

Ch0. Introduction and Overview

ST4240, 2016/2017

Alexandre Thiéry

Department of Statistics and Applied Probability

Outline

1 Practicalities

2 Ressources

3 Some applications

4 In this course...










- Class: Mondays (10-12) + Thursdays (10-12)
- Class = Lecture + **Class Exercises** + Tutorials
- Class exercises: participate!
- Webcast available: if you prefer to sleep, stay at home!
- I answer emails pretty quickly!

- **Graded Homework** Assignments. (15%)
- No late assignment will be accepted ([online submission](#))
- Only **typed** homework will be considered.
- **Midterm** (25%)
- **Final Exam** (60 %)

Assignments groups

- Assignments: group of maximum 5 students per groups
- Same group for the whole semester.
- 1 group = 1 mark (i.e. choose wisely)

Data Mining Prediction Assignment

	 8	Random Forest Benchmark	0.86141		
45	 8	dickoa	0.86141	1	Tue, 22 May 2012 12:07:36
45	 8	Rohit	0.86141	2	Fri, 25 May 2012 21:00:14
45	 8	squawkboxed	0.86141	1	Fri, 08 Jun 2012 14:57:28
45	new	BLetson	0.86141	3	Fri, 29 Jun 2012 14:49:38
50	 9	testing	0.86135	4	Sat, 16 Jun 2012 05:18:44 (-26.1h)
51	 9	schappi	0.86130	7	Sat, 16 Jun 2012 12:53:13
52	 8	Greg Park	0.86116	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
53	 8	Glen	0.86111	35	Tue, 05 Jun 2012 23:44:06 (-3.3d)

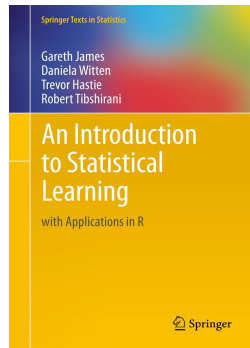
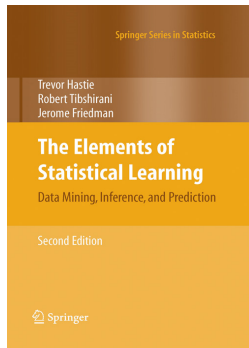
- 5-10 min pause at 11am: come **ask questions!**
- Scans of exercises corrected in classes will be made available.
- There will be loads of class exercises: bring notebook / rough paper/ etc...
- Release time \approx 11 : 30
- Come **ask questions** after the class!

- Attending ST4240 without implementing examples is **useless**.
- We will be using the *R* programming language
- Tools Recommendation:
 - **R-studio**: R-environment
 - **R-markdown**: reproducible research
 - **ggplot2**: plotting package
 - **dplyr**: data manipulation package
- you are expected to learn yourself how to use these packages: Google is your friend...
- I will be posting source-codes ...
- **Kaggle.com**: Data Science competition, discussion, datasets, job advertisement, ...

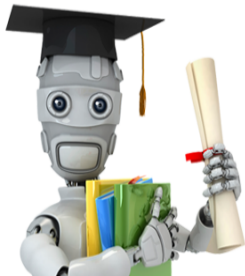


Outline

- 1 Practicalities
- 2 Ressources
- 3 Some applications
- 4 In this course...



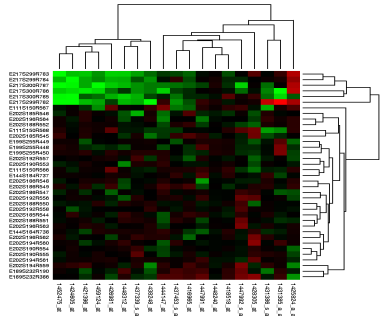
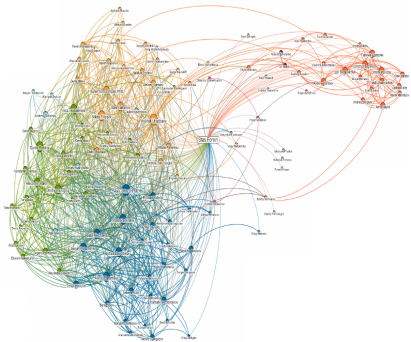
Machine Learning online lecture by Andrew Ng



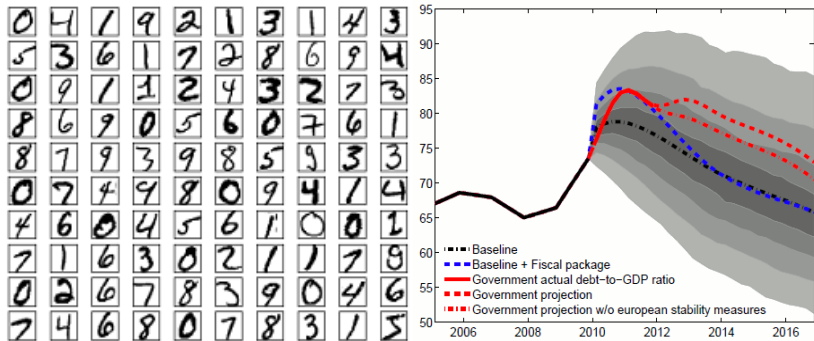
Outline

- 1 Practicalities
- 2 Ressources
- 3 Some applications**
- 4 In this course...

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.



- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy of those models and procedures when applied to fresh data.



Avalanche of data

- Financial transactions : billions of transactions per year
- Analysis of internet traffic data
- Human Genome Project has to deal: gigabytes
- remote-sensing satellite systems:gigabytes per hour
- U.S. census file: $\geq 10^{12}$ bytes



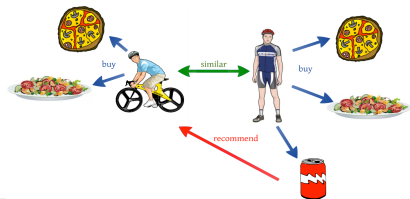
- **Marketing:** Predict new purchasing trends. Given customers who have purchased product A, B, or C, identify those who are likely to purchase product D.

Grant, Welcome to Your Amazon.com (if you're not Grant Ingersoll, click here.)

Today's Recommendations For You

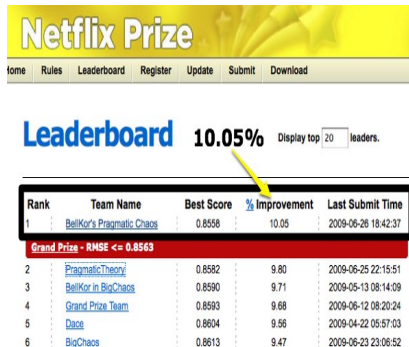
Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).

Book Title	Author	Rating	Price
Principles of Data Mining (A...	by David J....	★★★★☆ (17)	\$52.00
Python in a Nutshell, Secon...	by Alex Mart...	★★★★☆ (40)	\$26.39
Introductory Statistics wit...	by Peter Dai...	★★★★☆ (20)	\$48.56



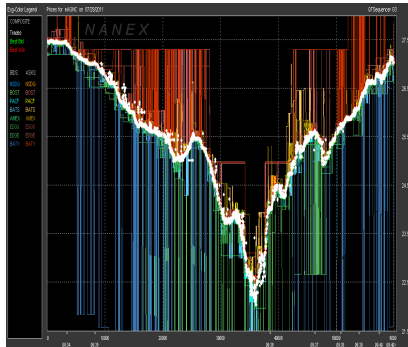
■ Netflix Challenge:

- 10^7 ratings for 17700 movies
- Goal: film recommendation
- One million dollar prize for a 10% improvement

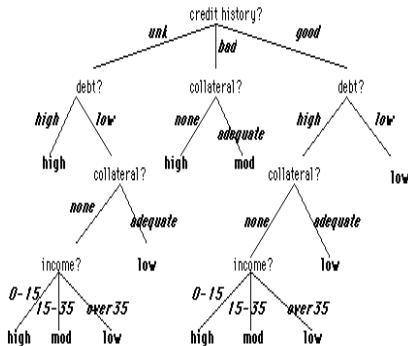


Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

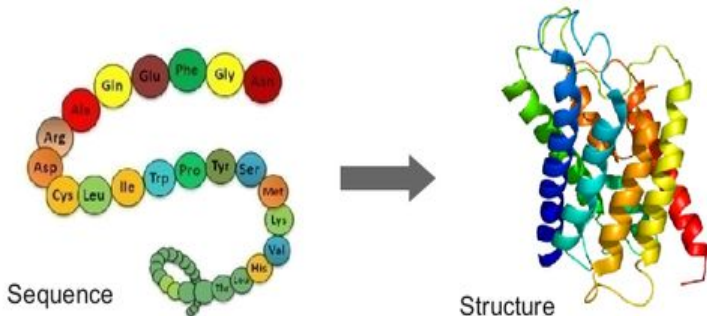
- **Financial Markets:** Analyse volatility patterns in high-frequency stock transactions using volume, price, and time of each transaction.



- **Insurance/credit scoring:** Identify characteristics of buyers of new policies. Find unusual claim patterns. Find "risky" customers.



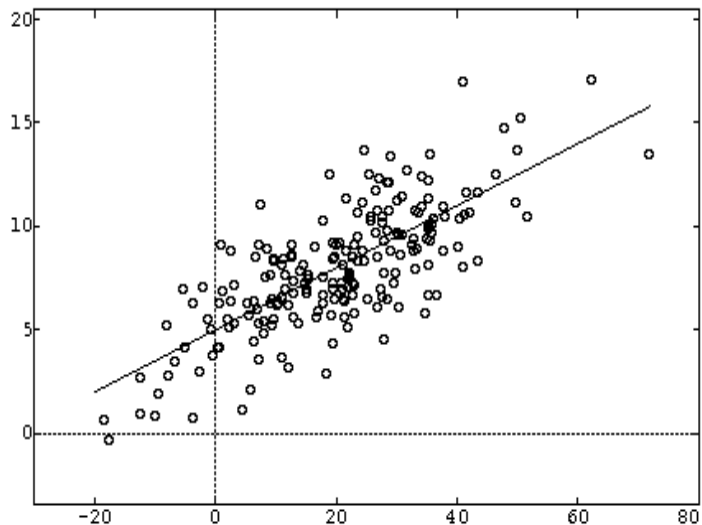
- **Molecular Biology:** Analyse amino acid sequences and DNA microarrays. Predict protein structure and identify related proteins.



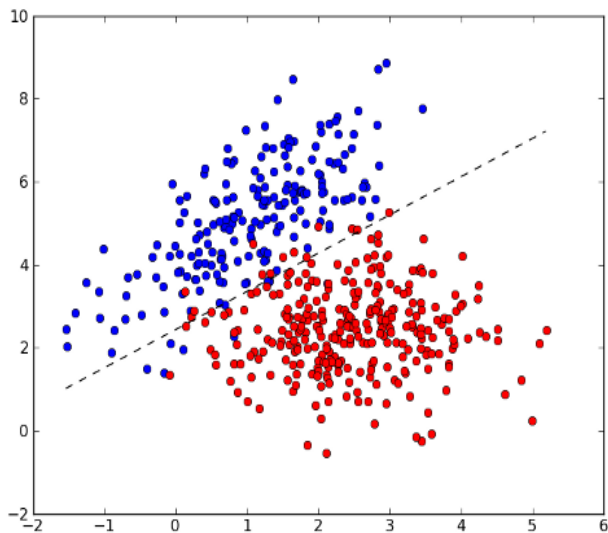
Outline

- 1 Practicalities
- 2 Ressources
- 3 Some applications
- 4 In this course...**

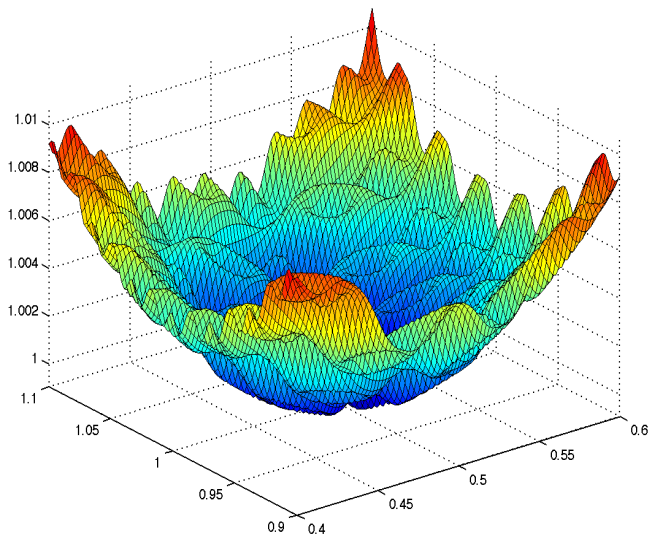
Regression



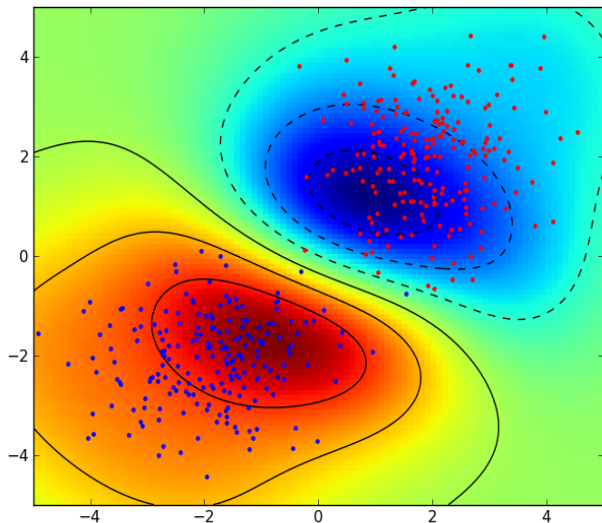
Classification



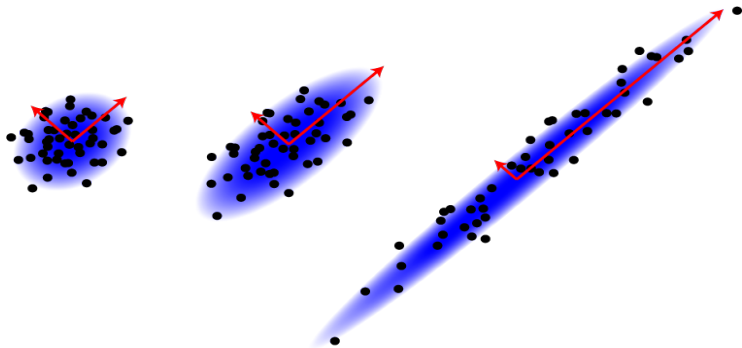
Optimization



Support Vector Machine (SVM) and Kernel Methods



Dimension reduction



Tree methods, Random Forest

