Data Science: introduction

Lionel Fillatre

fillatre@unice.fr

Polytech Nice Sophia

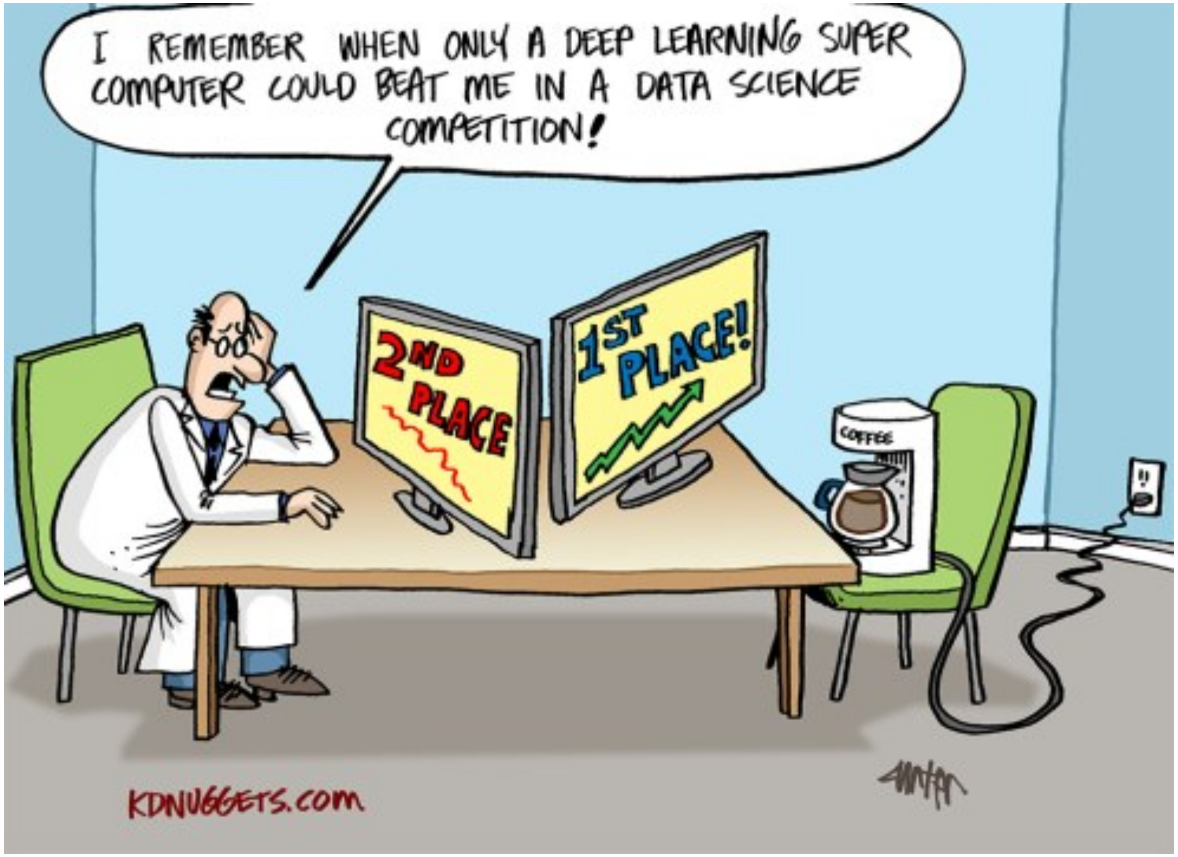
2019-2020

Outlines

- Introduction
- Course Logistics
- Kaggle
- 10 Best Practices in Data Science
- Programming tools
- Conclusion

Introduction

Data Science Automation



I remember when only a Deep Learning supercomputer could beat me in a Data Science competition

“Hard” Data Science Skills

python™

TensorFlow™

hadoop

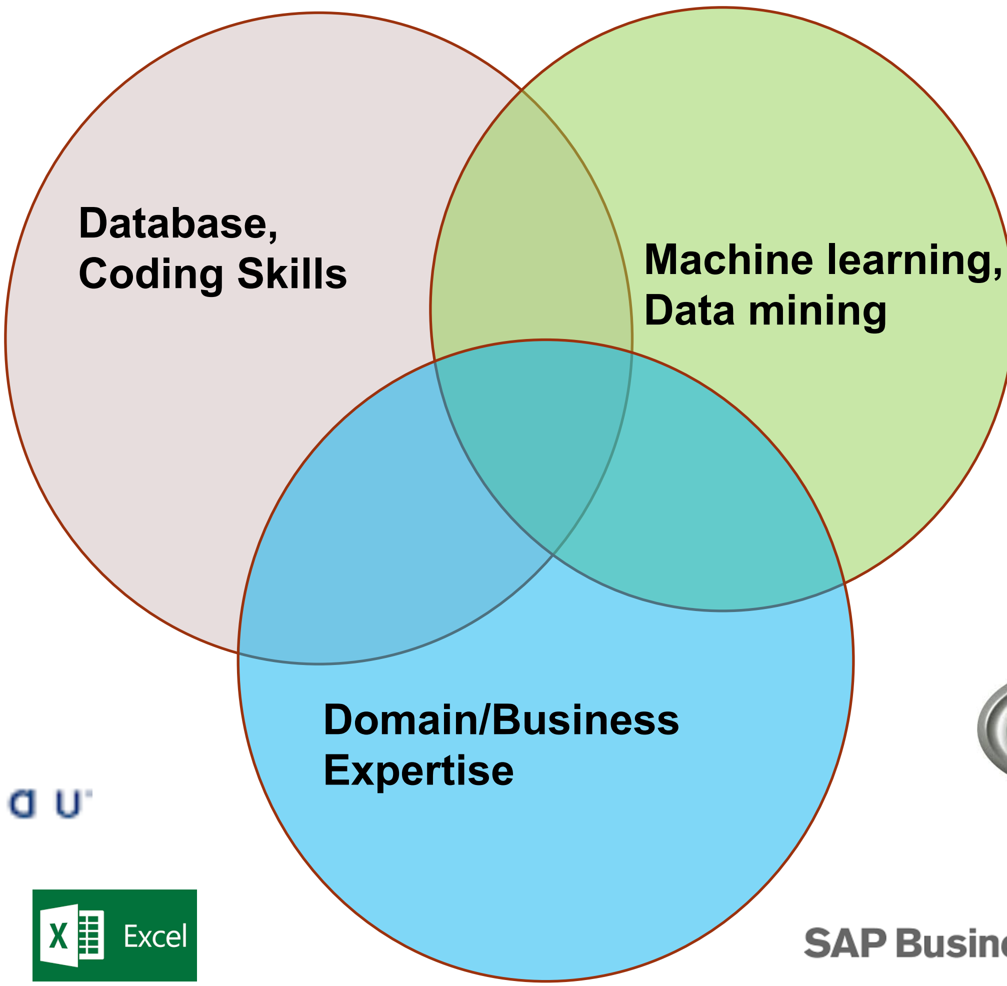
Math

Spark

scikit learn

NoSQL
Not Only SQL

SQL



R

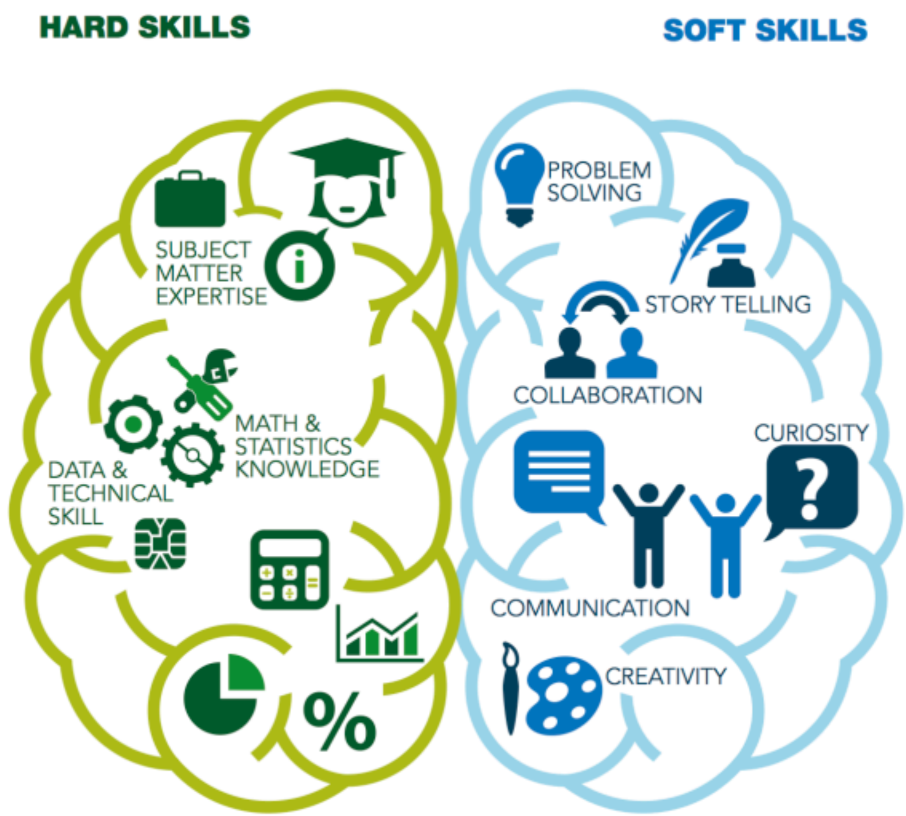
tableau

Excel

SAP BusinessObjects

"Soft" Data Science Skills: Harder to Automate

- Curiosity
- Intuition
- Business Knowledge
- Selecting a good metric
- Posing the right question
- Presentation Skills



Course Logistics

Course Output: What You Will Learn…

1. 20 September 2019: Initiation to Kaggle challenge (14H00-17h00)
2. 27 September 2019: INRIA (deep neural networks)
3. 4 October 2019: INRIA (deep neural networks)
4. 11 October 2019: INRIA (deep neural networks)
5. 18 October 2019 : IBM 1st seminar (13H30-15h30)
6. 25 October 2019 (to be confirmed): Tableau seminar (no exam)
7. 8 November 2019: IBM 2nd seminar (13H30-19h30)

All details will be given on the Moodle website!

https://lms.univ-cotedazur.fr/course/view.php?id=3548

Password: **zrp28KHY**

Grading (to be confirmed)

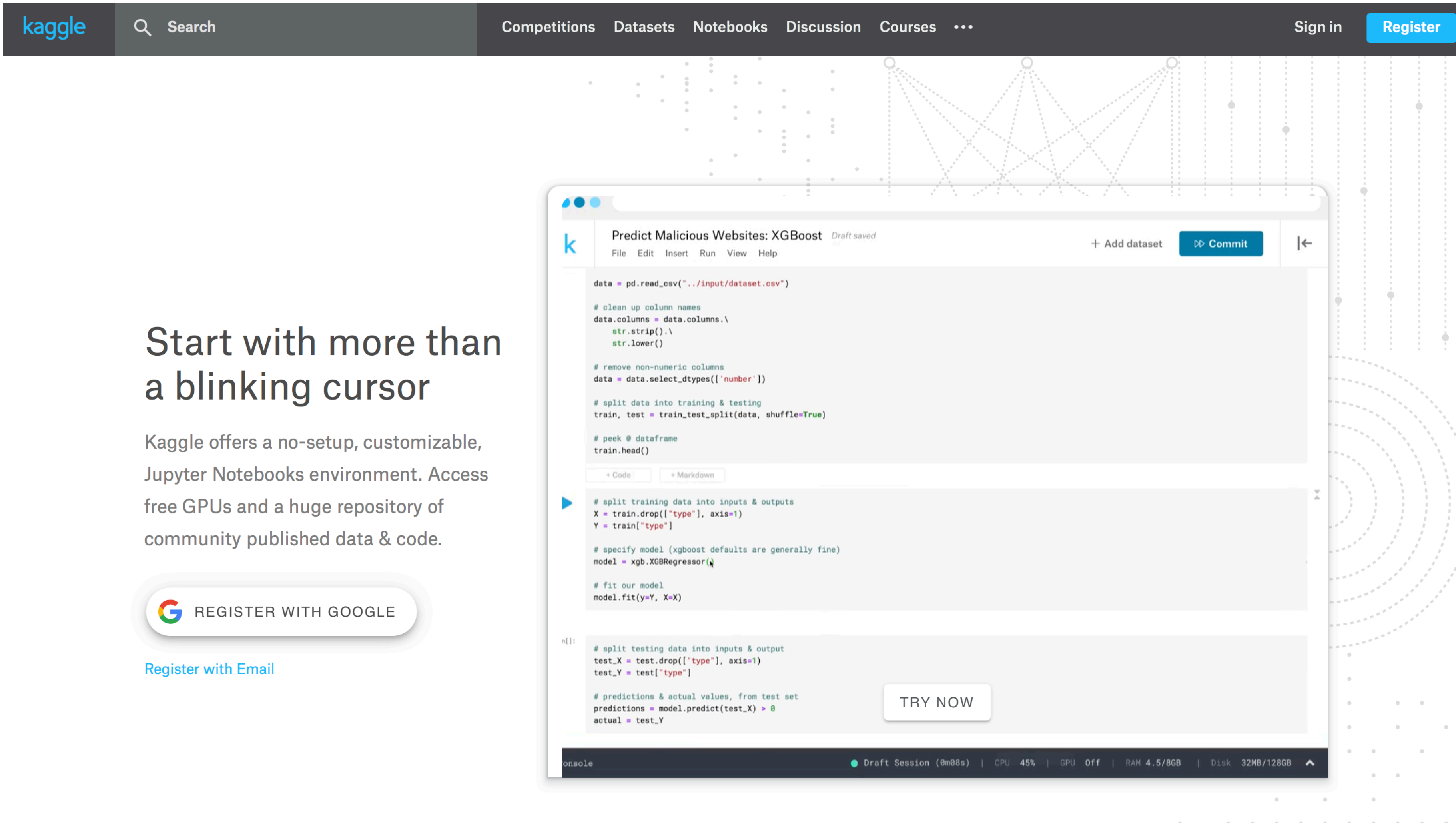
- Assiduity will be taken into account, especially for industrial lecturer (prepare your laptop and be on time!)
- First grade: Kaggle-like Challenge with INRIA
- Second grade: quizz on INRIA lectures
- Third grade: quizz on IBM lecture

Kaggle

Kaggle

- Kaggle is a platform for predictive modelling and analytics competitions
- Companies and researchers post their data
- Statisticians, data miners, data scientists (and others) from all over the world compete to produce the best models.
- Website: https://www.kaggle.com/

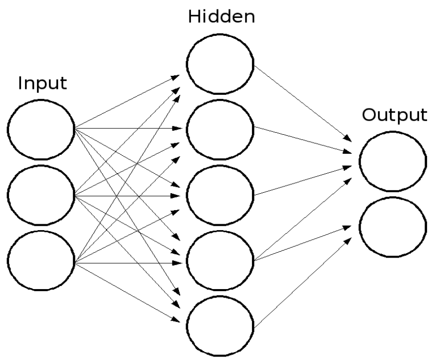
Kaggle website

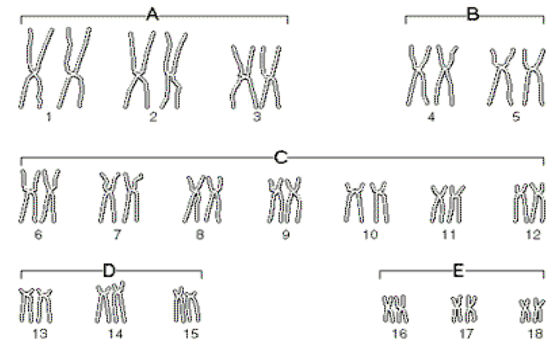


Kaggle website

- **Competitions**
  - The competition host prepares the data and a description of the problem
- **Datasets**
  - With or without competition
- **Kernels**
  - Kernels contain both the code needed for an analysis, and the analysis itself. It's the core of a work, what it needs to make it reproducible, to make it grow, and to invite collaboration.
- **Discussion**: forum of discussions
- **Jobs**: Hiring? Seeking?
- **Learn**: learn the basics to confidently start a new career or upgrade your skills.
- **Blog**: official blog of Kaggle.com
- **User rankings**: ranking of Kaggle users
- **Tags**: to find pages associated to a specific tags
- **Host a competition**: Kaggle can help you solve difficult problems, recruit strong teams, and amplify the power of the data science talent.

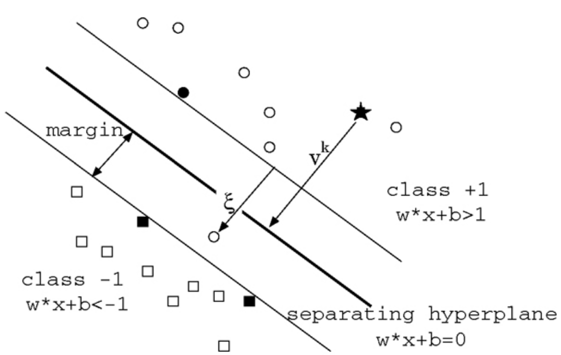
Many analytics methods

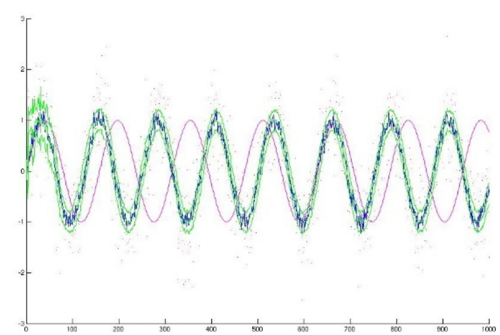




- Neural networks
- Logistic regression
- Support vector machine
- Decision trees
- Ensemble methods
- AdaBoost
- Bayesian networks

- Genetic algorithms
- Random forest
- Monte Carlo methods
- Principal component analysis
- Kalman filter
- Evolutionary fuzzy modeling





First Labs

- We will study the Kaggle Challenge « Titanic: Machine Learning from Disaster »

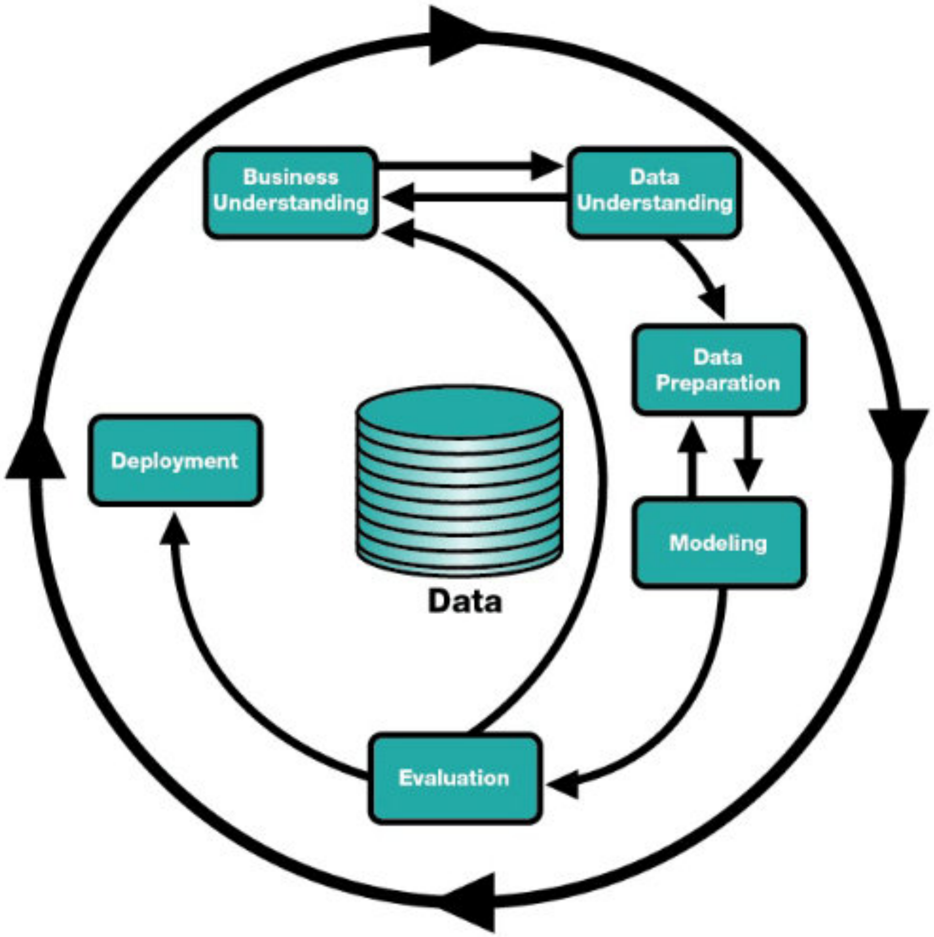
- More details on this challenge on https://www.kaggle.com/c/titanic

10 Best Practices in Data Science

Lesson 1: It is a Iterative, Circular Process

- Waterfall model does NOT work for Data Science

CRISP-DM: Iterative, Circular Process

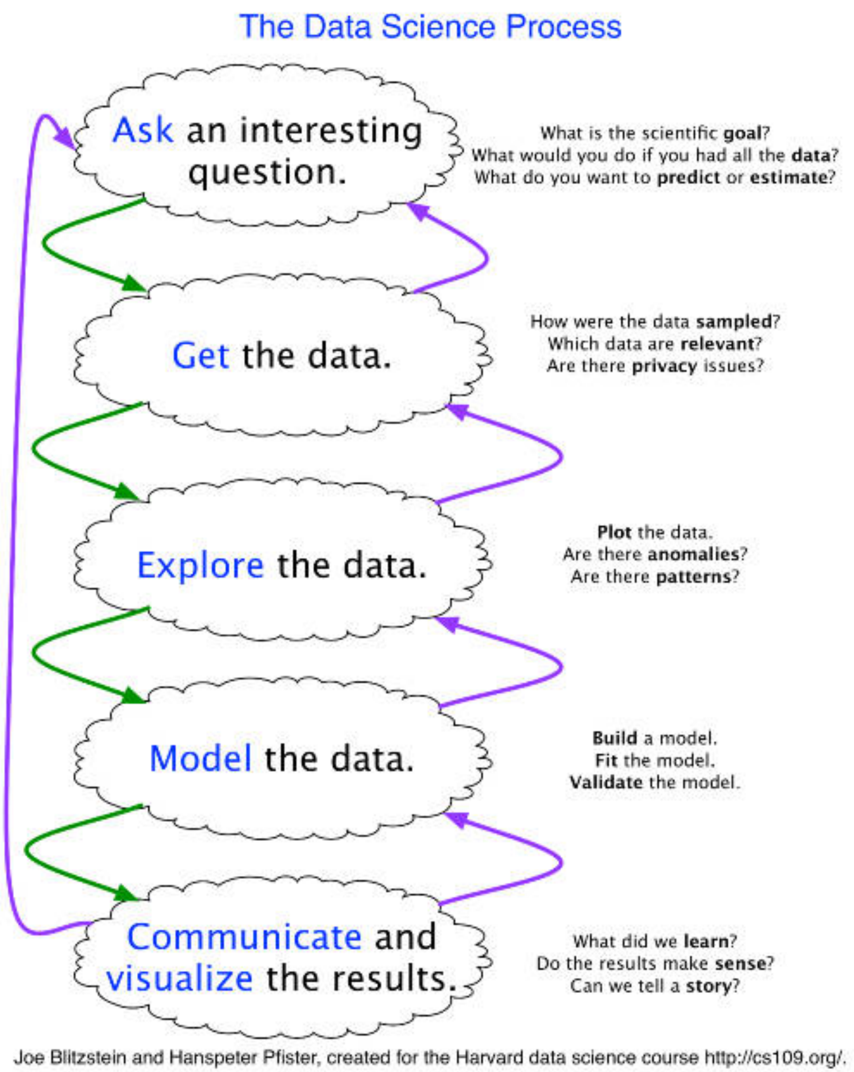


CRISP-DM
Cross Industry Standard Process
for Data Mining

CRISP-DM, 1998

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

Academic Data Science Process



Harvard, 2013

Lesson 2: Data Engineering Takes The Bulk of Time

- Building Machine Learning/Predicting Models is the key (and most fun) part, but only a small part of the whole process

- 60-80% spent on Data Preparation/Engineering

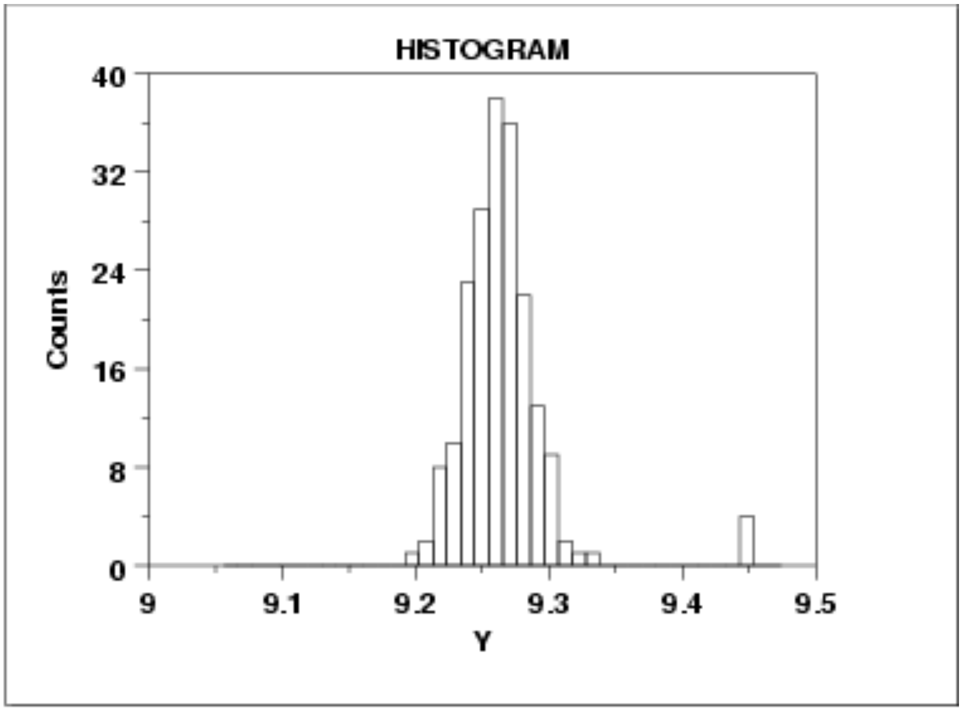
Competitions might be different

- How Kaggle winner spent time:
  - **35% read forums,**
  - 25% build models,
  - 25% evaluate results
  - 15% data preparation,

- See for example

http://blog.kaggle.com/2016/05/10/march-machine-learning-mania-2016-winners-interview-1st-place-miguel-alomar/

Lesson 3: Question Assumptions



- Problem:
  - Too many counts at 9.45

- Why?

Lesson 4: Focus on the Right Metric - Actionable

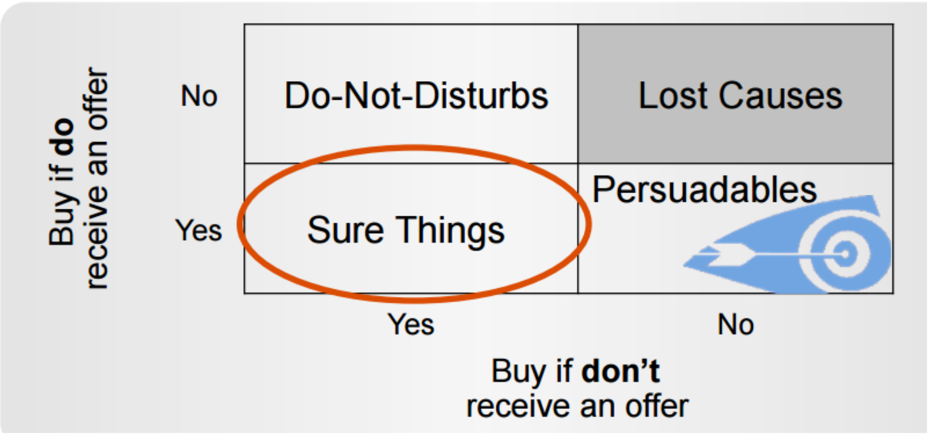
- Consumer:
  - Churn may depend on age, region, usage, and rate plan.
  - Rate plan easiest to change.

- Uplift Modeling in Marketing and Politics:
  - Focus on persuadables

Right Metric: Uplift Modeling

- Don't model if consumer will buy
- Model if consumer will buy **in response to an offer**

Response Uplift Modeling



From Eric Siegel presentation at PAW, 2011

- In Obama 2012 Campaign

www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory

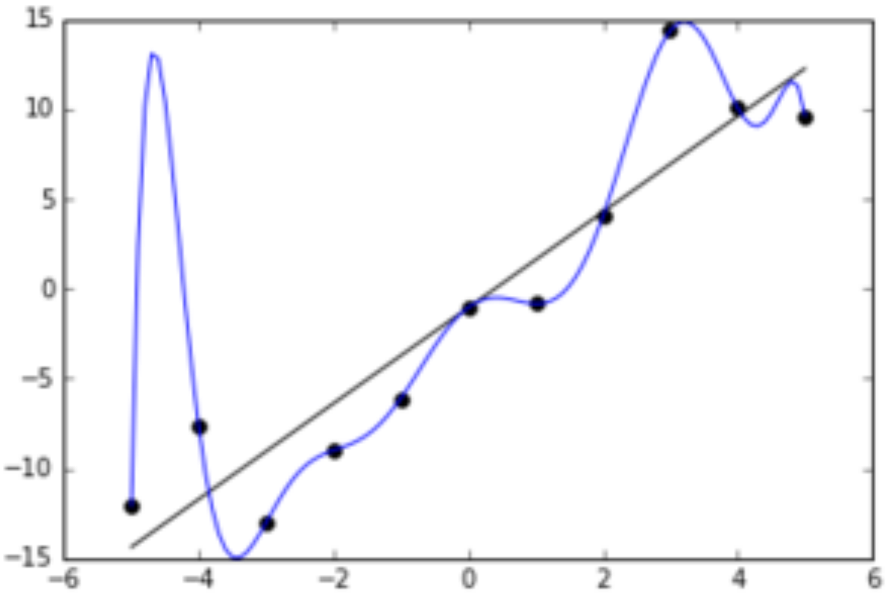
Lesson 5: Be a Fox, not a Hedgehog

A fox knows many things, but a hedgehog - one important thing.

Lesson 5: Modeling

- No Free Lunch Theorem – no method is universally the best (Wolpert)

Lesson 6: Avoid Overfitting



- Due to
  - Small samples
  - Testing too many hypotheses
  - Confirmation bias (explicit or implicit)
  - Poor training

http://www.kdnuggets.com/2014/06/cardinal-sin-data-mining-data-science.html

Lesson 7: Tell a story

- Combine facts into a story
- Combine visual and text presentation
- Explanation gives credibility
- Dynamic / Interactive

Lesson 8: Limits to Predicting Human Behavior?

- Inherent randomness, complexity in human behavior
- Individual predictions have limited accuracy (but can still be better than random and very useful for consumer analytics)
- Aggregate predictions (e.g., who will win the election?) more accurate, because individual randomness cancels out

Lesson 9: Deployment & Maintenance

- Netflix Prize winning algorithm not deployed

  > … the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment. Also, our focus on improving Netflix personalization had shifted to the next level by then.
  > http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

- In real-world, simpler is usually better

- Is model explainable ? (legal, acceptance reasons)

- Deployment Test and Monitor
  - Monitor assumptions
    - Do fields have the same value distributions?
  - Detect when model is no longer valid, needs rebuilding
  - Automatic model re-build

Lesson 10: Don’t just predict, optimize

- Prediction is usually just one part of making a decision
- Consider cost, frequency, latency, human behavior, etc
- Goal: Optimization
- From Data Science to Decision Science

Programming tools

Useful programming languages

- SQL (1970): querying and namaging data
- Python (1991): data processing, productivity, good learning curve
- R (1995): data analysis, oriented toward statistical analysis, more difficult to learn, free alternative to SAS, huge community
- And others: Java, Scala, SAS, Matlab, C/C++,…

Data Analysis Tools for Data Science

- MLLIB: MLlib is Apache Spark's scalable machine learning library.
  - logistic regression, linear support vector machine (SVM), classification, random forest, clustering via k-means, singular value decomposition (SVD), principal component analysis (PCA), linear regression with L1, L2, hypothesis testing

- MAHOUT: Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms
  - Collaborative Filtering, Matrix Factorization, Classification, Logistic Regression, Naive Bayes, Random Forest, Hidden Markov Models, Multilayer Perceptron, Clustering, k-Means Clustering, Spectral Clustering, Dimensionality Reduction, Singular Value Decomposition, PCA

- And many others: Rhadoop, H2O, Scikit-learn, Theano, Weka, LibSVM, etc.

Python Libraries for Data Science

- Many popular Python toolboxes/libraries:
  - NumPy
  - SciPy
  - Pandas
  - SciKit-Learn

- Visualization libraries
  - Matplotlib
  - Seaborn

- And many more…

Python Libraries for Data Science

NumPy:

- Introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects
- Provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance
- Many other python libraries are built on NumPy

Link: http://www.numpy.org/

Python Libraries for Data Science

SciPy:

- Collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more

- Built on NumPy

Link: https://www.scipy.org/scipylib/

Python Libraries for Data Science

Pandas:

- Adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)

- Provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

- Allows handling missing data

Link: http://pandas.pydata.org/

Python Libraries for Data Science

SciKit-Learn:

- Provides machine learning algorithms: classification, regression, clustering, model validation etc.

- Built on NumPy, SciPy and matplotlib

Link: http://scikit-learn.org/

Python Libraries for Data Science

matplotlib:

- Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- A set of functionalities similar to those of MATLAB
- Line plots, scatter plots, barcharts, histograms, pie charts etc.
- Relatively low-level; some effort needed to create advanced visualization

Link: https://matplotlib.org/

Python Libraries for Data Science

Seaborn:

- Based on matplotlib
- Provides high level interface for drawing attractive statistical graphics
- Similar (in style) to the popular ggplot2 library in R

Link: https://seaborn.pydata.org/

Python Libraries for Deep Learning

TensorFlow, Keras, Pytorch:

- Provides mid-level interface and high level interface for designing and visualizing deep neural networks

- Can exploit GPU (Graphics Processing Unit) cards automatically for high performance computing

Links:
https://keras.io
https://www.tensorflow.org
https://pytorch.org

Conclusion

Conclusion

- Keep an eye on the Moodle website, especially before the IBM lectures!

Homework to prepare the labs with INRIA

- Need a laptop with power cable

- Create a gmail account to be able to use https://colab.research.google.com/
  - Colaboratory is a Google research project created to help disseminate machine learning education and research.
  - It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.
  - It is possible to exploit a GPU