Data Mining

CS 5140 / CS 6140

Jeff M. Phillips

January 5, 2020

What is Data Mining?

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

- How to think about data analytics.

What is Data Mining?

- Finding structure in data?
- Machine learning on large data?
- Unsupervised learning?
- Large scale computational statistics?

- How to think about data analytics.

- *Principals* of converting from messy raw data to abstract representations.
- Algorithms of how to analyze data in abstract representations.
- Addressing challenges in scalability, error, and modeling.

Modeling versus Efficiency

Two Intertwined (and often competing) Objectives:

- Model Data Correctly
- Process Data Efficiently

Other Data Mining Courses

Every university teaches data mining differently!

Other Data Mining Courses

Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat

Other Data Mining Courses

Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat
- *no specific software tools / programming languages*

Other Data Mining Courses

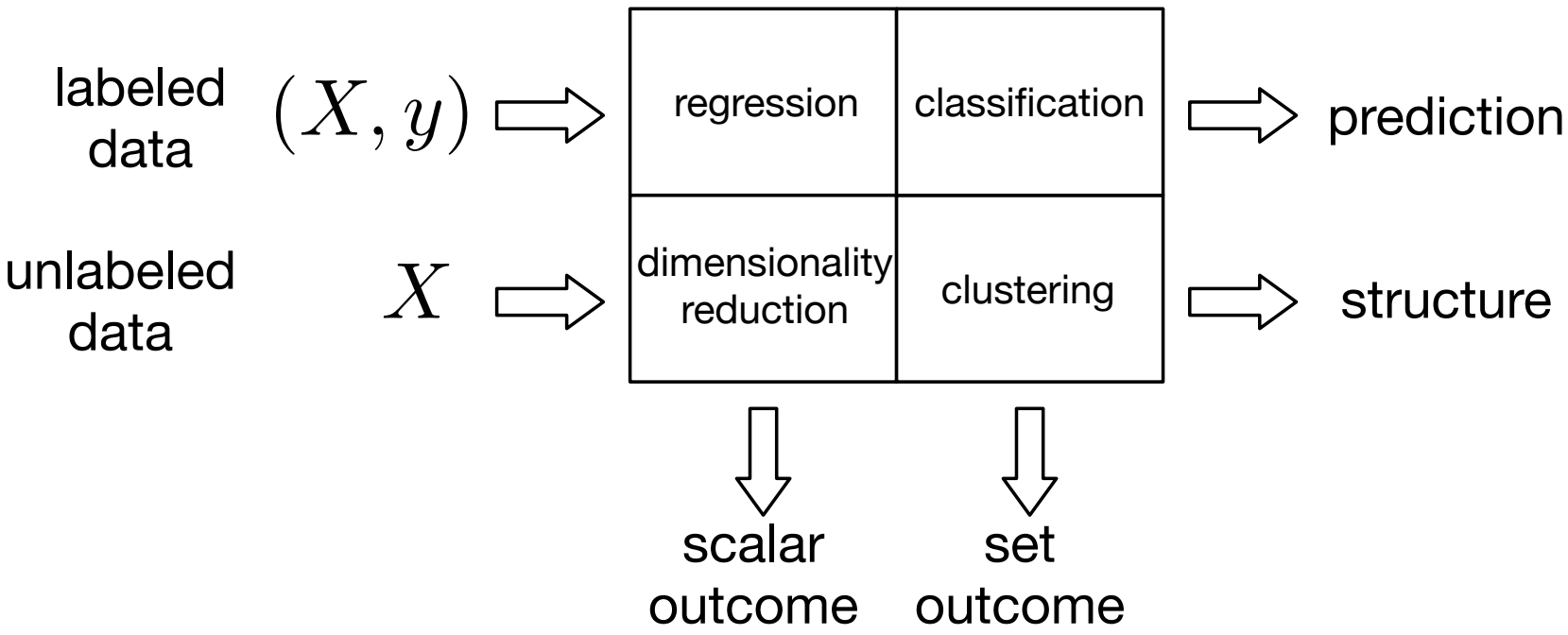
Every university teaches data mining differently!

What flavor is offered in this class:

- Focus on techniques for *very* large scale data
- Broad coverage ... with recent developments
- Formally and generally presented (proof sketches)
- ... but useful in practice (e.g. internet companies)
- Probabilistic algorithms: connections to CS and Stat
- *no specific software tools / programming languages*

Maths: Linear Algebra, Probability, High-dimensional geometry

Classic (Old) View of Data Mining



Outline

Statistical and Mathematical Principals:

- 1. **Hashing, Concentration of Measure**
- 2. **Similarity** (find duplicates and similar items)

Structure in Data:

- 3. **Clustering** (aggregate close items)
- 4. **Regression** (linearity of high-d data, sparsity)
- 5. **Dimensionality Reduction** (PCA, embeddings)
- 7. **Link Analysis** (prominent structure in large graphs)

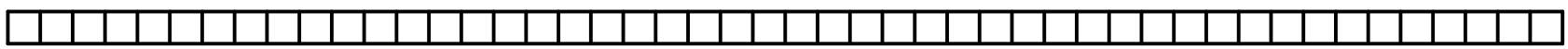
Controlling for Noise and Uncertainty:

- 6. **Noisy Data** (anomalies in data, ethics, privacy)

Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

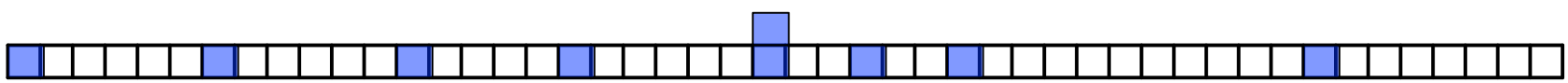
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

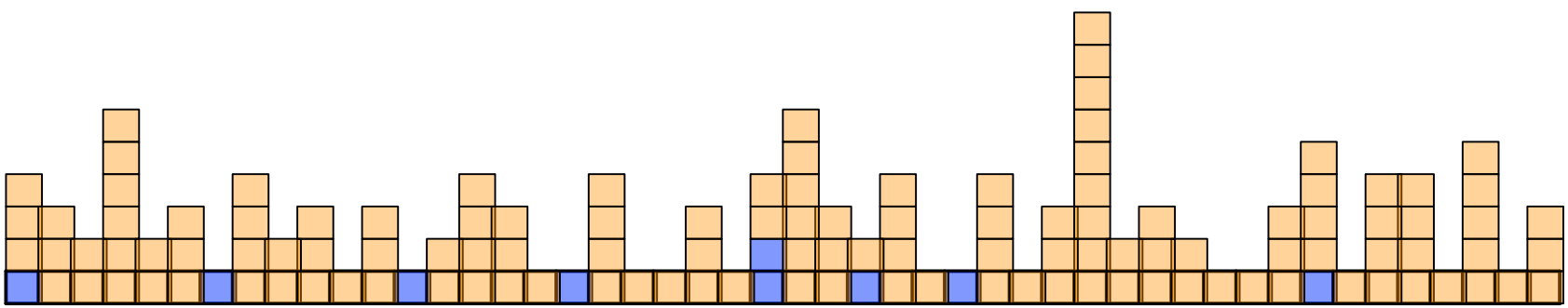
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

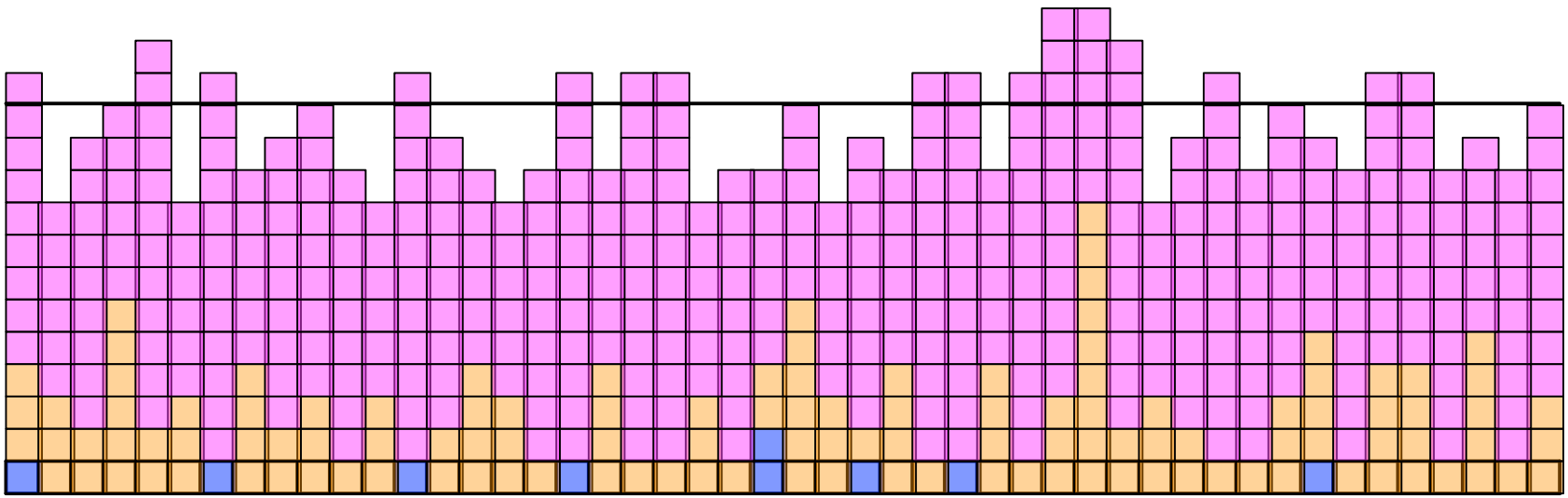
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Principals

What happens as data is generated with replacement
{IP addresses, words in dictionary, edges in graph, hash table}

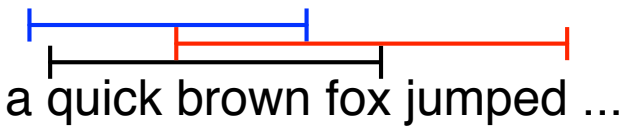
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?

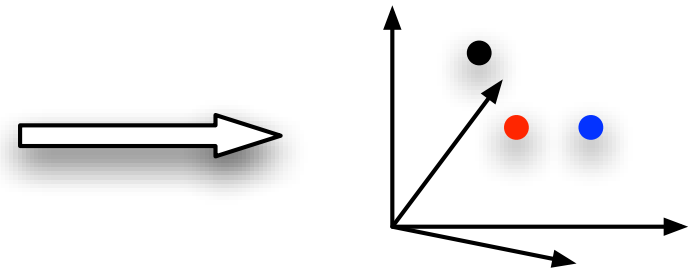


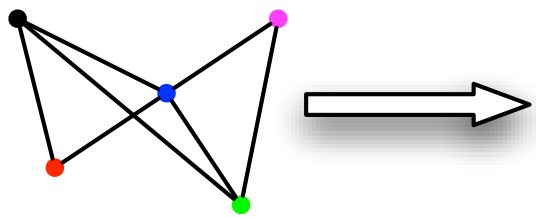
Raw Data to Abstract Representations

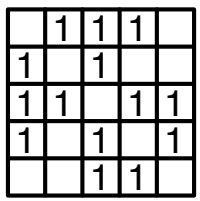
How to measure similarity between data?

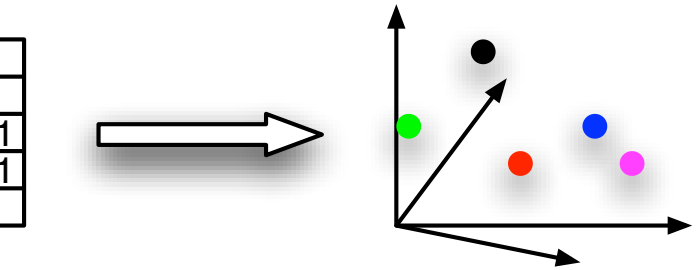
Key idea: data $\rightarrow$ point



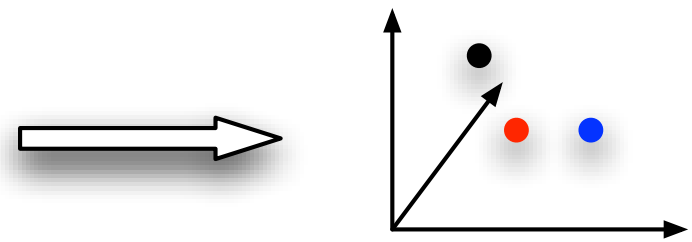








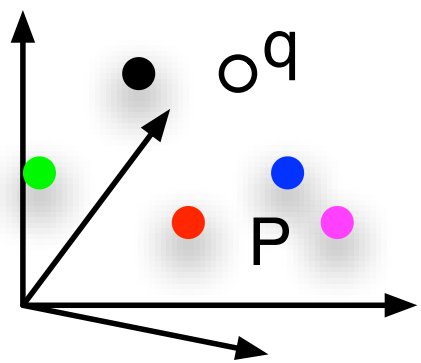
	age	income	height
joe	25	90K	1.85
bob	32	45K	1.52
sue	28	38K	1.61



Similarity

Given a large set of data P.
Given new point q, is q in P?

Given a large set of data P.
Given new point q, what is the *closest* point in P to q?



Clustering

How to find groups of similar data.

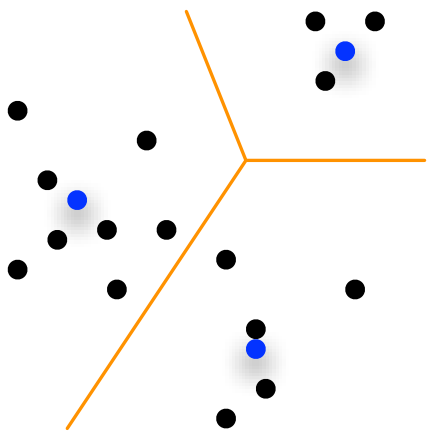
- do we need a representative?
- can groups overlap?
- what is structure of data/distance?

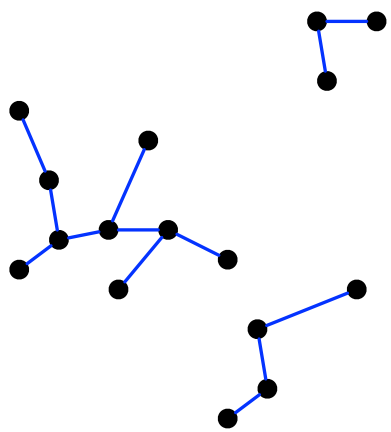
Clustering

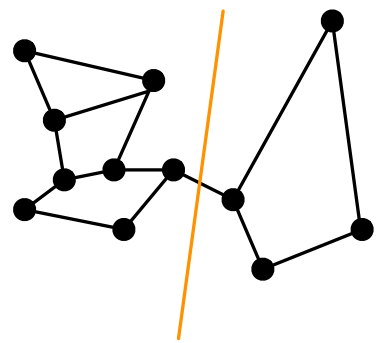
How to find groups of similar data.

- do we need a representative?
- can groups overlap?
- what is structure of data/distance?

- **Hierarchical clustering** : When to combine groups?
- k-**means clustering** : k-median, k-center, k-means++
- **Graph clustering** : modularity, spectral



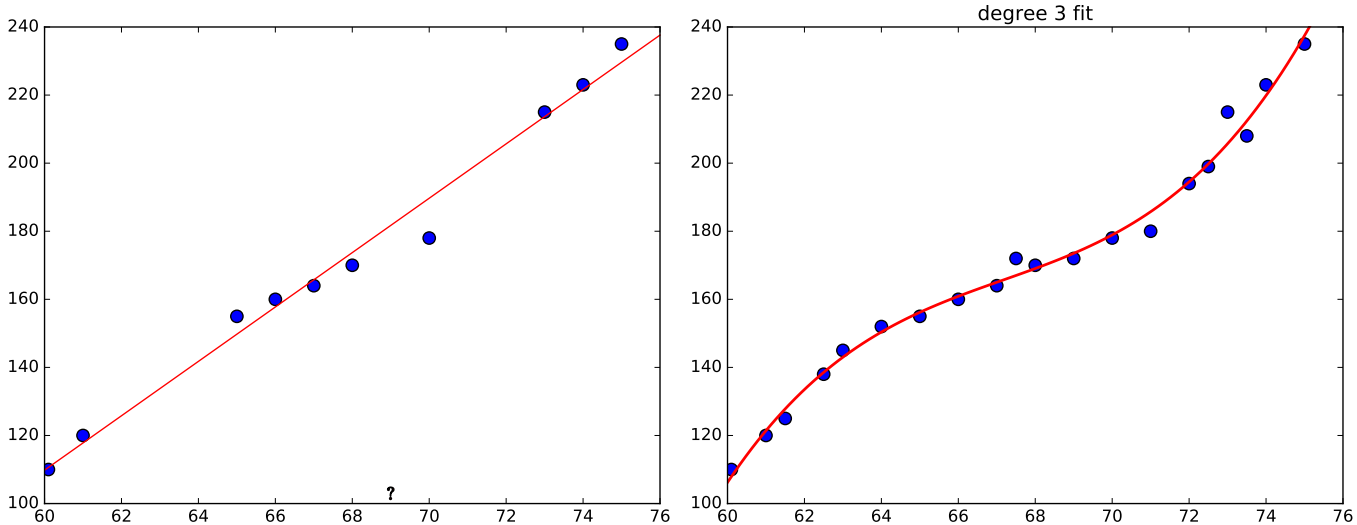




Regression

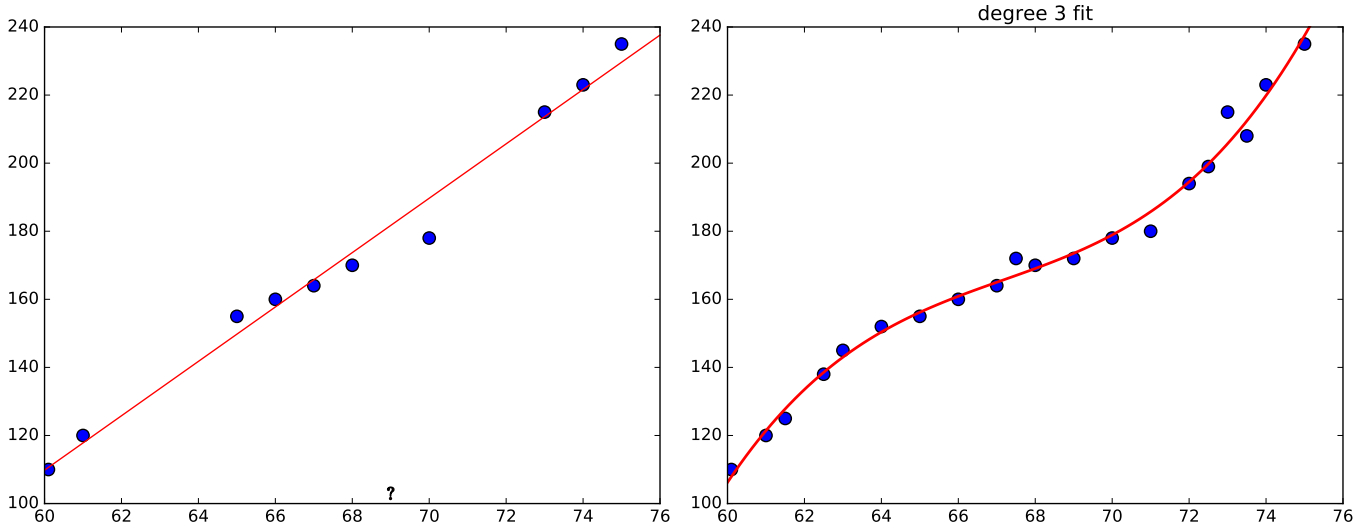
Consider a data set $P \in \mathbb{R}^d$, where d is BIG!

Want to find linear (or polynomial) function that represents P.



Regression

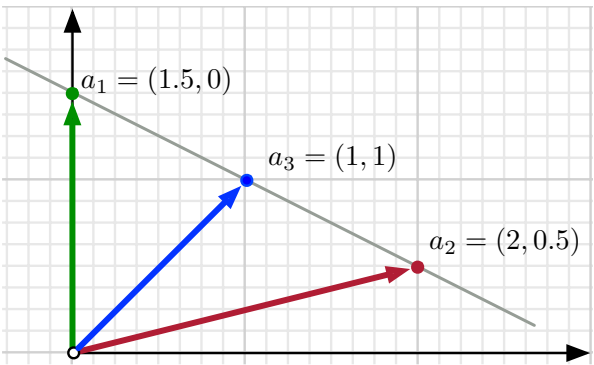
Consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find linear (or polynomial) function that represents P.

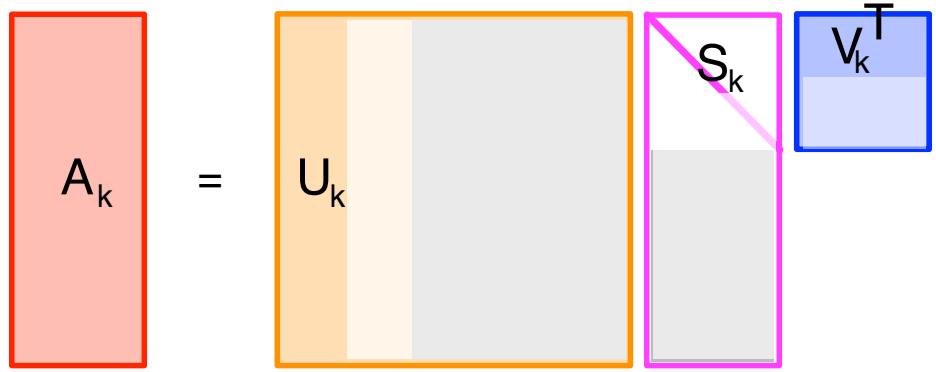


- **Least Squares** : Common easy approach (polynomial, high-dimensional)
- L_1 **Regression** : Sparser, generalizes better, Orthogonal Matching Pursuit
- **Info Recovery** : Compressed Sensing

Dimensionality Reduction

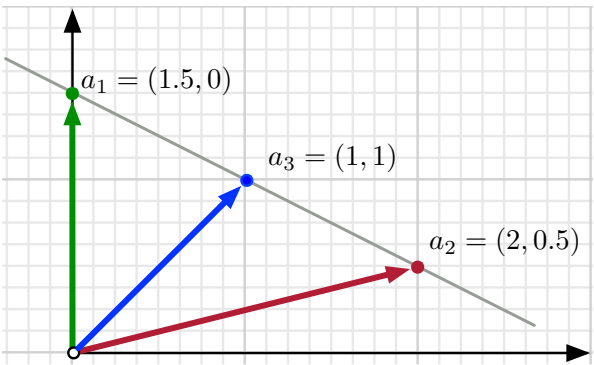
Again consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find linear subspace that represents P.

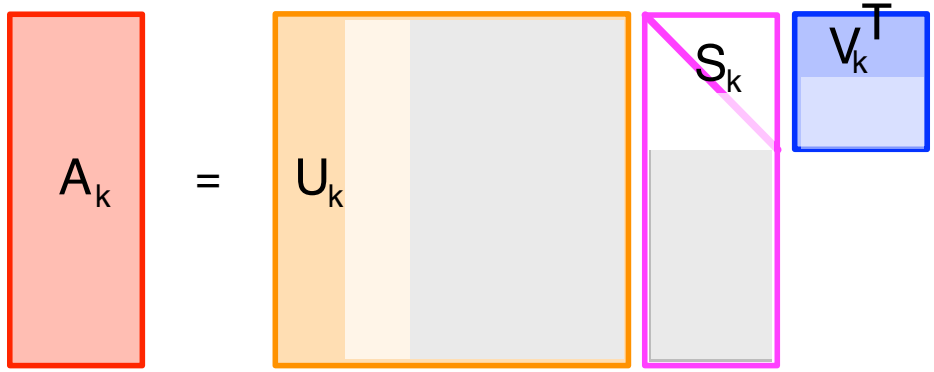




Dimensionality Reduction

Again consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find linear subspace that represents P.



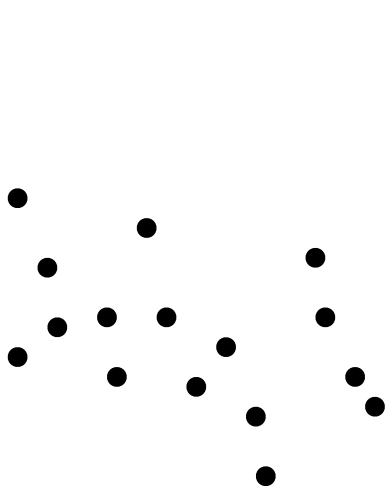


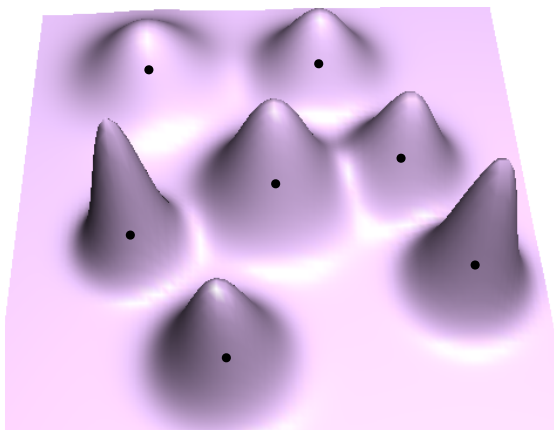
- **SVD** : Linear Algebra basis for PCA
- **Multidimensional Scaling** : Fits sets of distances in $\mathbb{R}^k$ with k small
- **Matrix Sketching**: Random Projections, Sampling, FD

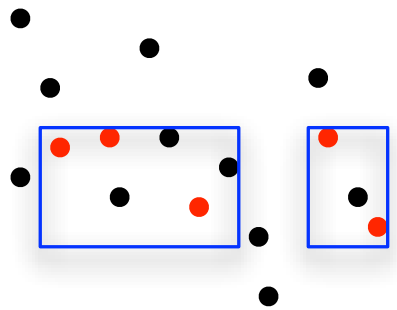
Noisy Data

What to do when data is noisy?

- **Identify it** : Find and remove outliers
- **Model it** : It may be real, affect answer
- **Exploit it** : Differential privacy, Ethics of Data Science



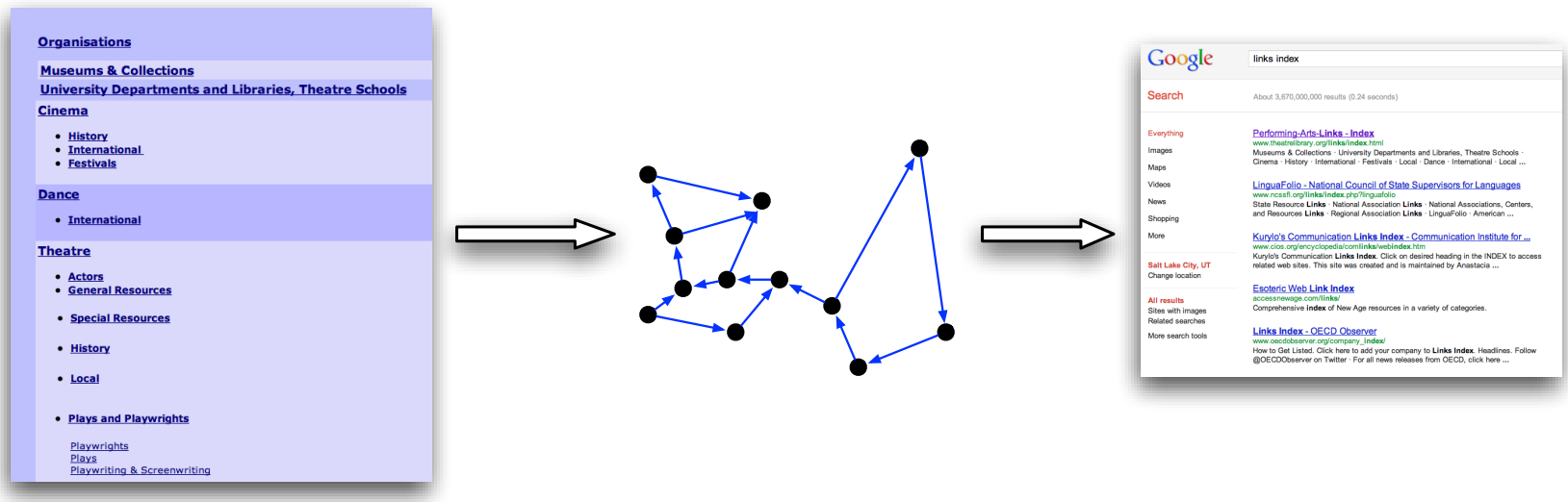




Link Analysis, Graphs

How does Google Search work?
Converts webpage links into directed graph.

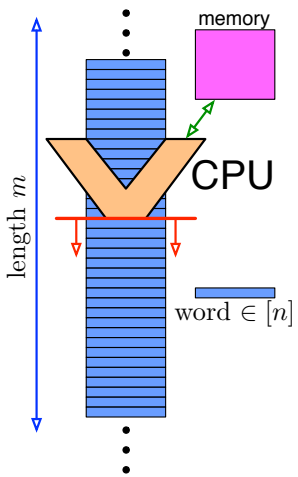
- **Markov Chains** : Models movement in a graph
- **PageRank** : How to convert graph into important nodes
- **MapReduce** : How to scale up PageRank
- **Communities** : Other important nodes in graphs

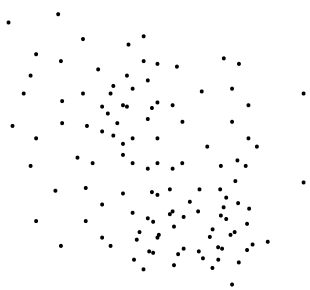


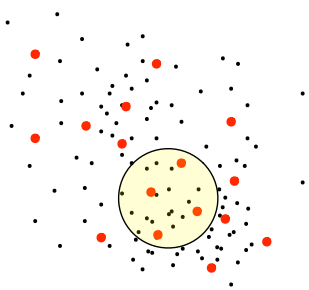
Summaries

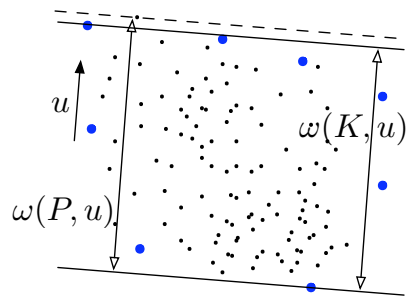
Reducing *massive* data to small space.
Want to retain as much as possible (not specific structure)
error guarantees

- **OnePass Sampling** : Reservoir Sampling
- **MinCount Hash** : Sketching data, $\rightarrow$ abstract features
- **Density Approximation** : Quantiles
- **Matrix Sketching** : Preprocessing complex data
- **Spanners** : graph approximations









Themes

What are course goals?

- Intuition for data analytics
- How to model data (convert to abstract data types)
- How to process data efficiently (balance models with algorithms)