Machine Learning

Introduction

Anisio Lacerda

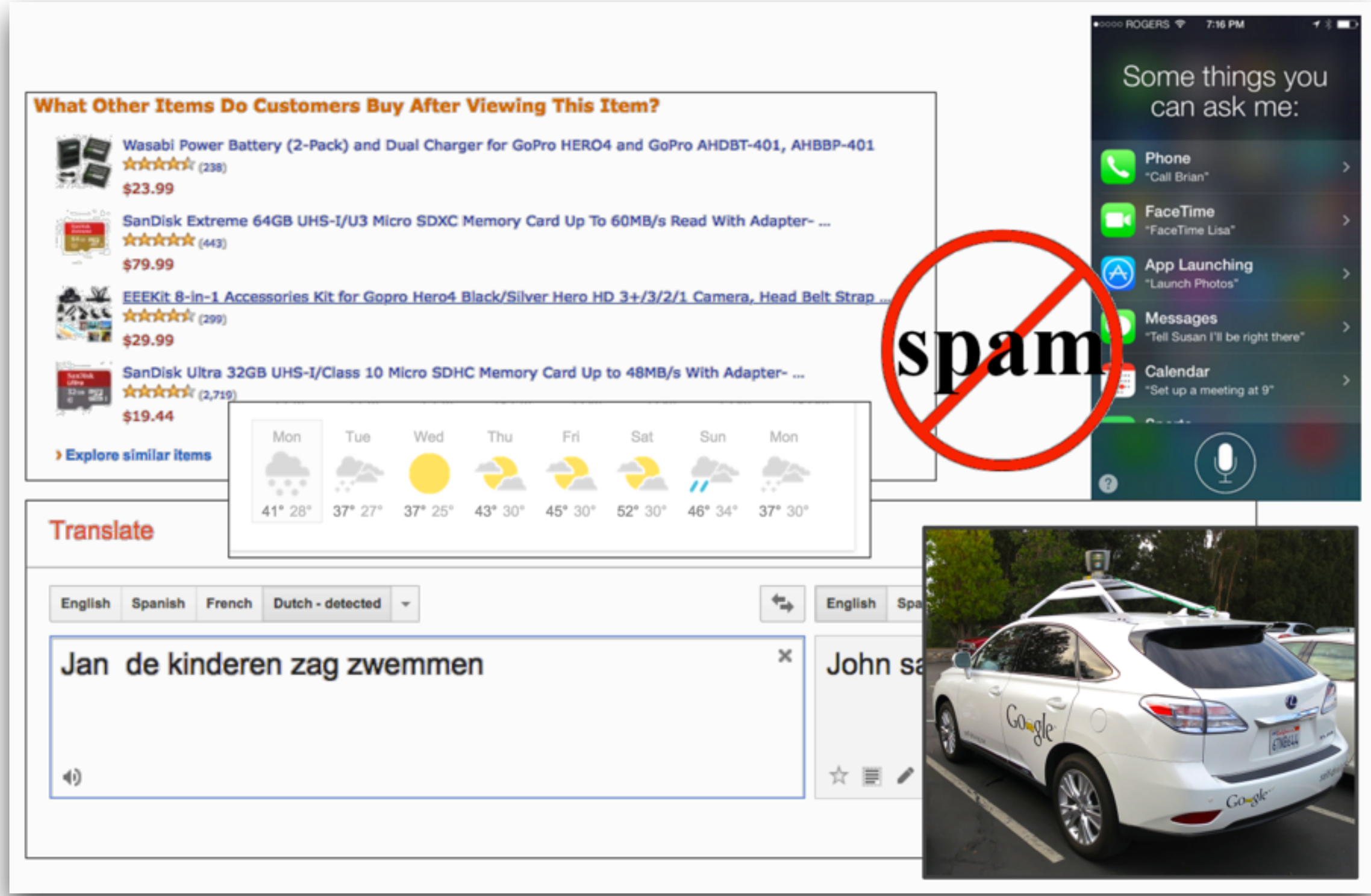
Learning from data

- Many applications require gaining insights from massive, noisy datasets
- **Science**
  - Physics (LHC, …), Neuroscience (fMRI, …), Geology (sensor arrays, …)
  - Social science, economics
- **Commercial / civil applications**
  - Consumer data (online advertising, viral marketing, …)
  - Health records (evidence based medicine, …)
- **Security / defense related applications**
  - Spam filtering / intrusion detection
  - Surveillance, …

Web-scale machine learning

- Predict relevance of search results from click data
- Personalization
- Online advertising
- Machine translation
- Spam filtering
- Fraud detection
- ….

Machine Learning is everywhere



So what is Machine Learning?

- Automation automation automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!

Traditional Programming

Machine Learning

Magic?

- No, more like gardening

seeds	algorithms
nutrients	data
gardener	you
plants	programs

What is Machine Learning?

"Field of study that gives computers the ability to learn without being explicitly programmed"

–Arthur Samuel (1959)

Learning as generalization

"Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task (or tasks drawn from the same population) more effectively the next time"

–Herbert Simon (1983)

Learning as generalization

"A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks **T**, as measured by **P**, improves with experience **E**"

–Tom Mitchell (1999)

Related fields

- The **artificial intelligence** dream: Computers that are as intelligent as humans
  - Machine learning closely tied to AI
- **Theoretical CS and mathematics**
  - Formalizing and understanding learning mathematically
  - Uses ideas from probability and statistics, linear algebra, theory of computation
- **Phylosophy, congnitive psychology, neuroscience, liguistics, robotics, …**
- Many, many application areas
  - **Medicine**, **engineering**, other areas of CS like **compilers, psychology, marketing, …**

Why Study Machine Learning: A Few Quotes

- "A breakthrough in machine learning would be worth ten Microsofts" (Bill Gates, Microsoft)
- "Machine learning is the next Internet" (Tony Tether, Former Director, DARPA)
- "Machine learning is the hot new thing" (John Hennessy, President, Stanford)
- "Web rankings today are mostly a matter of machine learning" (Prabhakar Raghavan, Dir. Research, Yahoo)
- "Machine learning is going to result in a real revolution" (Greg Papadopoulos, CTO, Sun)

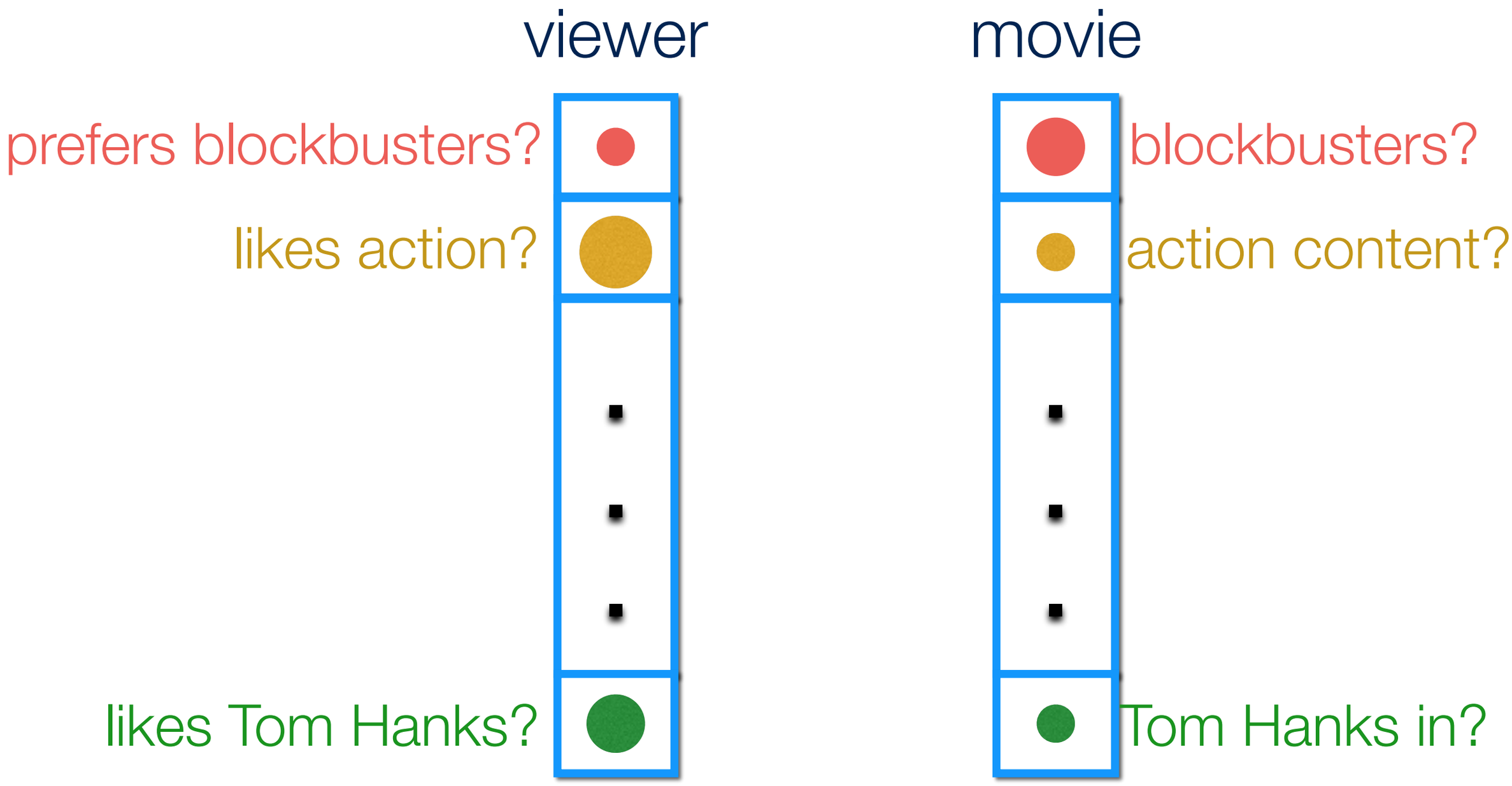
Overview of the course

Example: Predicting how a viewer will rate a movie

- Netflix Prize:
  - 10% improvement = 1 million dollar prize

- The essence of machine learning:
  - A **pattern** exists
  - We **cannot** pin it down **mathematically**
  - We have **data** on it

Movie rating - a solution



add contributions from each factor = predicted rating

Example: Credit approval

Attribute	Value
age	23 years
gender	male
annual salary	$30,000
years in job	1 year
current debt	$15,000
...	...

Approve credit?

Topics

- Defining models
- Different learning protocols
- Learning algorithms
- Representing data
- Evaluation

We will see different “models”

- Or: what kind of a function should a learner learn
  - Linear models (classifiers and regressors)
  - Decision trees
  - Non-linear classifiers, kernels
  - Ensembles of classifiers
  - …

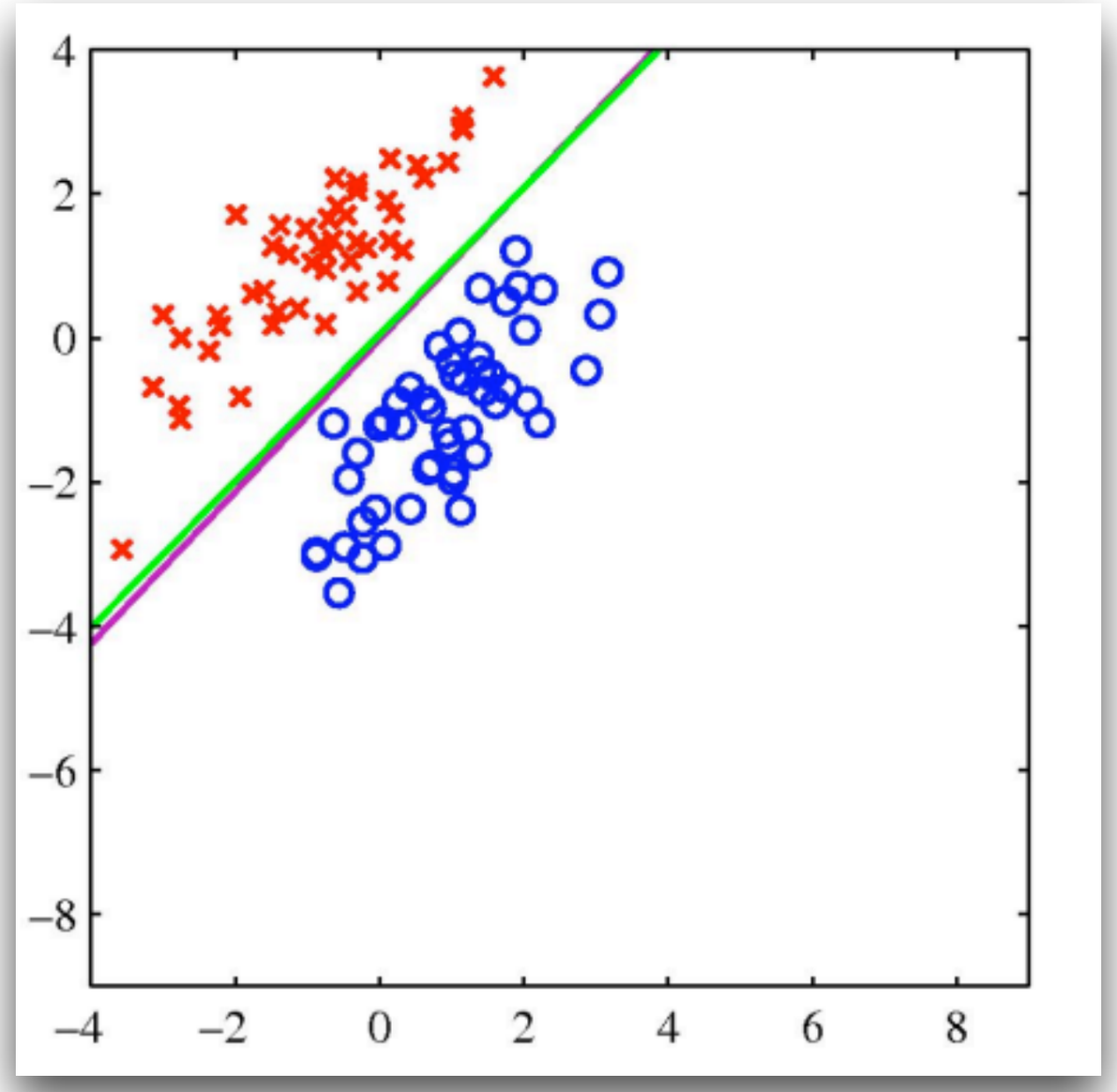
Different learning protocols

- Supervised learning
  - A *teacher* supplies a collection of examples with labels
  - The *learner* has to label new examples using this data
- Unsupervised learning
  - No *teacher*, *learner* has only unlabeled examples
- Reinforcement learning
  - *Learner* learns by interacting with the environment
- Active learning
  - *Learner* and *teacher* interact with each other
  - *Learner* can ask questions
- Semi-supervised learning
  - *Learner* has access to both labeled and unlabeled examples

Supervised learning

- **Classification:** target outputs $\mathbf{y_i}$ are discrete class labels. The goal is to correctly classify new inputs:

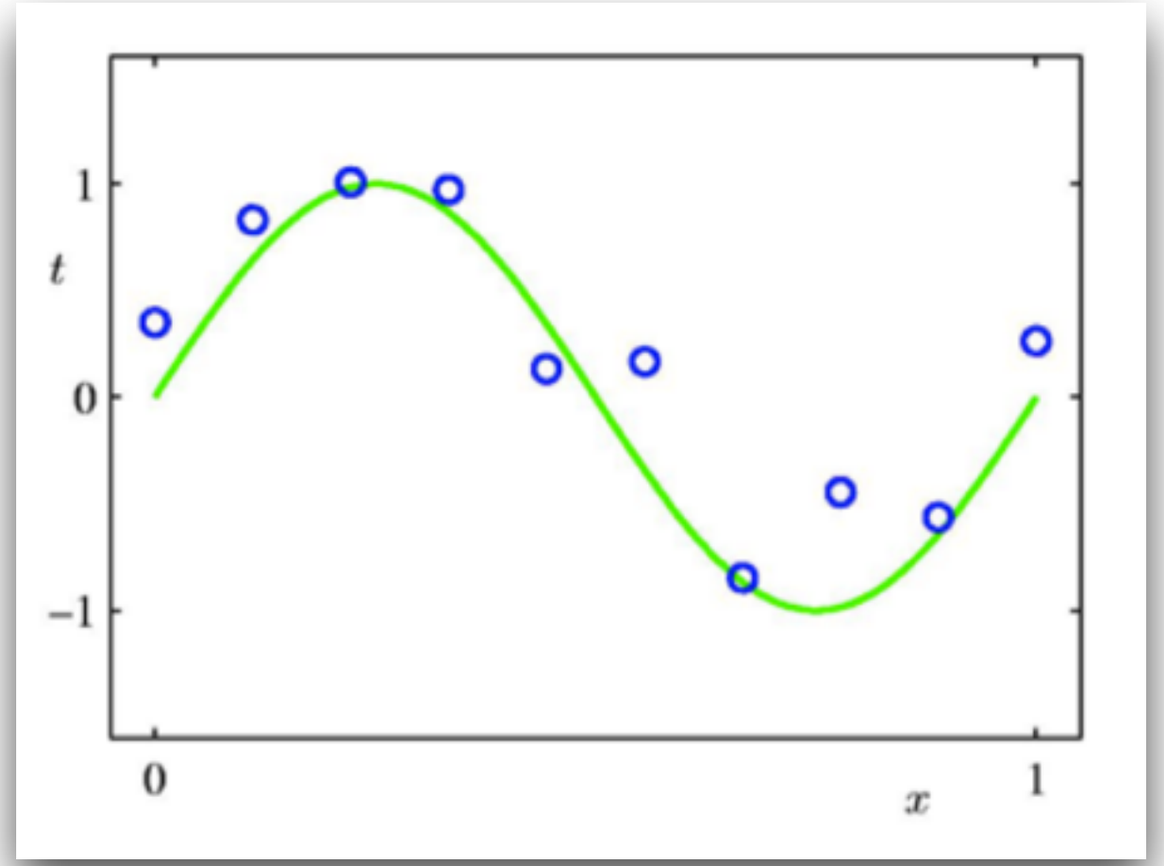
  - Ex: spam filter, credit, ….



Supervised learning

- **Regression:** target outputs $\mathbf{y_i}$ are continuous. The goal is to predict the output given new inputs.

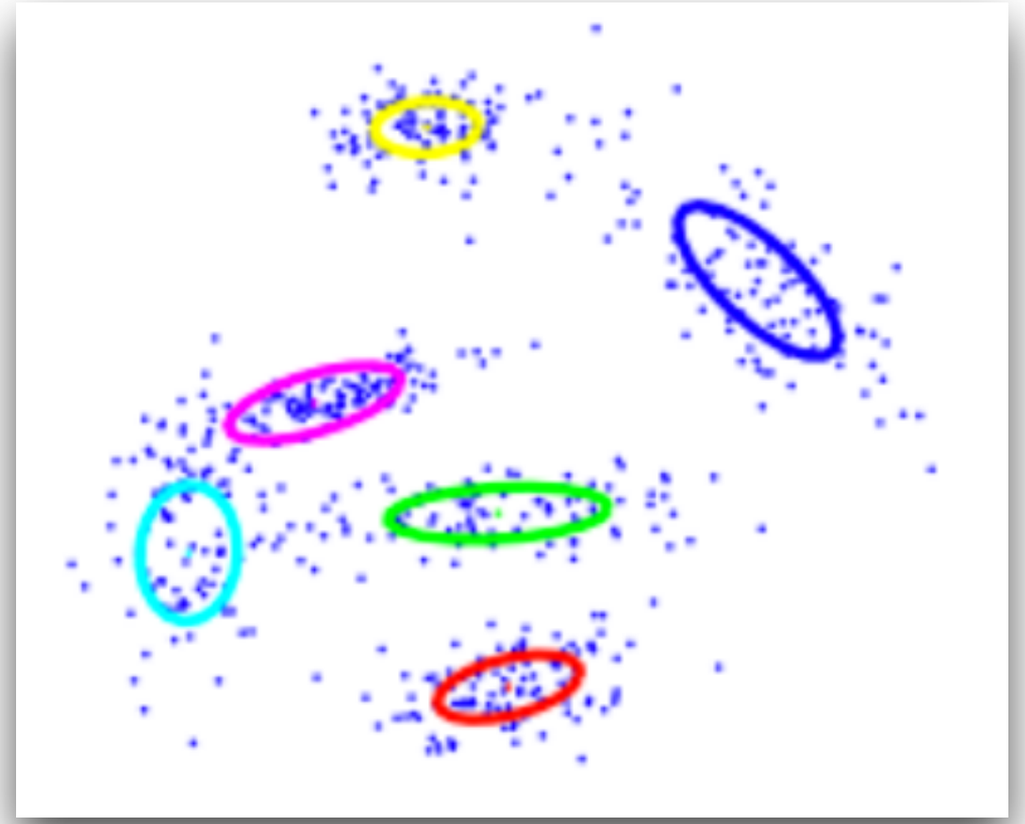
- Ex: price, blood pressure, …



Unsupervised learning

- **Clustering/Dimensionality reduction:** the goal is to construct statistical model that finds useful representation of data:

- Ex: anomaly detection, outlier detection, …



Learning algorithms

- Batch algorithms: Learner can access to the entire datasets

- Online algorithms: Learner can access only one labeled at a time

Representing data

- What is the best way to represent data for a particular task?
  - Features
  - Dimentionality reduction

Evaluation

- What is the best performance metric to our problem?

This course

- Focuses on the underlying concepts and algorithmic ideas in the field of machine learning

- This course is not about
  - Using a specific machine learning tool
  - Any single learning paradigm

Logistics

- **Instructor:** Anisio Lacerda
  - Email: **anisio@decom.cefetmg.br**
    - Office: DECOM - 310
- **Web**: To be announced
- **Discussion:** Piazza + Sistema Acadêmico + others

Pre-requisites

- Programming skills in **Python**
- Basic knowledge of probability/statistics and linear algebra
- Git (create an account on Bitbucket) and read tutorials
- Latex (reports)

Source Materials

- C. Bishop, **Pattern Recognition and Machine Learning (PRML)** (Required)
- J. Gareth, D. Witten, T. Hastie, Robert Tibshirani, **The Elements of Statistical Learning** (Required)
- D. Mackay, **Information Theory, Inference, and Learning Algorithms (ITILA)** (Recommended)
- Additional readings will be made available

Grading

- **Exams** (30 total - 15 each)
- **Programming assignments** (30 total - 15 each)
  - collaboration: **write alone, list collaborators**
- **Final project** (40)
  - Proposal (10), Midway report (20), Presentation (30), Final report (40)
  - Grad (Pós): 1 student
  - Undergrad (Graduação): 2 students
- **Class notes/presentations** (8)
- **Class participation** (2)

Course project

- “Get your hands dirty” with the course material
- Implement a solution to a practical problem
- Application of techniques you learnt here
- Ideas on the course website (soon)

Project: Timeline

- 11/04/2017: Project proposals due; feedback by instructor
- 16/05/2017: Midway report
- 22/06/2017: Final report
- 27-29/06/2017: Final project presentation

- We will have a Best Project Award!

Project proposal

- What is the idea of this project?
- Who will participate?
- What data will you use? Will you need time “cleaning up” the data?
- What code will you need to write?
- What existing code are you planning to use?
- What references are relevant? Mention 1-3 related papers
- What are you planning to accomplish by the milestone?
- See suggestions (soon)

Todo

- Read: PRML Ch. 1.1 / ESL Ch. 1
- Read: Pedro Domingos, ***A few useful things to know about Machine Learning***, Communications of the ACM (2012)
- Python + IPython notebook (see tutorials)
- Git tutorial
- Latex

It’s going to be hard work

Remember to have fun!