Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Use Blackboard´s forum if a question may be relevant to other students, too. Email always both joeran.beel@scss.tcd and doug.leith@scss.tcd.ie. Give a meaningful subject, starting with "*[ML1819]*". No file attachments.

Week 06 & 08: The Machine Learning Pipeline

CS7CS4/CS4404 Machine Learning
v5 2018-10-31

Dr Joeran Beel
Assistant Professor in Intelligent Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

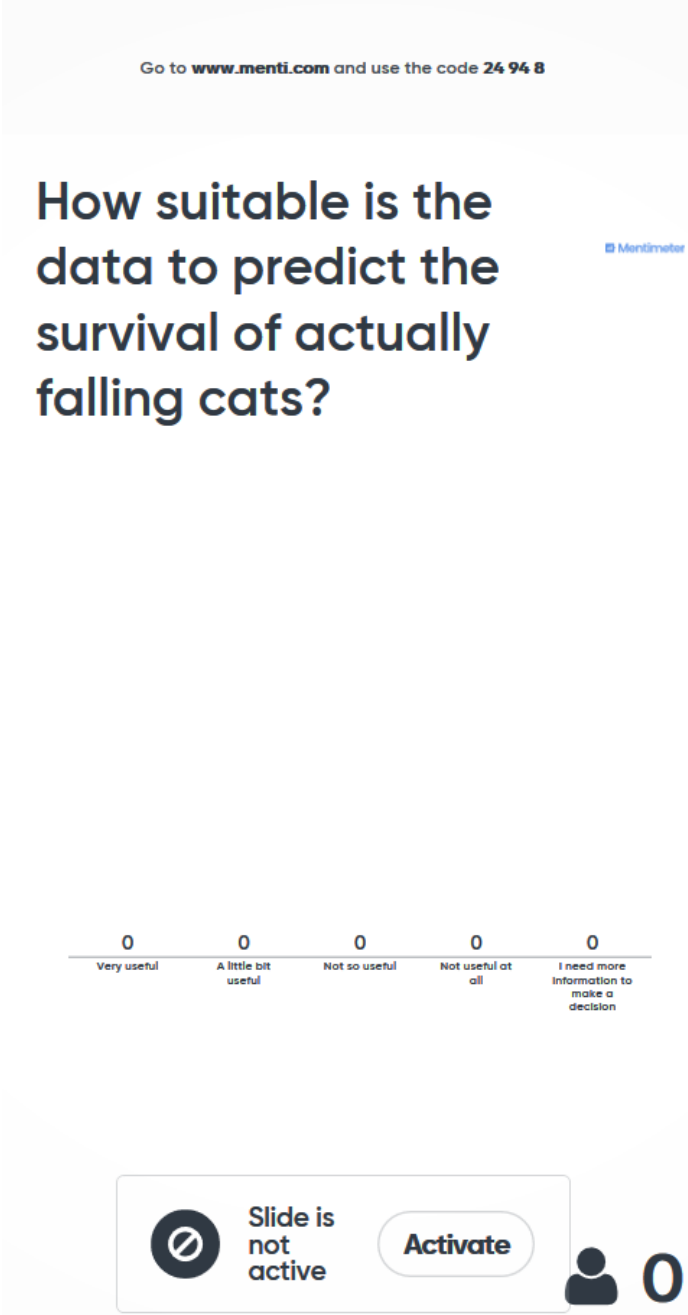
Dr Douglas Leith
Professor in Computer Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

Any questions from previous lecture?

https://kaiserhealthnews.files.wordpress.com/2017/02/khnoncall-2-2.jpg?w=1024

Case Study (Homework)

EVERY year, scores of cats fall from open windows in New York City. From June 4 through Nov. 4, 1984, for instance, 132 such victims were admitted to the Animal Medical Center on 62d Street in Manhattan. Most of the cats landed on concrete. Most survived. Experts believe they were able to do so because of the laws of physics, superior balance and what might be called the flying-squirrel tactic. In a study for the medical center, Dr. Wayne Whitney and Dr. Cheryl Mehlhaff recorded the distance of the fall for 129 of the 132 cats. The falls ranged from 2 to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect. Seventeen of the cats were put to sleep by their owners, in most cases not because of life-threatening injuries but because the owners said they could not afford medical treatment. Of the remaining 115, 8 died from shock and chest injuries. Even more surprising, the longer the fall, the greater the chance of survival. Only one of 22 cats that plunged from above 7 stories died, and there was only one fracture among the 13 that fell more than 9 stories. The cat that fell 32 stories on concrete, Sabrina, suffered a mild lung puncture and a chipped tooth. She was released from the hospital after 48 hours. The cat's ability to twist around while falling and land on its feet is well known. But why did cats from higher floors fare better than those on lower ones? One explanation is that the speed of the fall does not increase beyond a certain point, Dr. Mehlhaff and Dr. Whitney said in the December 1987 issue of The Journal of the American Veterinary Medical Association. This point, ''terminal velocity,'' is reached relatively quickly in the case of cats. Terminal velocity for a cat is 60 miles per hour; for an adult human, 120 m.p.h. Until a cat reaches terminal velocity, the two speculated, the cat reacts to acceleration by reflexively extending its legs, making it more prone to injury. But after terminal velocity is reached, they said, the cat might relax and stretch its legs out like a flying squirrel, increasing air resistance and helping to distribute the impact more evenly. ''Cats may be behaving like well-trained paratroopers,'' Dr. Jared Diamond, who teaches physiology at the University of California at Los Angeles Medical School, wrote in the August issue of the magazine Natural History.

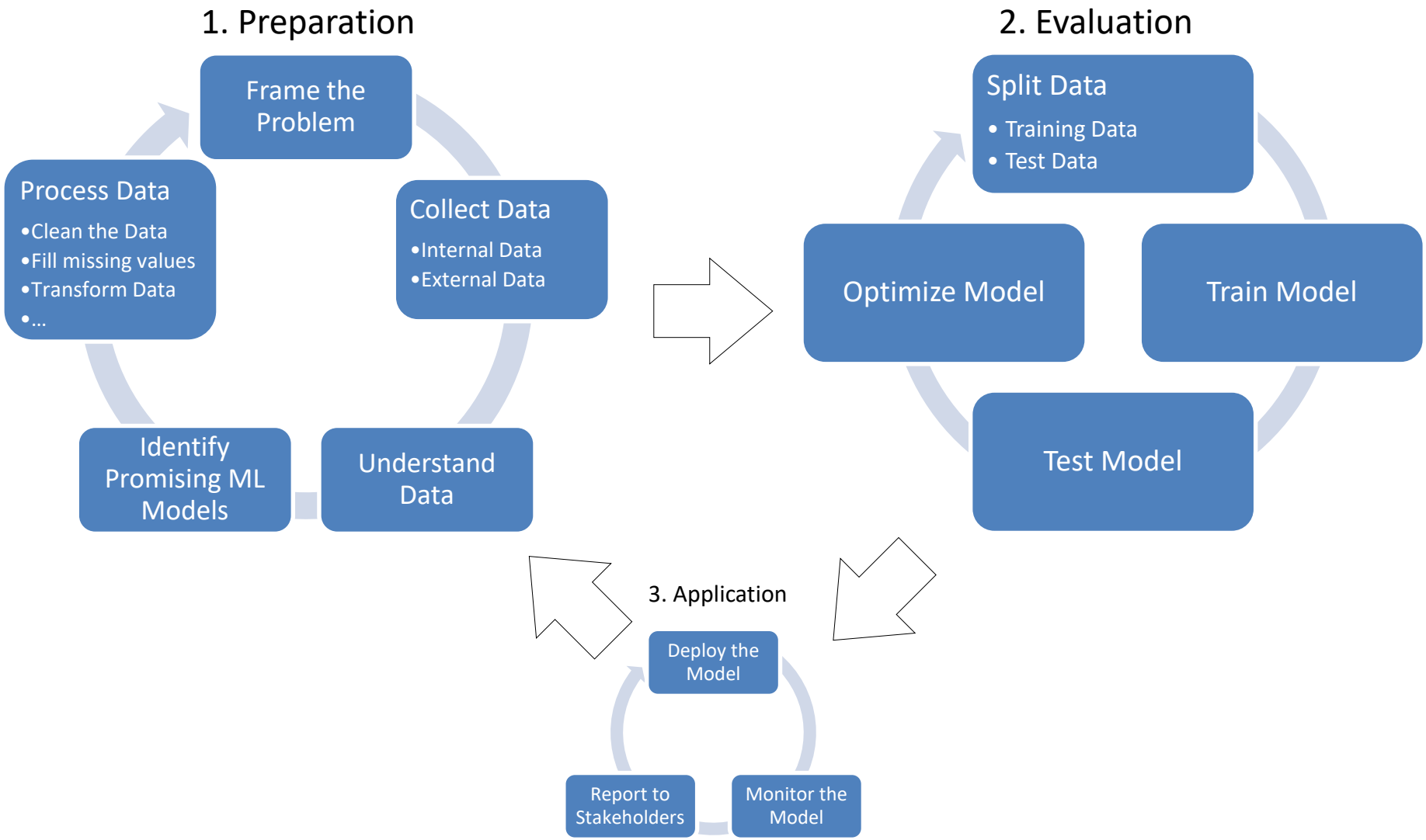


http://www.nytimes.com/1989/08/22/science/on-landing-like-a-cat-it-is-a-fact.html

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

The Machine-Learning Workflow

The Machine-Learning Workflow



“At large companies, machine learning is 80 percent infrastructure.”

https://techcrunch.com/2017/08/08/the-evolution-of-machine-learning/

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Frame the Problem

Frame the problem and Look at the Big Picture

1. **Define the objective in business terms**
2. **How will your solution be used**
3. **What are the current solutions/workarounds (if any)?**
4. **What machine learning models are appropriate to achieve the objective? (supervised vs. unsupervised; online vs. offline)**
5. **How should performance be measured?**
   - Is the measure aligned with the business objective?
   - What is the minimum performance that needs to be achieved
6. **What are comparable problems, and how have they been solved?**
7. **Which existing tools can you use to achieve the objective?**
8. **List the assumptions you (or others) have made so far**
9. **Verify assumptions if possible**

Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

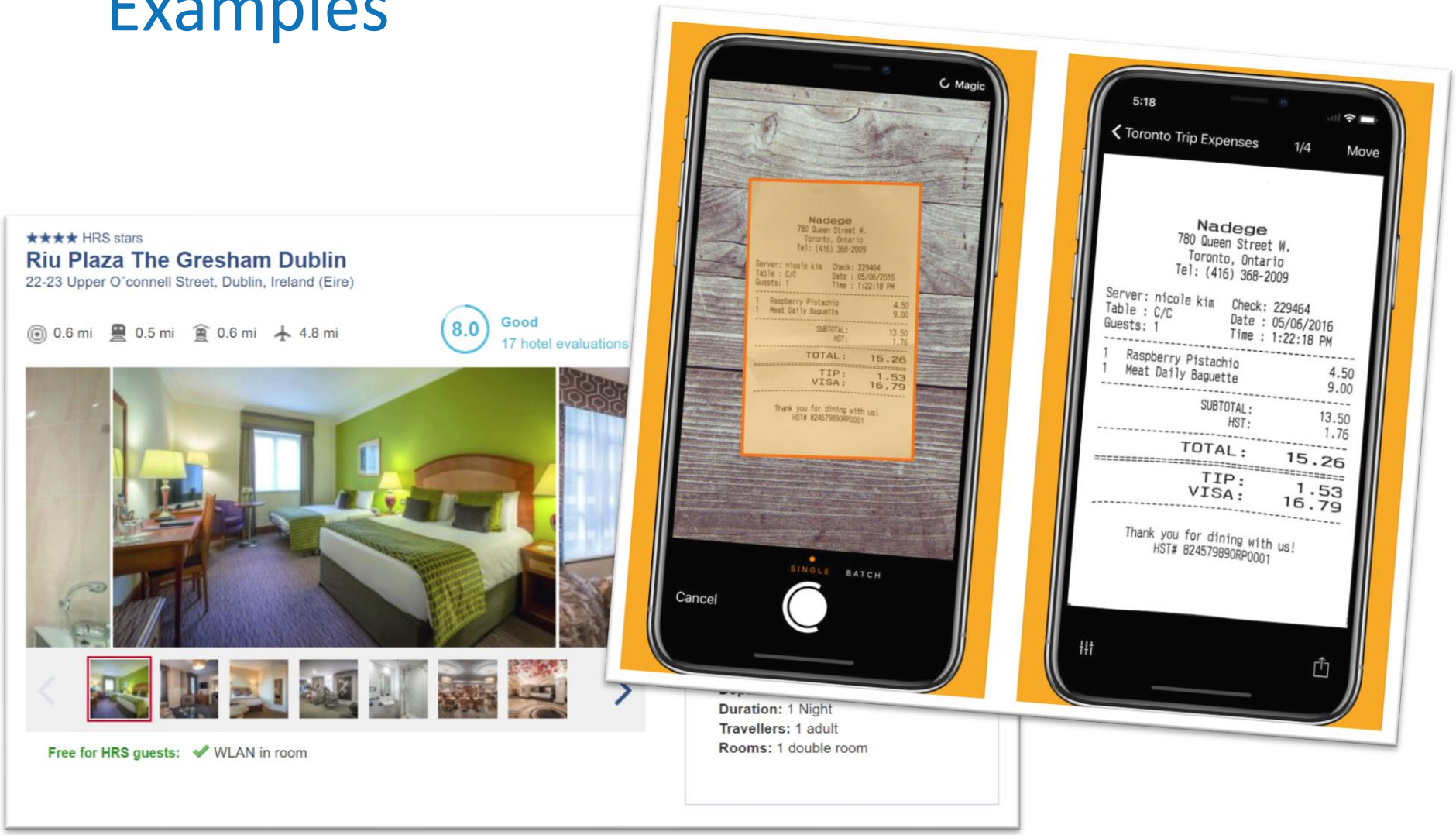
Get Data

Use Own Data

Nothing more to say ☺ (for „sampling“ see previous lecture)

Create Data

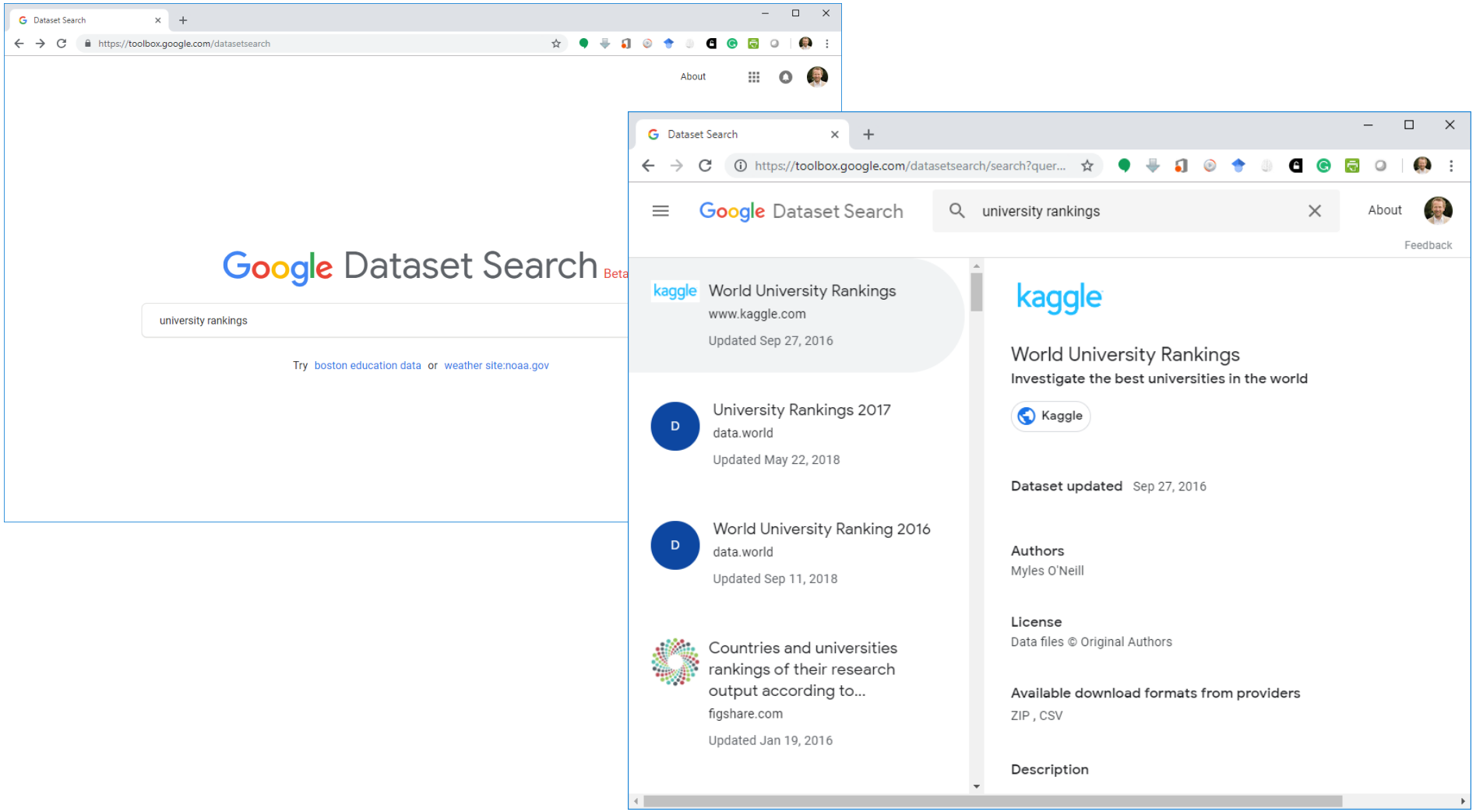
Examples


★★★★ HRS stars
Riu Plaza The Gresham Dublin
22-23 Upper O´connell Street, Dublin, Ireland (Eire)
0.6 mi
0.5 mi
0.6 mi
4.8 mi
8.0
Good
17 hotel evaluations
Free for HRS guests:
WLAN in room
Duration: 1 Night
Travellers: 1 adult
Rooms: 1 double room
Magic
Nadege
780 Queen Street W.
Toronto, Ontario
Tel: (416) 368-2009
Server: nicole kim
Table : C/C
Guests: 1
Check: 229464
Date : 05/06/2016
Time : 1:22:18 PM
1 Raspberry Pistachio 4.50
1 Meat Daily Baguette 9.00
SUBTOTAL: 13.50
HST: 1.76
TOTAL: 15.26
TIP: 1.53
VISA: 16.79
Thank you for dining with us!
HST# 824579890RP0001
SINGLE
BATCH
Cancel
5:18
Toronto Trip Expenses
1/4
Move

Use Third Party Data

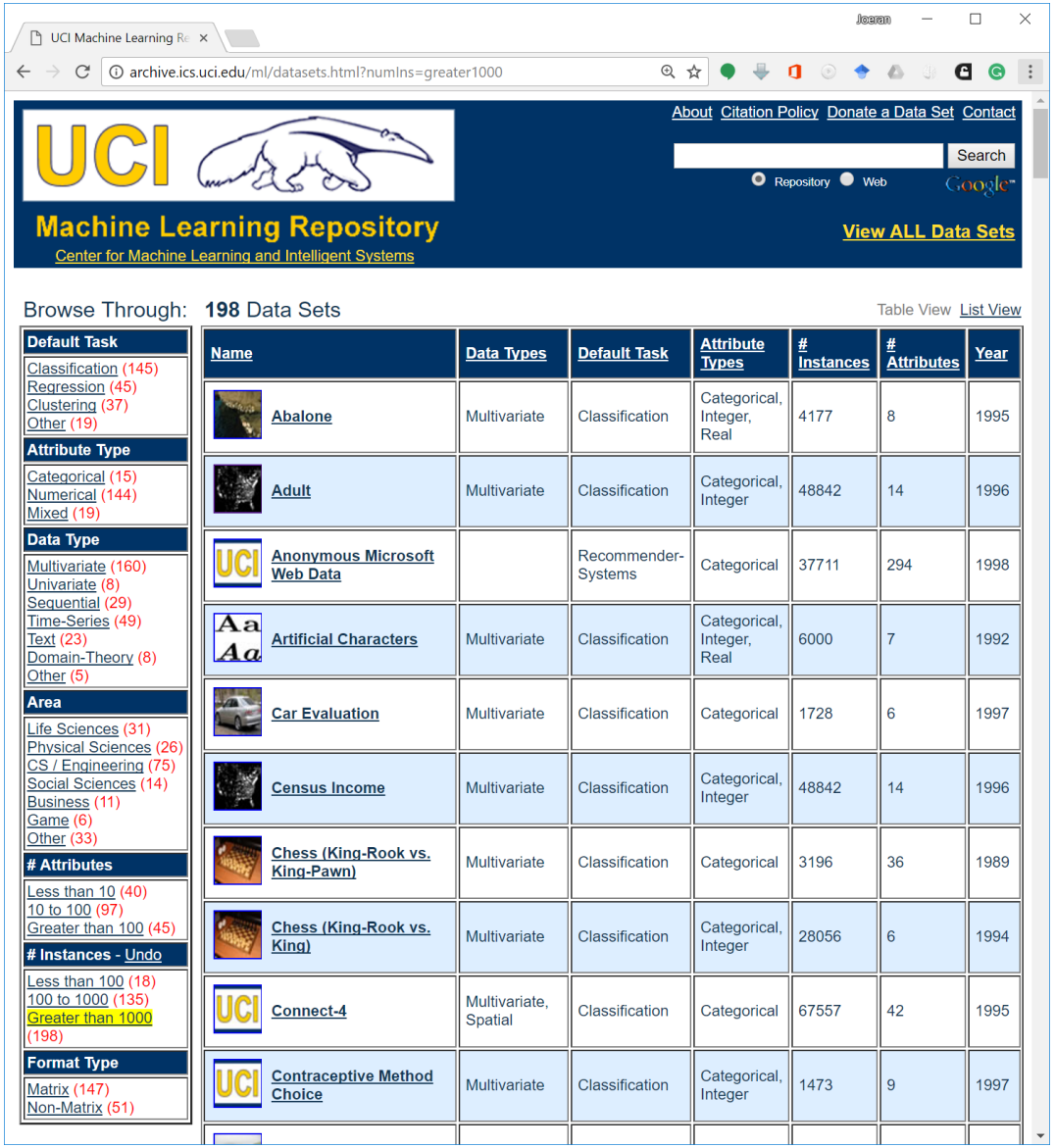
Google Dataset Search

https://toolbox.google.com/datasetsearch

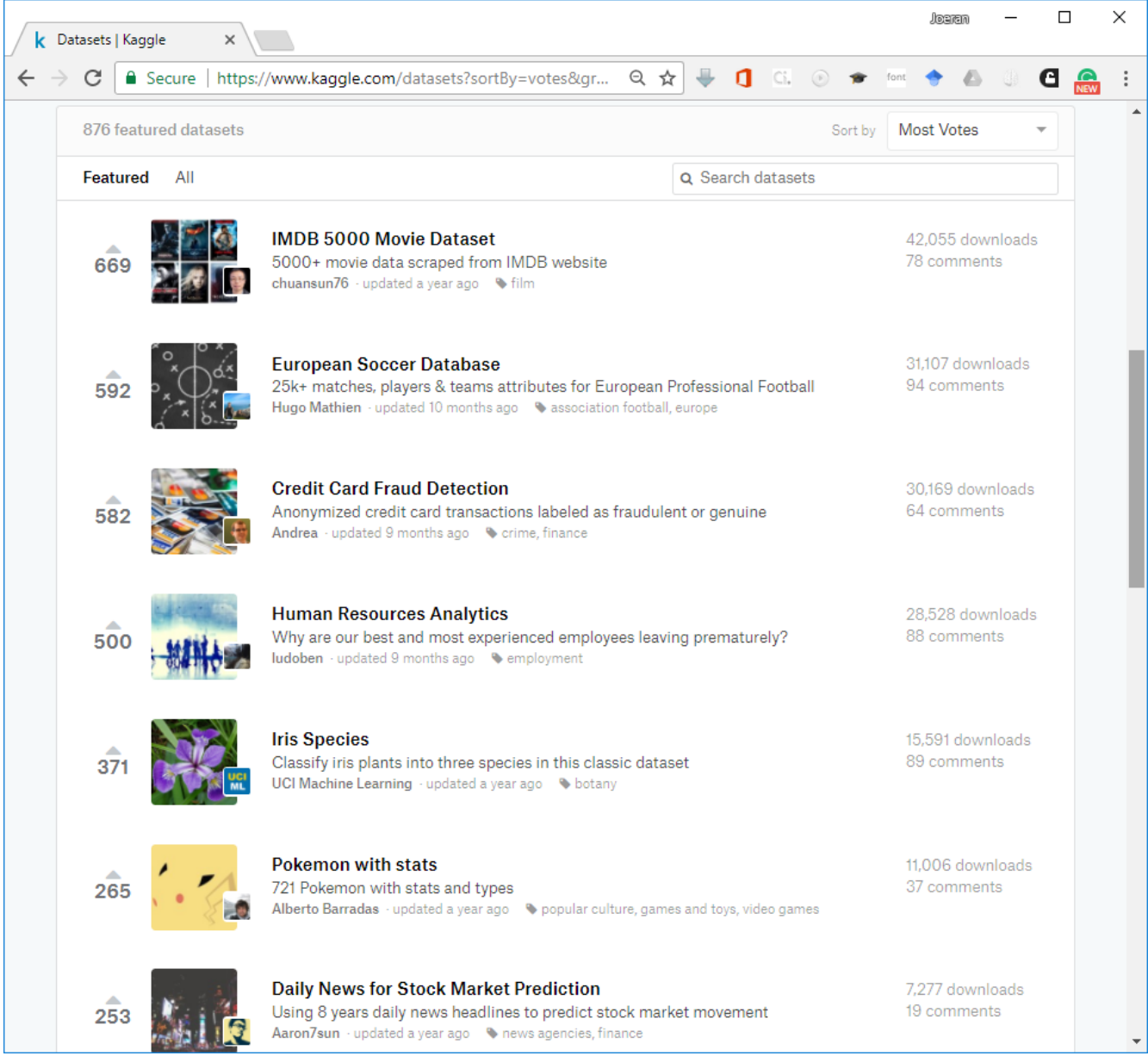


UC Irvine Machine Learning Repository

- **http://archive.ics.uci.edu/ml/datasets.html?numIns=greater1000**
  - Car Evaluation
  - Wine Quality
  - Human Activity Recognition Using Smartphones
  - Forest Fires
- **Has good filtering options**



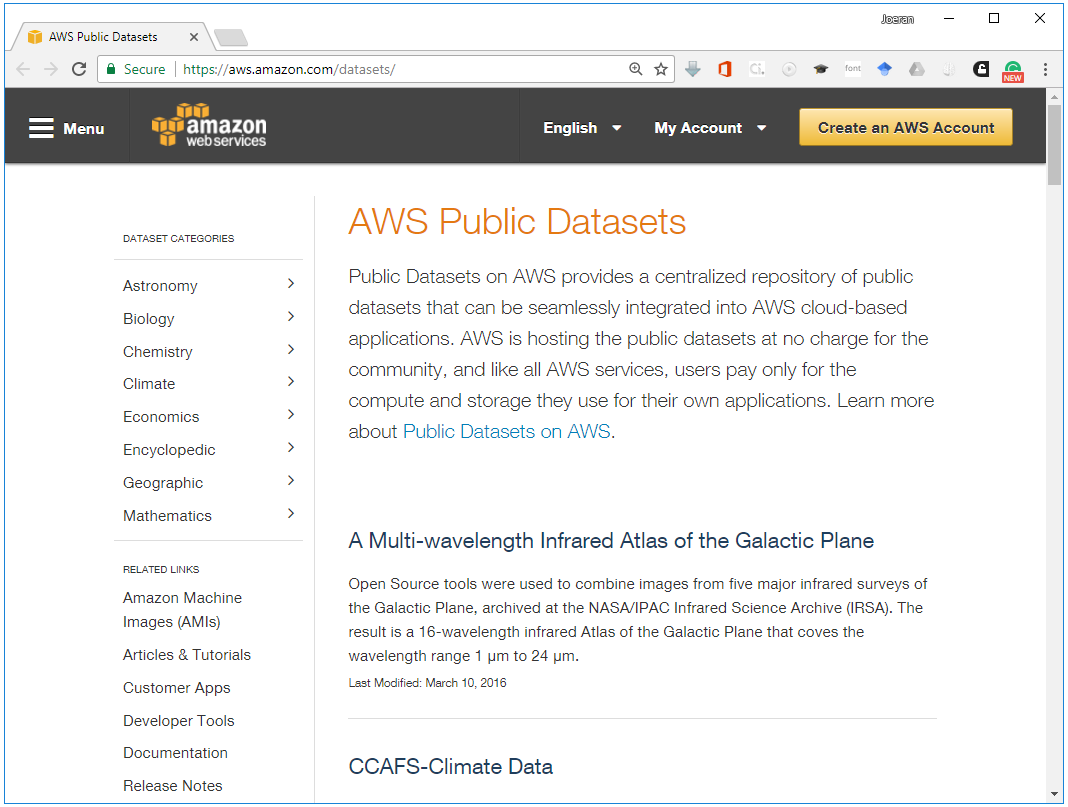
Kaggle



https://www.kaggle.com/datasets

Amazon's AWS Public Datasets

- https://aws.amazon.com/datasets/
- **Enron Email Data**
- **Japan Census Data**
- **Wikipedia Page Traffic Statistic V3**
- **Million Song Dataset**
- **Google Books Ngrams**

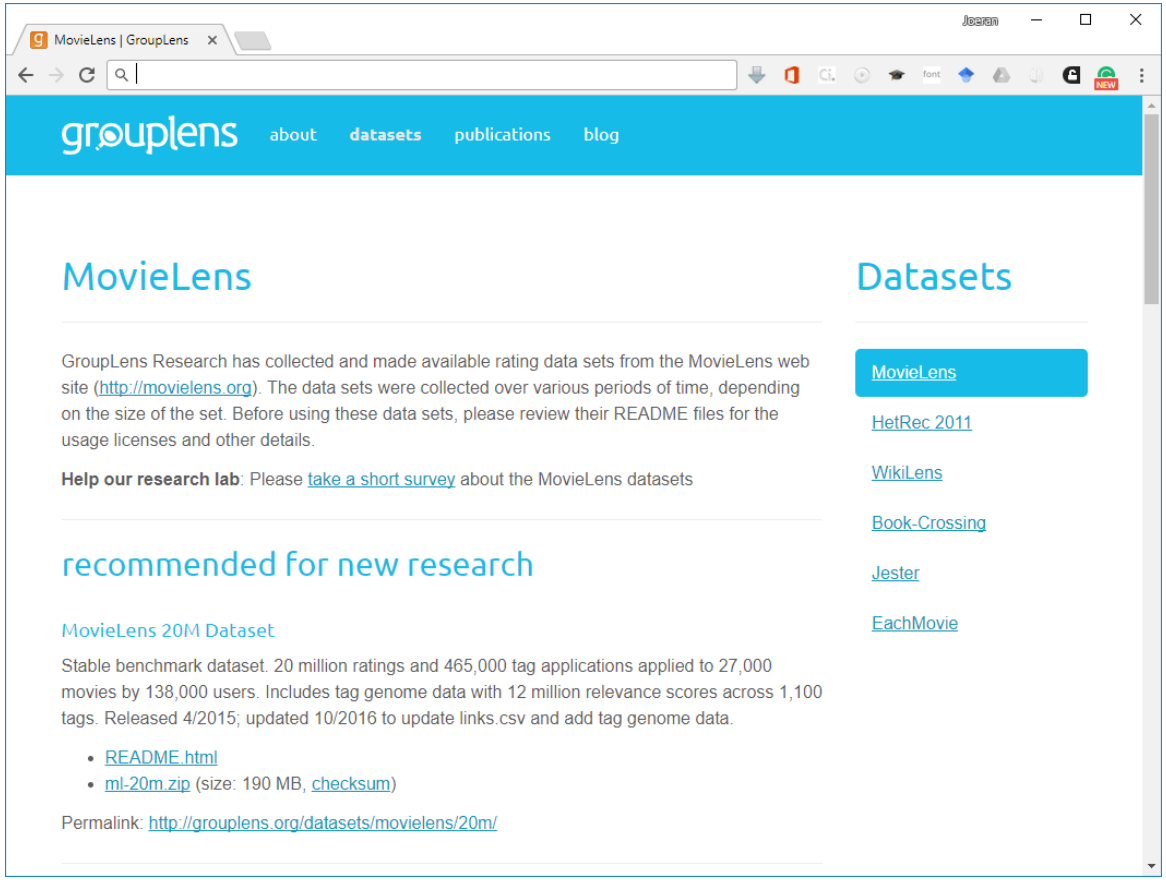


RecSys2013: Yelp Business Rating Prediction

- **https://www.kaggle.com/c/yelp-recsys-2013/data**
- **In the training set:**
  - 11,537 businesses
  - 8,282 checkin sets
  - 43,873 users
  - 229,907 reviews
- **In the testing set:**
  - 1,205 businesses
  - 734 checkin sets
  - 5,105 users
  - 22,956 reviews to predict

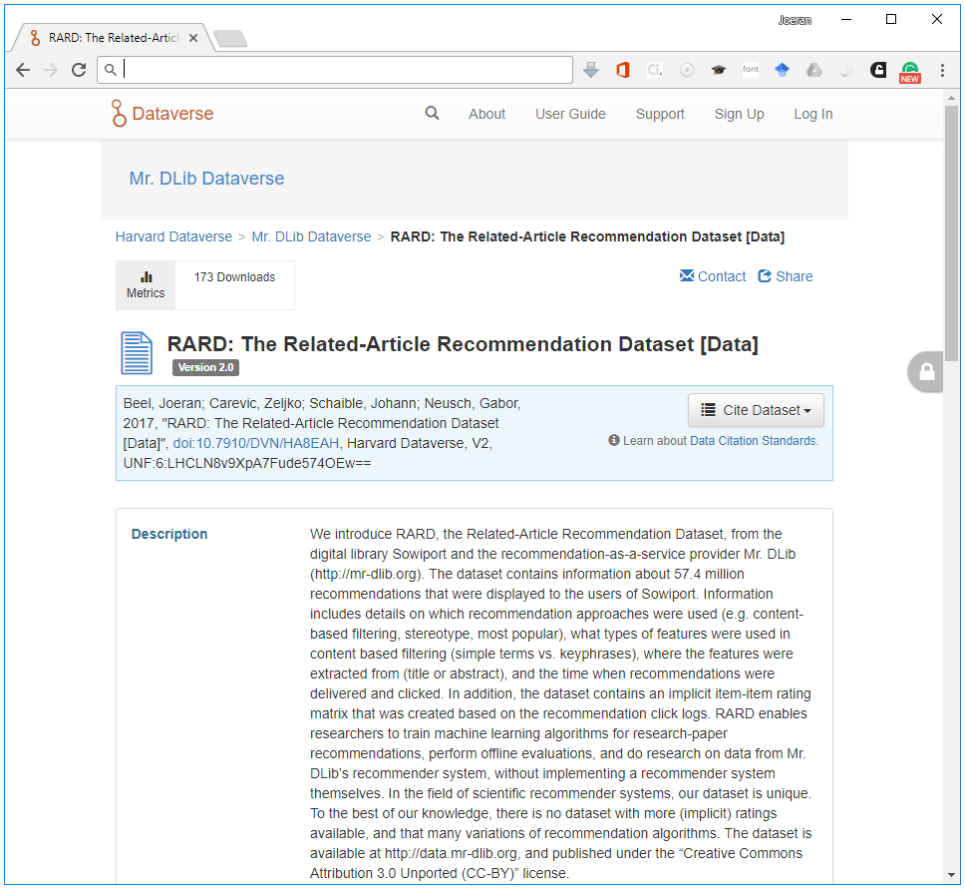
MovieLens Dataset

- https://grouplens.org/datasets/movielens/
- **27 thousand movies**
- **20 million ratings**
- **138 thousand users**
- **…**



RARD – The Related-Article Recommendation Dataset

- **https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HA8EAH**
- **Statistics about 57 million delivered recommendations**
- **2 million documents**
- **77 thousand clicks**



And more…

- FiveThirtyEight https://github.com/fivethirtyeight/data
- Buzz Feed https://github.com/BuzzFeedNews
- Socrata OpenData https://opendata.socrata.com/
- Google Public Data sets https://cloud.google.com/bigquery/public-data/
- Wikipedia https://en.wikipedia.org/wiki/Wikipedia:Database_download
- Data.World https://data.world/
- And even more…
  - https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/
  - https://www.dataquest.io/blog/free-datasets-for-projects/
  - http://dataportals.org/
  - https://www.quandl.com/
  - http://opendatamonitor.eu/

How much data to we need?

Sometimes, little data is sufficient

- **19 features**
- **250-3333 data points**

Features									Target
Cust. ID	State	Acct length	Area code	Int'l plan	Voicemail plan	Total messages	Total mins.	Total calls	Churned?
502	FL	124	561	No	Yes	28	251.4	104	False
1007	OR	48	503	No	No	0	190.4	92	False
1789	WI	63	608	No	Yes	34	152.2	119	False
2568	KY	58	606	No	No	0	247.2	116	True

Figure 2.2 Training data with four instances for the telecom churn problem

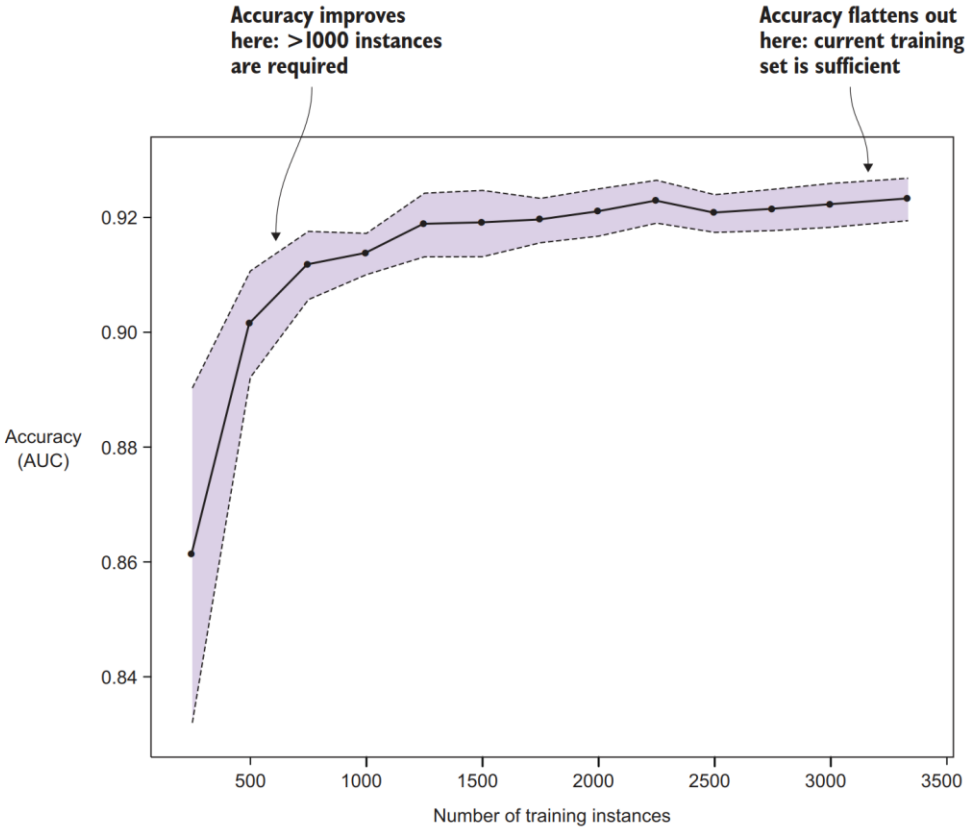
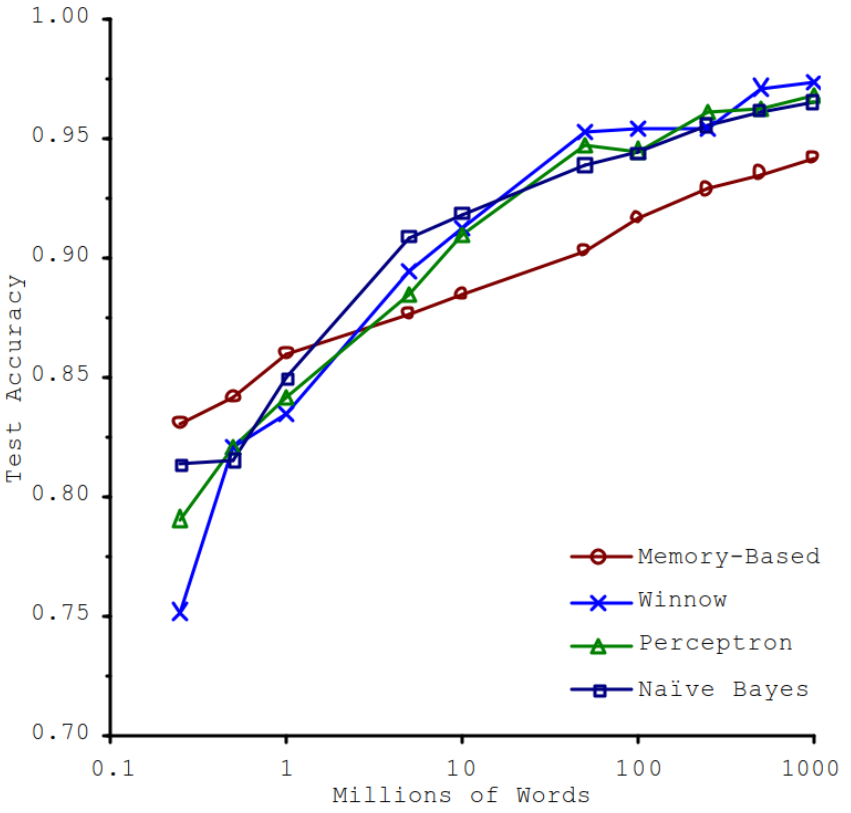


Figure 2.3 Testing whether the existing sample of 3,333 training instances is enough data to build an accurate telecom churn ML model. The black line represents the average accuracy over 10 repetitions of the assessment routine, and the shaded bands represent the error bands.

Henrik Brink, Joseph Richards, and Mark Fetherolf, *Real-world machine learning* (Manning Publications Co., 2016).

Sometimes, lots of data is necessary

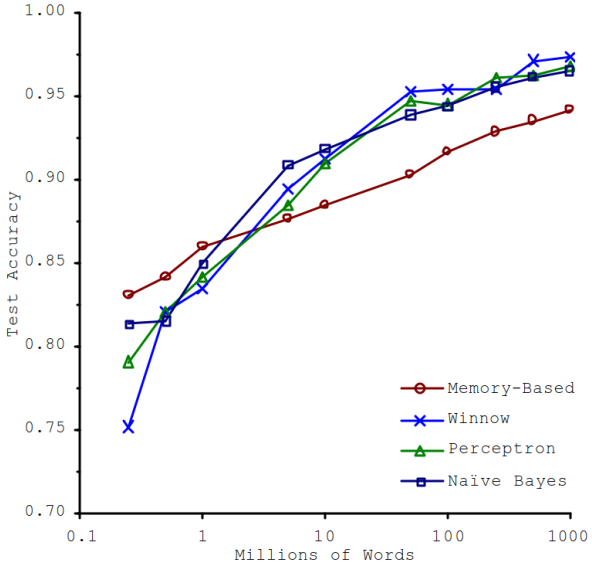
- **For medium-complex problems, (tens of) thousands of instances are needed (e.g. recognizing a title in a document)**
- **For complex problems, millions of instances are needed (e.g. understanding hand-writing)**
- **Different machine learning algorithms perform alike (for Natural Language Disambiguation), once the training data set is big enough**
- **"These results suggest that we may want to reconsider the trade-off between spending time and money on algorithm development versus spending it on corpus development. "**
  M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in Proceedings of the 39th annual meeting on association for computational linguistics, 2001, pp. 26–33.
- **See also A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," IEEE Intelligent Systems, vol. 24, no. 2, pp. 8–12, 2009.**

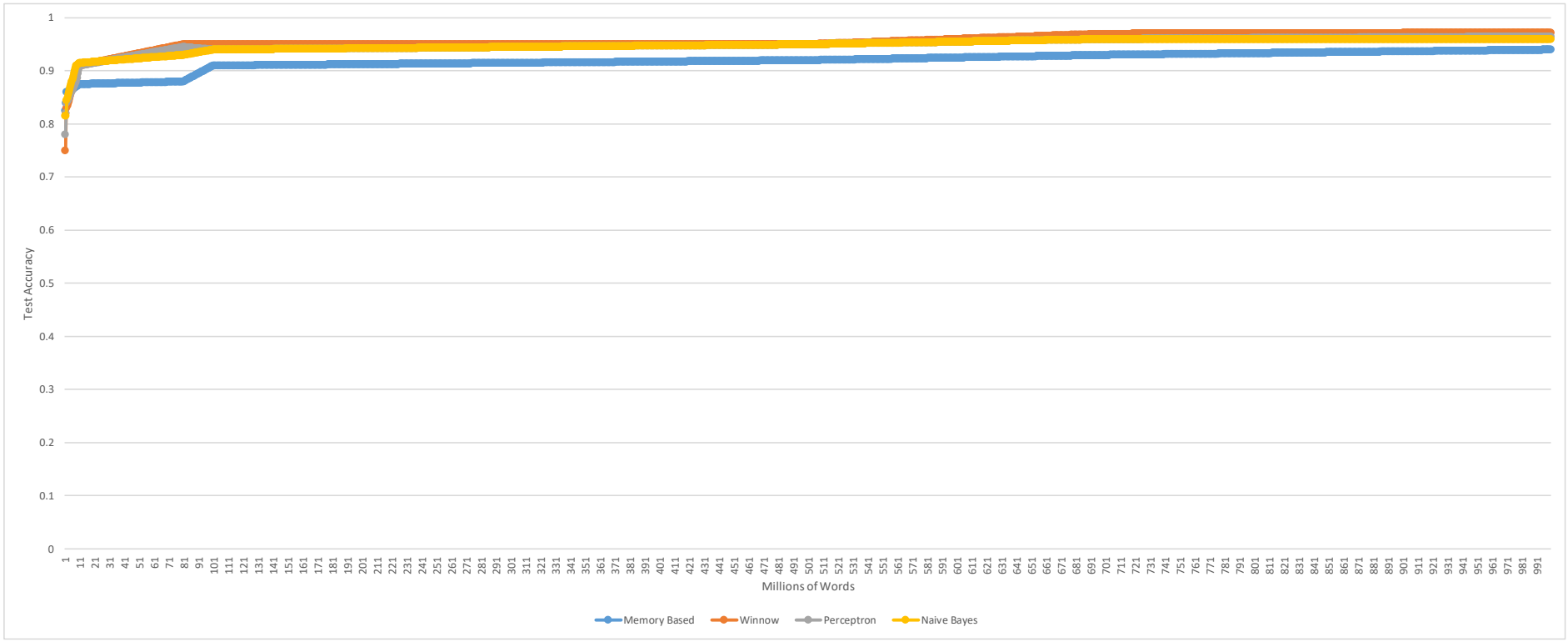


Michele Banko and Eric Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th annual meeting on association for computational linguistics* (Association for Computational Linguistics, 2001), 26–33.

And again

Proper x and y axis





What can go wrong, even if there is lots of data?

Low quality data

- **Noise**
- **Missing values**
- **Redundant/Irrelevant Features**
- **…**

https://www.incimages.com/uploaded_files/image/970x450/rotten-apple-1725x810_12112.jpg

Noise Example

CBC news | Technology & Science

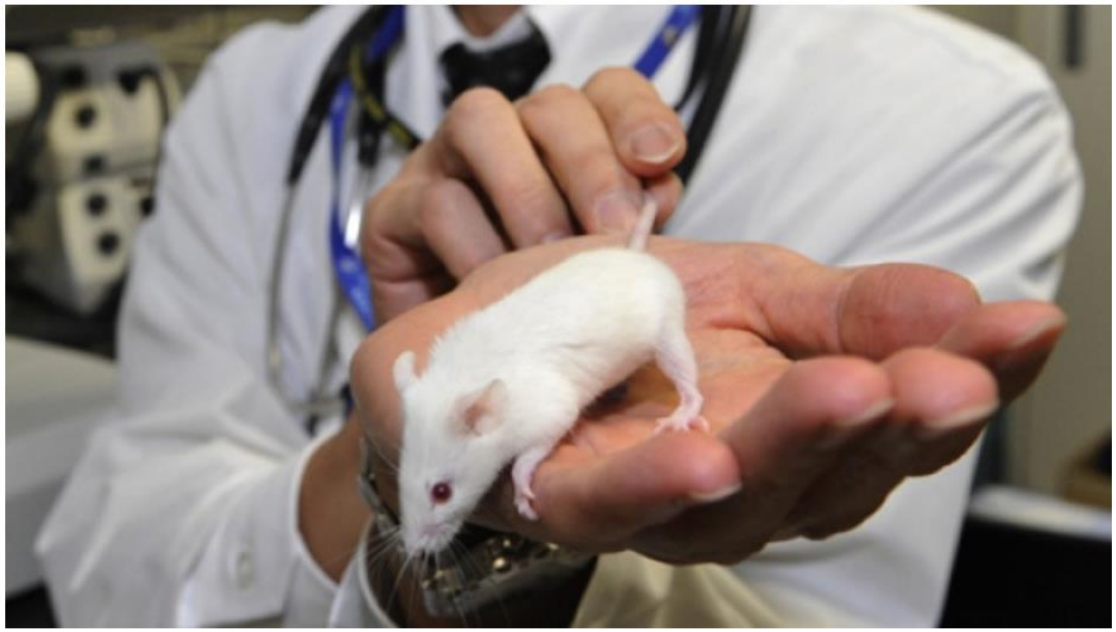
Home | Opinion | World | Canada | Politics | Business | Health | Entertainment | Technology & Science | Video

Technology & Science | Quirks & Quarks Blog | Spark | Photo Galleries

Lab mouse test results depend on scientist's gender

The smell of a man makes mice more stressed than the smell of a woman

The Canadian Press Posted: Apr 29, 2014 10:59 AM ET | Last Updated: Apr 29, 2014 1:50 PM ET

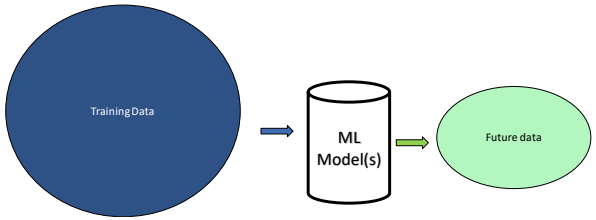


McGill University researchers showed that lab rodents become stressed in the presence of male researchers. (Yoshikazu Tsuno/AFP)

http://www.cbc.ca/news/technology/lab-mouse-test-results-depend-on-scientist-s-gender-1.2625550

Completely Bad (Unrepresentative) Data

- **The training data is not representative for the real population**
- **Conclusions („learnings“) drawn from the data are meaningless for generalizing about new instances**



Potential Reasons

- Too small sample (either for training or testing)
- Sampling errors
- Artificial data
- Third-party data
- Outdated data (too old data)
  - Population/instances changes over time (e.g. different types of customers with different habits/needs, different styles, different gender distributions, …)
  - Features change over time (e.g. more expensive items, different duration of movies, better quality of images)

Getting-Data Checklist

1. **List the data you need and how much you need**
2. **Find and document where you can get the data**
3. **Check how much space it will take**
4. **Check legal obligations and get authorization if necessary**
5. **Get access authorization**
6. **Create a workspace (with enough storage space)**
7. **Get the data**
8. **Convert the data to a format you can easily manipulate (without changing the data itself)**
9. **Ensure sensitive information is deleted or protected**
10. **Check the size and type of data (time series, sample, geographical, …)**
11. **Sample a test set, put it aside, and never look at it (no data snooping)**

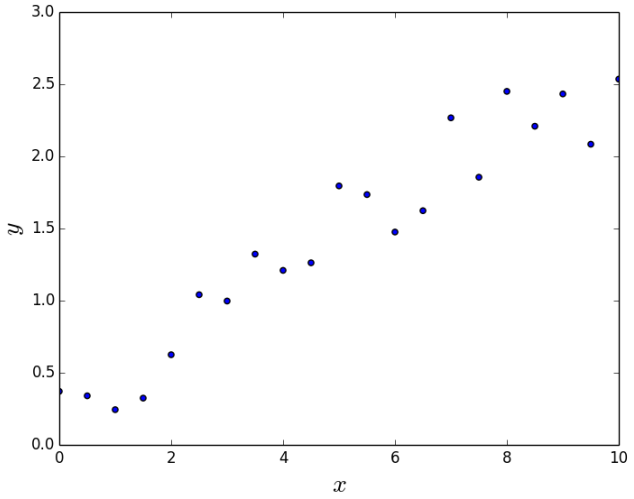
Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

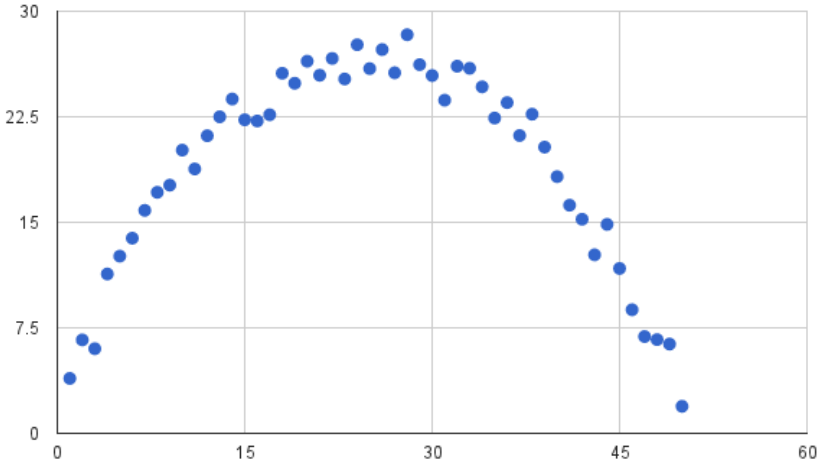
Understand / Visualize / Report Data

Purposes

- **Understand what data you have**
- **How the data quality is (noise, outliers, missing data)**
- **Understand what ML models might be appropriate**
- **Get ideas what features to use for the learning**



http://image.diku.dk/shark/sphinx_pages/build/html/_images/linearRegressionData.png



http://3.bp.blogspot.com/-wg7eovEn0GA/VEXW-DPhGuI/AAAAAAAAD4Y/hDWD4TZKRWU/s1600/quadratic_data.png

Scatter Plots

- **Shows relationship (correlation) between two variables (independent-dependent or independent-independent)**

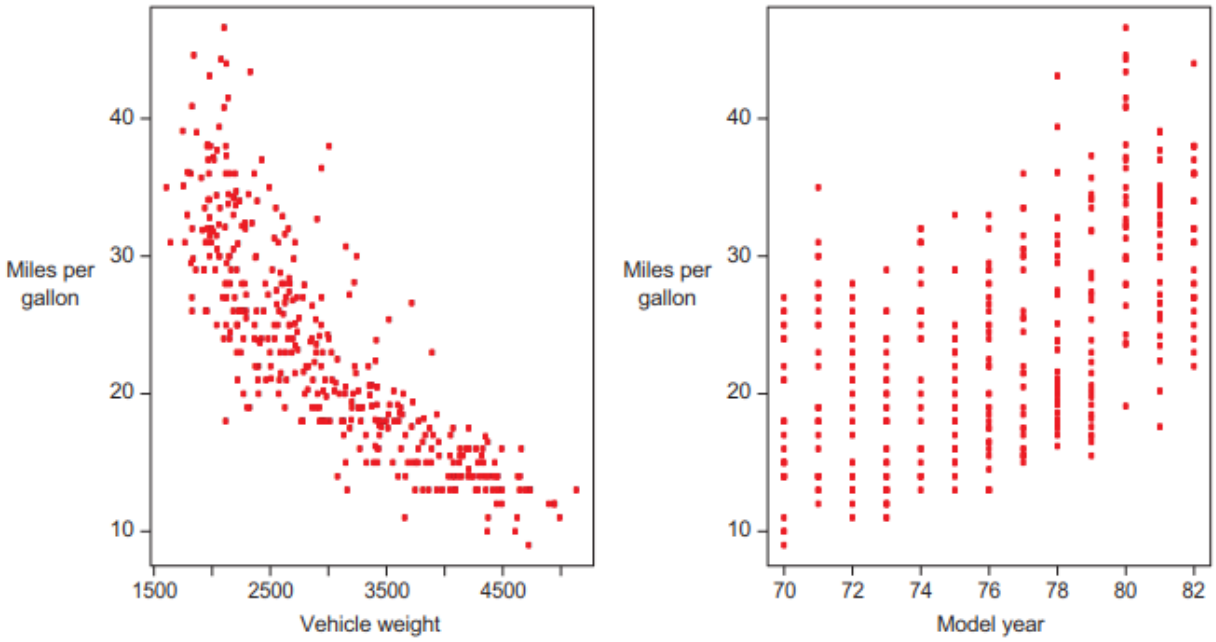
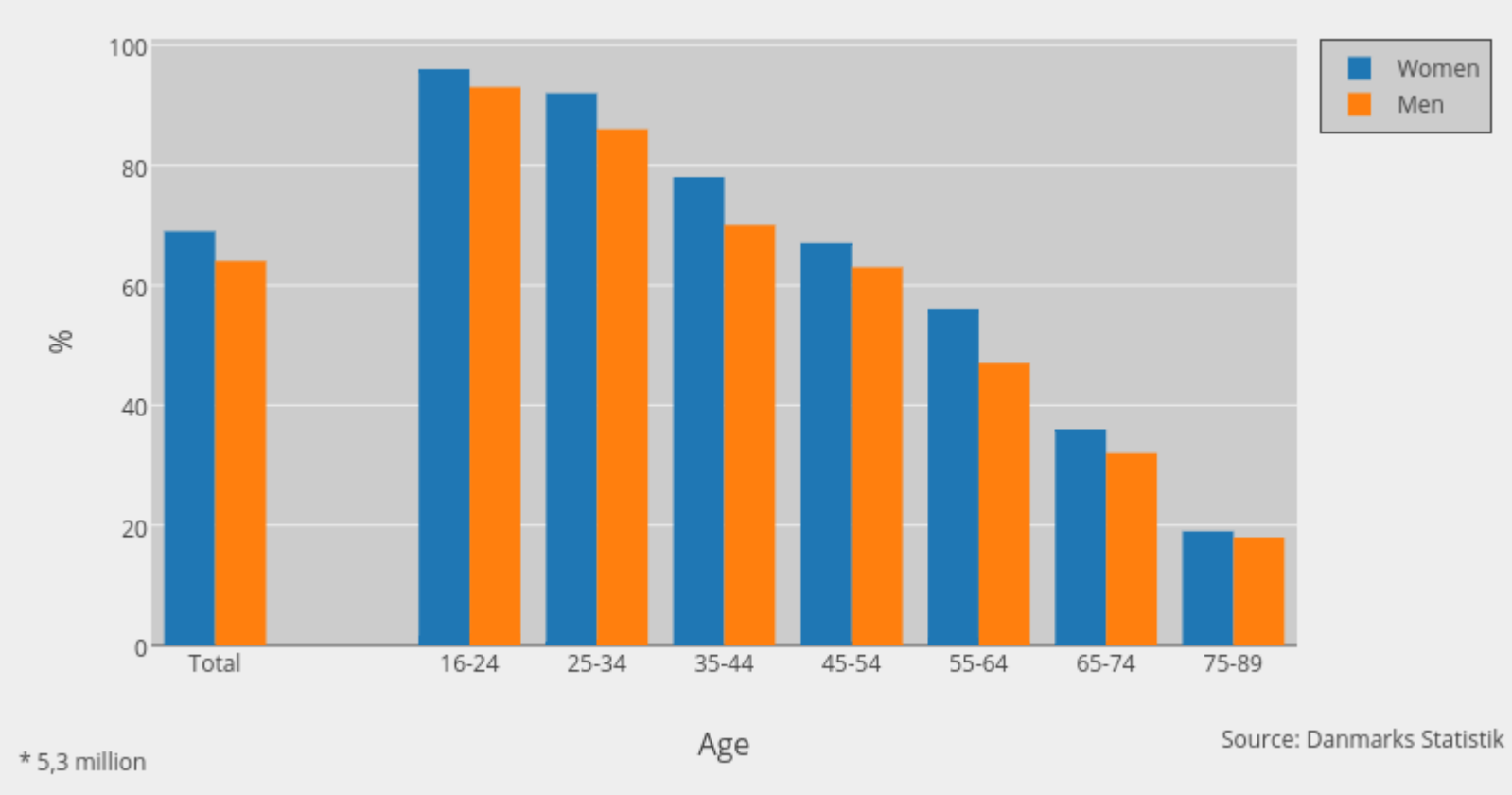


Figure 2.17 Scatter plots for the relationship of vehicle miles per gallon versus vehicle weight (left) and vehicle model year (right)

Henrik Brink, Joseph Richards, and Mark Fetherolf, Real-world machine learning (Manning Publications Co., 2016).

Bar charts

- **See distributions of instances and features (e.g. missing values; imbalanced datasets).**



https://plot.ly/~Lufferia1991/120/age-distribution-on-social-media-usage-for-internet-users-in-denmark-2014.png

Mosaic Plots

- **Can be used to identify important variables (in this case gender)**

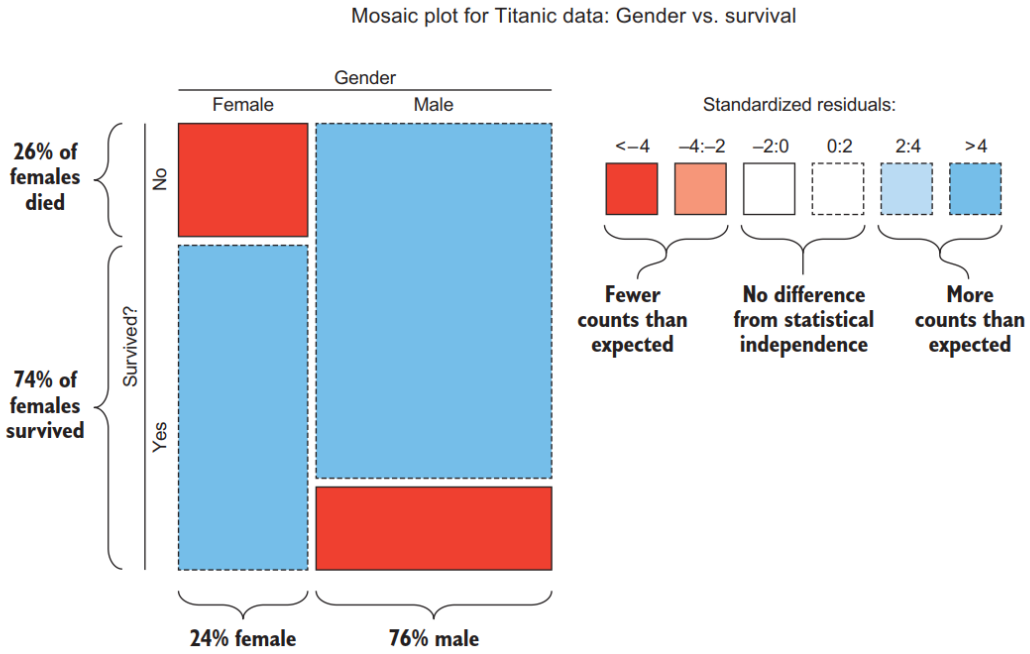


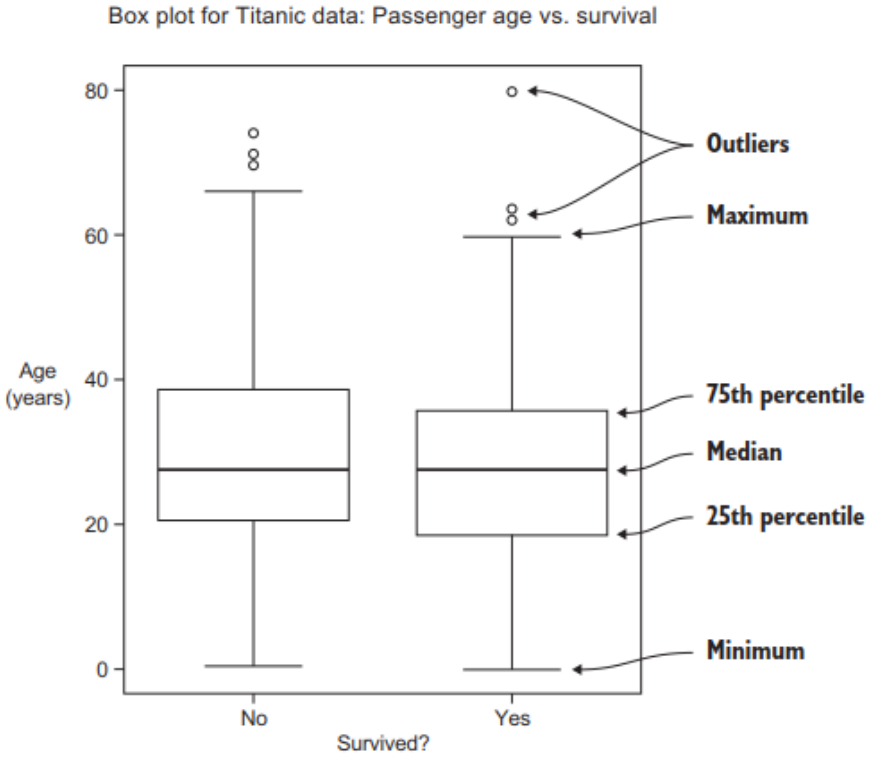
Figure 2.12 Mosaic plot showing the relationship between gender and survival on the Titanic. The visualization shows that a much higher proportion of females (and much smaller proportion of males) survived than would have been expected if survival were independent of gender. "Women and children first."

Confidence Level	z
99%	2.58
98%	2.33
95%	1.96
90%	1.65

Henrik Brink, Joseph Richards, and Mark Fetherolf, Real-world machine learning (Manning Publications Co., 2016).

Box plots

- **Displays distribution of values for a variable**
- **Example indicates that age had no impact on survival**



Henrik Brink, Joseph Richards, and Mark Fetherolf, Real-world machine learning (Manning Publications Co., 2016).

Decision Help

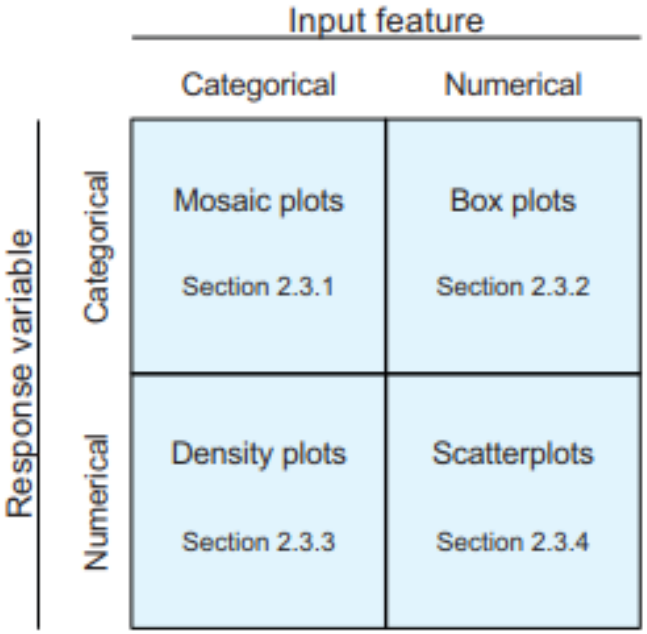
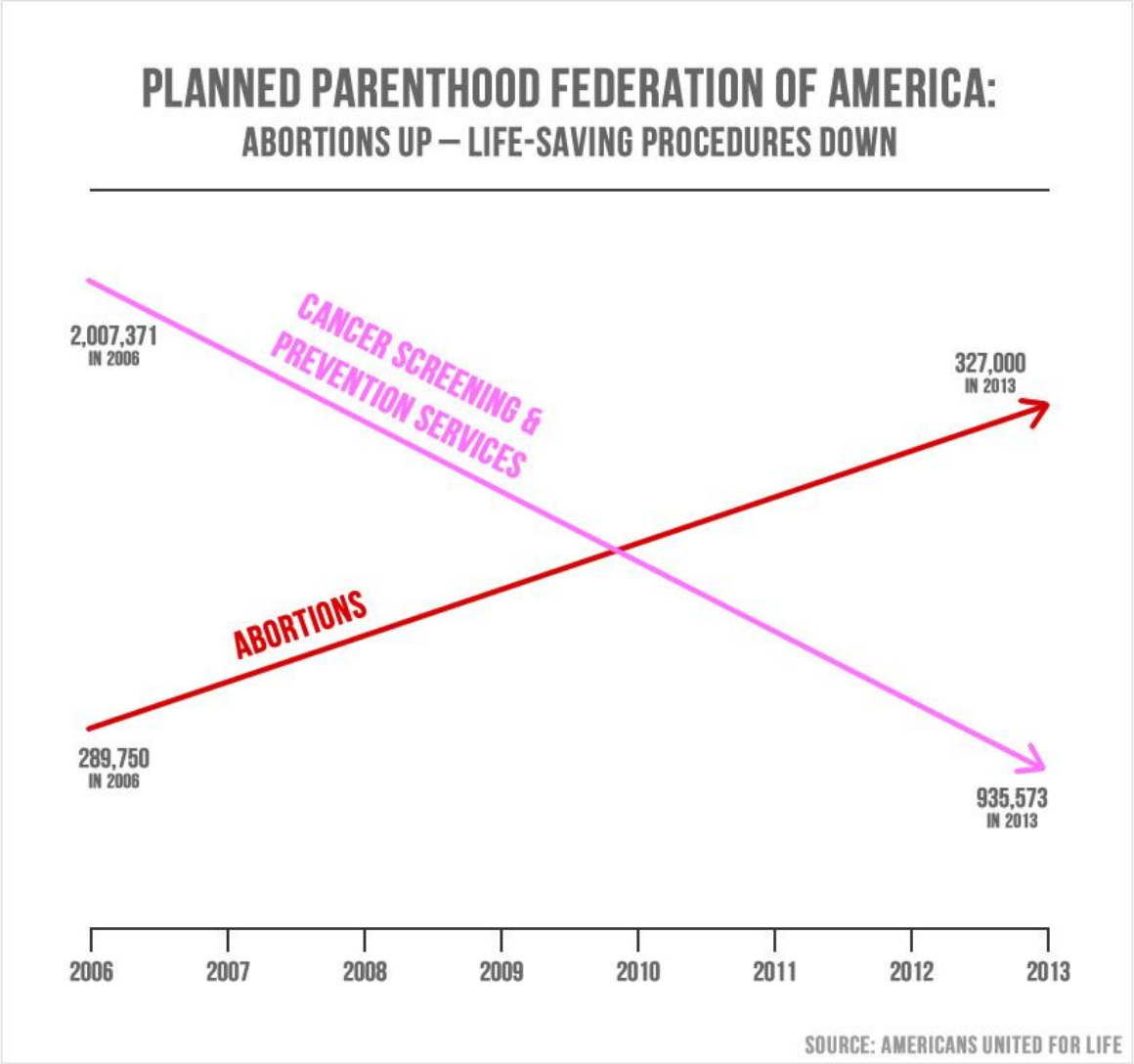


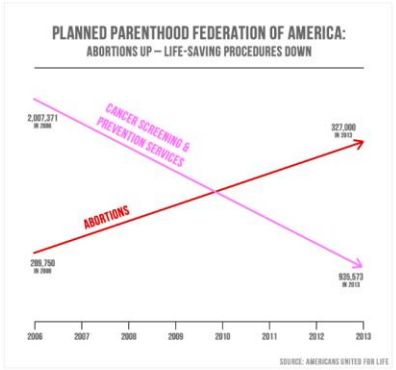
Figure 2.11 Four visualization techniques, arranged by the type of input feature and response variable to be plotted

Henrik Brink, Joseph Richards, and Mark Fetherolf, Real-world machine learning (Manning Publications Co., 2016).

Be critical and careful!

Example

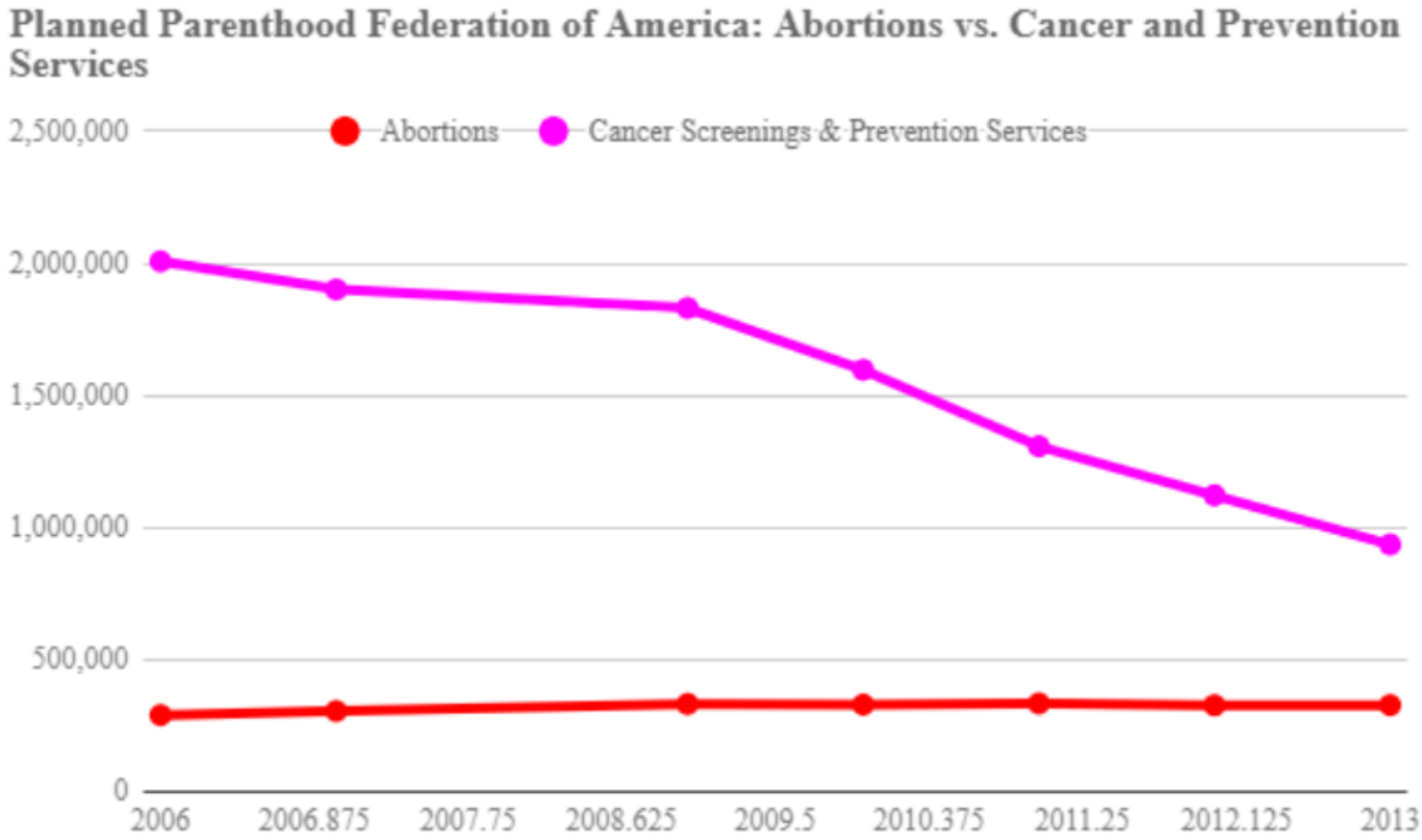




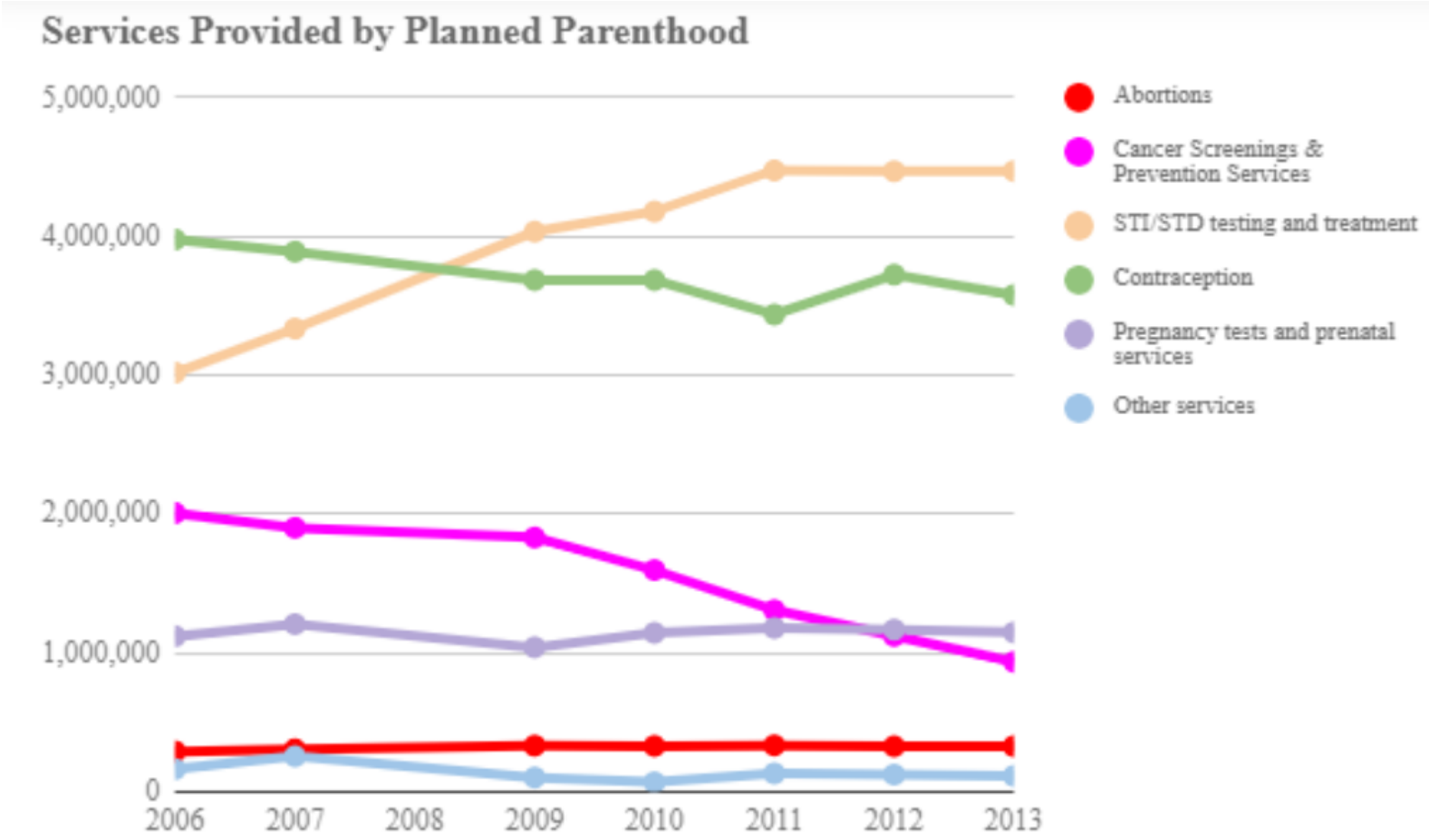
http://www.aul.org/blog/aul-releases-the-new-leviathan-the-mega-centers-report-how-planned-parenthood-has-become-abortion-inc/

https://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/

Corrected Scale



Cherry Picking

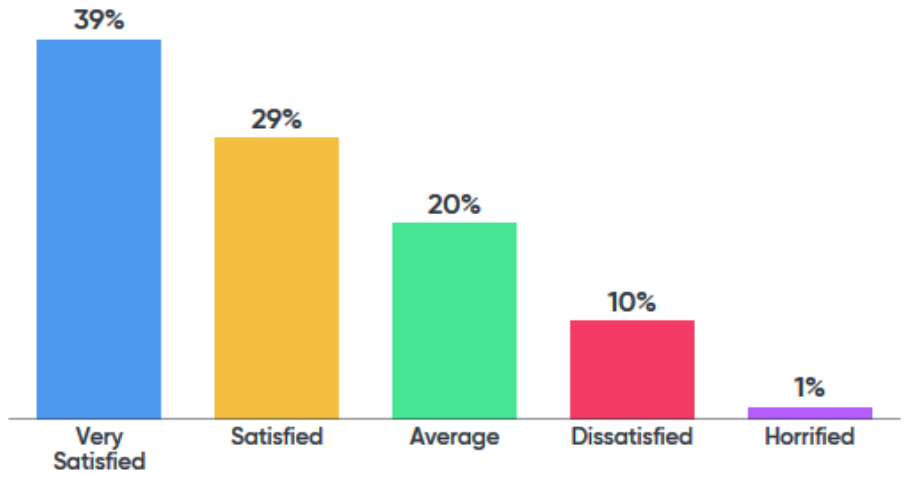


Any questions from previous lecture?

Go to **www.menti.com** and use the code **24 94 8**

How satisfied were you with your group members doing the assignment?

Mentimeter



Slide is not active **Activate**

Deadlines

Rules

- **Plagiarism: If we catch you, you fail the entire module (or worse). We run plagiarism checks with tools like Turnitin; Moss; …**
- **Deadlines are deadlines. Even 1 minute late is too late. Only exception: serious illness (medical certificate required).**

Schedule / Deliverables (2018-09-22)

Subject to changes

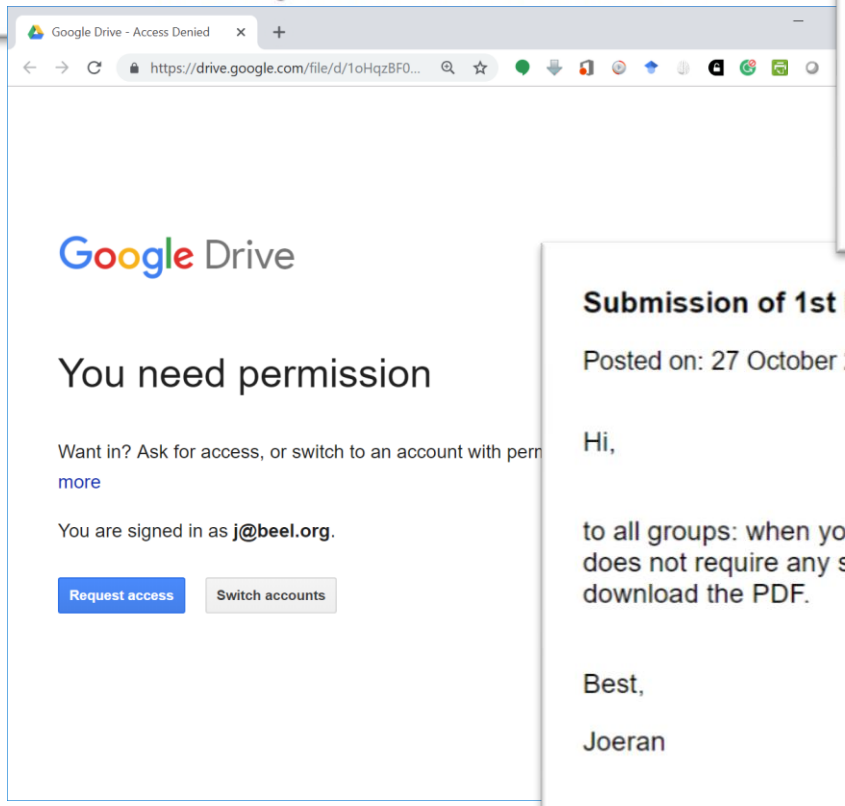
- **All deadlines are 8:00 o´clock in the morning (Irish time), if not stated otherwise**

Penalties


CS4404 Tean eport submission
Inbox ×
Mon, Oct 29, 10:11 AM (23 hours ago)
Reply to all
to joeran.beel,
Hi,
Mon, Oct 29, 6:43 PM (14 hours ago)
Reply to all
to joeran.beel,
Hi,
My apologies, I accidentally uploaded the wron
Here is the link for the right one.
Mon, Oct 29, 10:16 PM (11 hours ago)
to joeran.beel,
Hello,
Apologies yet again, the previous link was missing the cover sheet!
This should (hopefully) be our final link to the file.
Kind regards once more,
Mon, Oc
to joeran.beel,
Hi,
apologies once again.
This is the actual link to the file.
Kind regards,

Submission

- Store your report as PDF
- Name the PDF file "ML1819--task-$taskID-*team-05.pdf*.
- Send a gDrive/Dropbox/Onedrive/Web/... Joeran.beel@scss.tcd.ie. Do not attach the
- The subject of the email should be



Research Assignment 1: Missed Deadlines, Incorrect URLs

Posted on: 29 October 2018 09:25:34 o'clock GMT

Hello everyone,

the deadline for research assignment 1 was today at 8:00 am. You find a list of all submissions that I have received here: https://1drv.ms/u/s!AtfAgPR4VDcEqJM-ZZIVQx7ZNGNYcw . Please double check that your submission is listed. If not:

1. A few of you have sent me links to their reports that require a log-in (TCD) or an additional request for access (GDrive). I have explicitly mentioned that you need to submit a link that directly allows the download of the PDF. For every team who has not yet done so: You have another 24 hours (until 30th October, 8:00am) to send me a link that allows me directly to download your PDF. Otherwise, your assignment will be marked with 0 marks.
2. It could be that I simply missed your email (especially if the subject was not correct). If you submitted your report before the deadline with a directly downloadable link, but your report does not appear n the list above, please send me another email.

Somewhat surprisingly, I have only received around 30 submissions although there are 51 teams. I would be interested in hearing from the teams who have not submitted why they did not submit.

Best,

Joeran

Submission of 1st Research Assignment: URLs

Posted on: 27 October 2018 12:36:04 o'clock BST

Hi,

to all groups: when you send me the link to your report via email, please send me a link that does not require any sign-up, log-in, requesting permissions etc. I want to click the link and download the PDF.

Best,

Joeran

Again: Mosaic Plots

- **Can be used to identify important variables (in this case gender)**

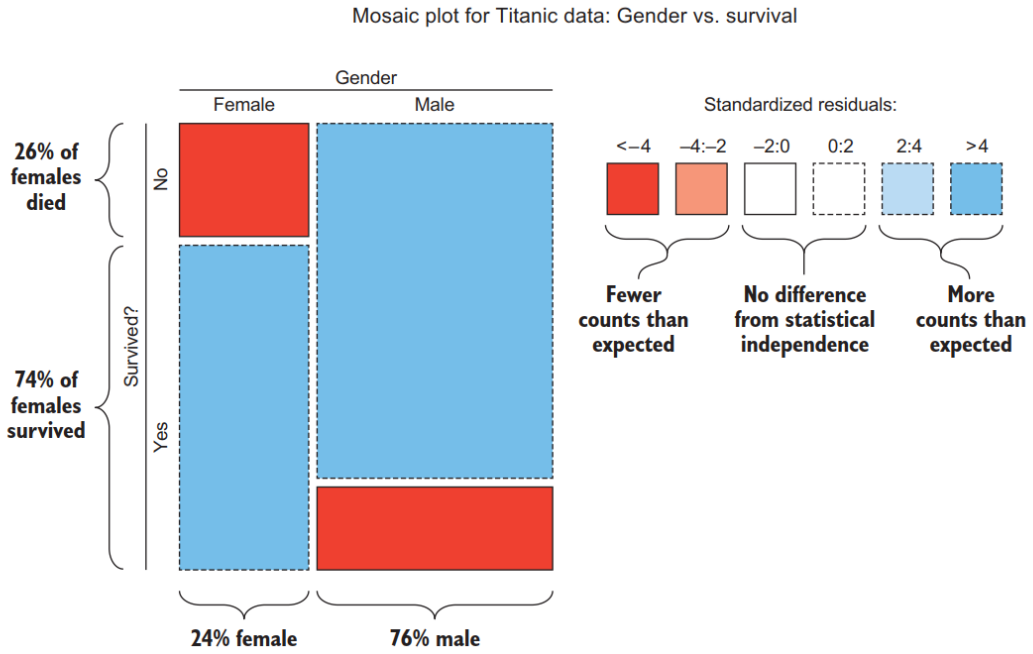
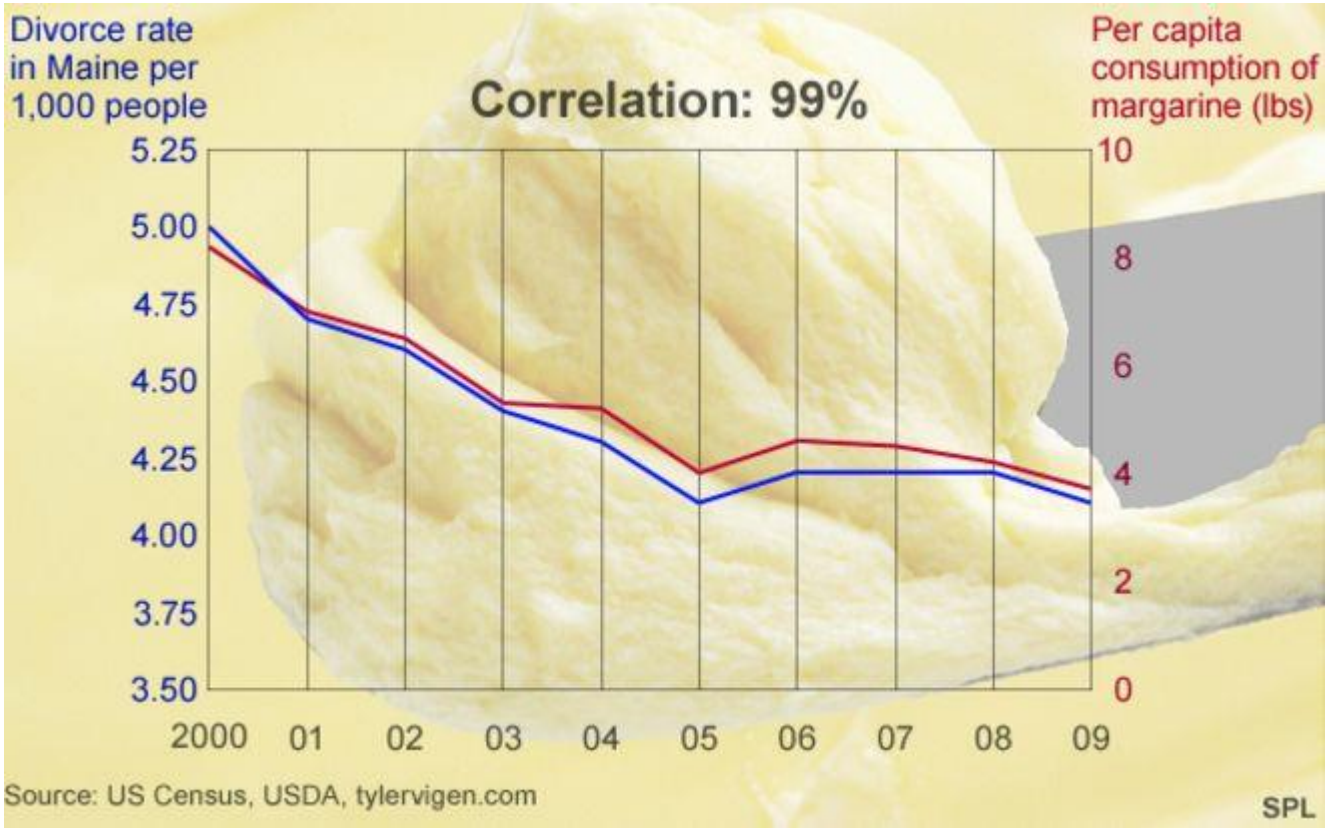


Figure 2.12 Mosaic plot showing the relationship between gender and survival on the Titanic. The visualization shows that a much higher proportion of females (and much smaller proportion of males) survived than would have been expected if survival were independent of gender. "Women and children first."

Confidence Level	z
99%	2.58
98%	2.33
95%	1.96
90%	1.65

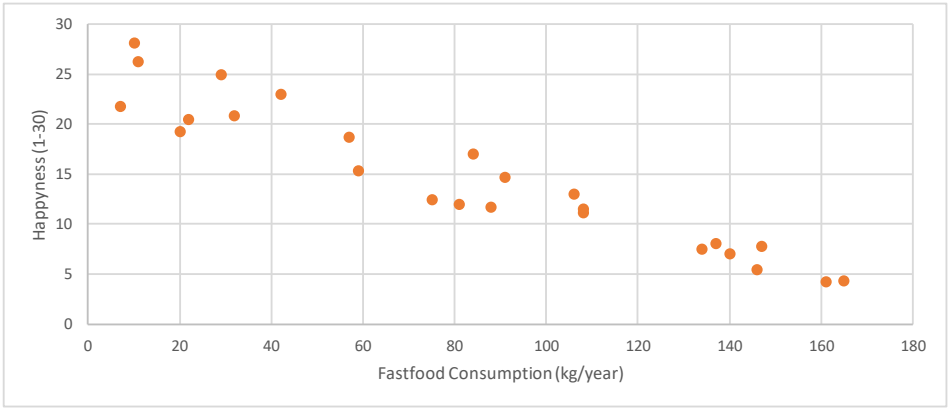
Henrik Brink, Joseph Richards, and Mark Fetherolf, Real-world machine learning (Manning Publications Co., 2016).

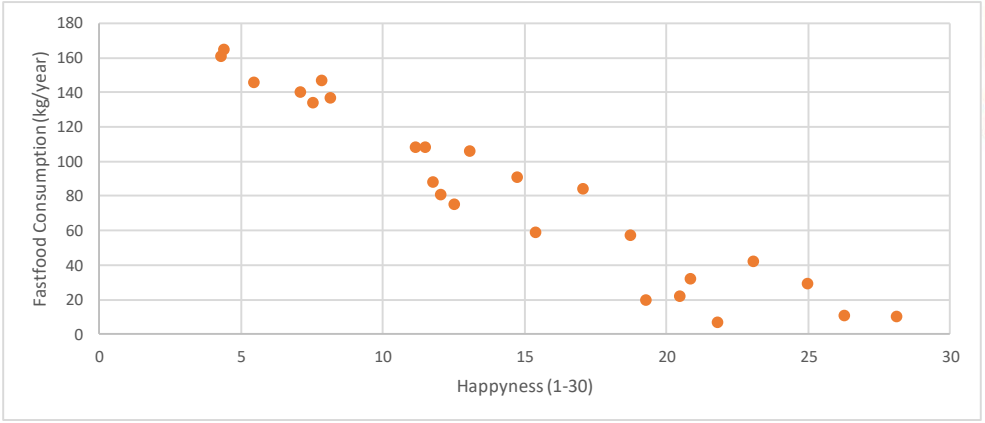
Correlation vs. Causality



http://www.bbc.com/news/magazine-27537142

Dependent vs. Independent Variables





Numbers are hypothetical to illustrate the issue

http://www.independent.co.uk/life-style/health-and-families/fast-food-can-make-you-depressed-and-unable-to-control-your-emotions-new-study-suggests-10339017.html

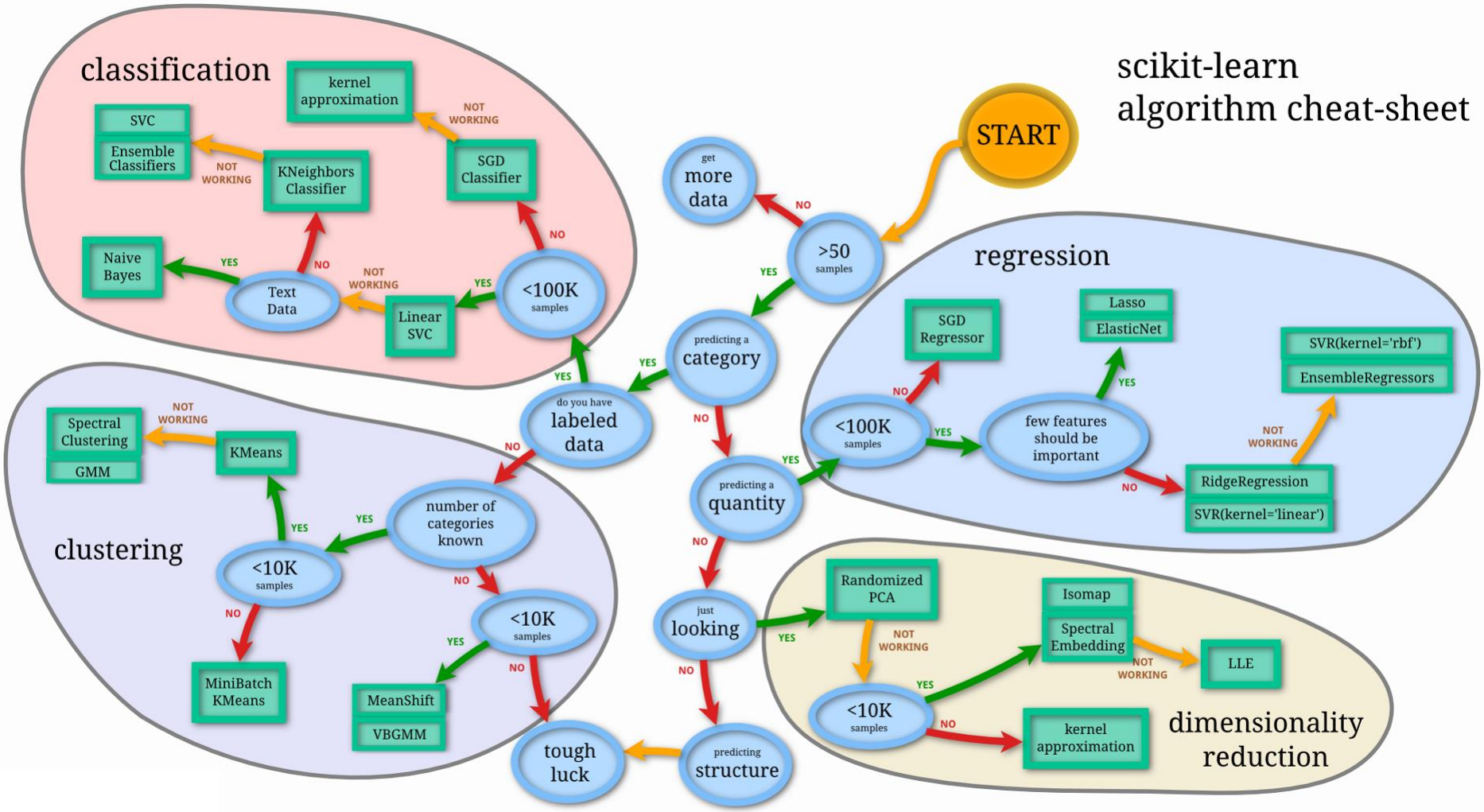
Explore the Data

1. **Create a copy of the data**
2. **Sample the data down to a manageable size if necessary**
3. **Study each attribute and its characteristics**
   - Name
   - Type (categorical, int, text, structured, …)
   - % of missing values
   - Noisiness and type of noise
   - Type of distribution
4. **For supervised learning, identify the target attributes**
5. **Visualize the data**
6. **Study correlation between attributed**
7. **Study how you would solve the problem manually**
8. **Identify the promising transformations you may want to apply**
9. **Identify extra data that would be useful**
10. **Document what you have learned**
11. **Consider how much data the system can handle**
    - Training easily can take days, weeks, or even longer
    - Amazon AWS largest instance (not necessarily Machine Learning):
      - 4 TB Memory
      - ~ 30,000€ per hour

Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).
https://techcrunch.com/2017/09/14/aws-now-offers-a-virtual-machine-with-over-4tb-of-memory/

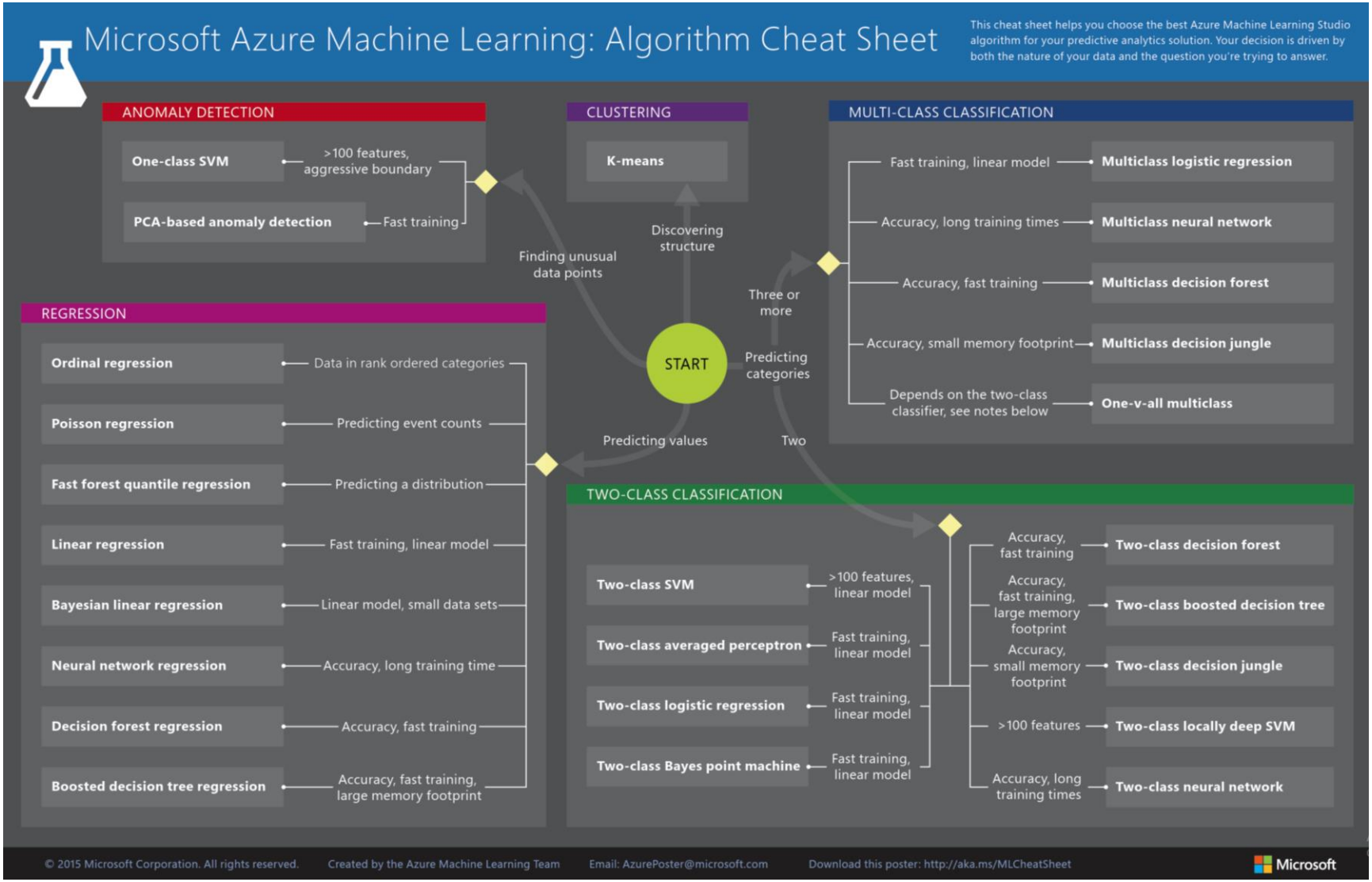
Identify promising models/algorithms

Scikit-learn



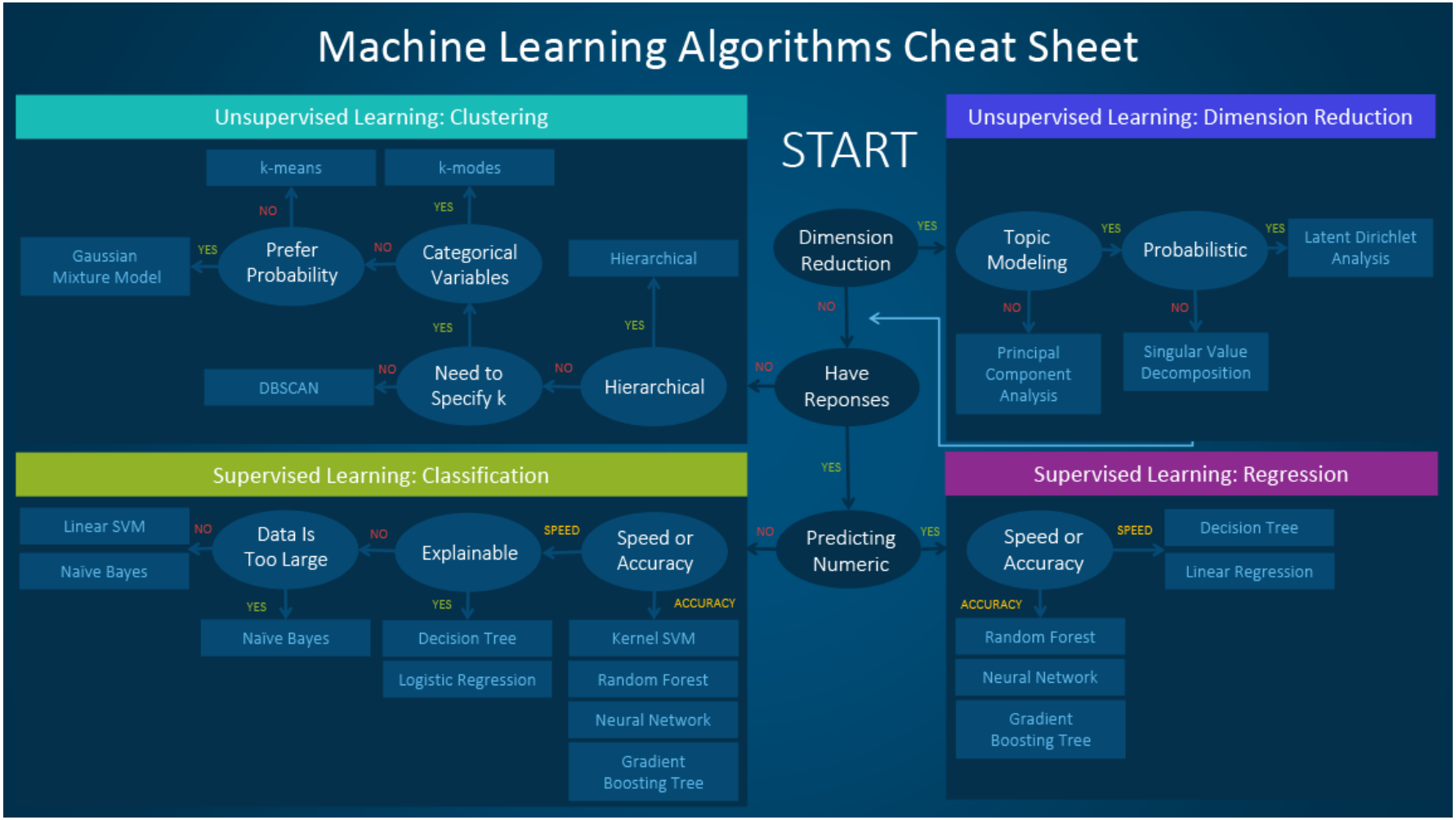
https://cdn-images-1.medium.com/max/1920/1*kJiLzEawYtmD7t-VQ3AXmw.png

Microsoft Azure



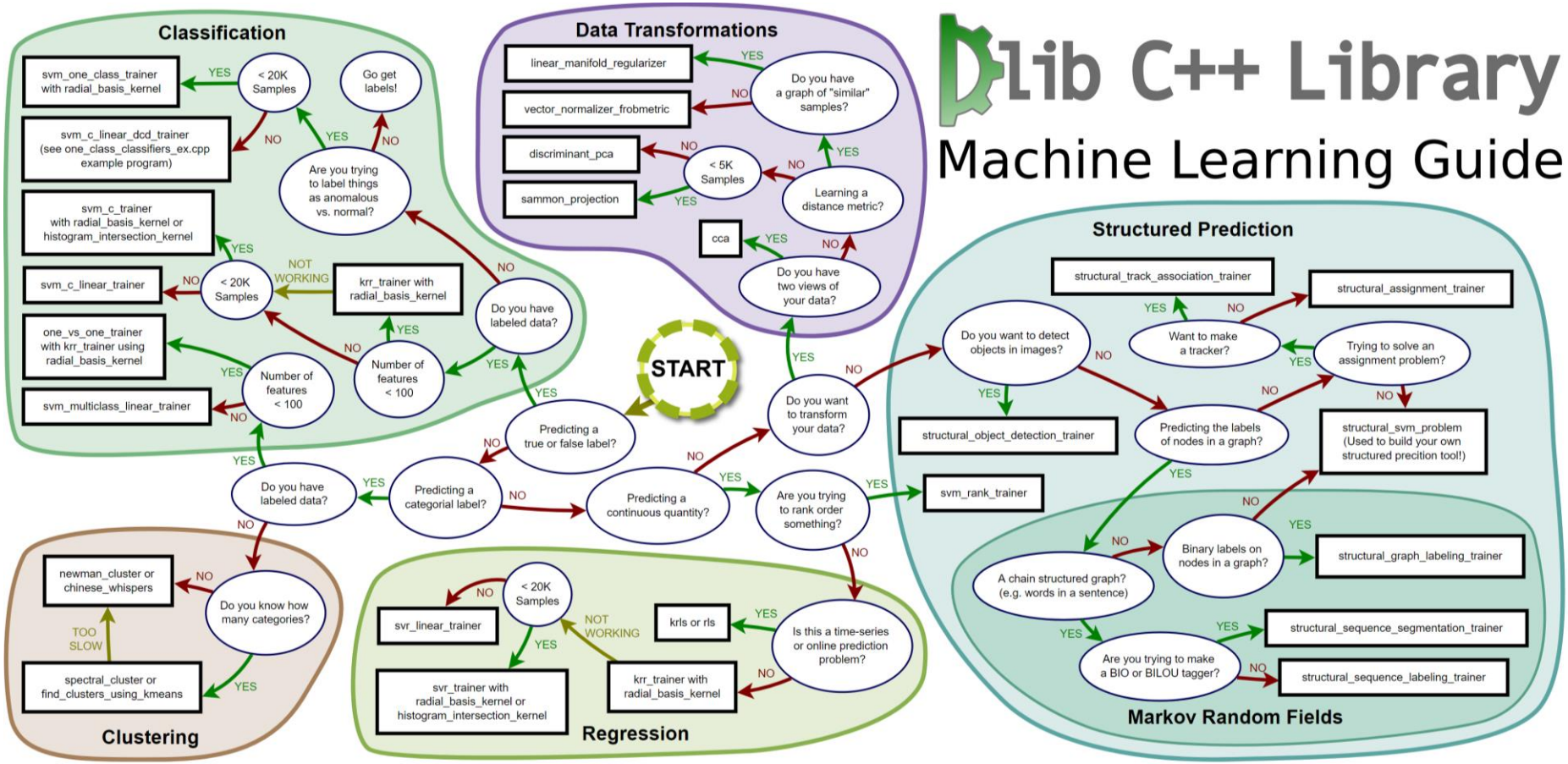
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-cheat-sheet

SAS



http://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

DLib



http://dlib.net/ml_guide.svg

Short-List Promising Models

1. **Train many quick and dirty models from different categories using standard parameters**
2. **Measure and compare their performance**
3. **Analyze the most significant variables for each algorithm**
4. **Analyze the types of errors the models make. What would a human have used to avoid the errors?**
5. **Have a quick round of feature selection and engineering**
6. **Repeat the previous steps two times or so.**
7. **Short-list the top three to five most promising models, preferring models that make different types of errors.**

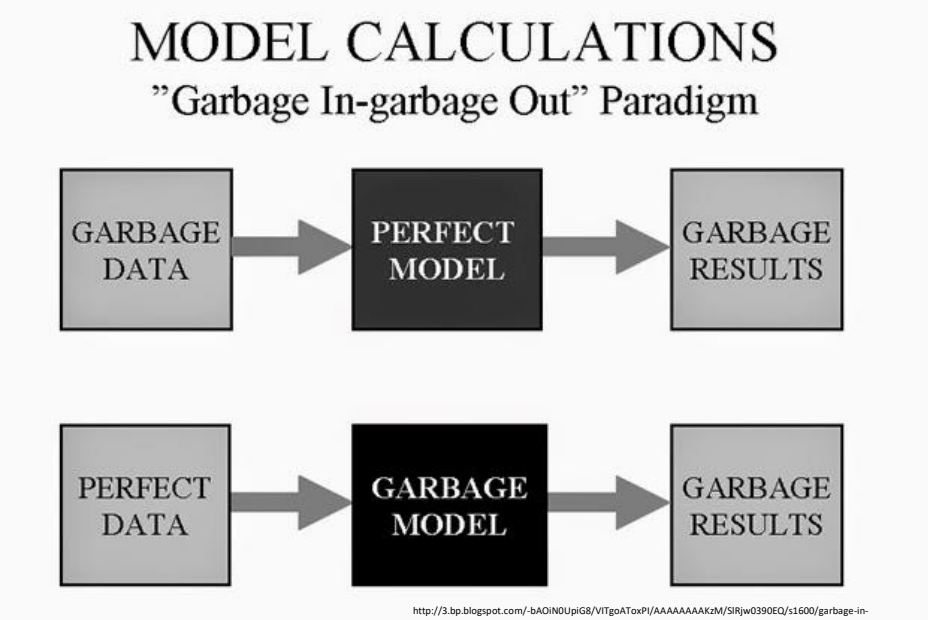
Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

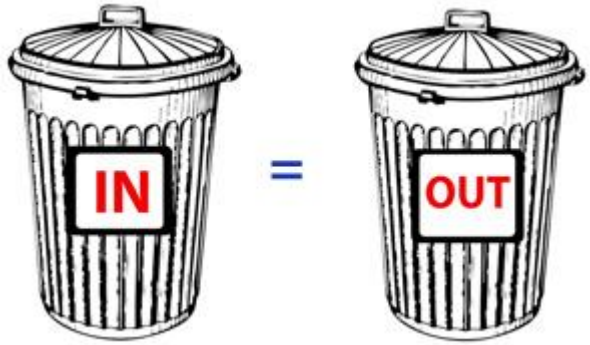
Process Data

Garbage in, garbage out

1. **Data Cleaning**
2. **Feature Selection/Removal: Select/Remove the most/least meaningful features**
3. **Feature Extraction / Dimensionality Reduction: Combining multiple features into one**
4. **Feature creation: Create new features by adding data e.g. from additional sources**
5. **Feature Transformation**
6. **Feature Scaling**



http://3.bp.blogspot.com/-bAOiN0UpiG8/VITgoATOxPI/AAAAAAAAKzM/SIRjw0390EQ/s1600/garbage-in-garbage-out.jpg



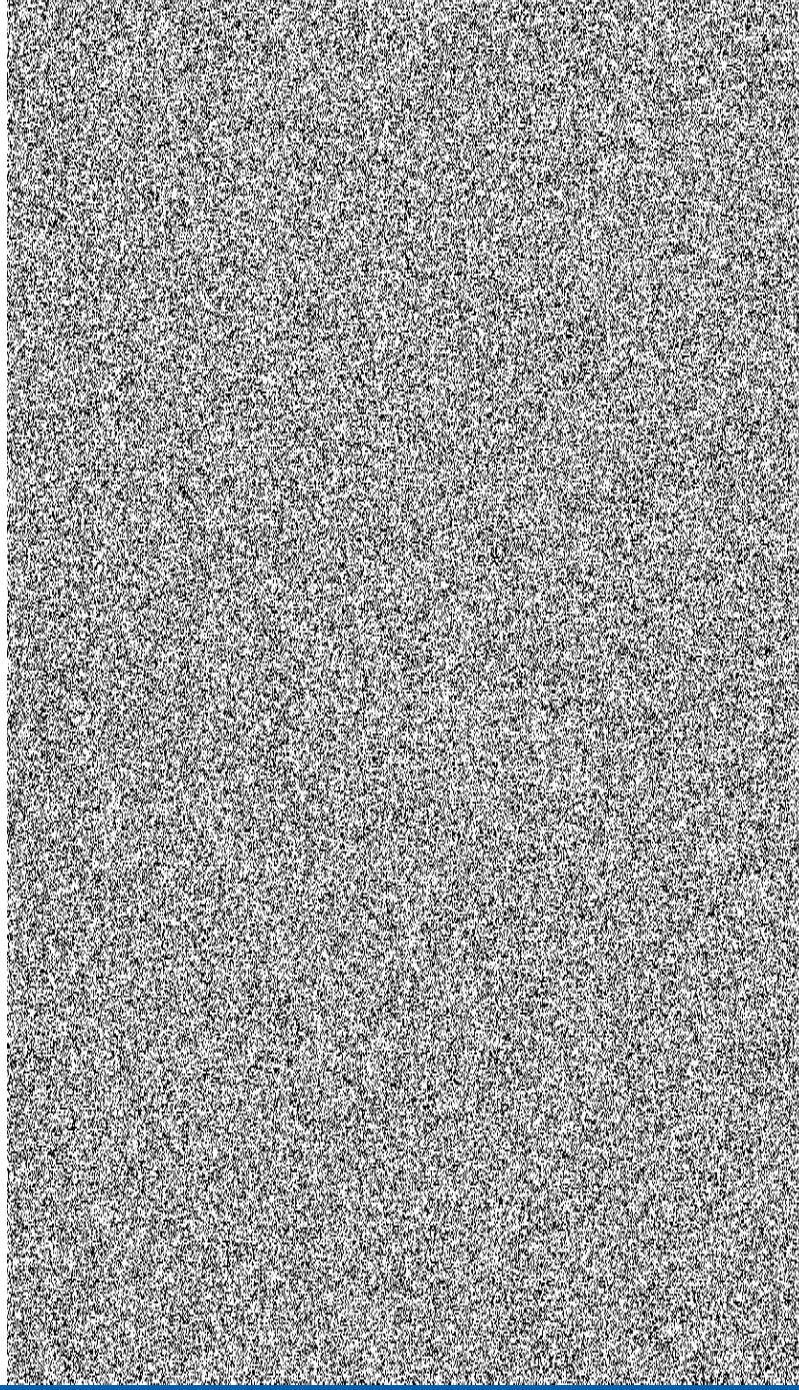
http://alifeoflight.com/wp-content/uploads/2012/01/gigo2.jpg

Aurélien Géron, Hands on Machine Learning with scikit-learn and Tensorflow (O'Reilly Media, 2017).

Clean the Data

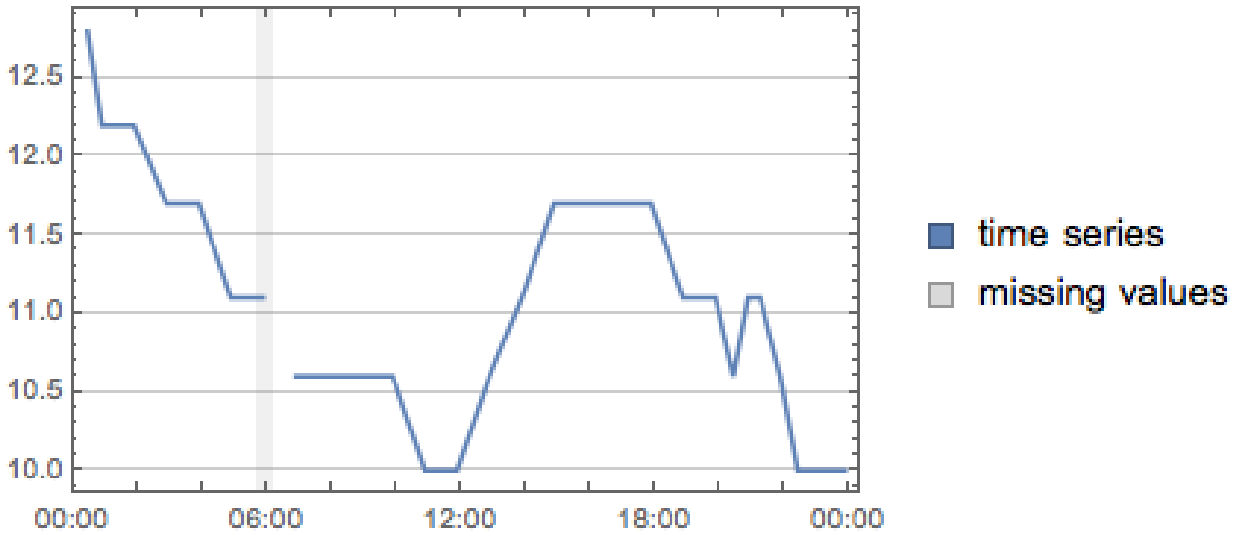
Noise

- **Noise = Errors = False/inaccurate information**
- **Affects both attributes and labels**
- **Stochastic noise**
  - Random
  - E.g. measuring body weight (varies during the day)
- **Human error**
  - Done to save e.g. time
  - E.g. a nurse not measuring blood pressure but just writing down the same value as yesterday
- **Systematic noise**
  - Drags all values in the same direction
  - E.g. a poorly calibrated thermometer
- **More examples**
  - Web statistics and web crawlers, bots, monitoring service, etc.
- **Solutions**
  - Delete or correct the data (if possible)



Miroslav Kubat, An Introduction to Machine Learning (Springer, 2015).

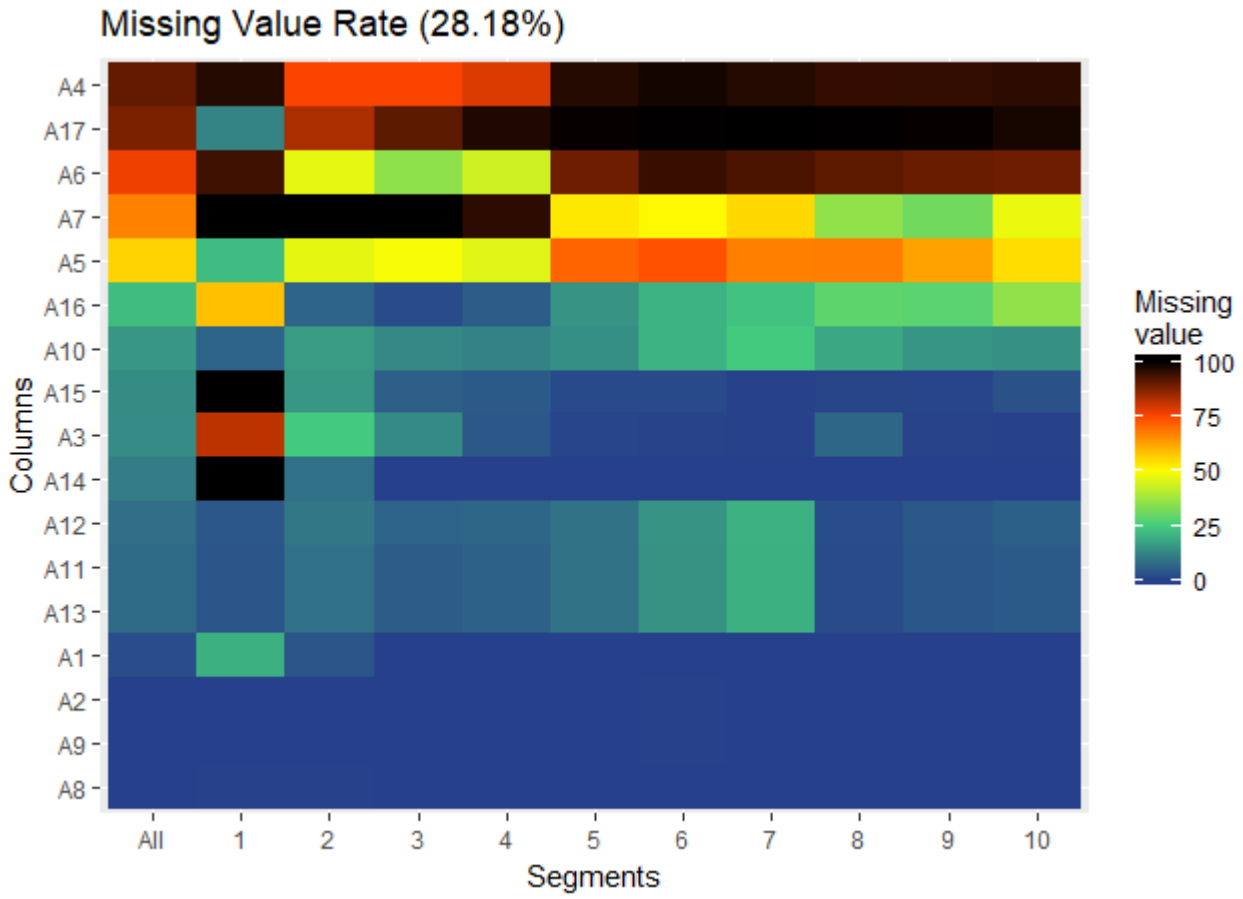
Missing Data



https://www.wolfram.com/mathematica/new-in-10/time-series/HTMLImages.en/work-with-time-series-containing-missing-data/O_29.png

Visualization

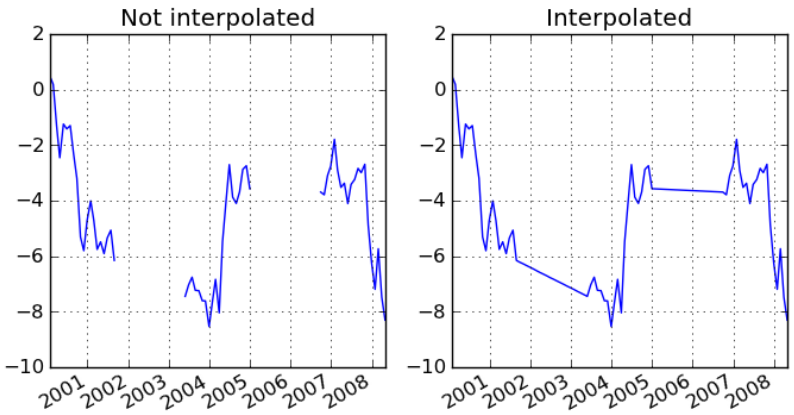
- **Visualization & Tables**



https://jev-pankov.com/2017/11/15/visualize-missing-values-in-r/

Handling Missing Data

1. Delete (if ~<5% or key data is missing)
2. Encode as missing (if the information seems to be important)
   - Discrete data: e.g. -1 (a number at the end of the spectrum)
   - Categorical data: dummy variable (missing = 0|1)
3. Replace/Estimate
   - Average (overall; for certain groups)
     - Mean
     - Median
     - Modal
   - Interpolate / Take from previous or next data point
   - Predict with Machine Learning ☺
     - http://mlg.eng.cam.ac.uk/zoubin/papers/nips93.pdf
     - http://www.gatsby.ucl.ac.uk/aistats/fullpapers/234.pdf
   - Beware if many data points have that same value!



http://pandas.pydata.org/pandas-docs/version/0.7.3/_images/series_interpolate.png

Types of Missing Data

1. **Not Missing at Random (NMAR) / Missing Not at Random (MNAR)**

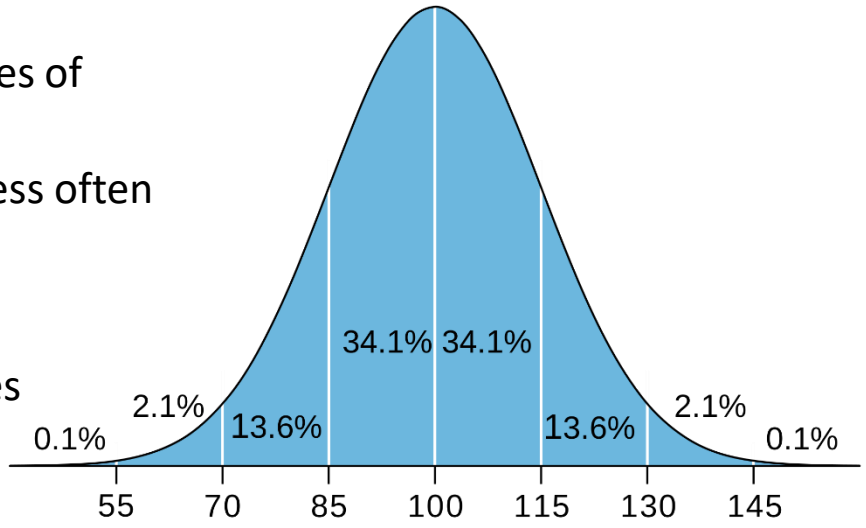
- Missing data relates to the feature itself
- E.g. people with low IQ are less likely to specify their IQ in a survey

2. **Missing at Random (MAR) / Missing Conditionally at Random (MCAR)**

- Probability of missing value relates to values of another variable
- E.g. older people tend to specify their IQ less often than younger people in a survey

3. **Missing Completely at Random (MCAR)**

- Missing data is not correlated to any features
- Occurs rarely
- E.g. surveys being lost in the mail.



https://commons.wikimedia.org/wiki/File:IQ_distribution.svg

Missing Data: Decision Diagram

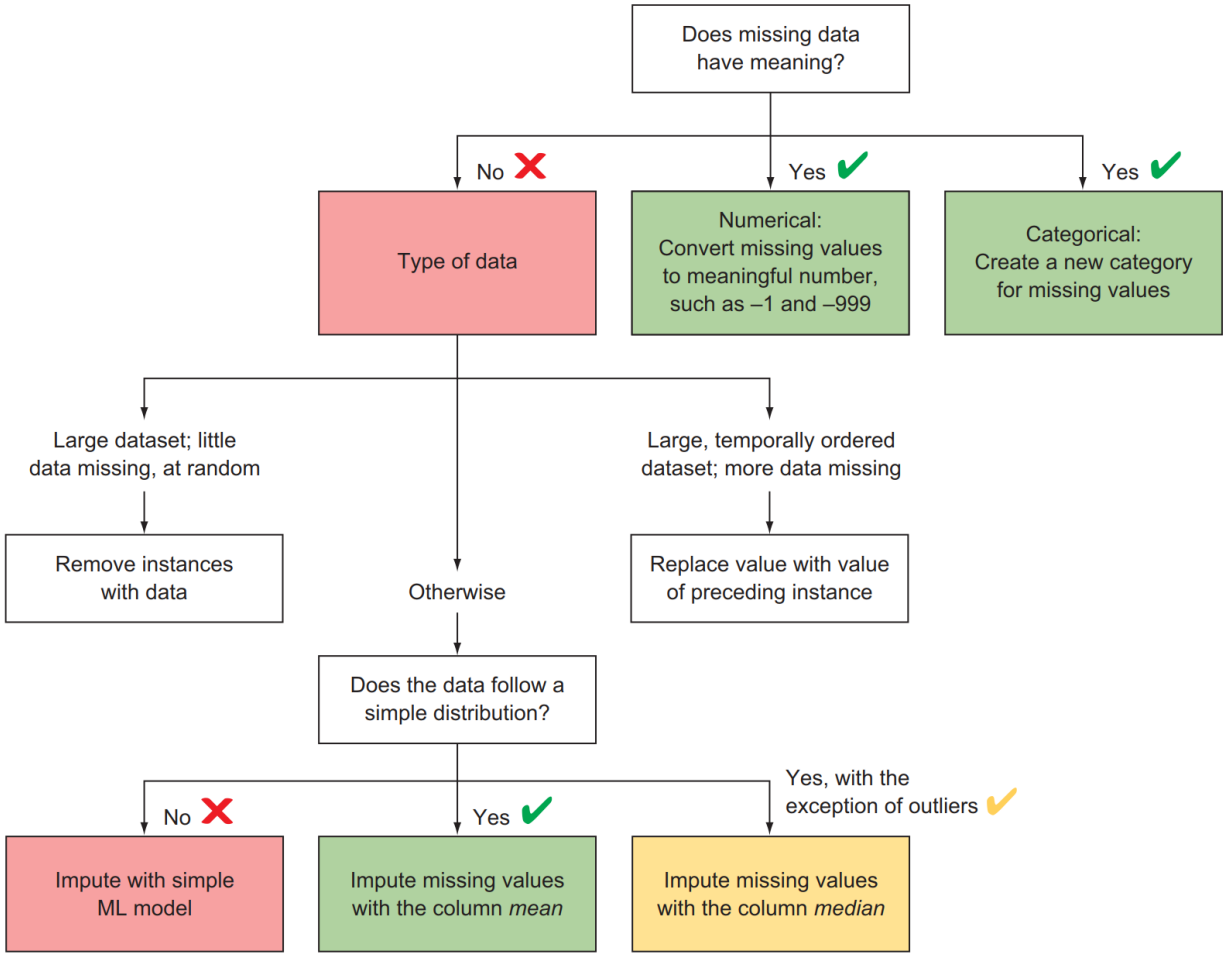
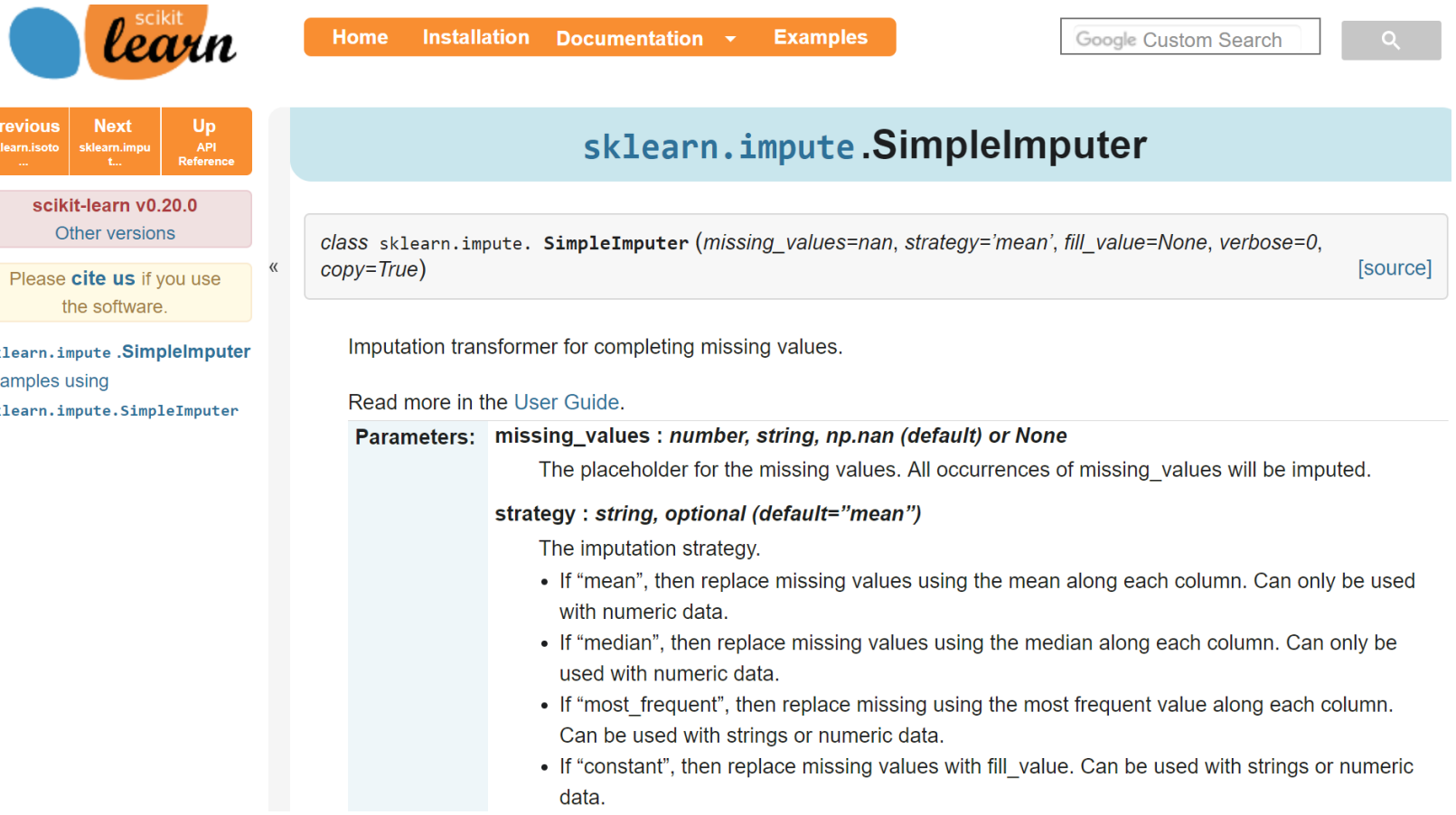


Figure 2.9 Full decision diagram for handling missing values when preparing data for ML modeling

Henrik Brink, Joseph Richards, and Mark Fetherolf, *Real-world machine learning* (Manning Publications Co., 2016)

Missing Data Handling in ML libraries



Effect of the Method to Handle Missing Data

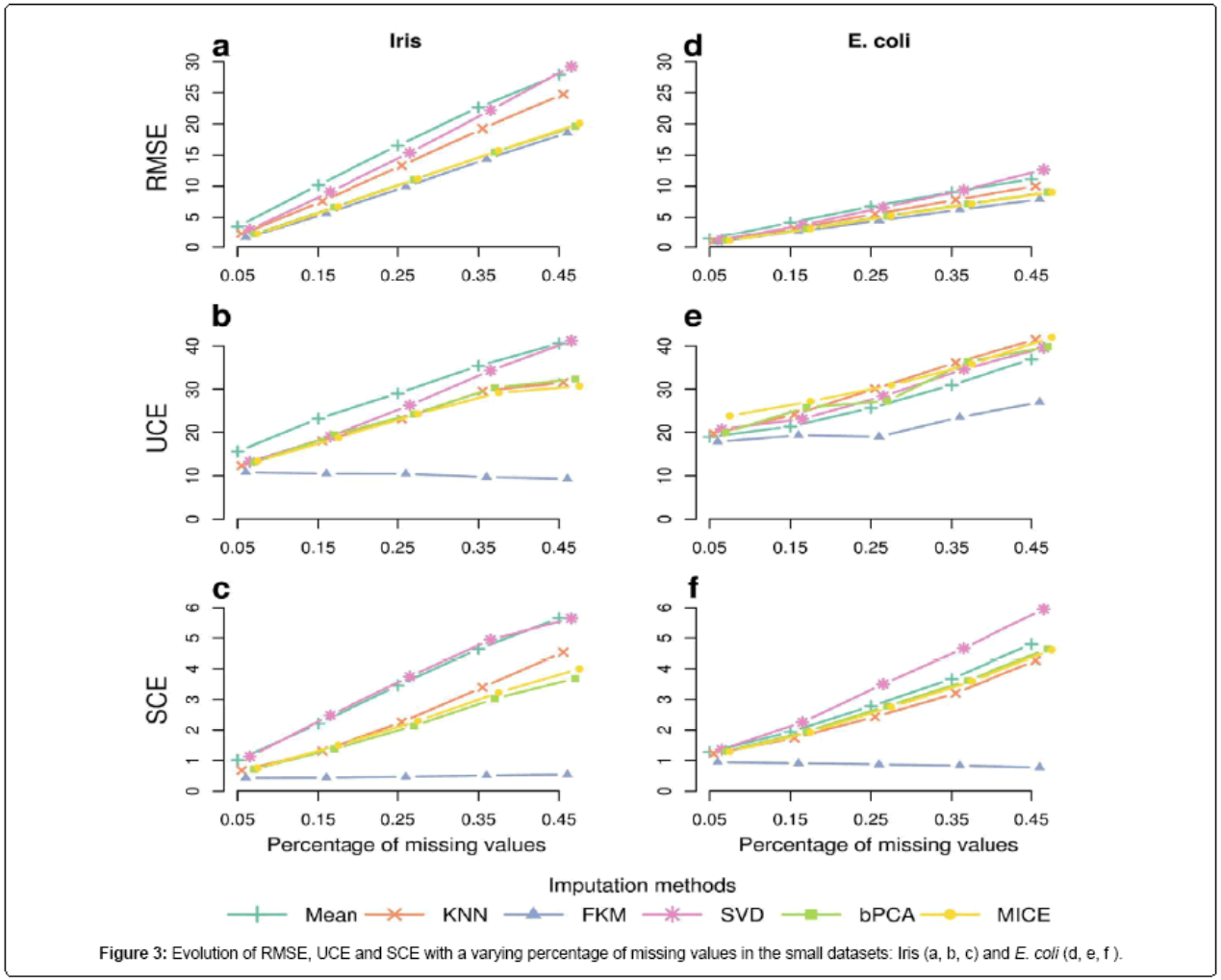


Figure 3: Evolution of RMSE, UCE and SCE with a varying percentage of missing values in the small datasets: Iris (a, b, c) and *E. coli* (d, e, f).

https://www.omicsonline.org/open-access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.php?aid=54590

„Missing Data“ is not a new phenomena

PSYCHOMETRIKA—VOL. 52, NO. 3, 431–462
SEPTEMBER 1987

ON STRUCTURAL EQUATION MODELING WITH DATA THAT ARE NOT MISSING COMPLETELY AT RANDOM

BENGT MUTHÉN

DAVID KAPLAN

GRADUATE SCHOOL OF EDUCATION

MICHAEL HOLLIS

GRADUATE SCHOOL OF ARCHITECTURE AND URBAN PLANNING
UNIVERSITY OF CALIFORNIA, LOS ANGELES

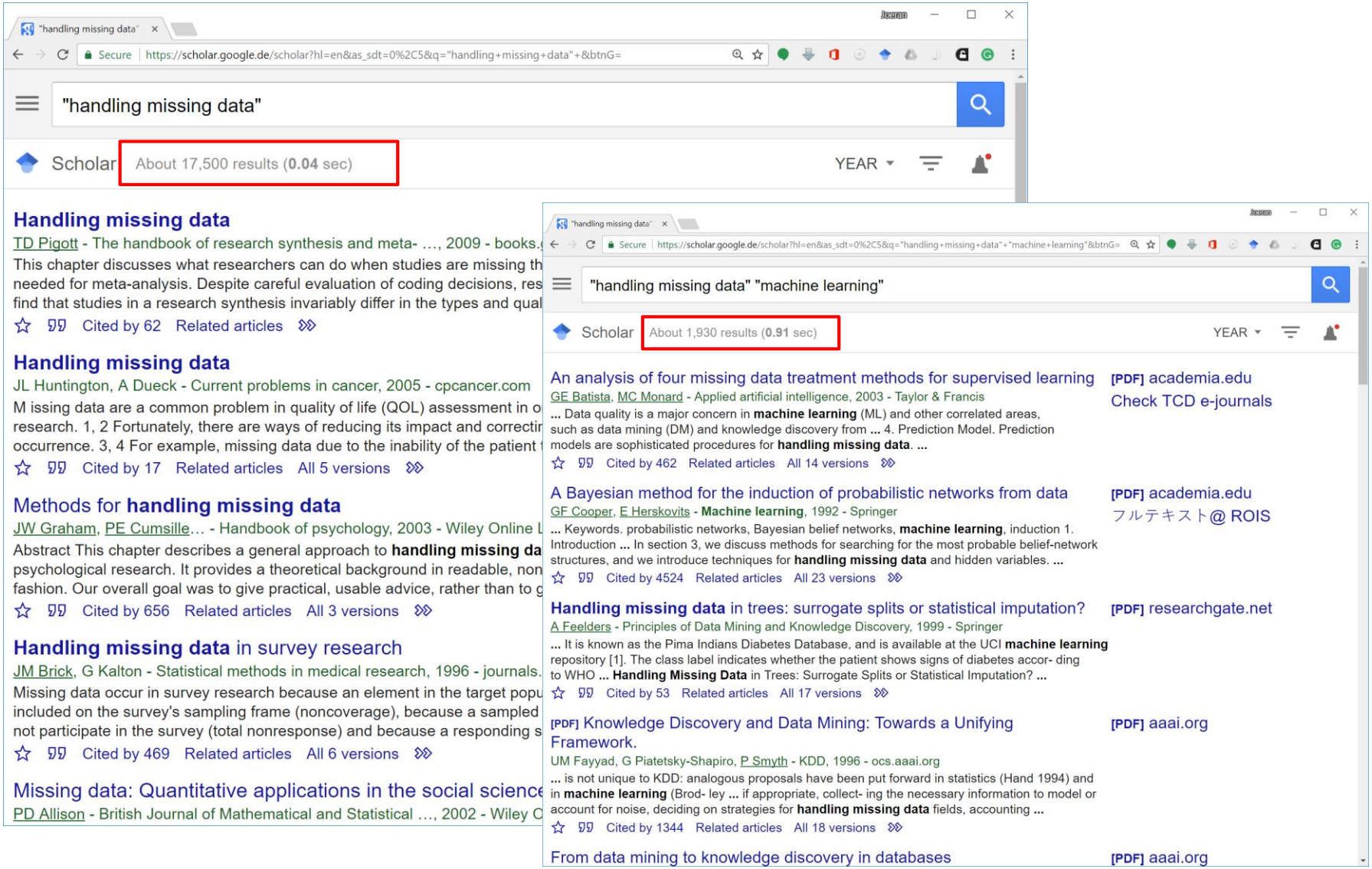
A general latent variable model is given which includes the specification of a missing data mechanism. This framework allows for an elucidating discussion of existing general multivariate theory bearing on maximum likelihood estimation with missing data. Here, missing completely at random is not a prerequisite for unbiased estimation in large samples, as when using the traditional listwise or pairwise present data approaches. The theory is connected with old and new results in the area of selection and factorial invariance. It is pointed out that in many applications, maximum likelihood estimation with missing data may be carried out by existing structural equation modeling software, such as LISREL and LISCOMP. Several sets of artifical data are generated within the general model framework. The proposed estimator is compared to the two traditional ones and found superior.

Key words: maximum likelihood, ignorability, selectivity, factor analysis, factorial invariance, LISREL.

1. Introduction

Confirmatory factor analysis and structural equation modeling (see e.g., Jöreskog, 1969, 1977) need often be applied in situations where data are missing on certain variables and it cannot be realistically assumed that the data are missing completely at random. Ordinary methods would in these cases give estimates that are both inefficient and have large sample bias. Existing missing data theory that provide better alternatives (see e.g., Anderson, 1957; Beale & Little, 1975; Little & Rubin, 1987; Rubin, 1974, 1976) does not seem to have been adapted in factor analysis and structural equation modeling practice. Reasons for this may include lack of familiarity with missing data theory and the fact that general maximum likelihood estimation requires special computational routines as in Finkbeiner (1979); see also Dempster, Laird, & Rubin (1977).

Many many more

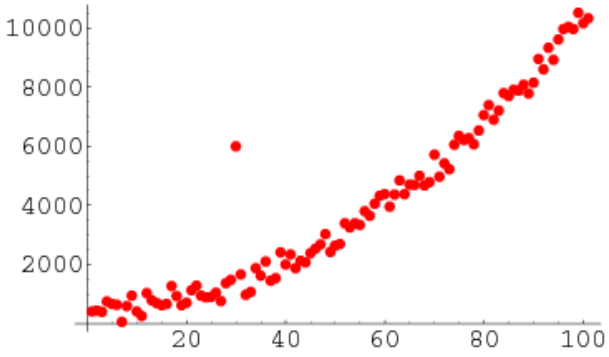


Inconsistent Data

- **Different data types (e.g. sometimes ‚1', sometimes ‚true' and sometimes ‚yes'**
- **Different data structures (different features saying the same, but used at different times during data collection)**
- **Identification**
  - Create a list of unique values and analyse
  - Validate against schema if available
- **Solution: Make it consistent**

ID	Birthyear	Year of birth	Gender	Income
1	1968		m	45,393 €
2	1978		male	50,428 €
3	1984		1	88,262 €
4	1983		f	86,138 €
5	1998		female	86,302 €
6	1967		woman	50,425 €
7	1969		0	56,093 €
8	1955		female	69,821 €
9	1959		m	31,287 €
10	1987		1	79,561 €
11	1964		1	86,124 €
12	1983		m	86,370 €
13	1974		female	82,764 €
14	1989		f	52,051 €
15		1997	woman	65,375 €
16		1966	male	75,785 €
17		1971	m	35,913 €
18		1963	woman	70,965 €
19		1984	f	86,599 €
20		1964	male	79,818 €

Outliers (!= Noise)

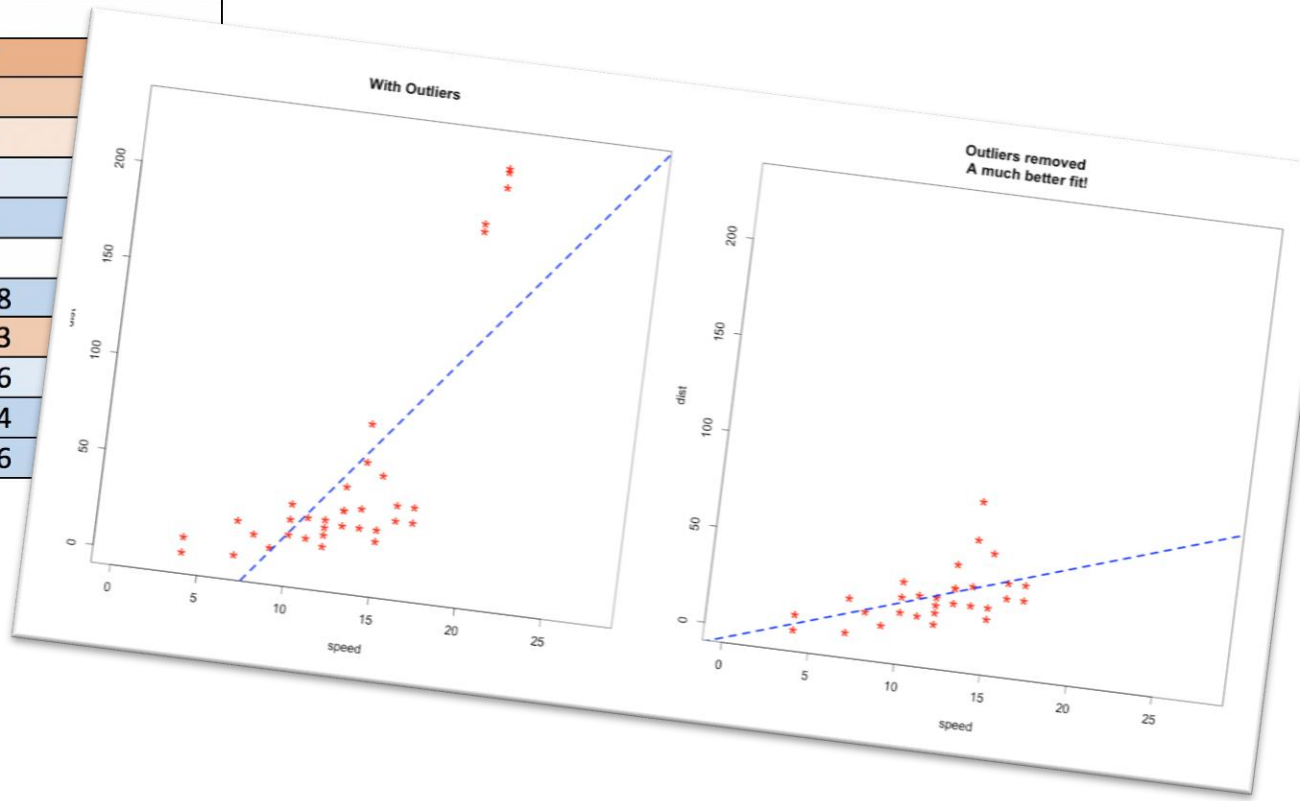


http://mathworld.wolfram.com/images/eps-gif/OutlierScatterplot_1000.gif

Potential Effect Of Outlier Detection

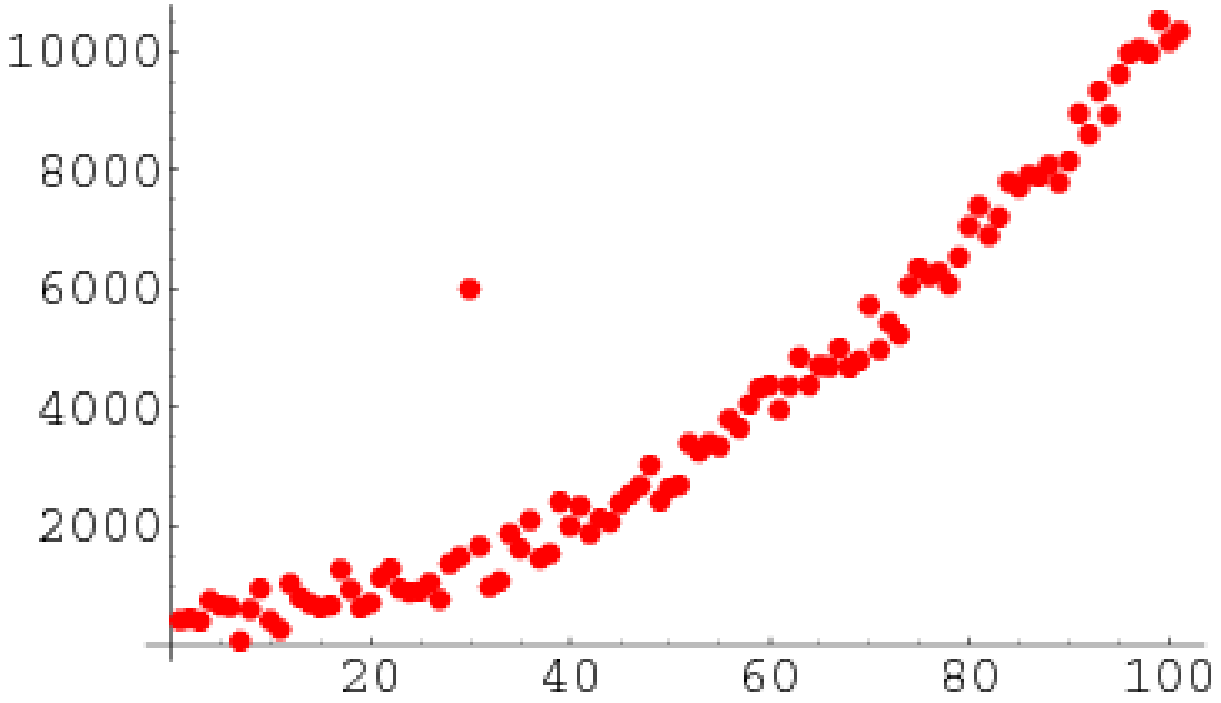
Method	RMSE
Linear Regression Models	
With Outlier	0.93
Winsorizing (0.05,0.95)	0.44
Removal - Z-Score	0.22
Removal - IQR	0.20
Log-Transformation	0.18
Random Forest Regressor	
With Outlier (Default Criteria -MSE)	0.188
Winsorizing (0.05,0.95)	0.753
Removal - IQR	0.206
Log-Transformation	0.184
With Outlier (Criteria - MAE)	0.186

https://www.kdnuggets.com/2018/08/make-machine-learning-models-robust-outliers.html



https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/

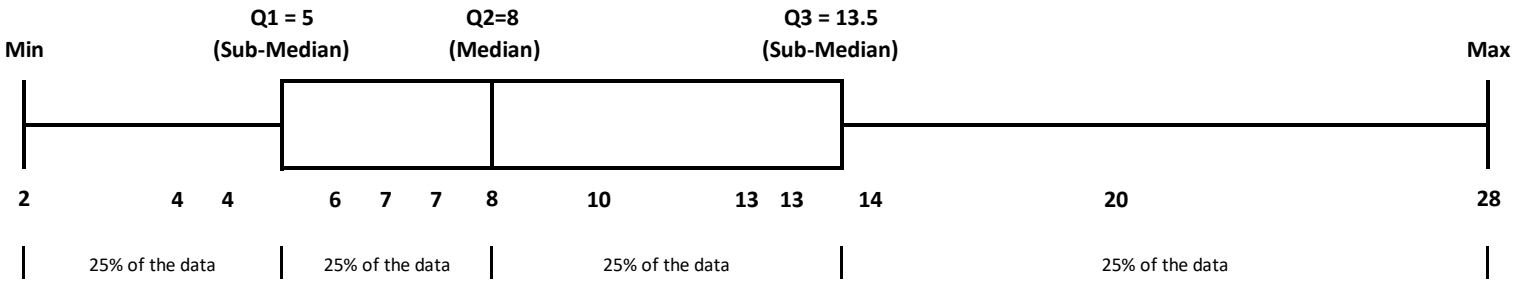
Identifying Outliers: Scatter Plot



http://mathworld.wolfram.com/images/eps-gif/OutlierScatterplot_1000.gif

Identifying Outliers: Box Plots

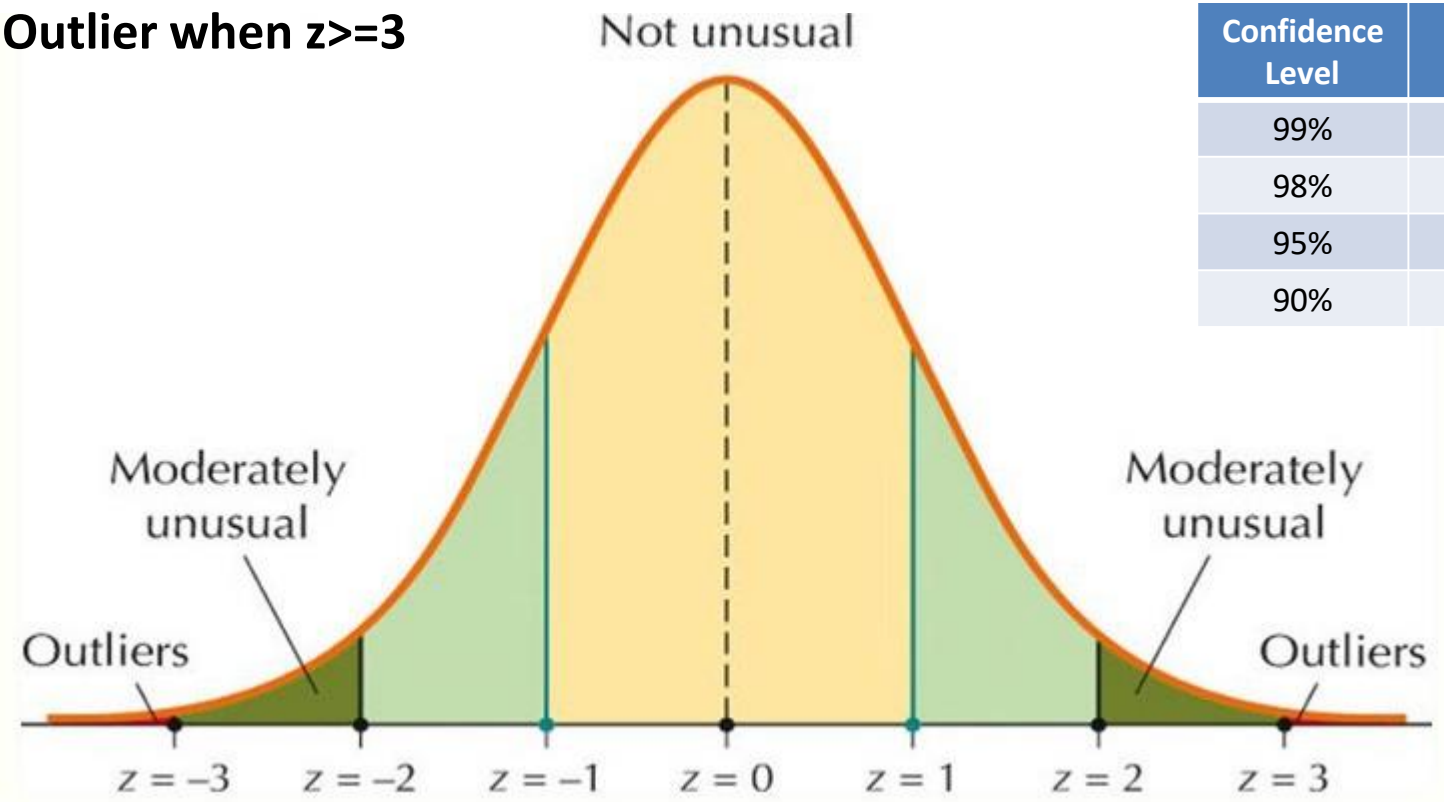
- **Sort values**
- **Find Quartile Ranges, i.e. Median and Sub-Medians (when the amount of values is even, average the two nearest values)**
- **Calculate Inter Quartile Range**
  - IQR = Q3 – Q1 [default for our lecture and calculations]
  - Or IRQ = Diff_Mid_50%(Smallest;Largest)
- **Calculate Upper Limit** *(Q3 + 1.5 * IQR)* **and Lower Limit (***Q1 – 1.5 * IQR***)**
- **Example: 10, 2, 4, 7, 8, 13, 20, 28, 4, 7, 6, 13, 14**



Outliers	Lower Limit (Q1 - 1.5 * IQR) = -7.75	IQR = Q3 - Q1 = 13.5 - 5 = 8.5 or IQR = Diff_Mid_50%(Smallest;Largest) = 13-6 = 7	Upper Limit (Q3 + 1.5 * IRQ) = 26.25	Outliers

Identifying Outliers: Z-Scores

- **Required: Gaussian Distribution**
- **z = number of standard deviations from the mean**
- **Outlier when z>=3**



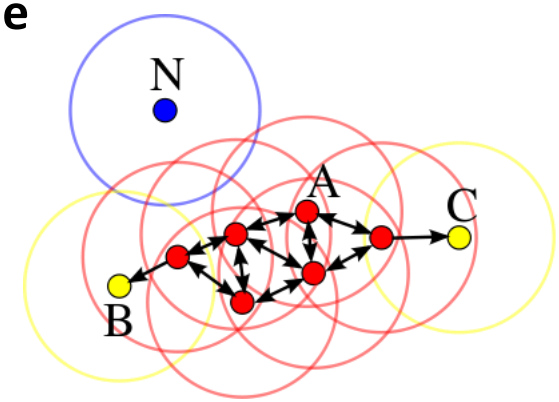
Confidence Level	z
99%	2.58
98%	2.33
95%	1.96
90%	1.65

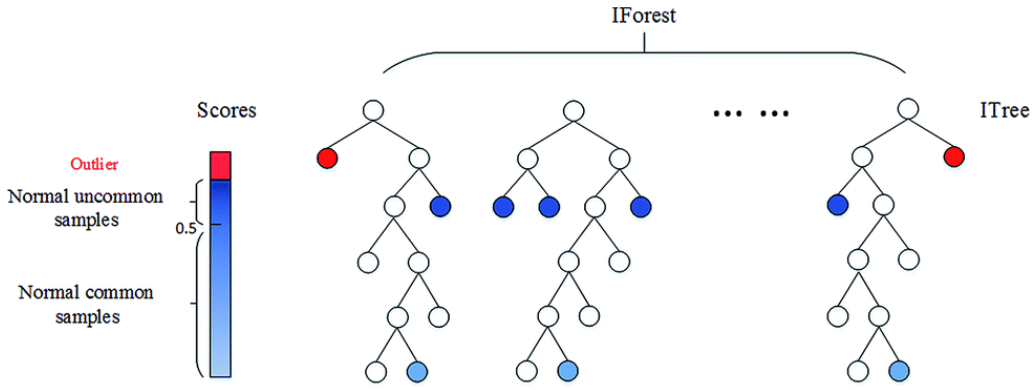
https://www.kdnuggets.com/2018/08/make-machine-learning-models-robust-outliers.html

Identifying Outliers: More

https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

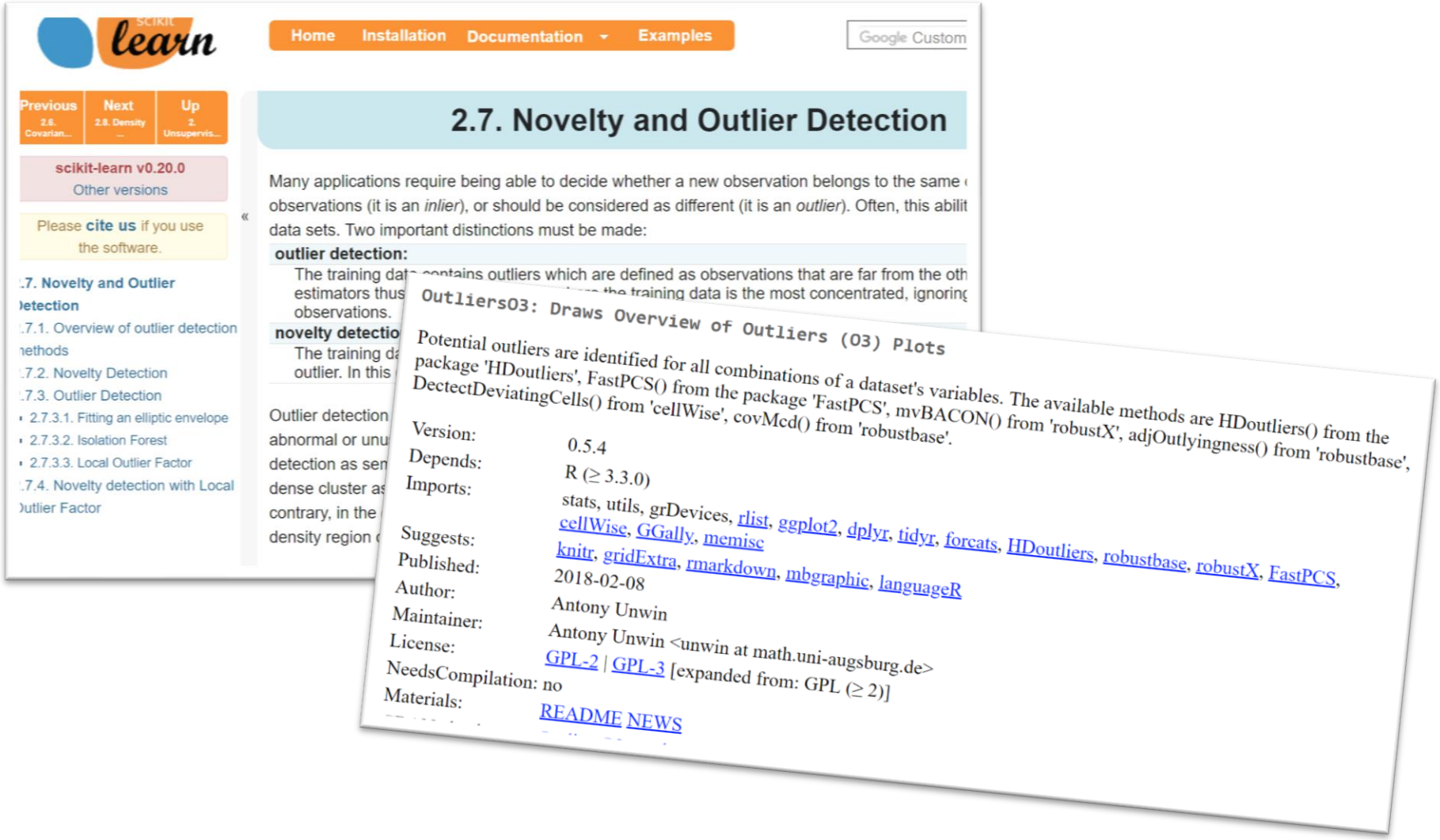
- **Cook's Distance (only for regression): Measures the effect that a single observation has on the model**
- **Principle Component Analysis**
- **Isolation Forests**
- **LOF (Local Outlier Factor)**
- **HiCS: High Contrast Subspaces for Density-Based Outlier Ranking**
- **Dbscan (Density Based Spatial Clustering of Applications with Noise)**
- **…**





https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

Outlier Detection in ML libraries



Dealing with Outliers

- **Remove**
- **Winsorize (Cap at Threshold)**
  - Decide on top and bottom threshold (e.g. top/bottom 5% of data = 90% Winsorization)
  - Replace values within the top/bottom range with the threshold value
  - Example (80% Winsorization)
    - Before: {**0.1,1**,12,14,16,18,19,21,24,26,29,32,33,35,39,40,41,44,**99,125**} Mean = 33.405.
    - After: {**12,12**,12,14,16,18,19,21,24,26,29,32,33,35,39,40,41,44,**44,44**} 80% Winsorized mean = 24.95.
- **Transform (log / normalize / standardize)**
- **Binning**

https://www.statisticshowto.datasciencecentral.com/winsorize/

Feature Transformation

The need for Transformation

- **Many machine learning algorithms need quantitative data (e.g. linear regression and SVMs)**
- **Categorical and ordinal data needs to be transformed**

Student Number	Number of Lectures Attended (x1)	Number of hours spent for exam preperation (x2)	IQ (x3)	Gender (x4) ?
1	20	14	129	Male
2	6	40	87	Trans
3	14	10	90	Male
4	19	28	140	Female
5	22	21	116	Male
6	3	17	128	Female
7	17	22	118	Trans
8	20	32	84	Female
9	12	13	116	Male

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \overset{?}{\boldsymbol{x_4}}$$

What to do (for multiple regression)?

Student Number	Number of Lectures Attended (x1)	Number of hours spent for exam preperation (x2)	IQ (x3)	Gender (x4)
1	20	14	129	Male
2	6	40	87	Trans
3	14	10	90	Male
4	19	28	140	Female
5	22	21	116	Male
6	3	17	128	Female
7	17	22	118	Trans
8	20	32	84	Female
9	12	13	116	Male

Option A

Gender (x4)
0
2
0
1
0
2
2
1
0

Option B

Gender_Male (x4)	Gender_Female (x5)	Gender_Trans (x6)
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

Option A: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \boldsymbol{x_4}$

Option B: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \boldsymbol{x_4} + \beta_5 \boldsymbol{x_5} + \beta_6 \boldsymbol{x_6}$

Student Number	Number of Lectures Attended (x1)	Number of hours spent for exam preperation (x2)	IQ (x3)	Gender (x4)
1	20	14	129	Male
2	6	40	87	Trans
3	14	10	90	Male
4	19	28	140	Female
5	22	21	116	Male
6	3	17	128	Female
7	17	22	118	Trans
8	20	32	84	Female
9	12	13	116	Male

Option A

Gender (x4)
0
2
0
1
0
2
2
1
0

Option B

Gender_Male (x4)	Gender_Female (x5)	Gender_Trans (x6)
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

Go to **www.menti.com** and use the code **24 94 8**

Feature Transformation: Transform "gender" into one or three new variables?

Mentimeter

0	0	0	0
Option A (one variable)	Option B (three variables)	Neither A nor B	I have no clue at all

Slide is not active Activate

0

Dummy Encoding

- **Binary representation of categorical variables**

Student Number	Number of Lectures Attended (x1)	Number of hours spent for exam preperation (x2)	IQ (x3)	Gender (x4)
1	20	14	129	Male
2	6	40	87	Trans
3	14	10	90	Male
4	19	28	140	Female
5	22	21	116	Male
6	3	17	128	Female
7	17	22	118	Trans
8	20	32	84	Female
9	12	13	116	Male

Option A

Gender (x4)
0
2
0
1
0
2
2
1
0

Option B

Gender_Male (x4)	Gender_Female (x5)	Gender_Trans (x6)
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \boldsymbol{x_4} + \beta_5 \boldsymbol{x_5} + \beta_6 \boldsymbol{x_6}$$

→ **Any problems (for regression)?**

One-Hot Encoding / Dummy Variables

- **The Dummy Variable Trap**
- **Occurs when two variables are highly correlated**
- **For regression, always use n-1 dummy variables („Dummy Encoding“)**
- **For e.g. SVM, use n dummy variables**

Student Number	Number of Lectures Attended (x1)	Number of hours spent for exam preperation (x2)	IQ (x3)	Gender (x4)
1	20	14	129	Male
2	6	40	87	Trans
3	14	10	90	Male
4	19	28	140	Female
5	22	21	116	Male
6	3	17	128	Female
7	17	22	118	Trans
8	20	32	84	Female
9	12	13	116	Male

Gender (x4)
0
2
0
1
0
2
2
1
0

Option A

Gender_Male (x4)	Gender_Female (x5)	Gender_Trans (x6)
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0
0	1	0
0	0	1
0	1	0
1	0	0

Option B

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \boldsymbol{x_4} + \beta_5 \boldsymbol{x_5} + \beta_6 \boldsymbol{x_6}$$

Make features more meaningful (for both humans and ML)

- **Some variables have no immediate meaning**
- **For instance, Titanic dataset (predict survival)**
  - Cabin number itself is rather meaningless
  - When transformed to e.g. deck (upper, lower, …) or where in the ship the cabin is (front, middle, back), this might change

PassengerId	Survived	Pclass	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Male	22	1	0	A/5 21171	7.25		S
2	1	1	Female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Female	26	0	0	STON/02. 3101282	7.925		S
4	1	1	Female	35	1	0	113803	53.1	C123	S
5	0	3	Male	35	0	0	373450	8.05		S
6	0	3	Male		0	0	330877	8.4583		Q

Henrik Brink, Joseph Richards, and Mark Fetherolf, *Real-world machine learning* (Manning Publications Co., 2016)

Discretize continuous features

- **Sometimes, continuous numbers have no „typical“ meaning (the higher the better/worse)**
  - E.g. GPS coordinates

→ a GPS coordinate might become a country name, which would be encoded in multiple dummy variables

	Longitude	Latitude
Instance 1	-6.2603097	53.3498053
Instance 2	139.6917064	35.6894875

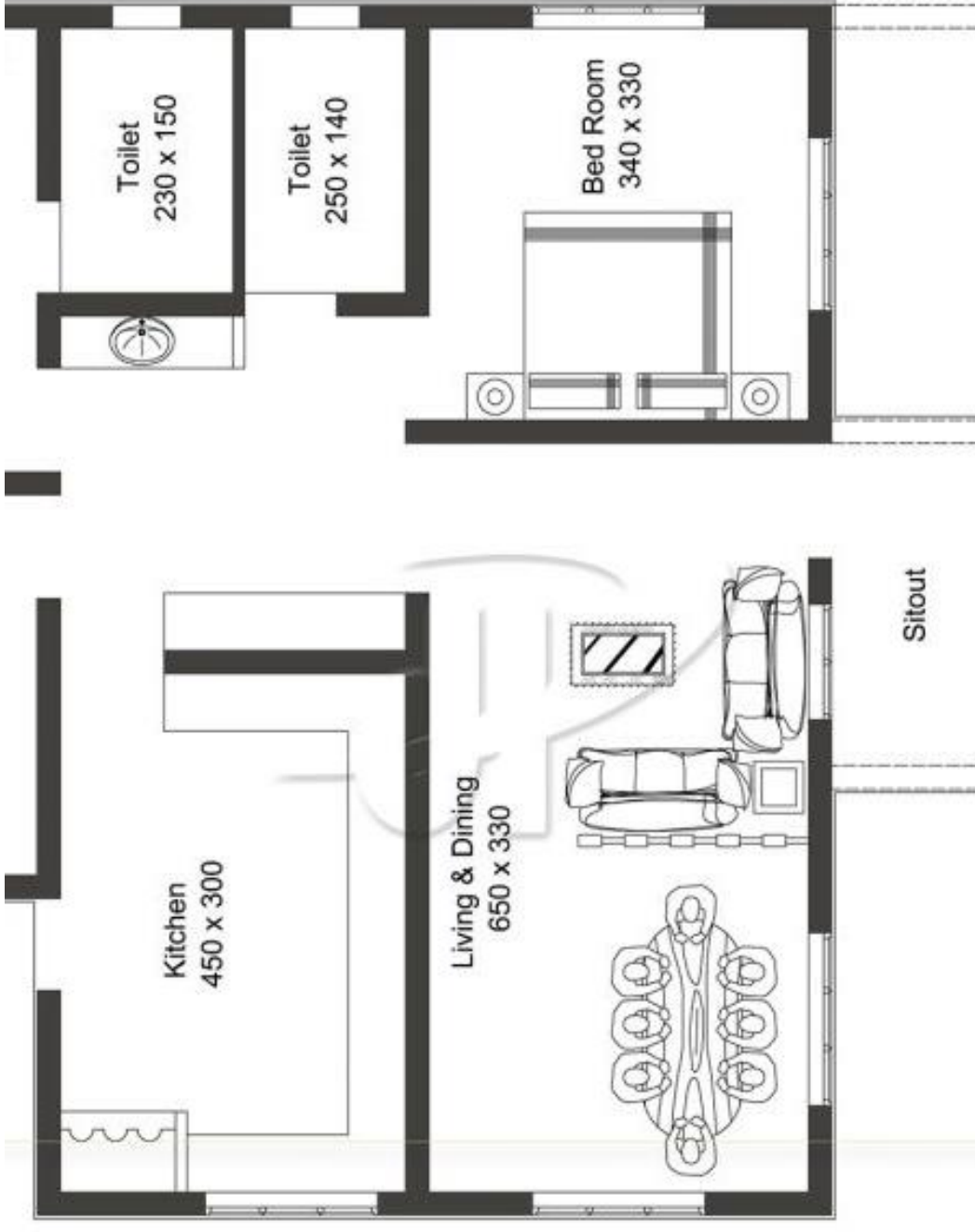
	Country	City
Instance 1	Ireland	Dublin
Instance 2	Japan	Tokyo

	Country (Ireland)	Country (Japan)	City (Dublin)	City (Tokyo)
Instance 1	1	0	1	0
Instance 2	0	1	0	1

Data/Feature Creation/Merging

Merge Features

- **Sometimes, individual variables make more sense when they are merged**
- **Example**
  - Length and width of a room → size (m^2)
  - Multiply Noise and Price of a Car to predict how much someone likes the car (noisy expensive car is probably a sport car; noisy cheap car is probably just trash).



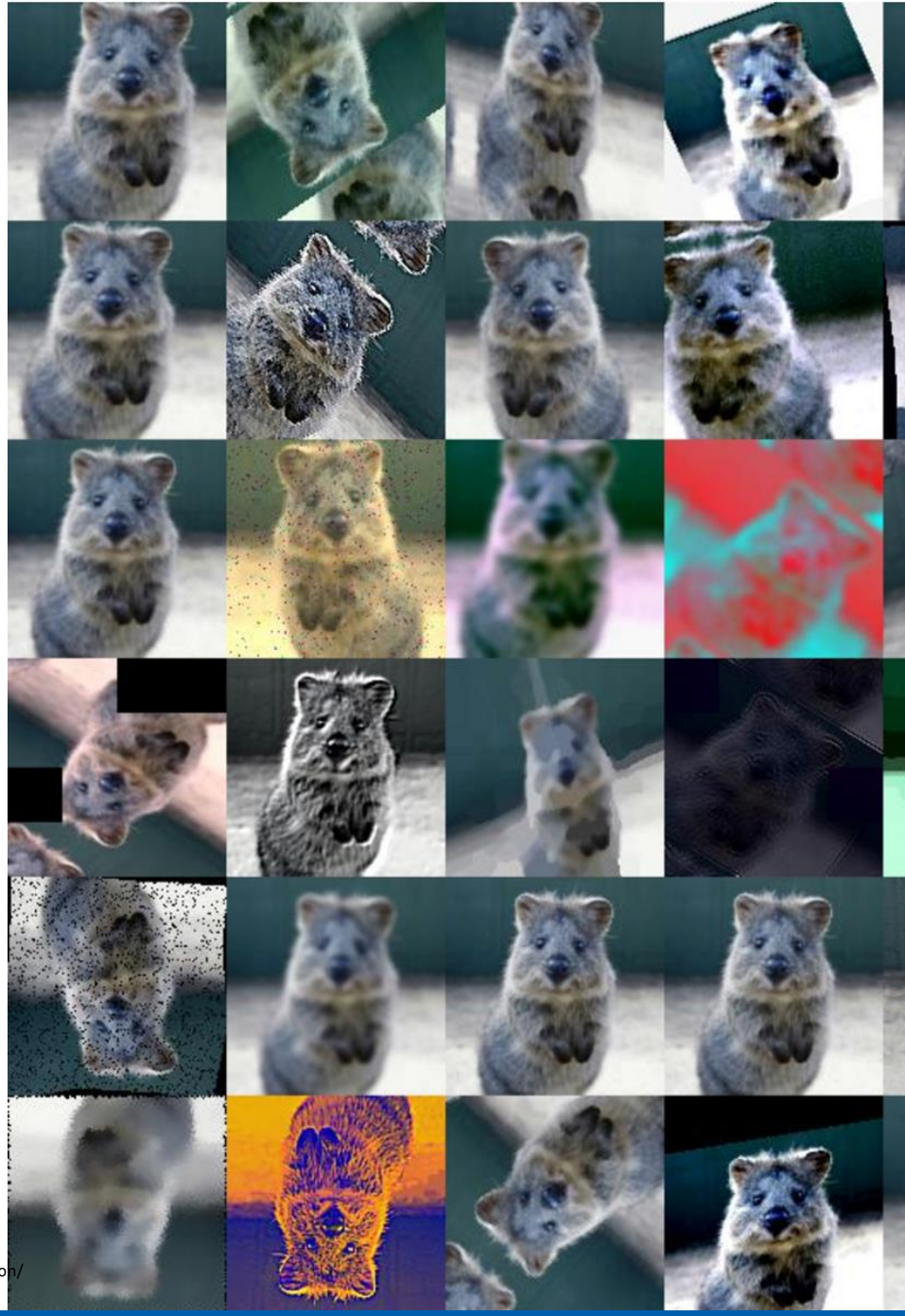
https://2.bp.blogspot.com/-HKDyCL1F52I/WYONeaZ0NKI/AAAAAAAAHZE/ZfEqQ7fG6089KmyfobPkUntW7n7UYIs8QCLcBGAs/s0/cc1.jpg

Automatic Feature Creating

- **Multiply**
- **Power**
- **Sum**
- **…**
- **More details**
  - https://www.dummies.com/programming/big-data/data-science/machine-learning-creating-features-data/
  - https://towardsdatascience.com/why-automated-feature-engineering-will-change-the-way-you-do-machine-learning-5c15bf188b96

Dataset Augmentation and Expansion

- **Create new instances, e.g. via**
  - Flipping (both vertically and horizontally)
  - Rotating
  - Zooming and scaling
  - Cropping
  - Translating (moving along the x or y axis)
  - Adding Gaussian noise (distortion of high frequency features)
- **More details on using ML for Creating Augmented Datasets**
  - AutoAugment: Learning Augmentation Policies from Data
  - https://arxiv.org/abs/1805.09501v1

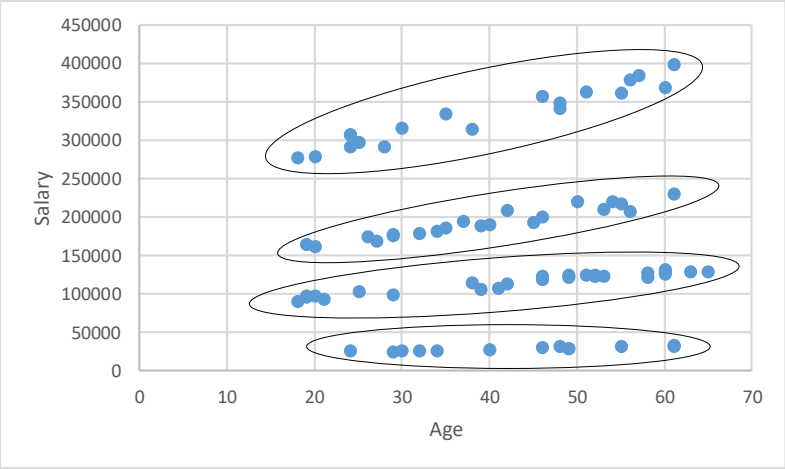


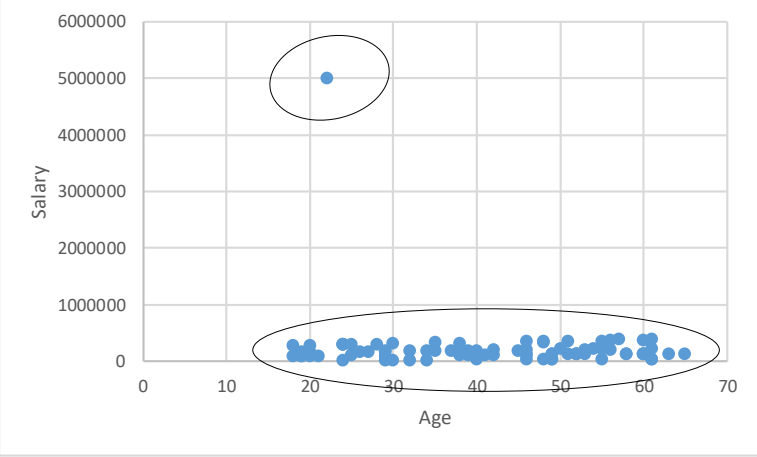
https://blog.algorithmia.com/introduction-to-dataset-augmentation-and-expansion/

Feature Scaling

Problem with Different Scales

- **Variables with larger scales have stronger impact („dominate“) if ML algorithm uses e.g. Euclidian distance or clustering**





Solutions

- **Simple Methods**
  - Square Root
  - Log
  - Square (opposite effect)
- **Min-Max Normalization / Rescaling**
  - Scale original value to a value between 0 and 1
  - As a variant for image processing (RGB range 0-255).
  - Required by some neural networks

$$x_{MinMaxNorm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Standardization/Z-score normalization**
  - Assumes Gaussian Distribution
  - Goal: Mean = 0 and Standard Deviation = 1
  - Typically used for
    - Neural Networks
    - SVM
    - Linear/Logistic Regression
    - Cluster Analysis
    - Principle Component Analysis

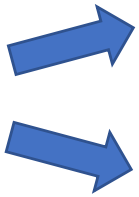
$$x_Z = \frac{x - \mu}{\sigma}$$

σ = standard deviation
μ = average (mean)

Example (Standardization vs. Normalization)

	Age	Income
Person 1	18	5,000 €
Person 2	20	25,000 €
Person 3	25	40,000 €
Person 4	30	50,000 €
Person 5	30	55,000 €
Person 6	28	60,000 €
Person 7	46	65,000 €
Person 8	58	70,000 €
Person 9	62	80,000 €
Person 10	61	100,000 €
Person 11	19	150,000 €

Min	18	5000
Max	62	150000
Mean	36.09	63636
Std Dev	16.54	36748



Normalized (min-max)

	Age	Income
Person 1	0.00	0.00
Person 2	0.05	0.14
Person 3	0.16	0.24
Person 4	0.27	0.31
Person 5	0.27	0.34
Person 6	0.23	0.38
Person 7	0.64	0.41
Person 8	0.91	0.45
Person 9	1.00	0.52
Person 10	0.98	0.66
Person 11	0.02	1.00

Standardize (z Score)

	Age	Income
Person 1	-1.09	-1.60
Person 2	-0.97	-1.05
Person 3	-0.67	-0.64
Person 4	-0.37	-0.37
Person 5	-0.37	-0.24
Person 6	-0.49	-0.10
Person 7	0.60	0.04
Person 8	1.32	0.17
Person 9	1.57	0.45
Person 10	1.51	0.99
Person 11	-1.03	2.35

Checklist

1. **Work on copies of the data**
2. **Write functions for all data transformation you apply**
3. **Clean the data (fix or remove outliers; fill in missing values)**
4. **Select the Features (drop not-needed features)**
5. **Feature Enginering, where necessary**

1. Discretize continuous features
2. Decompose features (e.g. categorical, date/time, etc.)
3. Add promising transformations of features (e.g. log(x); sqrt(x); x^2; …)
4. Aggregate features into promising new features
5. Scale features: Standardize or normalize features

Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Training/Adjusting ML Models

Batch/Offline vs. Incremental Learning

- **Batch/Offline Learning**
  - Training must be performed on *all* data
  - When new data is available, the old and new data must be used
  - Can be mostly automated but still needs a lot of time and resources (CPU, RAM, HDD, …)
- **Incremental Learning / Online learning**
  - New/updated training is done on new data only
  - Updates can be sequentially, or in „mini batches“
  - Data that was used for learning is not needed any more→ can be disposed
  - Good for systems with
    - continuously new data (e.g. stock prices),
    - limited computing resources (e.g. mobile phones)
    - huge amounts of data (too much to fit into memory)
  - Sometimes called „Online learning“, which is misleading

A. Géron, *Hands on Machine Learning with scikit-learn and Tensorflow*. O'Reilly Media, 2017.

Incremental Learning (Illustration)

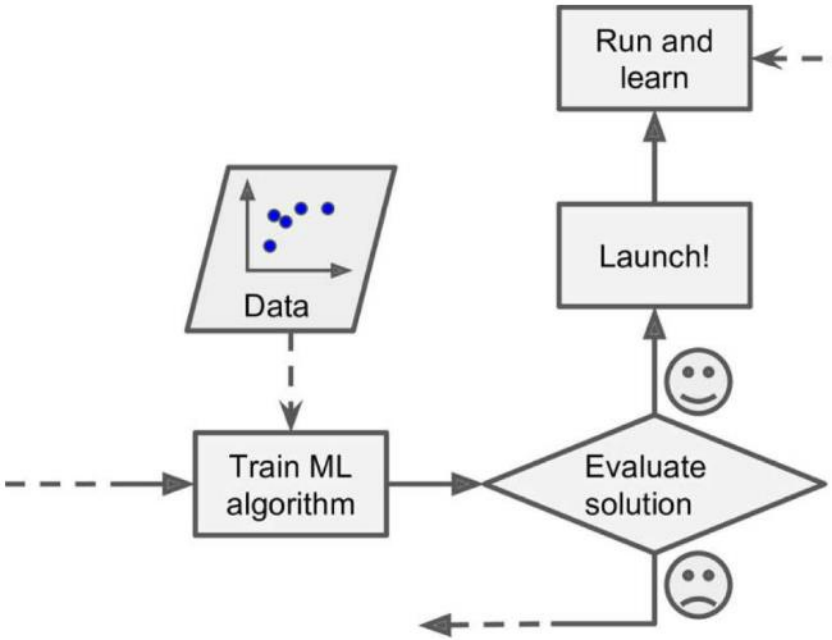


Figure 1-13. Online learning

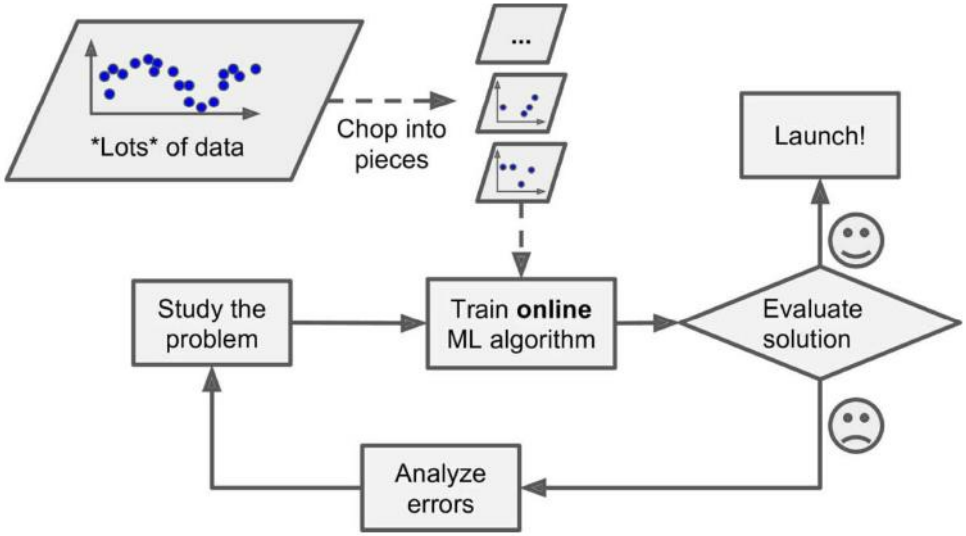


Figure 1-14. Using online learning to handle huge datasets

A. Géron, *Hands on Machine Learning with scikit-learn and Tensorflow*. O'Reilly Media, 2017.

Feature Selection/Removal

The more the better? No

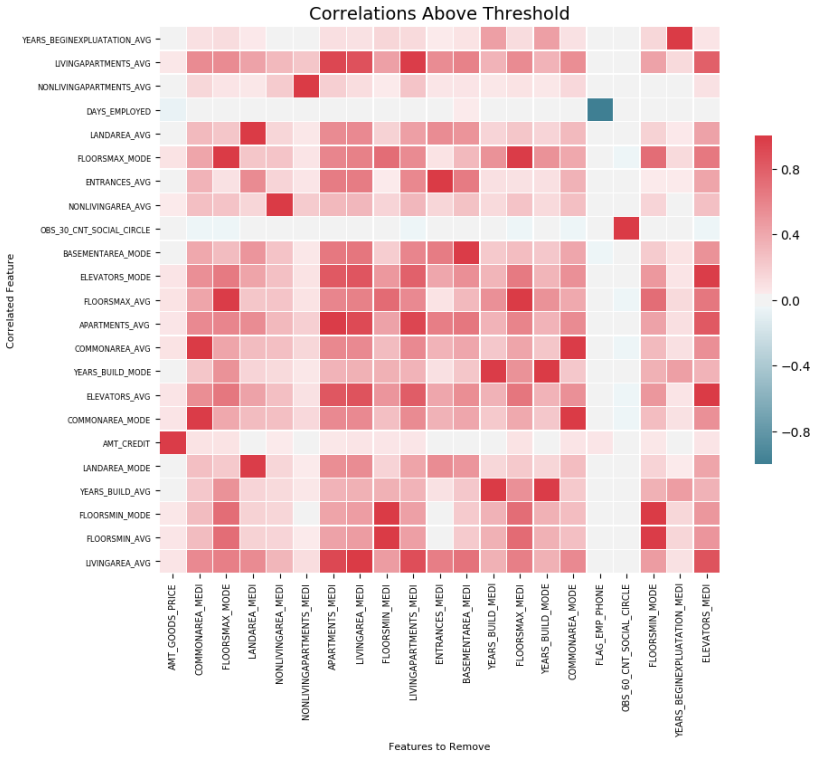
More features → more complexity → more difficult to understand, more memory/CPU needed (i.e. more expensive/slower), higher risk for overfitting (i.e. lower performance eventually)

Common sense

- **Only sensible for obviously irrelevant or redundant features**
- **For instance**
  - Time stamp when data was exported
  - Birthyear *and* age
  - …

Filter Methods / Correlation-based Feature Selection

- **High correlation to the target and low correlation to each other**
- **Potential correlation measures**
  - Pearson's Correlation
  - Linear discriminant analysis
  - ANOVA
  - Chi-Square

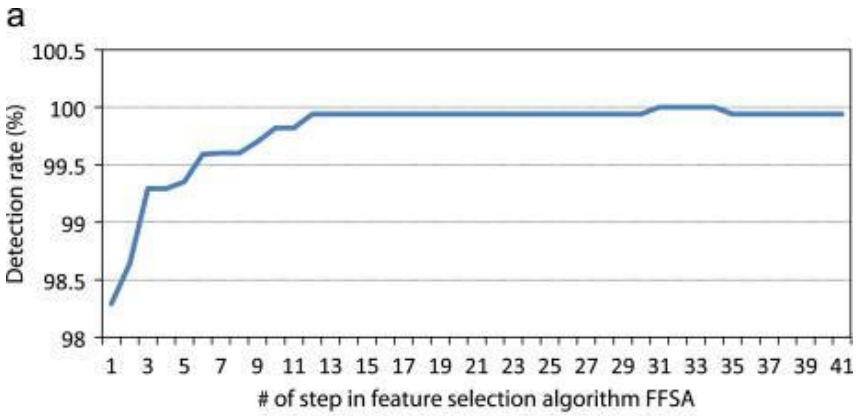


https://towardsdatascience.com/a-feature-selection-tool-for-machine-learning-in-python-b64dd23710f0

Mark Andrew Hall, "Correlation-based feature selection for machine learning" (1999).

Wrapper Methods

- **Methods**
  - Forward Feature Selection
    - Start with 0 features
    - Find and add single-best feature
    - Add next best feature, until performance decreases
  - Backward Feature Elimination
    - Start with all features
    - Find and remove worst performing feature
    - Continue until performance decreases
- **Characteristics**
  - Use Cross Validation
  - Computational Expensive
  - Solves the „real“ problem (optimize model performance)



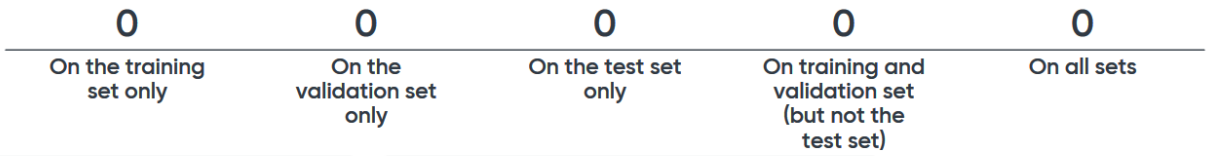
http://www.sciencedirect.com/science/article/pii/S1084804511000038#f0005

On what data would you perform...

Go to **www.menti.com** and use the code **24 94 8**

What data would you use for feature selection with a wrapper method (e.g. forward selection)?

Mentimeter



Results are hidden **Show results**

Slide is not active **Activate**

1

Embedded Methods

- **Maybe later**

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Fine-tune the system

No free lunch theorem

NO FREE LUNCH THEOREM

A model is a simplified version of the observations. The simplifications are meant to discard the superfluous details that are unlikely to generalize to new instances. However, to decide what data to discard and what data to keep, you must make *assumptions*. For example, a linear model makes the assumption that the data is fundamentally linear and that the distance between the instances and the straight line is just noise, which can safely be ignored.

In a famous 1996 paper,[11] David Wolpert demonstrated that if you make absolutely no assumption about the data, then there is no reason to prefer one model over any other. This is called the *No Free Lunch* (NFL) theorem. For some datasets the best model is a linear model, while for other datasets it is a neural network. There is no model that is *a priori* guaranteed to work better (hence the name of the theorem). The only way to know for sure which model is best is to evaluate them all. Since this is not possible, in practice you make some reasonable assumptions about the data and you evaluate only a few reasonable models. For example, for simple tasks you may evaluate linear models with various levels of regularization, and for a complex problem you may evaluate various neural networks.

Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

Hyperparameters

- **Cannot be learned**
- **Are set before the learning starts**
- **For instance, the gradient clipping threshold**
- **Rule of thumb: If a machine-learning framework asks you to specify a parameter manually, then it‘s (probably) a hyperparameter**

Fine-tune the system

1. **Fine-tune the hyperparameters using cross validation**

1. Treat data transformations as hyperparameters when you are not sure about them (e.g. replacing missing values with zero or mean, or just dropping the rows)
2. Unless there are very few hyperparameters, prefer random search over grid search

2. **Try ensemble methods**

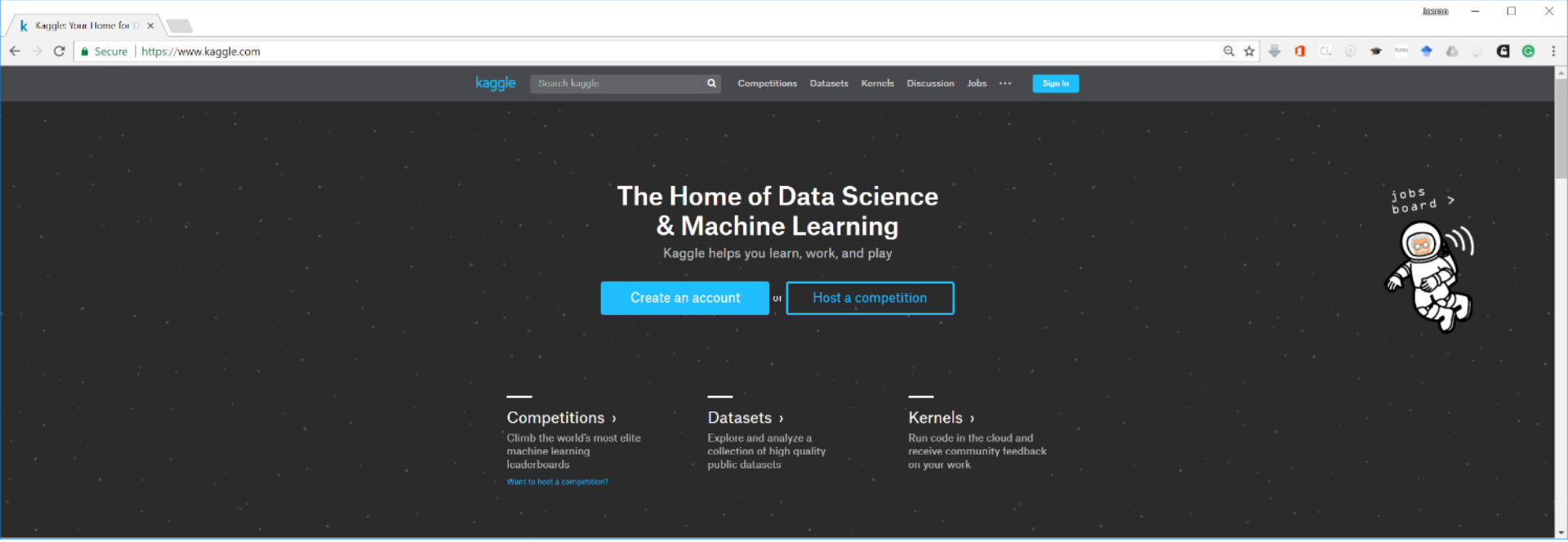
3. **Measure the performance on the test set.**

Aurélien Géron, *Hands on Machine Learning with scikit-learn and Tensorflow* (O'Reilly Media, 2017).

- **Hopefully more in one of the later lectures**

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

General Resources

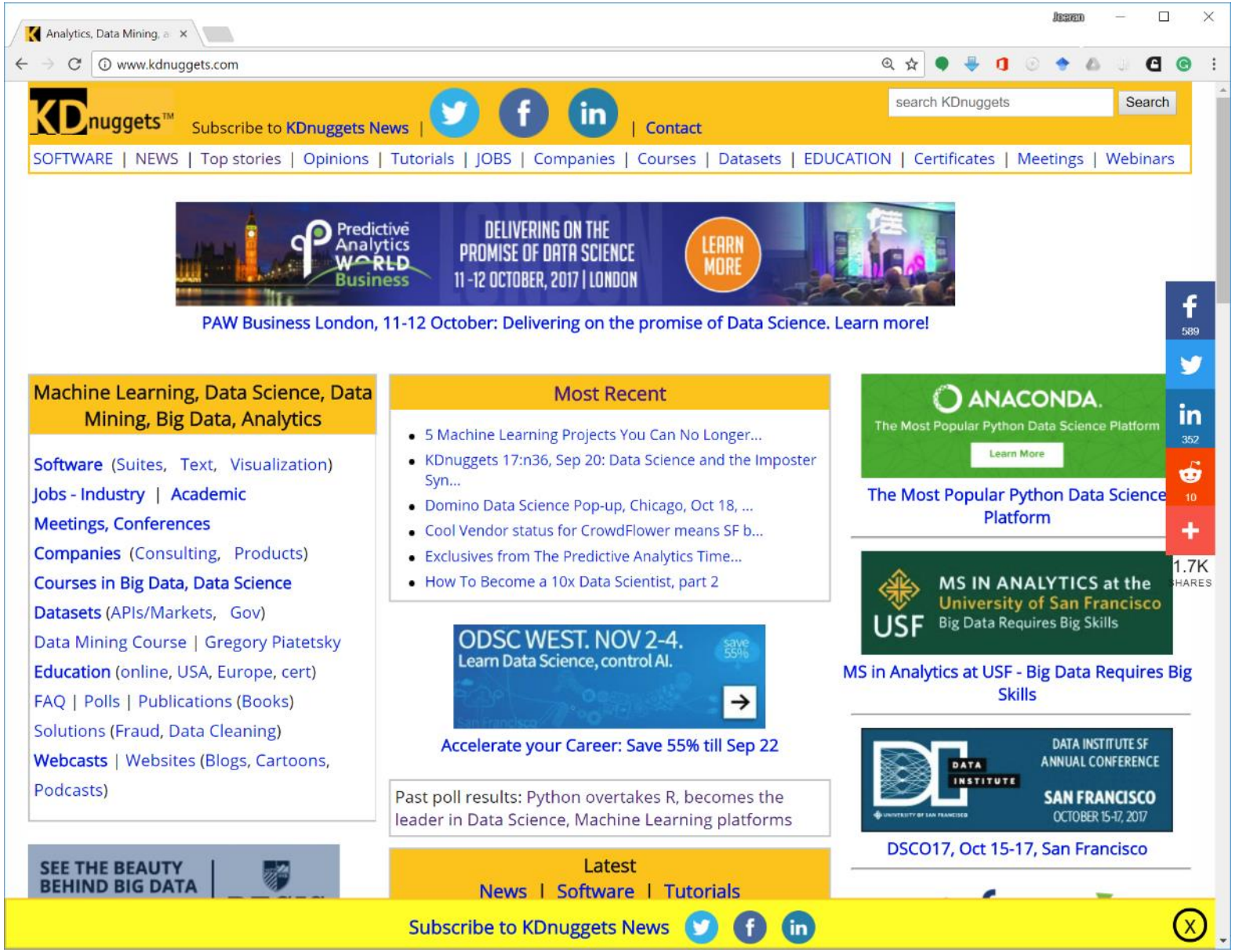


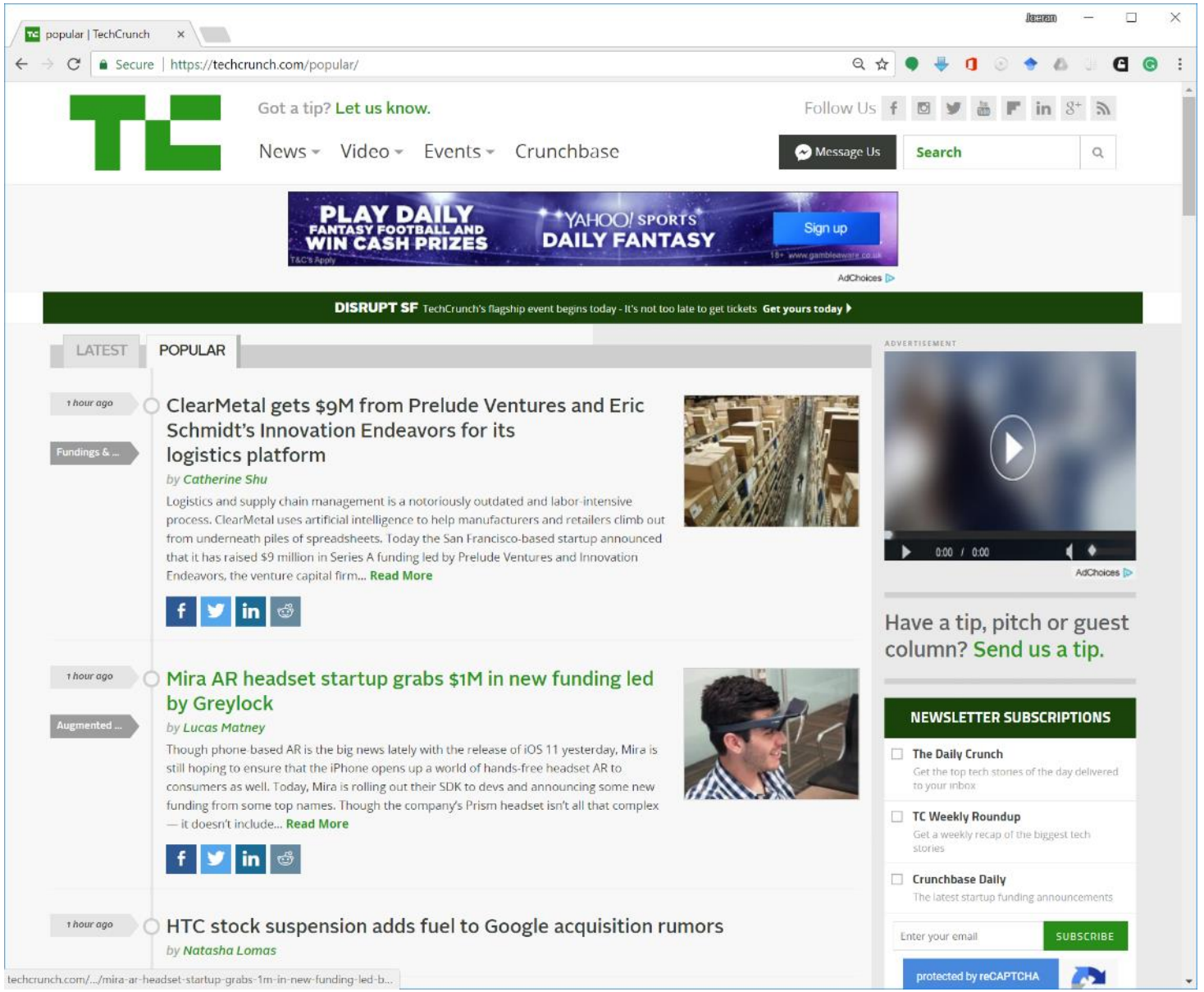
Kaggle

http://kaggle.com

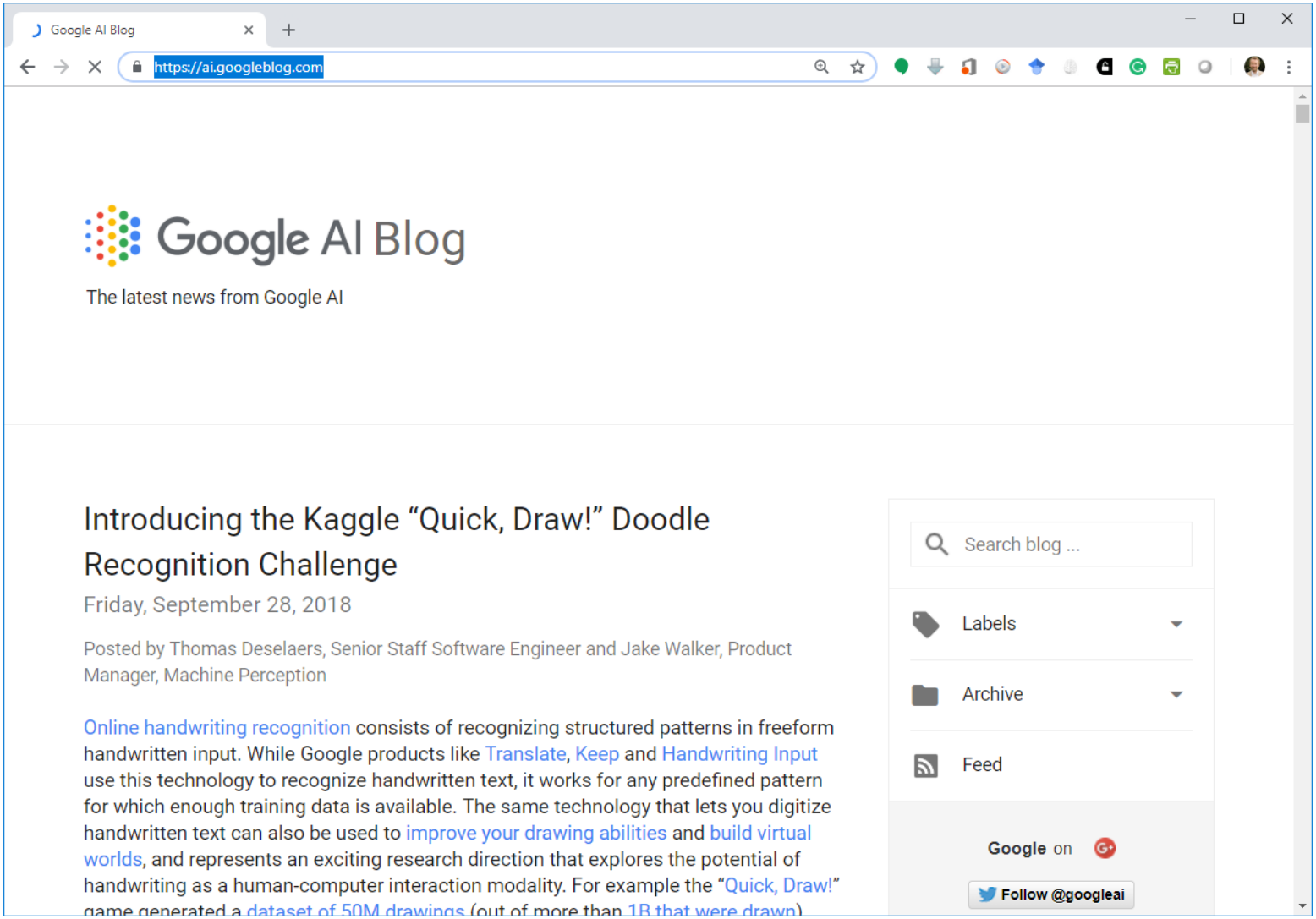
- **News**
- **Competitions**
- **Datasets**
- **Forums**

- **Tutorials/Kernels/Scripts**
  - e.g. https://www.kaggle.com/selfishgene/spatio-temporal-patterns-of-new-york-city

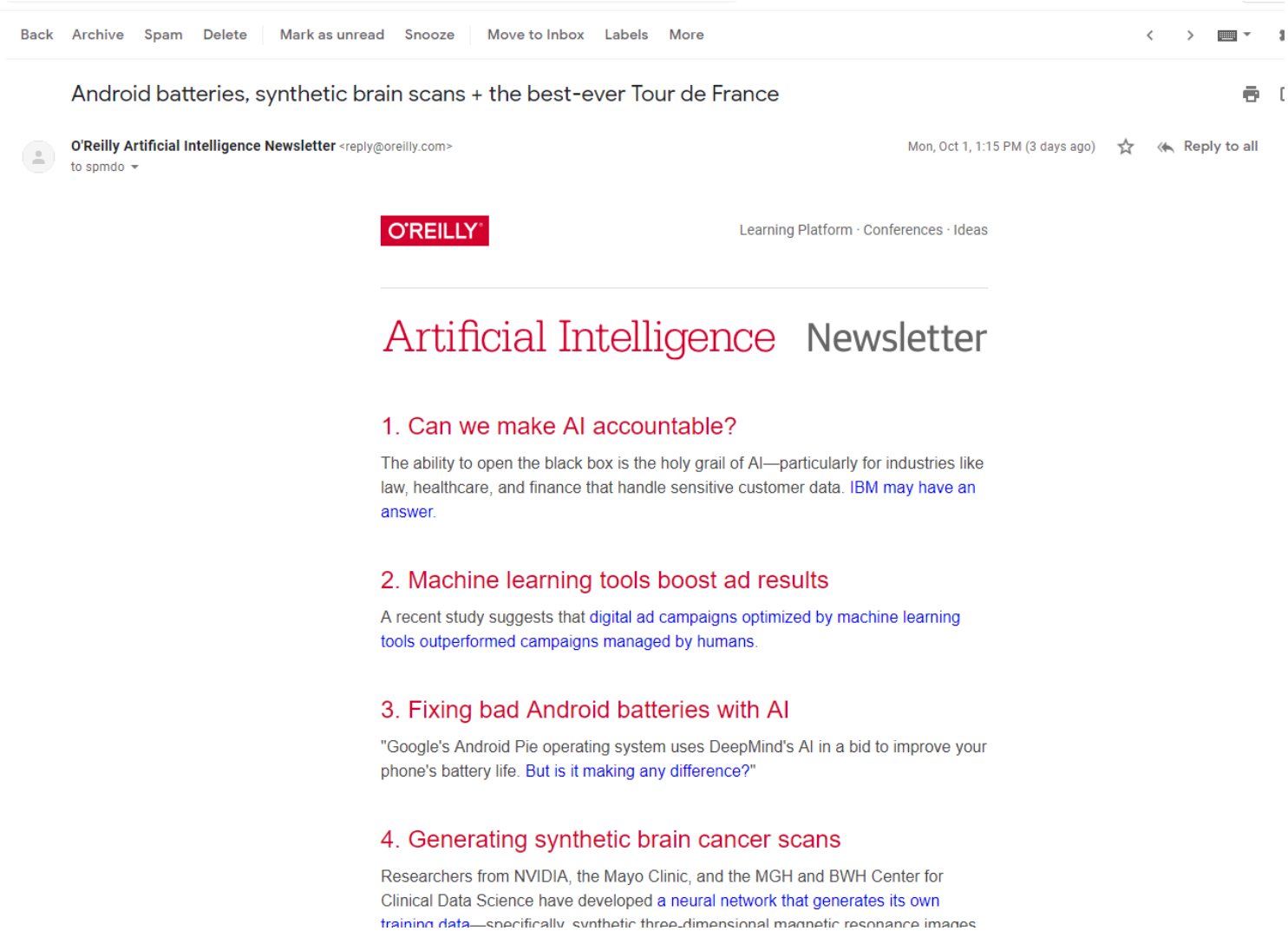

Analytics, Data Mining, a
www.kdnuggets.com
KDnuggets™
Subscribe to KDnuggets News
Contact
search KDnuggets
Search
SOFTWARE | NEWS | Top stories | Opinions | Tutorials | JOBS | Companies | Courses | Datasets | EDUCATION | Certificates | Meetings | Webinars
Predictive Analytics WORLD Business
DELIVERING ON THE PROMISE OF DATA SCIENCE
11-12 OCTOBER, 2017 | LONDON
LEARN MORE
PAW Business London, 11-12 October: Delivering on the promise of Data Science. Learn more!
589
352
10
1.7K SHARES
Machine Learning, Data Science, Data Mining, Big Data, Analytics
Software (Suites, Text, Visualization)
Jobs - Industry | Academic
Meetings, Conferences
Companies (Consulting, Products)
Courses in Big Data, Data Science
Datasets (APIs/Markets, Gov)
Data Mining Course | Gregory Piatetsky
Education (online, USA, Europe, cert)
FAQ | Polls | Publications (Books)
Solutions (Fraud, Data Cleaning)
Webcasts | Websites (Blogs, Cartoons, Podcasts)
Most Recent
5 Machine Learning Projects You Can No Longer...
KDnuggets 17:n36, Sep 20: Data Science and the Imposter Syn...
Domino Data Science Pop-up, Chicago, Oct 18, ...
Cool Vendor status for CrowdFlower means SF b...
Exclusives from The Predictive Analytics Time...
How To Become a 10x Data Scientist, part 2
ANACONDA.
The Most Popular Python Data Science Platform
Learn More
The Most Popular Python Data Science Platform
ODSC WEST. NOV 2-4.
Learn Data Science, control AI.
save 55%
Accelerate your Career: Save 55% till Sep 22
USF
MS IN ANALYTICS at the University of San Francisco
Big Data Requires Big Skills
MS in Analytics at USF - Big Data Requires Big Skills
DATA INSTITUTE
DATA INSTITUTE SF ANNUAL CONFERENCE
SAN FRANCISCO
OCTOBER 15-17, 2017
DSCO17, Oct 15-17, San Francisco
Past poll results: Python overtakes R, becomes the leader in Data Science, Machine Learning platforms
SEE THE BEAUTY BEHIND BIG DATA
Latest
News | Software | Tutorials
Subscribe to KDnuggets News


popular | TechCrunch
Secure | https://techcrunch.com/popular/
Got a tip? Let us know.
Follow Us
News
Video
Events
Crunchbase
Message Us
Search
PLAY DAILY FANTASY FOOTBALL AND WIN CASH PRIZES
YAHOO! SPORTS DAILY FANTASY
Sign up
T&C's Apply
18+ www.gambleaware.co.uk
AdChoices
DISRUPT SF TechCrunch's flagship event begins today - It's not too late to get tickets Get yours today
LATEST
POPULAR
1 hour ago
Fundings & ...
ClearMetal gets $9M from Prelude Ventures and Eric Schmidt's Innovation Endeavors for its logistics platform
by Catherine Shu
Logistics and supply chain management is a notoriously outdated and labor-intensive process. ClearMetal uses artificial intelligence to help manufacturers and retailers climb out from underneath piles of spreadsheets. Today the San Francisco-based startup announced that it has raised $9 million in Series A funding led by Prelude Ventures and Innovation Endeavors, the venture capital firm... Read More
1 hour ago
Augmented ...
Mira AR headset startup grabs $1M in new funding led by Greylock
by Lucas Matney
Though phone-based AR is the big news lately with the release of iOS 11 yesterday, Mira is still hoping to ensure that the iPhone opens up a world of hands-free headset AR to consumers as well. Today, Mira is rolling out their SDK to devs and announcing some new funding from some top names. Though the company's Prism headset isn't all that complex — it doesn't include... Read More
1 hour ago
HTC stock suspension adds fuel to Google acquisition rumors
by Natasha Lomas
techcrunch.com/.../mira-ar-headset-startup-grabs-1m-in-new-funding-led-b...
ADVERTISEMENT
0:00 / 0:00
AdChoices
Have a tip, pitch or guest column? Send us a tip.
NEWSLETTER SUBSCRIPTIONS
The Daily Crunch
Get the top tech stories of the day delivered to your inbox
TC Weekly Roundup
Get a weekly recap of the biggest tech stories
Crunchbase Daily
The latest startup funding announcements
Enter your email
SUBSCRIBE
protected by reCAPTCHA

https://ai.googleblog.com/



O´Reilly AI Newsletter



Back Archive Spam Delete Mark as unread Snooze Move to Inbox Labels More

Android batteries, synthetic brain scans + the best-ever Tour de France

O'Reilly Artificial Intelligence Newsletter <reply@oreilly.com>
to spmdo

Mon, Oct 1, 1:15 PM (3 days ago) Reply to all

O'REILLY

Learning Platform · Conferences · Ideas

Artificial Intelligence Newsletter

1. Can we make AI accountable?

The ability to open the black box is the holy grail of AI—particularly for industries like law, healthcare, and finance that handle sensitive customer data. IBM may have an answer.

2. Machine learning tools boost ad results

A recent study suggests that digital ad campaigns optimized by machine learning tools outperformed campaigns managed by humans.

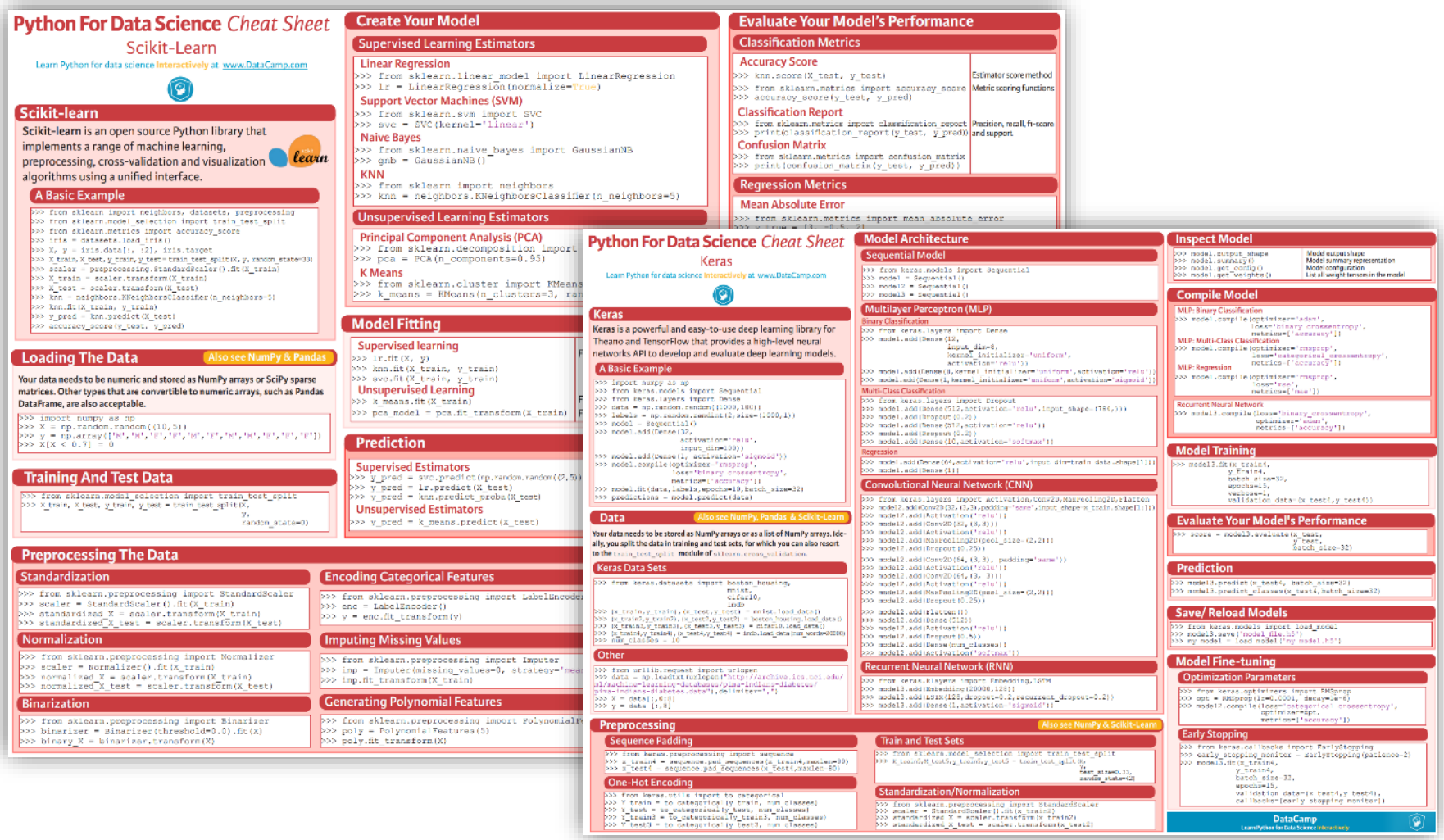
3. Fixing bad Android batteries with AI

"Google's Android Pie operating system uses DeepMind's AI in a bid to improve your phone's battery life. But is it making any difference?"

4. Generating synthetic brain cancer scans

Researchers from NVIDIA, the Mayo Clinic, and the MGH and BWH Center for Clinical Data Science have developed a neural network that generates its own training data—specifically, synthetic three-dimensional magnetic resonance images

„Cheat Sheets“



http://www.kdnuggets.com/2017/09/essential-data-science-machine-learning-deep-learning-cheat-sheets.html

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Machine-Learning Frameworks

a.k.a Machine-Learning Libraries

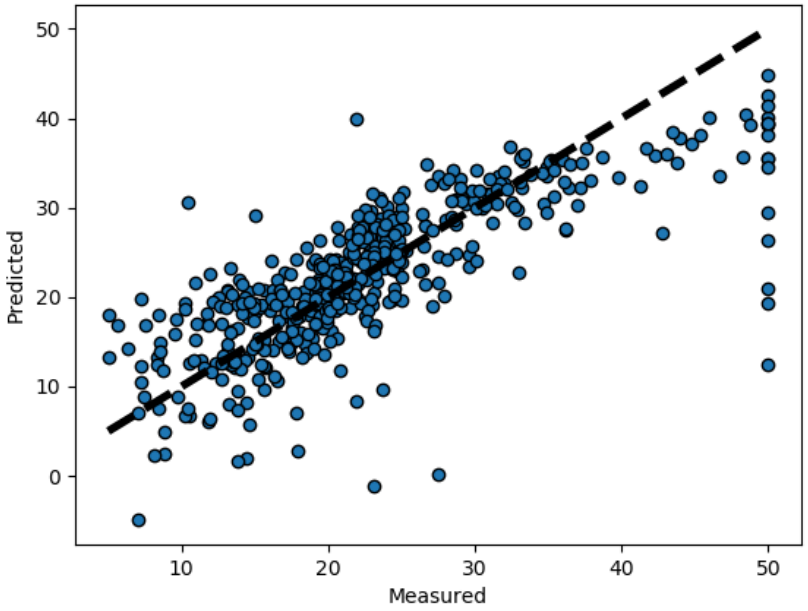
(Some) Popular Machine-Learning Frameworks

- **"Traditional" Machine Learning**
  - Scikit-learn (Python)
  - Weka (JAVA)
  - ML Pack (C++)
  - Shogun (C#, Python, Java, …)
  - Apache Mahout
  - Apache Singa (also deep learning)
  - Apache MLlib (Spark)
  - R… https://www.r-bloggers.com/what-are-the-best-machine-learning-packages-in-r/
- **Deep Learning**
  - TensorFlow by Google (Python)
  - Theano (Python)
  - Caffee(2Go)
  - Torch & pyTorch
  - MXNet (Python, R, Scala, …)

Using the Frameworks

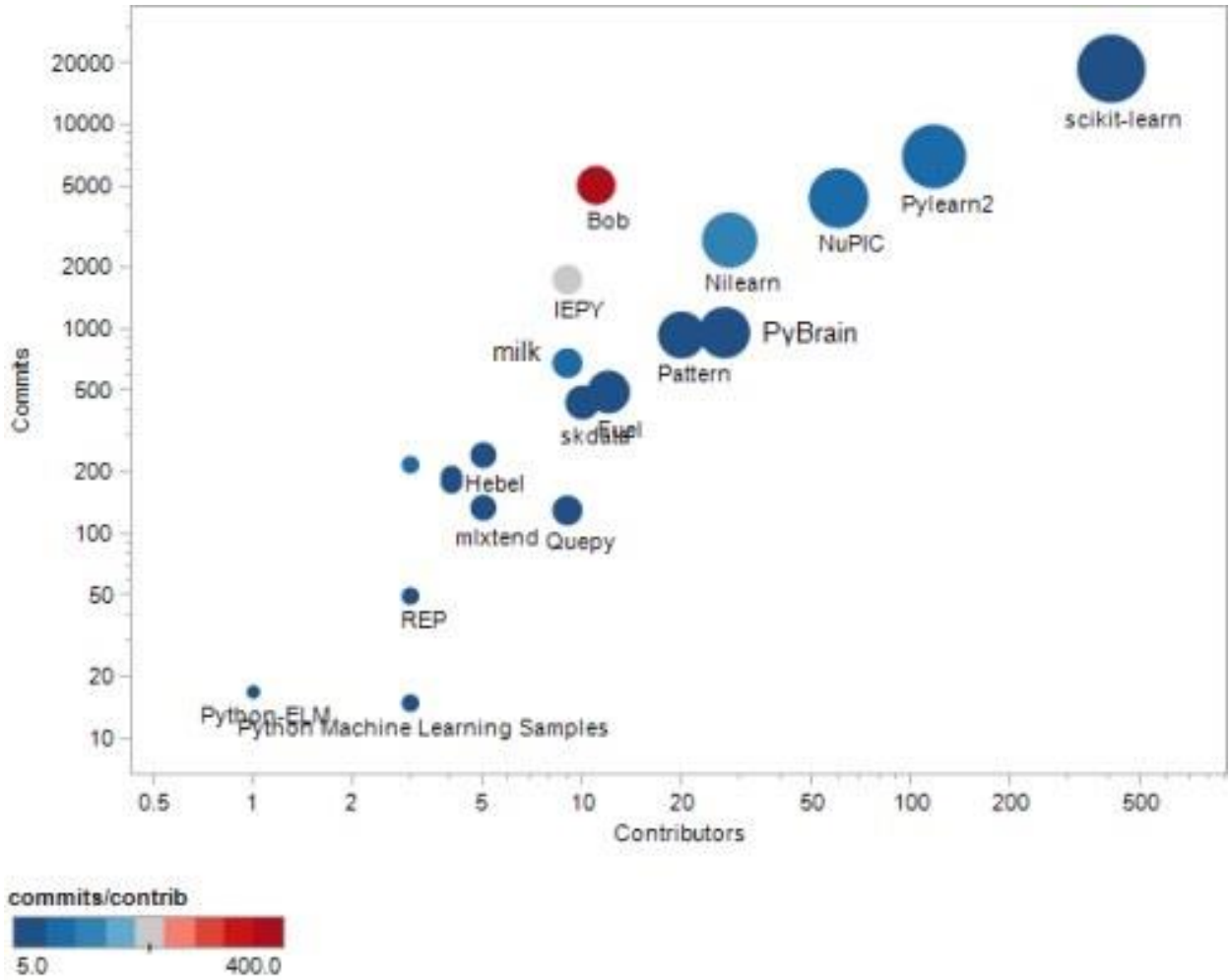
Example

```
from sklearn import datasets
from sklearn.model_selection import cross_val_predict
from sklearn import linear_model
import matplotlib.pyplot as plt

lr = linear_model.LinearRegression()
boston = datasets.load_boston()
y = boston.target

# cross_val_predict returns an array of the same size as
# `y` where each entry is a prediction obtained by cross
# validation:
predicted = cross_val_predict(lr, boston.data, y, cv=10)

fig, ax = plt.subplots()
ax.scatter(y, predicted, edgecolors=(0, 0, 0))
ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw=4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.show()
```

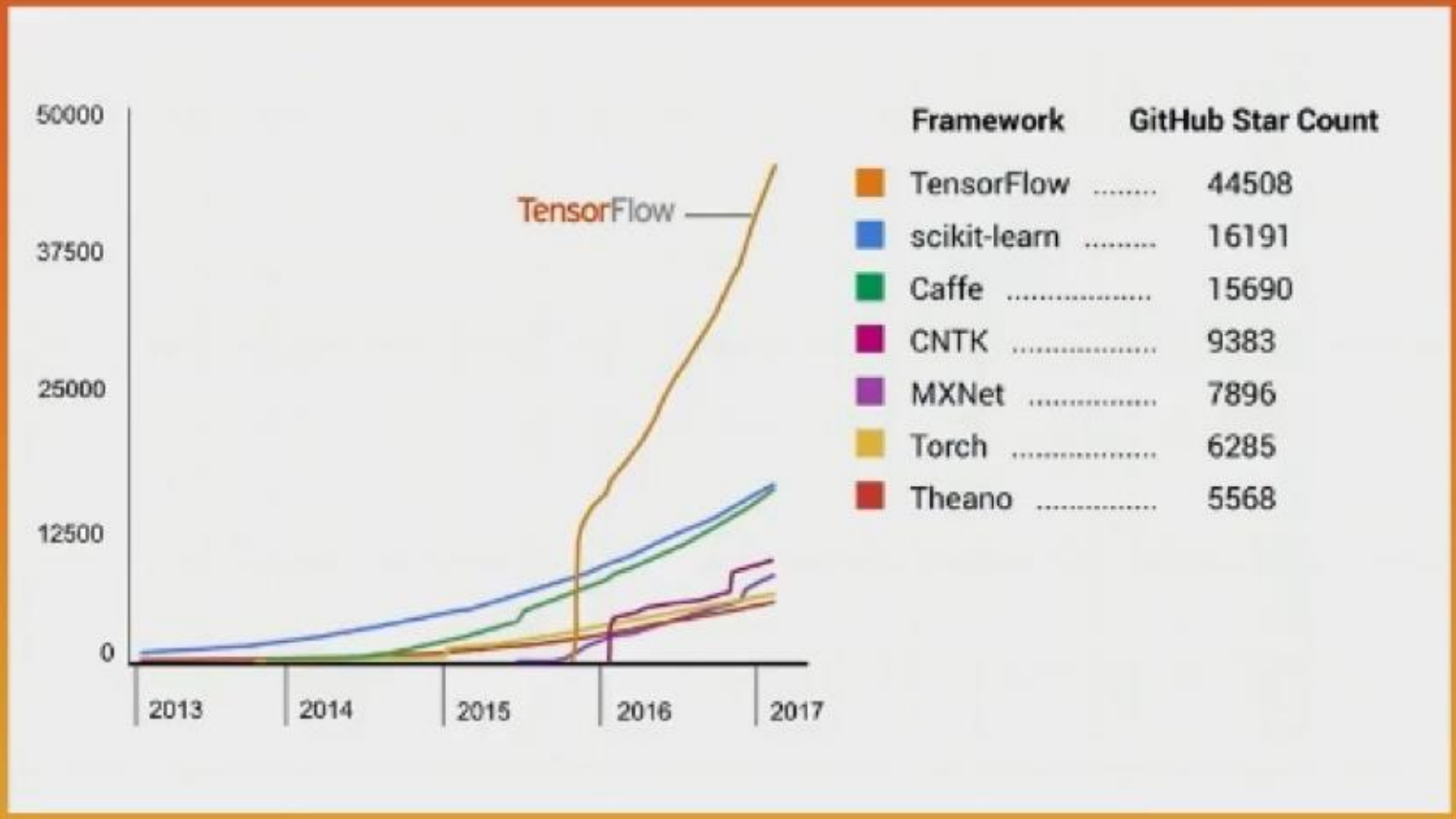


http://scikit-learn.org/stable/auto_examples/plot_cv_predict.html

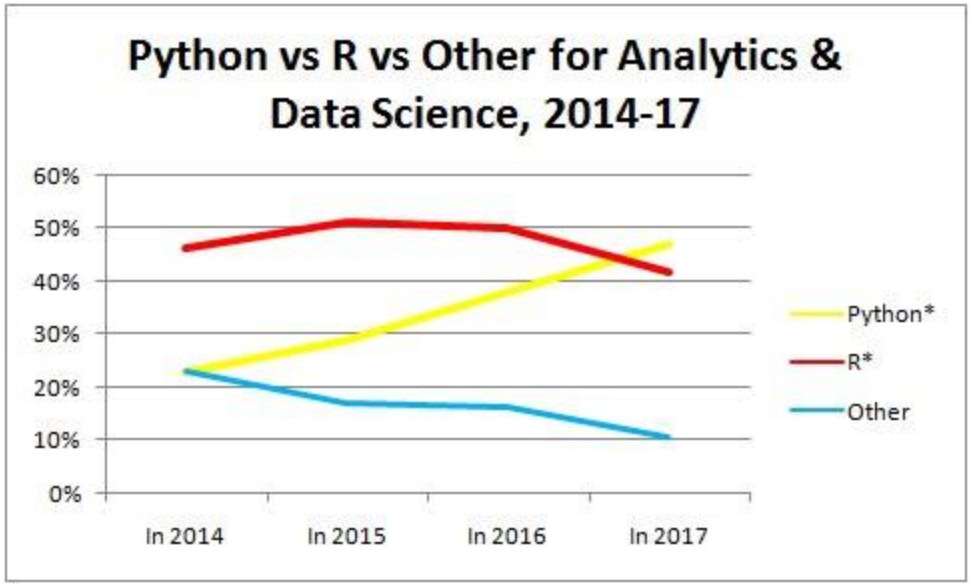
Popularity of Python Open-Source Frameworks



http://www.kdnuggets.com/2015/06/top-20-python-machine-learning-open-source-projects.html



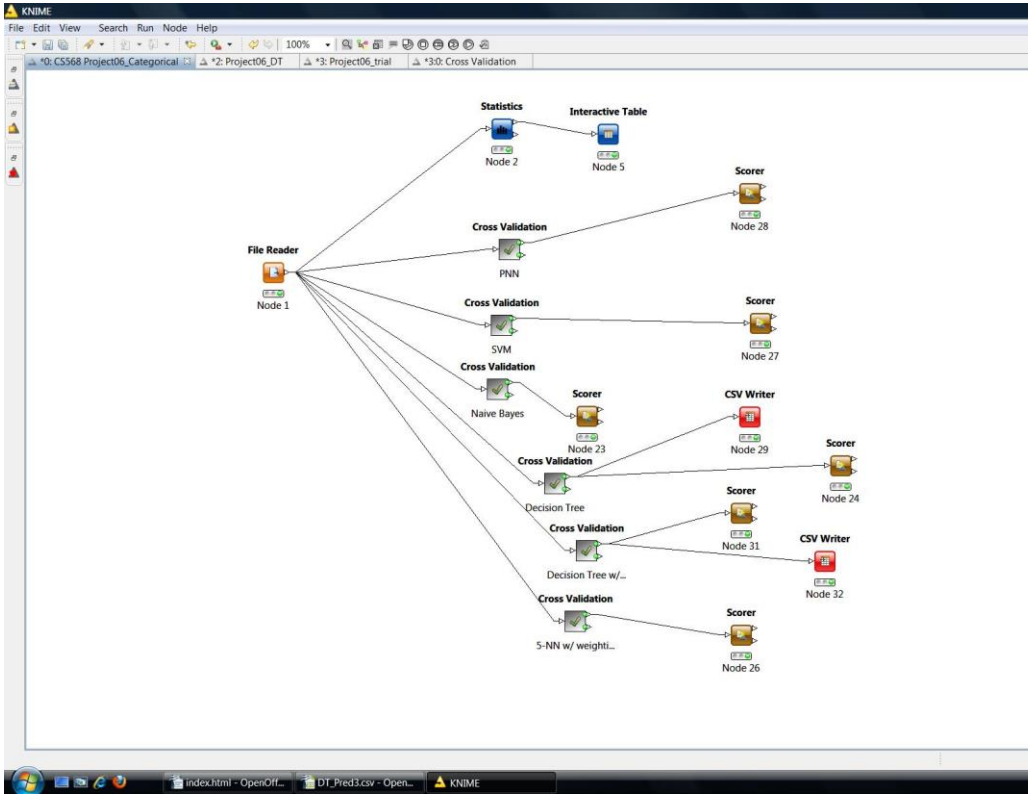
https://www.slideshare.net/jiangjun1990/gdgshanghai-2017-tensorflow-summit-recap



http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

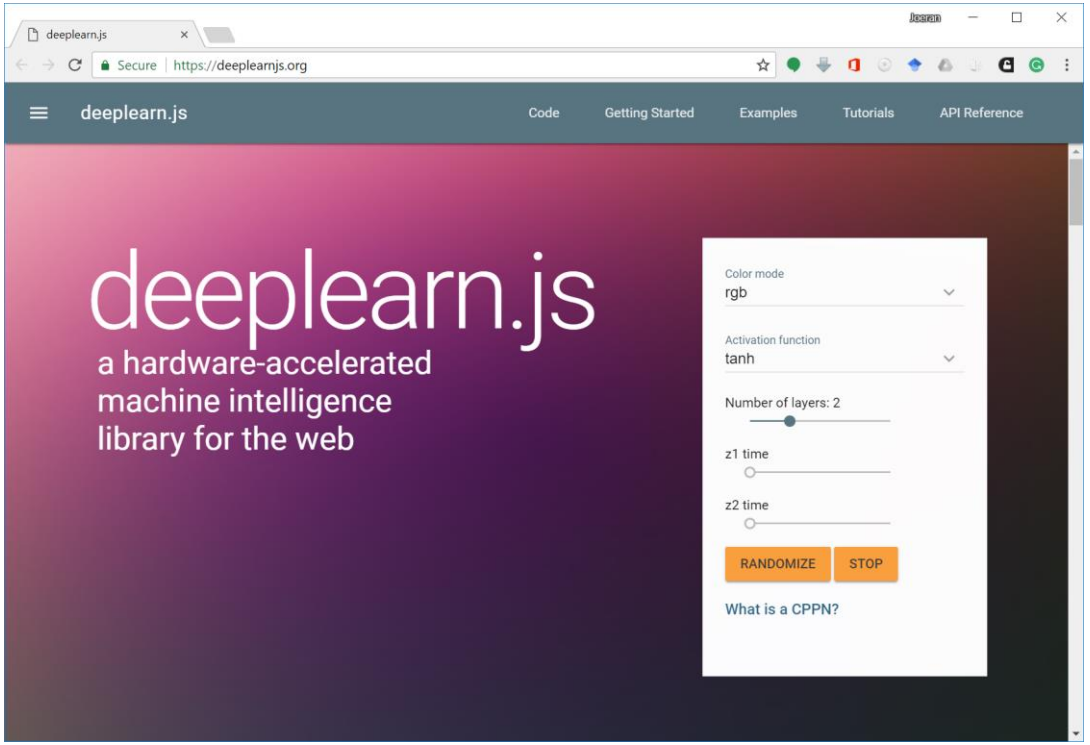
GUIs

- **Design the entire workflow in a GUI**
- **No coding skills need**
- **(Some) ML knowledge still needed**
- **Tools**
  - Rapidminer
  - Knime
  - …



JavaScript Frameworks

- http://cs.stanford.edu/people/karpathy/convnetjs/
- https://pair-code.github.io/deeplearnjs/

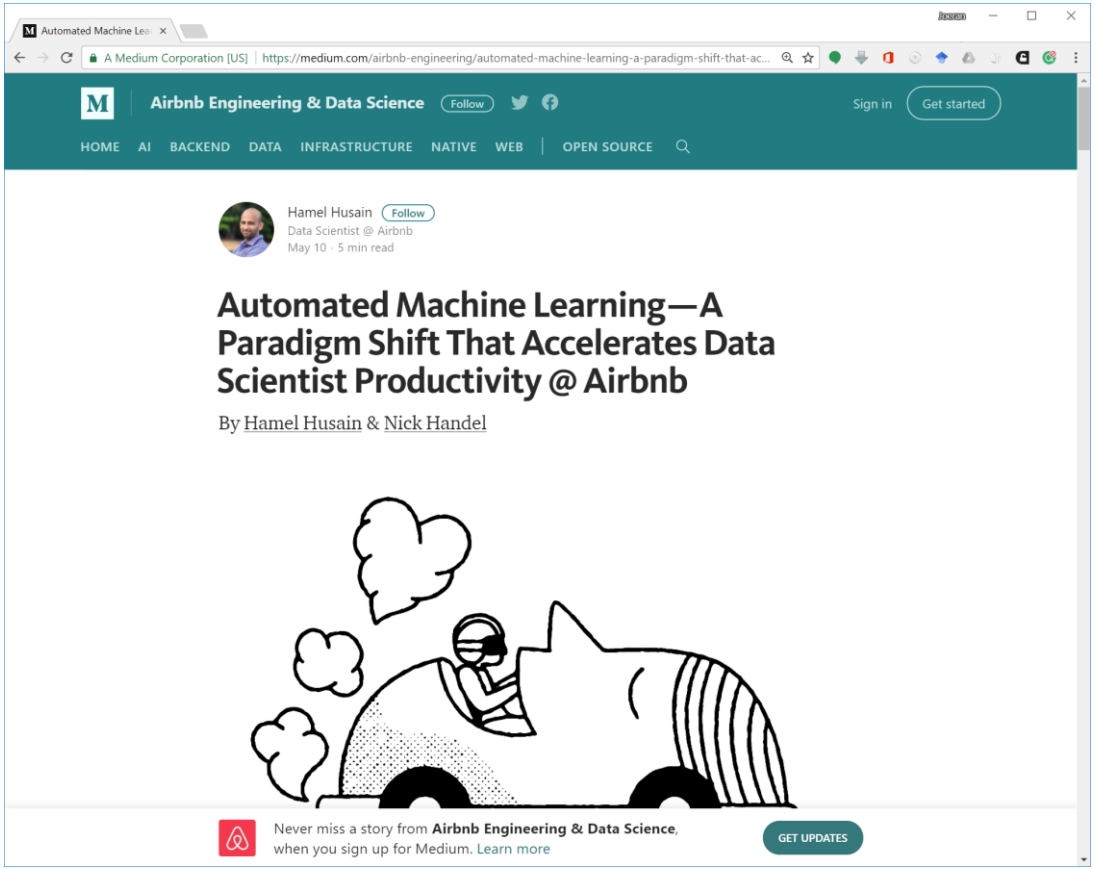


Cloud Frameworks / APIs

- **Microsoft‘s Azure Machine Learning**
- **Google Cloud Prediction/ML/GPU API**
- **Amazon Machine Learning**
- **Nvidea GPU Cloud**
- **…**

Automated Machine Learning

- **Automate the entire process of Machine Learning**
- **Frameworks**
  - TPOT
  - auto-sklearn
  - Auto-Weka
  - machineJS
  - DataRobot
  - …



https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8

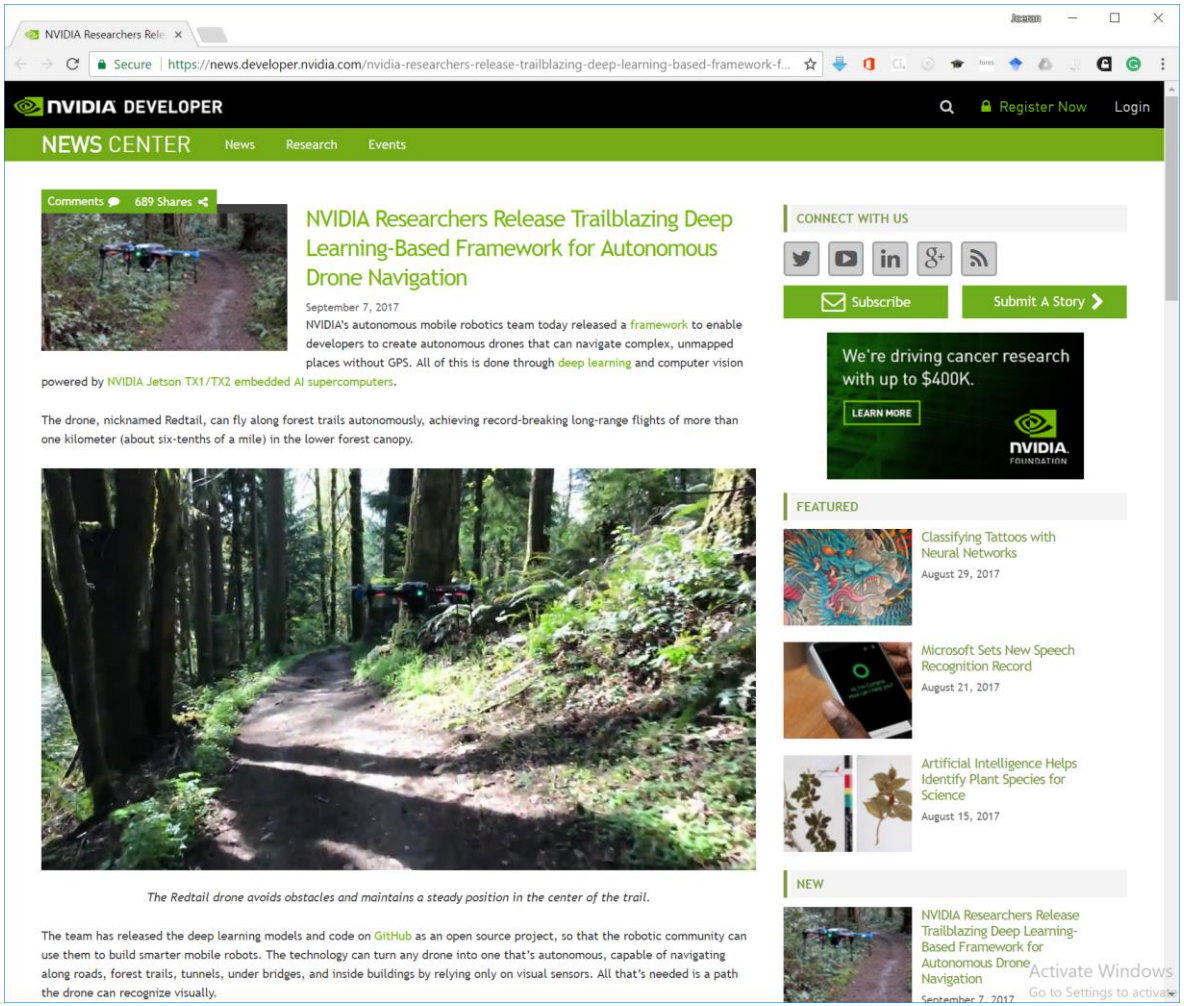
More…

- **https://www.techleer.com/articles/434-apple-to-simplify-machine-learning-development-by-turi-create/**

Domain Specific Frameworks

NVIDIA Redtail project: Autonomous Drone Navigation

https://github.com/NVIDIA-Jetson/redtail/



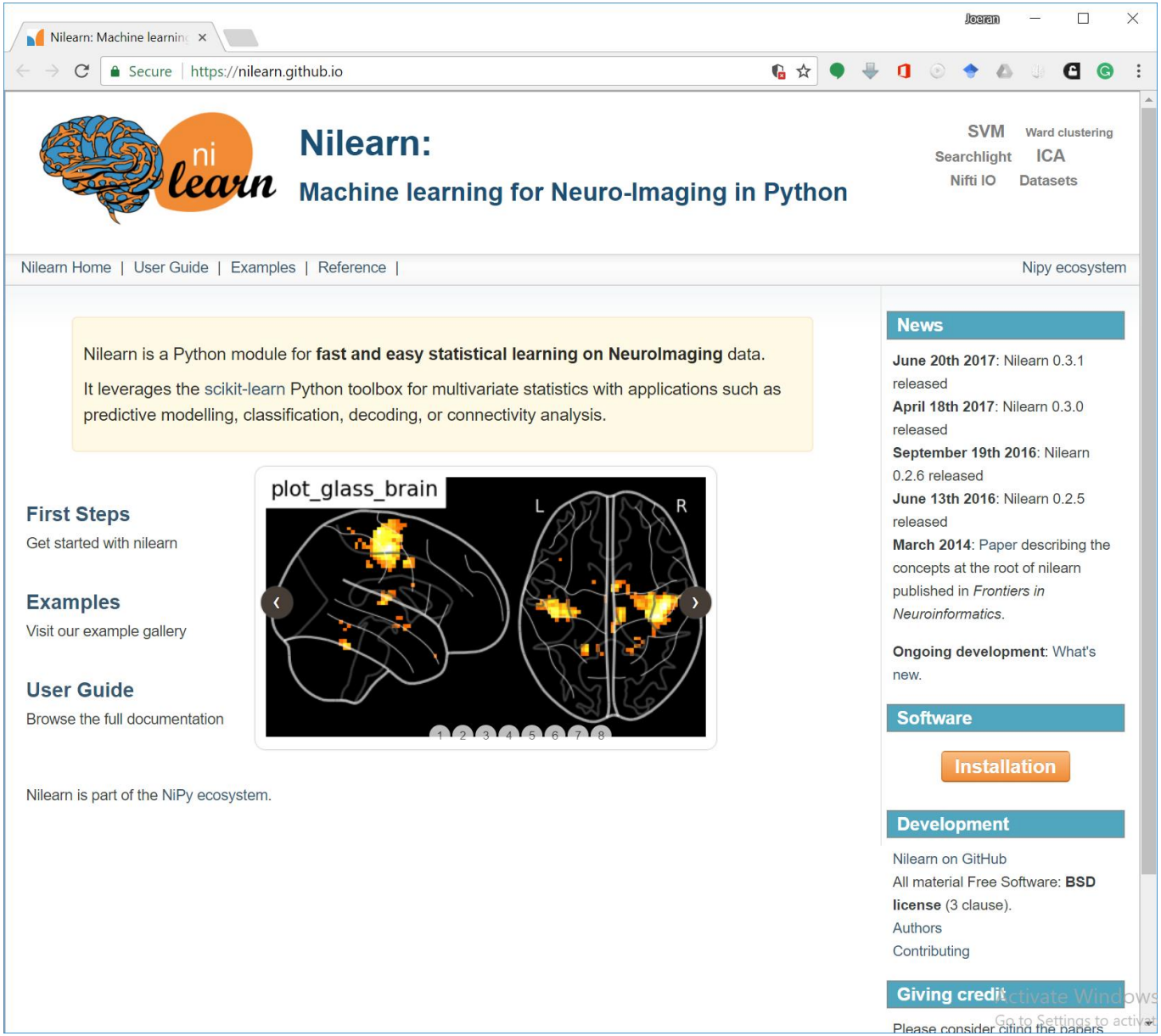
https://news.developer.nvidia.com/nvidia-researchers-release-trailblazing-deep-learning-based-framework-for-autonomous-drone-navigation/

VectorFlow (Netflix)

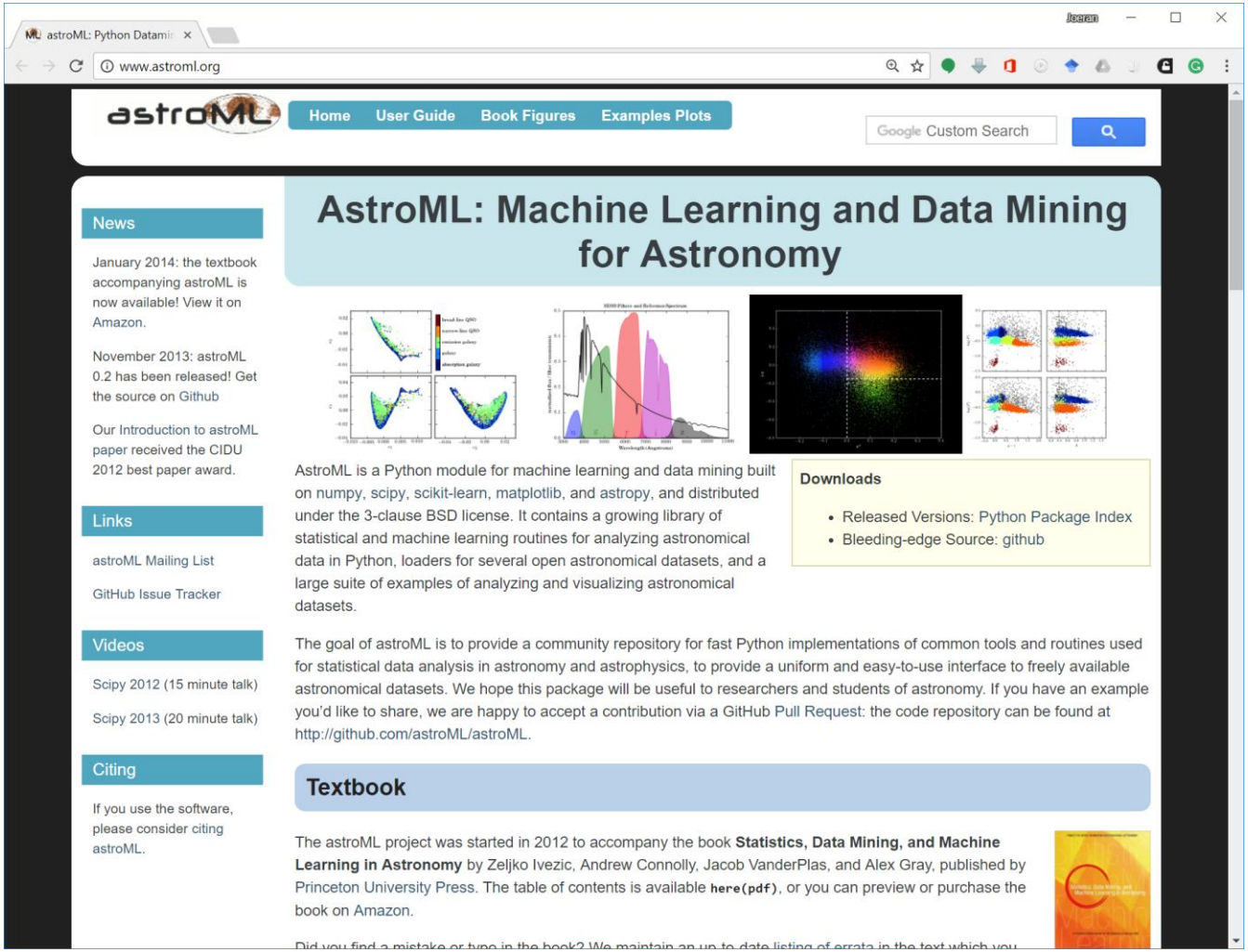
- **“Very very” sparse data (millions of features; some having only a dozens of entries)**
- **Released in July 2017**

https://medium.com/@NetflixTechBlog/introducing-vectorflow-fe10d7f126b8
https://github.com/Netflix/vectorflow

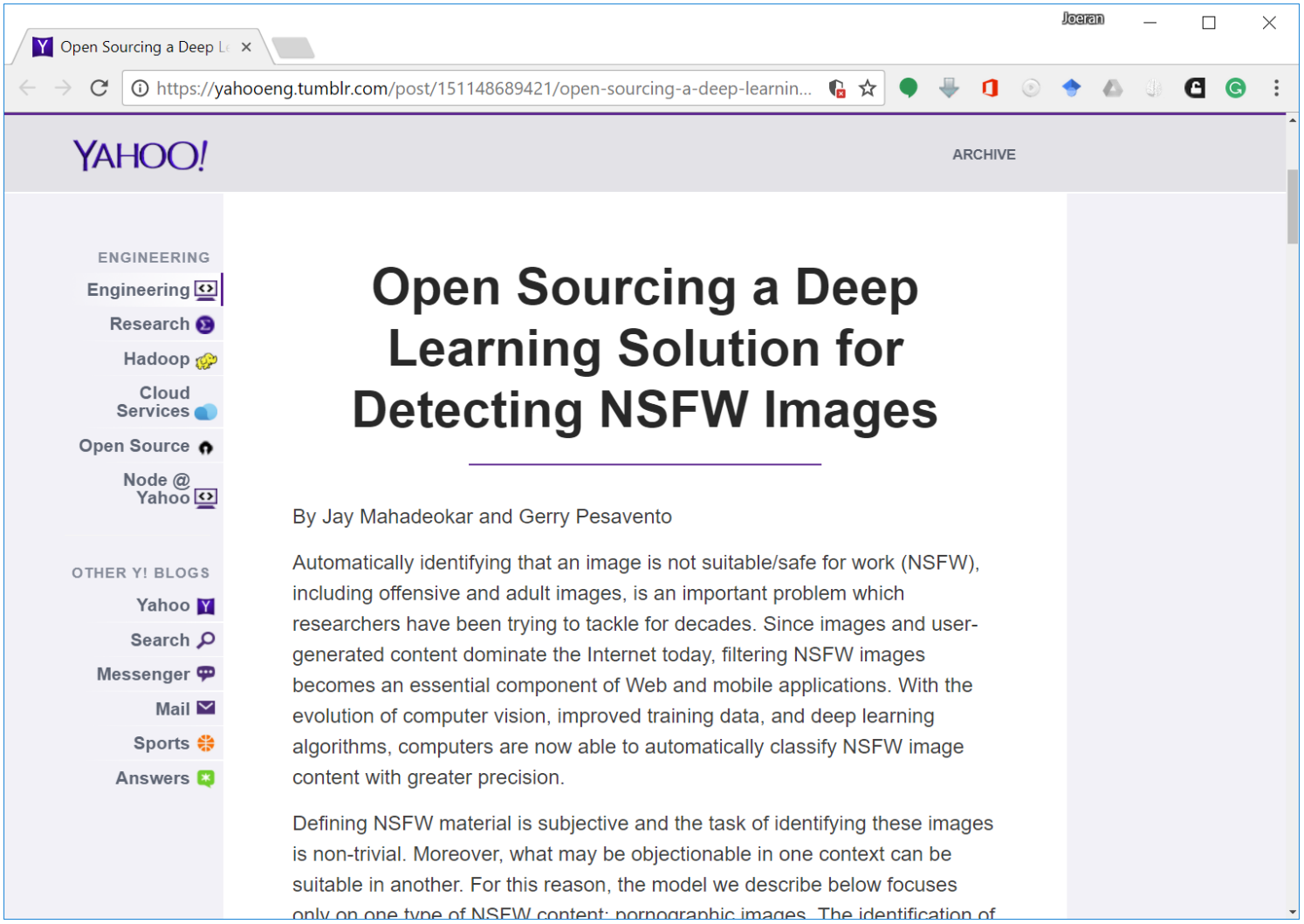
Nilearn


Nilearn: Machine learning
Secure
https://nilearn.github.io
Nilearn:
Machine learning for Neuro-Imaging in Python
SVM
Ward clustering
Searchlight
ICA
Nifti IO
Datasets
Nilearn Home | User Guide | Examples | Reference |
Nipy ecosystem
Nilearn is a Python module for fast and easy statistical learning on NeuroImaging data.
It leverages the scikit-learn Python toolbox for multivariate statistics with applications such as predictive modelling, classification, decoding, or connectivity analysis.
First Steps
Get started with nilearn
Examples
Visit our example gallery
User Guide
Browse the full documentation
plot_glass_brain
L
R
1 2 3 4 5 6 7 8
Nilearn is part of the NiPy ecosystem.
News
June 20th 2017: Nilearn 0.3.1 released
April 18th 2017: Nilearn 0.3.0 released
September 19th 2016: Nilearn 0.2.6 released
June 13th 2016: Nilearn 0.2.5 released
March 2014: Paper describing the concepts at the root of nilearn published in Frontiers in Neuroinformatics.
Ongoing development: What's new.
Software
Installation
Development
Nilearn on GitHub
All material Free Software: BSD license (3 clause).
Authors
Contributing
Giving credit
Please consider citing the papers

AstroML

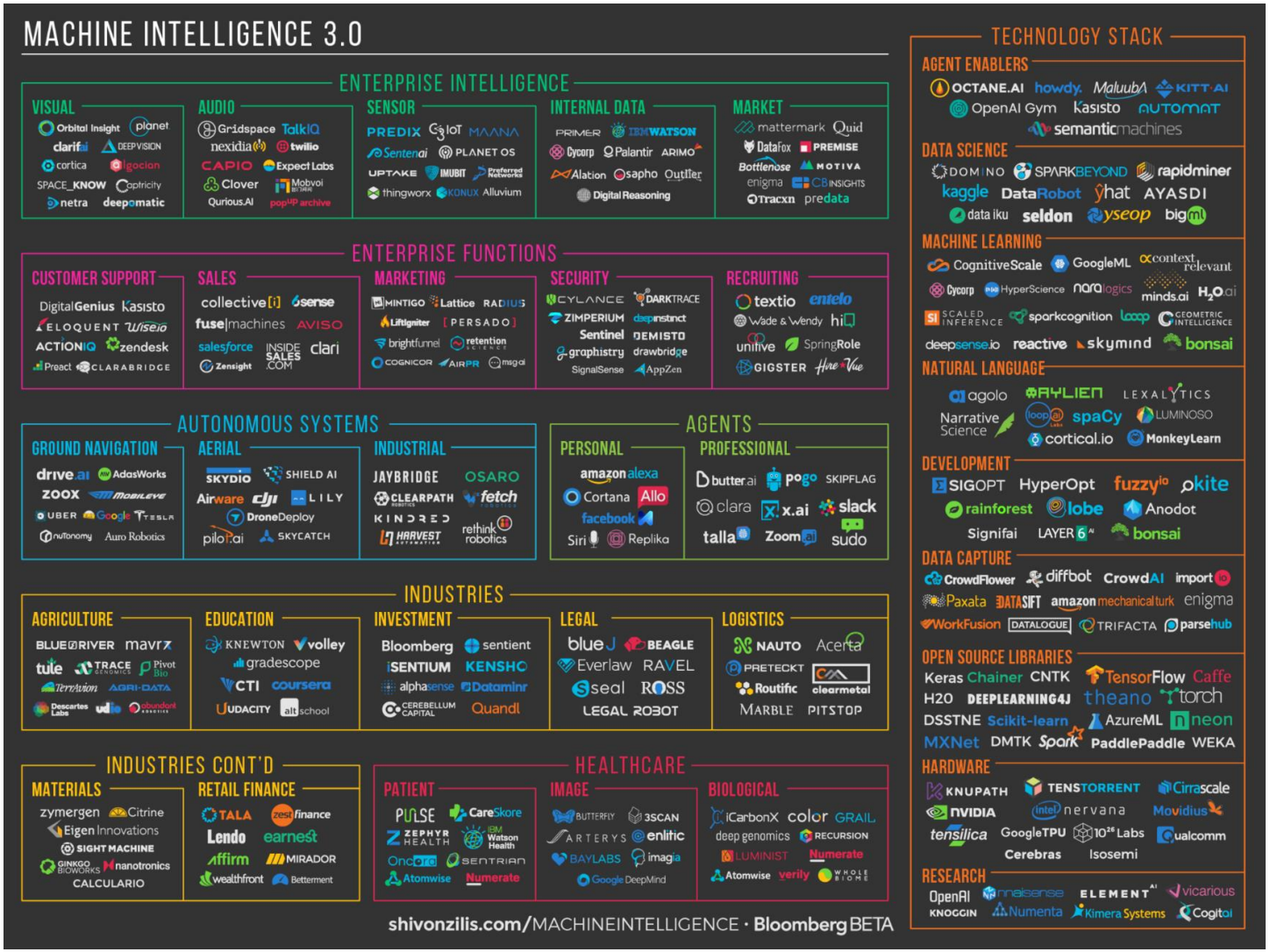

astroML: Python Datami
www.astroml.org
astroML
Home
User Guide
Book Figures
Examples Plots
Google Custom Search
News
January 2014: the textbook accompanying astroML is now available! View it on Amazon.
November 2013: astroML 0.2 has been released! Get the source on Github
Our Introduction to astroML paper received the CIDU 2012 best paper award.
Links
astroML Mailing List
GitHub Issue Tracker
Videos
Scipy 2012 (15 minute talk)
Scipy 2013 (20 minute talk)
Citing
If you use the software, please consider citing astroML.
AstroML: Machine Learning and Data Mining for Astronomy
AstroML is a Python module for machine learning and data mining built on numpy, scipy, scikit-learn, matplotlib, and astropy, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in Python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.
Downloads
Released Versions: Python Package Index
Bleeding-edge Source: github
The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. If you have an example you'd like to share, we are happy to accept a contribution via a GitHub Pull Request: the code repository can be found at http://github.com/astroML/astroML.
Textbook
The astroML project was started in 2012 to accompany the book Statistics, Data Mining, and Machine Learning in Astronomy by Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alex Gray, published by Princeton University Press. The table of contents is available here(pdf), or you can preview or purchase the book on Amazon.

“Not Suitable/Safe For Work” (NSFW) Images



https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for

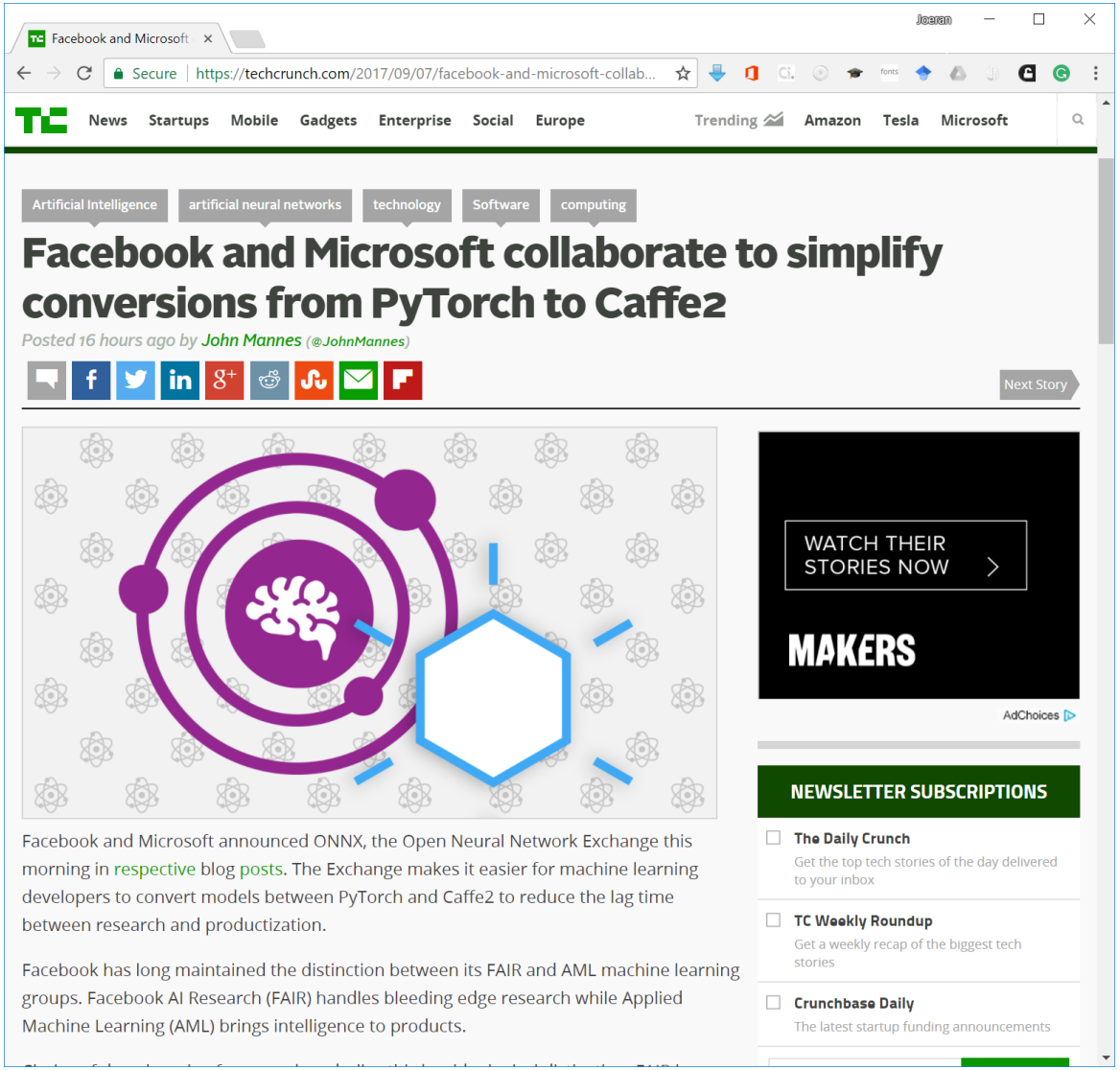
And much more…



https://format-com-cld-res.cloudinary.com/image/private/s--RCb7PzQR--/c_crop,h_1500,w_2000,x_0,y_0/c_fill,g_center,h_855,w_1140/a_auto,dpr_2,fl_keep_iptc.progressive,q_95/v1/19575bcc040a6dcff3097618ec9c585e/MI-Landscape-3_7.png

Onnx: Open Neural Network Exchange

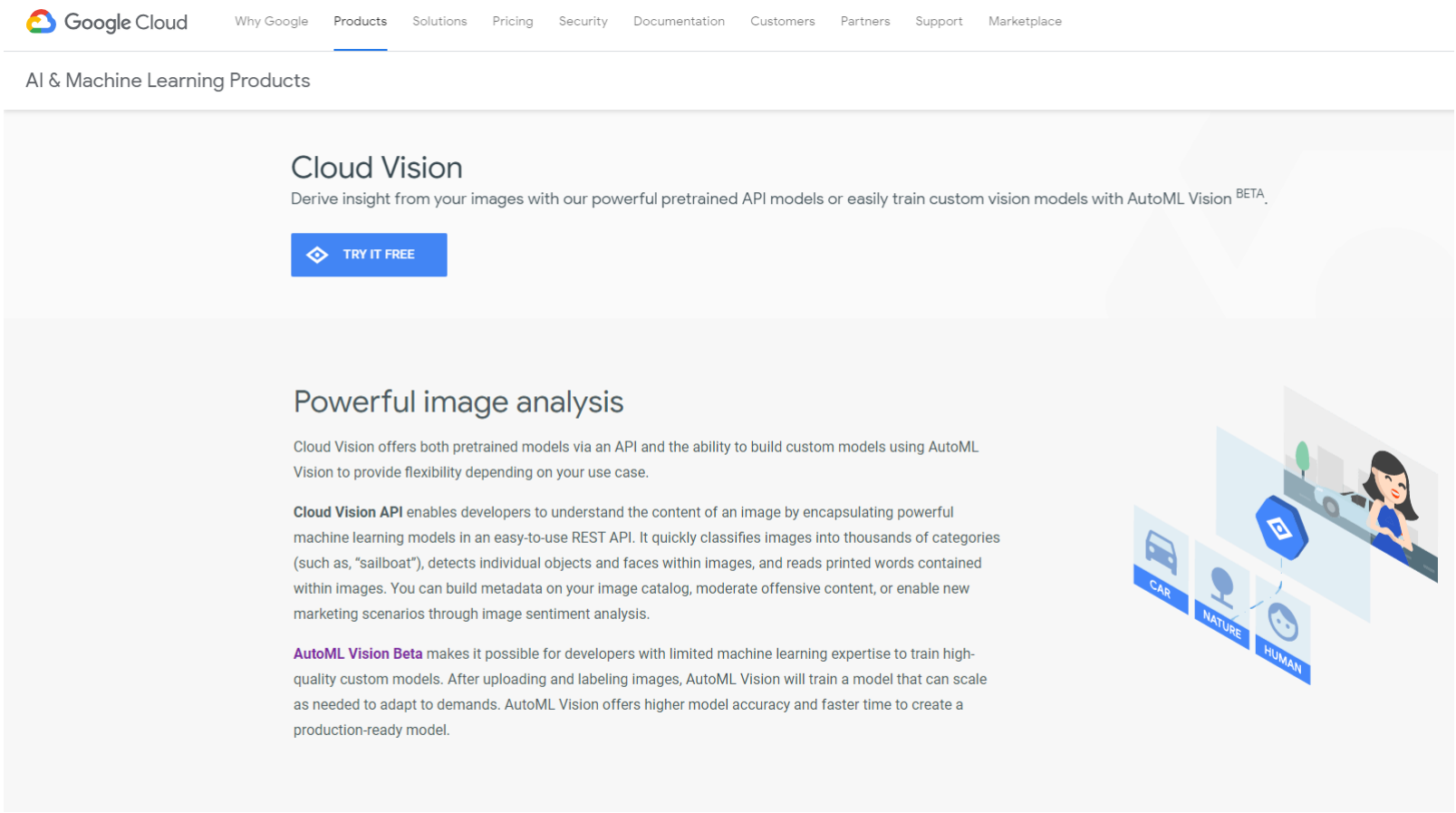
https://github.com/onnx



https://techcrunch.com/2017/09/07/facebook-and-microsoft-collaborate-to-simplify-conversions-from-pytorch-to-caffe2/

Cloud Vision

- **https://cloud.google.com/vision/?hl=en**



Free GPU

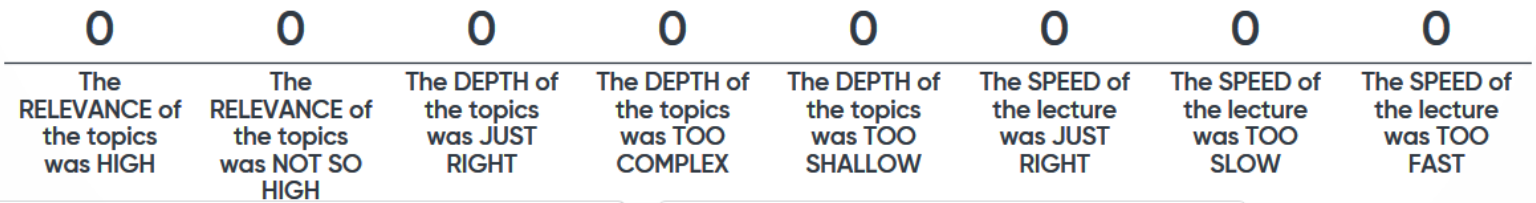
- **https://colab.research.google.com/notebooks/welcome.ipynb**

QUIZ

Go to **www.menti.com** and use the code **24 94 8**

Lecture Evaluation

Mentimeter



Results are hidden **Show results**

Slide is not active **Activate**

0