Introduction to Data Science (for physics)

Outline of the course:

1. Statistics and Data Analysis
2. Multivariate Techniques and Machine Learning
3. Physics Modeling, Simulation and Monte Carlo Methods
4. Regression, Classification, Clustering and Retrieval

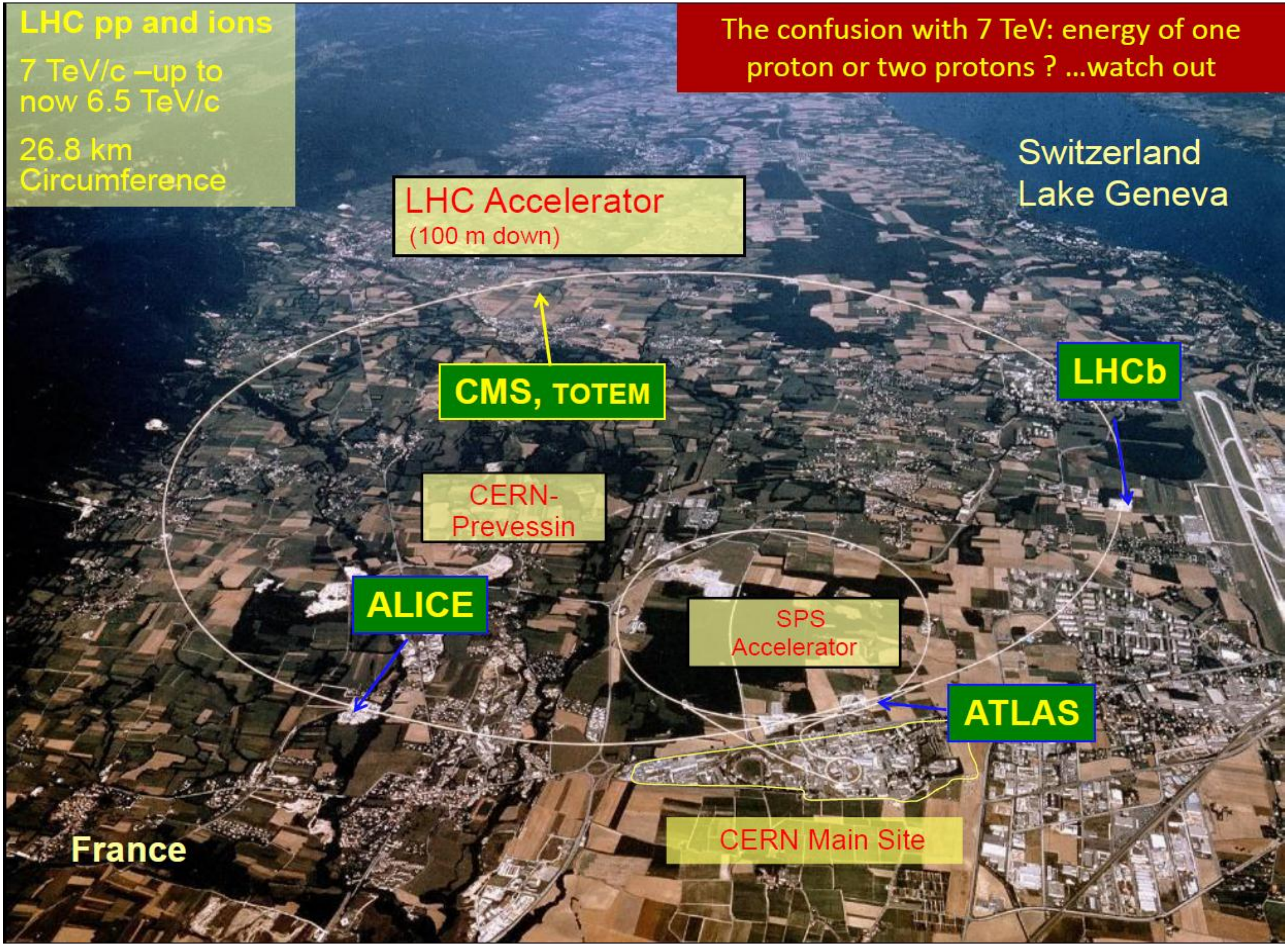
First three parts will focus on applications in physics (mostly in High Energy Physics)

The last part will discuss more typical „Data Science” problems and solutions.

Acknowledgement: slides below „borrowed” fron different courses in HEP and Data Science.

Prof. dr hab. Elżbieta Richter-Wąs

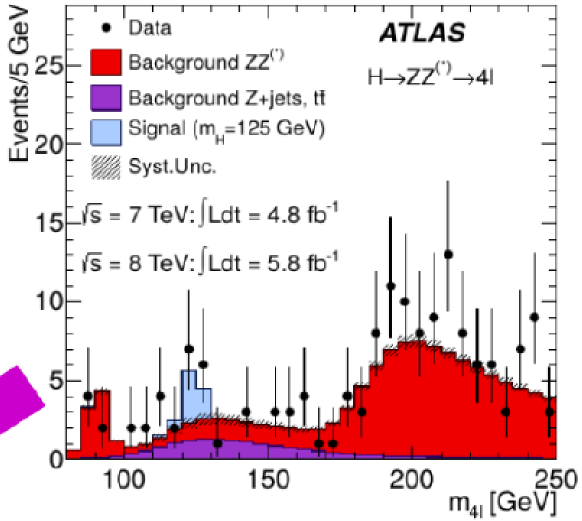
Part 1: Statistics and Data Analysis



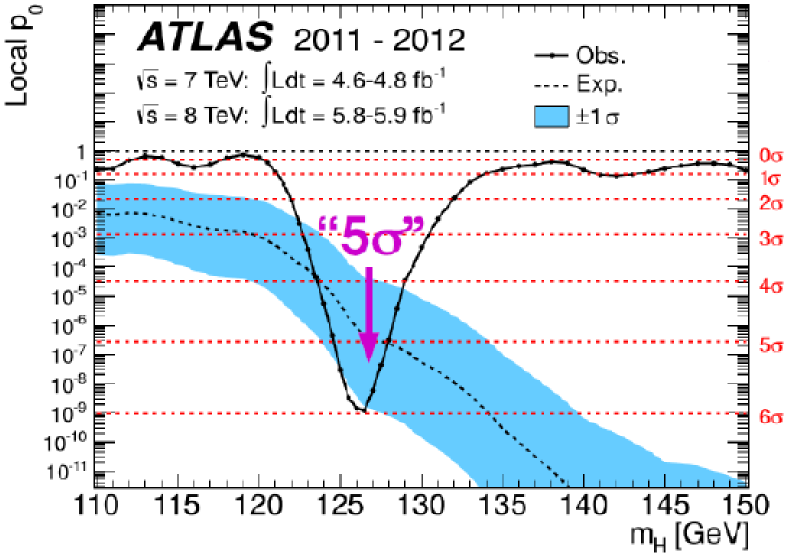
Part 1: Statistics and Data Analysis

Statistical methods play a critical role in many areas of physics

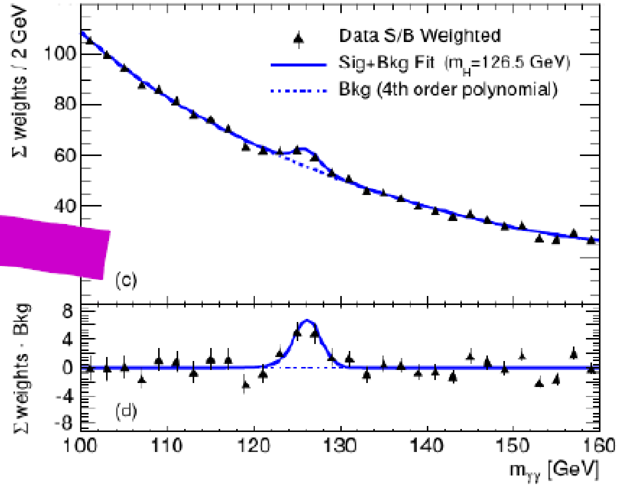
Higgs discovery : "We have 5σ" !



Events/5 GeV
● Data
🟥 Background ZZ$^{(*)}$
🟪 Background Z+jets, $t\bar{t}$
🟦 Signal (m_H=125 GeV)
▨ Syst.Unc.
ATLAS
$H \rightarrow ZZ^{(*)} \rightarrow 4l$
$\sqrt{s}$ = 7 TeV: $\int Ldt$ = 4.8 fb^{-1}
$\sqrt{s}$ = 8 TeV: $\int Ldt$ = 5.8 fb^{-1}
0
5
10
15
20
25
100
150
200
250
m_{4l} [GeV]



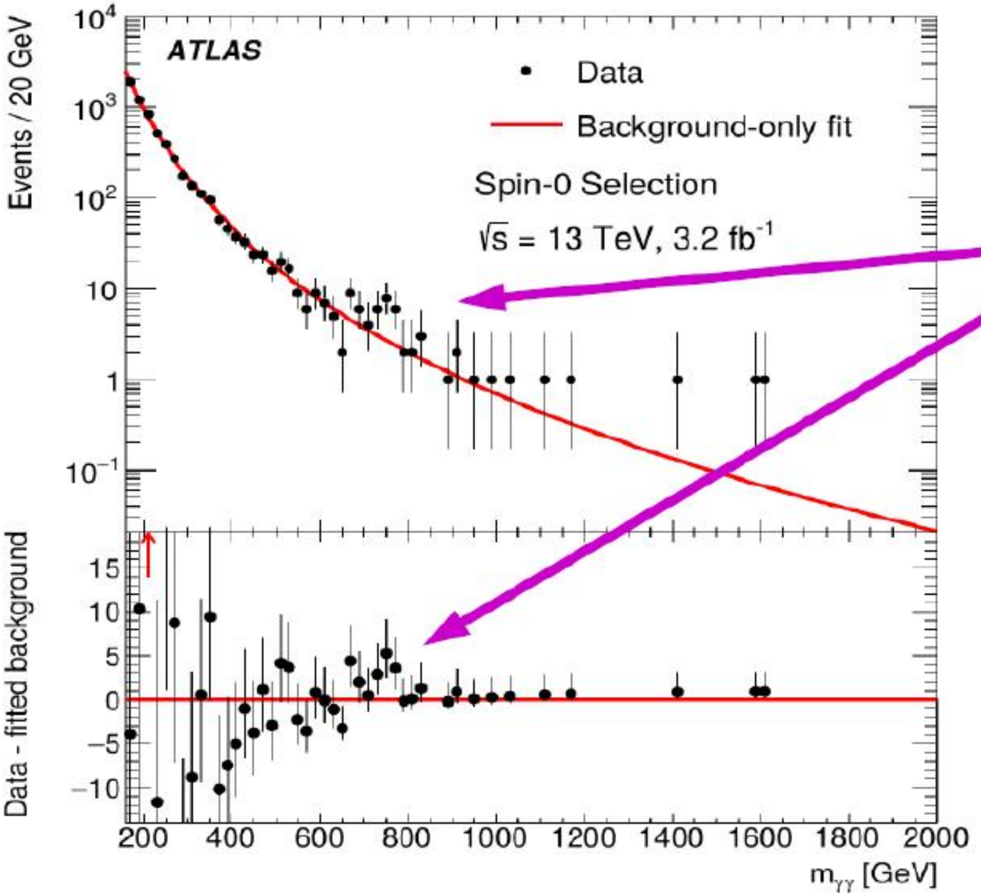
Phys. Lett. B 716 (2012) 1-29



From N. Berger, CERN Summer School, 2019

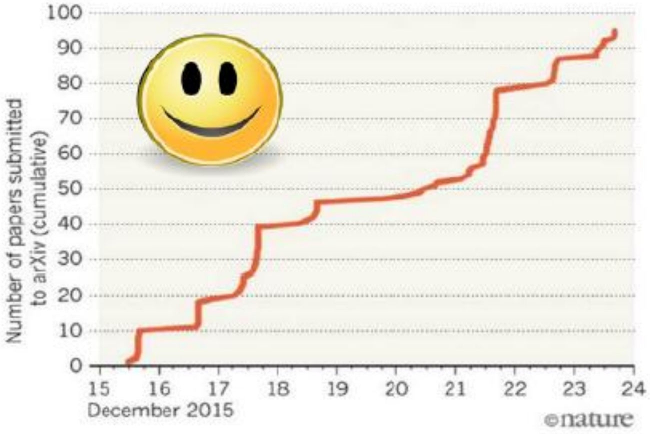
Part 1: Statistics and Data Analysis

Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**...



New Physics ?

3.9σ ? 2.1σ ?

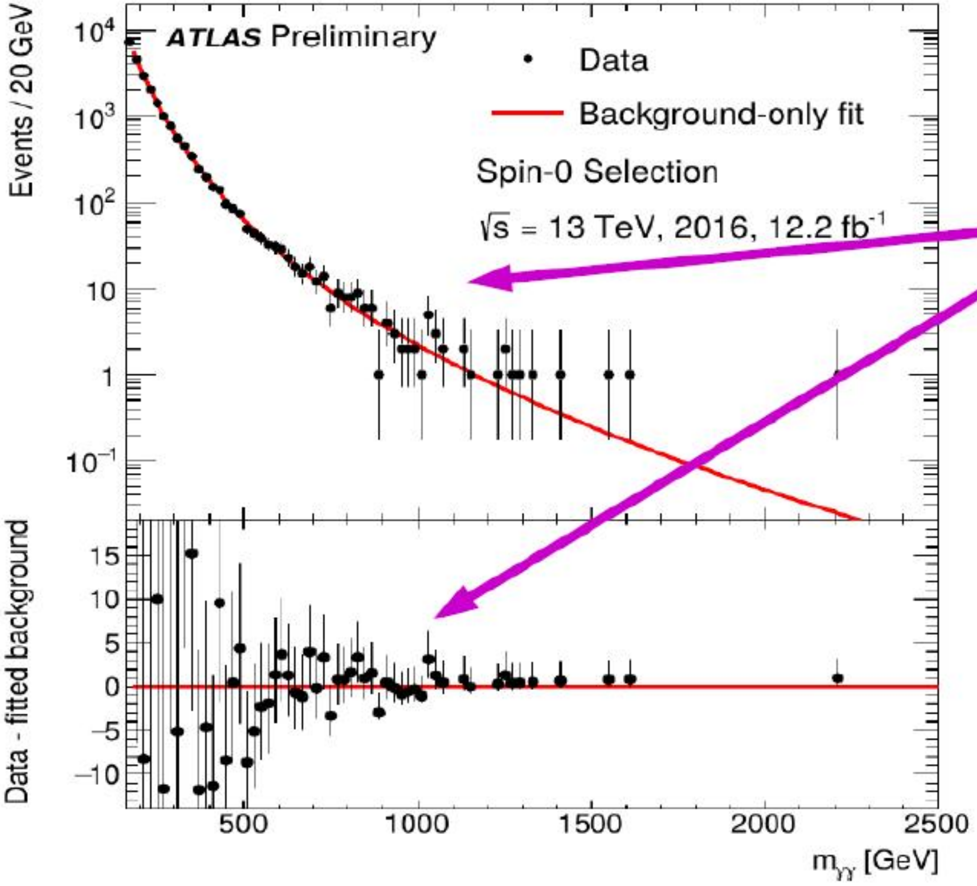


JHEP 09 (2016) 1

From N. Berger, CERN Summer School, 2019

Part 1: Statistics and Data Analysis

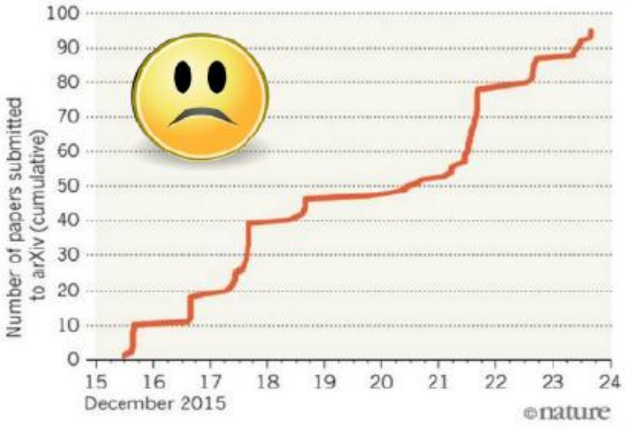
Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**…



A few months later…

~~**New Physics ?**~~

~~3.9σ ? 2.1σ ?~~



JHEP 09 (2016) 1

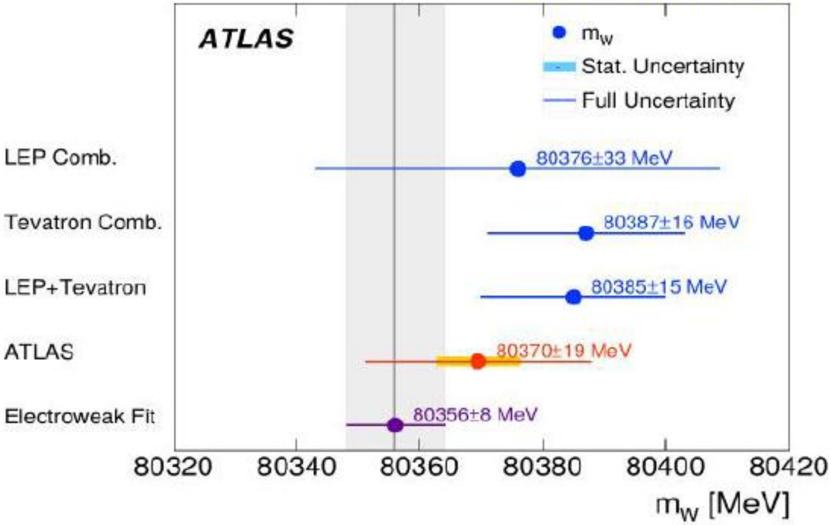
From N. Berger, CERN Summer School, 2019

Part 1: Statistics and Data Analysis

Many important questions answered by **precision measurements**, especially if no new peaks found at high mass…

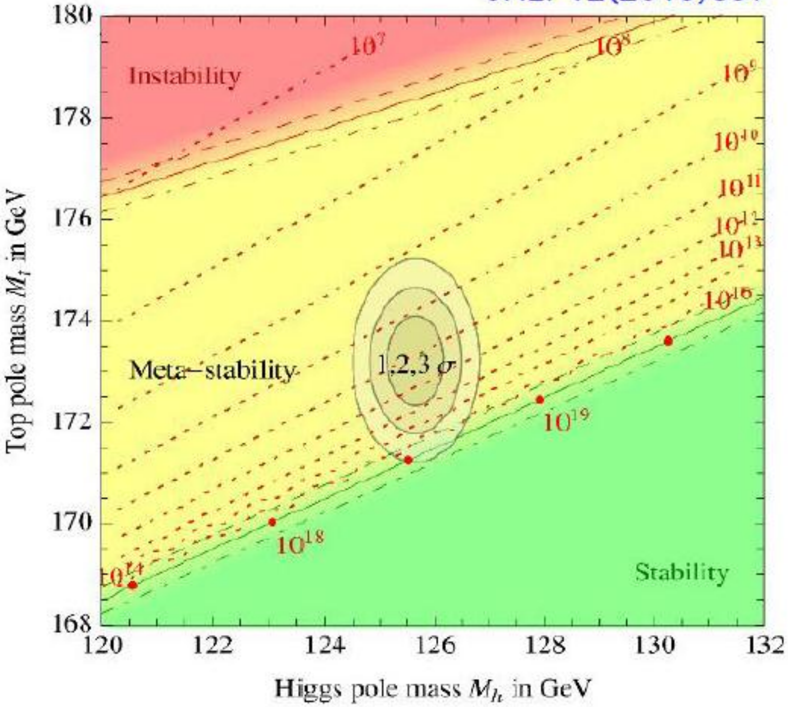
Key point = determination of **uncertainties**

arXiv:1701.07240



Consistency of the SM…

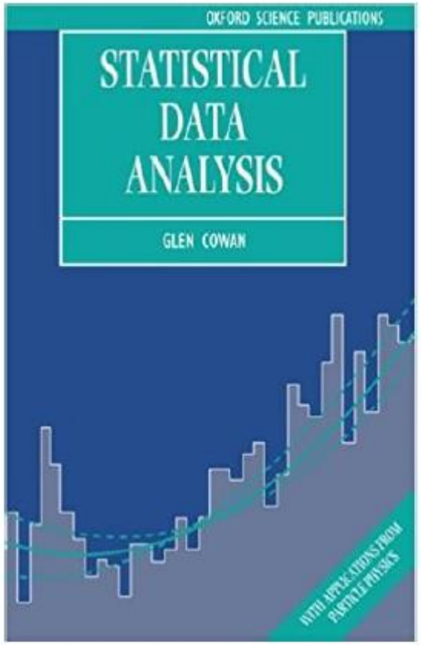
JHEP12(2013)089

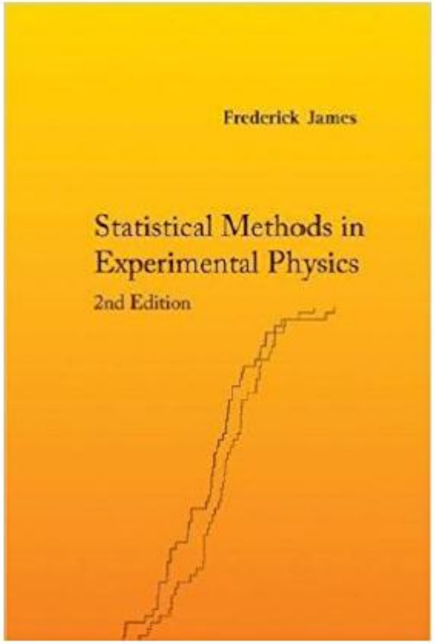


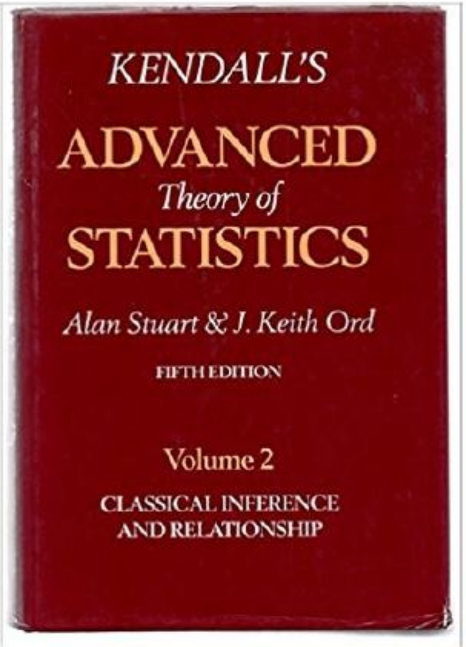
… or the fate of the universe

From N. Berger, CERN Summer School, 2019

Part 1: Statistics and Data Analysis


OXFORD SCIENCE PUBLICATIONS
STATISTICAL DATA ANALYSIS
GLEN COWAN
WITH APPLICATIONS FROM PARTICLE PHYSICS




KENDALL'S
ADVANCED Theory of STATISTICS
Alan Stuart & J. Keith Ord
FIFTH EDITION
Volume 2
CLASSICAL INFERENCE AND RELATIONSHIP

Some other courses available online:

Glen Cowan's Cours d'Hiver and 2010 CERN Academic Training lectures

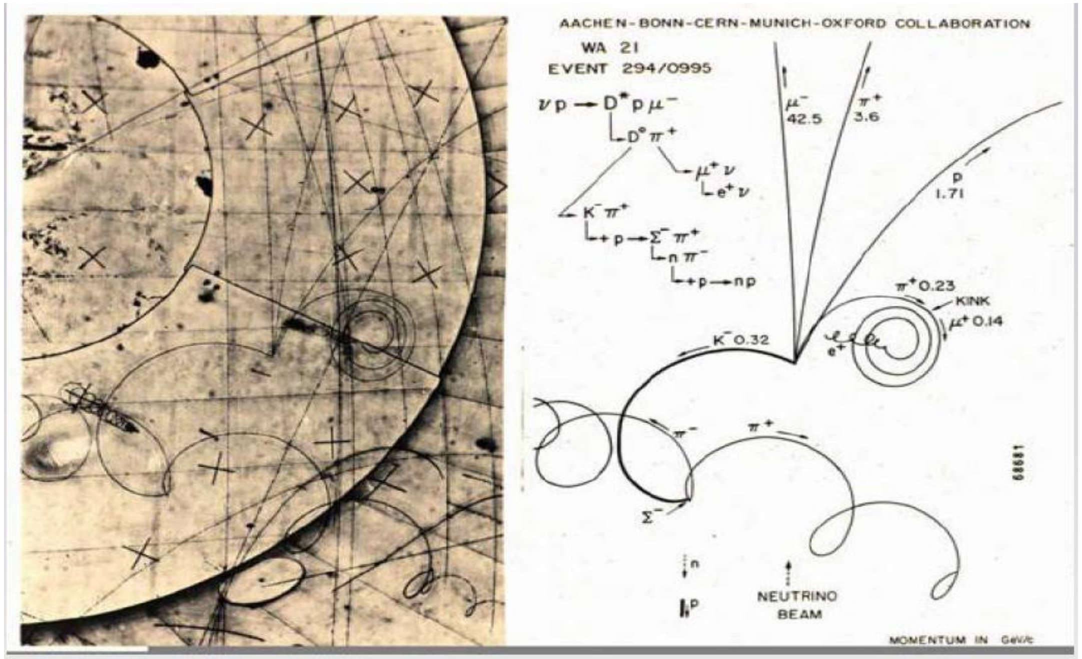
Kyle Cranmer's CERN Academic Training lectures

Louis Lyons'and Lorenzo Moneta's CERN Academic Training Lectures

Part 2: Multivariate Analysis and Machine Learning

In HEP everything started multivariate.

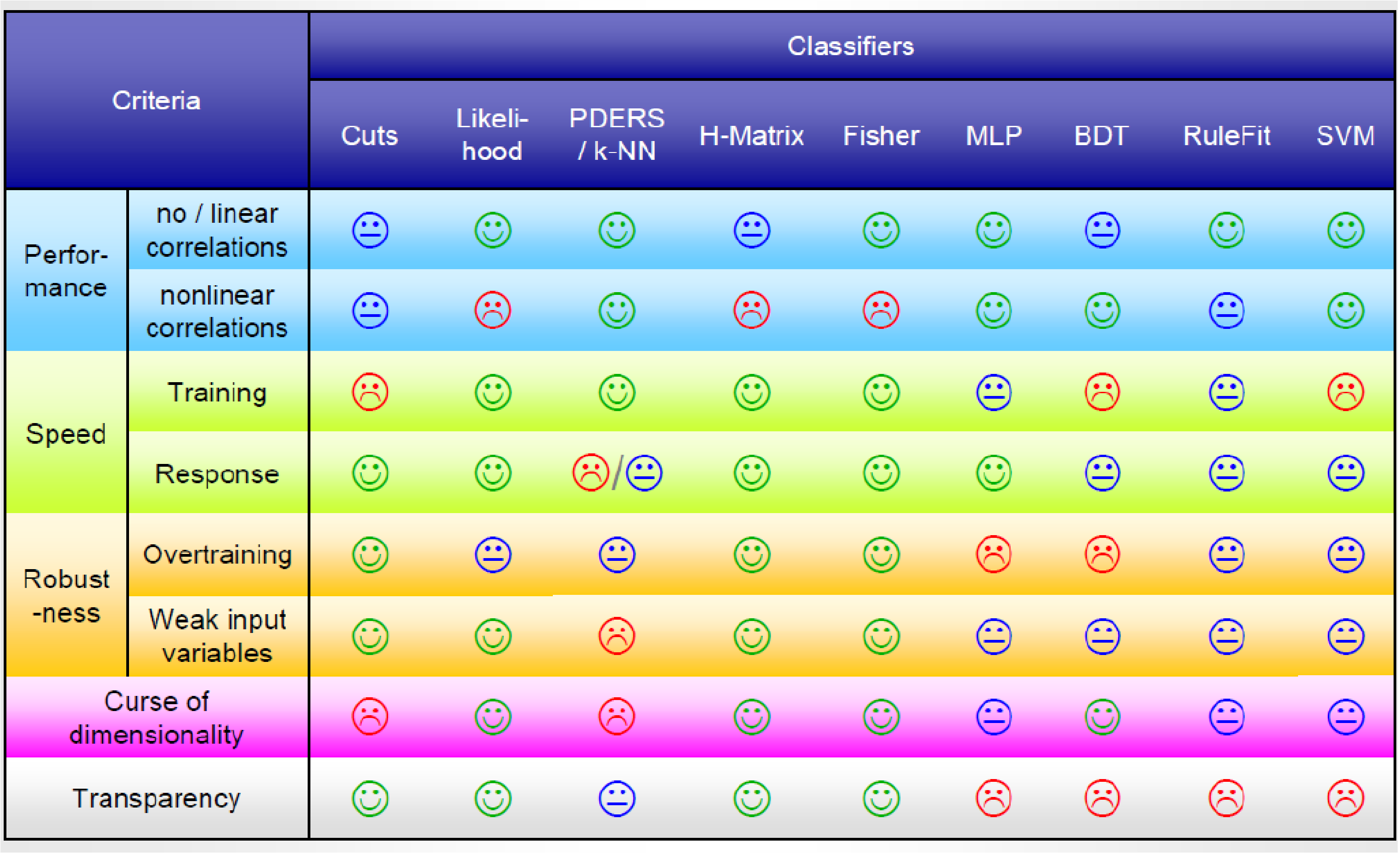
Below: inteligent „Multivariate Pattern Recognition" used to identify particles



Nowdays: let computer help you.

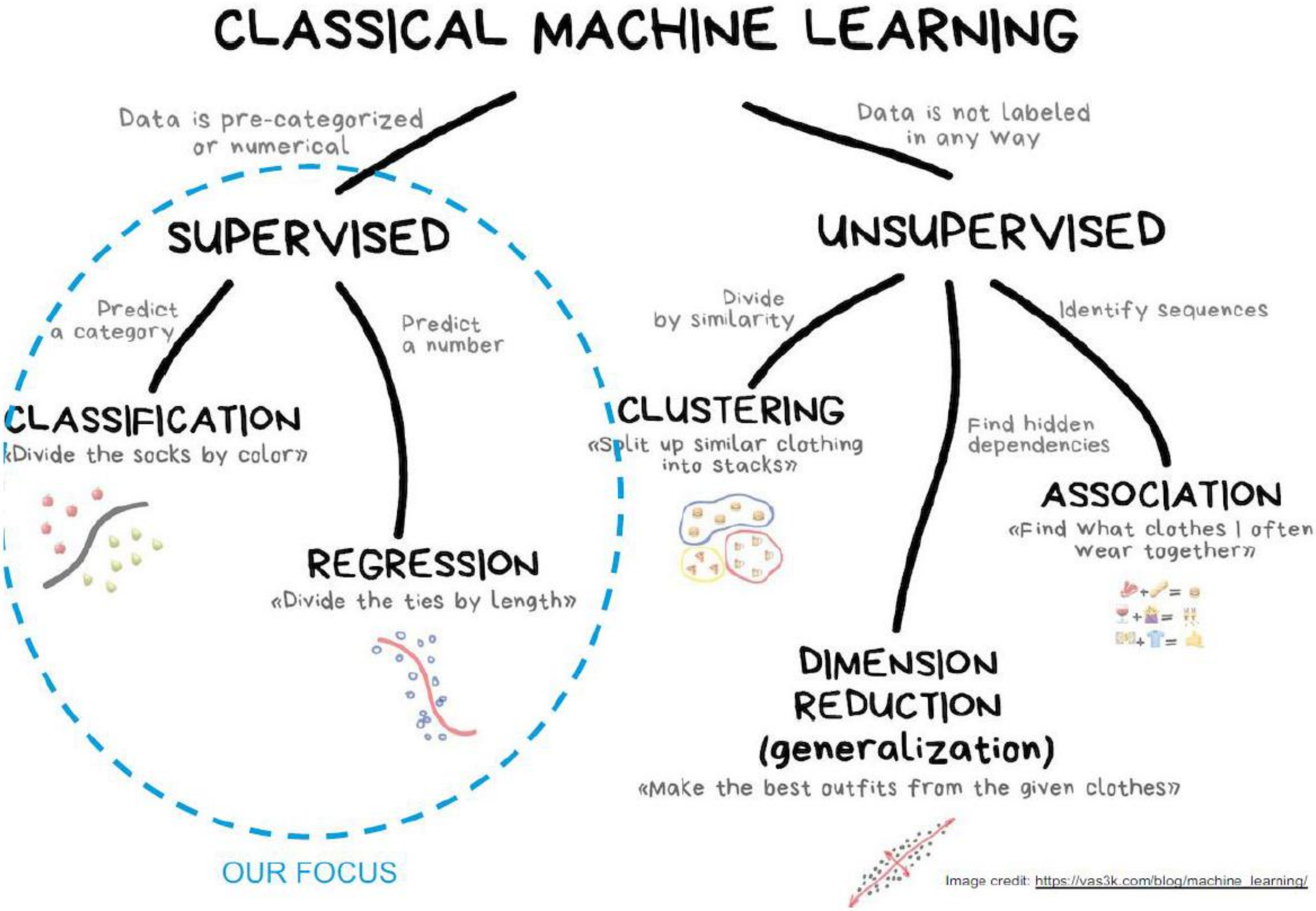
Classifiers and their properties

H. Voss, Multivariate Data Analysis and Machine Learning in High Energy Physics
http://tmva.sourceforge.net/talks.shtml



Criteria		Classifiers								
		Cuts	Likeli-hood	PDERS / k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
Perfor-mance	no / linear correlations	😐	☺	☺	😐	☺	☺	😐	☺	☺
	nonlinear correlations	😐	☹	☺	☹	☹	☺	☺	😐	☺
Speed	Training	☹	☺	☺	☺	☺	😐	☹	😐	☹
	Response	☺	☺	☹/😐	☺	☺	☺	😐	😐	😐
Robust-ness	Overtraining	☺	😐	😐	☺	☺	☹	☹	😐	😐
	Weak input variables	☺	☺	☹	☺	☺	😐	😐	😐	😐
Curse of dimensionality		☹	☺	☹	☺	☺	😐	☺	😐	😐
Transparency		☺	☺	😐	☺	☺	☹	☹	☹	☹

Classical Learning


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

OUR FOCUS

Image credit: https://vas3k.com/blog/machine_learning/

Machine Learning

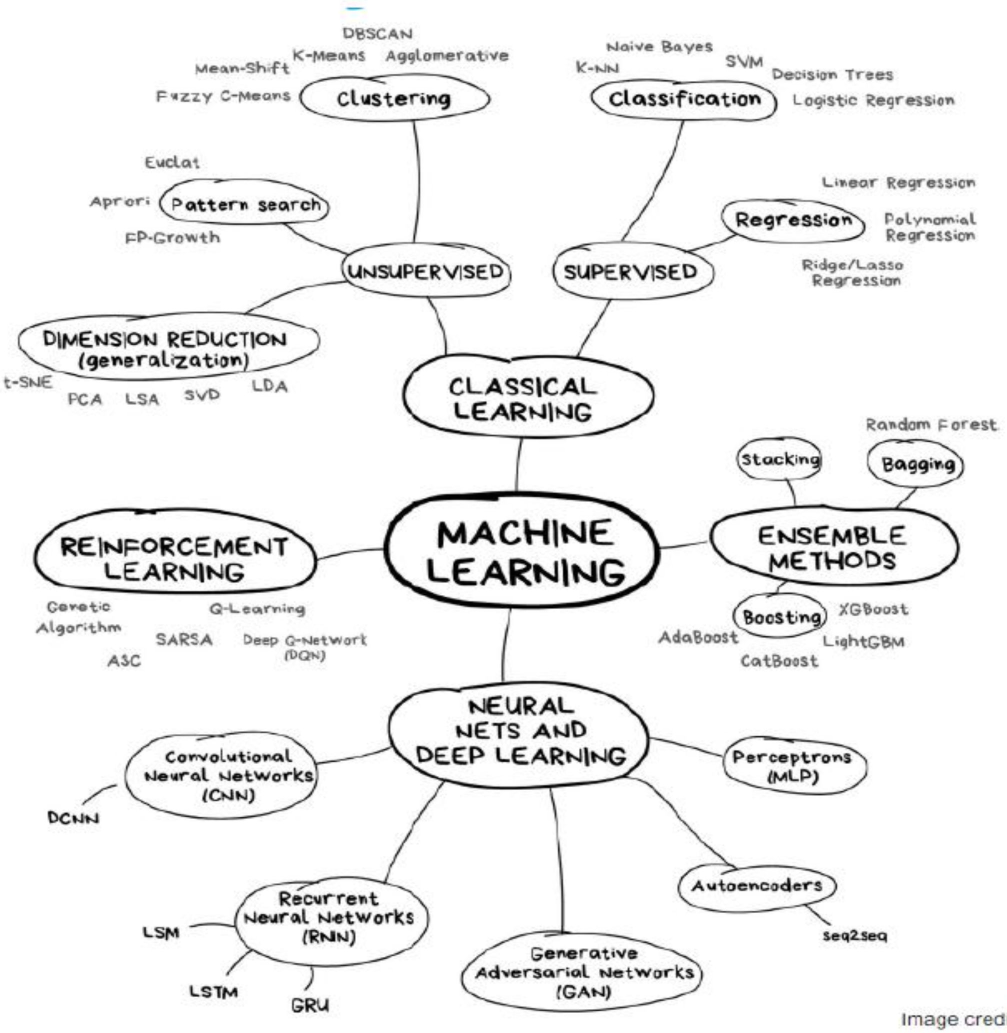


Image credit: https://vas3k.com/blog/machine_learning/

What is the model?

Ceci n'est pas une pomme

- This is not an apple just its graphical representation

Many skills are needed to build a new model, to run it and analyze its results.

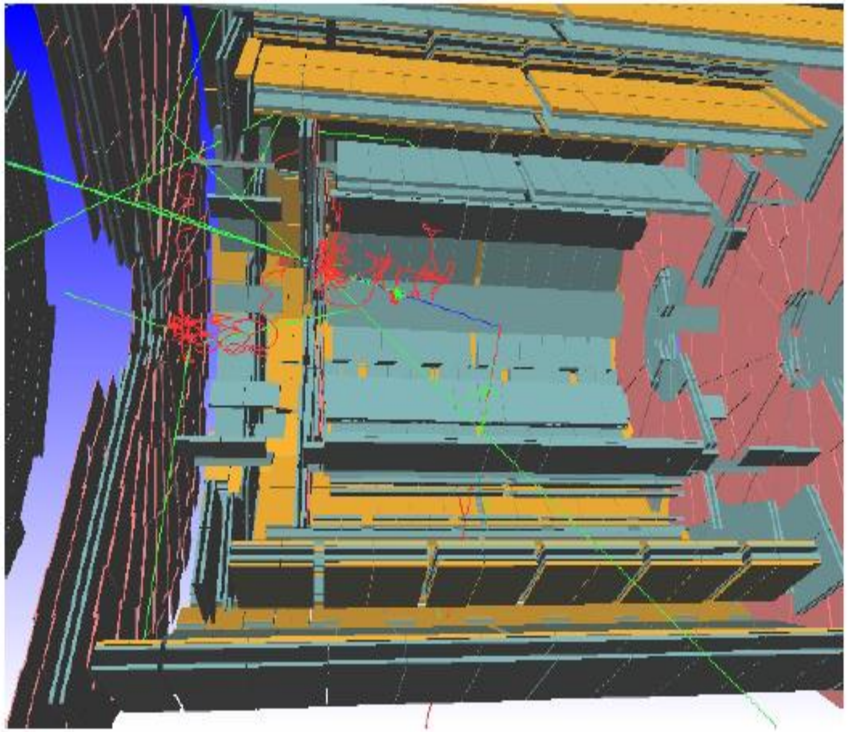
- Computational Science is an emerging, multidisciplinary domain, based on the idea of **“computational thinking”**.
- A computer-based description offers a new language, a new methodology to address scientific challenges, far beyond the scope of traditional numerical methods, and in fields where these classical approaches hardly apply.

B. Chopard et al., coursera lectures, University of Geneva

Part 3: Physics modeling, simulation and and Monte Carlo methods

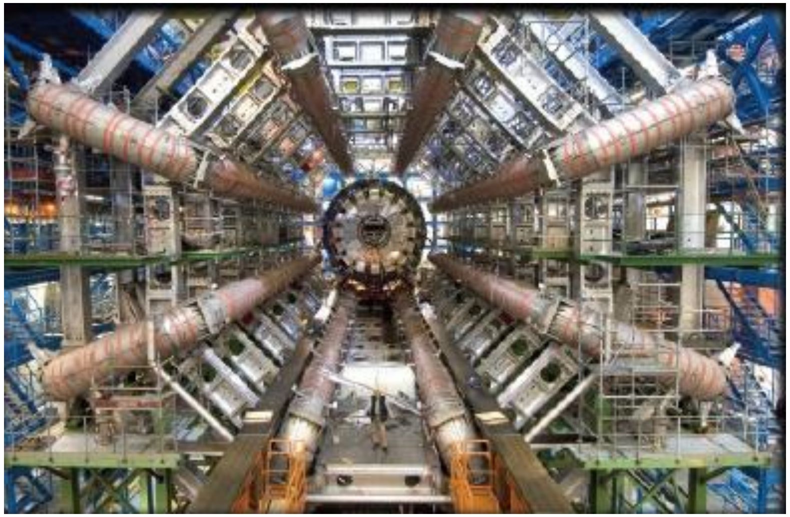
GEANT4

Visualised model of the detector used for simulation



GEANT4 is also used to determine the performance of X-ray and gamma-ray detectors for astrophysics

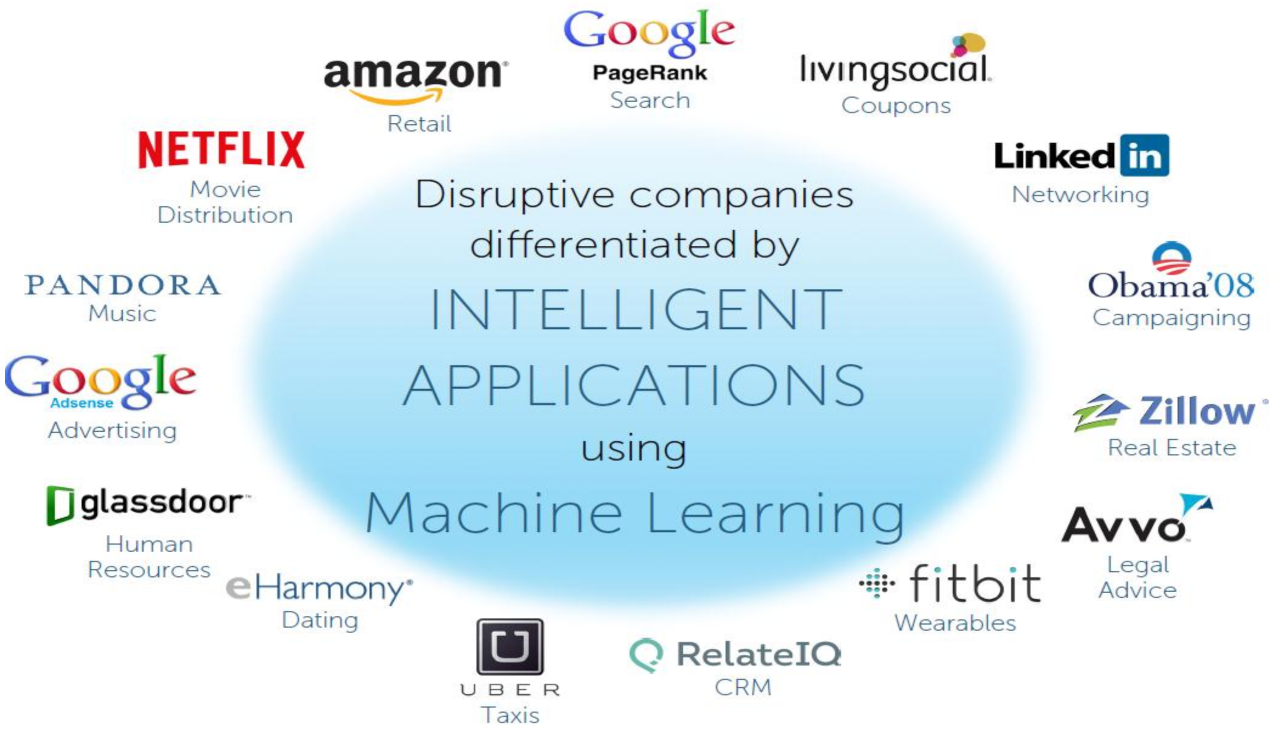
Detector



B. Chopard et al., coursera lectures, University of Geneva

Part 4: Regression, Classification, Clustering

- **Current view on Machine Learning : disruptive inteligent applications are used by leading comercial companies**

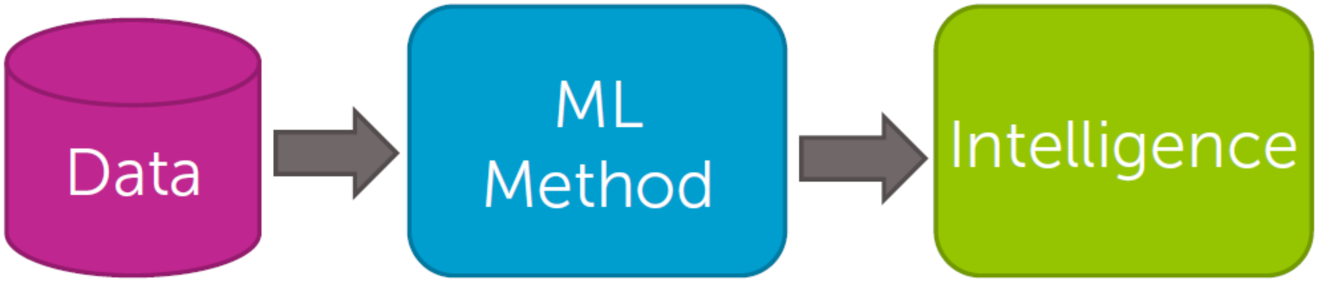


course by E. Fox and C. Guestrin, Univ of Washington

Part 4: Regresion, Classification, Clustering

- **Data → inteligence pipeline**

New kind of analysis
which brings inteligence
how to solve a problem

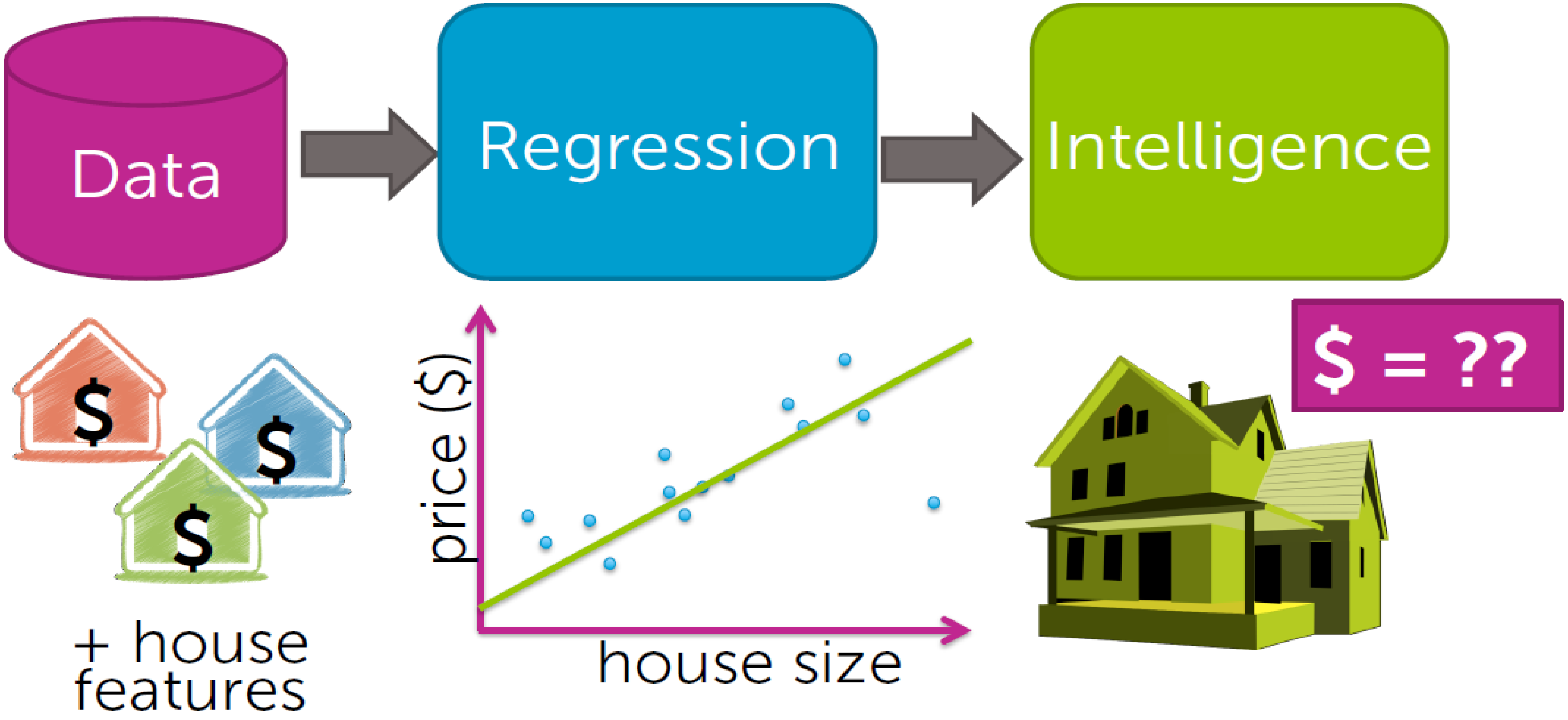


Eg. which product to buy
which film to chose
connect people and taxi driver

course by E. Fox and C. Guestrin, Univ of Washington

Regression

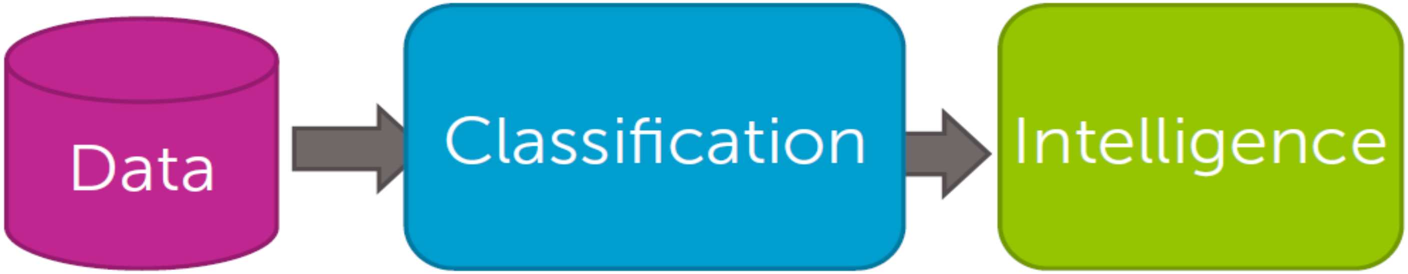
Case study: prediction for the house price



course by E. Fox and C. Guestrin, Univ of Washington

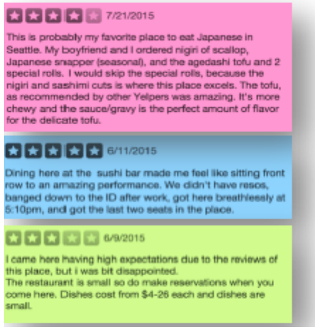
Classification

Case study: Score of the restaurant



Sushi was <u>awesome</u>,
the food was <u>awesome</u>,
but the service was <u>awful</u>.

All reviews:



★★★★☆ 7/21/2015

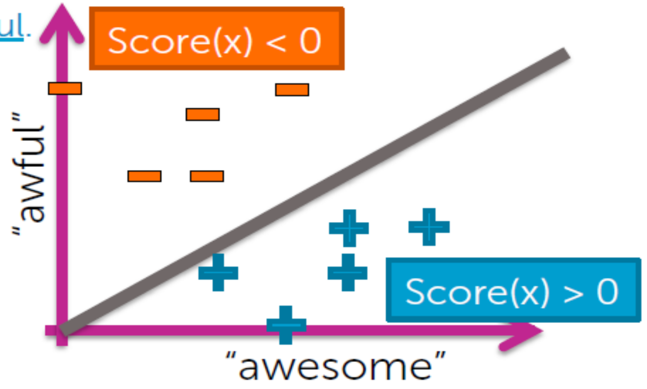
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

★★★★★ 6/11/2015

Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have resos, banged down to the ID after work, got here breathlessly at 5:10pm, andl got the last two seats in the place.

★★★☆☆ 6/9/2015

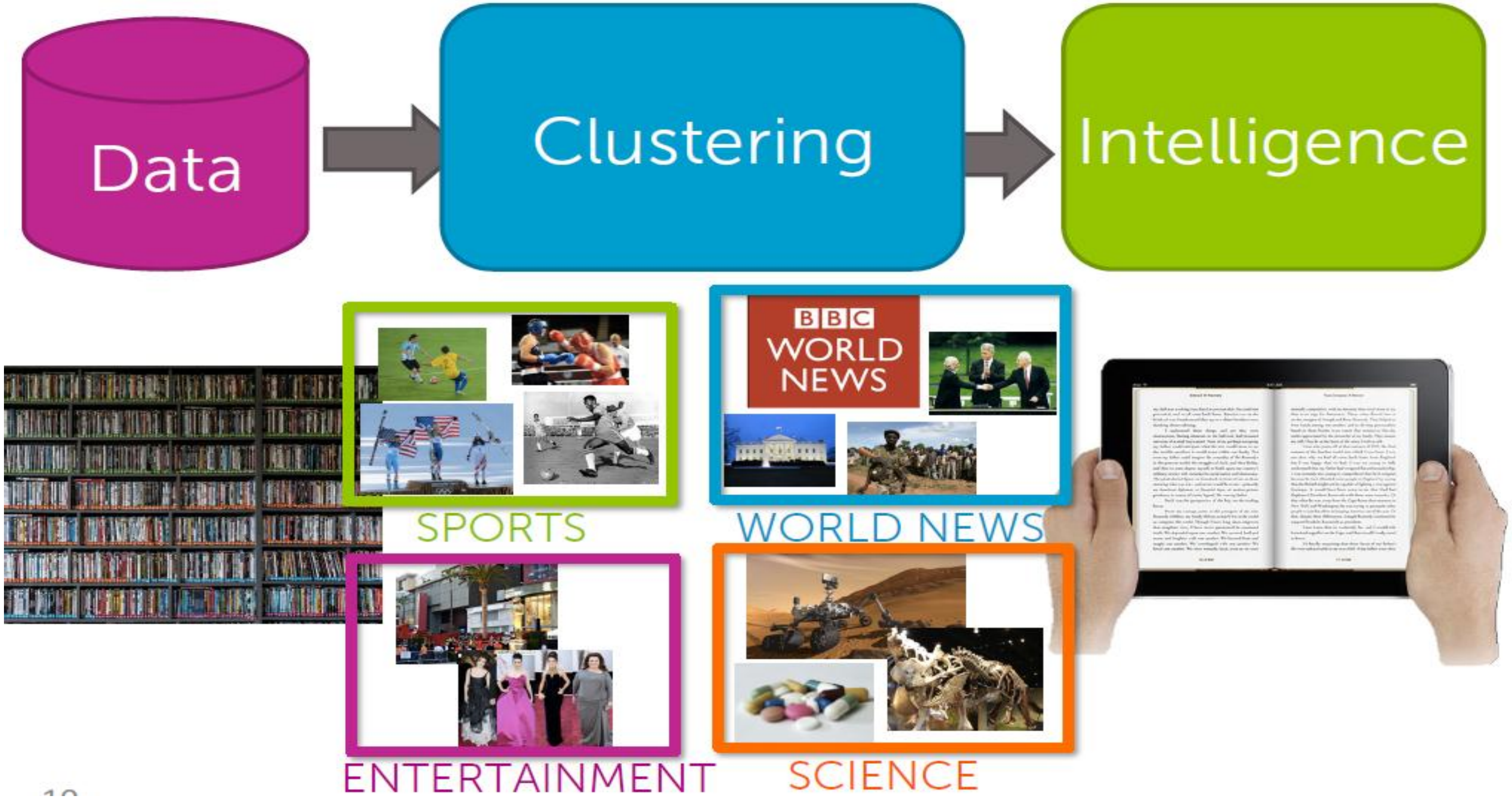
I came here having high expectations due to the reviews of this place, but i was bit disappointed.
The restaurant is small so do make reservations when you come here. Dishes cost from $4-26 each and dishes are small.



course by E. Fox and C. Guestrin, Univ of Washington

Clustering

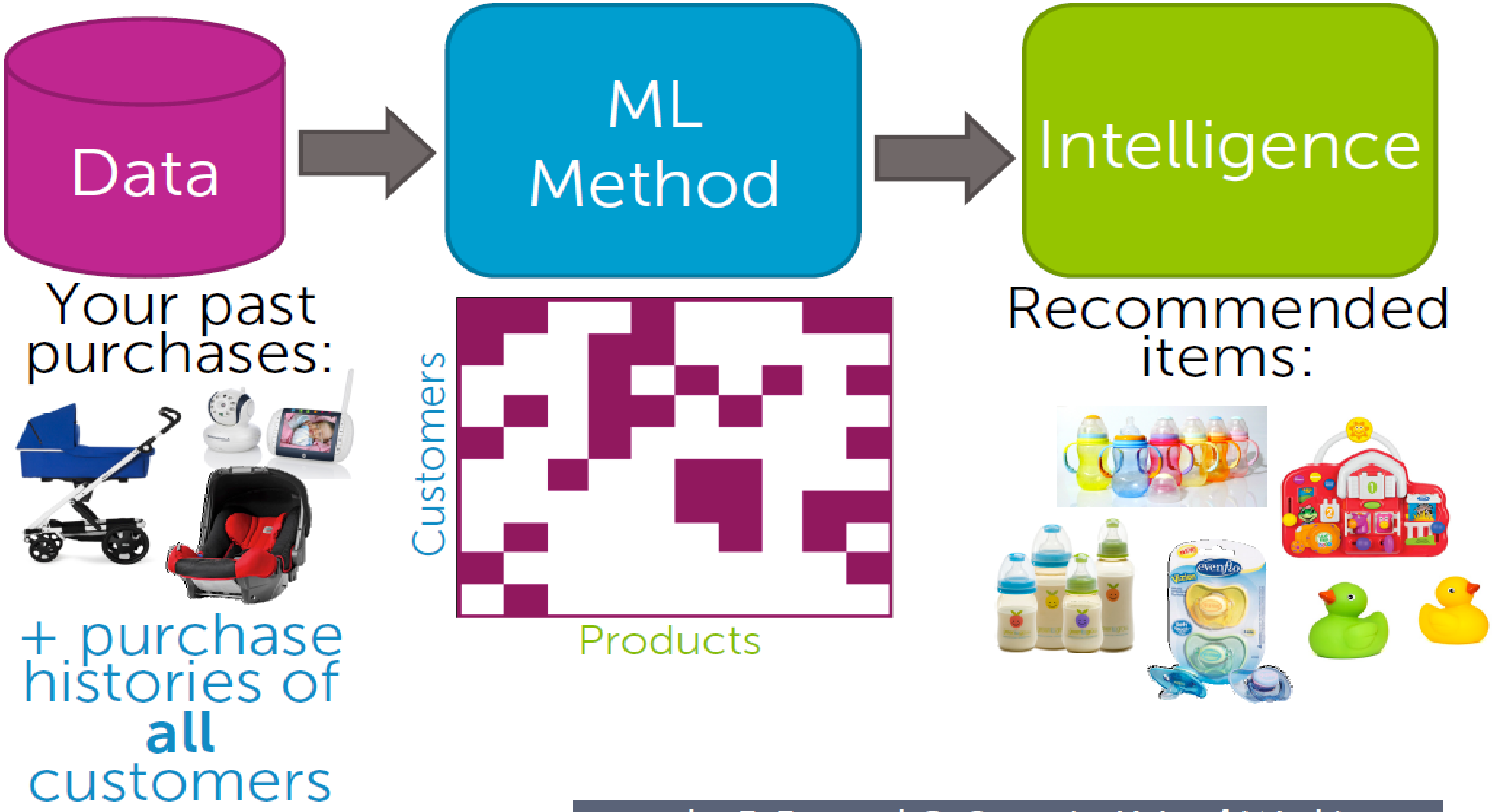
Case study: assigning books to groups by topics



course by E. Fox and C. Guestrin, Univ of Washington

Recommendation

Case study: personalisation of recommending items



Data

ML Method

Intelligence

Your past purchases:

+ purchase histories of **all** customers

Customers

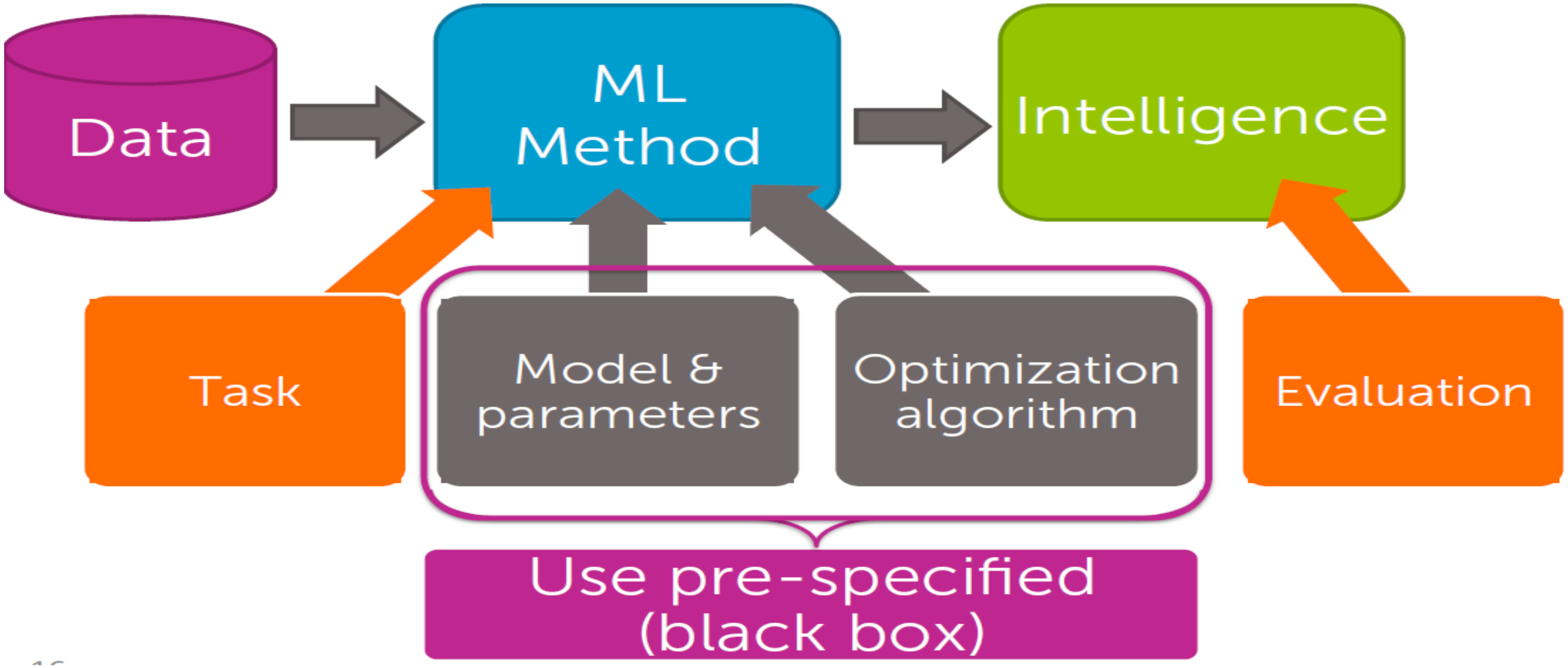
Products

Recommended items:

course by E. Fox and C. Guestrin, Univ of Washington

Deploying inteligence module

Case studied are about building, evaluating, deploying inteligence in data analysis.



course by E. Fox and C. Guestrin, Univ of Washington

Regression: Predicting house prices

Models

- Linear regression
- Regularization:
  Ridge (L2), Lasso (L1)

Algorithms

- Gradient descent
- Coordinate descent

Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

course by E. Fox and C. Guestrin, Univ of Washington

Classification: Sentiment analysis

Models

- Linear classifiers (logistic regression, SVMs, perceptron)
- Kernels
- Decision trees

Algorithms

- Stochastic gradient descent
- Boosting

Concepts

- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning

course by E. Fox and C. Guestrin, Univ of Washington

Clustering: Finding documents

Models

- Nearest neighbors
- Clustering, mixtures of Gaussians
- Latent Dirichlet allocation (LDA)

Algorithms

- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)

Concepts

- Distance metrics, approximation algorithms, hashing, sampling algorithms, scaling up with map-reduce

course by E. Fox and C. Guestrin, Univ of Washington

Getting your ETCs for lectures

- I foresee written exam on the theory part.
- List of topical questions will be available before Xmass break.
- You will be asked to answer 5 questions out of 25-30 on the list.