MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 11: Clustering to discover disease subtypes and stages

Prof. David Sontag

MIT EECS, CSAIL, IMES

Massachusetts Institute of Technology

Outline of today's class

1. Overview of clustering (k-means algorithm)
   - **Application: discovering asthma subtypes**

2. Overview of latent variable models and Bayesian networks
   - **Application: learning disease progression models**

Clustering

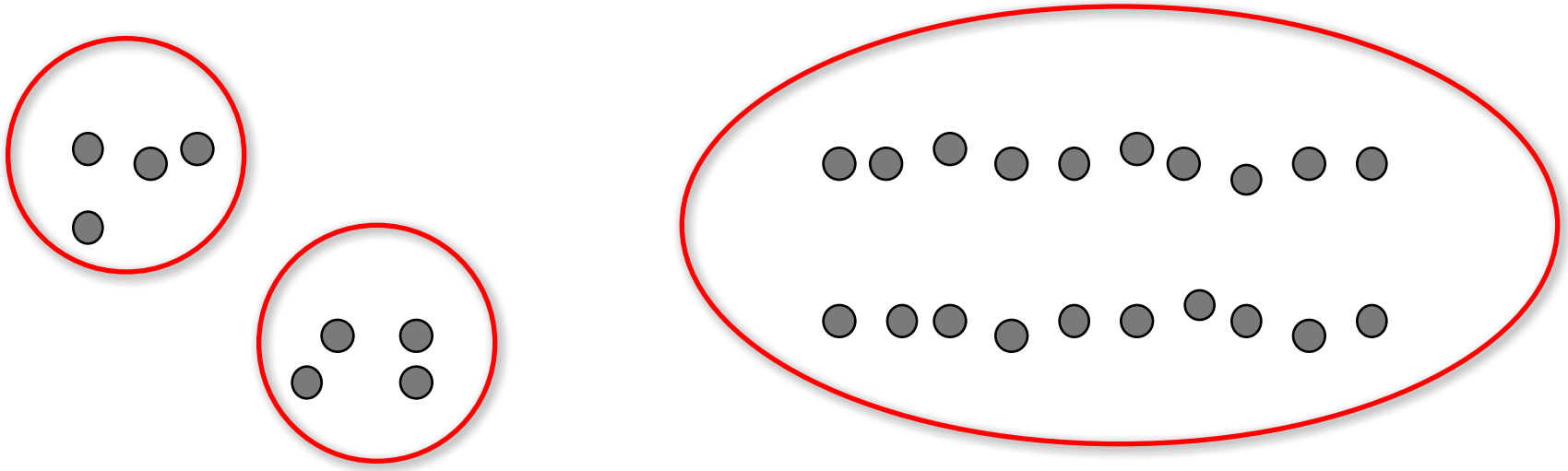
Clustering:

- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish

[This & next few slides adapted from Luke Zettlemoyer, Vibhav Gogate, Carlos Guestrin, Andrew Moore, Dan Klein]

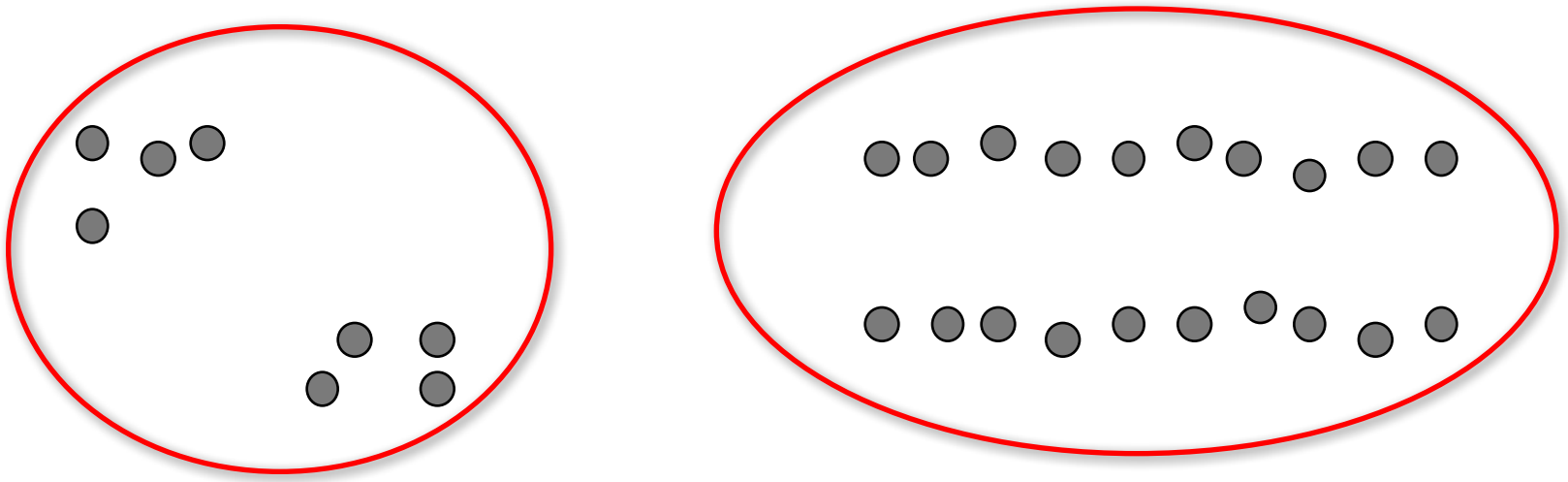
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



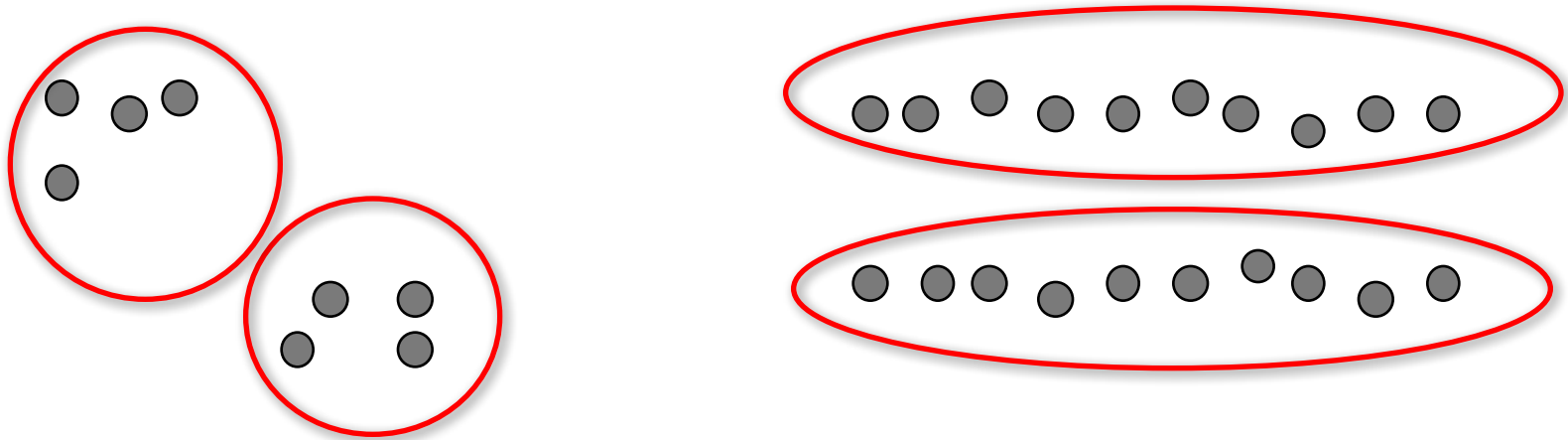
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



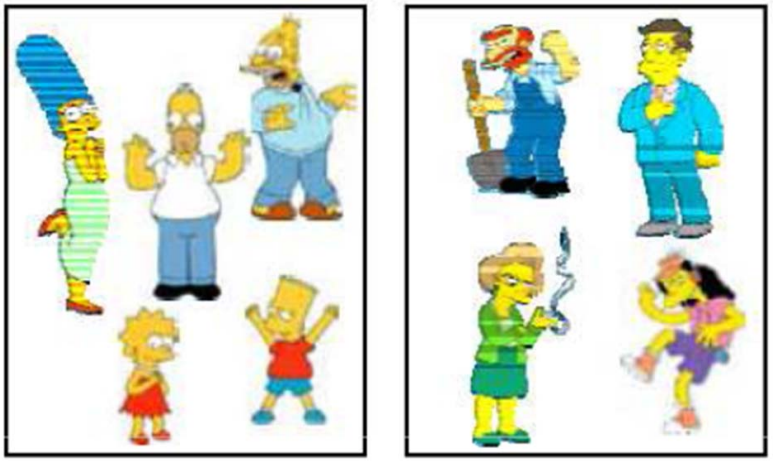
- What could "similar" mean?
  - One option: small Euclidean distance (squared)

$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

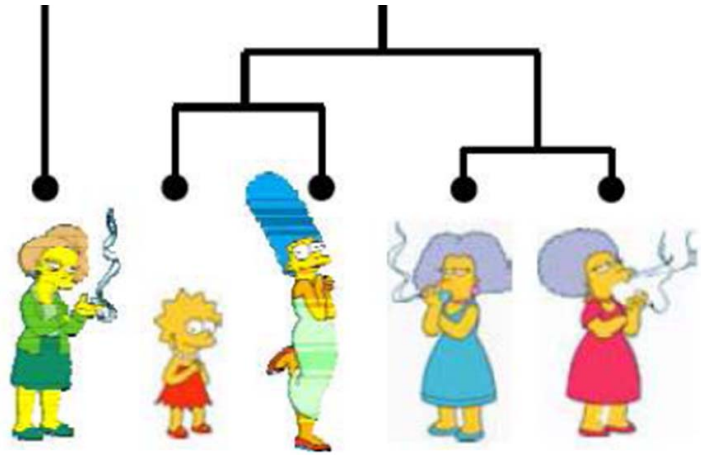
  - Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered

Clustering algorithms

- Partition algorithms (Flat)
  - K-means
  - Mixture of Gaussian
  - Spectral Clustering

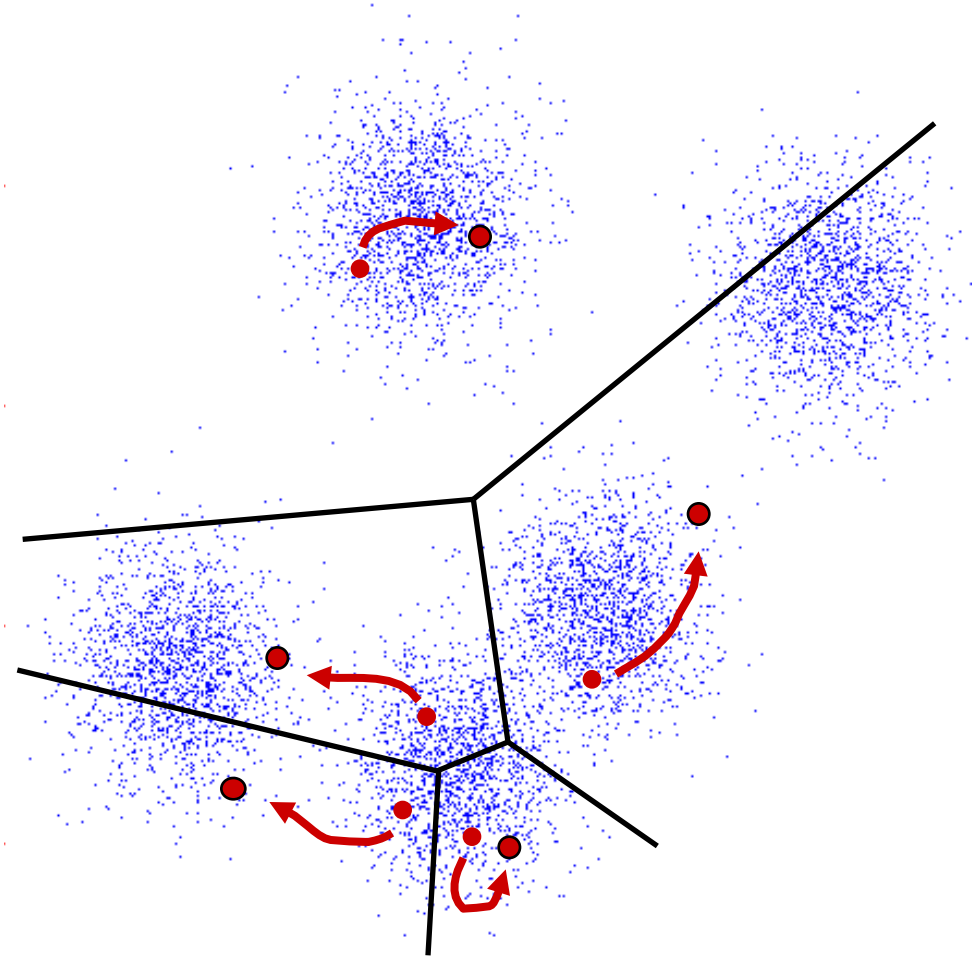


- Hierarchical algorithms
  - Bottom up – agglomerative
  - Top down – divisive

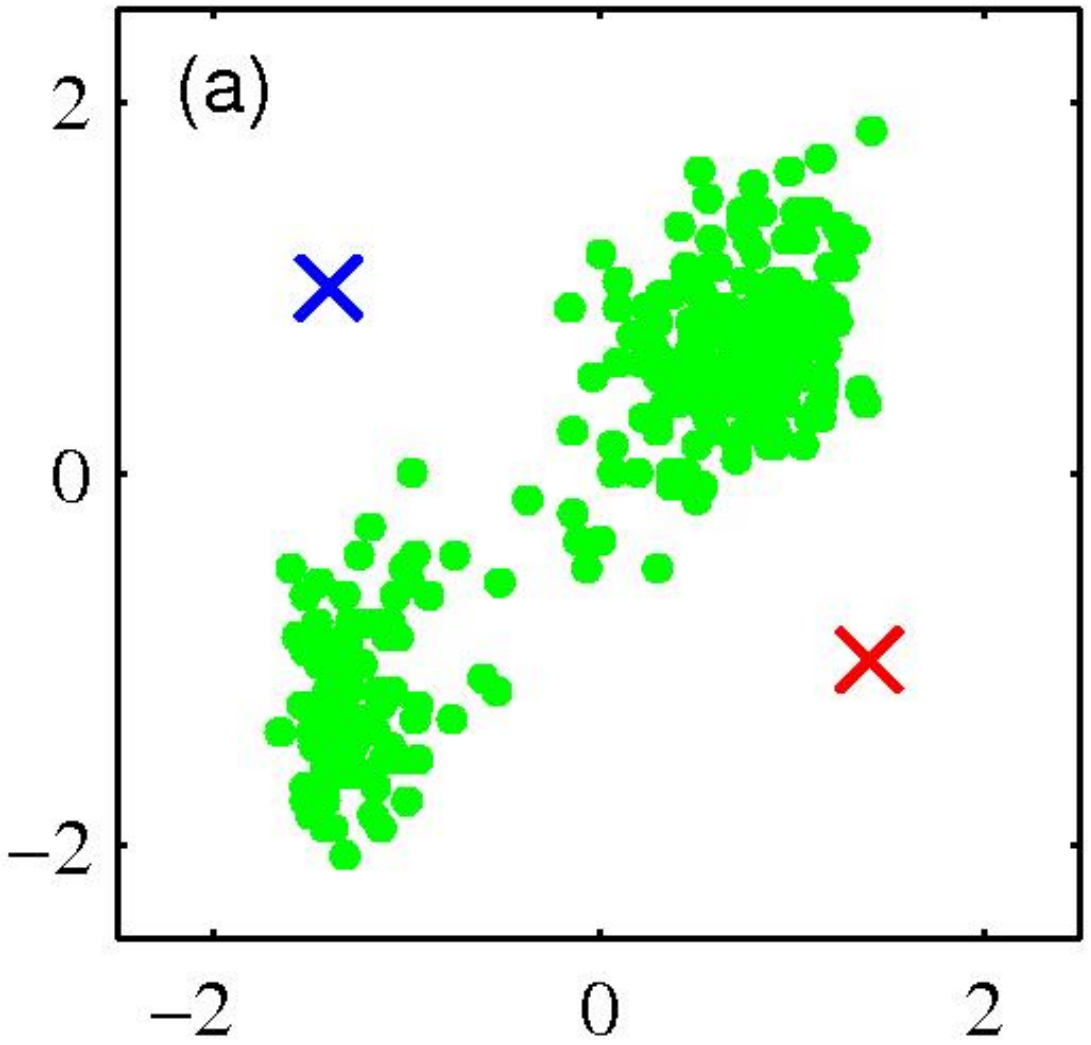


K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



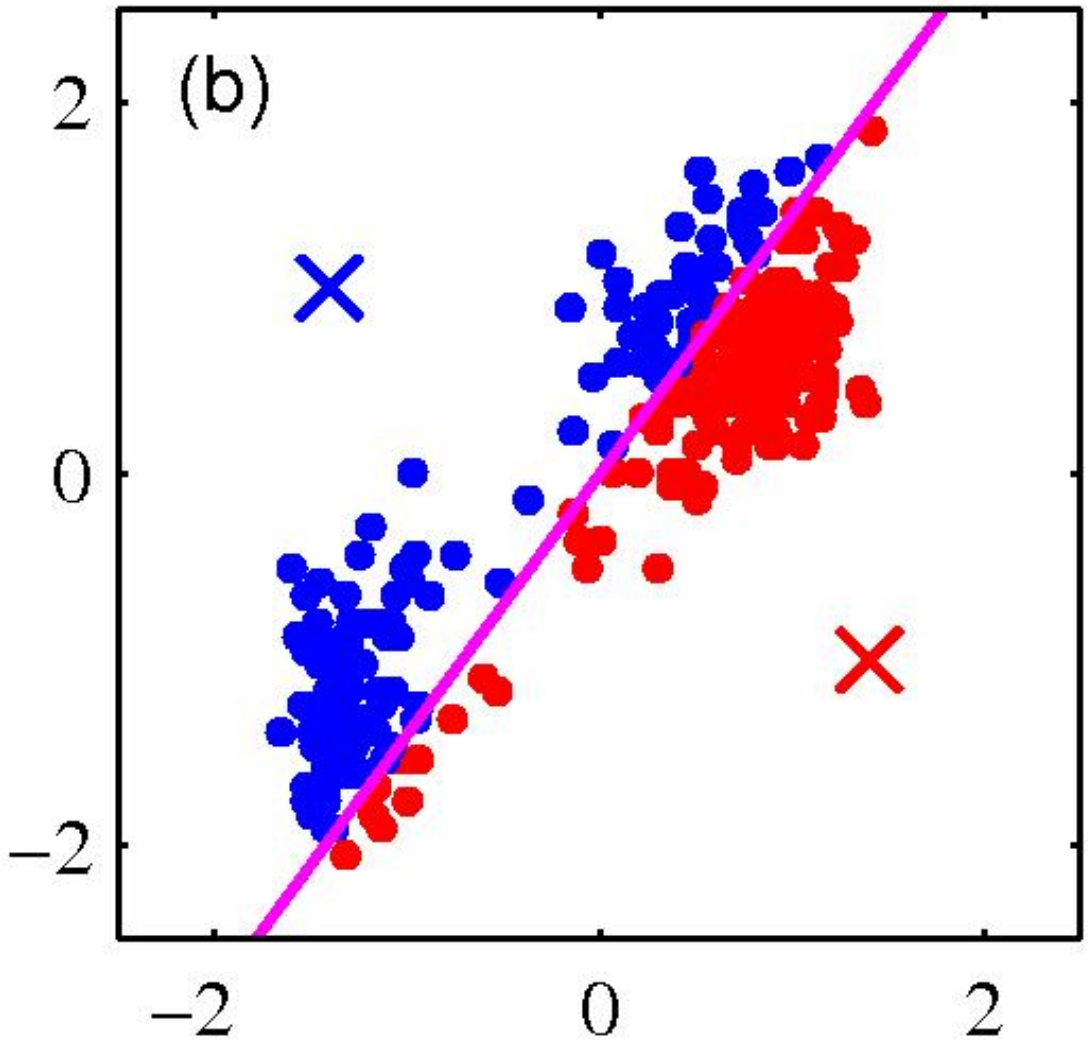
K-means clustering: Example



- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

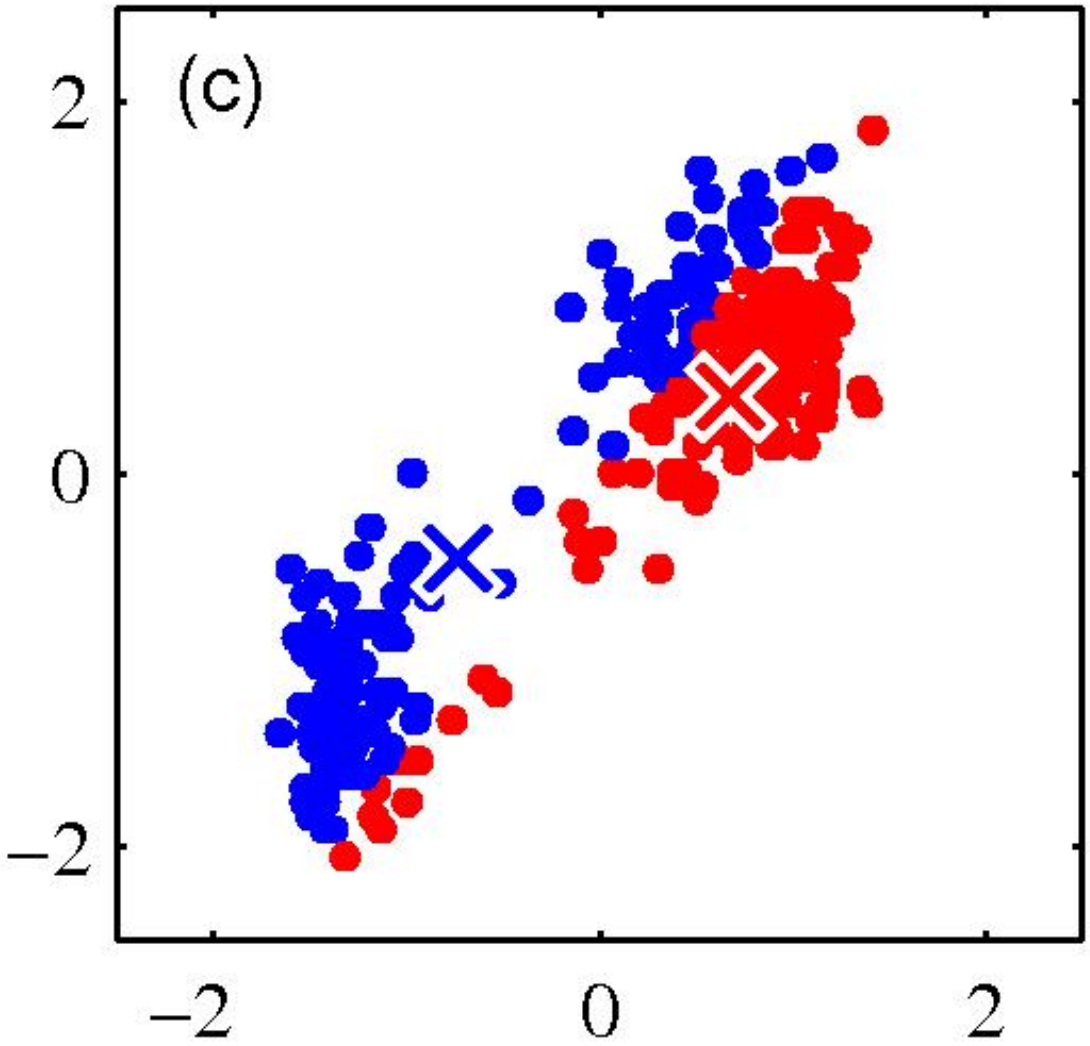
K-means clustering: Example



Iterative Step 1

- Assign data points to closest cluster center

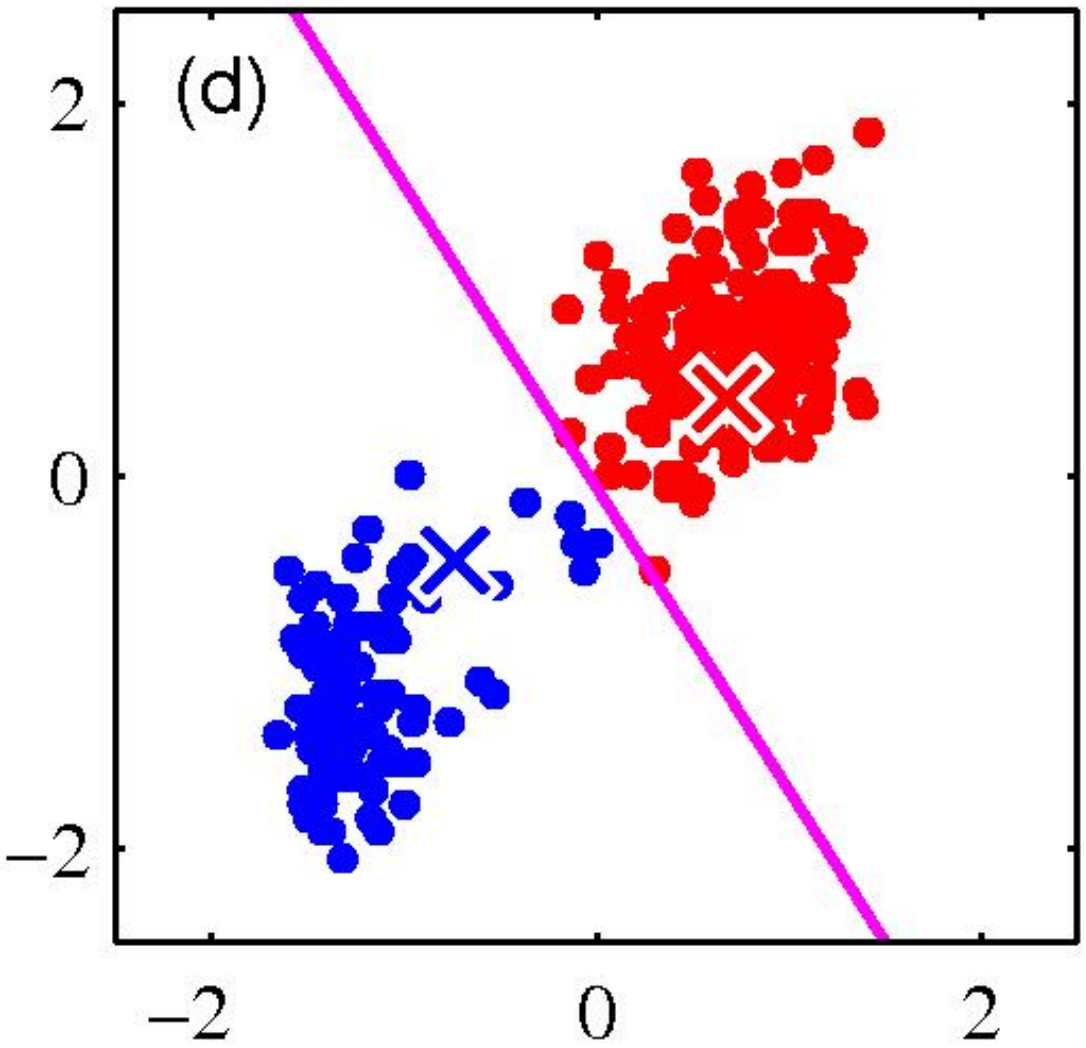
K-means clustering: Example



Iterative Step 2

- Change the cluster center to the average of the assigned points

K-means clustering: Example



- Repeat until convergence

Asthma: the problem

- 5 to 10% of people with severe asthma remain poorly controlled despite maximal inhaled therapy

[Holgate ST, Polosa R. The mechanisms, diagnosis, and management of severe asthma in adults. Lancet. 2006; 368:780–793]

[whatasthmais.com]

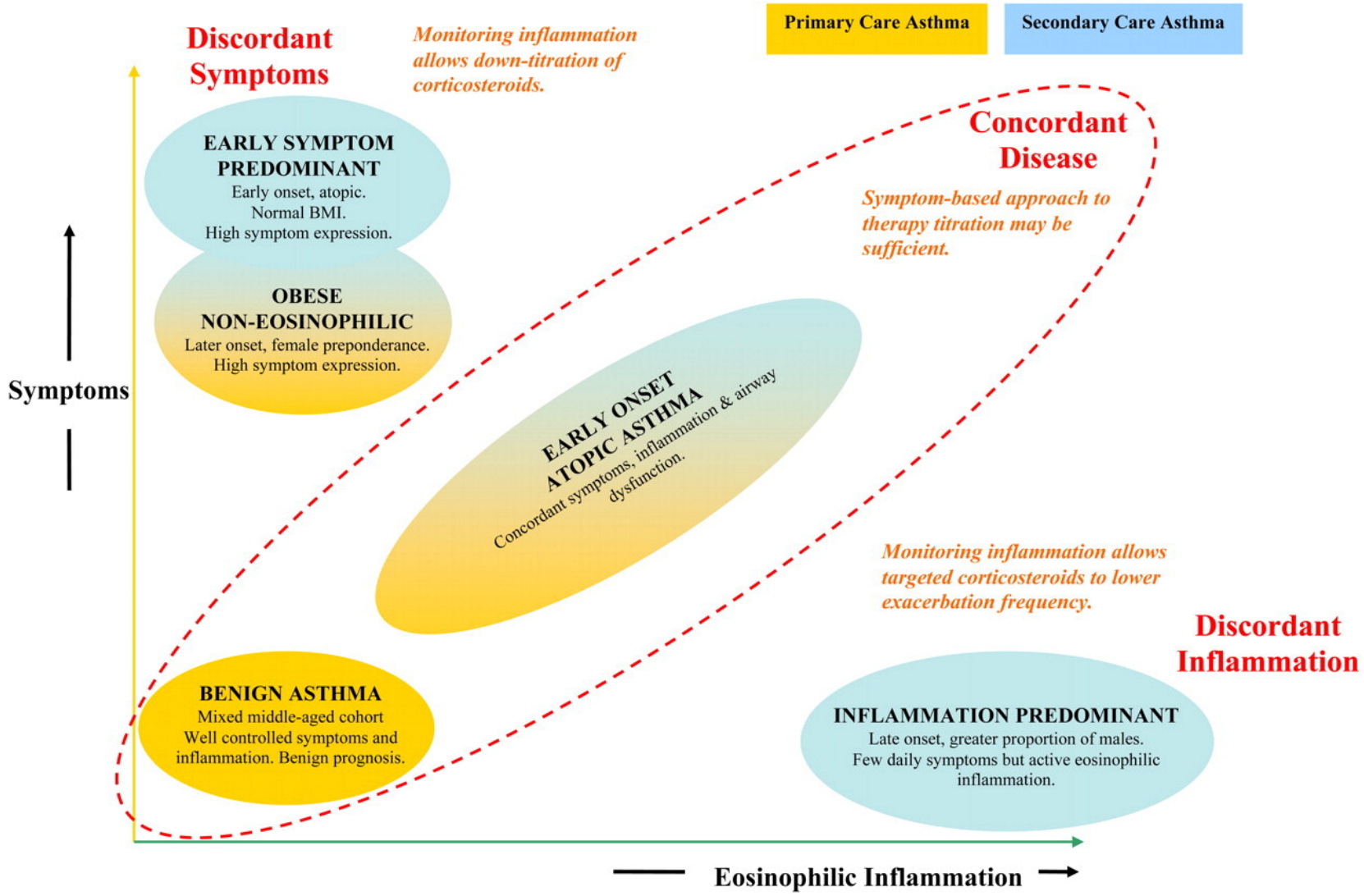
Asthma: the question

"It is now recognised that there are distinct asthma phenotypes and that distinct therapeutic approaches may only impinge on some aspects of the disease process within each subgroup"

- What are the processes (genetic or environmental) that underlie different subtypes of asthma?
- Which aspects of airway remodelling are important in disease subtypes?
- What are the best biomarkers of disease progression or treatment response?
- Why are some patients less responsive to conventional therapies than others?

[Adcock et al., "New targets for drug development in asthma". The Lancet, 2008]

Discovering subtypes from data



[Haldar et al., *Am J Respir Crit Care Med*, 2008]

The data

- All patients had physician diagnosis of asthma and one prescription for asthma therapy
- All were current nonsmokers
- *Data set #1*: 184 patients recruited from primary-care practices in the UK
- *Data set #2*: 187 patients from refractory asthma clinic in the UK
- *Data set #3*: 68 patients from 12 month clinical study
- Features: *z* scores for continuous variables, 0/1 for categorical

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Comparison of Baseline Characteristics in the three Asthma Populations

Variable	Primary Care (*n = 184*)	Secondary Care (*n = 187*)	Longitudinal Cohort (*n = 68*)
Sex, % female	54.4	65.8	47.1
Age, yr (SD)	49.2 (13.9)	43.4 (15.9)	52.4 (14.6)
Age of onset, yr (SD)	24.7 (19)	20.3 (18.4)	31.1 (23.7)
Atopic status, % positive	72.8	73.8	57.4
Body mass index, kg/m^2 (SD)	27.5 (5.4)	28.5 (6.5)	28.0 (5.9)
PC_{20} methacholine†, mg/ml	1.04 (1.13)	†	0.67 (0.68)
Peak flow variability, amp % mean	17 (0.38)	32.2 (0.48)	13.8 (0.29)
FEV_1 change with bronchodilator, %	1.63 (1.16)	12.8 (0.41)	3.2 (1.04)
Post-bronchodilator FEV_1, % predicted	91.4 (21)	82.1 (21.1)	80.2 (20.6)
Sputum eosinophil count, %	1.32 (0.62)	2.9 (0.99)	2.4 (0.81)
F_{ENO}‡, ppb	31.6 (0.33)	43 (0.32)	4.32 (0.64)‡
Sputum neutrophil count, %	55.09 (0.31)	46.7 (0.32)	41.1 (0.35)
Modified JACS§ (SD)	1.36 (0.74)	2.02 (1.16)	1.42 (1.26)
Dose of inhaled corticosteroid, BDP equivalent/μg (SD)	632 (579)	1,018 (539)	1,821 (1,239)
Long-acting bronchodilator use, %	40.2	93	86.7

Definition of abbreviations: amp = amplitude; BDP = beclomethasone dipropionate; JACS = Juniper Asthma Control Score

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Clusters in primary care

Variable	Primary Care (*n* = *184*)	Cluster 1 Early-Onset Atopic Asthma (*n* = *61*)	Cluster 2 Obese Noneosinophilic (*n* = 27)	Cluster 3 Benign Asthma (*n* = 96)	Significance (*P* Value)*
Sex†, % female	**54.4**	45.9	81.5	52.1	0.006
Age, yr (SD)	**49.2 (13.9)**	44.5 (14.3)	53.9 (14)	50.8 (13)	0.003
Age of onset†, yr (SD)	**24.7 (19)**	14.6 (15.4)	35.3 (19.6)	28.2 (18.3)	<0.001
Atopic status†, % positive	**72.8**	95.1	51.9	64.6	<0.001
Body mass index†, kg/m^2 (SD)	**27.5 (5.4)**	26.1 (3.8)	36.2 (5.5)	26 (3.6)	<0.001
PC_{20} methacholine†‡, mg/ml	**1.04 (1.13)**	0.12 (0.86)	1.60 (0.93)	6.39 (0.75)	<0.001
PC_{20} >8 mg/ml, n (%)	**64 (34.7)**	2 (3.3)	6 (22.2)	56 (58.3)	<0.001
Peak flow variability†‡, amp % mean	**17 (0.38)**	20 (0.47)	21.9 (0.32)	14.8 (0.32)	0.039
FEV_1 change with bronchodilator‡, %	**1.63 (1.16)**	4.5 (0.91)	1.82 (1.16)	0.83 (1.22)	<0.001
Post-bronchodilator FEV_1, % predicted	**91.4 (21)**	86.9 (20.7)	91.5 (21.4)	94.2 (20.7)	0.107
Sputum eosinophil count†‡, %	**1.32 (0.62)**	3.75 (0.64)	1.55 (0.51)	0.65 (0.44)	<0.001
F_{ENO}‡§, ppb	**31.6 (0.33)**	57.5 (0.27)	25.8 (0.29)	22.8 (0.27)	<0.001
Sputum neutrophil count‡, %	**55.09 (0.31)**	45.87 (0.24)	72.71 (0.13)	57.56 (0.36)	0.038
Modified JACS† (SD)	**1.36 (0.74)**	1.54 (0.58)	2.06 (0.73)	1.04 (0.66)	<0.001
Dose of inhaled corticosteroid, BDP equivalent/μg (SD)	**632 (579)**	548 (559)	746 (611)	653 (581)	0.202
Long-acting bronchodilator use, %	**40.2**	34.4	48.2	41.7	0.442
Previous hospital admission or emergency attendance, no. per patient	**0.60 (1.57)**	1.04	0.26	0.20	0.037
Previous outpatient attendance, % attended	**15%**	22%	19%	6%	0.121
Severe asthma exacerbations (requiring oral corticosteroids) in past 12 mo, no. per patient	**1.25 (1.94)**	1.86 (0.32)	1.07 (0.32)	0.39 (0.18)	0.002

Clusters in secondary care

Variable	Secondary Care (*n = 187*)	Cluster 1 Early Onset, Atopic (*n = 74*)	Cluster 2 Obese, Noneosinophilic (*n = 23*)	Cluster 3 Early Symptom Predominant (*n = 22*)	Cluster 4 Inflammation Predominant (*n = 68*)	Significance (*P* Value)*
Sex†, % female	**65.8**	75.7	87	68.2	47.1	<0.001
Age, yr (SD)	**43.4 (15.9)**	39.4 (15.7)	42.7 (11.1)	35.5 (15.5)	50.6 (15.1)	<0.001
Age of onset†, yr (SD)	**20.3 (18.4)**	12.7 (12.9)	15.4 (15.2)	12.6 (15)	32.6 (19.1)	<0.001
Atopic status†, % positive	**73.8**	83.8	65.2	81.8	63.2	0.024
Body mass index†, kg/m^2 (SD)	**28.5 (6.5)**	27.6 (4.5)	40.9 (6.5)	23.6 (3.1)	27 (3.9)	<0.001
Peak flow variability‡, amp % mean	**32.2 (0.48)**	46.1 (0.35)	21.2 (0.76)	24.2 (0.65)	27.6 (0.36)	0.002
FEV_1 change with bronchodilator‡, %	**12.8 (0.41)**	24.5 (0.31)	9.3 (0.35)	4.5 (0.33)	9.8 (0.34)	<0.001
Post-bronchodilator FEV_1, % predicted (SD)	**82.1 (21.1)**	79.0 (21.9)	79.0 (18.5)	79.5 (26.1)	87.2 (18.5)	0.093
Sputum eosinophil count†‡, %	**2.9 (0.99)**	4.2 (0.76)	1.3 (1.01)	0.1 (0.9)	8.4 (0.64)	<0.001
F_{ENO}‡§, ppb	**43 (0.32)**	51.2 (0.36)	24.2 (0.27)	22.6 (0.30)	53.1 (0.32)	<0.001
Sputum neutrophil count, %‡	**46.7 (0.32)**	45.4 (0.39)	49.3 (0.22)	51.3 (0.23)	45.9 (0.29)	0.892
Modified JACS† (SD)	**2.02 (1.16)**	2.63 (0.93)	2.37 (1.09)	2.11 (1.11)	1.21 (0.95)	<0.001
Dose of inhaled corticosteroid, BDP equivalent/μg (SD)	**1,018 (539)**	1,168 (578)	1,045 (590)	809 (396)	914 (479)	0.008
Long-acting bronchodilator use, %	**93.0**	91.9	95.4	90.9	94.1	0.999

Patients in different clusters respond differently to treatment!

Cluster (found using *baseline* data)	Outcomes	Treatment strategy: Clinical (*n* = *10*)	Treatment strategy: Sputum (*n* = *8*)	Significance
1: Obese female	Δ Inhaled corticosteroid dose*/μg per day (SEM)	−400 (328)	−462 (271)	0.89
	Severe exacerbation frequency over 12 mo (SEM)	1.40 (0.78)	1.50 (0.80)	0.93
	Number commenced on oral corticosteroids	2	1	0.59
		Clinical (*n* = *15*)	Sputum (*n* = *24*)	
2: Inflammation predominant	Δ Inhaled corticosteroid dose*/μg per day (SEM)	+753 (334)	+241 (233)	0.22
	Severe exacerbation frequency over 12 mo (SEM)	3.53 (1.18)	0.38 (0.13)	0.002
	Number commenced on oral corticosteroids	2	9	0.17
		Clinical (*n* = *7*)	Sputum (*n* = *4*)	
3: Early symptom predominant	Δ Inhaled corticosteroid dose*/μg per day (SEM)	+1,429 (429)	−400 (469)	0.022
	Severe exacerbation frequency over 12 mo (SEM)	5.43 (1.90)	2.50 (0.87)	0.198
	Number commenced on oral corticosteroids	6	0	Undefined

[Haldar et al., *Am J Respir Crit Care Med*, 2008]

Outline of today's class

1. Overview of clustering (k-means algorithm)
   - **Application: discovering asthma subtypes**

2. Overview of latent variable models and Bayesian networks
   - **Application: learning disease progression models**

Bayesian networks

- A **Bayesian network** is specified by a directed *acyclic* graph G=(V,E) with:
  - One node *i* for each random variable X_i
  - One conditional probability distribution (CPD) per node, $p(x_i \mid \boldsymbol{x}_{Pa(i)})$, specifying the variable's probability conditioned on its parents' values

- Corresponds 1-1 with a particular factorization of the joint distribution:

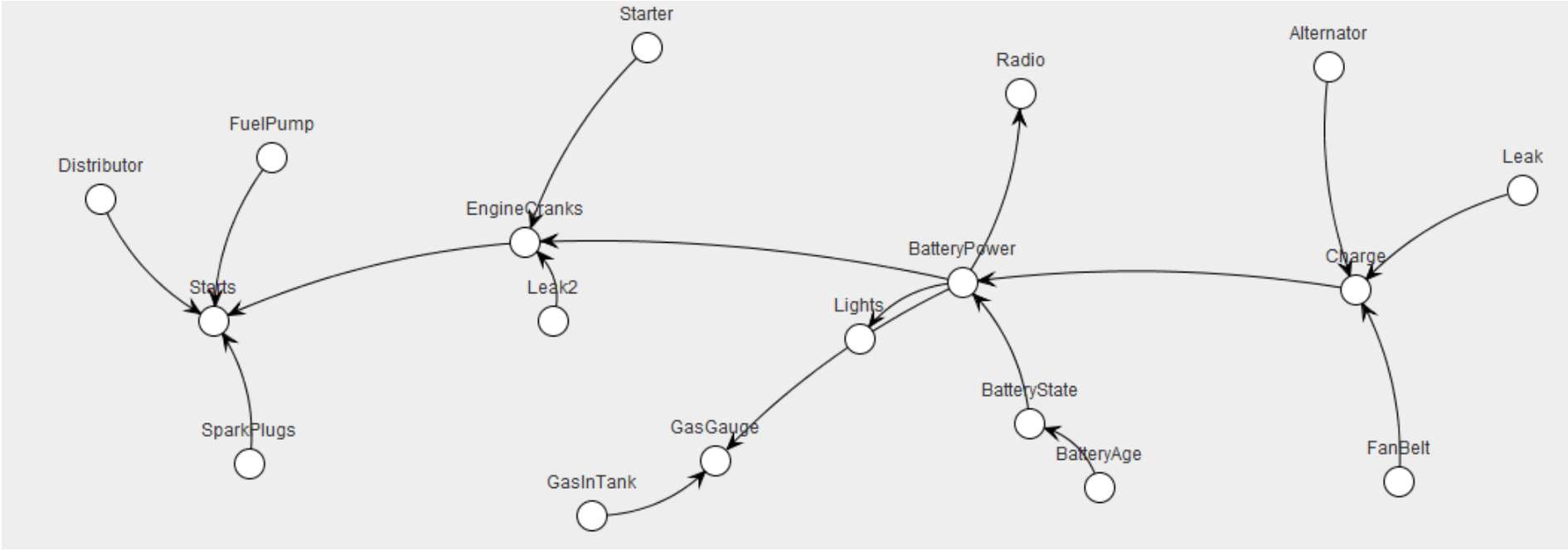
$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations

Bayesian networks enable use of domain knowledge

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Will my car start this morning?



Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

Bayesian networks enable use of domain knowledge

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

What is the differential diagnosis?

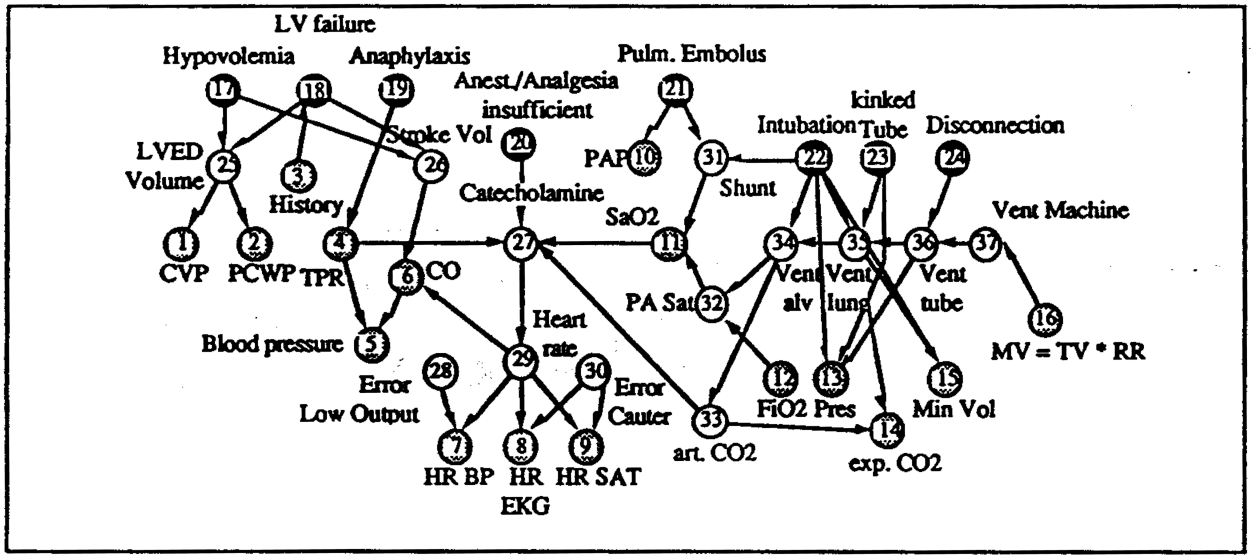
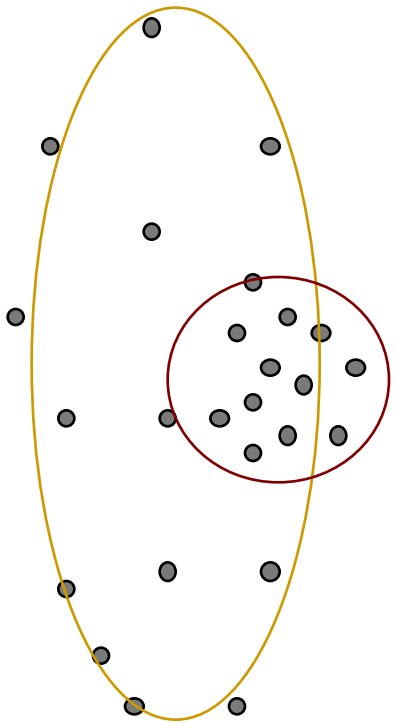


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◎) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

Beinlich *et al.*, The ALARM Monitoring System, 1989

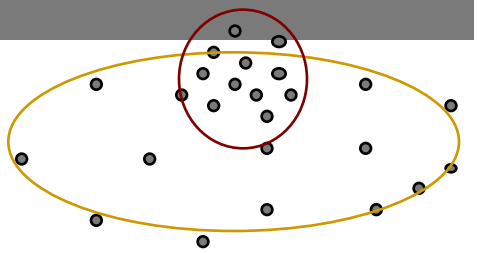
Returning to clustering example…



- Clusters may overlap
- Some clusters may be "wider" than others
- Can we model this explicitly?
- With what **probability** is a point from a cluster?

[Next few slides adapted from Carlos Guestrin, Dan Klein, Luke Zettlemoyer, Dan Weld, Vibhav Gogate, and Andrew Moore]

Clustering as latent variable model



- Try a probabilistic model!
  - allows overlaps, clusters of different size, etc.
- Can tell a *generative story* for data
  - P(Y)P(X|Y)
- Challenge: we need to estimate model parameters without labeled Ys

Y	X_1	X_2
??	0.1	2.1
??	0.5	-1.1
??	0.0	3.0
??	-0.1	-2.0
??	0.2	1.5
…	…	…

Gaussian Mixture Models

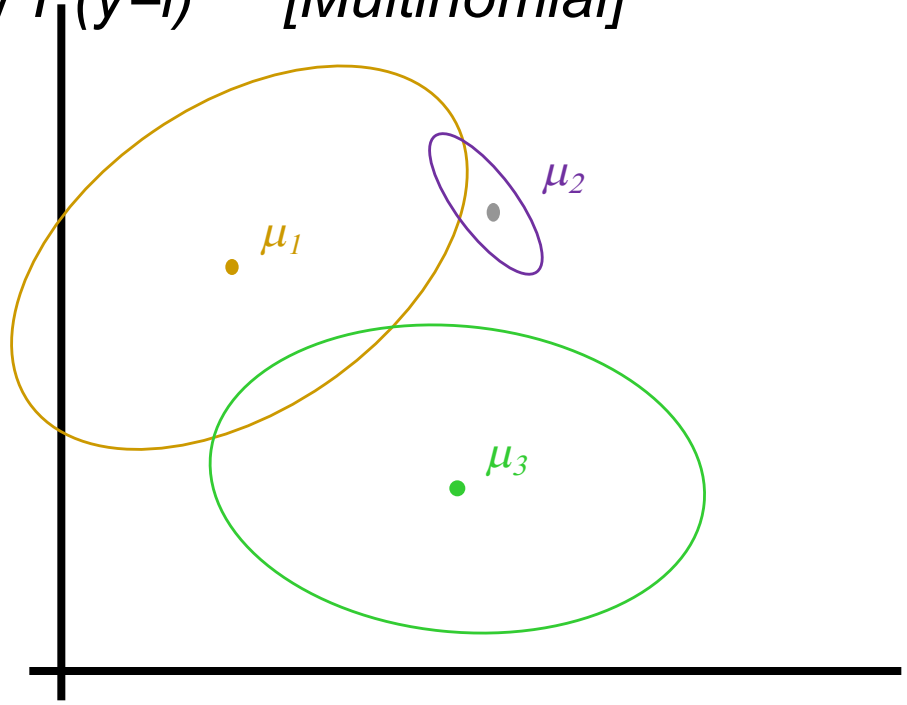
- P(Y): There are k components
- P(X|Y): Each component generates data from a **multivariate Gaussian** with mean μ_i and covariance matrix Σ_i

Each data point assumed to have been sampled from a ***generative process***:

1. Choose component i with probability *P(y=i)* *[Multinomial]*
2. Generate datapoint ~ N(m_i, Σ_i)

$$P(X = \mathbf{x}_j \mid Y = i) = \frac{1}{(2\pi)^{m/2} \|\Sigma_i\|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}_j - \mu_i\right)^T \Sigma_i^{-1}\left(\mathbf{x}_j - \mu_i\right)\right]$$

By fitting this model (unsupervised learning), we can learn new insights about the data



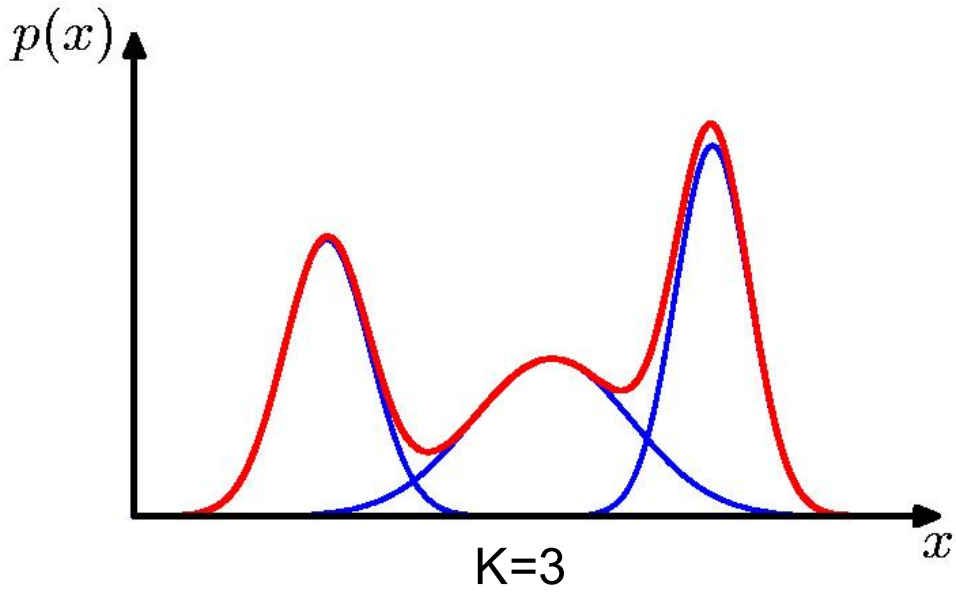
Marginal likelihood for mixture of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



Unsupervised learning is computationally challenging

- Maximize ***marginal likelihood***:
  - $\text{argmax}_\theta \prod_j P(x_j) = \text{argmax} \prod_j \sum_{k=1} P(Y_j=k, x_j)$
- Almost always a hard problem!
  - Usually no closed form solution
  - Even when lgP(X,Y) is convex, lgP(X) generally isn't…
  - Many local optima
- Common approaches are gradient ascent and expectation maximization (EM) – both will just reach a local optima

The burden of chronic disease

- Chronic disease is a global burden
  - Hundreds of millions of people
  - Trillions of dollars spent
  - Loss in life expectancy
  - Loss in quality of life

- **Example:** Chronic Obstructive Pulmonary Disease (COPD)
  - Impacts low-income population
  - Key risk factors: smoking and air pollution
  - Causes systemic illness

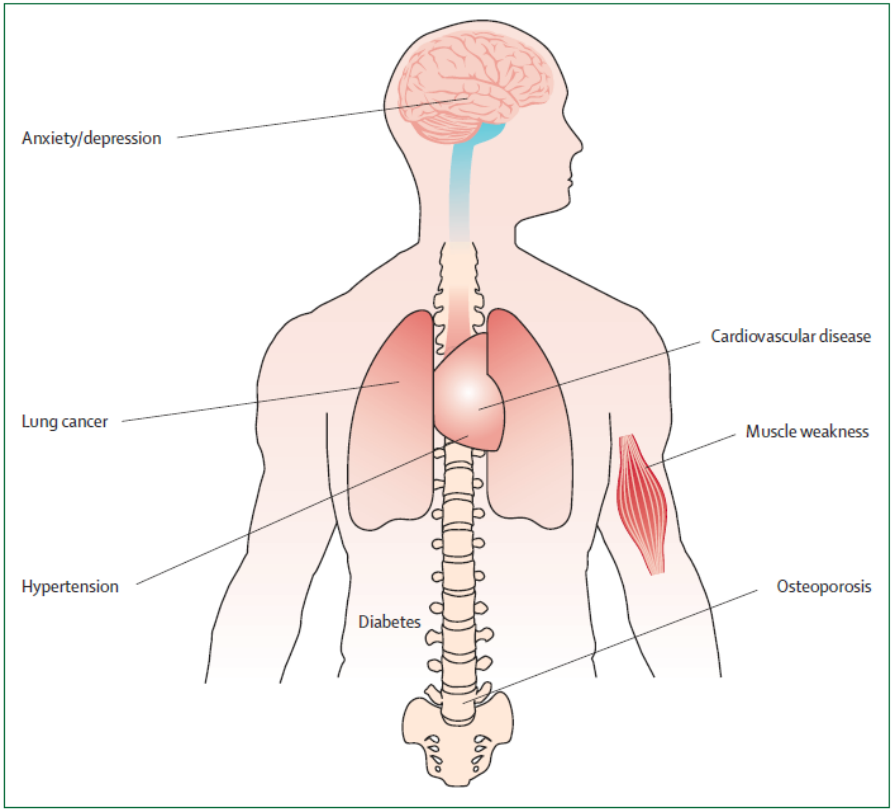


Figure 4: Comorbidities of chronic obstructive pulmonary disease

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

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]**

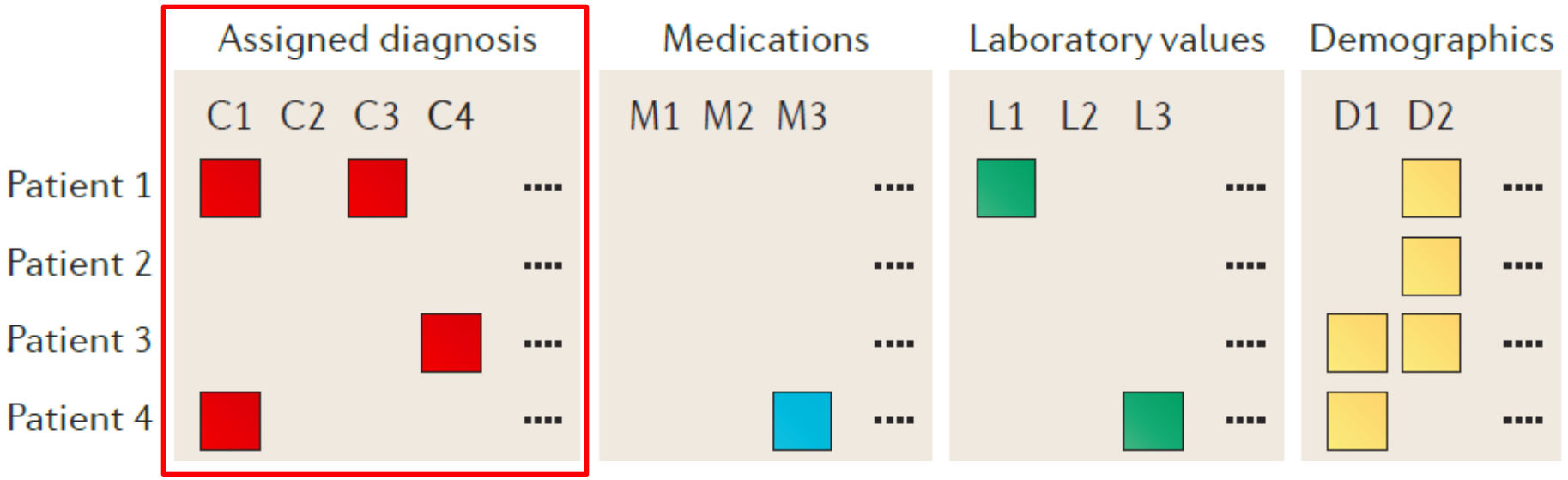
Chronic obstructive pulmonary disease. The Lancet, Volume 379, Issue 9823, Pages 1341 - 1351, 7 April 2012

Unsupervised learning of disease progression models

- Algorithm to learn a disease progression model from EHR data
  - Generative model
- We demonstrate its use in
  - Deriving a meaningful characterization of disease progression and stages
  - Identifying the progression trajectory of individual patients
- More broadly, these models will be used to
  - Provide decision support for early intervention
  - Develop data-driven guidelines for care plan management
  - Align patients across time, by disease stage, to enable comparative effectiveness research (e.g., of medications)

[Wang, Sontag, Wang, KDD 2014]

Goal: Learn from Electronic Health Records (EHR)



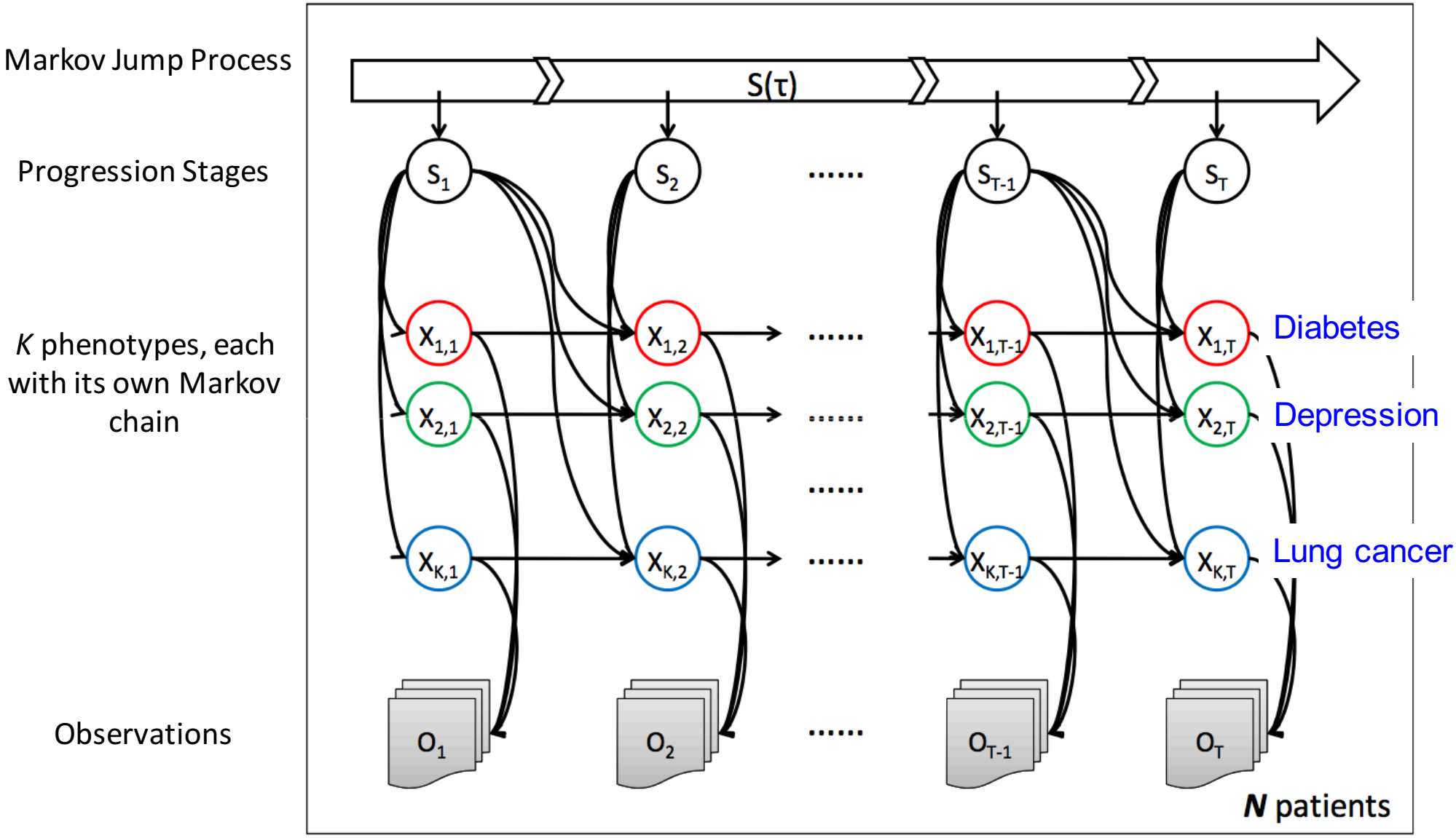
PID	DAY_ID	CLINICAL_EVENT	ICD9_LONGNAME
000000	74053	305.1	Tobacco Use Disorder
000000	74053	496	Chronic Airway Obstruction, Not Elsewhere Classified
000000	74053	733	Osteoporosis, Unspecified
000000	74053	724.2	Lumbago
000000	74091	733	Osteoporosis, Unspecified
000000	74148	733	Osteoporosis, Unspecified
000000	74148	782.3	Edema
000000	74148	780.79	Other Malaise And Fatigue

Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 May 2;13(6):395-405.

Challenges of disease progression modeling from EHRs

- Multiple covariates
- Progression heterogeneity
  - No natural alignment between records with varied progression rates
- Missing data
  - Doctors only document the relevant clinical context
- Incomplete records
  - Might only be 3-6 years of data available for any one person
- Irregular visits
  - Continuous-time model is needed
- Limited supervision
  - No ground truth regarding the current stage of progression

The big picture: generative model for patient data



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