Data Mining

DS 4140 / CS 5140 / CS 6140

Jeff M. Phillips

August 18, 2025

What is Data Mining?

What is Data Mining?

- Finding structure in data?
- Unsupervised machine learning?
- Large scale computational statistics?
- Anomaly and outlier detection/removal?
- Exploratory data analysis?

What is Data Mining?

- Finding structure in data?
- Unsupervised machine learning?
- Large scale computational statistics?
- Anomaly and outlier detection/removal?
- Exploratory data analysis?

- Toolbox for data analytics.

What is Data Mining?

- Finding structure in data?
- Unsupervised machine learning?
- Large scale computational statistics?
- Anomaly and outlier detection/removal?
- Exploratory data analysis?

- Toolbox for 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 libraries / python*

Other Data Mining Courses

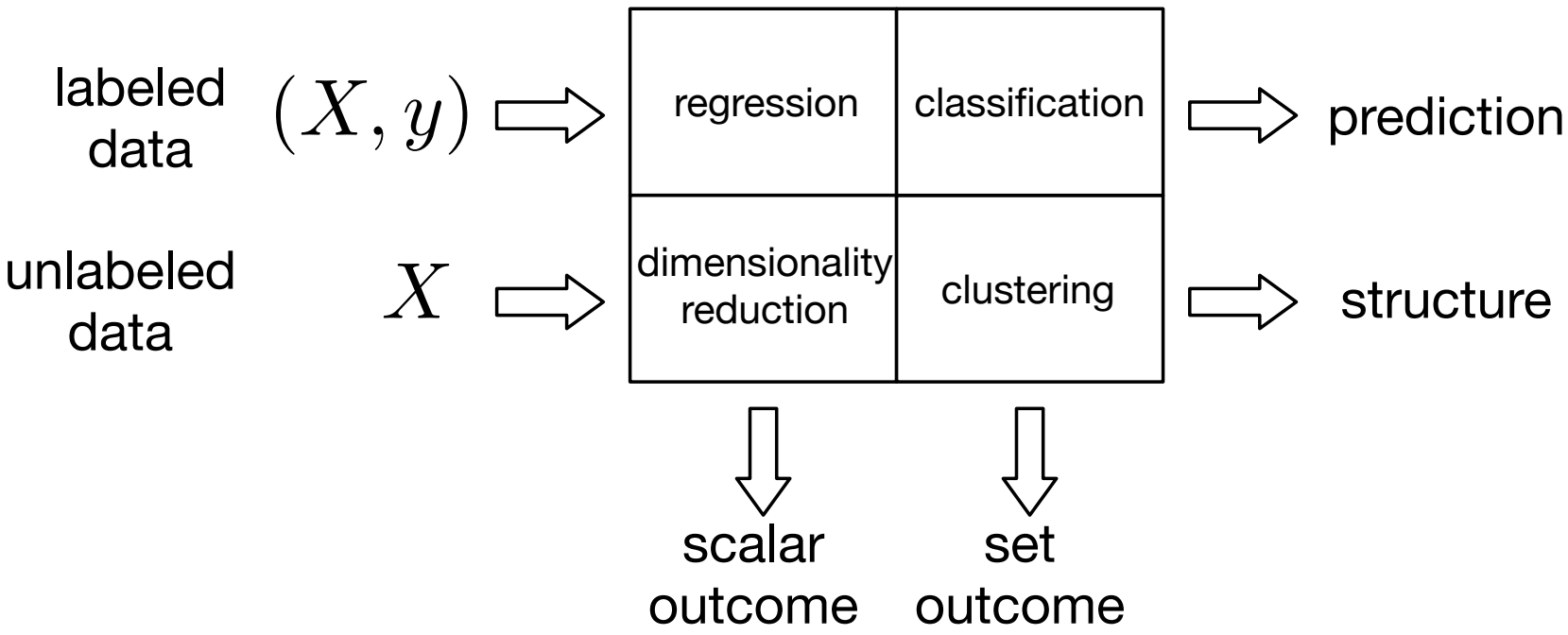
Every university teaches data mining differently!

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 libraries / python*

Maths: Linear Algebra, Probability, High-dimensional geometry

Classic View of Supervised and Un-Supervised Learning



Outline

Statistical, Anomalies, Uncertainty:

- 1. **Hashing & Anomalies** (& concentration of measure)
- 6. **Noisy Data** (outliers in data, ethics, privacy)

Structure in Data:

- 2. **Similarity** (find duplicates and similar items, NN search)
- 3. **Clustering** (aggregate close items, finding groups)
- 4. **Summaries** (exemplars, data reduction, streams)
- 5. **Dimensionality Reduction** (PCA, embeddings, metrics)
- 7. **Link Analysis** (prominent structure in large graphs)

Anomalies

When you see something strange ...

How do we know if it is unusual?
How do we quantify it?

Anomalies

When you see something strange ...

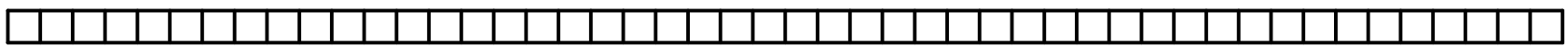
How do we know if it is unusual?
How do we quantify it?

Need to model what we expect – baseline distribution.
Can we *simulate* such data?
How unusual is it compared to baseline.

Statistical Phenomena

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

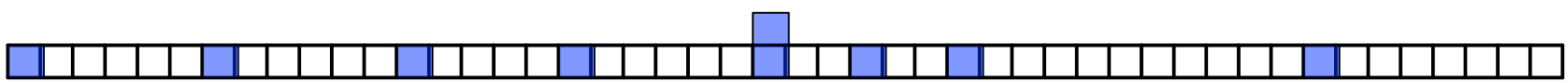
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Phenomena

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

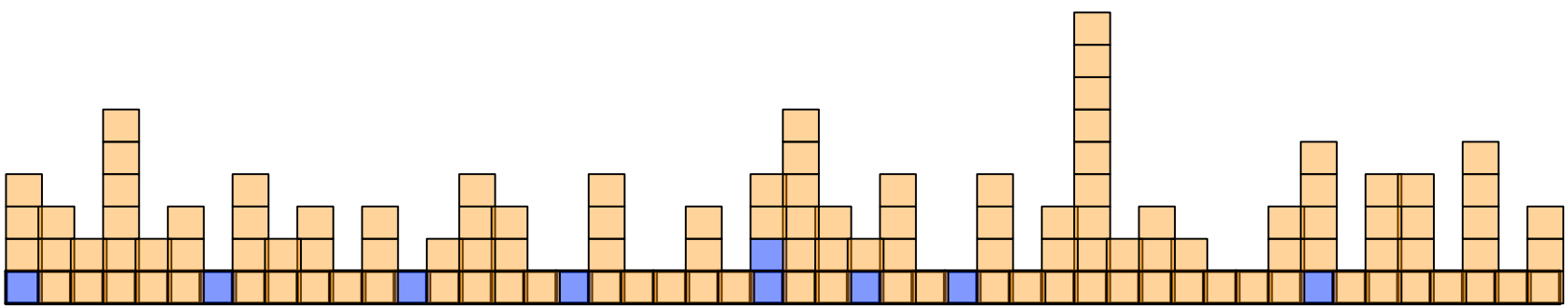
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Phenomena

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

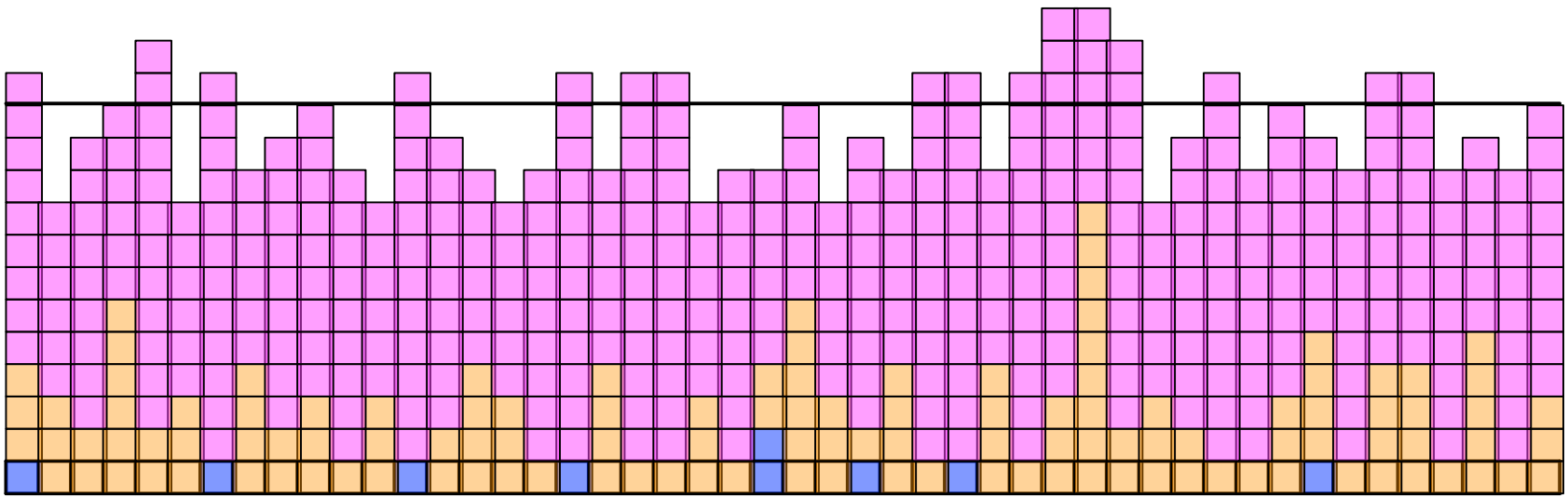
- When do items collide?
- When do you see all items?
- When is the distribution almost uniform?



Statistical Phenomena

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?



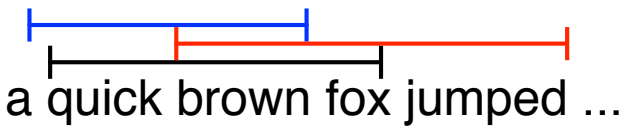
Raw Data to Abstract Representations

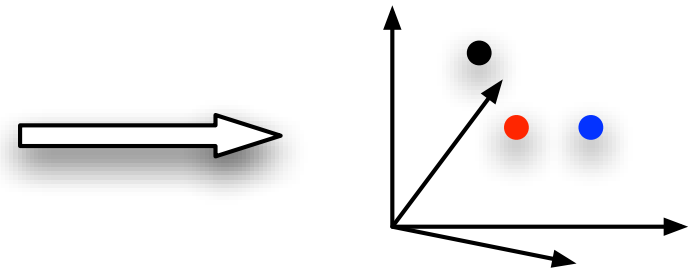
How to measure similarity between data?

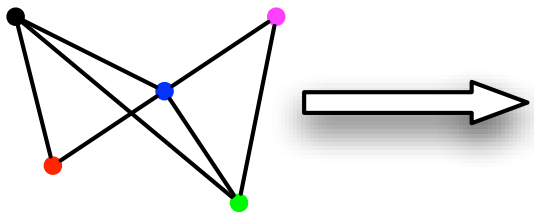
Raw Data to Abstract Representations

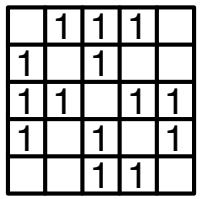
How to measure similarity between data?

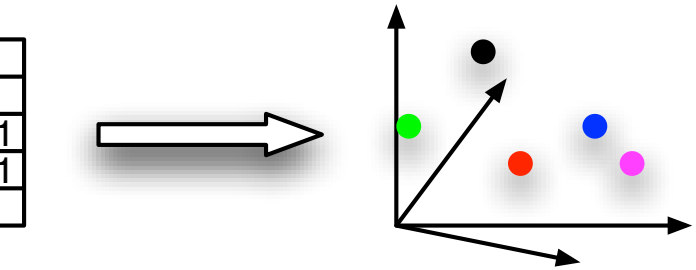
Key idea: data $\rightarrow$ point



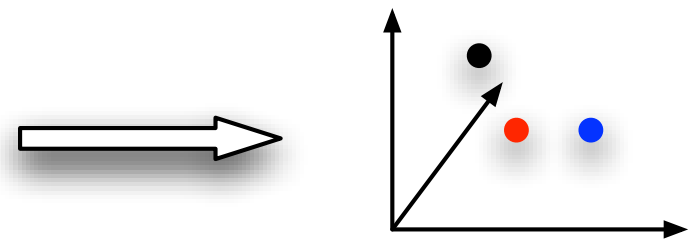








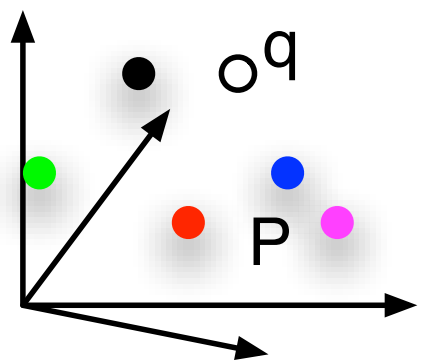
	age	income	height
joe	25	90K	1.85
bob	32	45K	1.52
sue	28	38K	1.61



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.

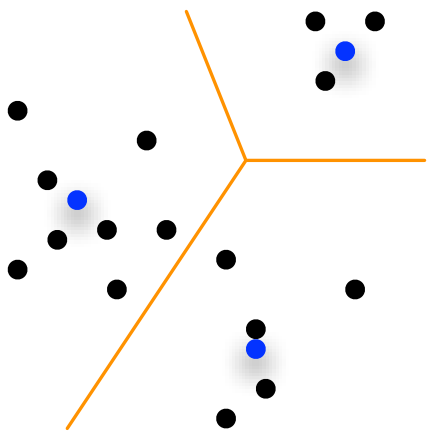
- do we need a representative?
- can groups overlap?
- what is structure of data/distance?

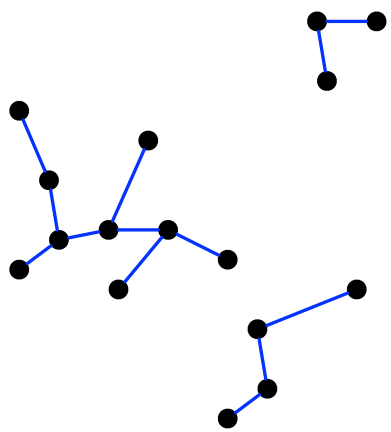
Clustering

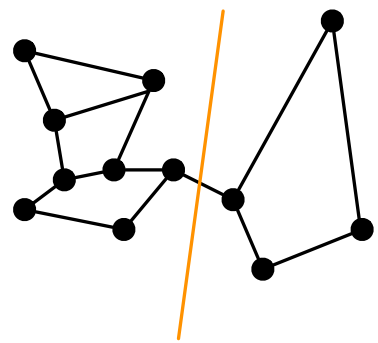
How to find groups of similar data.

- 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

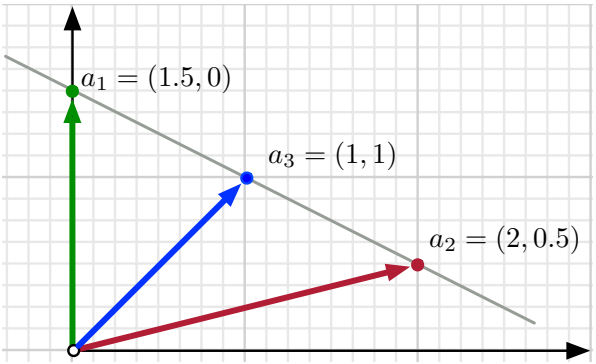


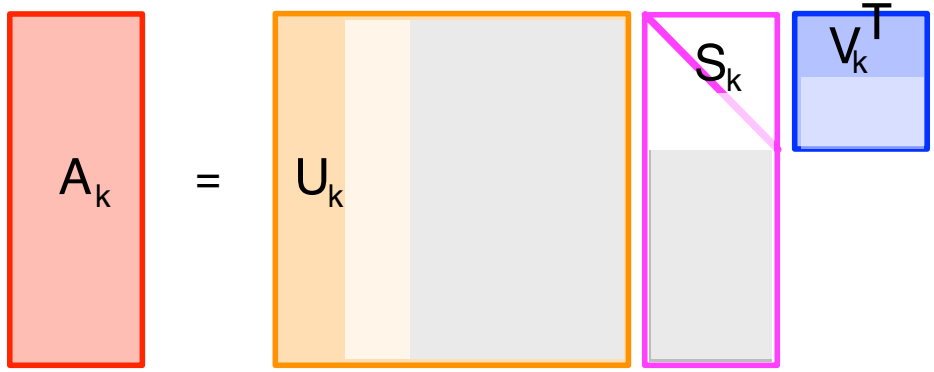




Dimensionality Reduction

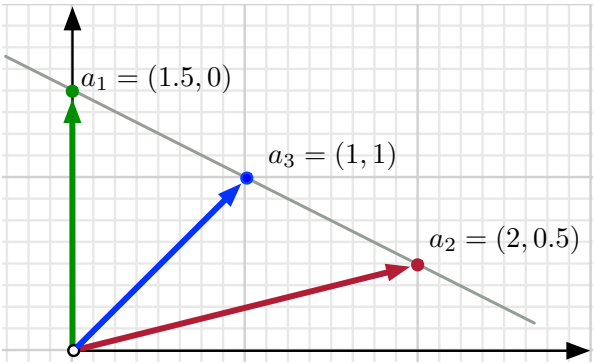
Again consider a data set $P \in \mathbb{R}^d$, where d is BIG!
Want to find linear subspace that represents P.

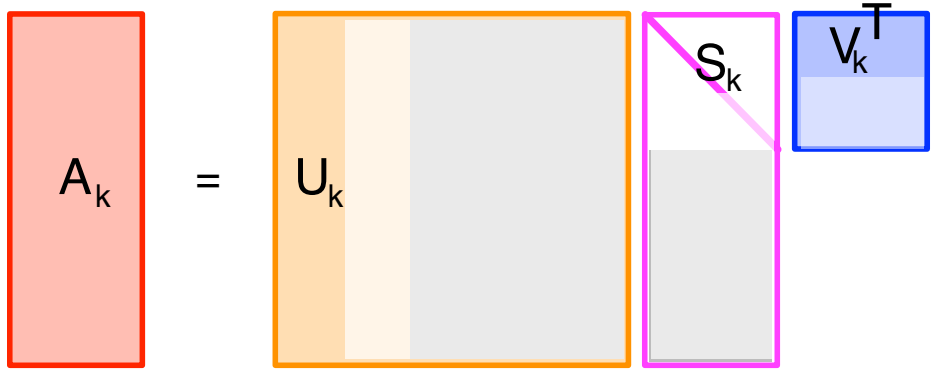




Dimensionality Reduction

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



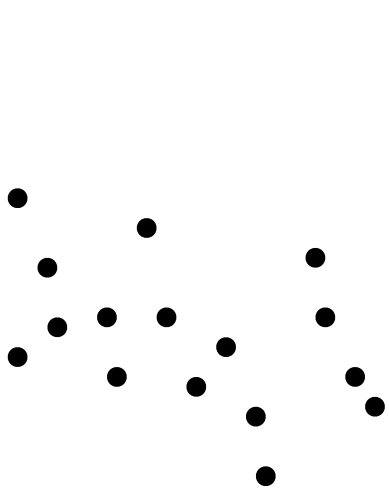


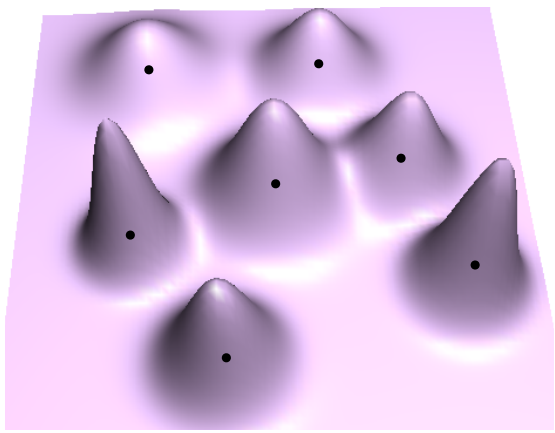
- **SVD** : Linear Algebra basis for PCA
- **Multidimensional Scaling** : Fits sets of distances in $\mathbb{R}^k$ with k small
- **Metric Learning** : Can labels help?
- **Matrix Sketching**: *Random Projections*, Sampling, FD
- **Feature Selection**: Which dimensions are useful?

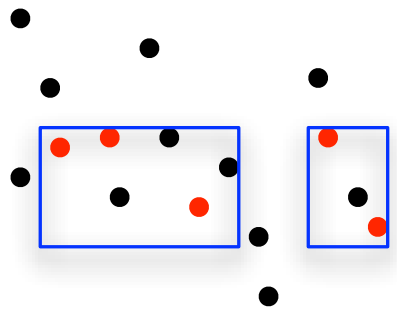
Noisy Data

What to do when data is noisy?

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



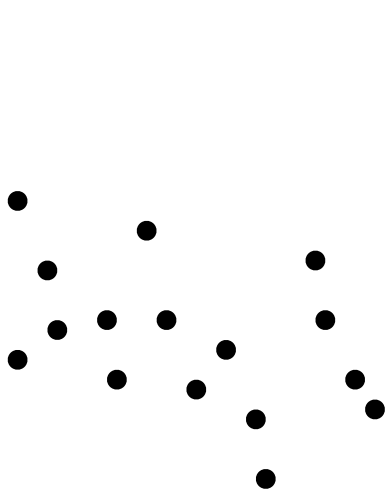


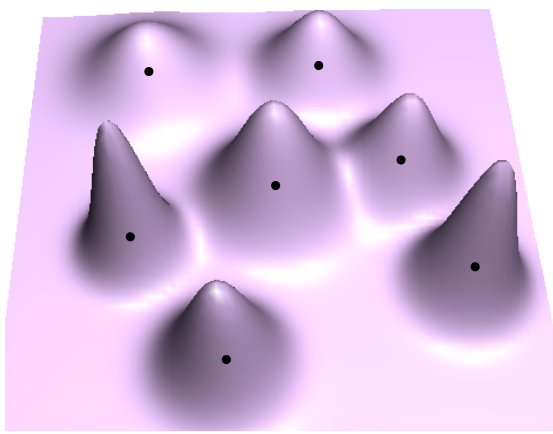


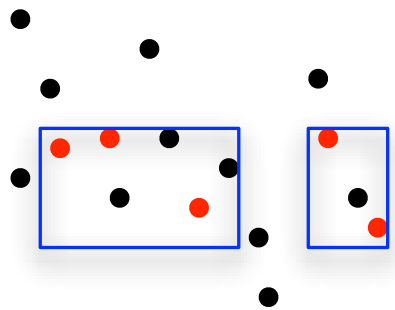
Noisy Data

What to do when data is noisy?

- **Identify it** : Find and remove outliers, Robust estimation
- **Model it** : It may be real, affect answer
- **Exploit it** : Differential privacy, Ethics of Data Science
- **Find Anomalies** : Data different (enough) from expected



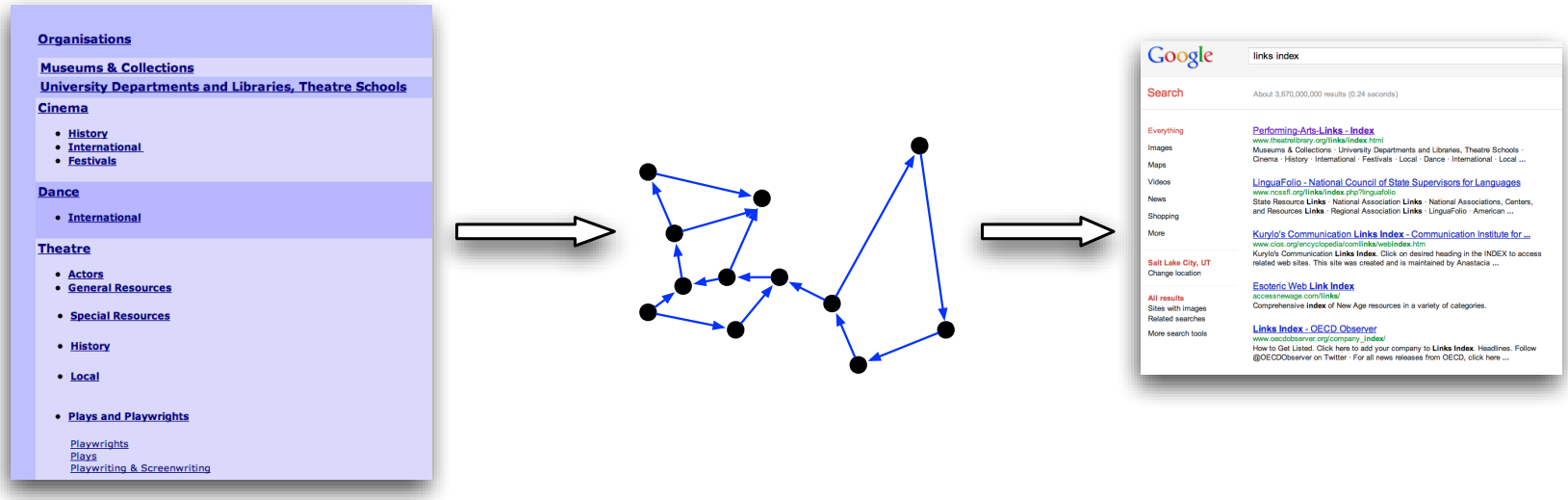




Link Analysis, Graphs

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

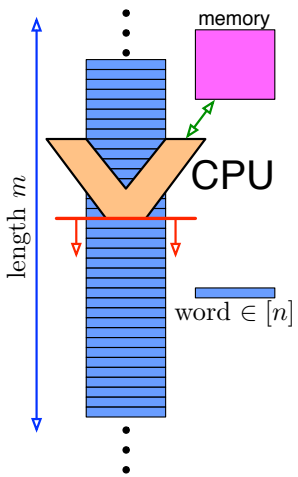
- **Markov Chains** : Models movement in a graph
- **PageRank** : How to convert graph into important nodes
- **Embeddings** : Map graphs to vector representations
- **Communities** : Other important nodes in graphs

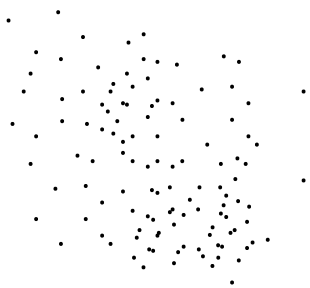


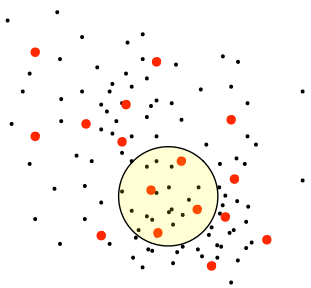
Summaries

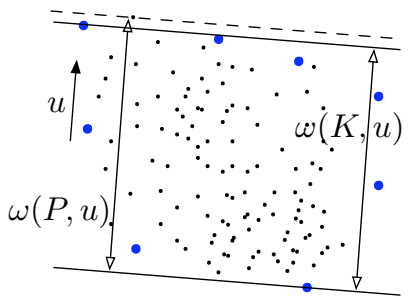
Reducing *massive* data to small space.
Want to retain as much as possible (not specific structure)
error guarantees

- **OnePass Sampling** : Reservoir Sampling
- **MinCount Hash** : Sketching data $\rightarrow$ abstract features
- **Density Approximation** : Quantiles
- **Frequency Approximation** : Identify most common items
- **Matrix Sketching** : Preprocessing complex data









Themes

What are course goals?

- Intuition for data analytics
- Develop toolbox for modeling data in many settings.
- How to convert to abstract data types
- How to process data efficiently (balance models with algorithms)
- Understand difference between regular noise in data vs. anomalies