Managing Big Data

Introduction

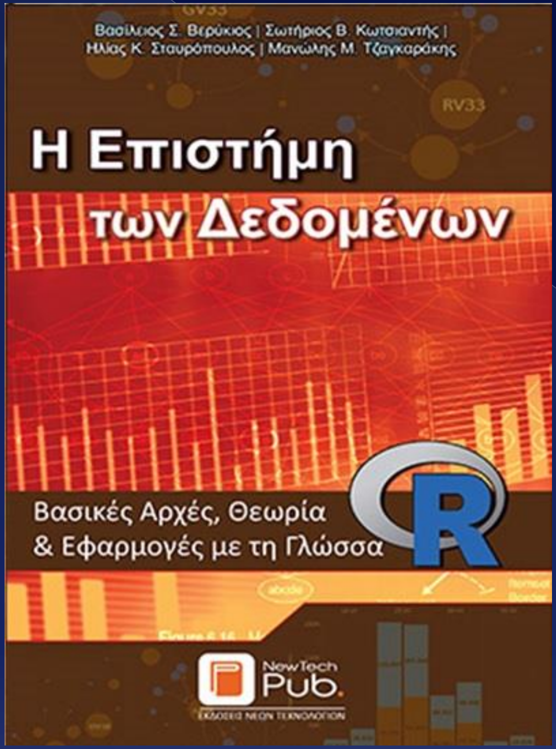
Manolis Tzagarakis
Assistant Professor
Department of Economics
University of Patras

tzagara@upatras.gr
2610 962588
blogs.upatras.gr/tzagara
github.com/deusExMac
google:tzagara
Facebook: tzagara
SkypeID: tzagara
QuakeLive: DeusEx
CoD: CoDFather

About

- **When:** Thursday, 17:00 – 20:30
  - Saturday WHENEVER ANNOUNCED.
- **Where:** PAM2
- **Who:** Manolis Tzagarakis
- **PhD student:** Andreas Retouniotis
- **Contact hours:** Tuesdays 10-12, 15-17 or after appointment (drop email or call)
- **Contact Info:** tzagara@upatras.gr
  Facebook: tzagara
  SkypeID: tzagara
  Steam: xmachina1
  QL: DeusEx
  CoD: CoDfather
  Tel: 2610 962588
  www: blogs.upatras.gr/tzagara
  GitHub: github.com/deusExMac
- **Office:** 2.22

About – Course book

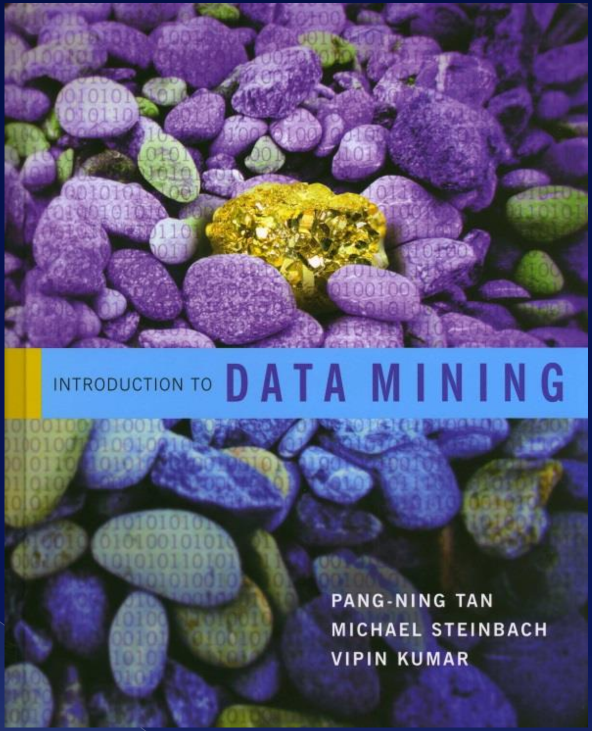

Βασίλειος Σ. Βερύκιος | Σωτήριος Β. Κωτσιαντής |
Ηλίας Κ. Σταυρόπουλος | Μανώλης Μ. Τζαγκαράκης
Η Επιστήμη
των Δεδομένων
Βασικές Αρχές, Θεωρία
& Εφαρμογές με τη Γλώσσα
R
NewTech Pub.
ΕΚΔΟΣΕΙΣ ΝΕΩΝ ΤΕΧΝΟΛΟΓΙΩΝ

Βερύκιος, Σταυρόπουλος, Κοτσιαντής, Τζαγκαράκης: Η ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ Βασικές Αρχές, Θεωρία, & Εφαρμογές με τη Γλώσσα R,

Εκδόσεις Νέων Τεχνολογιών

ISBN-13: 978-960-578-043-2

2019



Tan, Steinbach, Kumar: Introduction to Data Mining,
Addison-Wesley,
ISBN-13: 978-0321321367
2007

About

- Resources
  - **elcass**
    - https://eclass.upatras.gr/courses/ECON1332/
  - **More resources?**
    - See list at the end of this presentation (Appendices A and B)
  - **More?**
    - Ask us
  - **Even more resources? Don’t like our answers?**
    - Google it! Many, many related videos on youtube.com

About

- **Course assessment/grading**
  - (Some) Term projects – 3 or 4 (contributing 30%)
    - Mandantory
    - In teams
    - Implementation in **R**, **Python**
    - Managing codebase with **Git, Github**
  - Online quizzes (10% - not yet decided)
  - Final exam (60% - 70%)

Q?

Course goals

Add to your arsenal **NEW** ways of analyzing data from the field of Machine Learning that

1) Can be applied in Big Data settings, where traditional ones cannot
2) Allow you to see and experience how machine learning algorithms can be used to address questions in economics
3) Work with data other than numbers
4) Get experience in processing bigdata and applying these algorithms using popular languages like **R and Python**
5) Lay down the **foundations** so that you may further your knowledge in using machine learning algorithms in the economics domain.

Big Data

Big Data?

“Big data is a broad term for **data sets so large or complex** that **traditional data processing (analysis) applications and methods are inadequate**.“

- “data processing applications/methods”?
  - *Ways to analyze quantitatively the data.*
- “inadequate methods”?
  - Traditional processing/analysis methods are incapable of (e.g. data does not fit into main memory) or are very, very slow in calculating results.

Big Data

- Four Vs that characterize Big Data
  - Volume: huge amounts/scale of data
  - Variety: different types and sources of data (text, images, videos, streams of unstructured data)
  - Velocity: great pace of data flows
  - Veracity: biases, noise and abnormality in data. Uncertainty of data

Big Data

- Torrents of data everywhere
  - **Uncontrolled human activities in the World Wide Web – the Web 2.0 era**
    - Million of web-pages, blog posts, comments
    - Everyone creates content anywhere!

Facebook
30 billion pieces of information (links, posts, photos etc) every month

Twitter
55 billion tweets every day

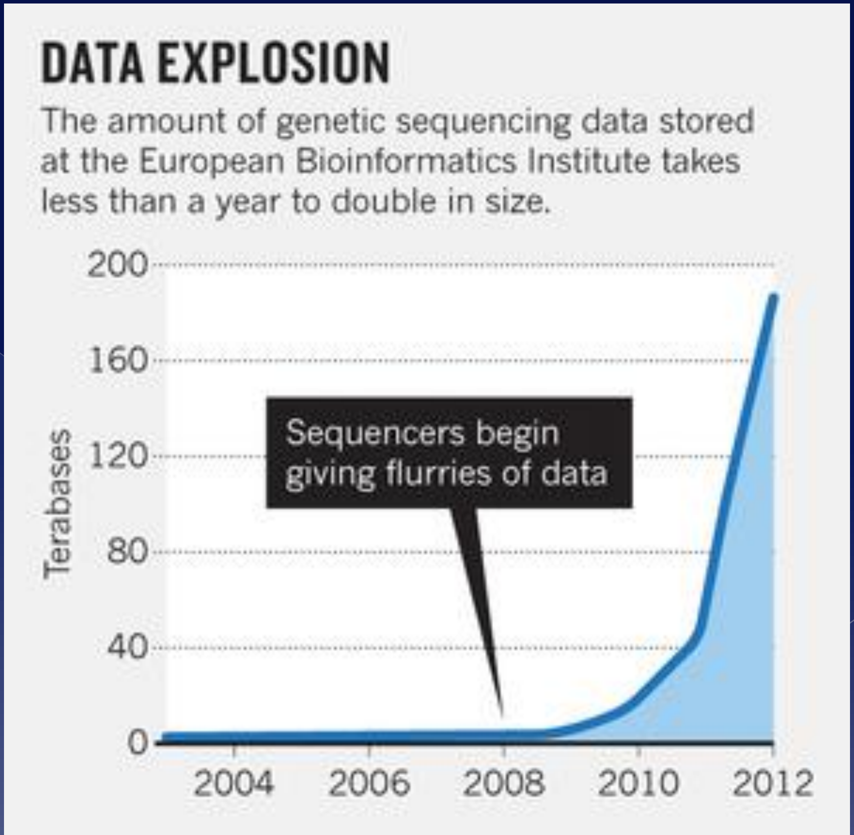
Youtube
35 hours of video uploaded every minute (eq. 176000 Hollywood movies per week)

Big Data

- Torrents of data everywhere (cont.)
  - **Medicine** (electronic patient records – US)
    - 1.6 billion outpatient encounters per year
    - 9 million hospital admissions per year
    - 2 billion text notes per year enriched with lot of information
    - Each day…
      - 420.000 patient encounters in hospitals
      - 2.4 million lab results
      - 553000 pharmacy fills
    - **One paper** added to PubMed **every minute** (2010)

Big Data

- Torrents of data everywhere (cont.)
  - **Biotechnology**
    - 20 petabytes (1 petabyte = 10^{15} bytes) of data about genes, proteins and small molecules at the European Bioinformatics Institute (EBI).
    - 2 Petabytes of genomic data (**doubling every year**) - EBI



DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.

Big Data

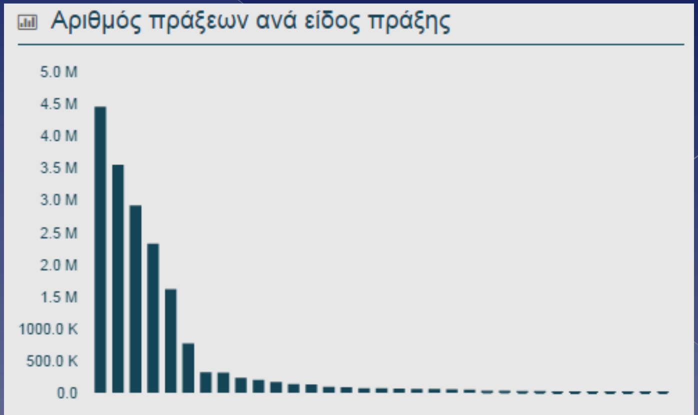
- Torrents of data everywhere (cont.)
  - **Electronic marketplaces – eBay**
    - **30000** product categories
    - **157 million** worldwide active buyers (Q2 2015)
    - **10 million** new items offered every day
    - **1 billion** transactions daily
    - **~$2000** merchandise value traded every second
    - Vehicle changes owner **every 2 minutes**
    - Processing **50TB of data each day**
    - ++ comments, reviews, ratings for each item

Big Data

- Torrents of data everywhere (cont.)
  - **Governmental acts (Transparency program initiative) – diavgeia**
    - Avg. ~10000 publications every day
    - 62372 subscribed active users
    - 4251 institutions publishing acts

Δι@ύγεια

διαφάνεια στο κράτος


Αριθμός πράξεων ανά είδος πράξης

Big Data

- Torrents of data everywhere (cont.)
  - **In many (many) more fields around us every day**
    - Cameras (e.g. traffic)
    - Sensors (e.g. cars, airplanes etc)
    - RFID (use of electromagnetic fields to transfer data, automatically identifying and tracking tags attached to objects) – **Internet of Things**
    - Logs (e.g. bank transactions)
    - Geolocation (identification of the real-world geographic location of an object)
    - GPS (e.g. data related where you are – any time)
    - …

Big Data

- In general, today **huge amounts of data** are not only produced by nuclear reactors or the Large Hadron Collider (LHC) at CERN with high tech sensors…
- **…but are produced in almost any human activity (and you are part of it).**
- Availability of these huge amounts of data helps in **gaining insights on assumptions, models and processes**

Motivation

Motivation?

- **Improving Decision making**
  - In retail, banking and electronic marketplaces, collected data from sales (e.g. bar code) can provide insights and improve
    - Services
    - Addressing of customer needs (CRM)
  - The idea: **knowledge and useful information** related to the improvement of services and customer needs **lurks in such kind of data!**

Motivation?

- **Supporting and facilitating research**
  - More data to assess existing models and shape new theories in
    - Economics
    - Medicine
    - Genome research
    - Environmental studies (e.g. global warming)
    - …

Perspectives on Big Data

Perspectives on Big Data

- You can view Big Data from many different viewpoints and the concerns that it raises

Perspectives on Big Data

- **Data storage and archiving**
  - Where and how to store the data?
  - Traditional data storage technologies (e.g. Relational DBMSs) can't handle Big Data
    - Good for millions of rows, not billions of rows ☹
  - For Big Data: use of massive distributed storage e.g. Google's BigTable, HDFS, Apache HBase

Perspectives on Big Data

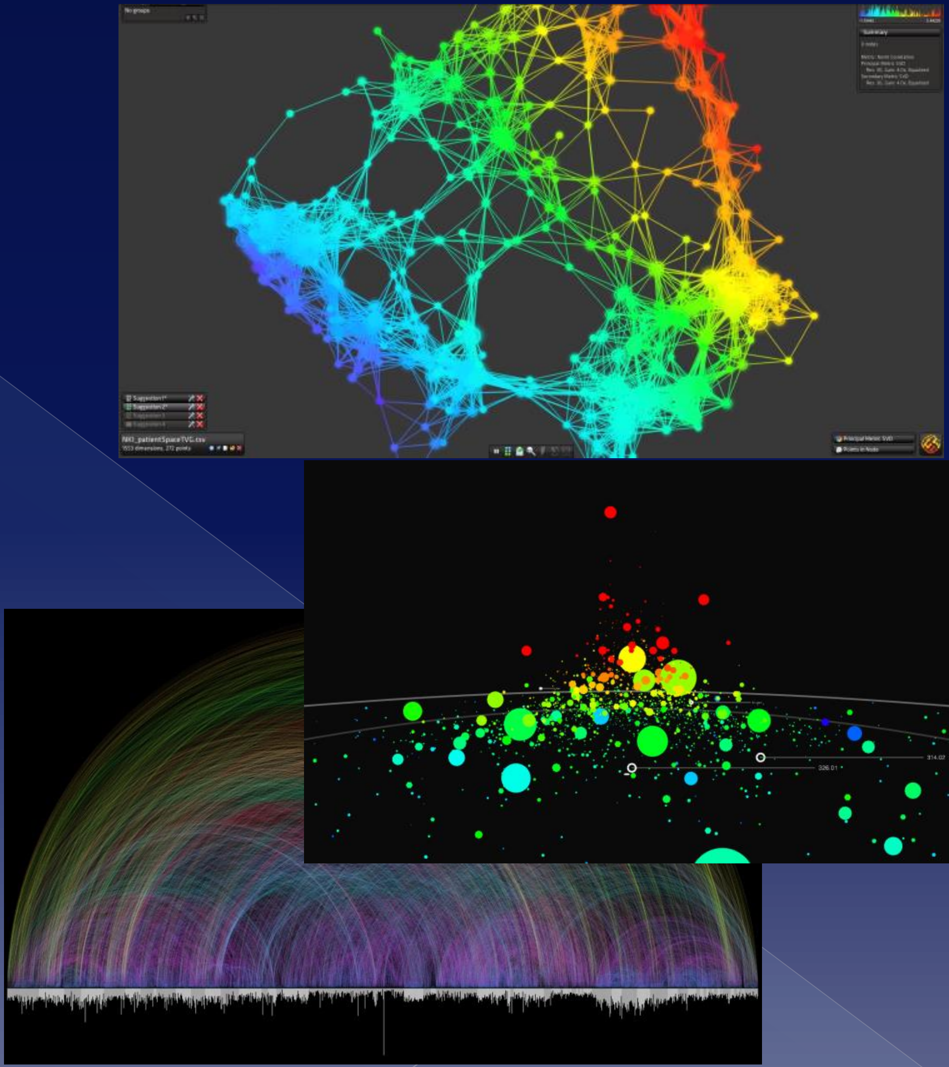
- **Data preparation (or data preprocessing)**
  - manipulation of data into a form suitable for further analysis and processing
  - Huge amount of data from different sources and in different formats.
  - Essential step for data processing!

Perspectives on Big Data

- **Real-time event and stream processing**
  - Processing data as it arrives - without delay- in order to get insights on demand

Perspectives on Big Data

- **Visualizing Data**
  - Clearly communicating information in data
  - Facilitates analysis and reasoning of data and evidence
  - Make complex data more accessible, understandable and usable



Perspectives on Big Data

- **Discovering useful information/Learning from data**
  - Discovering patterns, identifying relationships and drawing inferences that lurk in the data and that are useful (i.e. actionable)
    - (aka) **Machine Learning**
  - **Note: Our focus in this course!**

Machine Learning

Machine Learning

- What is Machine Learning?
  - **Computational process (i.e algorthims executed by machines i.e. computers) that is** capable of improving automatically itself or a model through experience and the use of data.
  - Part of Knowledge Discovery process
  - In general, aims at **identifying interesting, useful information and patterns** that are hidden, lurk in the data
    - Such Information and **patterns not necessarily known beforehand (unknown information/patterns)**

Machine Learning

- What is Machine Learning?
  - Identify and build models **FROM DATA** in terms of…
    - **Associations** (e.g. butter, bread => milk)
    - **Classification and clustering** (e.g. building (predefined or not) groups of things that share common properties)
    - **Series** (e.g. time series and events related to financial markets)
  - "**Interesting information**" ?
    - Semantics of information, trusted and supported, unexpected but **useful in the decision making process**.

Machine Learning

- Why Machine Learning?
  - There is **one immense problem** when dealing with the **processing of huge amounts of data** and trying to find relationships and/or patterns

The limitations of the human brain:

Very GOOD at identifying a dog, lion and run.

Very GOOD at ducking when something is thrown at you

Very BAD at looking at huge amount of data and extract patterns

Very BAD at solving equations, integrals etc.

Machine Learning

- Don’t you get **angry** about the fact that you can easily recognize and solve this problem without thinking…

© Charles Hotham

Machine Learning

- …but not these problems?

$$\mathbf{a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right) = ???}$$

$$\mathbf{\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = ???}$$

$$\mathbf{\frac{0.78912}{0.912289} = ???}$$

$$\mathbf{1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = ???, -\infty < x < \infty}$$

"Let a and b be positive integers and $k = \frac{a^2+b^2}{1+ab}$. Show that if k is an integer then k is a perfect square."

Machine Learning

- But humans are **very good at augmenting** their biological capabilities if they are not up to circumstances! Like…

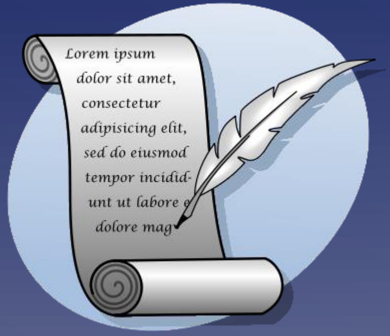
Improved skin

Improved ears

Machine Learning

- But humans are **very good at augmenting** their biological capabilities if they are not up to circumstances! Like…

Improved arm


Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incidid- unt ut labore e dolore mag

Improved memory

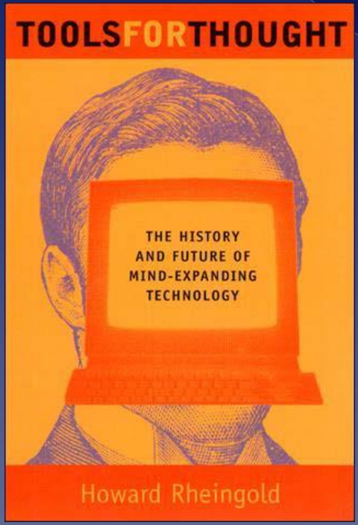
Machine Learning

- But humans are **very good at augmenting** their biological capabilities if they are not up to circumstances! Like…

Improved cognitive capabilities and overcome the processing limitations of the human brain

Machine Learning

- Consider the **computer an extension and augmentation of your brain**, to overcome its limitations
  - Data mining helps in **augmenting your brain** with respect to its **pattern/relationship finding capabilities** allowing humans to see the world differently



Machine Learning

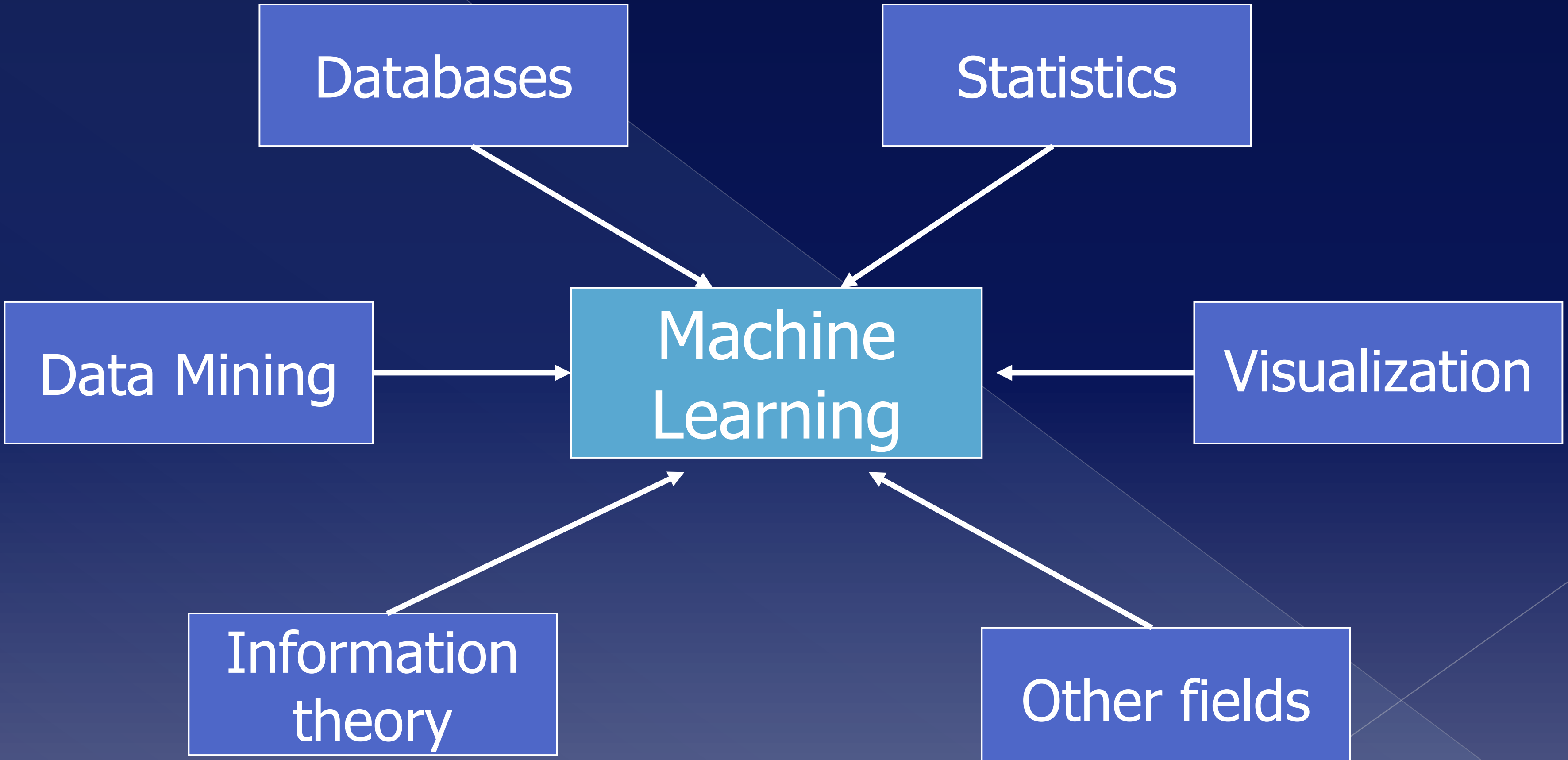
- **Yep,** we are already on our way to becoming **cyborgs** (first, draft prototypes)
  - Not as cool looking as Iron Man though, but still.

Machine Learning

- "A rose by any other name would smell as sweet"
  - Machine Learning important in many contexts
    - Knowledge Discovery
    - Business Intelligence
    - …
- Closely related to but distinct from Data Mining

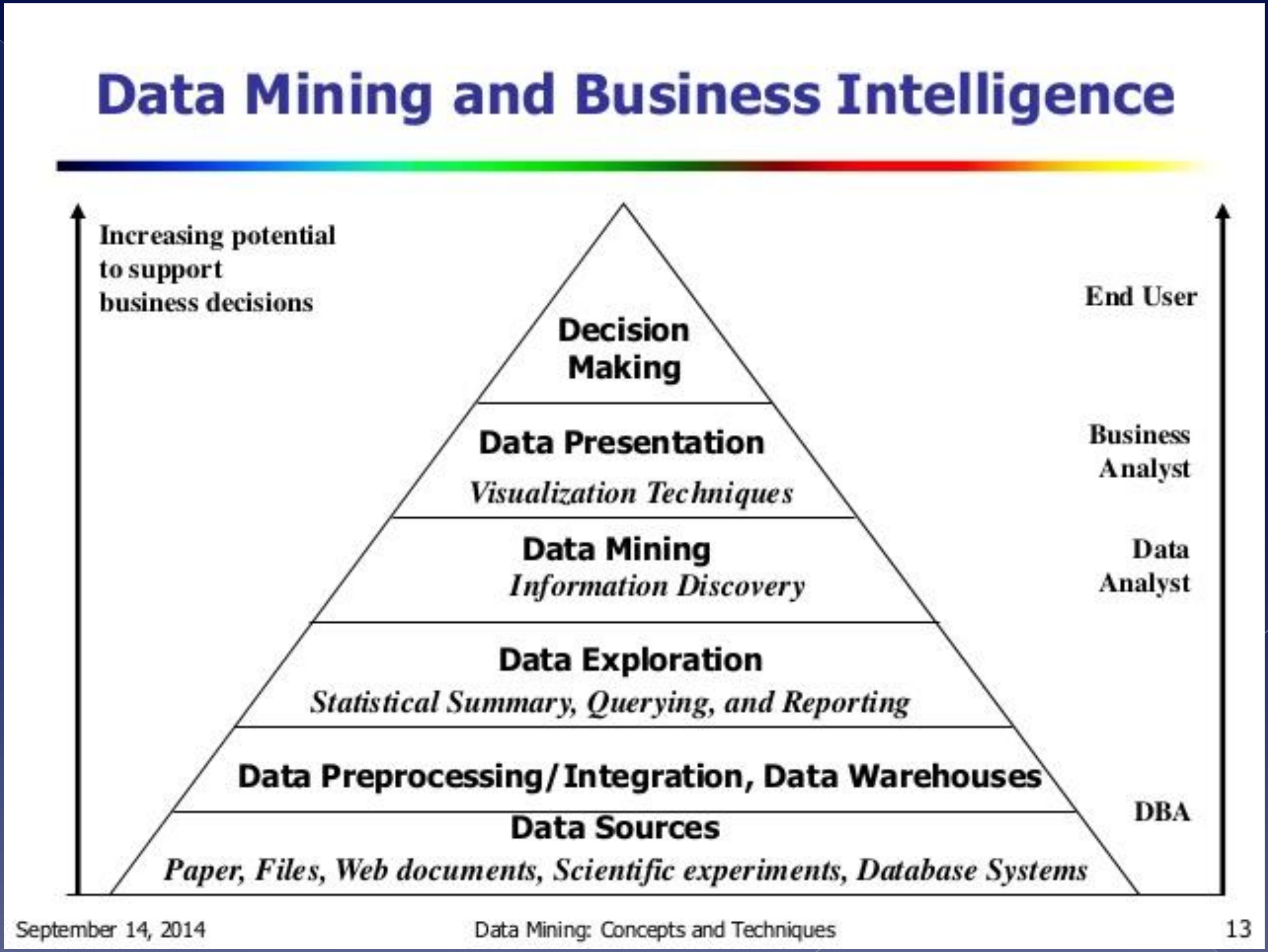
Machine Learning

- Contributions from many different areas



Machine Learning

- Position of data mining in the context of Business Intelligence



Data Mining and Business Intelligence

Increasing potential to support business decisions

Decision Making — End User

Data Presentation *Visualization Techniques* — Business Analyst

Data Mining *Information Discovery* — Data Analyst

Data Exploration *Statistical Summary, Querying, and Reporting*

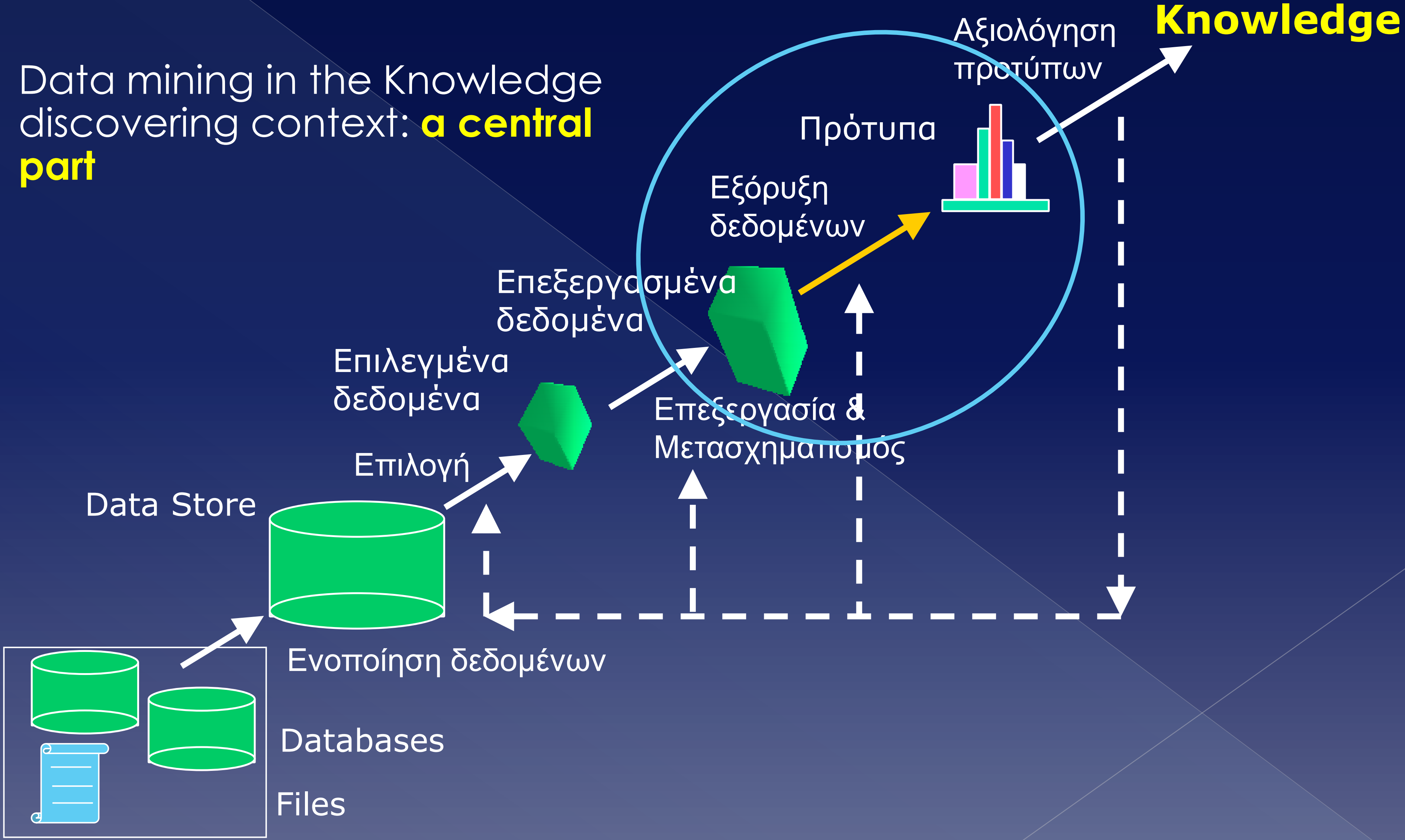
Data Preprocessing/Integration, Data Warehouses — DBA

Data Sources *Paper, Files, Web documents, Scientific experiments, Database Systems*

September 14, 2014 Data Mining: Concepts and Techniques 13

Machine Learning

Data mining in the Knowledge discovering context: **a central part**



Machine Learning in use today

Where is Machine Learning used today?

- **Market analysis**
  - Finding target groups for products based on income, frequent buys etc
  - Discovering consumer patters in relation to time
  - Cross-market analysis e.g. associate/correlate product consumption with forecasts
  - Consumer profiling e.g. products customer buy
  - Customer needs e.g. determining best product for different customers

Where is Machine Learning used today?

- **Risk assessment**
  - Economic planning
    - Analysis and forecasting cash flows
    - Analysis of time series, cross-sectional (different subjects same point in time), to identify trends
  - Competition
    - Assess competitors and market trends
    - Grouping/clustering of customers and determine price of products for each group
    - Pricing strategy in very competitive markets

Where is Machine Learning used today?

- **Financial fraud**
  - Health and car insurance, e.g. locating groups of people that deliberately cause accidents to claim insurance, groups of “professional” patients
  - Credit cards e.g. determine, based on previous consumer behavior, whether card has been stolen or not
  - Money laundering e.g. by locating suspicious transactions

Where is Machine Learning used today?

- **Medicine**
  - Mapping human genome e.g. associate genes with illnesses
  - Causal relationships e.g. to find pathological or environmental causes of illnesses
  - Assessment of treatments/therapies

Where is Machine Learning used today?

- **And many many more**
  - **Astronomy**
    - Discovering type of celestial body (planet?, star?, quasar?, black hole?). Using data mining successful discovery of 22 quasars by JPL and Palomar Observatory
  - **Sports**
    - Improve tactics based on statistics e.g. New York Nicks analyzing data (shots blocked, assists, fouls etc) to get comparative advantage over Miami Heat
  - **Improve the design of Websites**
    - Data mining on logs (i.e. which pages the users visited) to discover customer preferences and behavior and improve the design of the Website
  - **Biology**
    - Classify animals
    - Finding nests of birds

In this course

- You will investigate how **Machine Learning** is used in the field of **economics**
- **Overall aim**
  - Improve/extend your toolset analyzing data with algorithms from the field of Machine Learning that work in big data settings.

A quick look at the contents of this course

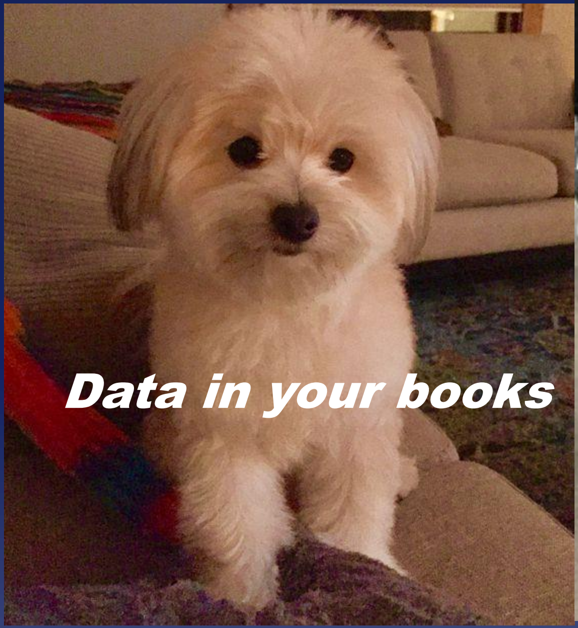
Assumptions

always, always keep in mind two things:

1. **(data) size matters**
2. **problem MUST be solved by a machine (i.e. computers)**

Assumptions

- Data not always in well formatted



A quick look at Machine Learning

- **Preprocessing**
  - Principal Component Analysis (PCA)
  - Distance and Similarity measures
- **Classification**
  - k-Nearest Neighbor (k-NN)
  - Decision Trees
  - Regression
  - Naïve Bayes
- **Regression**
  - Gradient Descent
- **Clustering**
  - K-Means
  - BIRCH
  - PAM
  - RICK
- **Mining association rules**
  - Apriori
  - FP-Growth

Classification

Classification

- Data classification
  - **Goal:** examining characteristics of a data item and deciding in which **predefined class/category it belongs** (takes the form of predicting labels i.e. values for specific attributes).
  - Items are usually records in a database
  - **Important:** The idea is to add records of information in predefined categories
  - Methods (algorithms) to achieve classification:
    - Decision trees
    - k Nearest Neighbor (k-NN)
    - Regression
    - Naïve Bayes

Yep, you probably know this. It's basically a classification method where the dependent var is continuous.

Classification

- Typical application domains
  - **Credit approval** (e.g. should customer x with the feature set z get loan?)
  - **Target marketing** (e.g. do customer x with features z belong in our target group?)
  - **Medical diagnosis** (e.g. are symptoms x consistent with disease y?)
  - **Treatment effectiveness analysis** (e.g. is patient x who takes medicine y healthy?)

Classification

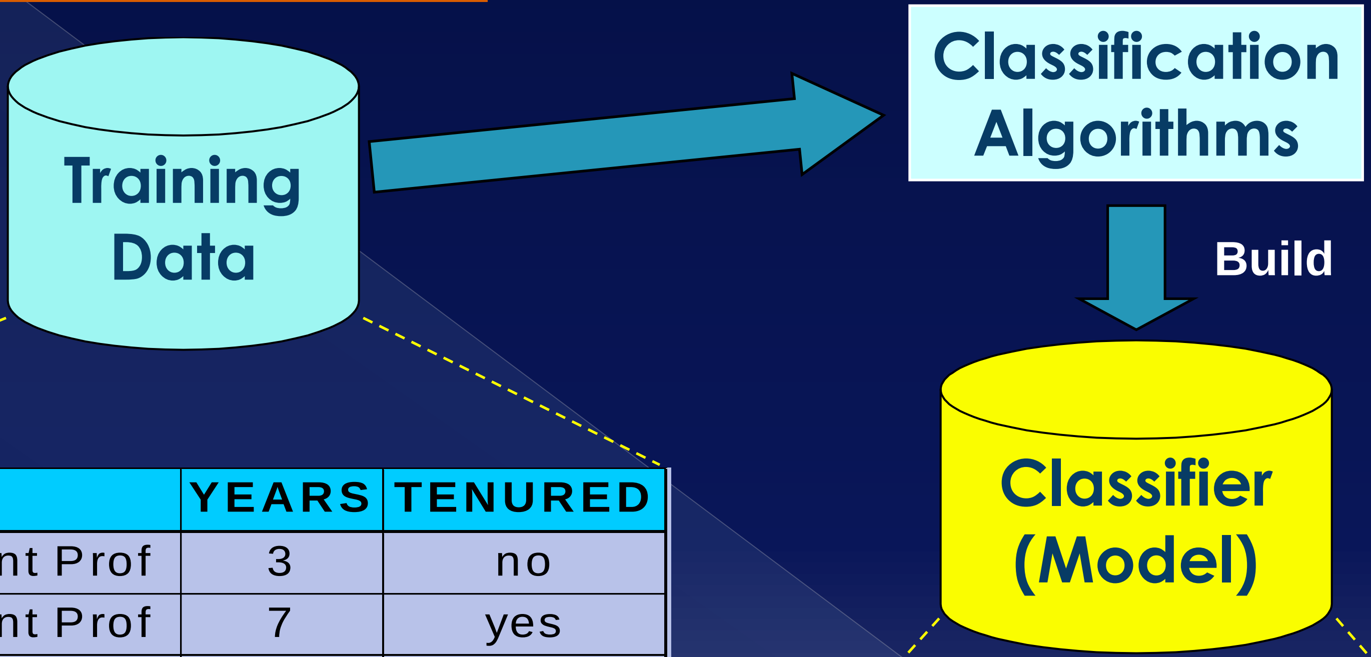
- Classification, a 2 step process
  - **Step 1: Building the classifier (or the classification model)**
    - Get a data set (records) where the label has been already defined for each record and is correct (**supervised learning**)
    - The set of data/records to build the model from is called the **training set**
    - The model is represented as **classification rules, Decision Tree or mathematical equation**
    - Model also known as the **classifier**.

Classification

- Classification, a 2 step process
  - **Step 2: Once the model has been built, use it to classify unknown records (i.e. records for which we don’t know the class they belong to)**
    - Assessing the validity and effectiveness of the model with a test set
      - The known label of the test set is compared with the output of the model
      - Metric: measuring the pct of the correctly classified test data
      - **Test data ≠ training set** . Otherwise **over-fitting** may occur (and this is bad ☹)

Classification: overview

STEP 1: Building the model



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

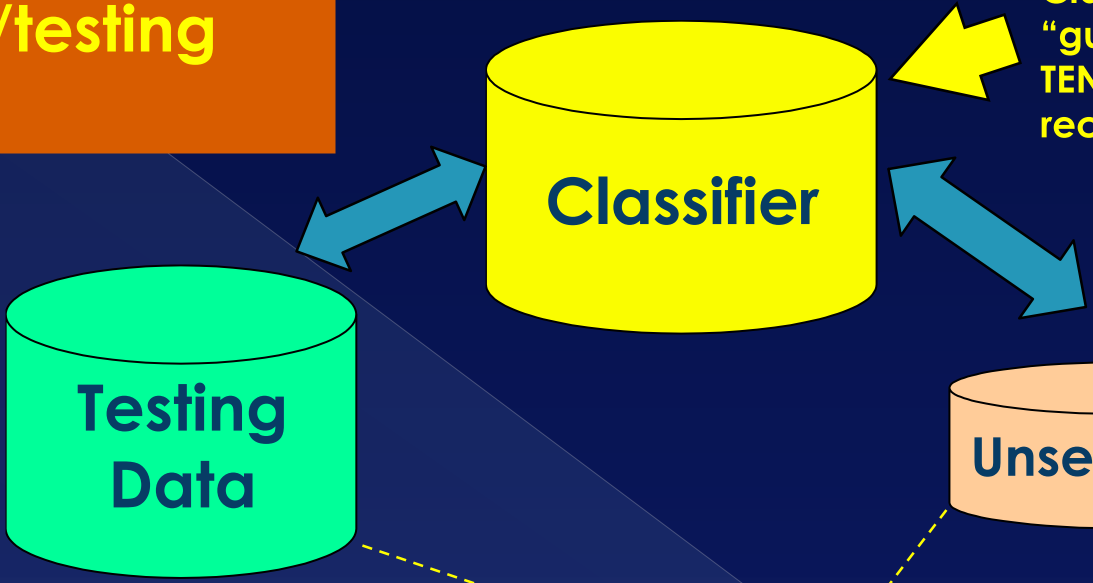
Value “no”, “yes” of TENURED classifies each record. Predefined.

Goal: find rules to assign value for TENURED!

IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Classification: Overview

STEP 2: Using/testing the model



Classifier: Tries to “guess” value for TENURED for each record

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

(Jeff, Professor, 4, ???)

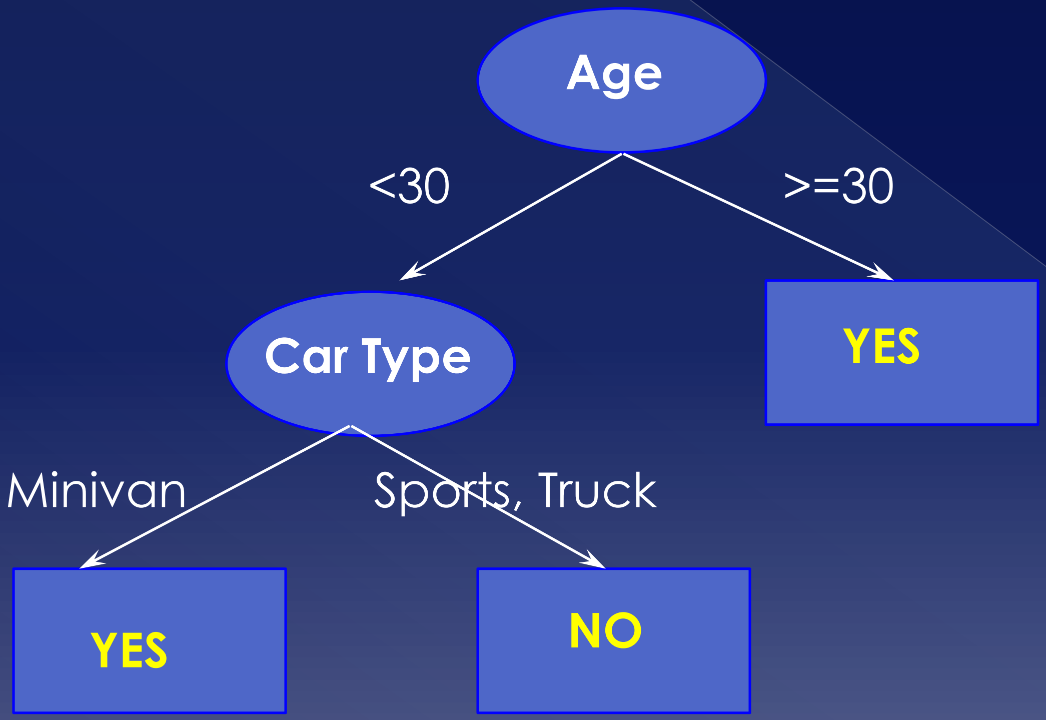
Tenured?

Yes

(Jeff, Professor, 4, Yes)

Classification

- Hey pal, what does a classifier look like?
  - Example: Decision Tree

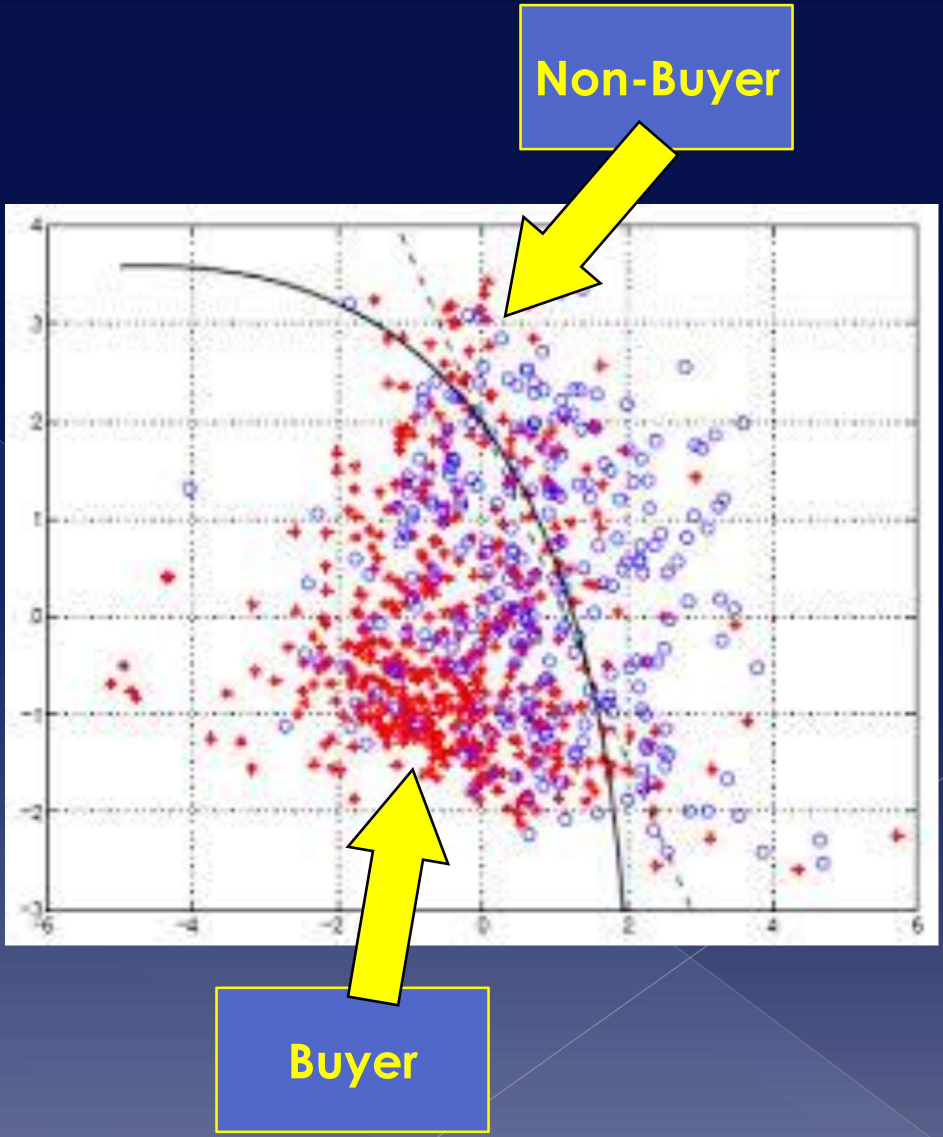


- A *decision tree* T encodes d (a classifier or regression function) in form of a tree.
- A node t in T without children is called a *leaf node*. Otherwise t is called an *internal node*.
- Each internal node has an associated *splitting predicate*. Most common are binary predicates.
  Example predicates:
  - Age <= 20
  - Profession in {student, teacher}
  - 5000*Age + 3*Salary – 10000 > 0

Decition tree to determine e.g whether people wear a tie or not (category “YES”, “NO”) based on predicates Age and CarType as it resulted from the training set.

Classification

- Example domains
- **Marketing**
  - Data about customers
  - 2 classes/categories {**Buyer, Non-Buyer**}
  - Data from demographics, questionnaires
  - Model/classifier creation using training set
  - Classify unknown customers



Regression

Regression

- What it is
  - Estimate the relationship between variables.
  - **Goal:** Understand how one variable changes if another variable (or variables) changes
  - Involves **variables with continuous values** (called the dependent) and a number of other variables of any data type (called independent)
  - Usually relationships come in the form of equations that are interpreted statistically (not deterministically).
  - Regression equations are estimated based on a sample data set (training set). Estimation methods
    - Normal equations
    - Gradient Descent (and variations)

Regression

- The most widely used statistical method to investigate relationships between variables
- Typical application domains (used almost everywhere)
  - **Predictive analysis :** financial forecasting, predicting future values of a variable e.g. predict number of items a consumer will purchase, predict the number of shoppers passing by a billboard, predict number of insurance claims in a time period, predicting crime rates based on drug usage, human trafficking, killings
  - **Operation efficiency:** e.g. optimize business processes like understanding factors influencing product quality, wait times of callers and number of complaints etc
  - **Supporting decisions:** use regression models as evidence to support decisions
  - **Correcting errors:** e.g. does extending shopping hours increase sales? Business manager may intuitively believe so. Regression analysis may support different claim.
  - **New insights:** Collect data that may give new insights and ideas. Regression can be used to identify relationships between variables or patterns that were previously unnoticed.

Regression

- Goals of a regression model
  - **Predict** the value of the dependent variable based on the values of the independent variables
  - **Explain** the variance of the dependent variable
  - Example regression model:

$$FoodConsumption = \beta_1 Income + \beta_0$$

The model aims to express the fact that the amount of food consumed by a family per year (in Euros) depends on the family's income. It captures the relationship between the variables FoodConsumption (dependent variable) and Income (independent). It assumes that food consumption increases proportionally to Income, and hence is an example of a simple linear regression model. The unknowns are the betas (coefficients) which must be estimated based on the training set. Who comes up with such model? Economists studying the literature.

Regression

- How to estimate the coefficients (betas)?
  - Different ways possible, depending on the type of the regression model and the data
  - For **linear regression models** (the most commonly used) methods to estimate the parameters are:
    - **OLS** or **The normal equation** (you know/have seen this!)
    - **Gradient Descent** and its variations

Regression

- **Normal equation** (the default)
  - Closed form
  - Very nice method, but only for small datasets and few variables
  - Works only if the entire training set is available
- **Gradient Descent (and variations)**
  - Not a closed form. Iterative method for estimating coefficients.
  - Suitable for big data sets with huge number of variables in the regression model
  - Many variations which speed-up estimation process
  - Variations work for cases where training set is not yet available (online algorithm)

Clustering

Clustering

- Clustering
  - **Goal:** the process of partitioning a heterogeneous set of data in a set of clusters (**not predefined!**)
  - **Important:** In contrast to classification, in clustering there are **no predefined categories/classes/clusters**
  - Data is **partitioned in clusters** based on their **similarity**. Assigning semantics to each cluster is the job of the analyst (i.e. human)
  - Methods (algorithms) for clustering:
    - K-Means
    - PAM
    - BIRCH
    - RICK
    - CURE

Clustering

- Different types of clustering
  - Partitioning clustering
  - Hierarchical clustering (i.e. create clusters of clusters etc)
  - Fuzzy clustering
  - Crisp clustering
  - Kohonen Net clustering
  - Density-based clustering
  - Grid-based clustering
  - Subspace clustering

Clustering

- Assumptions about data
  - Let x be a data item (record)
  - x is considered a vector of d metrics:

$$x = (x_1, x_2, x_3, \dots, x_d)$$

where x_i **is the ith feature of the data item** and **d the dimension** of the data item or the space created by the data items.
Clustering attempts to group such data items together (in clusters), based on their similarity.

Clustering

- Many different metrics for similarity (note: in dm usually "**distance**" and "**dissimilarity**" synonyms)
  - E.g. Minkowski distance

$$d(x, y) = \sqrt[r]{\sum_{i=1}^{n} |x_i - y_i|^r}$$

A generalization of the Euclidean distance

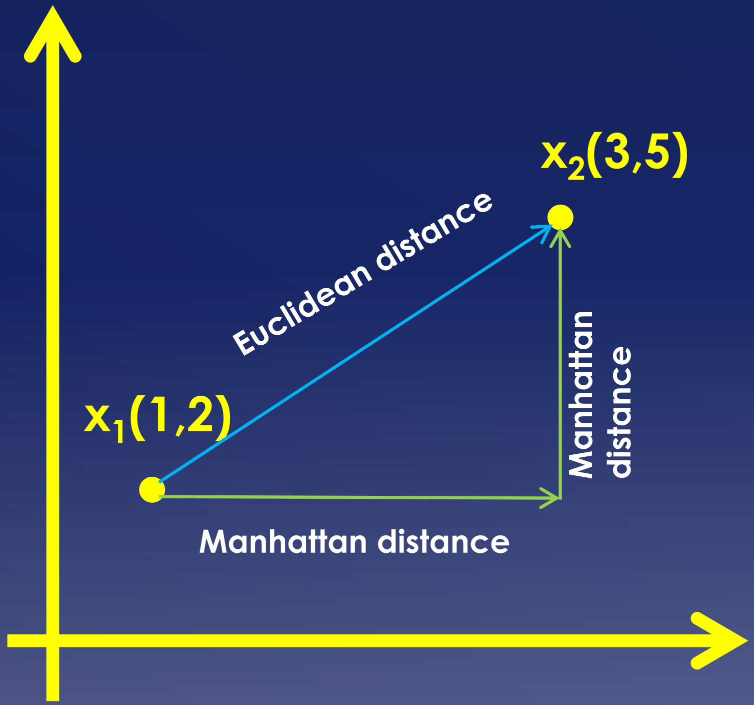
r =1, Manhattan distance

r = 2, Euclidean distance

r = ∞, Maximum distance between features (L_{max} *or* L_∞ norm)

Clustering

- Euclidean vs Manhattan distance



Euclidean distance (x_1, x_2)
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance (x_1, x_2)
$= 2 + 3 = 5$

Clustering

- Cosine distance (similarity)

$$\boldsymbol{cos(x, y) = \frac{x \cdot y}{\|x\| \; \|y\|}}$$

where:

x, y vectors

$$x \cdot y = \sum_{k=1}^{n} x_k y_k$$

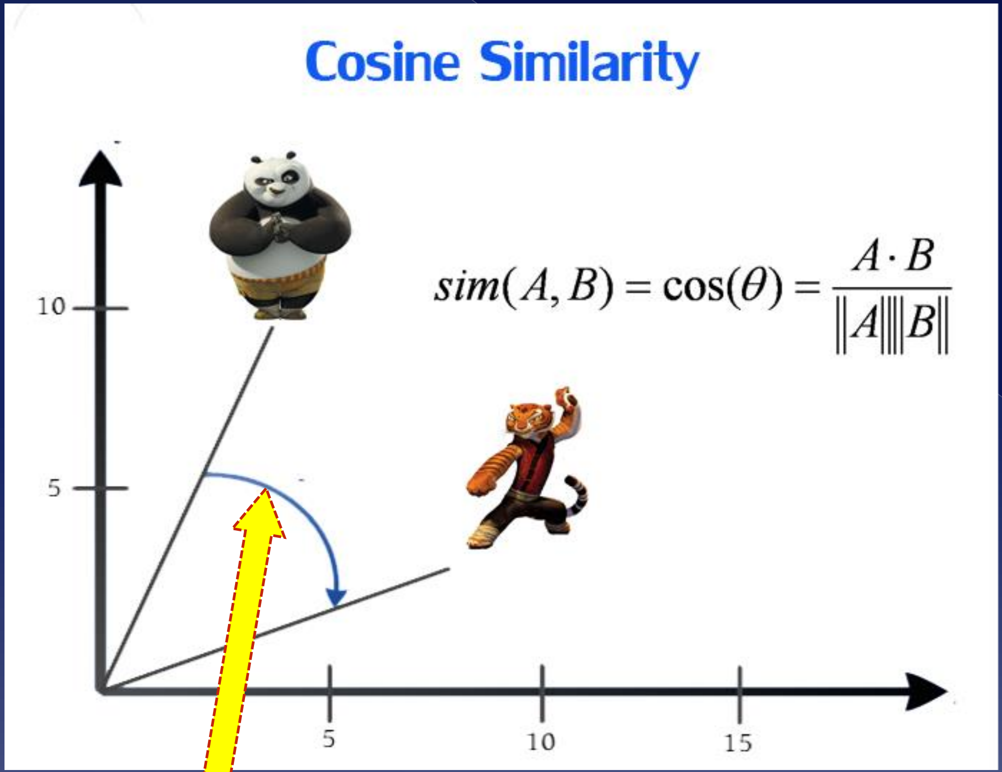
Euclidean dot product / inner product of vectors x, y

$$\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x \cdot x}$$

Length/magnitude/norm of vector x

Clustering

- Cosine distance/similarity

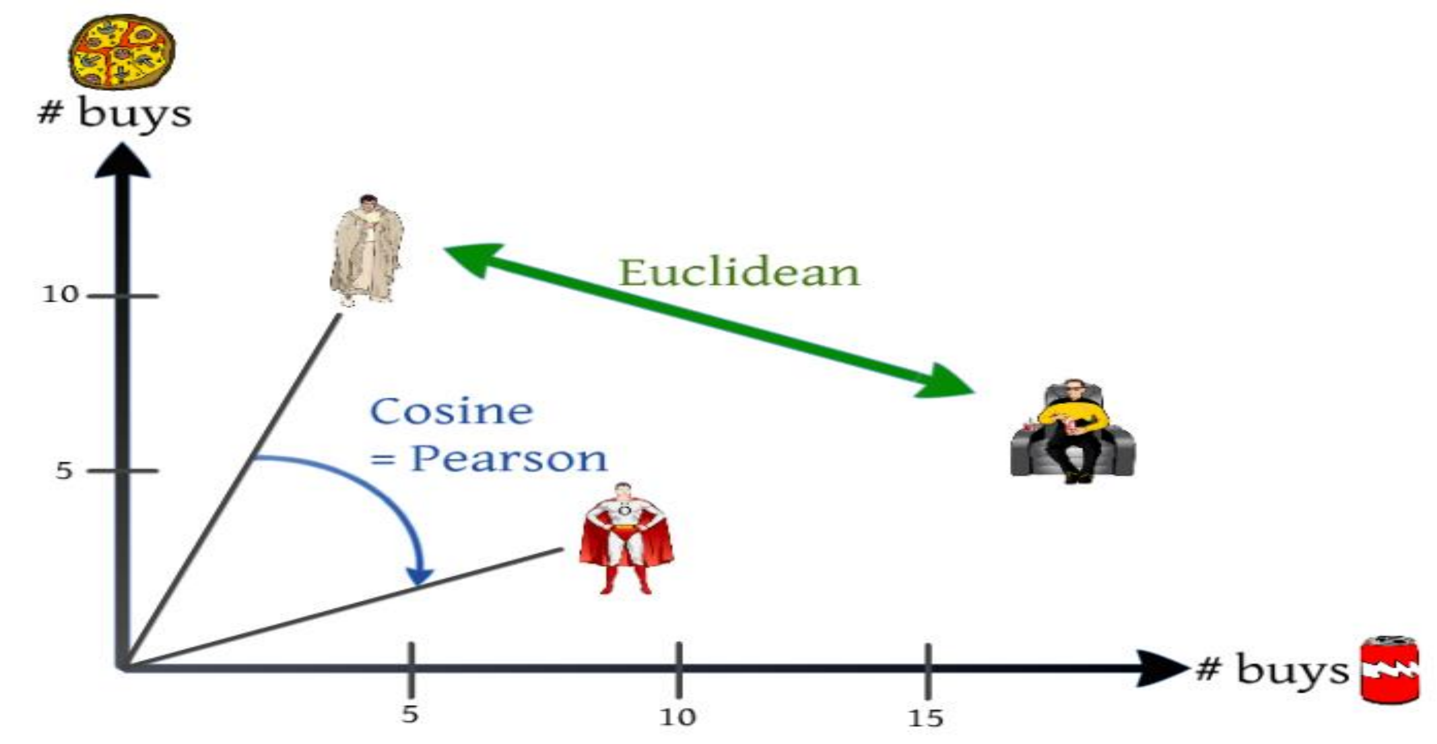


- Cosine similarity **measures the cosine of the angle** between two vectors x, y
- If angle = 0° then this means cosine similarity =1 i.e. greatest similarity score.
- If angle <> 0° then cosine similarity < 1 (at 90° it is 0)
- Opposed vectors: cosine similarity = -1

Cosine similarity is expressed in terms of this angle!

Clustering

- You **already know cosine similarity** (but with a different name)
  - **Pearson correlation coefficient (ρ or R)**
    - **Cosine distance is simply a geometric interpretation of ρ/R**

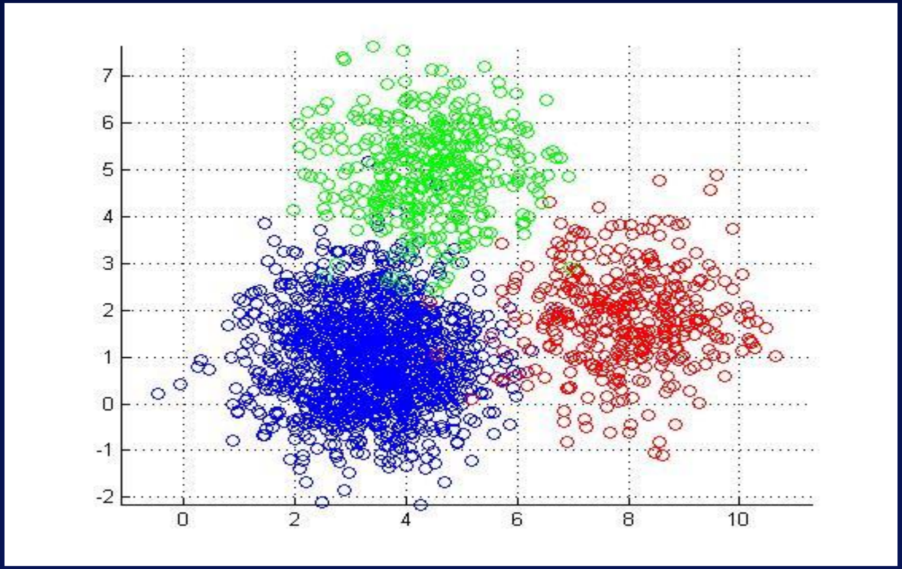


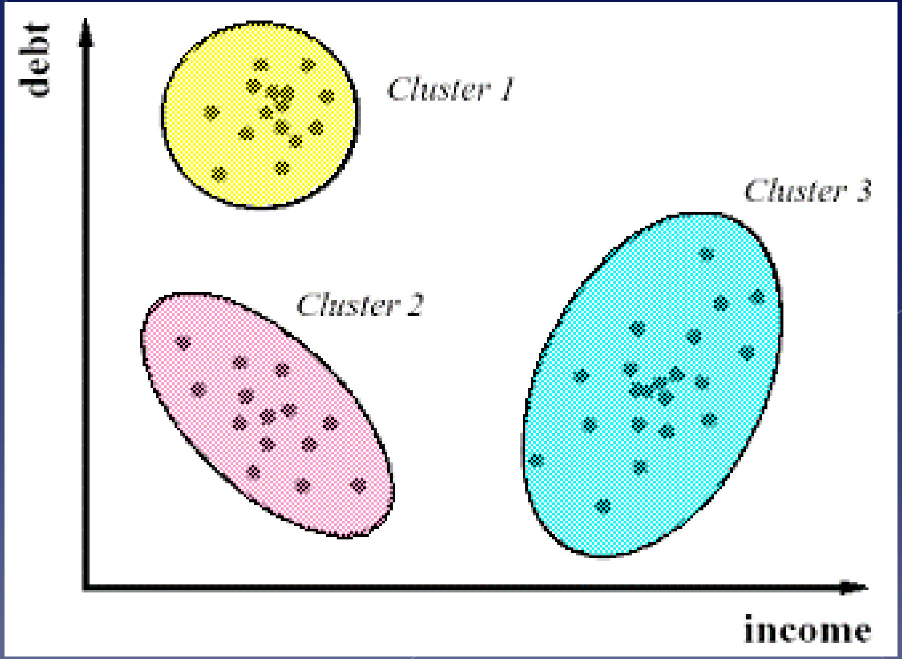
Clustering

- Clustering algorithms can be categorized in different ways
- E.g. based on the certainty with which an item is assigned to a cluster/class
  - **Hard clustering techniques:** assign a class label l_i to a data item x_m which designates unambiguously the class/cluster
  - **Fuzzy clustering techniques:** assign to each data item x a probability of membership to each cluster j .

Clustering

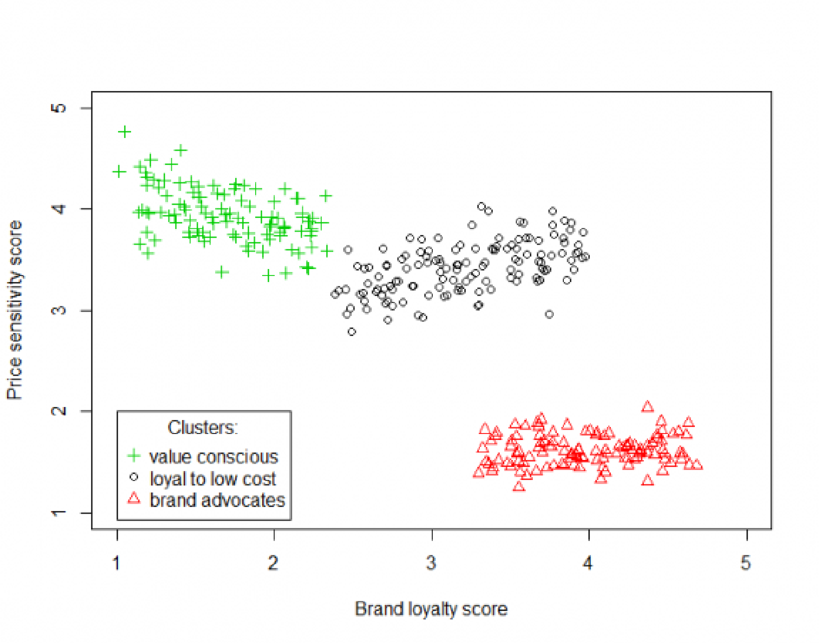
- The general idea of the clustering approach
  - **Minimize** distance of items belonging in the same cluster
  - **Maximize** distance between clusters

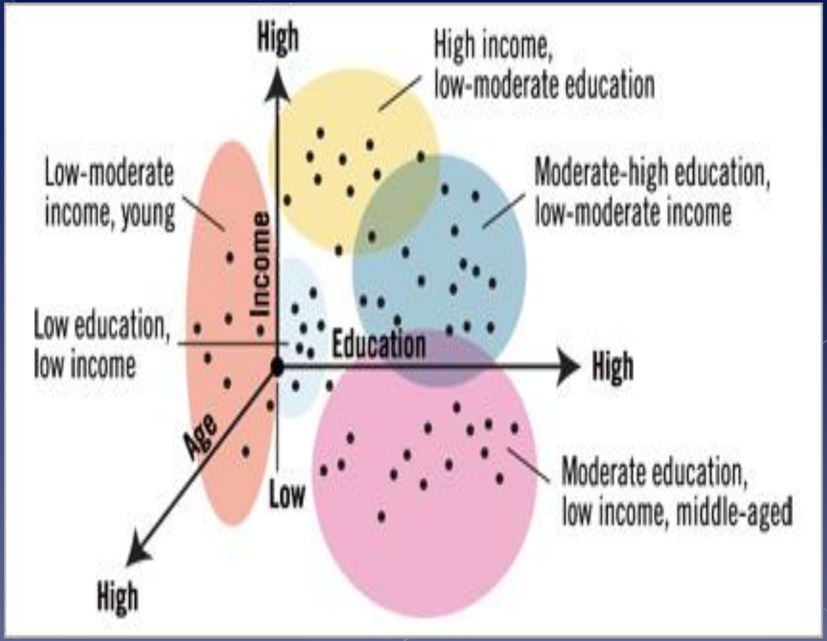




Clustering

- Example domains
- **Market segmentation**
  - **Separate customers into groups** so that they can be targeted differently (i.e. different deals, products etc)
  - Based on geography, demographics etc.

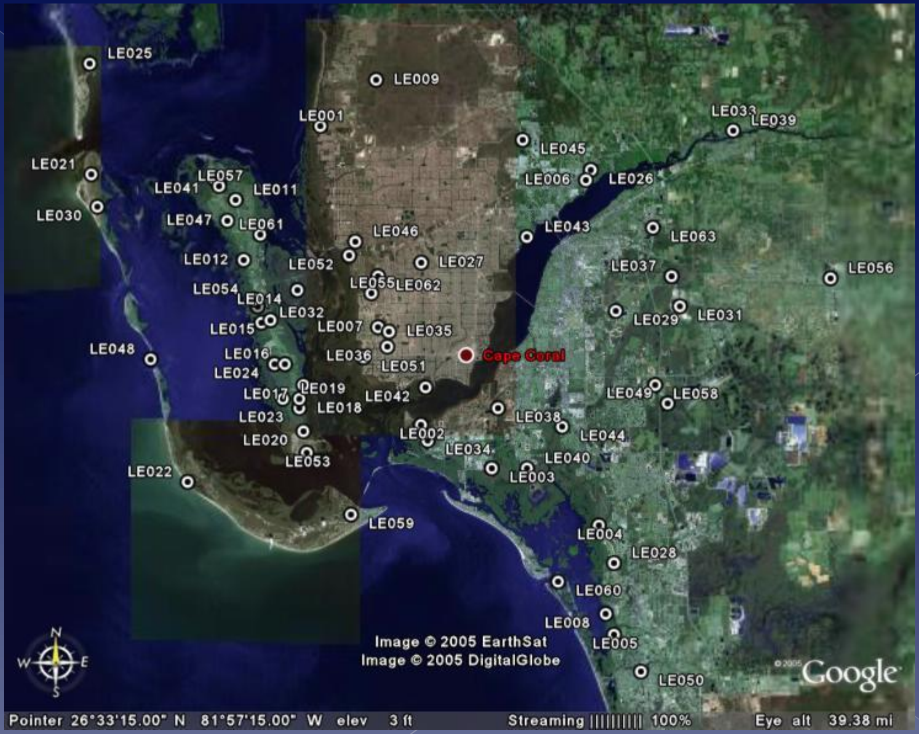



High
Income
Low
Education
High
Age
High
High income, low-moderate education
Low-moderate income, young
Moderate-high education, low-moderate income
Low education, low income
Moderate education, low income, middle-aged

Clustering

- **Ecology**
  - **Finding bird nests**
  - Data
    - Spatial
  - Each cluster of nests assessed based on e.g.
    - Distance from water
    - etc

CNDDB GOEA Element Occurrences
Quads with "unprocessed" CNDDB GOEA data
Desert Renewable Energy Conservation Plan (DRECP)



Mining Association Rules

Mining association rules

- Mining association rules
  - **Goal:** discover hidden associations (called association rules) existing between the features of the data items.
  - Association rules take the form of

    $$\mathbf{A \rightarrow B}$$

    where **A** and **B** are **feature sets** that exist in the data examined
  - One of the **most important processes** in data mining as it provides an easy way to express useful relationships, that are human understandable.

Mining association rules

- Concepts
  - **Set of objects/items:** $I = \{ I_1, I_2, I_3, \ldots, I_m \}$
    - Example: all items in a supermarket {bread, butter, toothpaste, cereal, milk, diapers, beer, vodka,…}
  - **Transactions:** $T = \{ t_1, t_2, t_3, \ldots, \}$, $t_j \subseteq I$
    - Example: each t_i represents what one customer buys e.g. if { bread, milk, butter } **∈** T=> one specific customer bought bread, milk and butter together. Customer's basket
  - **Itemset:** A subset of I with 0 or more items
    - *k-itemset*: itemset with k items in it
    - Example: {milk, diapers} => 2-itemset, {beer, milk, bread} => 3-itemset
  - Say that an transaction t_j contains itemset A, if A is subset of t_j
    - Example: Transaction {beer, milk, diapers} contains 2-itemset {beer, diapers}

Mining association rules

- The overall idea
  - You have many transactions
  - Extract from there association rules like

    {milk, beer} → {diapers}

    meaning whoever buys **milk and beer** also buys (with great prob., anyway) **diapers.**
  - Identifying such relationships based on three metrics
    - **Support of itemset, σ**
    - **Support of association rule, s**
    - **Confidence of association rule, c**

Mining association rules

- Support of itemset, **σ**
  - How frequent (count, pct, prob.) transactions t_i -in the set T- contain itemset X, σ(X) . More formally:

$$\sigma(\mathbf{X}) = |\{t_i \mid X \subseteq t_i, t_i \in T\}|$$

TxID	Transaction
1	{beer, milk, diapers}
2	{vodka, beer, cereal}
3	{beer, appel, knife, milk}
4	{apple, beer, diapers}
5	{shampoo, banana, coffee}
6	{beer}

$\sigma(\{\mathbf{milk, beer}\}) = 2$ OR

$\sigma(\{\mathbf{milk, beer}\}) = {}^{1}/_{3}$

$\sigma(\{\mathbf{beer}\}) = {}^{5}/_{6}$

Mining association rules

- Support of association rule, **s**
  - How frequent a rule of the form **X** → **Y** is observed in all known transactions T

$$\boldsymbol{Support, s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}}$$

Support of itemset resulting from the union of X with Y

where **N** is the total number of transactions and $\mathbf{X} \cap \mathbf{Y} = \emptyset$

Mining association rules

- Confidence of association rule, **c**
  - How frequent the items of itemset Y appears in transaction that also contain itemset X

$$\boldsymbol{Confidence, c}(\mathbf{X} \rightarrow \mathbf{Y}) = \frac{\boldsymbol{\sigma}(\mathbf{X} \cup \mathbf{Y})}{\boldsymbol{\sigma}(\mathbf{X})}$$

Support of itemset resulting from the union of X with Y

Support of itemset X

with $\mathbf{X} \cap \mathbf{Y} = \emptyset$

Mining association rules

- Using support and confidence
  - Items sets with support and confidence above some minimum (minsup, minconf) are called **frequent itemsets**.
  - **Goal:** Find (quickly!) **association rules** that have **above some minimum support (minsup) and above some minimum confidence (minconf)** based on frequent itemsets

Mining association rules

- How difficult can that be finding such association rules?
  - Very difficult because of size of problem space
  - **Problem:** “brute force”/exhaustive algorithms **take a very long, long, long, long time** finding association rules that meet these criteria.
  - E.g. for **d items**, the total number of association rules is $\boldsymbol{R = 3^d - 2^{d+1} + 1}$ i.e. with **6 items** we can come up with a total of **602 association rules (size of problem space)**
  - In today’s supermarkets easily, **d > 50 meaning R > 7178979854400527775085000 association rules must be checked (support, confidence)**

Mining association rules

- **Better algorithms** to find association rules with support and confidence above a minimum
  - E.g. not consider some association rules
  - E.g. reducing problem space
- Existing methods (algorithms)
  - Apriori
  - FP-Growth
- Application domains
  - Supermarkets, predicting consumer behavior
  - Voting, predict what voters will vote based on previous preferences

Mining association rules

- **Supermarkets**
  - Input: transactions – what people buy
  - Output: associations between items in transactions

TxID	Transaction
1	{bread, flower, milk}
2	{beer, bread}
3	{beer, diaper, milk, bread}
4	{beer, bread, diapers, milk}
5	{flower, diapers, milk}

Rules discovered:

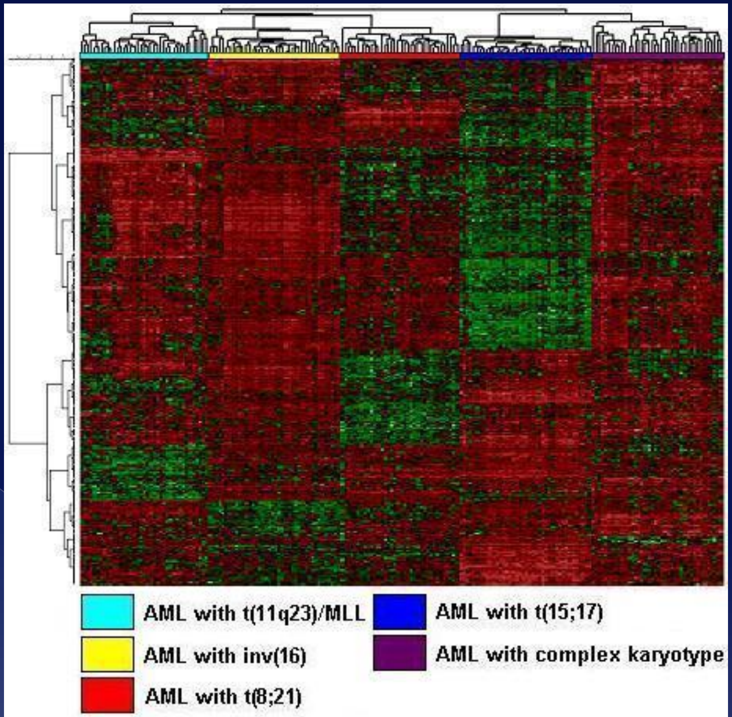
{flower} → **{milk }** , p(milk | flower)=1

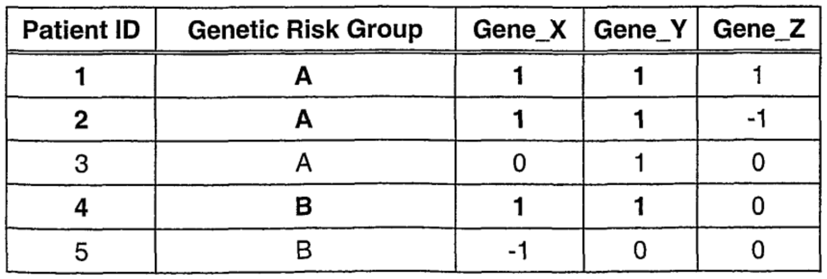
{milk}→ **{flower}** , p(flower | milk)=0.5

{beer, bread}→ **{diaper }** ,
p(diaper | beer, bread)= 0.66

Mining association rules

- **Biology**
  - DNA microarray data
  - Many experiments with many involved genes in each
  - Measuring: < 0 or >0 with respect to two different forms of leukemia (AML, ALL)
  - Genes which coappear => interact
  - Rules: **{desease}=> {gene A ↑ gene B ↓ gene C ↑}**



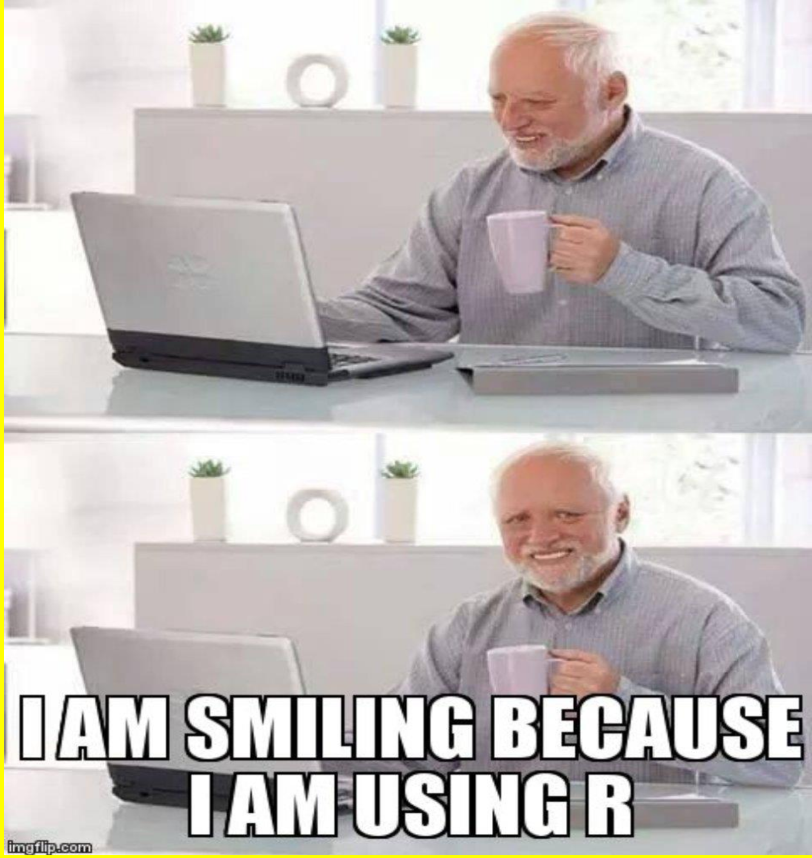


Patient ID	Genetic Risk Group	Gene_X	Gene_Y	Gene_Z
1	A	1	1	1
2	A	1	1	-1
3	A	0	1	0
4	B	1	1	0
5	B	-1	0	0

Tools used in this course

R

When your stats prof starts talking in R commands:

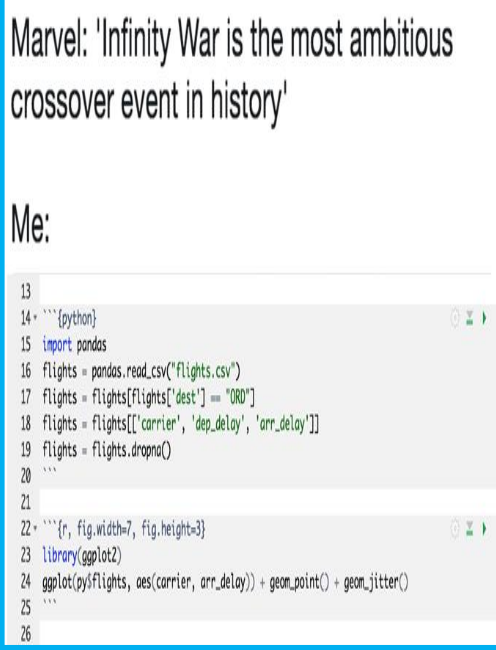


When you can't find an example of R code to steal from stackoverflow

BASE R
R Studio
R Studio + ggplot2
R Studio + sjPlot ggplot2
R Studio + Shiny sjPlot ggplot2 knitr
R Studio + Shiny sjPlot knitr ggplot2
R Studio + The tidyverse
Components

Python

When you replace a for loop with a vectorized numpy function and see the speed improvement



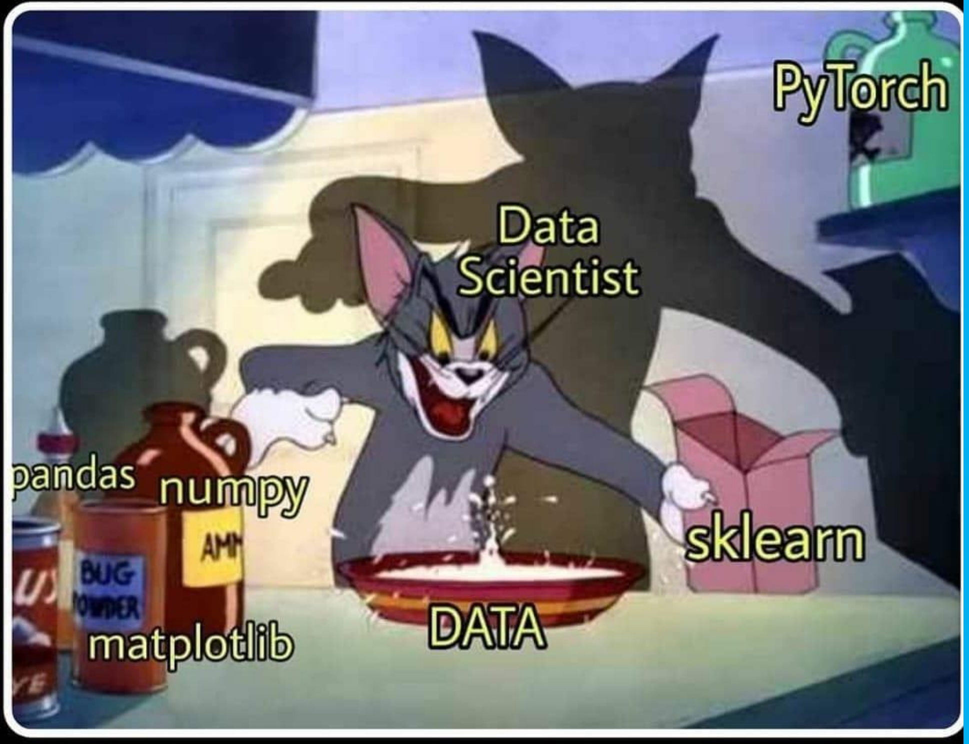
Marvel: 'Infinity War is the most ambitious crossover event in history'

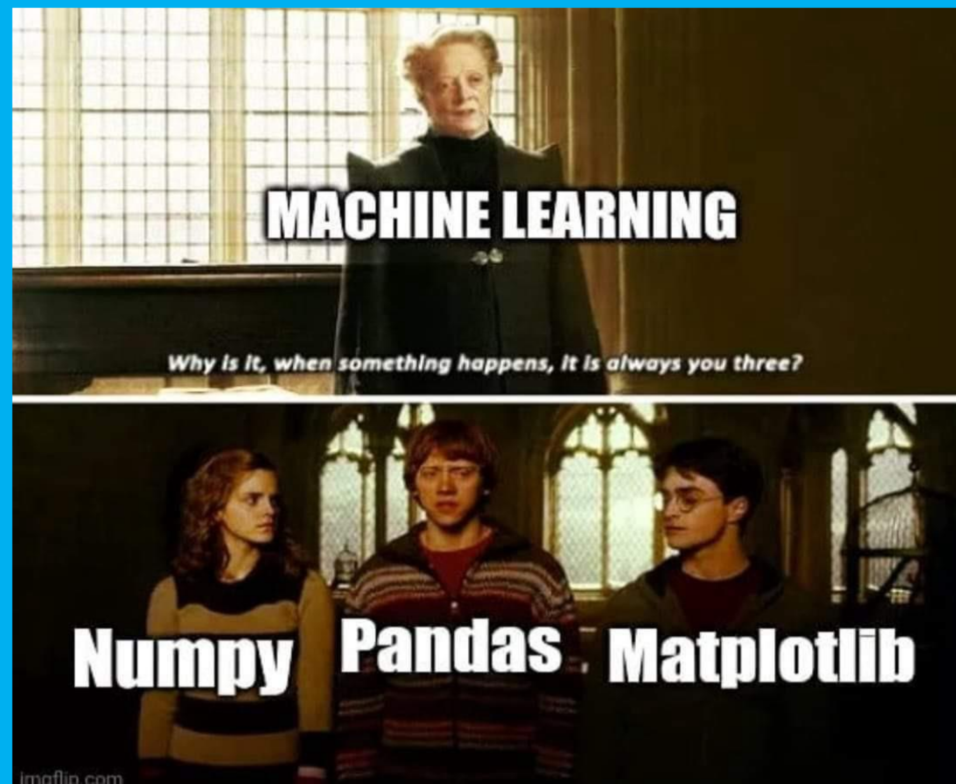
Me:

~~~
```{python}
import pandas
flights = pandas.read_csv("flights.csv")
flights = flights[flights['dest'] == "ORD"]
flights = flights[['carrier', 'dep_delay', 'arr_delay']]
flights = flights.dropna()
~~~

```{r, fig.width=7, fig.height=3}
library(ggplot2)
ggplot(py$flights, aes(carrier, arr_delay)) + geom_point() + geom_jitter()
```

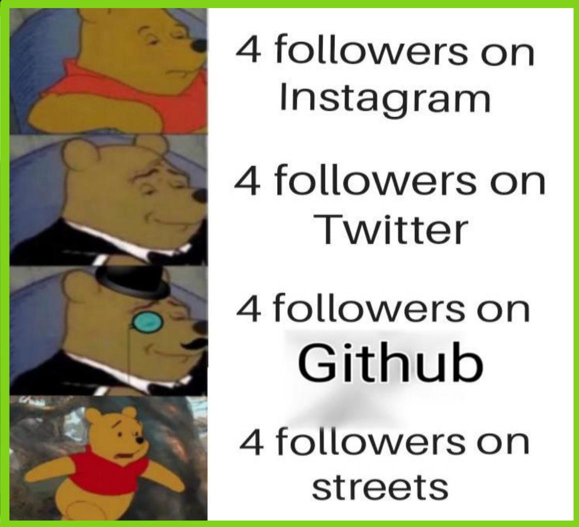
```




```

Git/GitHub

WHAT THEY TELL YOU GITHUB IS
WHAT GITHUB IS
GitHub


4 followers on Instagram
4 followers on Twitter
4 followers on Github
4 followers on streets

USING AIRDROP
USING GOOGLE DRIVE
USING A PRIVATE DISCORD
USING GITHUB

Stackoverflow discussions
@redco
Github discussions

When your GitHub repo gets 1 star
Google
how to handle fame
Google Search
I'm Feeling Lucky

Thnx!

APPENDIX A: Bibliography

- Μ. Βαζιργιάννης, Μ. Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων, *Τυπωθήτω*, Δαρδάνος, 2003.
- Margaret Dunham, Data Mining Introductory and Advanced Topics, 2003, Pearson Education.
- Αλέξανδρος Νανόπουλος, Ιωάννης Μανωλόπουλος, Εισαγωγή στην Εξόρυξη και στις Αποθήκες Δεδομένων
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997.
- I.H. Witten, E. Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, October, 1999.
- Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd Edition, ISBN 1-55860-901-6, 2006.

APPENDIX A: Bibliography

- David J. Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining , MIT Press, Fall 2000.
- S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan-Kaufmann Publishers 2003
- Cathy O'Neil and Rachel Schutt, Doing Data Science: Straight Talk from the Frontline, 1st Edition, ISBN-13: 978-1449358655, 2013
- Nate Silver : The Signal and the Noise: Why So Many Predictions Fail — but Some Don’t, New York: Penguin Press (2013)
- Foster Provost and Tom Fawcett, Data Science for Business: What you need to know about data mining and data-analytic thinking, 1st Edition, ISBN 978-1-4493-6132-7, O'Reilly Media, 2013

APPENDIX B: Conferences/Journals

- **KDD Conferences**
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

- **Other related conferences**
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- **Journals**
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDE (Knowledge and Data Engineering)
  - KDD Explorations
  - ACM Trans. on KDD

APPENDIX C: Online resources

- ***R***
  - **MarinStatsLectures**
    - https://www.youtube.com/c/marinstatlectures/videos?view=0&sort=da&flow=grid
  - **Machine Learning Plus - Learn R by Intensive Practice**
    - https://www.youtube.com/playlist?list=PLFAYD0dt5xCwDNFdrqeNoU9t-nhAWkbKe
  - **Twotorials.com**
    - https://www.youtube.com/playlist?list=PLcgz5kNZFCkzSyBG3H-rUaPHoBXgijHfC

APPENDIX C: Online resources

- **R**
  - **Using the Data Frame in R**
    - https://www.youtube.com/watch?v=9f2g7RN5N0I
  - **R Programming Tutorial - Learn the Basics of Statistical Computing**
    - https://www.youtube.com/watch?v=_V8eKsto3Ug
  - **R Tutorial For Beginners 2022**
    - https://www.youtube.com/watch?v=KlsYCECWEWE

APPENDIX C: Online resources

- **R**
  - **Introduction to R: Plotting in Base R**
    - https://www.youtube.com/watch?v=8HD4riFaqYs
  - **R Tutorials - Learn ggplot2**
    - https://www.youtube.com/playlist?list=PLjgj6kdf_snaBCTJEi53DvRVgOuVbzyku

APPENDIX C: Online resources

- **Python**
  - **Python Programming Fundamentals | Data Analysis with Python**
    - https://www.youtube.com/playlist?list=PLyMom0n-MBrpzC91Uo560S4VbsiLYtCwo
  - **Complete Python Pandas Data Science Tutorial! (Reading CSV/Excel files, Sorting, Filtering, Groupby)**
    - https://www.youtube.com/watch?v=vmEHCJofslg
  - **Economic Data Analysis Project with Python Pandas - Data scraping, cleaning and exploration!**
    - https://www.youtube.com/watch?v=R67XuYc9NQ4

APPENDIX C: Online resources

- **Python**
  - **NumPy and Pandas Tutorial | Data Analysis With Python | Python Tutorial for Beginners | Simplilearn**
    - https://www.youtube.com/watch?v=FniLzpaSFGk
  - **Data Analysis with Python - Full Course for Beginners (Numpy, Pandas, Matplotlib, Seaborn)**
    - https://www.youtube.com/watch?v=r-uOLxNrNk8

APPENDIX C: Online resources

- Git/GitHub
  - **Git and GitHub for Beginners - Crash Course**
    - https://www.youtube.com/watch?v=RGOj5yH7evk
  - **Git Tutorial for Beginners: Learn Git in 1 Hour**
    - https://www.youtube.com/watch?v=8JJ101D3knE
  - **Git Tutorial For Dummies**
    - https://www.youtube.com/watch?v=mJ-qvsxPHpY
  - **Learn how to use GitHub for Beginners | GitHub Tutorial**
    - https://www.youtube.com/watch?v=HJAwAKwFX-A

APPENDIX C: Online resources

- **Git/GitHub**
  - **Git Tutorial for Beginners - Git & GitHub Fundamentals In Depth**
    - https://www.youtube.com/watch?v=DVRQoVRzMIY

APPENDIX C: Online resources

- Other
  - **StatQuest**
    - https://www.youtube.com/c/joshstarmer
  - **3Blue 1Brown**
    - https://www.youtube.com/c/3blue1brown

APPENDIX C: Online resources

- Online courses
  - **Data Science Specialization**
    - https://www.coursera.org/specializations/jhu-data-science
  - **Machine Learning Specialization**
    - https://www.coursera.org/specializations/machine-learning-introduction
  - **Introduction to Machine Learning for Data Science**
    - https://www.udemy.com/course/machine-learning-for-data-science/
  - **Introduction to Data Science in Python**
    - https://www.datacamp.com/courses/introduction-to-data-science-in-python
  - **Machine Learning with Python**
    - https://www.coursera.org/learn/machine-learning-with-python

APPENDIX C: Online resources

- **Data Science Foundations**
  - https://www.codecademy.com/learn/paths/data-science-foundations
- **Data Science: Machine Learning**
  - https://pll.harvard.edu/course/data-science-machine-learning?delta=1
- **Python for Data Science and Machine Learning Bootcamp**
  - https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/