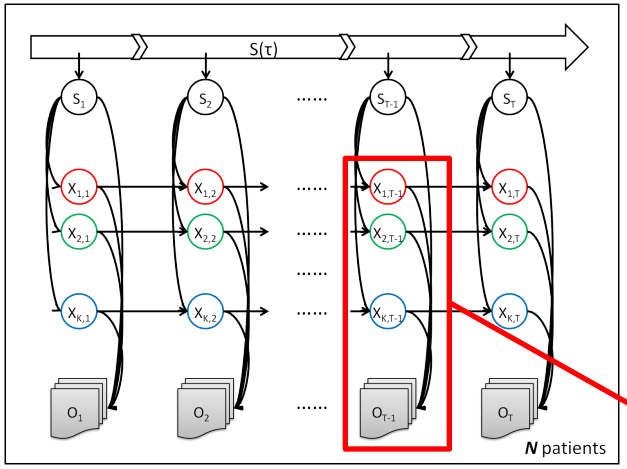
296.3

···· Methotrexate

O_D

Clinical findings
(observable)

Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. '91)

Comorbidities / Phenotypes
(hidden)

"Everything else"
(always on)

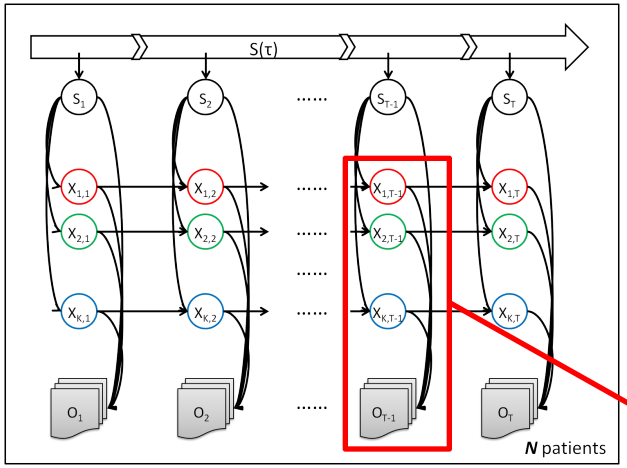
Diabetes Depression Lung cancer X_K L

We also learn which edges exist

205.02 296.3 .... Methotrexate O_D

Clinical findings
(observable)

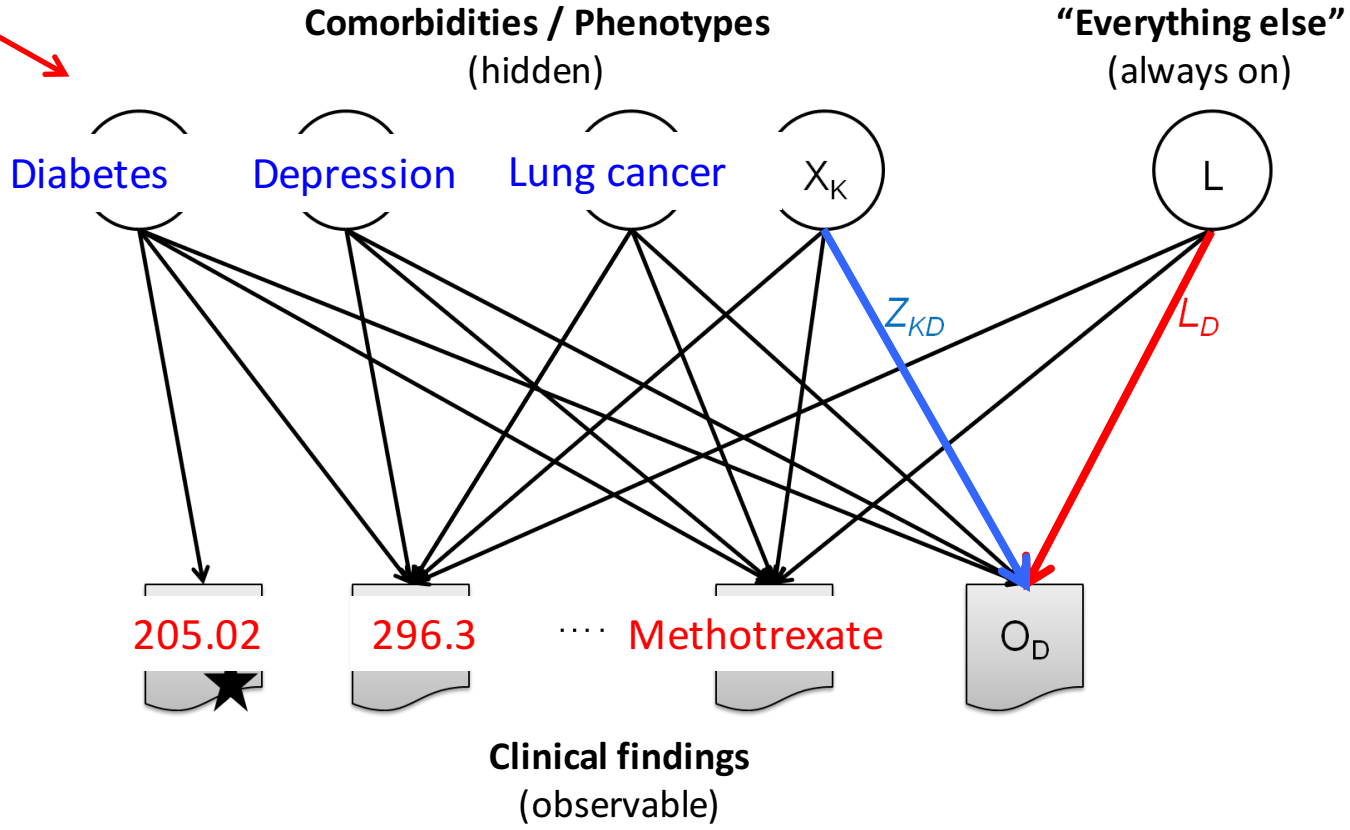
Model for data at single point in time: Noisy-OR network



Previously used for medical diagnosis, e.g. QMR-DT (Shwe et al. '91)

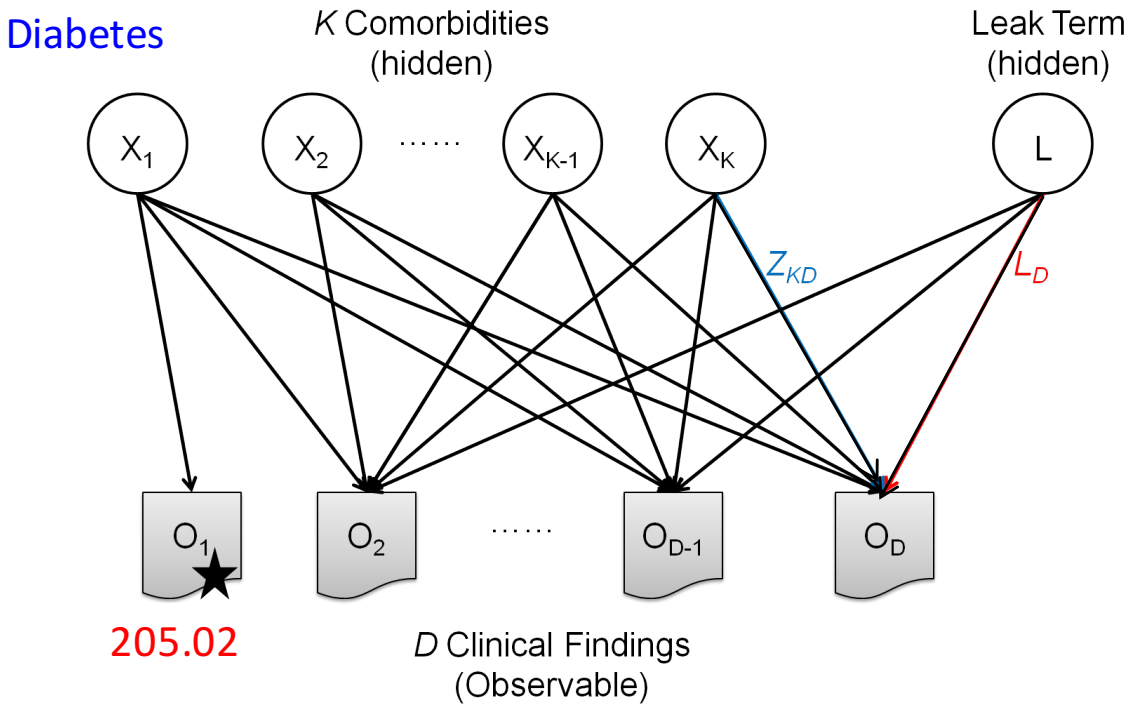
We also learn which edges exist

Associated with each edge is a *failure probability*



Using anchors to ground the hidden variables

- An *anchor* is a finding that can only be caused by a single comorbidity (discussed in Lecture 8)



Y. Halpern, YD Choi, S. Horng, D. Sontag. Using Anchors to Estimate Clinical State without Labeled Data. To appear in the American Medical Informatics Association (AMIA) Annual Symposium, Nov. 2014

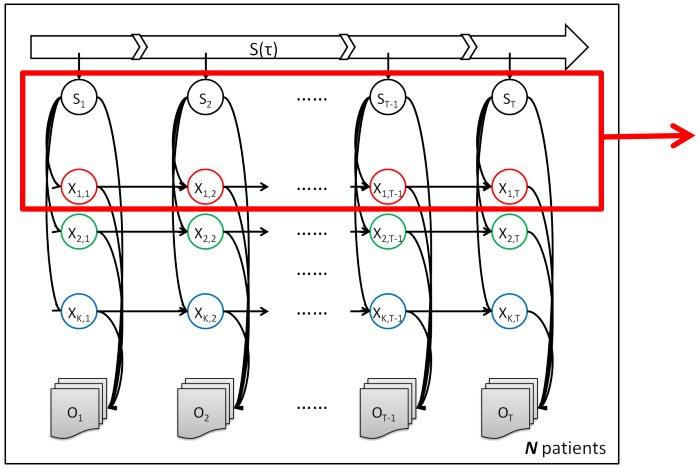
Using anchors to ground the hidden variables

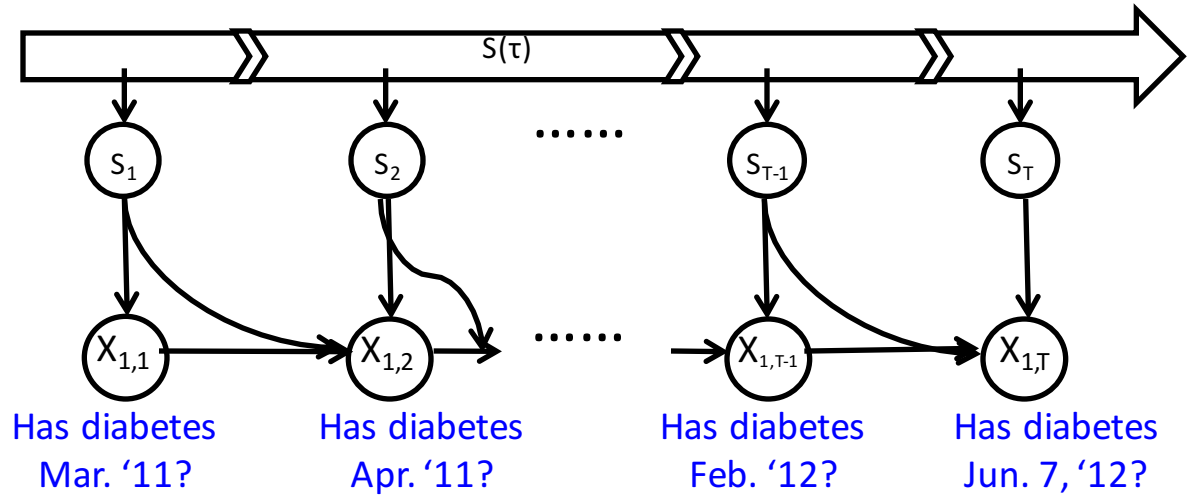
- Provide anchors for each of the comorbidities:

Comorbidity	Representative Conditions (Anchor ICD-9 Codes)
COPD	Chronic Bronchitis (491), Emphysema (492, 518), Chronic Airway Obstruction (496)
Asthma	Asthma (493)
Cardiovascular	Hypertension (401), Congestive Heart Failure (428), Arrhythmia (427), Ischemic Heart Disease (414)
Lung Infection	Pneumonia (481, 485, 486)
Lung Cancer	Malignant Neoplasm of Upper/Lower Lobe, Bronchus or Lung (162)
Diabetes	Diabetes with Different Types and Complications (250)
Musculoskeletal	Spinal Disorders (724), Soft Tissue Disorders (729), Osteoporosis (733)
Kidney	Acute Kidney Failure (584), Chronic Kidney Disease (585), Renal Failure (586)
Psychological	Anxiety (300), Depression (296, 311)
Obesity	Morbid Obesity (278)

- Can be viewed as a type of weak supervision, using clinical domain knowledge
- Without these, the results are less interpretable

Model of comorbidities across time





- Presence of comorbidities depends on value at previous time step and on disease stage
- Later stages of disease = more likely to develop comorbidities
- Make the assumption that once patient has a comorbidity, likely to always have it

Experimental evaluation

- We create a COPD cohort of 3,705 patients:
  - At least one COPD-related diagnosis code
  - At least one COPD-related drug
- Removed patients with too few records
- Clinical findings derived from 264 diagnosis codes
  - Removed ICD-9 codes that only occurred to a small number of patients
- Combined visits into 3-month time windows
- 34,976 visits, 189,815 positive findings

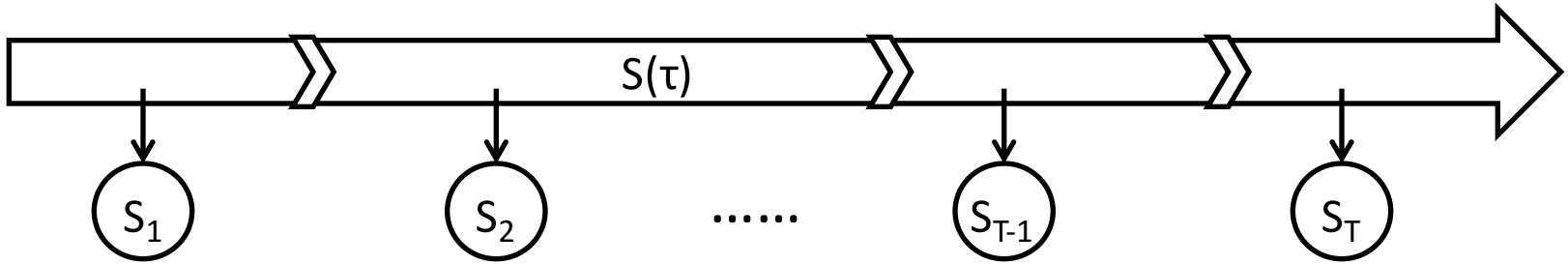
Inference

- Outer loop
  - EM
  - Algorithm to estimate the Markov Jump Process is borrowed form recent literature in physics
- Inner loop
  - Gibbs sampler used for approximate inference
  - Perform block sampling of the Markov chains, improving the mixing time of the Gibbs sampler
- **If I were to do it again… would do variational inference with a recognition network (as in VAEs)**

P. Metzner, I. Horenko, and C. Schutte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. Physical Review E, 76(6):066702, 2007.

Customizations for COPD

- Enforce monotonic stage progression, i.e. $S_{t+1} \geq S_t$:



- Enforce monotonicity in distributions of comorbidities in first time step, e.g. $\Pr(X_{j,1} \mid S_1 = 2) \geq \Pr(X_{j,1} \mid S_1 = 1)$
  - To do this, we solve a tiny convex optimization problem within EM
- Enforce that transitions in X can only happen at the same time as transitions in S
- Edge weights given a Beta(0.1, 1) prior to encourage sparsity

Edges learned for *kidney disease*

Diagnosis code	Weight	
*585.3	0.20	Chronic Kidney Disease, Stage Iii (Moderate)
285.9	0.15	Anemia, Unspecified
*585.9	0.10	Chronic Kidney Disease, Unspecified
599.0	0.08	Urinary Tract Infection, Site Not Specified
*585.4	0.08	Chronic Kidney Disease, Stage Iv (Severe)
*584.9	0.07	Acute Renal Failure, Unspecified
*586	0.07	Renal Failure, Unspecified
782.3	0.06	Edema
*585.6	0.05	End Stage Renal Disease
593.9	0.04	Unspecified Disorder Of Kidney And Ureter
272.4	0.04	Other And Unspecified Hyperlipidemia
272.2	0.03	Mixed Hyperlipidemia

Edges learned for *kidney disease*

Diagnosis code	Weight	
***585.3**	**0.20**	**Chronic Kidney Disease, Stage Iii (Moderate)**
285.9	0.15	Anemia, Unspecified
***585.9**	**0.10**	**Chronic Kidney Disease, Unspecified**
599.0	0.08	Urinary Tract Infection, Site Not Specified
***585.4**	**0.08**	**Chronic Kidney Disease, Stage Iv (Severe)**
***584.9**	**0.07**	**Acute Renal Failure, Unspecified**
***586**	**0.07**	**Renal Failure, Unspecified**
782.3	0.06	Edema
***585.6**	**0.05**	**End Stage Renal Disease**
593.9	0.04	Unspecified Disorder Of Kidney And Ureter
272.4	0.04	Other And Unspecified Hyperlipidemia
272.2	0.03	Mixed Hyperlipidemia

Edges learned for *kidney disease*

Diagnosis code	Weight	
*585.3	0.20	Chronic Kidney Disease, Stage Iii (Moderate)
285.9	**0.15**	**Anemia, Unspecified**
*585.9	0.10	Chronic Kidney Disease,
599.0	**0.08**	**Urinary Tract Infection,**
*585.4	0.08	Chronic Kidney Disease,
*584.9	0.07	Acute Renal Failure, Uns
*586	0.07	Renal Failure, Unspecifie
782.3	**0.06**	**Edema**
*585.6	0.05	End Stage Renal Disease
593.9	**0.04**	**Unspecified Disorder Of**
272.4	**0.04**	**Other And Unspecified**
272.2	**0.03**	**Mixed Hyperlipidemia**

Why do people with kidney disease get anemia?

Your kidneys make an important hormone called *erythropoietin (EPO)*. Hormones are secretions that your body makes to help your body work and keep you healthy. EPO tells your body to make red blood cells. When you have kidney disease, your kidneys cannot make enough EPO. This causes your red blood cell count to drop and anemia to develop.

Edges learned for *lung cancer*

Diagnosis code	Weight	
*162.9	0.60	Malignant Neoplasm Of Bronchus And Lung
518.89	0.15	Other Diseases Of Lung, Not Elsewhere Classified
*162.8	0.15	Malignant Neoplasm Of Other Parts Of Lung
*162.3	0.15	Malignant Neoplasm Of Upper Lobe, Lung
786.6	0.15	Swelling, Mass, Or Lump In Chest
793.1	0.10	Abnormal Findings On Radiological Exam Of Lung
786.09	0.07	Other Respiratory Abnormalities
*162.5	0.06	Malignant Neoplasm Of Lower Lobe, Lung
*162.2	0.04	Malignant Neoplasm Of Main Bronchus
702.0	0.03	Actinic Keratosis
511.9	0.03	Unspecified Pleural Effusion
*162.4	0.03	Malignant Neoplasm Of Middle Lobe, Lung

Edges learned for *lung cancer*

Diagnosis code	Weight	
***162.9**	**0.60**	**Malignant Neoplasm Of Bronchus And Lung**
518.89	0.15	Other Diseases Of Lung, Not Elsewhere Classified
***162.8**	**0.15**	**Malignant Neoplasm Of Other Parts Of Lung**
***162.3**	**0.15**	**Malignant Neoplasm Of Upper Lobe, Lung**
786.6	0.15	Swelling, Mass, Or Lump In Chest
793.1	0.10	Abnormal Findings On Radiological Exam Of Lung
786.09	0.07	Other Respiratory Abnormalities
***162.5**	**0.06**	**Malignant Neoplasm Of Lower Lobe, Lung**
***162.2**	**0.04**	**Malignant Neoplasm Of Main Bronchus**
702.0	0.03	Actinic Keratosis
511.9	0.03	Unspecified Pleural Effusion
***162.4**	**0.03**	**Malignant Neoplasm Of Middle Lobe, Lung**

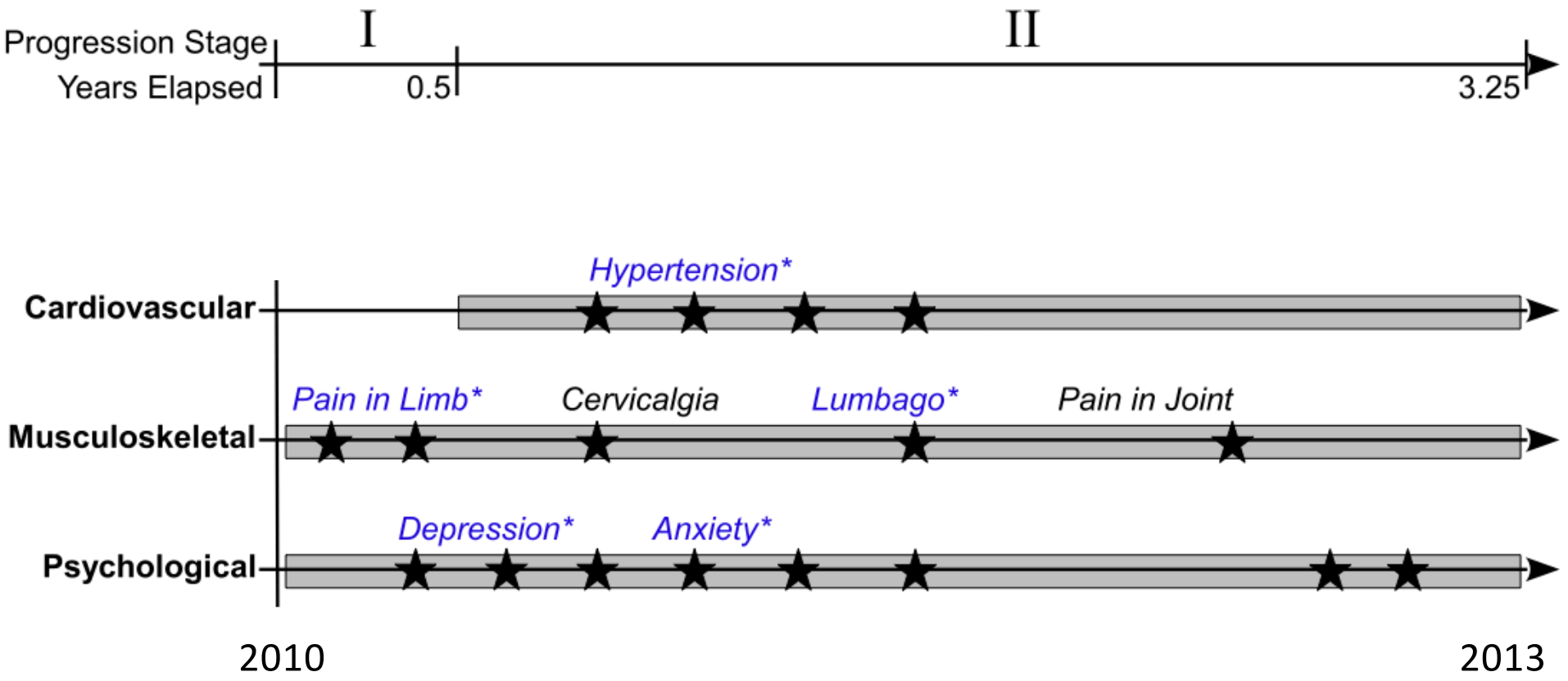
Edges learned for *lung cancer*

Diagnosis code	Weight	
*162.9	0.60	Malignant Neoplasm Of Bronchus And Lung
518.89	**0.15**	**Other Diseases Of Lung, Not Elsewhere Classified**
*162.8	0.15	Malignant Neoplasm Of Other Parts Of Lung
*162.3	0.15	Malignant Neoplasm Of Upper Lobe, Lung
786.6	**0.15**	**Swelling, Mass, Or Lump In Chest**
793.1	**0.10**	**Abnormal Findings On Radiological Exam Of Lung**
786.09	**0.07**	**Other Respiratory Abnormalities**
*162.5	0.06	Malignant Neoplasm Of Lower Lobe, Lung
*162.2	0.04	Malignant Neoplasm Of Main Bronchus
702.0	**0.03**	**Actinic Keratosis**
511.9	**0.03**	**Unspecified Pleural Effusion**
*162.4	0.03	Malignant Neoplasm Of Middle Lobe, Lung

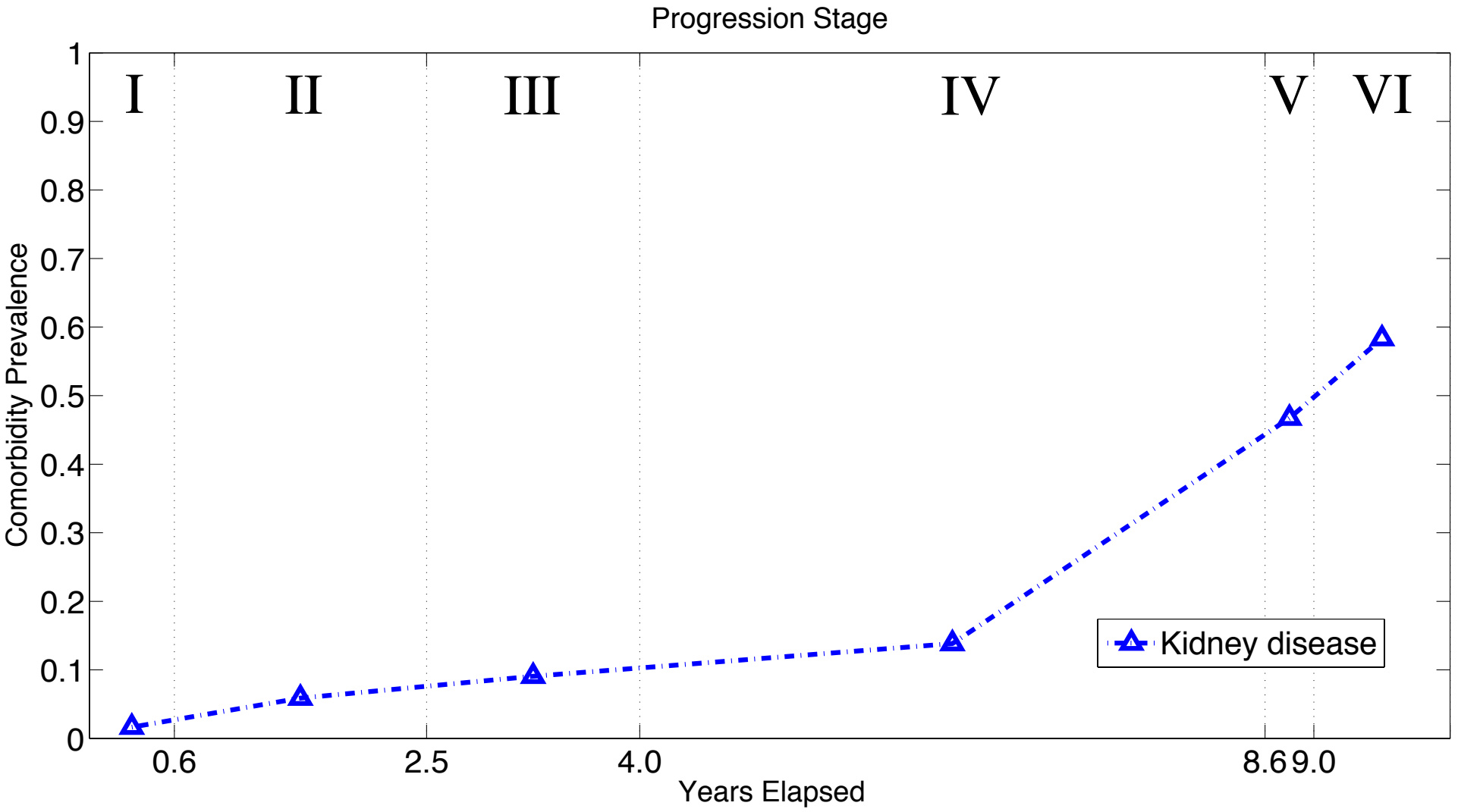
Edges learned for *lung infection*

Diagnosis code	Weight	
***486**	**0.30**	**Pneumonia, Organism Unspecified**
786.05	0.10	Shortness Of Breath
786.09	0.10	Other Respiratory Abnormalities
786.2	0.10	Cough
793.1	0.06	Abnormal Findings On Radiological Exam Of Lung
285.9	0.05	Anemia, Unspecified
518.89	0.05	Other Diseases Of Lung, Not Elsewhere Classified
466.0	0.05	Acute Bronchitis
799.02	0.05	Hypoxemia
599.0	0.04	Urinary Tract Infection, Site Not Specified
V58.61	0.04	Long-Term (Current) Use Of Anticoagulants
786.50	0.04	Chest Pain, Unspecified

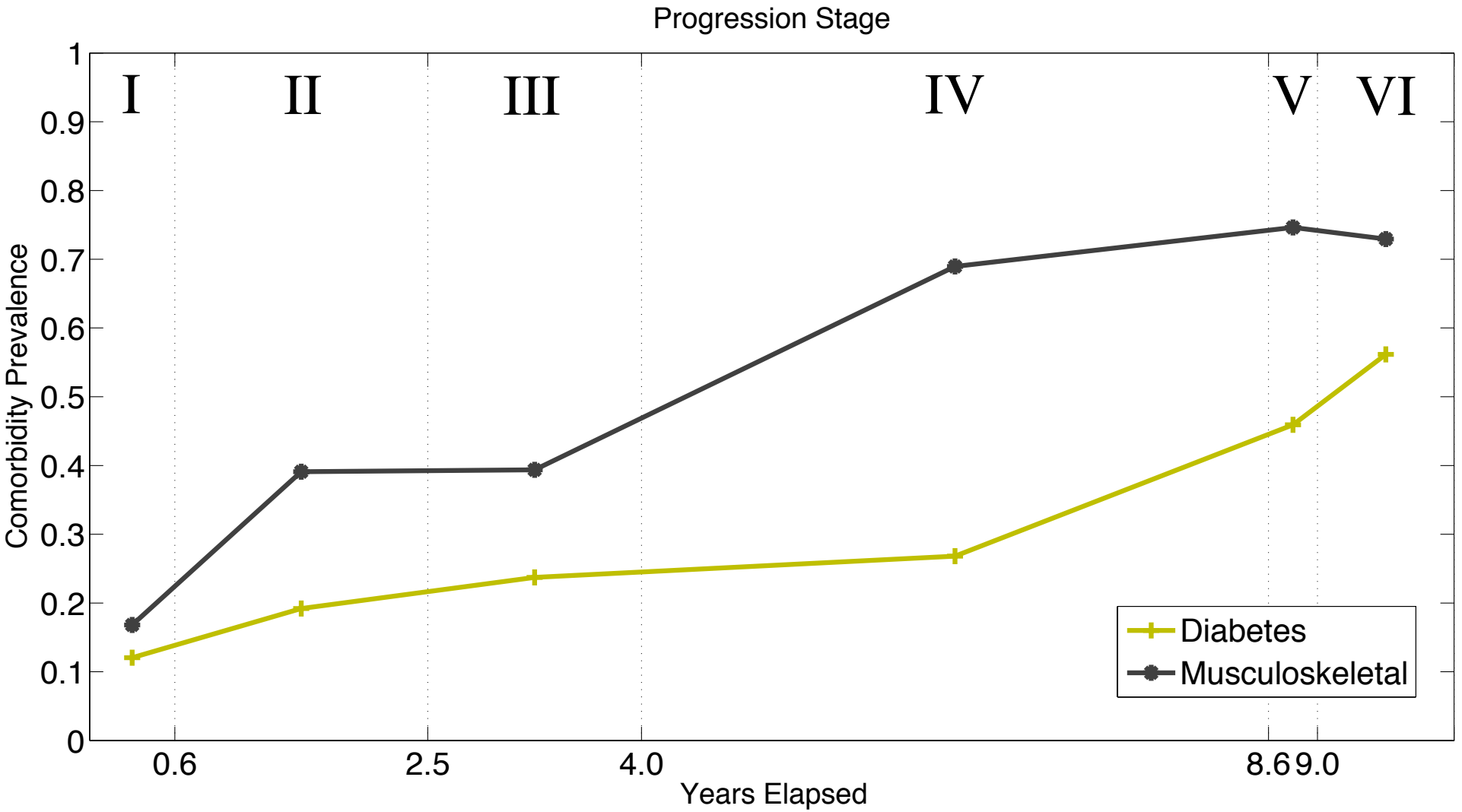
Progression of a single patient



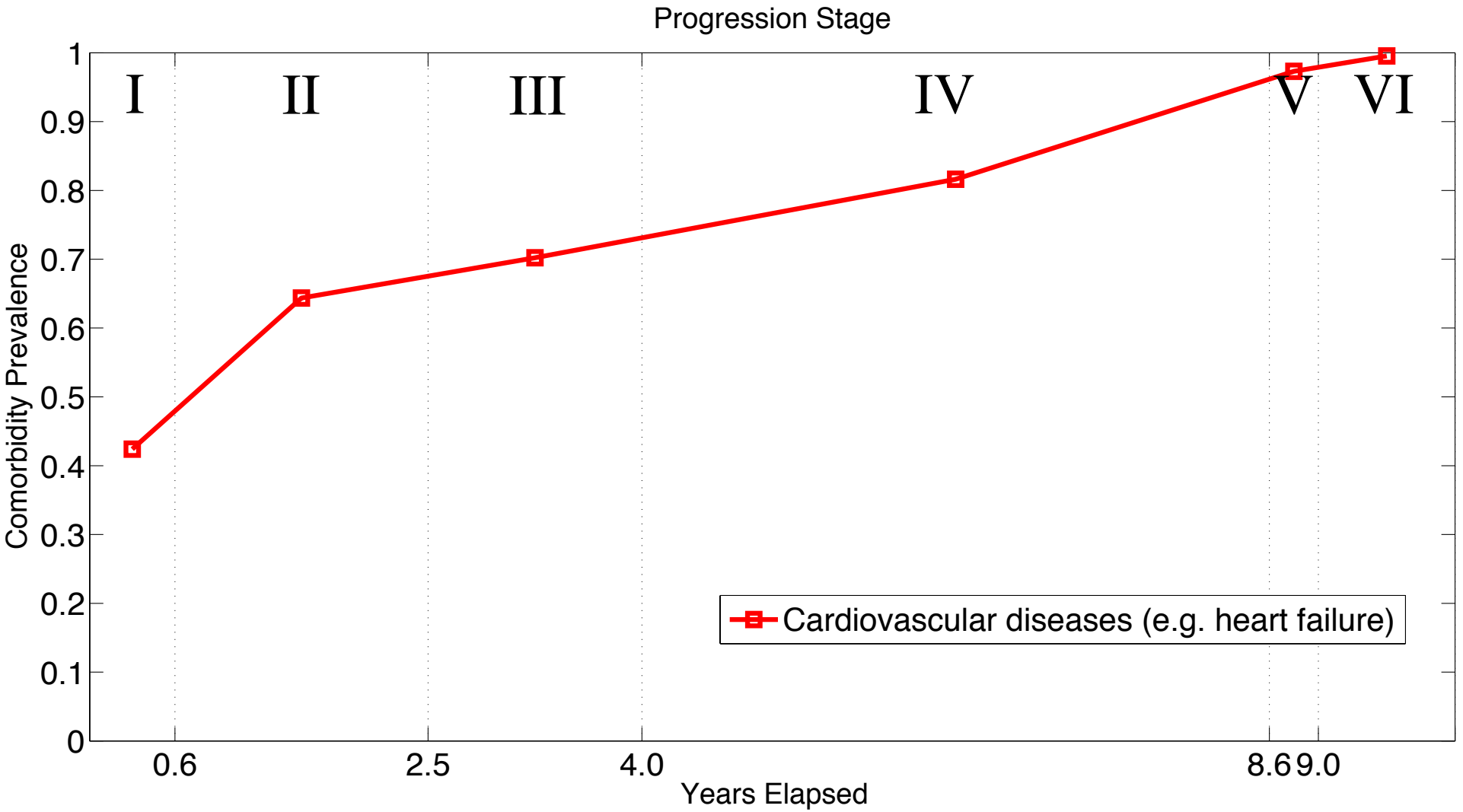
Prevalence of comorbidities across stages (Kidney disease)



Prevalence of comorbidities across stages (Diabetes & Musculoskeletal disorders)



Prevalence of comorbidities across stages (Cardiovascular disease)



Pr

CHEST JOURNAL
Official Publication of the American College c

es

August 2009, Vol 136, No. 2

< Previous in this issue

Next in this issue >

Editorials | August 2009

Is COPD Really a Cardiovascular Disease?

FREE TO VIEW

Don D. Sin, MD, FCCP

Author and Funding Information

Chest. 2009;136(2):329-330. doi:10.1378/chest.09-0808

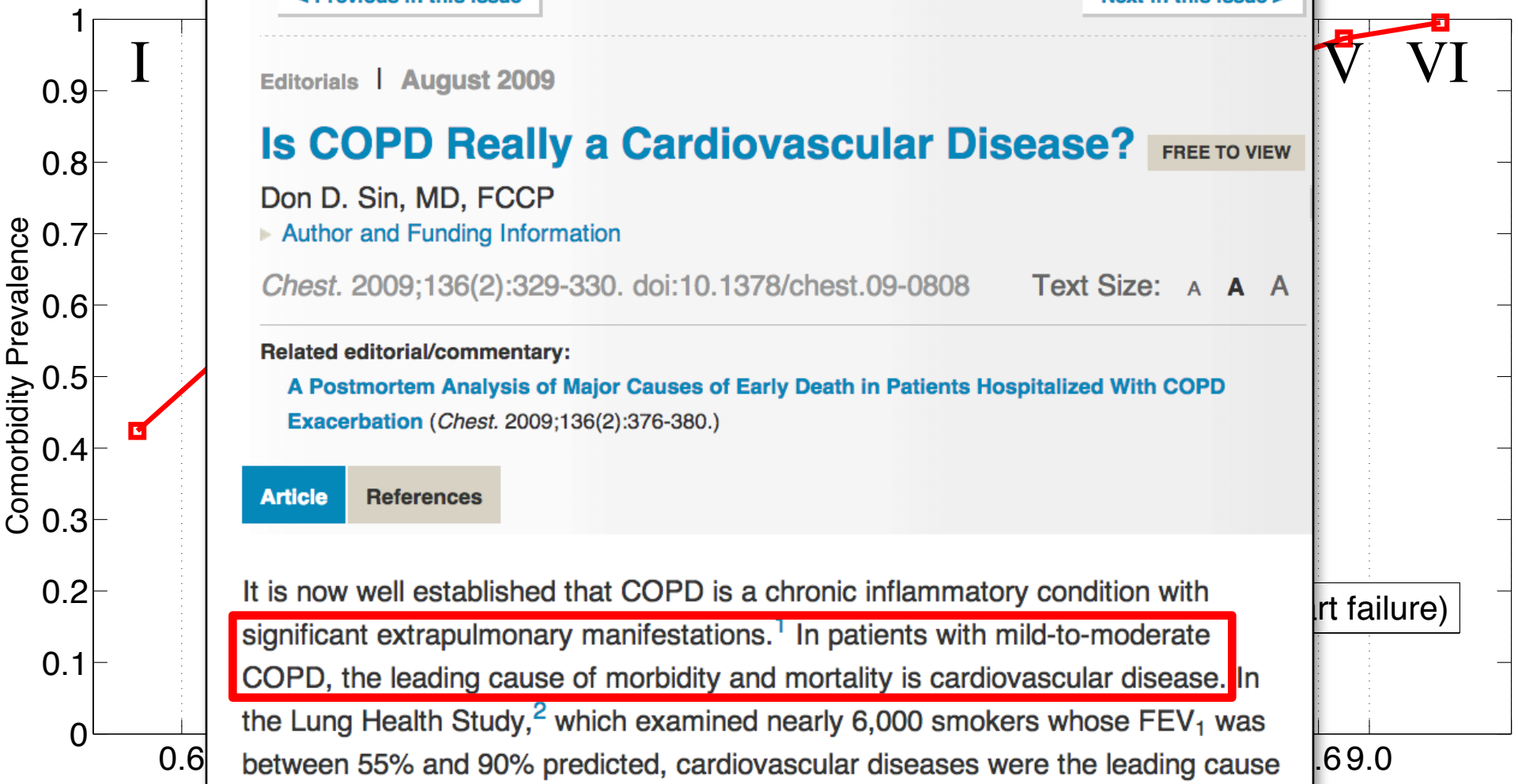
Text Size: A A A

Related editorial/commentary:

A Postmortem Analysis of Major Causes of Early Death in Patients Hospitalized With COPD Exacerbation (*Chest*. 2009;136(2):376-380.)

Article | References

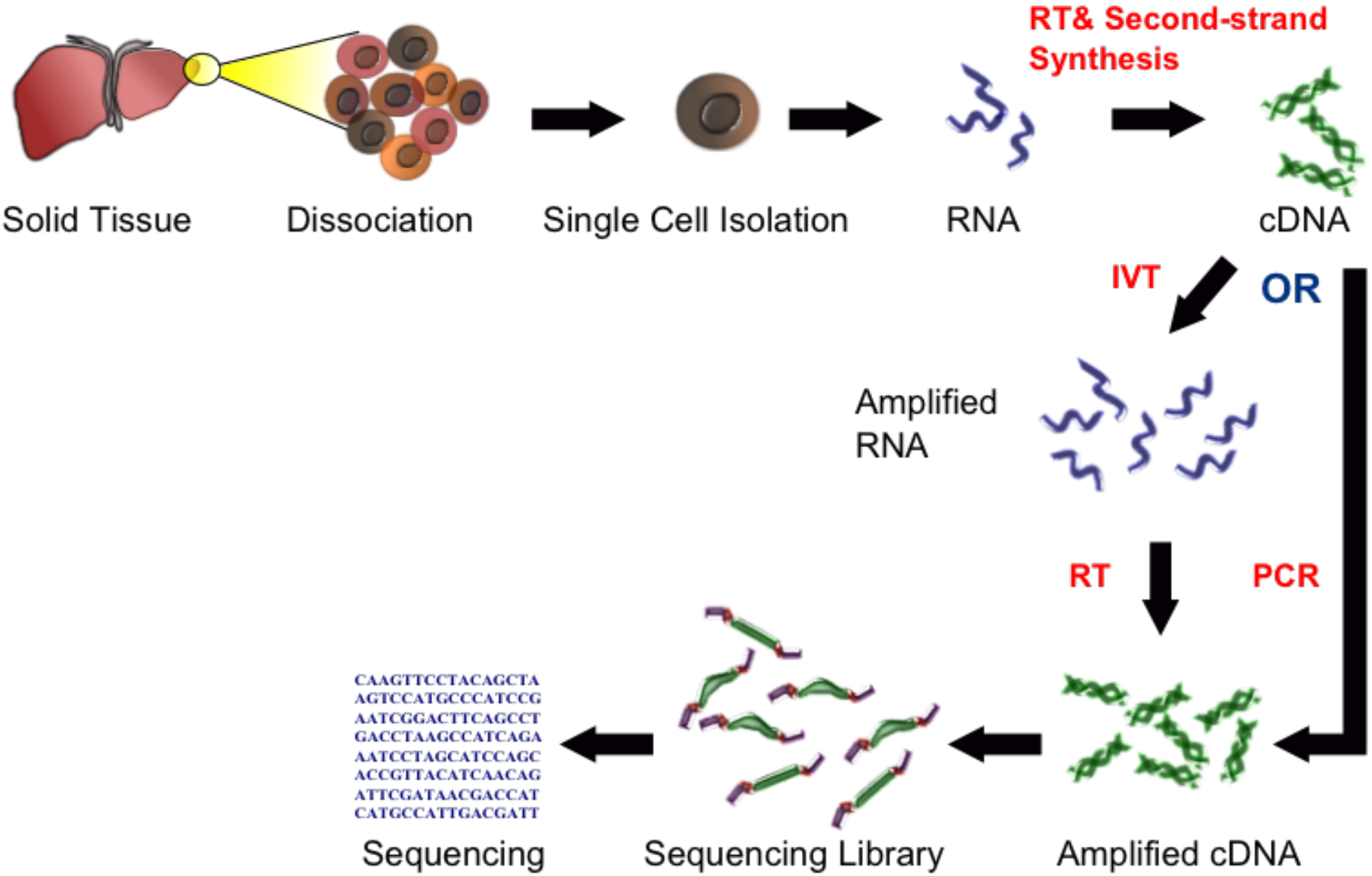
It is now well established that COPD is a chronic inflammatory condition with significant extrapulmonary manifestations.[1] In patients with mild-to-moderate COPD, the leading cause of morbidity and mortality is cardiovascular disease. In the Lung Health Study,[2] which examined nearly 6,000 smokers whose FEV_1 was between 55% and 90% predicted, cardiovascular diseases were the leading cause of hospitalization, accounting for nearly 50% of all hospital admissions, and the second leading cause of mortality, accounting for a quarter of all deaths.



Outline of today's lecture

1. Staging from cross-sectional data
   - Wang, Sontag, Wang, *KDD* 2014
   - **Pseudo-time methods from computational biology**
2. Simultaneous staging & subtyping
   - Young et al., *Nature Communications* 2018

Single-cell sequencing



[Figure source: https://en.wikipedia.org/wiki/Single_cell_sequencing]

Inferring original trajectory from single-cell data

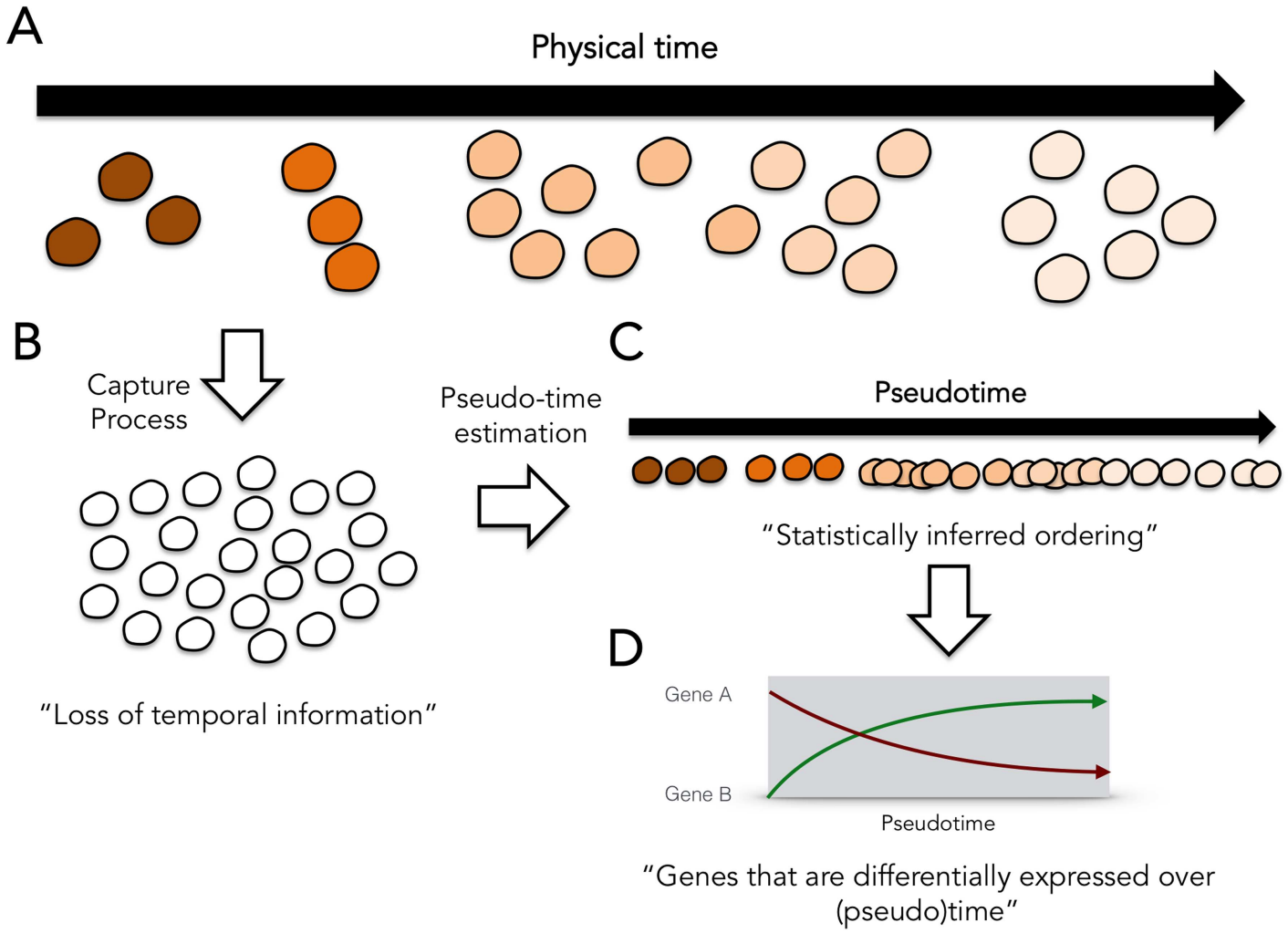
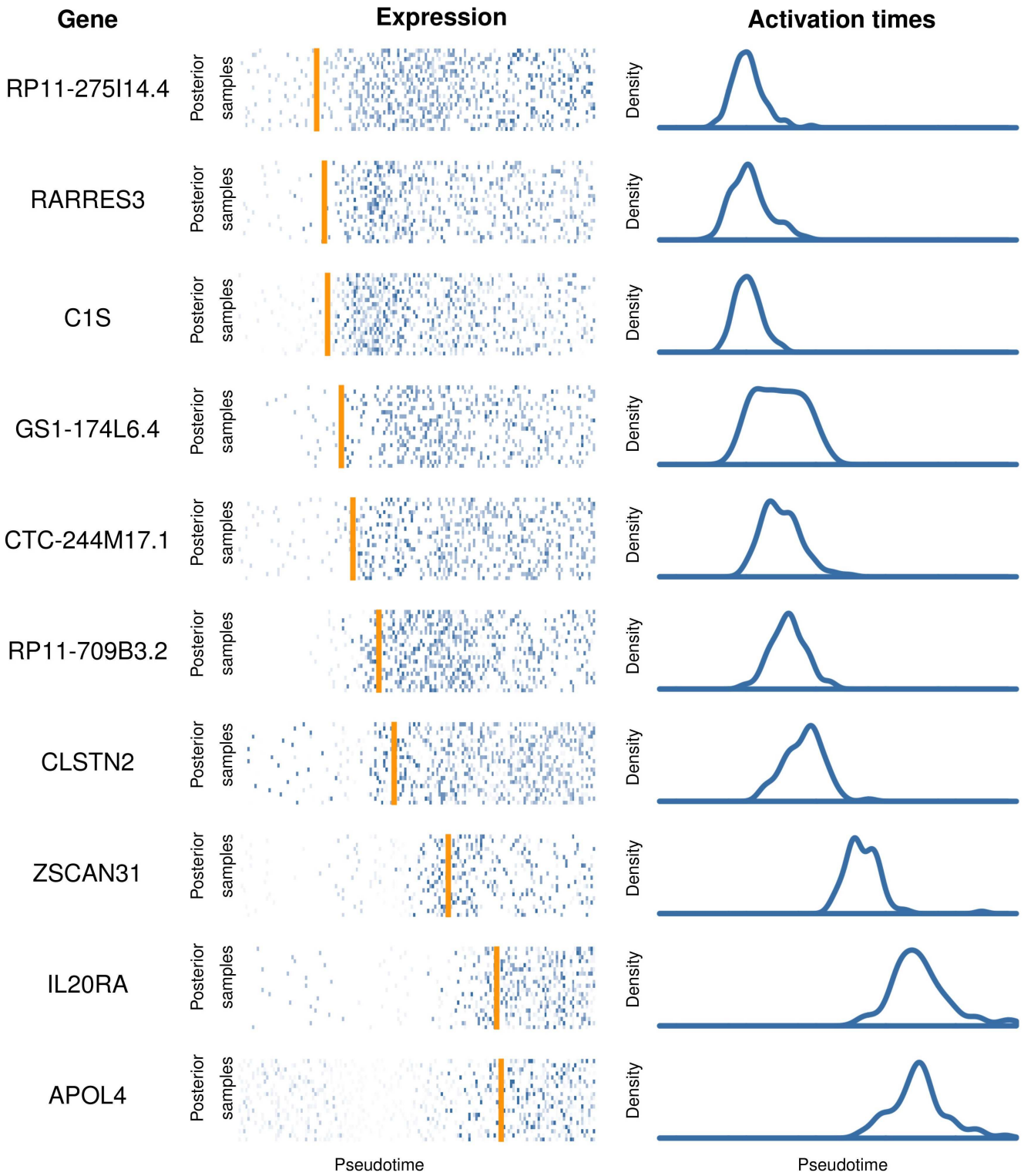
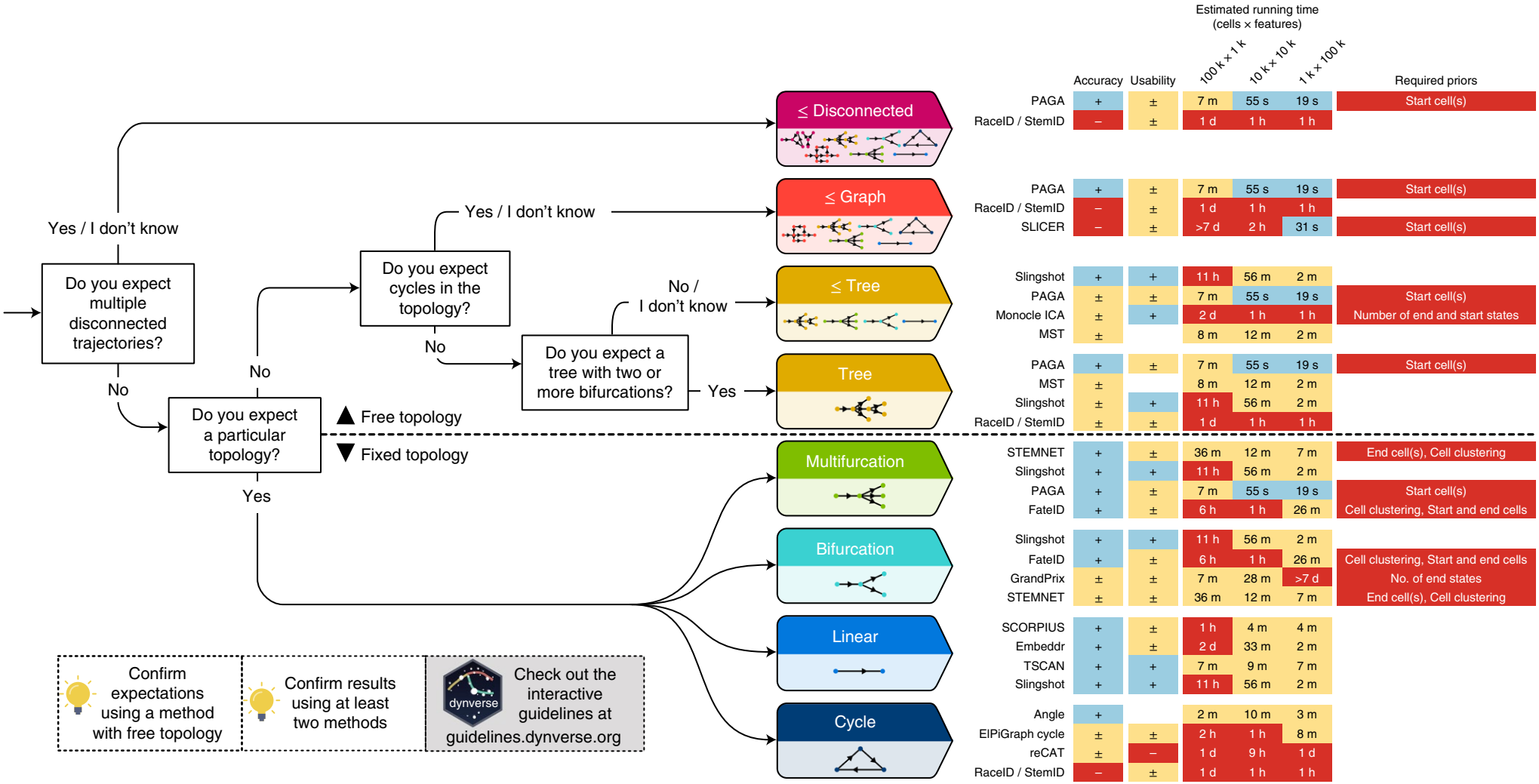


Fig 1. The single cell pseudotime estimation problem. (A) Single cells at different stages of a temporal process. (B) The temporal labelling information is lost during single cell capture. (C) Statistical pseudotime estimation algorithms attempt to reconstruct the relative temporal ordering of the cells but cannot fully reproduce physical time. (D) The pseudotime estimates can be used to identify genes that are differentially expressed over (pseudo)time.

[Figure from: Campbell & Yau, PLOS Computational Biology, 2016]



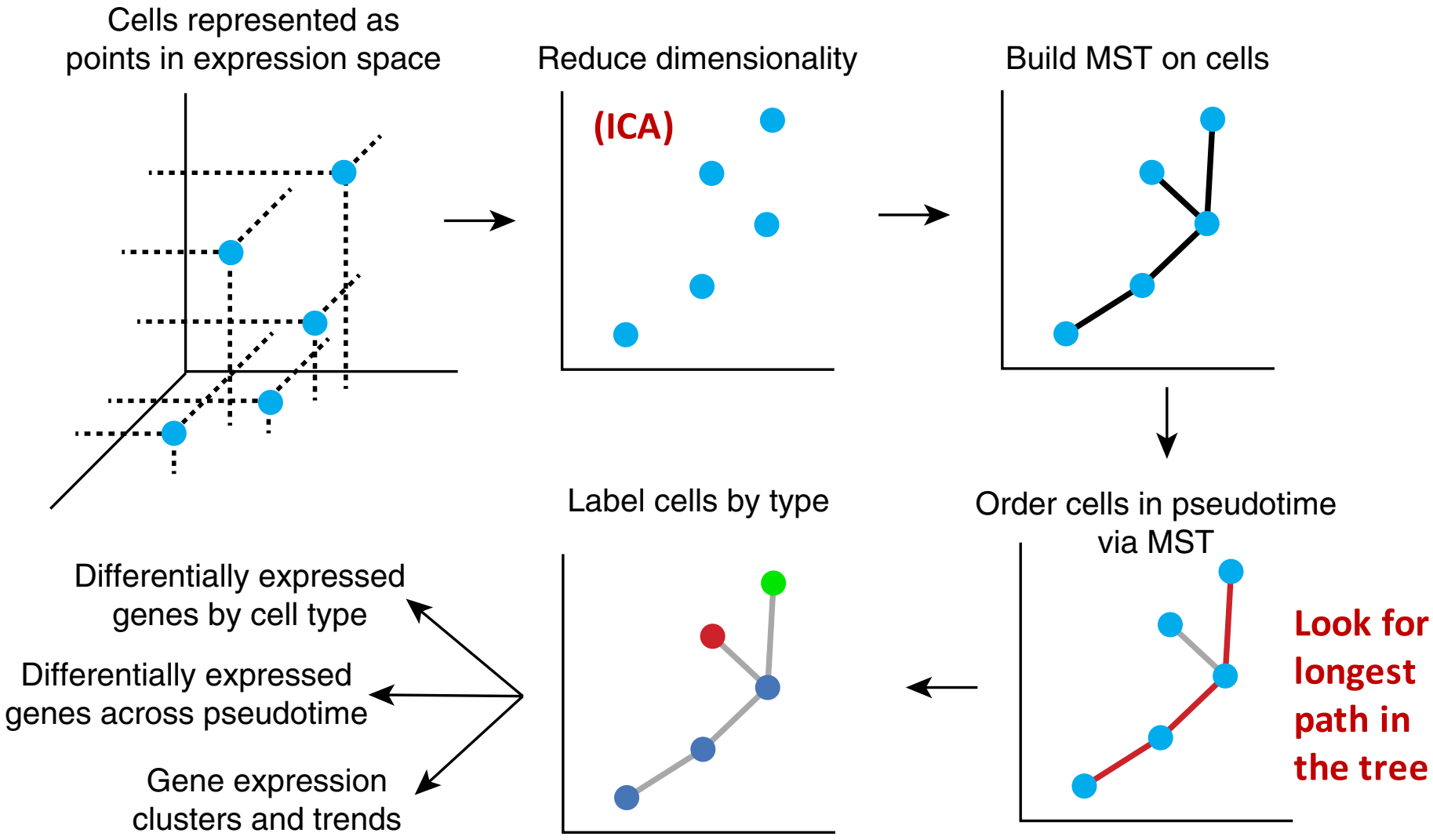
[Campbell & Yau, PLOS Computational Biology, 2016]



[Saelens, Cannoodt, Todorov, Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, 2019]

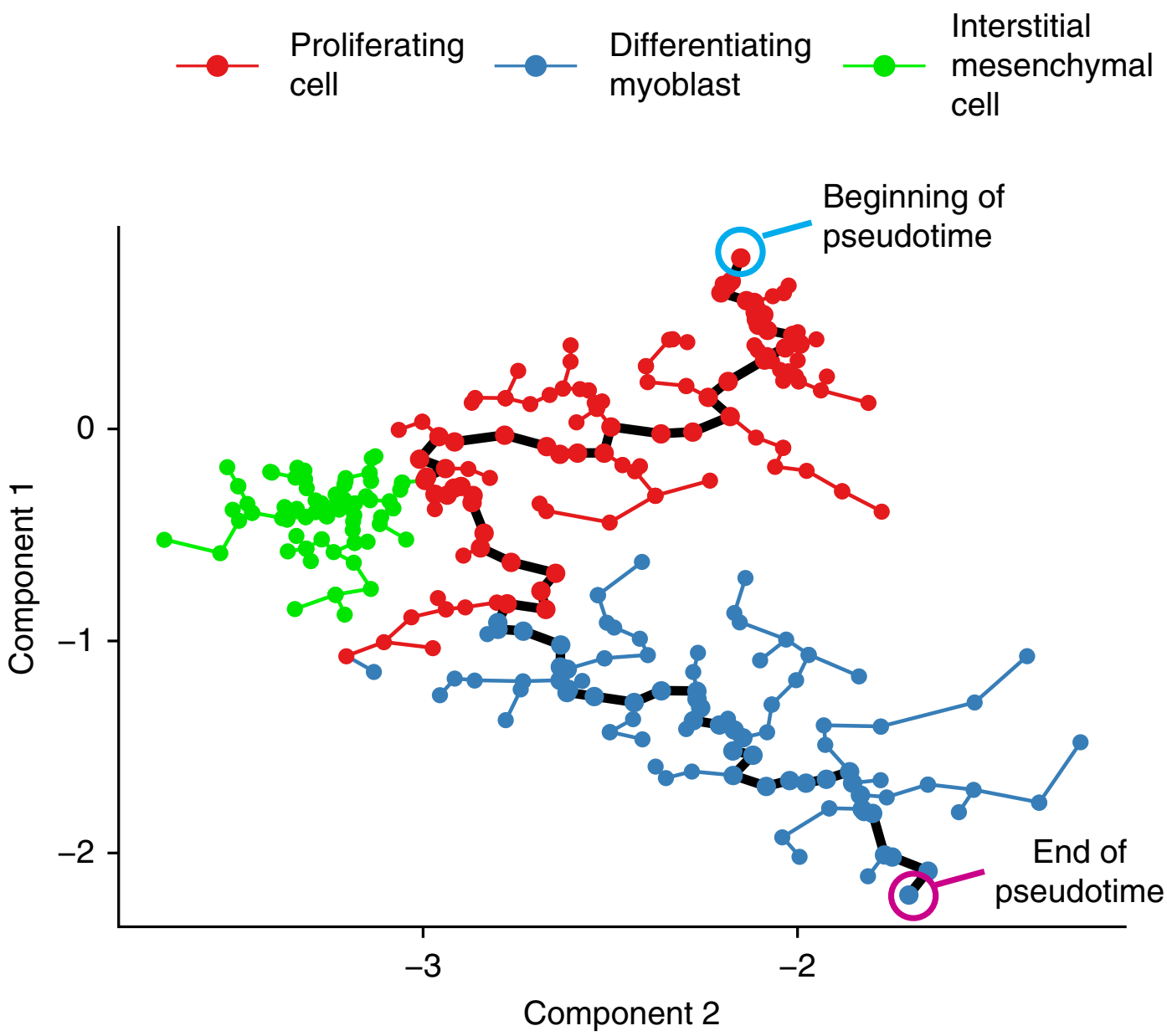
https://github.com/dynverse/dynbenchmark/

MST-based approach (Monocle)



[Magwene et al., *Bioinformatics*, 2003; Trapnell et al., *Nature Biotechnology*, 2014]

MST-based approach (Monocle)



[Trapnell et al., *Nature Biotechnology*, 2014]

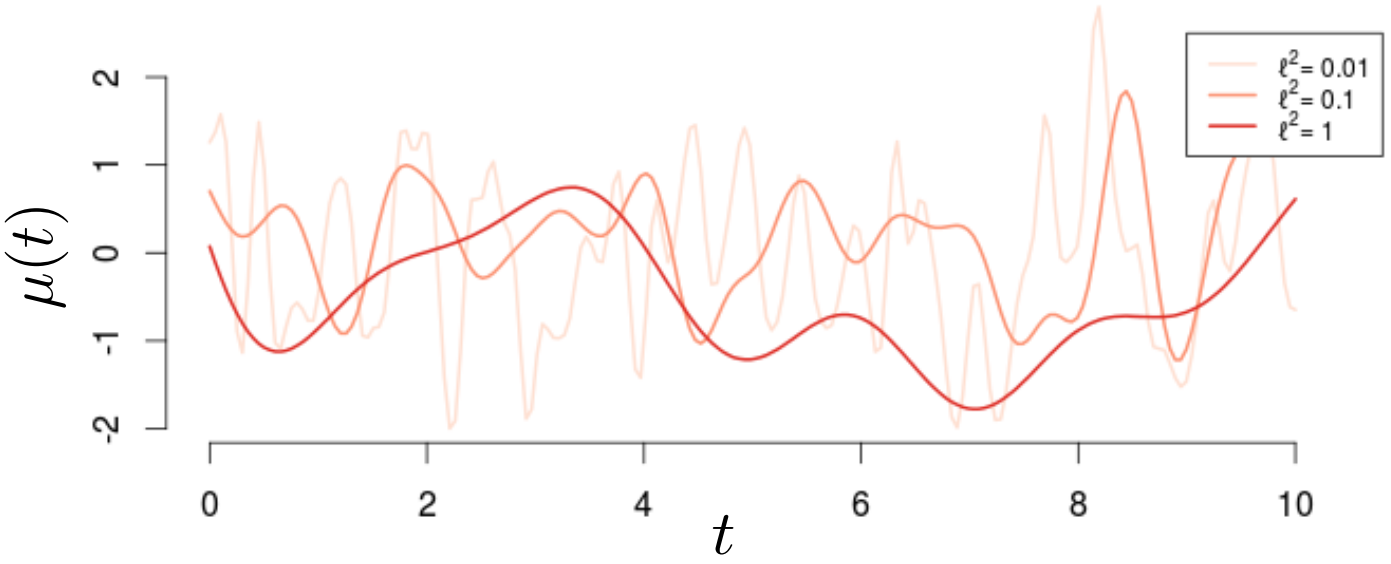
Statistical model for probabilistic pseudotime

Definition

μ is a Gaussian process if for any collection $\mathbf{T} = \{t_i, i = 1, \ldots, N\}$,

$$\begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_N) \end{pmatrix} \sim \mathcal{N}(0, K(\mathbf{T}, \mathbf{T}))$$

$$k(t_{i_1}, t_{i_2}) = \tau^2 \exp\left(-\frac{||t_{i_1} - t_{i_2}||^2}{2\ell^2}\right) \text{ (squared exponential)}$$



Statistical model for probabilistic pseudotime

$$t_i \sim \text{TruncNormal}_{[0,1)}(\mu_t, \sigma_t^2), \;\; i = 1, \ldots, N,$$

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_P^2)$$

P: dimension (e.g. 2)

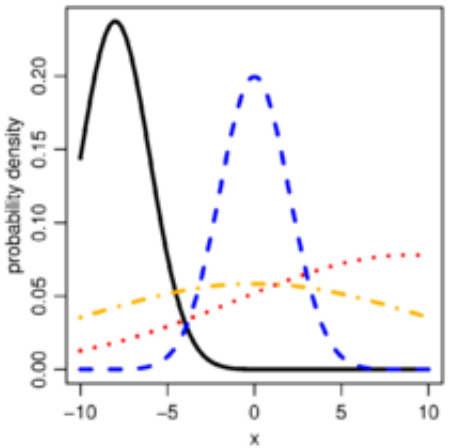
$$K^{(j)}(t, t') = \exp\left(-\lambda_j (t - t')^2\right), \;\; j = 1, \ldots, P,$$

$$\mu_j \sim \text{GP}(0, K^{(j)}), \;\; j = 1, \ldots, P,$$

GP: Gaussian Process (1-D)

$$\mathbf{x}_i \sim \text{MultiNorm}(\boldsymbol{\mu}(t_i), \boldsymbol{\Sigma}), \;\; i = 1, \ldots, N.$$

N: number of data points



Truncated normal distribution

[Campbell & Yau, PLOS Computational Biology, 2016]

Outline of today's lecture

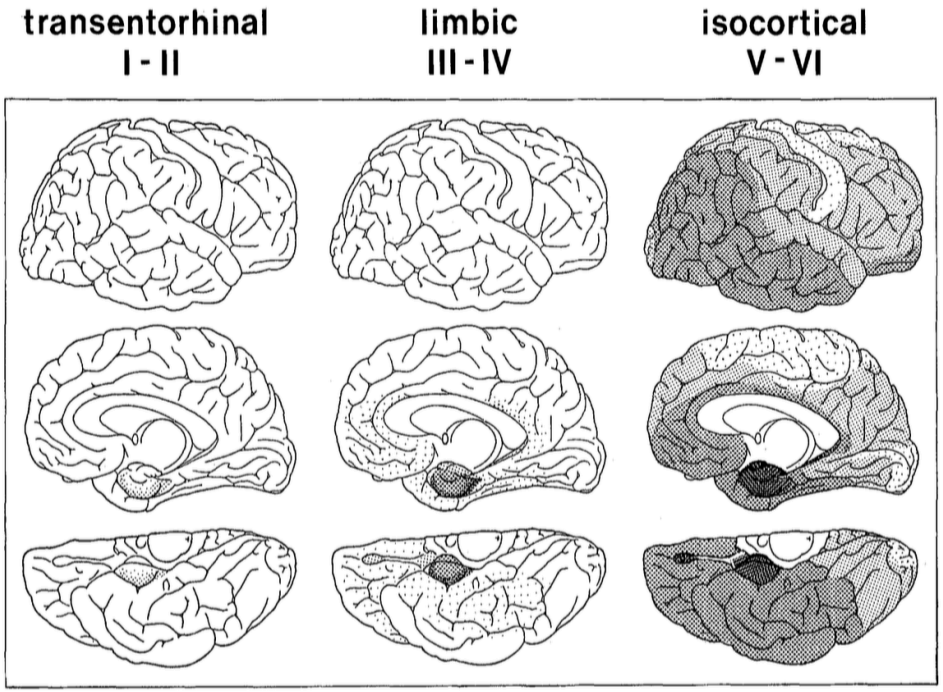
1. Staging from cross-sectional data
   - Wang, Sontag, Wang, *KDD* 2014
   - Pseudo-time methods from computational biology
2. **Simultaneous staging & subtyping**
   - **Young et al., *Nature Communications* 2018**

Acknowledgement: Subsequent slides adapted from Daniel Alexander

Temporal heterogeneity

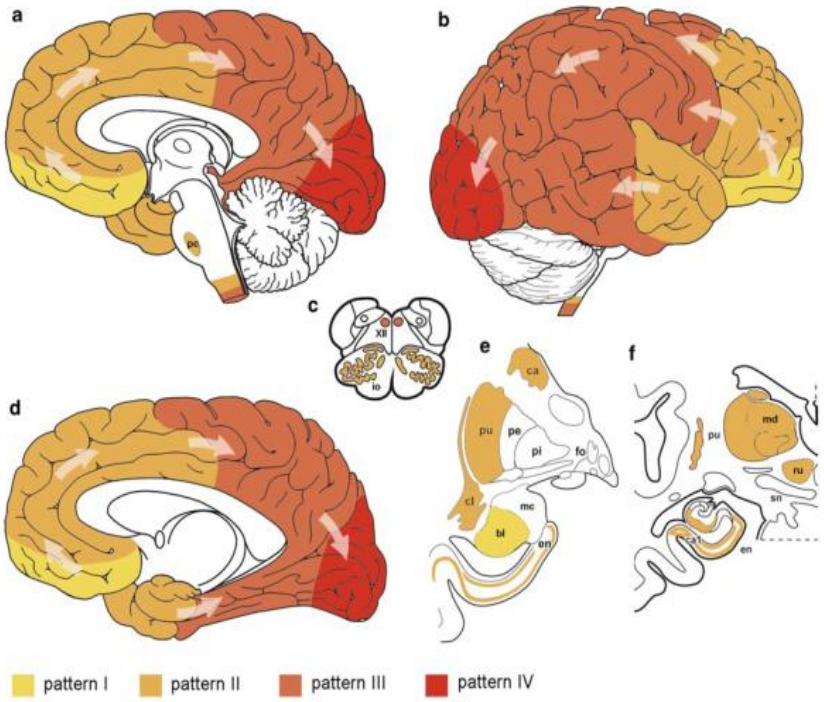
Patients show various disease stages through which patterns of pathology evolve

Alzheimer's disease



Braak and Braak 1991

Frontotemporal dementia

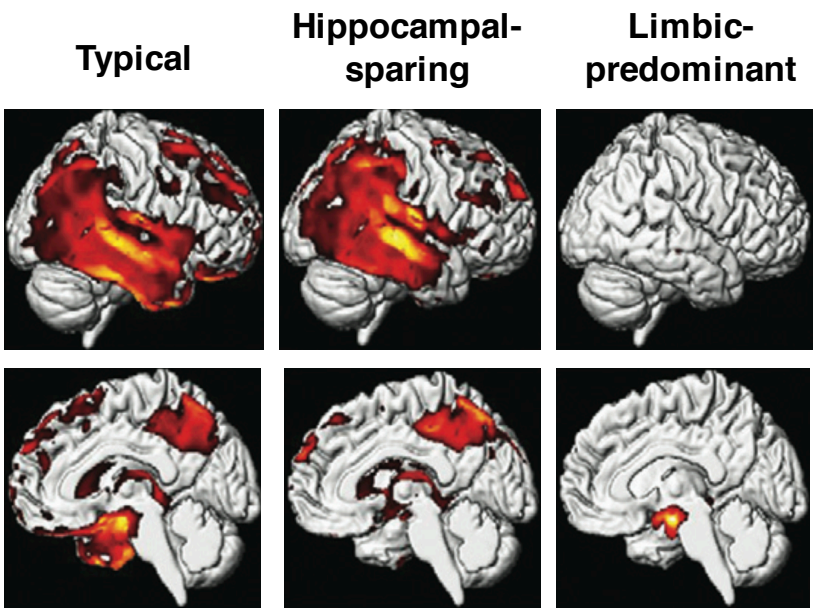


Brettschneider et al. 2014

Phenotypic heterogeneity

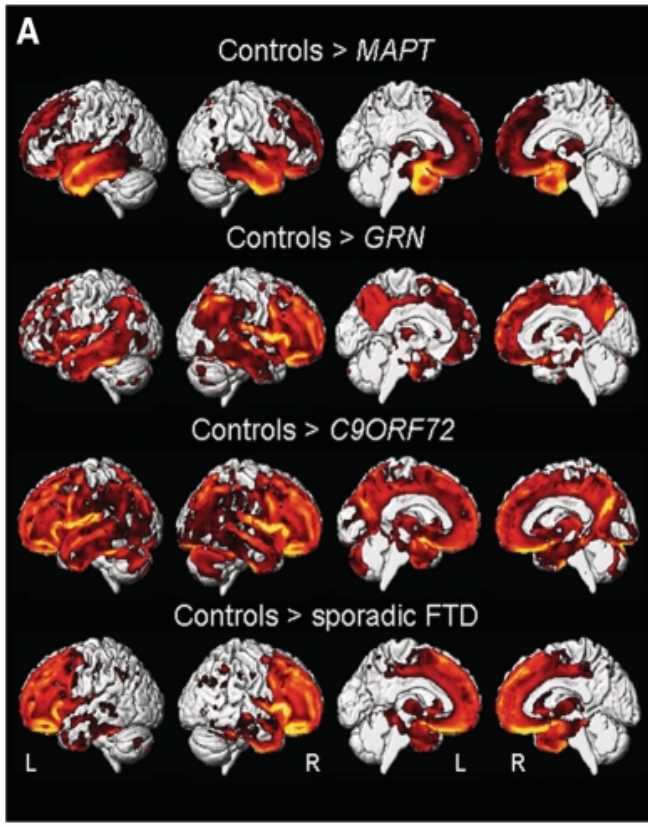
Individuals have different disease subtypes with distinct patterns of pathology

Alzheimer's disease



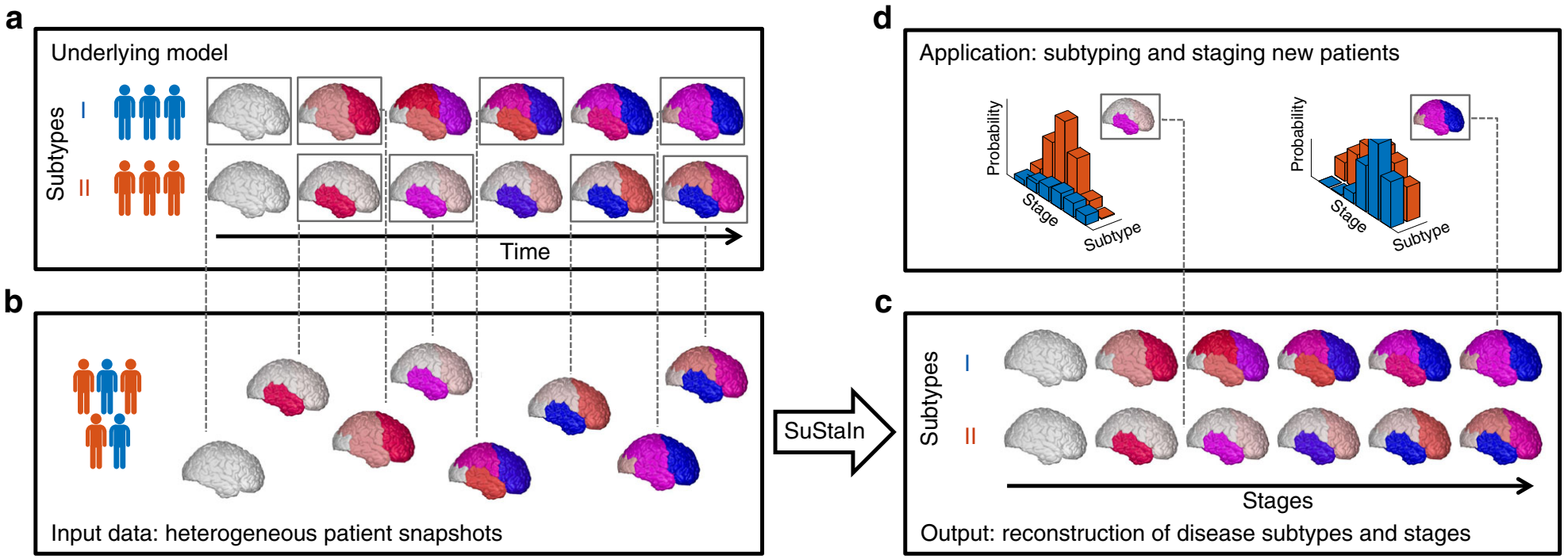
Murray et al. 2011, Whitwell et al. 2012

Frontotemporal dementia



Whitwell et al. 2012

Subtype and Stage Inference (SuStaIn)



[Young et al., *Nature Communications* 2018]

Subtype and Stage Inference (SuStaIn)

- Generative model for a data point:
  - Sample subtype *c* ~ Categorical(f_1, …, f_C)
  - Sample stage *t* ~ Categorical(uniform)
  - For each biomarker *i*, sample $x_i \sim \mathcal{N}(g_{c,i}(t), \sigma_i)$
- Means are enforced to be monotonically increasing and piece-wise linear:

$$g(t) = \begin{cases} \frac{z_1}{t_{E_{z_1}}} t, 0<t \leq t_{E_{z_1}} \\ z_1 + \frac{z_2 - z_1}{t_{E_{z_2}} - t_{E_{z_1}}} \left(t - t_{E_{z_1}}\right), t_{E_{z_1}} < t \leq t_{E_{z_2}} \\ \vdots \\ z_{R-1} + \frac{z_R - z_{R-1}}{t_{E_{z_R}} - t_{E_{z_{R-1}}}} \left(t - t_{E_{z_{R-1}}}\right), t_{E_{z_{R-1}}} < t \leq t_{E_{z_R}} \\ z_R + \frac{z_{max} - z_R}{1 - t_{E_{z_R}}} \left(t - t_{E_{z_R}}\right), t_{E_{z_R}} < t \leq 1 \end{cases}.$$

Shown here for one choice of *c,i* – no parameter sharing across biomarkers or subtypes

[Young et al., *Brain* 2014; Young et al., *Nature Communications* 2018]