Doing Stats and ML **wrong** in the Big Data age

Berlin Machine Learning Meetup
Gerrit Gruben
4th September 2017

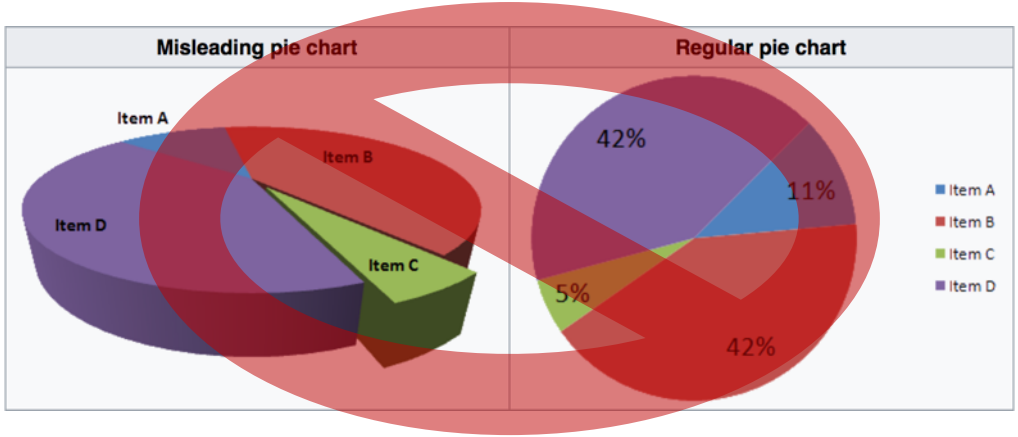
about.me

- Freelance DS
- Training people in a 3-month boot camp to be DS →
- Organizer of Kaggle meetup
  Looking for rooms!
- Degrees in math. & CS

Goals

~~Here is the silver-bullet for all your problems~~

~~Correlation does not imply causation~~

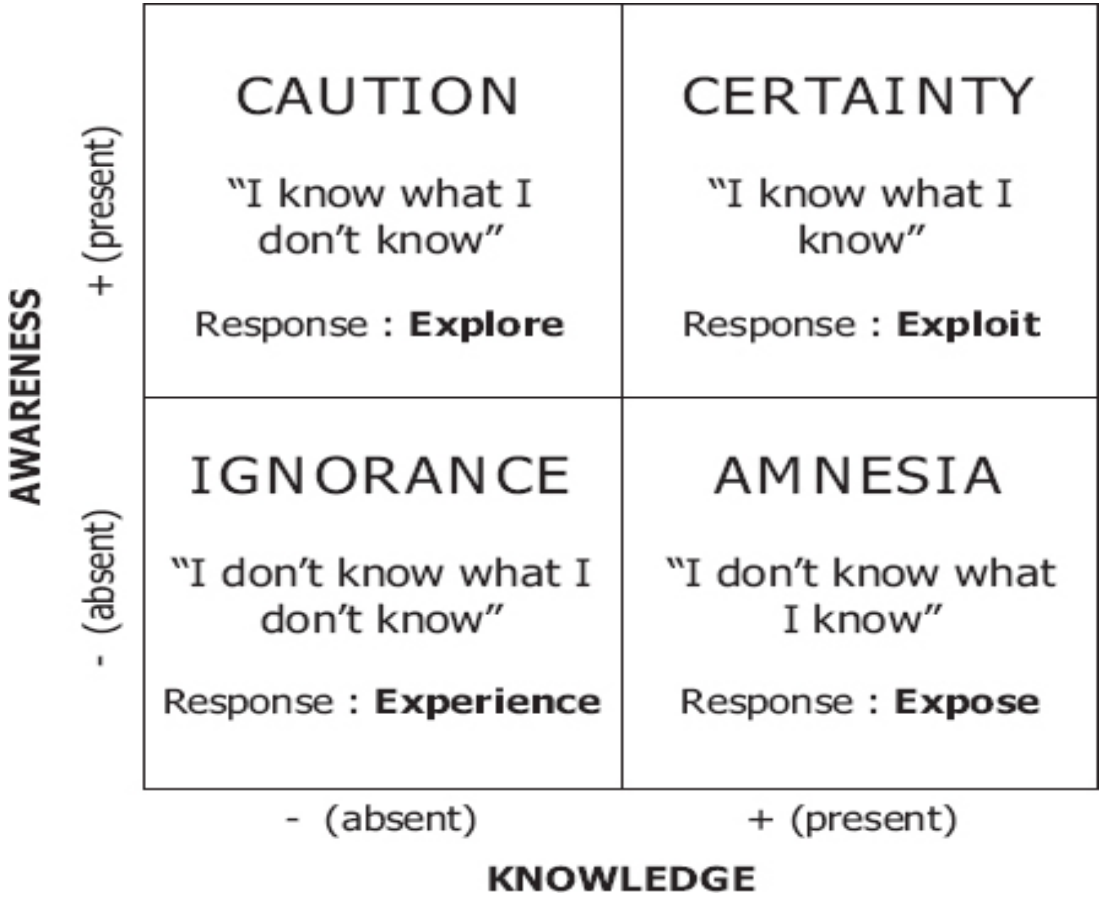


~~Use the right test~~

Limits & Biases


Benevolent or evil?

”Absence of Evidence is not Evidence of Absence" --- **Data Scientist’s Proverbs**



Source: http://www.gpmfirst.com/books/exploiting-future-uncertainty/risk-concepts

*”I beseech you, in the bowels of Christ, think it possible that you may be mistaken**"*** **--- Oliver Cromwell**

Dennis Lindley: avoid prior probabilities of 0 and 1.

Problem of Induction

- More general as the black swan problem.
- ML models have an **inductive bias.**

” When you have two competing theories that make exactly the same predictions, the simpler one is the better.*"* --- ***Ockham’s Razor***

Common Errors

A very quick walkthrough.

Hypothesis Testing

- Hypothesis testing is a statistical tool to empirically check whether one of two hypothesis is true
- One hypothesis favored, H_0 the null hypothesis
- Example: H_0: both models perform the same, H_1: model performance is different

$$H_0 : \mu_A = \mu_B, \quad H_1 : \mu_A \neq \mu_B$$

Hypothesis Testing II

- The way it works: compute some test statistic on the data sample.
- This test statistic has an associated distribution, compute **p-value** using it: probability that the data sample happened "or worse" (as-in more extreme) if H_0 is true.
- p-value thresholded on α **significance level**. p-value $< \alpha$, then reject H_0. Otherwise accept H_0.
- Depends on test: test statistic, test statistic distribution, *definition of "or worse"*.

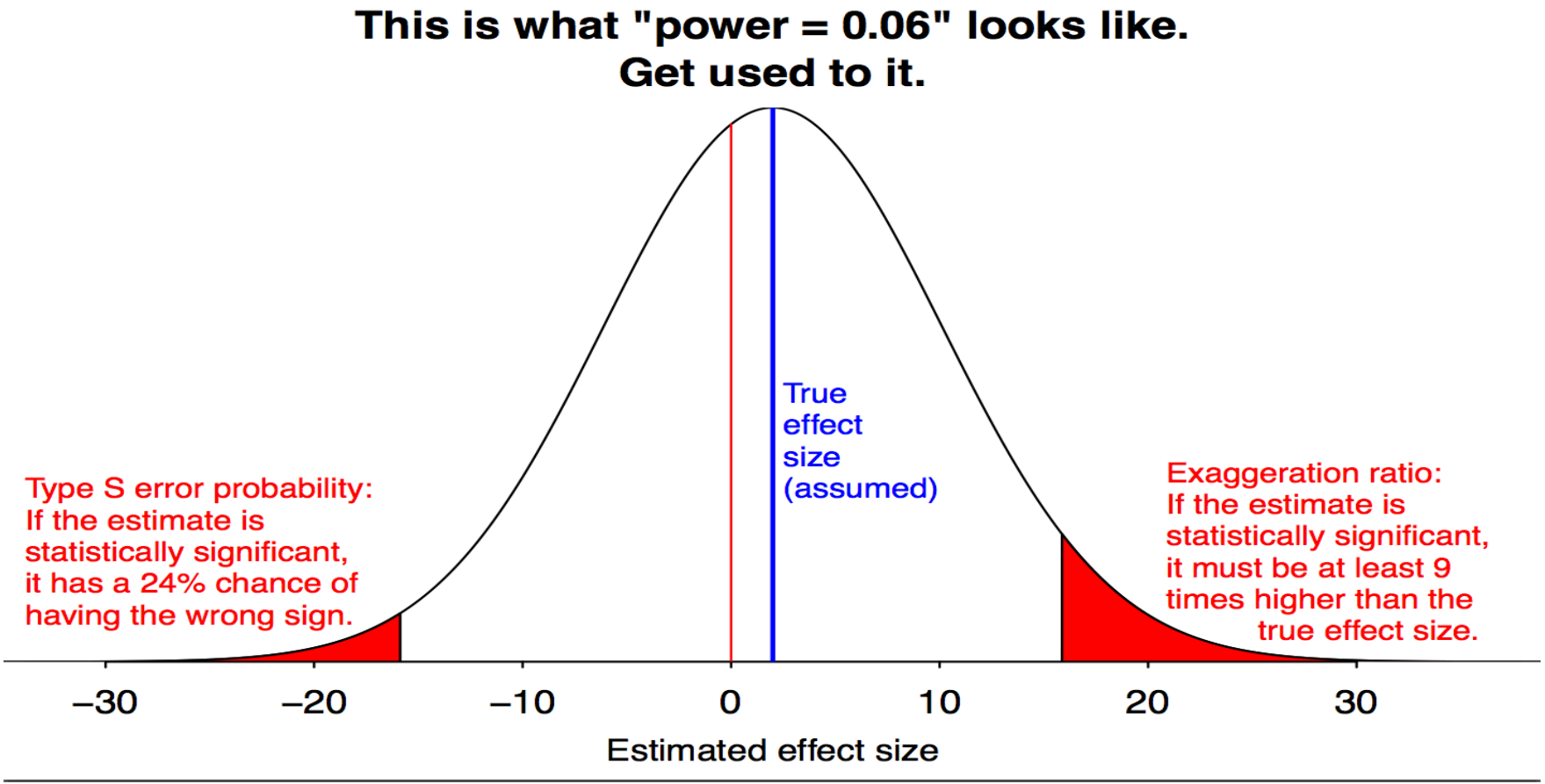
Multiple Testing

Retrying the tests so often, until "hitting" the significance level by chance.

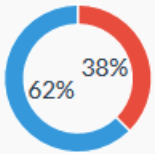
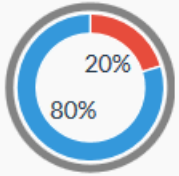
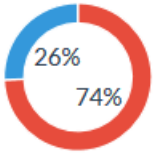
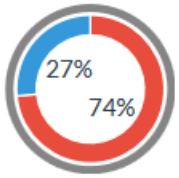
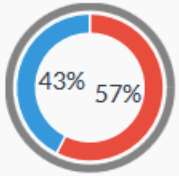
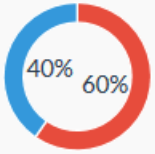
Solution: Bayesian or correction (e.g. Bonferroni correction) or different experimental design.

Data Snooping: http://bit.ly/2iWoFrV

Statistical Power



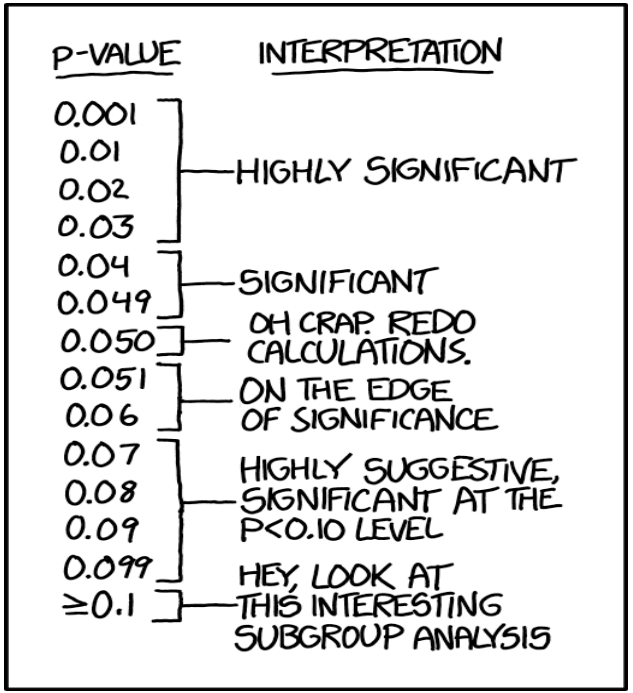
Simpson's Paradox

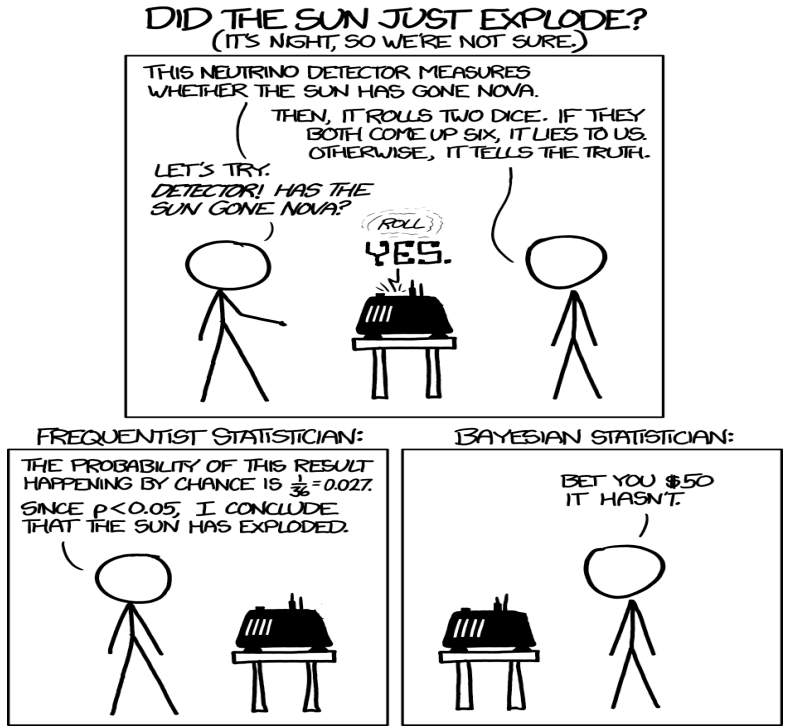
	# applied		# admitted		% admitted	
departments	men	women	men	women	men	women
"Easy"	1,238	477	772	380		
"Hard"	1,453	1,358	372	360		
Combined	2,691	1,835	1,144	740		

Simpson's paradox? yes

Let's try at: https://vudlab.com/simpsons/

Frequentist vs Bayesian





"P-hacking"

1. Stop collecting data once $p<.05$
2. Analyze many measures, but report only those with $p<.05$.
3. Collect and analyze many conditions, but only report those with $p<.05$.
4. Use covariates to get $p<.05$.
5. Exclude participants to get $p<.05$.
6. Transform the data to get $p<.05$.

"P-hacking" II

"*When a measure becomes a target, it ceases to be a good measure****"*** **--- Goodhart's law**

Machine Learning

ML is math intense

7.6.1 Computing the posterior

In linear regression, the likelihood is given by

$$p(\mathbf{y}|\mathbf{X},\mathbf{w},\mu,\sigma^2) = \mathcal{N}(\mathbf{y}|\mu+\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N) \quad (7.52)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mu\mathbf{1}_N-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mu\mathbf{1}_N-\mathbf{X}\mathbf{w})\right) \quad (7.53)$$

where μ is an offset term. If the inputs are centered, so $\sum_i x_{ij}=0$ for each j, the mean of the output is equally likely to be positive or negative. So let us put an improper prior on μ of the form $p(\mu)\propto 1$, and then integrate it out to get

$$p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}||\mathbf{y}-\overline{y}\mathbf{1}_N-\mathbf{X}\mathbf{w}||_2^2\right) \quad (7.54)$$

where $\overline{y}=\frac{1}{N}\sum_{i=1}^N y_i$ is the empirical mean of the output. For notational simplicity, we shall assume the output has been centered, and write $\mathbf{y}$ for $\mathbf{y}-\overline{y}\mathbf{1}_N$.

The conjugate prior to the above Gaussian likelihood is also a Gaussian, which we will denote by $p(\mathbf{w})=\mathcal{N}(\mathbf{w}|\mathbf{w}_0,\mathbf{V}_0)$. Using Bayes rule for Gaussians, Equation 4.125, the posterior is given by

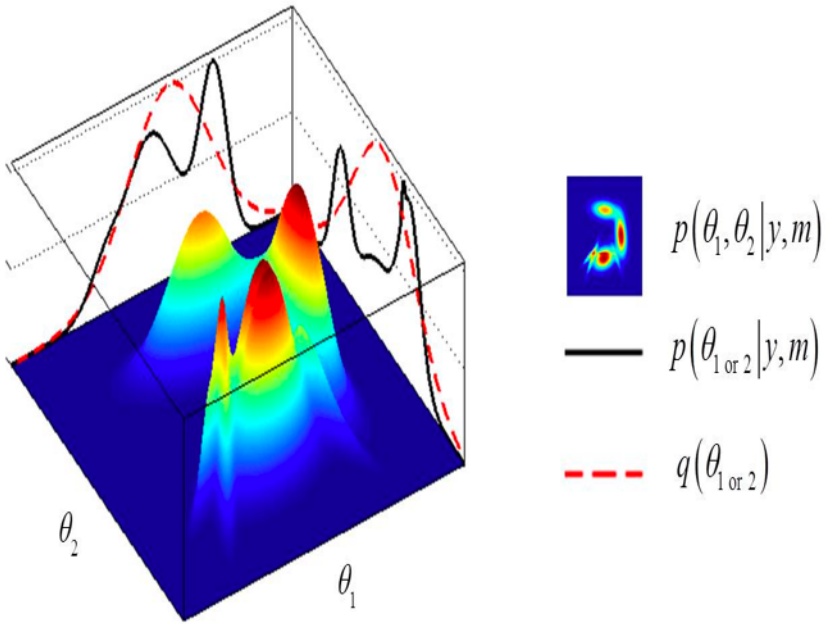
$$p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) \propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\mathbf{V}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N)=\mathcal{N}(\mathbf{w}|\mathbf{w}_N,\mathbf{V}_N) \quad (7.55)$$

$$\mathbf{w}_N = \mathbf{V}_N\mathbf{V}_0^{-1}\mathbf{w}_0+\frac{1}{\sigma^2}\mathbf{V}_N\mathbf{X}^T\mathbf{y} \quad (7.56)$$

$$\mathbf{V}_N^{-1} = \mathbf{V}_0^{-1}+\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X} \quad (7.57)$$

$$\mathbf{V}_N = \sigma^2(\sigma^2\mathbf{V}_0^{-1}+\mathbf{X}^T\mathbf{X})^{-1} \quad (7.58)$$

If $\mathbf{w}_0=\mathbf{0}$ and $\mathbf{V}_0=\tau^2\mathbf{I}$, then the posterior mean reduces to the ridge estimate, if we define $\lambda=\frac{\sigma^2}{\tau^2}$. This is because the mean and mode of a Gaussian are the same.



selection ≠ evaluation

Journal of Machine Learning Research 11 (2010) 2079-2107 Submitted 10/09; Revised 3/10; Published 7/10

On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation

Gavin C. Cawley GCC@CMP.UEA.AC.UK
Nicola L. C. Talbot NLCT@CMP.UEA.AC.UK
School of Computing Sciences
University of East Anglia
Norwich, United Kingdom NR4 7TJ

Editor: Isabelle Guyon

Abstract

Model selection strategies for machine learning algorithms typically involve the numerical optimisation of an appropriate model selection criterion, often based on an estimator of generalisation performance, such as k-fold cross-validation. The error of such an estimator can be broken down into bias and variance components. While unbiasedness is often cited as a beneficial quality of a model selection criterion, we demonstrate that a low variance is at least as important, as a non-negligible variance introduces the potential for over-fitting in model selection as well as in training the model. While this observation is in hindsight perhaps rather obvious, the degradation in performance due to over-fitting the model selection criterion can be surprisingly large, an observation that appears to have received little attention in the machine learning literature to date. In this paper, we

Prefer to call it “over-selection”

In “Learning with Kernels” from Smola & Schölkopf they name ex. 5.10. “overfitting on the test set”.

Paper: http://bit.ly/2gBIR1M

Empirical Risk Minimization

- Two assumptions:
  - Data follows a distribution $p(x, y)$ on $X \times Y$.
  - Learner sees data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ of **independent** and **identically sampled** (from p) data (*IID assumption*).

- **The IID assumption is almost always broken in practice.**

Empirical Loss

- Given a loss function l the *empirical loss/risk is,*

$$R_D[f] = \sum_{n=1}^{N} l(f(x_n), y_n, x_n).$$

- This is a statistical estimator if you vary the selection of D.

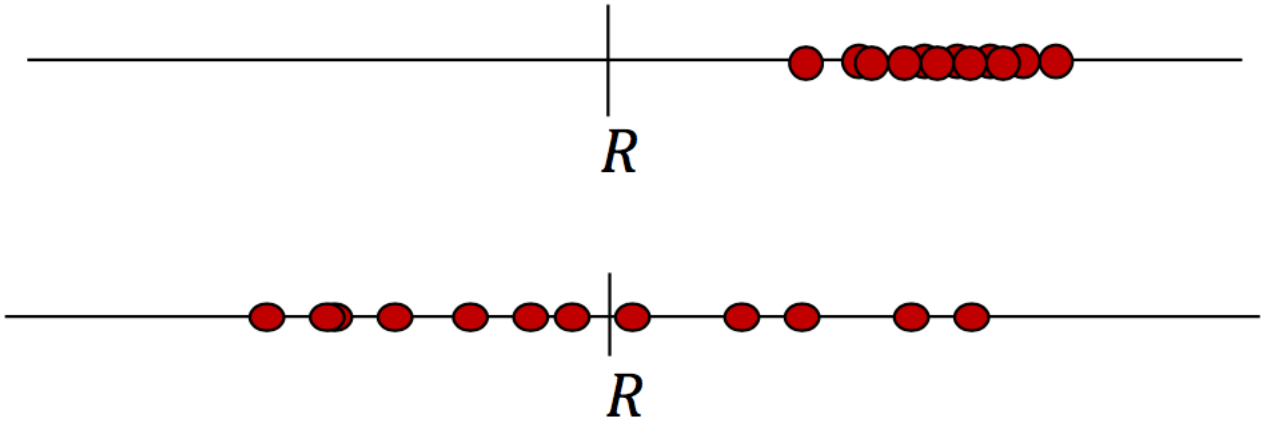
Empirical Risk Minimization II

- $R_D[f]$ is supposed to estimate the *generalization error/risk,*

$$R[f] = \int l(f(x), y, x) dp(x, y) .$$

- The larger $|D|$ the more accurate.
- $R_D[f]$ has a bias (distance to $\mathrm{R}[f]$) and a variance.

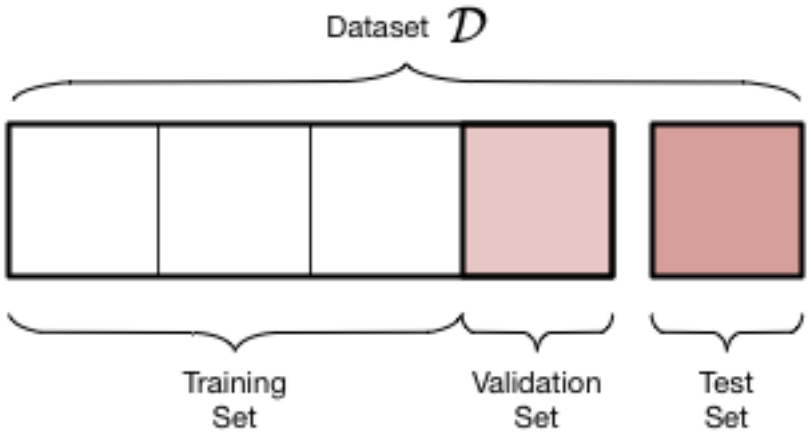
Bias / Variance



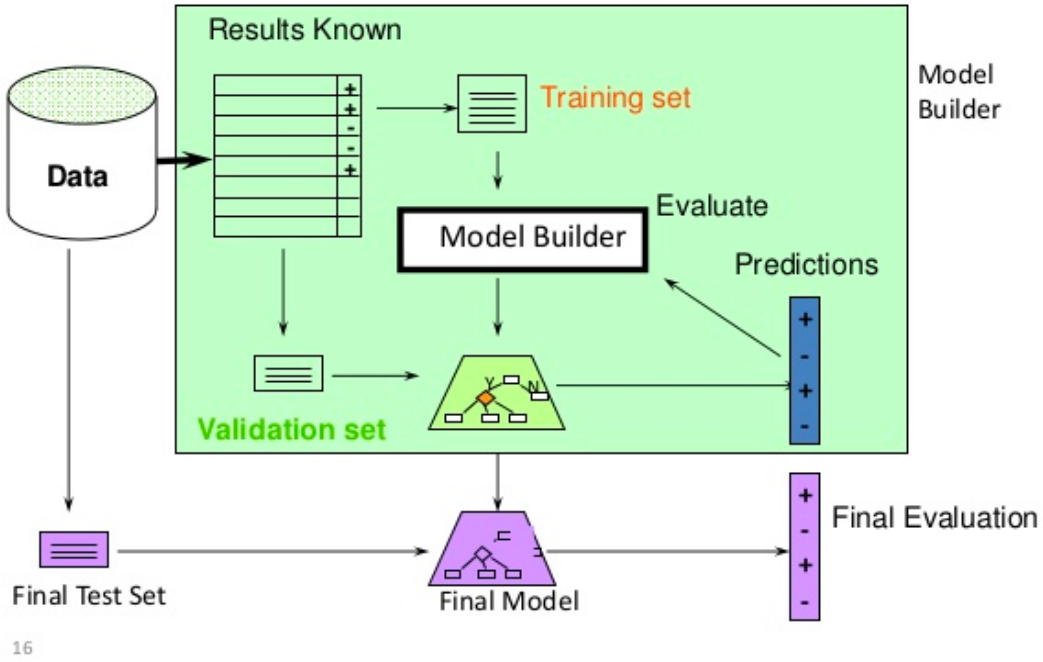
- **Model fitting:** finding the best model parameter:

$$\theta = argmin_{\theta} R_D[f_\theta]$$

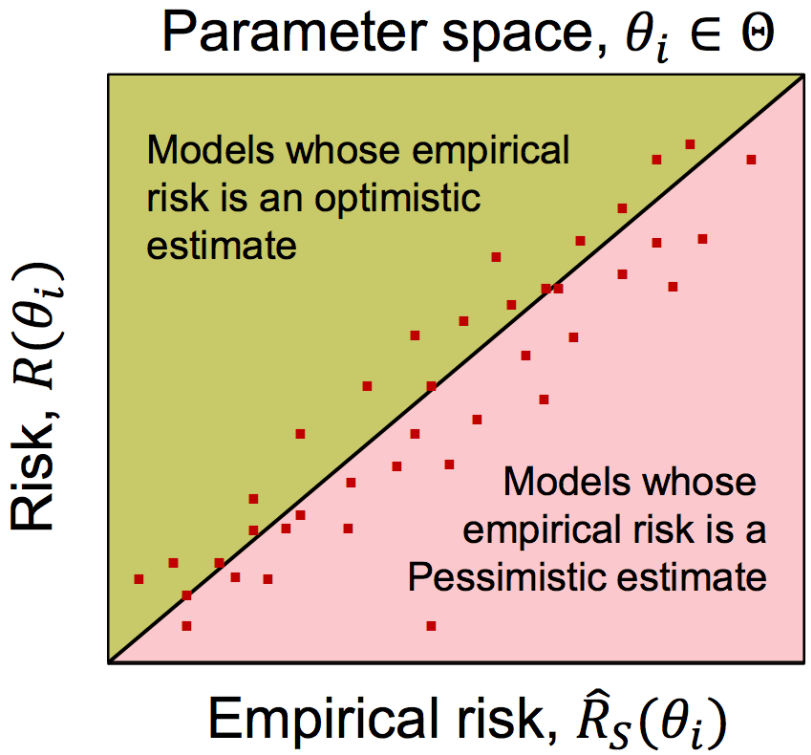
- **Model selection:** find the best fitting model family / hyperparams.
- **Model evaluation:** estimate the generalization risk.



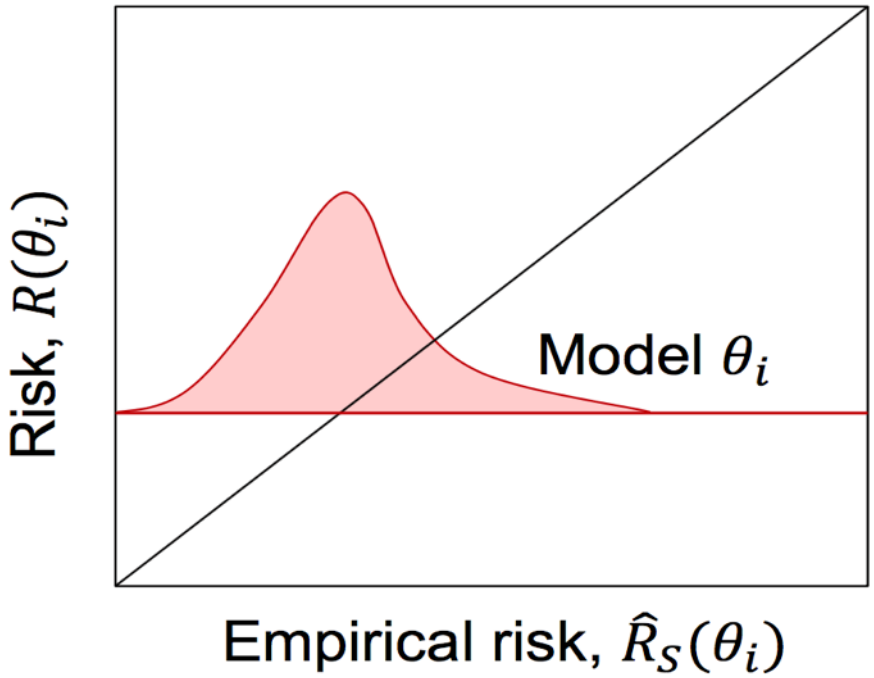
Classification:
Train, Validation, Test split



Source: http://bit.ly/2vDfoLp

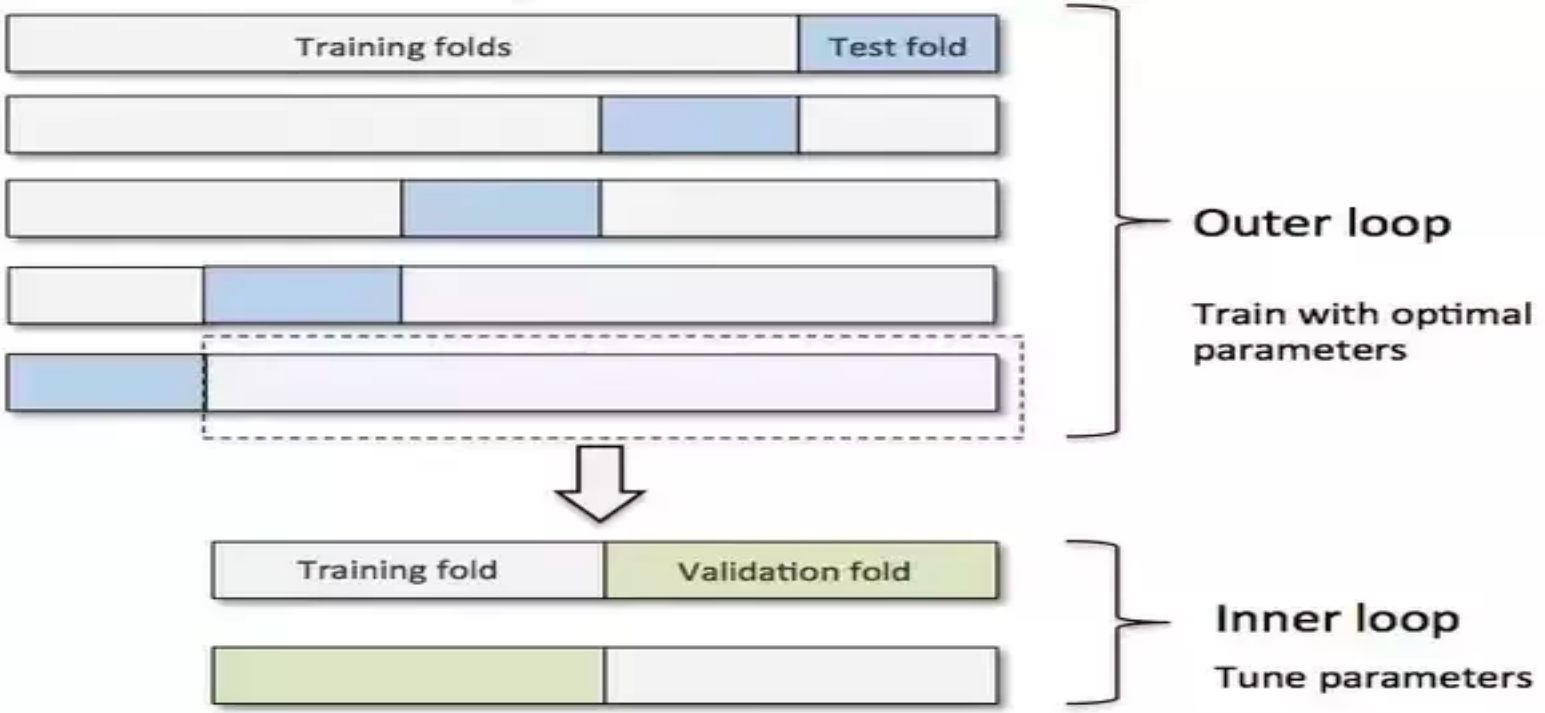


Source: University of Potsdam



Source: University of Potsdam

Nested CV



From Quora: http://bit.ly/2wvz2aZ

Messing up your experiments

- Data split strategy is part of experiment.
- Mainly care for:
  - Class distribution
  - Problem domain relevant issues such as time

> *”Validation and Test sets should model nature and nature is not accommodating.*" --- **Data Scientist’s Proverbs**

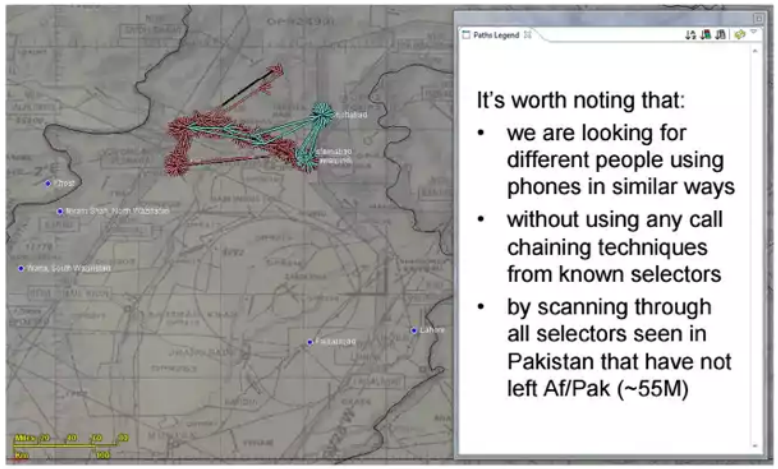
“Model evaluation, model selection…“
by Sebastian Raschka: http://bit.ly/2p6PGY0

“Approximate Statistical Tests For Comparing Supervised Class. Learning Algorithms” (Dietterich 98): http://bit.ly/2wyItF6

Courier/Terrorist detection in Pakistan

TOP SECRET//COMINT//REL TO USA, FVEY

Given a handful of courier selectors, can we find others that "behave similarly" by analyzing GSM metadata?



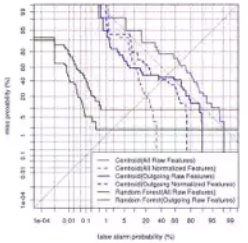
TOP SECRET//COMINT//REL TO USA, FVEY

TOP SECRET//COMINT//REL TO USA, FVEY

Preliminary results indicate that we're on the right track, but much remains to be done

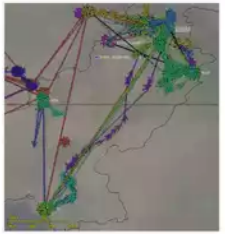
Cross Validation Experiment:

- Random Forest classifier operating at 0.18% false alarm rate at 50% miss
- Enhancing training data with Anchory selectors reduced that to 0.008%
- Mean Reciprocal Rank is ~1/10



Preliminary SIGINT Findings:

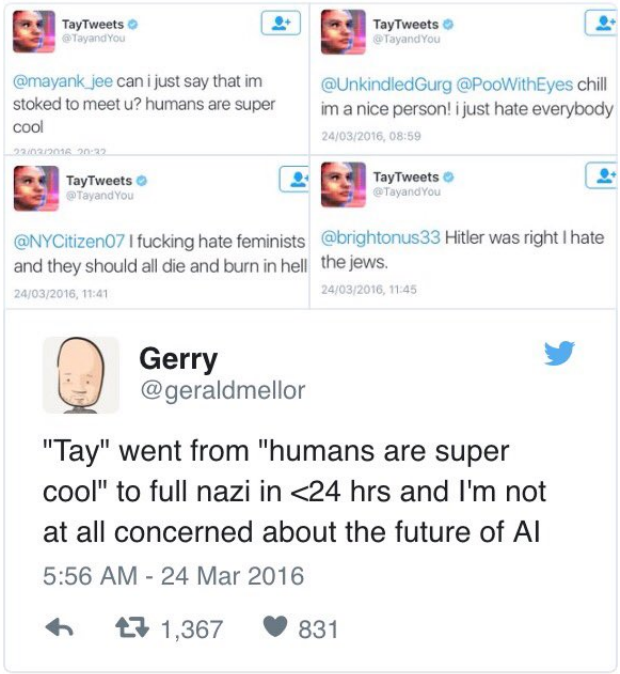
- Behavioral features helped discover similar selectors with "courier-like" travel patterns
- High number of tasked selectors at the top is hopefully indicative of the detector performing well "in the wild"



TOP SECRET//COMINT//REL TO USA, FVEY

Source: http://bit.ly/1KY4SQE

Feedback loops abused

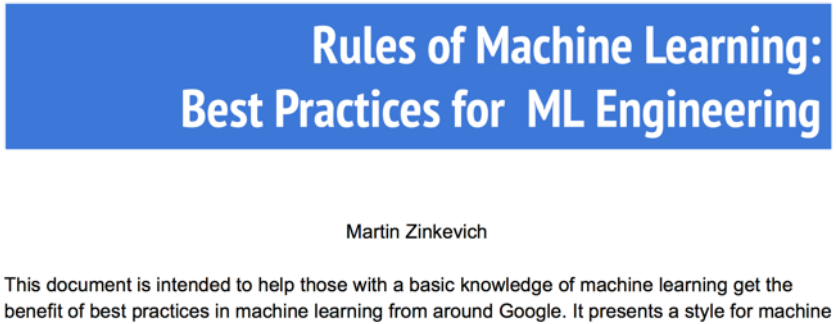


Tay.ai was a chat bot deployed on Twitter by Microsoft for just a day.

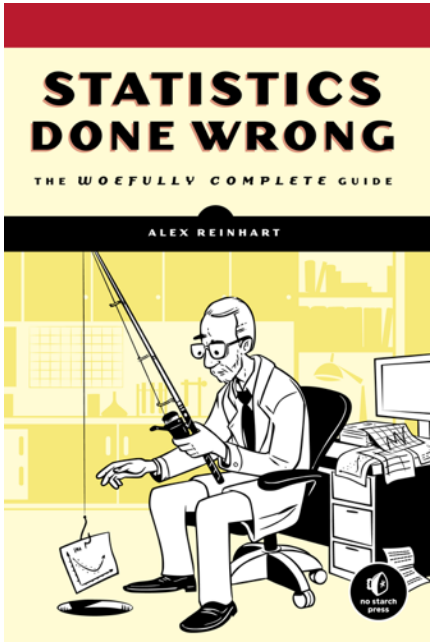
Trolls started to "subvert" the bot by "teaching" it to be politically incorrect by focussed exposure to extreme content.

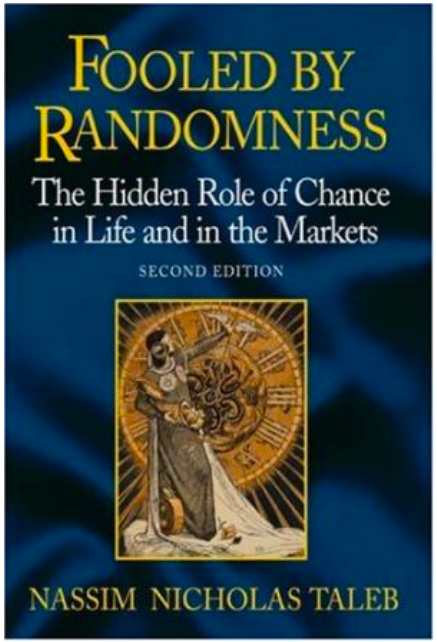
Smaller tips for ML

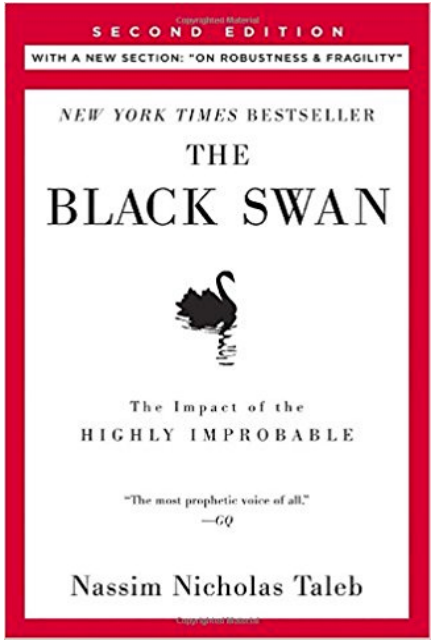
- Always model uncertainty.
- Read this
- Don’t mock values of a non-existant predictive model.



Sources


STATISTICS
DONE WRONG
THE WOEFULLY COMPLETE GUIDE
ALEX REINHART
no starch press


FOOLED BY
RANDOMNESS
The Hidden Role of Chance
in Life and in the Markets
SECOND EDITION
NASSIM NICHOLAS TALEB


SECOND EDITION
WITH A NEW SECTION: "ON ROBUSTNESS & FRAGILITY"
NEW YORK TIMES BESTSELLER
THE
BLACK SWAN
The Impact of the
HIGHLY IMPROBABLE
"The most prophetic voice of all."
—GQ
Nassim Nicholas Taleb

Other Links

- https://www.ma.utexas.edu/users/mks/statmistakes/StatisticsMistakes.html
- Quantopian Lecture Series: p-Hacking and Multiple Comparison bias https://www.youtube.com/watch?v=YiDfbYtgUPc