Best Practice Analytics

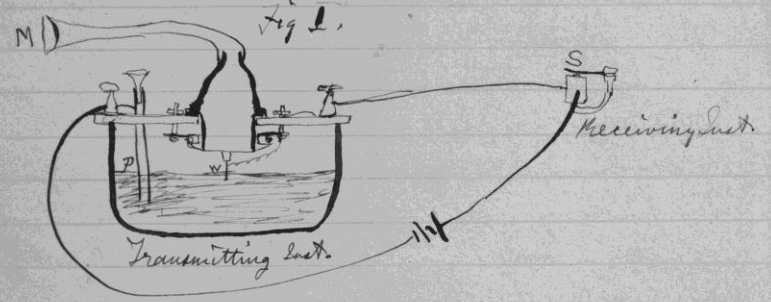
How to go about data analysis projects

Dr Detlef Nauck

Chief Research Scientist for Data Science
BT, Research & Innovation

March 10th 1876

Fig. I.



1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece — and S the armature of the Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting Instrument was placed in another room and the doors of both rooms were closed.

I then shouted into M the following sentence: "Mr. Watson — Come here — I want to

see you". To my delight he came and declared that he had heard and understood what I said.

I asked him to repeat the words — ~~He said~~ He answered "You said 'Mr. Watson — come here — I want to see you'." We then changed places and I listened at S while Mr. Watson read a few passages from a book into the mouth piece M. It was certainly the case that articulate sounds proceeded from S. The effect was loud but indistinct and muffled.

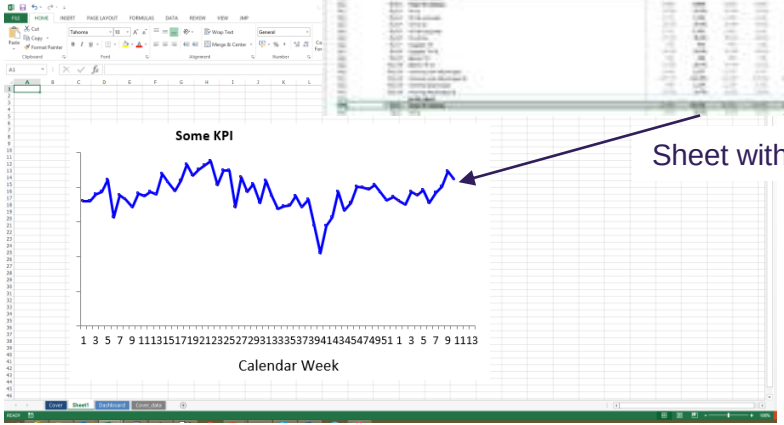
If I had read beforehand the passage given by Mr. Watson I should have recognized every word. As it was I could not make out the sense — but an occasional word here and there was quite distinct. I made out "to" and "out" and "further"; and finally the sentence "Mr. Bell Do you understand what I say? DO-YOU-UN-DER-STAND-WHAT-I-SAY" came quite clearly and intelligibly. No sound was audible when the armature S was removed.

Why We Need Best Practice Analytics

A Typical Spreadsheet-based Reporting Scenario

Decision maker

Weekly report in a spreadsheet

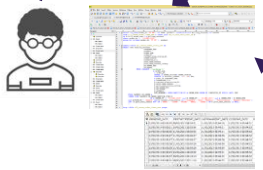


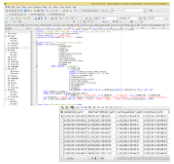
Cover sheet with selected charts

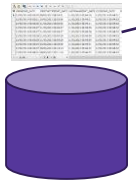
Sheet with hundreds of KPIs

IT personnel using undocumented scripts to produce tables in personal schemas

Business owner of KPI receives a weekly number

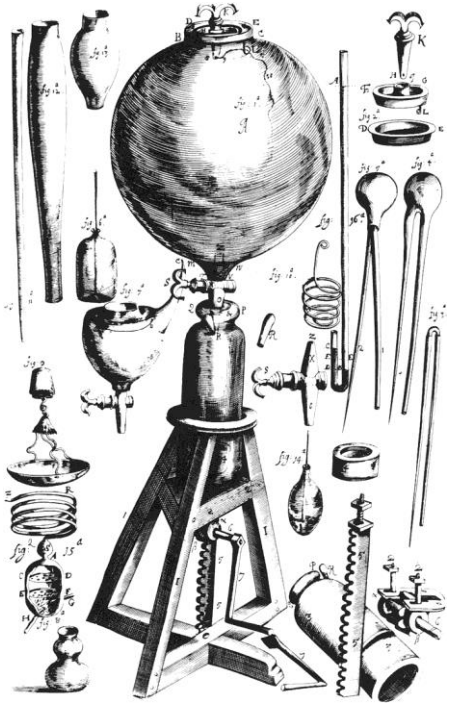






Undocumented table in enterprise data warehouse

Reproducibility



Boyle's air pump (see https://en.wikipedia.org/wiki/Reproducibility)

- Reproducibility is the ability of an entire experiment or study to be duplicated (by someone else).
- Reproducibility is one of the main principles of the scientific method.
- The term *reproducible research* refers to the idea that the ultimate product of academic research is the paper along with the *laboratory notebooks* and *full computational environment* used to produce the results in the paper such as the *code, data*, etc. that can be used to reproduce the results and create new work based on the research.

Reproducible Research

- Reproducible does not mean correct
- Your analysis/claims can be reproducible and can still be wrong
- Reproducibility is the only thing an analyst can guarantee about a study.
- Example: Thomas Piketty's book on 'Capital in the Twenty-First Century'. All the data and analysis has been made available on the web http://piketty.pse.ens.fr/en/capital21c2

Reproducible Research – Online Course

coursera

Catalog | Search catalog

Institutions Log In Sign Up

Overview

Syllabus

FAQs

Creators

Pricing

Ratings and Reviews

Reproducible Research

Enroll Now
Started Jan 09

Financial Aid is available for learners who cannot afford the fee. Learn more and apply.

Home > Data Science > Data Analysis

Reproducible Research

About this course: This course focuses on the concepts and tools behind reporting modern data analyses in a reproducible manner. Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. The need for reproducibility is increasing dramatically as data analyses become more complex, involving larger datasets and more sophisticated computations. Reproducibility allows for people to focus on the actual content of a data analysis, rather than on superficial details reported in a written summary. In addition, reproducibility makes an analysis more useful to others because the data and code that actually conducted the analysis are available. This course will focus on literate statistical analysis tools which allow one to publish data analyses in a single document that allows others to easily execute the same analysis to obtain the same results.

Show less

Created by: Johns Hopkins University

JOHNS HOPKINS UNIVERSITY

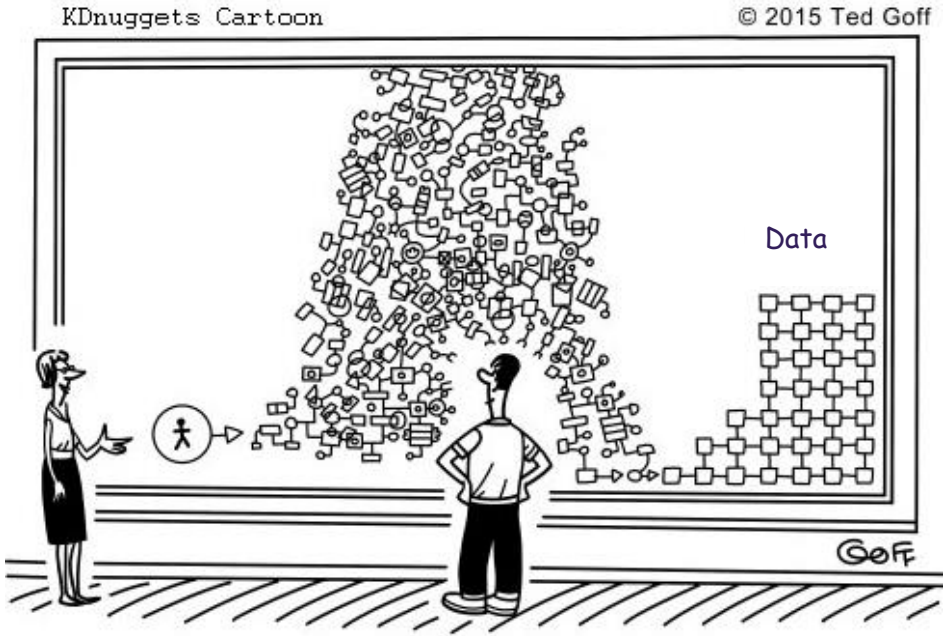
https://www.coursera.org/learn/reproducible-research

BT

Test Driven Analytics – Learn from Software Development

- How do you make sure that your analysis is correct?
  - Is your data any good?
  - Have you validated your model?
  - Do decisions taken have the expected business impact?
- Use a test-driven approach for data wrangling, modelling and deployment

- *Test-driven development (TDD) is an advanced technique of using automated unit tests to drive the design of software and force decoupling of dependencies. The result of using this practice is a comprehensive suite of unit tests that can be run at any time to provide feedback that the software is still working (Microsoft: Guidelines for test-driven development)*
  https://msdn.microsoft.com/en-us/library/aa730844(v=vs.80).aspx



“This is you, these are organisational and regulatory obstacles, and this is the operational data you want. Welcome to the Data Science Team!“

(Original at http://www.kdnuggets.com/2015/03/cartoon-us-chief-data-scientist-challenge.html)

Data Quality Issues in Enterprises

- Operational data schemas are not designed for analytics.
- Data is distributed across silos and access can be difficult (administrative and computational obstacles).
- Data inconsistencies make joining data across operational domains difficult and error prone.
- It won't get better (human nature, operational pressure, lack of business case) - actually, Big Data makes it worse.
- Effort in data preparation is and will remain typically 80%-90% of project effort.

Making Assertions about Data

- While you pre-process your data or while you conduct exploratory data analysis you should gain some insight into properties of the data.
- Think about which properties are essential and which you may want to check or guarantee before you conduct any further analysis.
- Write some code to demonstrate that your data is correct.

Data Quality Test in a Jupyter Notebook Using R

jupyter Data Assertions Last Checkpoint: 22 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help

Markdown Cell Toolbar: None

Function to test data assertions on the RFT Churn data set

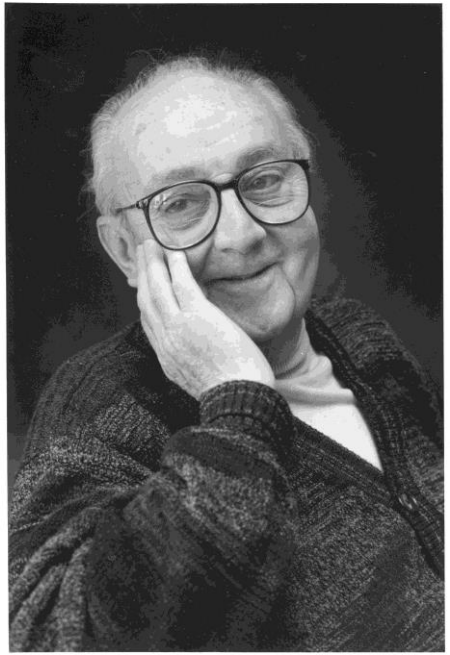
In [12]:
```
test_rft = function(df){
  tests = 4
  testframe = data.frame(test=integer(tests+1), label=character(tests+1), stringsAsFactors = FALSE)
  testframe$label[1] = "Difference between number of churn rows and number of rows with a cease product"
  testframe$test[1] = sum(df$churn) - nrow(df[df$CEASE_PROD != "",])

  testframe$label[2] = "Rows where cease product is empty, but days to cease is not"
  testframe$test[2] = nrow(df[df$CEASE_PROD=="" & df$DAYS_TO_CEASE != "",])

  testframe$label[3] = "Rows where days to cease is empty, but cease product is not"
  testframe$test[3] = nrow(df[df$CEASE_PROD=="" & df$DAYS_TO_CEASE != "",])

  testframe$label[4] = "Number of unexpected install product labels"
  testframe$test[4] = sum(!levels(rft_churn$INSTALL_PROD) == c("","B","BL","BLV","BV","L","LV","V" ))

  testframe$label[tests+1] = "Overall Test Result"
  testframe$test[tests+1] = sum(testframe$test, na.rm = TRUE)
  testframe
}
```

Running the Test Script

In [11]:
```
test_rft(rft_churn)
```

Out[11]:

	test	label
1	0	Difference between number of churn rows and number of rows with a cease product
2	0	Rows where cease product is empty, but days to cease is not
3	0	Rows where days to cease is empty, but cease product is not
4	0	Number of unexpected install product labels
5	0	Overall Test Result



George Edward Pelham Box
(18 October 1919 – 28 March 2013)
British statistician

All models are wrong, but some are useful.

George E.P. Box, 1978

https://en.wikipedia.org/wiki/All_models_are_wrong

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.
George E.P. Box, 1976

Model Building: Cross-Validation

1 Train	2 Train	3 Train	4 Train	5 Test	6 Train	7 Train	8 Train	9 Train	10 Train

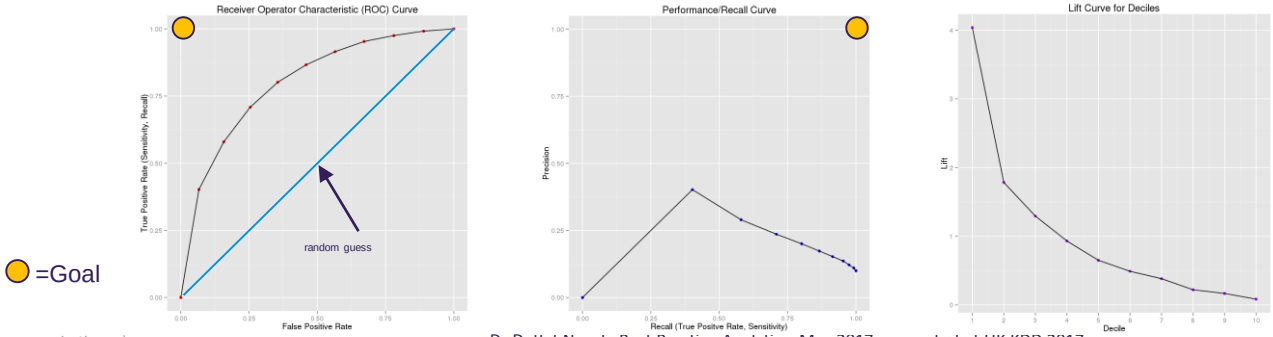
- Randomly split your data into K subsets (folds). Typically, we pick K=10.
- Keep one fold for testing and train the model on the other (K-1) folds. Repeat K times, each time keep another fold for testing (picture above shows step 5 out of 10).
- The average error across all runs is the estimate for the average expected error on unseen data.
- Build the final model using all data.
- *Careful*: all modelling steps, including feature selection must be part of the cross-validation. That includes any step that used the target information. Otherwise the error is underestimated.
- You can execute unsupervised steps before cross-validation, i.e. pick features with high variance. “Unsupervised step” means that the target information is not used.

Cross-Validation Caveats

- If you have only a small amount of data, cross-validation can over-estimate the error.
- If K=N (size of data) you have 1-leave-out validation. This is useful for small data sets and does not over-estimate the error, but the error variance can be over-estimated.
- Cross validation tells you something about the expected error of your modelling method. You can use it to pick a preferred method.
- “Method” can mean the way you pick the model parameters or different modelling approaches (e.g. tree vs. neural network).
- If you have vastly different outcomes in different runs, your data selection strongly influences the modelling results. Check if you have enough data.

Understanding if a Predictive Model is Useful

- Just quoting the accuracy of a model can be misleading especially in cases where the data is extremely skewed and the target class is rare (e.g. crime, churn, fraud, faults, …).
  Example: if the churn probability is 1% then a model that classifies everyone as non-churner is 99% correct but 100% pointless.
- Ideally, we want to understand the trade-off between benefit and cost of a model which can be expressed in different ways depending on context.
- Always use cross-validation when building a model.
- Look at ROC, Performance/Recall and Lift charts

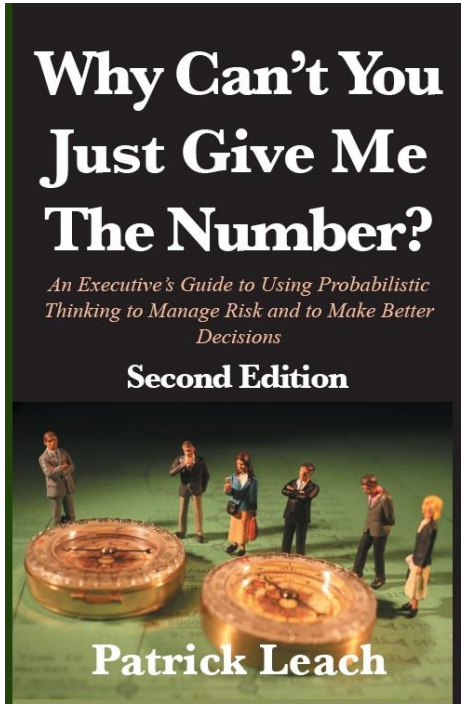


=Goal

Challenges in Analytics (especially Machine Learning)

- Data Provenance
  Where does the data come from, how was it collected and can I trust it?
- Data Quality
  Is the data error free?
- Data Bias
  Does the data accurately reflect the population/situation I want to model? What is missing?
- Model Bias
  Driven by Data Bias, or bad sampling – to what extent is my model biased by the training data?
- Model Comprehension
  Why are the outputs of my (black box) models the way they are?

Decision Making Based on Analytics

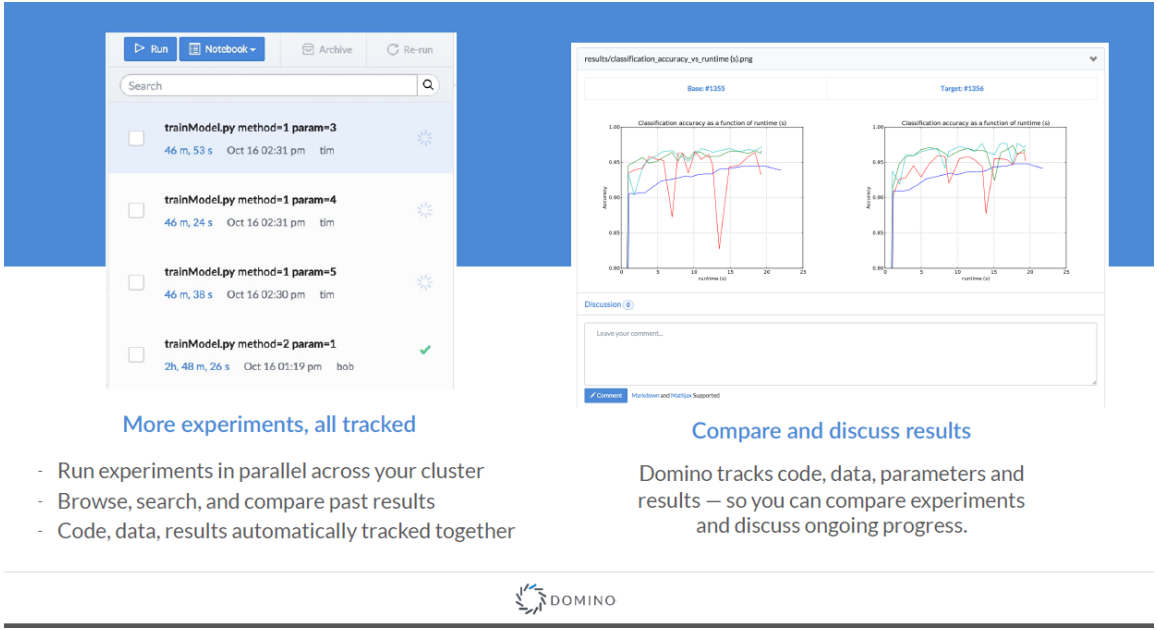


- Executives are not necessarily familiar with analytics
- Analysts can find it difficult to present results as compelling business cases
- Managers must become comfortable with systematic experimentation with rigorous controls to determine cause and effect.
- Use prescriptive analytics (outcome + action)
- Revisit deployments in regular intervals

Challenges in Enterprises – Enable Best Practice Analytics

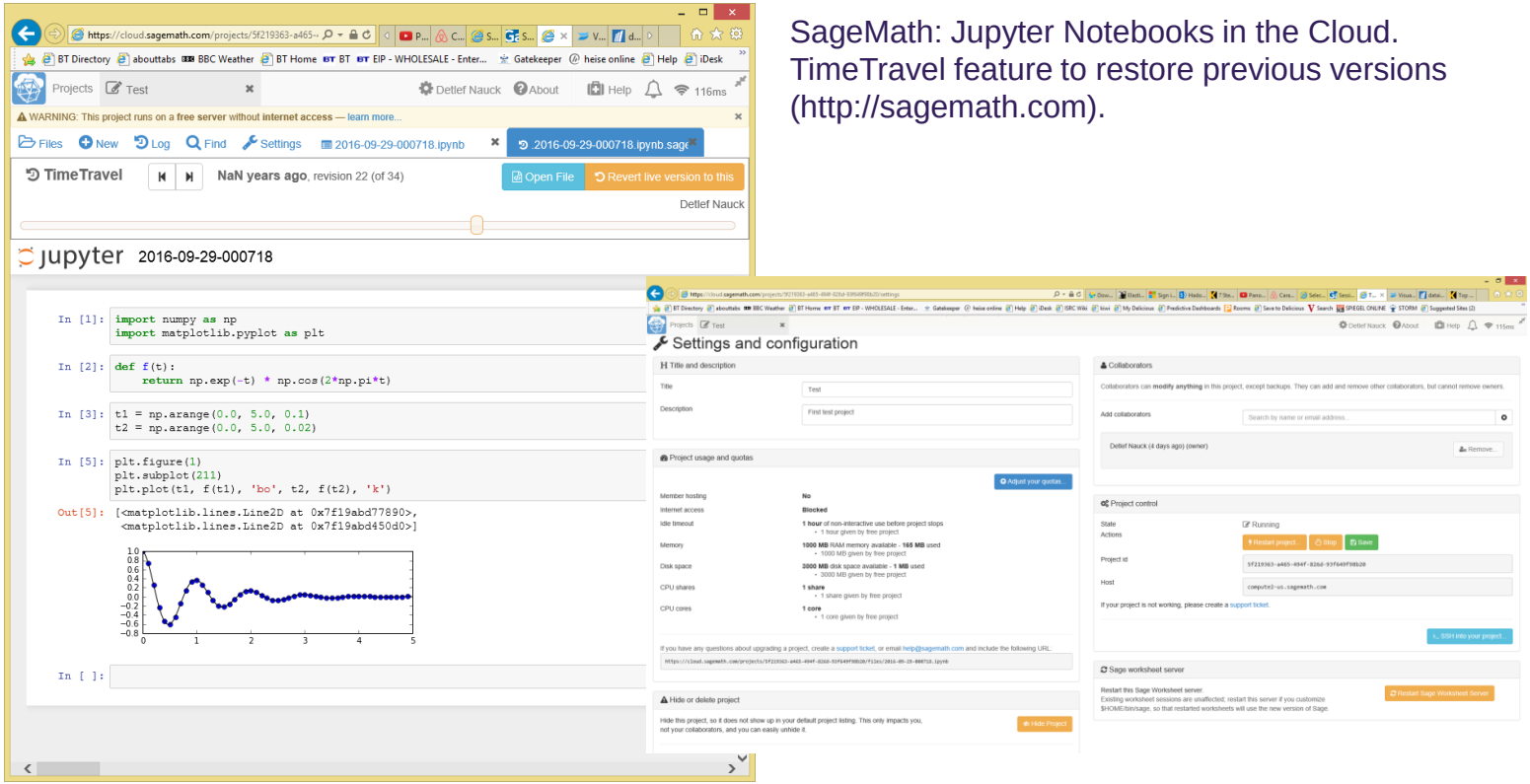
- Compute power for analysts (internal cloud instead of corporate laptops)
- Centrally configured environments with flexibility for individual requirements (allow installation of tools and libraries)
- Use of Open Source and COTS tools for different analytic communities
- Support for teams and collaboration
- Reusability and sharing of work
- Preserve, version control, and quickly reproduce work
- Fast deployment of results and models
- Automation

Cloud-based Tools – Versioning and Collaboration



Domino – cloud based environment for analytics (Notebooks, RStudio), versioning, collaboration, and deployment (http://dominodatalabs.com)

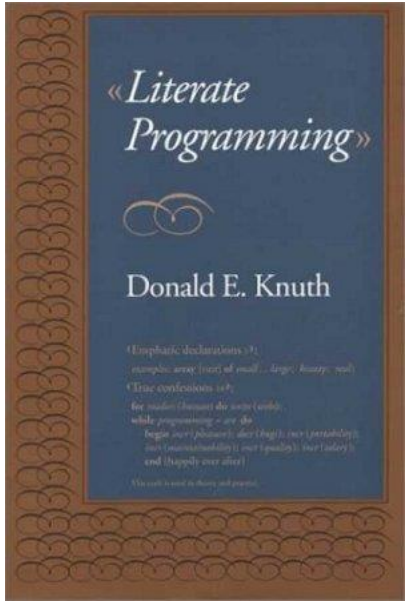
Cloud-based Tools – Versioning and Collaboration



SageMath: Jupyter Notebooks in the Cloud. TimeTravel feature to restore previous versions (http://sagemath.com).

Literate Programming

The program logic is described in natural language interspersed with traditional code from which compilable code can be generated (Donald Knuth, 1992).



Tools:

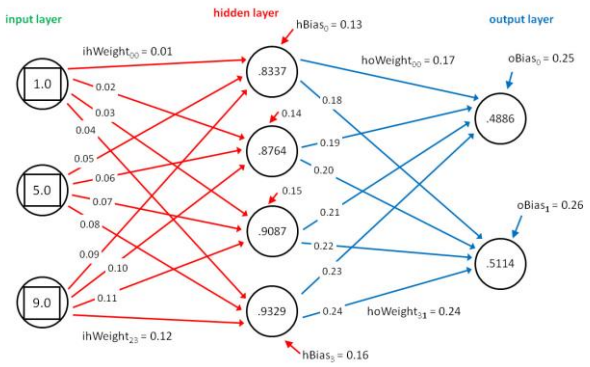
- Jupyter Notebooks (or other flavours) based on Python supporting 50+ languages
- Knitr (R package for dynamic report generation, supported in Rstudio)
- R Markdown (based on knitr and pandoc) allows you to combine R code with markdown (shorthand for HTML, like you can use in Jupyter) and compile the documents into reports of different formats
- Rstudio 1.0 now provides a Notebook interface

In Conclusion

- Write tests and assertions about your data
- Use cross-validation for modelling and evaluate models comprehensively
- Use prescriptive analytics and systematic experimentation for deployment
- Version, document and keep everything (data, code, tools, libraries, successful and unsuccessful experiments, …)

Trends in ML: Deep Networks

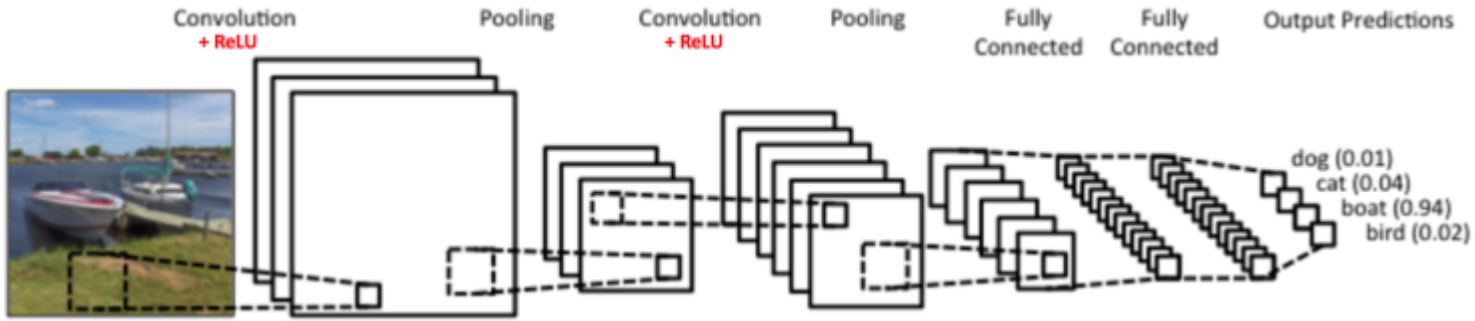
From:



Multilayer Perceptron

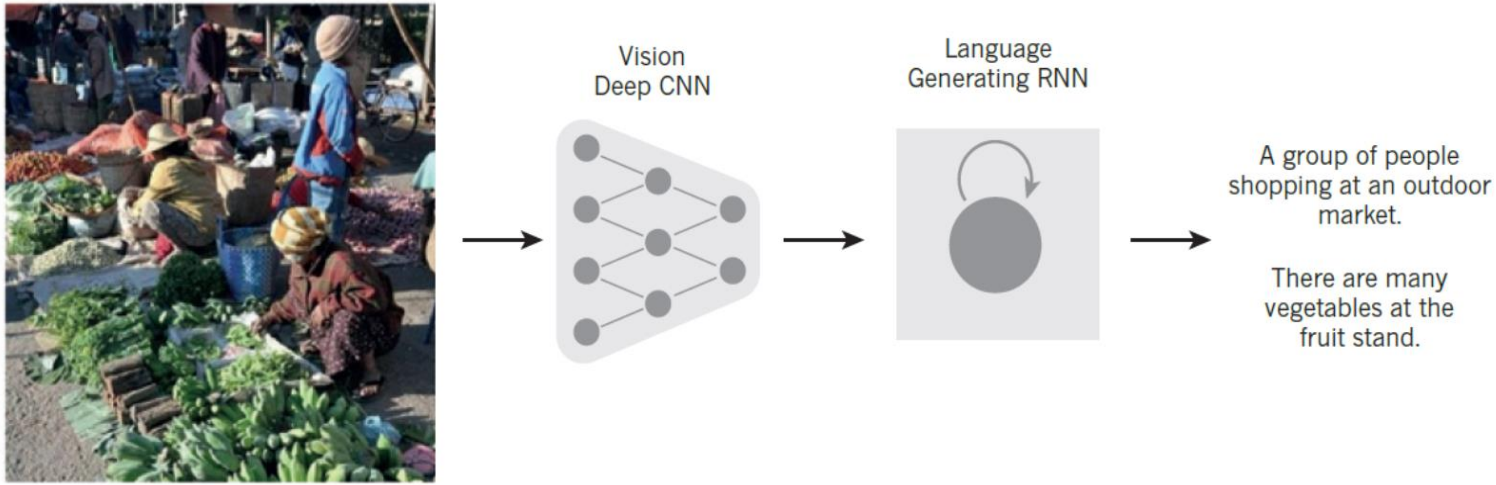
- Trained through Backpropagation (invented in 1980s)
- For large number of hidden layers training is slow
- Real world problems need many hidden units and a lot of labelled data
- Advances through structured hidden layers, improved algorithms, specialised hardware (GPUs) and Big Data.

To:



Machine-learning "programmers" design the network structure with experience and by trial and error

Example: Combining Deep Networks



Yann LeCun, Yoshua Bengio, & Geoffrey Hinton (2015). Deep Learning, Nature, Vol. 521, (pp. 436-444)

A deep convolution neural net (CNN) produces a set of outputs (abstract "words")

A language-generating recurrent neural net (RNN) "translates" the abstract "words" into captions

Challenges with Deep Networks

Statistically impressive,
but individually unreliable

Challenges with Deep Networks

“Panda”

+

<1%
targeted distortion

=

“Gibbon”
(99.3% confidence)

Inherent flaws can be exploited

Challenges with ML – New Application Areas

Example: Algorithms to Predict Likelihood of Criminal Conduct

DJIA Futures ▲ 18854 0.11% S&P 500 F ▲ 2176.00 0.15% Stoxx 600 ▲ 338.92 0.13% U.S. 10 Yr ▲ 6/32 Yield 2.200% Crude Oil ▲ 46.09 1.14% Euro ▲ 1.0735 0.41%

THE WALL STREET JOURNAL.

Subscribe Now | Sign In

SPECIAL OFFER: JOIN NOW

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

Search

Minnesota Officer Faces Manslaughter Charges in Shooting of Black Man

In Maine's Tight Job Market, Businesses Look to Immigrants

Sign-Ups Under Affordable Care Act So Far Seem Not Hurt by Trump Win

Trump's Made-in-America Hurdle: Asia

U.S.

Wisconsin Supreme Court to Rule on Predictive Algorithms Used in Sentencing

Ruling would be among first to speak to legality of risk assessments as aid in meting out punishments

The Wisconsin Supreme Court, located in the State Capitol, is set to rule on whether algorithms used to predict criminality can be used in sentencing. PHOTO: GETTY IMAGES/SCIENCE SOURCE

By JOE NAVARRO

June 5, 2016 5:30 p.m. ET

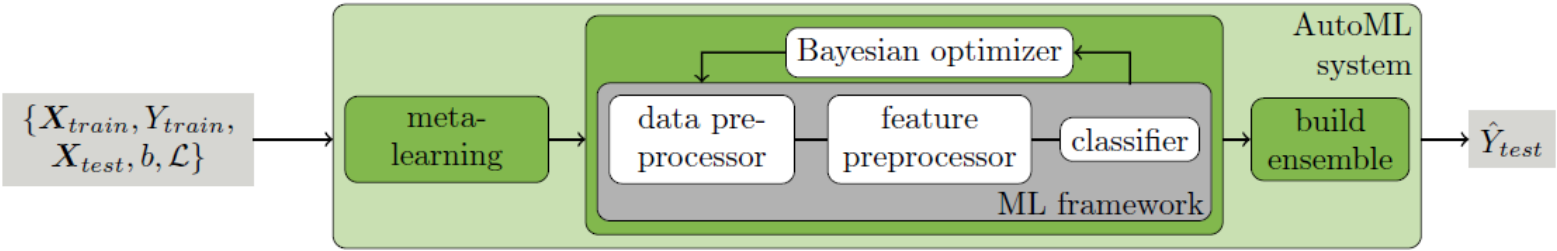
0 COMMENTS

http://www.wsj.com/articles/wisconsin-supreme-court-to-rule-on-predictive-algorithms-used-in-sentencing-1465119008

Models assessing individuals are highly problematic

- Independent tests show that these algorithms only about 60%-70% correct.
- False positives can be devastating for individuals
- High risk for machine bias if features like race are used.
- Models are opaque and cannot be challenged.
- It is illegal in the UK to use algorithms for assigning risk to individuals (exception: credit score).
- New legislation: GDPR

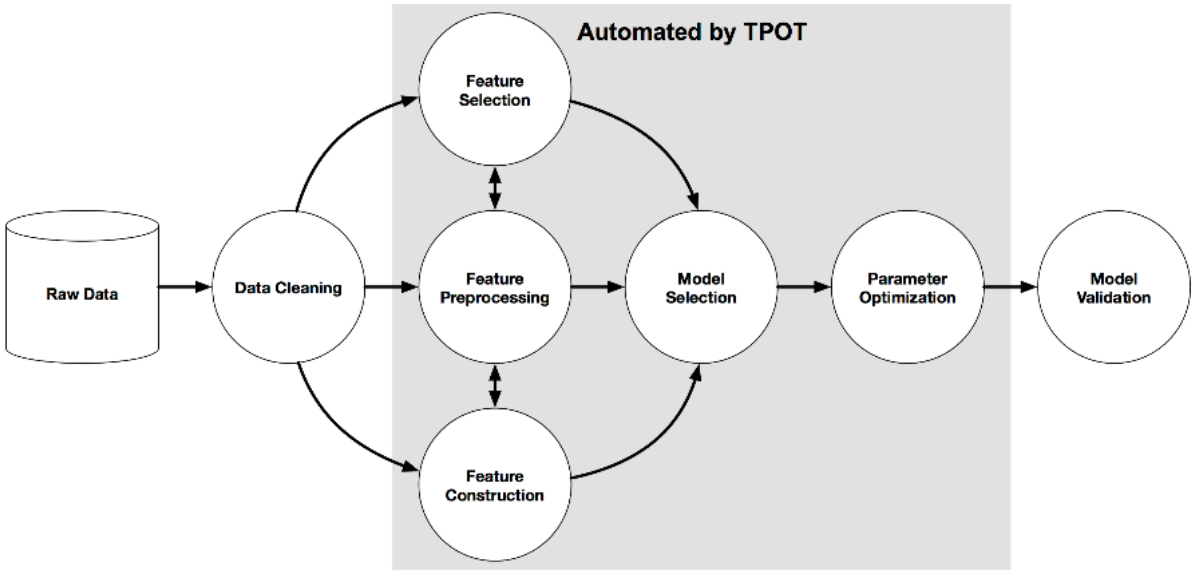
Trends in ML: AutoML – Automatic (Automated) ML



Auto-sklearn: a python package based on scikit-learn

- Bayesian optimisation is used to optimise hyper-parameters of learning algorithms

Trends in ML: AutoML – Automatic (Automated) ML



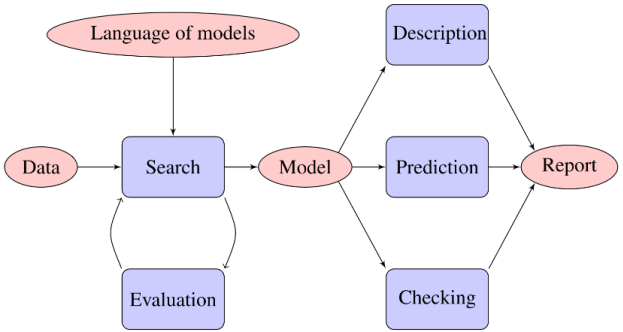
TPOT: a python package based on scikit-learn

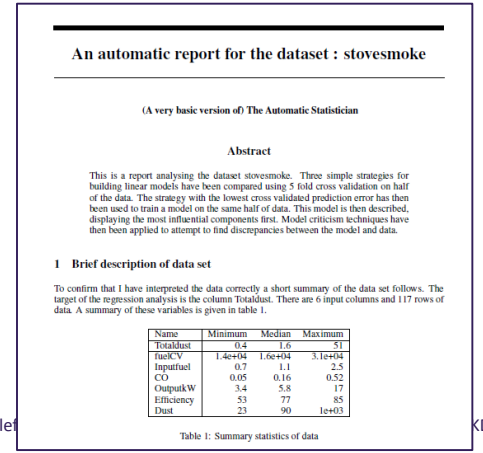
- Uses genetic programming to test combinations of feature representations, algorithms and hyper-parameters.
- Produces scikit-learn pipeline of the best performing solution.

The Automatic Statistician

(Zoubin Gharamani's group, Cambridge, www.automaticstatistician.com)

- Idea:
  - have a language that can describe arbitrarily complicated models
  - a method to search over those models
  - a procedure to check model fit
- So far: Implementation for a grammar of Gaussian Processes which can be used for Bayesian Regression
- Generates natural language report.





An automatic report for the dataset : stovesmoke

(A very basic version of) The Automatic Statistician

Abstract

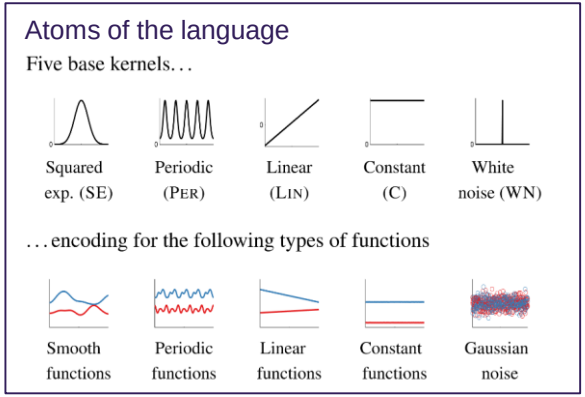
This is a report analysing the dataset stovesmoke. Three simple strategies for building linear models have been compared using 5 fold cross validation on half of the data. The strategy with the lowest cross validated prediction error has then been used to train a model on the same half of data. This model is then described, displaying the most influential components first. Model criticism techniques have then been applied to attempt to find discrepancies between the model and data.

1 Brief description of data set

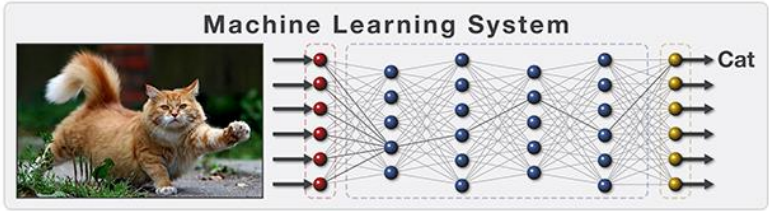
To confirm that I have interpreted the data correctly a short summary of the data set follows. The target of the regression analysis is the column Totaldust. There are 6 input columns and 117 rows of data. A summary of these variables is given in table 1.

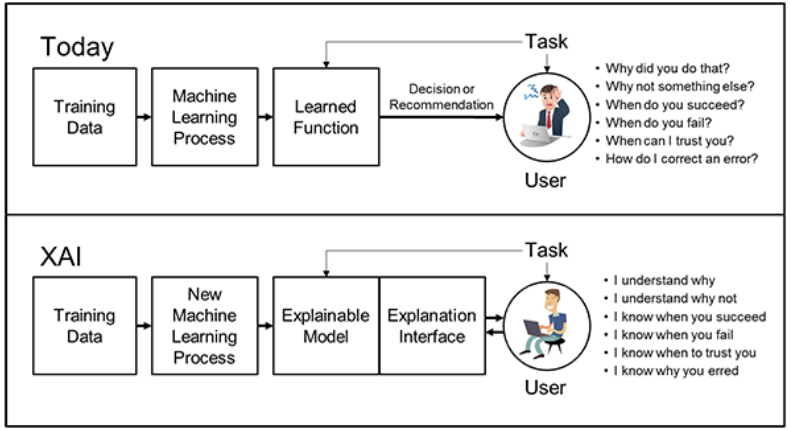
Name	Minimum	Median	Maximum
Totaldust	0.4	1.6	51
fuelCV	1.4e+04	1.6e+04	3.1e+04
Inputfuel	0.7	1.1	2.5
CO	0.05	0.16	0.52
OutputkW	3.4	5.8	17
Efficiency	53	77	85
Dust	23	90	1e+03

Table 1: Summary statistics of data



The Third Wave in AI: XAI – Explainable AI



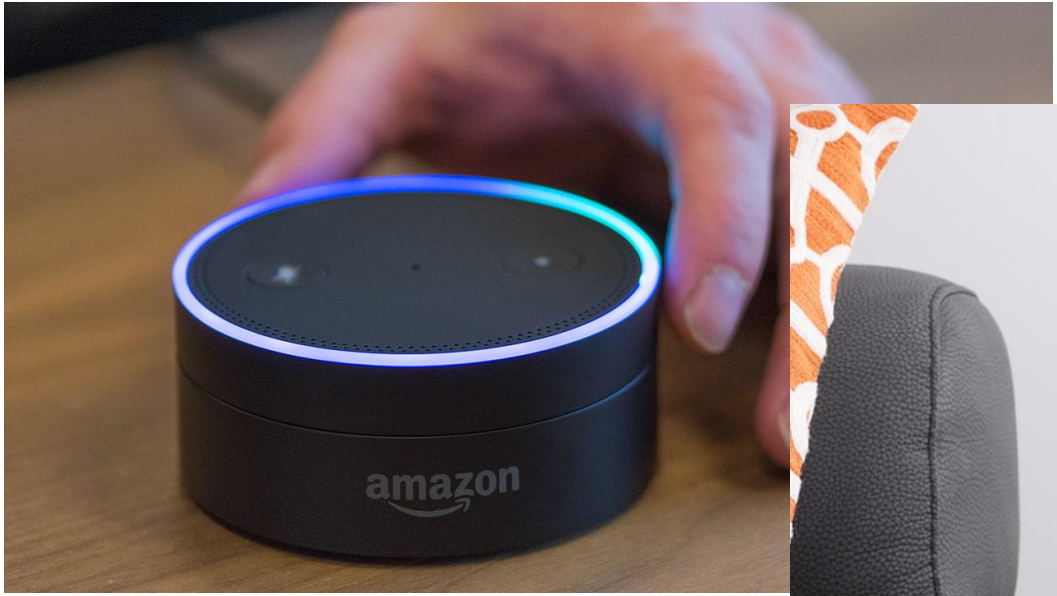


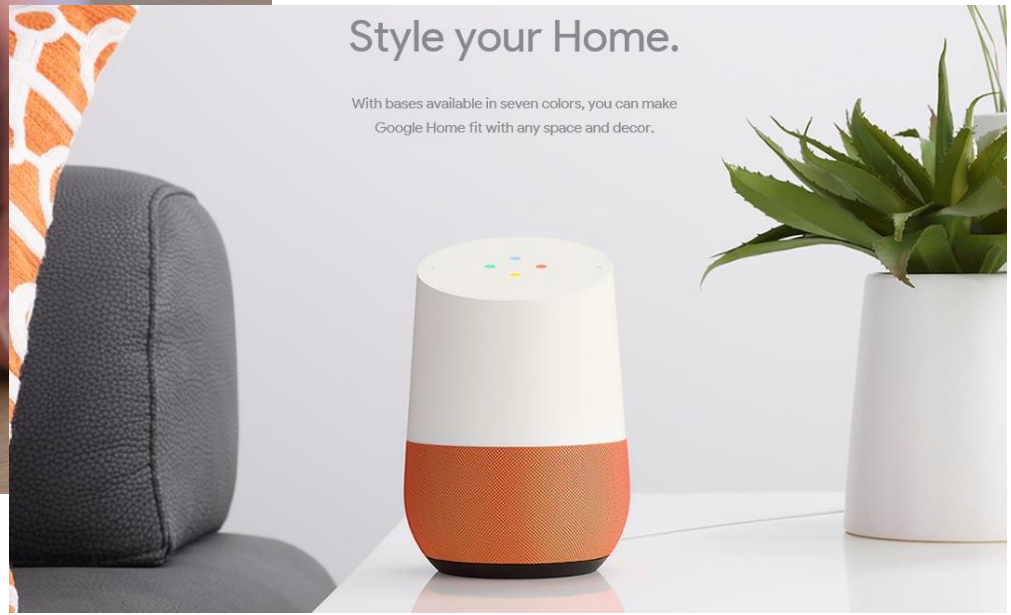
Dr Detlef Nauck: Best Practice Analytics, May 2017, presented at UK KDD 2017

http://www.darpa.mil/program/explainable-artificial-intelligence

The Immediate Future

Home Automation and Intelligent Virtual Assistants





Context aware question-answering virtual assistants
with plugin-interfaces for apps that provide new capabilities