ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό

Ενότητα # 1: Εισαγωγή

Διδάσκων: Μιχάλης Βαζιργιάννης

Τμήμα: Προπτυχιακό Πρόγραμμα Σπουδών "Πληροφορικής"

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

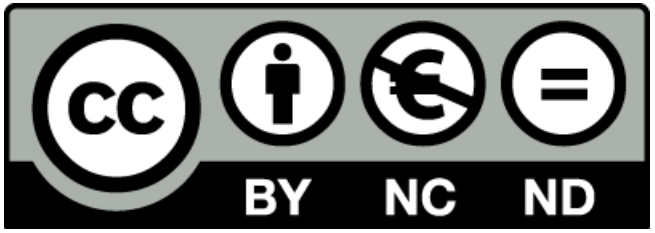
Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Οικονομικό Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Οι εικόνες προέρχονται … .



Σκοποί ενότητας

Εισαγωγή στις μεθόδους Dimesionality Reduction and Feature Selection, Supervised learning, Unsupervised Learning (Clustering), Community Detection and Evaluation in Social and Information Networks, Web Mining, Introduction to Big Data που θα δούμε πιο αναλυτικά στις επόμενες ενότητες.

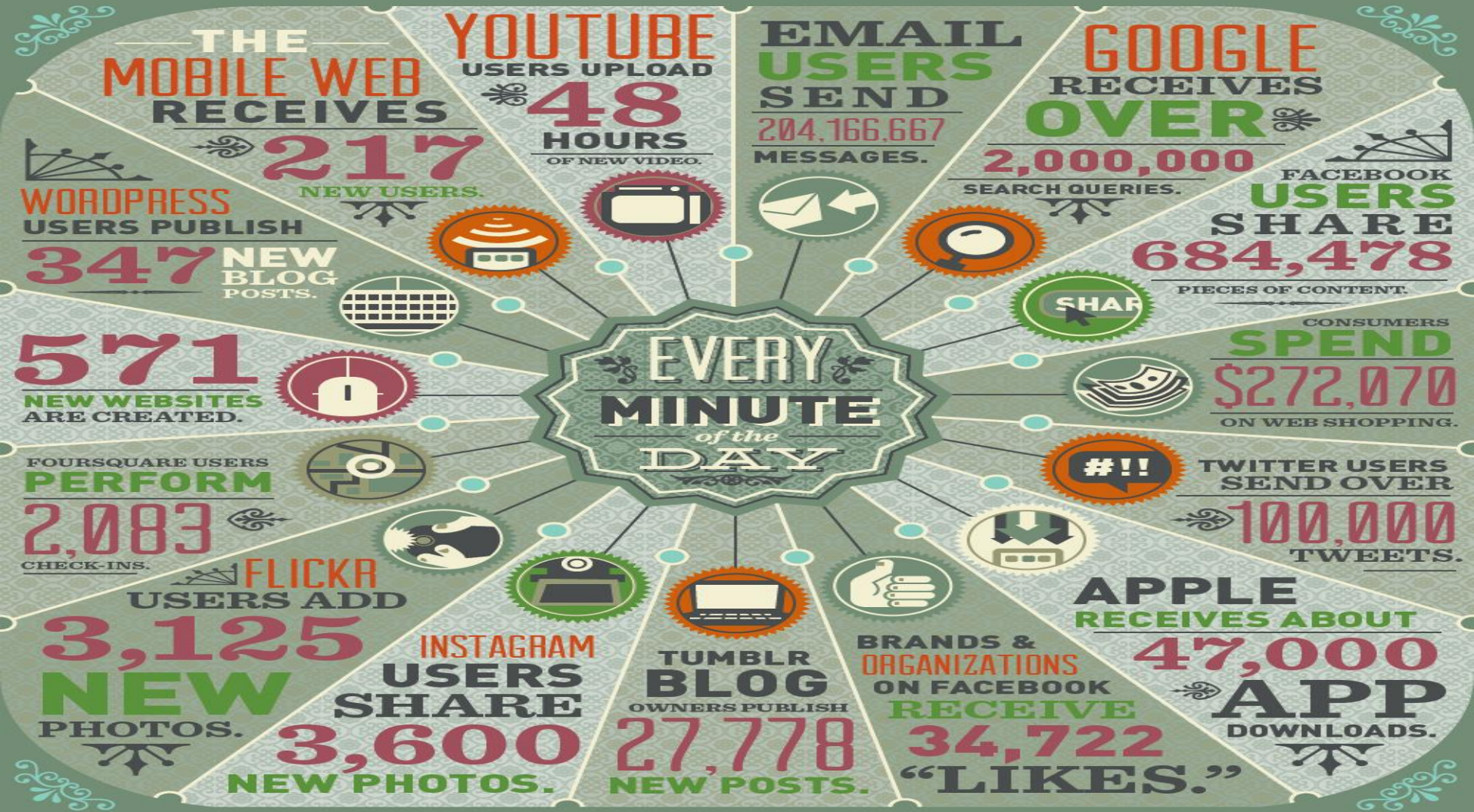
Περιεχόμενα ενότητας

- Data Science
- Text/Web Mining
- Graph Mining
- Big Data

DATA NEVER SLEEPS

How Much Data Is Generated **Every Minute**?

Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital trails that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.



WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global Internet population grew 6.59 percent from 2010 to 2011 and now represents **2.1 BILLION PEOPLE.**

These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.

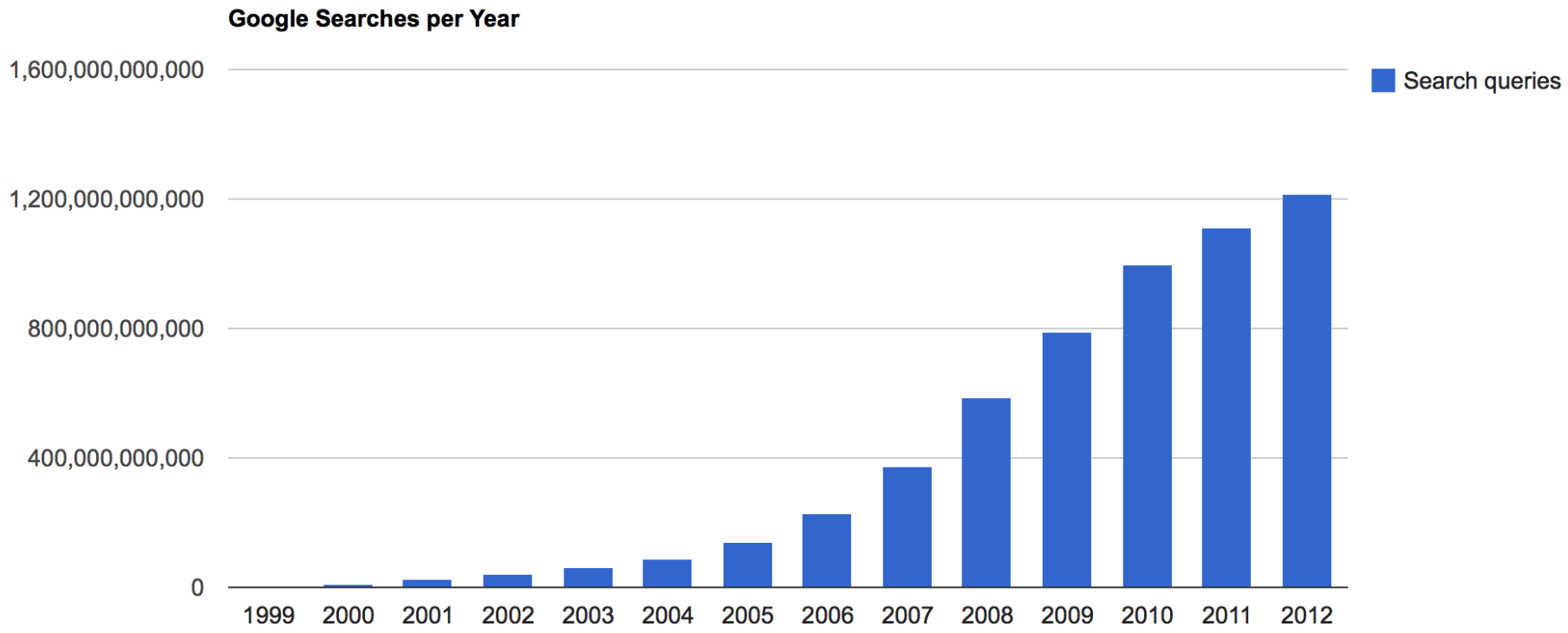
New kinds of data

- Traditional: numerical, categorical, or binary
- Text: emails, tweets, *New York Times* articles
- Records: user-level data, time-stamped event data, json formatted log files
- Geo-based location data
- Network
- Sensor data
- Images

The Bigdata challenge

Volume

- Billions of web pages, Goggle queries/day



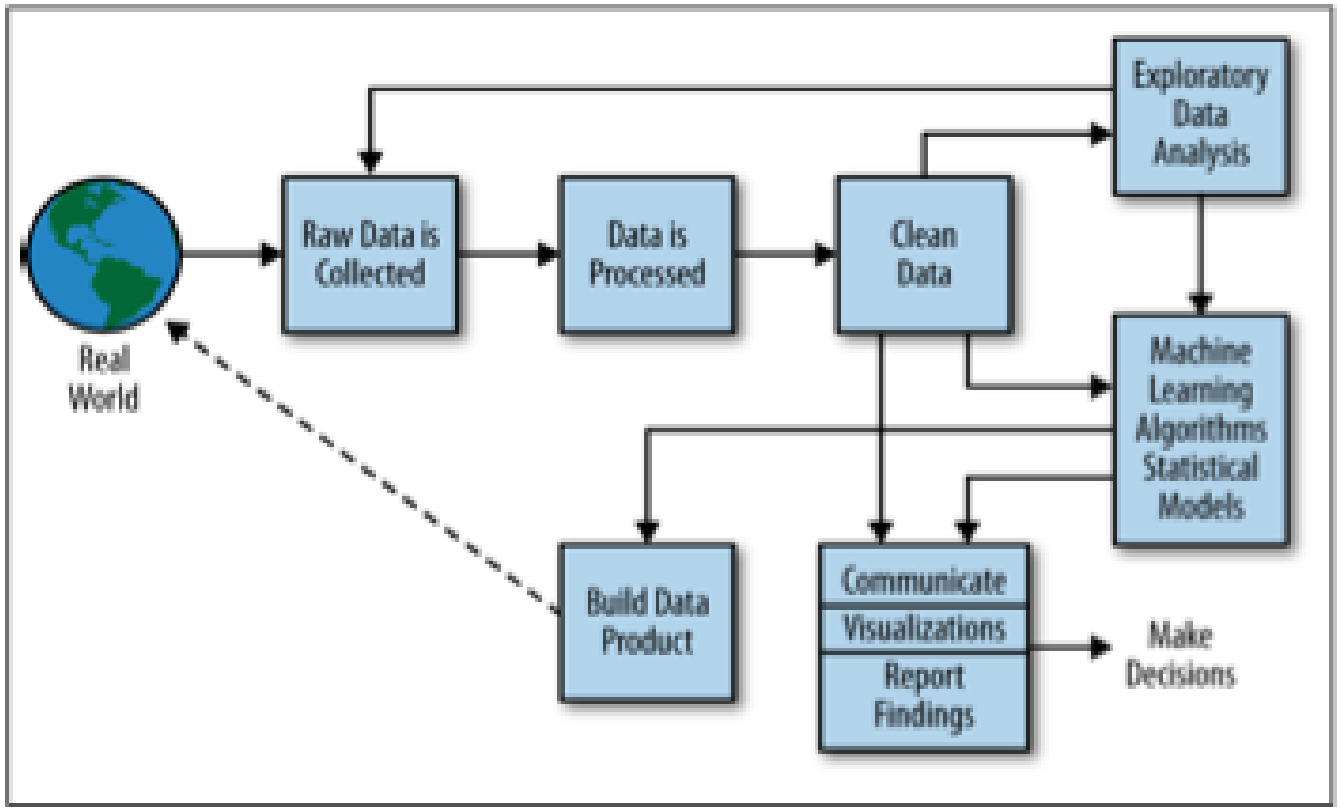
Data Science

Datafication[1]

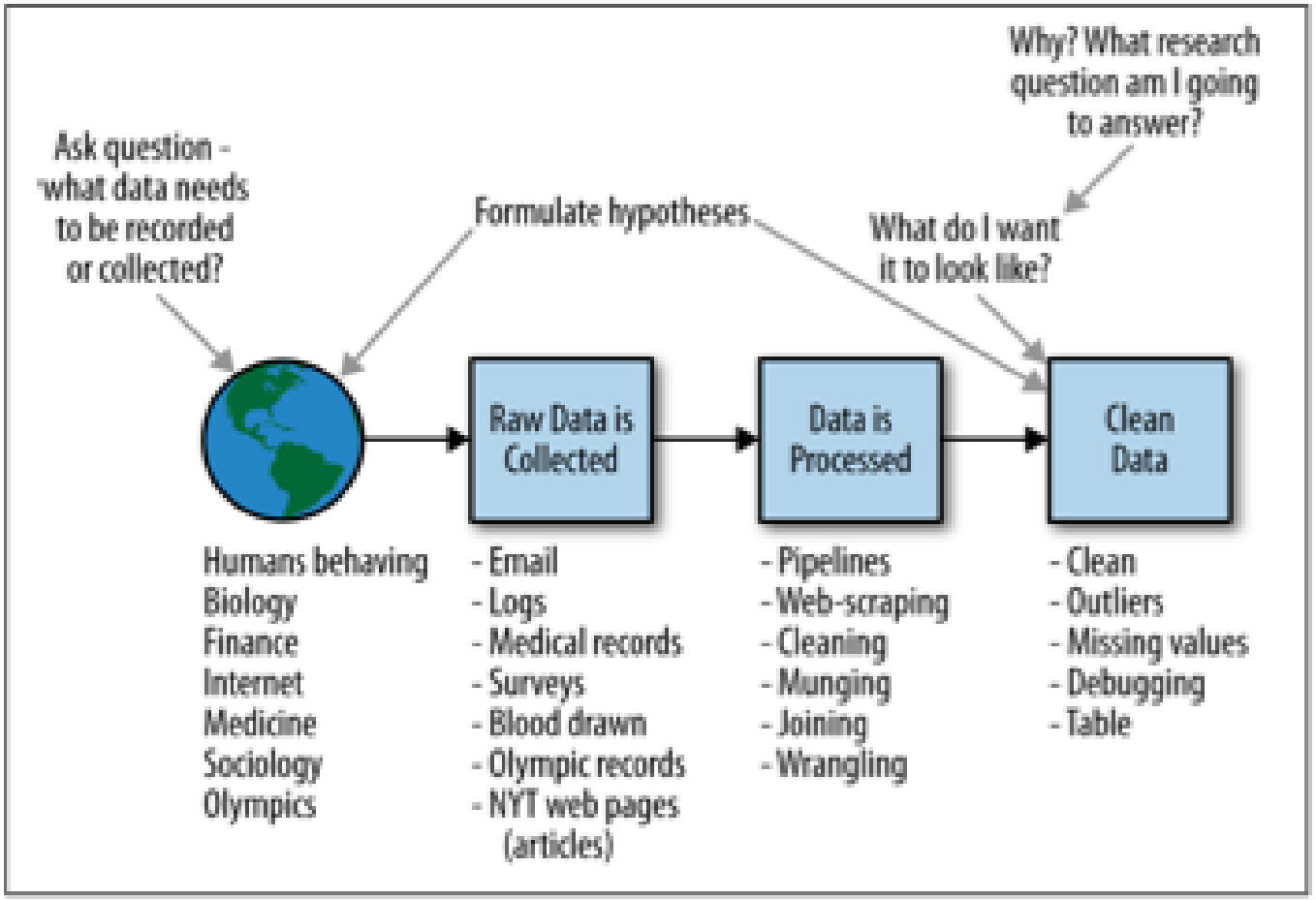
- massive amounts of data for of our lives
  - Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions …**tracked online**
- not just Internet data, though
  - finance, the medical industry, pharmaceuticals, bioinformatics, social welfare, government, educa tion, retail, ….
- growing influence of data in most sectors and most industries.
- abundance of inexpensive computing power.

[1] May/June 2013 issue of Foreign Affairs, Kenneth Neil Cukier and Viktor Mayer-Schoenberger wrote an article called “The Rise of Big Data”

The Data Science Process



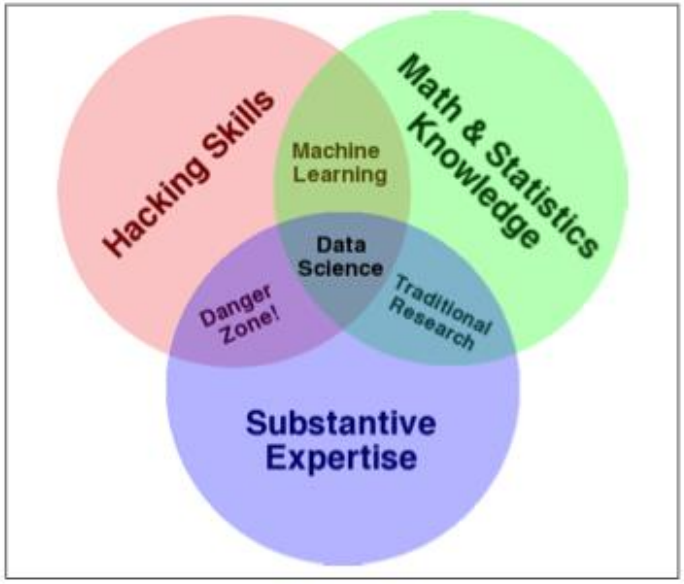
The role of data scientist



Data Science profile

A Data Science Profile

skill levels in the following domains:



- **Computer science (INF553)**
- Math
- Statistics
- **Machine learning (INF582)**
- Domain expertise
- Communication and presentation skills
- Data visualization

Course objectives

To equip the student with

- introduction, overview and hands on experience on the fundamental - as well as advanced – methods, algorithms and technologies for pre-processing potentially large data sets and learn patterns and knowledge from them

Course Syllabus

Introduction to Machine Learning

- Pre-processing, Exploration, Feature selection, Dimensionality reduction, feature extraction and evaluation
- Supervised Learning (k-nn, regression, logistic regression, decision trees
- Unsupervised learning (Clustering, K-means, EM, Spectral Clustering)

Text/Web Mining

- Text indexing and retrieval, Novelty detection
- Recommendations/collaborative filtering

Syllabus

Graph Mining

- Graph Ranking algorithms and evaluation measures graph clustering and classification
- Degeneracy (k-core & extensions)
- community mining methods & applications in social networks

Bigdata

- Map reduce - distributed processing, technologies (Hadoop, Map Reduce, NoSQL storage - Hbase)
- Query languages (HIVE)

Course Logistics

Assignments (A1, A2, …)

Course project (CP)

Grading scheme

- Final Grade = A*~10% + CP*~25% + Exams*~65%

Course/Lab Material:

https://eclass.aueb.gr/courses/INF131/

Suggested textbooks

- **Learning from Data –** Y. Abu-Mostafa, M. Magdon-Ismail, Hsuan-Tien Lin,
- **Doing Data Science, Straight Talk from the Frontline,** Cathy O'Neil, Rachel Schutt
- Pattern Recognition and Machine Learning (Information Science and Statistics) Hardcover – October 1, 2007, Christopher M. Bishop
- **Hadoop: The Definitive Guide, 3rd Edition,** T. WhiteSlides

+ course note/slides on advanced issues

Big data

- Capacity to store information has doubled every 40 months since the 1980s
- In 2012, 2.5 exabytes (2.5 × 1018) created per day
- Big internet companies such as Google, Amazon, Facebook, but also industries from pharmaceuticals, insurance, banks, telecoms, personalized medicine, marketing, bioinformatics

Examples of data volumes

- MEDLINE text database
  - 12 million published articles
- Google
  - 4.2 billion Web pages indexed
  - 80 million site visitors per day
- CALTRANS loop sensor data
  - Every 30 seconds, thousands of sensors, 2Gbytes per day
- NASA MODIS satellite
  - Coverage at 250m resolution, 37 bands, whole earth, every day
- Walmart transaction data
  - Order of 100 million transactions per day

Competitions | Kaggle - Mozilla Firefox

e Edit View History Bookmarks Yahoo! Tools Help

Mozilla Firefox Start Page | CIKM 2013, Burlingame, CA, USA | eClass του Οικονομικού Πανεπιστημί... | Competitions | Kaggle | Kaggle Member FAQ

www.kaggle.com/competitions

καγγλε

YAHOO! καγγλε

kaggle

Customer Solutions ▾ Competitions Community ▾ Sign Up Login

Welcome to Kaggle, the leading platform for predictive modeling competitions. Here's how to jump into competing on Kaggle —

New to Data Science? Visit our Wiki »
Learn about hosting a competition »
in-Class & Research competitions »

Enter

Find a competition & download the training data. You don't need new software/skills to submit.

Build

Build a model using whatever methods you prefer and upload your predictions to Kaggle.

...Win!

Kaggle scores your solution in real time and you'll see your place on the live leaderboard.

Active Competitions

All Competitions

116 found, 14 active

Search competitions

- **All competitions**
- Enterable

Status

- **Active**

Competition Name	Reward	Teams	▾ Deadline
 **Titanic: Machine Learning from Disaster** Predict survival on the Titanic (with tutorials in Excel, Python and an introduction to Random Forests)	Knowledge	6876	11 months
 **Digit Recognizer** Classify handwritten digits using the famous MNIST data	Knowledge	1947	9 months
 **Data Science London + Scikit-learn** Scikit-learn is an open-source machine learning library for Python. Give it a try here!	Knowledge	501	4 months
 **Dogs vs. Cats**			

Find: Vaziq Next Previous Highlight all Match case

EL 3:50 μμ 9/10/2013

Two Types of Data

- Experimental Data
  - Hypothesis H
  - design an experiment to test H
  - collect data, infer how likely it is that H is true
  - e.g., clinical trials in medicine

- Observational or Retrospective or Secondary Data
  - massive non-experimental data sets
    - e.g., human genome, atmospheric simulations, etc
  - assumptions of experimental design no longer valid
  - how can we use such data to do science?
    - data must support model exploration, hypothesis testing

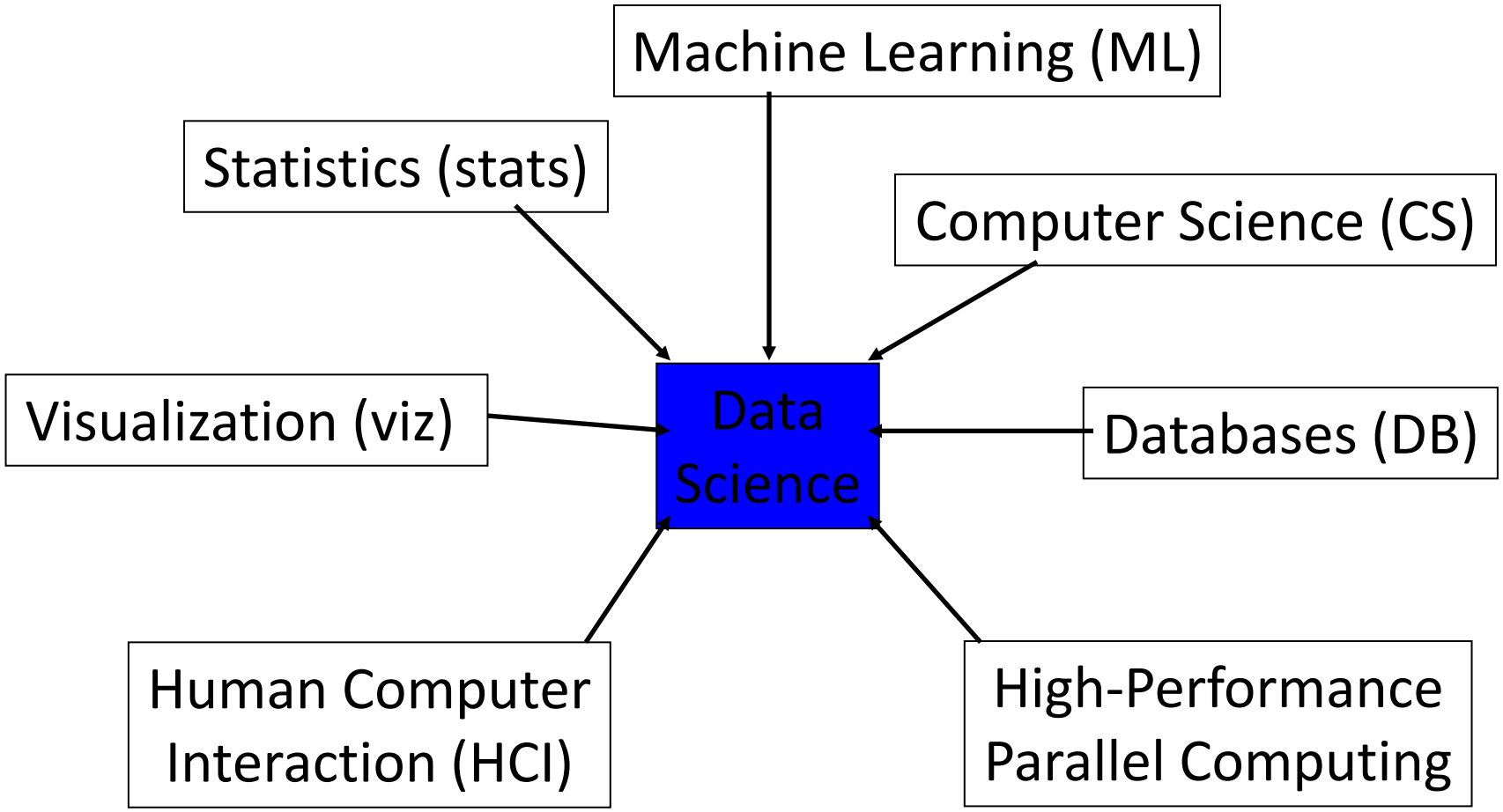
Data-Driven Discovery

- Observation data
  - cheap relative to experimental data
    - Examples:
      - Transaction data archives for retail stores, airlines, etc
      - Web logs for Amazon, Google, etc
      - The human/mouse/rat genome
      - Etc., etc

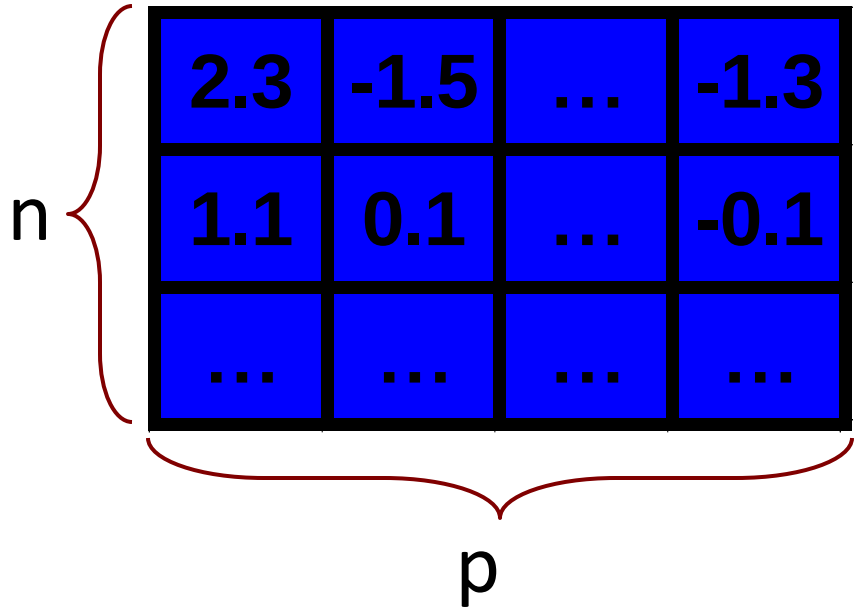
    $\Rightarrow$ makes sense to leverage available data

    $\Rightarrow$ useful (?) information may be hidden in vast archives of data
- Contrast data mining with traditional statistics
  - traditional stats: first hypothesize, then collect data, then analyze
  - data mining:
    - few if any a priori hypotheses,
    - data is usually already there
    - analysis is typically data-driven not hypothesis driven
  - Nonetheless, statistical ideas are very useful in data mining, e.g., in validating whether discovered knowledge is useful

DS: Intersection of Many Fields

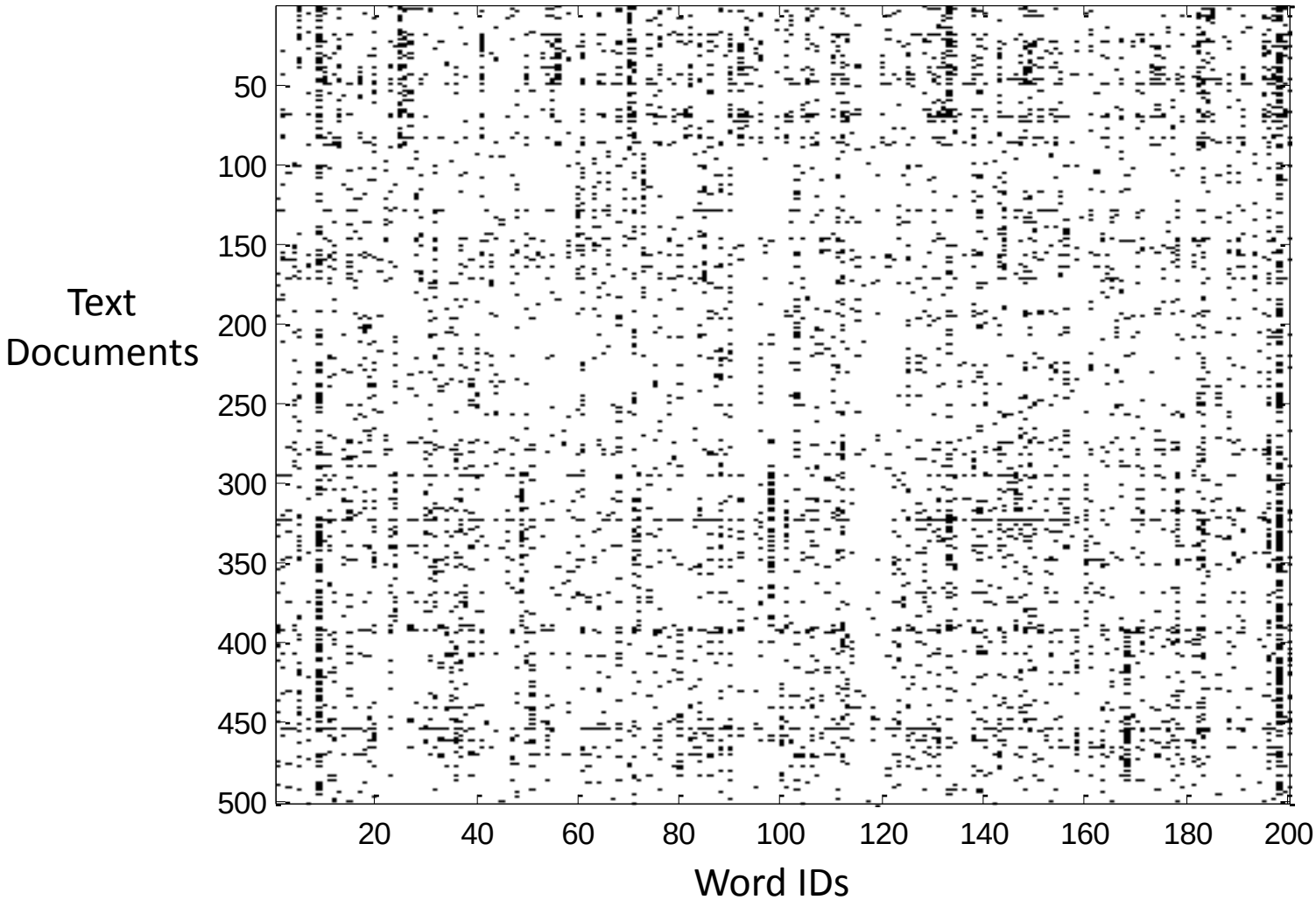


Flat File or Vector Data

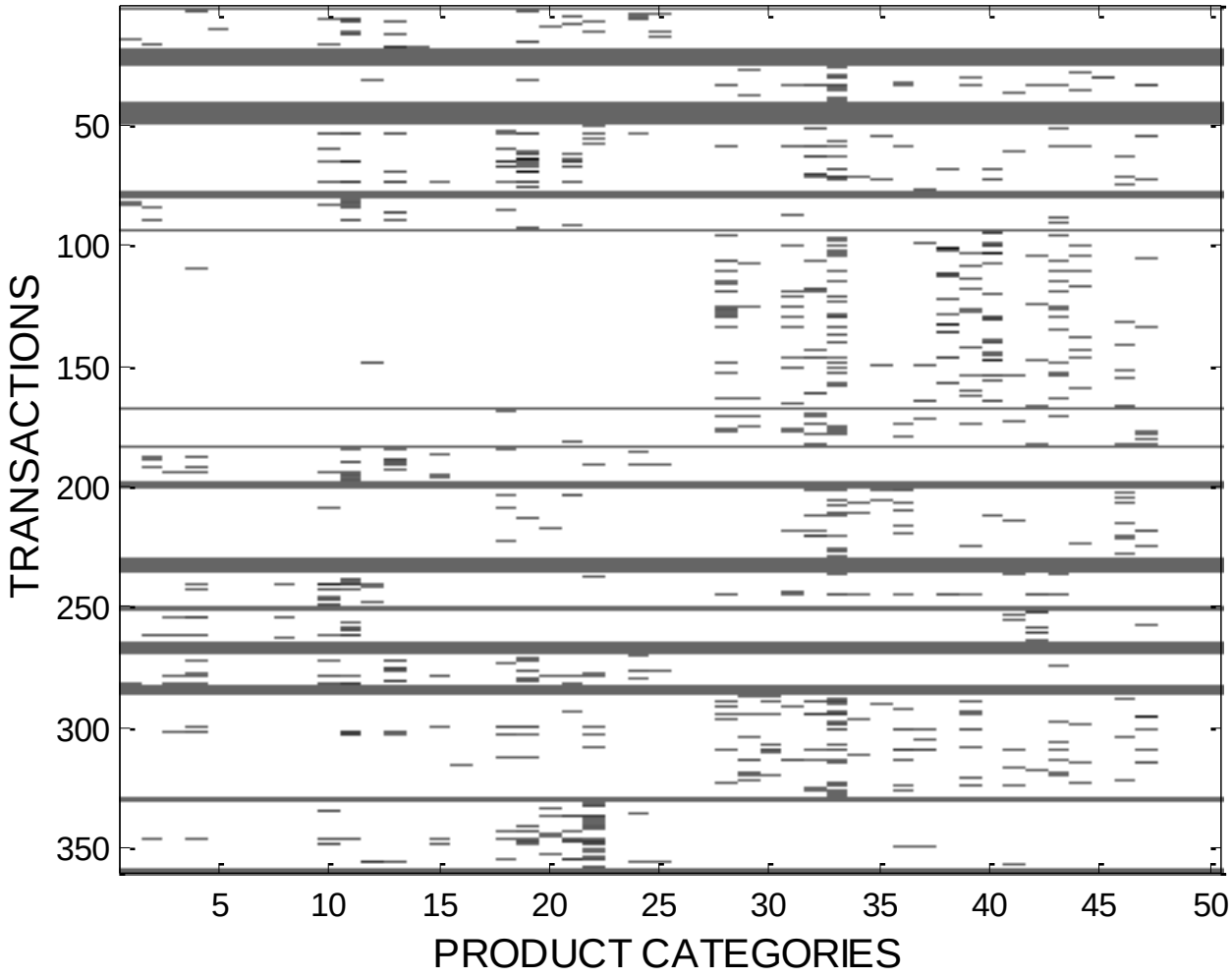


- Rows = objects
- Columns = measurements on objects
  - Represent each row as a p-dimensional vector, where p is the dimensionality
    - In efffect, embed our objects in a p-dimensional vector space
    - Often useful, but always appropriate
- Both n and p can be very large in certain data mining applications

Sparse Matrix (Text) Data

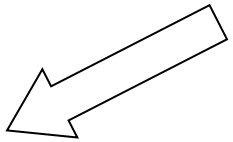


“Market Basket” Data



Sequence (Web) Data

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,



User 1	2 3 2 2 3 3 3 1 1 1 3 1 3 3 3 3
User 2	3 3 3 1 1 1
User 3	7 7 7 7 7 7 7 7
User 4	1 5 1 1 1 5 1 5 1 1 1 1 1 1
User 5	5 1 1 5
…	…

Time Series Data

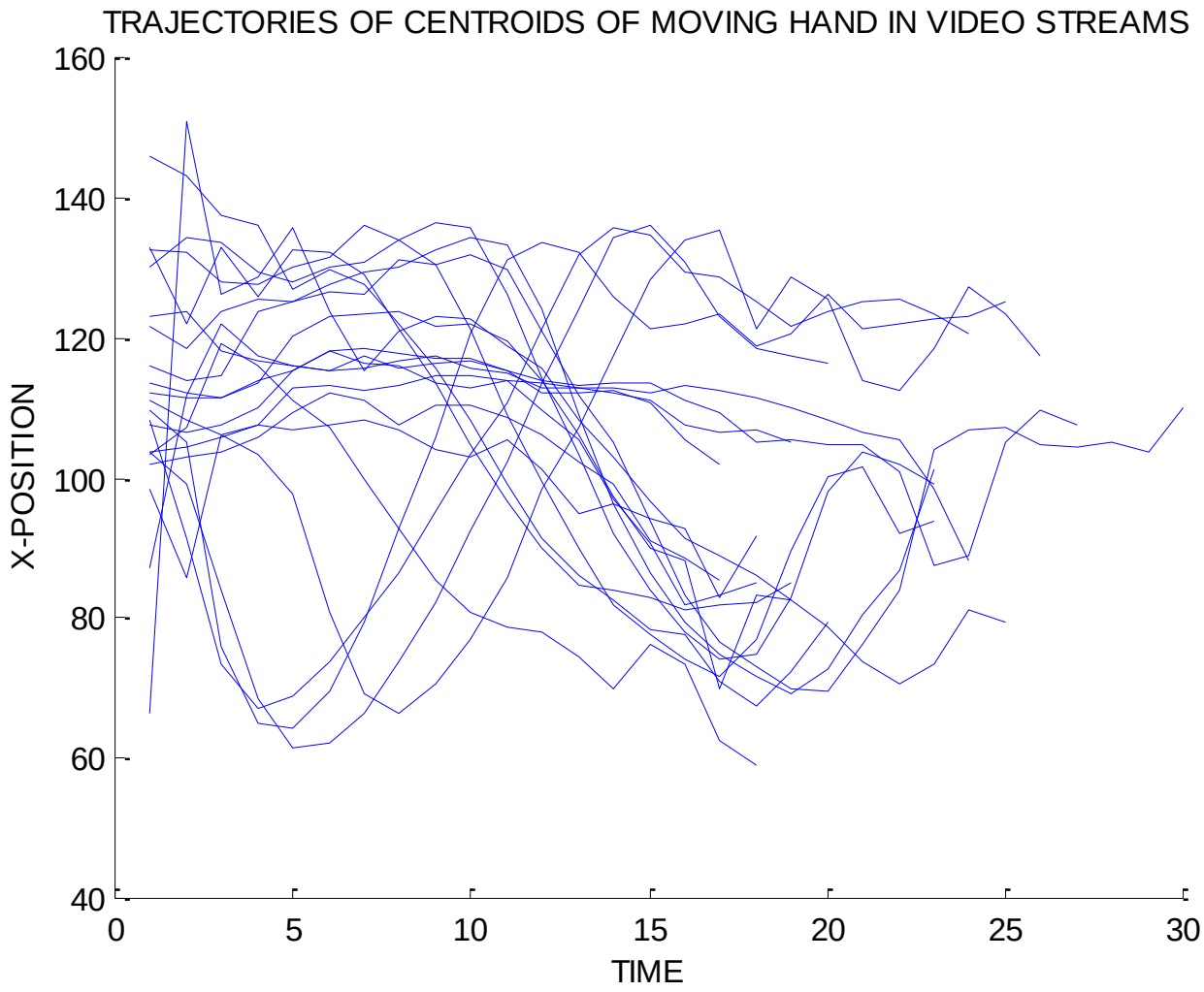
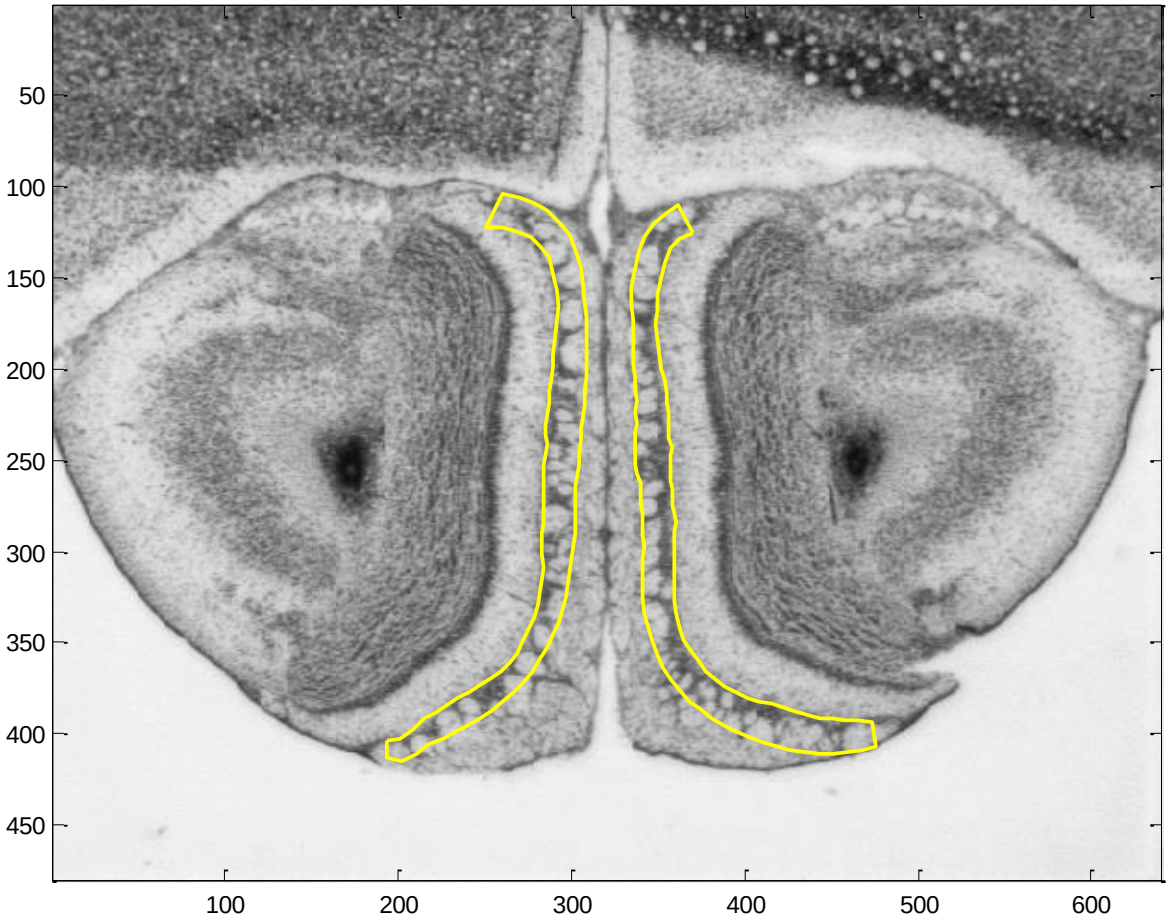
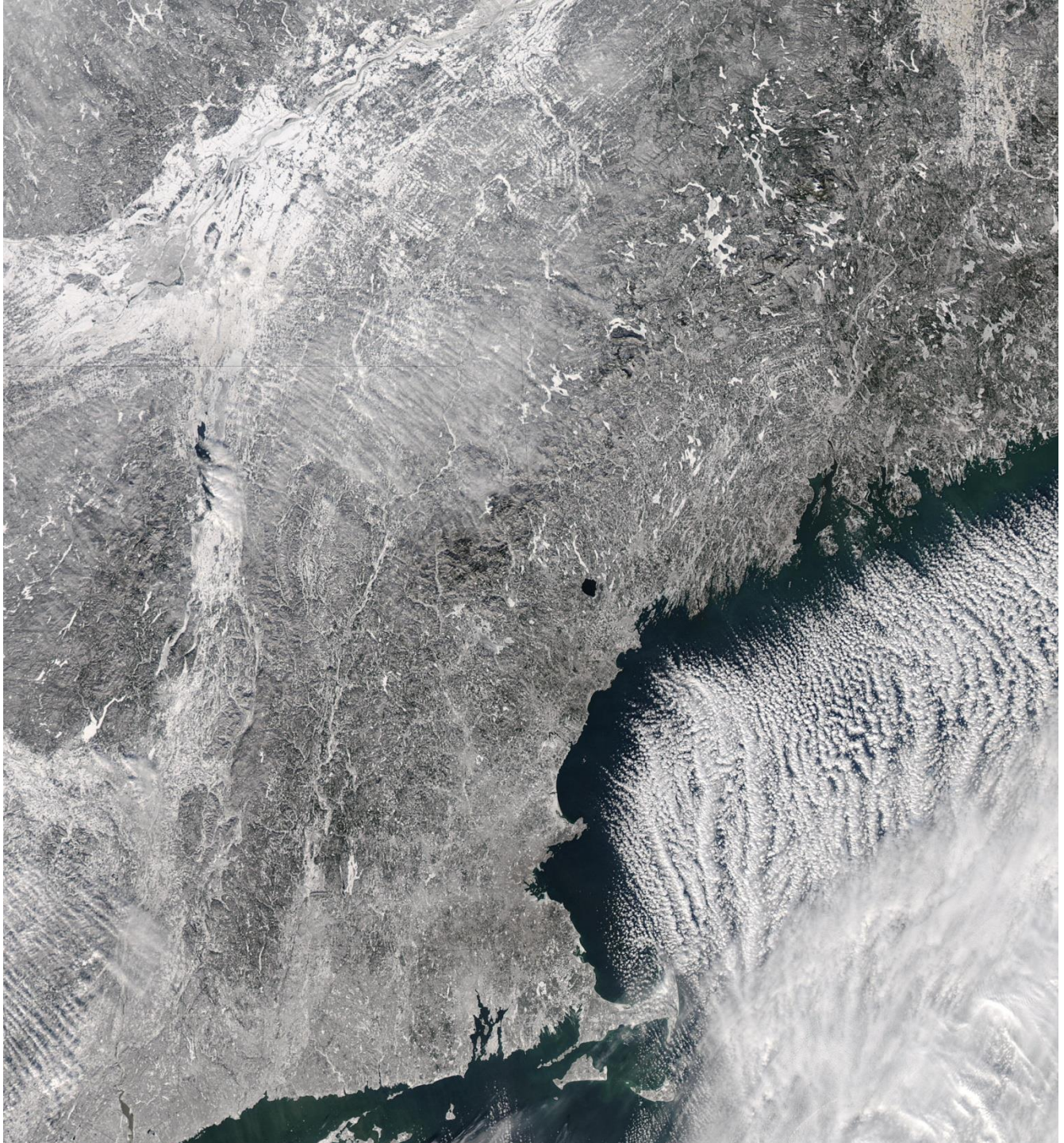
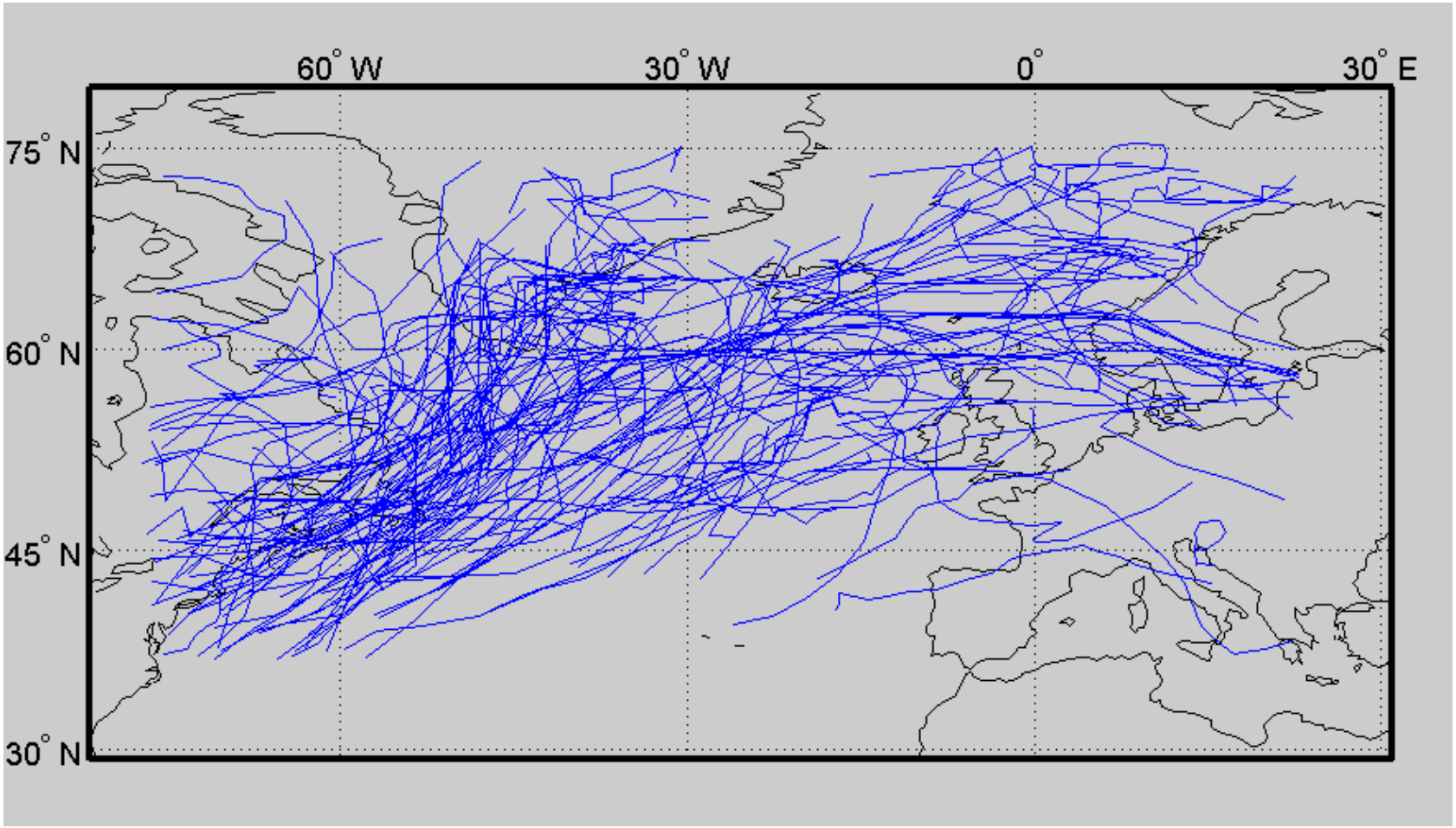


Image Data

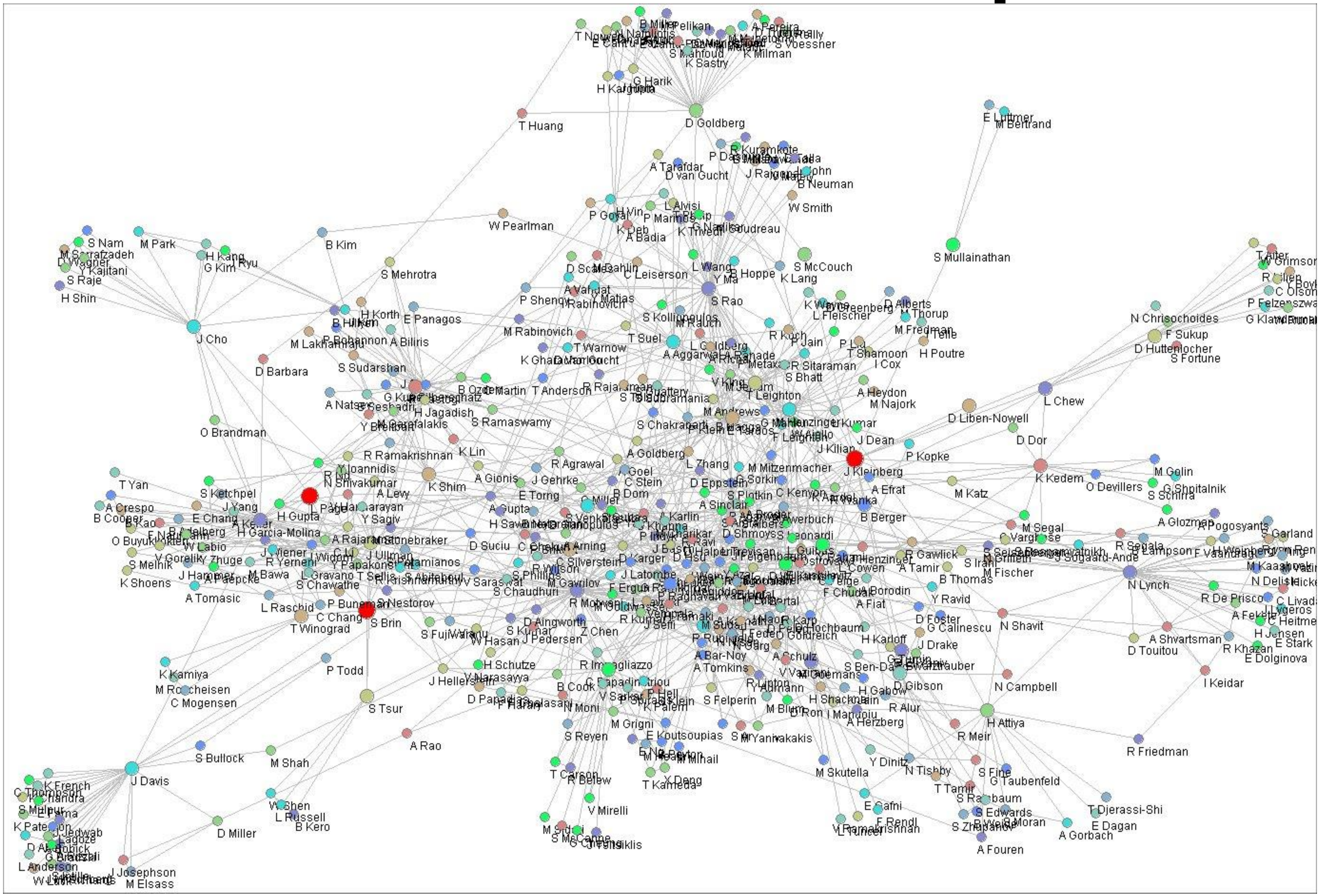




Spatio-temporal data



Social Networks – Graphs



Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others….

Exploratory Data Analysis

- Getting an overall sense of the data set
  - Computing summary statistics:
    - Number of distinct values, max, min, mean, median, variance, skewness,..

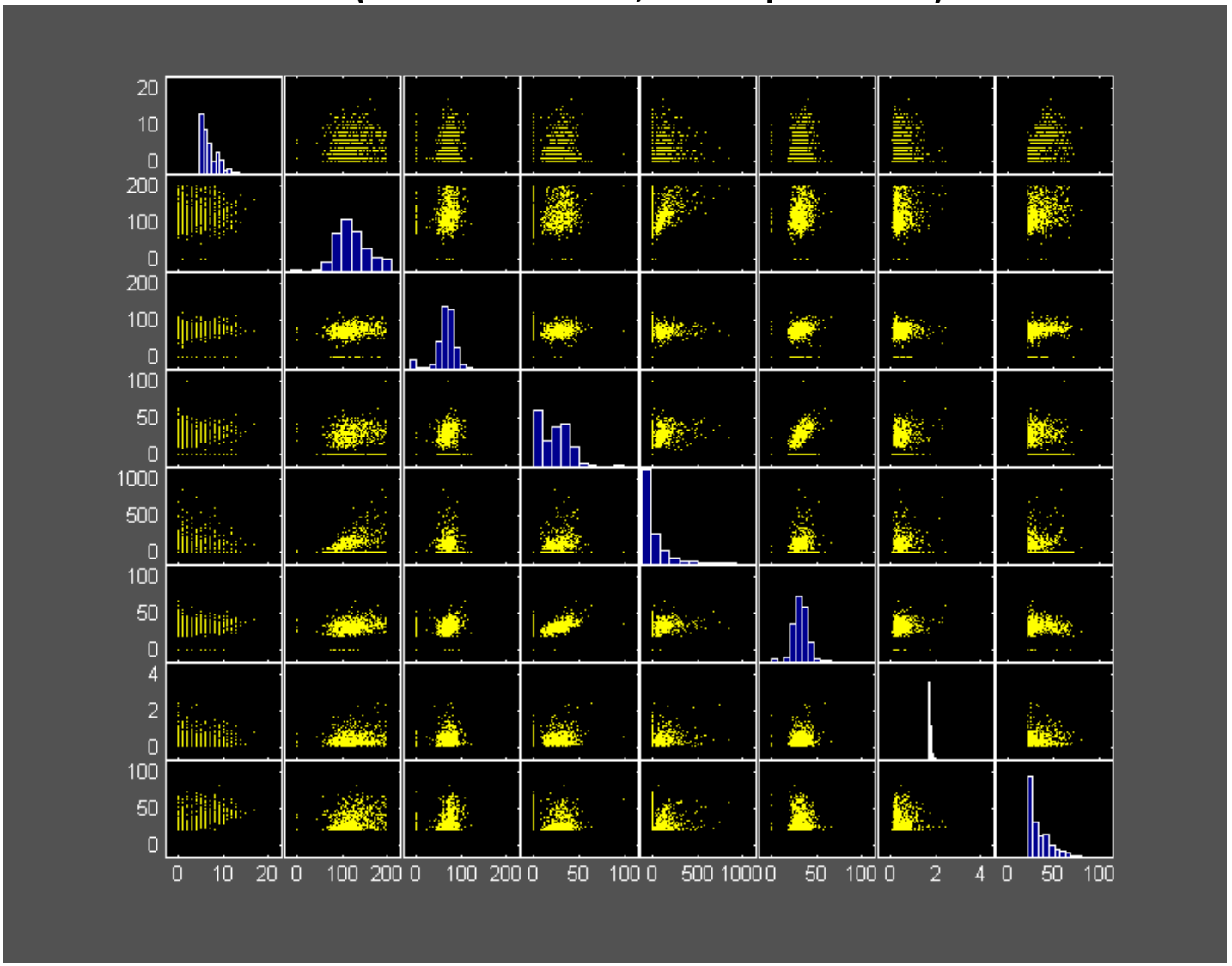
- Visualization is widely used
  - 1d histograms
  - 2d scatter plots
  - Higher-dimensional methods

- Useful for data checking
  - E.g., finding that a variable is always integer valued or positive
  - Finding the some variables are highly skewed

- Simple exploratory analysis can be extremely valuable
  - You should always “look” at your data before applying any data mining algorithms

Example of Exploratory Data Analysis

(Pima Indians data, scatter plot matrix)

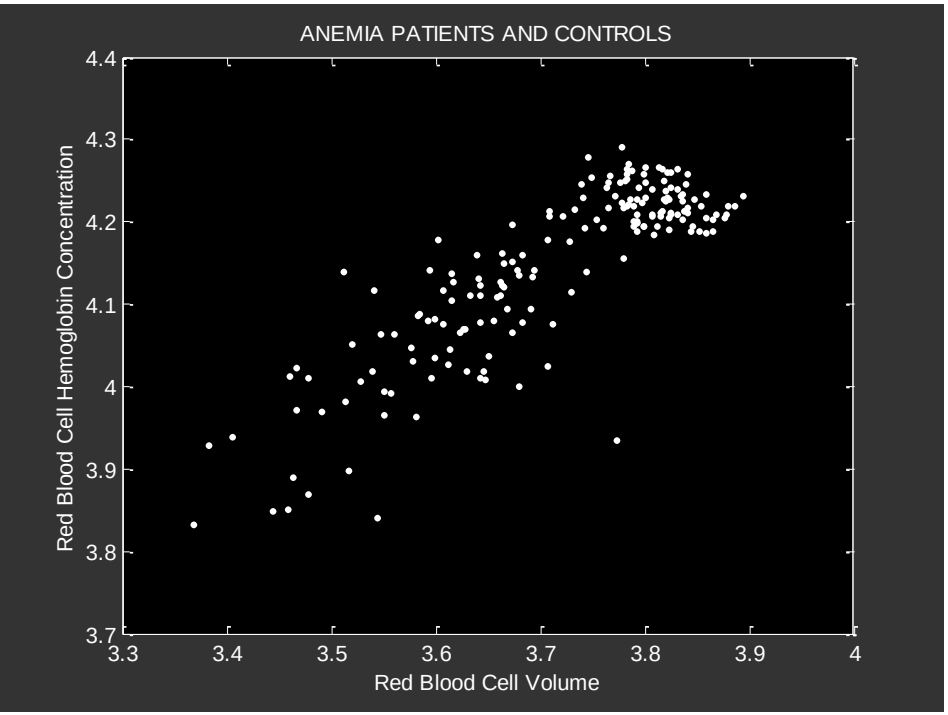


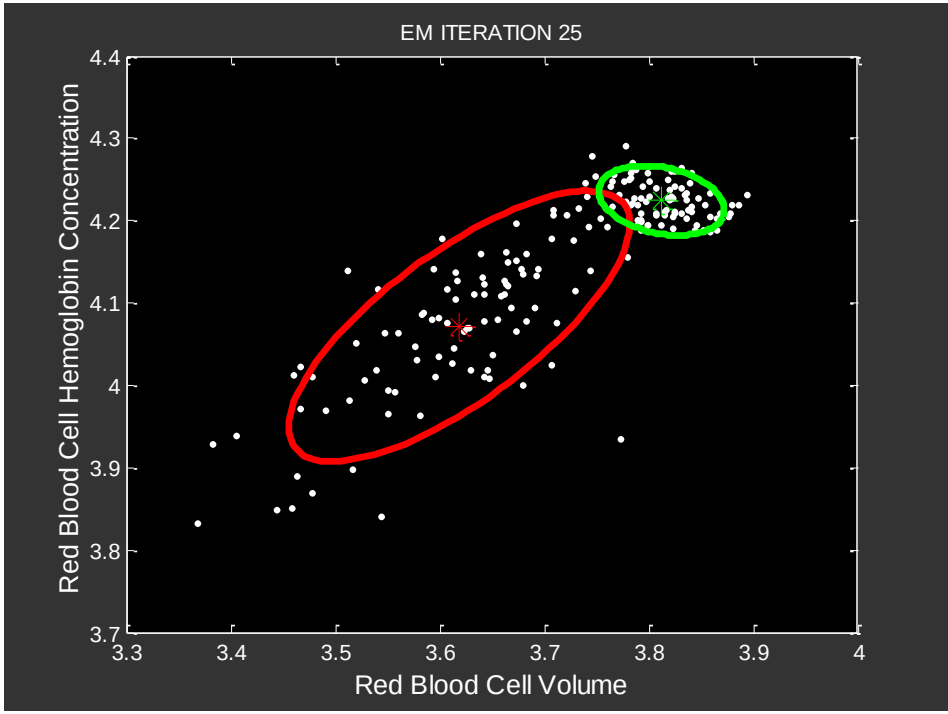
Descriptive Modeling

- Goal is to build a "generative" or "descriptive" model,
  - E.g., a model that could simulate the data if needed
  - models the underlying process

- Examples:
  - Density estimation:
    - estimate the joint distribution $P(x_1,\ldots\ldots x_p)$
  - Cluster analysis:
    - Find natural groups in the data
  - Dependency models among the p variables
    - Learning a Bayesian network for the data

Example of Descriptive Modeling

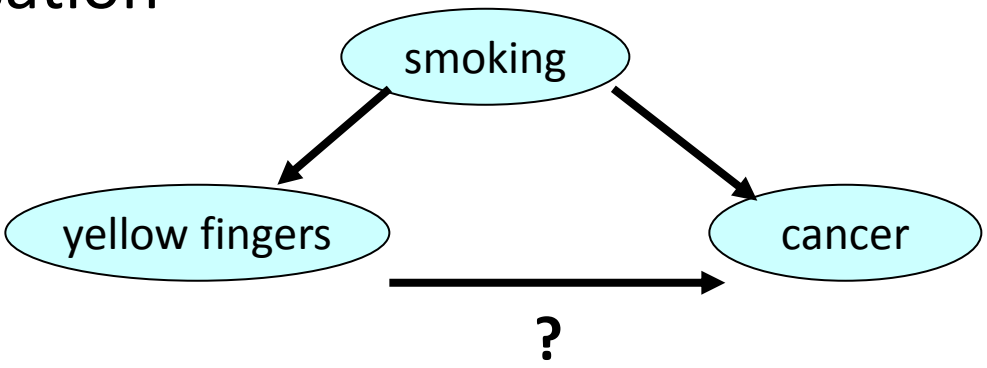




Another Example of Descriptive Modeling

- Learning Directed Graphical Models (aka Bayes Nets)
  - goal: learn directed relationships among p variables
  - example: Do yellow fingers cause lung cancer?
  - techniques: directed (causal) graphs
  - challenge: distinguishing between correlation and causation

hidden cause: smoking



Predictive Modeling

- Predict one variable Y given a set of other variables $\underline{X}$
  - Here $\underline{X}$ could be a p-dimensional vector
- Classification: Y is categorical
- Regression: Y is real-valued
- In effect this is function approximation, learning the relationship between Y and $\underline{X}$
- Many, many algorithms for predictive modeling in statistics and machine learning
- Often the emphasis is on predictive accuracy, less emphasis on understanding the model

Example of Predictive Modeling

- Background
  - AT&T has about 100 million customers
  - It logs 200 million calls per day, 40 attributes each
  - 250 million unique telephone numbers
  - Which are business and which are residential?
- Solution (Pregibon and Cortes, AT&T,1997)
  - Proprietary model, using a few attributes, trained on known business customers to adaptively track p(business|data)
  - Significant systems engineering: data are downloaded nightly, model updated (20 processors, 6Gb RAM, terabyte disk farm)
- Status:
  - running daily at AT&T
  - HTML interface used by AT&T marketing

Pattern Discovery

- Goal is to discover interesting “local” patterns in the data rather than to characterize the data globally
- given market basket data we might discover that
    - If customers buy wine and bread then they buy cheese with probability 0.9
    - These are known as “association rules”
- Given multivariate data on astronomical objects
    - We might find a small group of previously undiscovered objects that are very self-similar in our feature space, but are very far away in feature space from all other objects

Example of Pattern Discovery

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDDD
CDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBBBDCA
BDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCADCB
ACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDABDDBCADAD
BCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADCCBBADABBA
AADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDCCDBBC
DBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCDCCADAADACABDA
BAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBCBDBDBADBBBBCDAD
ABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBCCDBAAADDDBDDCABAC
BCADCDCBAAADCADDADAABBACCBB

Example of Pattern Discovery

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDDD
CDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBBBDCA
BDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCADCB
ACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDABDDBCADAD
BCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADCCBBADABBA
AADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDCCDBBC
DBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCDCCADAADACABDA
BAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBCBDBDBADBBBBCDAD
ABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBCCDBAAADDDBDDCABAC
BCADCDCBAAADCADDADAABBACCBB

Structure: Models and Patterns

- Model = abstract representation of a process

  e.g., very simple linear model structure

  $$Y = a X + b$$

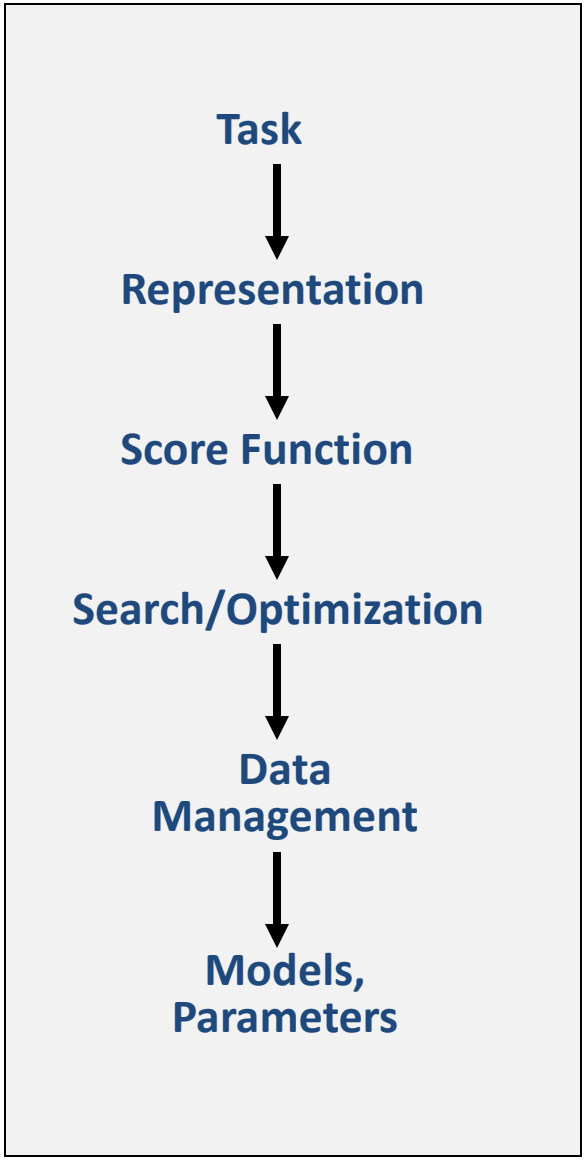
  - a and b are parameters determined from the data
  - Y = aX + b is the model structure
  - Y = 0.9X + 0.3 is a particular model
  - "All models are wrong, some are useful" (G.E. Box)

- Pattern represents "local structure" in a data set
  - E.g., if X>x then Y >y with probability p
  - or a pattern might be a small cluster of outliers in multi-dimensional space

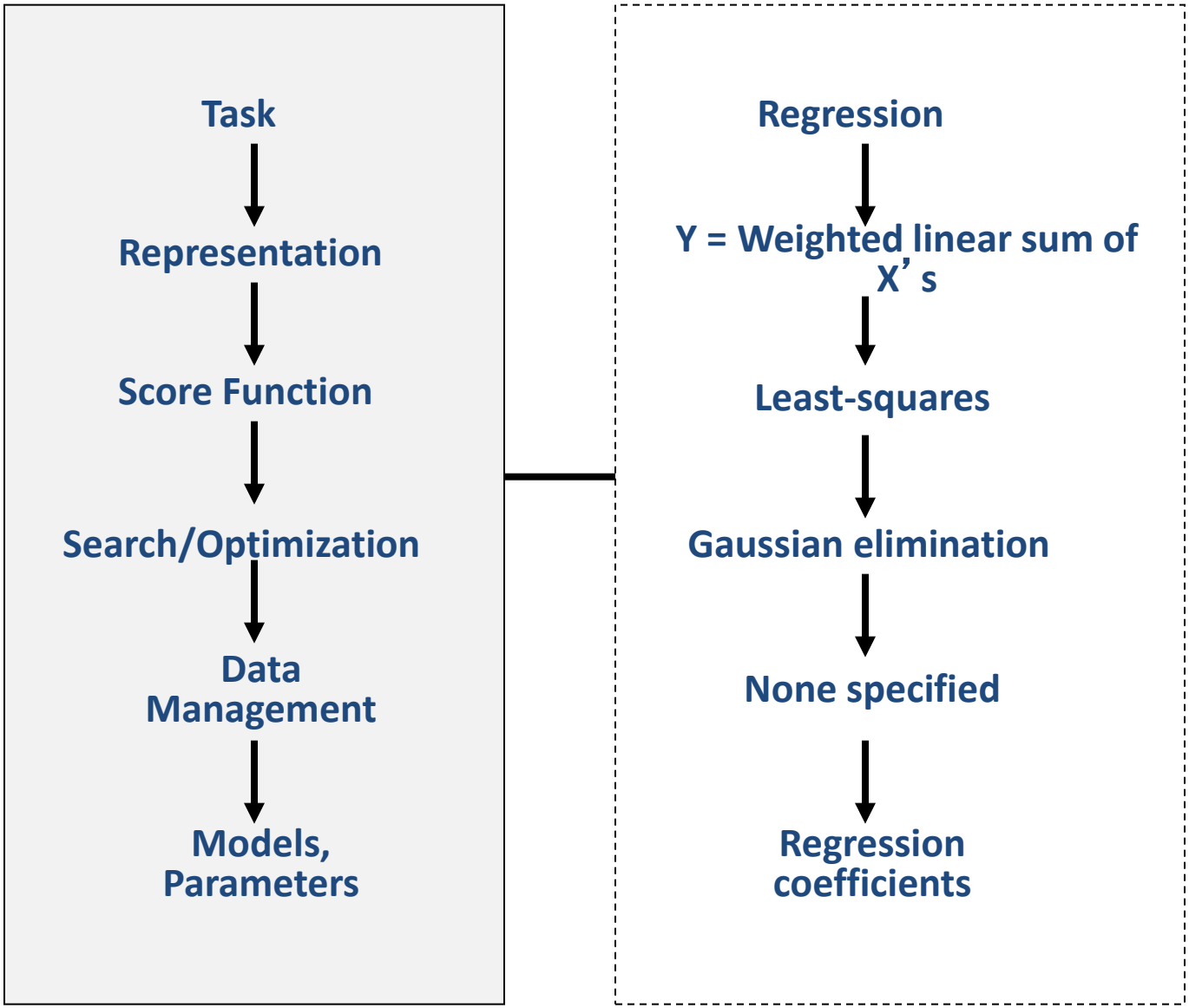
Components of Data Mining Algorithms

- Representation:
  - Determining the nature and structure of the representation to be used;
- Score function
  - quantifying and comparing how well different representations fit the data
- Search/Optimization method
  - Choosing an algorithmic process to optimize the score function; and
- Data Management
  - Deciding what principles of data management are required to implement the algorithms efficiently.

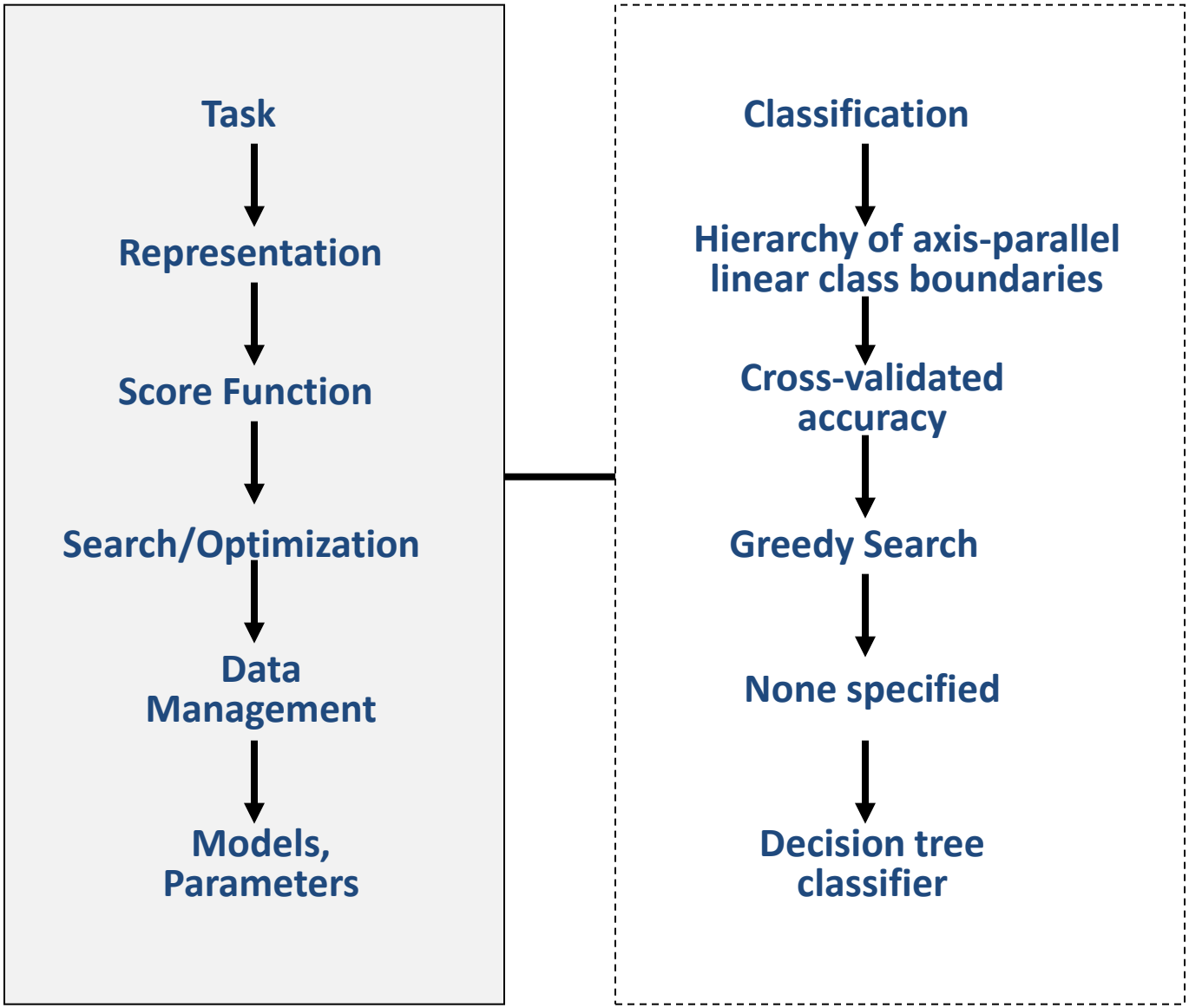
What's in a Data Mining Algorithm?



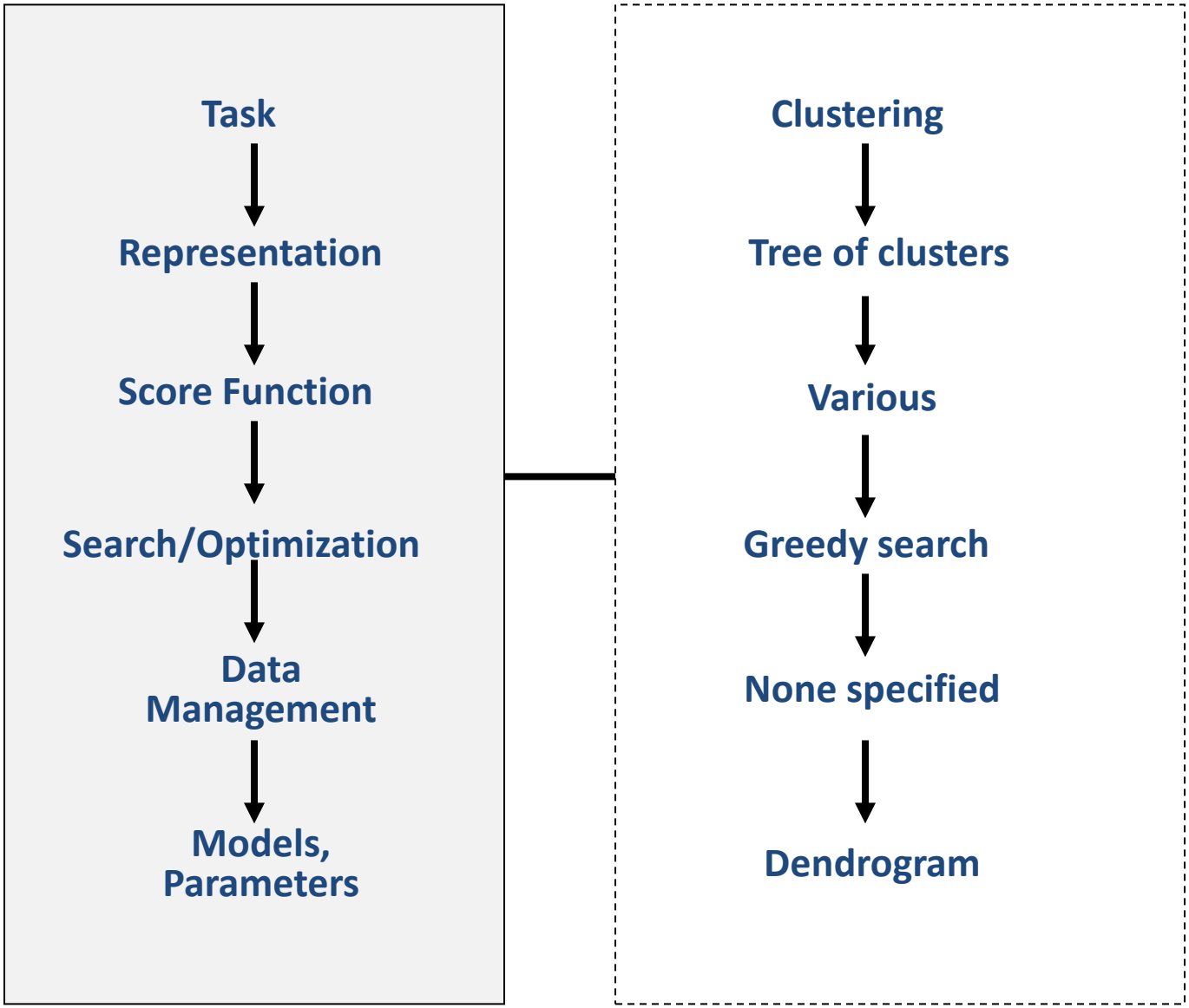
An Example: Multivariate Linear Regression



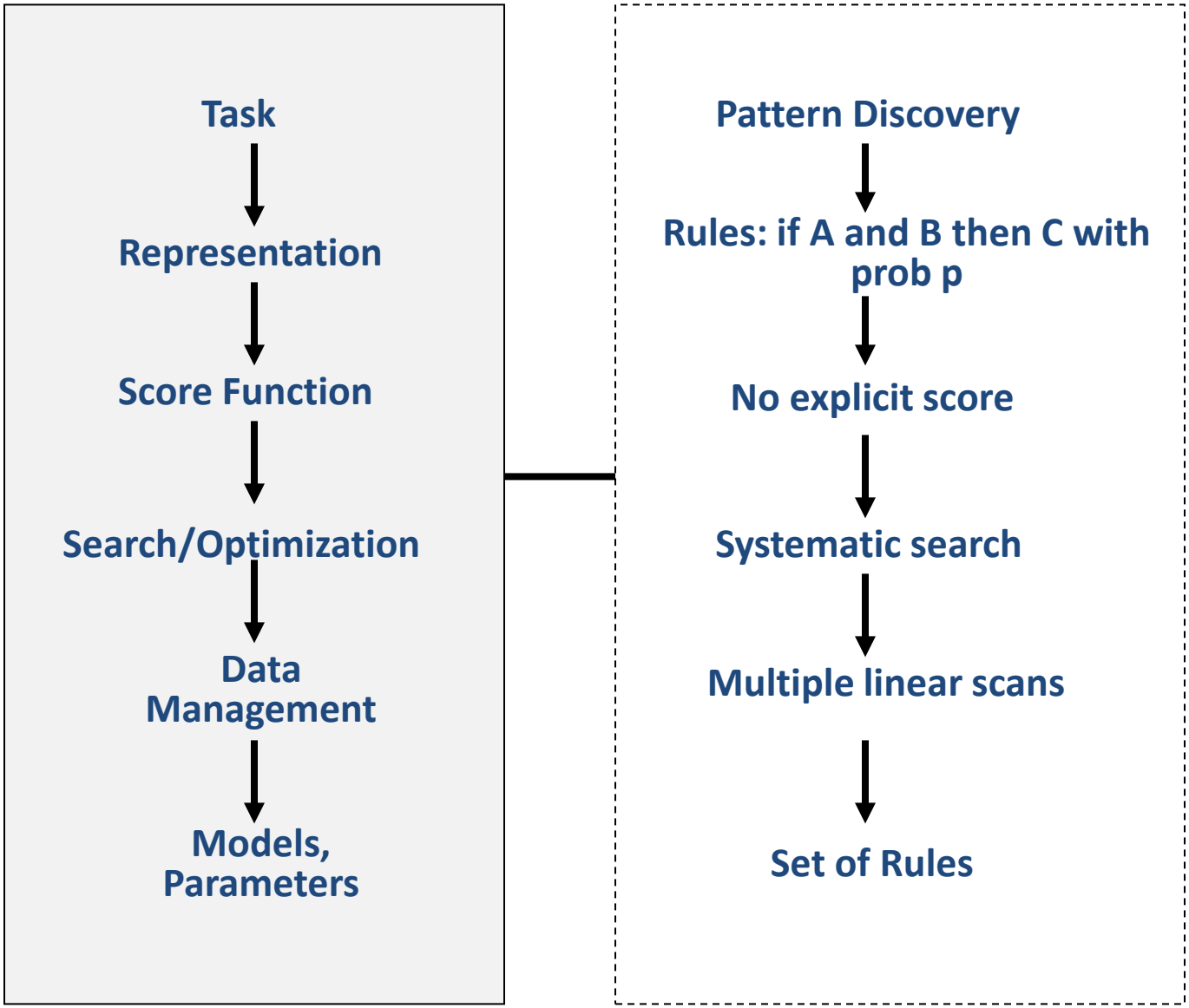
An Example: Decision Trees (C4.5 or CART)



An Example: Hierarchical Clustering



An Example: Association Rules



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Τέλος Ενότητας # 1

Μάθημα: Εξόρυξη γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, **Ενότητα # 1:** Εισαγωγή

Διδάσκων: Μιχάλης Βαζιργιάννης**, Τμήμα:** Προπτυχιακό Πρόγραμμα Σπουδών “Πληροφορικής”