DATA PREPROCESSING

Hari Sundaram

hs1@illinois.edu

http://sundaram.cs.illinois.edu

adapted from slides by Jiawei Han and Kevin Chang

AN OVERVIEW

completeness

accuracy

consistency

why preprocess?

interpretability

believability

timeliness

cleaning

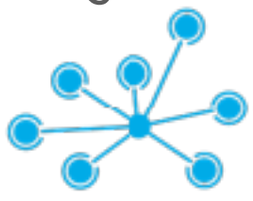
- missing data
- smooth noisy data
- outliers
- resolve inconsistencies

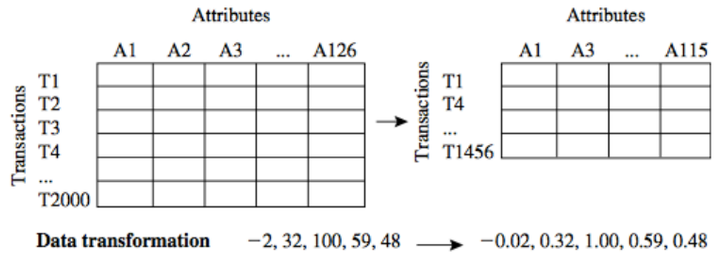
reduction

- dimensions
- number
- compression

major tasks

integration





transformation and discretization

- normalization
- concept hierarchy generation

DATA CLEANING

REAL WORLD DATA IS MESSY!

Many reasons: faulty instruments, human or computer error, transmission error, etc.

incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregated data

e.g., Occupation = " " (missing data)

noisy: containing noise, errors, or outliers

e.g., Salary = "−10" (an error)

inconsistent: containing discrepancies in codes or names, e.g.,

Age = "42", Birthday = "03/07/2010"

Was rating "1, 2, 3", now rating "A, B, C"

discrepancy between duplicate records

Intentional (e.g., disguised missing data)

Jan. 1 as everyone's birthday?

MISSING DATA

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

WHAT CAN BE DONE?

Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably

Fill in the missing value manually: usually tedious + infeasible

Fill in it automatically with:

- a global constant : e.g., "unknown", a new class?!
- the attribute mean
- the attribute mean for all samples belonging to the same class: smarter
- the most probable value: inference-based such as Bayesian formula or decision tree

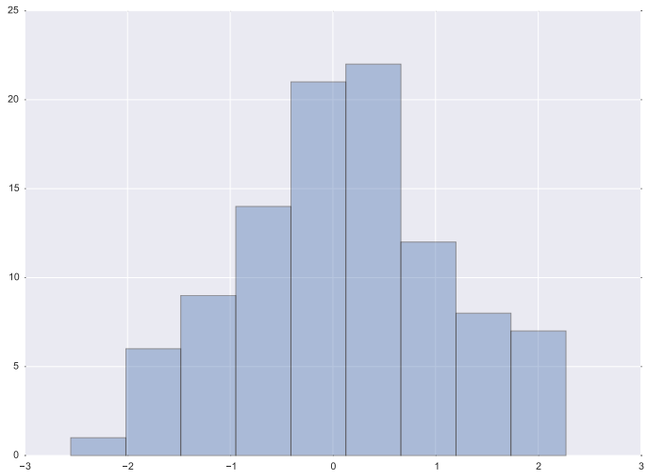
Incorrect attribute values may be due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

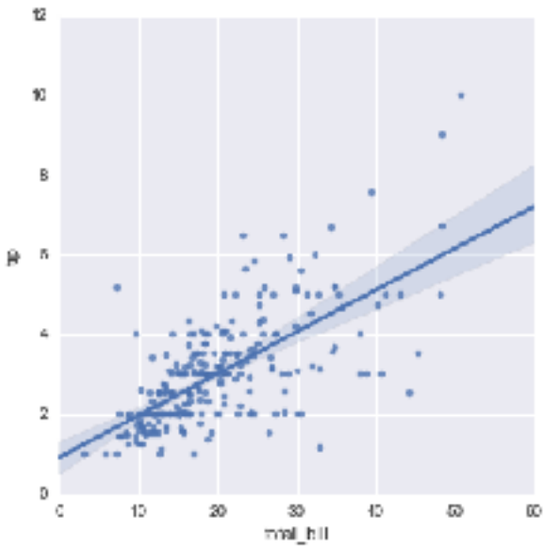
noise

random error or variance
in a measured variable

HANDLING NOISY DATA

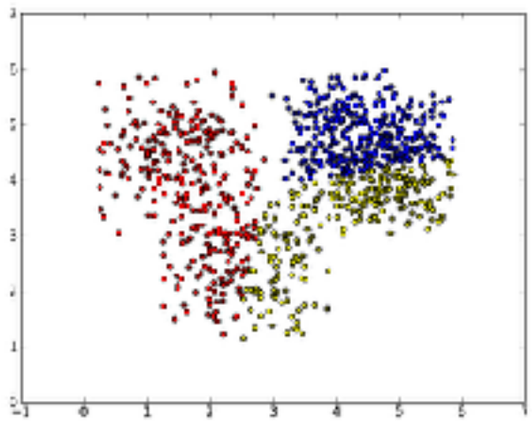


duplicate records



inconsistent data

incomplete data



Binning

first sort data and partition into (equal-frequency) bins

then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Regression

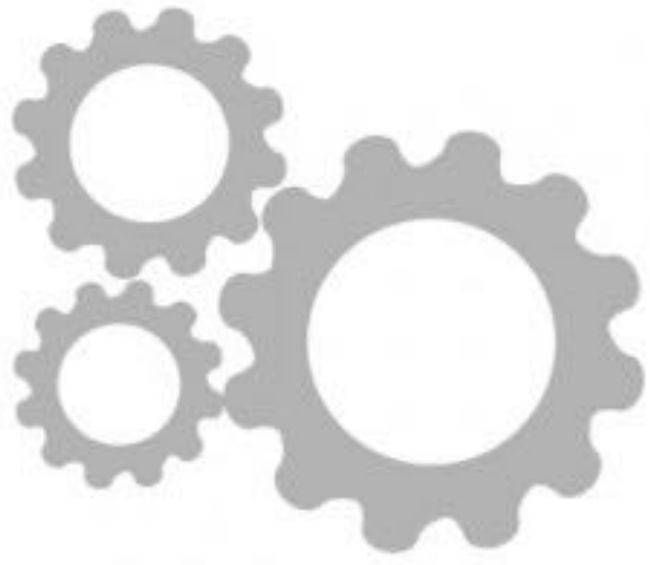
smooth by fitting the data into regression functions

Clustering

detect and remove outliers

Semi-supervised

detect suspicious values and check by human (e.g., deal with possible outliers)



DATA CLEANING AS A PROCESS

Step 1: Discrepancy detection

- Use metadata (e.g., domain, range, dependency, distribution)
- Check field overloading
- Check uniqueness rule, consecutive rule and null rule
- Use commercial tools
  - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
  - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Step 2: Data transformation (to correct the discrepancies)

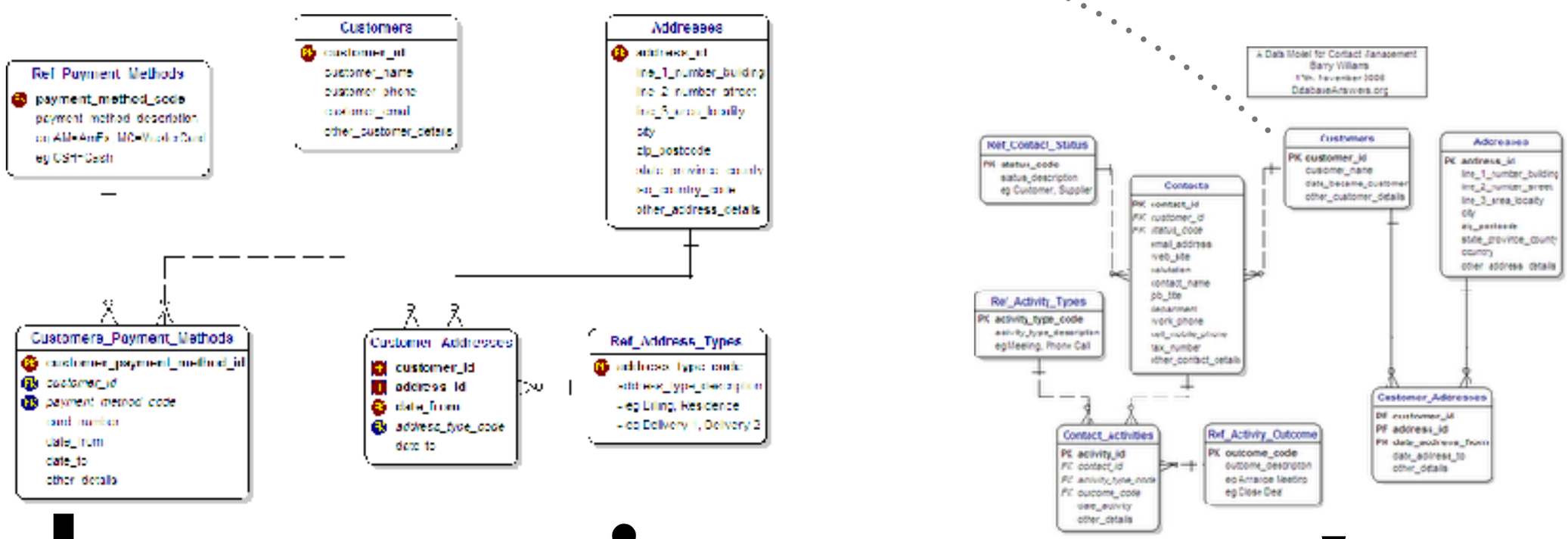
- Data migration tools: allow transformations to be specified
- ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Data cleaning process: the two steps iterate and reinforce

DATA INTEGRATION

data integration

Combines data from multiple sources into a coherent store



schema integration

e.g. A.cust-id B.cust-#

Integrate
metadata from
different sources

entity identification

Identify real world entities from multiple data sources, e.g., Bill Clinton = William Jefferson Clinton

For the same real world entity, attribute values from different sources are different

data conflict resolution

Possible reasons: different representations, different scales, e.g., metric vs. British units

1999

NASA lost its $125-million Mars Climate Orbiter because spacecraft engineers **failed to convert from English to metric measurements** when exchanging vital data before the craft was launched, space agency officials said Thursday.

REDUNDANCIES & INCONSISTENCIES

Redundant data often occur when integrating multiple databases

- Object identification: The same attribute or object may have different names in different databases
- Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue

Redundant attributes may be able to be detected by correlation analysis and covariance analysis

Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies to improve mining speed and quality

Exercise!

why should redundant data affect pattern discovery?

null hypothesis: the two distributions are independent (i.e. uncorrelated)

chi square test $\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$

observed

expected

The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count

correlation analysis

why does this equation have n-1 degrees of freedom?

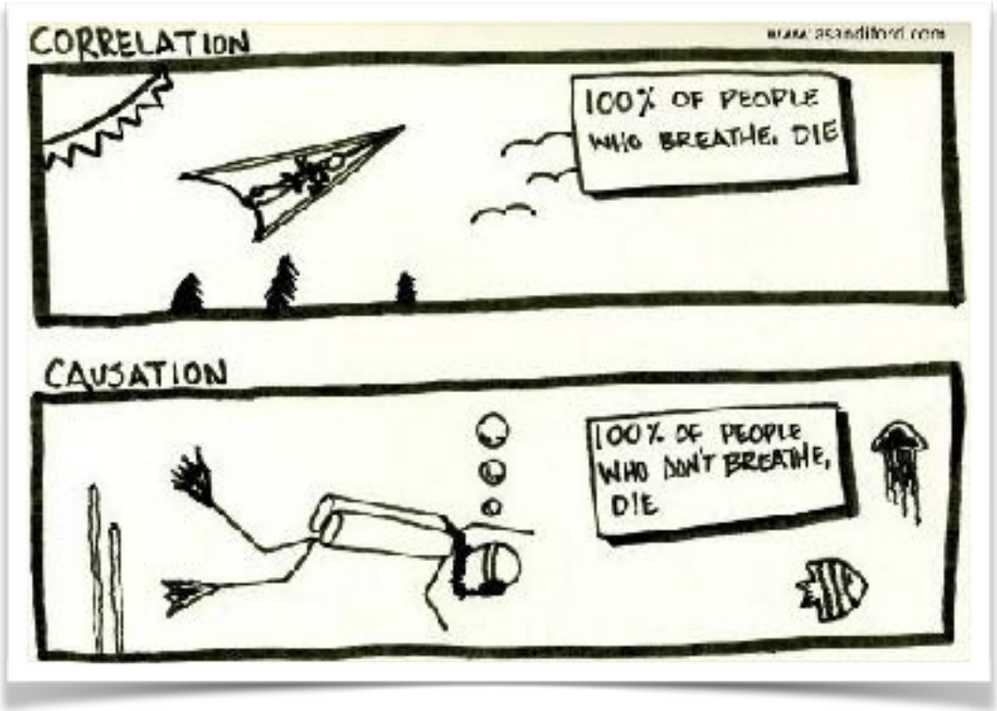
$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

"the number of observations minus the number of necessary relations among these observations."

correlation

≠

causation



	play chess	don't play chess	sum (row)
like science fiction	250 (90)	200 (360)	450
dislike science fiction	50 (210)	1000 (840)	1050
Sum (column)	300	1200	1500

expected counts

$$1500 \times \frac{450}{1500} \times \frac{300}{1500} = 90$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

0.001

we can reject the null hypothesis of independence at a confidence level of

degrees of freedom = (2 -1) x (2-1) = 1

correlation analysis

number of tuples

Pearson's coefficient $$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

$$= \frac{\sum_{i=1}^{n} a_i b_i - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

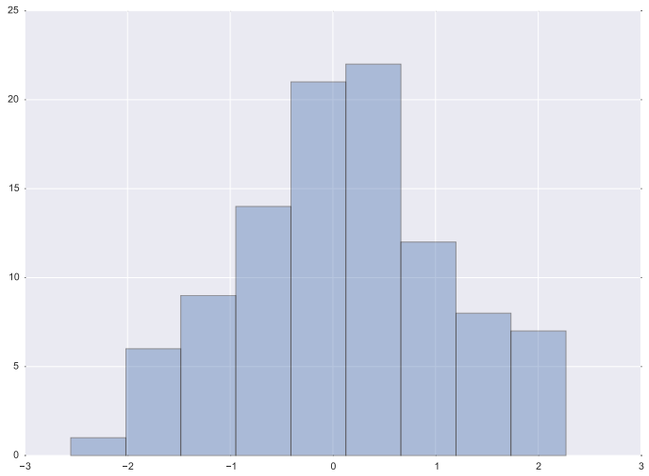
<0 >0 =0

Review

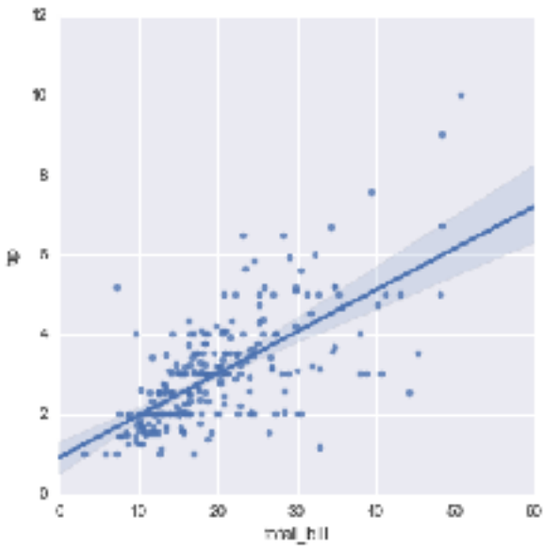
How to choose the right distance?

$$L_1, L_2, L_\infty, \cos(\theta)$$

HANDLING NOISY DATA

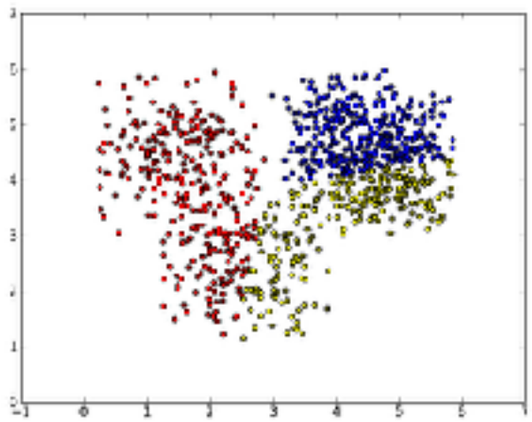


duplicate records



inconsistent data

incomplete data



Binning

first sort data and partition into (equal-frequency) bins

then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Regression

smooth by fitting the data into regression functions

Clustering

detect and remove outliers

Semi-supervised

detect suspicious values and check by human (e.g., deal with possible outliers)

why does this equation have n-1 degrees of freedom?

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

"the number of observations minus the number of necessary relations among these observations."

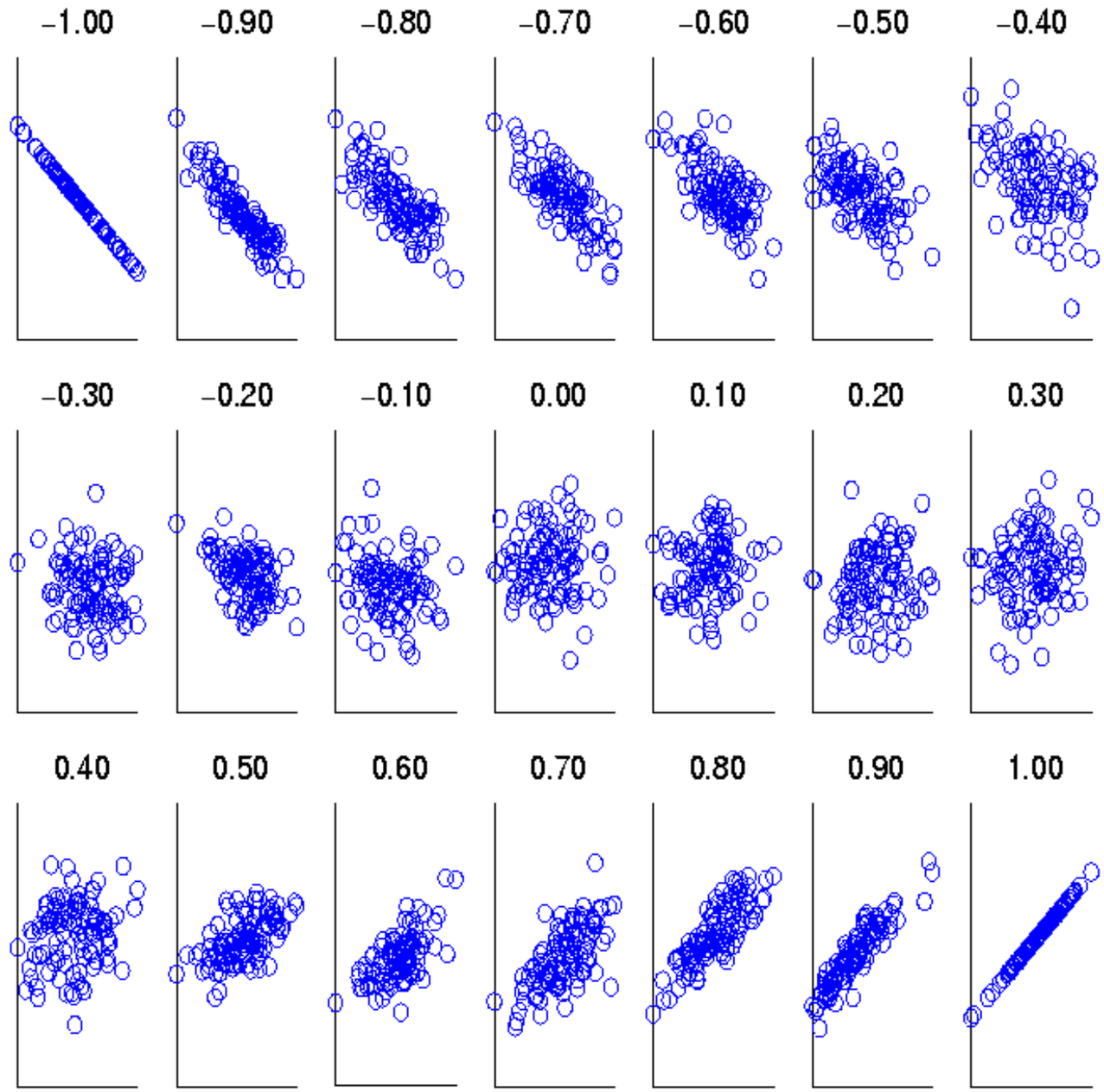
correlation analysis

number of tuples

Pearson's coefficient $$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

$$= \frac{\sum_{i=1}^{n} a_i b_i - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

<0 >0 =0



correlation varying between -1 and 1

covariance

covariance and correlation are related

$$cov(A,B) = \mathbb{E}((A-\bar{A})(B-\bar{B}))$$

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n\sigma_A\sigma_B}$$

$$= \frac{cov(A,B)}{\sigma_A\sigma_B}$$

<0 >0 =0

covariance matrices are positive semi-definite

positive semidefinite

$$u^t C u \geq 0$$

for any vector u

Proof:

column vector

$$u^t C u \geq 0$$

$$u^t \mathbb{E}(XX^t) u \geq 0$$

$$\mathbb{E}(Z_u^2) \geq 0, \text{ where } Z_u = u^t X$$

scalar

independence implies no correlation

but zero correlation does not imply independence!

it depends on the joint distribution

$$f(x, y) = f(x) \times f(y)$$

$$\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) = \exp\left(-\frac{x^2}{2\sigma^2}\right) \times \exp\left(-\frac{y^2}{2\sigma^2}\right)$$

easy to check for multivariate Gaussian distributions

code

COVARIANCE

$$cov(A,B) = \mathbb{E}((A-\bar{A})(B-\bar{B}))$$

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n\sigma_A\sigma_B}$$

$$= \frac{cov(A,B)}{\sigma_A\sigma_B}$$

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n\sigma_A\sigma_B}$$

$$= \frac{\sum_{i=1}^{n}a_ib_i - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

E(A)=(2+3+5+4+6)/5 = 20/5 = 4

E(B)=(5+8+10+11+14)/5 = 48/5 = 9.6

Cov(A,B)=(2×5+3×8+5×10+4×11 +6×14)/5 − 4 × 9.6 = 4

Thus, A and B rise together since cov(A, B) > 0.

DATA REDUCTION

data reduction

Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

A database/data warehouse may store hundreds of terabytes of data.

why?

Complex data analysis may take a very long time to run on the complete data set.

Dimensionality reduction
(remove unimportant attributes)

Wavelet transforms
Principal Components Analysis (PCA)
Feature subset selection, feature creation

Data compression

lossy, lossless

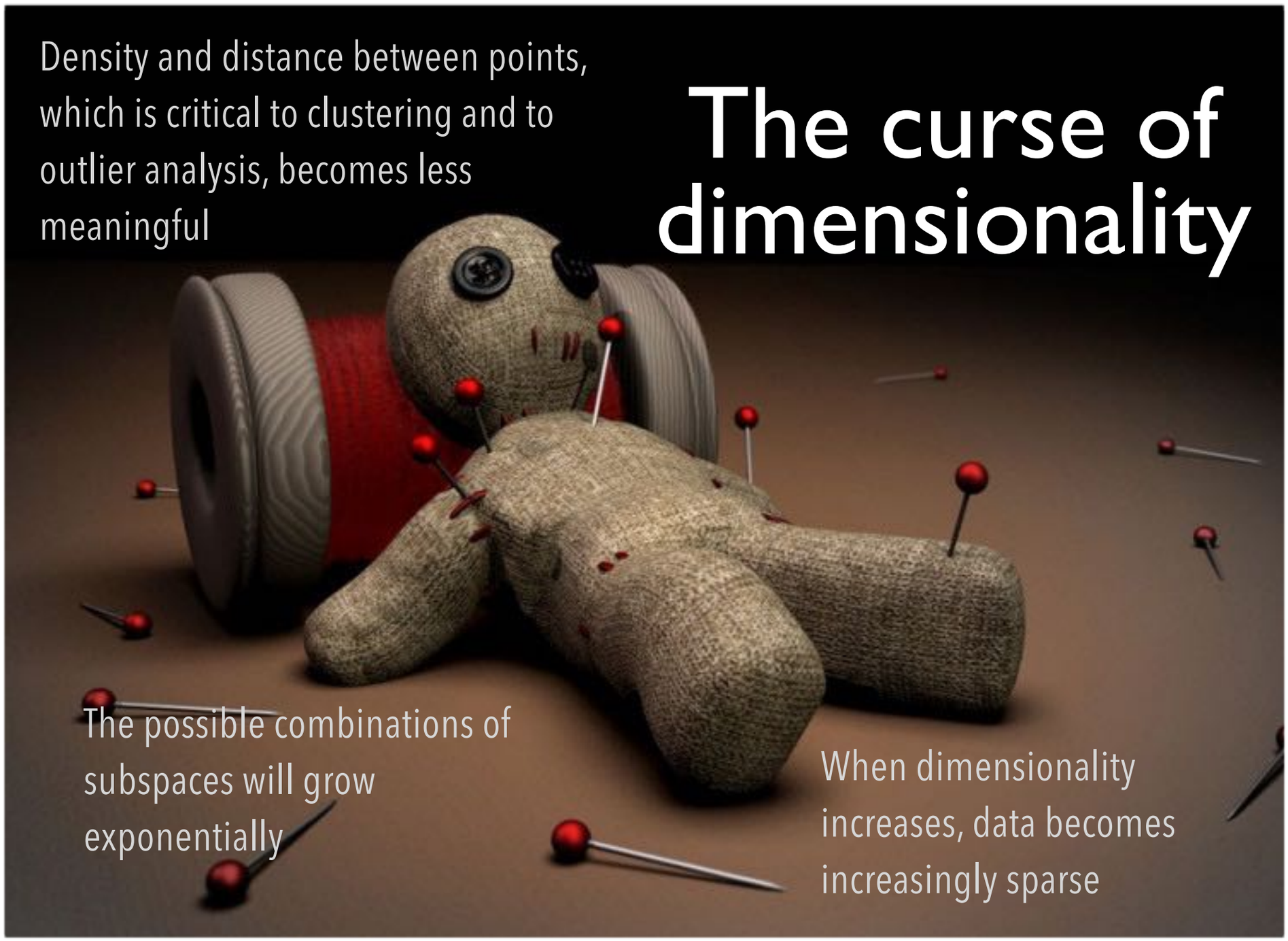
strategies

Data Reduction

Regression and Log-Linear Models
Histograms, clustering, sampling
Data cube aggregation

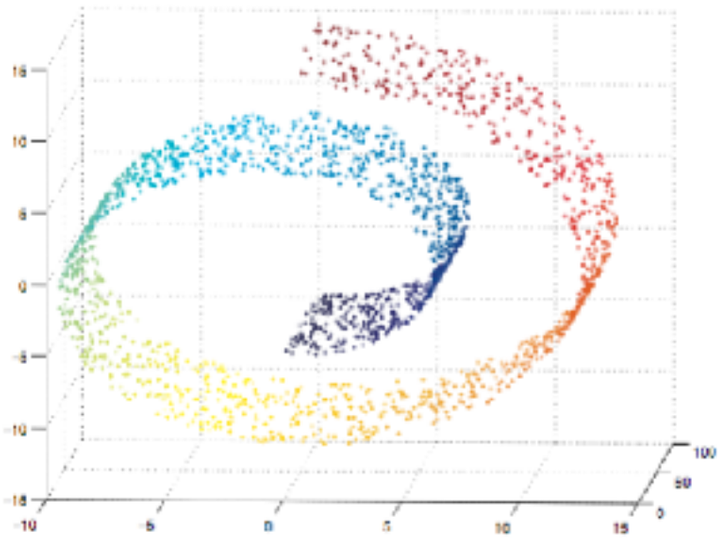
The curse of dimensionality

Density and distance between points, which is critical to clustering and to outlier analysis, becomes less meaningful

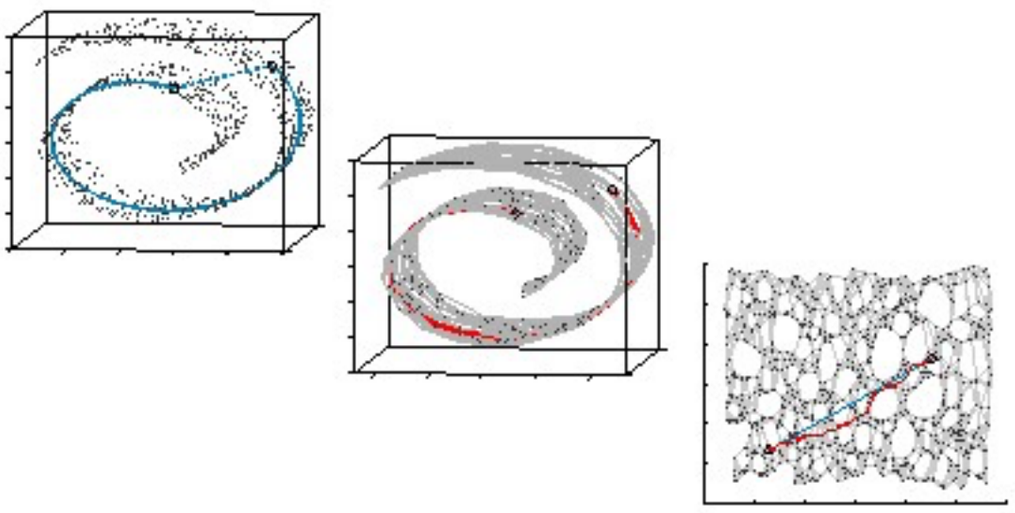


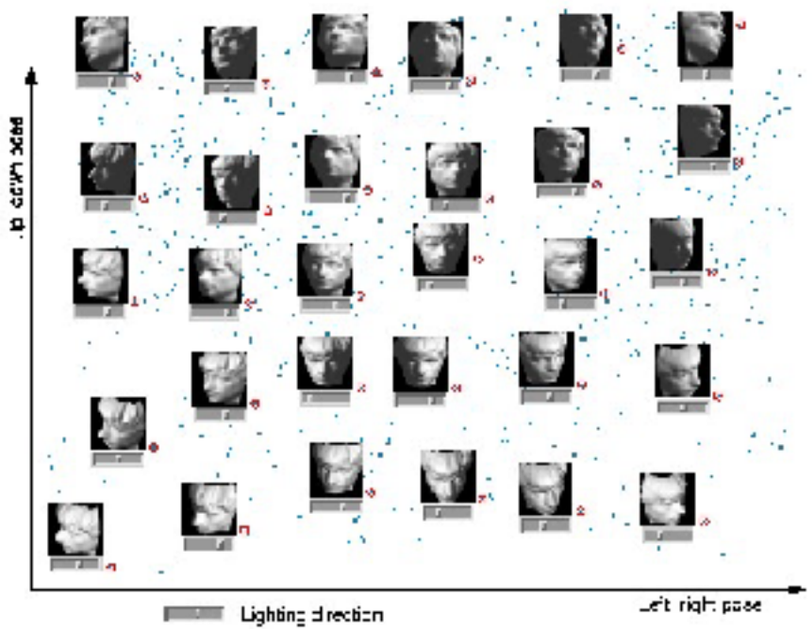
The possible combinations of subspaces will grow exponentially

When dimensionality increases, data becomes increasingly sparse



how many dimensions?

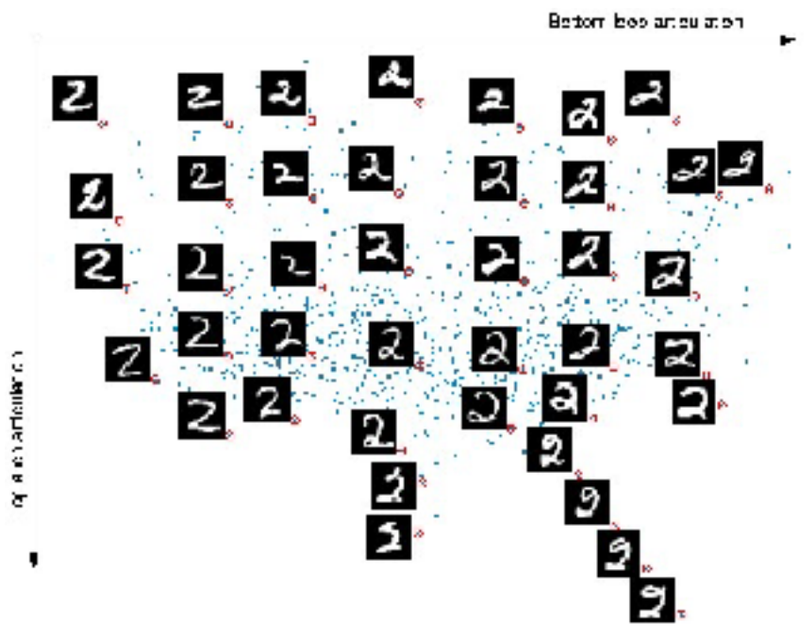




J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000.

dimensionality reduction

isomap



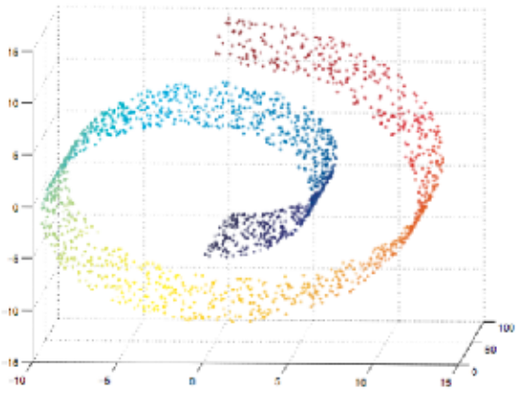
Reduces time and
space required for
data mining

dimensionality reduction

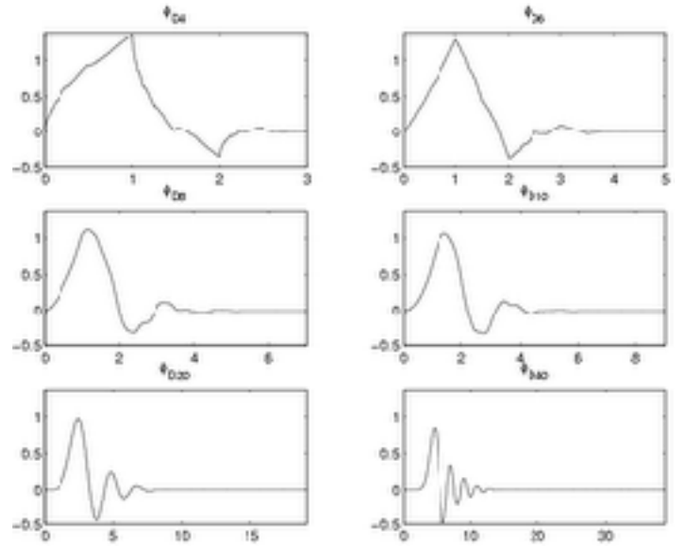
Helps eliminate
irrelevant features
and reduce noise

Avoids the curse
of dimensionality

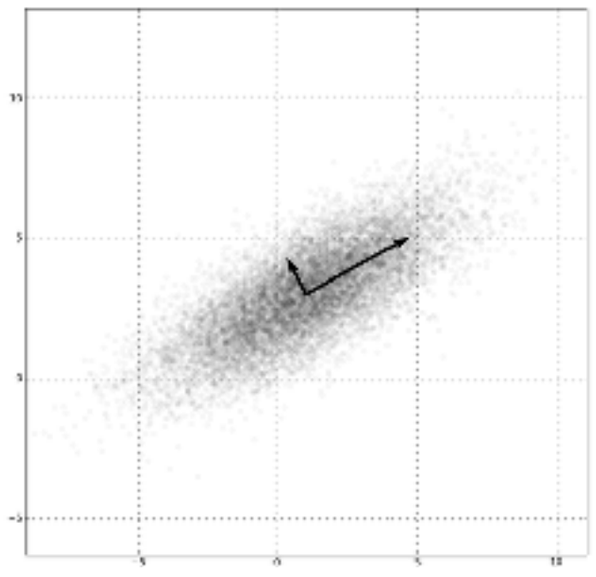
Supervised and nonlinear techniques



Wavelet transforms

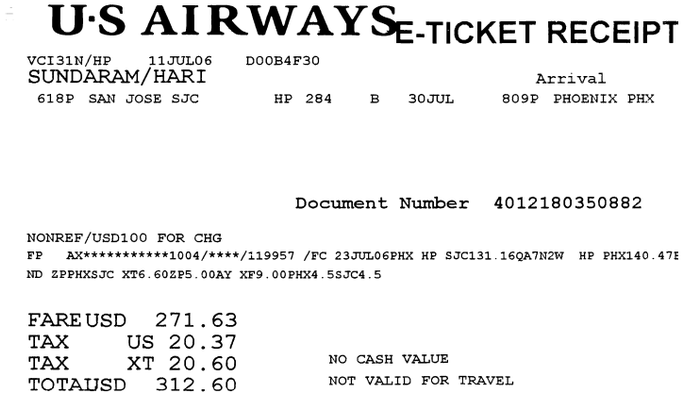


techniques



Principal Component Analysis

ATTRIBUTE SUBSET SELECTION



U·S AIRWAYS E-TICKET RECEIPT

VCI31N/HP 11JUL06 D00B4F30
SUNDARAM/HARI

					Arrival	
618P	SAN JOSE SJC	HP 284	B	30JUL	809P	PHOENIX PHX

Document Number 4012180350882

NONREF/USD100 FOR CHG
FP AX**********1004/****/119957 /FC 23JUL06PHX HP SJC131.16QA7N2W HP PHX140.47E
ND ZPPHXSJC XT6.60ZP5.00AY XF9.00PHX4.5SJC4.5

FARE	USD 271.63	
TAX	US 20.37	
TAX	XT 20.60	NO CASH VALUE
TOTAL	USD 312.60	NOT VALID FOR TRAVEL

Redundant attributes

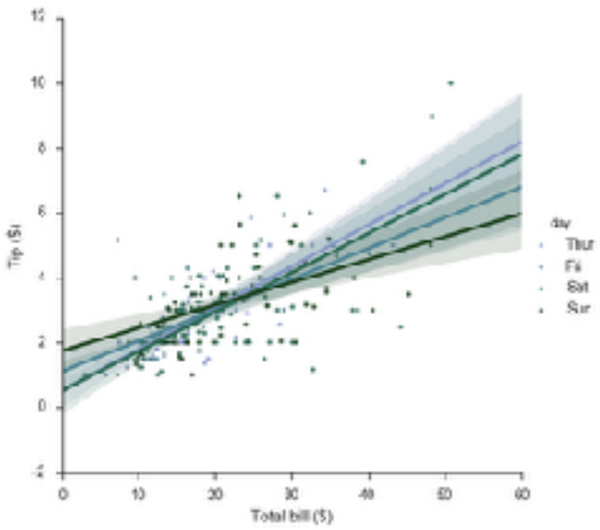
Duplicate much or all of the information contained in one or more other attributes

E.g., purchase price of a product and the amount of sales tax paid

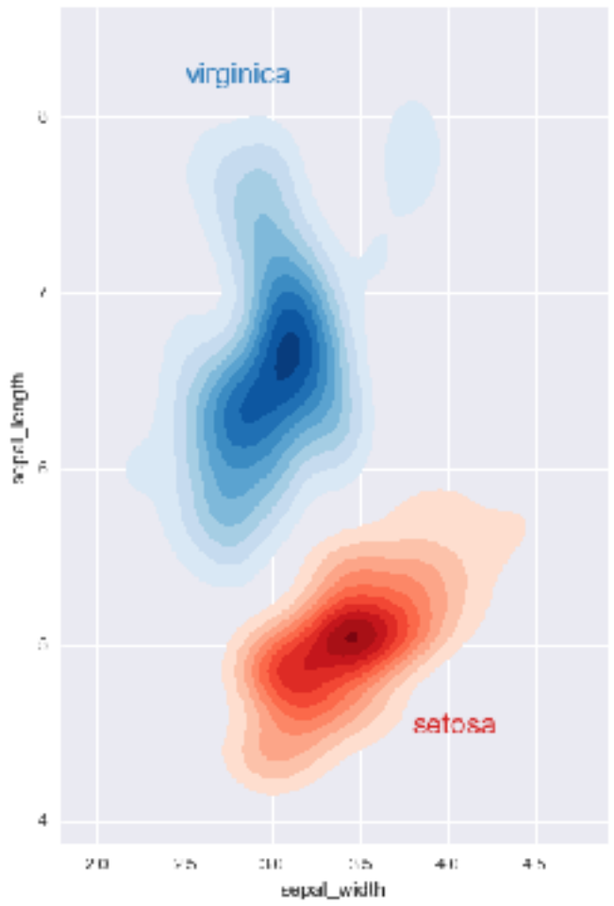
Irrelevant attributes

Contain no information that is useful for the data mining task at hand

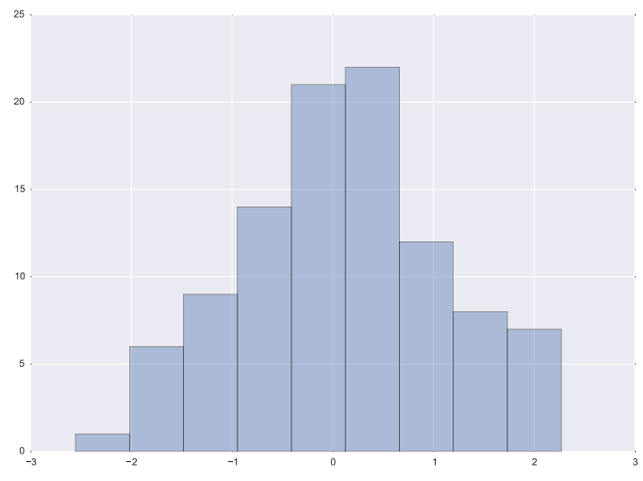
E.g., students' ID is often irrelevant to the task of predicting students' GPA



tip vs bill



leaf length vs. width



DATA SIZE REDUCTION

Reduce data volume by choosing alternative, smaller forms of data representation

Parametric methods (e.g., regression)

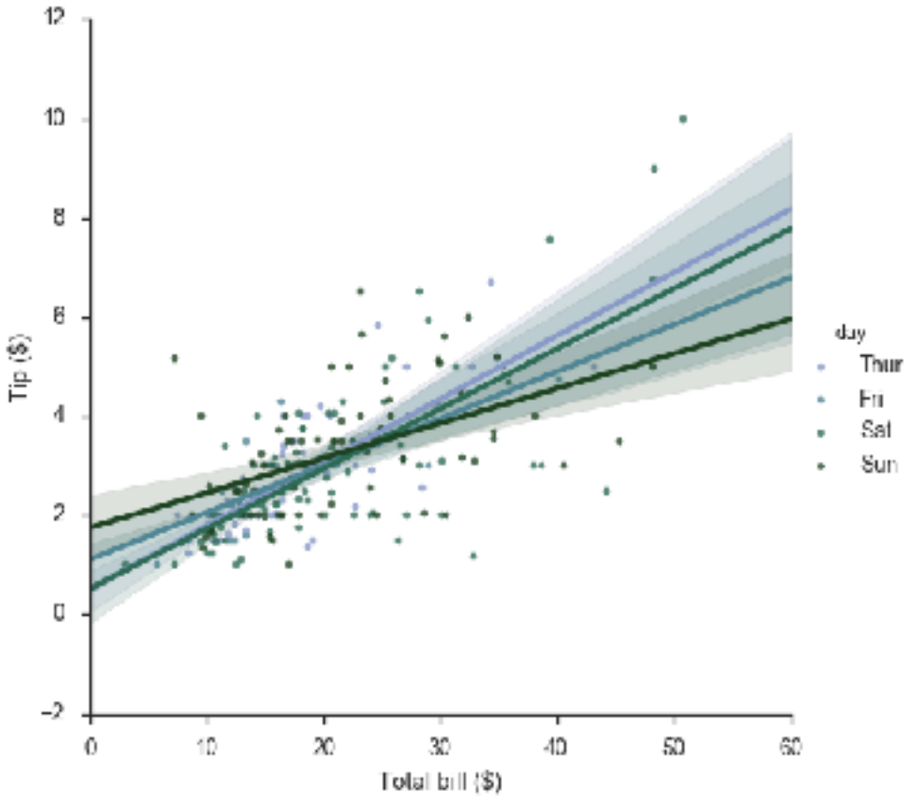
Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)

Ex.: Log-linear models—obtain value at a point in m-D space as the product on appropriate marginal subspaces

Non-parametric methods

Do not assume models

Major families: histograms, clustering, sampling,



tip vs bill

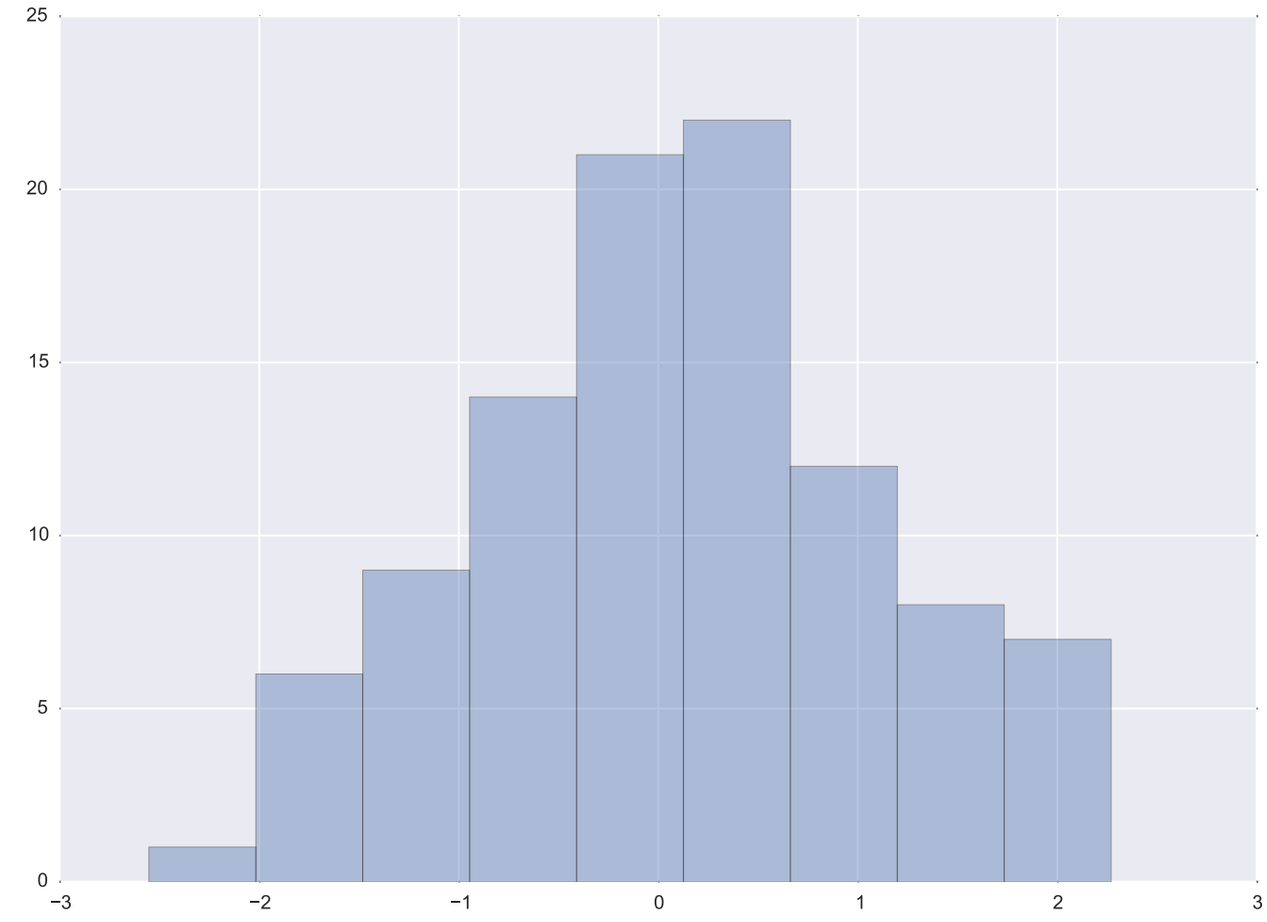
REGRESSION

Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a dependent variable (also called response variable or measurement) and of one or more independent variables (aka. explanatory variables or predictors)

The parameters are estimated so as to give a "best fit" of the data

The best fit is usually evaluated by using the least squares method, but other criteria have also been used

HISTOGRAMS



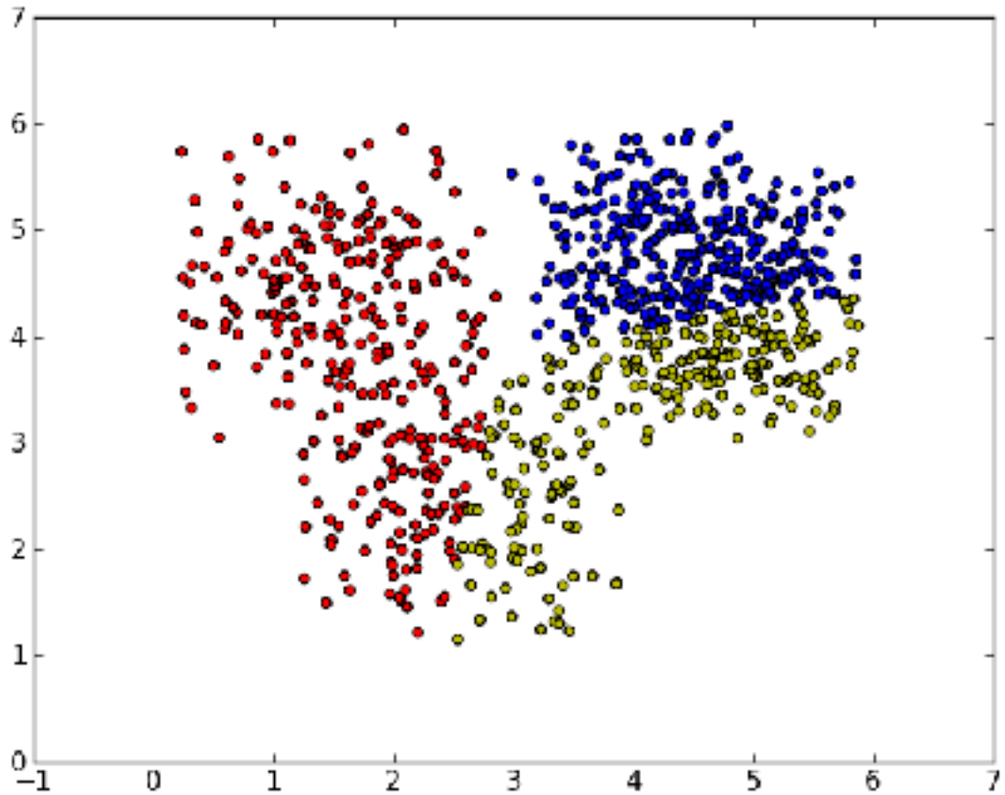
Divide data into buckets and store average (sum) for each bucket

Partitioning rules:

Equal-width: equal bucket range

Equal-frequency (or equal-depth)

CLUSTERING



Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only

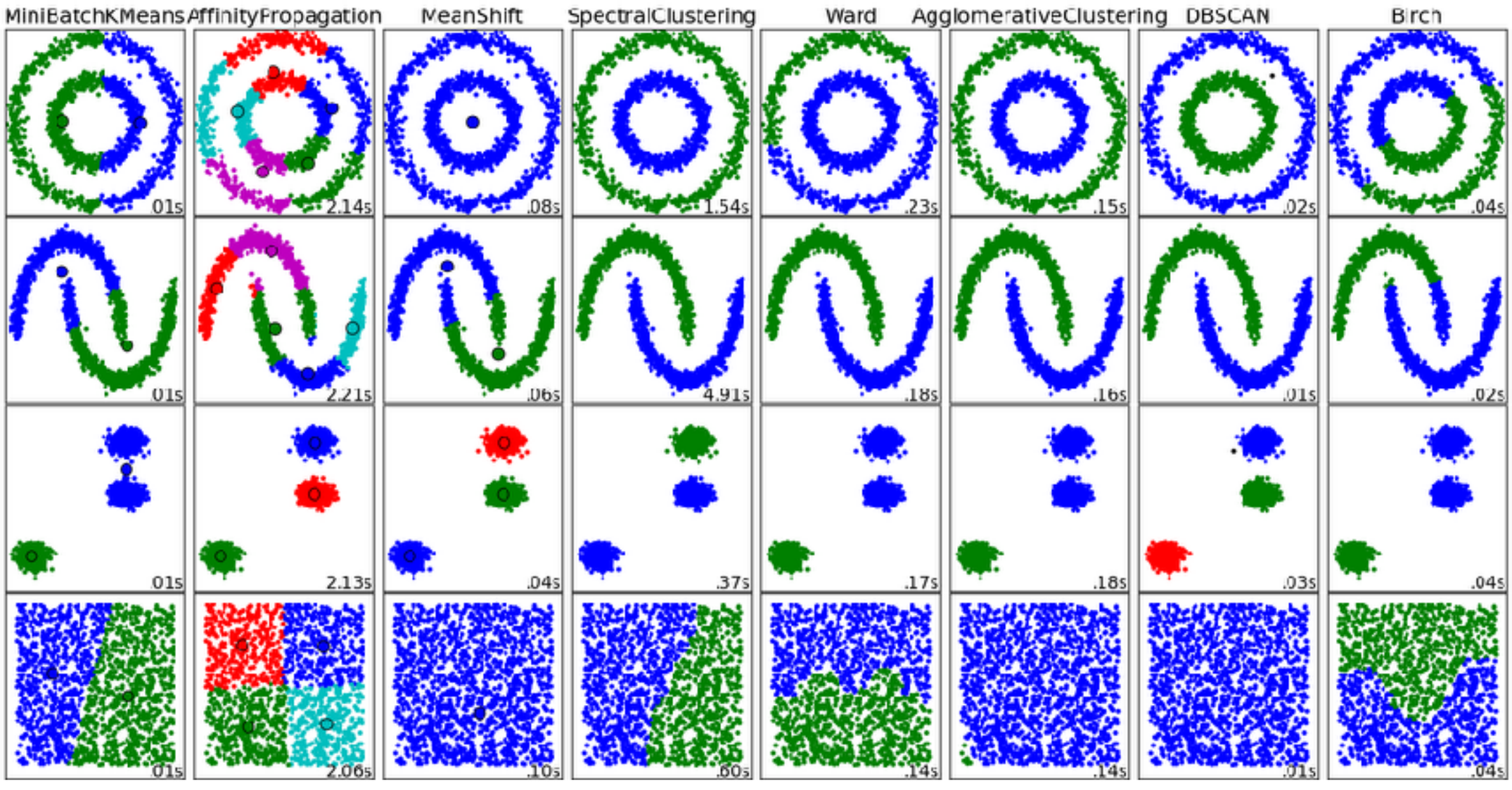
Can be very effective if data is clustered but not if data is "smeared"

Can have hierarchical clustering and be stored in multi-dimensional index tree structures

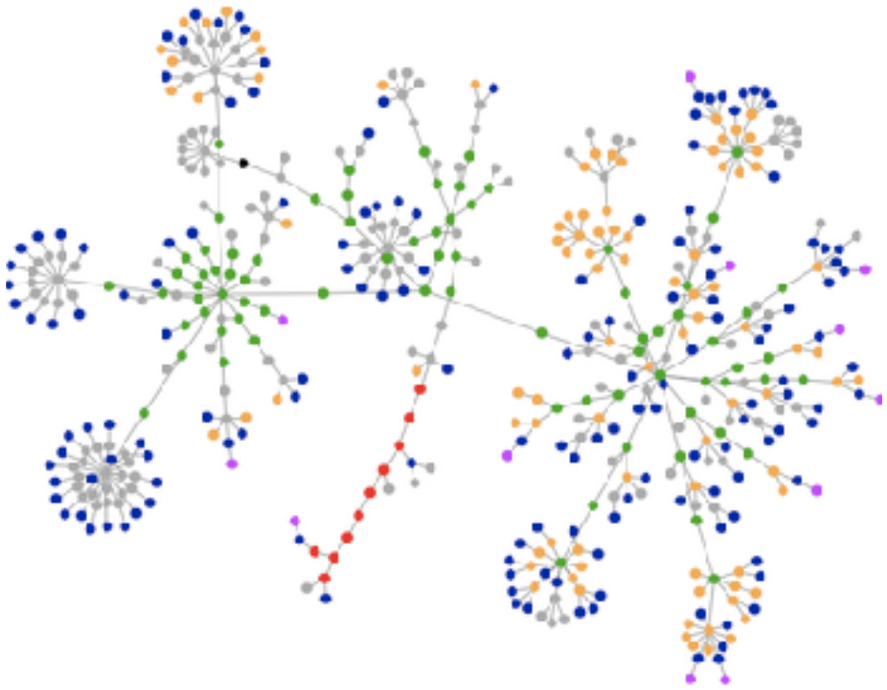
There are many choices of clustering definitions and clustering algorithms

Will study clustering in depth later on in the course

clustering comparison



http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html



SAMPLING

Sampling: obtaining a small sample s to represent the whole data set N

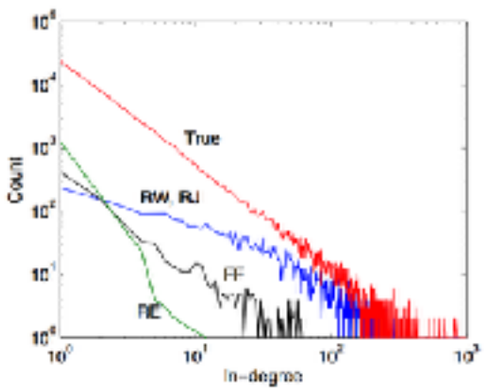
Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data

Key principle: Choose a representative subset of the data

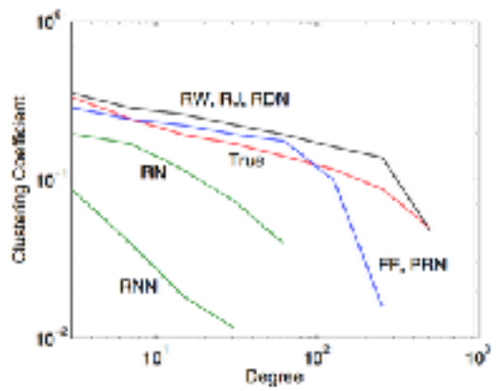
Simple random sampling may have very poor performance in the presence of skew

Develop adaptive sampling methods, e.g., stratified sampling:

Note: Sampling may not reduce database I/Os (page at a time)



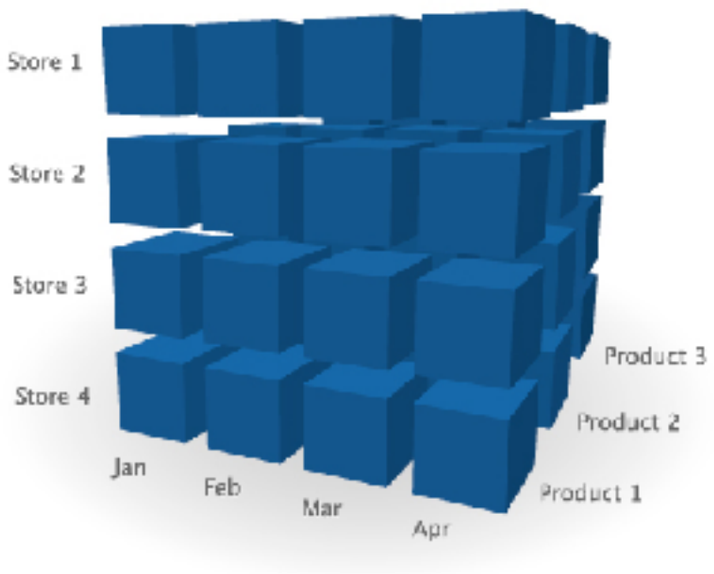
(a) S1: In-degree

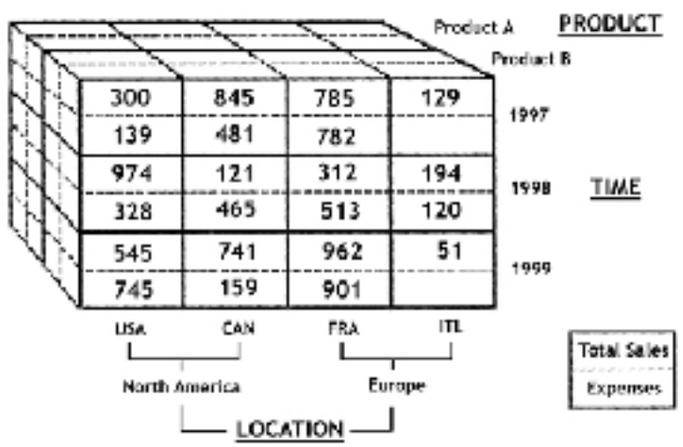


(b) S9: Clustering coef.

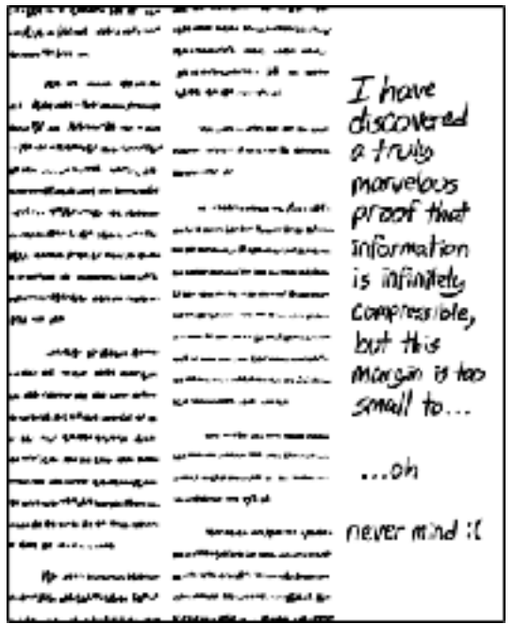
J. Leskovec and C. Faloutsos. Sampling from large graphs. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 631–636. ACM, 2006.

DATA CUBE AGGREGATION





- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible



progressive jpeg

DATA COMPRESSION

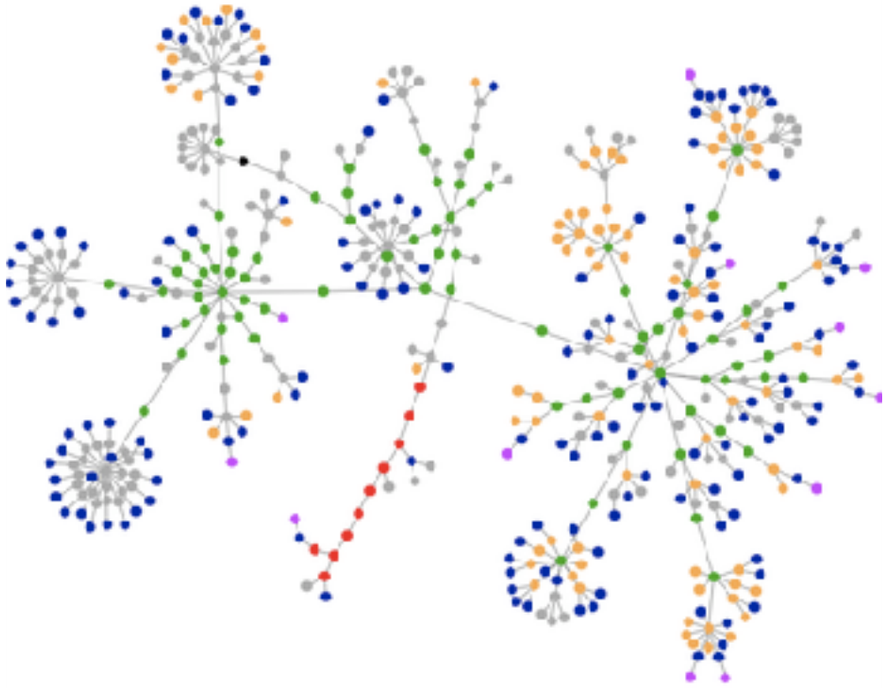
String compression

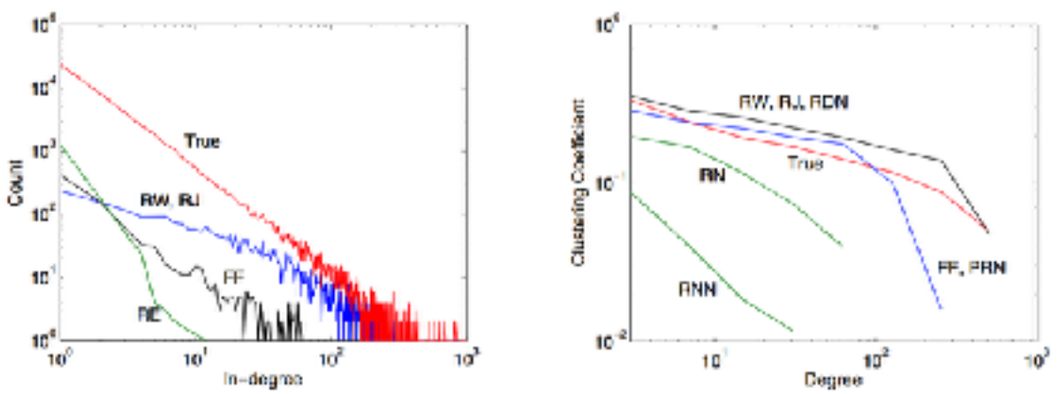
- There are extensive theories and well-tuned algorithms
- Typically lossless, but only limited manipulation is possible without expansion

Audio/video compression

- Typically lossy compression, with progressive refinement
- Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Dimensionality and data size reduction may also be considered as forms of data compression





(a) S1: In-degree (b) S9: Clustering coef.

J. Leskovec and C. Faloutsos. Sampling from large graphs. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 631–636. ACM, 2006.

TYPES OF SAMPLING

Simple random sampling

There is an equal probability of selecting any particular item

Sampling without replacement

Once an object is selected, it is removed from the population

Sampling with replacement

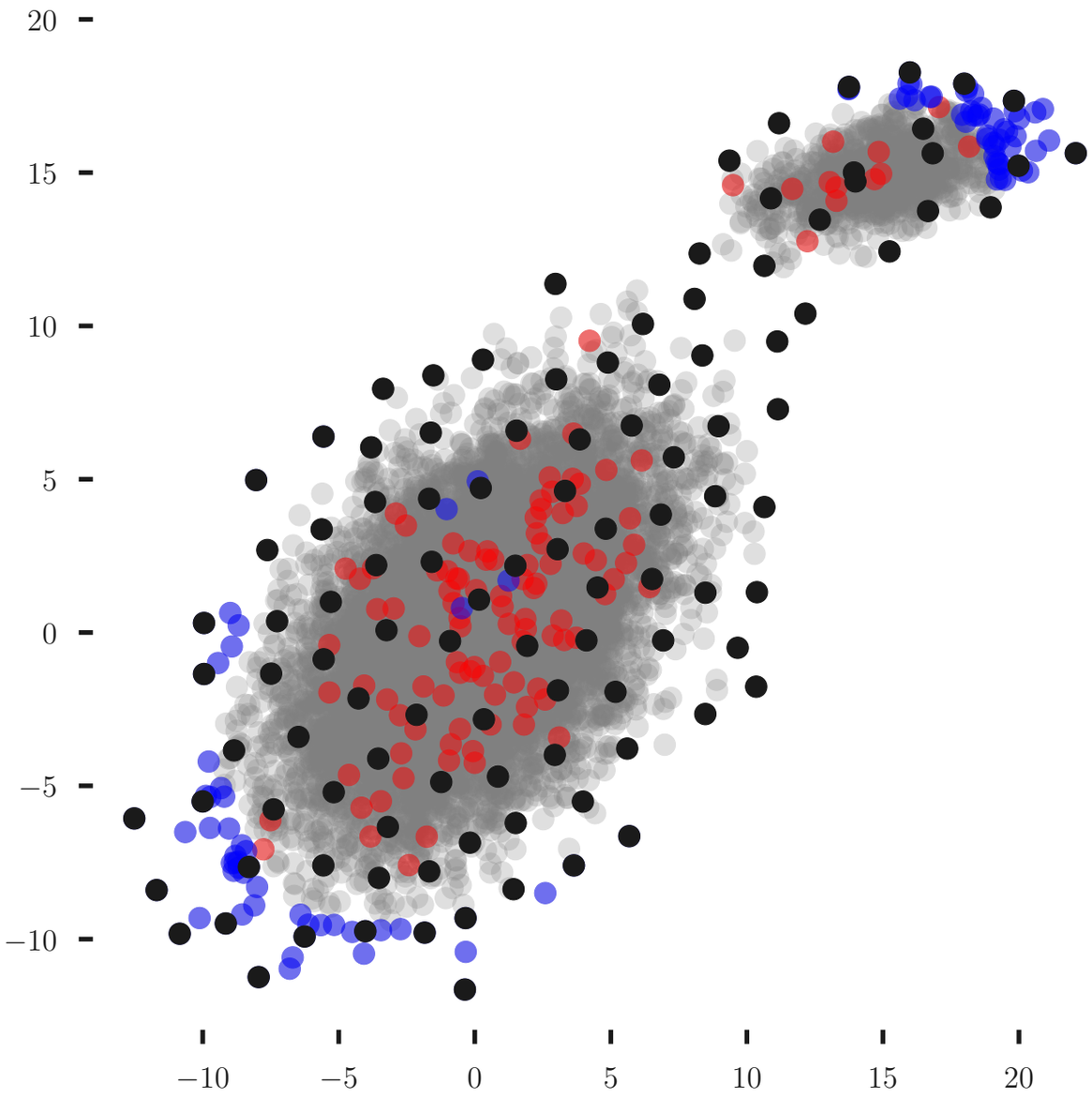
A selected object is not removed from the population

Stratified sampling:

Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)

Used in conjunction with skewed data

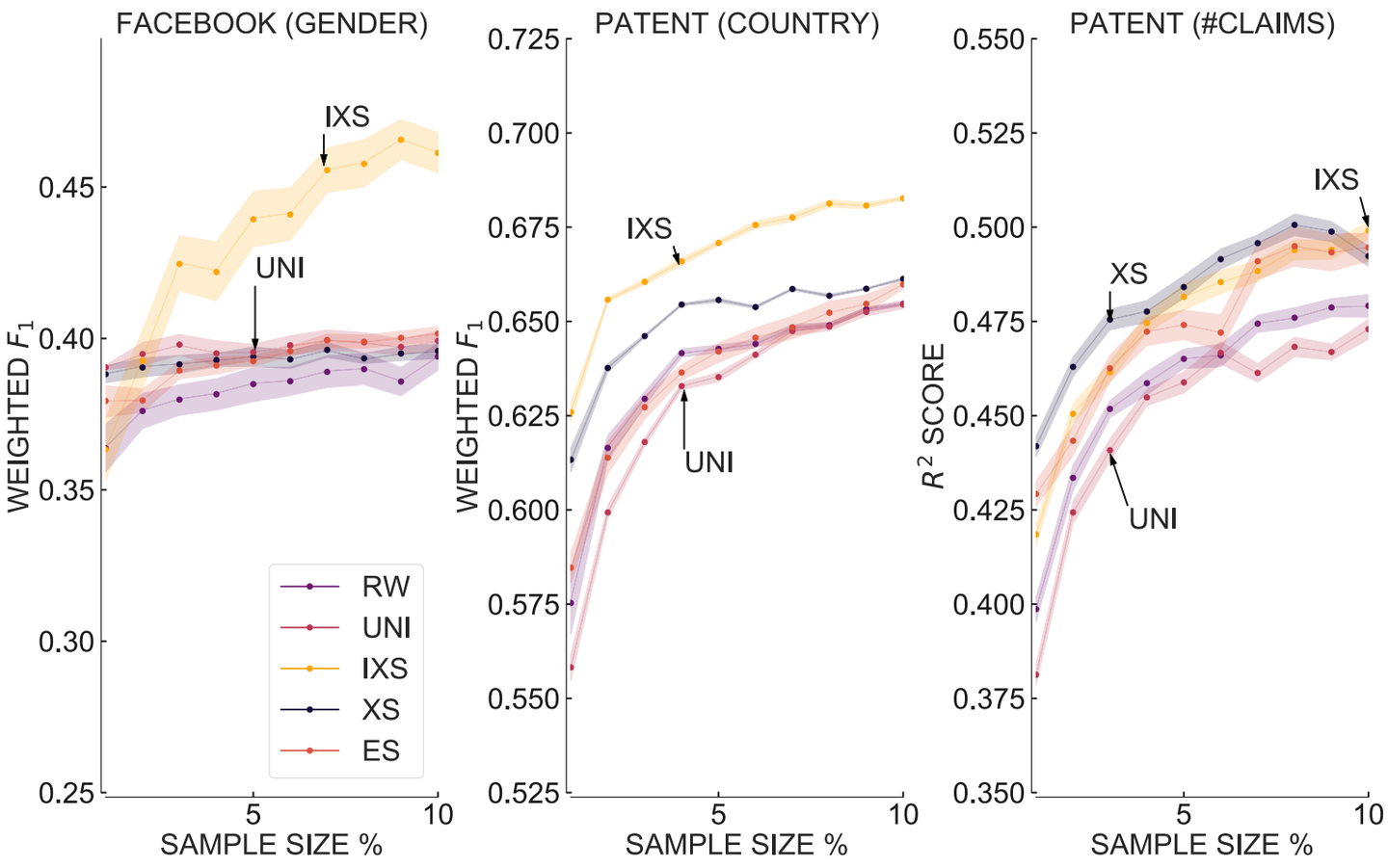
Deep Dive!



We sample two skewed classes (gray) using ideal uniform random sampling (red), our proposed link-trace sampler based on surprise (**black**) and a third sampling method that focuses on extreme nodes (**blue**). Our sampler first captures informative samples at the boundary of the classes and then samples the class interior, whereas the uniform sampler (**red**) captures samples from the center of the distribution and the extremal sampler samples extrema nodes. Notice that the clusters from both the extrema sampler and the uniform are well separated, but these samples do not cover the cluster boundary.

Current Work: Suhansanu Kumar, Hari Sundaram

Deep Dive!

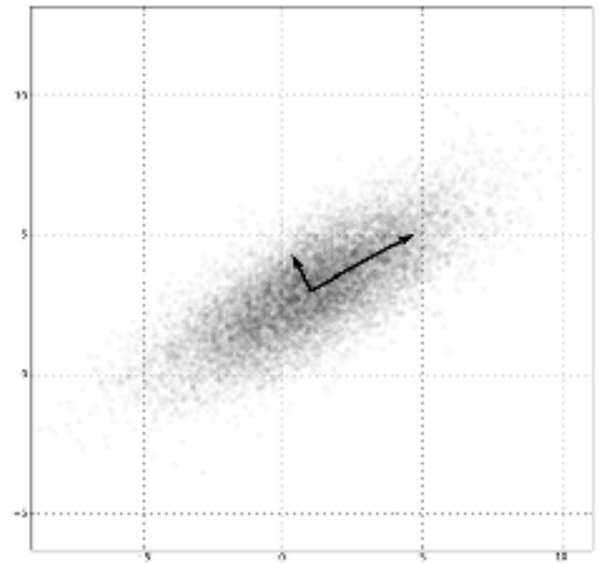


Information expansion sampler performs classification tasks well in Facebook (task: identify gender) and Patent (task: country of inventor) data.

There is no significant difference in sampler performance for a regression task in Patent data (number of claims). CI = 95%.

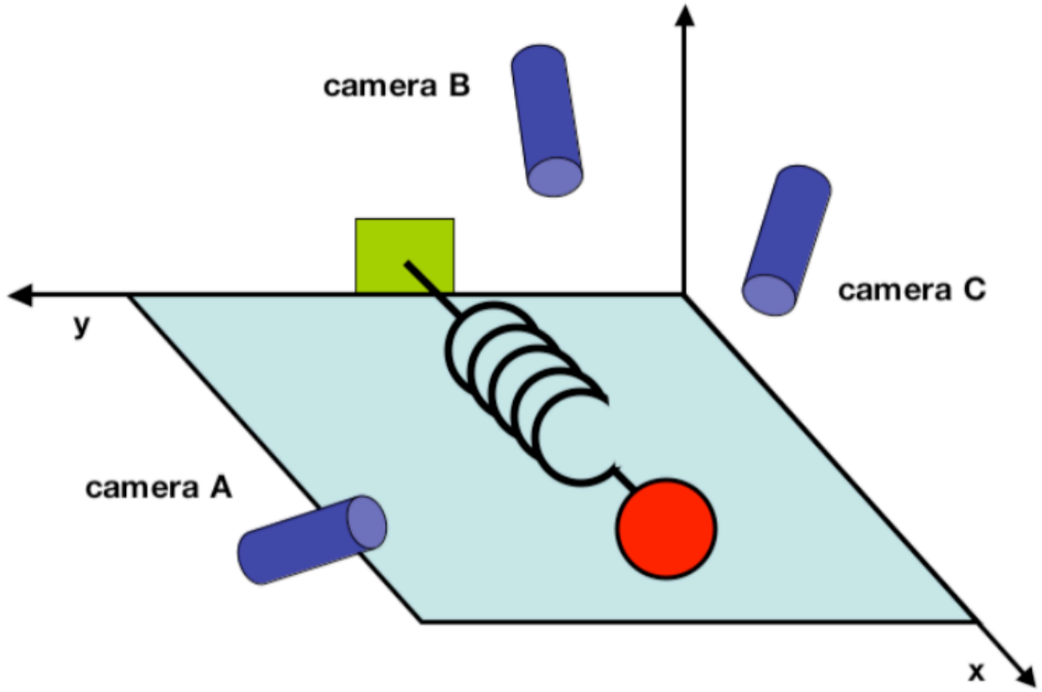
Current Work: Suhansanu Kumar, Hari Sundaram

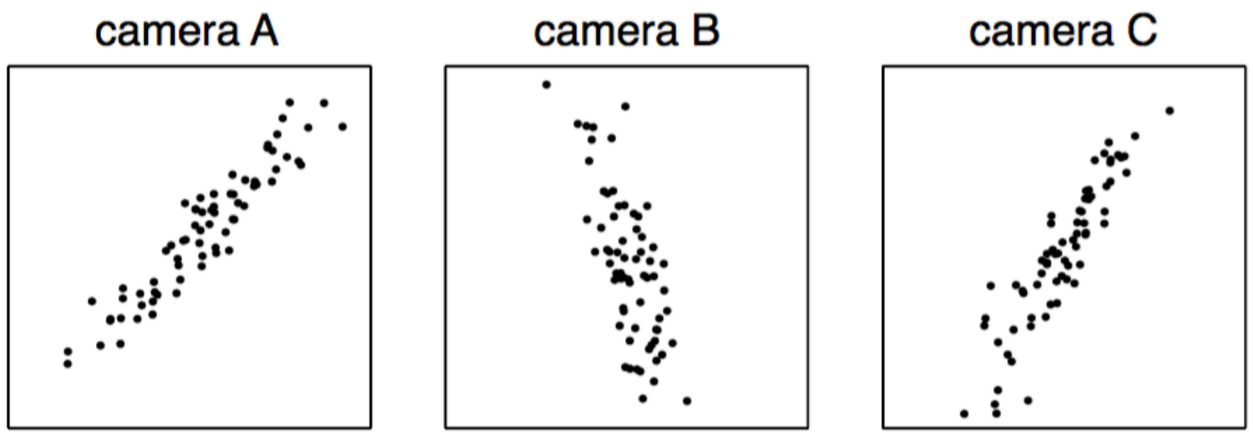
PRINCIPAL COMPONENT ANALYSIS



Jonathon Shlens. A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100, 2014.

TOY EXAMPLE: BALL & SPRING





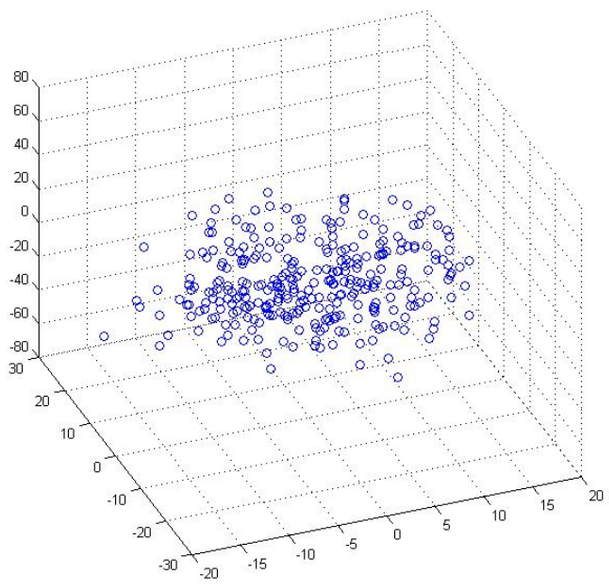
A ball of mass m attached to massless, frictionless spring

The ball moved away from equilibrium results in spring oscillating indefinitely along x-axis

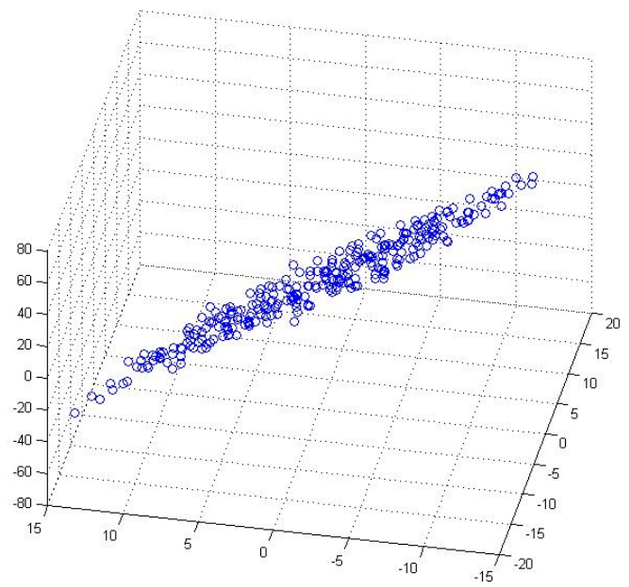
Three cameras: three dimensions

However, all dynamics can be a function of only a single variable x

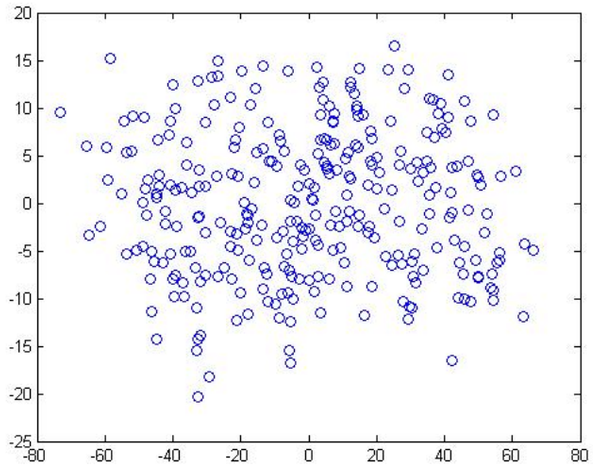
original 3D data



side view, reveals a plane!

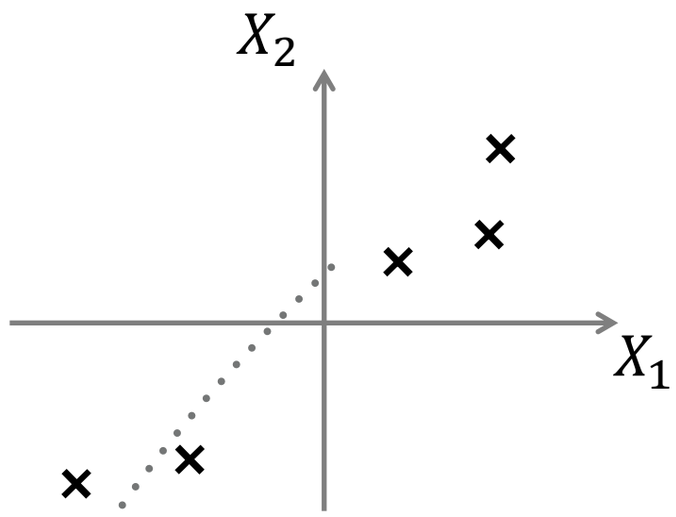


2D projection

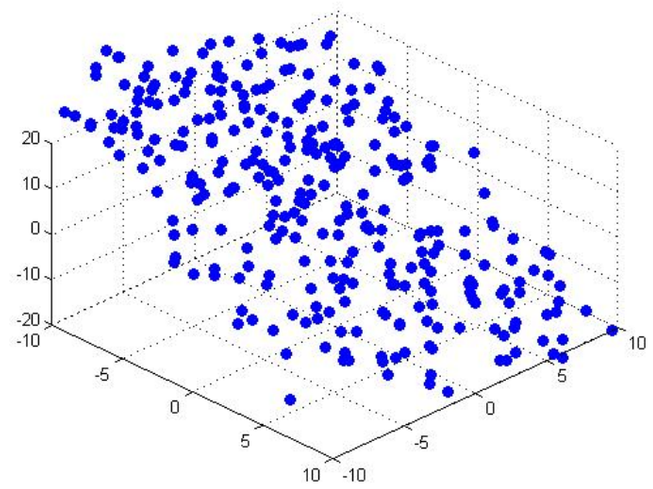


3D → 2D

2D → 1D



3D → 2D



problem formulation

2→1: Find one direction onto which to project the data so as to minimize the projection error.

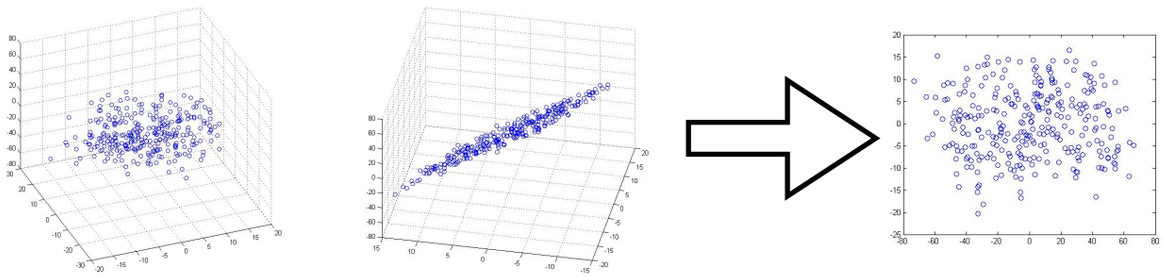
$$P_1 \in \mathbb{R}^2$$

1 vector, 2 dimensional

n→k: Find k vectors onto which to project the data, so as to minimize the projection error.

$$P_1, P_2, \ldots, P_k \in \mathbb{R}^n$$

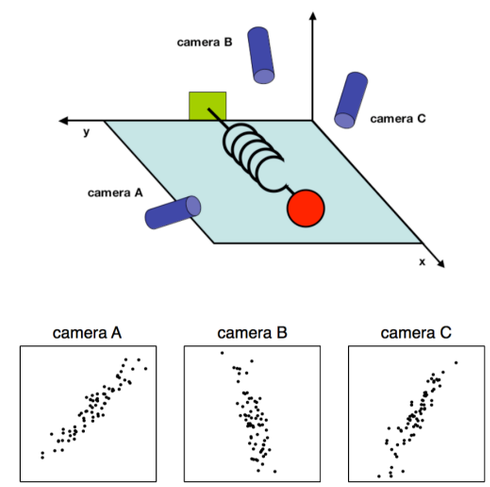
k vectors, n dimensional



Compute the most meaningful basis to re-express a noisy data set

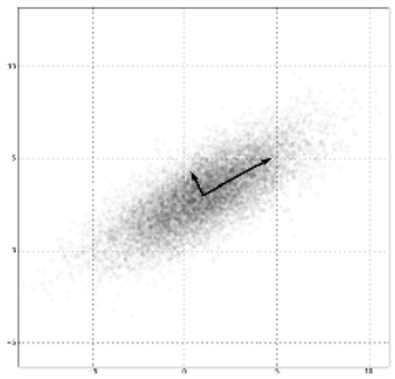
In the toy example:

Determine that the dynamics are along the x-axis

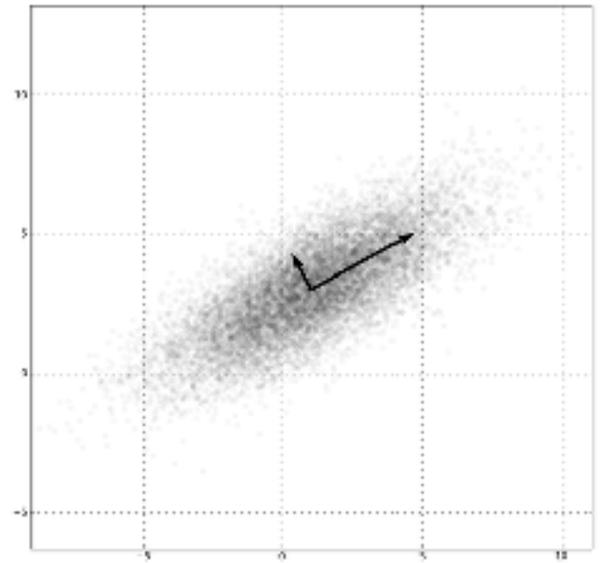


summary

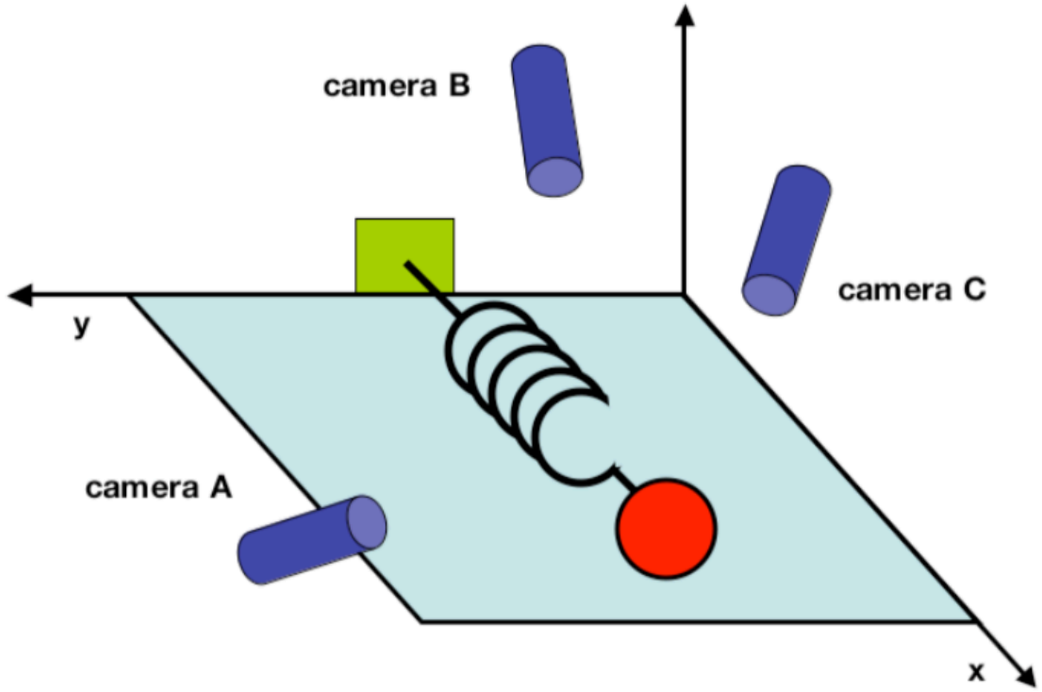
Hope that this new basis will filter out the noise and reveal hidden structure

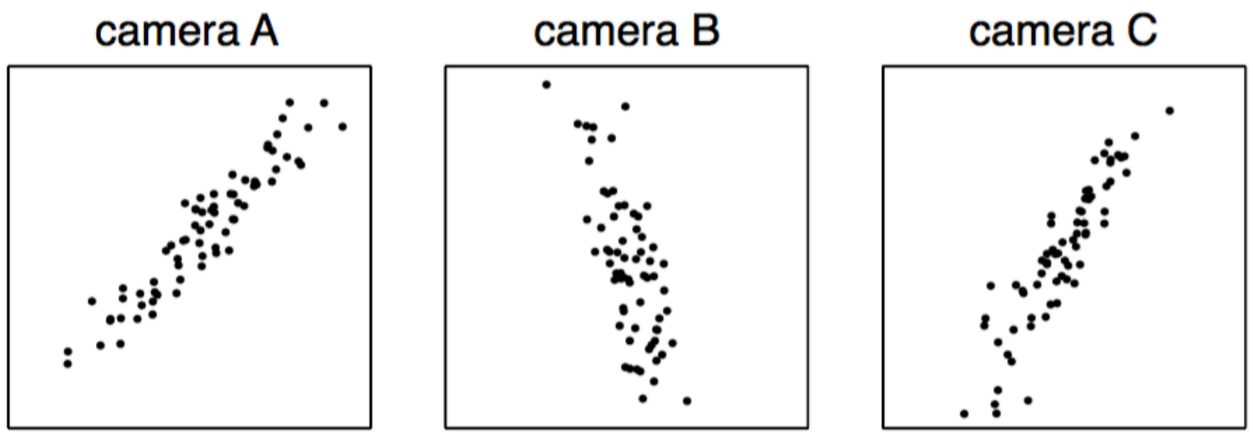


PCA ALGORITHM



NAÏVE BASIS





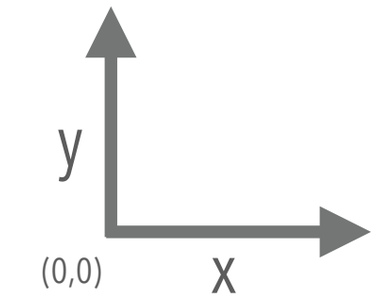
At each point in time, record 2 coordinates of ball position in each of the 3 images

After 10 minutes at 120Hz, we have 10×60×120=7200 6-dimensional vectors

These vectors can be represented in arbitrary coordinate systems

$$\vec{X} = \begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \\ x_C \\ y_C \end{bmatrix}$$

background!



standard Cartesian Basis

linear combination

basis vectors

$$(5,6) = 5 * (1,0) + 6 * (0,1)$$

$$(x,y) = x * (1,0) + y * (0,1)$$

each coefficient is just the inner product of the
original vector with the corresponding basis vector

$$x =< (x,y),(1,0) >$$

Is there a basis **better than the standard** Cartesian basis that expresses the data as a **linear** combination?

$$\vec{x} = \sum_{i}^{n} \alpha_i \vec{p}_i$$

X: m x n

$$\mathbf{PX} = \begin{bmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{bmatrix} \cdots\cdots\cdots\cdots \text{basis vectors} \quad \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{bmatrix}$$

transforms X to Y

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1} \cdot \mathbf{x_1} & \cdots & \mathbf{p_1} \cdot \mathbf{x_n} \\ \vdots & \ddots & \vdots \\ \mathbf{p_m} \cdot \mathbf{x_1} & \cdots & \mathbf{p_m} \cdot \mathbf{x_n} \end{bmatrix}$$

Y: m x n

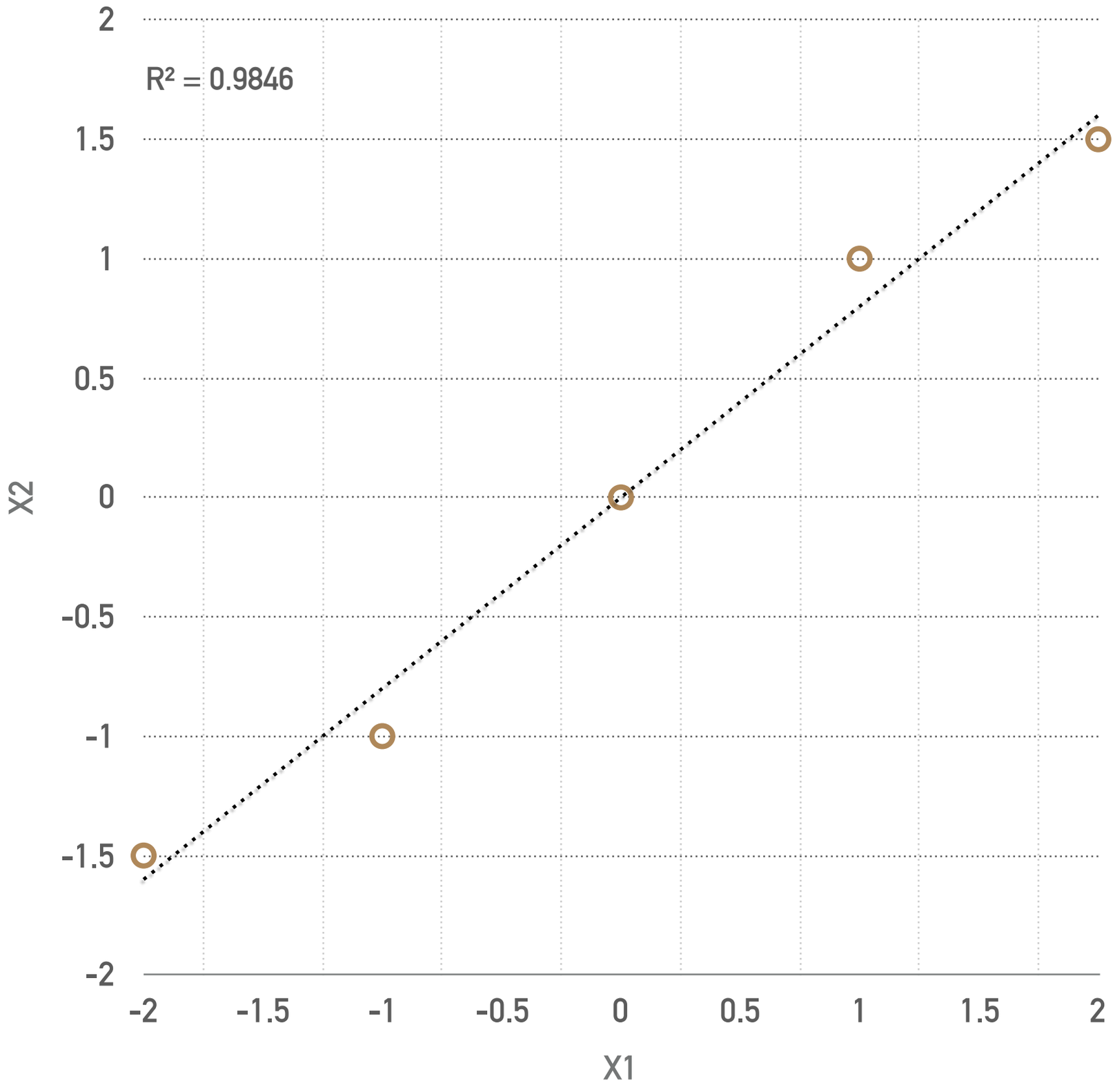
n dimensions → m dimensions

change of basis

each coefficient of y_i is a dot-product of x_i with the corresponding row in P

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{p_1} \cdot \mathbf{x_i} \\ \vdots \\ \mathbf{p_m} \cdot \mathbf{x_i} \end{bmatrix}$$

running example of two measurements



X1	X2
-2.00	-1.50
-1.00	-1.00
0.00	0.00
1.00	1.00
2.00	1.50

$$p_1 = (0.90, 0.45)$$

$$p_2 = (-0.45, 0.90)$$

assume that we know the new basis

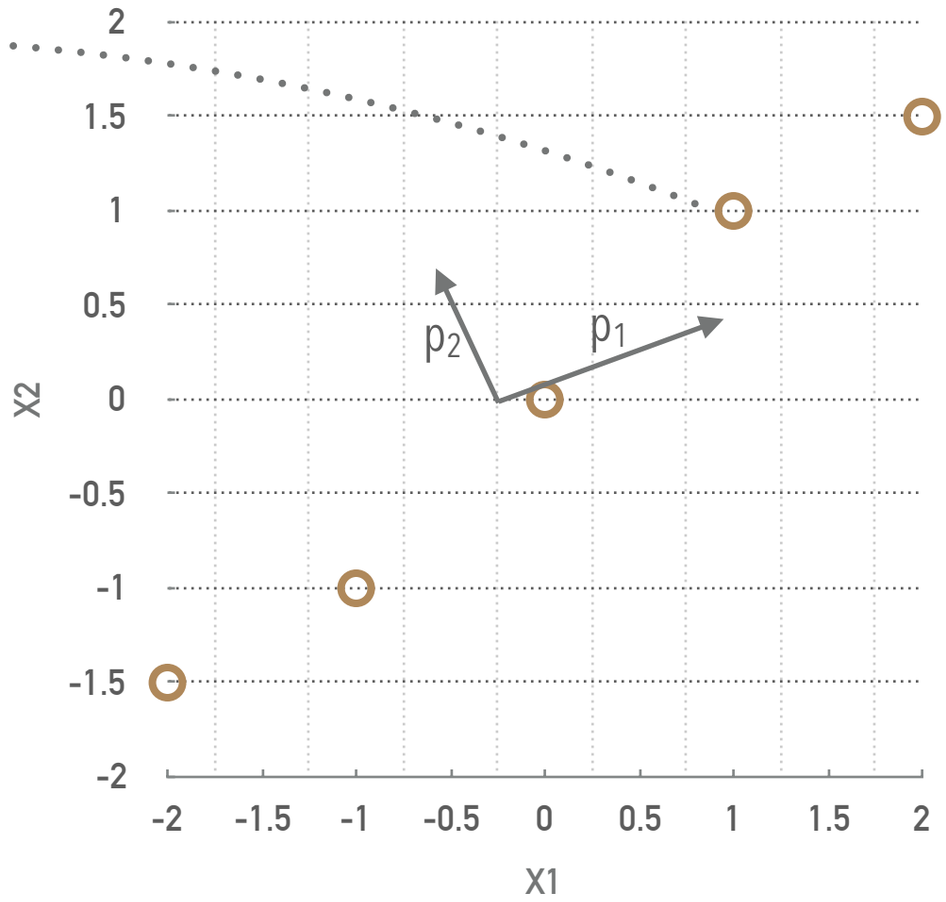
$$P = \begin{bmatrix} 0.9 & 0.45 \\ -0.45 & 0.9 \end{bmatrix}$$ orthogonal

$$x'_4 = P * x_4 = \begin{bmatrix} 0.9 & 0.45 \\ -0.45 & 0.9 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.35 \\ 0.45 \end{bmatrix}$$

$$Y = PX = \begin{bmatrix} 0.9 & 0.45 \\ -0.45 & 0.9 \end{bmatrix} * \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -1.5 & -1 & 0 & 1 & 1.5 \end{bmatrix}$$

change to the whole dataset

$$Y = \begin{bmatrix} -2.48 & -1.35 & 0 & 1.35 & 2.48 \\ -0.45 & -0.45 & 0 & 0.45 & 0.45 \end{bmatrix}$$



X1	X2
-2.00	-1.50
-1.00	-1.00
0.00	0.00
1.00	1.00
2.00	1.50

$p_1 = (0.90, 0.45)$
$p_2 = (-0.45, 0.90)$

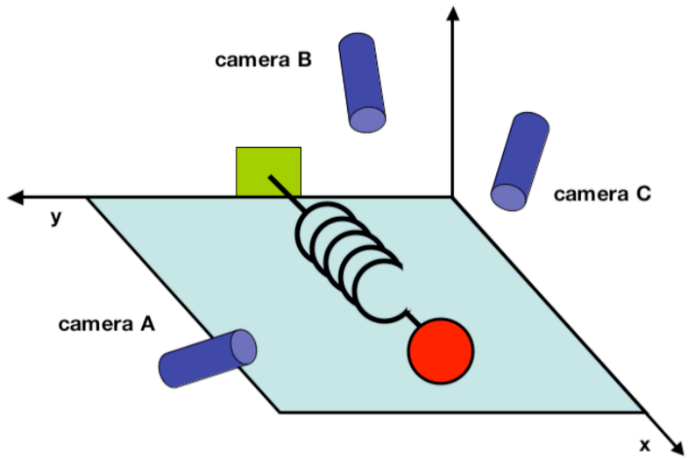
how do we discover the best transformation?

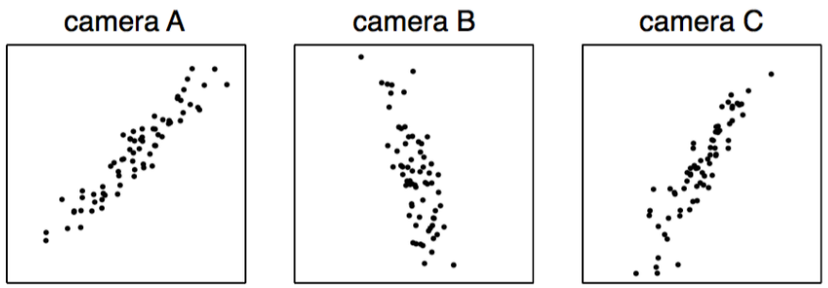
optimum basis

what criteria should we use?

Signal to Noise Ratio
Redundancy

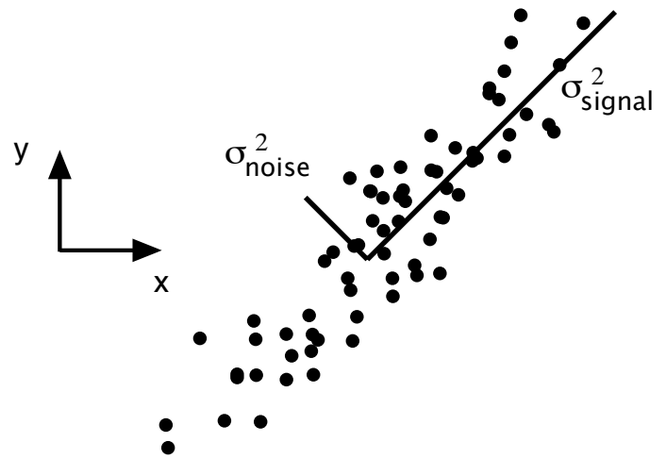
SIGNAL TO NOISE RATIO





$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$$

Signal to Noise Ratio



Measurement noise in any reasonable data set should be low.

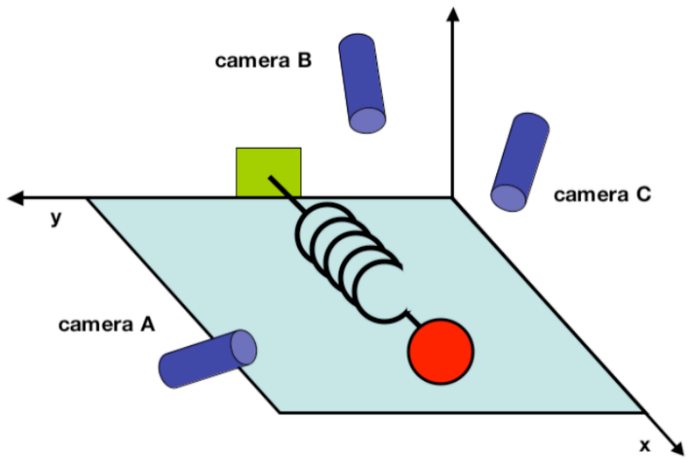
In the toy example: ball travels in straight line

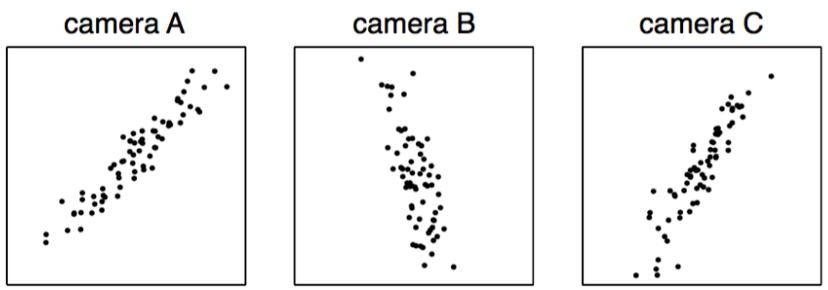
- Any deviation must be noise.
- The variance due to the signal and noise are indicated by each line in the diagram.

Assumption: directions with largest variances in our measurement space contain the dynamics of interest.

Goal 1: Maximize variances of new dimensions.

MINIMIZE REDUNDANCY



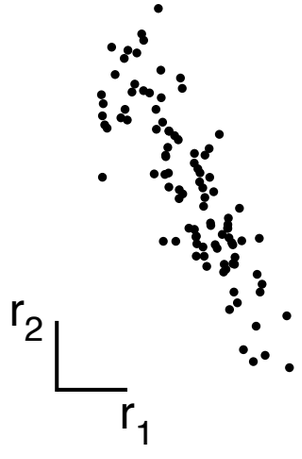


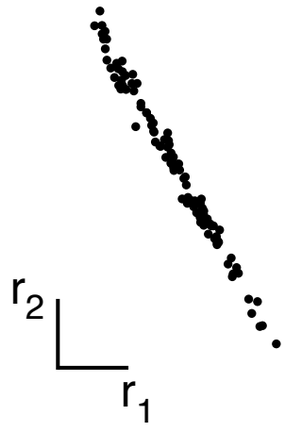
Is it necessary to record 2 variables for the ball-spring system?

Is it necessary to use 3 cameras?

Covariance Cov(X_1, X_2) (or correlation) reveals redundancy between X_1 and X_2.

low redundancy





high redundancy

Redundancy should be removed.

The new basis can use smaller number of dimensions than the naïve basis does.

Goal 2: Minimize co-variance between new dimensions.

covariance matrix

$$\Sigma = \begin{bmatrix} \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathbb{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathbb{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathbb{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathbb{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

Off-diagonal elements of $\mathbf{\Sigma}$: variances of dimensions

Large: high redundancy
Small: low redundancy

Combined goal: Covariance matrix $\mathbf{\Sigma}_Y$ of the new space should be diagonal, sorted by the diagonal values!

Diagonal elements of $\mathbf{\Sigma}$: variances of dimensions

Large: interesting dynamics
Small: noise

special case: μ=0

$$\Sigma = \begin{bmatrix} \mathbb{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathbb{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathbb{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathbb{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathbb{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathbb{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathbb{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathbb{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

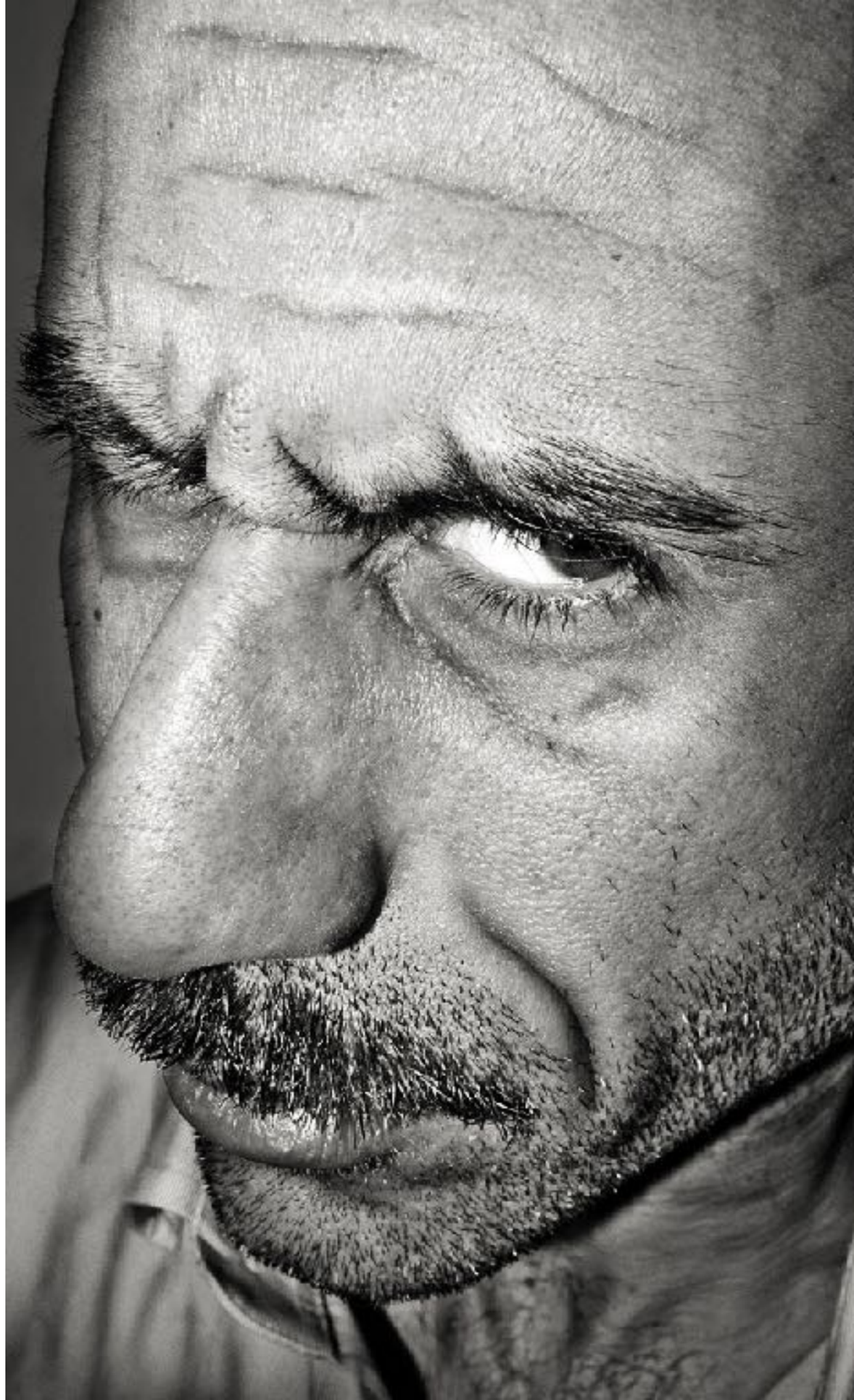
$$\Sigma_X = \frac{1}{n} X X^t$$

first simplify, by zero mean normalization

toy example is already zero mean normalized

$$C_x = \frac{1}{5}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -1.5 & -1 & 0 & 1 & 1.5 \end{bmatrix} * \begin{bmatrix} -2 & -1.5 \\ -1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 1.5 \end{bmatrix} = \begin{bmatrix} 2.0 & 1.6 \\ 1.6 & 1.3 \end{bmatrix}$$

what is the mistake in the previous slide?



the PCA objective:

Find some orthonormal matrix $\mathbf{P}$ in $\mathbf{Y} = \mathbf{PX}$ such that $\mathbf{C_Y} \equiv \frac{1}{n}\mathbf{YY}^T$ is a diagonal matrix. The rows of $\mathbf{P}$ are the *principal components* of $\mathbf{X}$.

New assumption:
P must be orthonormal

But if P is orthonormal, there is an efficient solution to find P:

Eigenvector decomposition

C_Y need not be diagonal

$$P = \begin{bmatrix} 0.9 & 0.45 \\ -0.45 & 0.9 \end{bmatrix}$$ orthogonal

$$C_Y = \frac{1}{5} * Y * Y^t = \begin{bmatrix} 3.18 & 0.69 \\ 0.69 & 0.16 \end{bmatrix}$$

How to determine P to diagonalize C_Y?

Relationship between C_X and C_Y

$$
\begin{aligned}
\mathbf{C_Y} &= \frac{1}{n}\mathbf{Y}\mathbf{Y}^T \\
&= \frac{1}{n}(\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T \\
&= \frac{1}{n}\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T \\
&= \mathbf{P}(\frac{1}{n}\mathbf{X}\mathbf{X}^T)\mathbf{P}^T \\
\mathbf{C_Y} &= \mathbf{P}\mathbf{C_X}\mathbf{P}^T
\end{aligned}
$$

what will make C_Y diagonal?

$$
\begin{aligned}
\mathbf{C_Y} &= \mathbf{P C_X P}^T \\
&= \mathbf{P}(\mathbf{E}^T \mathbf{D E})\mathbf{P}^T \\
&= \mathbf{P}(\mathbf{P}^T \mathbf{D P})\mathbf{P}^T \\
&= (\mathbf{P P}^T)\mathbf{D}(\mathbf{P P}^T) \\
&= (\mathbf{P P}^{-1})\mathbf{D}(\mathbf{P P}^{-1}) \\
\mathbf{C_Y} &= \mathbf{D}
\end{aligned}
$$

diagonalization

set P to be eigenvector of C_X

diagonalizing C_Y

Make X a mean-normalized m x n data matrix

$$\text{Compute } \underset{m \times m}{C_x} = \frac{1}{n} X X^t$$

PCA summary

eigenvectors or SVD

$$E = [e_1, e_2, \ldots, e_m]$$

compute eigenvectors of C_X

$$D = Diag(\Lambda^2)$$

$C_Y = D$

$$P = \begin{bmatrix} e_1^t \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ e_m^t \end{bmatrix}$$

projection matrix

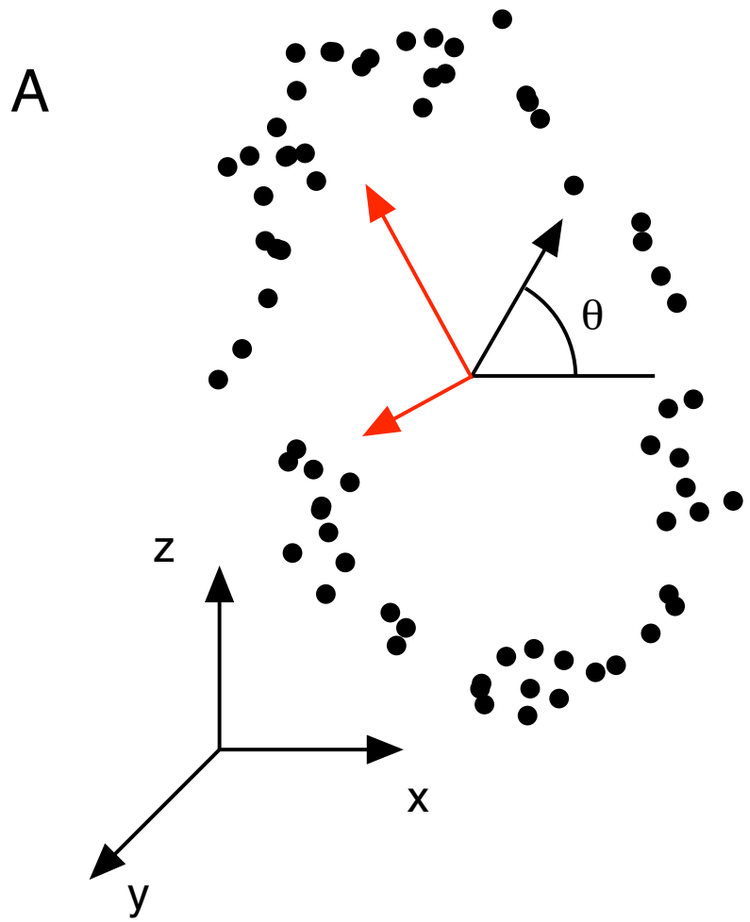
$$Y = P_k * X$$

first k eigenvectors

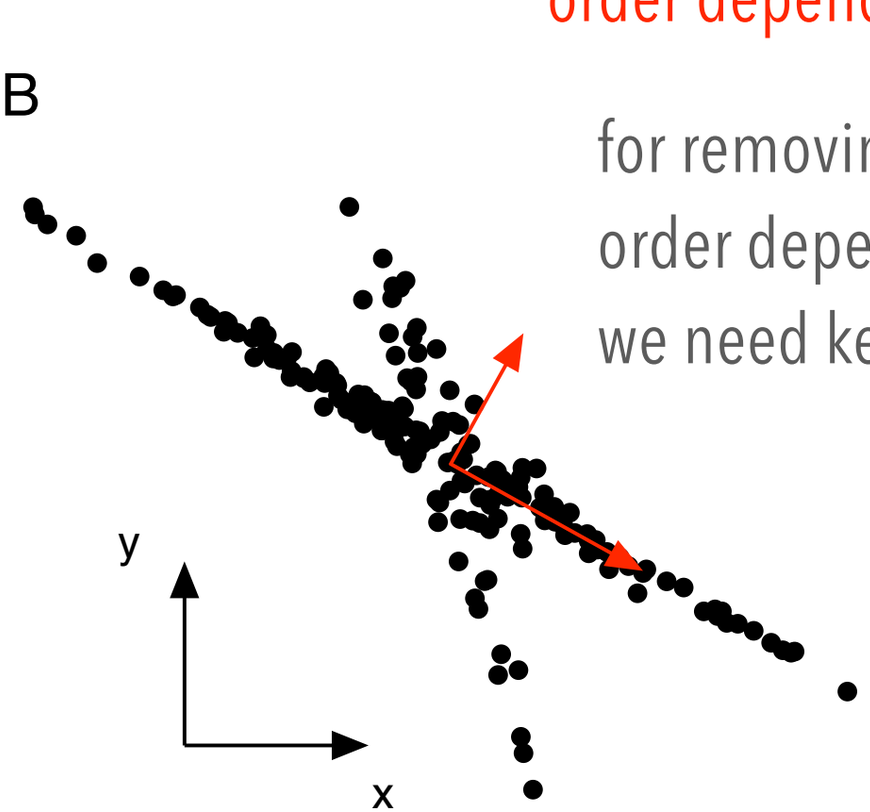
PCA failure cases

PCA de-correlates data well, when the data only contains **up to** 2nd order dependencies

for removing higher order dependencies, we need kernel PCA

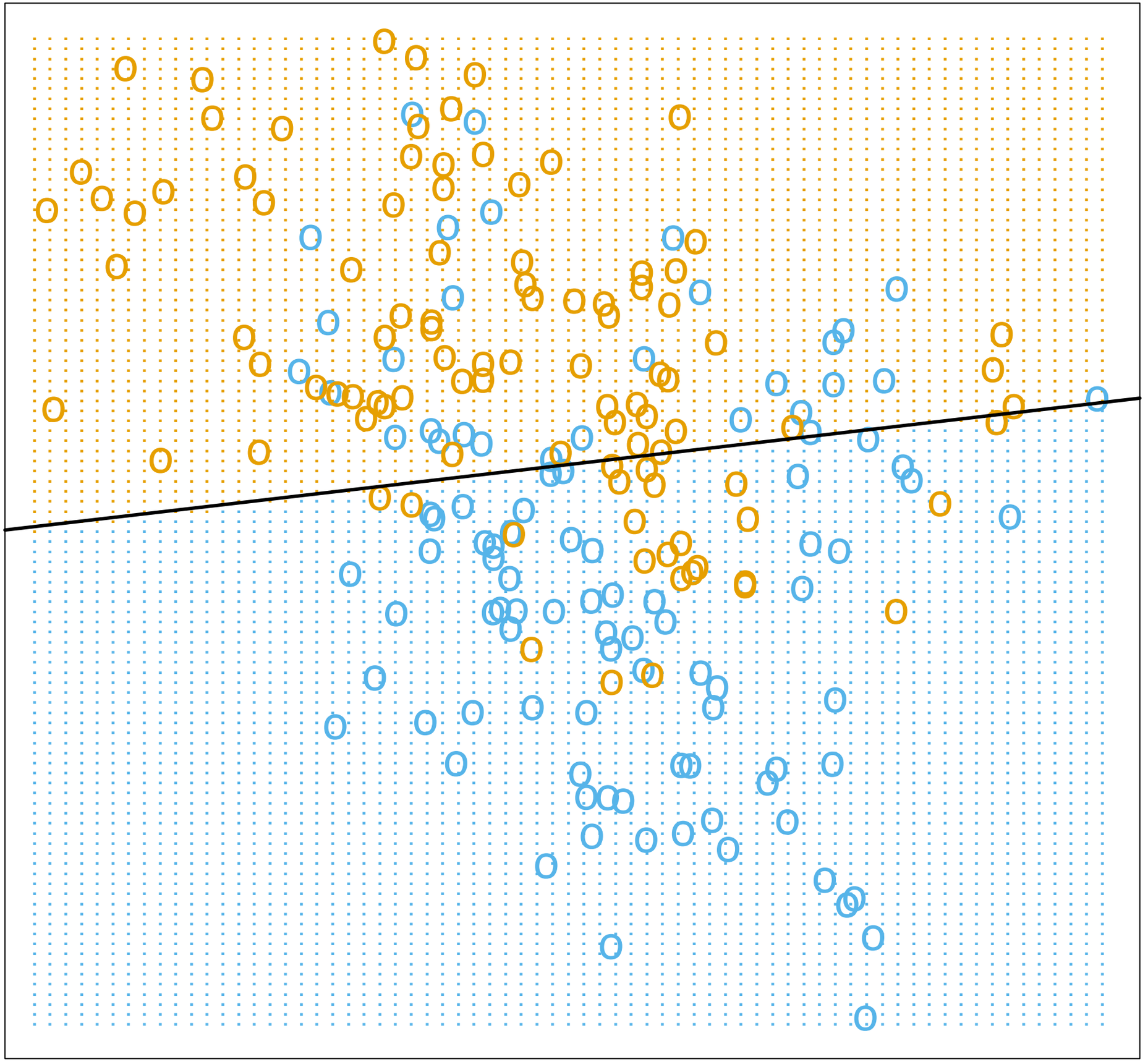


information in the phase



orthogonality is too stringent

Kernel PCA



dimensions to which we map

number of data points

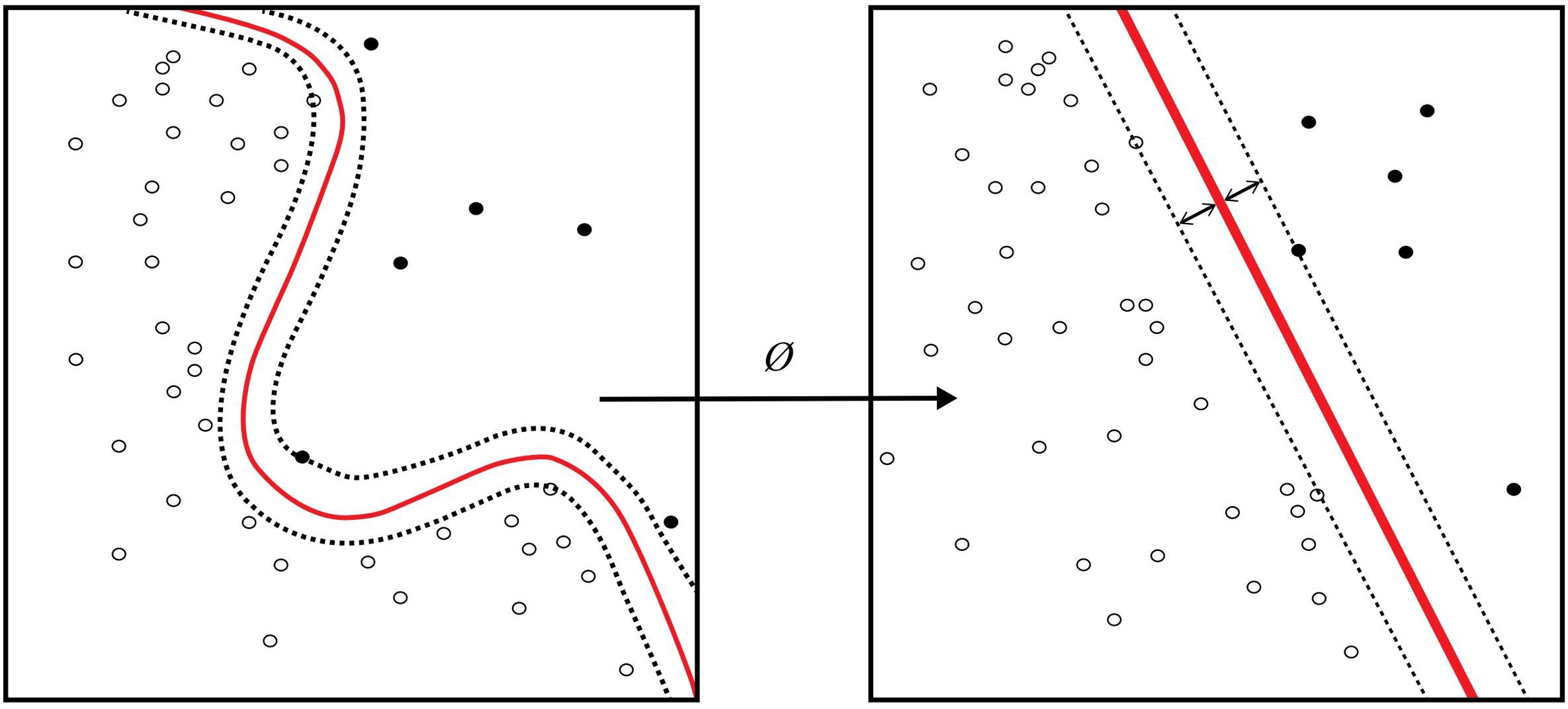
if $d \geq N$, we are guaranteed to separate the data

Deep Dive!

we will use a non linear mapping to a high dimensional space

$$Z_i = \phi(X_i)$$

Deep Dive!



$$Z_i = \phi(X_i)$$


use PCA in the
mapped space!

Deep Dive!

how do we find these kernels?

enter the kernel trick

what if:

$$K(X_i, X) = \phi(X_i) \cdot \phi(X)$$

K is a function

$$K(X_i, X) = \phi(X_i) \cdot \phi(X)$$

there exists some function K, which is equivalent to the dot product in some transformed space

Deep Dive!

now not only we don't have to find ϕ, but we <u>don't</u> have to compute the dot product!

$$K(X_i, X) = \phi(X_i) \cdot \phi(X)$$

this sounds too good to be true!

is there a catch?

$$K(X_i, X) = \phi(X_i) \cdot \phi(X)$$

there exists some function K, which is equivalent to the dot product in some transformed space

polynomial

$$K(\boldsymbol{X}_i, \boldsymbol{X}) = (\boldsymbol{X}_i \cdot \boldsymbol{X} + 1)^h$$

$$K(\boldsymbol{X}_i, \boldsymbol{X}) = \exp\left(\frac{-||\boldsymbol{X}_i - \boldsymbol{X}||^2}{2\sigma^2}\right)$$

Gaussian

DATA TRANSFORMATION AND DATA REDUCTION

A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

DATA TRANSFORMATION METHODS

Smoothing: Remove noise from data

Attribute/feature construction

- New attributes constructed from the given ones

Aggregation: Summarization, data cube construction

Normalization: Scaled to fall within a smaller, specified range

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Discretization: Concept hierarchy climbing

$$v' = \frac{v - min_A}{max_A - min_A}(max_{new,A} - min_{new,A}) + min_{new,A}$$

min-max normalization

transforming one range to another

income range: [$12,000, $98,000]
normalized to [0,1]

$$\$73,600 \rightarrow \frac{73,600 - 12,000}{98,000 - 12,000} = 0.716$$

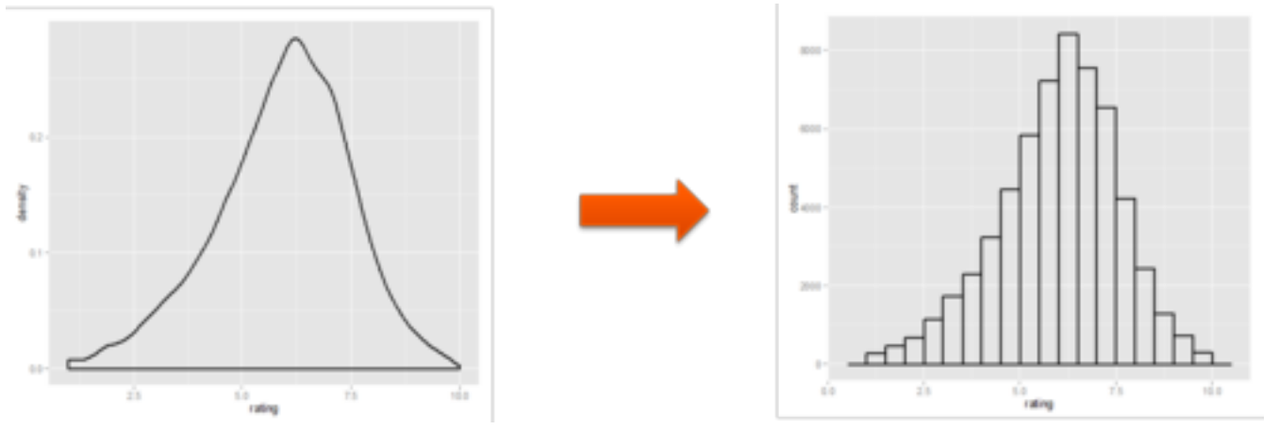
$$v = \frac{v - \mu_A}{\sigma_A}$$

z-score normalization

let $\mu = \$54{,}000$, $\sigma = \$16{,}000$

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

DISCRETIZATION

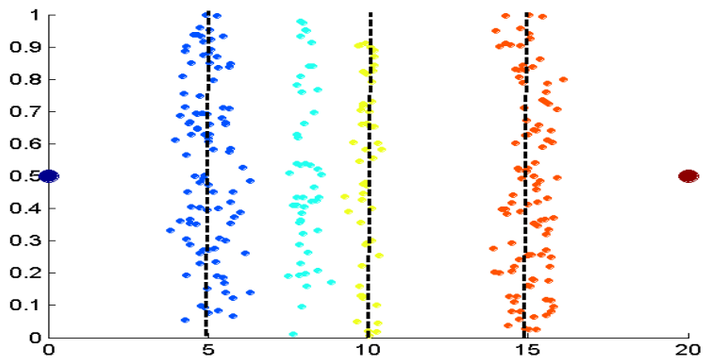


Three types of attributes

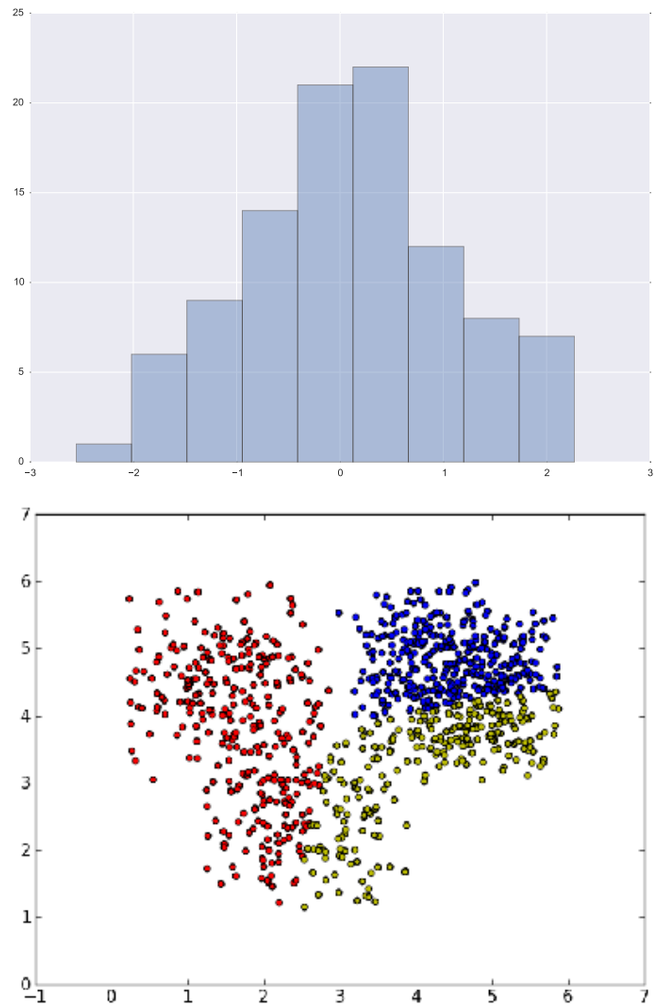
- Nominal—values from an unordered set, e.g., color, profession
- Ordinal—values from an ordered set, e.g., military or academic rank
- Numeric—real numbers, e.g., integer or real numbers

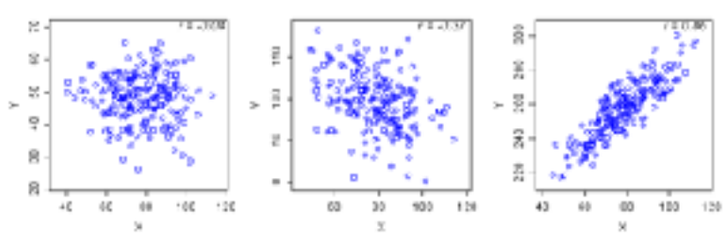
Discretization: Divide the range of a continuous attribute into intervals

- Interval labels can then be used to replace actual data values
- Reduce data size by discretization
- Supervised vs. unsupervised
- Split (top-down) vs. merge (bottom-up)
- Discretization can be performed recursively on an attribute
- Prepare for further analysis, e.g., classification



equal width binning





METHODS

Binning

Top-down split, unsupervised

Histogram analysis

Top-down split, unsupervised

Clustering analysis

(unsupervised, top-down split or bottom-up merge)

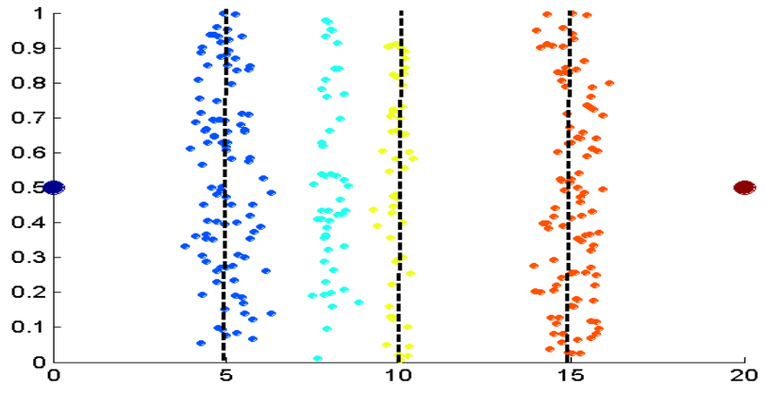
Decision-tree analysis

(supervised, top-down split)

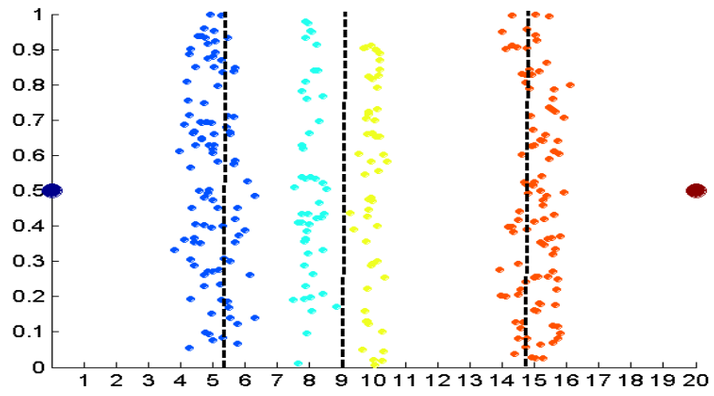
Correlation (e.g., χ^2) analysis

(unsupervised, bottom-up merge)

BINNING



equal width binning



equal frequency binning

Equal-width (distance) partitioning

Divides the range into N intervals of equal size: uniform grid

if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B – A)/N.

The most straightforward, but outliers may dominate presentation

Skewed data is not handled well

Equal-depth (frequency) partitioning

Divides the range into N intervals, each containing approximately same number of samples

BINNING & SMOOTHING

Sorted data for price (in dollars):
4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equal-frequency (equi-depth) bins:

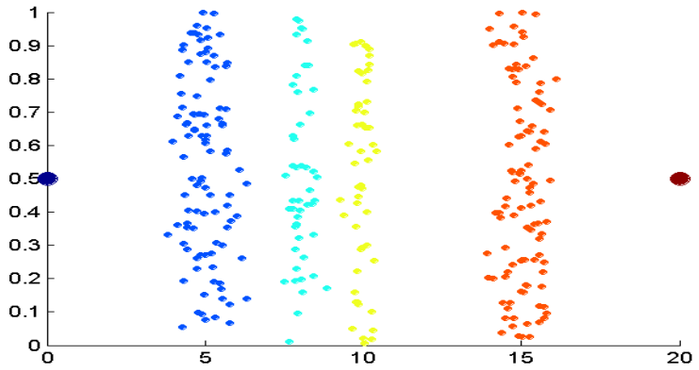
- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

Smoothing by bin means:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

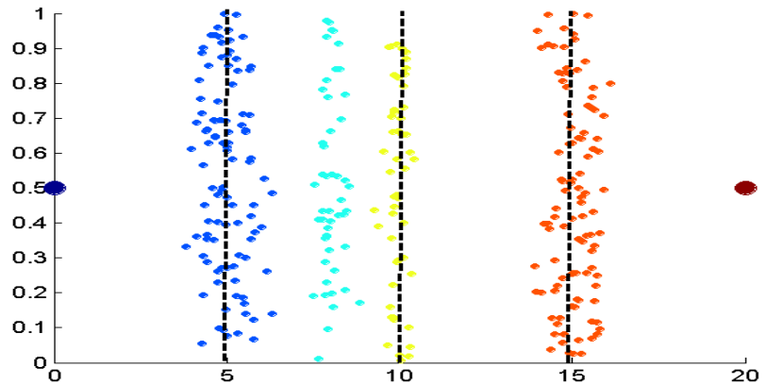
Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

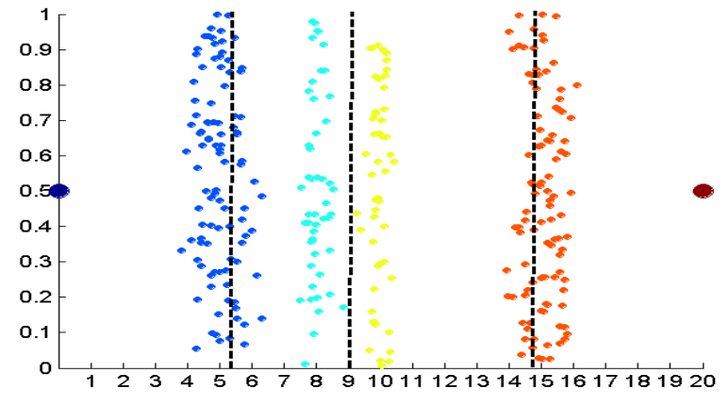


data

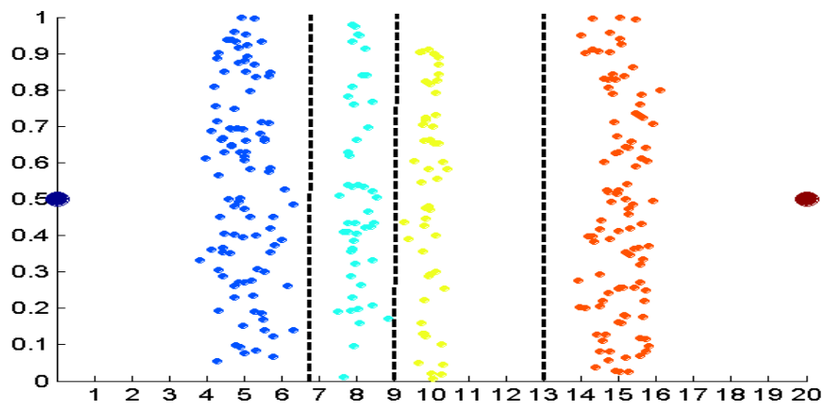
binning vs clustering



equal width binning

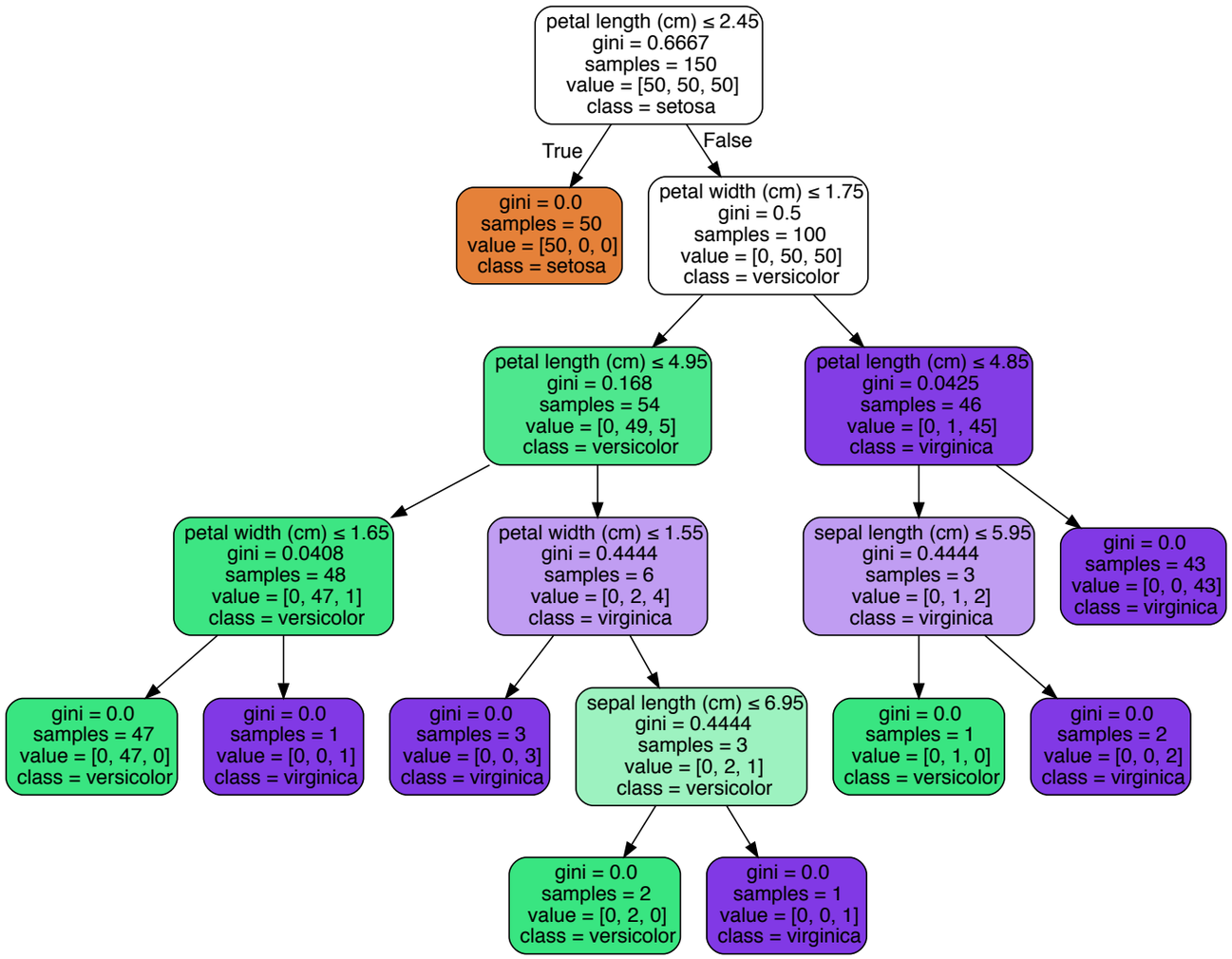


equal frequency binning



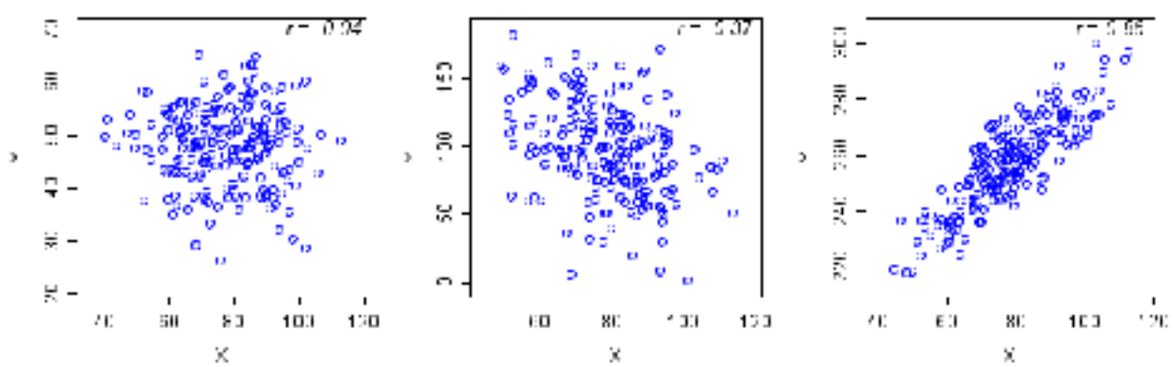
k-means clustering

CLASSIFICATION & CORRELATION



decision tree on iris dataset

http://scikit-learn.org/stable/modules/tree.html



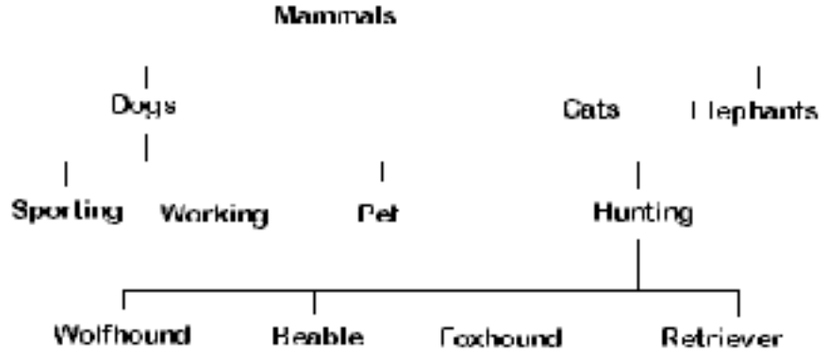
Classification (e.g., decision tree analysis)

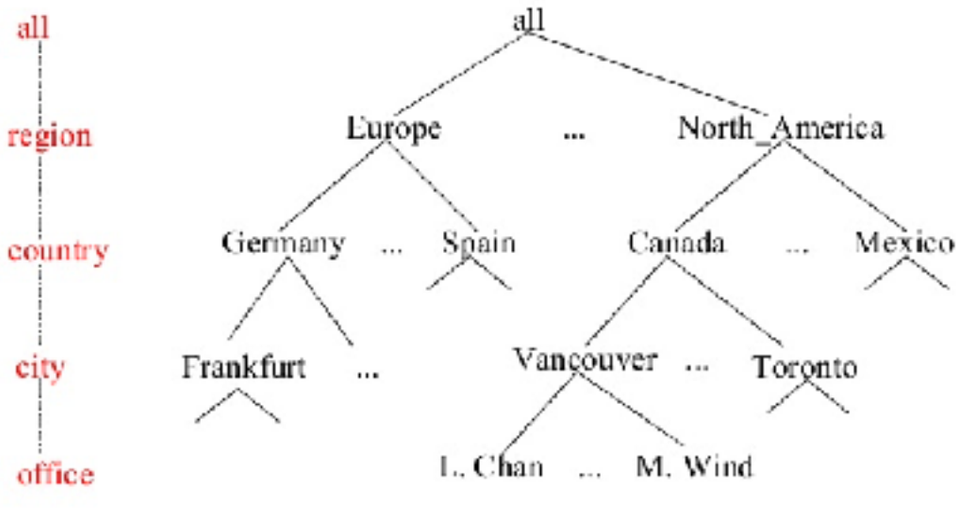
- Supervised: Given class labels, e.g., cancerous vs. benign
- Using entropy to determine split point (discretization point)
- Top-down, recursive split
- Details to be covered in Chapter "Classification"

Correlation analysis (e.g., Chi-merge: $\chi 2$-based discretization)

- Supervised: use class information
- Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low $\chi 2$ values) to merge
- Merge performed recursively, until a predefined stopping condition

CONCEPT HIERARCHY





what is the problem here?

Concept hierarchy organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse

Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts

street < city < state < country

Specification of a hierarchy for a set of values by explicit data grouping

{Urbana, Champaign, Chicago} < Illinois

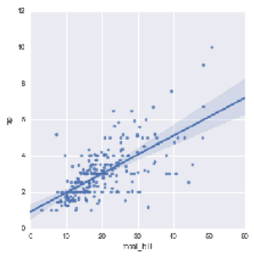
Specification of only a partial set of attributes

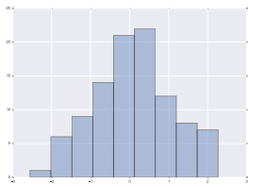
E.g., only street < city, not others

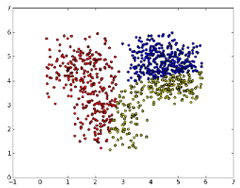
Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values

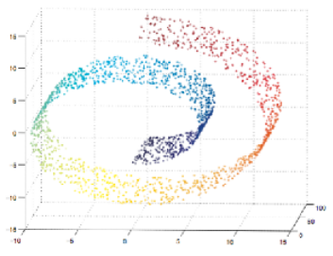
Country	15 distinct values
State	365 distinct values
City	3567 distinct values
Street	674339 distinct values

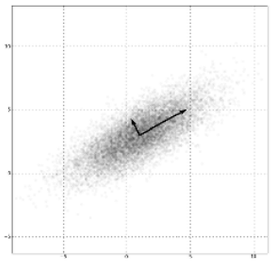
SUMMARY

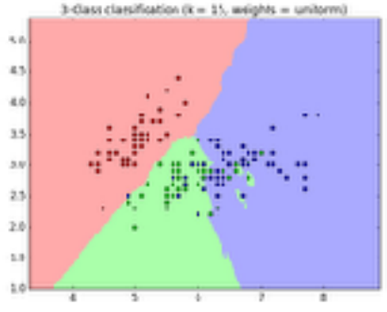


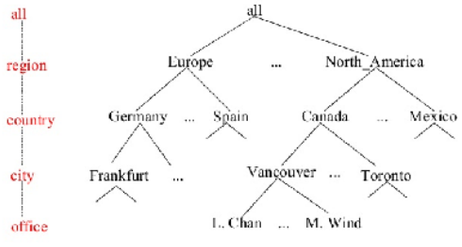












$$v = \frac{v - \mu_A}{\sigma_A}$$

Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability

Data cleaning: e.g. missing/noisy values, outliers

Data integration from multiple sources:

- Entity identification problem; Remove redundancies; Detect inconsistencies

Data reduction

- Dimensionality reduction; Numerosity reduction; Data compression

Data transformation and data discretization

- Normalization; Concept hierarchy generation