Machine Learning and Computational Statistics

(DSC6135)

1. Why Machine Learning?

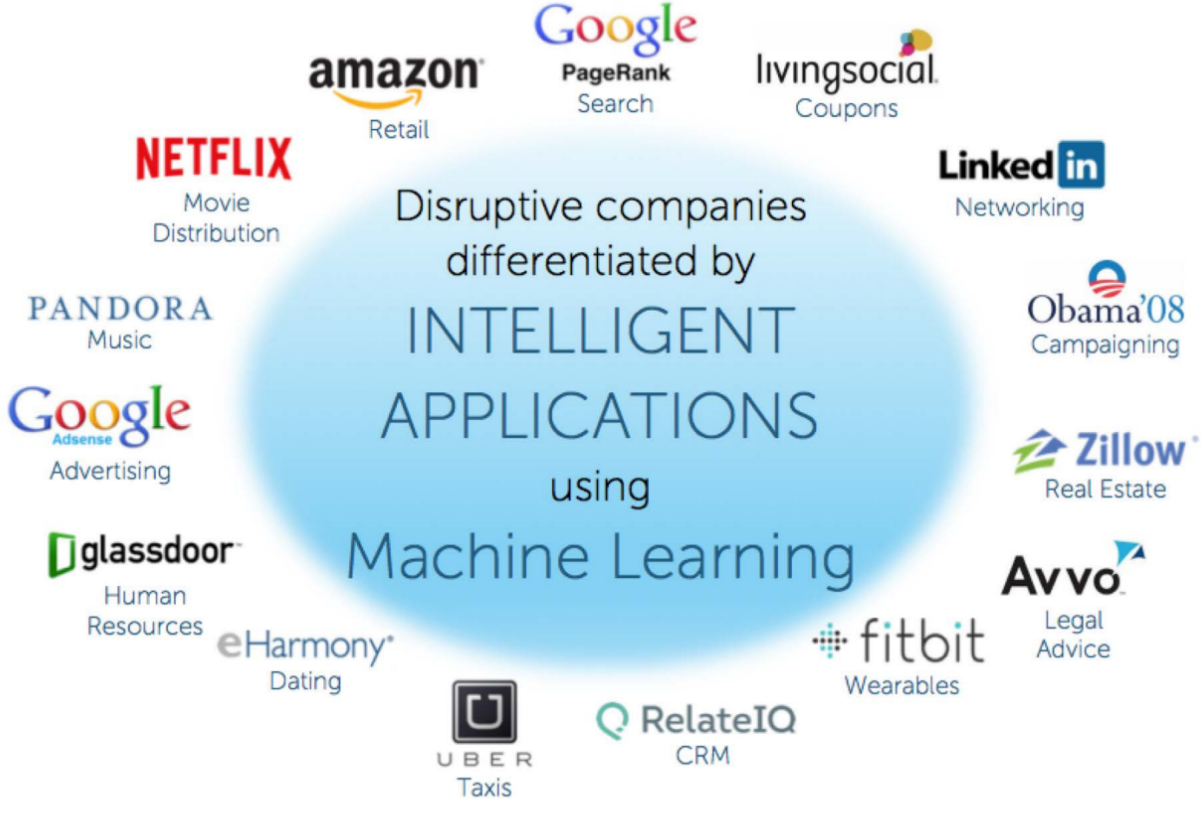
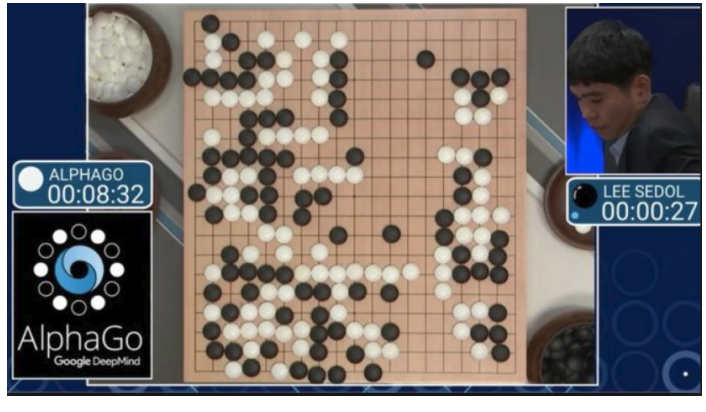


Image Credit: Emily Fox

Artificial Intelligence (AI)

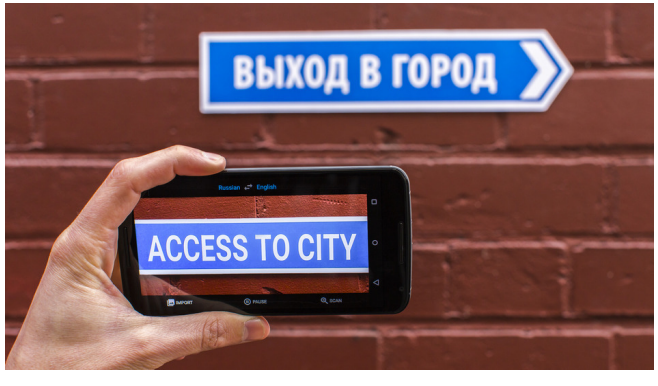
Study of “intelligent systems”, with many parts: logic, planning, search, probabilistic reasoning, **learning from experience**, interacting with other agents, etc

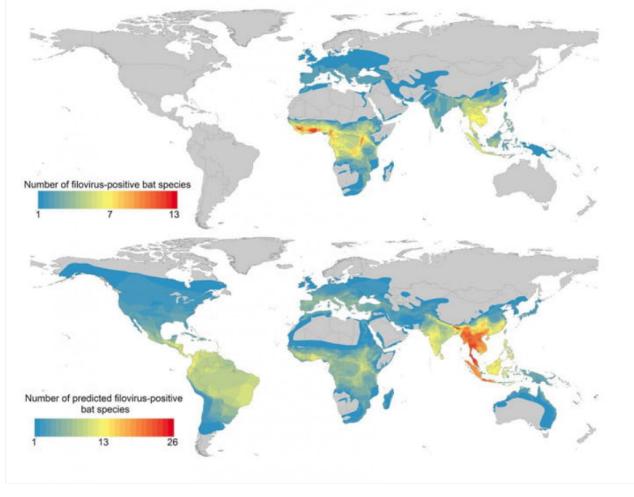


Machine Learning (ML)

- Study of algorithms that learn from experience/data to perform a task
- Task output: a prediction or a decision / underlying pattern in data

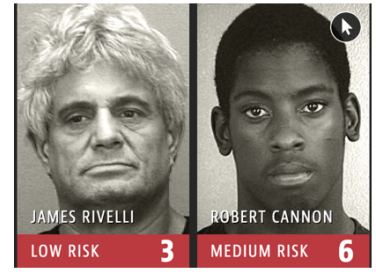
A wide range of possibilities...

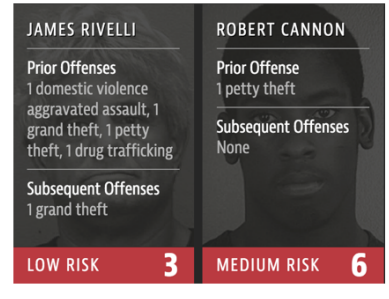




But problems can also arise:

* discrimination on race example: Compas
* Gender discrimination
* Biases in healthcare





Machine learning needs to be applied responsibly!

Objectives of this course

"With great power comes great responsibility".

Goals: After this course, you will...

```
* make appropriate model choices
* understand why and how machine learning works
* identify sources of error
* evaluate carefully
```

Objectives of this course

Our goal is to prepare students to **deeply understand and effectively apply machine learning methods** to problems that might arise in "the real world" -- in industry, medicine, education, and beyond.

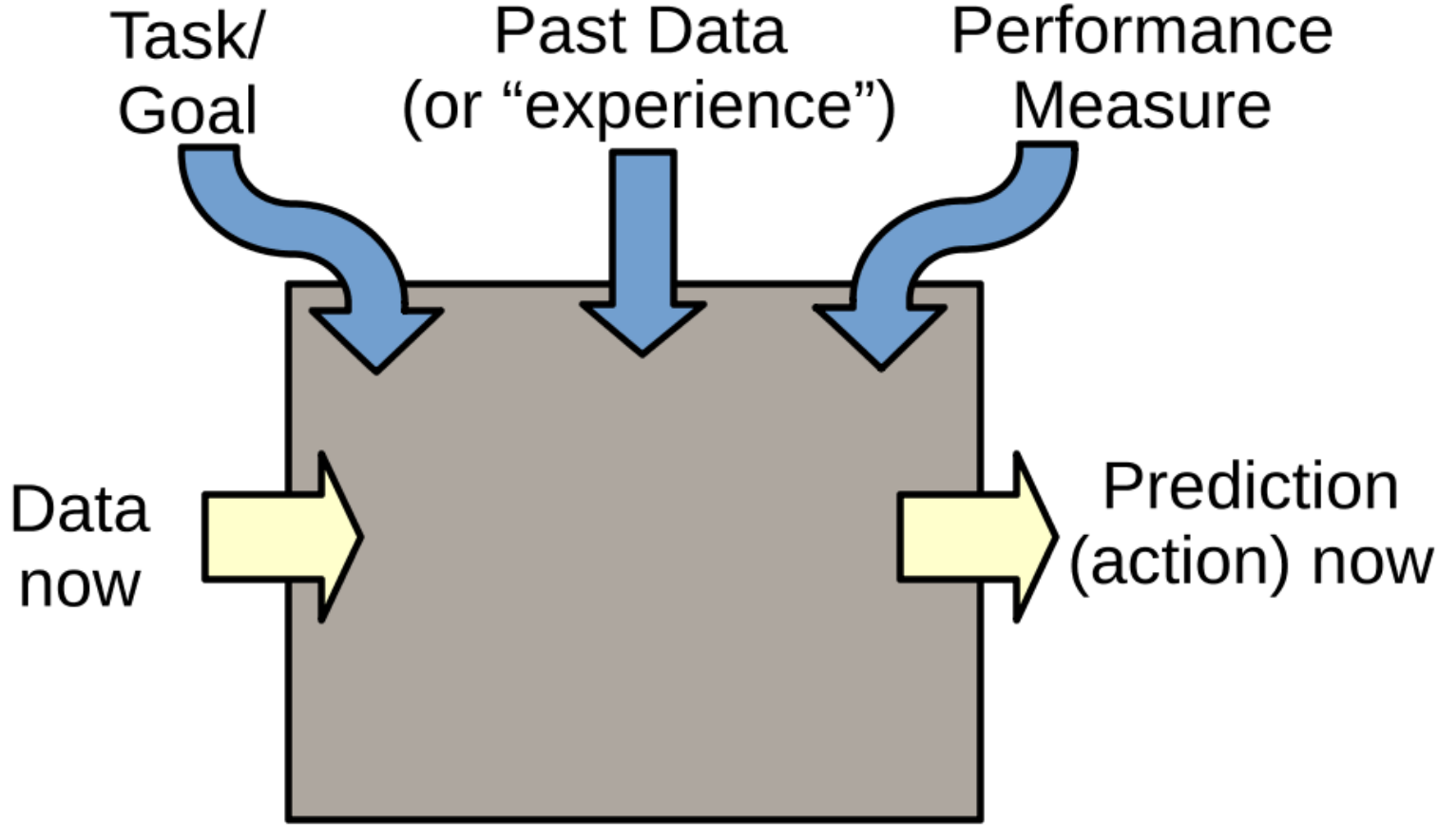
You will gain skills and understanding for a future as:

- **Developer** using ML “out-of-the-box”
- **Researcher**, either using or developing ML methods

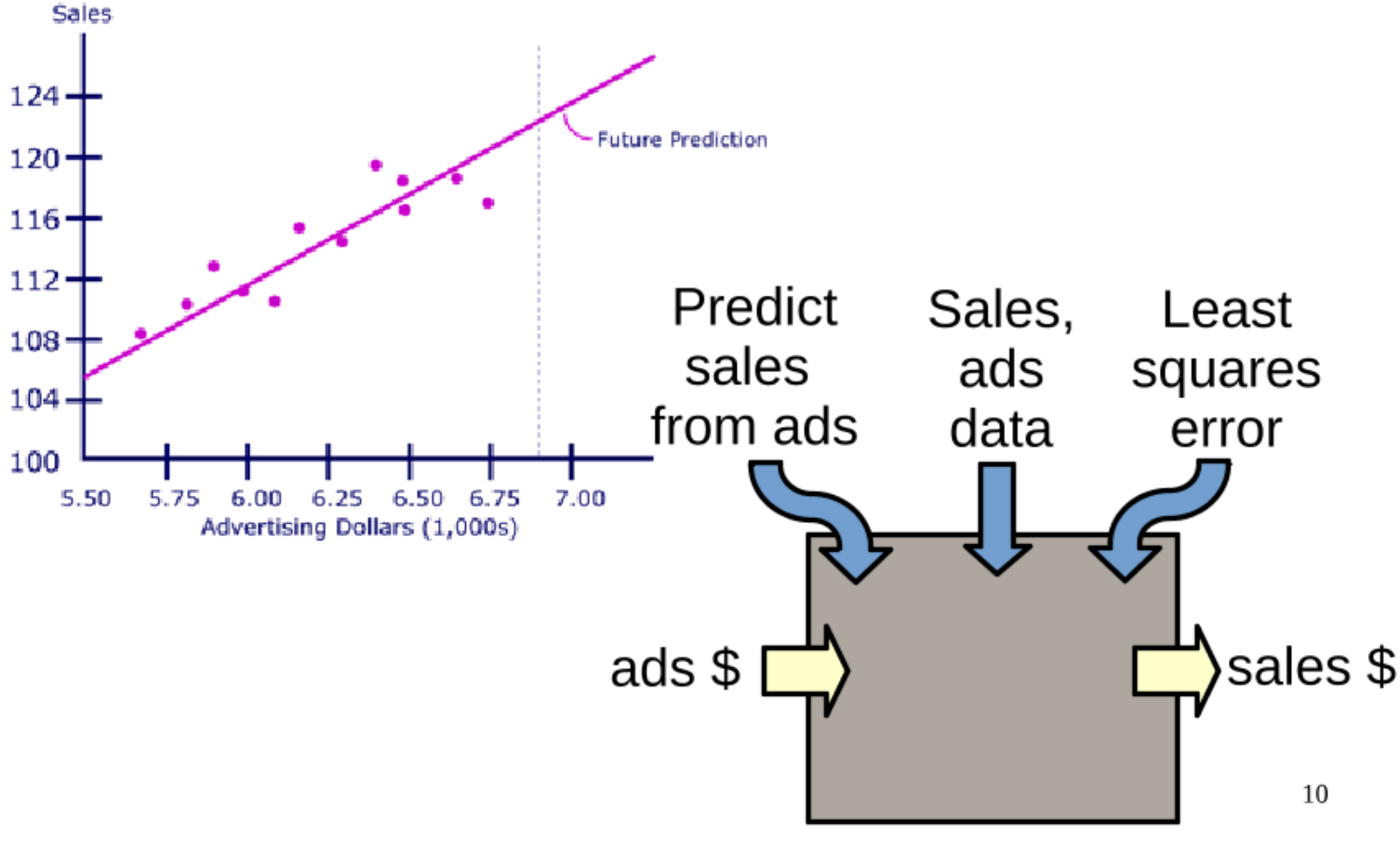
Focus on both, **theory** and **practice**.

Let’s get to it!

2. The Machine Learning (ML) Process



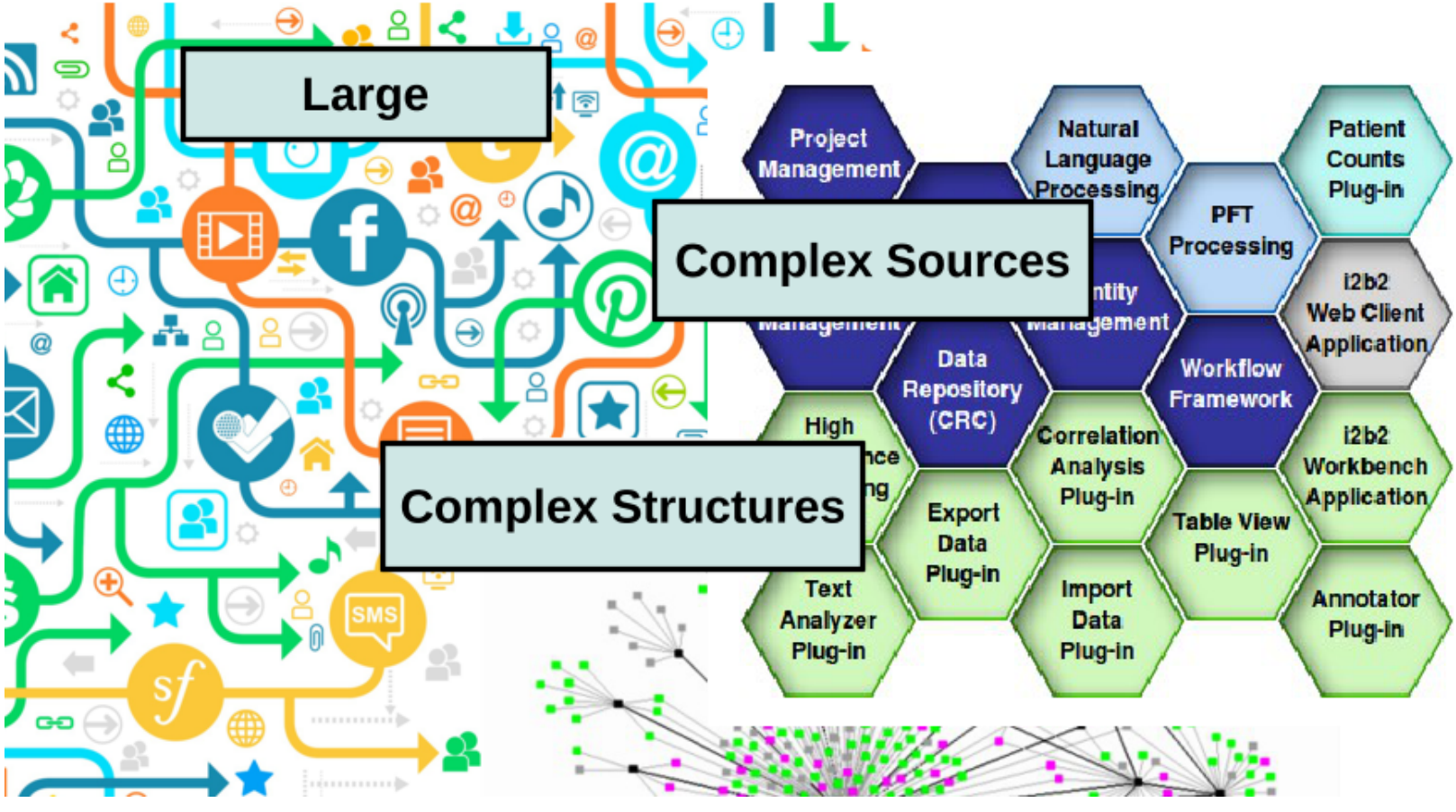
The starting point...



http://ci.columbia.edu/ci/premba_test/c0331/images/s7/7176267017.gif

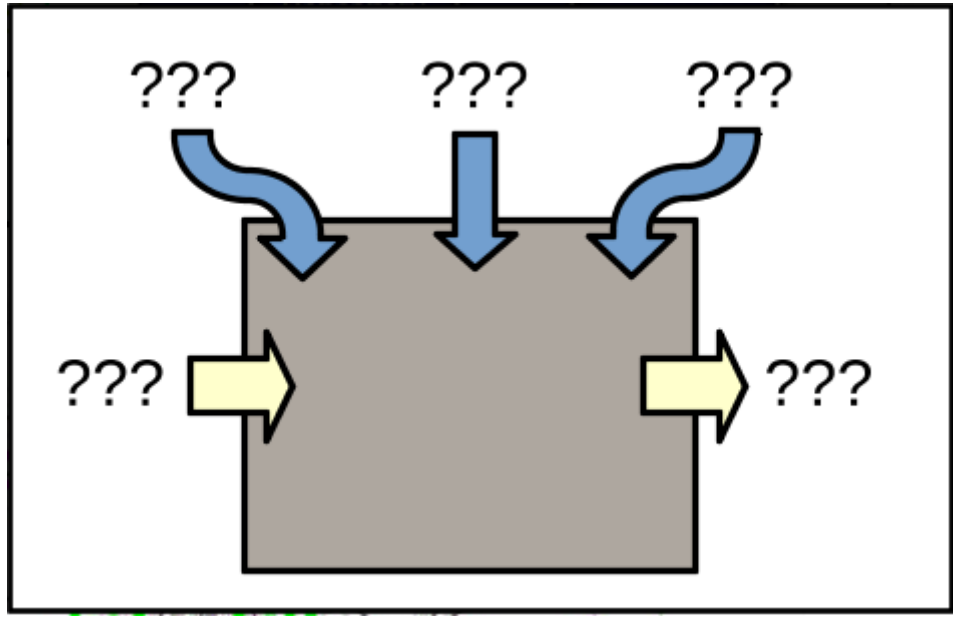
(slide from CS-181 2019 Harvard course of Prof. Finale Doshi-Velez)

... where we are now



(slide from CS-181 2019 Harvard course of Prof. Finale Doshi-Velez)

In general, hard to formalize ML process!



Data Science Pipeline

Step 1. Exploratory Data Analysis (EDA): `Understand your data and ML task`

Step 2. Choose model: `Make reasonable assumptions!`

Step 3. Train model (also called `Inference`)

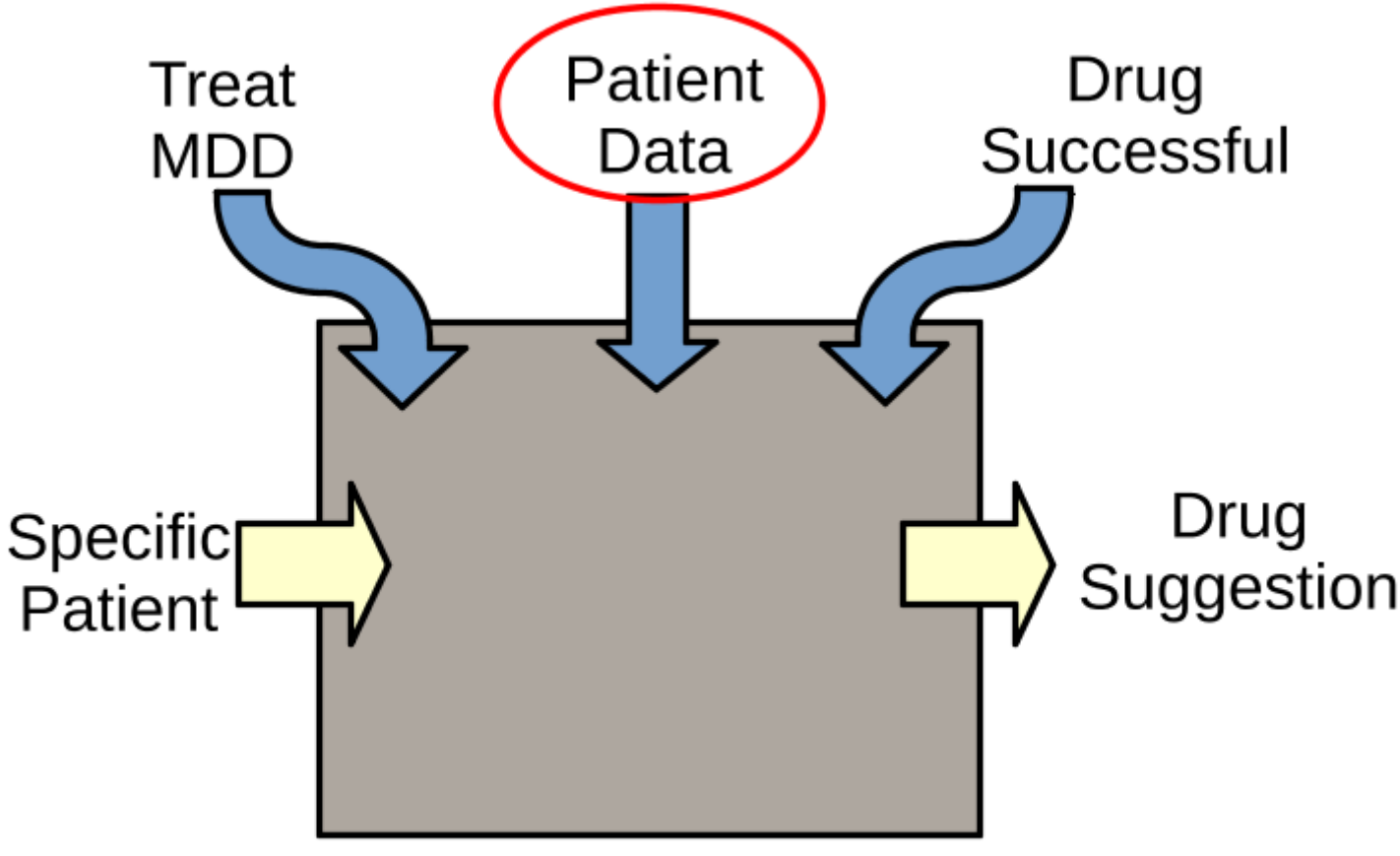
Step 4. Criticize your model (repeat Steps 1,2,3)

3. Case-study: Treating Depression

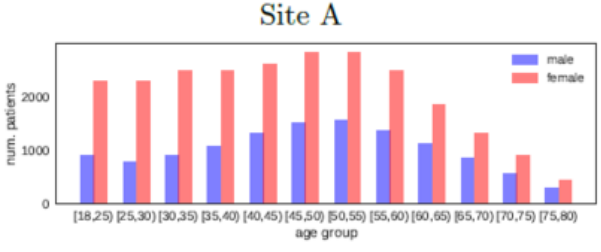
Clinical question: what meds to give to which depression patients?

Also, can we *explain* who responds to what drugs?

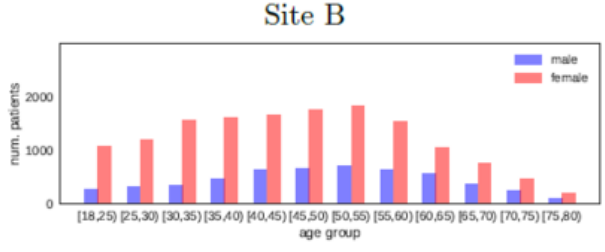
Formalizing...



Understanding your data – preprocessing



	female	male	total	frac.
Asian	603	222	825	0.022
Black	959	412	1371	0.037
Hispanic	1006	402	1408	0.038
Other	2153	885	3038	0.082
White	20158	10415	30575	0.822
total	24879	12336	37217	
frac.	0.668	0.331		

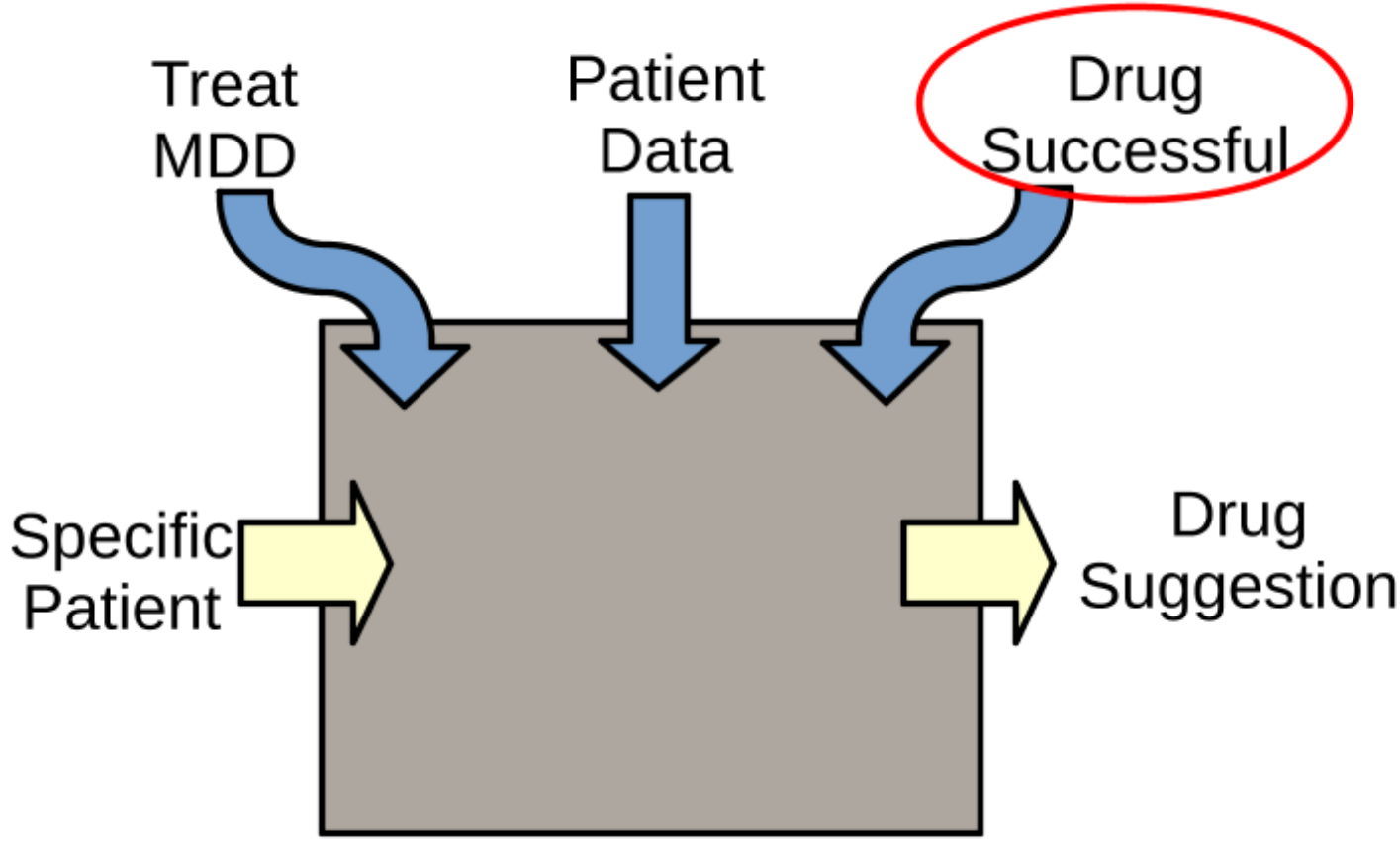


	female	male	total	frac.
Asian	233	56	289	0.014
Black	1604	419	2023	0.100
Hispanic	2264	657	2921	0.145
Other	1026	386	1413	0.070
White	9662	3888	13551	0.671
total	14789	5406	20197	
frac.	0.732	0.268		

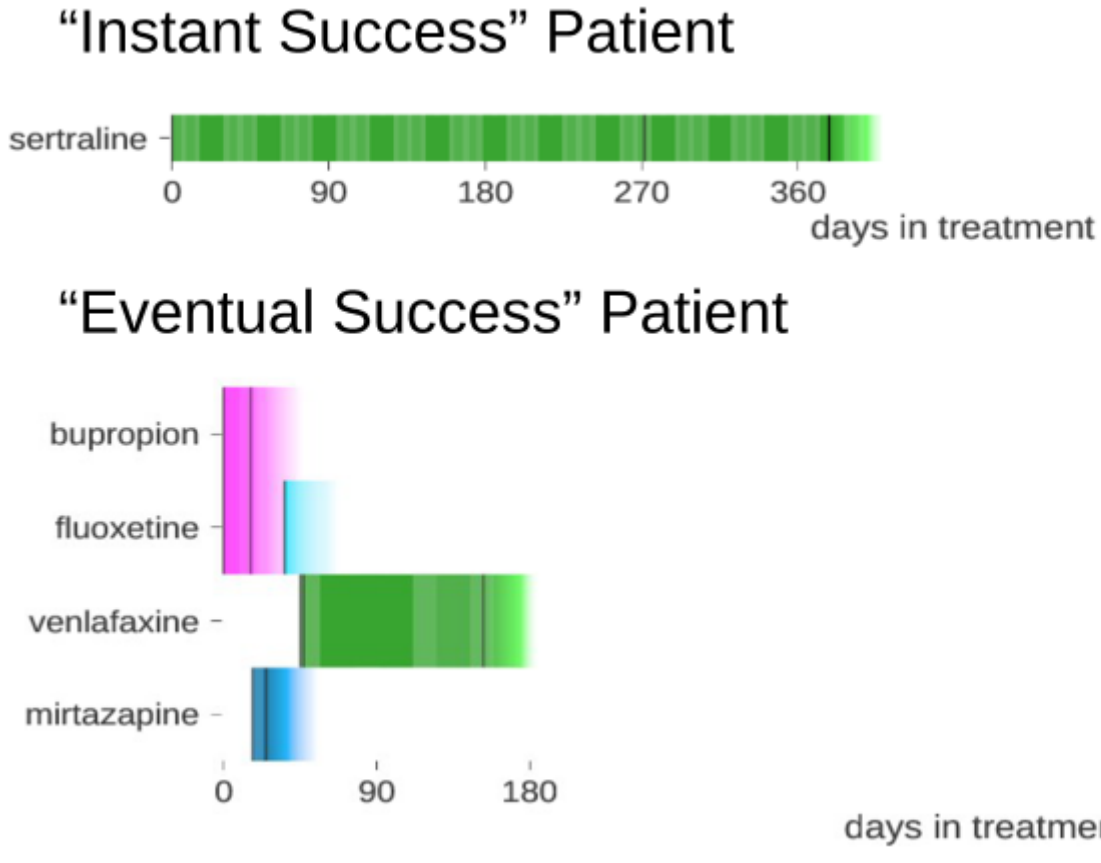
Electronic Health Records (EHR)

- Originally ~32k codes (diagnosis, procedures, drugs, etc...)
- Focus on most common antidepressants

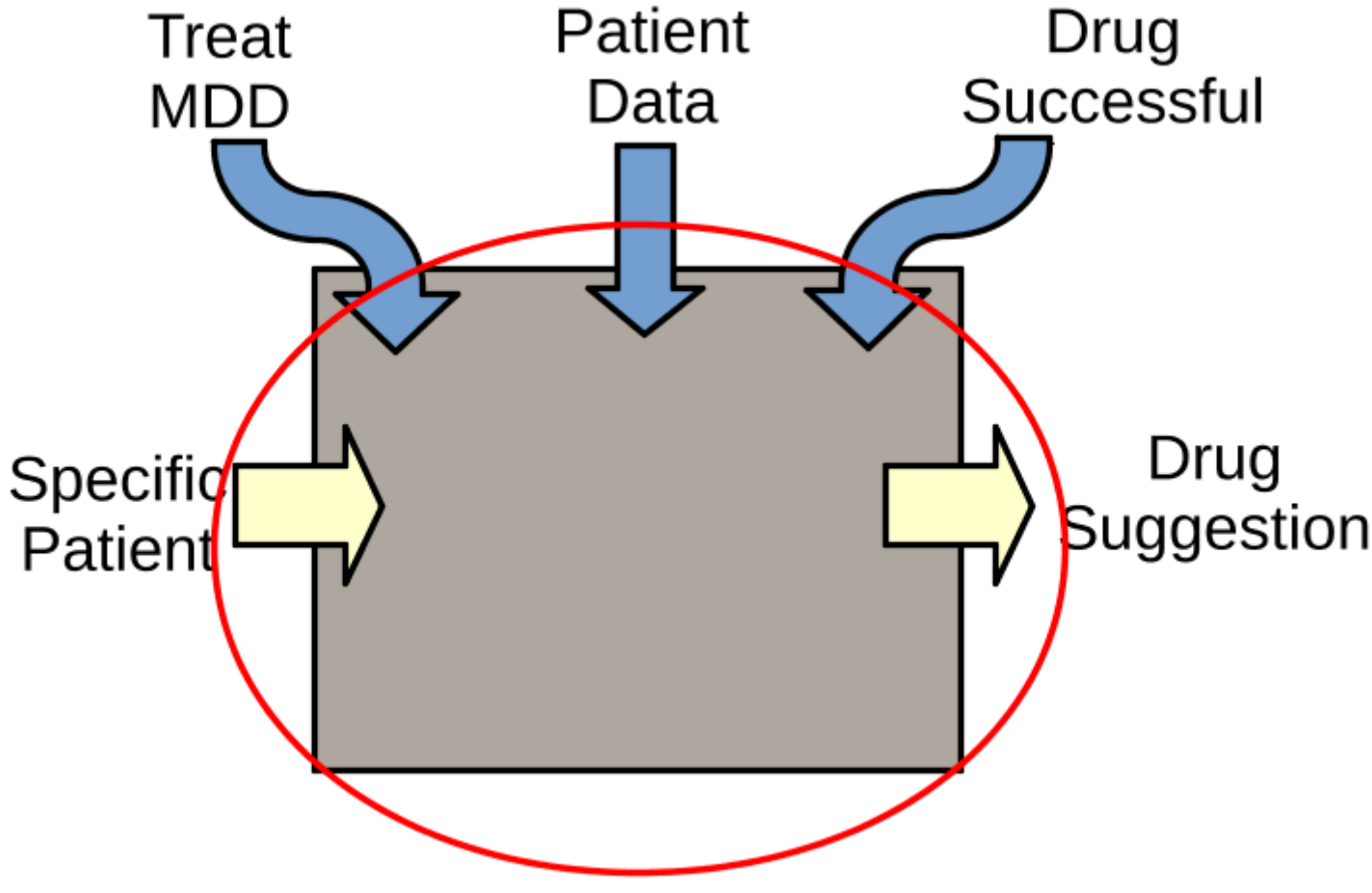
Formalizing...



Performance Metric



Formalizing...



Choices for Classifiers

(all these covered in this course)

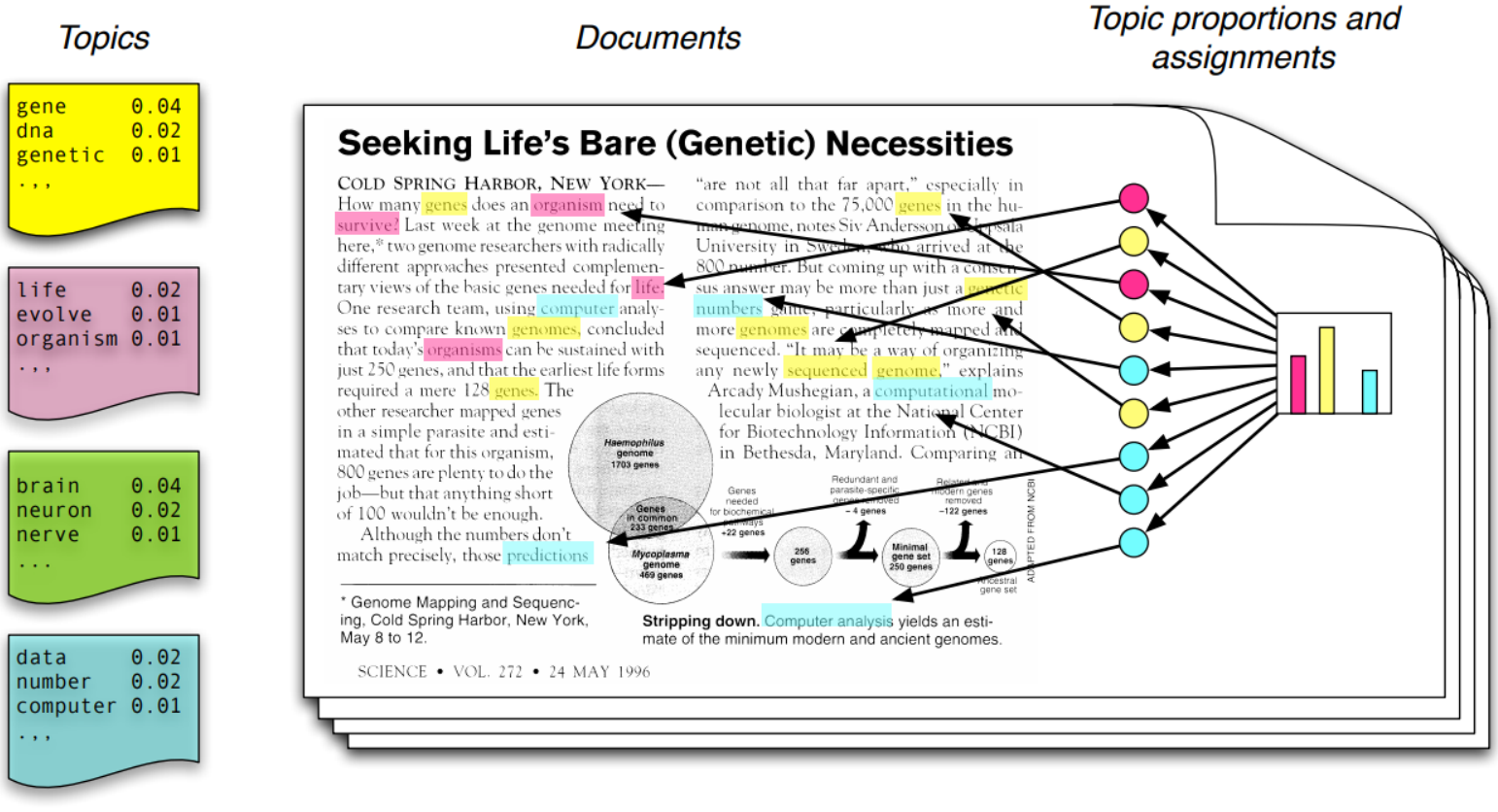
- logistic regression
- random forest
- neural networks
- ...

Which classifier to choose for the "`treating depression`" task?

Challenge: high-dimensionality of EHR codes

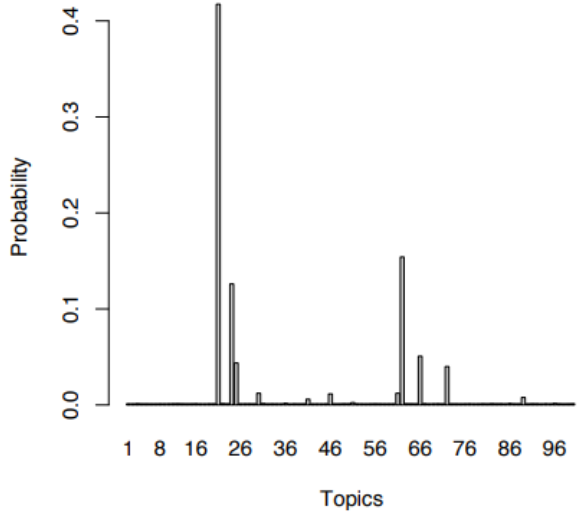
- Observed codes might encode a higher-level concept: – “Hip fracture” could code for “elderly” – “Pregnant” could code for “female”
- One option is to **apply topic models for dimensionality reduction**, and then **predict** based on the doc-topic probabilities.
    - Topics are distributions over words (codes)
    - Patients are distributions over topics

Topic Models



Blei, David M.. "Introduction to Probabilistic Topic Models." (2010).

Topic Models



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Blei, David M.. "Introduction to Probabilistic Topic Models." (2010).

Topic Models: back to Treating Depression

- Topics are distributions over words (codes)
- Patients are distributions over topics

Example topic

```
1.0000   29650:bipolar_affective_disorder,_depres
0.9999    2967:bipolar_affective_disorder,_unspec
0.9999   29570:schizo-affective_type_schizophreni
0.9999   29660:bipolar_affective_disorder,_mixed,
0.9998  c90870:electroconvulsive_therapy_(include
0.9998  c00104:anesthesia_for_electroconvulsive_t
0.9997   29560:residual_schizophrenia,_unspecifie
0.9996   p9427:other_electroshock_therapy
0.9993  d00061:lithium
0.9993   29653:bipolar_affective_disorder,_depres
0.9985   29651:bipolar_affective_disorder,_depres
0.9985  d04825:aripiprazole
```

Patient has bipolar disorder

Your turn!

Fighting Climate Change: Car Milleage Prediction

Your turn!

Protecting Endangered Specie: Predicting Age of White Abalone

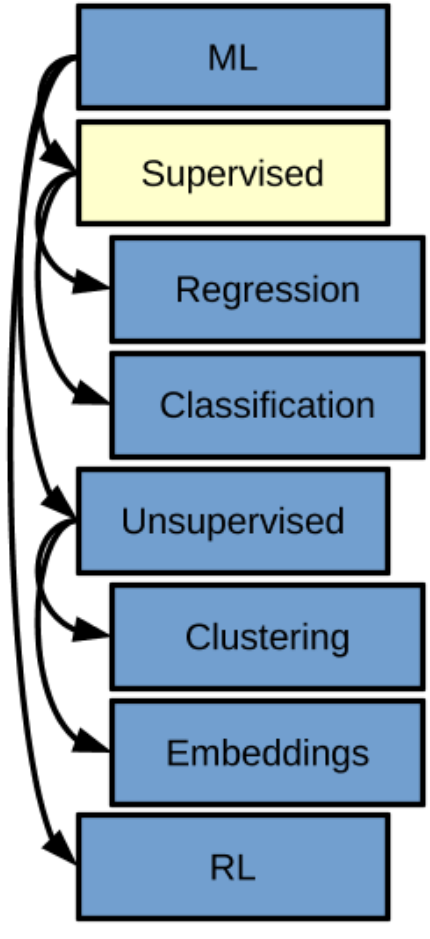
3. Machine Learning Taxonomy

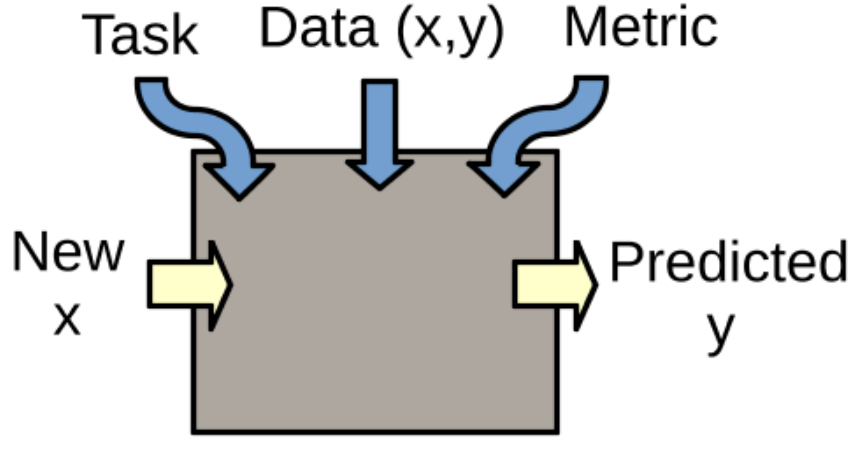
Healthcare Diagnosis: Classifying Patients at Risk of Cancer

3. Machine Learning Taxonomy

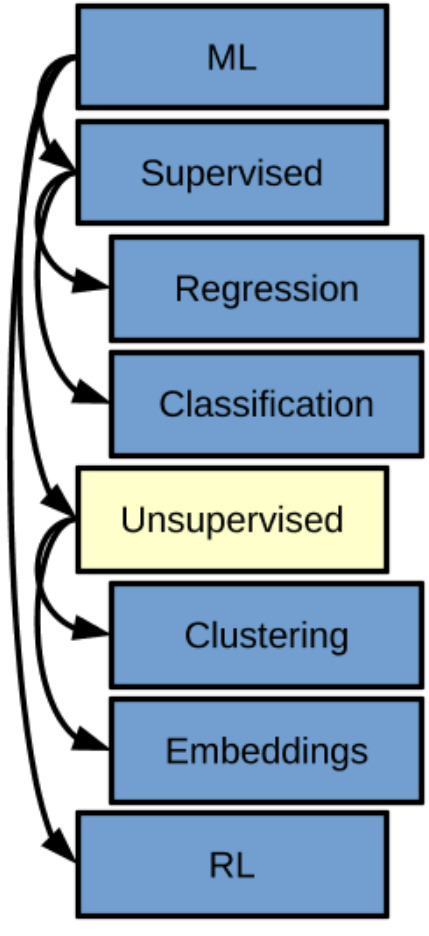
Building a Recommendation System: Predicting Movie Ratings

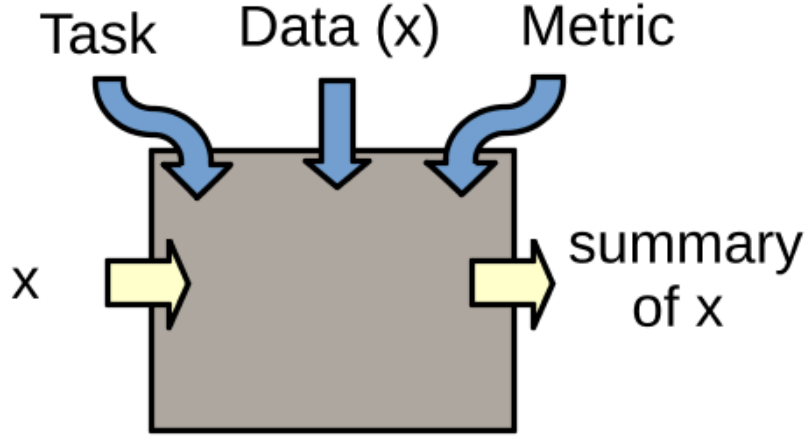
3. Machine Learning Taxonomy



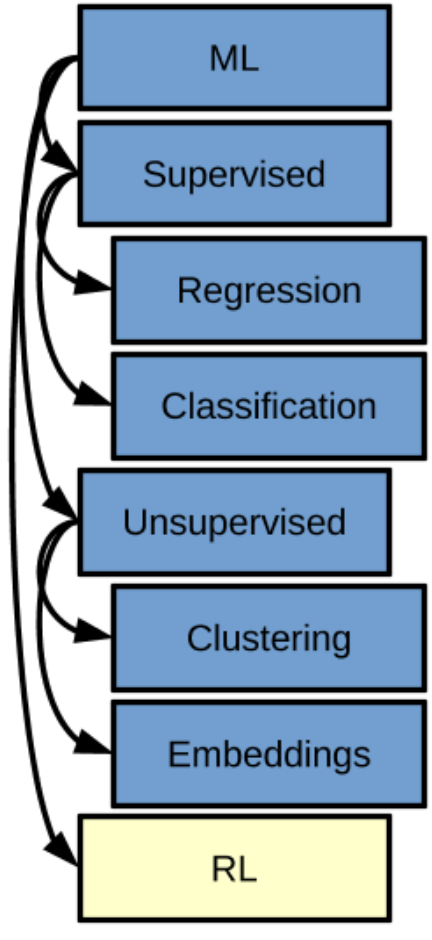


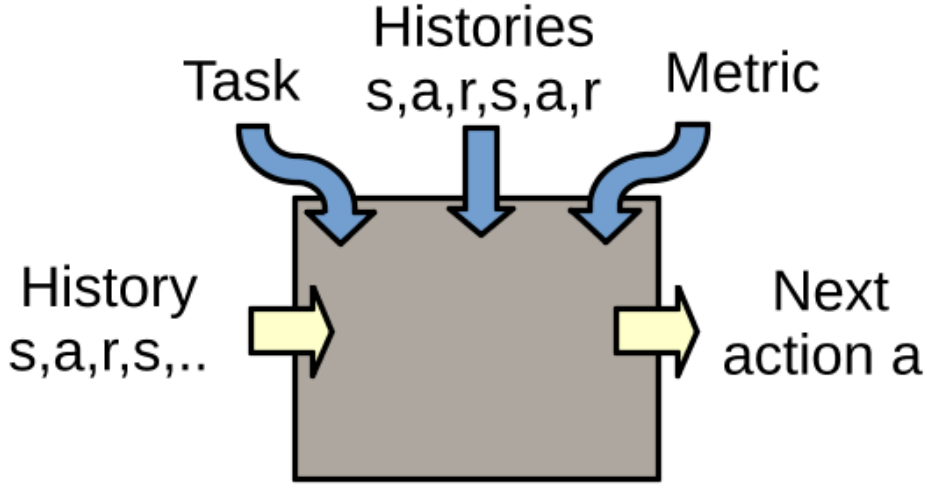
3. Machine Learning Taxonomy





3. Machine Learning Taxonomy





4. Course Logistics

Prerequisites

- Python programming
- Background in stats, and simple distributions
- Some linear algebra, multivariate calculus

We will review concepts on the way.

Connection with other ACE-DS courses

This is a master level course, it connects with the following courses:

- Principles of Data Science (DSC6131)
- Probability and Statistical models (DSC6132)
- Programming for Data Scientists (DSC6133)

What is ahead (content)

Tasks to be discussed, what is ahead.

```
* Regression
* Classification
* Recommendation systems
* Dimensionality reduction
* Clustering
* Probabilistic Modeling
* Statistical Decision Theory
```

- **Emphasis**: hands-on course, learn by doing! Lectures will include real world stories and concept exercises.

Structure of the course

- Intensive course of 10 days, 4h/day
- Approximate schedule for each day:
    - Quizz or homework correction + recap (30 min)
    - Lecture (1h)
    - Practical (30 min)
    - Break (15 min)
    - Lecture (1h)
    - Practical (30 min)
    - Introduction to Homework or Clarifications (15 min)

- Homeworks will be released/described at the end of class, every two days.
- Short quizzes at the beginning of each day.

Grading: what is expected from you

- (50 pts) 5 mandatory homeworks (every 2 days)
- (30 pts) quizzes
- (15 pts) paper presentation
- (5 pts) participation

Come prepared every day!

Infrastructure

- Instructors: Melanie F. Pradier (melanie@seas.harvard.edu), Javier Zazo (jzazo@seas.harvard.edu), and Weiwei Pan (weiweipan@g.harvard.edu)
- Office hours: 16h30 to 18h00 (send us an email)
- Webpage: https://melaniefp.github.io/intro_to_ML_DSC6135/ (https://melaniefp.github.io/intro_to_ML_DSC6135/)

Many slides in this course attributable to/inspired by:

- Mike Hughes (Tufts)
- Erik Sudderth (UCI)
- Finale Doshi-Velez (Harvard)
- James, Witten, Hastie, Tibshirani (ISL/ESL books)

Questions?