ANALYTICAL TECHNIQUES FOR DATA VISUALIZATION

CSE – 537
Artificial Intelligence
Professor Anita Wasilewska

GROUP 2

TEAM MEMBERS:

SAEED BOOR BOOR - 110564337

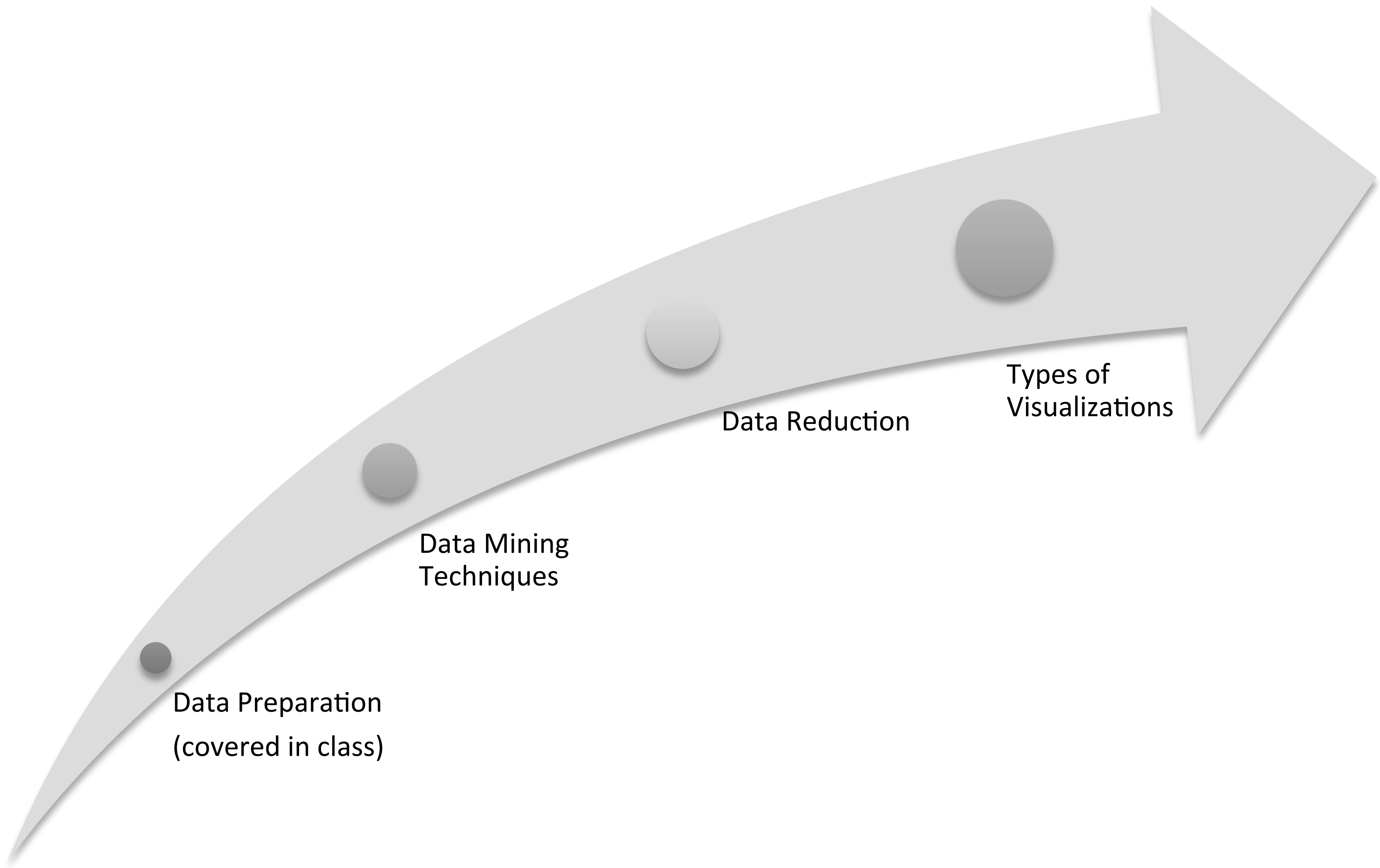
SHIH-YU TSAI - 110385129

HAN LI – 110168054

SOURCES

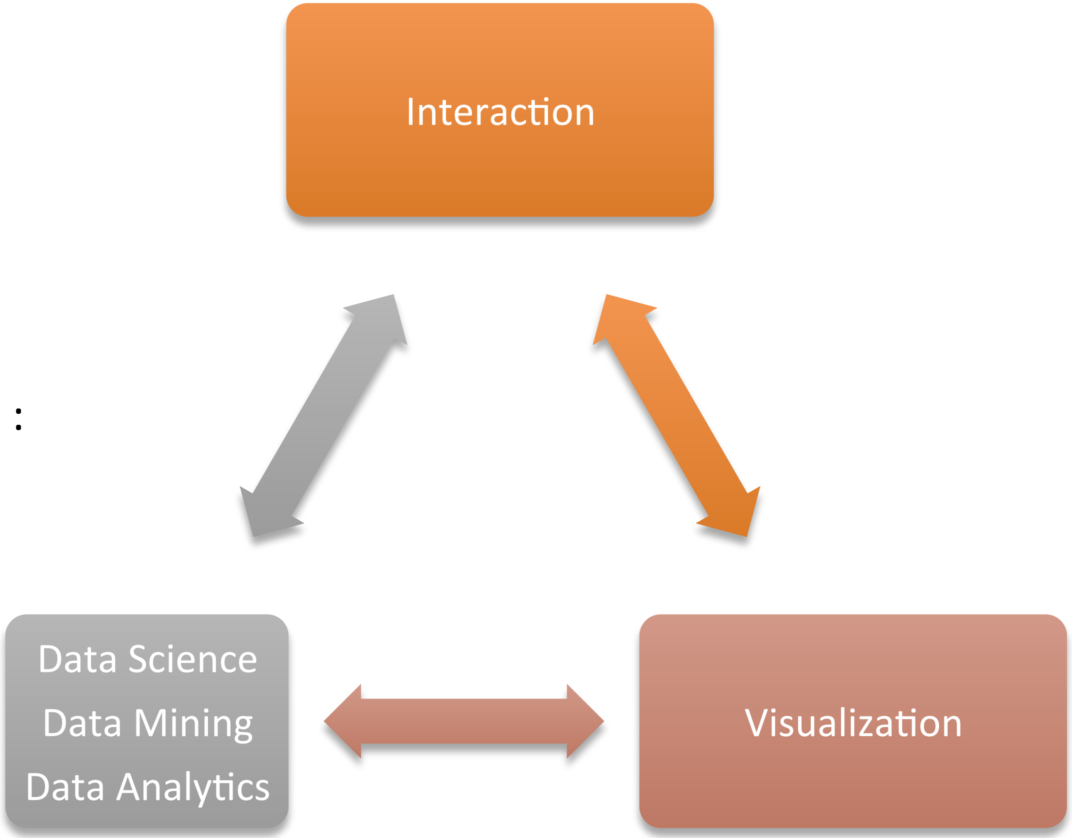
- Slides 9 – 10: Visualization and Visual Analytics, Fundamental Tasks, Klaus Muller, SBU
- Slides 13 – 17: Clustering, Eamonn Keogh, UC Riverside
- Slides 19 – 26 : Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9
- Slides 28 – 30: Visualization and Visual Analytics, Fundamental Tasks, Klaus Muller, SBU
- Extra Sources:
  - Lecture Notes by Anita Wasilewska, Chapter 2: Preprocessing, Chapter 6: Classification
  - http://bl.ocks.org/ for types of visualizations
  - https://www.youtube.com/watch?v=_aWzGGNrcic_for K-Means algorithm

STEPS (PRESENTATION OUTLINE)



PROLOGUE

- Definition of Visual Analytics :



PROLOGUE

- What: Analytical Techniques for **Data Visualization**
- Why: so much data => visualize (see, perceivable)
- How:
  - AI or Machine Learning Methods to read, sample, **clustering**
    - K-means clustering algorithms/E-m algorithms/Nearest neighbor
  - Eliminating dimensions to 2D : **Principal component analysis (PCA)**
  - Types of aesthetic visualizations

DATA CLEANING

- fill in **missing values**
- smooth **noisy data**
- identify or **remove outliers**
- **resolve inconsistencies**

MISSING VALUES

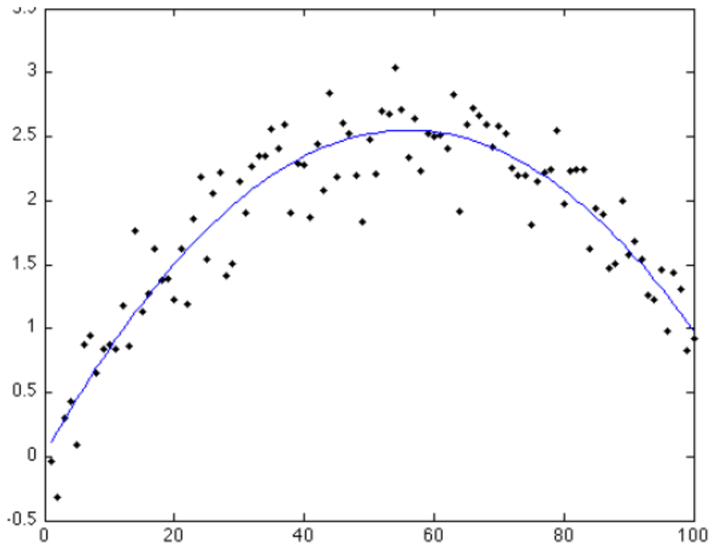
- Data is not always available
  - e. g, many tuples have no recorded value for several attributes, such as customer income in sales data

Age	Income	Team	Gender
23	24,200	Red Sox	M
39	$?_1$	Yankees	F
45	45,390	$?_2$	F

- Ways to solve:
  - Ignore tuple
  - Fill manually
  - Use Global Constant
  - Attribute Mean
  - Most probable value

NOISY DATA

- Why?
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- What to do?
  - Binning Method
  - Clustering
  - Regression
  - Semi Automatic: Computer Algorithm + Human input

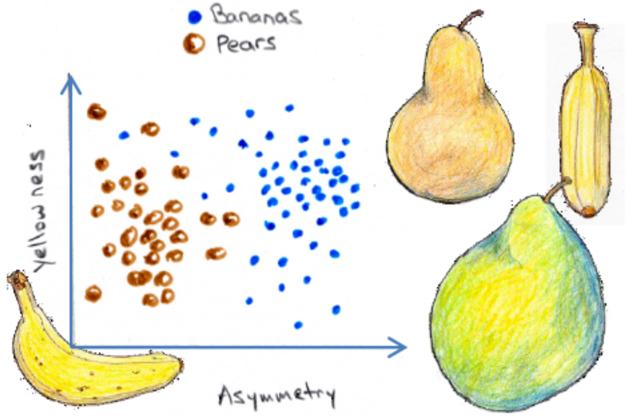


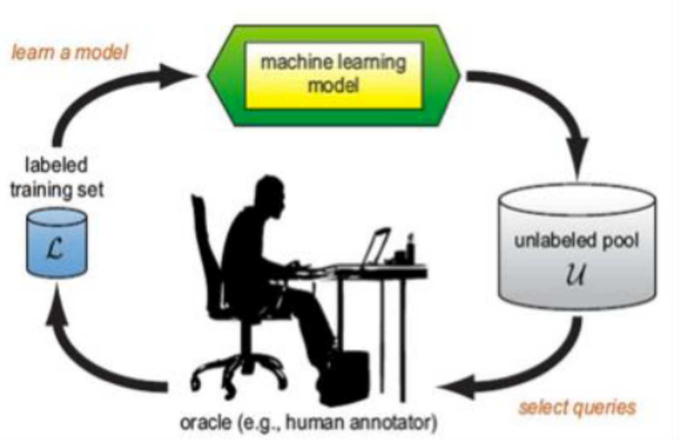
DATA MINING TECHNIQUES – ANALYZE DATA

CLASSIFICATION -> REGRESSION -> CLUSTERING -> SIMILARITY MATCHING -> LINK PREDICTION

TASK # 1: CLASSIFICATION (COVERED IN CLASS)

- Predict which class a member of a certain population belongs to
  - absolute
  - probabilistic
- Requires Classification Model
  - Supervised Learning
  - Unsupervised Learning
- Scoring with a Model
  - each population member gets a score for a particular class/category
  - sort each class or member scores to assign
  - scoring and classification are related



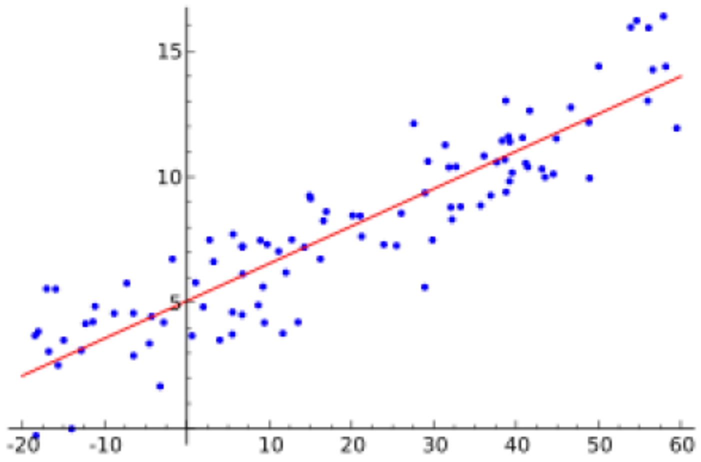


TASK # 2: REGRESSION

- Regression = value estimation

- Fit the data to a function
  - often linear, but does not have to be
  - quality of fit is decisive

- Regression vs. classification
  - classification predicts that **something will happen**
  - regression predicts **how much of it will happen**



TASK # 3: CLUSTERING

- Group individuals in a population together by their similarity
- Cluster methods
  - K-means algorithms
  - E-m algorithms

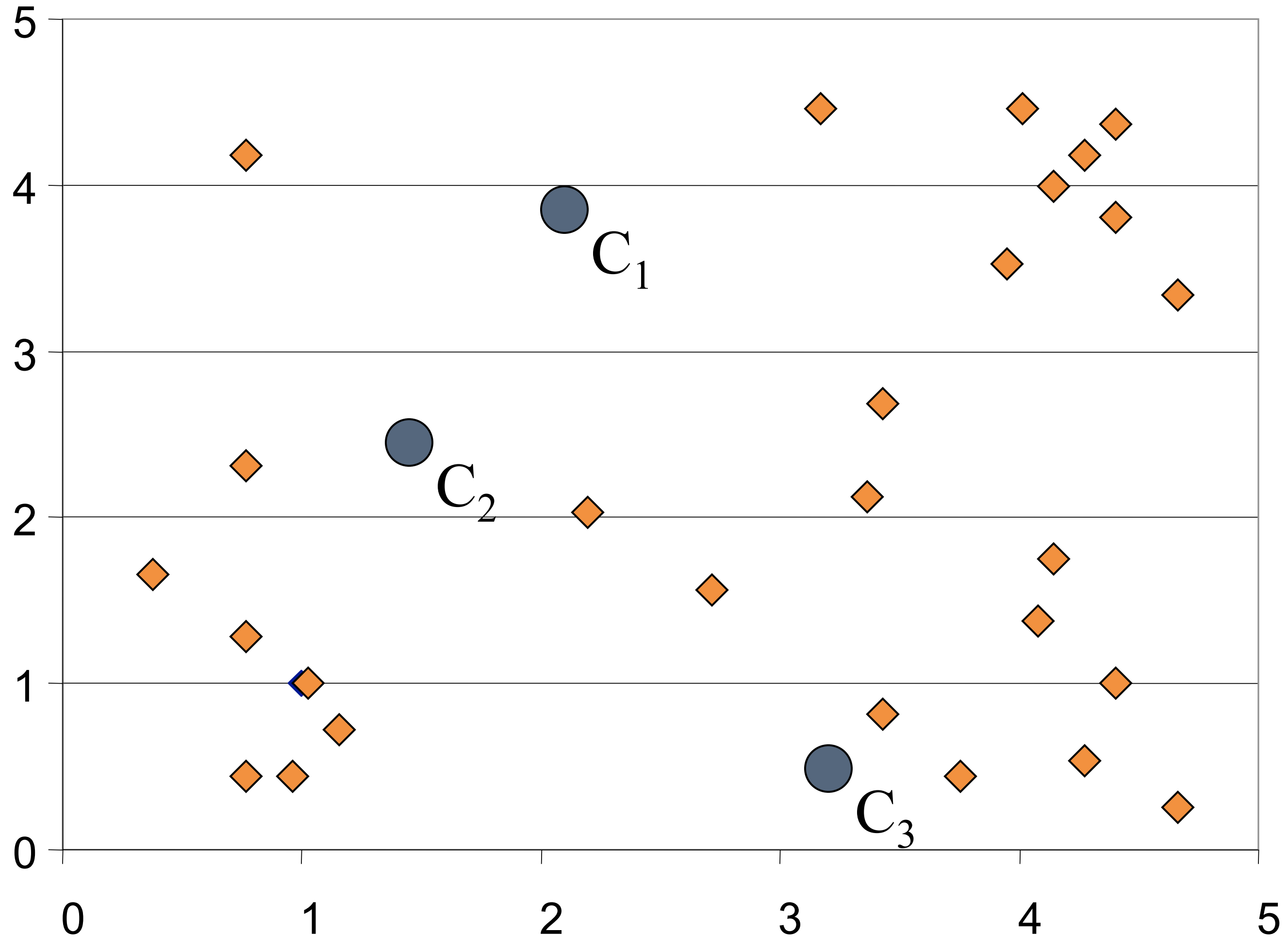
K-MEANS ALGORITHMS

- Input: K, set of points x_1 … x_n
- Place centroids c_1 … c_K at random positions
- Repeat the following procedure until it converges
  - For each x_i
    - Find nearest centroid c_j
    - Assign x_i to cluster j
  - For each cluster j
    - New c_j is the mean of all point x_i in cluster j in previous step

Source: https://www.youtube.com/watch?v=_aWzGGNrcic

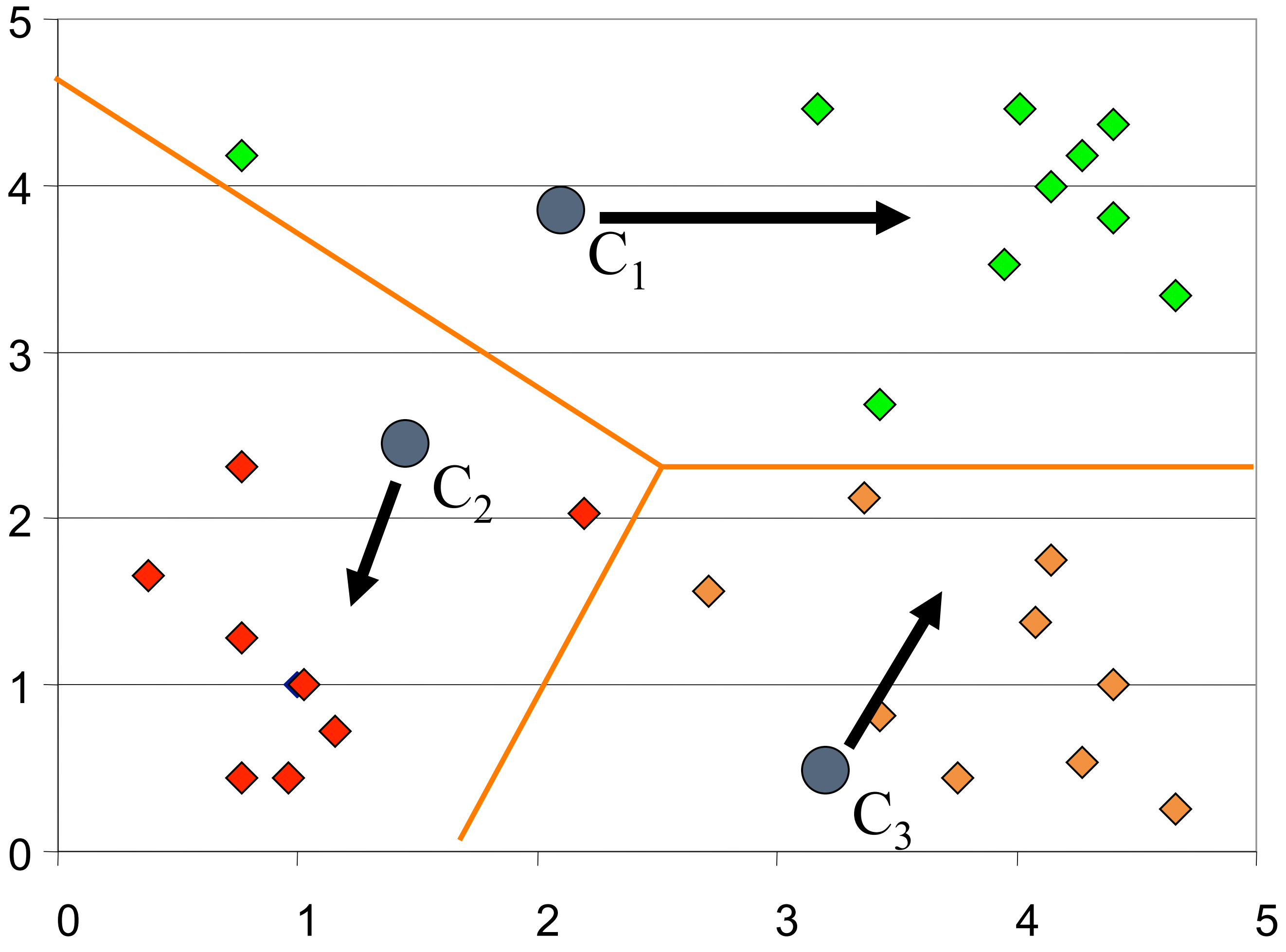
K-MEANS CLUSTERING: STEP 1

Algorithm: K-means, K =3, Distance Metric: Euclidean Distance



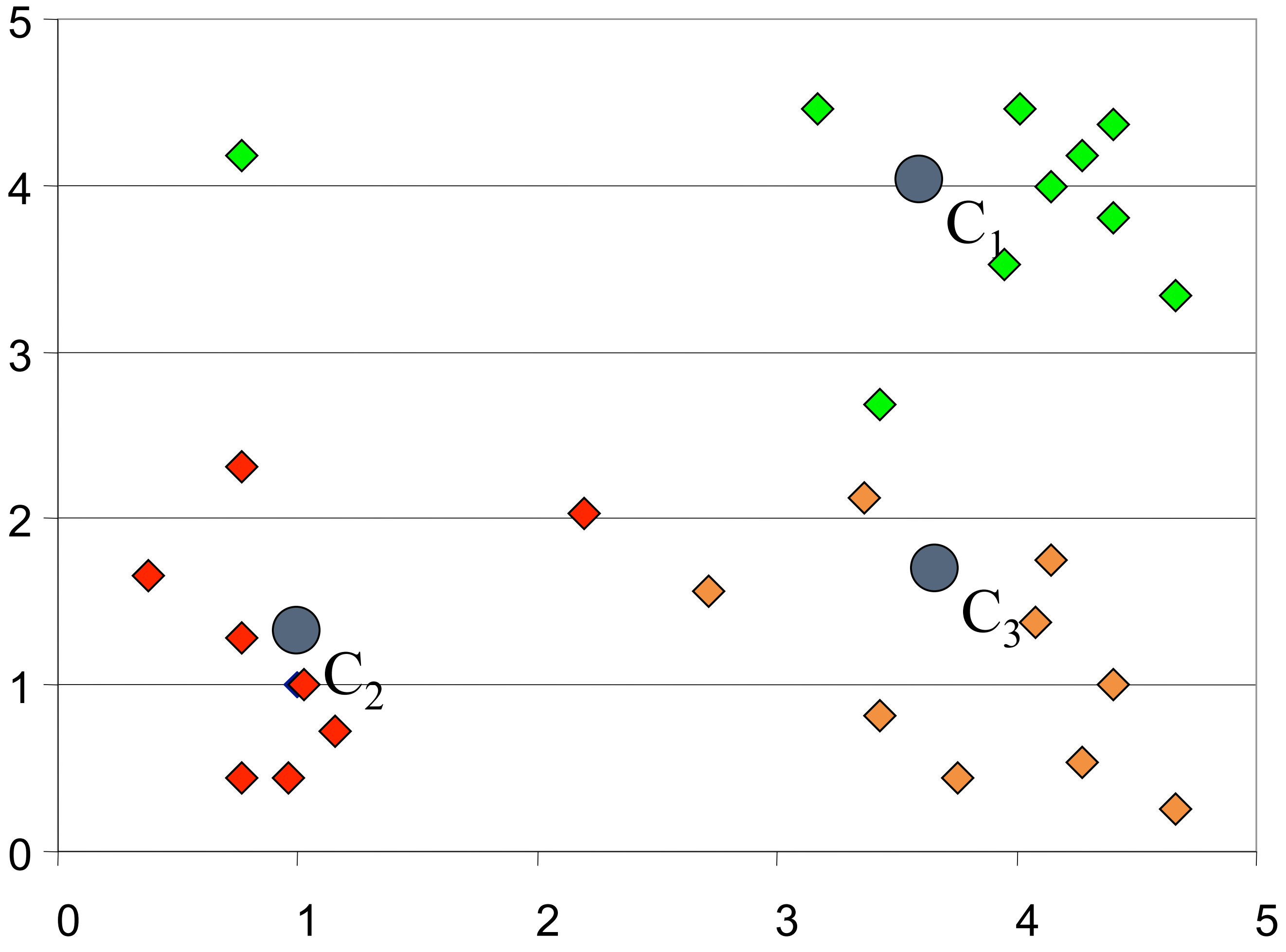
K-MEANS CLUSTERING: STEP 2

Algorithm: K-means, K =3, Distance Metric: Euclidean Distance



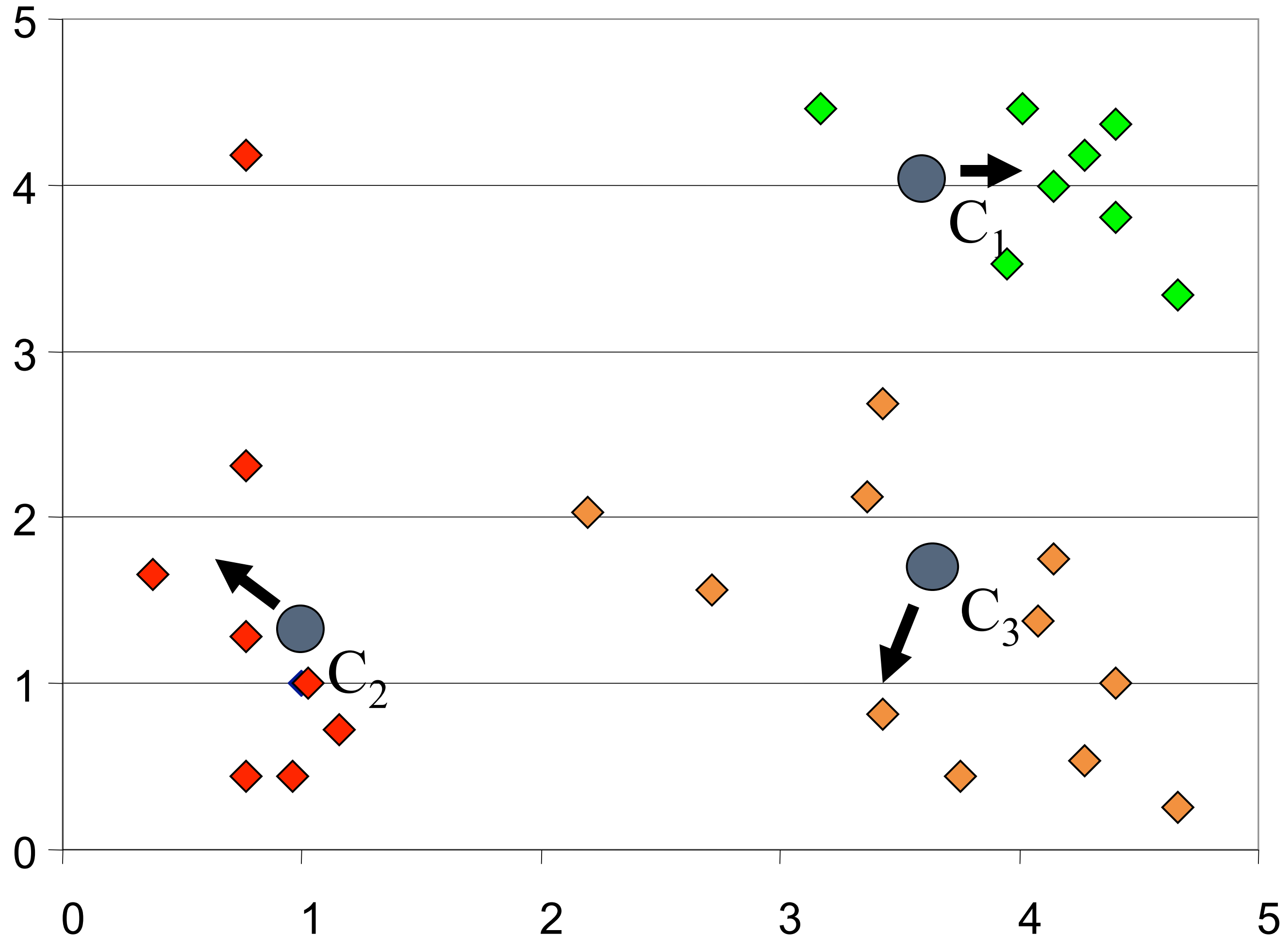
K-MEANS CLUSTERING: STEP 3

Algorithm: K-means, K =3, Distance Metric: Euclidean Distance



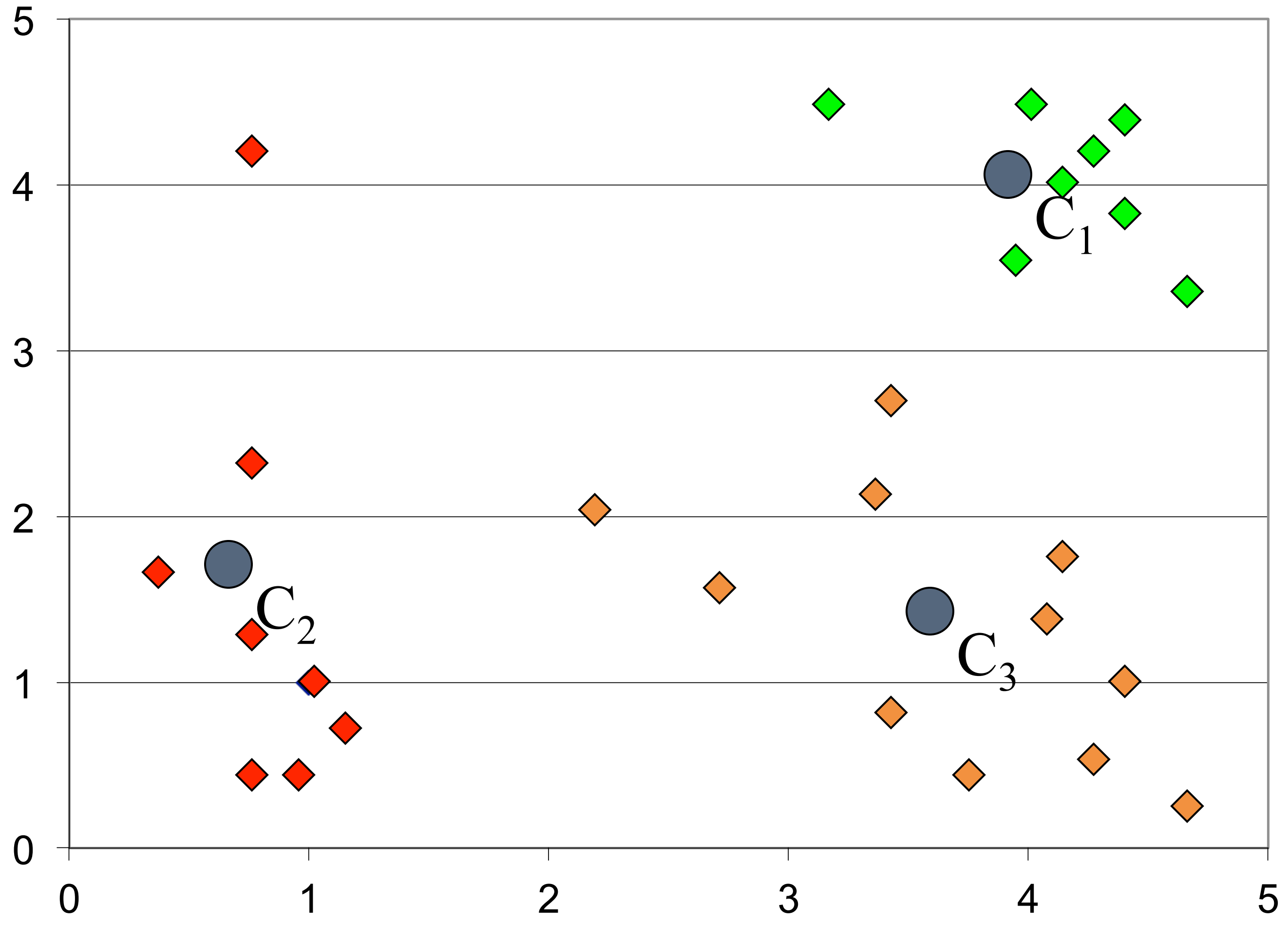
K-MEANS CLUSTERING: STEP 4

Algorithm: K-means, K =3, Distance Metric: Euclidean Distance



K-MEANS CLUSTERING: STEP 5

Algorithm: K-means, K =3, Distance Metric: Euclidean Distance



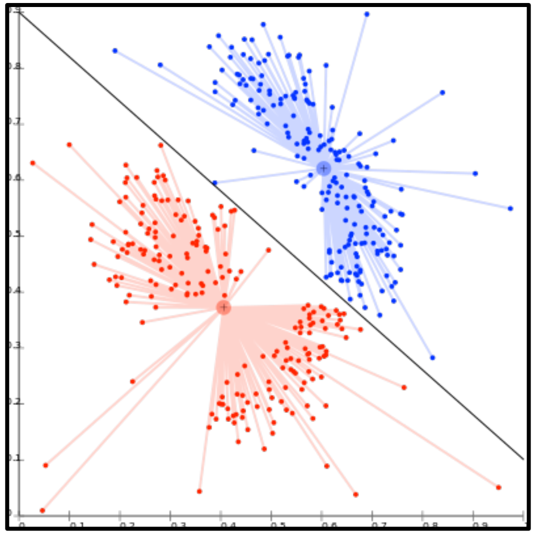
K- MEANS - COMMENTS

- **Strength**
  - *Relatively efficient*: *O*(*tKn*), where *t* is # iterations. Normally, *K*, *t* << *n*.
  - Often terminates at a *local optimum*.
- **Weakness**
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *K,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

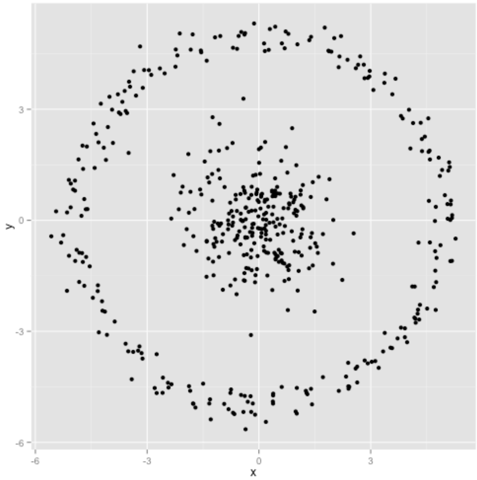
E-M ALGORITHMS

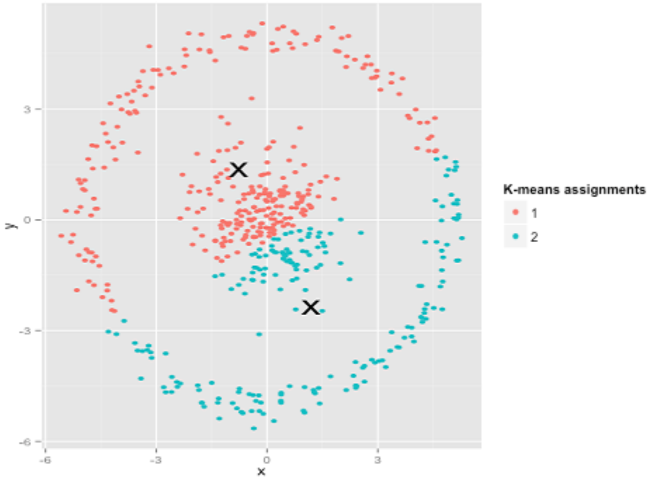
Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9

- Sometimes K-mean is not intuitive



- Horrible result from K-mean





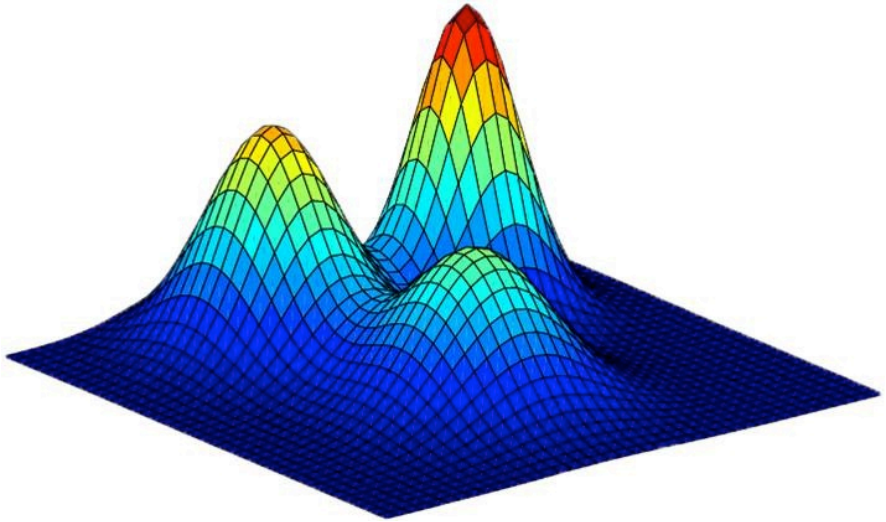
EM APPROACH -- MIXTURE MODEL

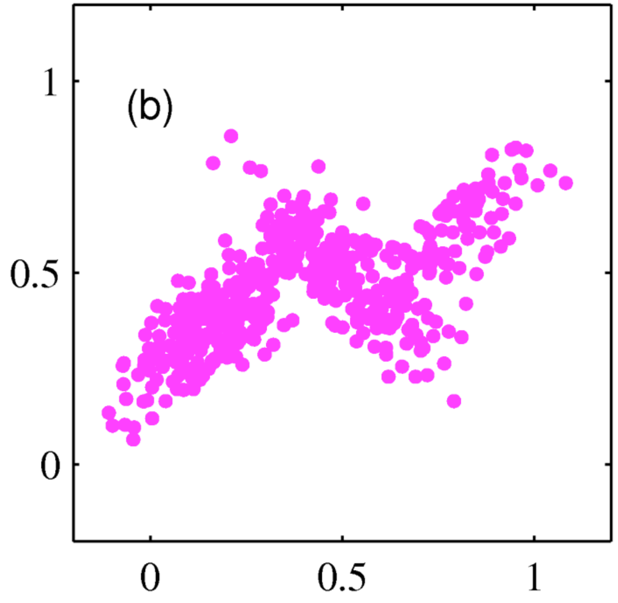
Source:
Pattern Recognition and Machine Learning,
Christopher M. Bishop,
Chapter 9

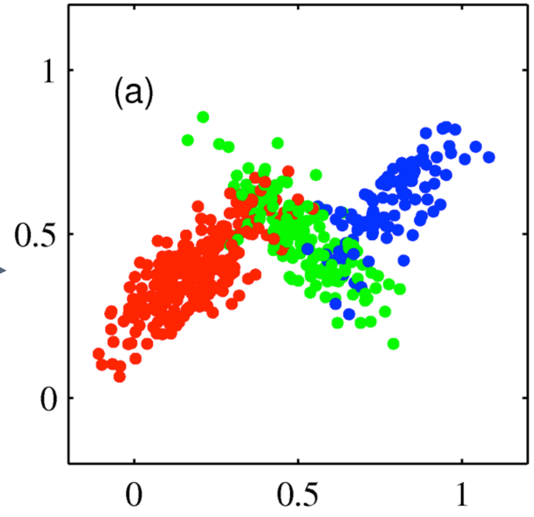
weighted sum of a number of pdfs where the weights are determined by a distribution

$$p(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x) + \ldots + \pi_k f_k(x)$$

$$\text{where } \sum_{i=0}^{k} \pi_i = 1$$







EM APPROACH -- ADD HIDDEN VARIABLES

Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9

Suppose we are told that the mixture model is…..

- Observed data $x = (x_1, x_2, \ldots, x_N)$
- Hidden distribution $z = (z_1, z_2, \ldots, z_N)$
- Each point, for example mixture of gaussian

$$p(x) = \pi_0 f_0(x) + \pi_1 f_1(x) + \pi_2 f_2(x) + \ldots + \pi_k f_k(x)$$

$$p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k) = \sum_{z} p(z)p(x|z)$$

EM APPROACH -- ADD HIDDEN VARIABLES

Suppose we are told that the mixture model is…..

- Log likelihood distribution for whole data set, n points

Target:

$$\ln p(x|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}$$

Use MLE to to get optimal solution

But hard to calculate, not concave or convex function

Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9

- WE KNOW:

$$L(\theta) = \log P(X;\theta) = \log \sum_Z P(X,Z;\theta)$$

$$L(\theta) = \log \sum_Z P(X,Z;\theta) = \log \sum_Z Q(Z)\frac{P(X,Z;\theta)}{Q(Z)} = \log E_{Z\sim Q}\left[\frac{P(X,Z;\theta)}{Q(Z)}\right]$$

- WE HAVE: JENSEN'S INEQUALITY

$$L(\theta) = \log E_{Z\sim Q}\left[\frac{P(X,Z;\theta)}{Q(Z)}\right] \geq E_{Z\sim Q}\left[\log \frac{P(X,Z;\theta)}{Q(Z)}\right]$$

When \ $\frac{P(X,Z;\theta)}{Q(Z)}$ is constant, the equality holds, we get the boundary

$$Q(Z) = \frac{P(X,Z;\theta_t)}{\sum_Z P(X,Z;\theta_t)} = P(Z|X;\theta_t)$$

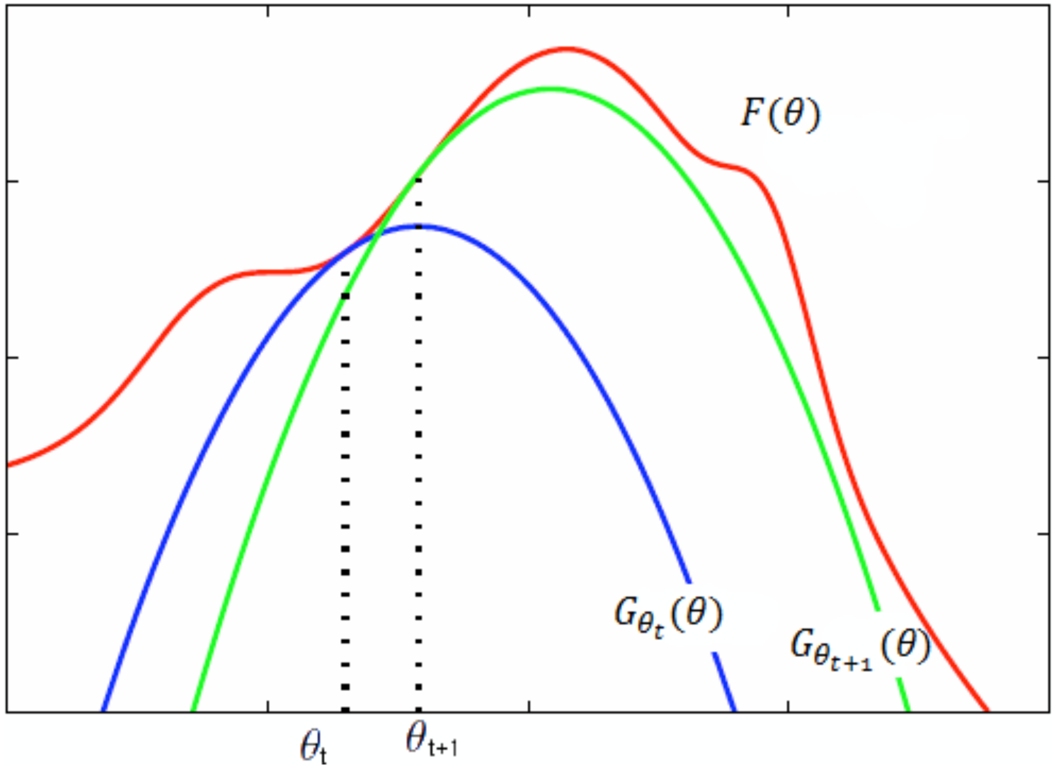
Then optimize the boundary: that is what we can do easily, take the derivative and equal to 0

$$\theta_{t+1} := \arg\max_\theta E_{Z|X;\theta_t}\left[\log \frac{P(X,Z;\theta)}{P(Z|X;\theta_t)}\right]$$

Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9

TOO ABSTRACT.

Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9



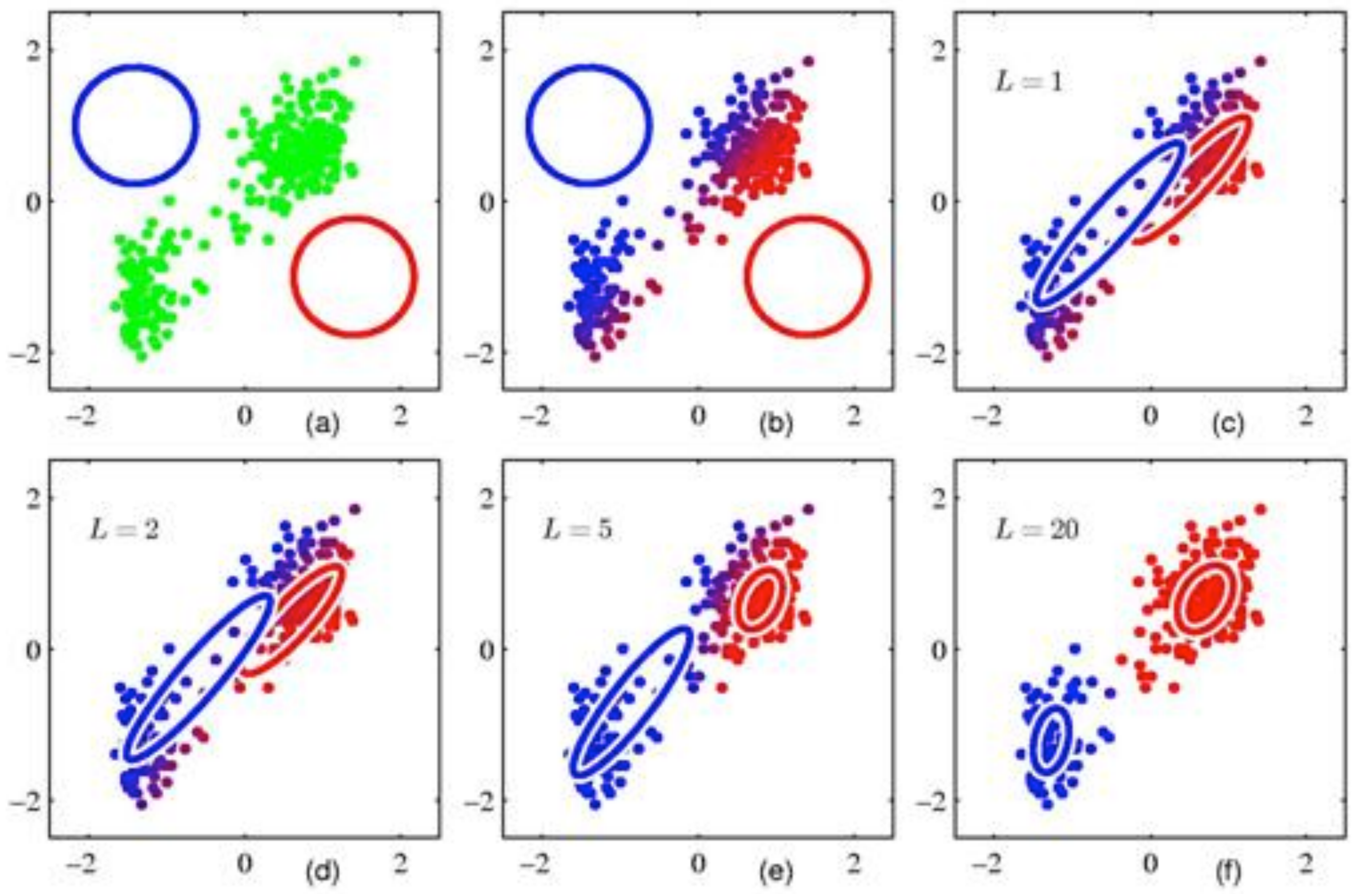
E-M ALGORITHMS

- Initialize K cluster centers
- Iterate between two steps
  - Expectation step: assign points to clusters (Bayes) $P(d_i \in c_k) = w_k \Pr(d_i | c_k) \Big/ \sum_j w_j \Pr(d_i | c_j)$
  - $w_k = \dfrac{\sum_i \Pr(d_i \in c_k)}{N}$ = probability of class c_k
  - Maximation step: estimate model parameters (optimization) $\mu_k = \dfrac{1}{m} \sum_{i=1}^{m} \dfrac{d_i P(d_i \in c_k)}{\sum_k P(d_i \in c_j)}$

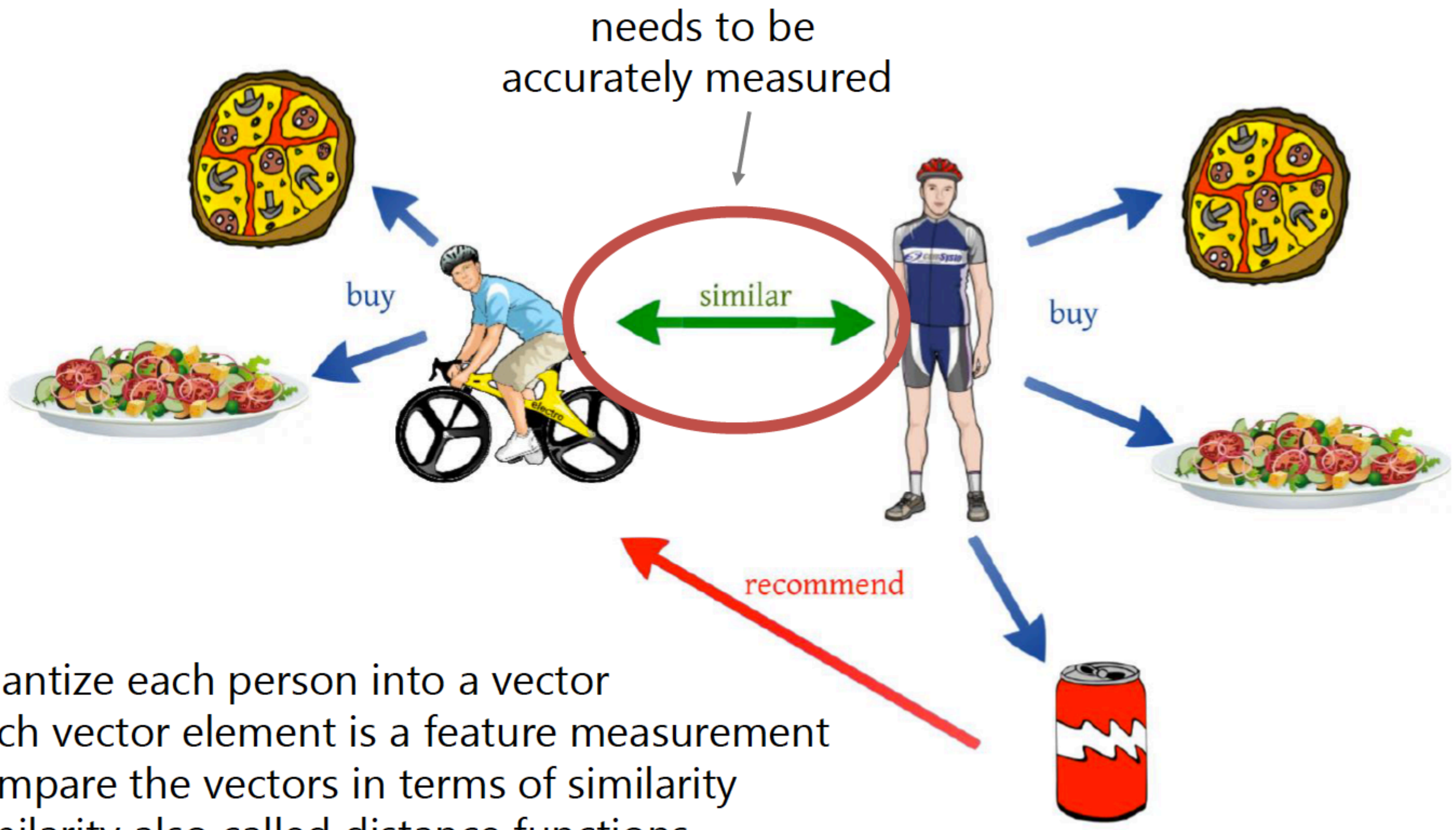
probability that d_i is in class c_j

EM APPROACH -- ILLUSTRATION

Source:
Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 9



TASK # 4: SIMILARITY MATCHING

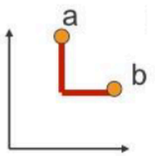


quantize each person into a vector
each vector element is a feature measurement
compare the vectors in terms of similarity
similarity also called distance functions

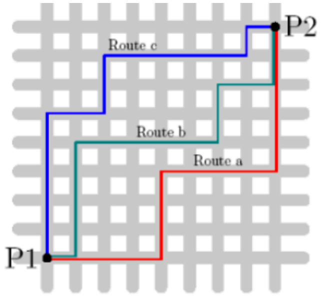
Picture from Klaus Muller

SIMILARITY MEASURE - DISTANCES

Manhattan distance

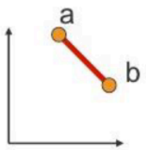


$$\text{dist(a,b)} = \|a - b\|_1 = \sum_i |a_i - b_i|$$

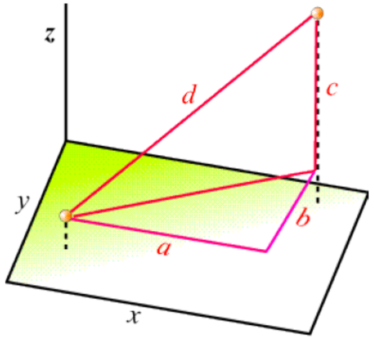


-

Euclidian distance

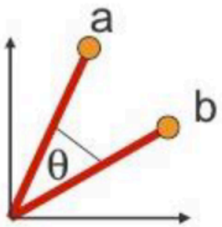


$$\text{dist(a,b)} = \|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$



Cosine Similarity

-



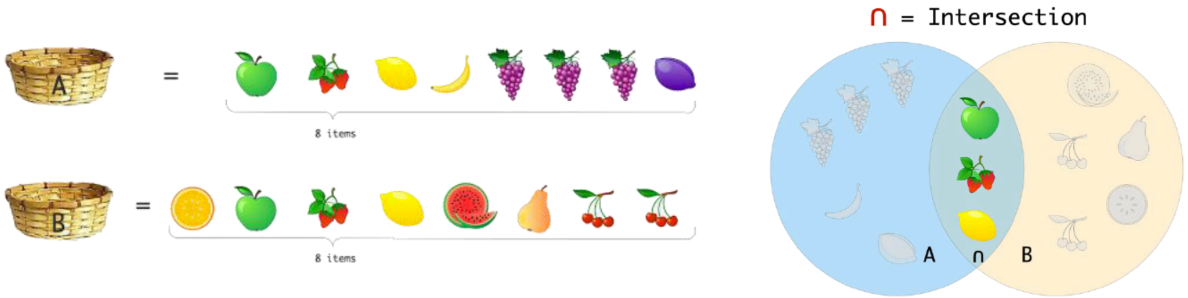
$$\text{dist(a,b)} = cos^{-1} \frac{\langle a, b \rangle}{\|a\| \|b\|}$$

how is this related to correlation?

SIMILARITY MEASURES

- Jaccard Distance

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

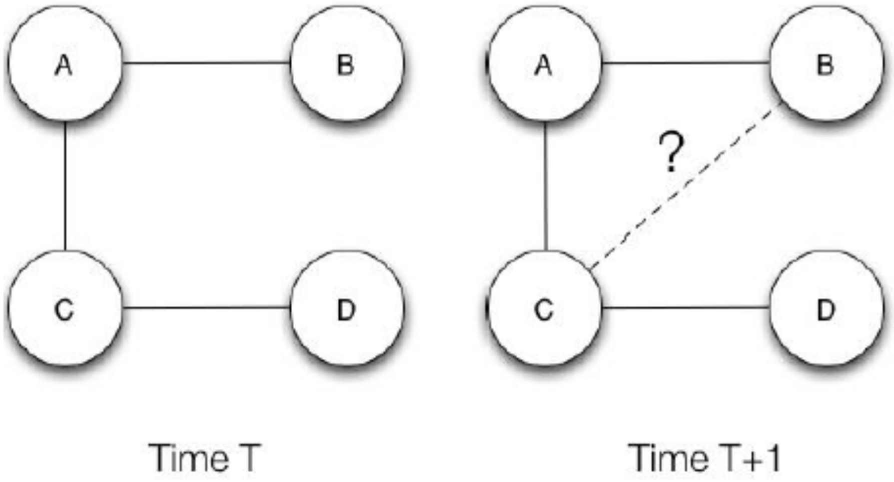


Picture from Klaus Muller

TASK # 5 : LINK PREDICTION

- Predict connections between data items
  - usually works within a graph
  - predict missing links
  - estimate link strength

- Applications
  - in recommendation systems
  - friend suggestion in Facebook (social graph)
  - link suggestion in LinkedIn (professional graph)
  - movie suggestion in Netflix (bipartite graph people – movies)



DATA REDUCTION

DATA REDUCTION

- Why?
  - Tuples are multi dimensional – Humans can “see” only 3 dimensions on the screen
- How?
  - PCA Principal Component Analysis

PCA PRINCIPAL COMPONENT ANALYSIS

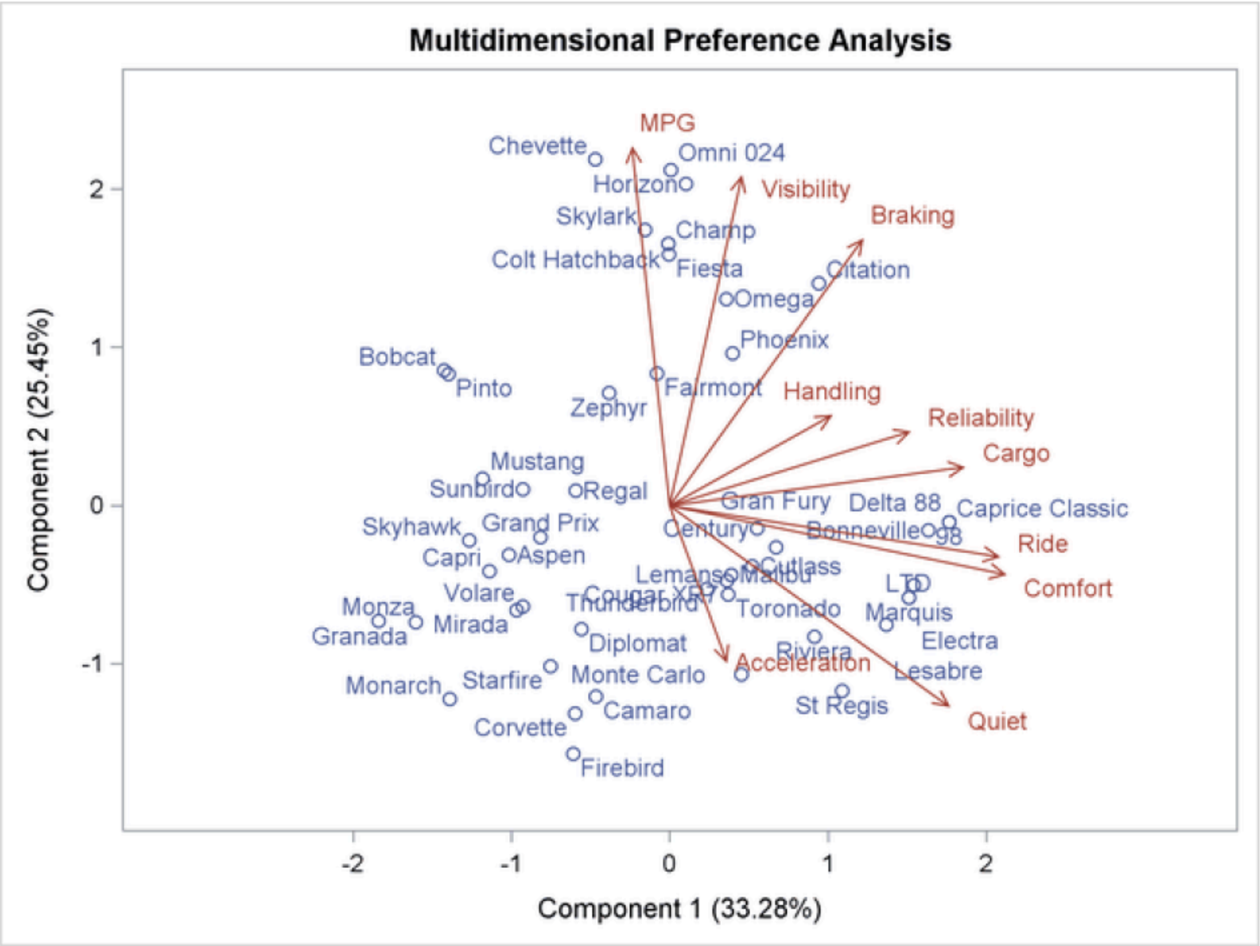
- Ultimate goal:
  - find a coordinate system that can represent the variance in the data with as few axes as possible
- Steps:

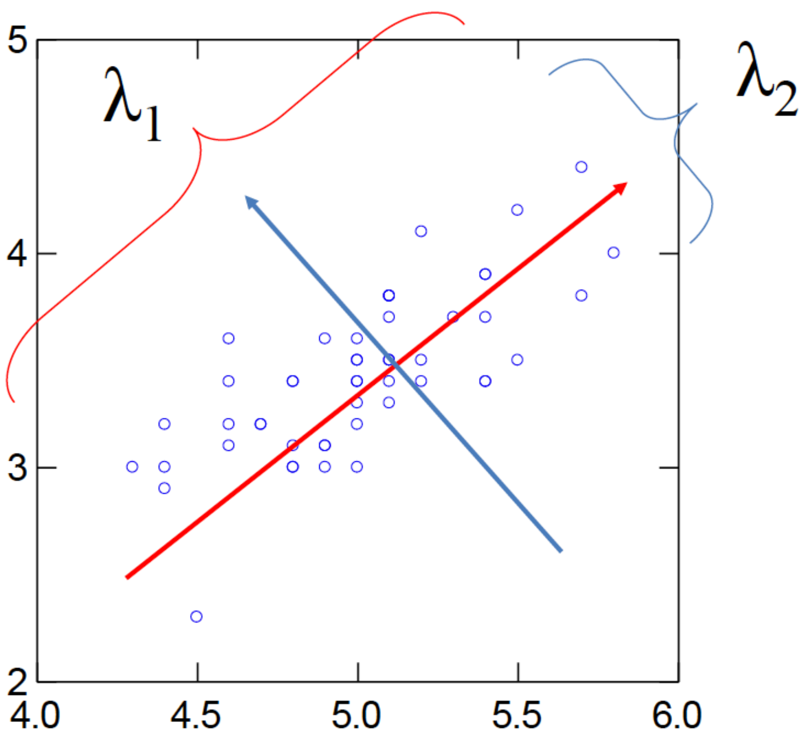
- First characterize the distribution by
  - covariance matrix Cov
  - correlation matrix Corr
  - lets call it C
- perform QR factorization or LU decomposition to get :

$$\boldsymbol{C} = Q \Lambda Q^{-1}$$

Q: matrix with Eigenvectors
Λ: diagonal matrix with Eigenvalues λ

- now order the Eigenvectors in terms of their Eigenvalues
- drop the axes that have the least variance


Multidimensional Preference Analysis
MPG
Visibility
Braking
Handling
Reliability
Cargo
Ride
Comfort
Acceleration
Quiet
Chevette
Omni 024
Horizon
Skylark
Champ
Colt Hatchback
Fiesta
Citation
Omega
Phoenix
Bobcat
Pinto
Zephyr
Fairmont
Mustang
Sunbird
Regal
Gran Fury
Delta 88
Caprice Classic
Skyhawk
Grand Prix
Century
Bonneville
98
Capri
Aspen
Lemans
Malibu
Cutlass
LTD
Monza
Volare
Thunderbird
Cougar XR7
Toronado
Marquis
Granada
Mirada
Diplomat
Electra
Riviera
Monarch
Starfire
Monte Carlo
Lesabre
Camaro
St Regis
Corvette
Firebird
Component 2 (25.45%)
Component 1 (33.28%)
2
1
0
-1
-2
-1
0
1
2

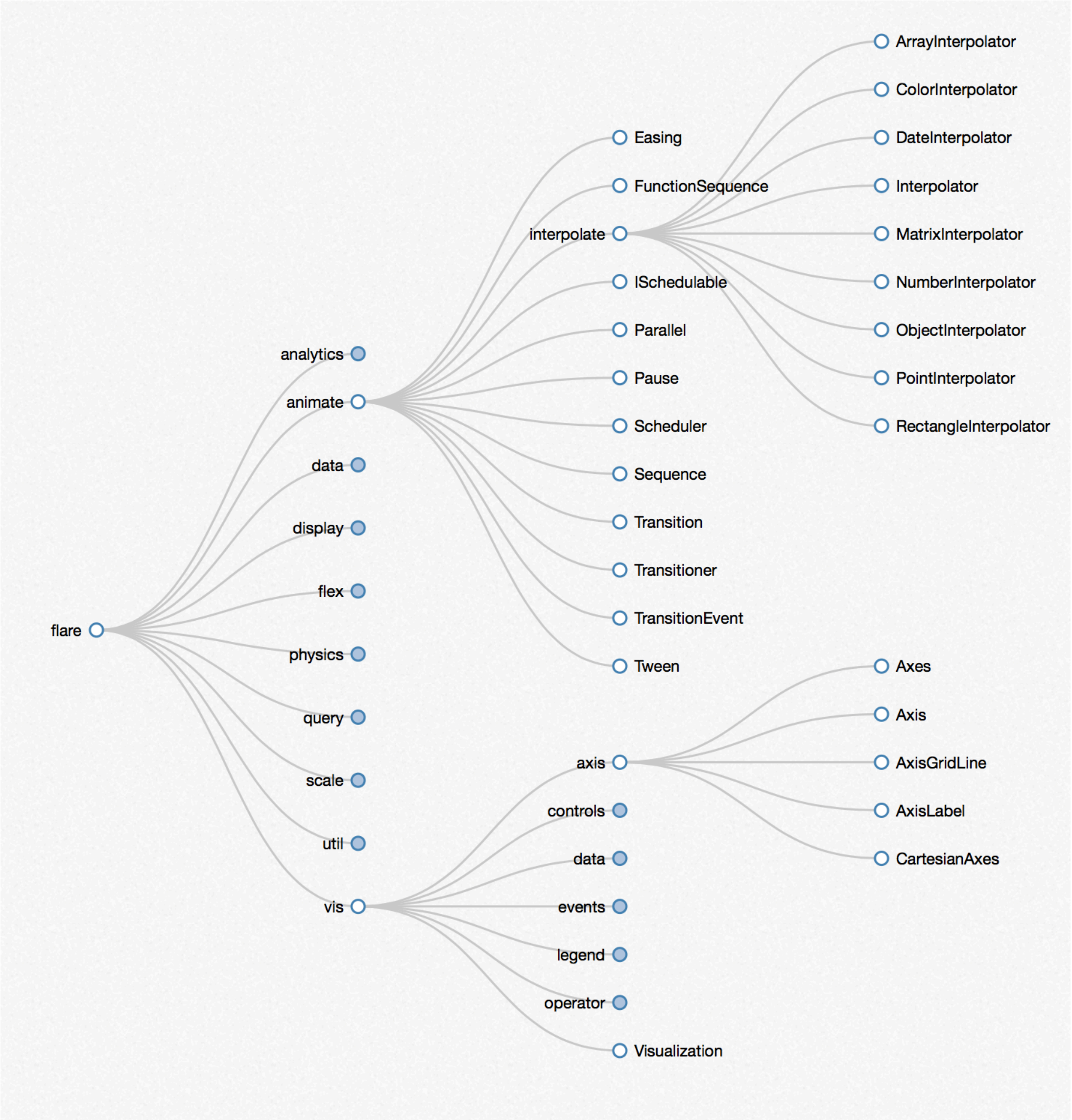

λ_1
λ_2
5
4
3
2
4.0
4.5
5.0
5.5
6.0

TYPES OF VISUALIZATION

AESTHETIC AND **CLEAR AND UNDERSTANDABLE** GRAPHICS FOR REPRESENTATION OF DATA

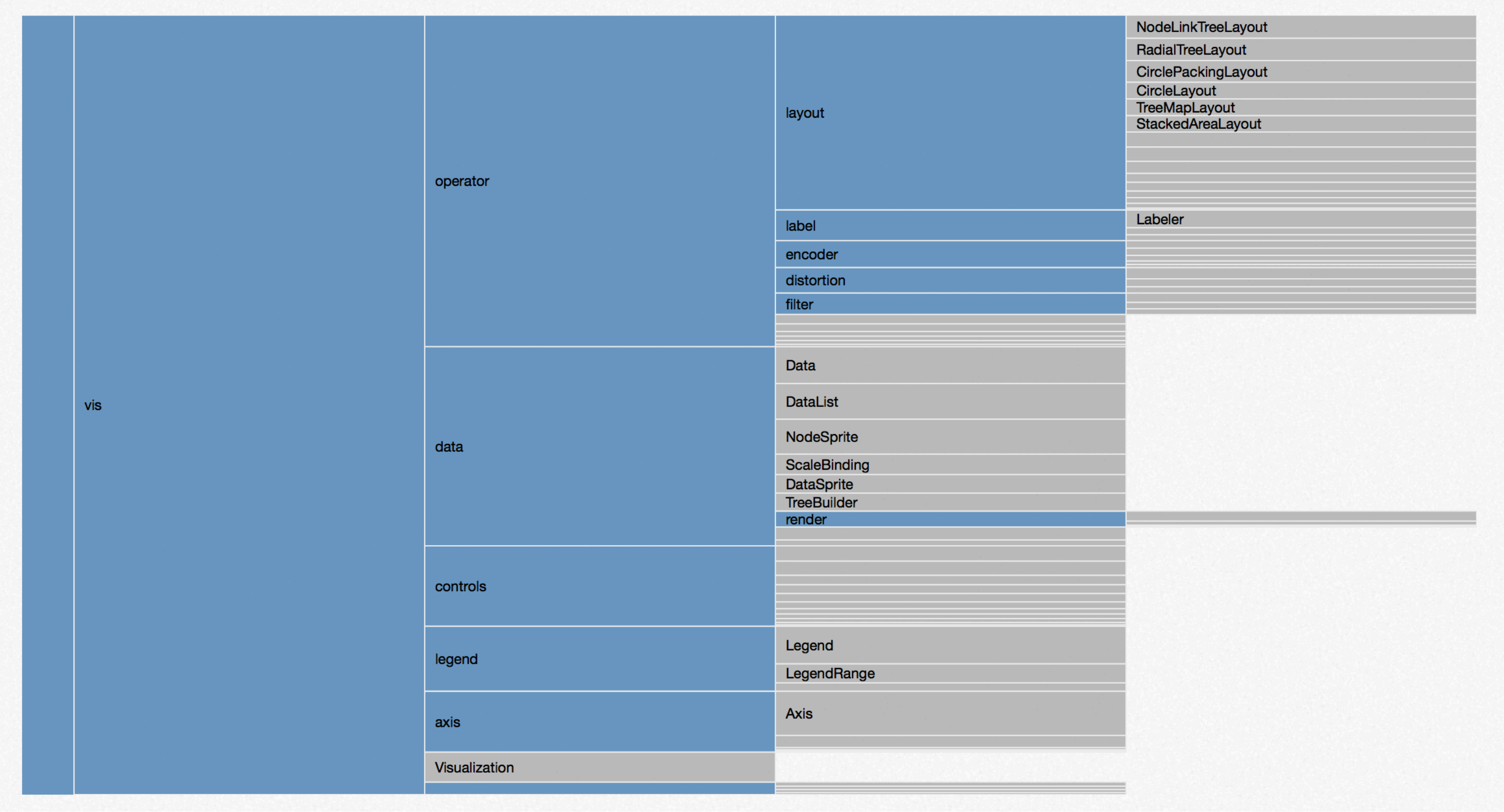
COLLAPSIBLE TREE

- A standard tree, but one that is scalable to large hierarchies
- http://mbostock.github.io/d3/talk/20111018/tree.html



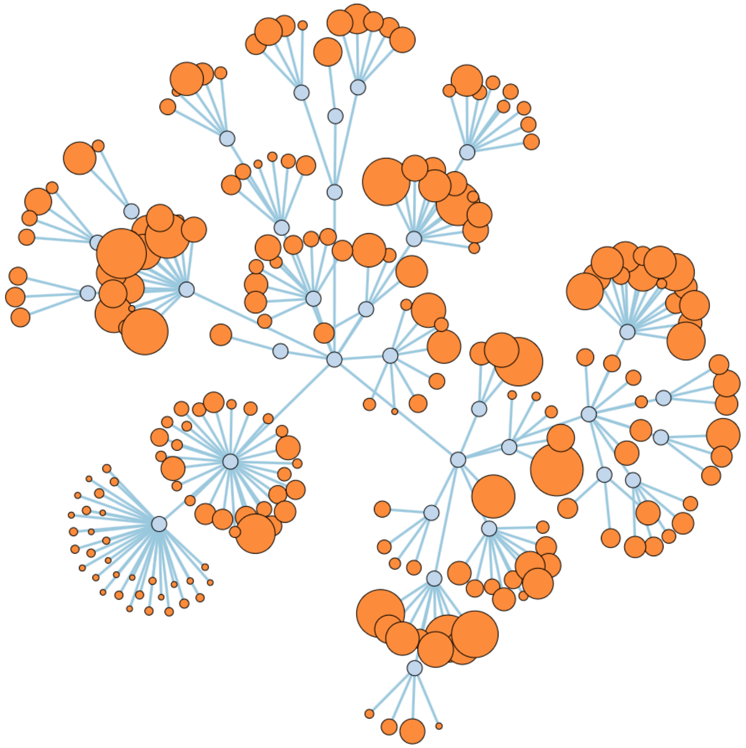
ZOOMABLE PARTITION LAYOUT

- A tree that is scalable and has partial partition of unity
- http://mbostock.github.io/d3/talk/20111018/partition.html



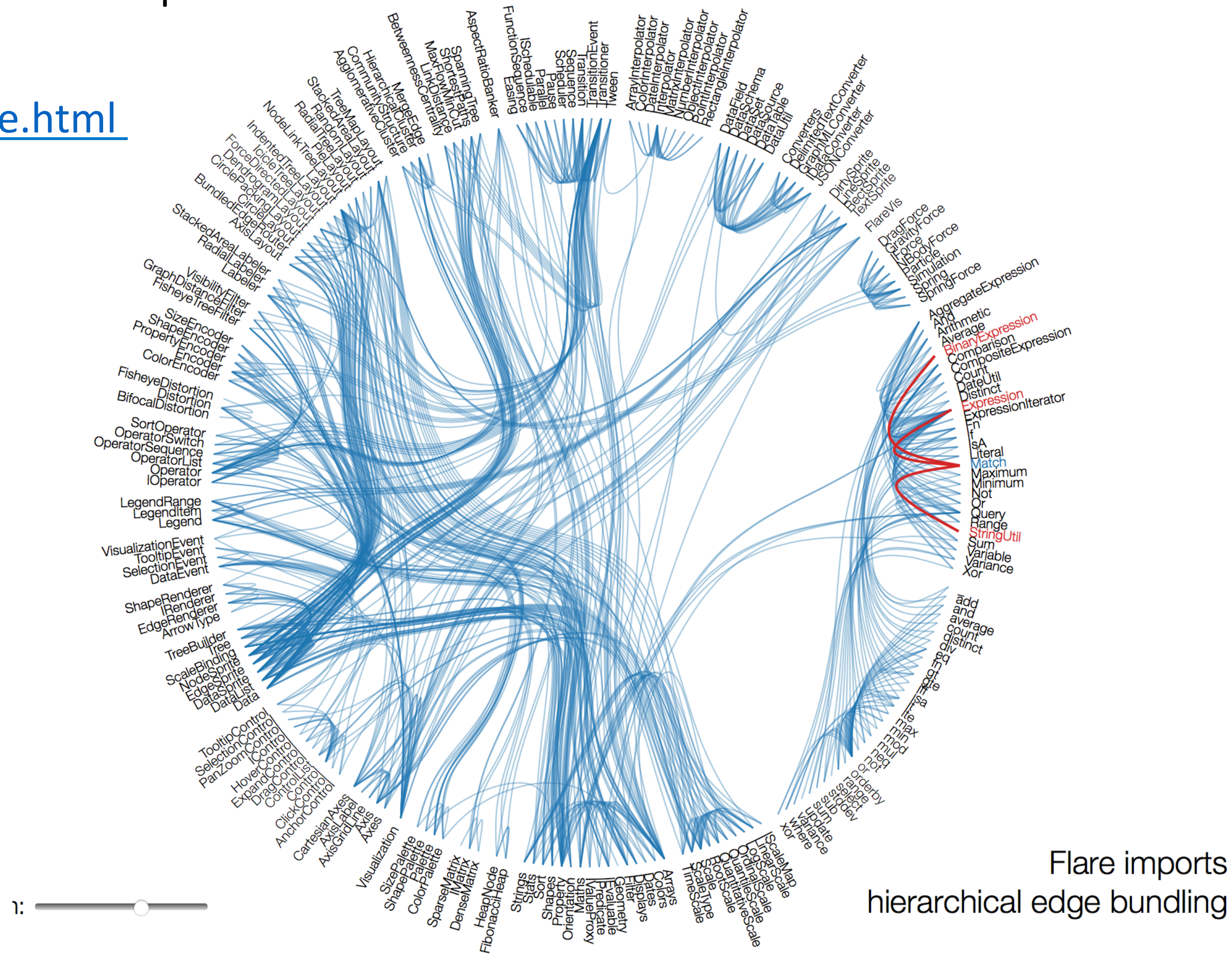
COLLAPSIBLE FORCE LAYOUT

- Relationships within organization members expressed as distance and proximity
- http://mbostock.github.io/d3/talk/20111116/force-collapsible.html



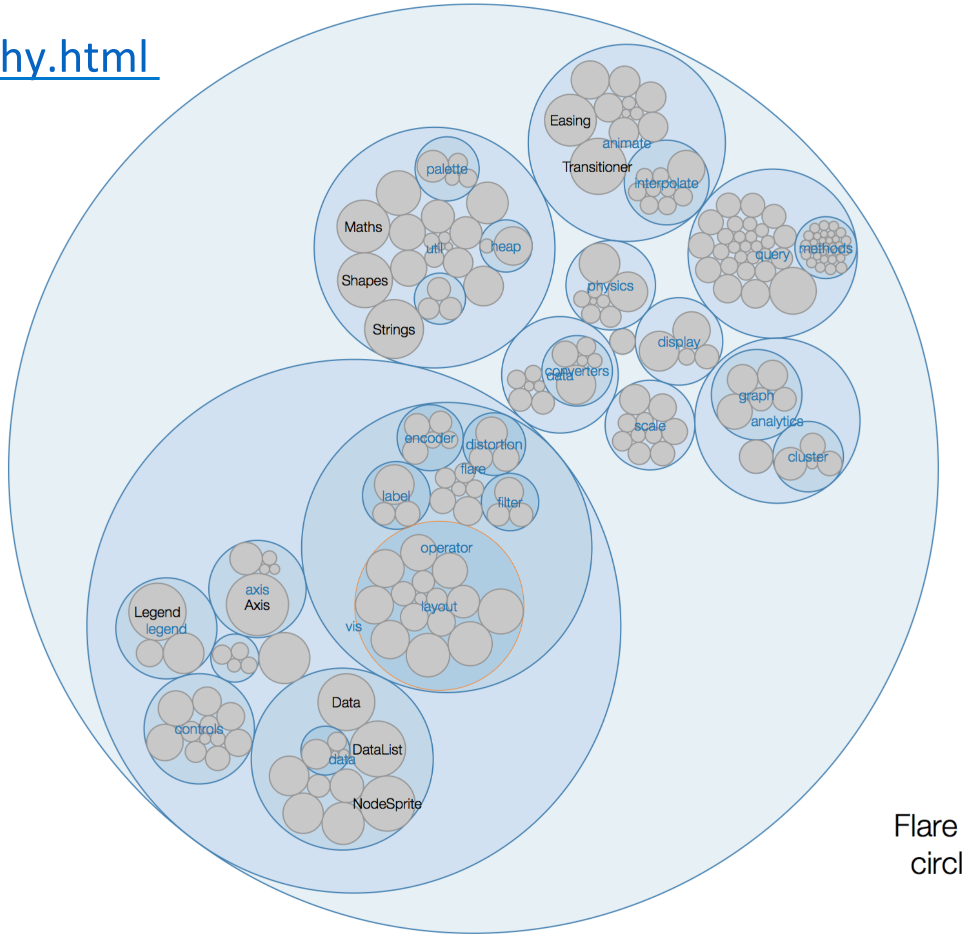
HIERARCHICAL EDGE BUNDLING

- Relationships of individual group members, also in terms of quantitative measures such as information flow
- http://mbostock.github.io/d3/talk/20111116/bundle.html



CIRCLE MAPPING

- Quantities and containment, but not partition of unity
- http://mbostock.github.io/d3/talk/20111116/pack-hierarchy.html



Flare code size
circle packing

DEMO

THANK YOU!

QUESTION AND ANSWERS