Yandex

Yandex

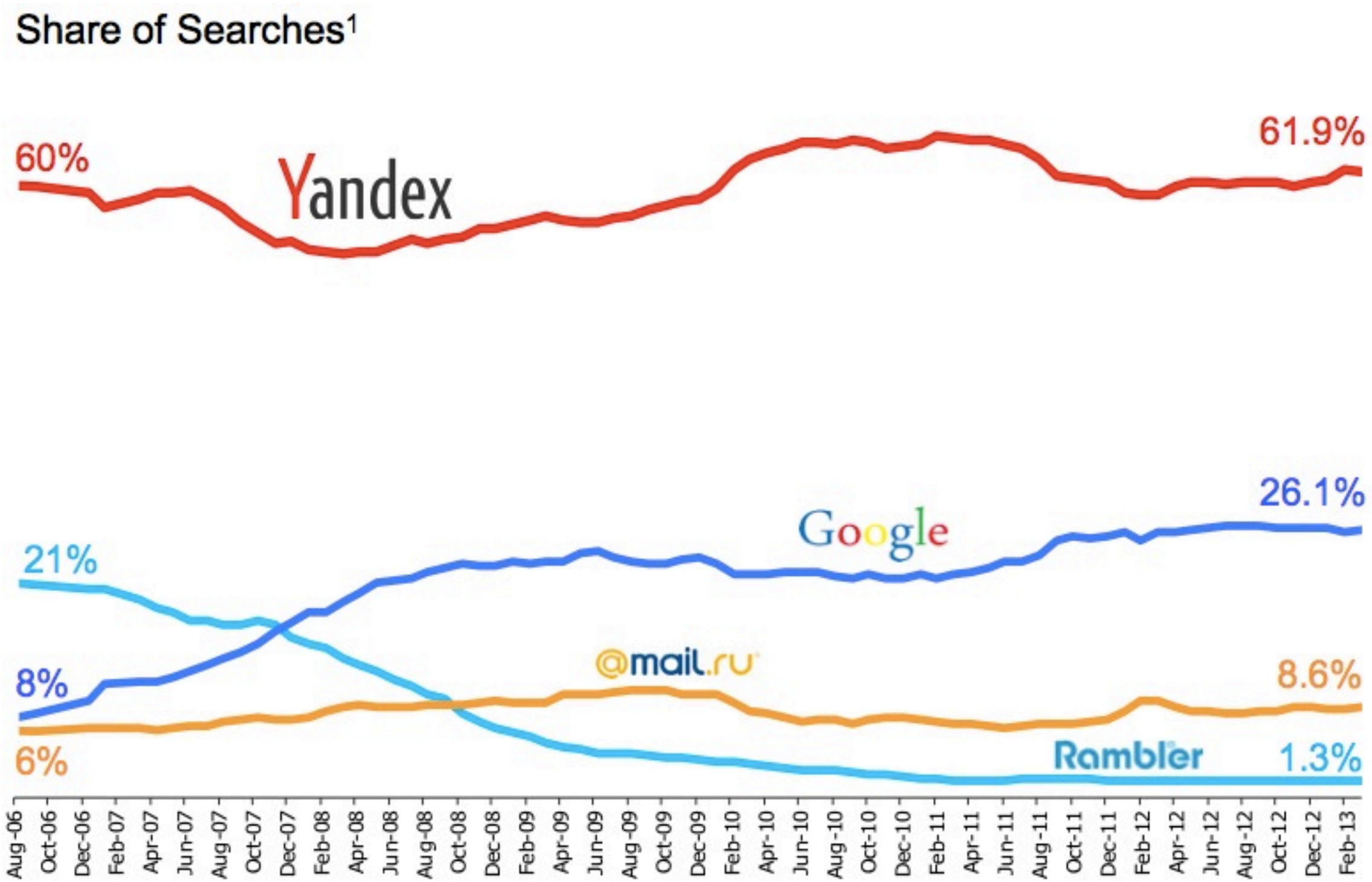
Introduction to Machine Learning

Andrey Ustyuzhanin
Head of CERN-Yandex joint projects

Overview

1. Introduction
2. History of Machine Learning
3. Problems
4. Approaches
5. Demo

Quick intro



[1] Source: LiveInternet.ru, February 2013. Search traffic reflects Russian users to Russian websites and includes desktop and mobile

Brief History of ML

1. Statistics
2. Artificial Intelligence
3. Expert Systems

...

4. Machine Learning

Math & Stats

There is a e-commerce website:

10000 users

100 clients (buying something)

Test (T):

Predict that user is a client - 99%

Predict that user is not a client - 99%

What is probability that U is a client if T == True?

Заголовок (не длинней одной строки)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

P(x is client | T == 1) = 1/2

Math & Stats

You have a website

People visit it regularly

Probability that someone comes in during 3sec = 0.992

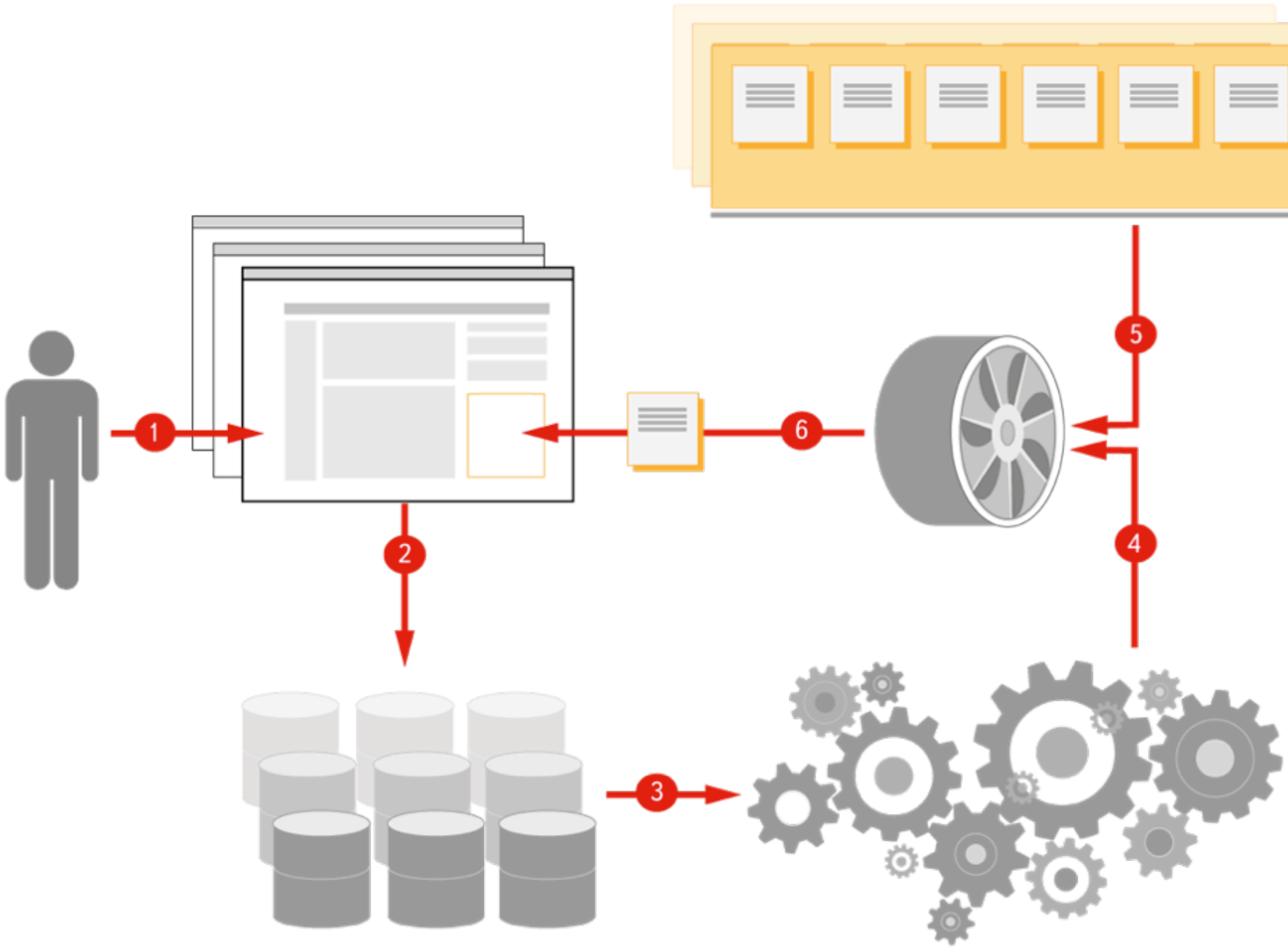
Question:

What is probability of someone coming in during 1 sec?

Meta transition: Statistics → Data Science

"How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"

-- Tom Mitchell, CMU


1
2
3
4
5
6

Applications

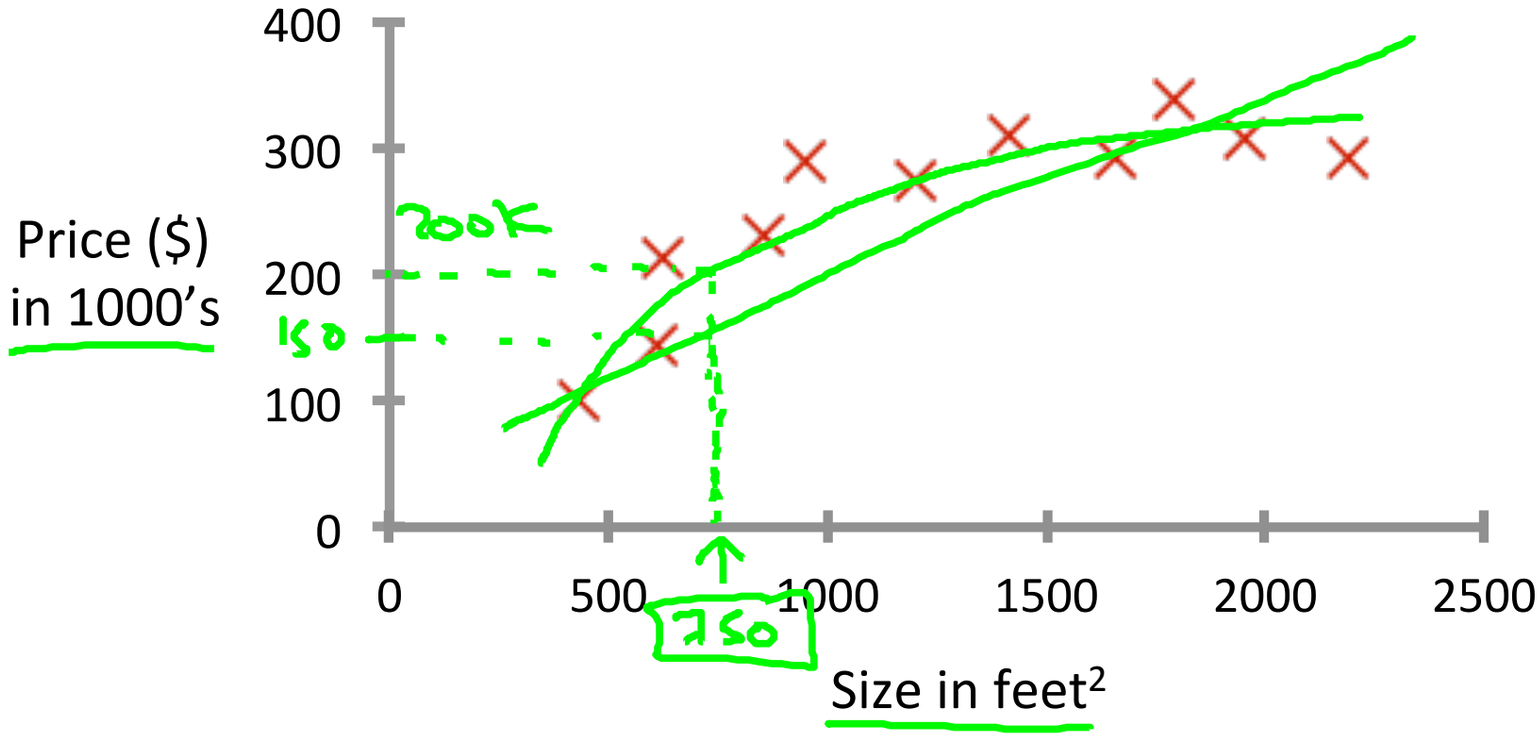
1. Webpage search ranking (as well as news, images and mail search),
2. Advertisement selection,
3. User behavior modeling,
4. Spam filtering,
5. Social-demographics

Large scale

1. Consists of
   - Math
   - Tools
   - Infrastructure
2. Pipeline:
   - Get Data
   - Scrub
   - Explore
   - Model
   - Interpret

Regression Problem

Housing price prediction.



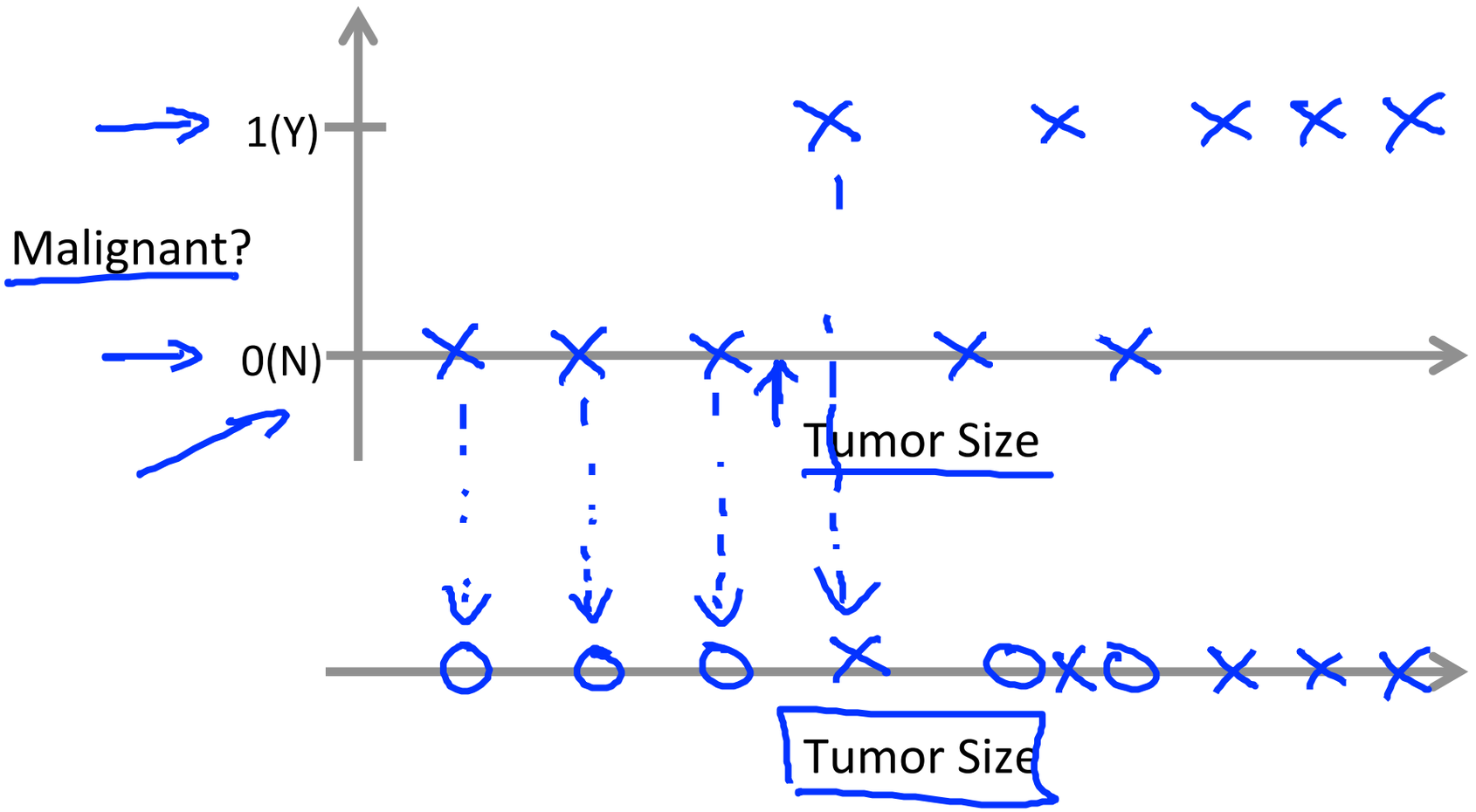
Supervised Learning

"right answers" given

Regression: Predict continuous valued output (price)

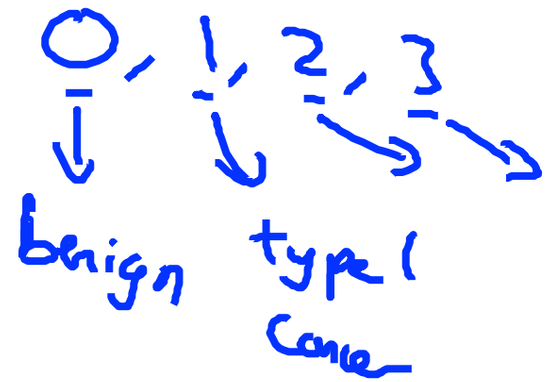
Classification Problem

Breast cancer (malignant, benign)



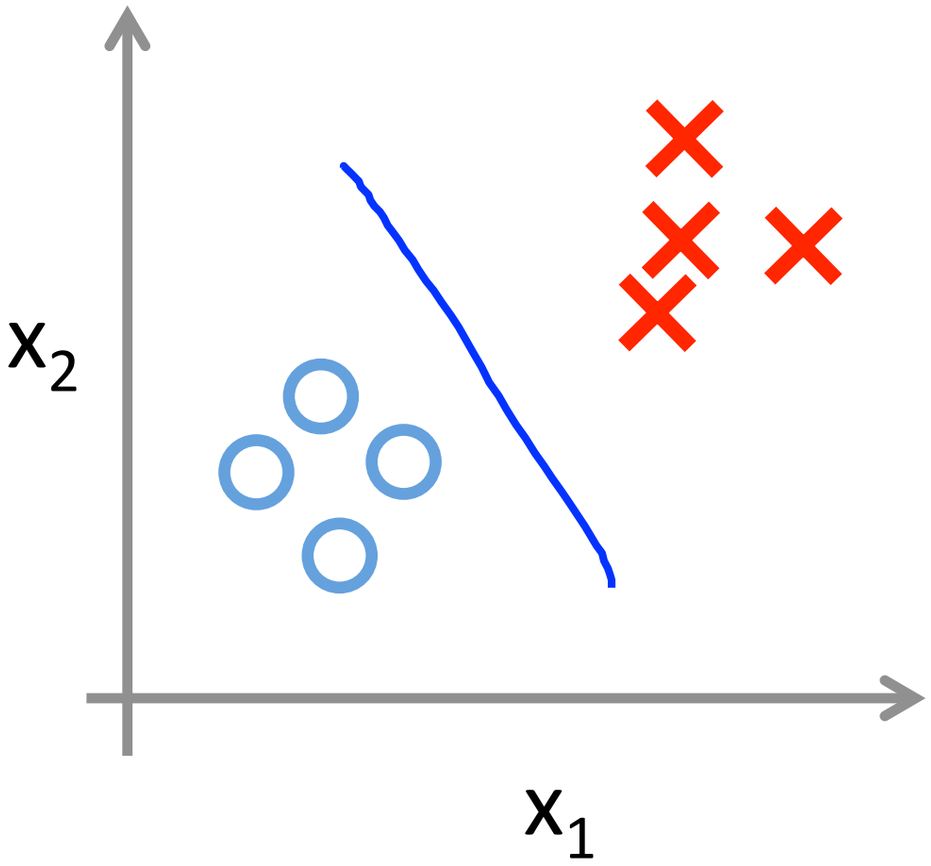
Classification

Discrete valued output (0 or 1)

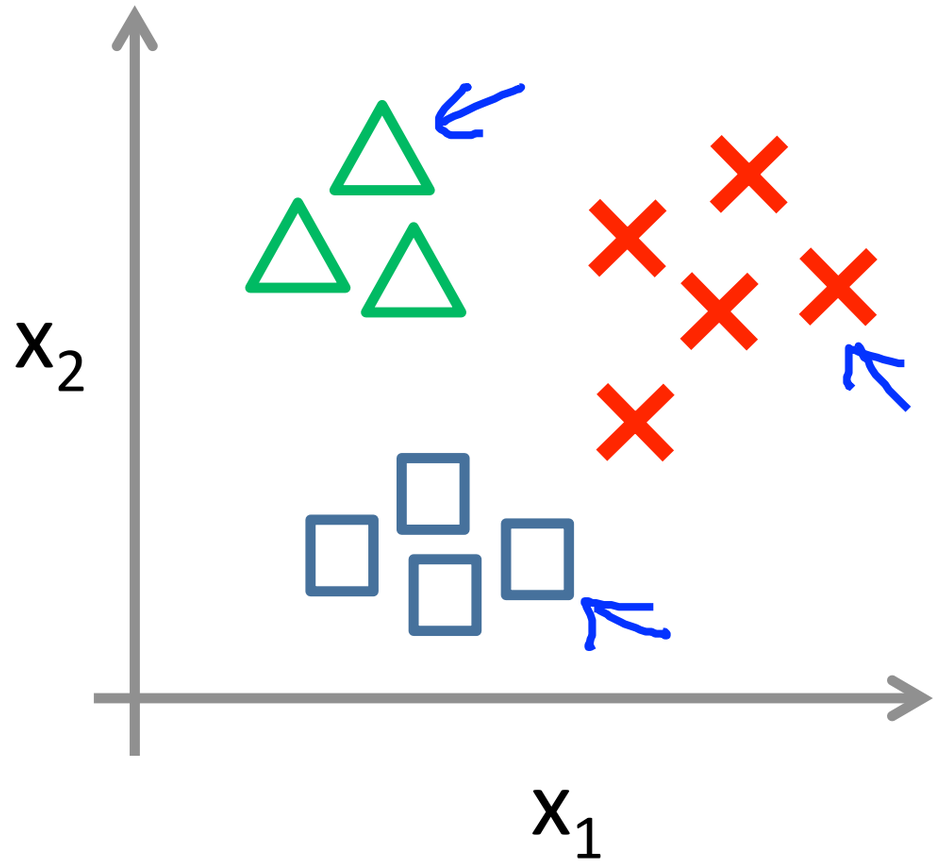


Multiclassification Problem

Binary classification:

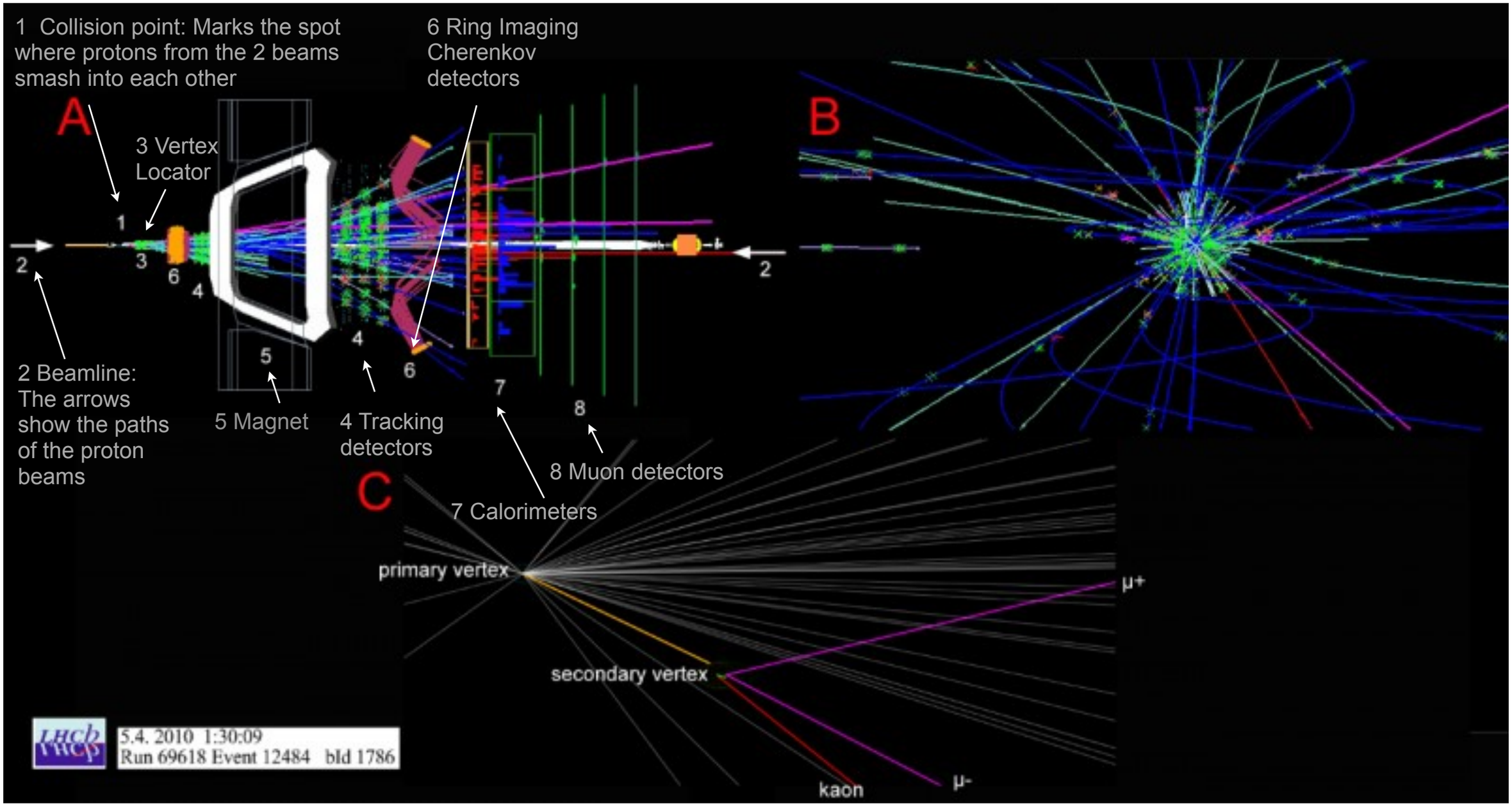


Multi-class classification:

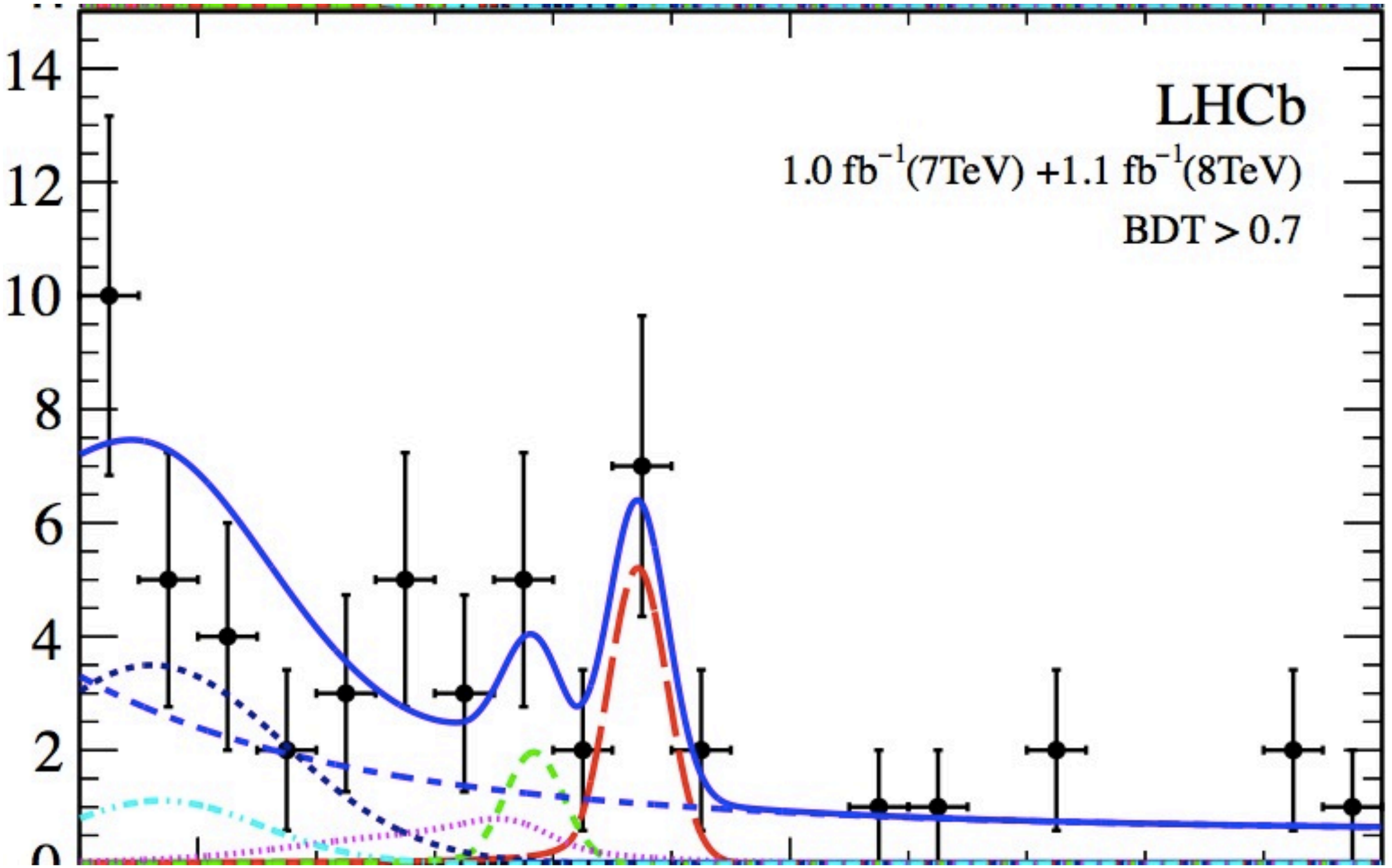


Andrew Ng

Event is...


1 Collision point: Marks the spot where protons from the 2 beams smash into each other
6 Ring Imaging Cherenkov detectors
A
B
3 Vertex Locator
1
2
3
6
4
2
2 Beamline: The arrows show the paths of the proton beams
5
4
6
7
8
5 Magnet
4 Tracking detectors
8 Muon detectors
7 Calorimeters
C
primary vertex
μ+
secondary vertex
kaon
μ-
5.4. 2010 1:30:09
Run 69618 Event 12484 bld 1786

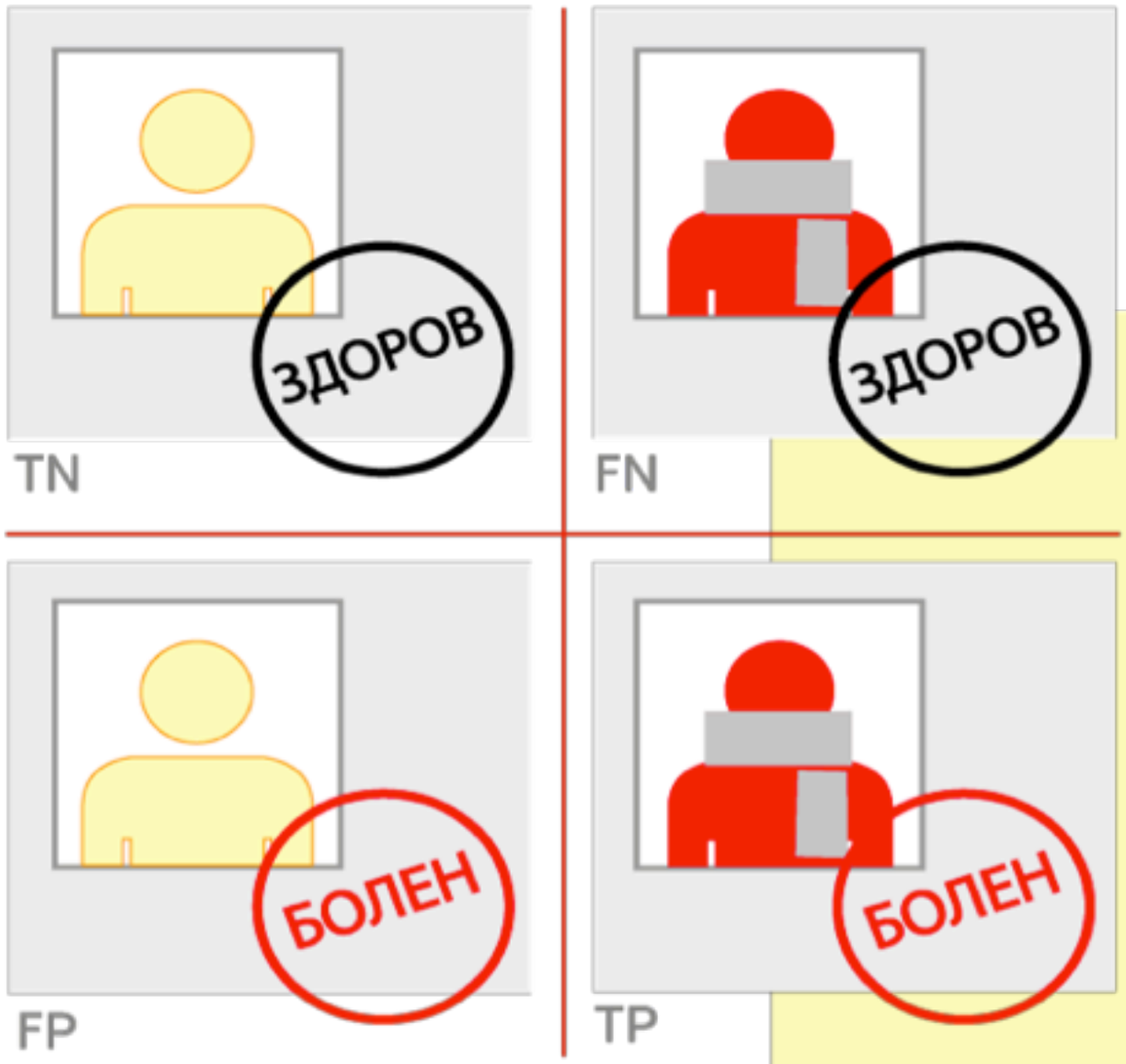
Event Classification (binary)

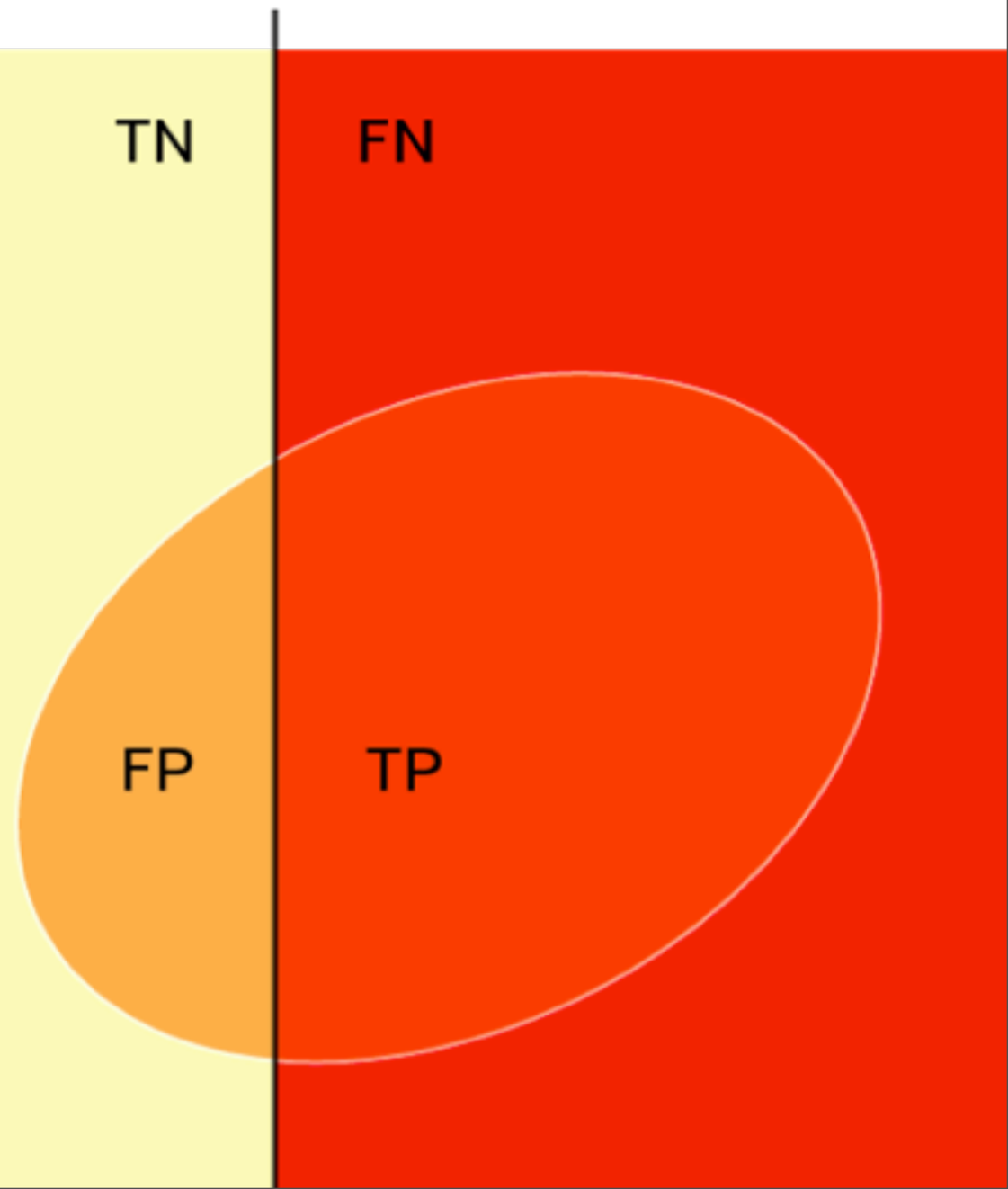


Training Process

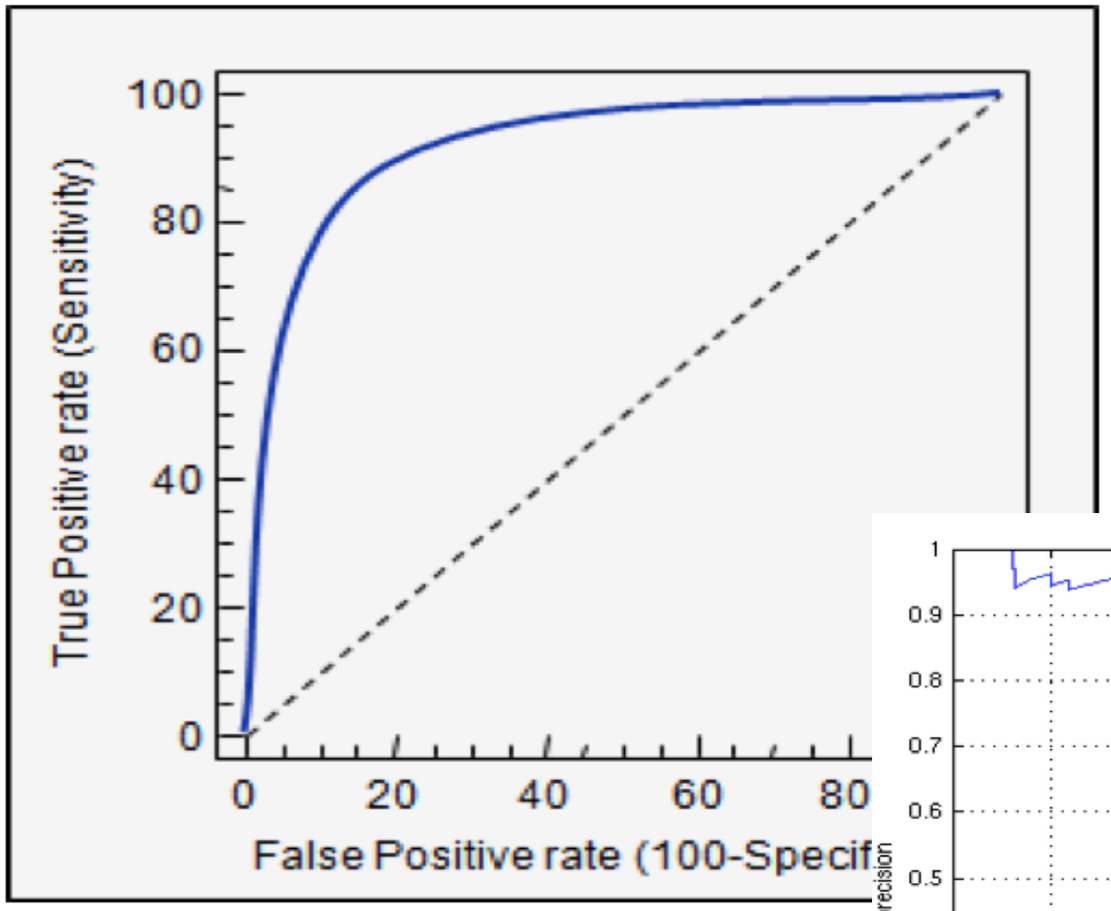
1. Get Train/Test Data
2. Chose set of features
3. Define Figure of Merit function
4. Define Cost function
5. Chose classifier parameters
6. Train
7. Evaluate
8. Repeat

Confusion matrix

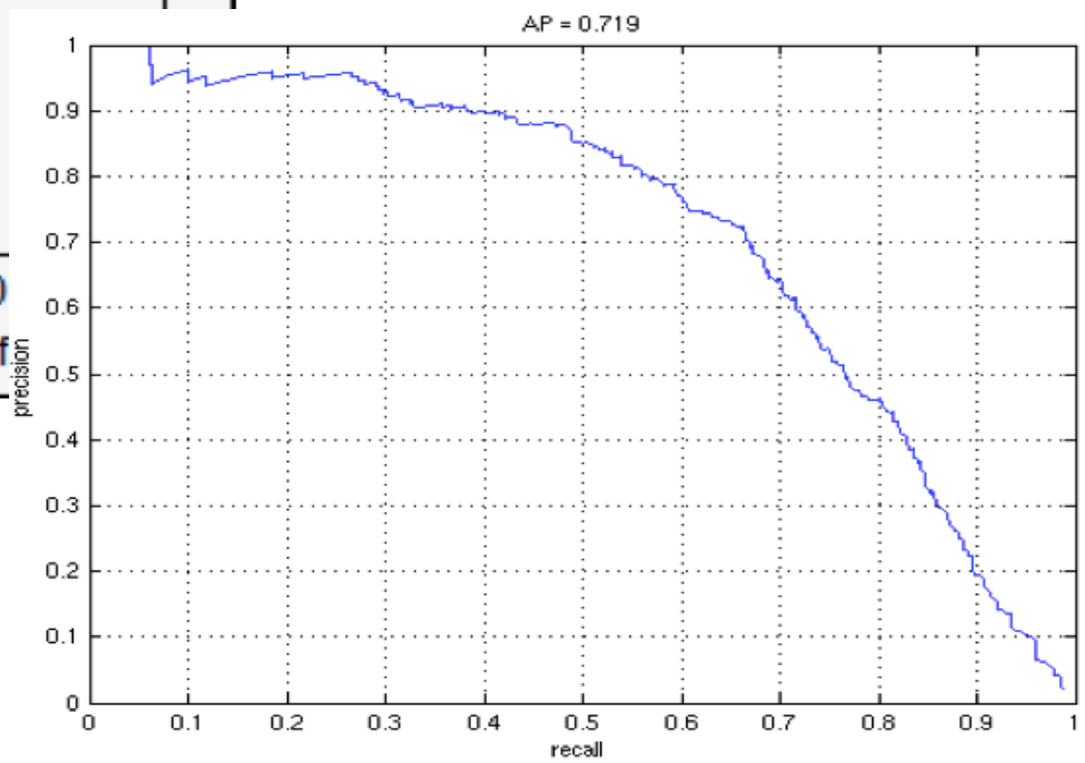




Quality Measures - 1



ROC -- AUC

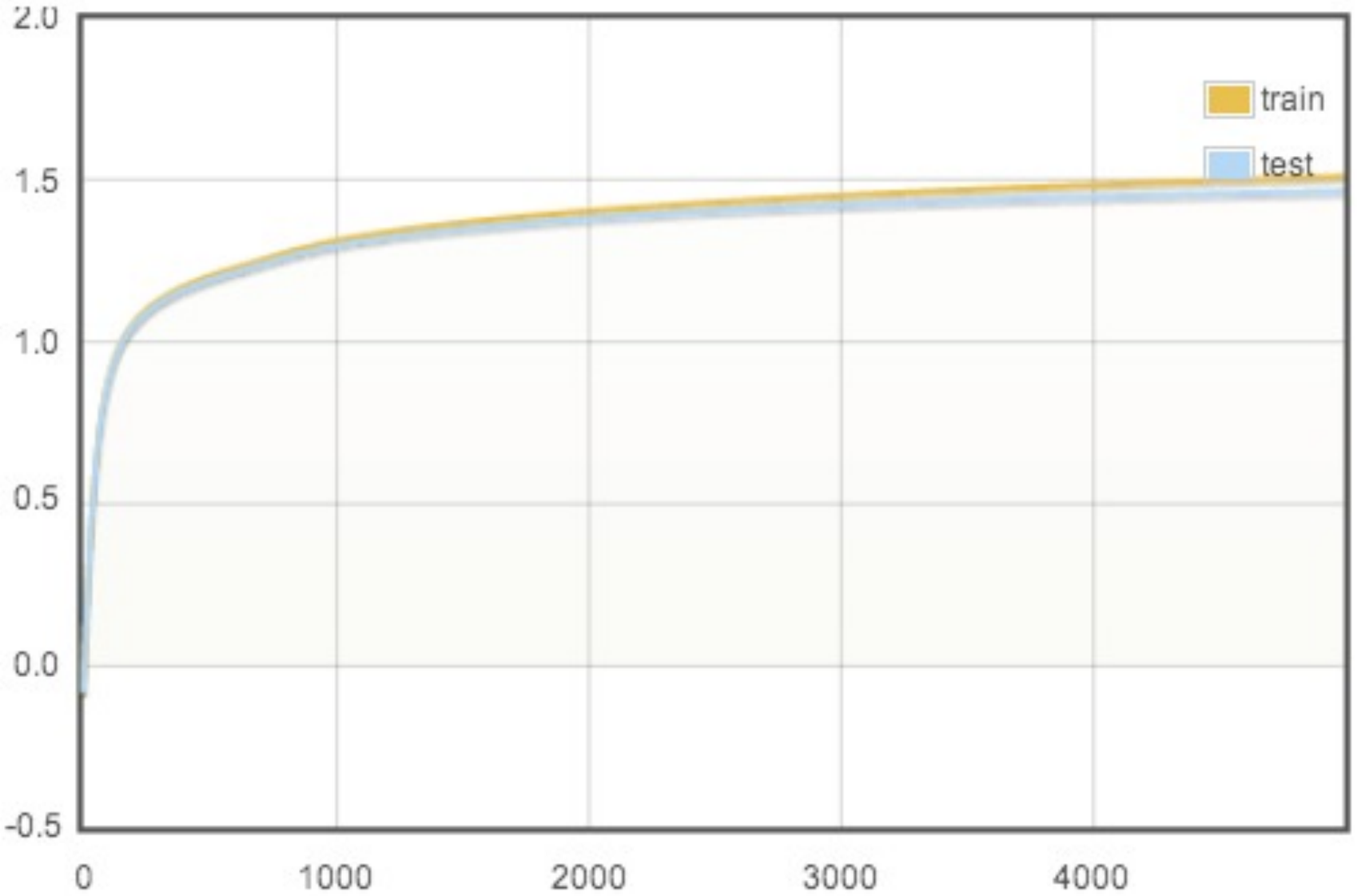


Precision/Recall -- BEP

Quality Measures - 2

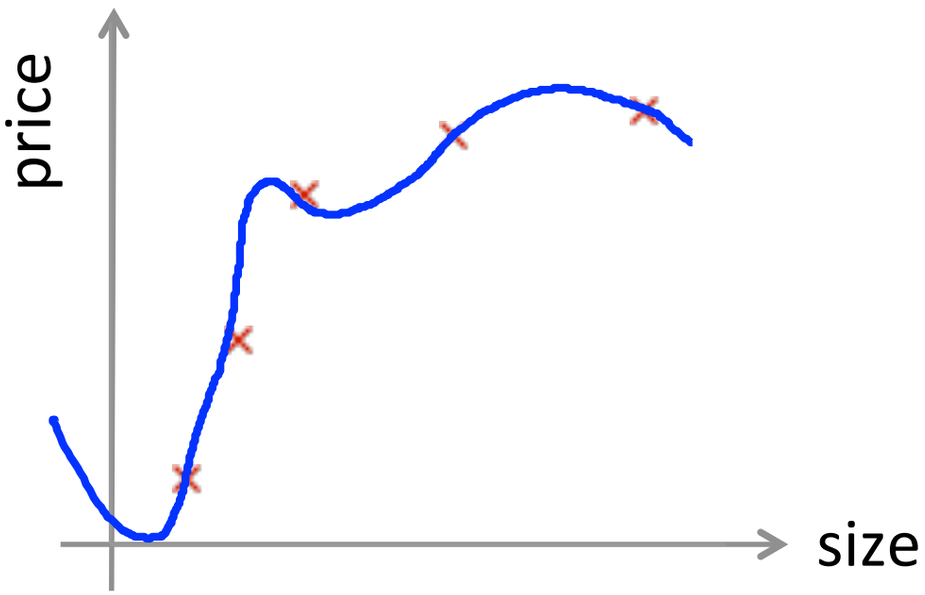
1. F-Score = 2 * Precision x Recall / (Precision + Recall)
2. LogLikelihood = sum {log P}
   - Convex function with derivatives
   - Used as a proxy for non-continuous functions like AUC/ BEP etc

Training diagnostics. Learning curve



Training problems

Overfitting example



$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

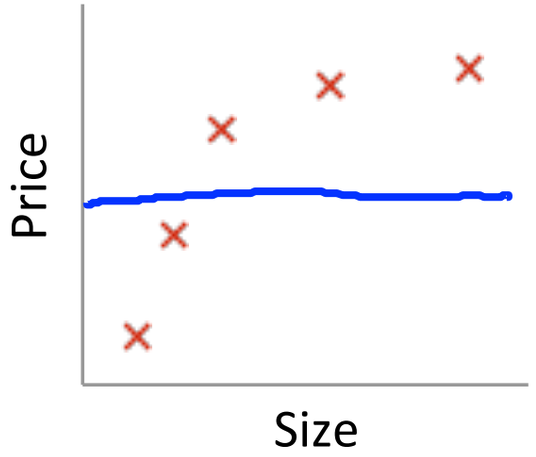
Once parameters $\theta_0, \theta_1, \ldots, \theta_4$ were fit to some set of data (training set), the error of the parameters as measured on that data (the training error $J(\theta)$) is likely to be lower than the actual generalization error.

Заголовок (не длинней одной строки)

Linear regression with regularization

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^{m} \theta_j^2$$

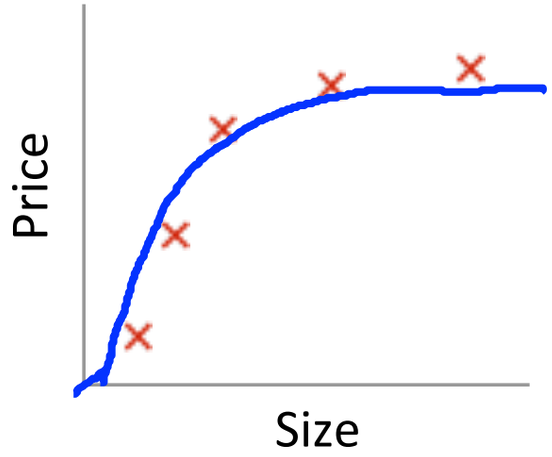


Large λ

High bias (underfit)

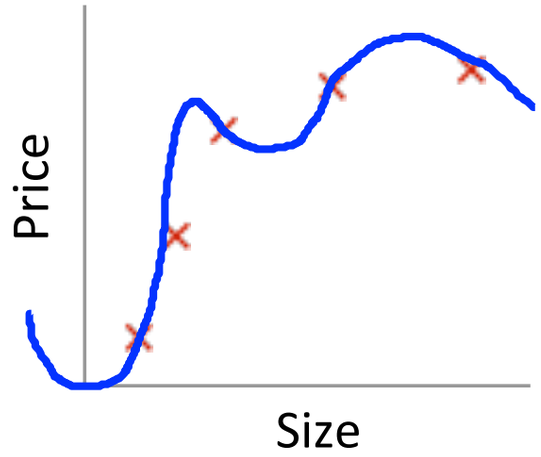
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$



Intermediate λ

"Just right"



Small λ

High variance (overfit)

$\lambda = 0$

Заголовок (не длинней одной строки)

Choosing the regularization parameter λ

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

$$J_{test}(\theta) = \frac{1}{2m_{test}}\sum_{i=1}^{m_{test}}(h_\theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

J(θ)

J train

J cv

J test

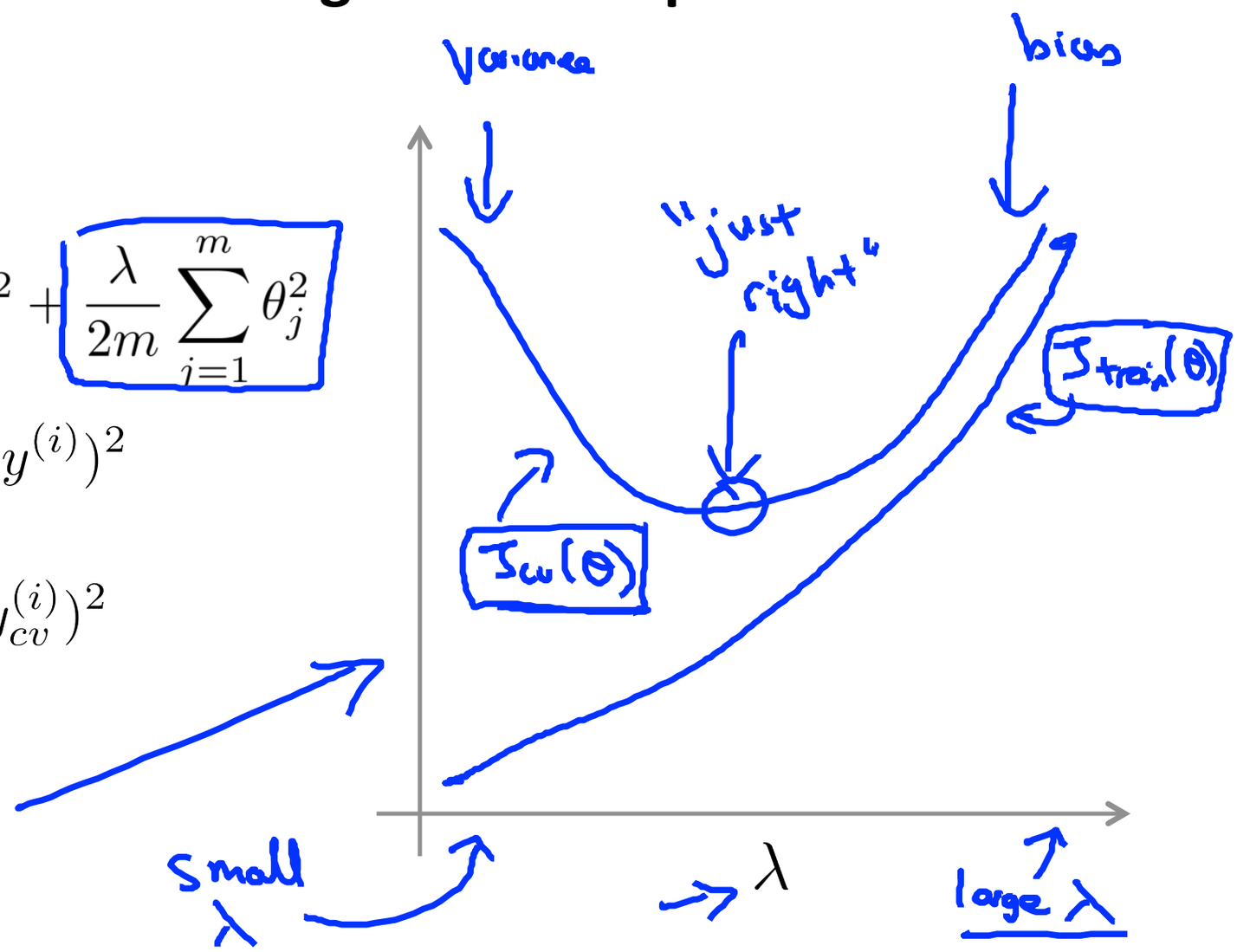
Заголовок (не длинней одной строки)

Bias/variance as a function of the regularization parameter λ

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



Debugging algorithm

1. Get more training examples
2. Try smaller sets of features
3. Try getting additional features
4. Try adding polynomial features
5. Try decreasing regularization (λ)
6. Try increasing (λ)

The Alchemist.

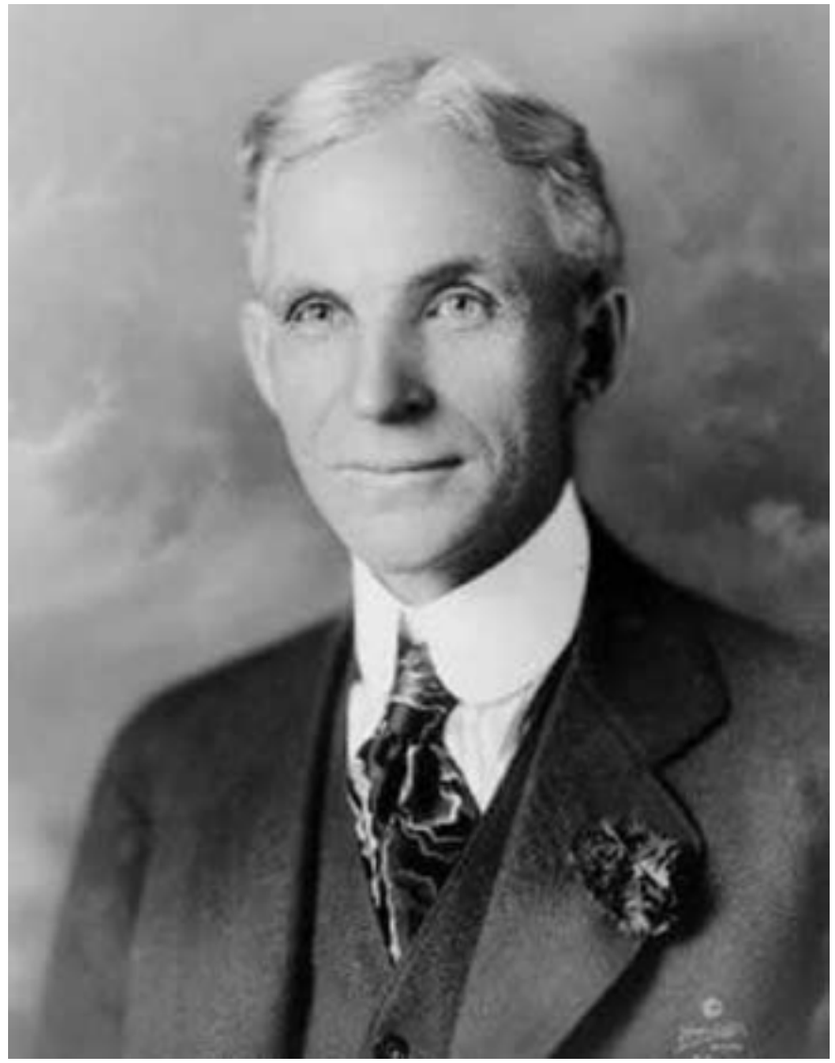
Event Filter Demo

Some links to follow

1. http://bigml.com
2. http://about.wise.io/
3. http://scikit-learn.org/stable/
4. http://orange.biolab.si/
5. http://tmva.sourceforge.net/
6. R
7. Coursera Machine Learning course
   (thanks to Andew Ng for a couple of slides)

Another meta transition awaiting...

«How research & learning can be automated?»



Machine Learning Way

Yandex	**CERN**
Gathering data for testing, learning, verification	
Logs, user models	Experimental data, simulation
Learning	
MatrixNet	Cut-based analysis, TMVA
Application	
Automatic	Manual mode

Machine Learning Way, Continued

Yandex	**CERN**
Analytics, quality monitoring	
Quality metric definition, automated verification & monitoring	Manual mode
Feature assembly line	
Yes	Manual mode

IPython demo.

Every 14 minutes,
somewhere in the world,
an ad exec strides on stage
with breathless declaration:

Every 14 minutes,
somewhere in the world,
an ad exec strides on stage
with breathless declaration:

«Data is the new oil!»

Yandex

Andrey Ustyuzhanin

anaderi@yandex-team.ru

Thank you