Multivariate analysis techniques: Recent developments

HELMHOLTZ ASSOCIATION

Stefan Ohm

KSETA: Machine Learning, Karlsruhe, 15.-16.10.2015

DESY

First of all, many thanks for your input!

Expertise	None	Basic	Expert
python		IIIII	IIIIIII
C++	I	IIIIIII	IIII
ROOT	IIII	IIII	IIII
scikit-learn	IIIIIII	IIII	I
TMVA	IIIIIIIII	I	II
BDTs	IIIII	IIIII	II
Random Forests	IIIIIII	III	II
Neural Networks	IIII	IIIII	III

- Expertise
  - no-one is unfamiliar with python → good
  - less people than I thought are familiar with ROOT/TMVA
  - we have experts in the room → please correct me if I'm missing something
  - very diverse knowledge among you → will try to keep the balance

Outline

- What this course is *not*
  - a math and statistics class, a programming course
- What this course tries to be
  - a “pragmatic astroparticle physicists approach to data”
- PART 1 (~1 hour)
  - Data Science: the bigger picture, or, what we are all facing sooner or later
  - Statistics and machine learning: The basic basics
- PART 2 (~2 hours)
  - Multivariate analysis (MVA) techniques: decision trees, neural networks, deep learning
  - Real-life examples: Boosted Decision Trees in γ-ray astronomy, Random Forests in optical transient searches, other fun examples
- PART 3 (~3 hours)
  - Hands-on: apply MVA techniques to data
- If anything is unclear just ask

Literature and Sources

- Many of the slides are inspired by these books and lecture notes
  - Think Stats: Exploratory Data Analysis in Python, A.B. Downey, Green Tea Press, 2014
  - The Elements of Statistical Learning: Data Mining, Interference, and Prediction, Hasti, Tibshirani, Friedman, 2009, Springer Series in Statistics
  - Data Science from Scratch: Joel Grus, O'Reilly, 2015
  - Practical Statistics for Astronomers, Jasper Wall
  - other references given throughout the presentation and at the end

PART 1

Data Science and the basics about Statistics and Machine Learning

HELMHOLTZ ASSOCIATION

DESY

Why a dedicated course on this topic?

> **DATA** EXPLOSION

- in science
- in industry
- in every-day life

Why a dedicated course on this topic?

> DATA EXPLOSION

- in science
- in industry
- in every-day life

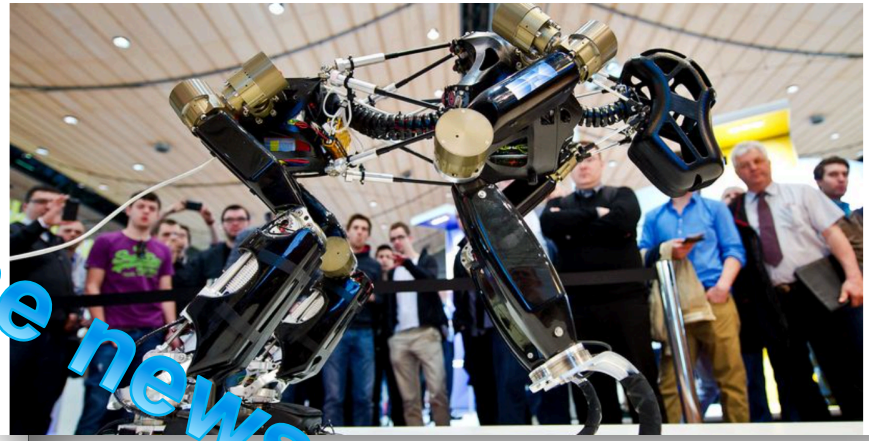
Künstliche Intelligenz

Google kauft sich bei deutschen KI-Forschern ein

Neuronale Netze, Roboter, Datenanalyse: Was am Deutschen Forschungszentrum für Künstliche Intelligenz entwickelt wird, interessiert Google. Nun beteiligt es sich daran.

Von **Axel Postinett**

7. Oktober 2015, 9:42 Uhr / Erschienen im Handelsblatt / 38 Kommentare



Datenanalyst

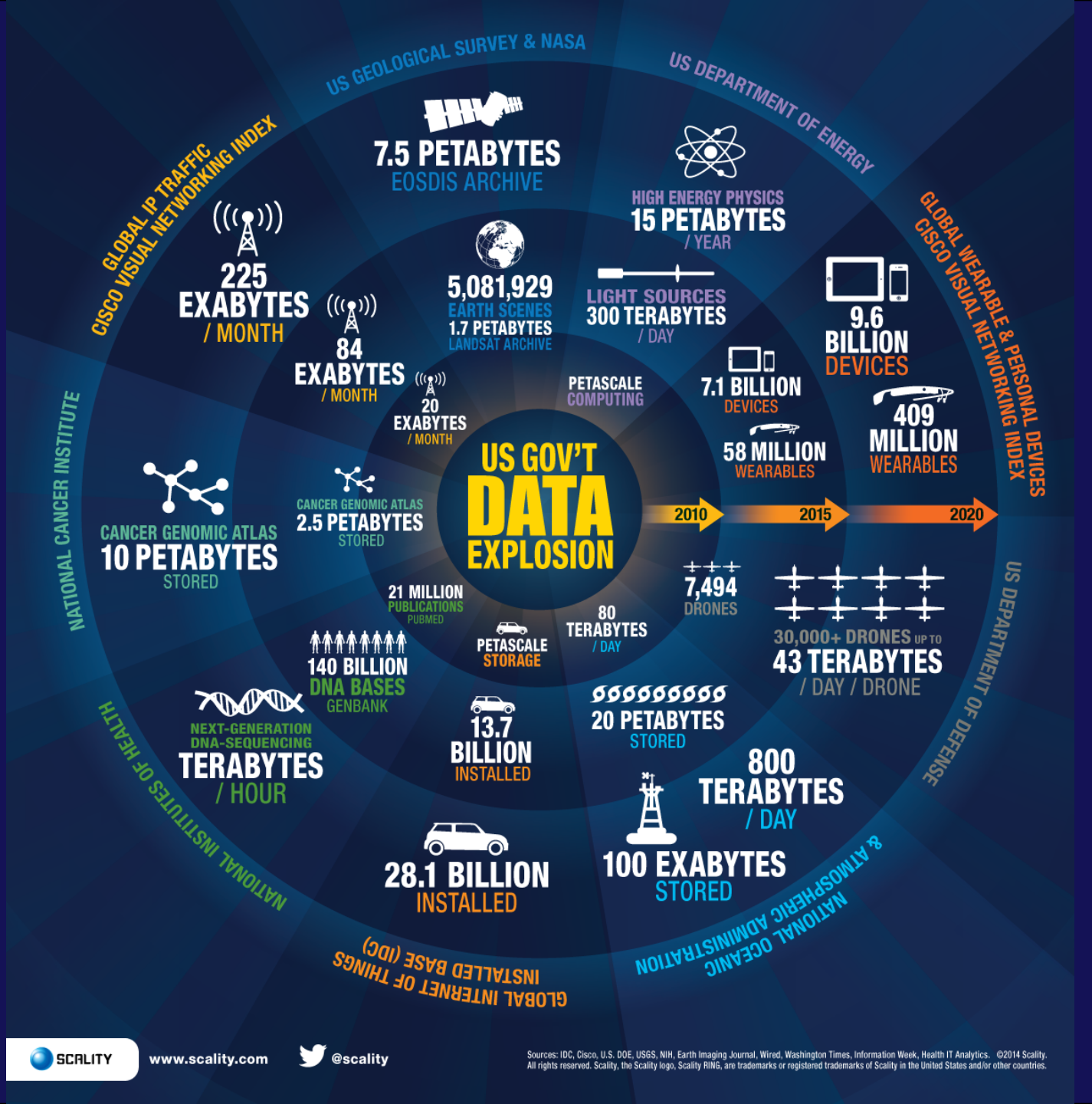
Mathefreaks im Glück

Astronomische Einstiegsgehälter und totaler Fachkräftemangel – so gut wie jede Branche sucht händeringend Datenanalysten. Warum es sich lohnt, in Mathe gut aufzupassen.

Von **Markus Schleufe**

22. September 2015, 21:57 Uhr / 10 Kommentare

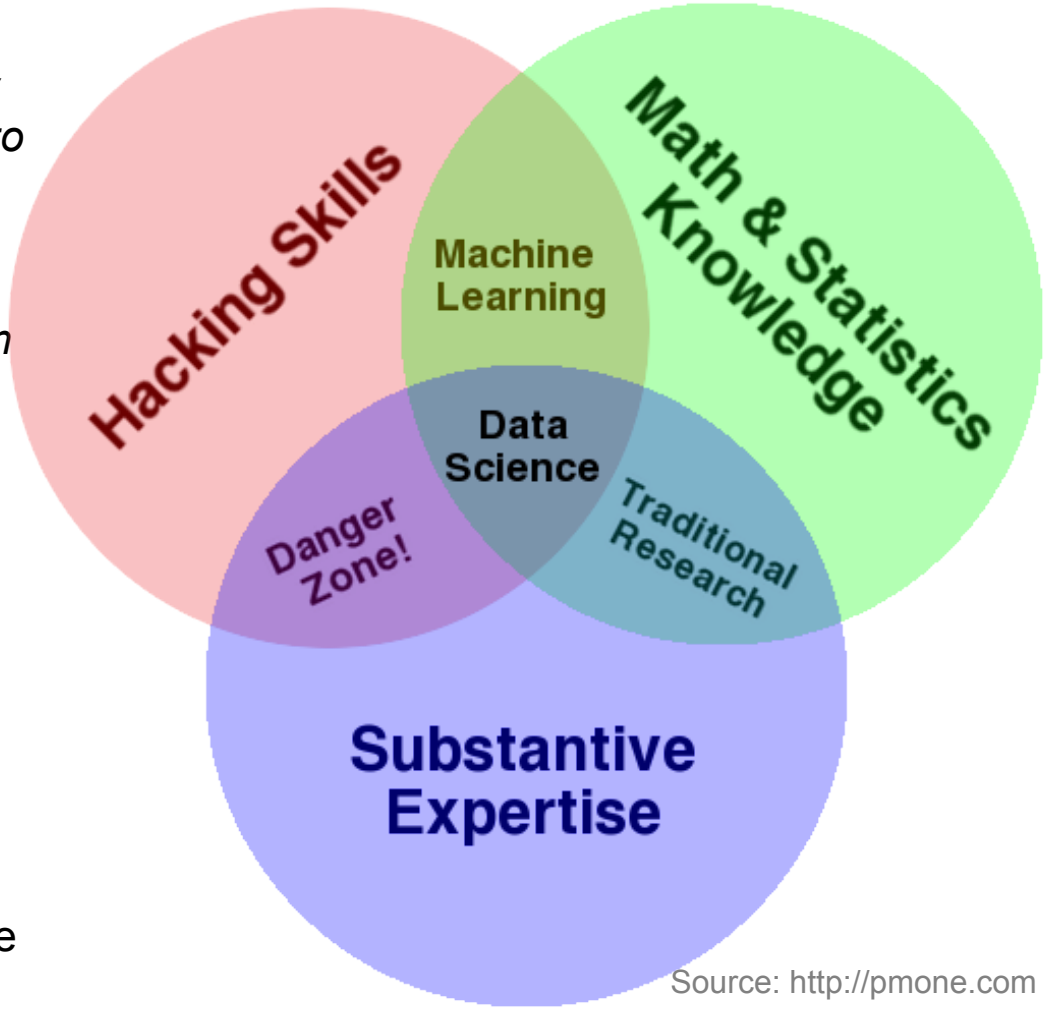
It's in the news



SCALITY www.scality.com @scality

Big Data & Data Science

- According to Wikipedia
  - *"Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as data mining and predictive analytics, as well as knowledge discovery in databases"*
- Venn Diagram
  - E.g. physicists and astronomers are the definition of Data Scientists
  - Fulfill all requirements and are developers of widely applicable code
  - Beware of the Danger Zone!



Source: http://pmone.com

Historical Background

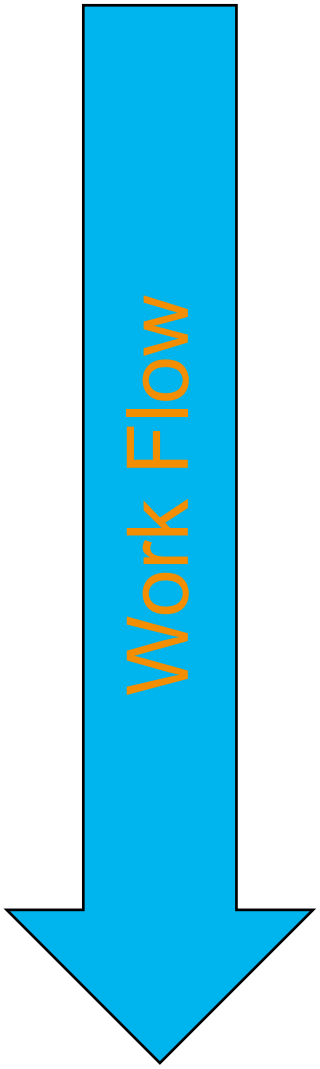
- The field of data science
  - 1960: Term used initially as substitute for computer science
  - Covered advanced analytics, simulations, data mining, machine learning and predictions
  - 1996: “Data Science” used for the first time in the title of a conference
  - 1997: C.F. Jeff Wu in inaugural lecture: “Statistics = Data Science”?
  - 2001: William S. Cleveland introduces data science as independent discipline
  - 2002/2003: First data science journals are published
  - 2008: DJ Patil and Jeff Hammerbacher used term “data scientist” to describe their jobs at LinkedIn and Facebook
  - 2010: Stefan gets his PhD for applying a boosted decision tree method to H.E.S.S. data ;-)

How does machine learning fit in?

- Techniques from many fields, including
  - signal processing, probability models, **machine learning**, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, […], high performance computing
- Difficult to talk about machine learning without talking about data
- A few examples
  - *Facebook* uses hometown and current location to identify global migration patterns
  - *Target* tracks purchases and interactions to predict which of its customers is pregnant
  - *2012 US election* – Obama employed hundreds of data scientists to identify potential voters
- Data Science for social good
  - Improve government
  - Help homeless people
  - Improve health care

Data analysis

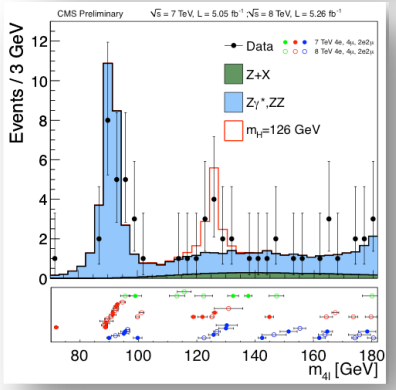
- Data retrieval
  - from experiments
  - generate your own data
  - from internet or an Application Programming Interface (API)
- Data preparation
  - calibrate raw data
  - remove outliers or noise
- Data pre-processing
  - Identify parameters with valuable information in calibrated data (signal/background classification, pattern recognition)
  - First-level selection cuts and data reduction (“outliers”)
- Data Mining (Machine Learning with multivariate analysis techniques)



→ *What are the tools to do that?*

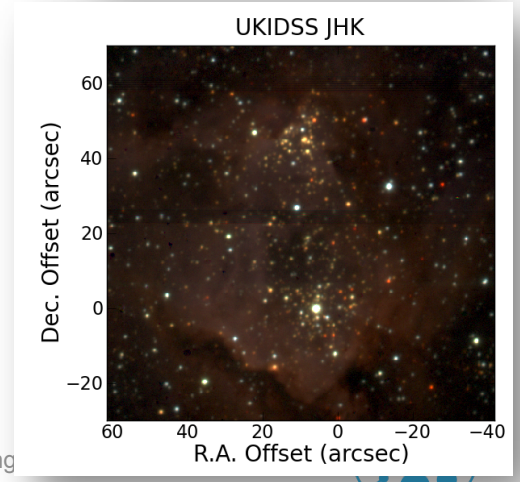
The toolkits

- ROOT
  - Mostly used in particle physics
  - Adapted in astronomy and astroparticle physics
  - C++, object-oriented
  - Higgs-discovery plots produced with ROOT



- IDL, MATLAB, R, ds9, …
  - standard in different fields for many years, now being replaced by python
  - many statistics features and numerical methods

- Python
  - the new standard in industry and science



ROOT vs. python

ROOT functionality	Python Equivalent
interactive interpreter	ipython, ipython notebook
TH1D,TH2D,TH3D	numpy.ndarray + numpy.histogramdd()
TProfile	numpy.ndarray
TGraph	numpy.ndarray + matplotlib plots
TTree	astropy.table or numpy.recarray
TMinuit	scipy.optimize or iminuit
2D/3D graphics	matplotlib
TMVA	scikit-learn
GUI	various (wxwidgets, Qt,gtk)
AstroROOT	astropy.io.fits or fitsio
ROOFit	astropy.modelling or lmfit
Reflex, CINT	not needed
TMatrix, etc	numpy.ndarray + scipy.linalg
TMatrixTSparse	scipy.sparse
TList	list
TArray, TObjArray	list
TMap, THashList	dict
TRandom	numpy.random + scipy.stats
Math, MathMore	numpy + scipy
Script compilation	Numba, Cython
ROOT::Math::VirtualIntegrator	scipy.integrate
–	astropy.coordinates
–	astropy.units
–	astropy.time
–	astropy.wcs (projections)
–	numpy.ndarray (n-dimensional tables)
–	scipy.interpolate.interpnd (n-dimensional interpolation)
–	scipy.signal (digital signal processing)

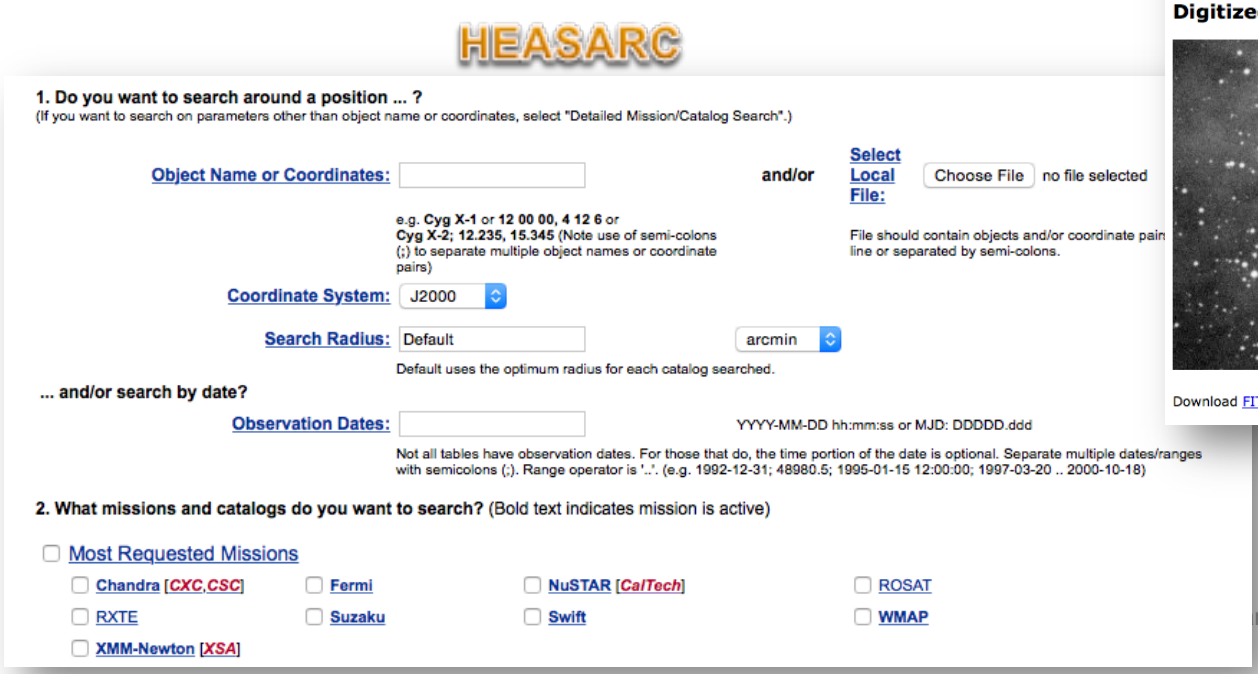
python

- provides more functionality than ROOT
- vastly more external developers
- shared by many communities (scientific and non-scientific)
- better/cleaner design
- quicker and easier to get things done
- more lightweight

DESY

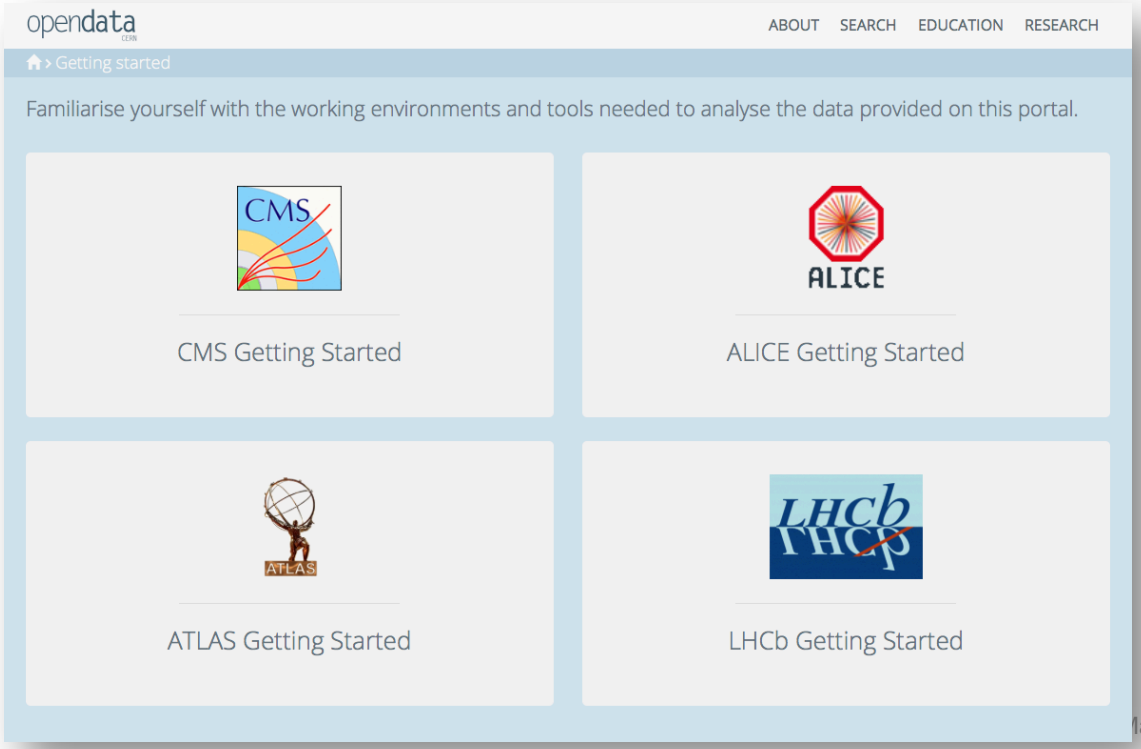
Data retrieval

- Open access is the way to go
  - Astronomical observatories after proprietary time
  - Provided via webpages and online archives
  - Mostly in FITS format



Data retrieval

- Open access is the way to go
  - Particle physics community is joining
  - Often via Virtual Machines (as you will likely use later)
  - No fiddling around with software versions, installations, etc (well, lets see how that goes tomorrow ☺)



Data retrieval

- Generate your own data
  - very useful in many circumstances
  - using random number generators
  - large-scale Monte Carlo simulations
  - Beware of the seed (!)
- Monte Carlo simulations
  - most of you know those
  - used for optimization
  - numerical integration
  - drawing from probability distributions (see later)

Generate function

```
TF1 parabola("parabola","[0]+[1]*x+[2]*x**2",0,20);
format_line(&parabola,kBlue,2);

TF1 gaussian("gaussian","[0]*TMath::Gaus(x,[1],[2])",0,20);
format_line(&gaussian,kRed,2);

TF1 gausppar("gausppar",the_gausppar,-0,20,6);
```

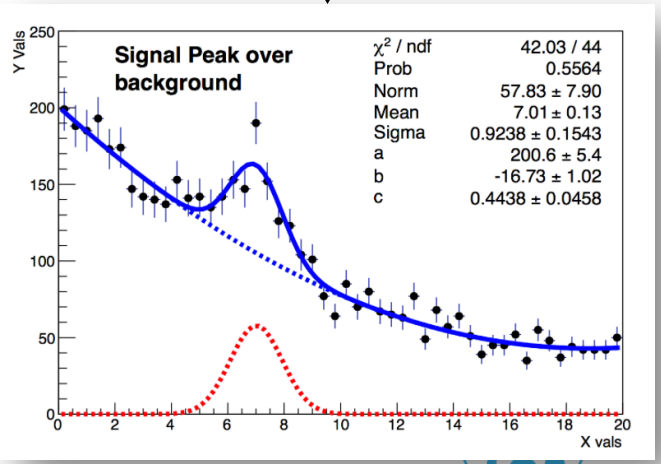
Fake data

```
// Fake the data
for (int i=1;i<=5000;++i) histo.Fill(gausppar.GetRandom());
```

Fit fake data

```
// perform fit ...
TFitResultPtr frp = histo.Fit(&gausppar, "S");
```

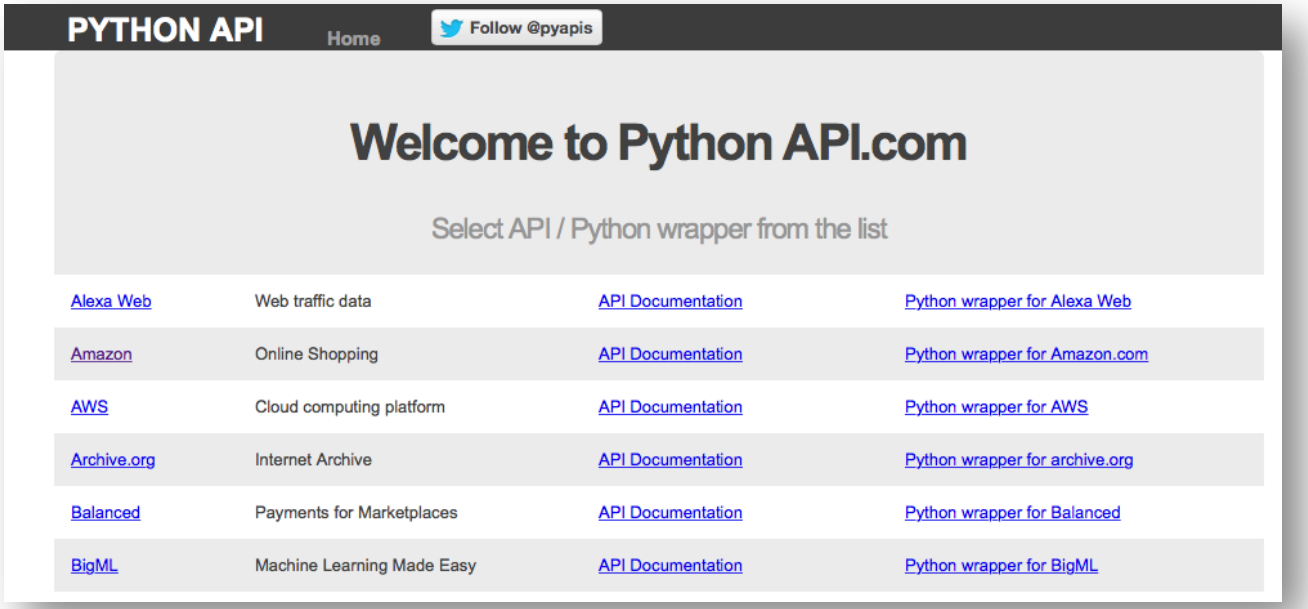
Draw everything



Data retrieval

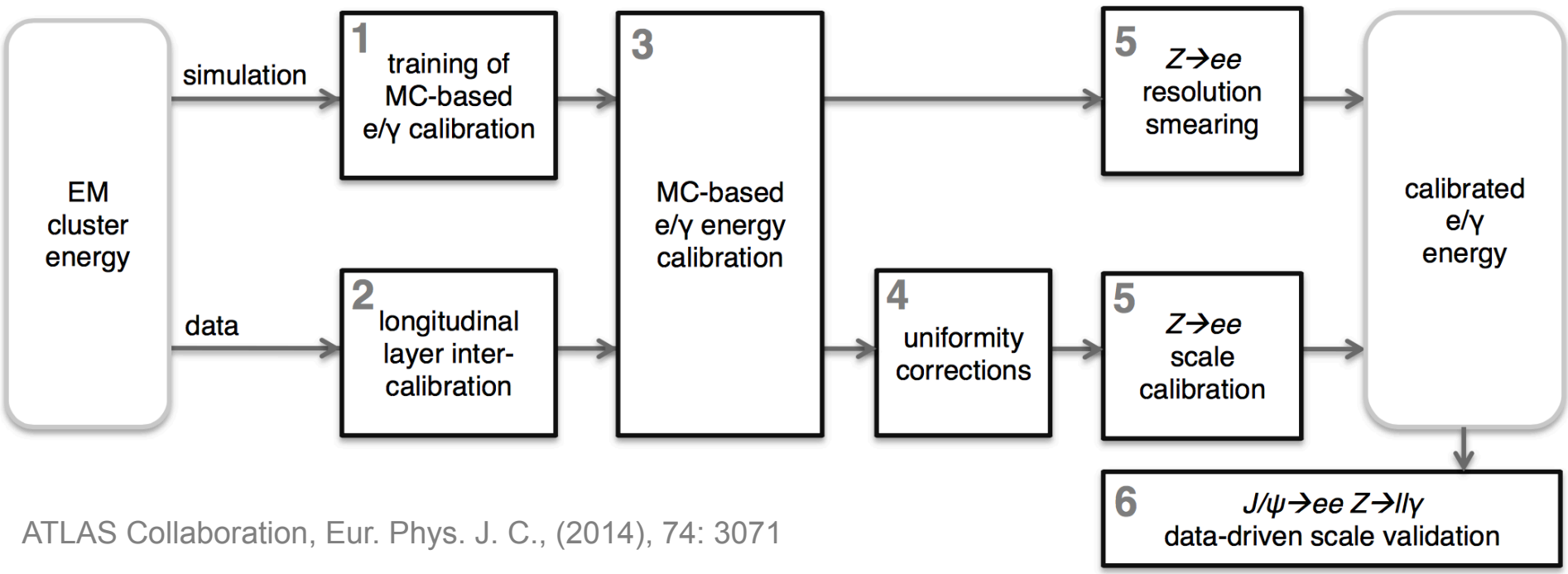
> Application Programming Interfaces

- Many webpages provide APIs
- Interface for programmers to receive data in structured format (XML, JSON)
- JSON looks like python dictionaries
- In python, many API libraries already exist (e.g. Amazon, Ebay, Facebook, Geopy, Google Maps, Last.fm, Rotten Tomatoes, Twitter, Wikipedia)



Data preparation

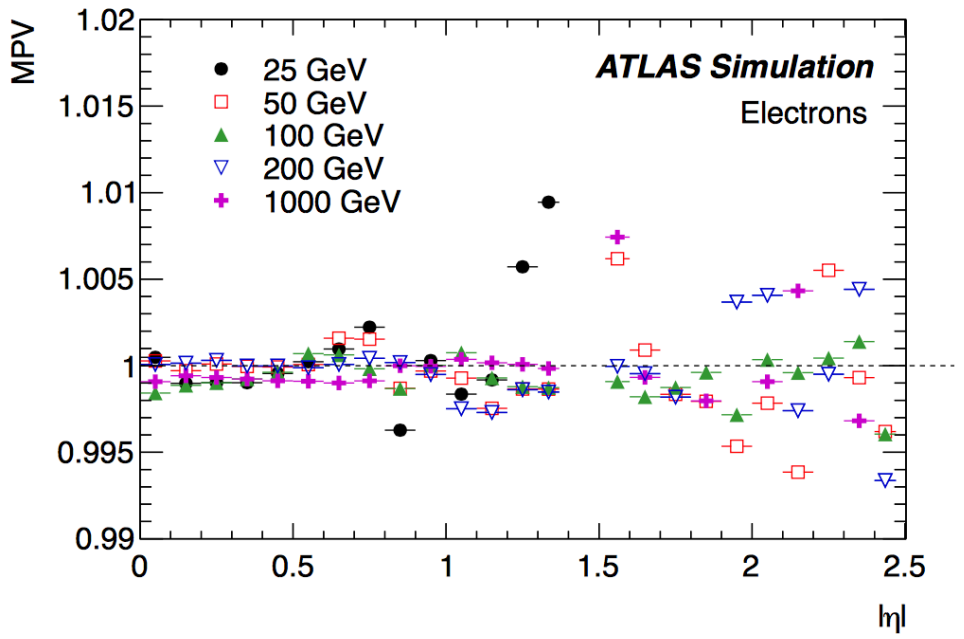
- The data you want to analyse is calibrated
  - Congratulations, someone has done the job for you
  - Don't forget to check also calibrated data
- The data you want to analyse is not calibrated
  - Lets roll up the sleeves and calibrate the data
  - Illustration of how important calibration (and simulations) are (ATLAS EM calorimeter)

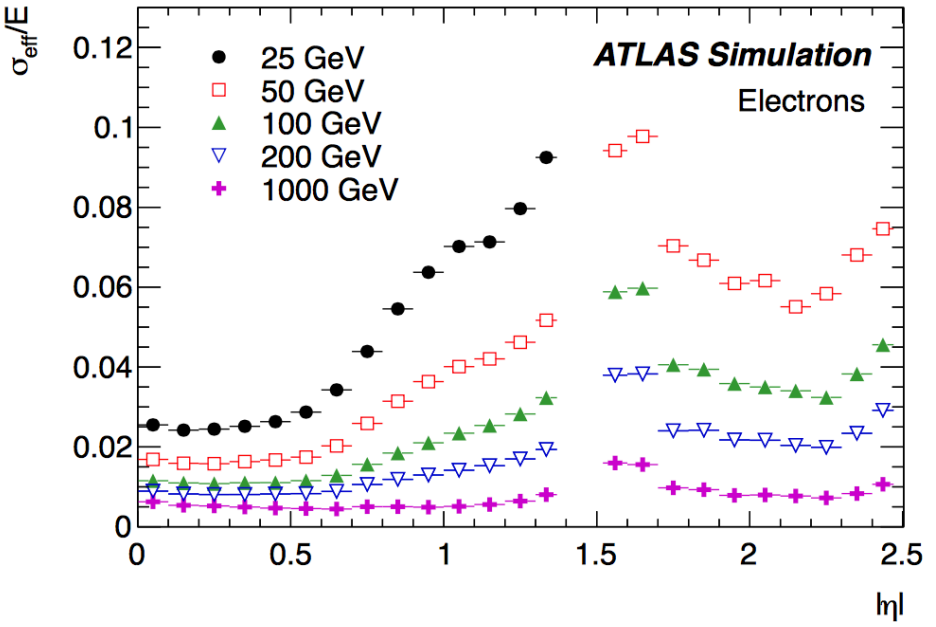


ATLAS Collaboration, Eur. Phys. J. C., (2014), 74: 3071

Data preparation

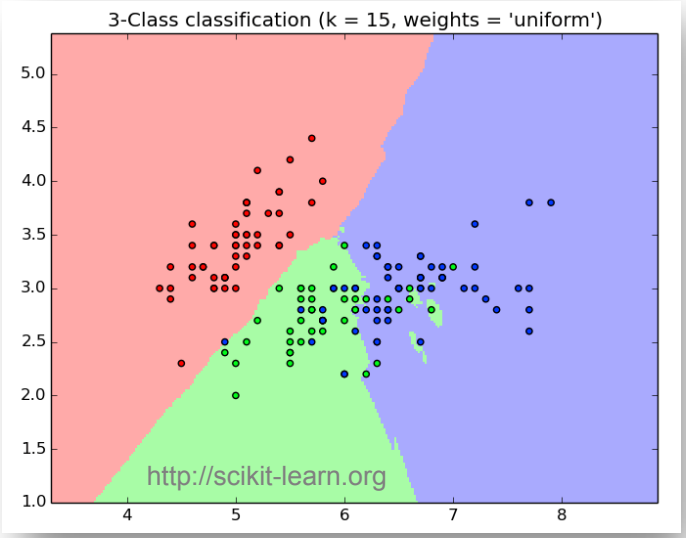
- The data you want to analyse is calibrated
  - Congratulations, someone has done the job for you
  - Don’t forget to check also calibrated data
- The data you want to analyse is not calibrated
  - Lets roll up the sleeves and calibrate the data
  - Illustration of how important calibration (and simulations) are (ATLAS EM calorimeter)

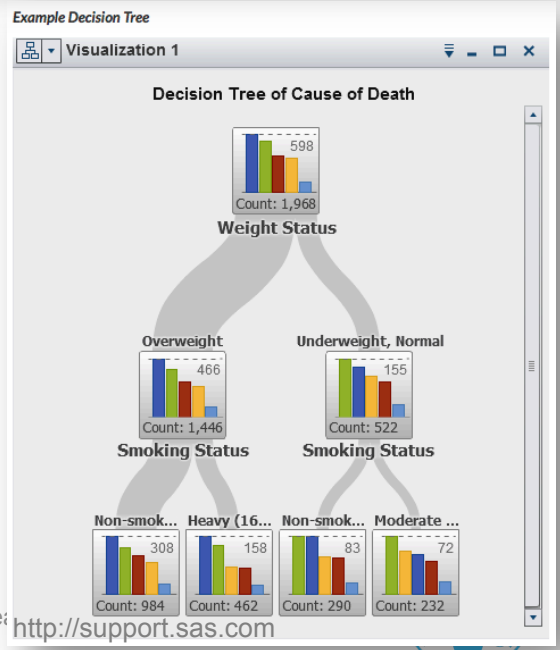




Data visualization

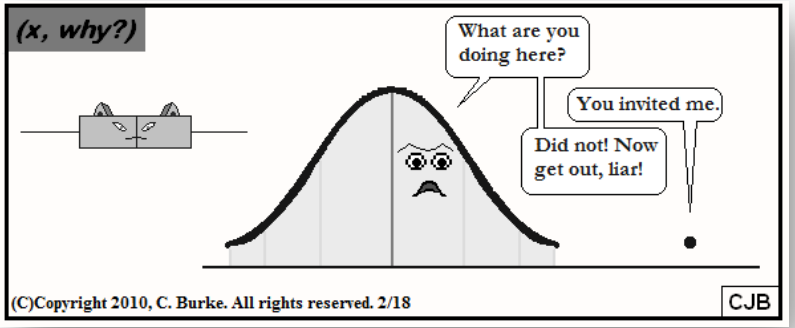
- Scatter plots, line charts, bar charts, histograms
- What is it good for?
  - inspection and exploration of acquired data
  - communicate via the data → knowledge transfer
- Histogram
  - is a complete description of the distributions of a data sample
  - we can reconstruct values in the sample based on the histogram
  - however, order of elements is lost
- Examples
  - scatter plots to visualize Nearest Neighbours, or check for correlations
  - histograms to check input of boosted decision trees
  - will chose one or the other depending on application
  - see later





Data pre-processing

- Identify outliers during data preparation
- Outliers
  1. Occur during data collection, i.e. human errors
  2. Malicious act (e.g. in questionnaires)
  3. Noise in data (temporal, spatial, etc.)
  4. Incorrect assumptions when looking at data or building model



- Importance
  - 1 & 2 very important in social sciences, medical studies
  - 3 & 4 very important in fields where data is collected with ‘instruments’ (e.g. physics, astronomy, geo-sciences)
- If kept
  - 1 & 2 may influence statistical analyses and outcome of statistical test (e.g. correlations)
  - 3 & 4 can fake signal, wrong understanding of background
- Remove or not to remove?
  - Remove only for good reasons!

Where are we now?

- We retrieved some data
  - from an instrument or from elsewhere
- We calibrated the data
  - or got calibrated data
- We inspected the data
  - identified outliers and removed them,
  - or know where they come from and include them in our model
- Building a model and interpreting the data

→ Dive a bit into Statistics

Statistics

- Statistics vs. Machine Learning?
  - both are strongly connected
- Terminology

Statistics	Machine Learning
model	network, graphs
parameters	weights
fitting	learning
test set performance	generalization
regression/classification	supervised learning
density estimation, clustering	unsupervised learning

- Some general remarks about Decision and Statistics
- Statistics and Expectations

Decision

- Science is Decision and we decide by comparing
- Example
  - “Does the event observed in the ATLAS detector look like a Higgs?”
  - Different from the question “Is the event […] a Higgs?
  - Measure particle tracks in collisions, infer secondary particle species and their properties (e.g. energy)
  - Do we observe more events than predicted by the standard *model* at a given energy?

  → Decision
- Decision is taken against background (the standard model in this case)
- Measurements, parameters and their values are useless without error estimates or given ‘range validity’

Statistic

- What is it?
  - statistic is a quantity that summarises data
  - it is a property of data (e.g. number, mean of distribution)
  - used to make decision
  - we need to know how to treat the data in a statistical sense, i.e. with a view to decision, to use in drawing statistical inference
- Common usage
  - Measuring a quantity or parameter estimation
  - Searching for correlations
  - Testing a model (“hypothesis testing”)
- Statistics allows us to
  - formulate the logic of **what** we are doing and **why**; to make **precise statements**
  - quantify the **uncertainty** in any measurement
  - avoid pitfalls such as **confirmation bias** (distortion of conclusions by preconceived beliefs)

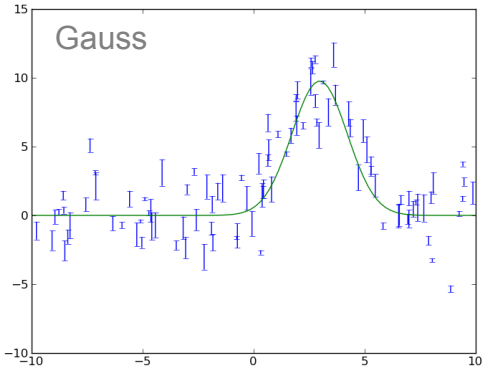
The basic basics: inspect and describe the data

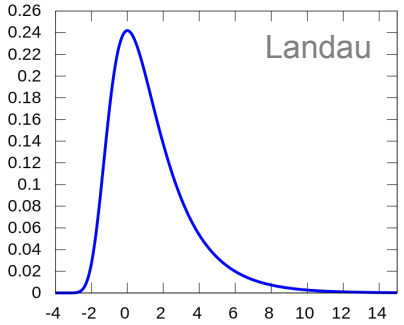
- How to describe the data?
  - by the data itself (may even be a good description of very sparse data)
  - not handy for a large data set
  - use statistics to distill and communicate ‘features’ of data
- Descriptions of large data sets
  - by inspecting the histogram and parametrising it
  - central tendency: mean or median of distribution
  - modes: is there clustering?
  - spread: Variance, RMS or FWHM of distribution
  - tails: Gaussian, Landau, Breit-Wigner
  - outliers: see before

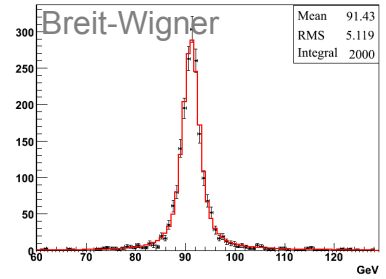
→ All relevant for individual distributions

→ Assume you all know your maths (if not, grab a textbook ☺)

→ *Although very important on its own, not really the focus here*

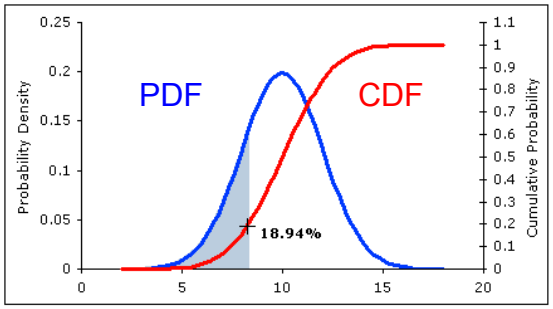






The basic basics: Probability Distributions

- Probability density functions
  - used for continuous random variables (e.g. radioactive decay, particle decay)
  - describe relative likelihood of a random variable to take on a given value
  - non-negative, integral is 1

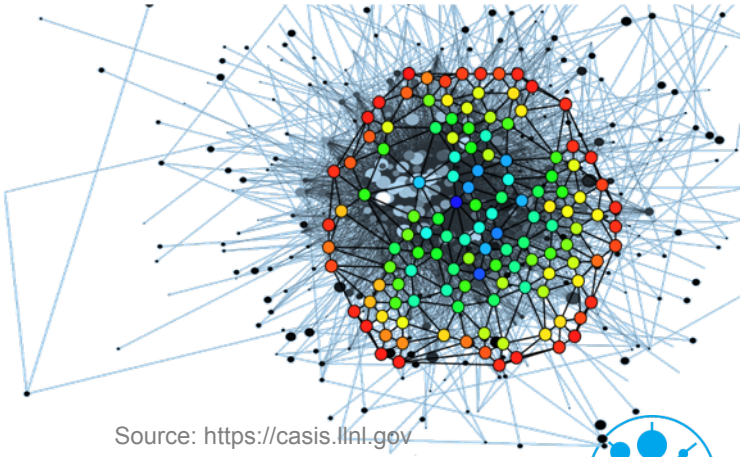


- Central Limit Theorem
  - says that: a random variable defined as the average of a large number of independent and identically distributed random variables is itself ~normally distributed.
  - or in other words: The distribution of a sample mean, drawn from a random distribution, tends toward the normal distribution as the sample size increases.
  - Why this is important: Many statistics have distributions that are ~normal for large sample sizes, even when sampling from a distribution that is not normal.
  - Implies that we can often use well-developed statistical inference procedures that are based on normal distributions, although underlying PDF is not normal.

Machine Learning

What is Machine Learning?

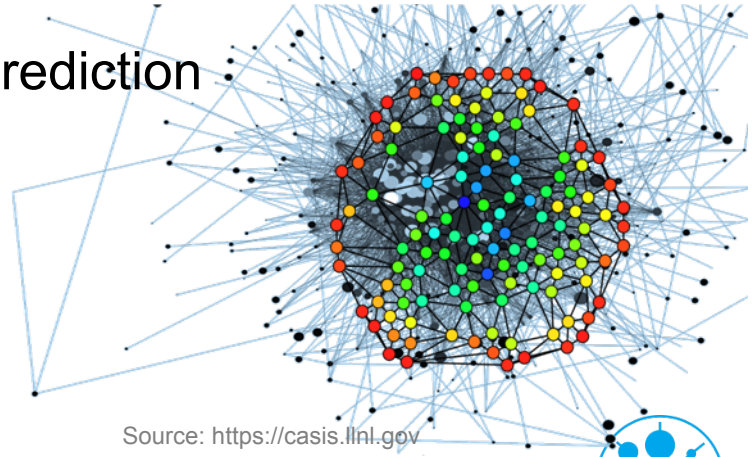
- Everyone has his/her own exact definition
- Wikipedia: "…*is a subfield of computer science, evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.*"



Source: https://casis.llnl.gov

Machine Learning

- What is Machine Learning?
  - Everyone has his/her own exact definition
  - Wikipedia: "*…is a subfield of computer science, evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.*"
- Algorithms are the tools that perform the learning
- Algorithms are building the model, not the user
- The better the input data, the better the prediction
- What are these algorithms?



Source: https://casis.llnl.gov

Machine Learning Categories

- Categorize, based on nature of the learning “signal” or “feedback” available to the system
- Supervised learning
  - algorithm is provided with example input and class labels (the desired output)
  - algorithms try to find a rule that maps inputs to outputs
- Unsupervised learning
  - algorithm is provided with example input data and no class labels
  - algorithm has to find structure in input
  - can be a task on its own (discover hidden patterns in data; search for correlations)
  - Learn at a specific task (find features) and learn the features themselves
- Reinforcement learning
  - Constant interaction of algorithm with dynamical environment to perform goal (e.g. drive car, don’t crash it).
  - Learn the rules of a game by playing it (goal: win game)

Machine Learning Categories

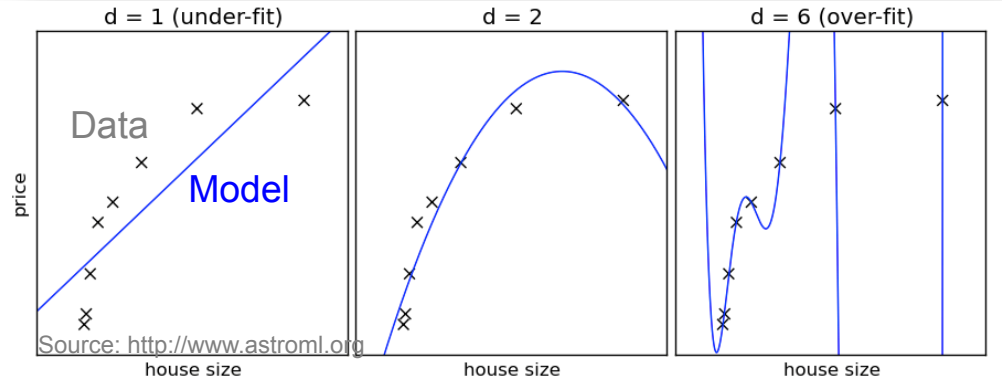
- Categorize, based on the desired output
- Classification
  - inputs are divided in two or more classes; produce model to map input to classes
  - Examples: spam filtering, signal/background classification, particle type
- Regression
  - the output is continuous rather than discrete as for classification
  - Example: What is the energy of an event with properties x,y,z
- Clustering
  - input data is to be divided into groups, which are not known beforehand
  - Where do sports-team supporters live and what is their typical age?
- Density estimation
  - finds the distributions of inputs
- Dimensionality reduction

Other types of tasks and problems

- Learning to learn
  - learns its own inductive bias based on past experience (frequent change of properties necessary to map input to output)
- Developmental learning
  - takes it one step further and generates own sequences of learning situations to acquire skill set
  - employing active learning, maturation, motor synergies, and imitation
  - that does sound like robots and AI, doesn't it?
- Relation to other fields
  - **ML** focuses on predicting, based on *known* properties – **data mining** focuses on discovering, based on *unknown* properties
  - **ML** and **Statistics** are closely related fields, now also with interdisciplinary approaches (Statistical Learning)
  - **Statistics** and ML now more or less combined in **Data Science**

Overfitting and Underfitting

- Overfitting
  - common danger in ML
  - build model that performs well on training sample but generalizes poorly to new data
  - common cause is learning on noisy data, a too small training sample, or learning to identify specific inputs, rather than predictive factors
- Underfitting
  - as overfitting, but does perform badly even on training data
  - typically this involves refactoring of model
- You will see both frequently
  - standard tests to check for over/underfitting are implemented for almost all methods



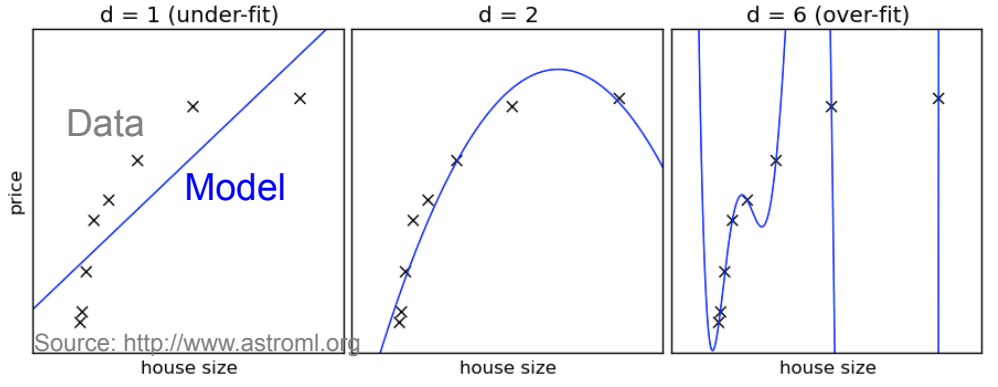
Overfitting and Underfitting

- Way out
  - split data in data in training and test data
  - if methods perform well on test data, first hint that model is generalized and predictive

  SOLUTION: Split data in *training* and *test* sample for one algorithm!

- Many models
  - often, performance of different algorithms should be compared
  - “choose a model that performs best on the test set” → meta-training

  SOLUTION: Split data in *training*, *validation*, and *test* for multiple algorithms



The Bias-Variance Trade-off

- Is another way to look at over/underfitting
- Trade-off between bias and variance
  - imagine randomly drawn points from a linear function f(x) = x
- Underfitting = large bias, small variance
  - fitting a constant will give similar mean, but points are all over the place

→ Solution: increase model complexity, increasing data sample doesn't help

- Overfitting = small bias, large variance
  - fitting a pol9 describes any given set perfectly, but repeating the random drawing will give a very different model

→ Solution: increase training set, or reduce model complexity

Correctness

- Predictive binary models can have 4 answers (example spam mails)
  - True positive: “Message is spam, and we predicted spam.”
  - False positive (type 1 error): “Message is not spam, but we predicted spam.”
  - False negative (type 2 error): “Message is spam, but we predicted not spam.”
  - True negative: “Message is not spam, and we predicted no spam.”
  - Often represented in a confusion matrix:

	Spam	not Spam
predict “Spam”	True positive	False positive
predict “no Spam”	False negative	True negative

  - What is a good measure of such a model?

Correctness

> Example

- Hypothesis/Model: “All kids named Luke will develop leukemia at some point in their life”
- Test: is 98% accurate
- Question: “Is accuracy a good measure for model?
- Lets look at the numbers
  - 0.5% of all babies born are named Luke, prevalence of leukemia is 1.4%
  - Assuming these two are independent the confusion matrix looks like this:

	Leukemia	No Leukemia	Total
“Luke”	70	4,930	5,000
not “Luke”	13,930	981,070	995,000
Total	14,000	986,000	1,000,000

- Fraction of correct predictions: **accuracy** = (tp + tn) / all = 98.114% → bad test
- How accurate were positive predictions: ***precision*** = tp / (tp + fp) = 1.4%
- What fraction of the positives were identified by the model: ***recall*** = tp / (tp + fn) = 0.5%
- Typical model is a trade-off between precision and recall

PART 1

> Summary

- Data explosion in the coming years
- Data is the bread and butter for physicists/scientists, but they also bring the skill set to do data science
- Data needs to be retrieved, prepared and cleaned before usage
- Data needs to be parameterized (e.g. PDFs) for generalization
- Plenty of toolkits available to do data analysis – some with more focus on statistical methods than others
- Machine Learning algorithms build a model and try to learn from data and make predictions on data
- Different classes of algorithms are used for different tasks (e.g. supervised vs. unsupervised learning)

→ PART 2: Discuss different ML algorithms