

Advanced Methods in Data Analysis

Outline of the course:

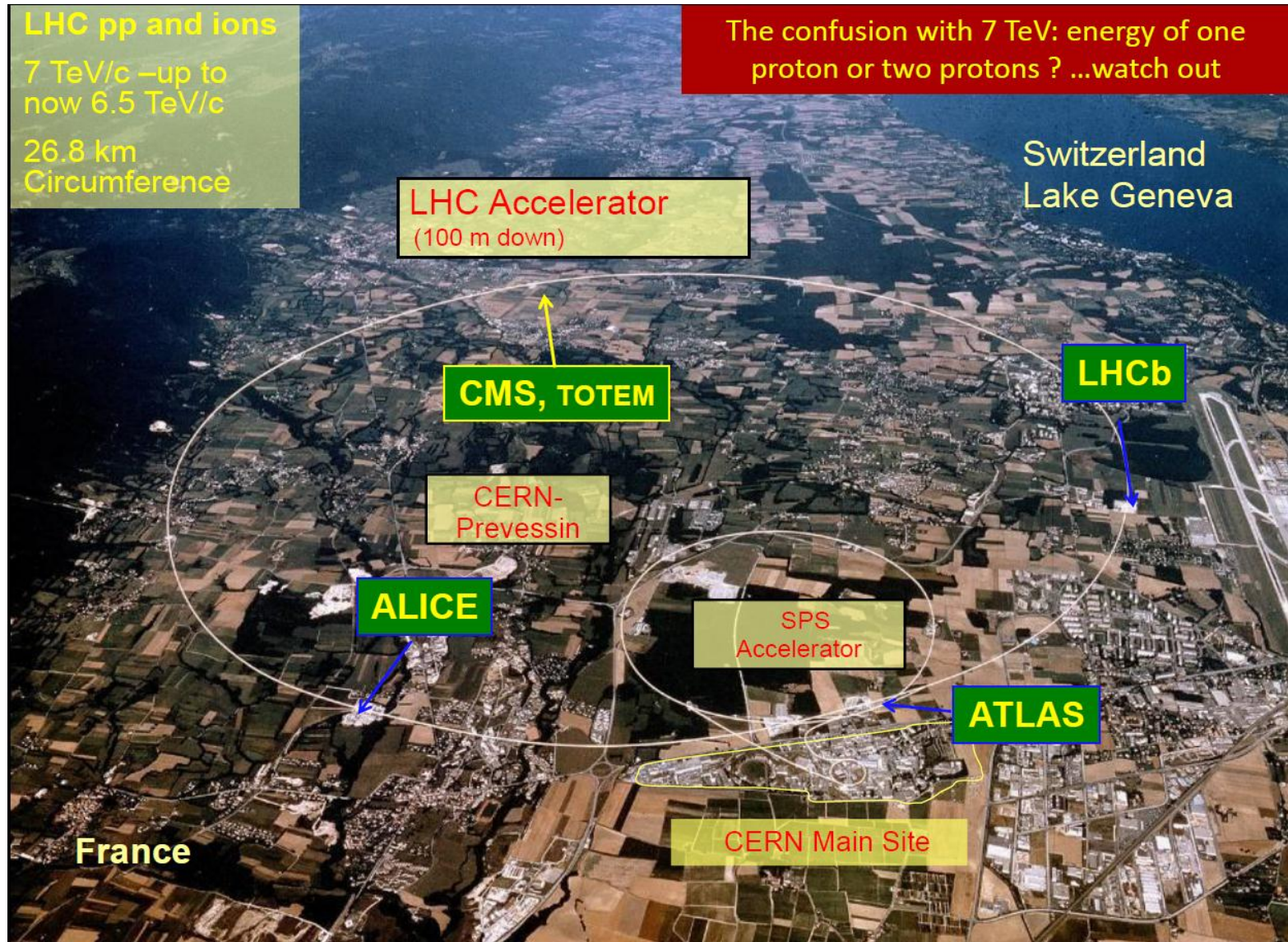
1. **Statistics and Data Analysis**
2. **Multivariate Techniques and Machine Learning**
3. **Physics Modeling, Simulation and Monte Carlo Methods**
4. **Regression, Classification, Clustering and Retrieval**

First three parts will focus on applications in physics, mostly in High Energy Physics.

The last part will cover few typical „Data Science” problems and solutions.

Acknowledgement: slides below „borrowed” from different courses on advanced analysis methods in HEP and Data Science.

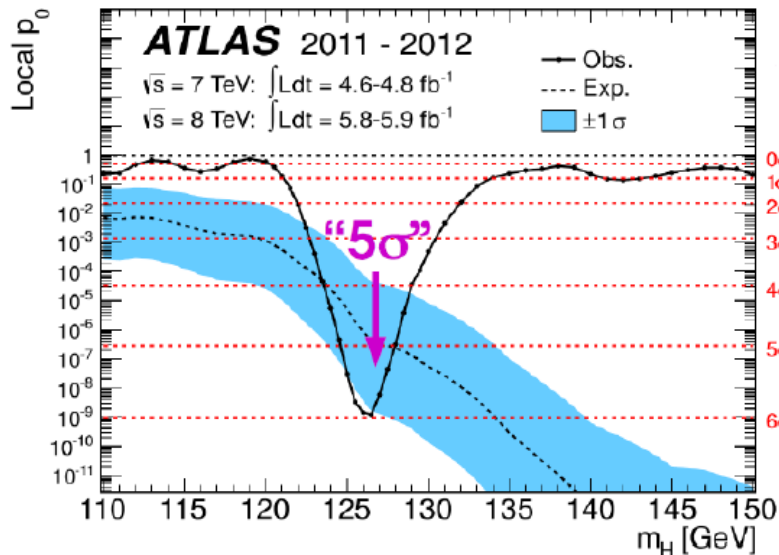
Part 1: Statistics and Data Analysis



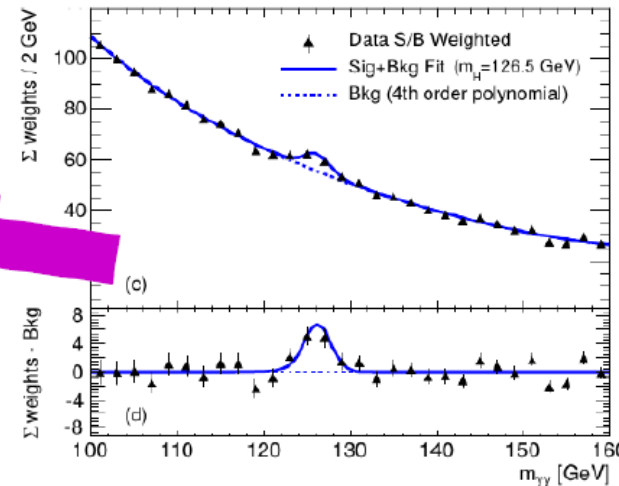
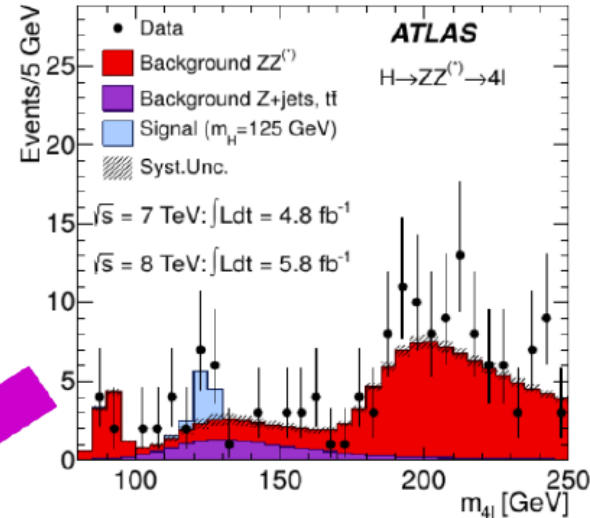
Part 1: Statistics and Data Analysis

Statistical methods play a critical role in many areas of physics

Higgs discovery : **“We have 5σ ”!**



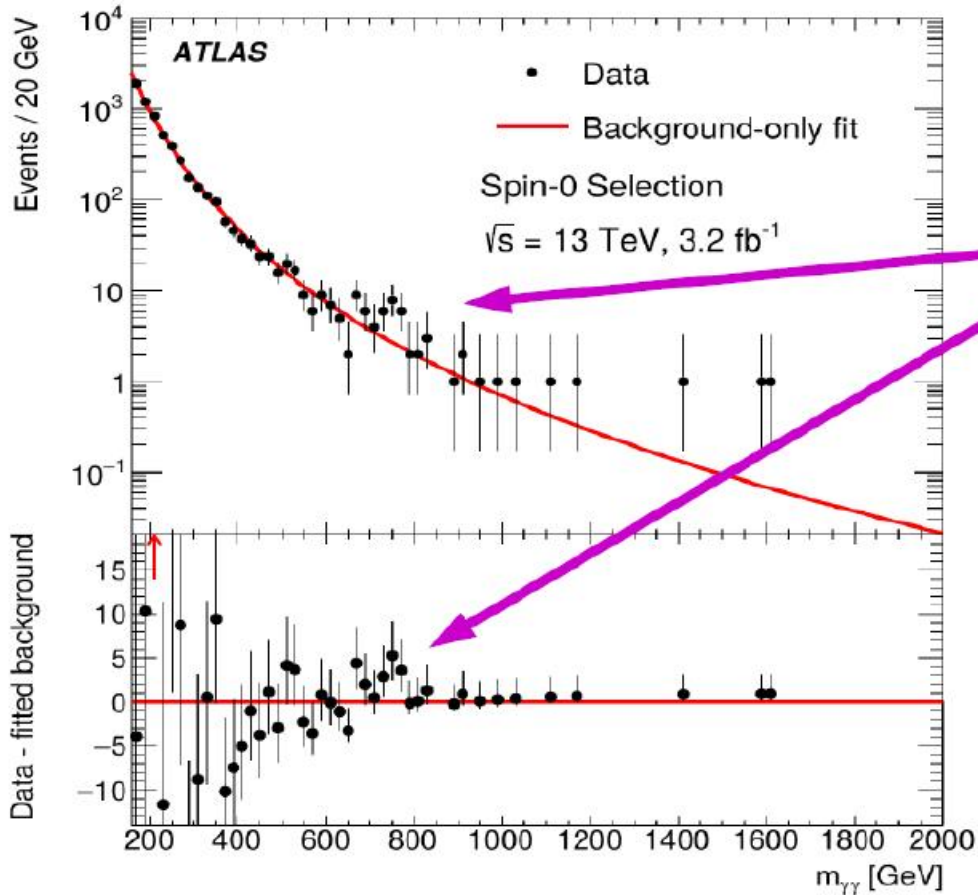
Phys. Lett. B 716 (2012) 1-29



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 ?

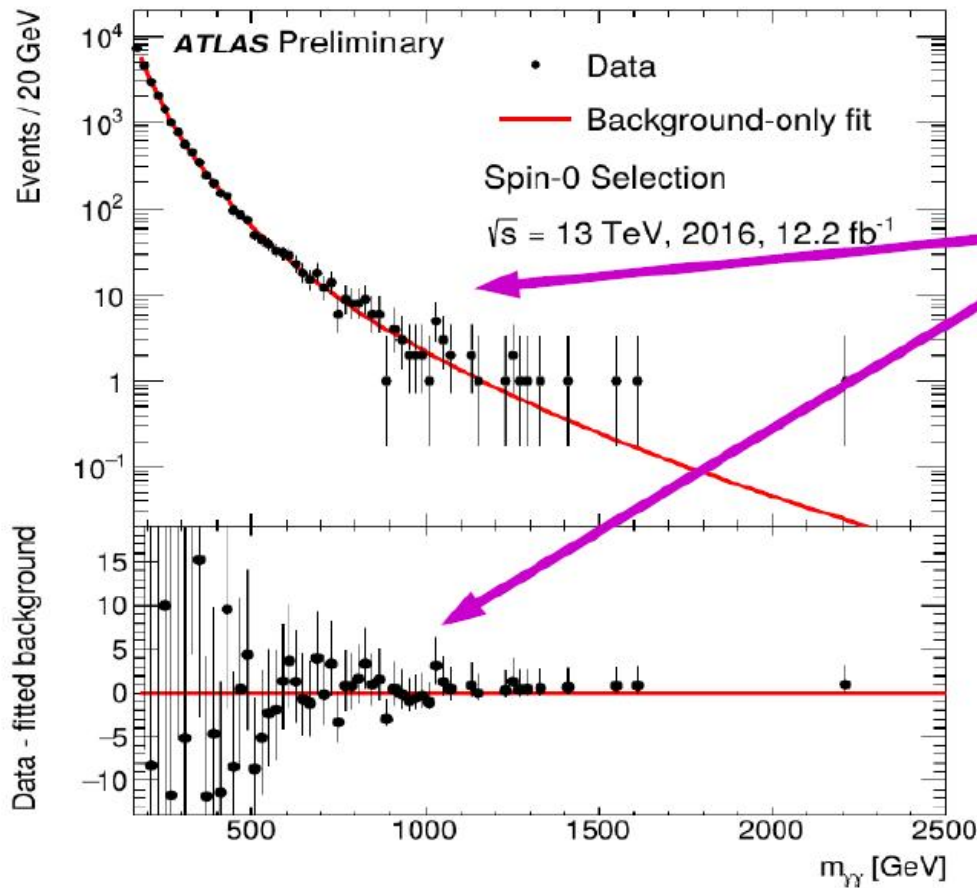
3.9 σ ? 2.1 σ ?



JHEP 09 (2016) 1

Part 1: Statistics and Data Analysis

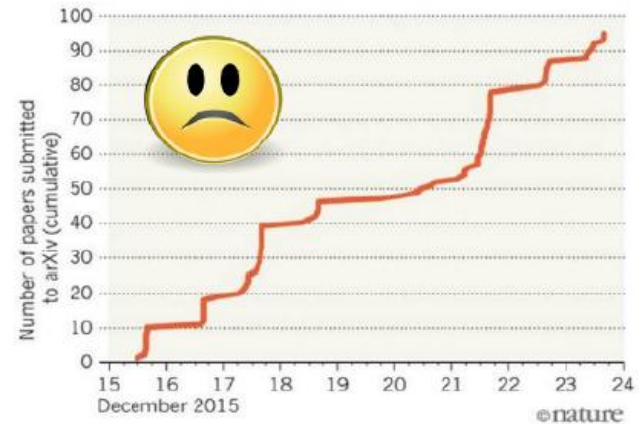
Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**...



A few months later...

~~New Physics ?~~

~~3.9 σ ? 2.1 σ ?~~

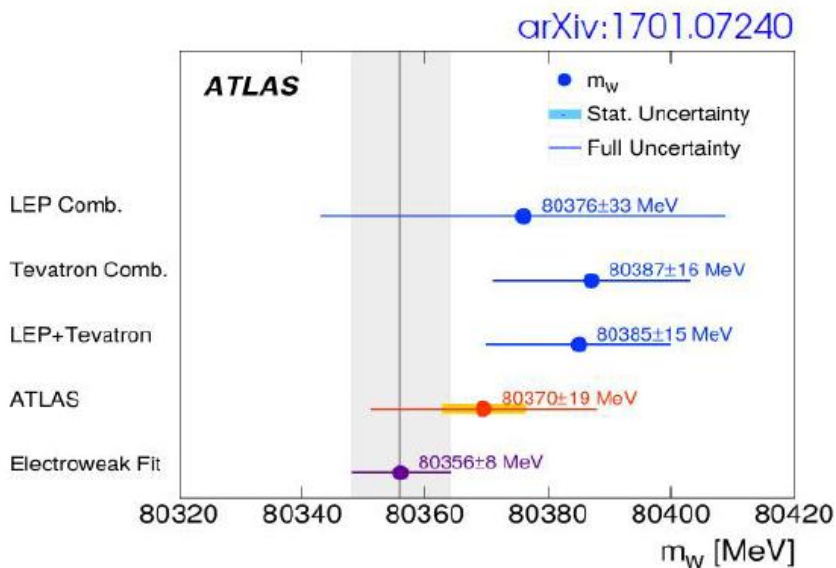


JHEP 09 (2016) 1

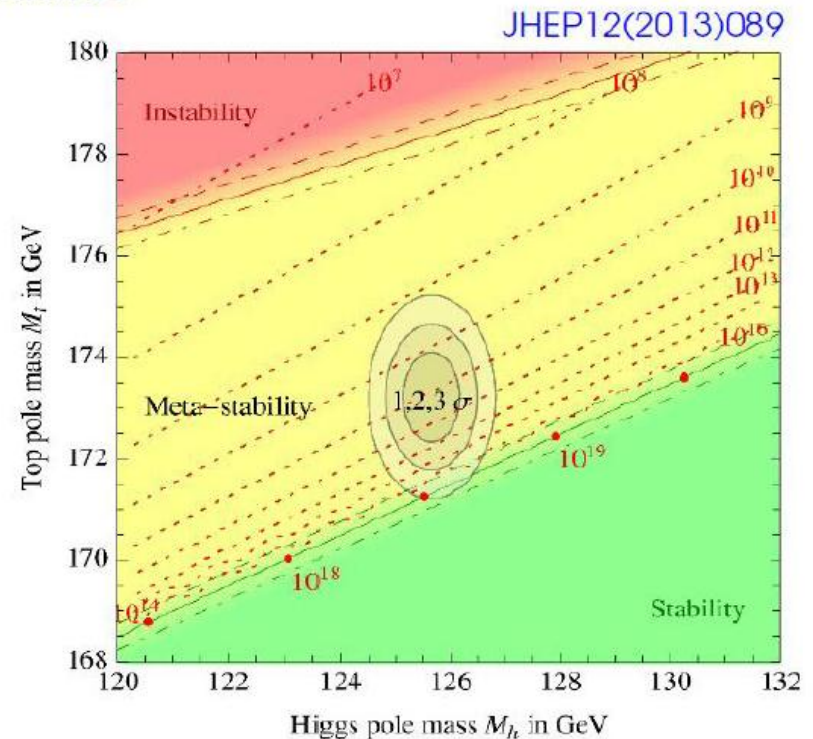
Part 1: Statistics and Data Analysis

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

Key point = determination of **uncertainties**

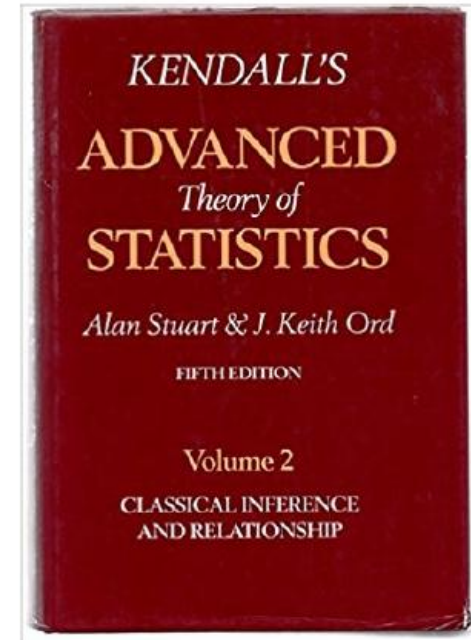
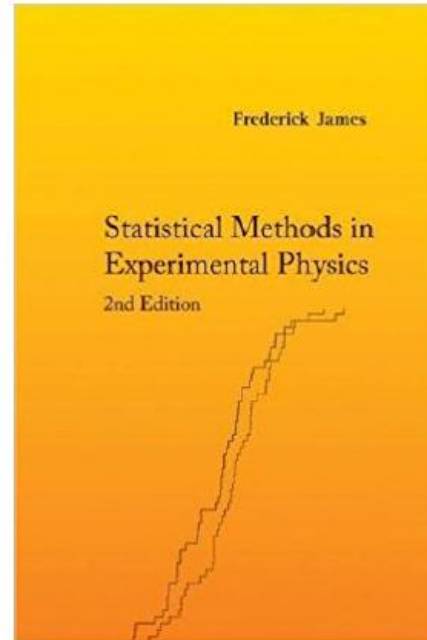
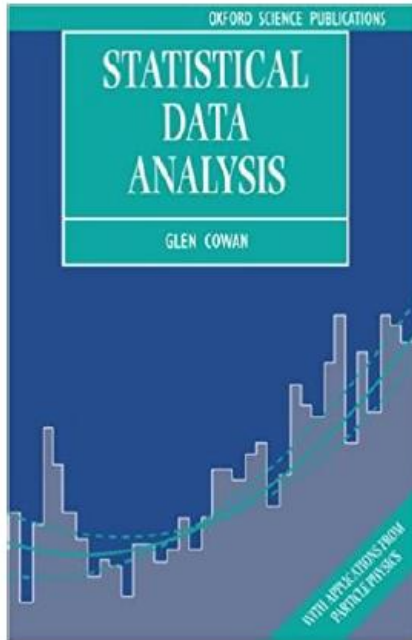


Consistency of the SM...



... or the fate of the universe

Part 1: Statistics and Data Analysis



Some other courses available online:

Glen Cowan's [Cours d'Hiver](#) and [2010 CERN Academic Training lectures](#)

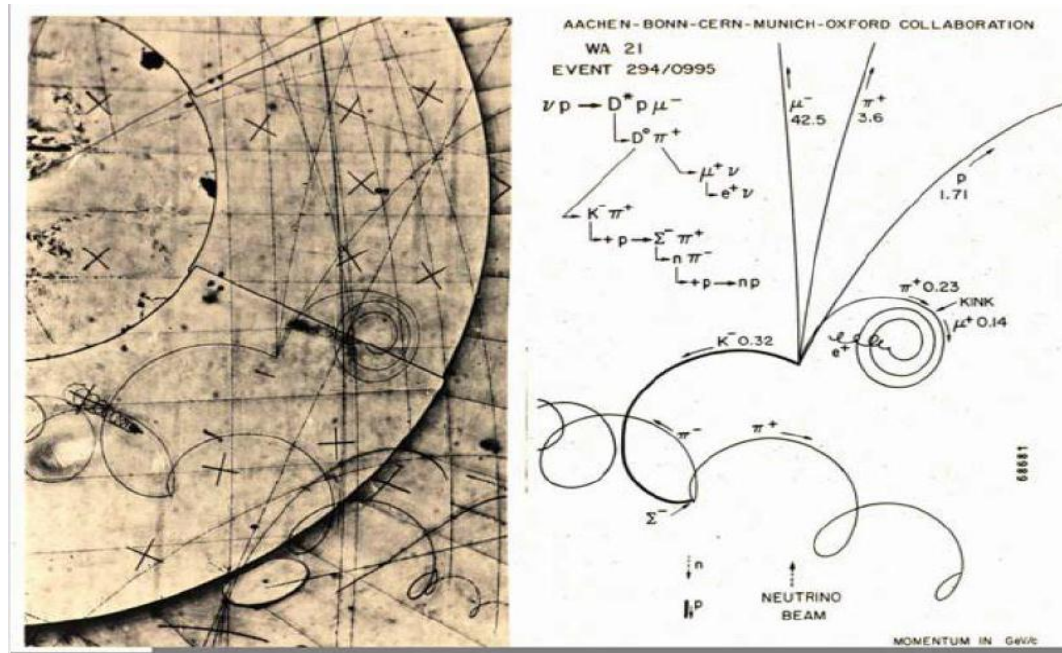
Kyle Cranmer's [CERN Academic Training lectures](#)

Louis Lyons' and Lorenzo Moneta's [CERN Academic Training Lectures](#)

Part 2: Multivariate Analysis and Machine Learning

In HEP everything started multivariate.

Below: intelligent „Multivariate Pattern Recognition” used to identify particles

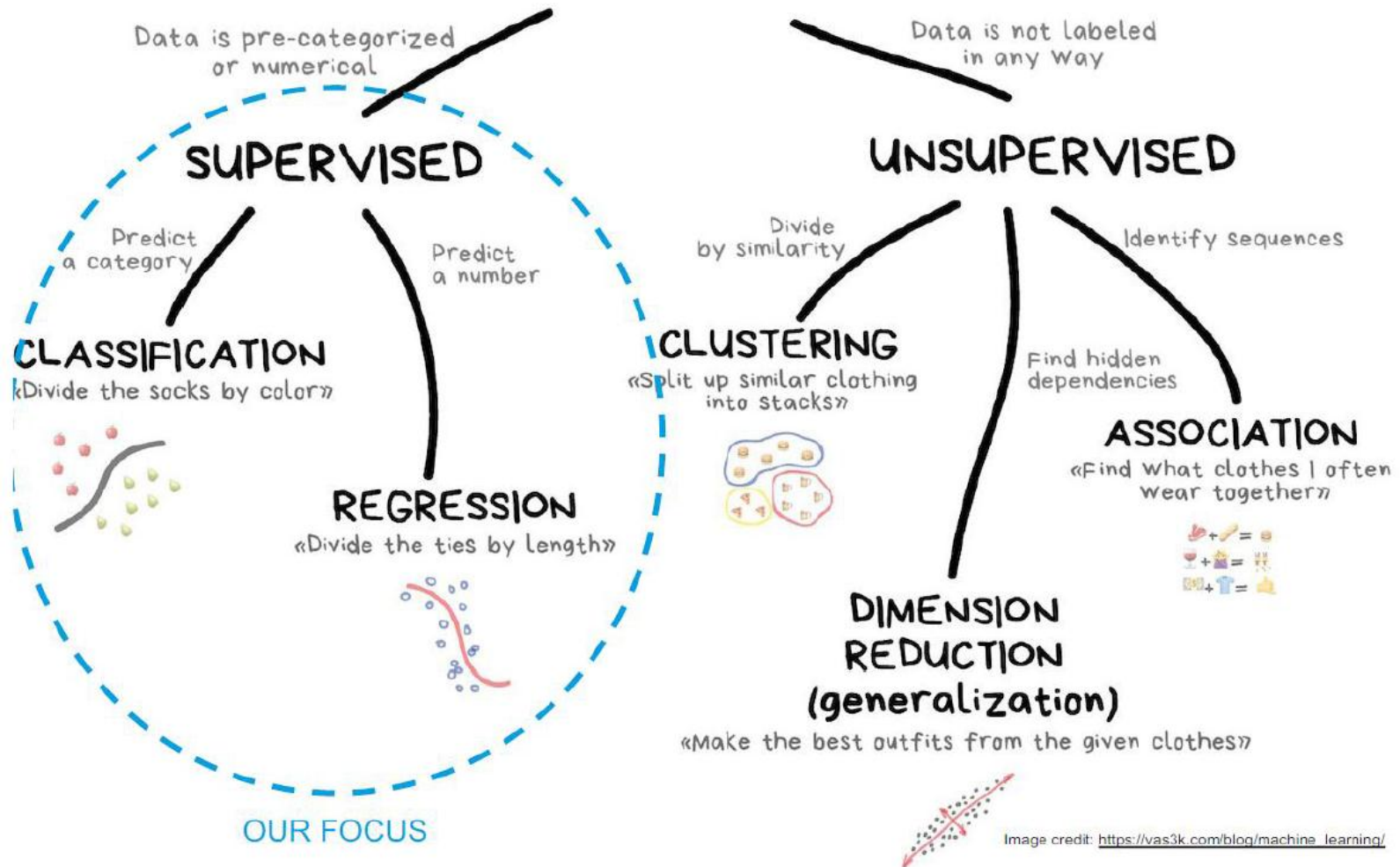


https://www.fehcom.net/WA21/wa21_01.html

Nowdays: let computer help you.

Classical Learning

CLASSICAL MACHINE LEARNING



Machine Learning

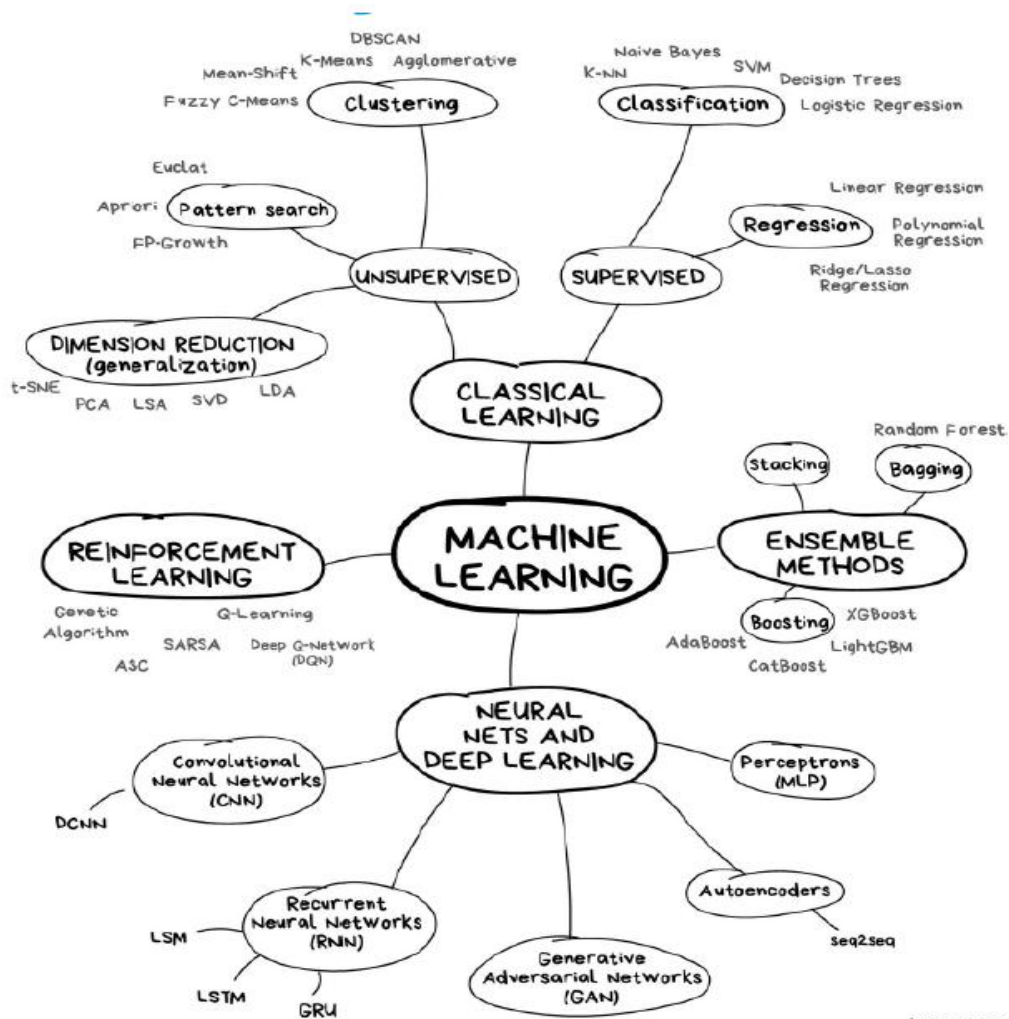


Image credit: https://vas3k.com/blog/machine_learning/

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	Likelihood	PDERS / k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
Performance	no / linear correlations	☹	😊	😊	☹	😊	😊	☹	😊	😊
	nonlinear correlations	☹	☹	😊	☹	☹	😊	😊	☹	😊
Speed	Training	☹	😊	😊	😊	😊	☹	☹	☹	☹
	Response	😊	😊	☹/☹	😊	😊	😊	☹	☹	☹
Robustness	Overtraining	😊	☹	☹	😊	😊	☹	☹	☹	☹
	Weak input variables	😊	😊	☹	😊	😊	☹	☹	☹	☹
Curse of dimensionality		☹	😊	☹	😊	😊	☹	😊	☹	☹
Transparency		😊	😊	☹	😊	😊	☹	☹	☹	☹

What is the model?



- ▶ This is not an apple just its graphical representation

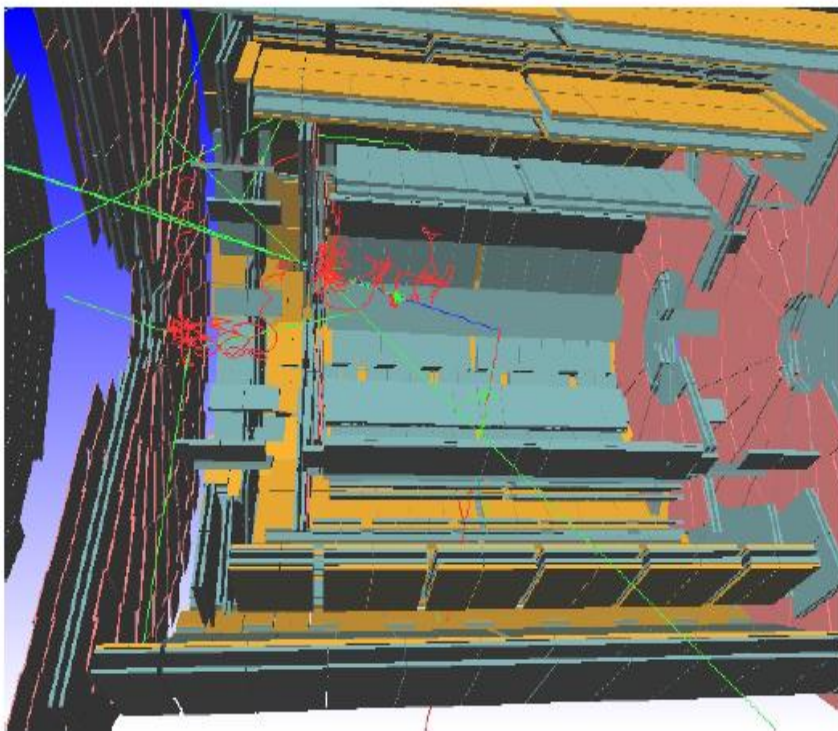
Many skills are needed to build a new model, to run it and analyze its results.

- ▶ 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.

Part 3: Physics modeling, simulation and Monte Carlo methods

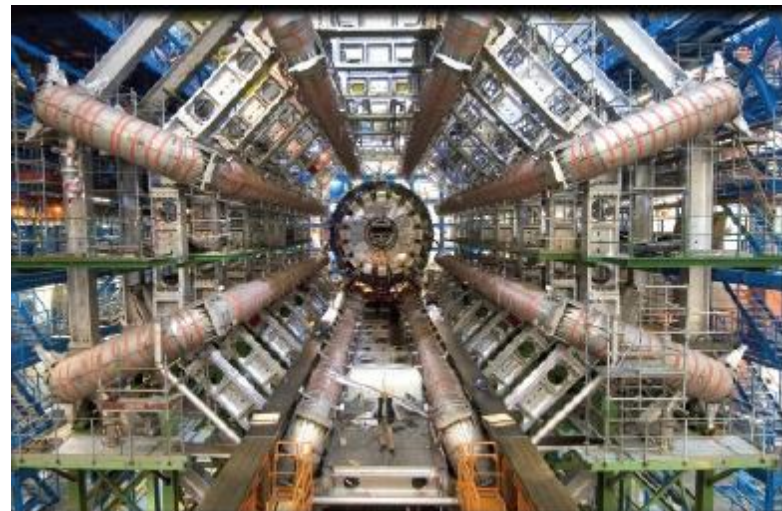
GEANT4

Visualised model of the detector used for simulation



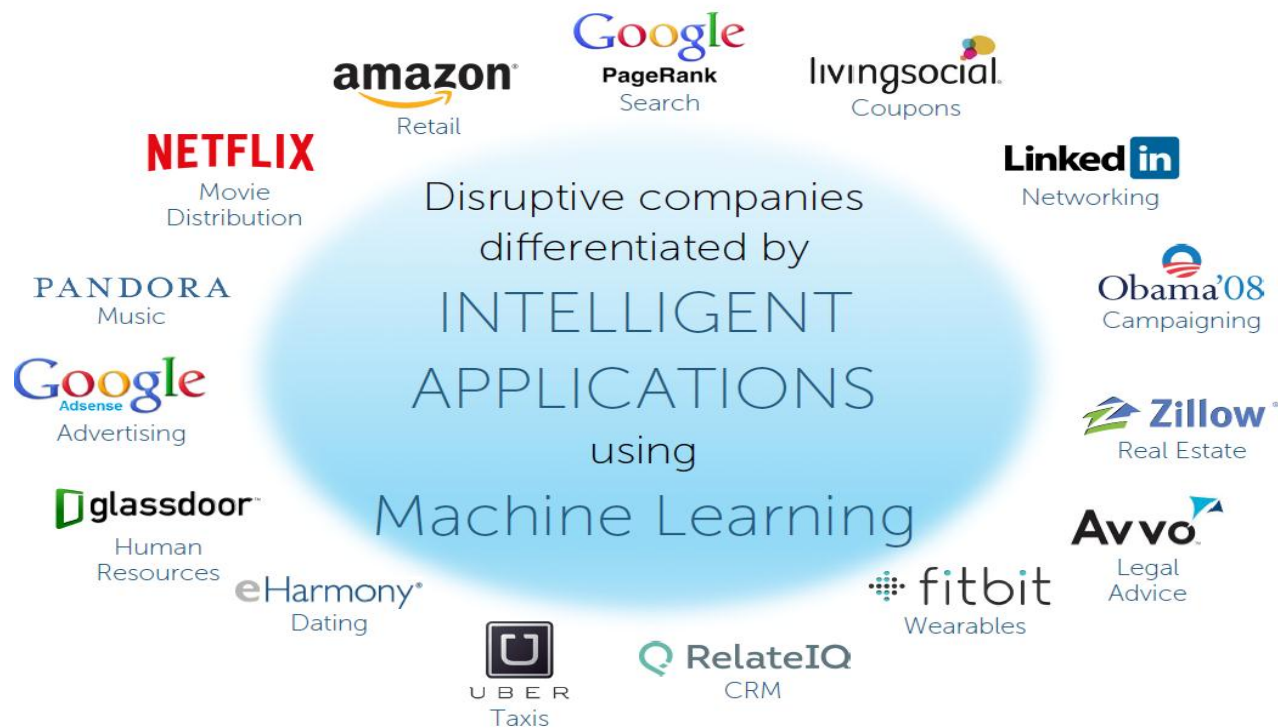
GEANT4 is also used to determine the performance of X-ray and gamma-ray detectors for astrophysics

Detector



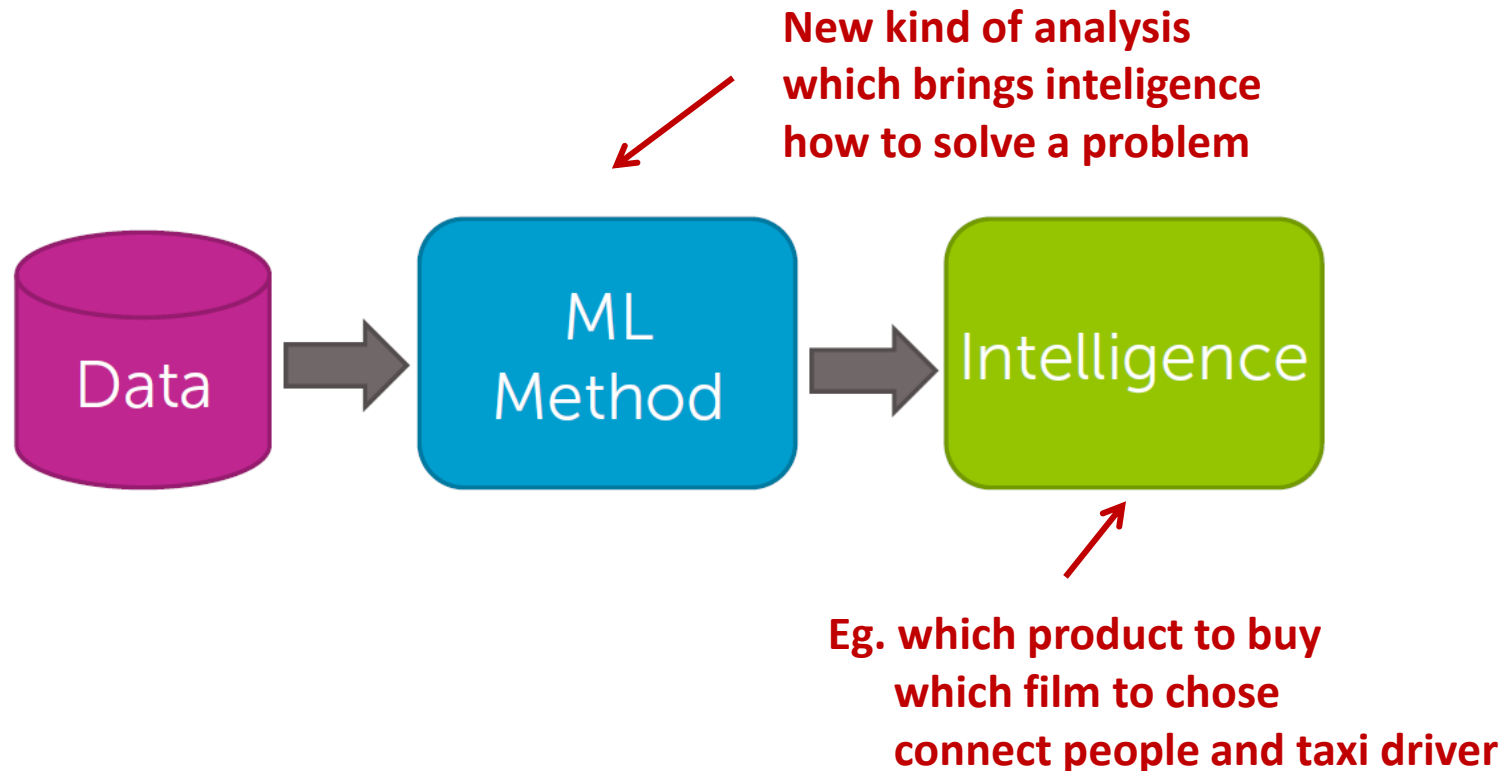
Part 4: Regression, Classification, Clustering

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



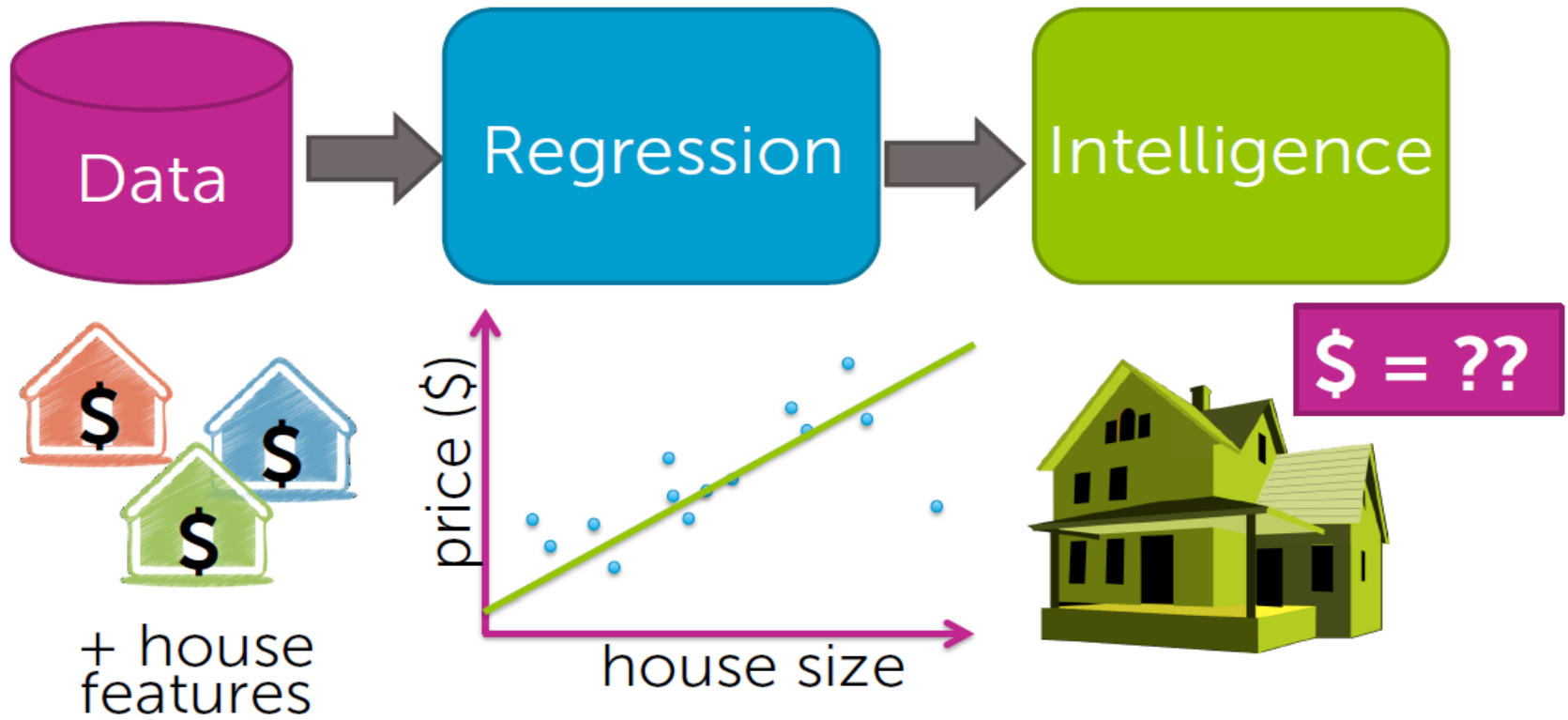
Part 4: Regression, Classification, Clustering

- **Data → intelligence pipeline**



Regression

Case study: prediction for the house price



Classification

Case study: Score of the restaurant



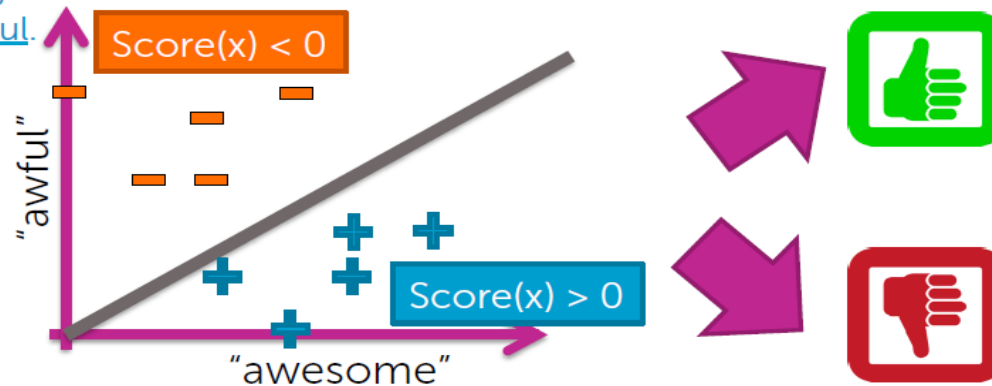
Sushi was awesome,
the food was awesome,
but the service was awful.

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 reservations, barged down to the ID after work, got here breathlessly at 5:10pm, and 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.



Clustering

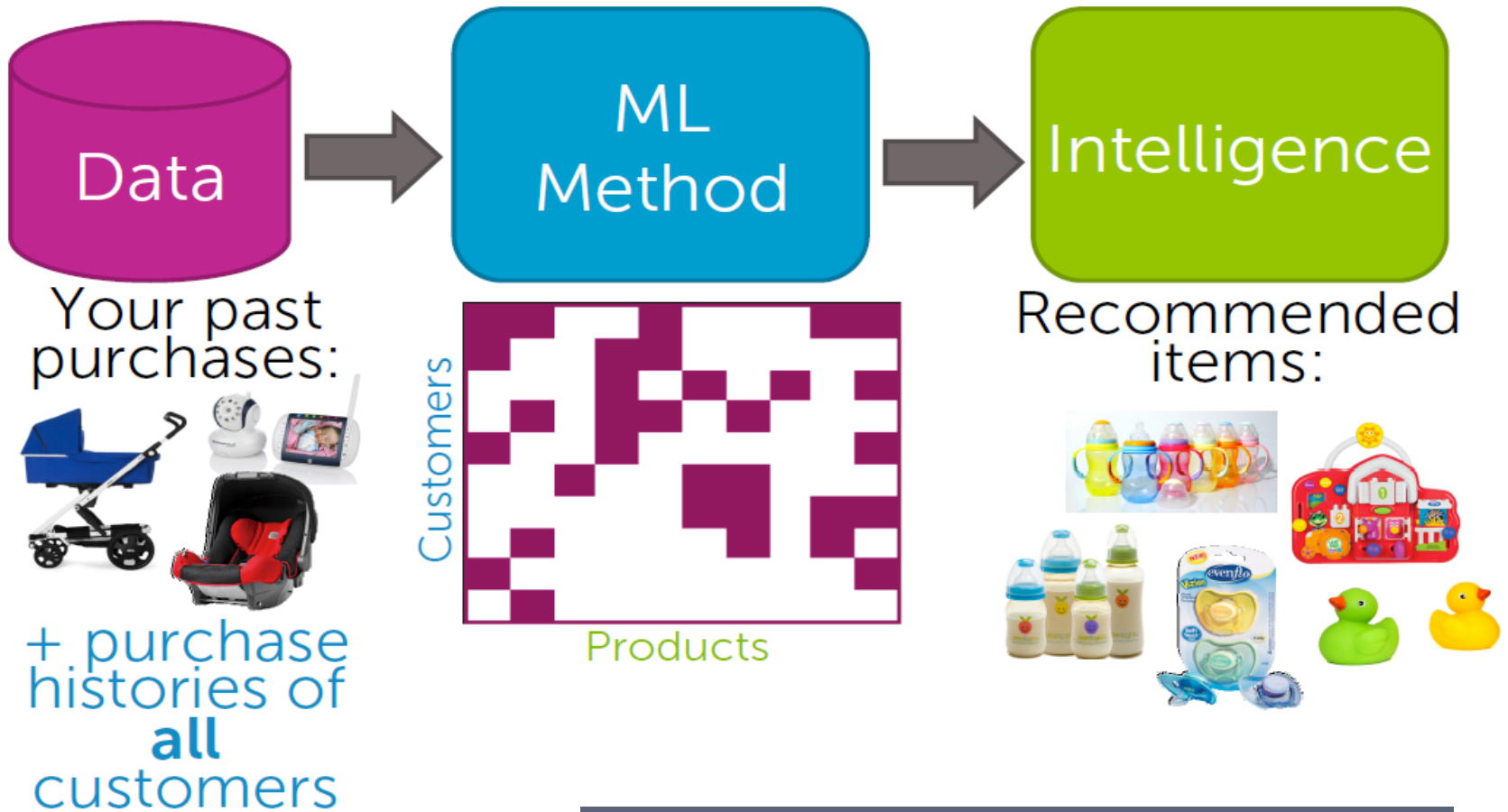
Case study: assigning books to groups by topics



10

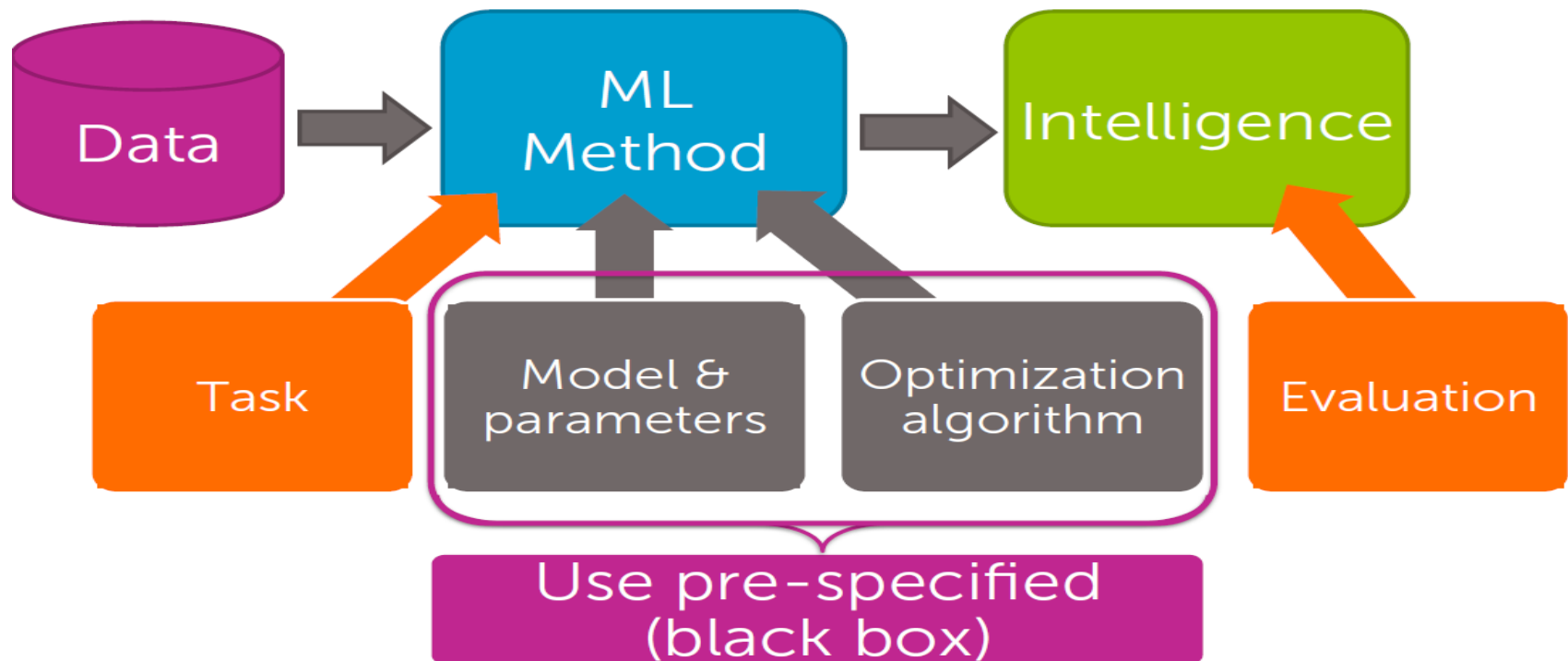
Recommendation

Case study: personalisation of recommending items



Deploying intelligence module

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



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

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

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

Getting your ETCs for the lectures

- **I foresee a 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.**