UC RIVERSIDE UNIVERSITY OF CALIFORNIA

# Introduction to Data Science

Evangelos (Vagelis) Papalexakis

UC Riverside

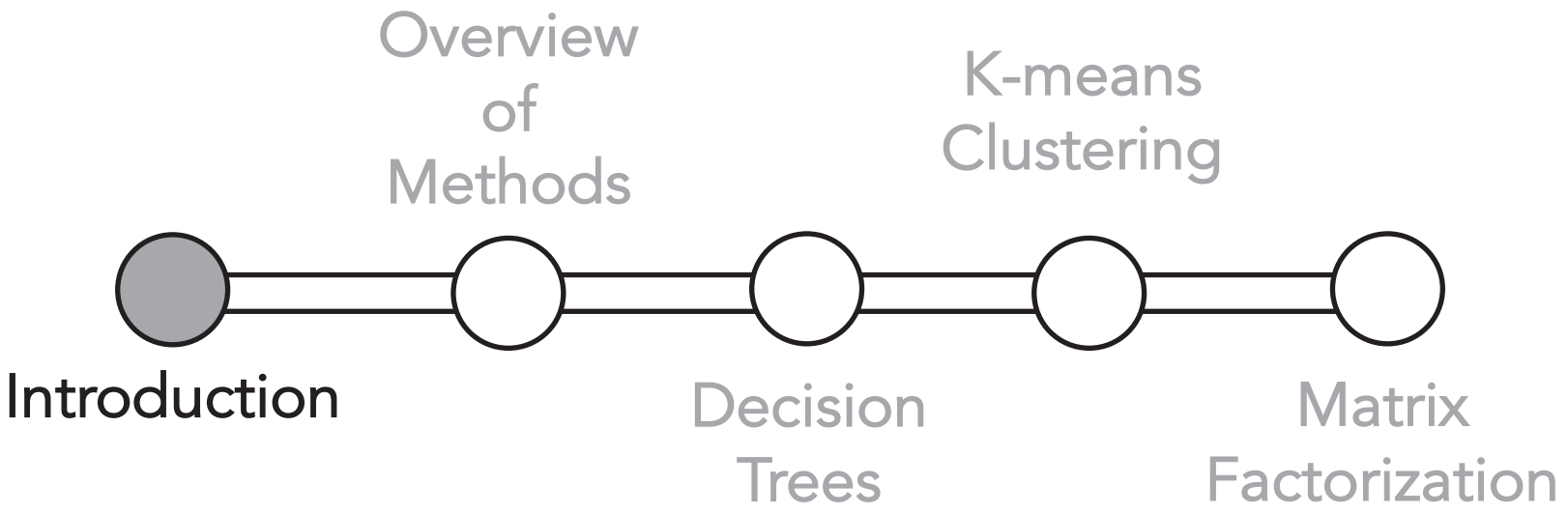
M Multi A Aspect D Data Lab @ UCR

NASA MIRO FIELDS 2020, UCR

Slides compiled from multiple sources including CS171 Winter 2020
https://www.cs.ucr.edu/~epapalex/teaching/171_W20/index.html

# Roadmap


ROUTE
66


Overview of Methods
K-means Clustering
Introduction
Decision Trees
Matrix Factorization

$$7 \pm 2$$

# Miller's Law

$$7 \pm 2$$

Number of **items** an average human holds in **working memory**
*George A. Miller, 1956*

More at: https://en.wikipedia.org/wiki/The_Magical_Number_Seven,_Plus_or_Minus_Two

# Data Science

# Data Science

Photos



# Data Science

Photos



Social Media



# Data Science

Photos



Social Media



Online News



# Data Science

Photos

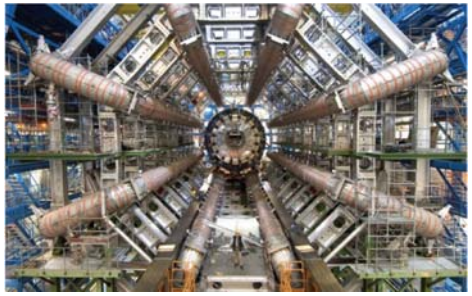


Social Media

Online News

Scientific Data

Large Hadron Collider (LHC)



LIGO

Image credits:
https://about.twitter.com/en_us/company/brand-resources.html, https://en.facebookbrand.com/assets
https://publicdomainvectors.org/en/free-clipart/Newspaper-vector-icon/75638.html
https://www.extremetech.com/extreme/276672-the-large-hadron-collider-came-online-10-years-ago-today
https://www.ligo.caltech.edu/image/ligo20150731a

# Data Science

Photos



Social Media

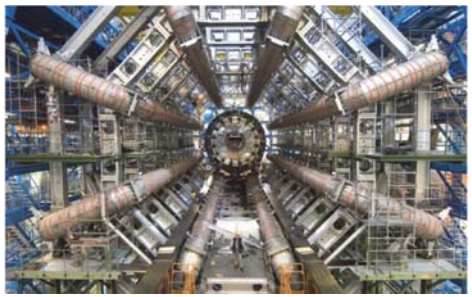


Online News



Scientific Data

Large Hadron Collider (LHC)



LIGO

Image credits:
https://about.twitter.com/en_us/company/brand-resources.html, https://en.facebookbrand.com/assets
https://publicdomainvectors.org/en/free-clipart/Newspaper-vector-icon/75638.html
https://www.extremetech.com/extreme/276672-the-large-hadron-collider-came-online-10-years-ago-today
https://www.ligo.caltech.edu/image/ligo20150731a

Credit to Duen Horng Chau (GaTech) for inspiring this point of view
*Chau, "Data Mining Meets HCI: Making Sense of Large Graphs" PhD Thesis, 2012*

# Data Science

Photos

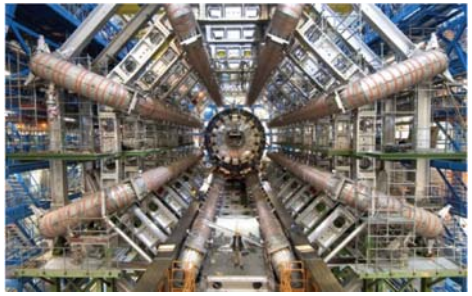


Social Media

Online News

Scientific Data

Large Hadron Collider (LHC)



LIGO

$7 \pm 2$

Image credits:
https://about.twitter.com/en_us/company/brand-resources.html, https://en.facebookbrand.com/assets
https://publicdomainvectors.org/en/free-clipart/Newspaper-vector-icon/75638.html
https://www.extremetech.com/extreme/276672-the-large-hadron-collider-came-online-10-years-ago-today
https://www.ligo.caltech.edu/image/ligo20150731a

# Data Science

Data

7 ± 2

Insights

Credit to Polo Chau (Georgia Tech)

# Data Science

Data

7 ± 2

Insights

coffee 100%
latte 95%
cortado 80%

# Data Science

Data

7 ± 2

Insights

coffee
latte
cortado

100%
95%
80%

iceberg
Greenland

100%
70%

# Data Science

Data

7 ± 2

Insights

coffee 100%
latte 95%
cortado 80%

iceberg 100%
Greenland 70%

temple 100%
Egypt 95%
museum 80%
MET 60%

# Data Science

Data

$7 \pm 2$

Insights

# Data Science

supervised

Data

$7 \pm 2$

Insights

unsupervised

# Data Science

supervised

Data

7 ± 2

Insights

semi-supervised

unsupervised

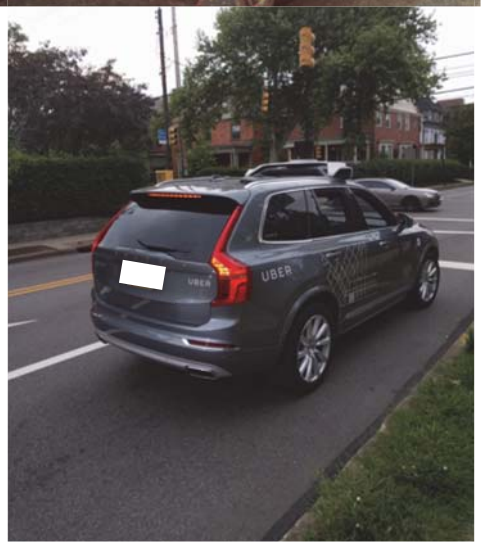
# Recommendation Systems


- **Problem**: Given a user's viewing/rating pattern, recommend new movies to watch
  - Based on similar movies
  - Based on similar users
- $1M prize!
- Collaborative Filtering

http://netflixprize.com/

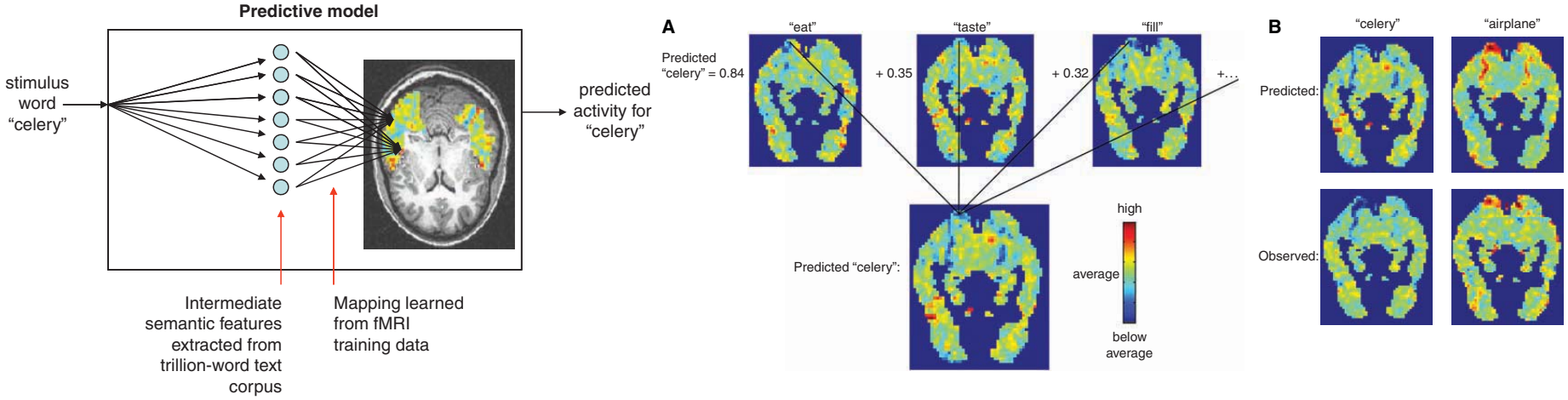
# Self-driving Car






http://www.businessinsider.com/30-companies-are-now-making-self-driving-cars-2016-4?r=UK&IR=T

# Mind Reading




http://www.cs.cmu.edu/afs/cs/project/theo-73/www/index.html

# Credit Card Fraud Detection

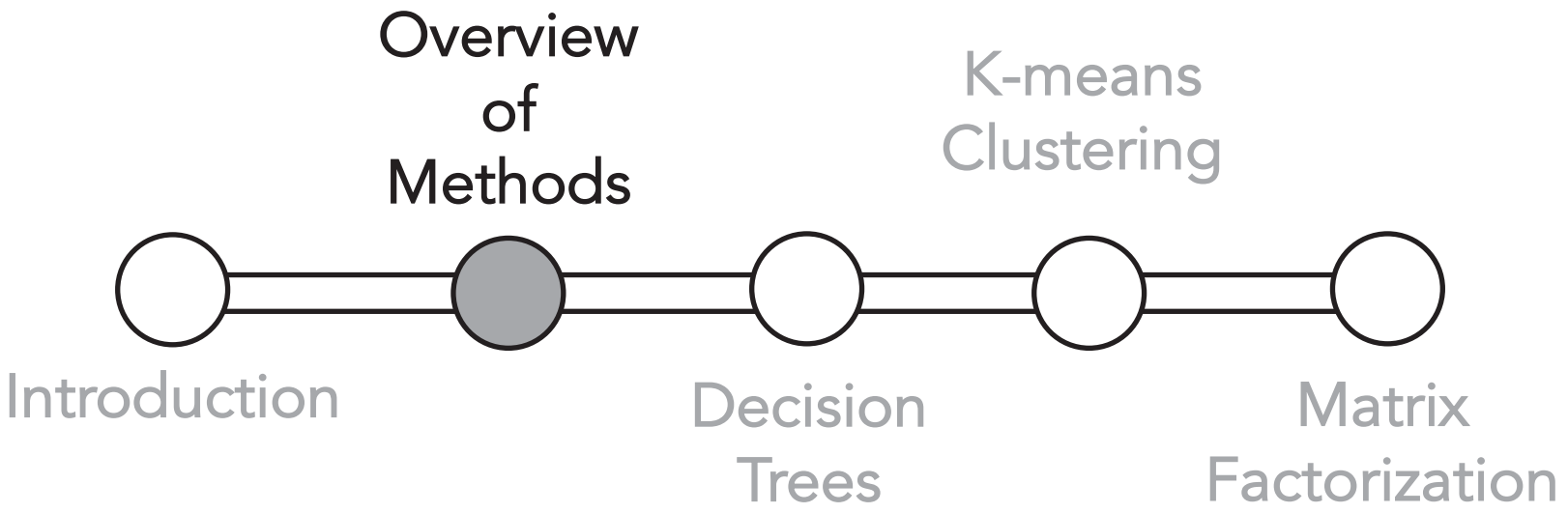
**Transactions**

| Vendor | Location | Amount | Timestamp |
|---|---|---|---|
| UCR Cafe | UCR | $4 | Tuesday 8am |
| UCR Hub | UCR | $8 | Tuesday 12pm |
| Grocery Store | Canyon Crest Twn. Ctr | $50 | Tuesday 9pm |
| SketchySite.com | Online | $5 | Wed. 3am |
| Ferrari Dealer | Not Riverside | $150,000 | Wed. 6am |

**Problem**: Given a transaction (and past transaction history) decide whether it is real or fraudulent. Preferably do that in *real time.*

# Roadmap


ROUTE
66


Overview of Methods
K-means Clustering
Introduction
Decision Trees
Matrix Factorization

# What classes of methods are there?

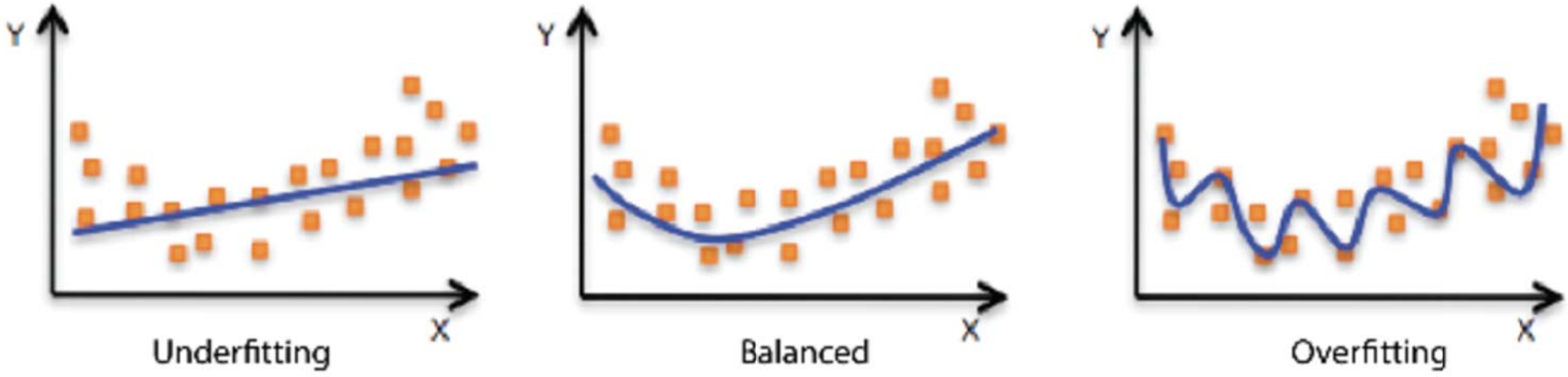
- Classification & regression (supervised)
- Clustering (unsupervised)
- Outlier & Anomaly Detection
- Sequential Pattern, Trend and Evolution Analysis
- Network Analysis  & Graph Mining
- Association and Correlation Analysis

# Classification & Regression

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …

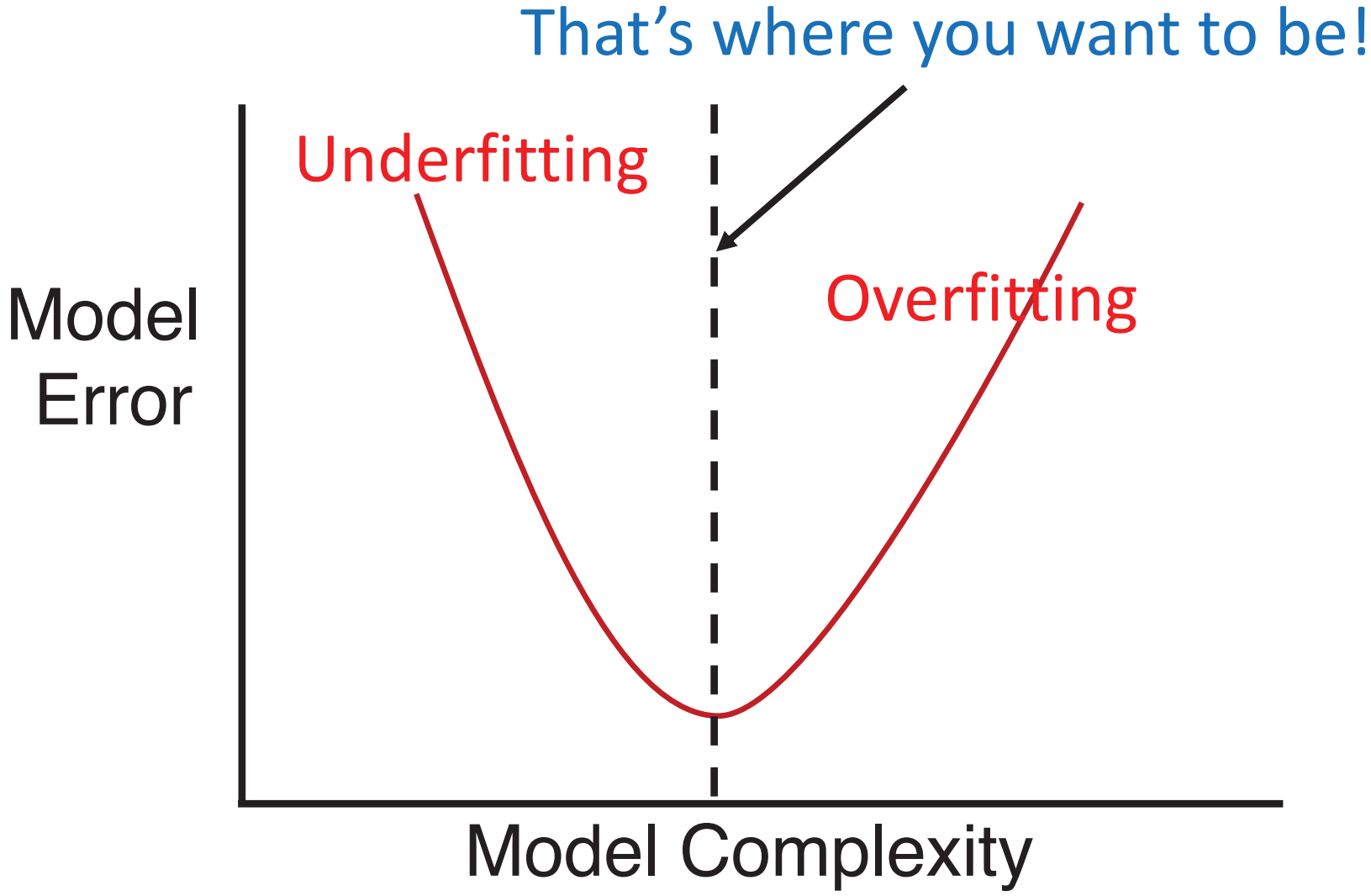
# Overfitting

- Happens when the model fits almost perfectly the training data
- May not be able to generalize to new **unseen** examples

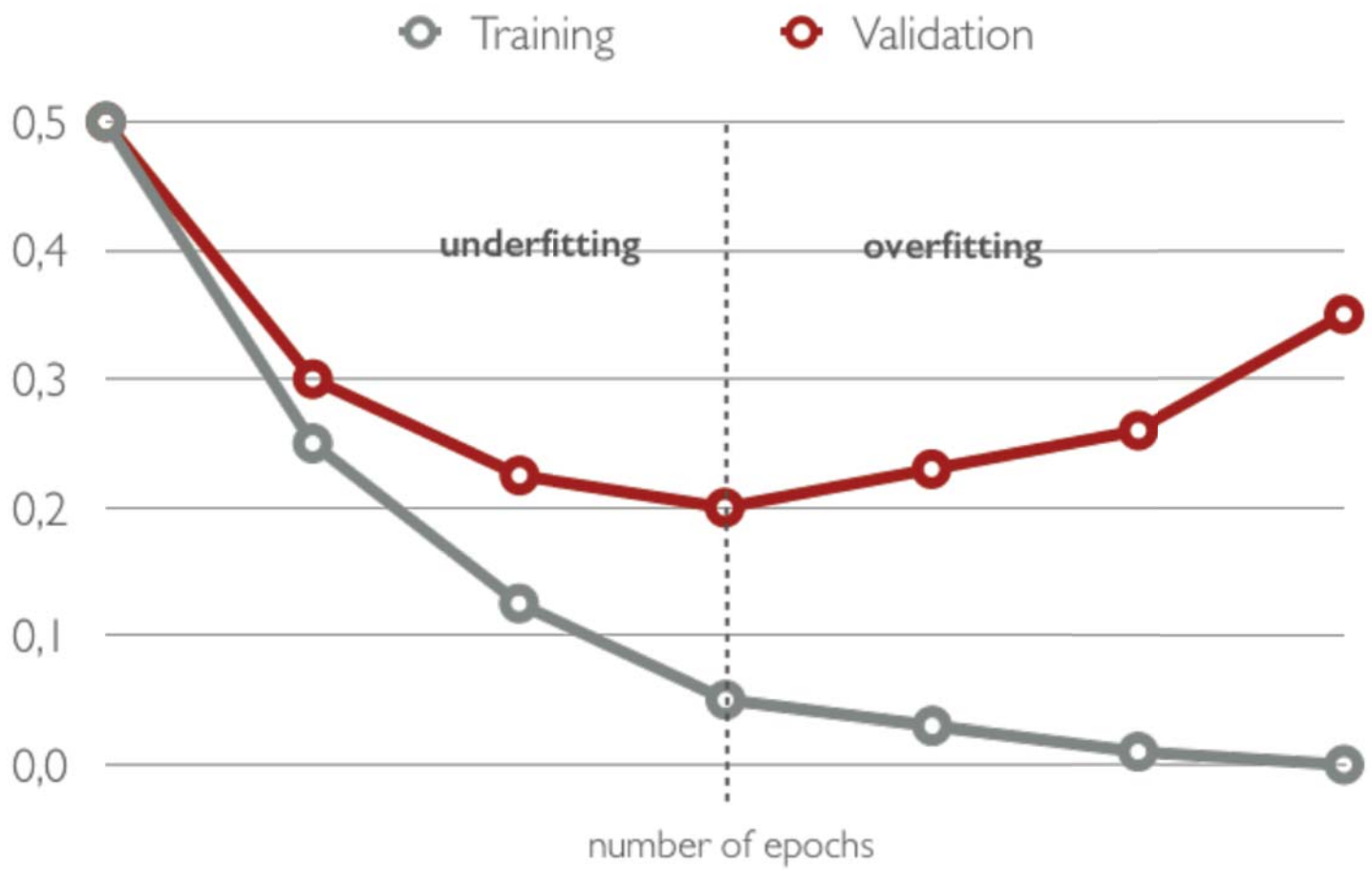



Source: http://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

# Overfitting Curve



# When to stop training?




Src: Russ Salakhutdinov https://www.cs.cmu.edu/~rsalakhu/10707/Lectures/Lecture_NN_Part2.pdf

# Clustering

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster documents according to topics
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

# Association and Correlation Analysis

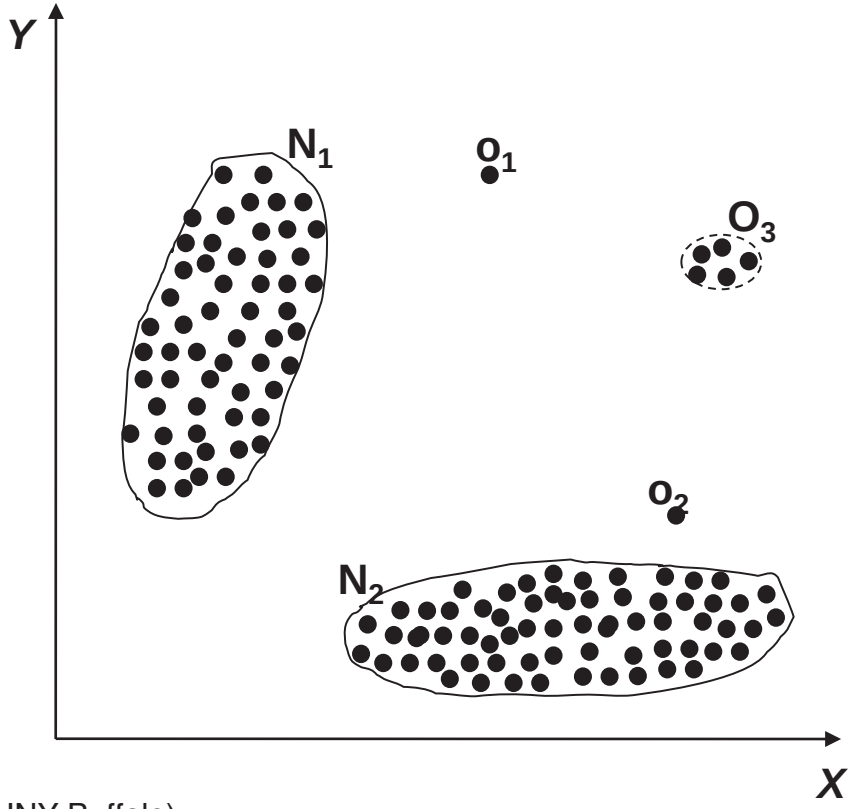
- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

# Outlier & Anomaly Detection

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

# Point Anomalies

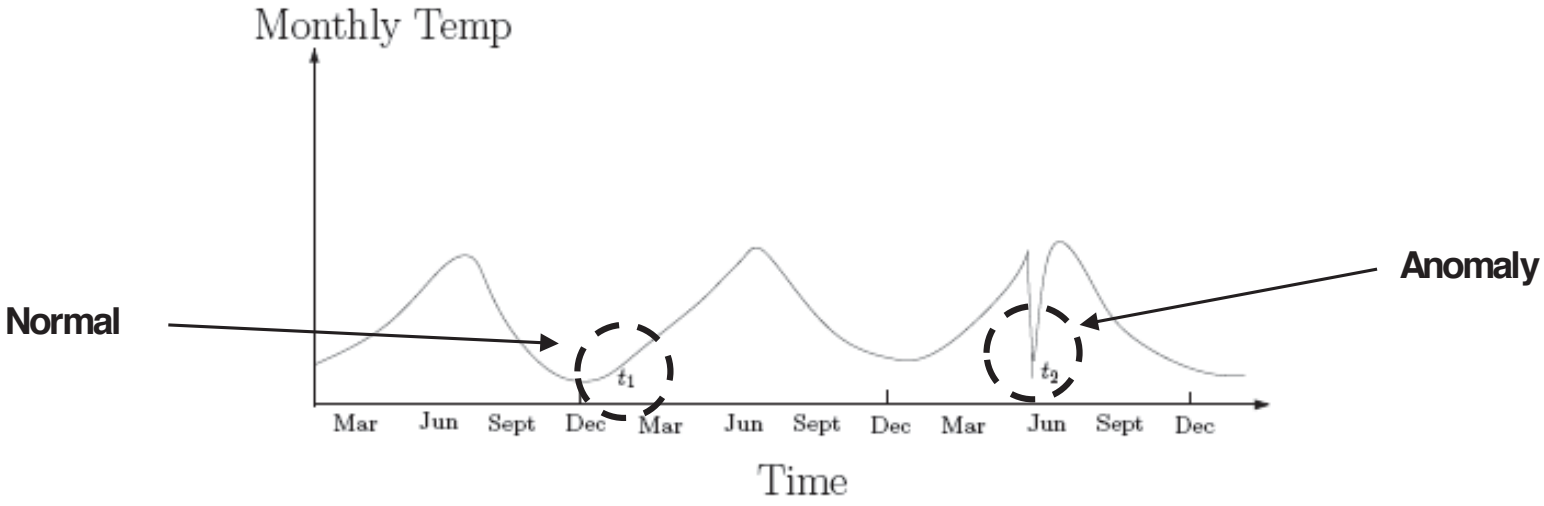
- A single data point being "far" from every other point




Slide adapted from Jing Gao (SUNY Buffalo)

# Contextual Anomalies

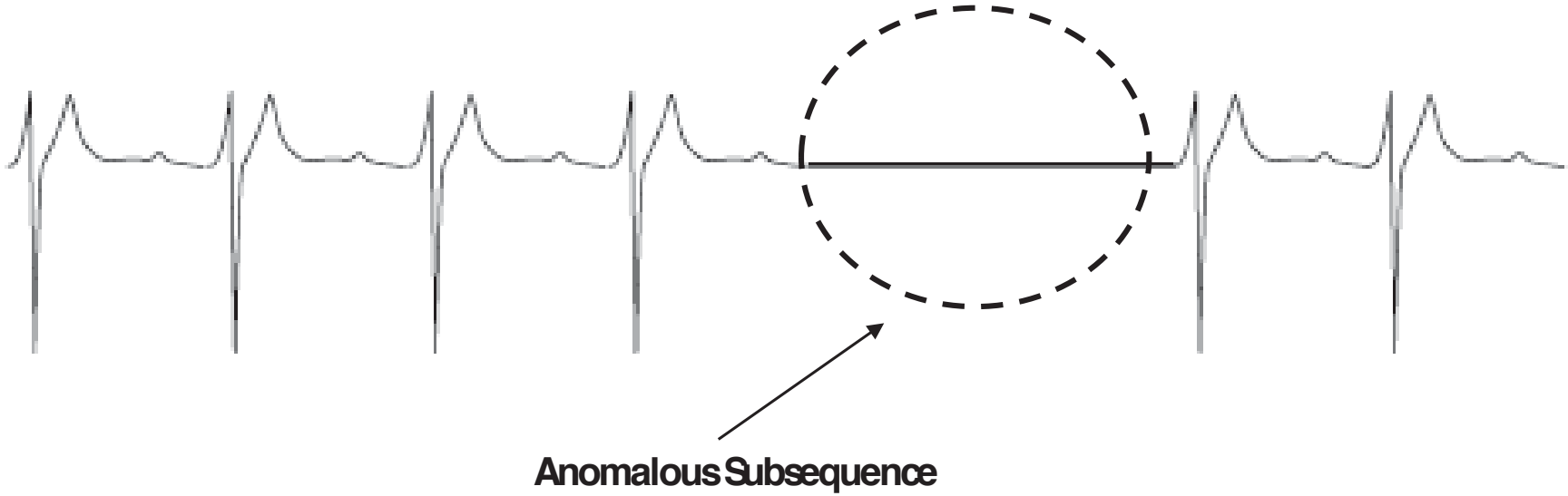
- The anomaly is within the *context* of the rest of the data




Slide adapted from Jing Gao (SUNY Buffalo)

# Collective Anomalies

- A collection/cluster of points is anomalous
- This is in contrast to the other, "normal" clusters



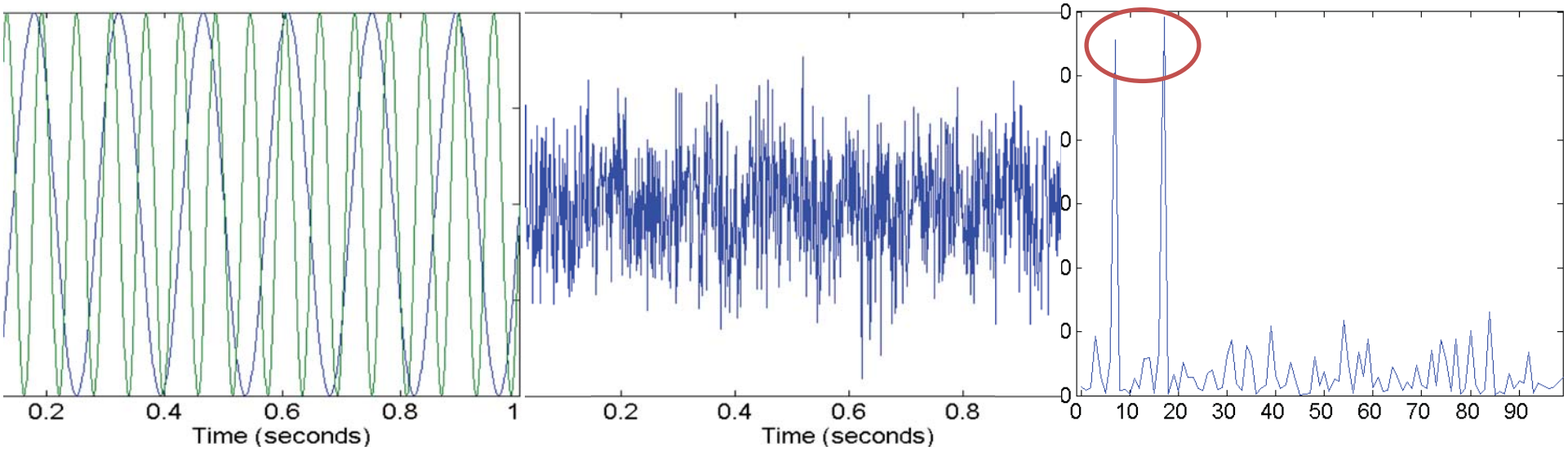

Slide adapted from Jing Gao (SUNY Buffalo)

# Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

# Fourier Transform

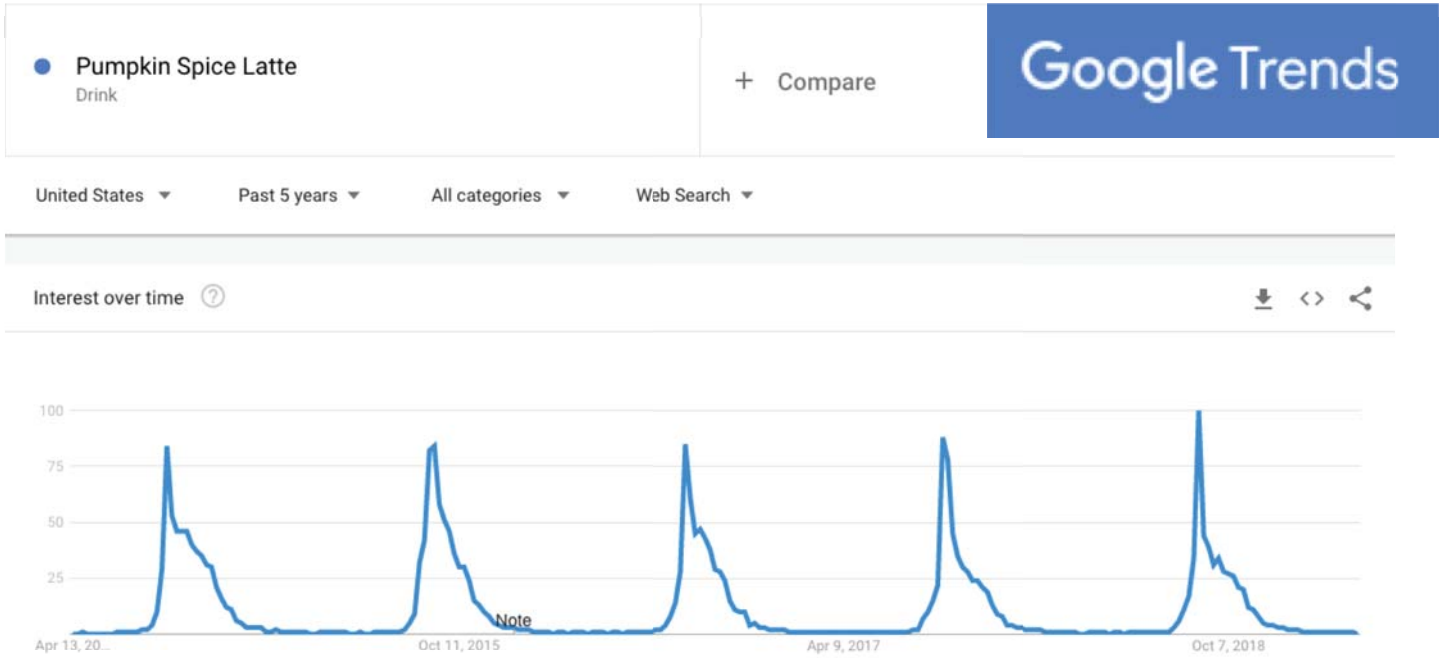
Keeping only those 2 peaks will give us the original sine waves!




**Two Sine Waves** **Two Sine Waves + Noise** **In Frequency Space**

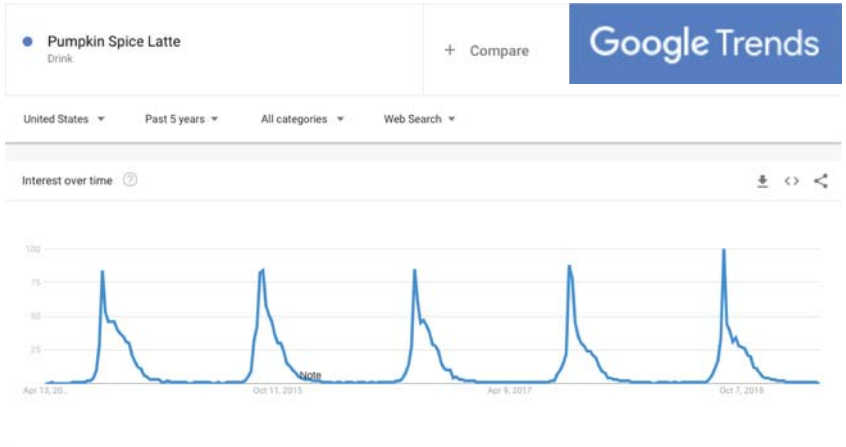
# Fourier Transform as Data Mining Tool

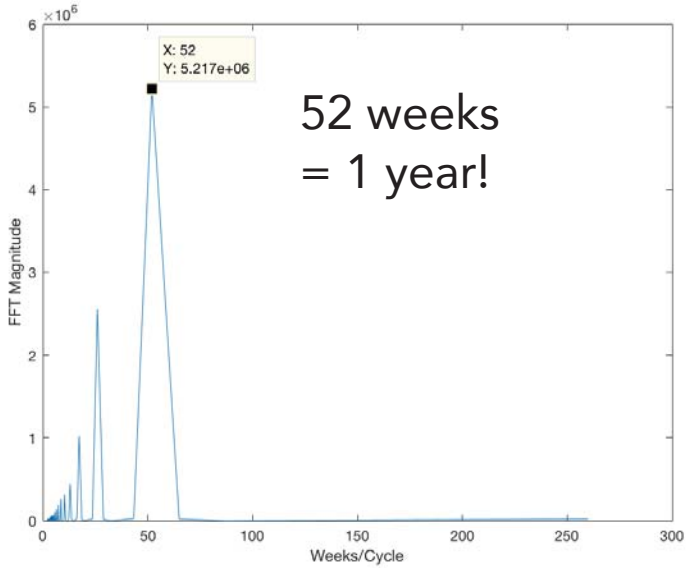
- The highest value in the Fourier domain is the main periodicity in our data!



# Fourier Transform as Data Mining Tool

- The highest value in the Fourier domain is the main periodicity in our data!




52 weeks
= 1 year!

# Network Analysis & Graph Mining

- **Graph mining**
  - Finding frequent subgraphs (e.g., chemical compounds), **Information network analysis**
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- **Web mining**
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

# The Anatomy of a Large-Scale Hypertextual Web Search Engine

## 2.1 PageRank: Bringing Order to the Web

*We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:*

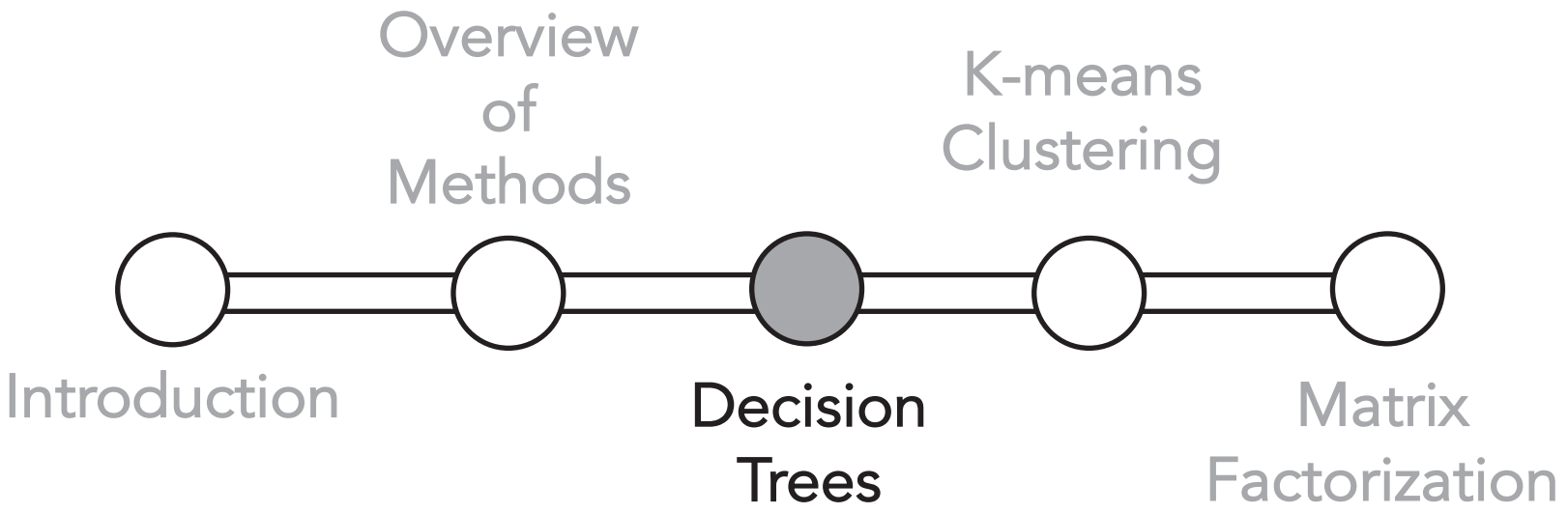
$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

*Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.*

http://ilpubs.stanford.edu:8090/361/

# Roadmap

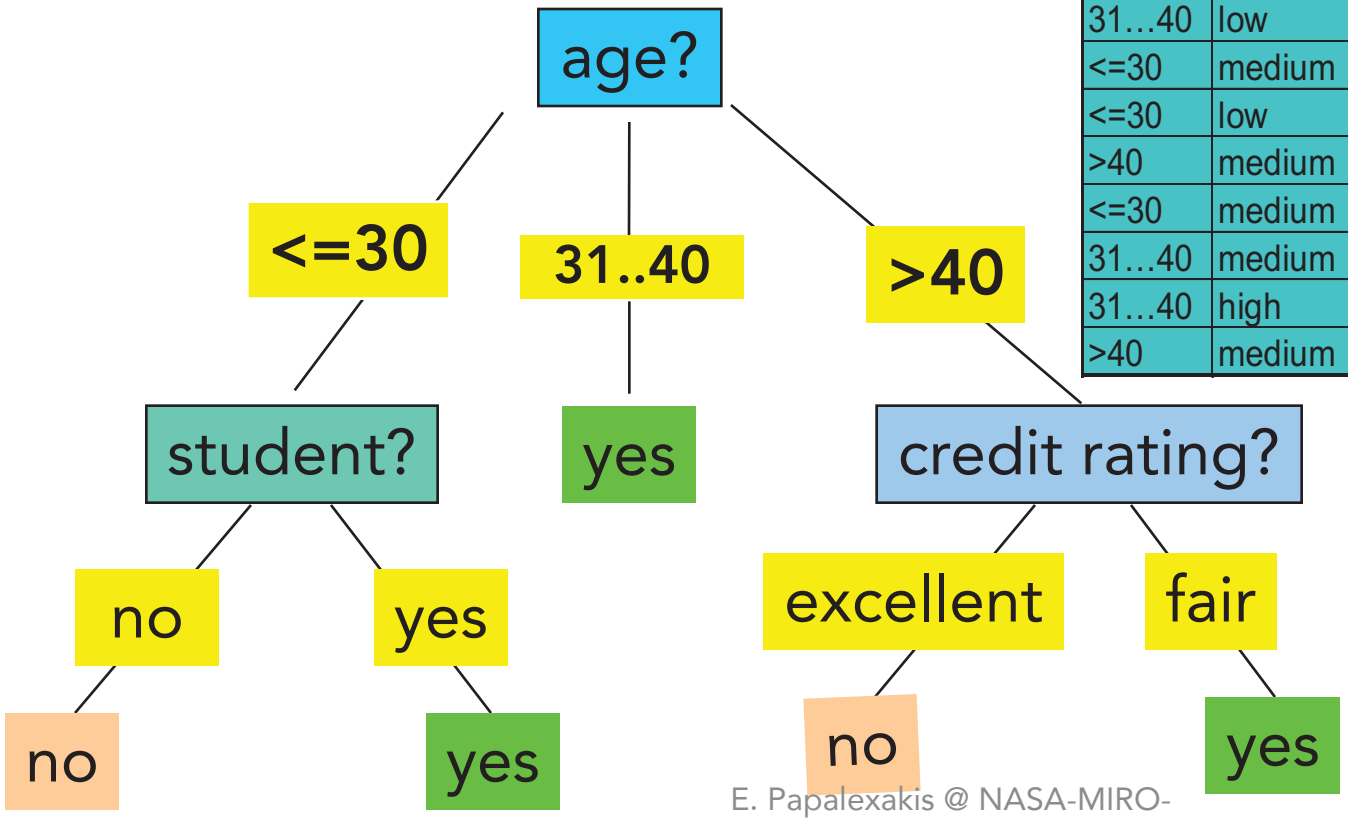

ROUTE
66


Overview of Methods
K-means Clustering
Introduction
Decision Trees
Matrix Factorization

# Decision Tree Induction: An Example

- Training data set: Buys_computer
- Resulting tree:

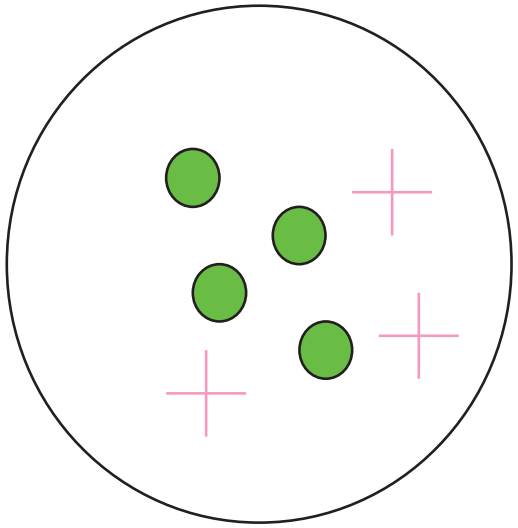
| age | income | student | credit_rating | buys_computer |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |



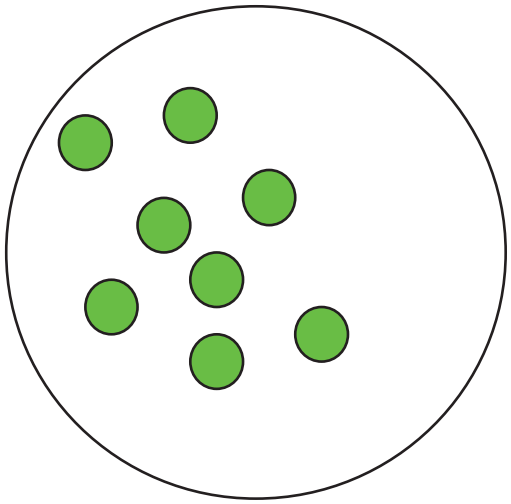
# Algorithm for Decision Tree Induction

- Basic (greedy) algorithm
  - Tree is constructed in a **top-down recursive divide-and-conquer manner**
  - At start, all the training examples are at the root
  - Features are typically categorical
    (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected features
  - Features to partition/split on are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- Most DTs use variations of this algorithm
  - What changes is how the split is made (criterion and number of children nodes)

# How to split?

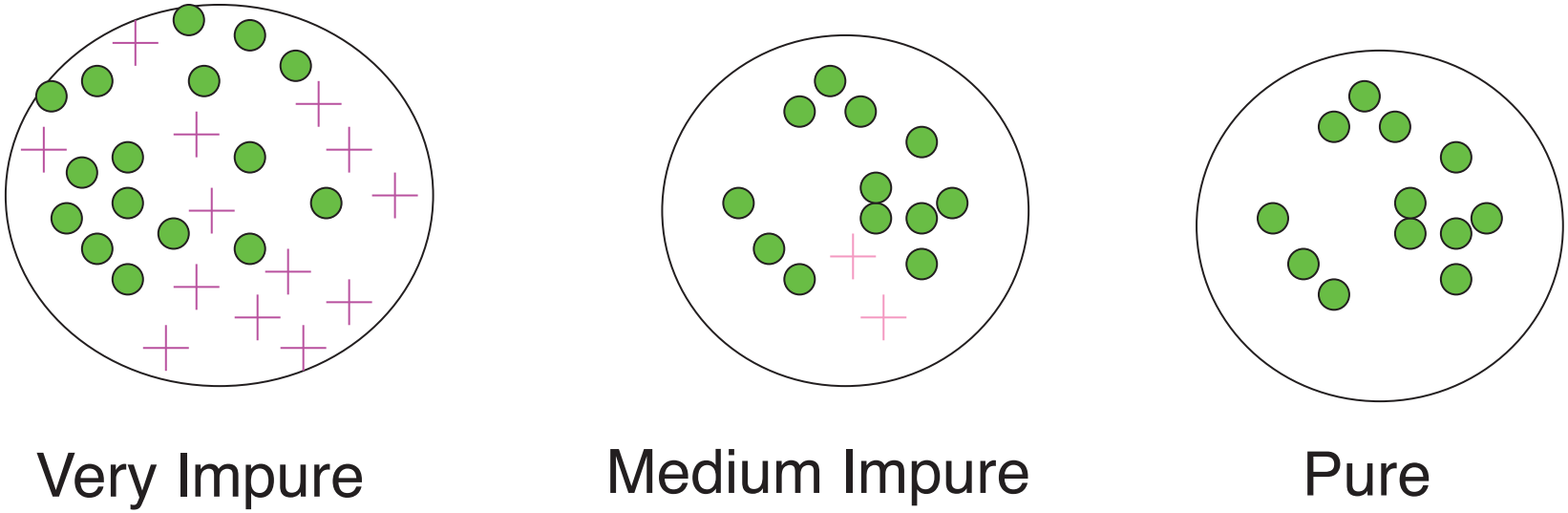


Need to create a “pure” set of data points



Source: http://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

# Levels of “impurity"



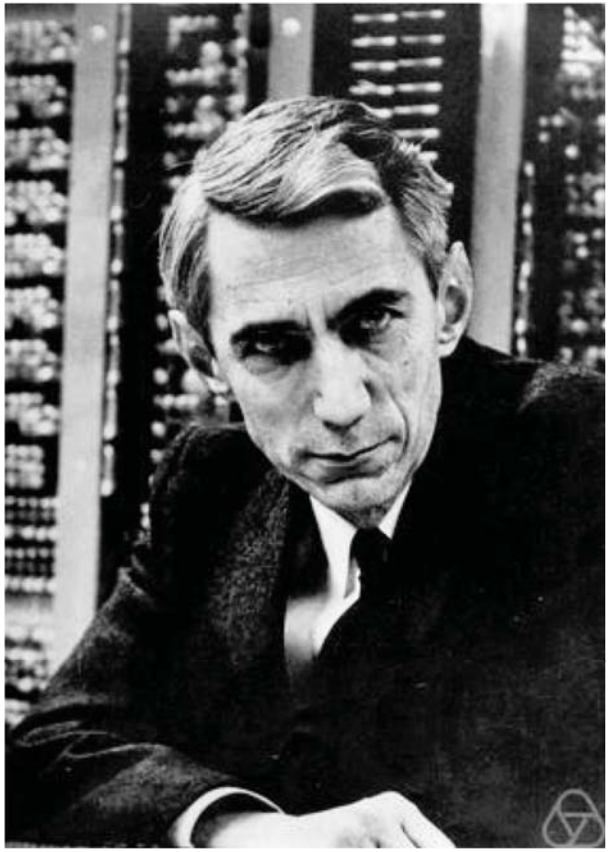
Very Impure

Medium Impure

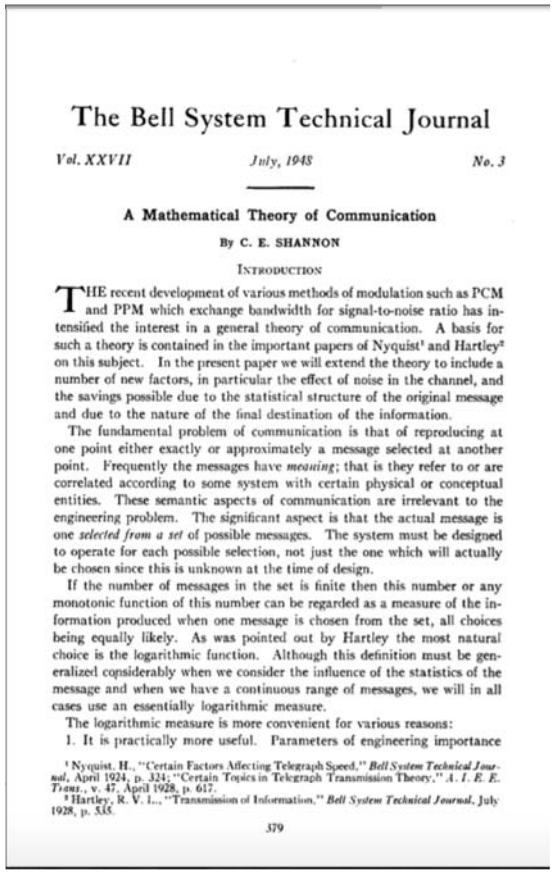
Pure

Source: http://homes.cs.washington.edu/~shapiro/EE596/notes/InfoGain.pdf

# Information Theory



Img from wikipedia.org



The Bell System Technical Journal

Vol. XXVII July, 1948 No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

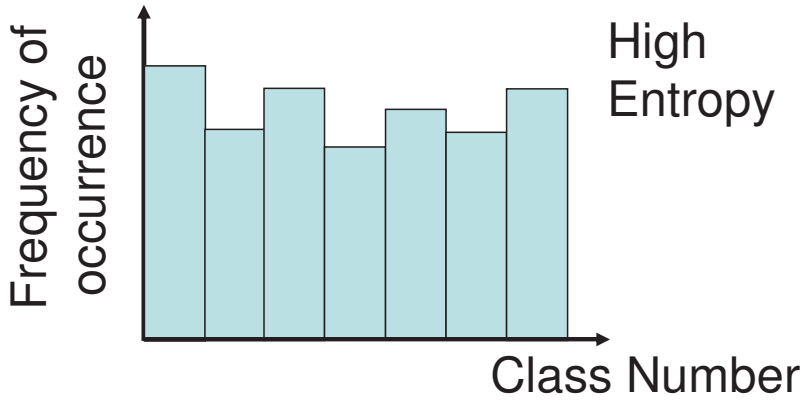
[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

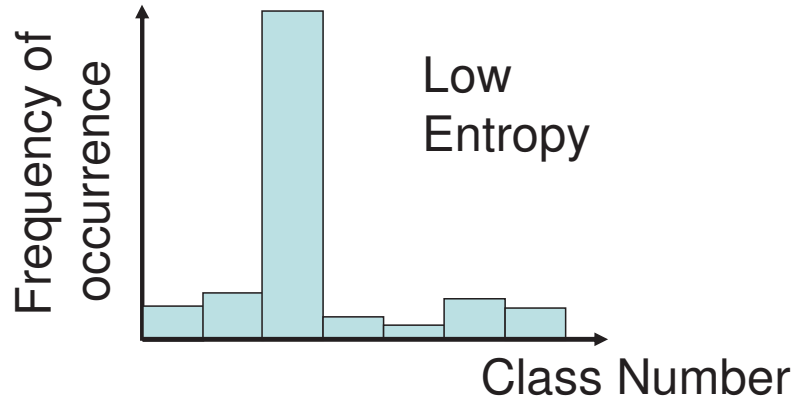
[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

379

http://ieeexplore.ieee.org/document/6773024/?arnumber=6773024

# Entropy as "impurity" measure




$$H = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Entropy quantifies uncertainty

Source: http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15381-s06/www/DTs.pdf

# Attribute Selection Measure: Information Gain (ID3)

- Select the feature with the highest information gain
  - Measure how "informative" a split is
- Let $p_i$ be the probability that an arbitrary data point in D belongs to class $C_i$, estimated by $|C_{i,D}|/|D|$

# Attribute Selection Measure: Information Gain (ID3)

- **Expected information** (entropy) needed to classify a data point in D (*info of parent node*):

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- **Information** needed (after using A to split D into v partitions) to classify D (*avg. entropy of children nodes*):

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

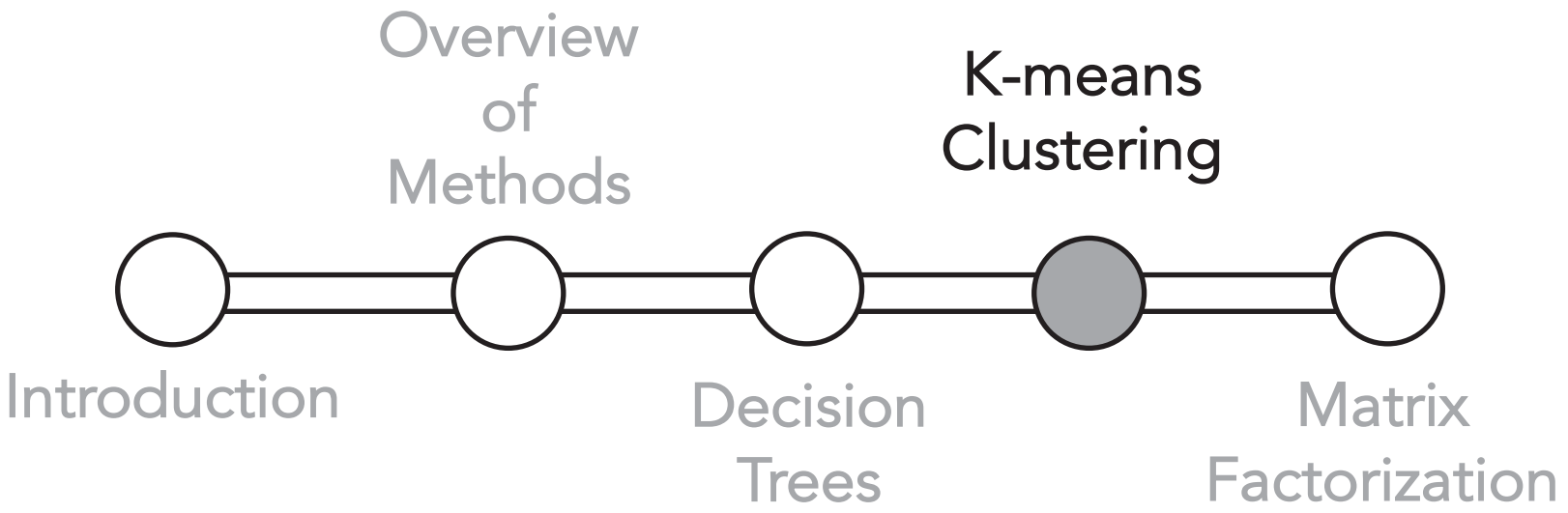
(Dj is partition j)

- **Information gained** by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

# Roadmap


ROUTE
66

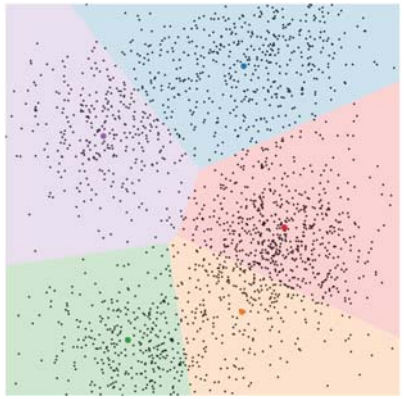

Overview of Methods
K-means Clustering
Introduction
Decision Trees
Matrix Factorization

# Partitioning Clustering

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where $c_i$ is the centroid or medoid of cluster $C_i$)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2$$

**E** (sometimes called **SSE** from Sum Squared Error) measures the total error of approximating every data point p by its assigned centroid $c_i$



$(p\text{-}c_i)^2$: Euclidean dist of point p from its assigned centroid $c_i$

# Partitioning Clustering

- Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - **Optimal Solution**:
    - Exhaustively enumerate all partitions
    - NP-Hard

# Partitioning Clustering

- Given *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - **Heuristic methods**: *k-means* algorithm
    - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster

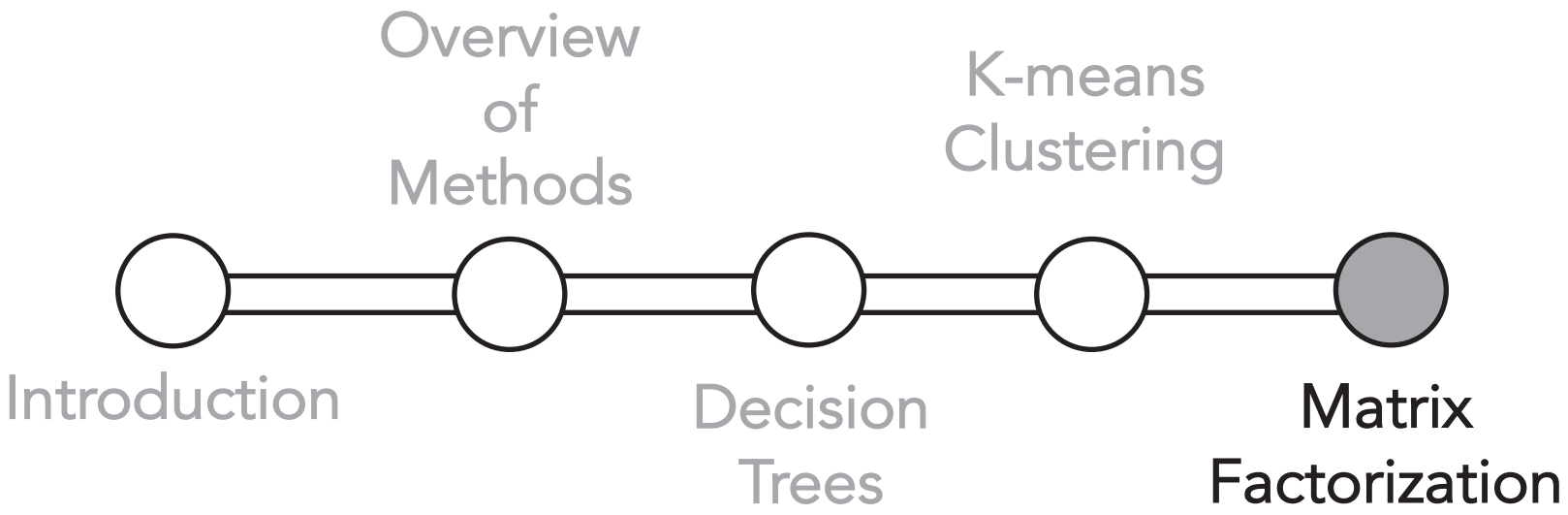
# The K-means Algorithm

Also known as Lloyd's Algorithm

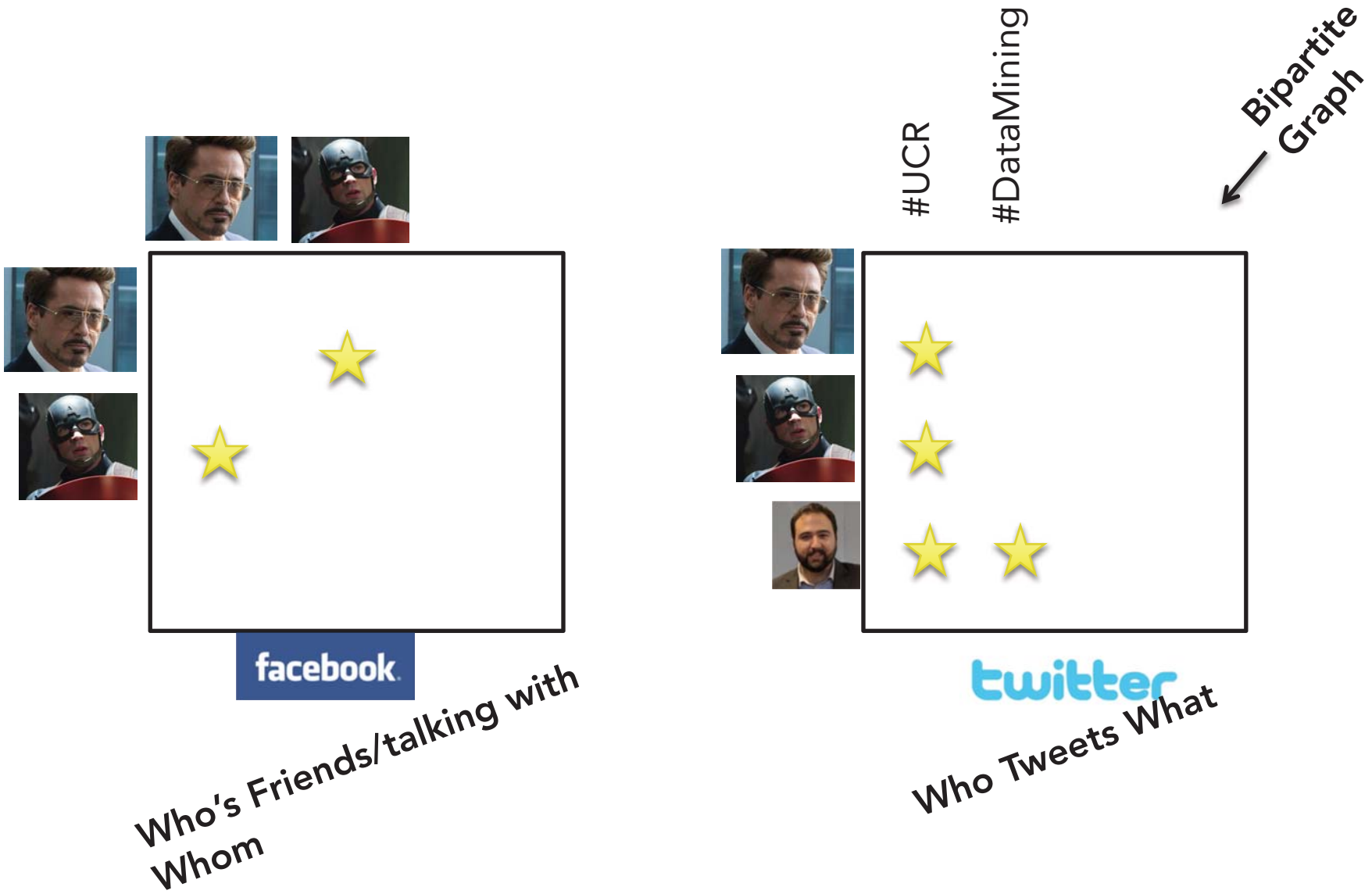
1. Initialize cluster centroids:
   - ✧ Pick *k* points at random and set as centroids representing each cluster
2. Repeat while cluster assignments don't change:
   a) Assign each point to the nearest centroid
   b) Given new assignments, compute new cluster centroids as mean of all points in cluster

# Roadmap

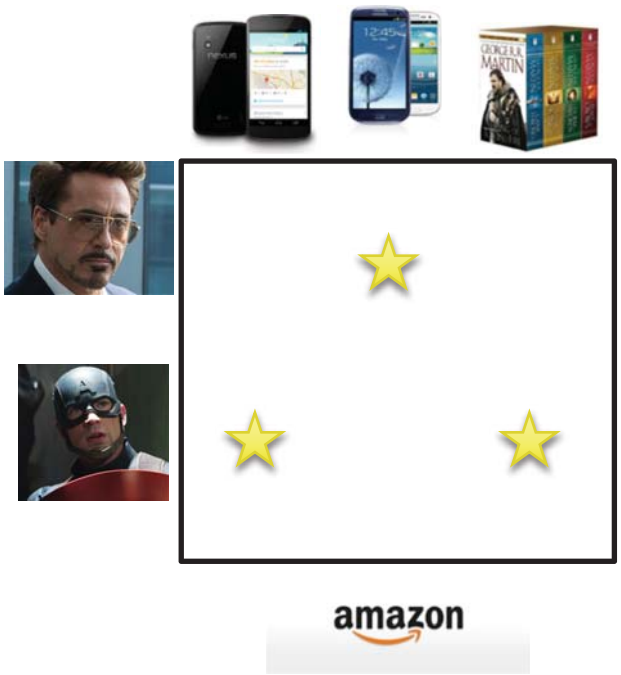

ROUTE
66


Introduction
Overview of Methods
Decision Trees
K-means Clustering
Matrix Factorization

# Matrix Data Representation



# Matrix Data Representation





★ may be binary or a rating

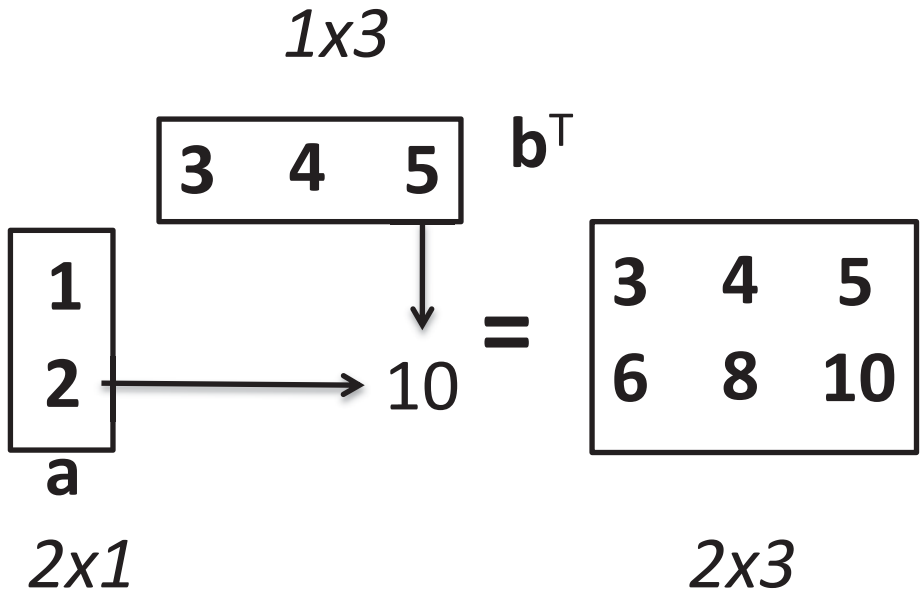
# Recommendation Systems


- **Problem**: Given a user's viewing/rating pattern, recommend new movies to watch
  - Based on similar movies
  - Based on similar users
- $1M prize!
- Collaborative Filtering

http://netflixprize.com/

# What is the *rank* of a matrix?

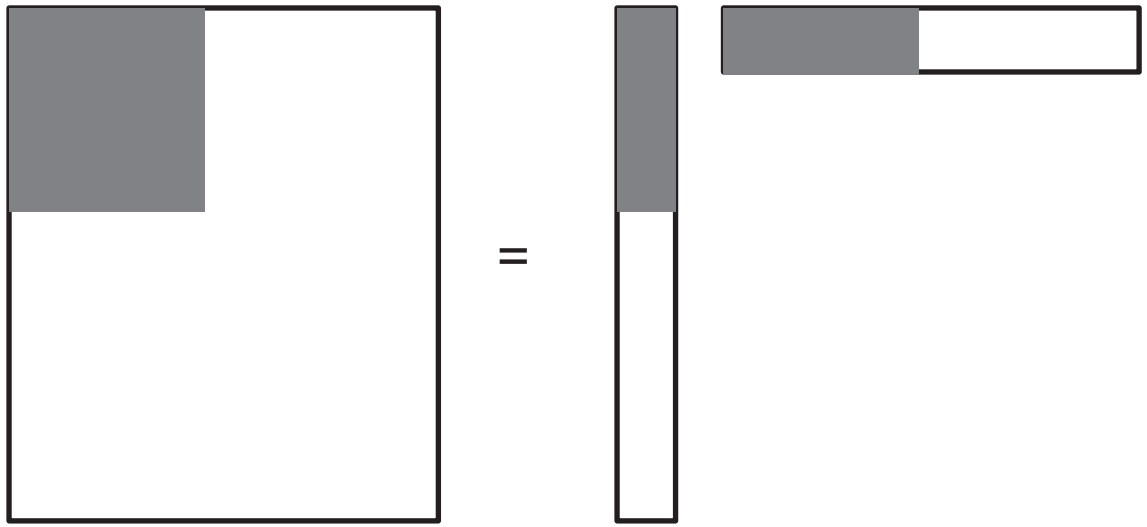
# Outer Product




By definition:
**Rank-one matrix**

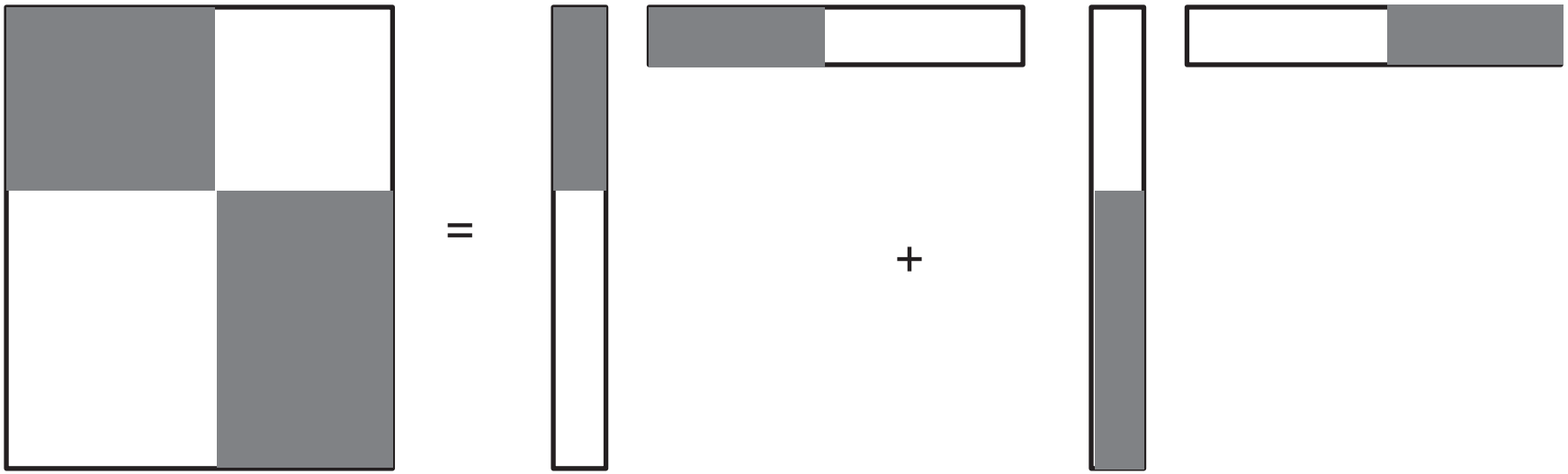
**Matrix Rank** = Min # of rank-one matrices that add up to that matrix

# Matrix Factorization/Decomposition



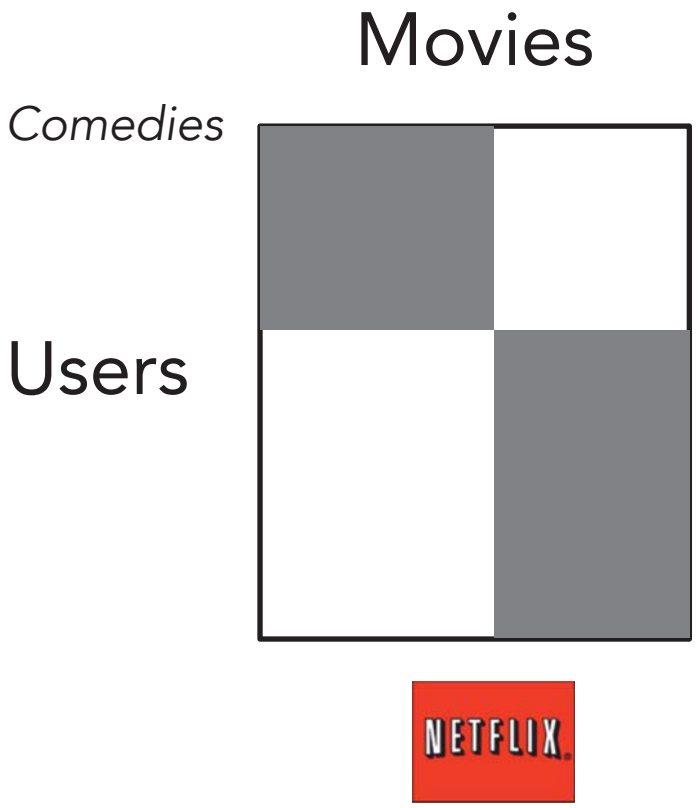

Rank 1

# Matrix Factorization/Decomposition



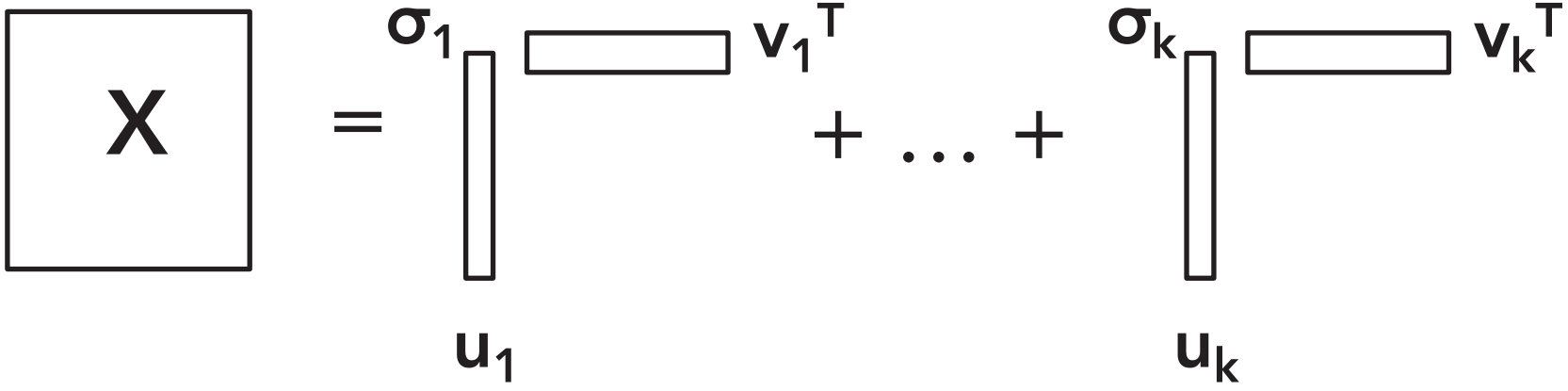

Rank 2

# Example



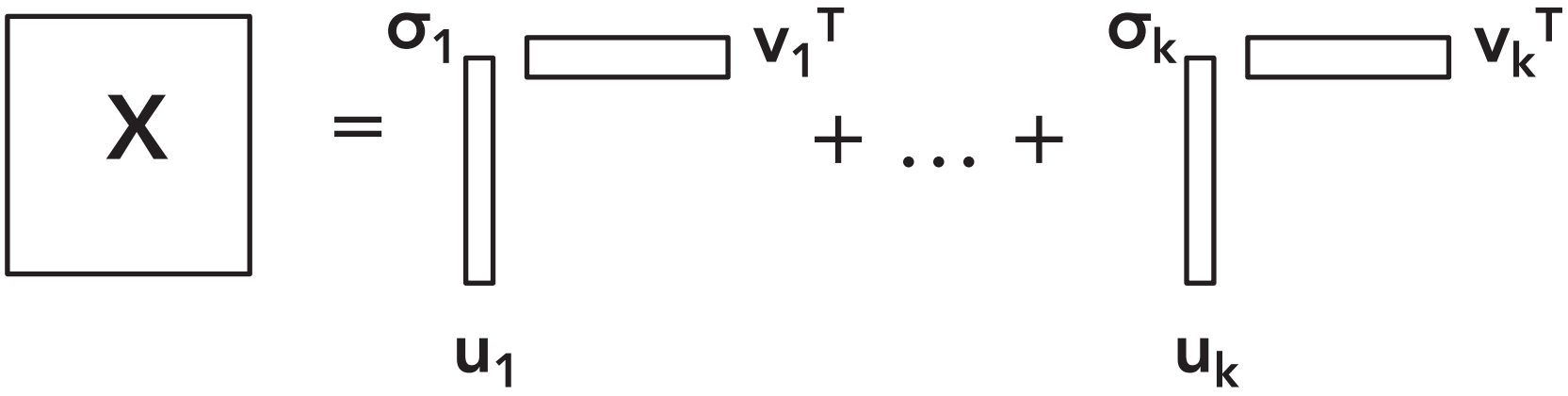

Each block in the data is a latent (“hidden”) concept

# Singular Value Decomposition



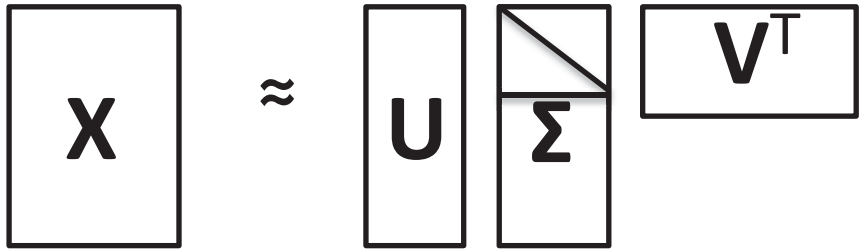

- $\mathbf{u}_i$, $\mathbf{v}_i$ are left and right singular vectors
- $\mathbf{u}_i$ are orthogonal: $\mathbf{u}_i^T \mathbf{u}_j = 0$ for diff i,j (same for $\mathbf{v}_i$)
- $\boldsymbol{\sigma}_i$ are the singular values (non-negative - sorted in desc. order)

# Singular Value Decomposition



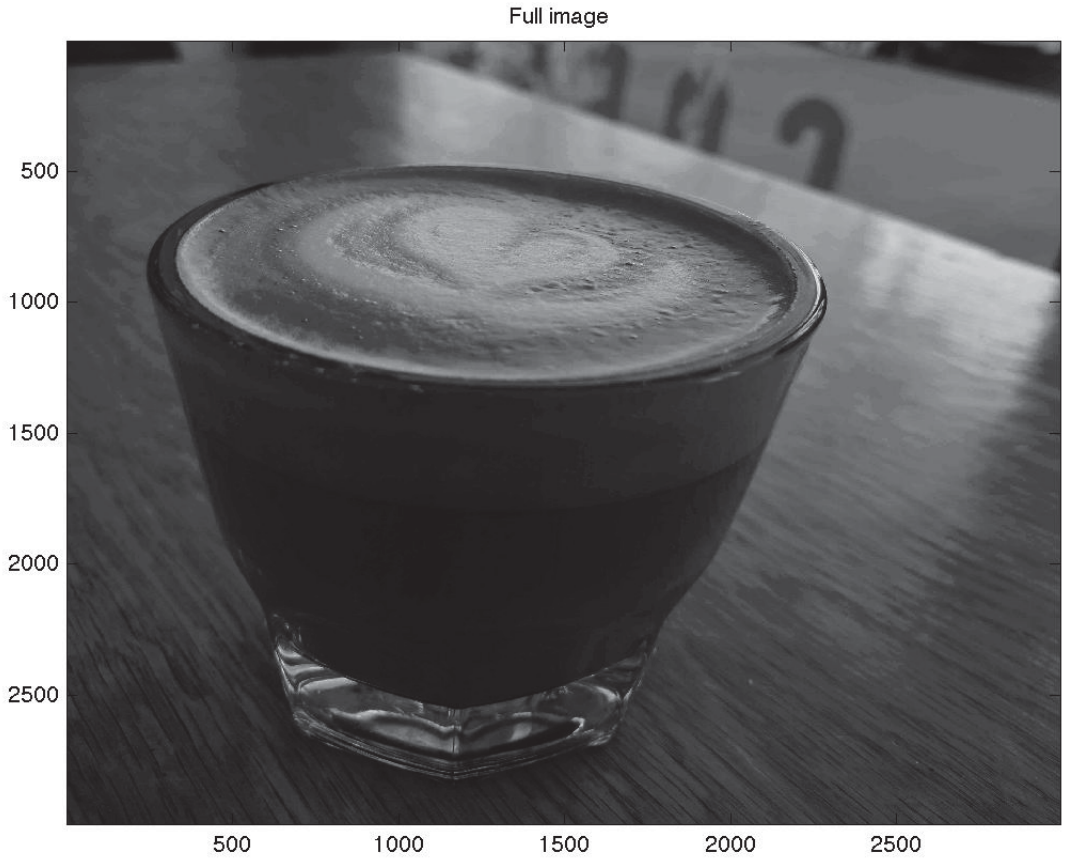

- If k = rank(X) then we have equality
- If k < rank(X) we have the best rank k approximation that minimizes the squared error (Eckart Young Theorem)

# Singular Value Decomposition



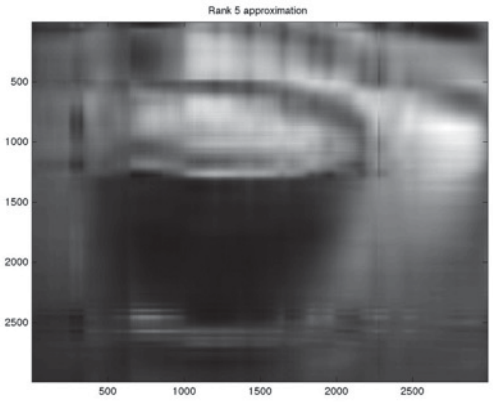

- Best low rank approx. of **X**
- Gives us PCA in centered data
- "Workhorse" of data analysis
  - Best projection of arbitrary dimensional points to 2 dims
    - Take first 2 cols of U and plot(**U**(:,1),**U**(:,2))

# SVD for Image Compression

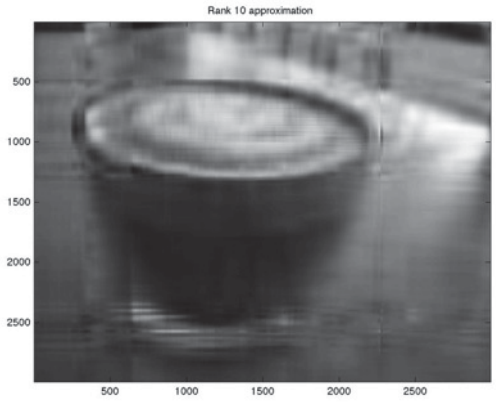


# SVD for Image Compression

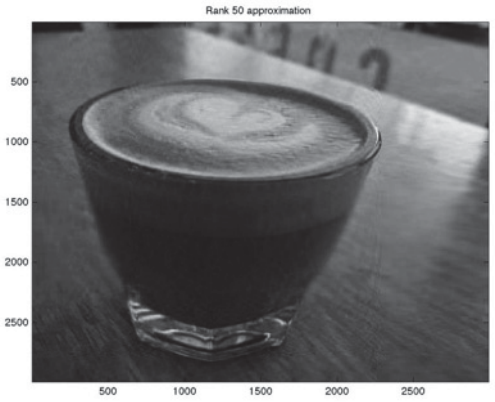
k = 5



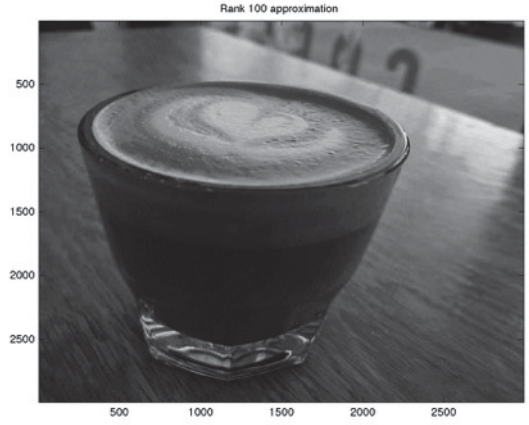

k = 10



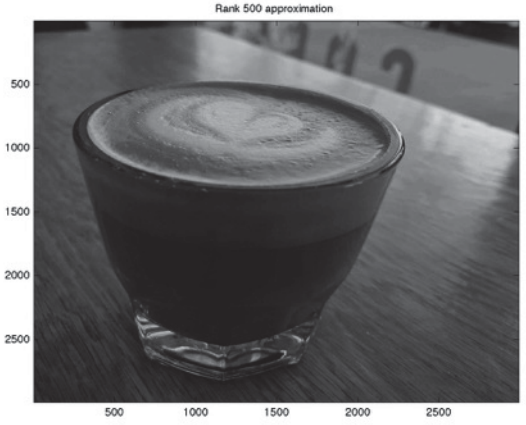

k = 50



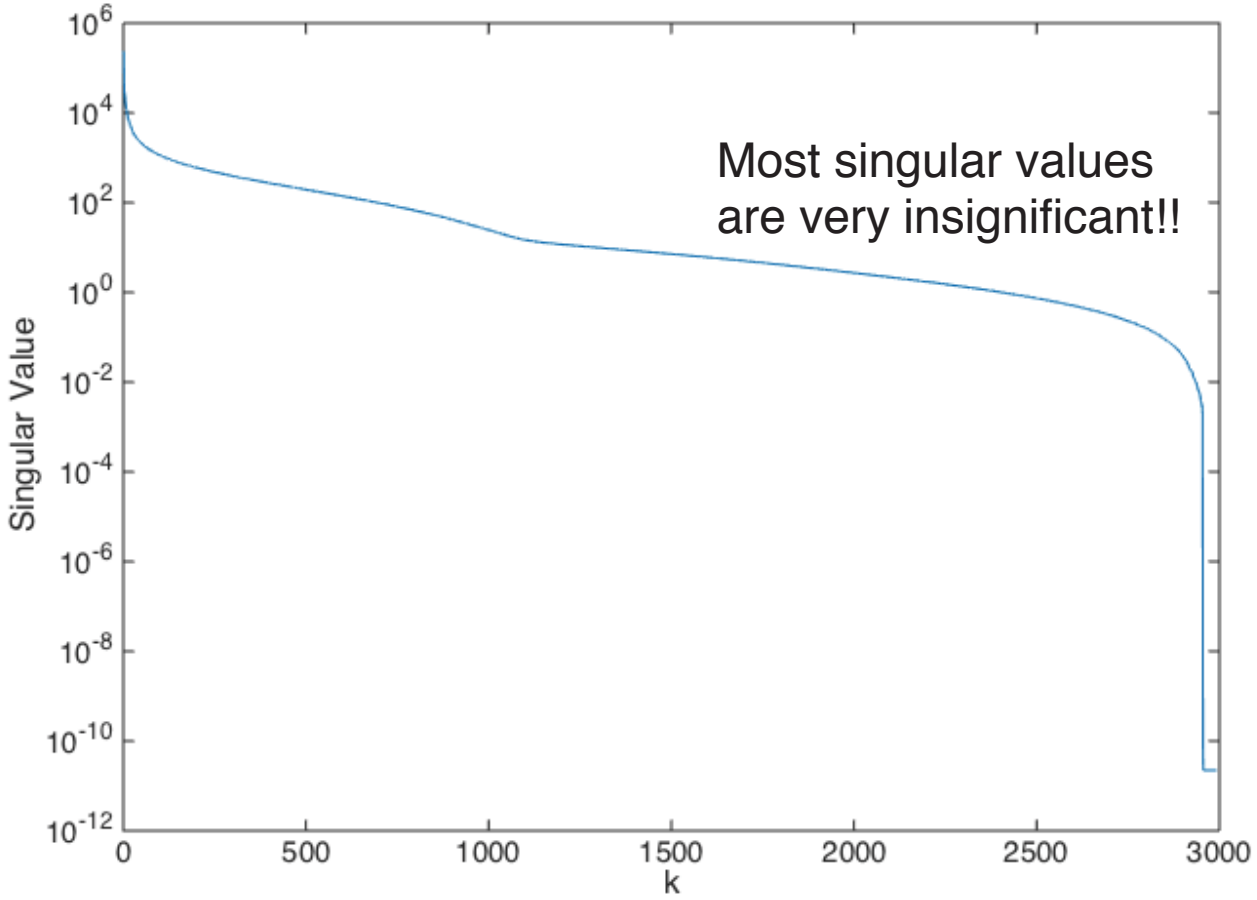

k = 100




k = 500



# A closer look

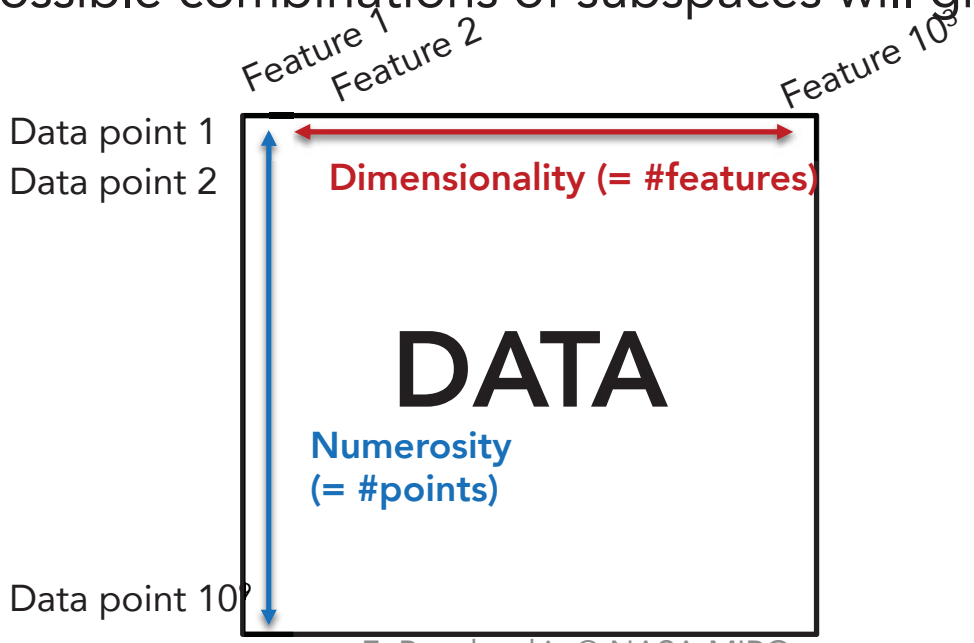


# You can try this at home!

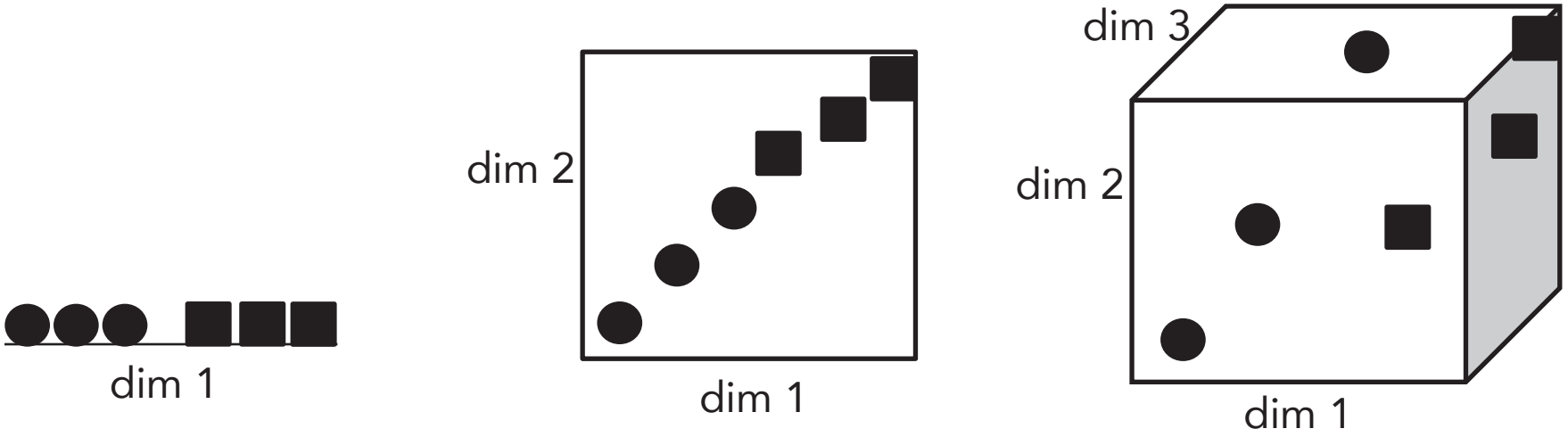
```
clear all;close all;clc
X = imread('cortado.jpg');
X = rgb2gray(X);
imagesc(X);colormap(gray)
title('Full image')
saveas(gcf,'img/full.jpg','jpg')
[U,S,V] = svd(double(X));
for k = [5 10 50 100 500]
    X_new = U(:,1:k)*S(1:k,1:k)*V(:,1:k)';
     imagesc(X_new);colormap(gray)
     title(sprintf('Rank %d approximation',k))
     saveas(gcf,sprintf('img/rank-%d.jpg',k),'jpg')
end
```

# Curse of Dimensionality

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially

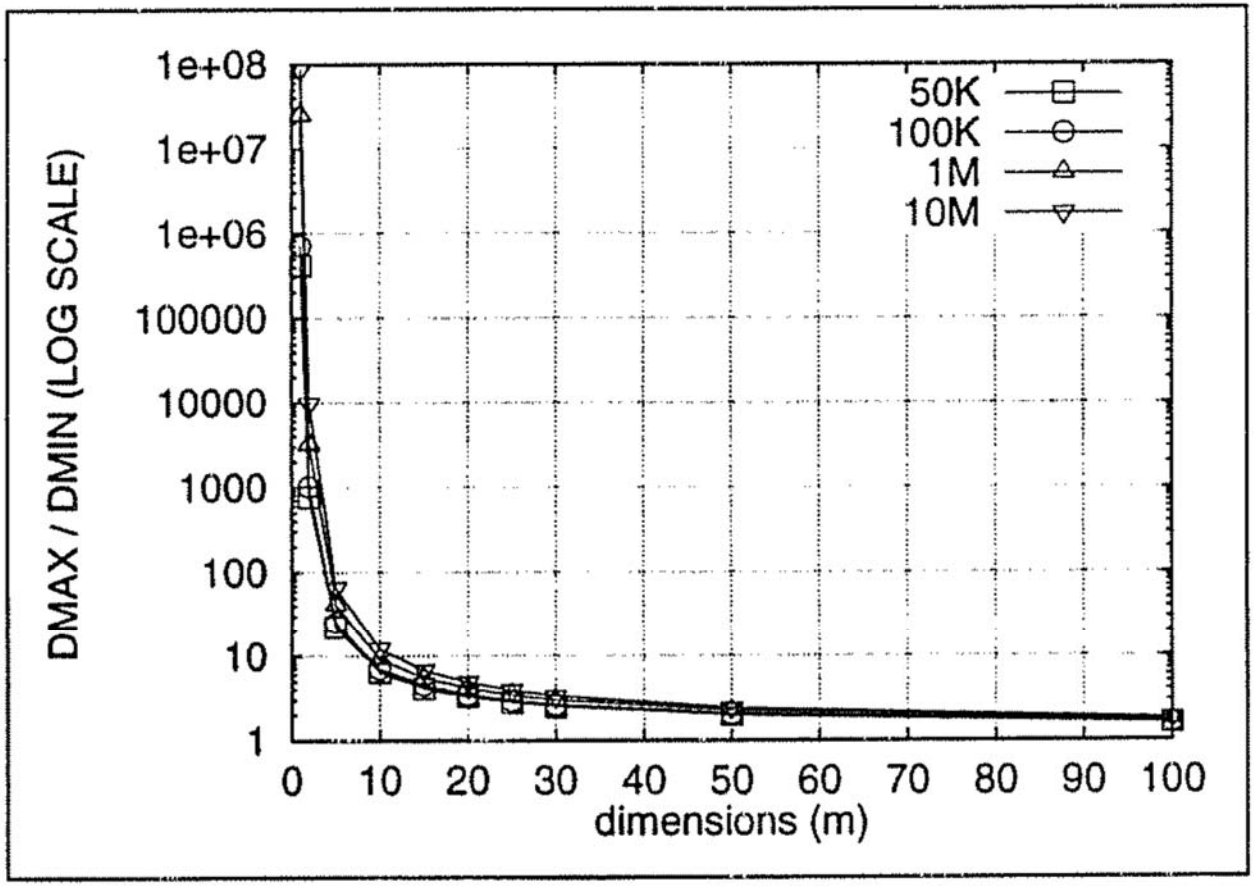


# Curse of Dimensionality




As dims grow, space becomes sparser!
Need exponentially more data points to be able to “learn” something useful

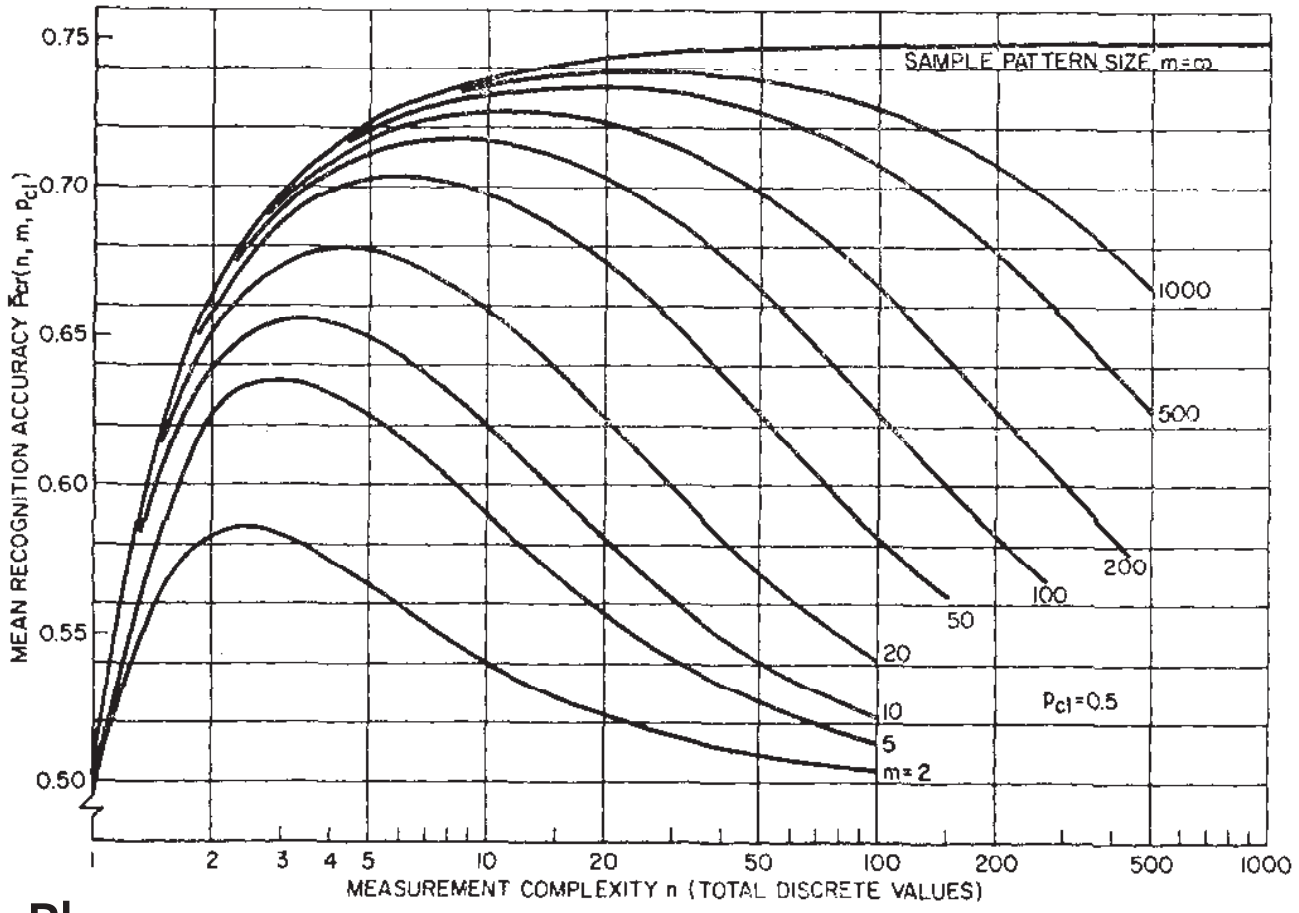
# Curse of Dimensionality




Ratio of furthest to closest point diminishes as the #dims grows!!!

Beyer et al., "When is Nearest Neighbor Meaningful?"
https://minds.wisconsin.edu/bitstream/handle/1793/60174/TR1377.pdf

# Curse of Dimensionality

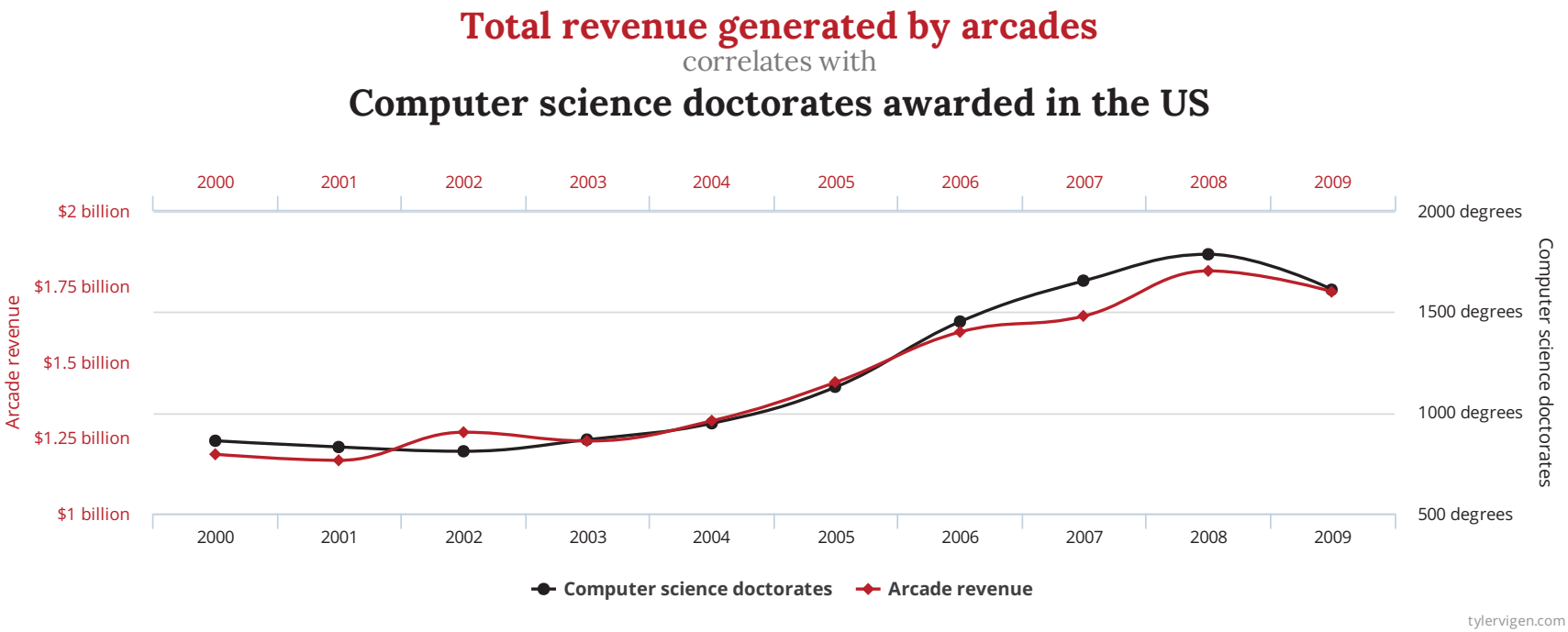



**Hughes Phenomenon:**

Model performance increases and then rapidly drops as dims increase!

Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers"

# Curse of Dimensionality

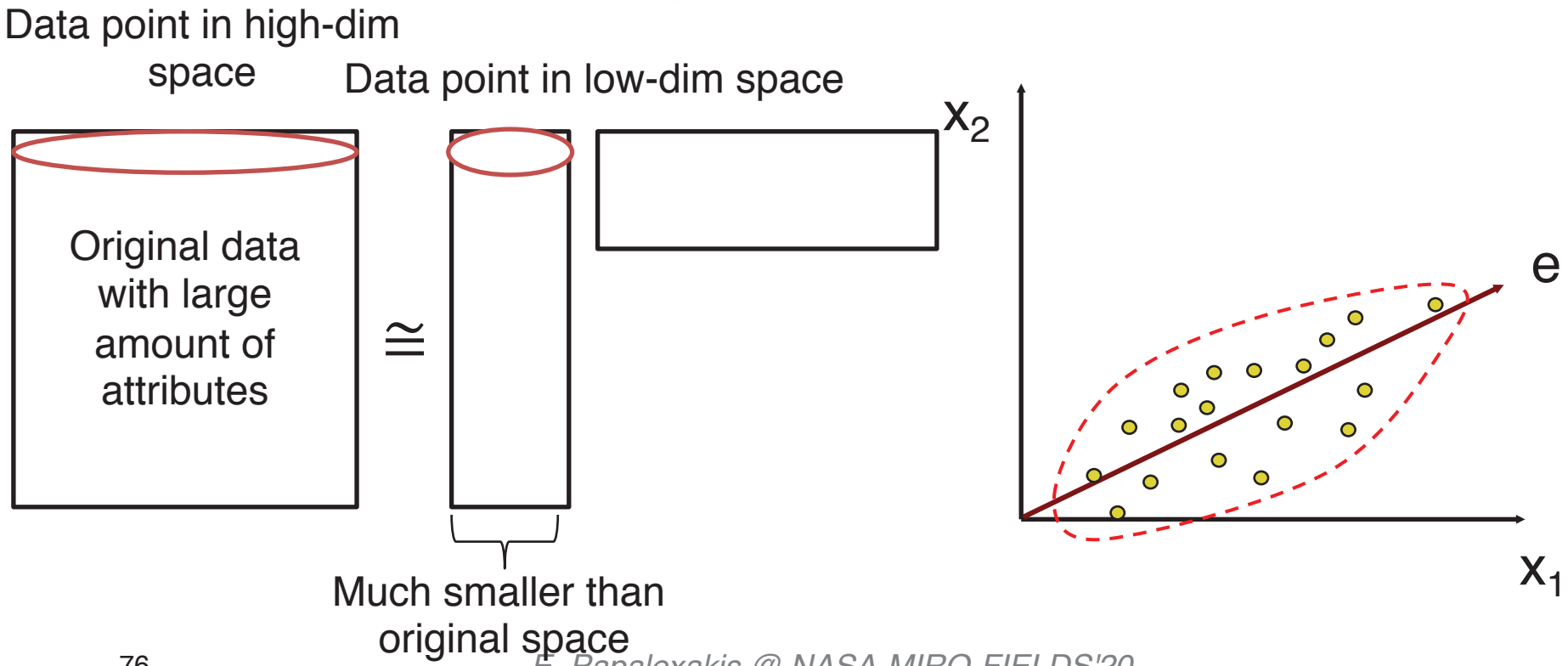



Combinations of dims grows exponentially
More likely to find spurious correlations!

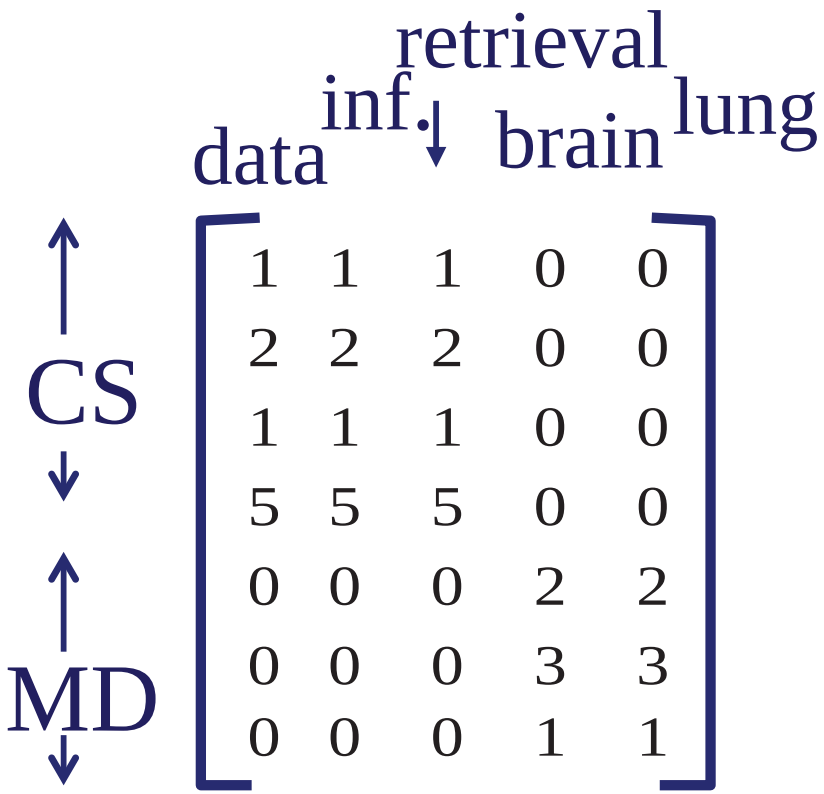
http://www.tylervigen.com/spurious-correlations

# Principal Component Analysis (PCA)

- Center (aka make zero-mean) and normalize the data
- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- Works for numeric data only

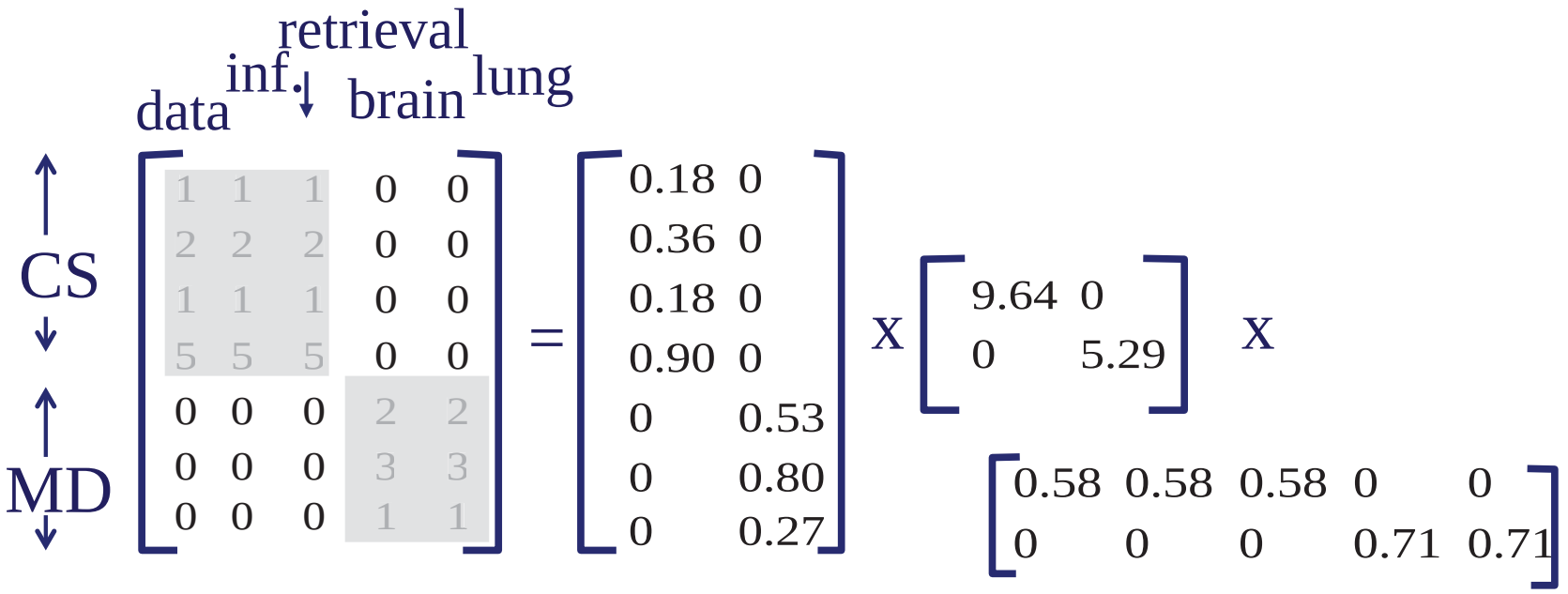


# Latent Semantic Indexing



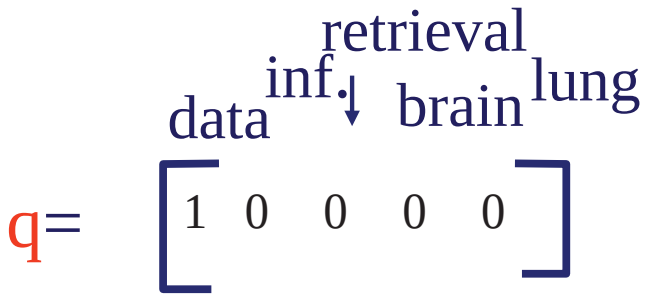

Slide by Christos Faloutsos

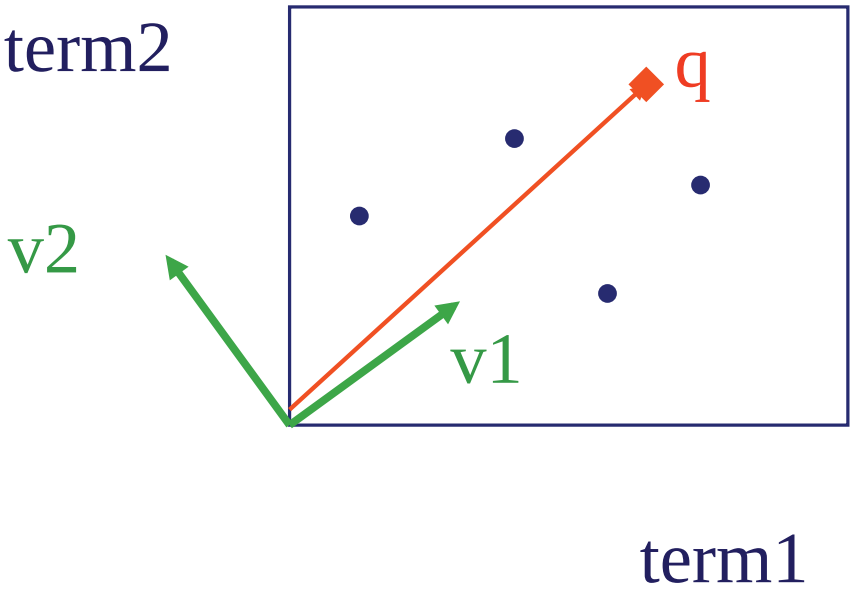
# Latent Semantic Indexing




Slide by Christos Faloutsos

# Querying in LSI

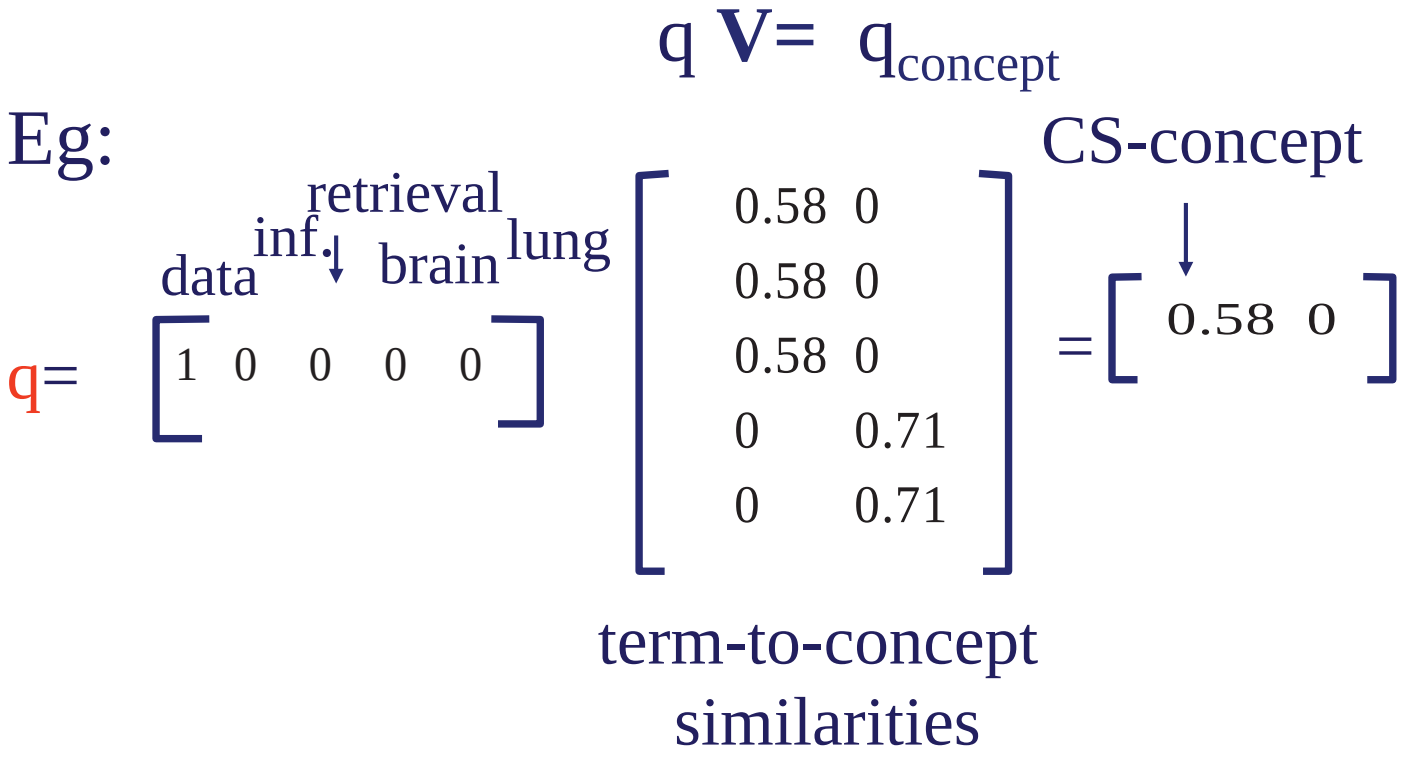



Need to transform the query from the keyword space to the LSI space

Slide by Christos Faloutsos

# Querying in LSI

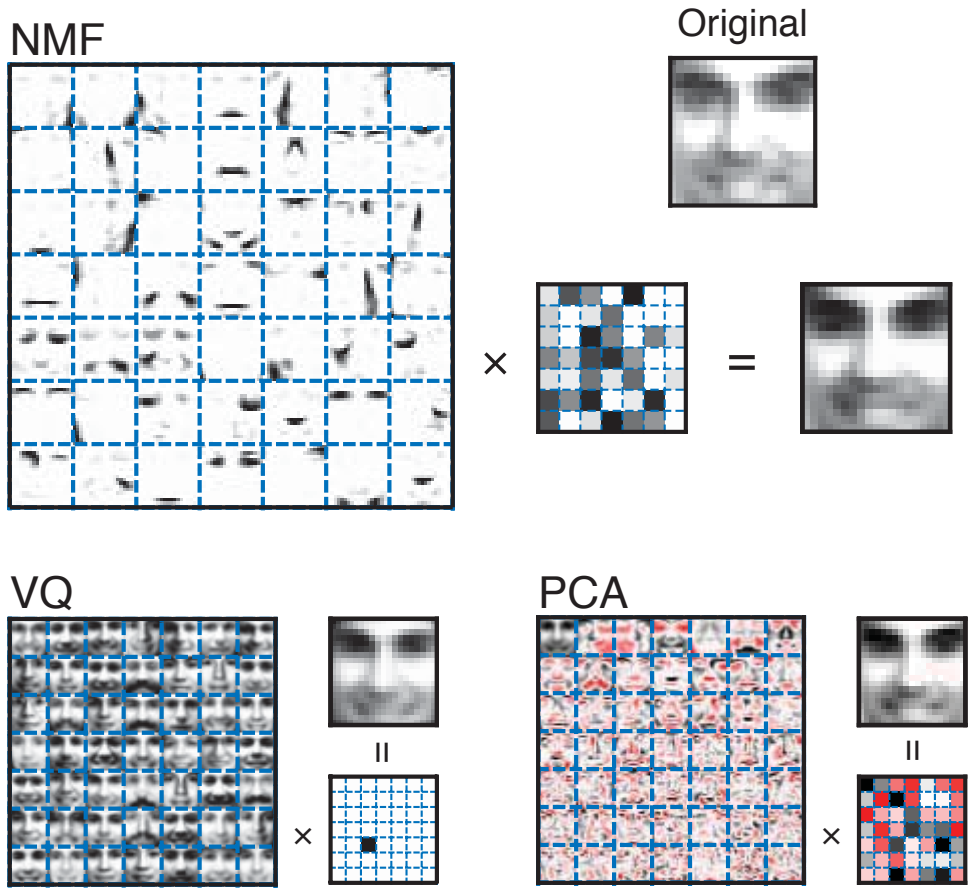



Slide by Christos Faloutsos

# Non-Negative Matrix Factorization (NMF)

$$X \approx U V^T$$

- **U**,**V** are element-wise >=0
- Helps for interpretation
- Good for **sum-of-parts** representation




Lee and Seung. "Learning the parts of objects by non-negative matrix factorization." Nature (1999)