Spot the Difference:

Statistics, Data Science, Machine Learning, AI

Vinny Davies

Who am I?

Dr Vinny Davies - *@Vinny_Davies89*

- BSc Maths & MSc Statistics – Lancaster University
- PhD Statistics – University of Glasgow

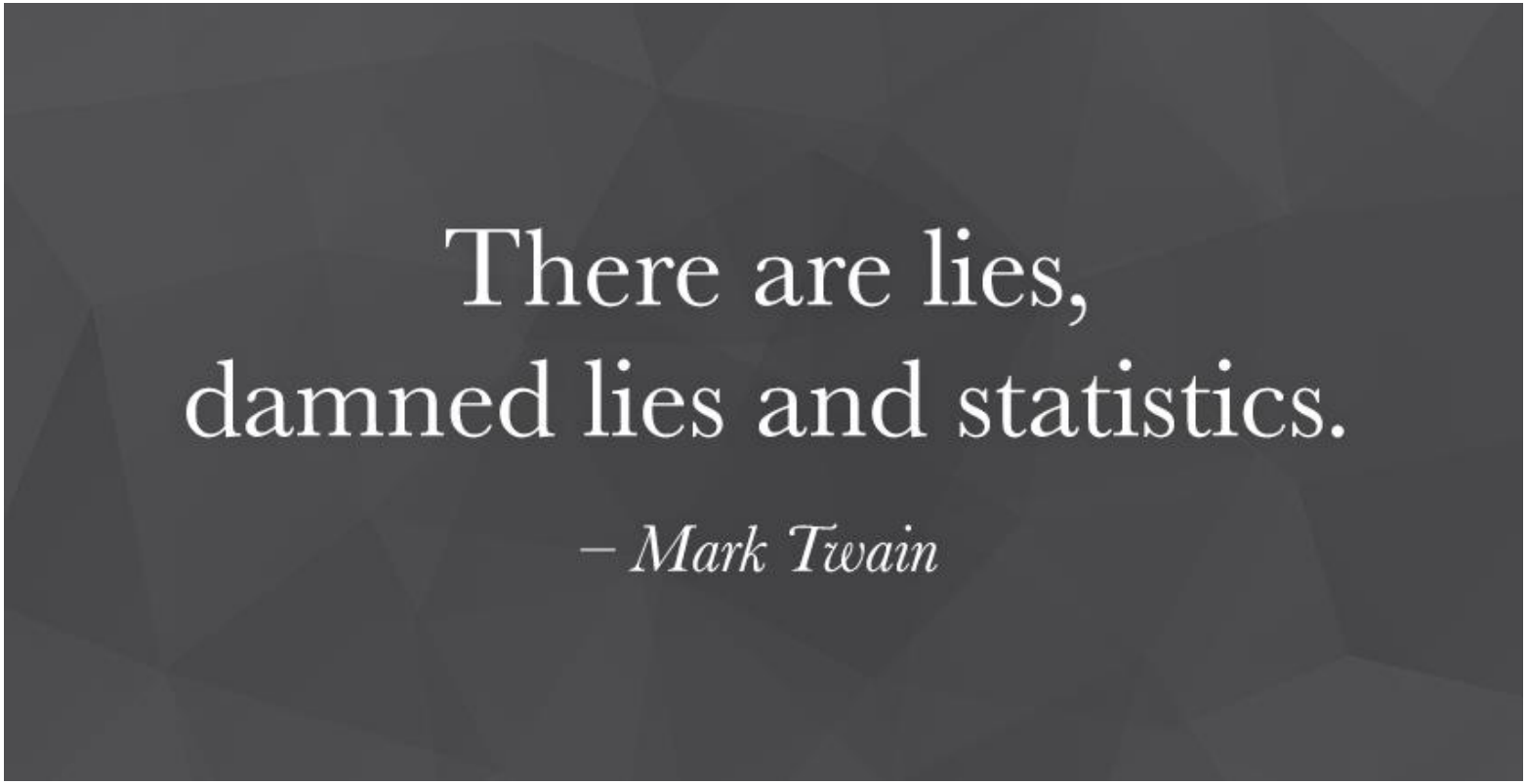
- Postdoc Statistics – University of Glasgow (2016-17)
- Research Biostatistician – University of Leeds (2017-18)
- Postdoc Computer Science – University of Glasgow (2018-Present)
- Consultant – Freelance (2014-Present)

Why am I qualified to “Spot the Difference”?

- Worked in Statistics, Applied Statistics and Machine Learning
- Worked in Statistics, Health and Computer Science departments
- Publications and conferences in both Statistics and Machine Learning
- Used and developed methods and techniques from all areas

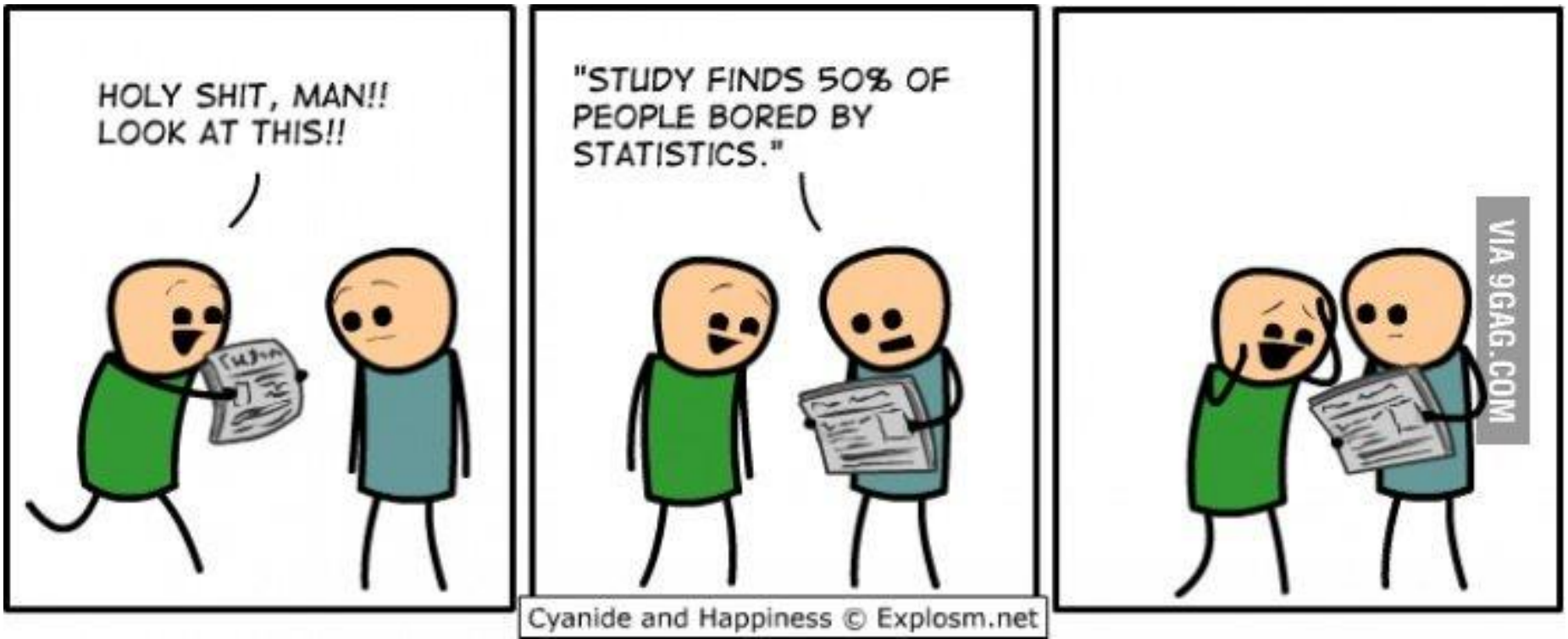
What do people think they are?

Statistics



Statistics

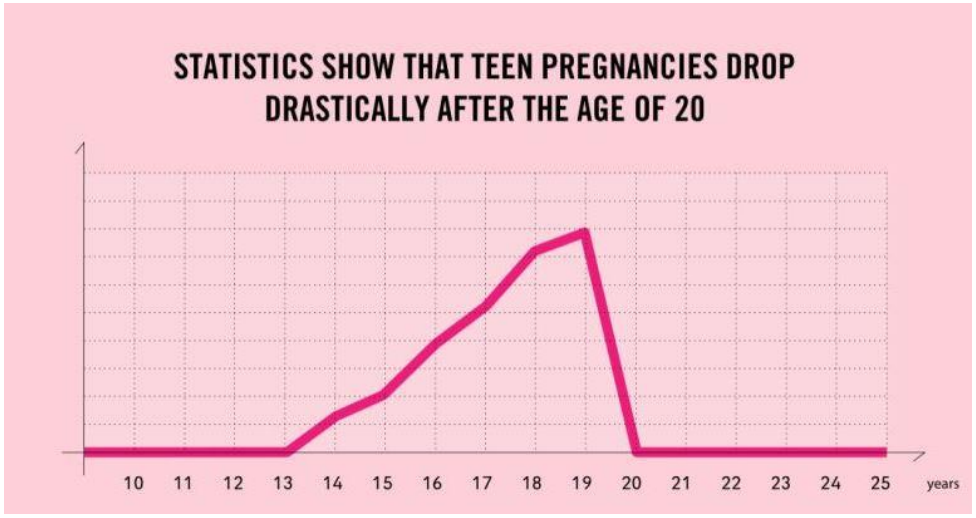
- Everyone in research needs Statistics, but not many people like it



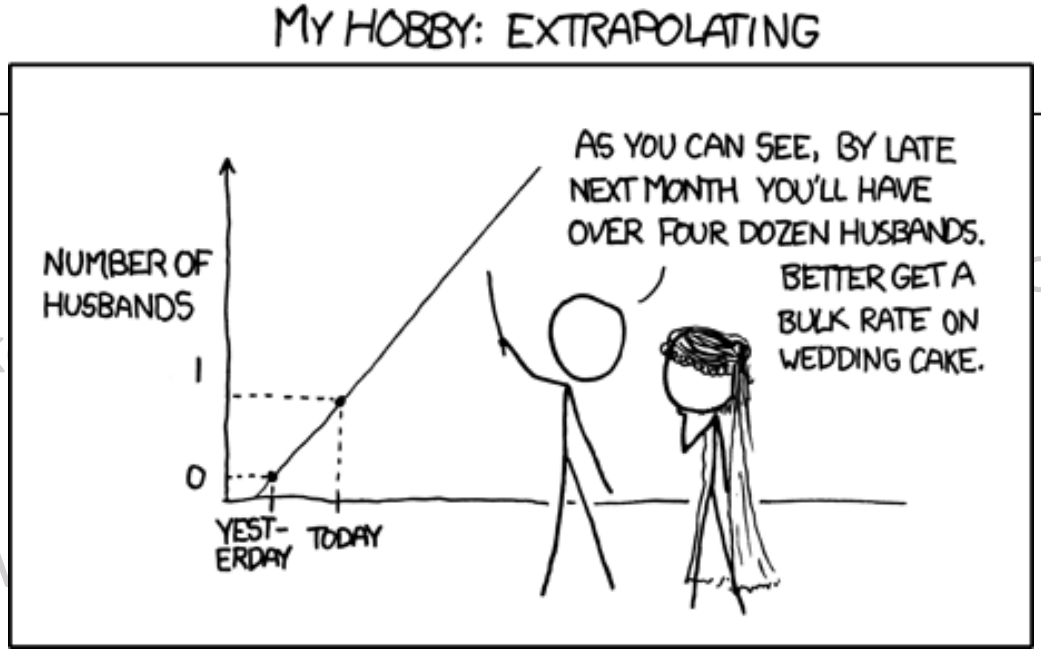
Statistics

- Everyone in research needs Statistics, but not many people like it

- Media coverage gives a bad impression



Statistics

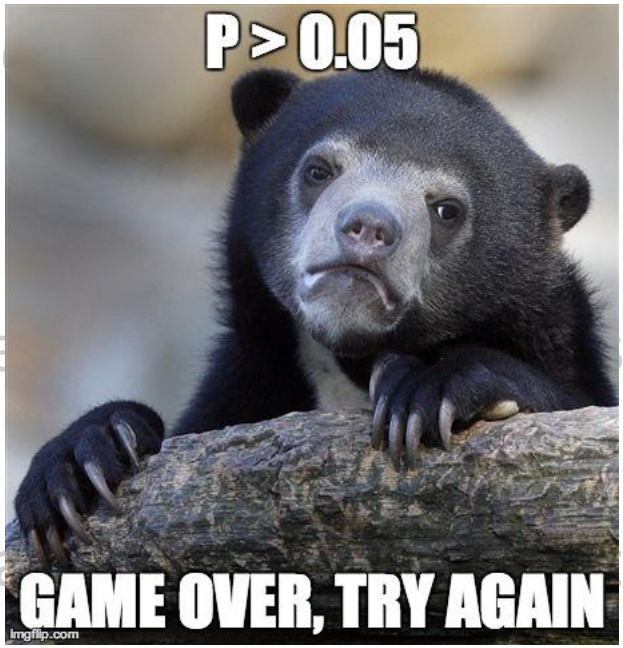


- Everyone ot many people lik
- Media co

- Statistics get blamed for terrible data summaries

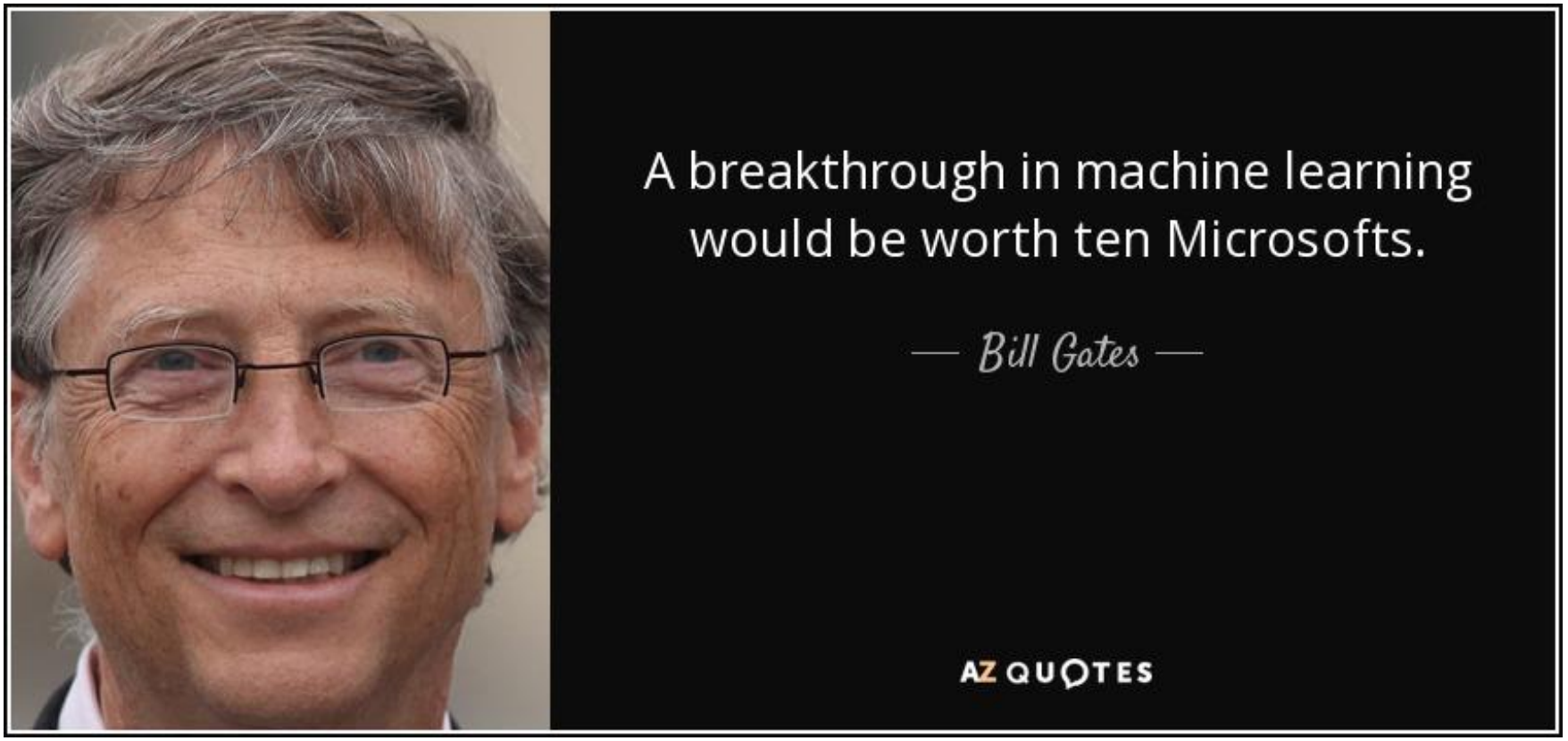
Statistics

- Everyone in res s, but not many people like it
- Media coverage sion
- Statistics get bla a summaries



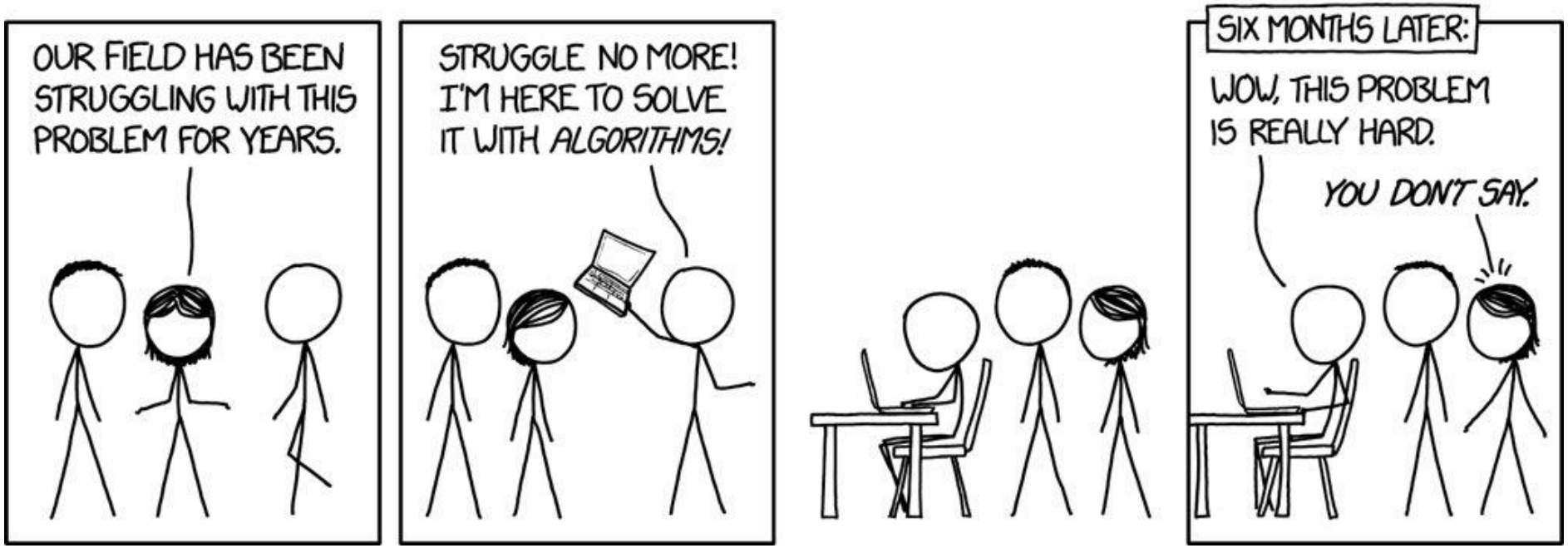
- Statistics has a reputation for being basic

Machine Learning



Machine Learning

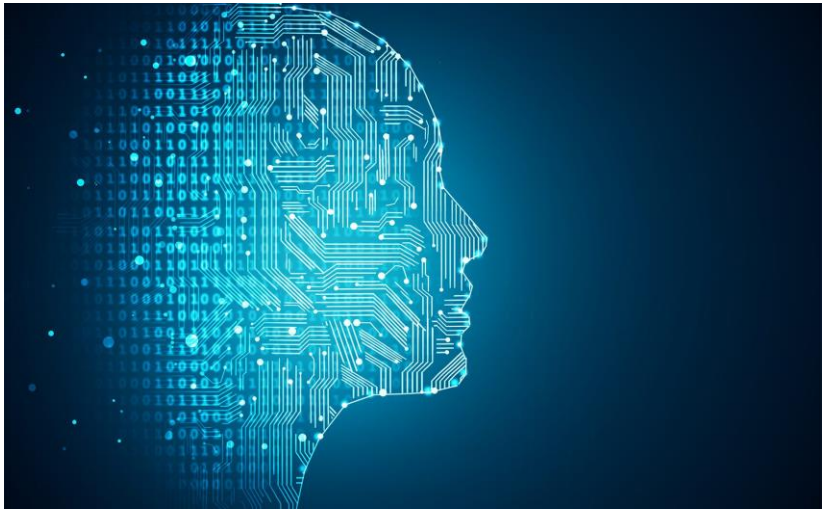
- People think Machine Learning can solve all their problems

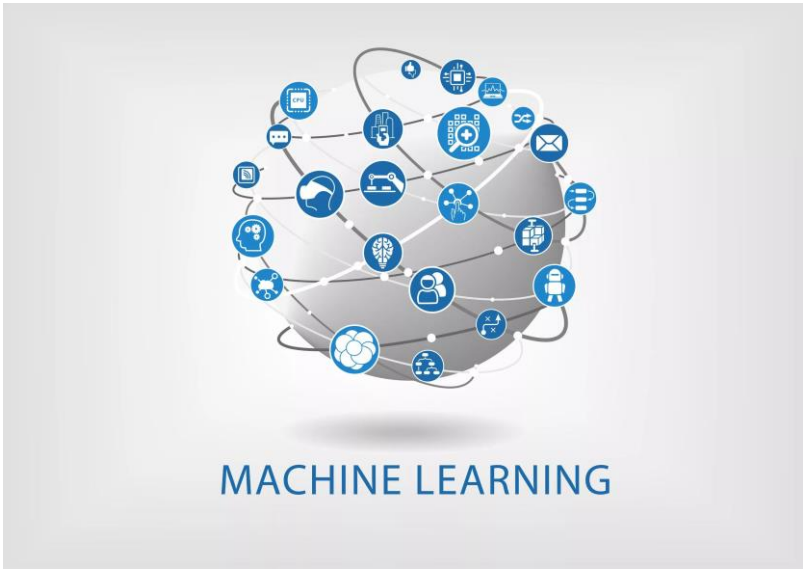


Machine Learning

- People think Machine Learning can solve all their problems

- People think Machine Learning is more advanced



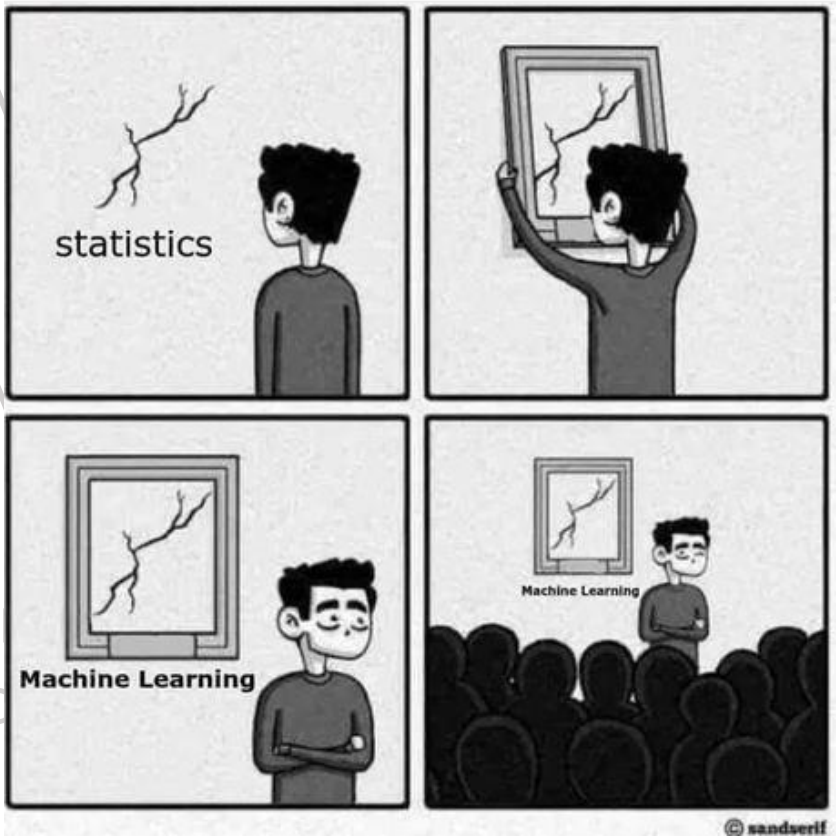


Machine Learnin

- People think M olve all their problems
- People think M re advanced
- Machine Learners don't always check assumptions – Good or Bad?

Machine Learning

- People think M… all their problems
- People think M… dvanced
- Machine Lear… umptions – Good or Bad

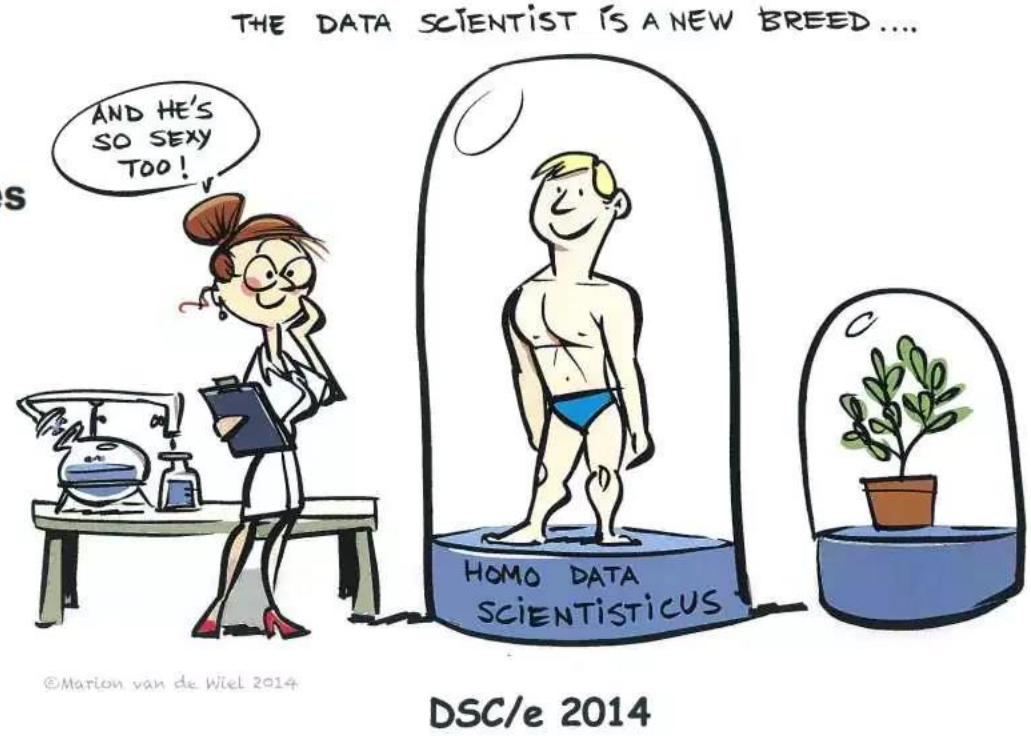


- Machine Learning gets you funding in some fields…

Data Science

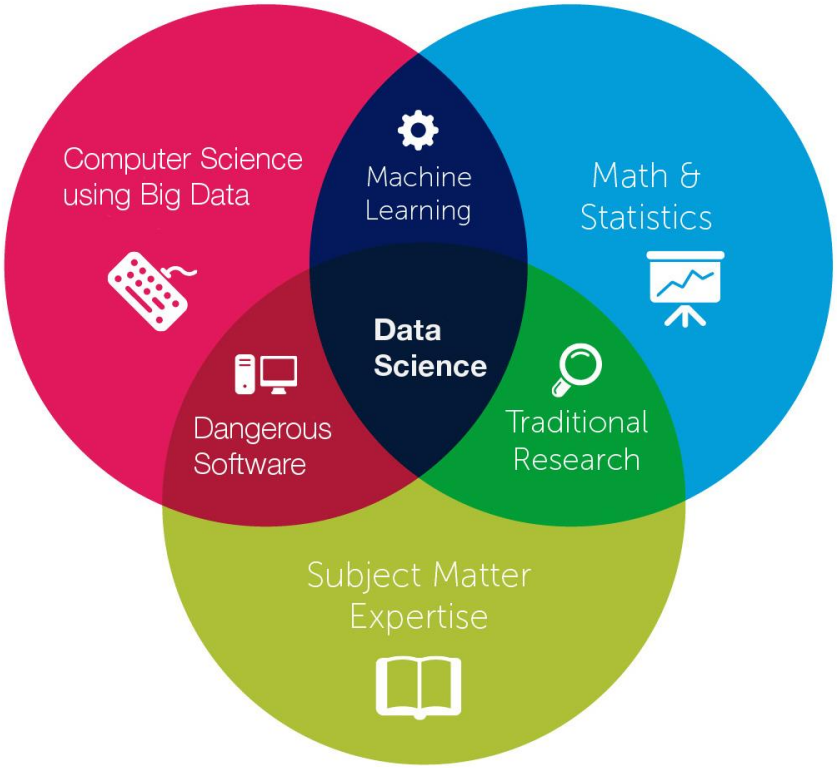
data science is

data science is **dead**
data science is **the future**
data science is **the sexiest job**
data science is **hard**
data science is **a branch of**
data science is **software**
data science is **statistics on a mac**
data science is **not taught at universities**
data science is **overrated**
data science is **a fad**



Data Science

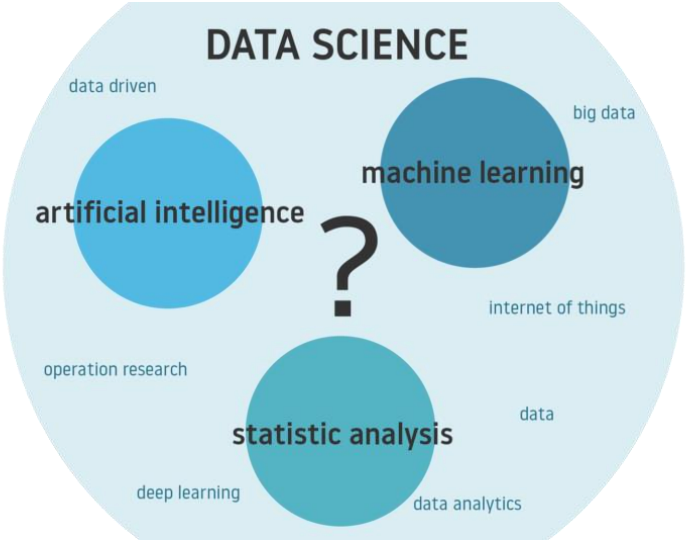
- Data Science encompasses aspects of both Statistics and Machine Learning



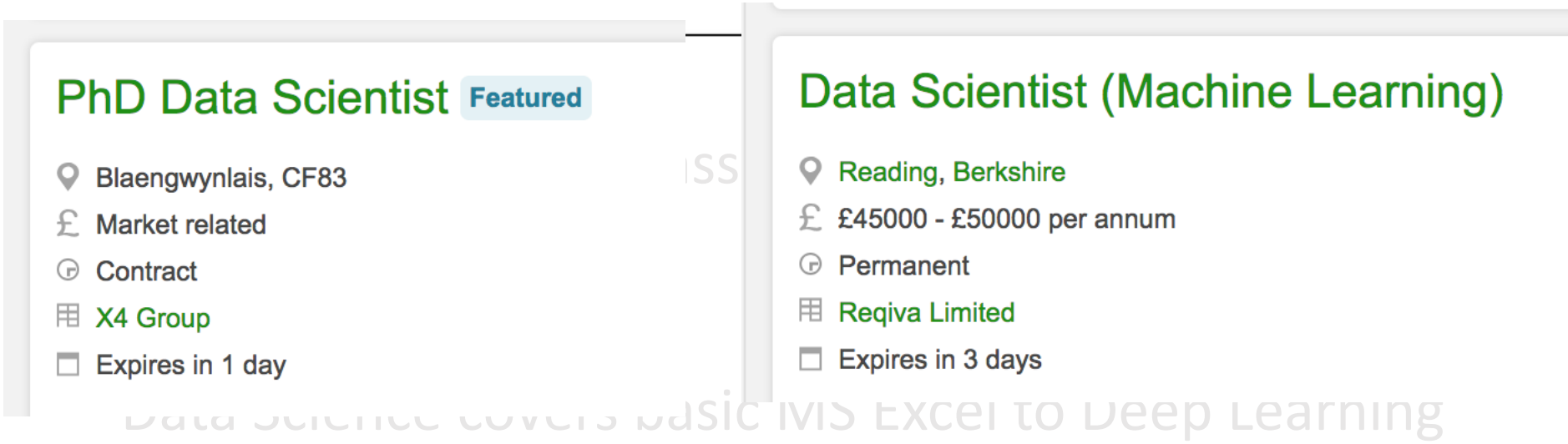
Data Science

- Data Science encompasses aspects of both Statistics and Machine Learning

- Data Science covers basic MS Excel to Deep Learning

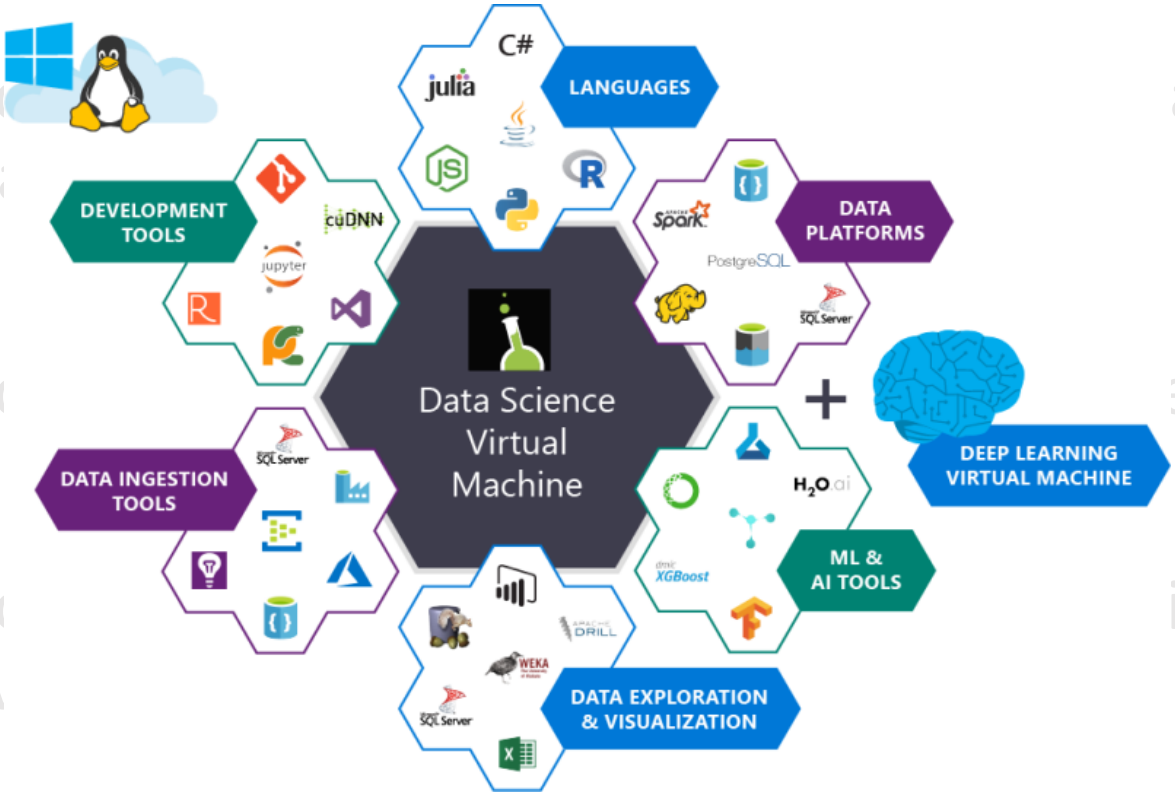


Data Science



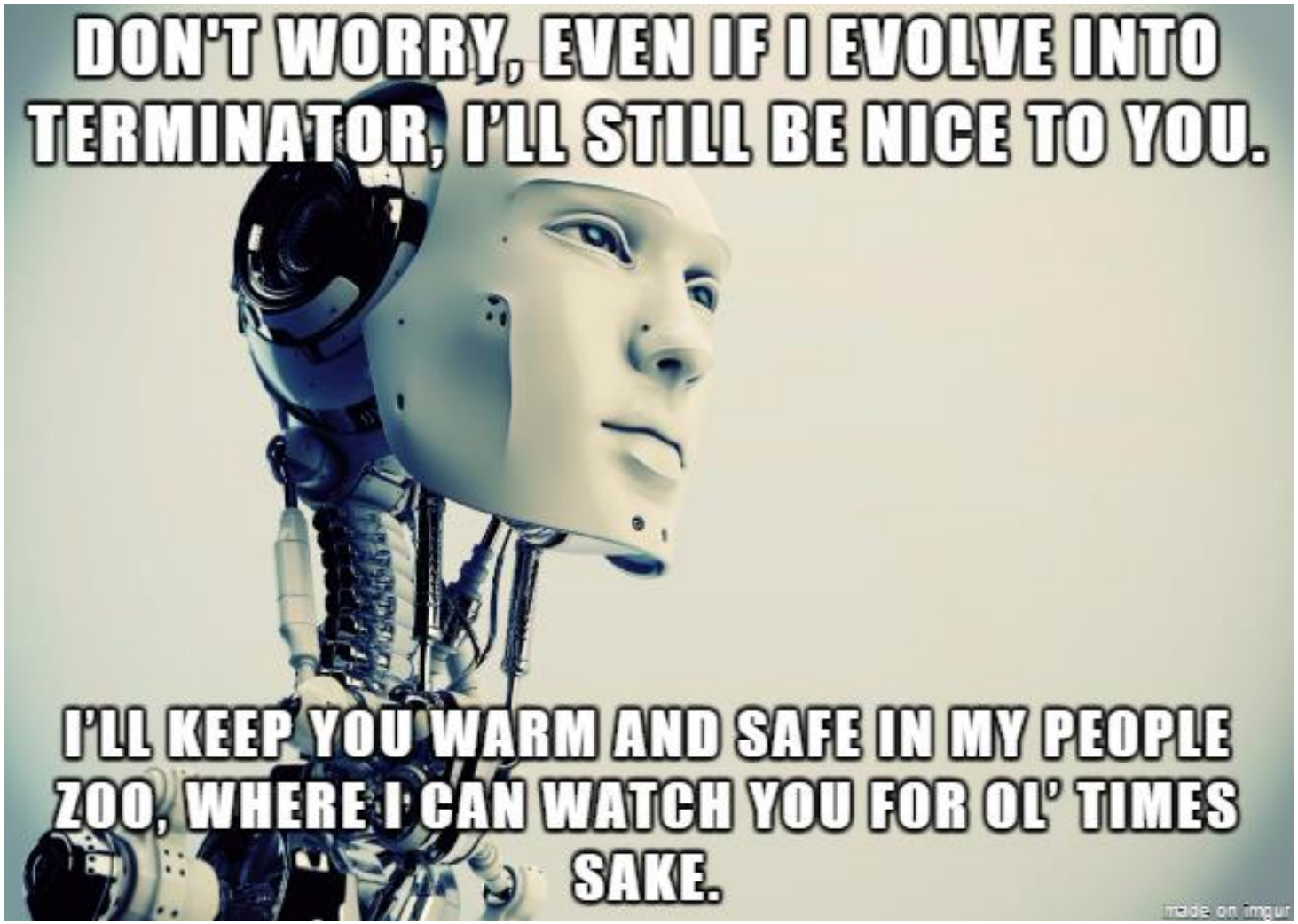
- Data Science employs people with a large variety of skill levels and job descriptions

Data Science



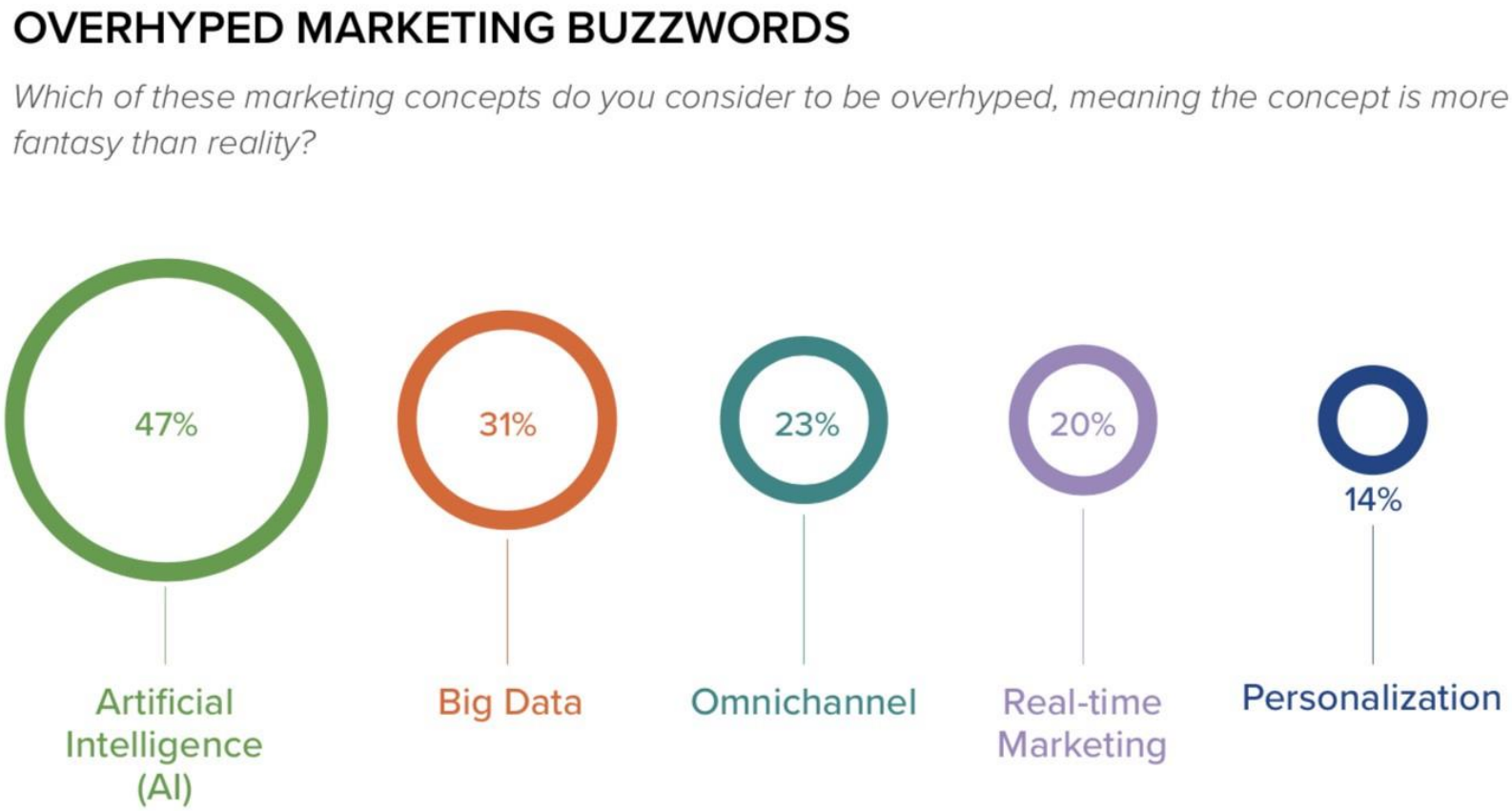
- Data Science has the reputation for being flexible, but potentially not done well

AI (Artificial Intelligence)



AI

- AI is the latest trend and one of the biggest buzzwords



AI

- AI is the latest trend and one of the biggest buzzwords

- AI is behind some of the newest technologies

AI

- AI is the latest trend and one of the biggest buzzwords
- AI is behind some of the newest technologies
- People believe that AI can solve all their problems

Clinician using basic statistics:
"I want to use AI on our data"

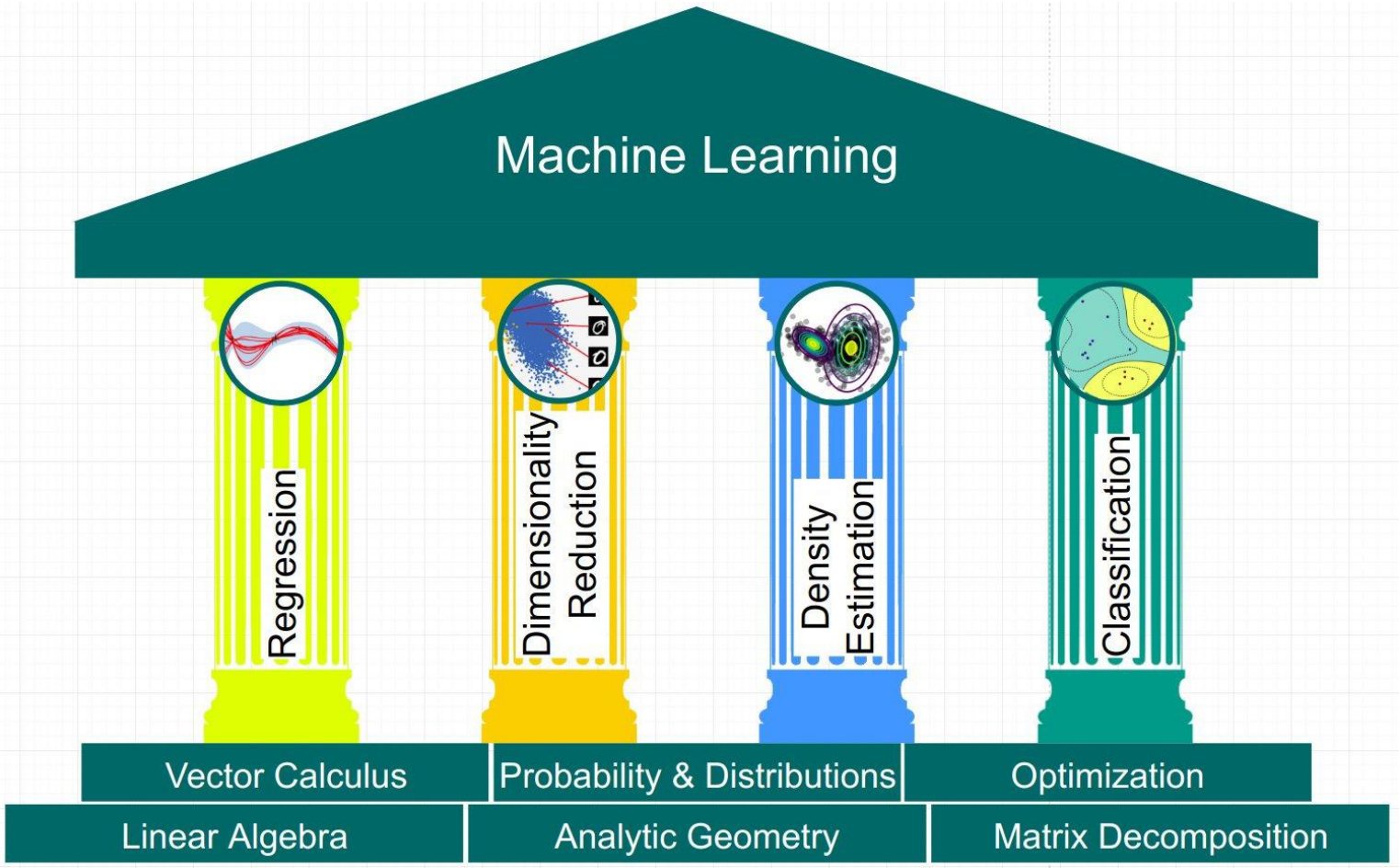
AI

- AI is the uzzwords
- AI is beh
- People b lems

- Applied researchers (not including applied statisticians etc) don’t generally understand when AI can be useful or what it is

What’s the actual difference?

Shared Ideas

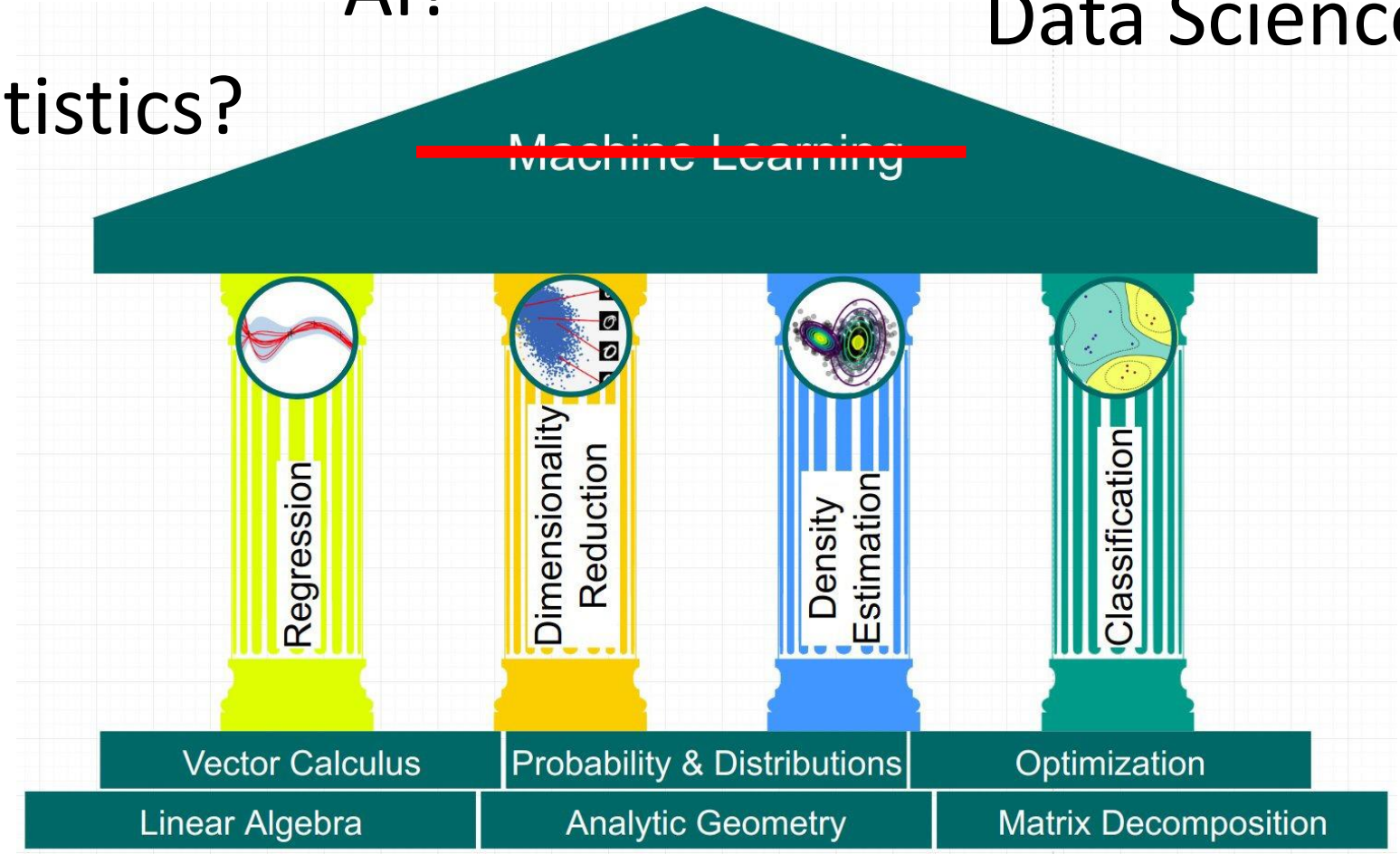


Shared Ideas

AI?

Data Science?

Statistics?



Shared Methods

- All fields have similar goals and therefore share methods
    - Prediction
    - Classification
    - Extrapolation
    - Etc….
- Regression is used across all fields in some form

Shared Methods

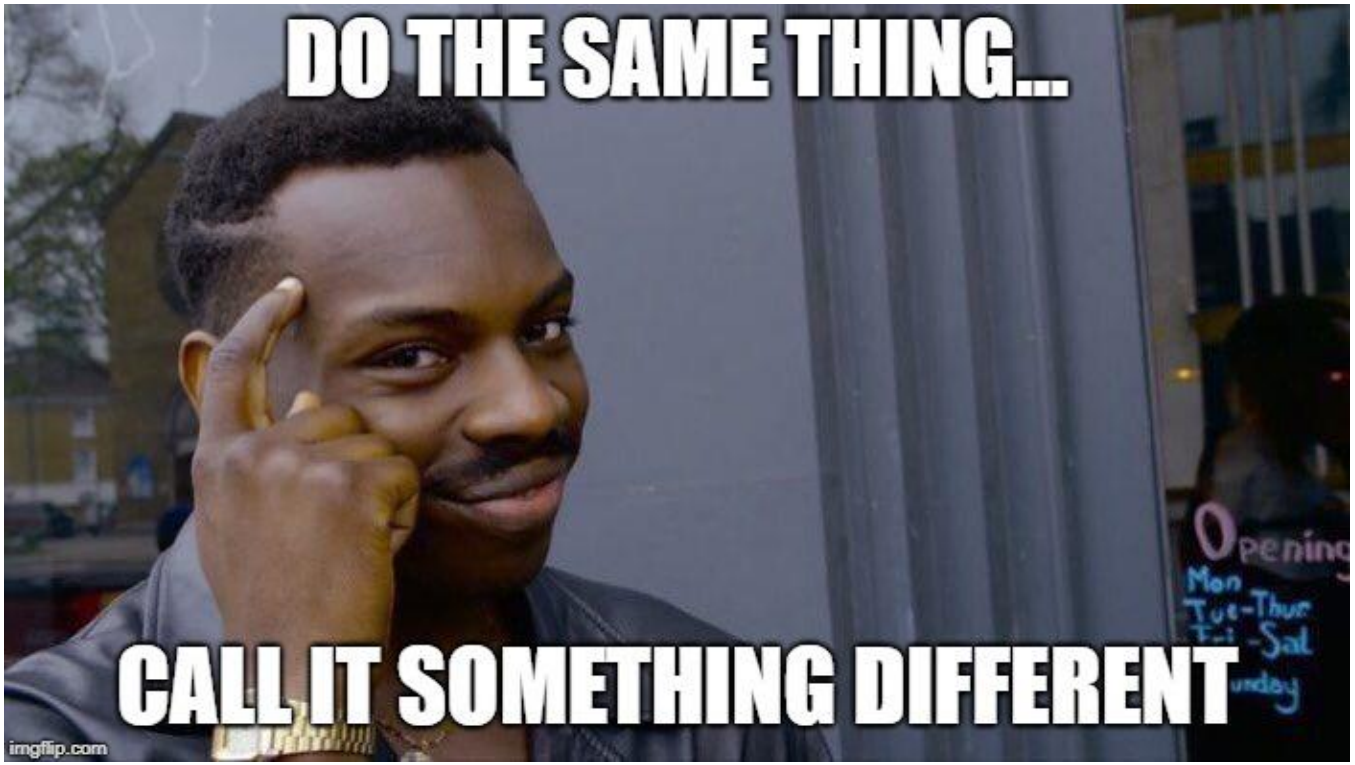
- Andrew Ng's popular Machine Learning course on Coursera spends 3 weeks on regression
- Andrew Ng's Deep Learning (AI?) course on Coursera again starts with regression
- Johns Hopkins University Data Science course of Coursera covers multiple aspects of Statistics and Machine Learning

Using Methods from Other “Disciplines”

- Applied Researchers have worked out that if they call Statistics “AI” or “Machine Learning” then they get grants, funding and papers

Using Methods from Other “Disciplines”

- Applied Researchers have worked out that if they call Statistics “AI” or “Machine Learning” then they get grants, funding and papers

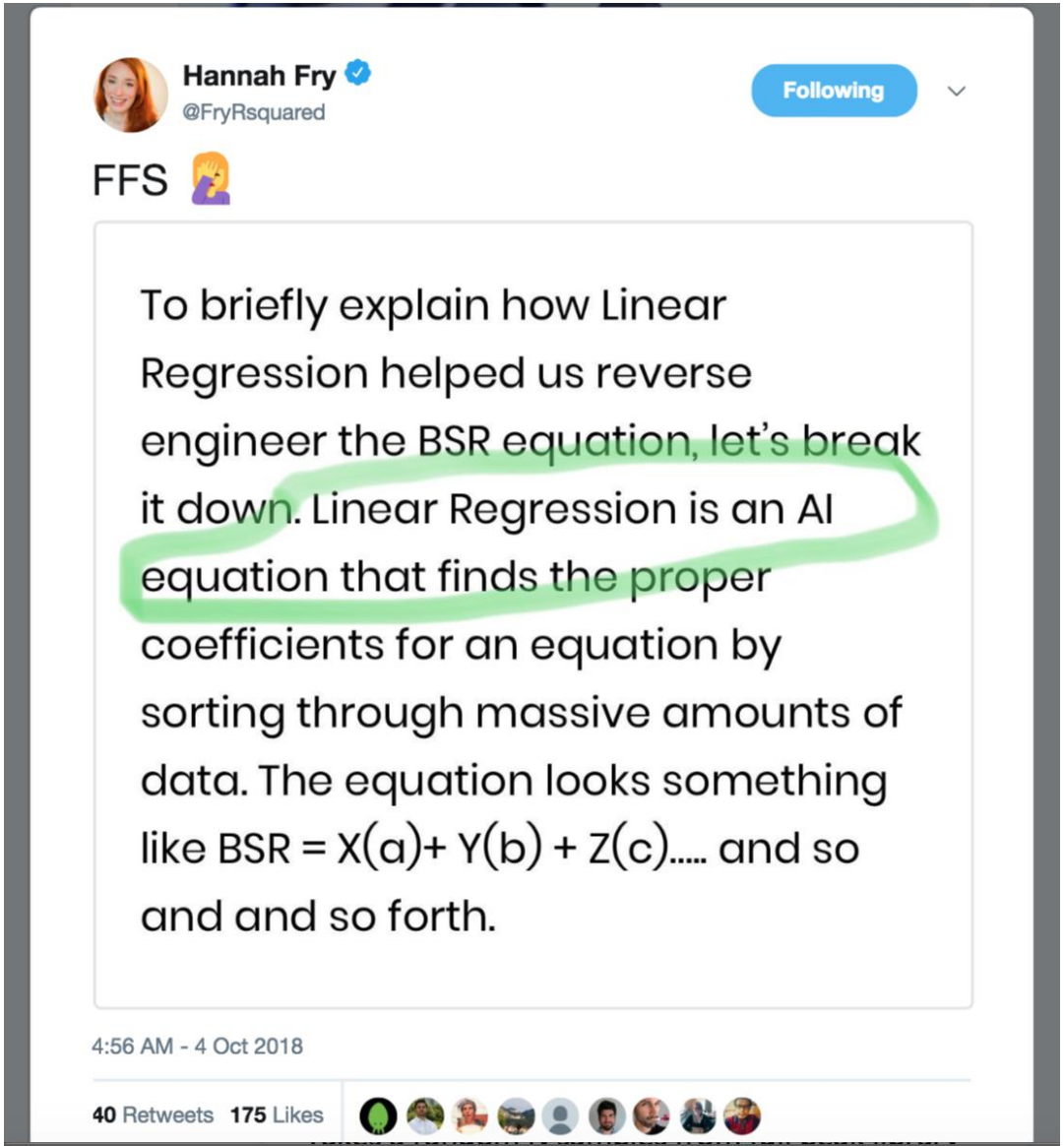


Using Methods from Other “Disciplines”

Using Methods from Other “Disciplines”

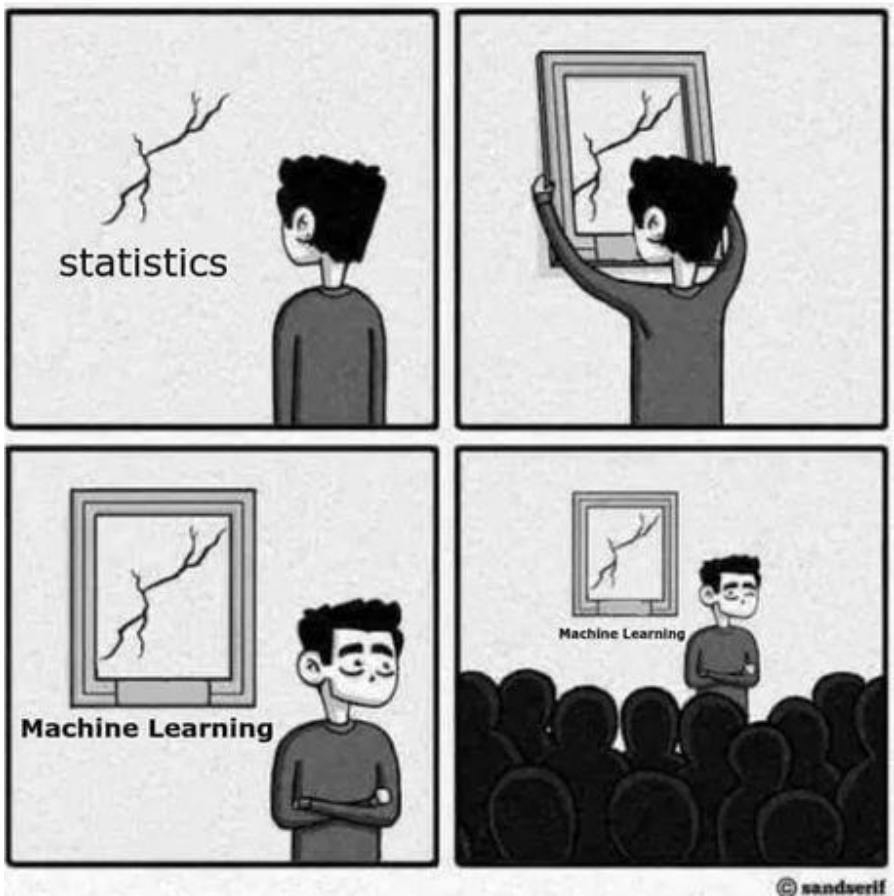
This comes from a website offering Search Engine Optimisation!!!

Hannah Fry did not say this, in case this isn’t clear!



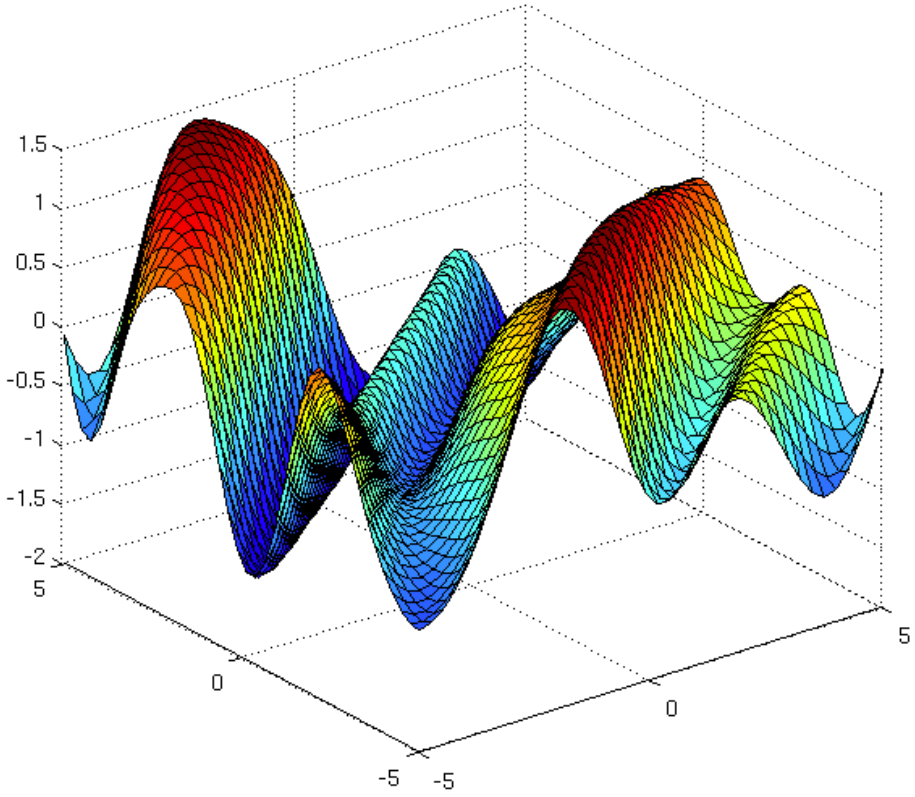
Using Methods from Other “Disciplines”

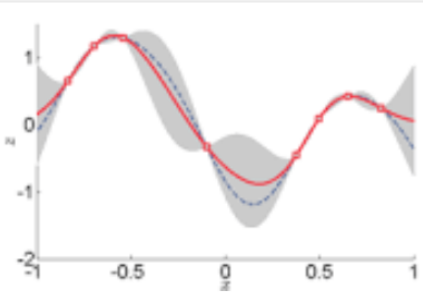
- Machine Learning and AI are often based on Statistical principles but rebranded

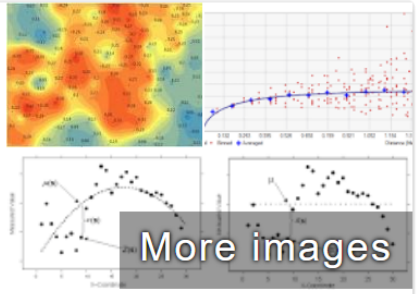


Same Methods, Different Names

Example 1: Gaussian Processes vs Kriging







Kriging

In statistics, originally in geostatistics, kriging or Gaussian process regression is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariances. Wikipedia

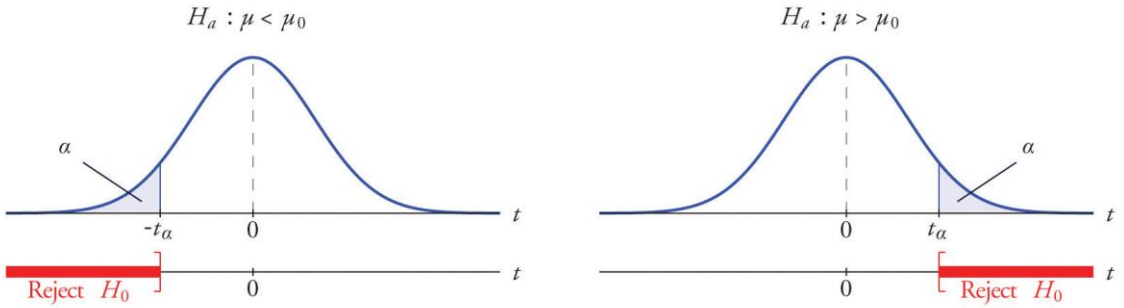
Same Methods, Different Names

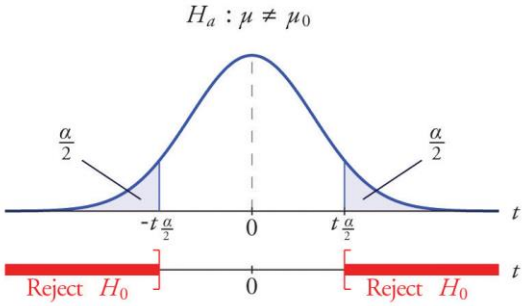
Example 1: Gaussian Processes vs Kriging

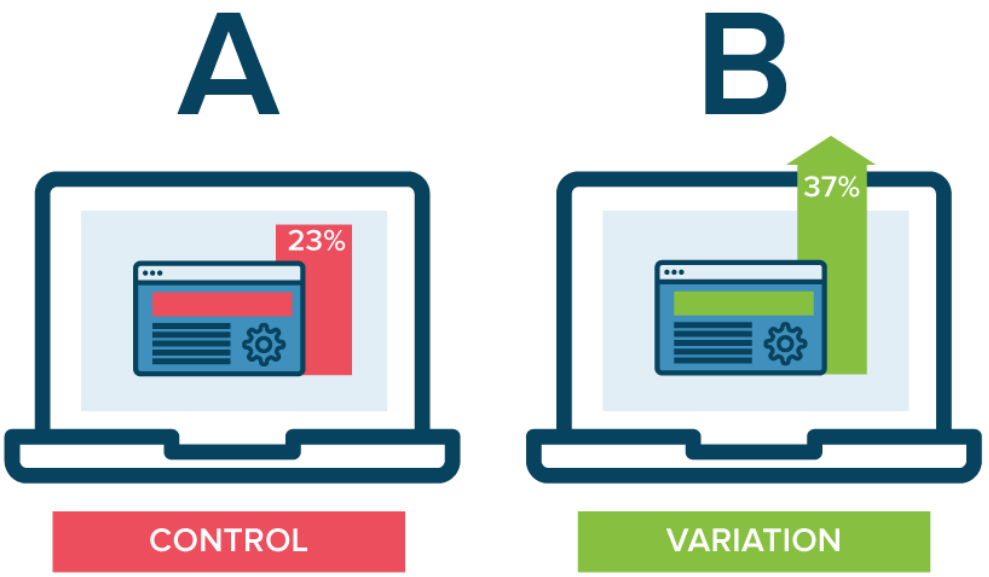
- Gaussian Processes are an extremely popular method in Machine Learning
- Kriging is often used in spatial statistics
- Which is better? Well they are identical…

Same Methods, Different Names

Example 2: Hypothesis Testing vs Ab Testing







Same Methods, Different Names

Example 2: Hypothesis Testing vs Ab Testing

- Statistics use hypothesis testing
    - Generally H_0 vs H_1
- Data Scientists use Ab Testing
    - Generally A vs b
- Ab testing is in principle Statistical Hypothesis Testing wrapped in a coding framework

Same Terms, Different Names

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

Same Terms, Different Names

Glossary

Machine learning	Statistics
large grant = $1,000,000	large grant= $50,000
nice place to have a meeti Snowbird, Utah, French A s	nice place to have a meeting: Las Vegas in August

What?!?! Where is my invite?!?!

Different Inference Techniques

Example 1: Estimating Linear Regression Parameters

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\beta} = \operatorname{argmin} (y - X\beta)^2$$

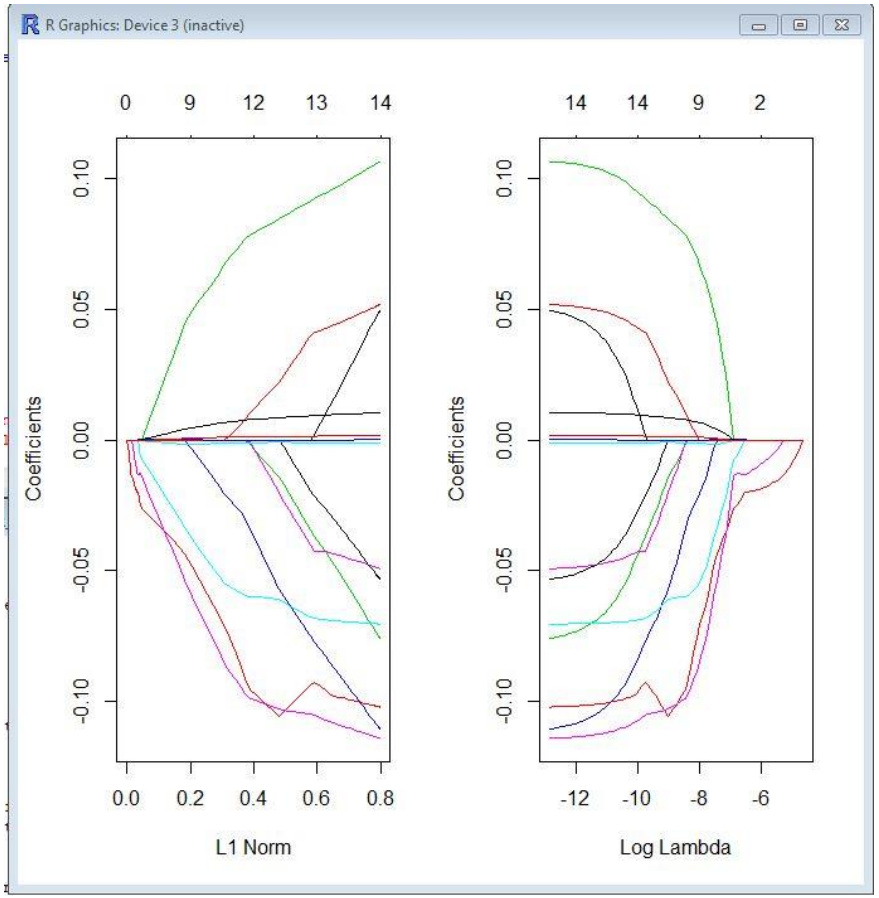
Different Inference Techniques

Example 1: Estimating Linear Regression Parameters

- Statistics traditionally uses OLS (Ordinary Least Squares)
- Machine Learning minimises model error
- Both give the same answer (up to a small error)
- In statistical software, OLS isn't used
    - Minimising model error is faster as no large Matrix inversions

Different Inference Techniques

Example 2: Parameter estimation



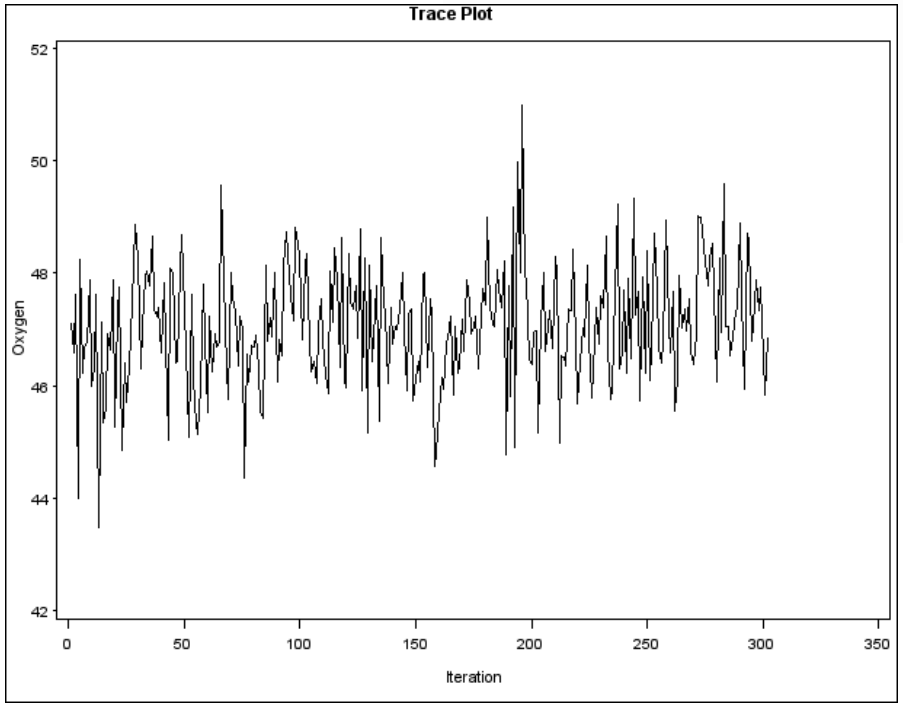
Different Inference Techniques

Example 2: Parameter estimation

- Statistics maximises likelihoods
- Machine Learning minimises cost functions (negative log likelihoods in this context)
- Results are the same…

Different Inference Techniques

Example 3: MCMC, Expectation Maximisation, Expectation Propagation, Variational Inference



Expectation Propagation for Approximate Bayesian Inference

Thomas P. Minka
Statistics Dept.
Carnegie Mellon University
Pittsburgh, PA 15213

Different Inference Techniques

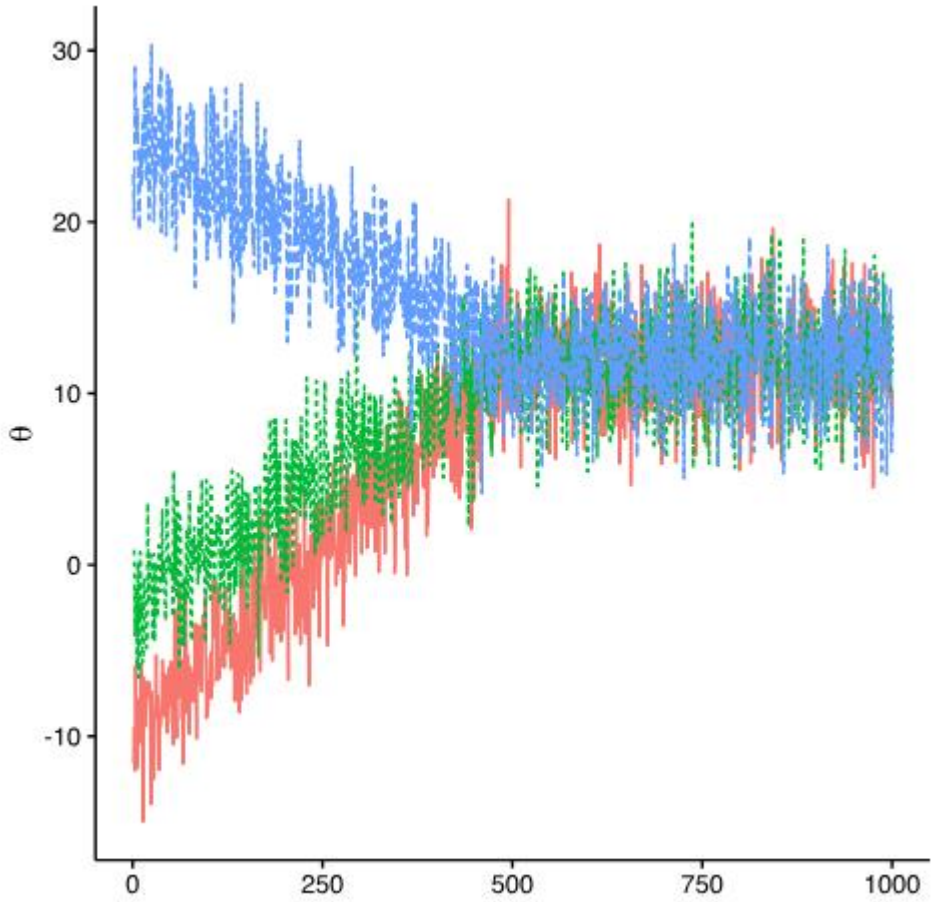
Example 3: MCMC, Expectation Maximisation, Expectation Propagation, Variational Inference

- Statistics uses MCMC and EM algorithm for complex parameter inference

- Machine Learning often prefers faster, potentially less reliable approximation based methods
    - Expectation Propagation, Variational Inference

Different Attitude to Assumptions / Checks

Example 1: MCMC Convergence Checks



Different Attitude to Assumptions / Checks

Example 1: MCMC Convergence Checks

- Statisticians check parameter convergence:

  PSRF / Gelman-Rubin / $\hat{R}$, Geweke, etc

- Machine Learning Professor at ML conference

  "You checked parameter convergence? Wow, you must be the only people here still doing that!"

Different Attitude to Assumptions / Checks

Example 2: Flexible vs Strict

Different Attitude to Assumptions / Checks

Example 2: Flexible vs Strict

- Statistics often want asymptotic results before using a method (i.e. method works in theory for large *N*)
- Machine Learner argue that you will never have large enough *N* and therefore don't require these proofs
- Machine Learners tend to take the attitude that whatever works works

Different Goals – Prediction & Explanation

- Goals can differ between Statistics, Machine Learning, Data Science and AI

Different Goals – Prediction & Explanation

- Goals can differ between Statistics, Machine Learning, Data Science and AI
- AI generally prioritises prediction, Statistics prioritises explanation, Machine Learning and Data Science do a bit of both
- Problems occur where Applied Scientists move from explanatory methods they semi-understand to predictive methods they 100% don’t

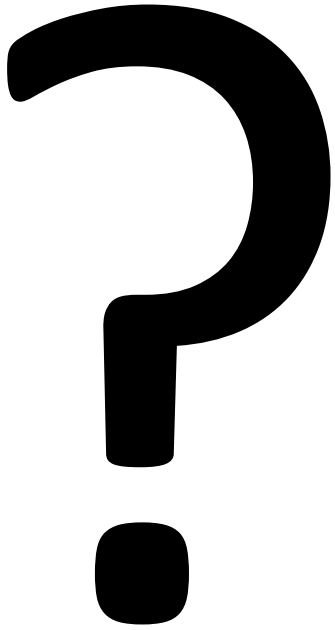
Different Goals – Prediction & Explanation

- Adding interpretability in Neural Networks (NNs) is a big area of research in Machine Learning now

- Why is does the NN think this image a cockatoo?

Zintgraf, L. M., Cohen, T. S., Adel, T., & Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595.

Different Programming Languages



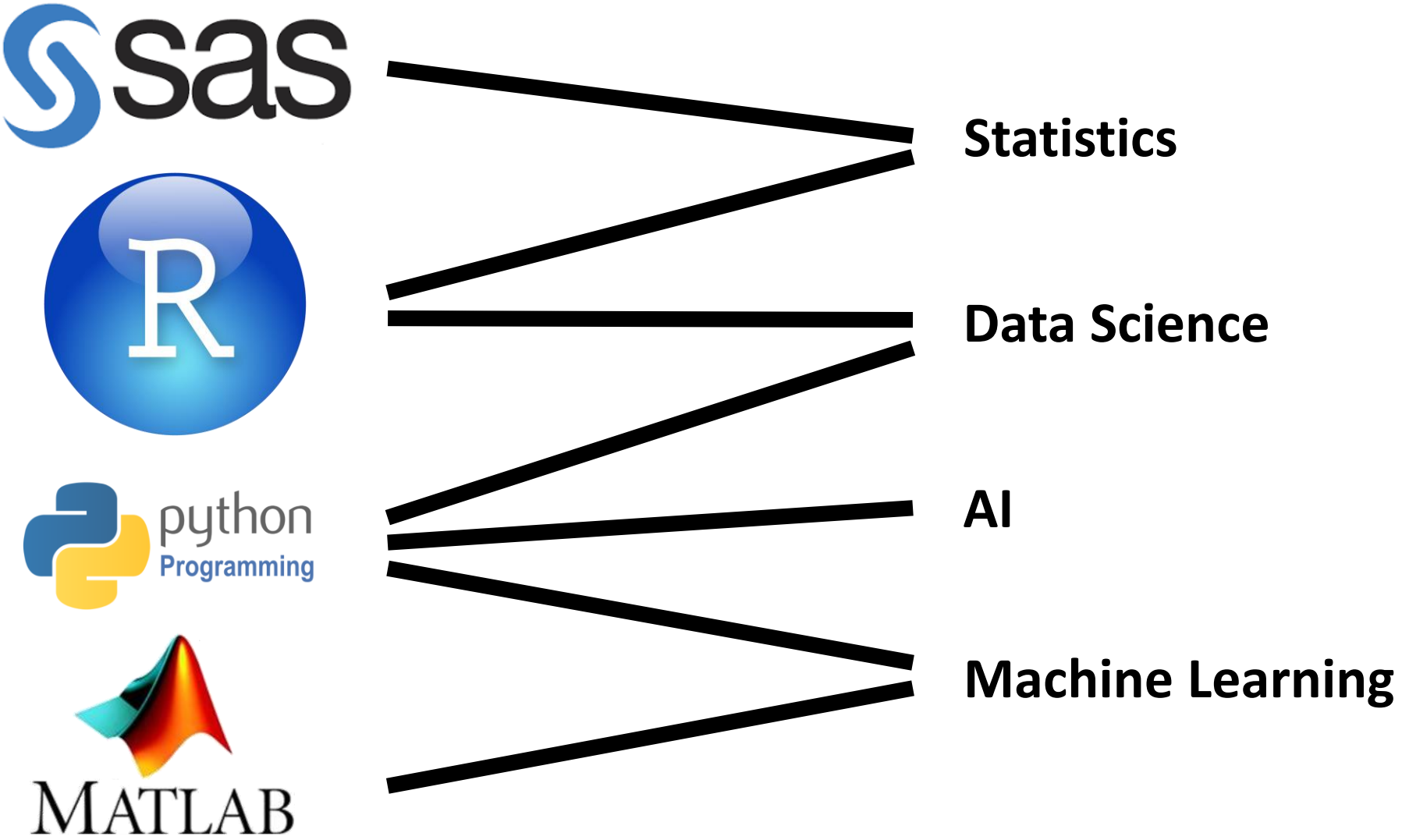
Statistics

Data Science

AI

Machine Learning

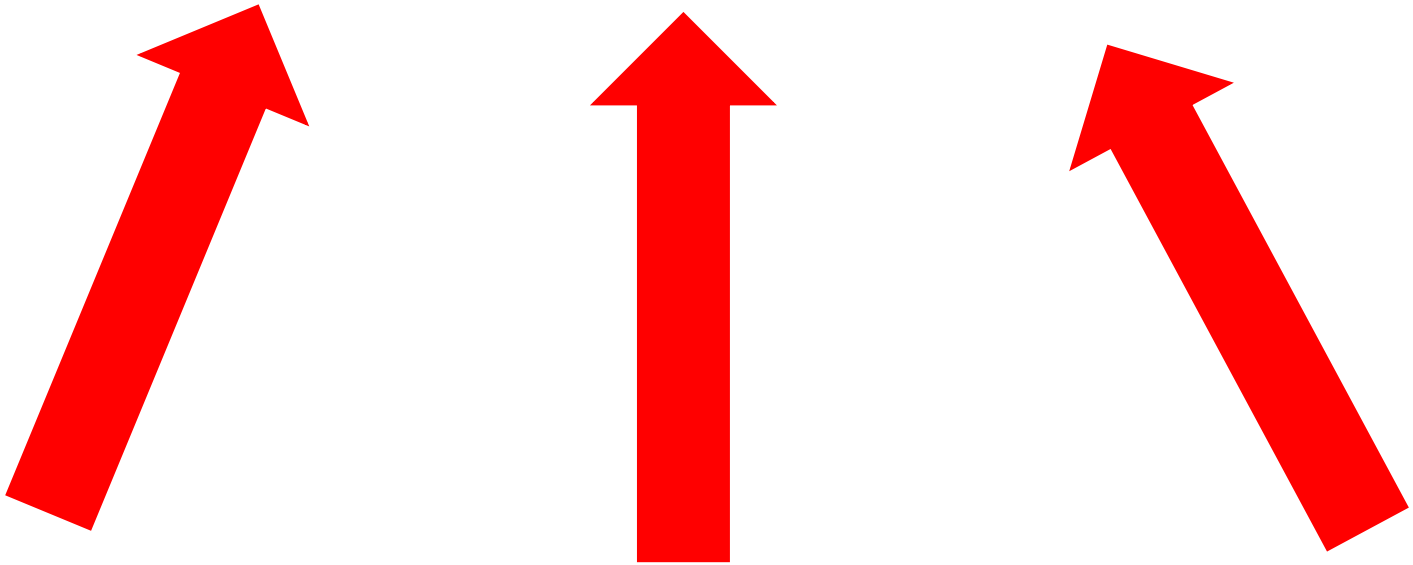
Different Programming Languages



Which is best?

Which is best?

It depends on the situation!!!



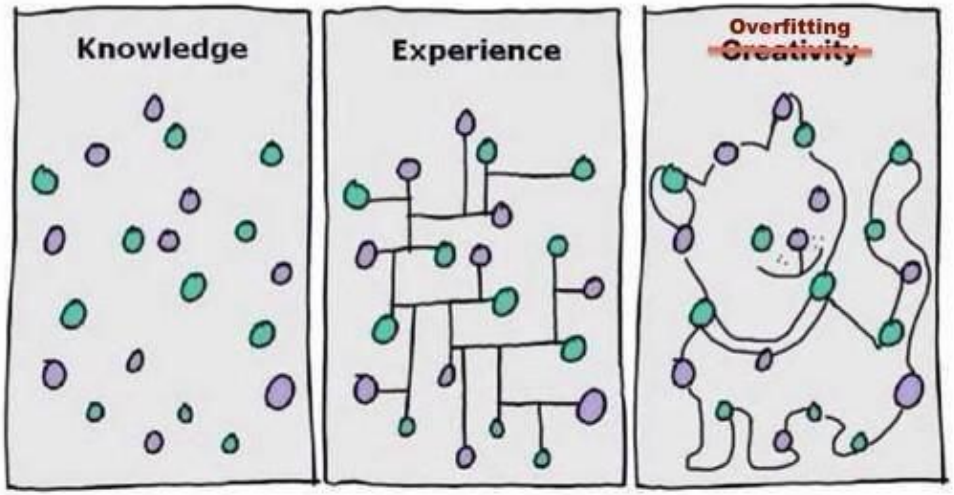
Which is best?

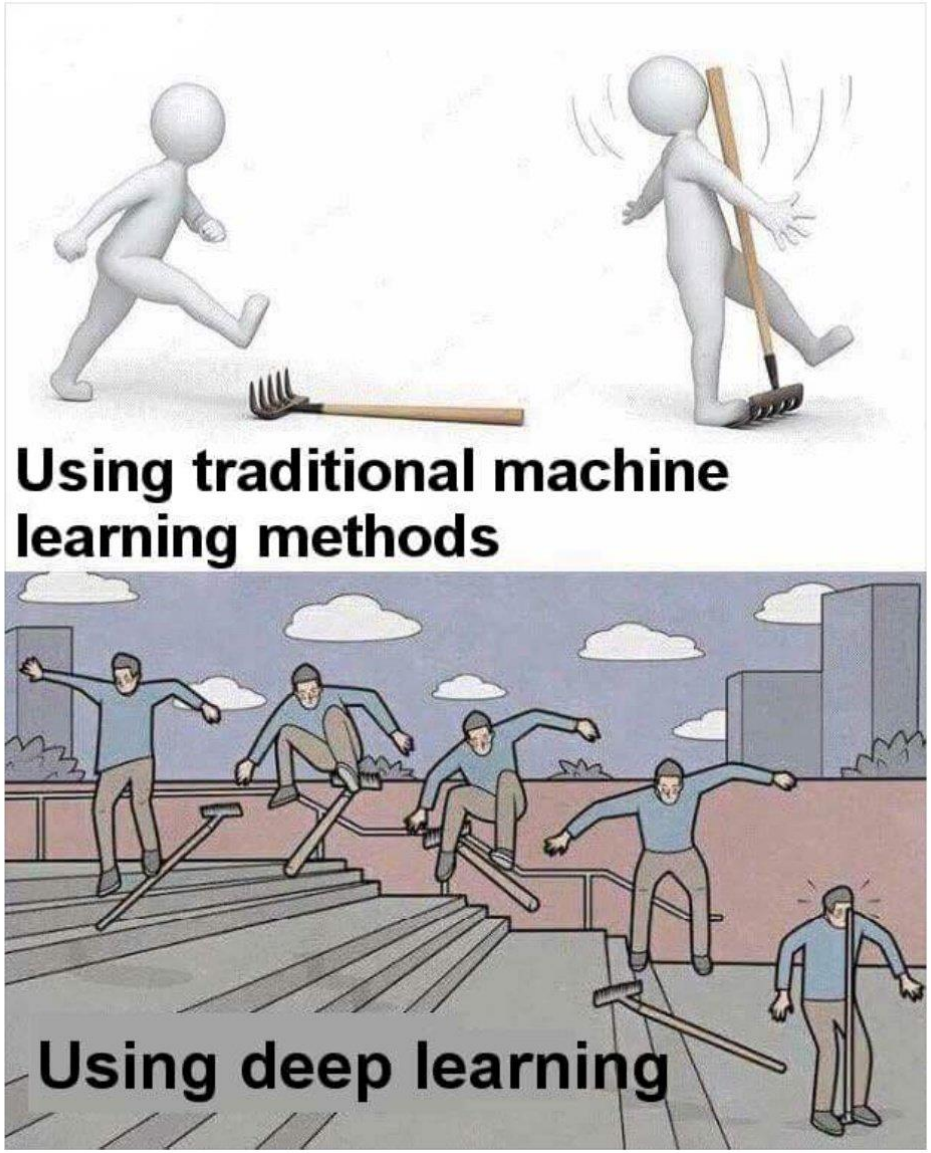
It depends on the situation!!!

- Just because it AI or Machine Learning doesn't mean its better than Statistics
    - Half of the time it's the same thing!

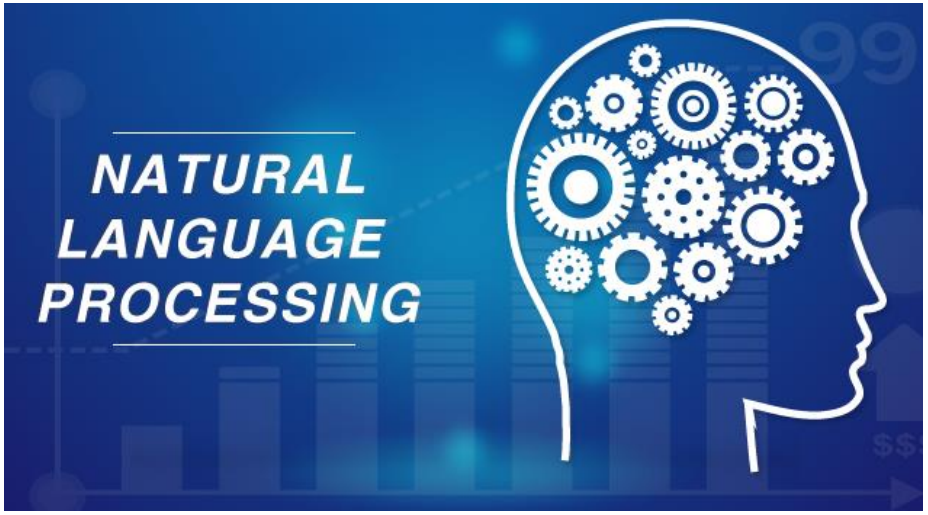
- A Statistical approach isn't always simpler than an AI or Machine Learning Approach

Complicated isn't always best



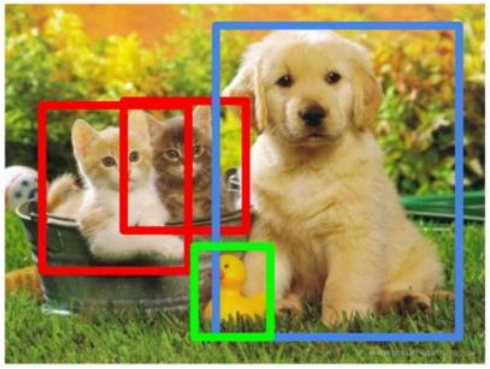


When is AI / Deep Learning traditionally used?



CAT

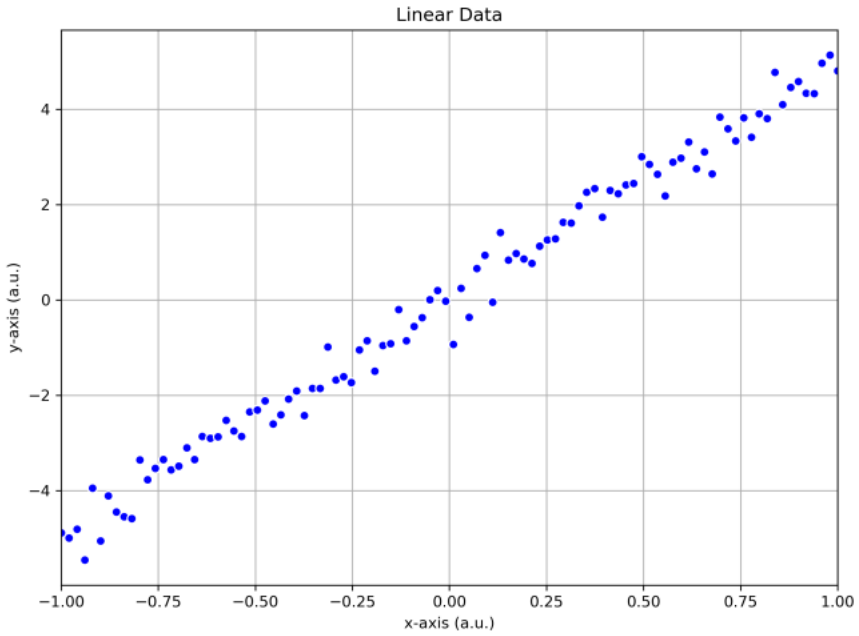
CAT

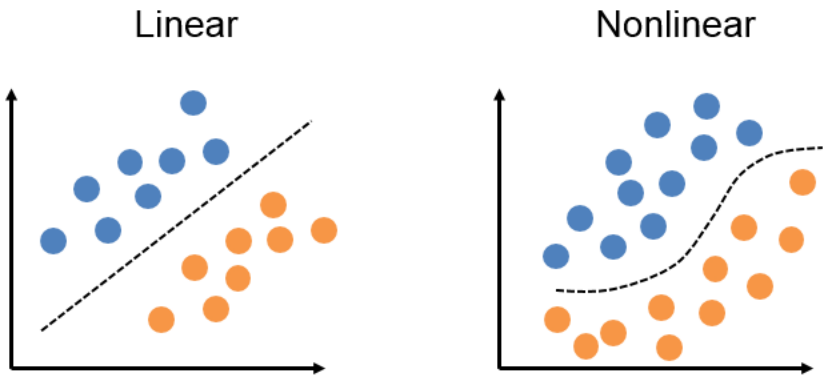


CAT, DOG, DUCK

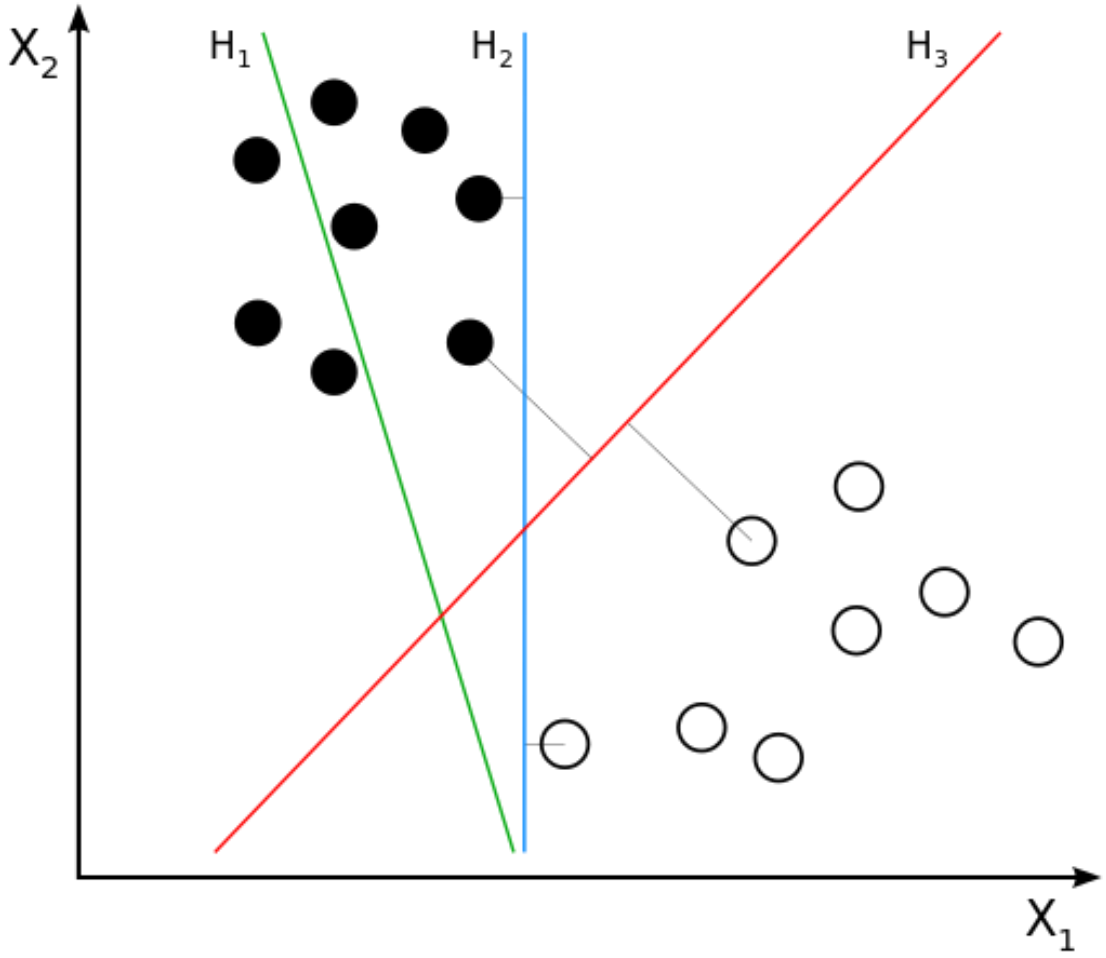
CAT, DOG, DUCK

When not to use AI / Deep Learning



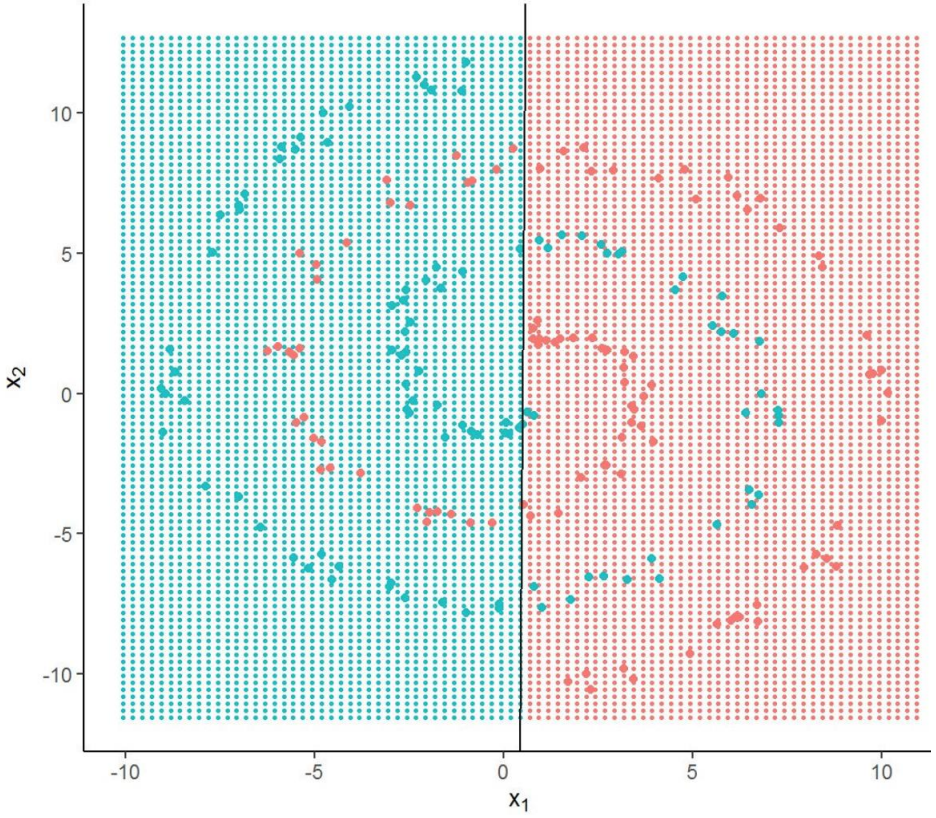


Sometimes Machine Learning does work best

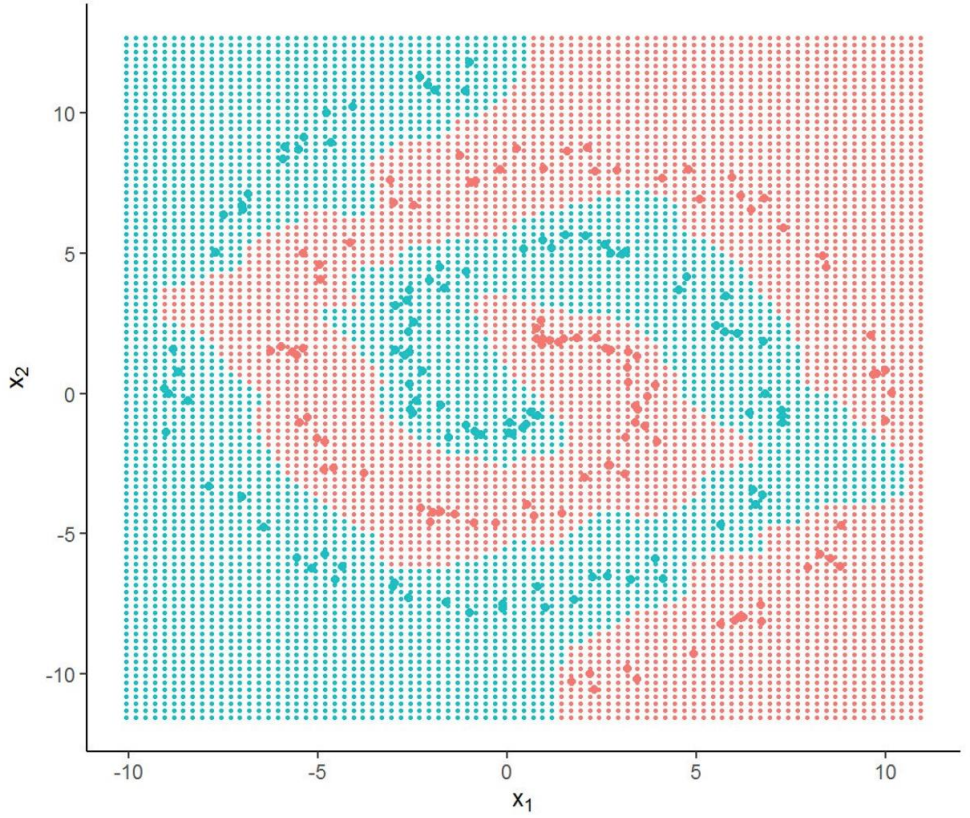


Sometimes Machine Learning does work best

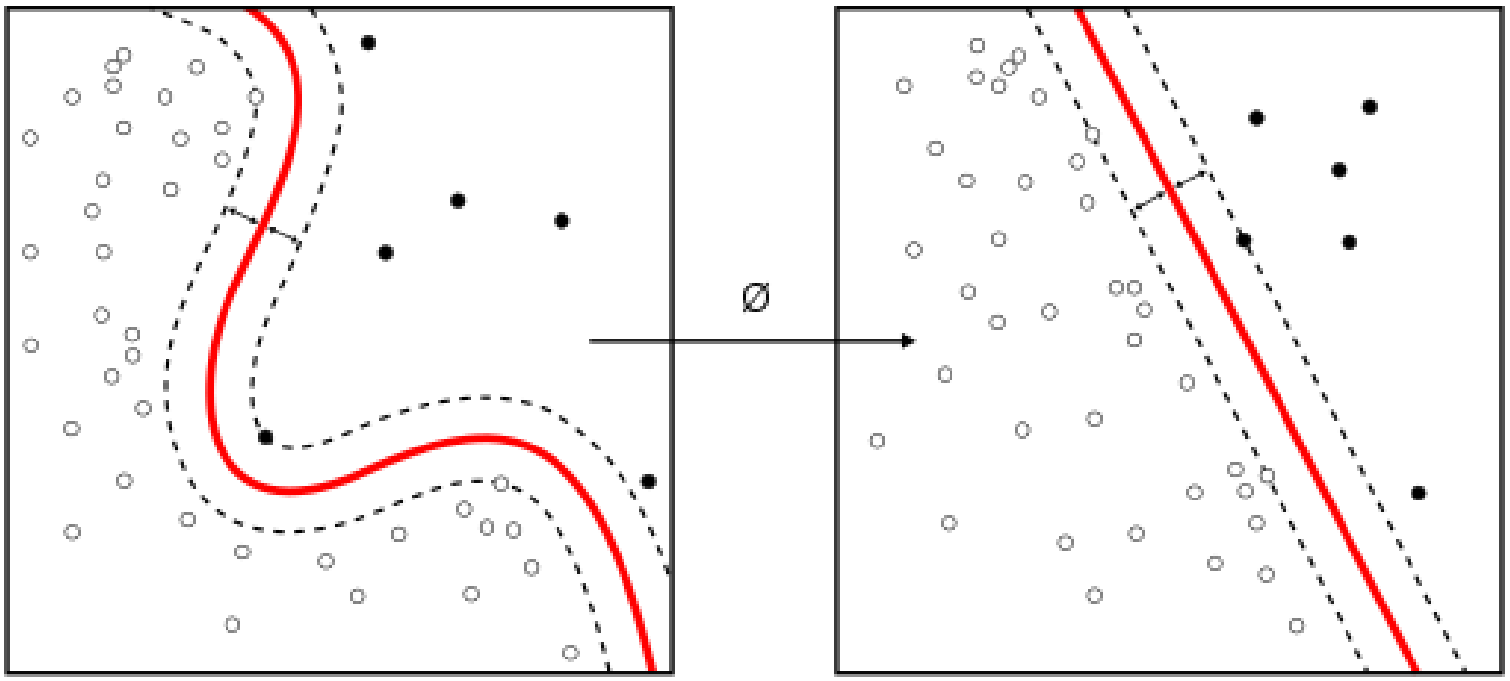
Logistic Regression



Neural Network

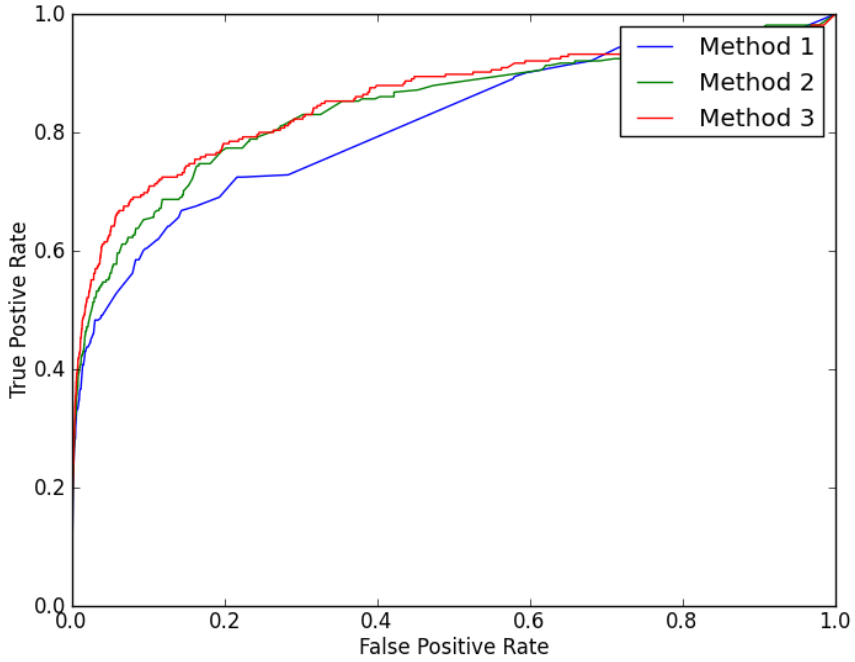


Sometimes a non-linear method is sufficient



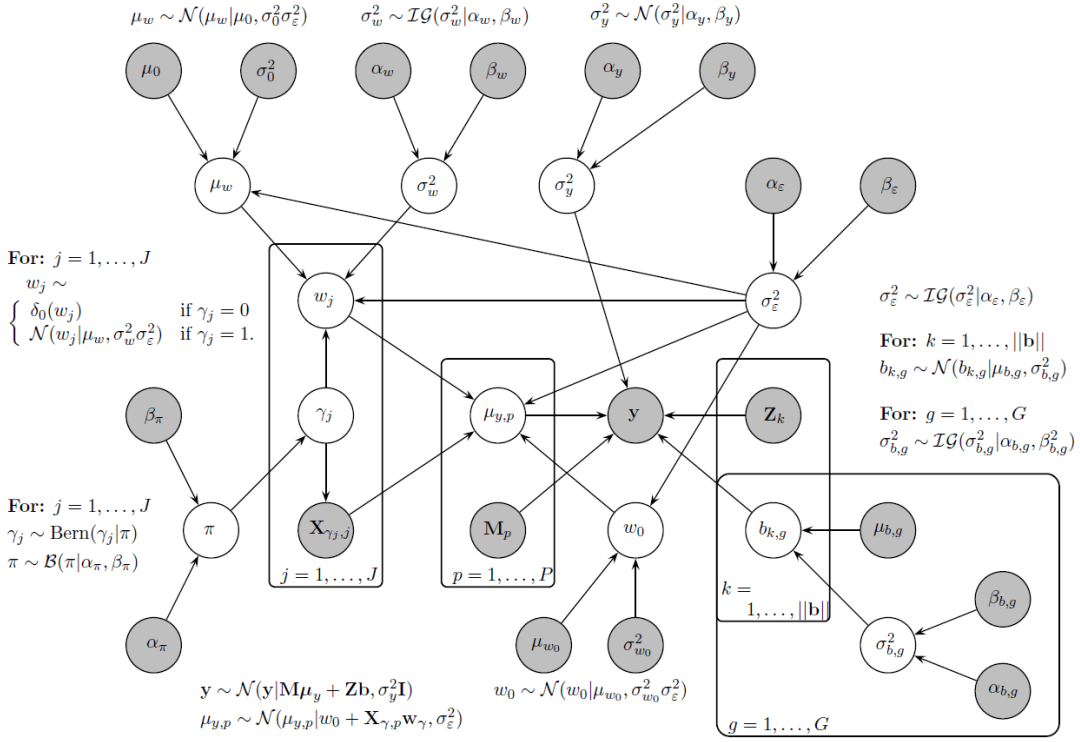
(Always?) Try Statistics First?

- Is it possible to fit a statistical model with some predictive ability? Yes? Then do it!
- At worst, Statistical models provide a performance baseline



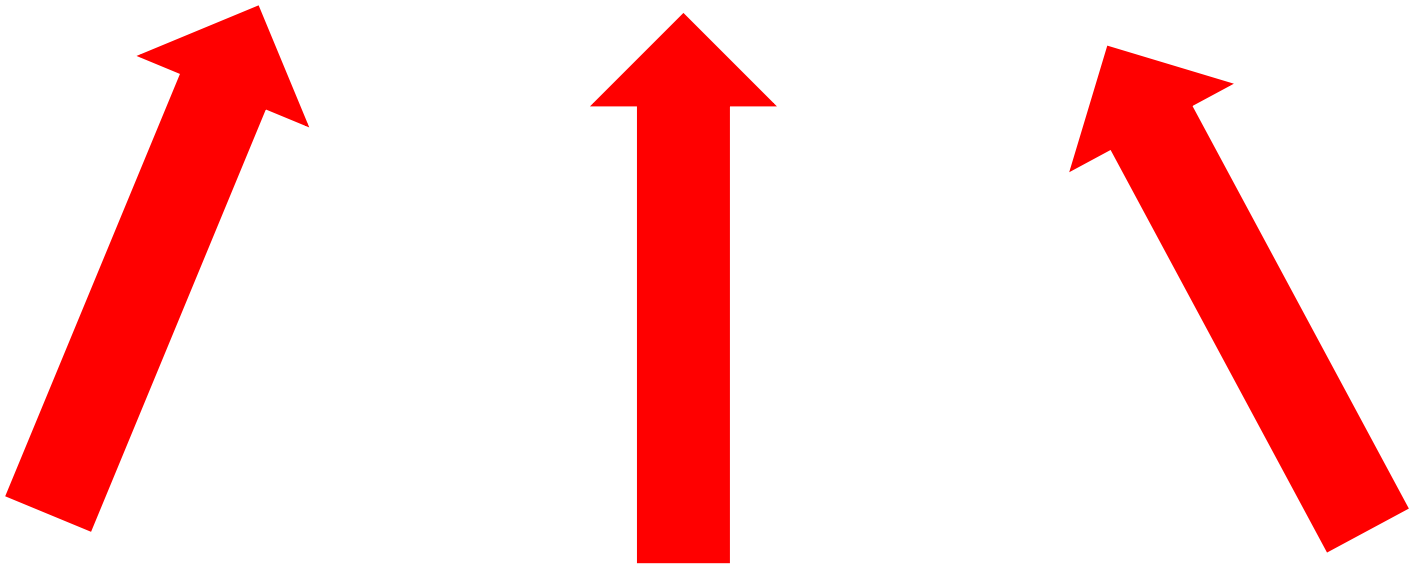
Sometimes Statistics can be complex

- There are many interpretable Statistical methods that can compete with AI / Machine Learning
  - Hierarchical Bayesian models
  - Generalised Additive Models
  - Elastic Net
  - Splines
  - Etc etc etc



Still want AI, but don't know how?

Talk to some experts!!!



Still want AI, but don't know how?

Talk to some experts!!!

- Important to teach people:
  - Talk to an expert
  - Don't just apply ML / AI
  - Higher AUC / MSE doesn't make it a better method if you have applied it incorrectly
  - Its easier to go wrong with ML / AI and the consequence are more severe

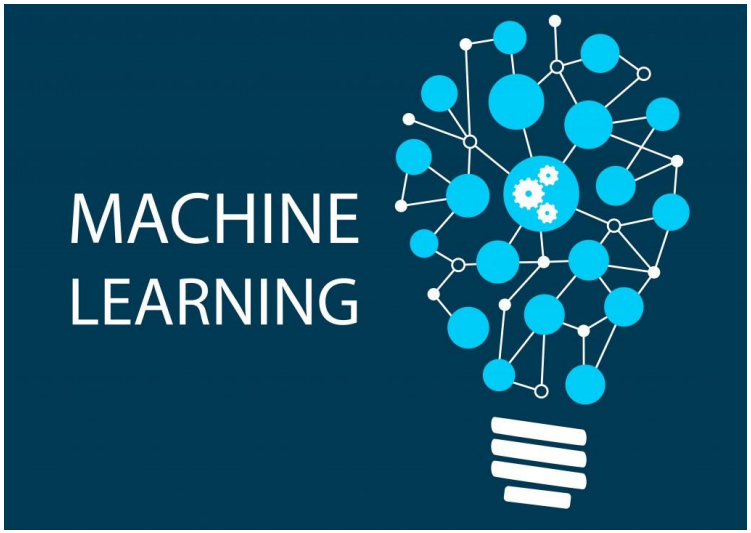
So what is the difference between the fields?

- All use the same methods, some methods are just more popular in one field than another
- Different attitude
    - ML, Data Science & AI are all more modern fields than statistics. They tends to be more relaxed about the assumptions
    - But would you want a Machine Learner analysing a clinical trial? No
- Different background / skillset
    - People in different fields tend to come from different backgrounds and therefore have different skills and expertise
- Different in practise more than different in theory

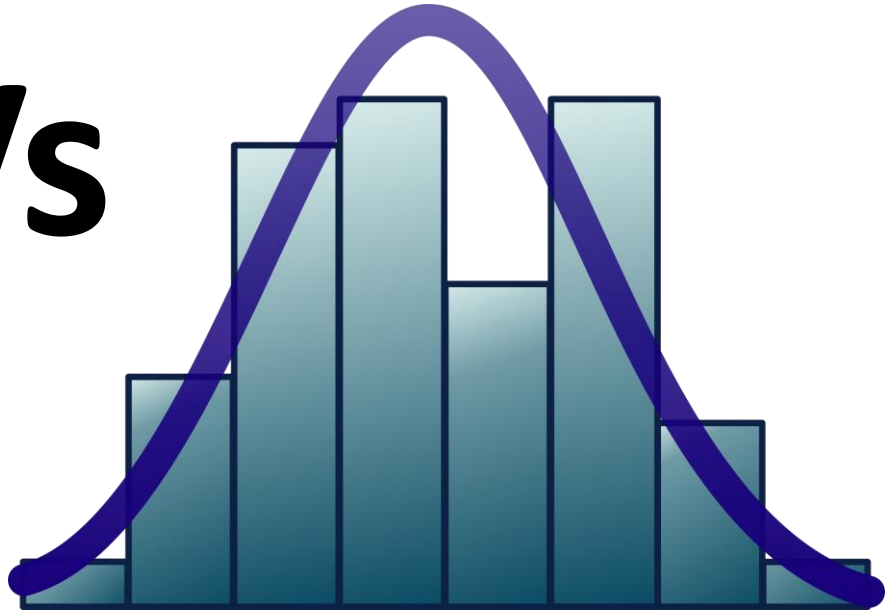
How can Statistics stay relevant?

Why Machine Learning, not Statistics?

- Strange question to ask, but it tells us what we do well and what we don't



Vs



Programming Skills

- To keep up with Machine Learners, Statisticians need to get better at programming!

Programming Skills

- To keep up with Machine Learners, Statisticians need to get better at programming!

- Statistics courses need to do more programming and at a higher level
    - Teach Python as well as R
    - Git version control?
    - Teach SQL? (One day extra course?)
    - Good coding practises?
    - C++, Java?

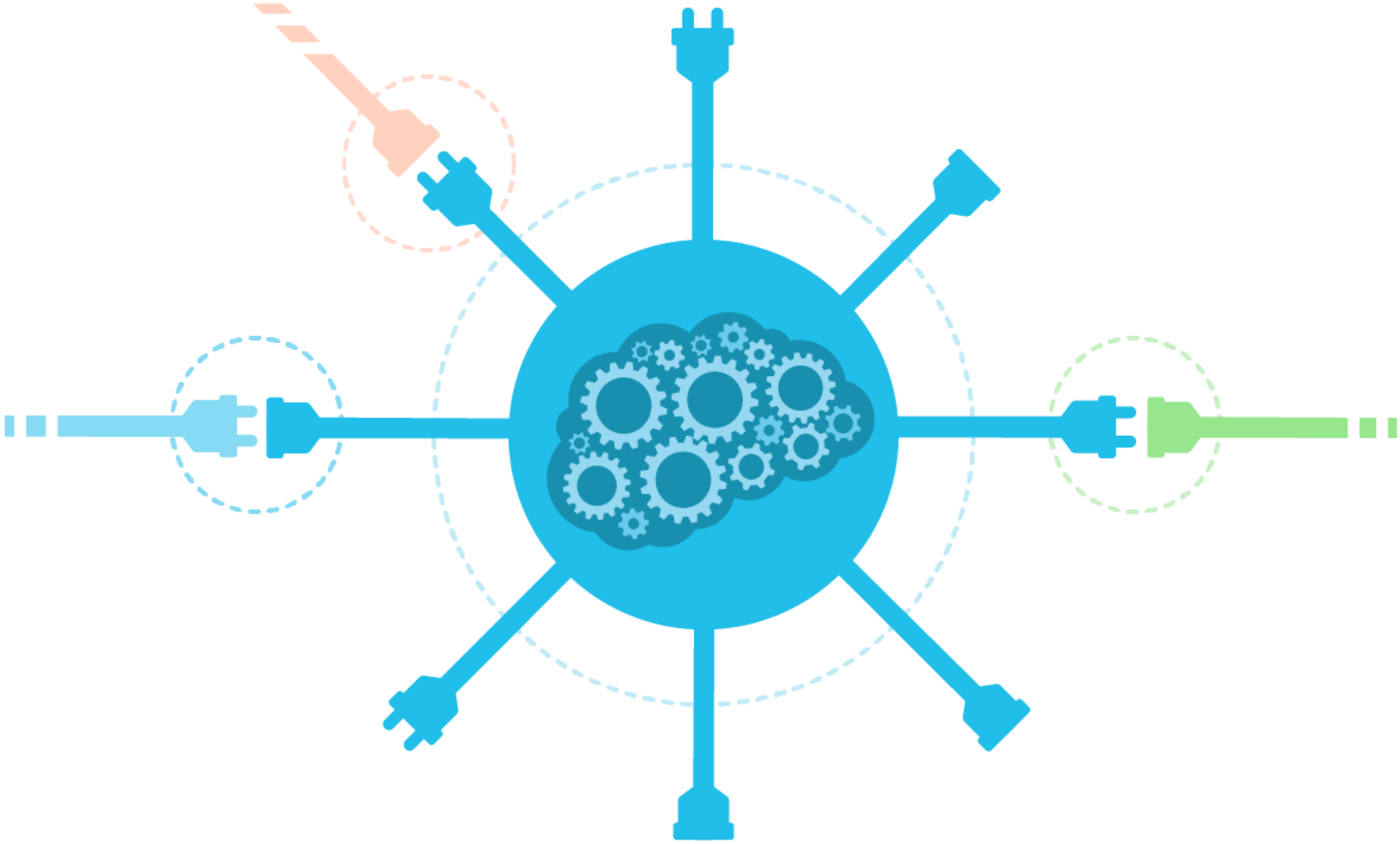
Outreach

- Have you done something cool with Statistics?
  – Tell us about it!

- How?
  - Interdisciplinary research article
  - Magazine Article
  - Social media

- Encourage students to talk about their work!

Implement our work



Implement our work

- Machine Learners fit a simple or non-specific complex model and then implement it in practise

- Statisticians often fit a better model, present the results, then…

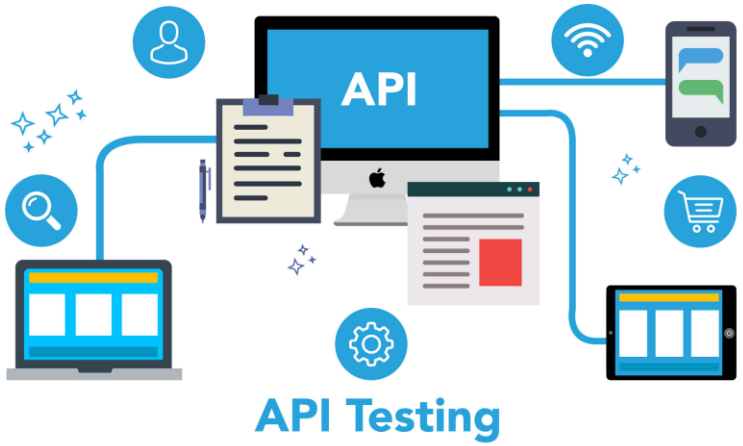
Implement our work

- Machine Learners fit a simple or non-specific complex model and then implement it in practise

- Statisticians often fit a better model, present the results, then…

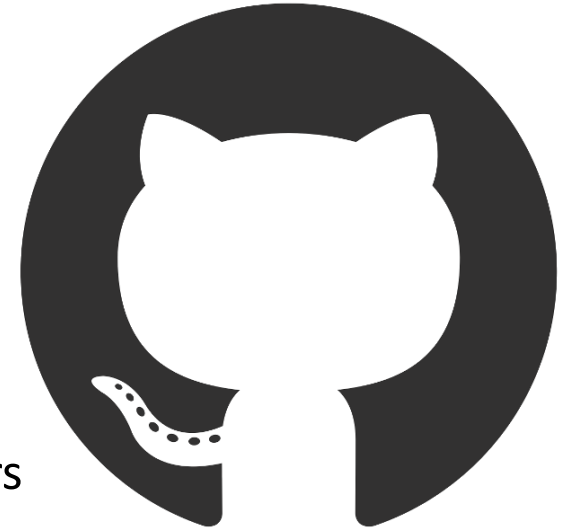
Implement our work

- **Example 1: Vinny is an idiot…**
- Two years after my Ph.D. finished and my code wasn't even on GitHub (it now is)
- 1 month in Computer Science and computer interface with the laboratory instrument has been discussed in every meeting



Implement our work – How can we get better at this?

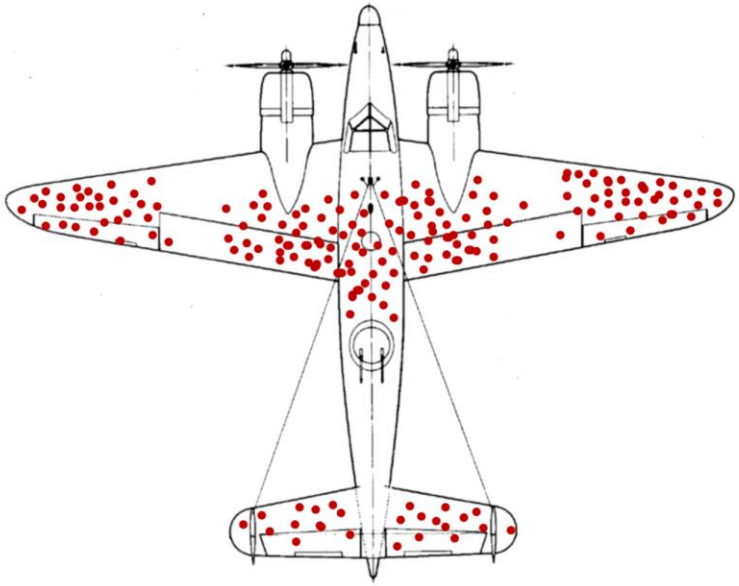
1. Put your code on GitHub
    - Requires writing cleaner code in the first place!
2. Help others implement it
    - Make sure its useable for collaborators
    - Answer GitHub queries



3. Allow time in project to implement the work in practise
4. Teach students Git version control

Why should companies employ Statisticians?

- Better understanding of data
    - Statisticians are more likely to understand censoring, missing data etc
- Better at describing data
    - More likely to give relevant summaries and plots
- Think more about causal implications of modelling
    - Models generated more likely to useable in practise



How can companies improve job adverts?

- Work out what they actually want, don't just write an unrealistically long wish list
    - Talk to someone if you are unsure
- Be sensible with your candidate expectations
    - If you don't offer much money, you wont get the top quality candidates (if you do, they are probably embellishing their CV…)
- Don't just put the most techy words you can think of…
    - It will be obvious you are clueless…
- Be clear how much software development is involved
    - You don't want to hire a statistician for a programming role

Am I a Statistician or A Machine Learner?

How I Feel in a Computing Science Department

What am I?

- I’m still a Statistician at heart!
- I’m a Machine Learner as well, but only because I think they are almost the same!
- Trying to get more coding / software development experience to become better at
- I would always use “Statistical Methods” first, “Machine Learning / AI Methods” second

Thanks for listening!