Data Mining

Lecture 1: Introduction to Machine Learning

Prof. M. Vazirgiannis

Data Science and Mining Team, LIX,
Ecole Polytechnique, IPP, France
http://www.lix.polytechnique.fr/dascim/

February, 2022

Course Syllabus

- **General Introduction to Machine Learning**
  - Machine Learning paradigms
  - The Machine Learning Pipeline
- **Supervised Learning**
  - Generative and non-generative methods
  - Naive Bayes, KNN and regressions
  - Tree based methods
- **Unsupervised Learning**
  - Dimensionality reduction
  - Clustering
- **Advanced Machine Learning Concepts**
  - Regularization
  - Model selection

  Kernels
  - Introduction to kernels
  - Support Vector Machines

Course Syllabus

- **Neural Networks**
  - Introduction to Neural Networks
  - Perceptrons and back-propagation
- **Deep Learning I**
  - Convolutional Neural Networks
  - Recurrent Neural Networks
  - Applications
- **Deep Learning II**
  - Modern Natural Language Processing
  - Unsupervised Deep Learning
  - Embeddings, Auto-Encoders, Generative Adversarial Networks
- Introduction to advanced topics
  - Deep Learning for NLP (word embeddings, attention)
  - Deep Learning for graphs (node, graph embeddings)
  - Deep Learning for recommendations

Course Logistics (important for 2021)

- Class and Labs: 15:00 – 18:00 – synchronous video conf:
  - Zoom link for ALL classes: https://zoom.com.cn/j/61306355467; Passcode: 633547
  - *Need to appear with your real surname@affiliation* (i.e. Smith@Ecole Polytechnique, Loison@M1_AI_IPP)
- Interaction/Q&As outside of course hours:
  - Would a slack workspace be useful?
  - Read our announcements carefully
- **Install the software requested (Anaconda)!**

Course Logistics

- Class Presense/Interaction (A), Course project (CP) – data challenge
- Grading scheme
  - Final Grade = A*20% + CP*80%
- Course/Lab Material will be uploaded to the Moodle course: “Data mining 数据挖掘”

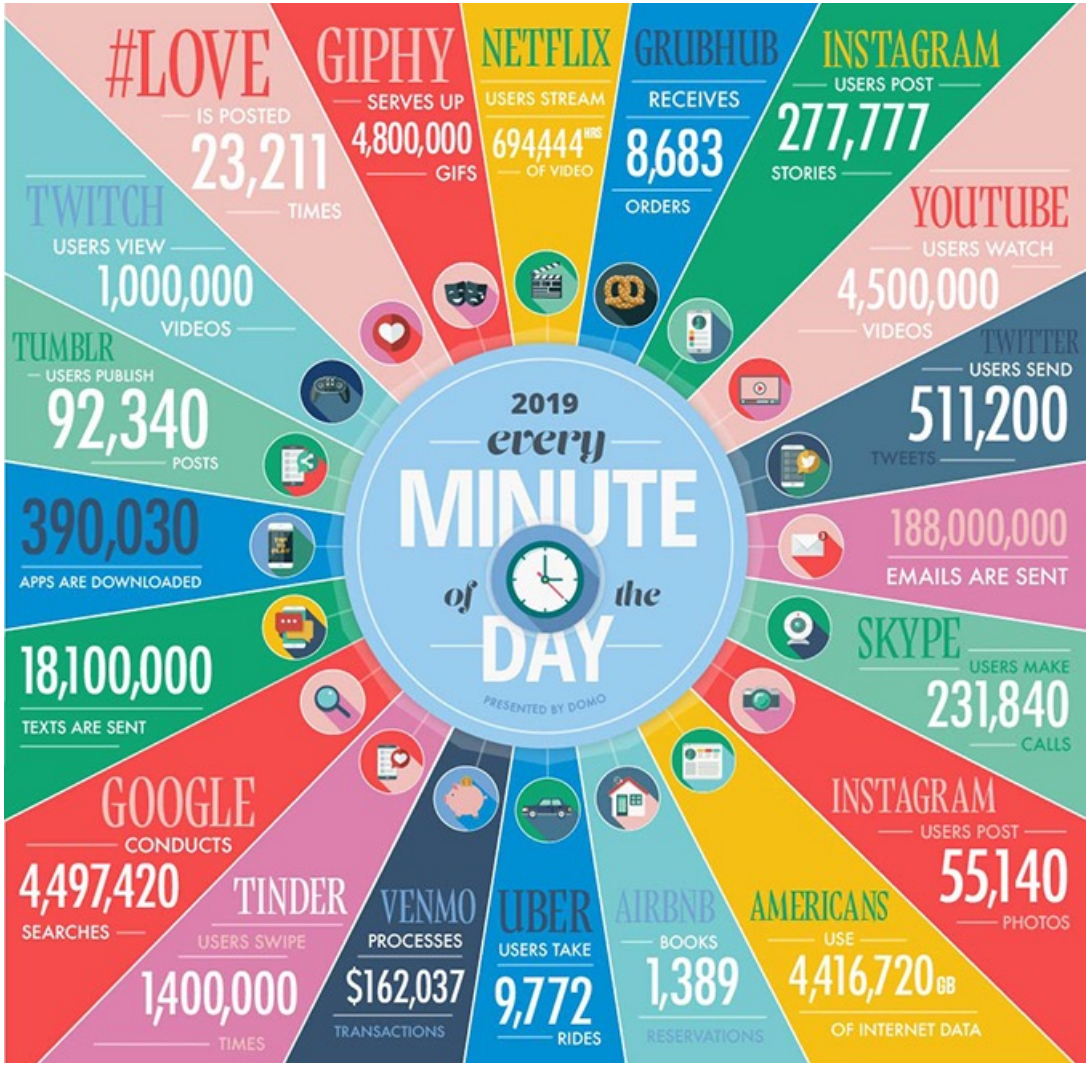
Suggested textbooks

Suggested textbooks

- **Foundations of Machine Learning**
  Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, 2012.
- **Learning from Data**
  Y. Abu-Mostafa, M. Magdon-Ismail, Hsuan-Tien Lin, 2012.
- **Pattern Recognition and Machine Learning**
  Christopher M. Bishop, 2007.

+ course note/slides

Big Data – Volume & Velocity



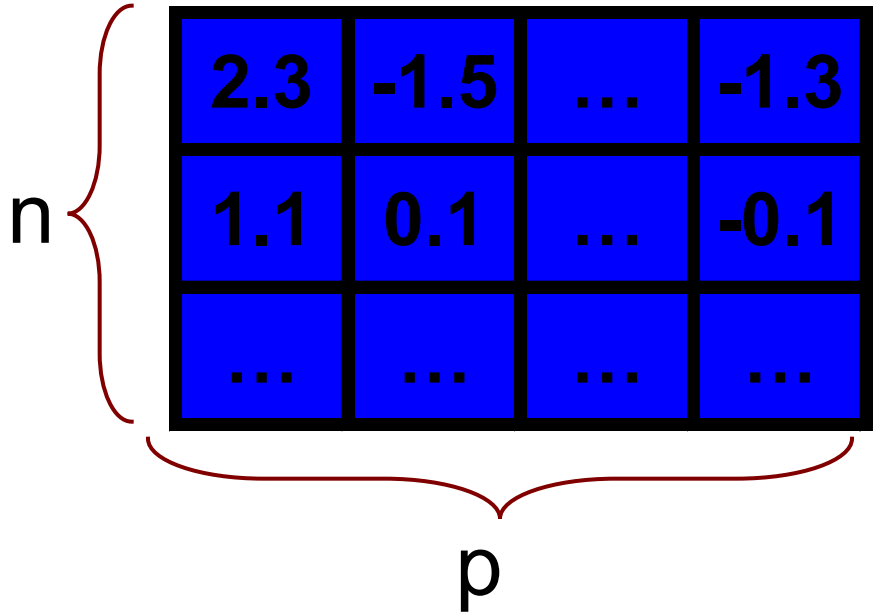
https://www.domo.com/learn/data-never-sleeps-7

Big Data – Variety & Value

Data is widely accepted as a resource of great value and arises in a great variety of formats:

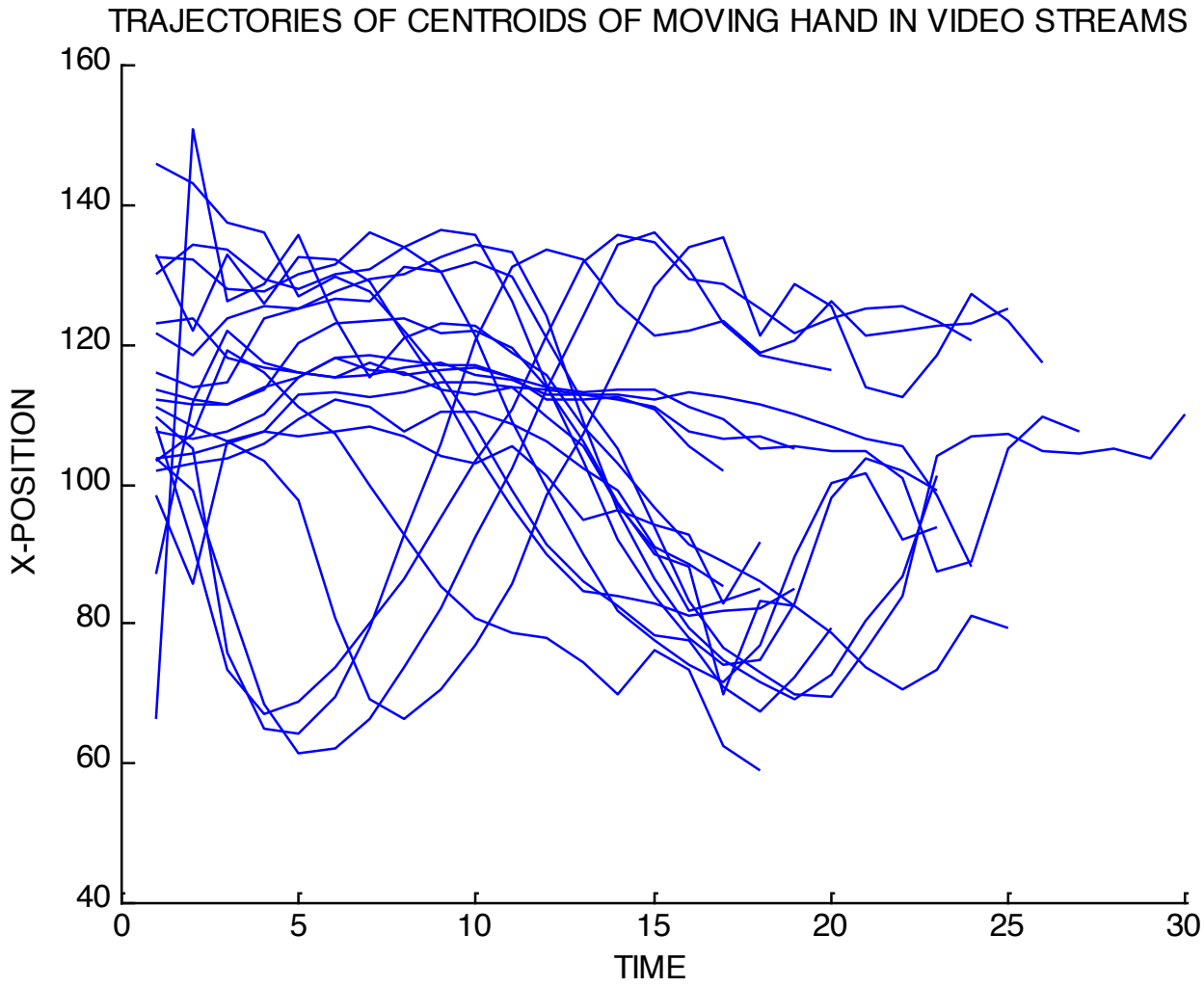
- Traditional: numerical, categorical, or binary
- Time Series data
- Text: emails, tweets, *New York Times* articles
- Geo-based location data
- Network, Sensor data
- Images, Video

Numeric Vector Data



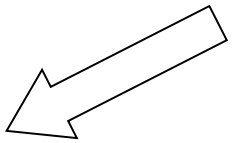
- Rows contain *data points*
- Columns contain the measurements of which a data point consists (*features* or *independent variables*)
- Both n and p can be very large in certain data mining applications

Time Series Data



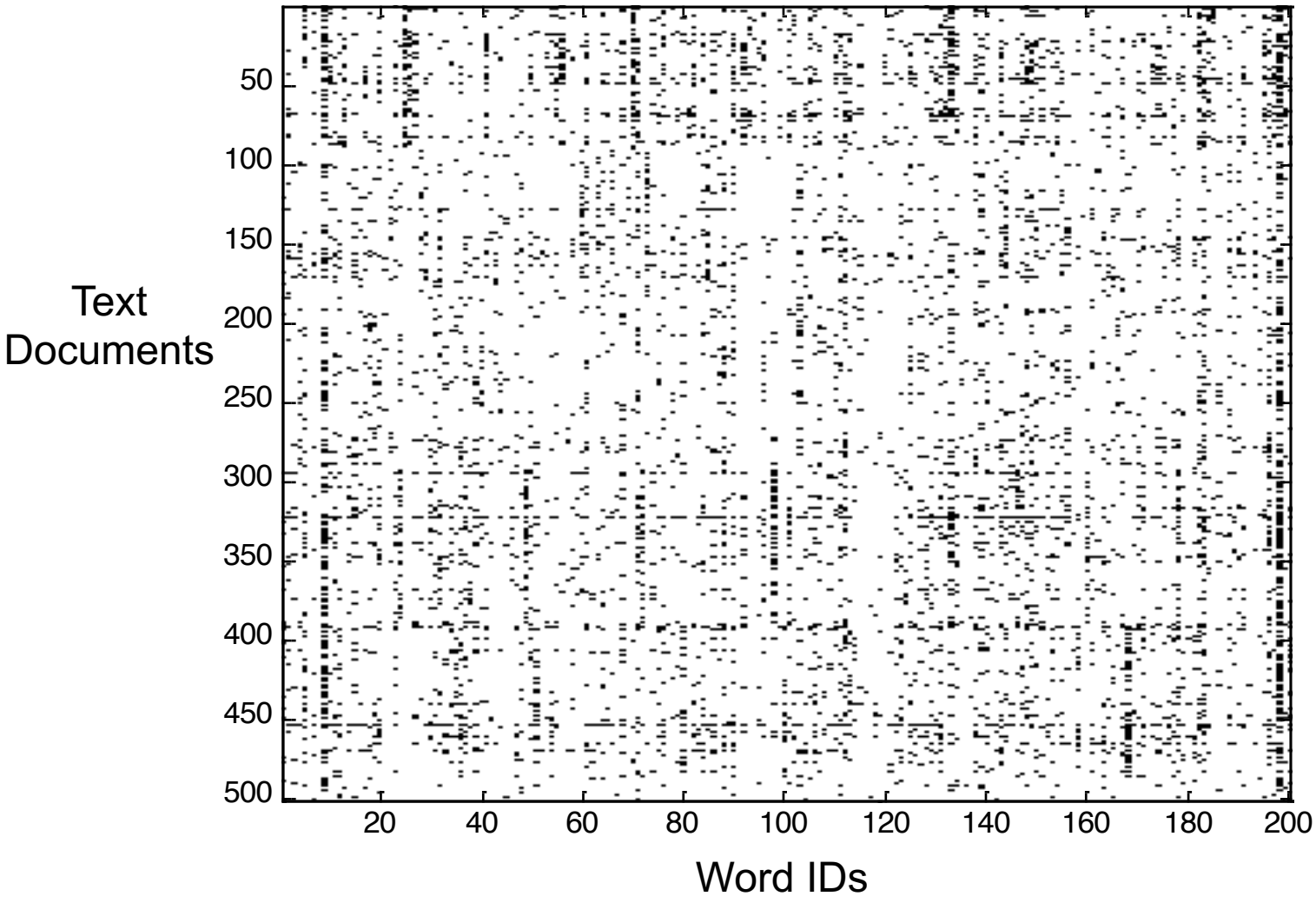
Sequence (Web) Data

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, -,

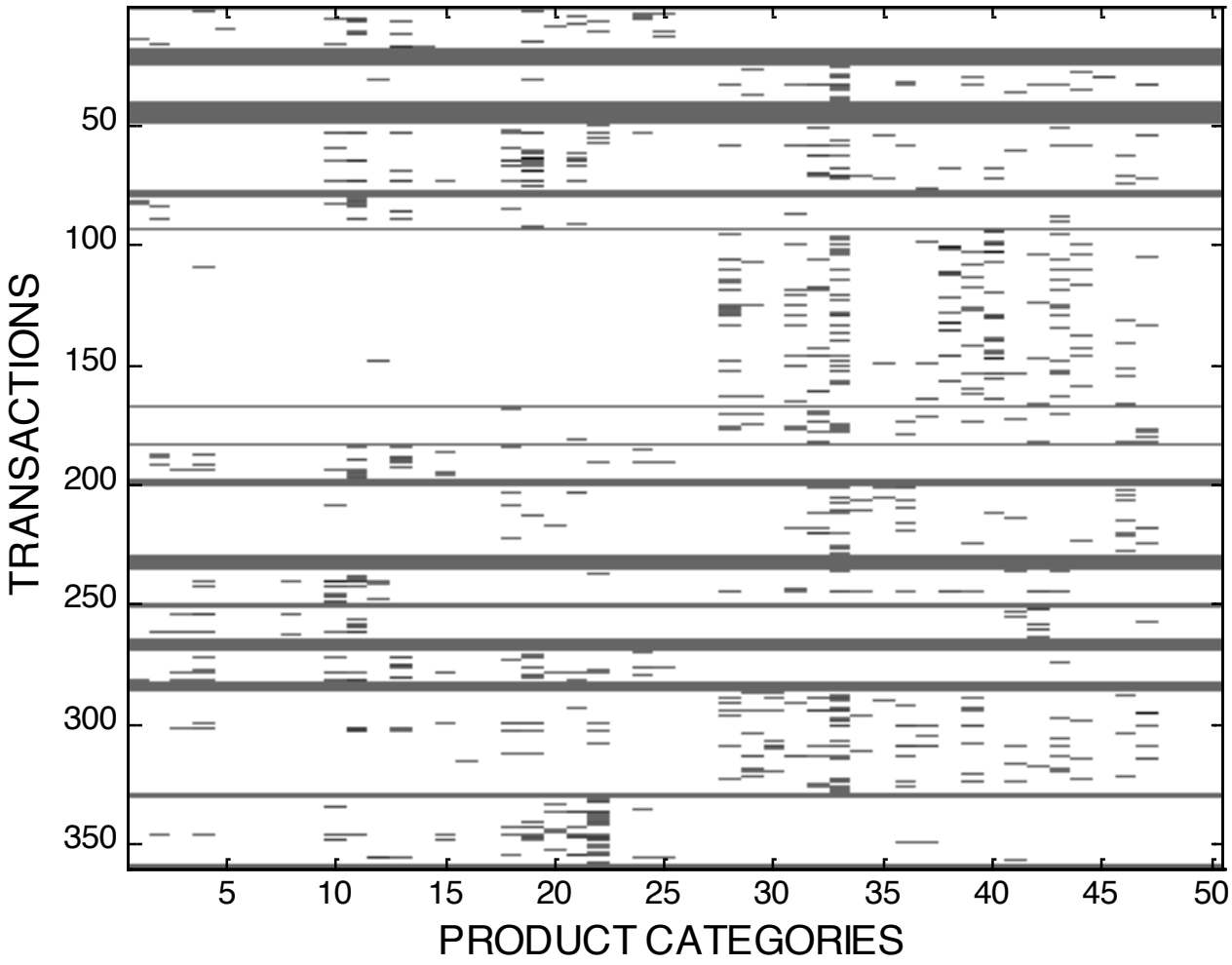


User 1	2 3 2 2 3 3 3 1 1 1 3 1 3 3 3 3
User 2	3 3 3 1 1 1
User 3	7 7 7 7 7 7 7 7
User 4	1 5 1 1 1 5 1 5 1 1 1 1 1 1
User 5	5 1 1 5
…	…

Sparse Matrix (Text) Data



“Market Basket” Data



Spatio-temporal data

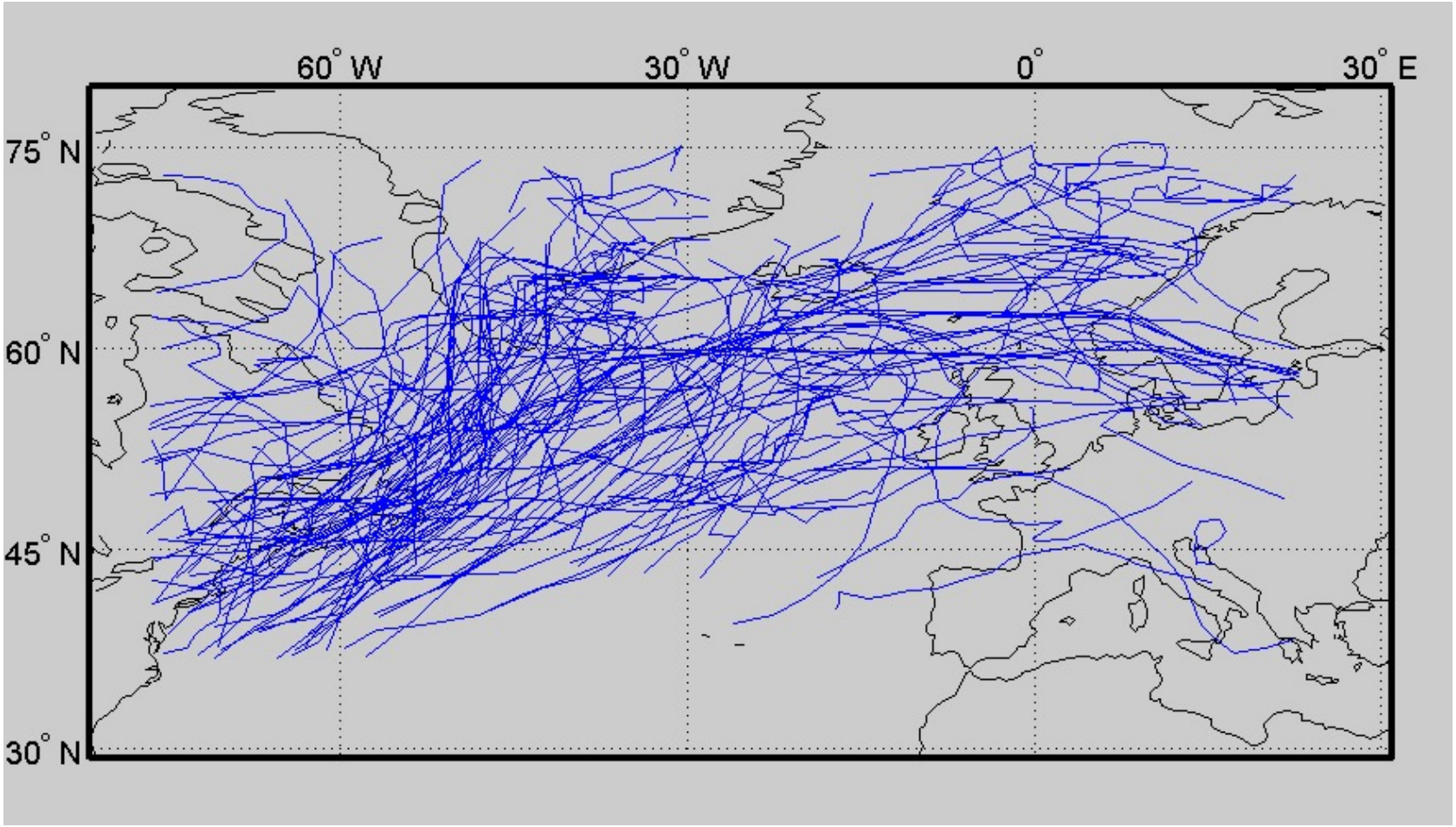
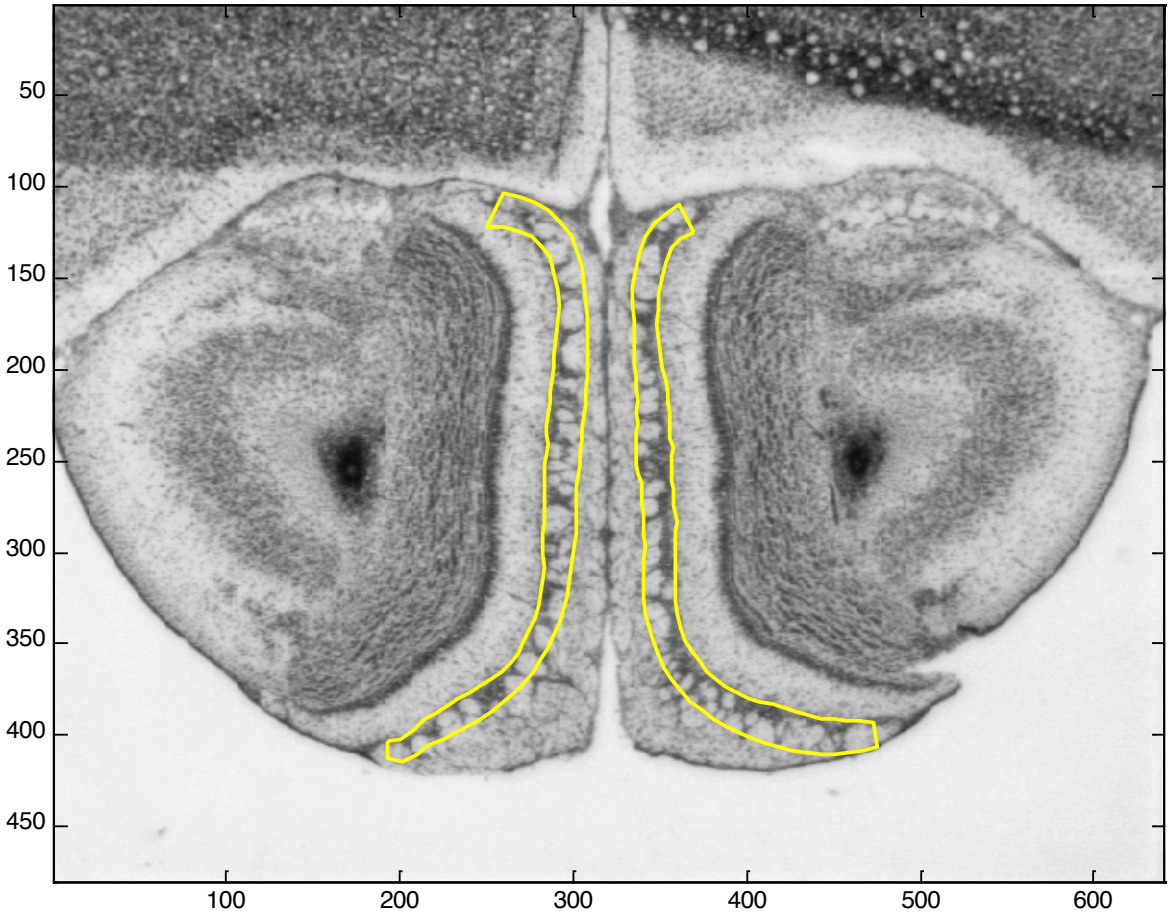
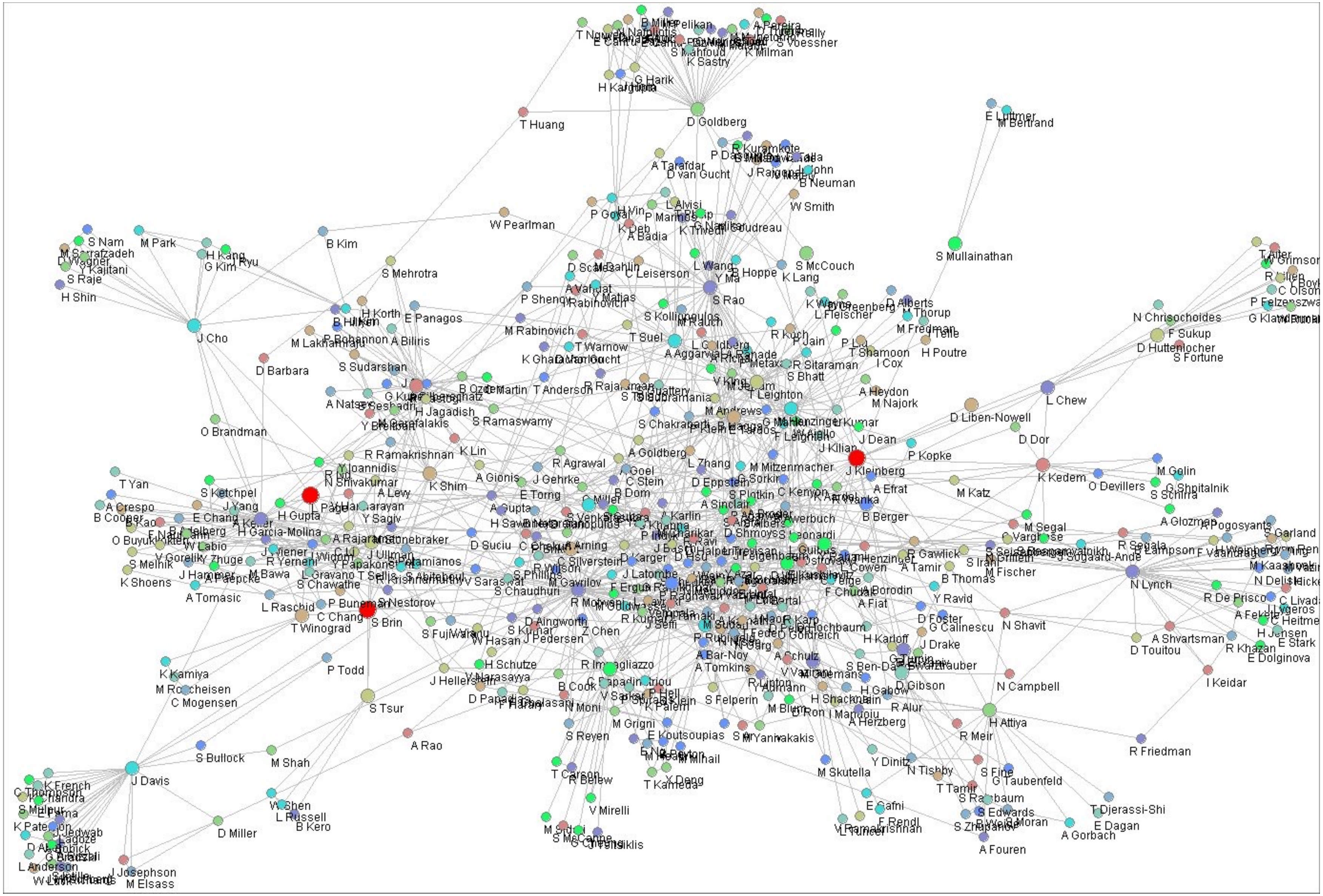


Image Data



Social Networks – Graphs

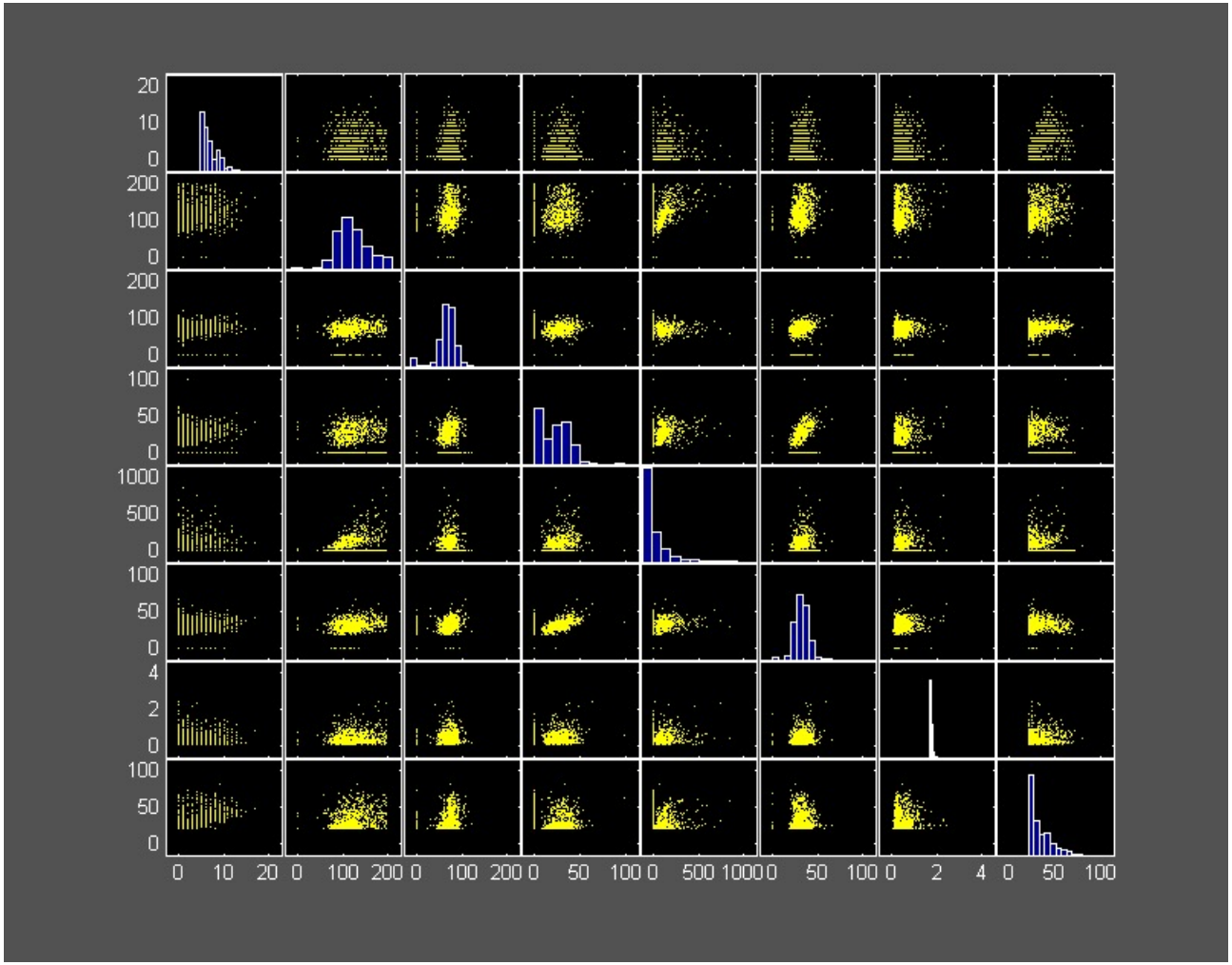


Exploratory Data Analysis

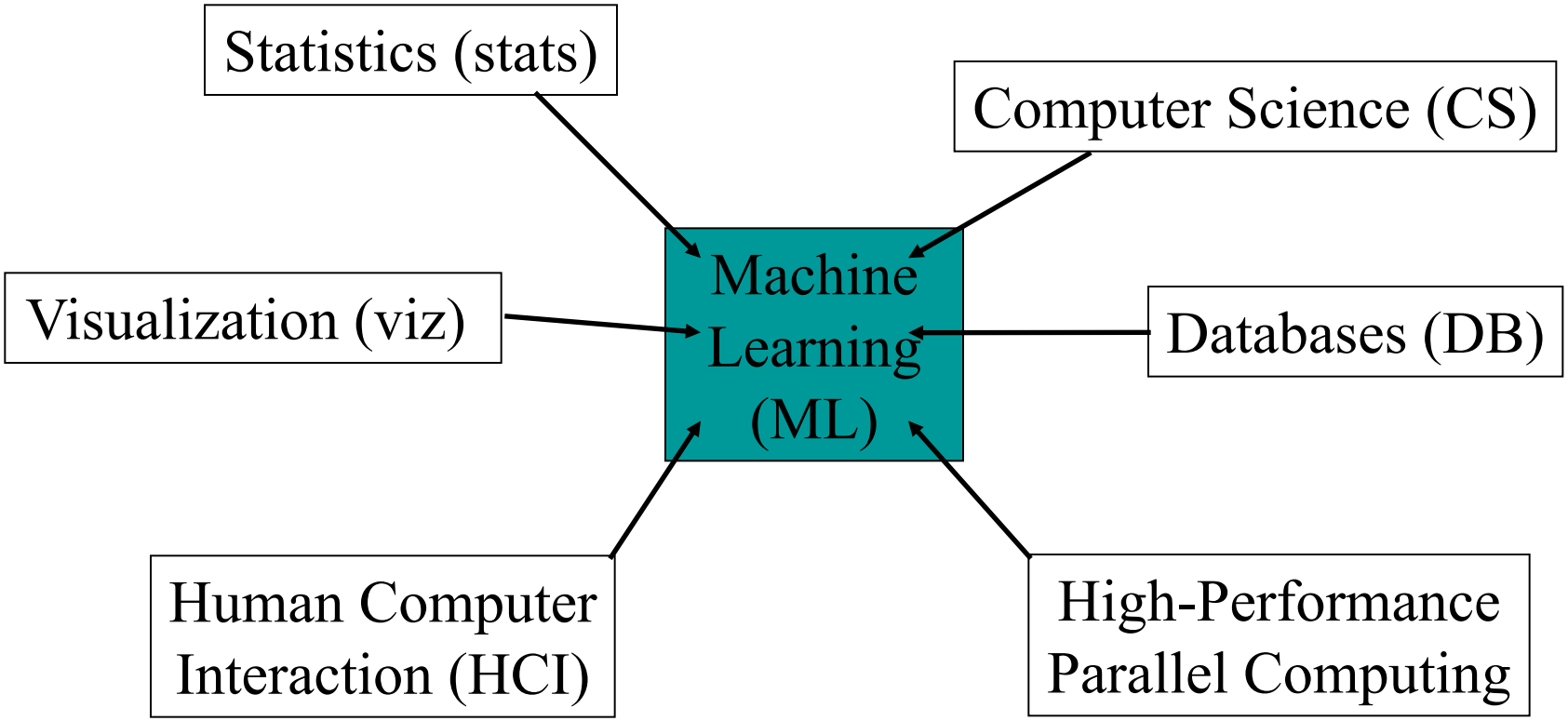
- Getting an overall sense of the data set
  - Computing summary statistics:
    - Number of distinct values, max, min, mean, median, variance, missing values...
- Visualization is widely used
  - 1d histograms
  - 2d scatter plots
  - Higher-dimensional methods
- Useful for data checking
  - Finding that some variables are highly skewed
- Simple exploratory analysis is extremely valuable
  - *You should always "look" at your data before applying any machine learning*

Example of Exploratory Data Analysis

(Pima Indians data, scatter plot matrix)



ML: Intersection of Many Topics



What is Machine learning

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.” Arthur Samuel (1959).

What is Machine learning

"Well-posed Learning Problem: A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on T, as measured by P, *improves with experience* E." Tom Mitchell (1998).

email spam prediction

- **Task**: email classification to spam/no-spam
- **Experience**: the user's action to characterize emails
- **Performance**: # of emails characterized as spam correctly.

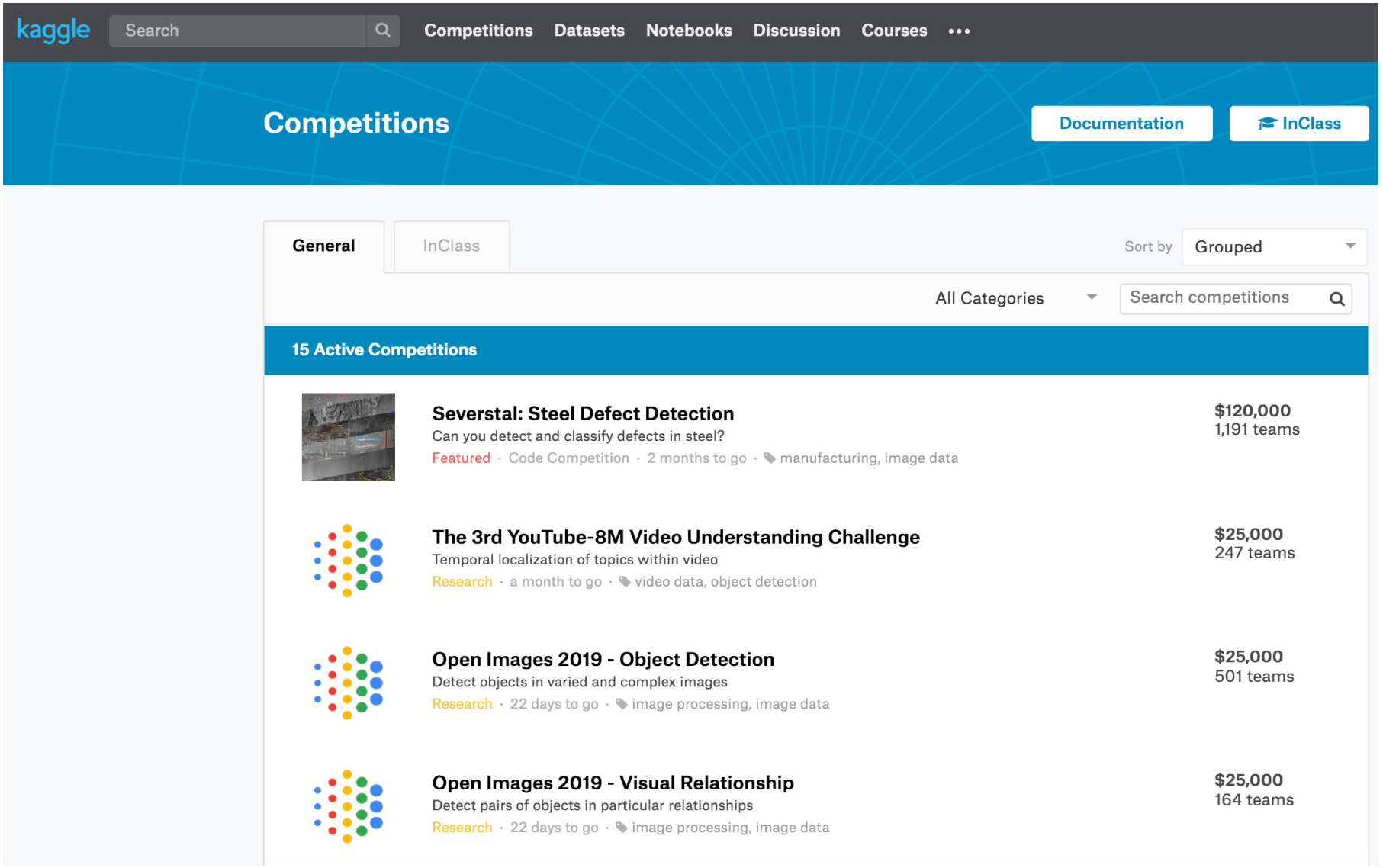
What is Machine Learning

- "…computational methods using experience to improve performance or to make accurate predictions" Mohri et. al. (2012)
- experience: past information available to the learner, in the form of electronic data collected and made available for analysis.
- Data quality and size are crucial to the success of the predictions made by the learner.
- Machine learning consists of
  - designing efficient & accurate prediction *algorithms* - time and space complexity.
- learning techniques are data-driven methods combining fundamental concepts in computer science with ideas from statistics, probability and optimization.

Applications of Machine/Deep Learning

- Text or document classification, e.g., spam detection;
- Natural language processing, e.g., morphological analysis, part-of-speech tagging, statistical parsing, named-entity recognition
- Recommendation systems, search engines, information extraction systems
- Fraud detection (credit card, telephone) and network intrusion
- Speech recognition, speech synthesis, speaker verification;
- Optical character recognition (OCR);
- Computational biology applications, e.g., protein function or structured prediction, Medical diagnosis;
- Computer vision tasks, e.g., image recognition, face detection;
- Games, e.g., chess, backgammon;
- Unassisted vehicle control (robots, navigation);
- …

Machine Learning for solving real problems



ML Terminology

Main distinction of *ML algorithms*:

- Supervised Learning – Labels y are known
- Unsupervised Learning –Labels y are unknown

The main *ML tasks* include:

- Classification (supervised) and Clustering (unsupervised), where group membership is inferred
- Regression (supervised), where a continous target variable is inferred.
- Anomaly detection, Text translation, link prediction...

ML Terminology

A *model M()* maps the data (input) *X* onto our target variable (output) *y,* i.e., *y′=M(X)*.

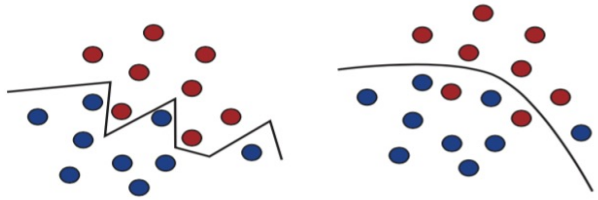
Typically our model has *parameters* θ, which are chosen by optimising a certain *loss function L (*e.g.*,* squared error loss: $(y-M(X))^2$*).* This loss is either optimised analytically (e.g., linear models) or via optimisation algorithms (e.g., Stochastic Gradient Descent in neural networks).

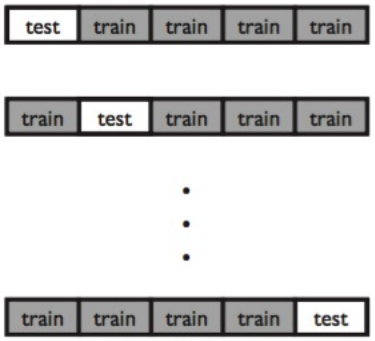
We split the data into:

- Training set (used to fit model parameters)
- Test set (used to assess model performance on data not available during training)

Overfitting vs generalisation

Machine Learning example





- Cross-validation: in many cases not enough training data.
  - Split the *m* data into *n* subsets(folds) and let θ the model parameters
  - Train the algorithm for *n-1* folds and test on the *n-th*
  - Compute the cross validation error
  - Choose parameters θ that

minimize the cv. error

$$\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{1}{m_i}\sum_{j=1}^{m_i} L(h_i(x_{ij}), y_{ij})}_{\text{error of } h_i \text{ on the } i\text{th fold}} .$$

Machine Learning – quiz..

Of the following examples, which would you address with *unsupervised/supervised* learning?

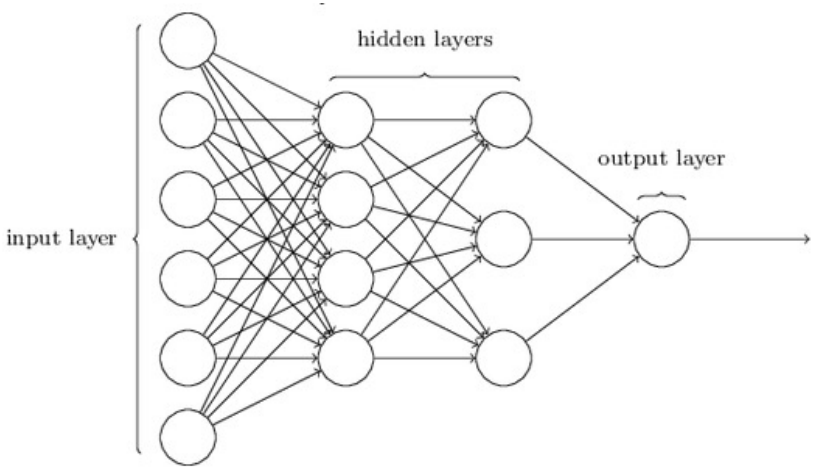
1. Given email labeled as spam/not spam, learn a spam filter.
2. Given a database of customer data, discover market segments and group customers into different market segments.
3. Given data of patients diagnosed with cancer, learn to classify new patients for this disease.
4. Given a set of news articles found on the web, group them into set of articles about the same story.

Machine Learning – quiz..

- Classification or regression?
  - Credit history -> offer a loan?
  - Human face picture -> {kid, adolescent, adult}
  - Human face picture -> age

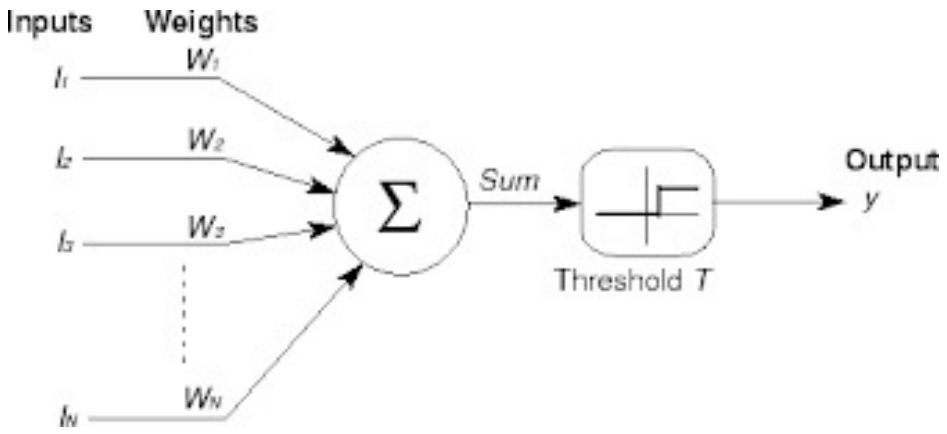
Deep Learning

- Based on perceptron – basic learning unit
- Layers of perceptrons learning a complex function *y=f(X)*



- *Learning features/embeddings*
- *Used for predictions*

Neural networks - perceptron



$$y = f(\sum_{j=1}^{n} w_j X_j + b)$$

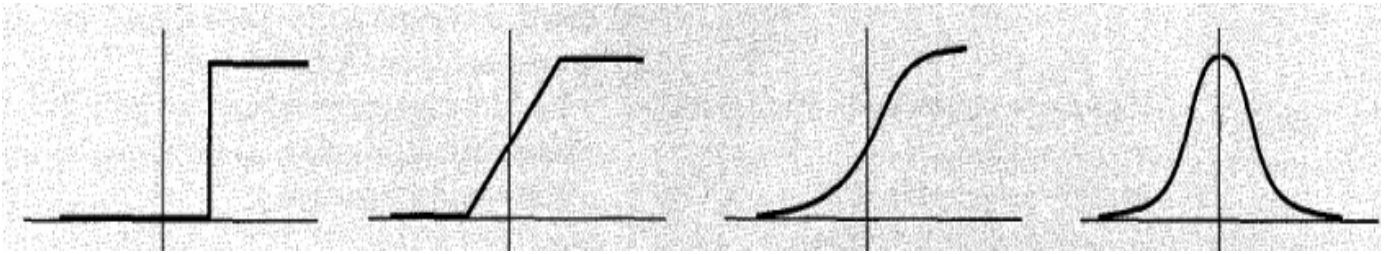
McCulloch-Pitts Neuron Model

-f: activation function

-w_j: weight of the j-th input Xj

-b : bias

- activation functions: piecewise linear, sigmoid, or Gaussian



Deep Learning

Dominant in recent years

Different architectures – non exhaustive

- Multilayer perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) – for sequential data
  - Recurrent Neural Nets (RNNs)
  - GRUs
  - LSTMs
- Attention based Architectures
  - Self Attention, Transformer (Bert), ...
- Autoencoders
- **Graph Neural Networks**