

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”?

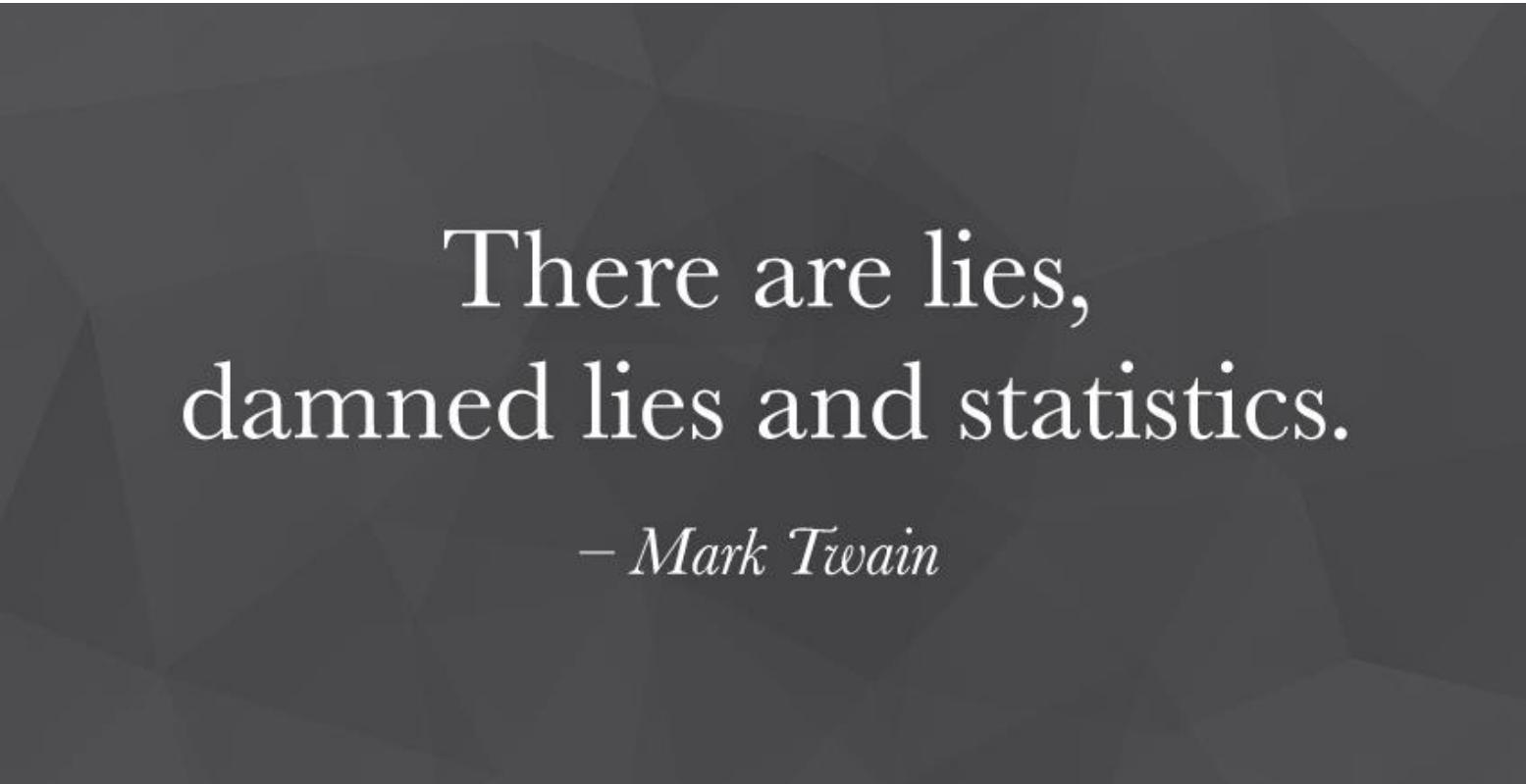
- Worked in Statistics, Applied Statistics and Machine Learning
- Publications and conferences in both Statistics and Machine Learning
- Used and developed methods and techniques from all areas



What do people think they are?



Statistics



There are lies,
damned lies and statistics.

– Mark Twain

Statistics

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

Statistics

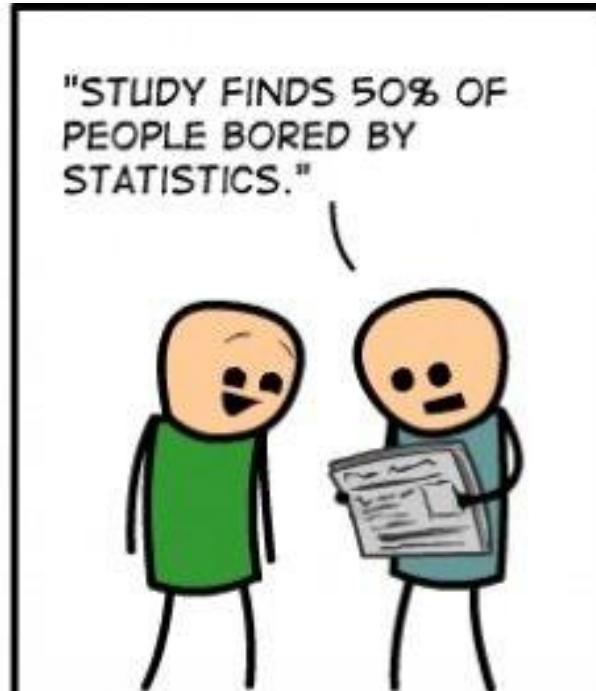
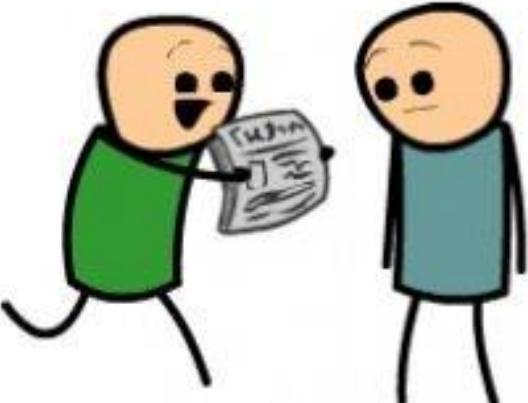
- Everyone in research people like it

I HATE 
STATISTICS

HOLY SHIT, MAN!!
LOOK AT THIS!!

"STUDY FINDS 50% OF
PEOPLE BORED BY
STATISTICS."

VIA 9GAG.COM



Statistics

- Everyone in research needs Statistics, but not many people like it
- Media coverage gives a bad impression

Statistics

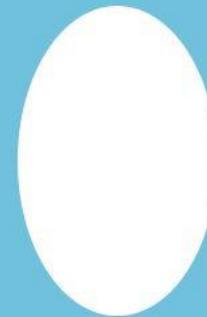
STATISTICS SHOW THAT TEEN PREGNANCIES DROP
DRASTICALLY AFTER THE AGE OF 20



is Statistics, but not many

I MISLEADING STATISTICS

THE AVERAGE ADULT HAS
ONE TESTICLE



Meaningful > Just True

CONTENT SHOULD BE USEFUL, NOT JUST PRETTY
vert.ms/Baddata



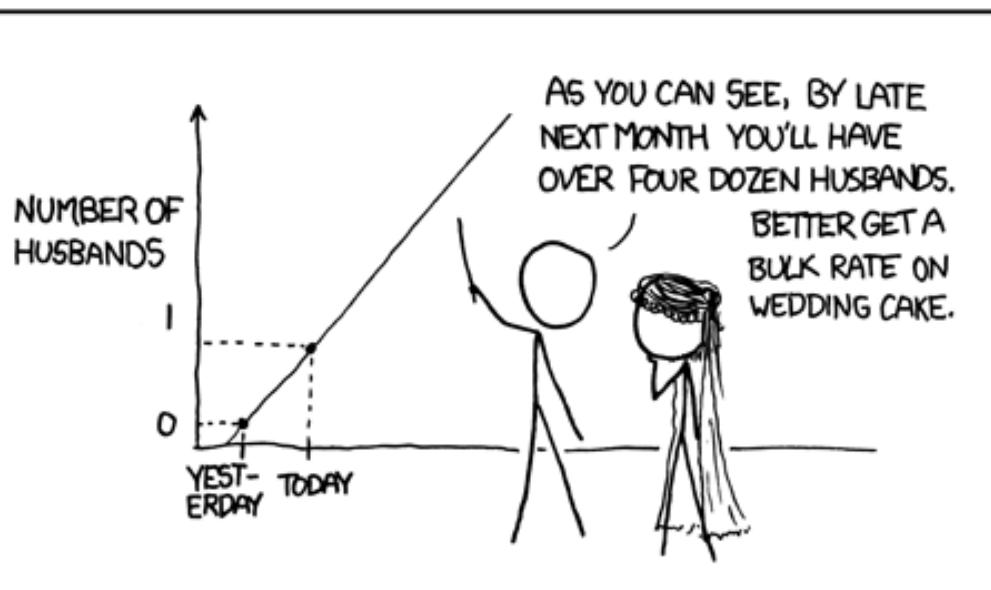
Statistics

- Everyone in research needs Statistics, but not many people like it
- Media coverage gives a bad impression
- Statistics get blamed for terrible data summaries

Statistics

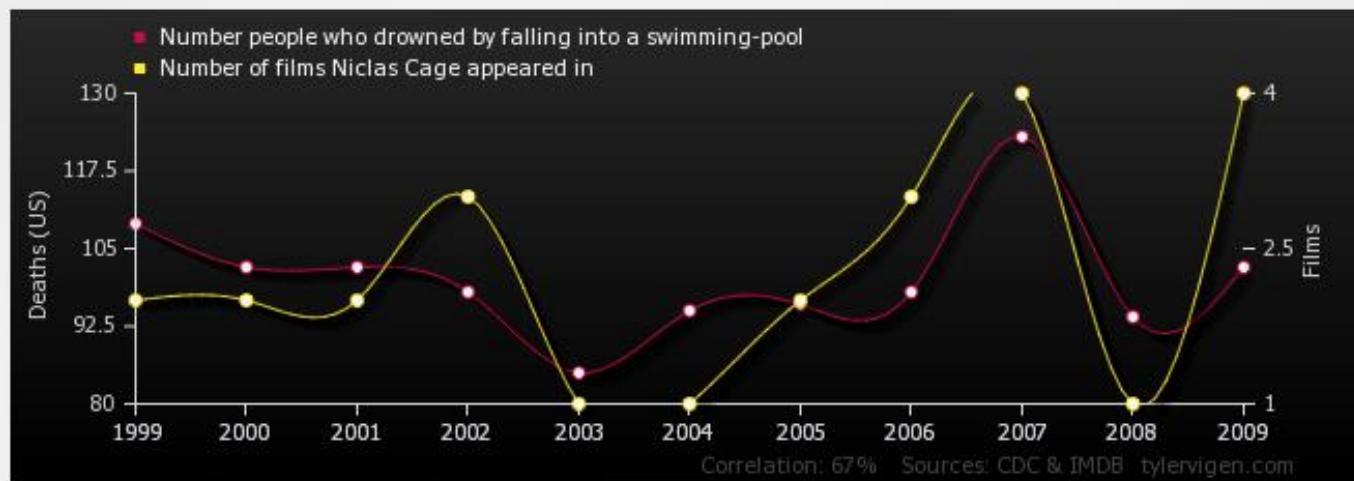
- Everyone
people

MY HOBBY: EXTRAPOLATING



not many

- Number people who drowned by falling into a swimming-pool correlates with
Number of films Nicolas Cage appeared in



Upload this image to imgur

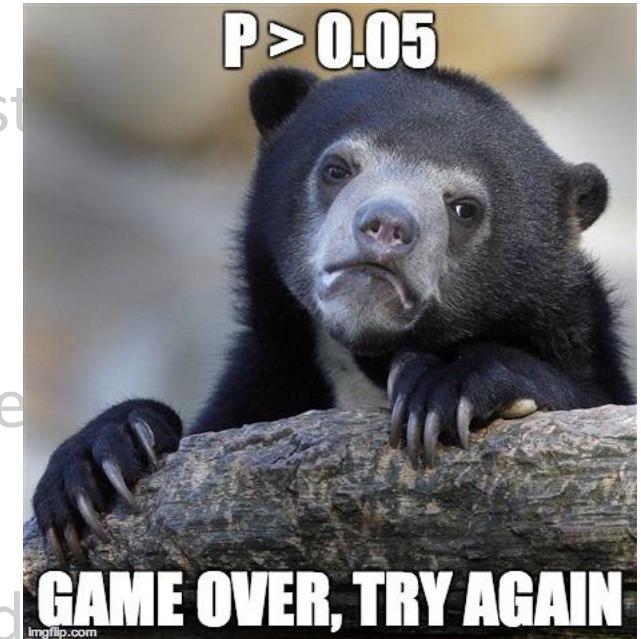
Statistics

- Everyone in research needs Statistics, but not many people like it
- Media coverage gives a bad impression
- Statistics get blamed for terrible data summaries
- Statistics has a reputation for being basic

Statistics

- Everyone in research needs Statistics
people like it

- ???
??
??
??
• ???
??
H₀ ?????????????????????????????????????
??
??
• ?????????????????????????????????????
????????????????????????????????????
H_a ?????????????????????????????????
??
??



Machine Learning



A breakthrough in machine learning
would be worth ten Microsofts.

— *Bill Gates* —

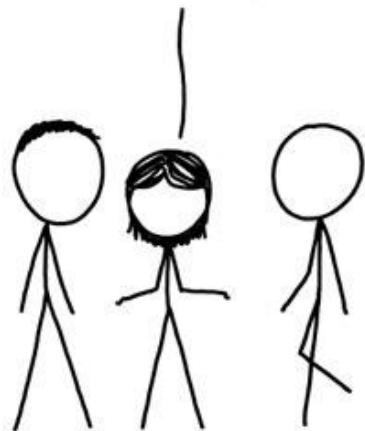
AZ QUOTES

Machine Learning

- People think Machine Learning can solve all their problems

Machine Learning

OUR FIELD HAS BEEN
STRUGGLING WITH THIS
PROBLEM FOR YEARS.



STRUGGLE NO MORE!
I'M HERE TO SOLVE
IT WITH ALGORITHMS!



SIX MONTHS LATER:

WOW, THIS PROBLEM
IS REALLY HARD.

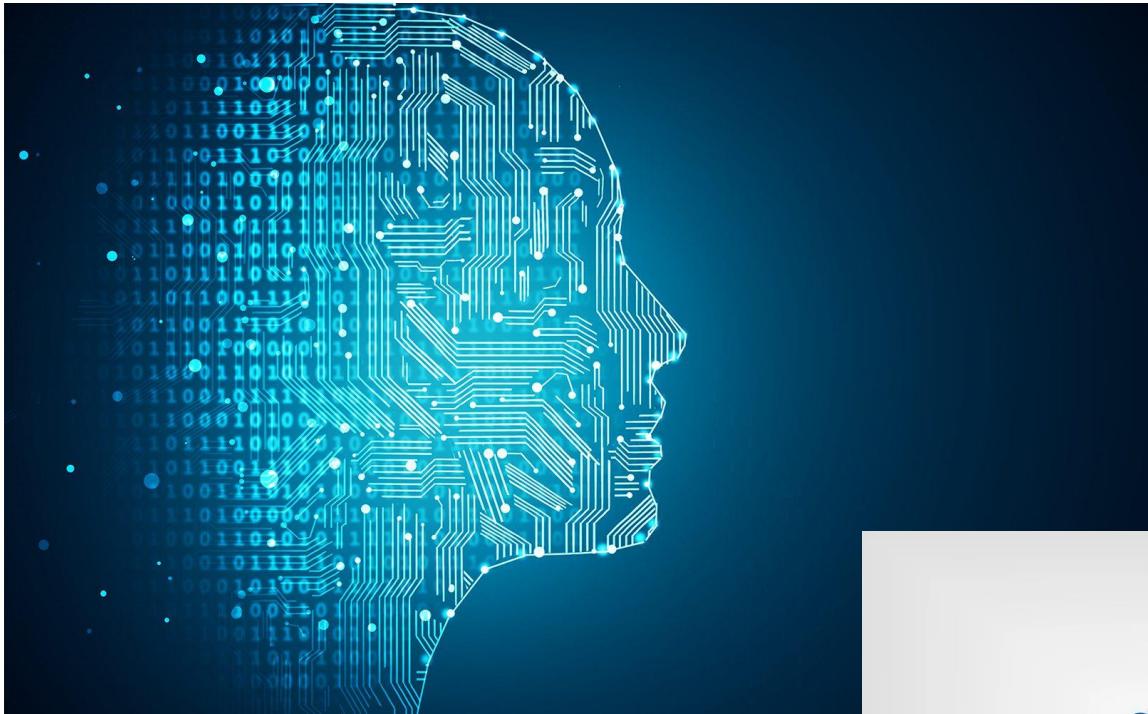
YOU DON'T SAY.



Machine Learning

- People think Machine Learning can solve all their problems
- People think Machine Learning is more advanced

Machine Learning



solve all their

more advanced



MACHINE LEARNING

Machine Learning

- People think Machine Learning can solve all their problems
- People think Machine Learning is more advanced
- Machine Learners don't always check assumptions
 - Good or Bad?

Machine

- People
problem
- People
- Machine
– Good

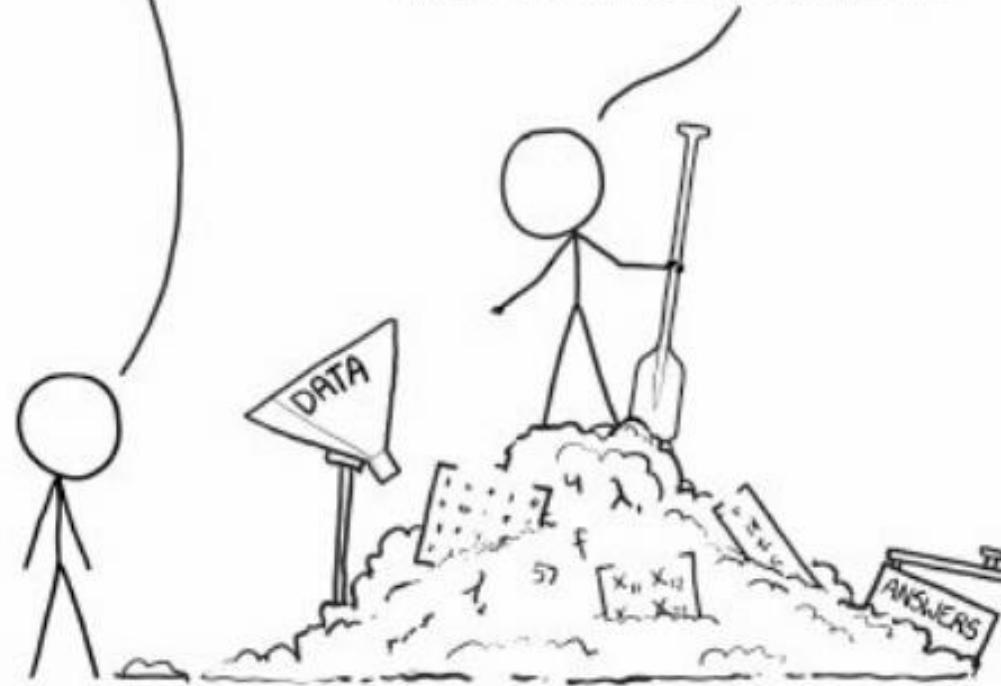
their
nced
ptions

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.

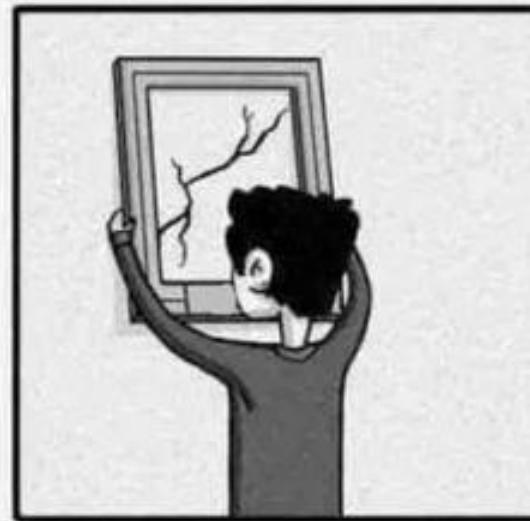
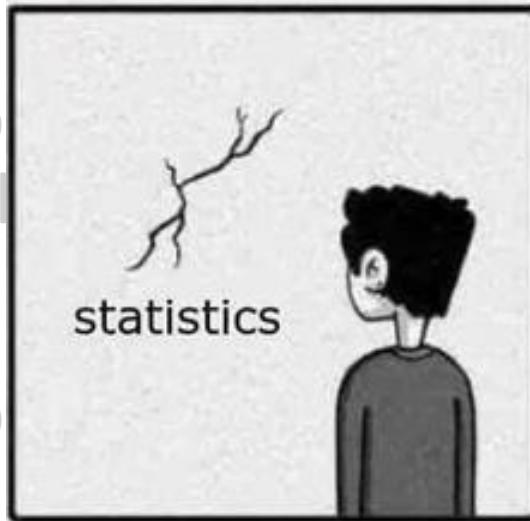


Machine Learning

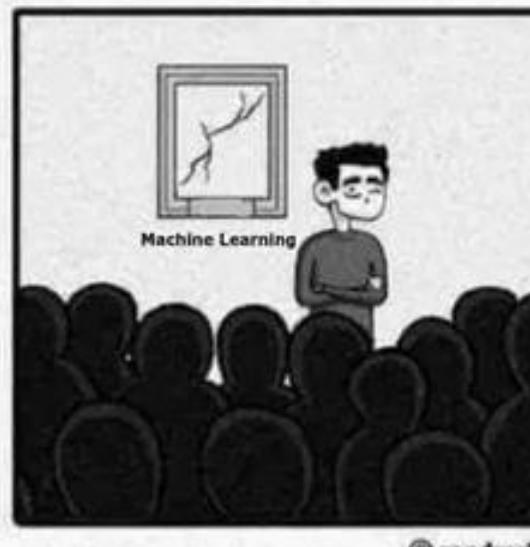
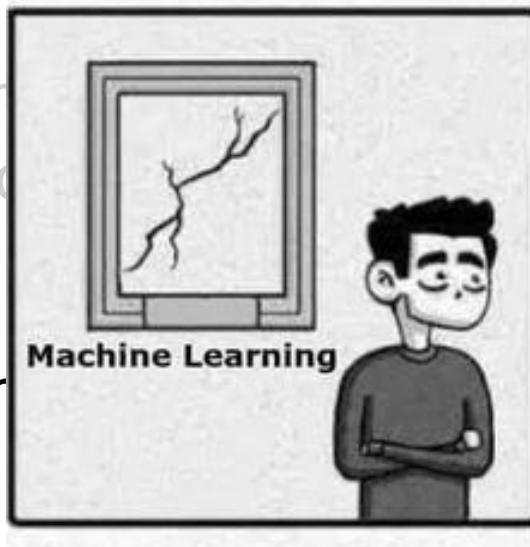
- People think Machine Learning can solve all their problems
- People think Machine Learning is more advanced
- Machine Learners don't always check assumptions
 - Good or Bad?
- Machine Learning gets you funding...

Machine Learning

- People prob



- People



- Machine – Good

all their

advanced

umptions

- Mach

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**

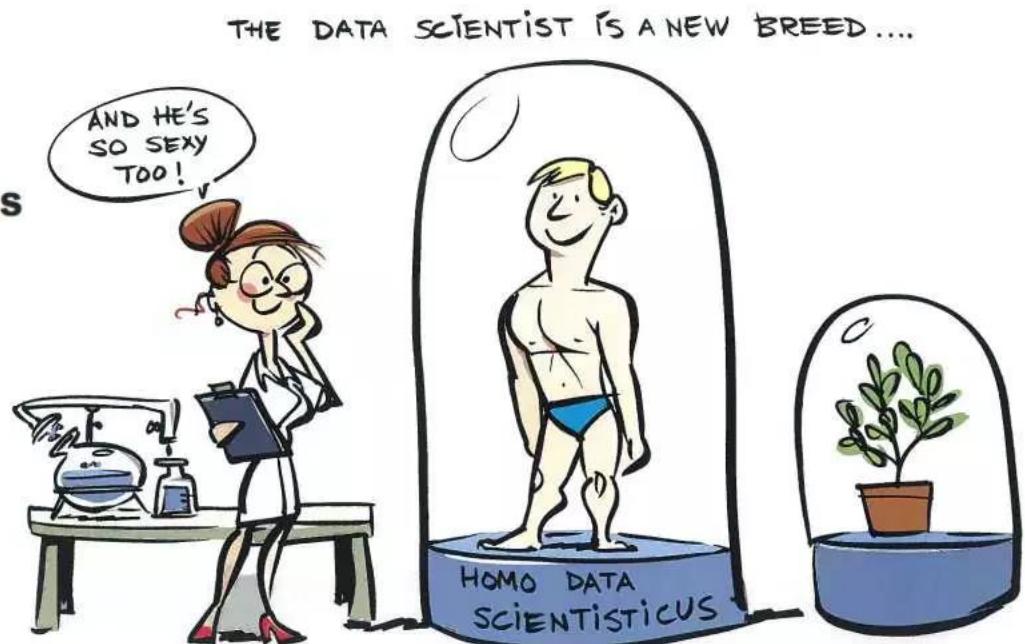
data science is **software**

data science is **statistics on a mac**

data science is **not taught at universities**

data science is **overrated**

data science is **a fad**

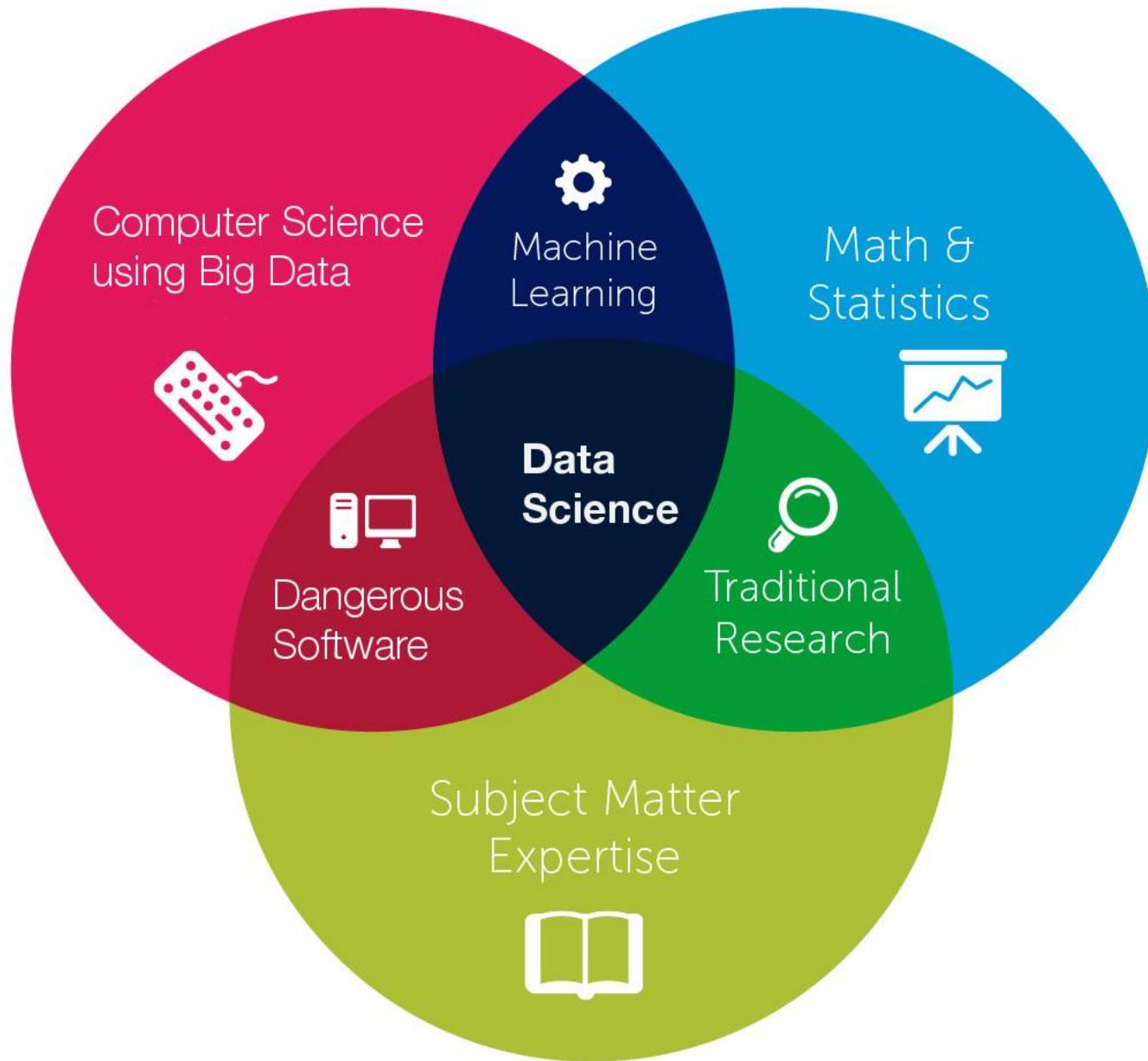


Data Science

- Data Science encompasses aspects of both Statistics and Machine Learning

Data

- Data
an



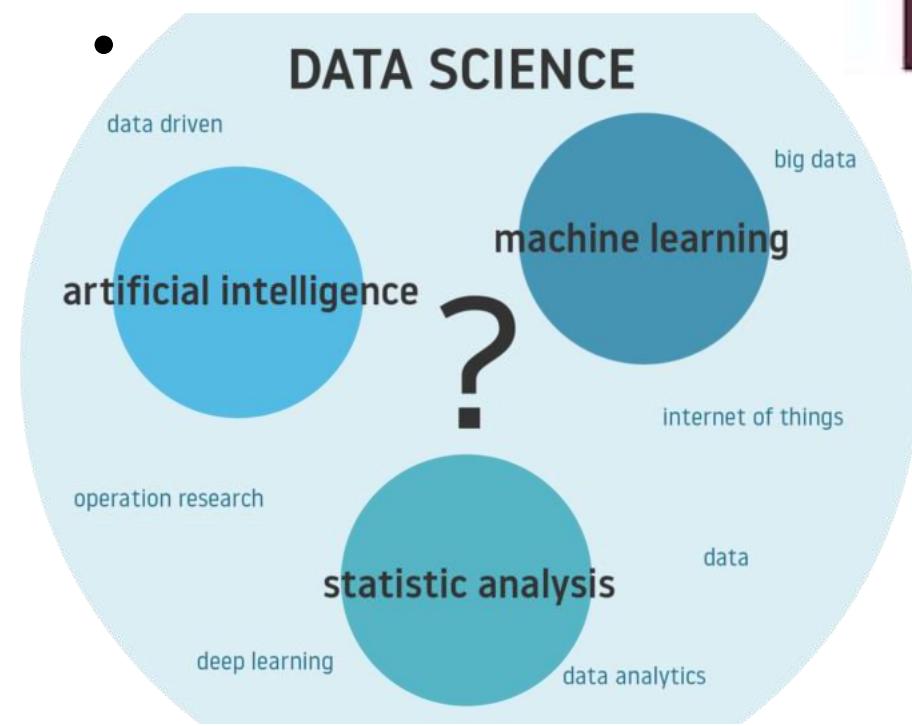
tics

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 encompasses Machine Learning and Machine Learning
- Data Science is a discipline that involves



Data Science

- Data Science encompasses aspects of both Statistics and Machine Learning
- Data Science covers basic MS Excel to Deep Learning
- Data Science employs people with a large variety of skill levels and job descriptions

Data Science

• Data Science encompasses aspects of both Statistics and Machine Learning.

PhD Data Scientist Featured

- 📍 Blaengwynlais, CF83
- 💷 Market related
- 🕒 Contract
- ⌘ X4 Group
- 📅 Expires in 1 day

Data Scientist (Machine Learning)

- 📍 Reading, Berkshire
- 💷 £45000 - £50000 per annum
- 🕒 Permanent
- ⌘ Reqiva Limited
- 📅 Expires in 3 days

Data Scientist - Python, R, SQL, Financial Services, Machine Learning

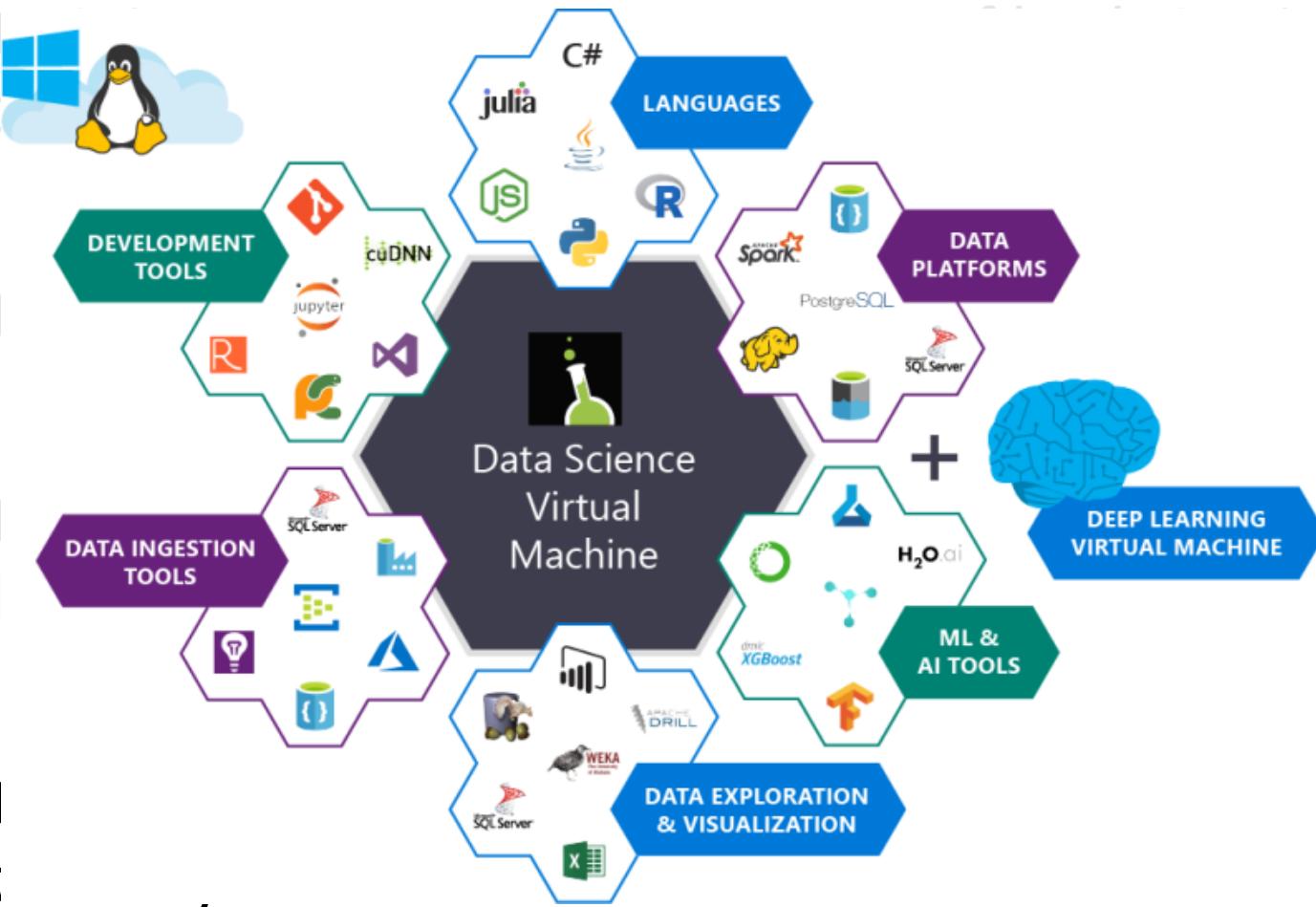
- 📍 London
- 💷 Unspecified
- 🕒 Permanent
- ⌘ Entasis Partners LLP
- 📅 Yesterday

Data Science

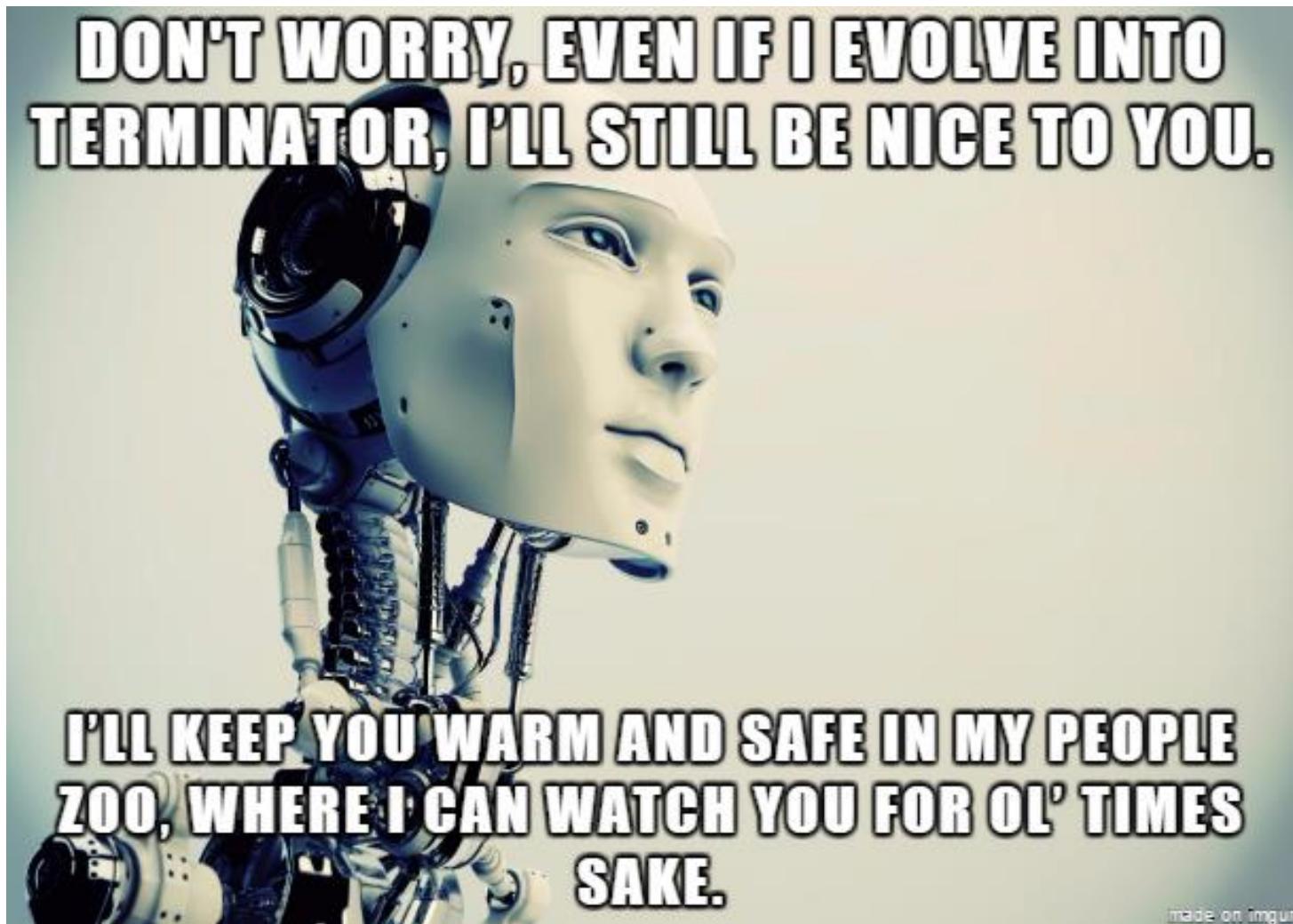
- Data Science encompasses aspects of both Statistics and Machine Learning
- Data Science covers basic MS Excel to Deep Learning
- Data Science employs people with a large variety of skill levels and job descriptions
- Data Science has the reputation for being flexible, but potentially not done well

Data Science

- Data science is a discipline that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply them to solve real-world problems.
- Data science is a multidisciplinary field that draws from various disciplines such as statistics, computer science, mathematics, and domain-specific knowledge.
- Data science requires a combination of technical skills, including programming, data analysis, machine learning, and domain expertise.
- Data science is often used in business intelligence, data mining, and predictive modeling.



AI (Artificial Intelligence)

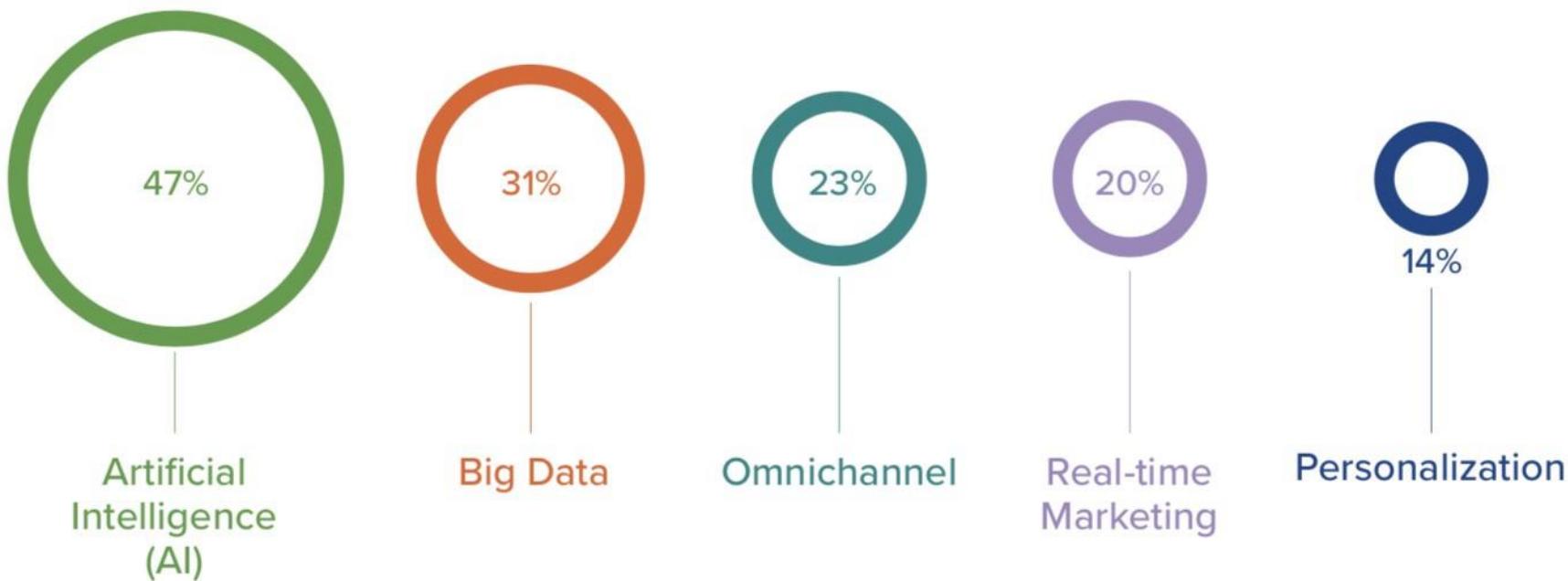


AI

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

OVERHYPED MARKETING BUZZWORDS

Which of these marketing concepts do you consider to be overhyped, meaning the concept is more fantasy than reality?



AI

- AI is the latest trend and one of the biggest buzzwords
- AI is behind some of the newest technologies



of the biggest buzzwords
est technologies

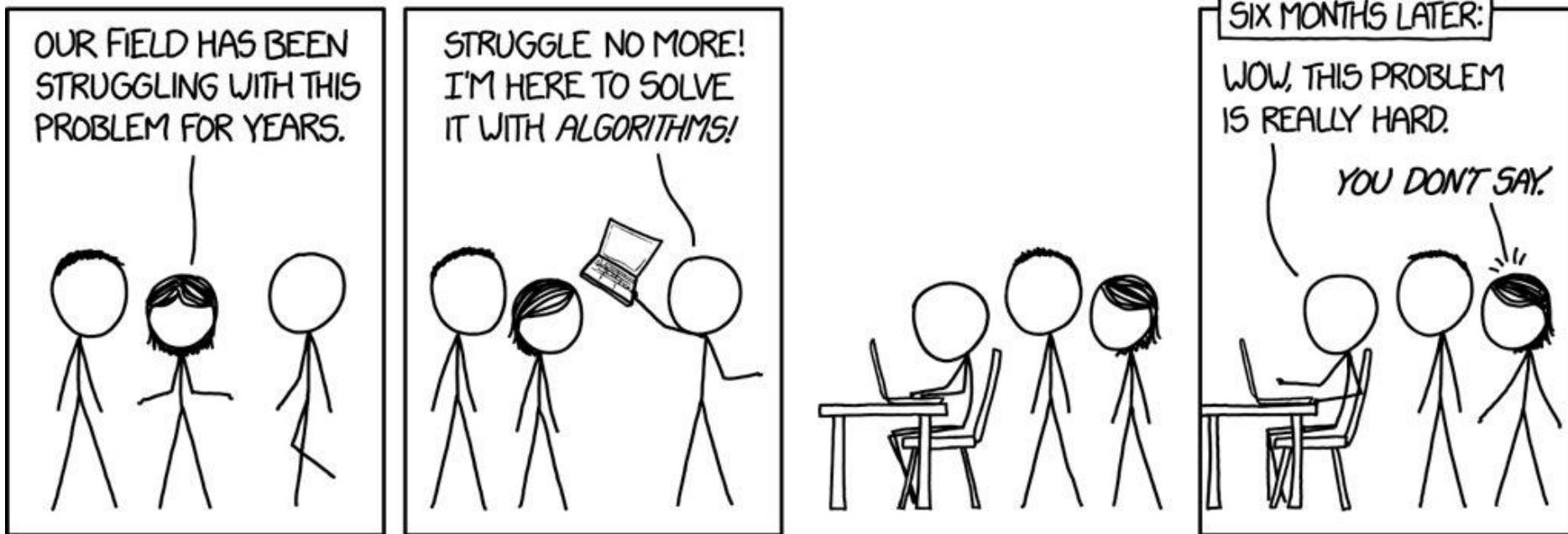


AI

- 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

AI

- AI is...
**Clinician using basic statistics:
“I want to use AI on our data”**
- 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
- Applied researchers (not including applied statisticians etc) don't generally understand when AI can be useful or what it is

AI

- AI is .
- AI is
- People
- Applications etc) or will

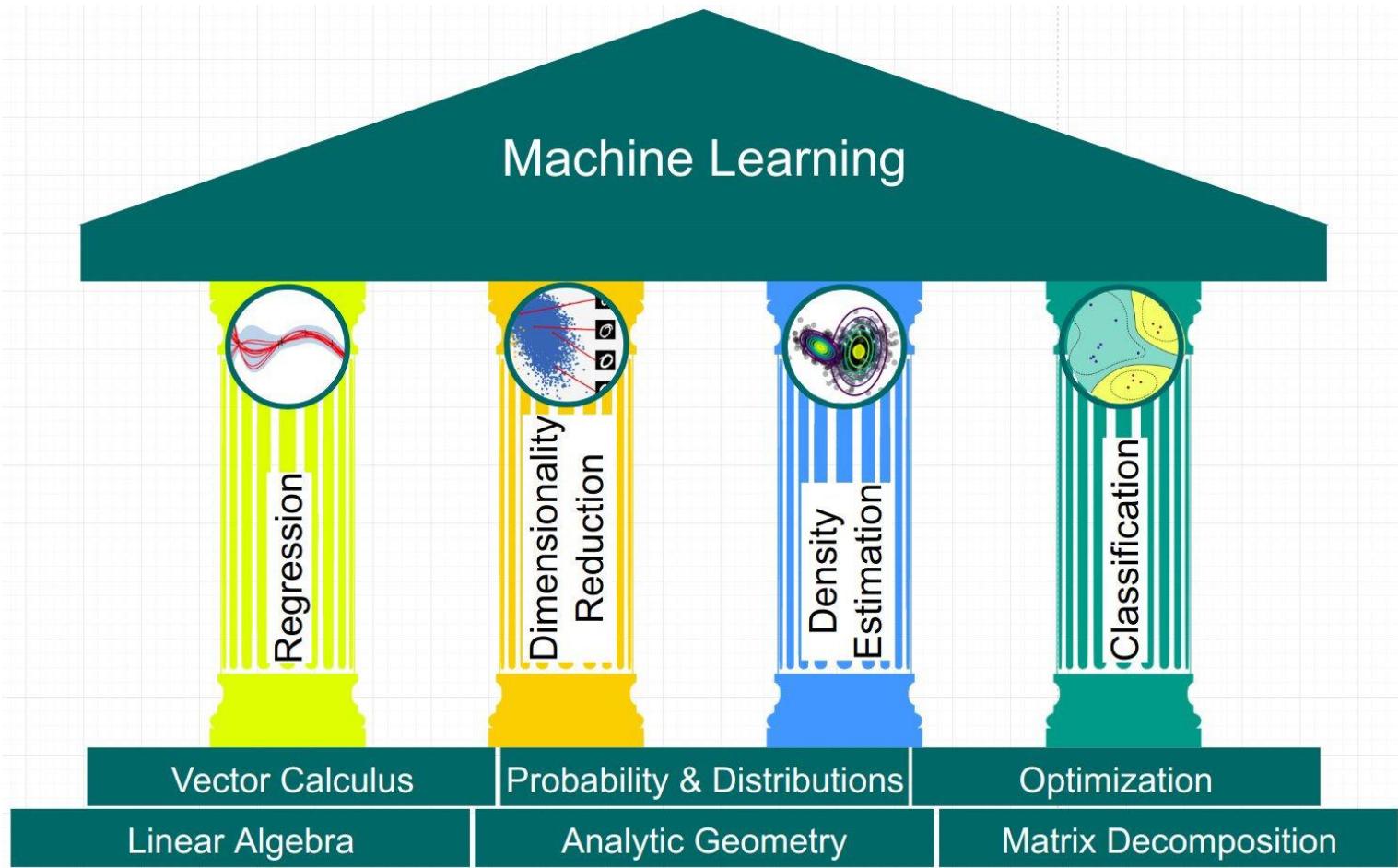


ds
ns
ul

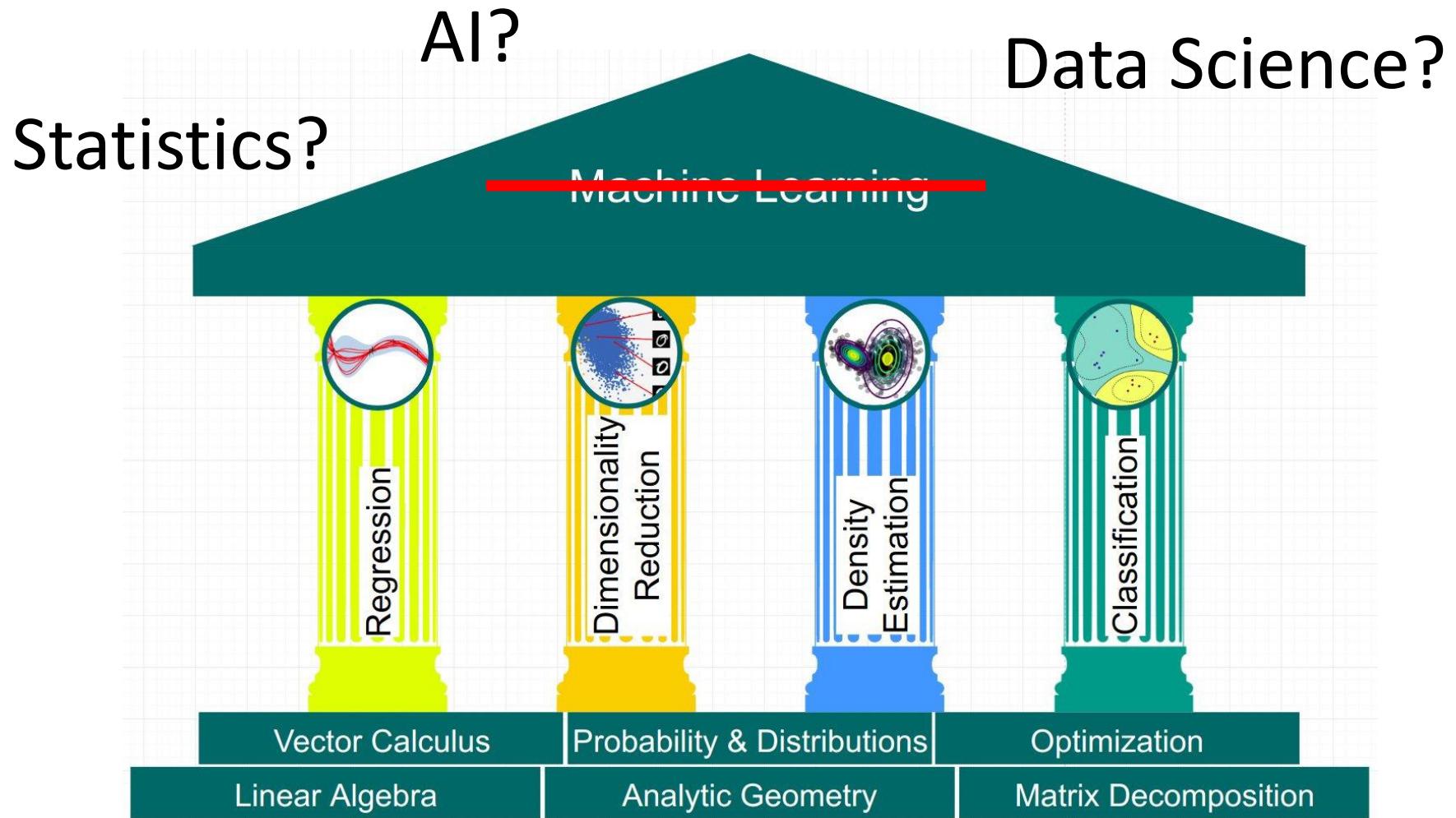
What's the actual difference?



Shared Ideas



Shared Ideas



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 new Deep Learning (AI?) course on Coursera again starts with regression
- Johns Hopkins University Data Science course on Coursera covers multiple aspects of Statistics and Machine Learning

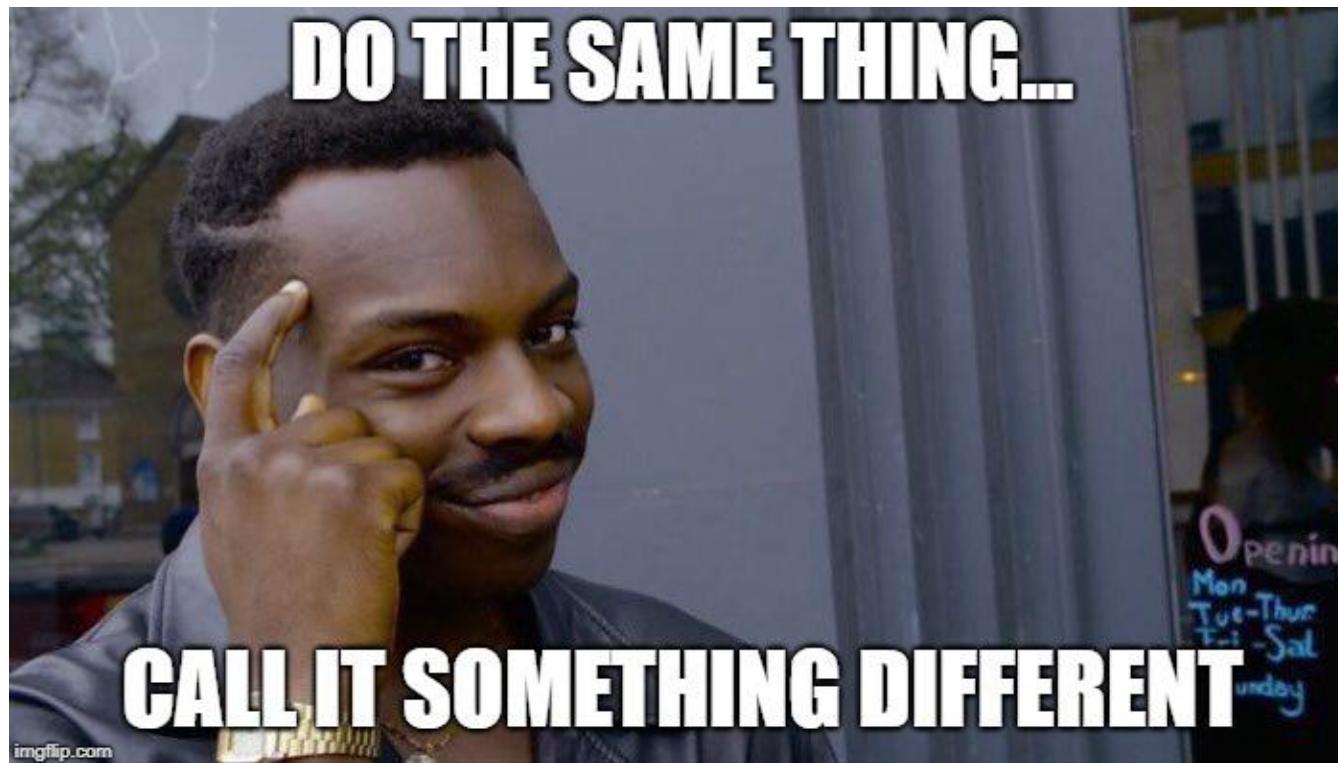


Using Methods from Other “Disciplines”

- 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”

Mark Girolami
@MarkGirolami

Following

AI is used in this really cool Nature article.
The AI deployed is a statistical method called
Linear Discriminant Analysis (LDA) developed
in 1936 for taxonomic analysis by R.A.Fisher.
:-0

Robot chemist uses AI to discover new
molecules epsrc.ukri.org/newsevents/news/2018/08/23/robot-chemist-uses-ai-to-discover-new-molecules/
via @epsrc

7:31 AM - 23 Aug 2018

3 Retweets 14 Likes

Find people you know

Trends for you
#YouChromebook
Save ££ and get a new Chromebook
before school starts
Promoted by Google UK

Tweet your reply

Google Ads

Using Methods from Other “Disciplines”

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

Hannah Fry @FryRsquared

FFS

To briefly explain how Linear Regression helped us reverse engineer the BSR equation, let's break it down. Linear Regression is an AI equation that finds the proper coefficients for an equation by sorting through massive amounts of data. The equation looks something like $BSR = X(a) + Y(b) + Z(c)$ and so and and so forth.

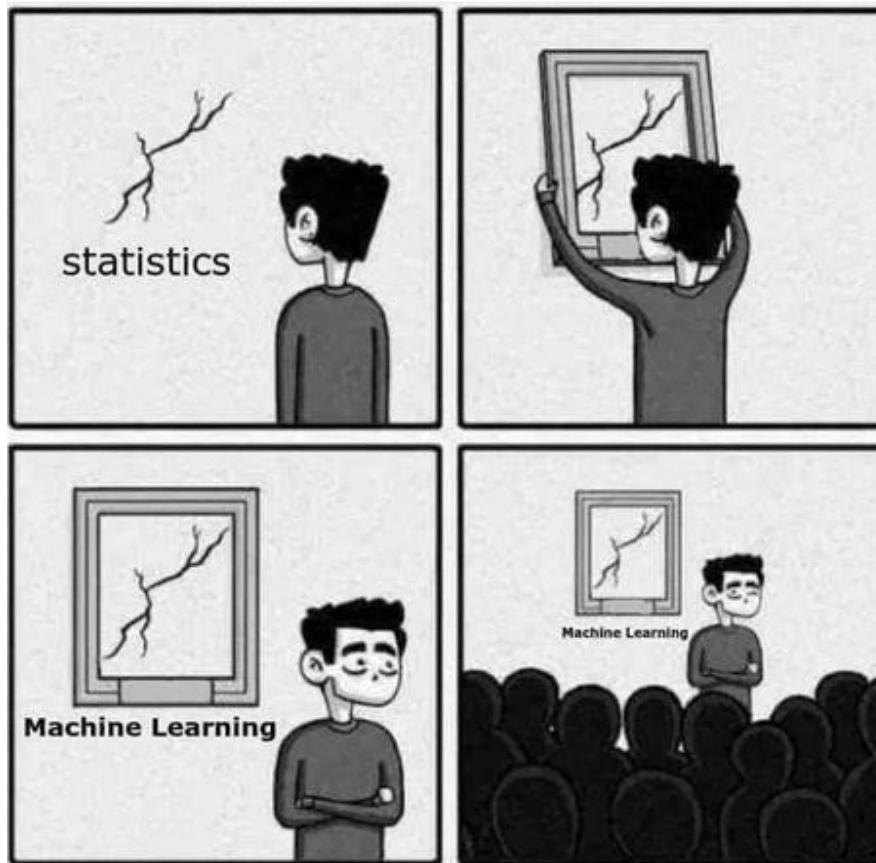
4:56 AM - 4 Oct 2018

40 Retweets 175 Likes

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

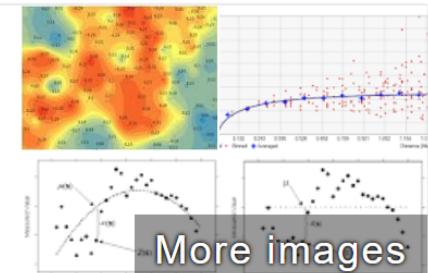
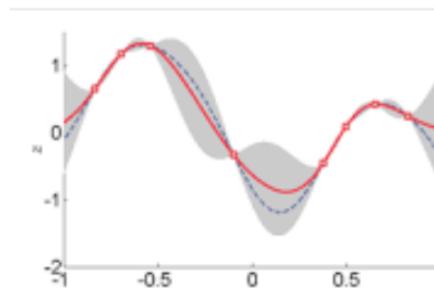
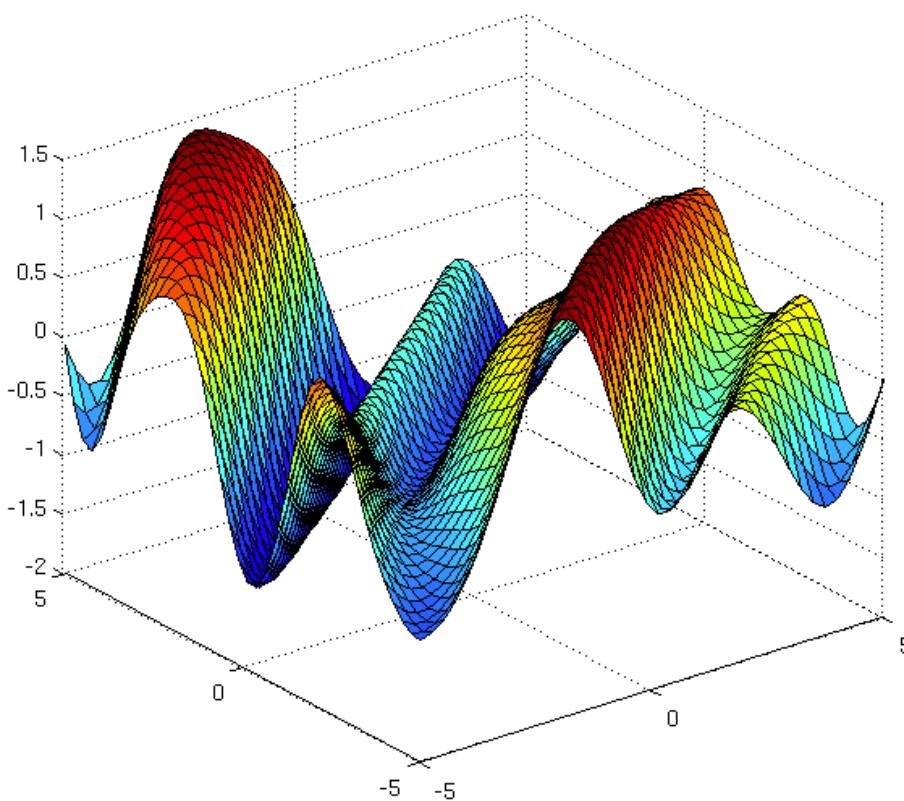
Using Methods from Other “Disciplines”

- 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](#)



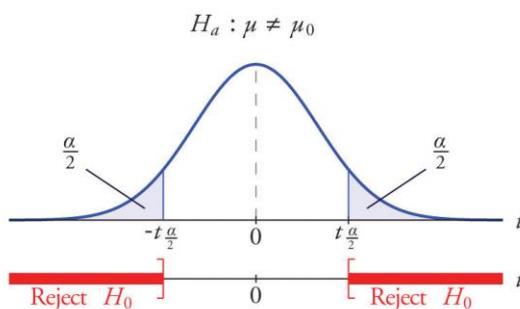
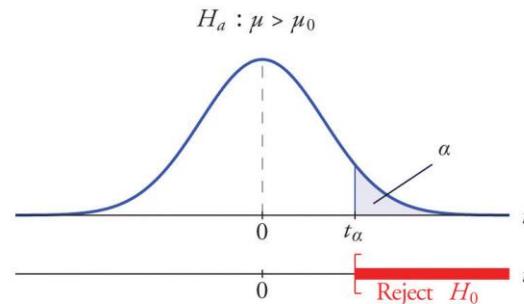
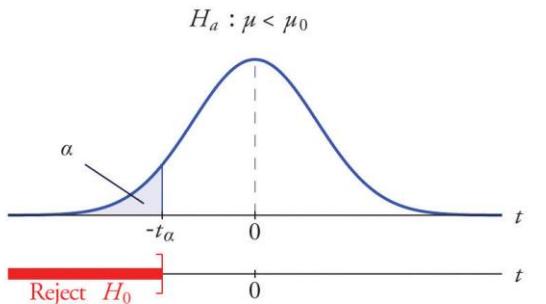
Same Methods, Different Names

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



A



B



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 basically Statistical Hypothesis Testing taught by different people

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

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

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

Different Inference Techniques

Example 1: Estimating Linear Regression Parameters

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

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

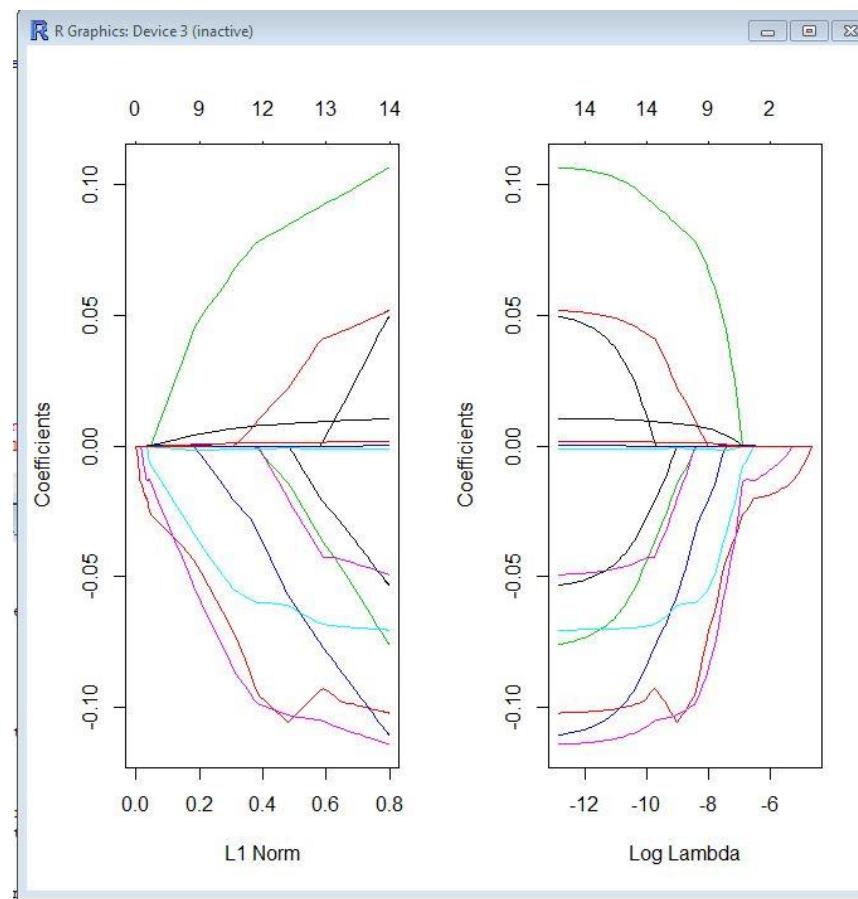
Different Inference Techniques

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

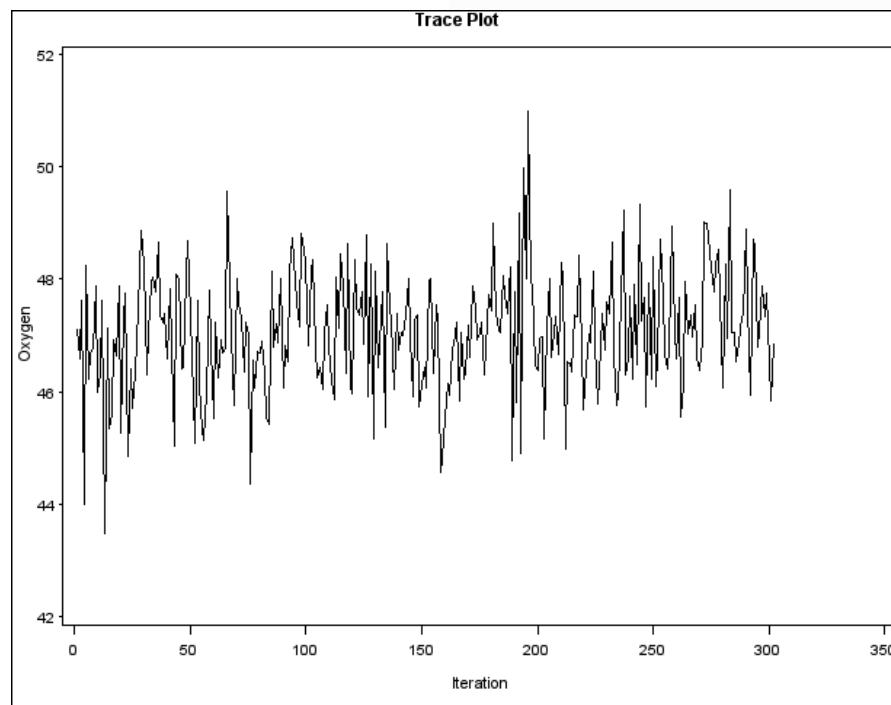
Example 2: Parameter estimation

- 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

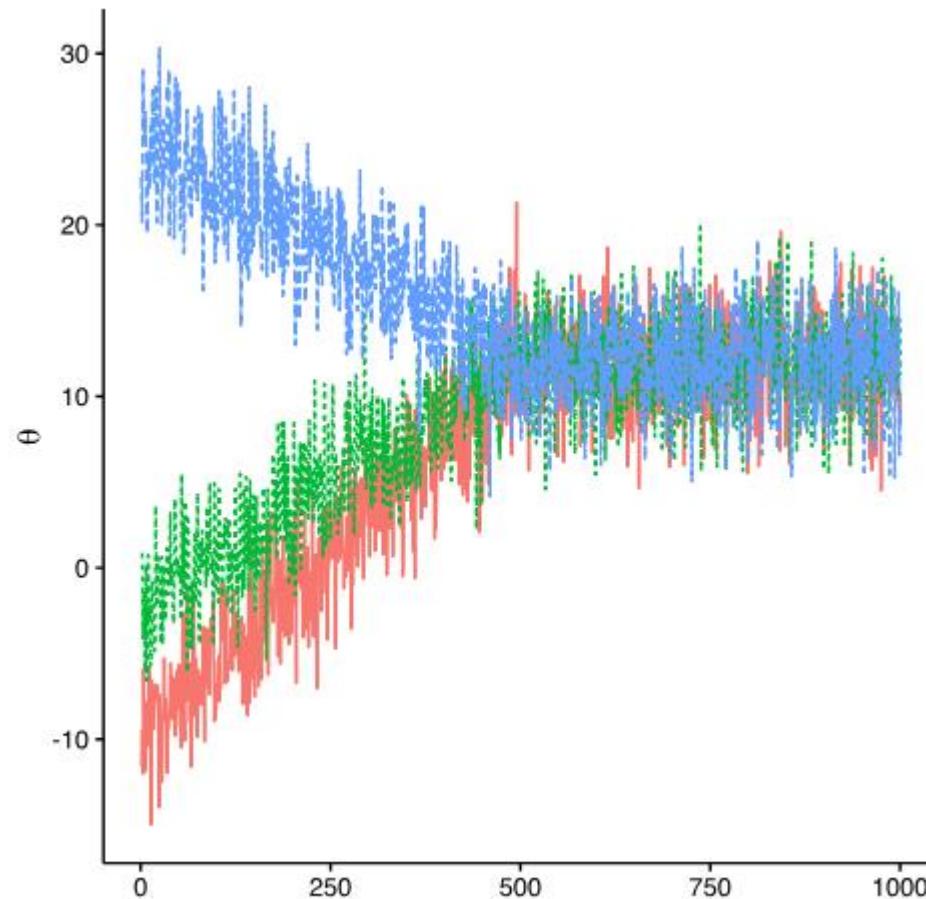
Different Inference Techniques

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, 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 take the attitude that whatever works works

Different Goals – Prediction & Explanation

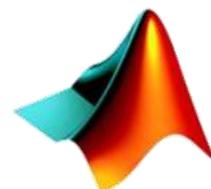
- Goals can differ between Statistics, Machine Learning, Data Science and AI



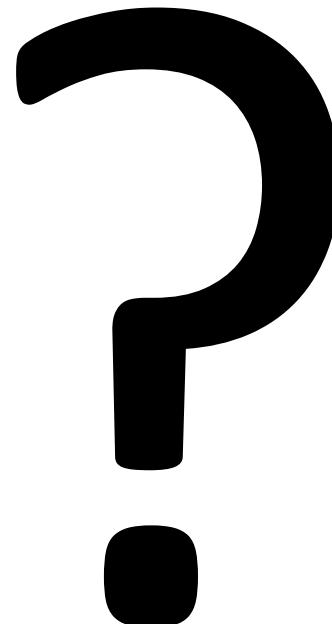
Different Goals – Prediction & Explanation

- 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 Programming Languages



MATLAB



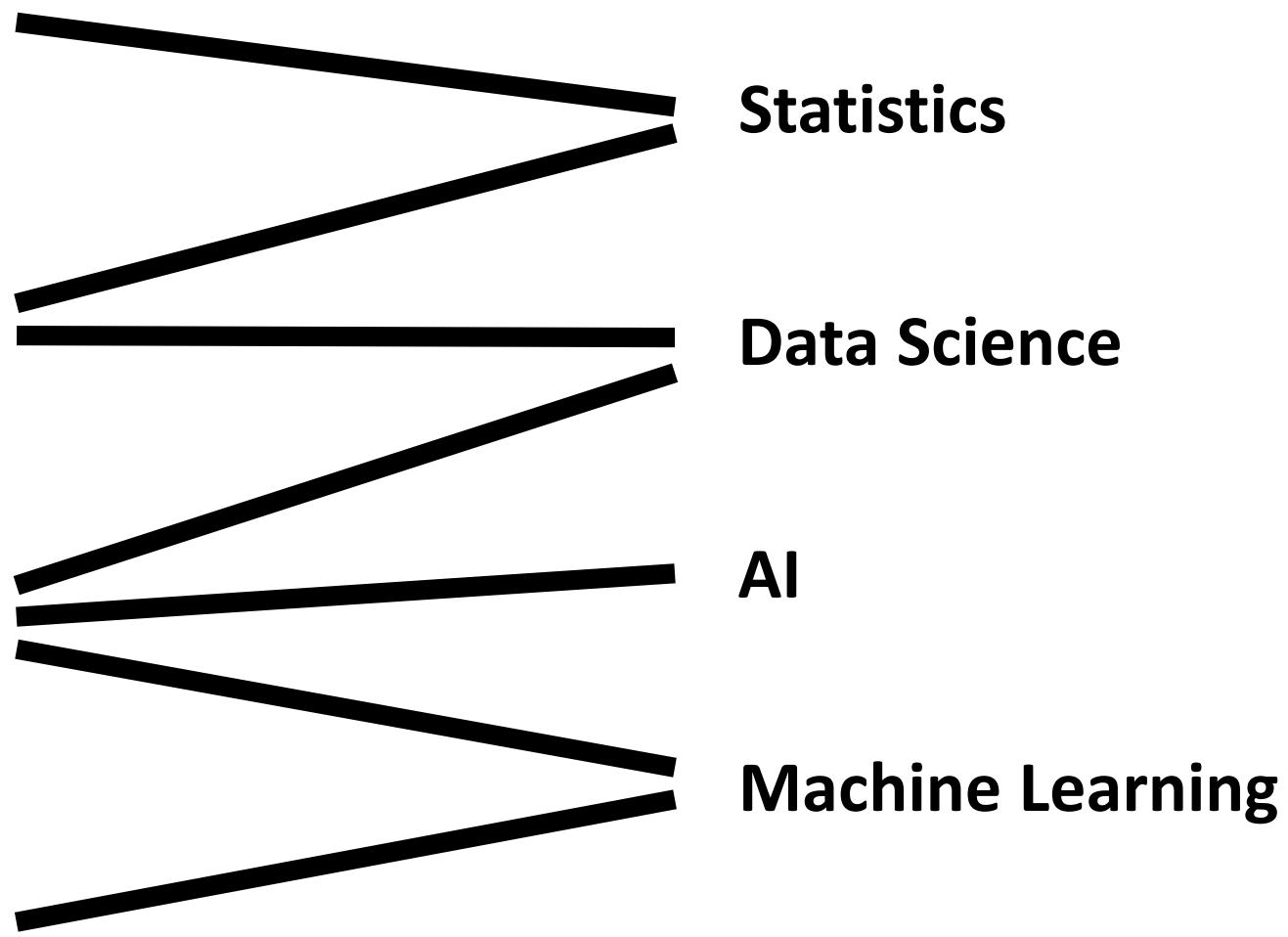
Statistics

Data Science

AI

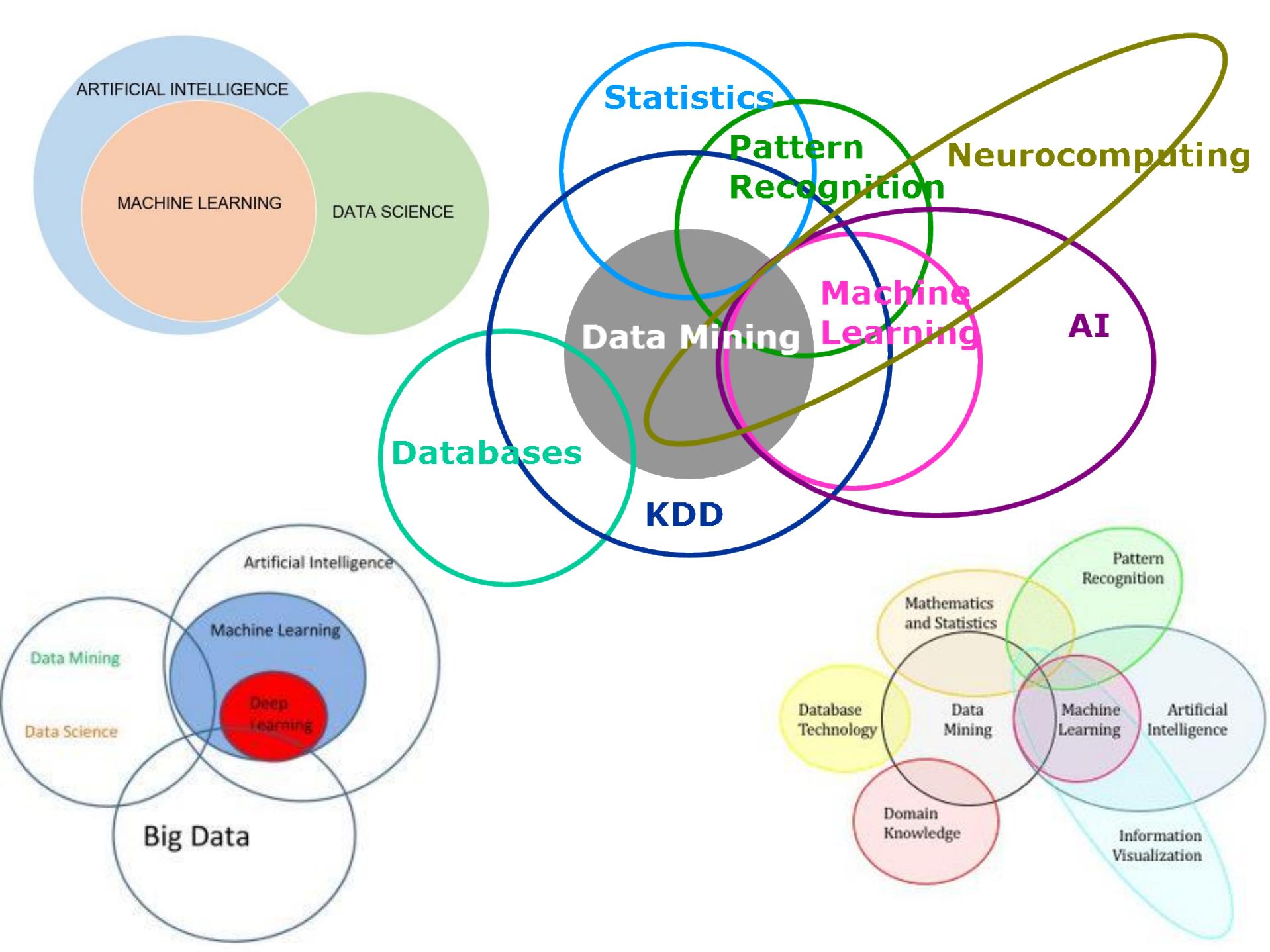
Machine Learning

Different Programming Languages

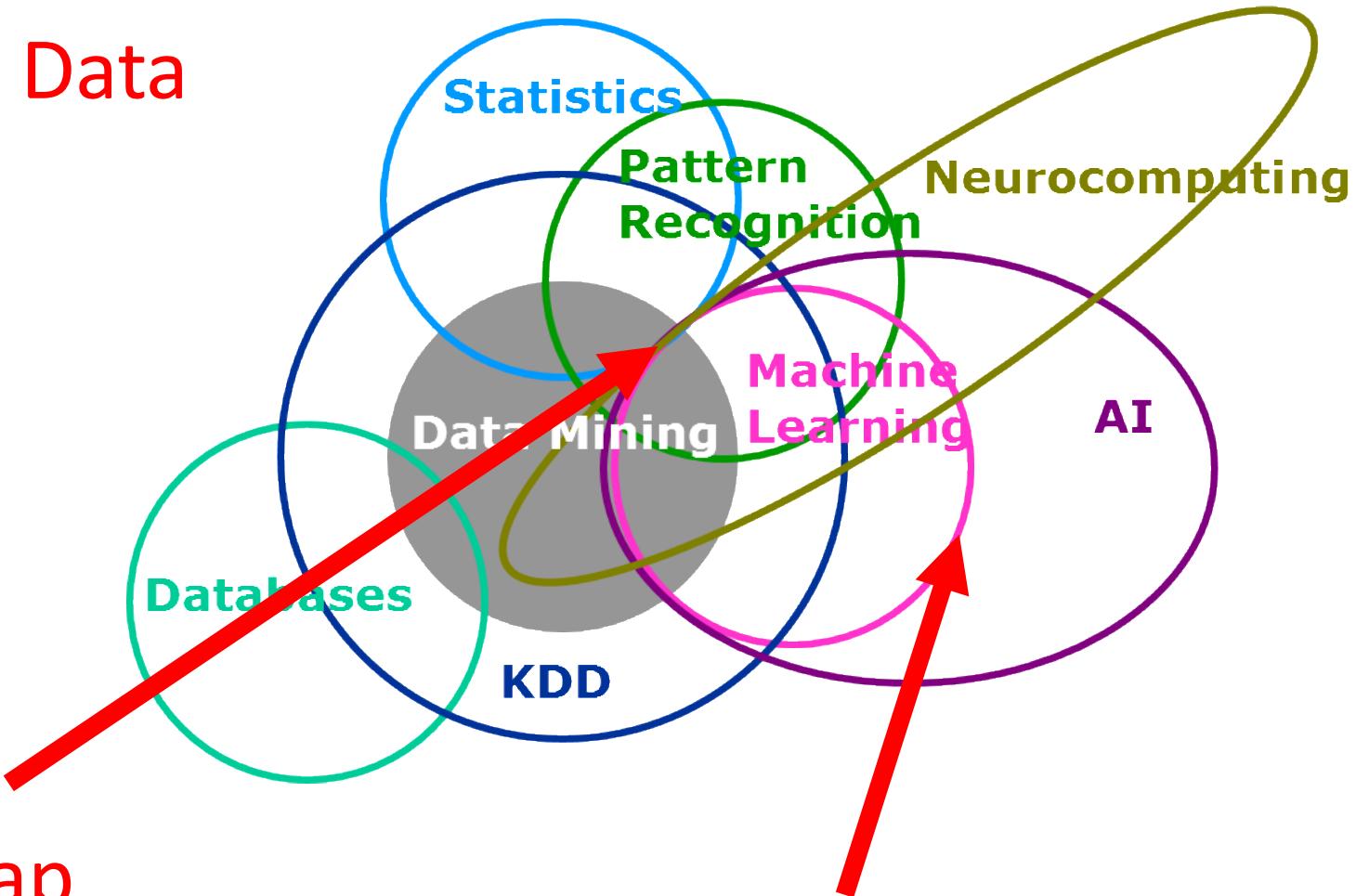


What is the overlap?



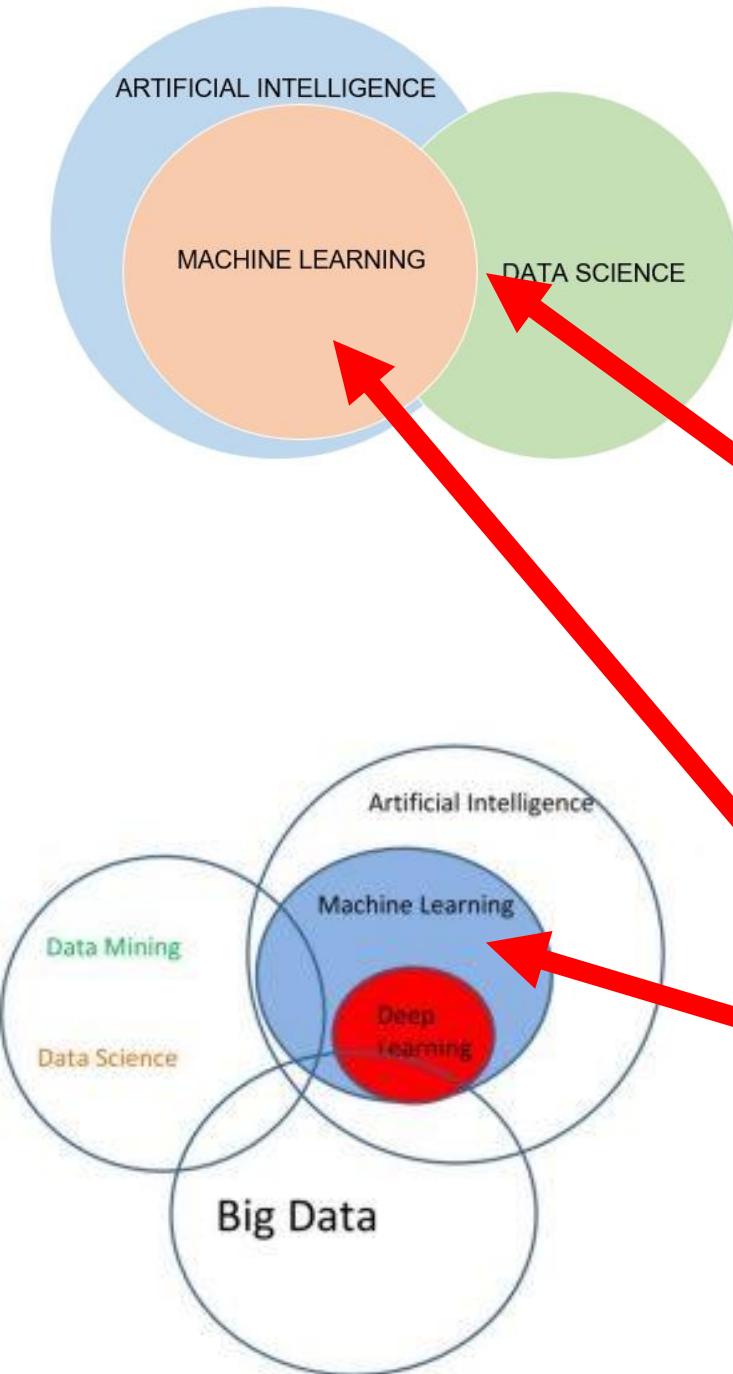


Where is Data Science?



No overlap
between Statistics
and Machine
Learning?

Machine Learning is
a subset of AI?

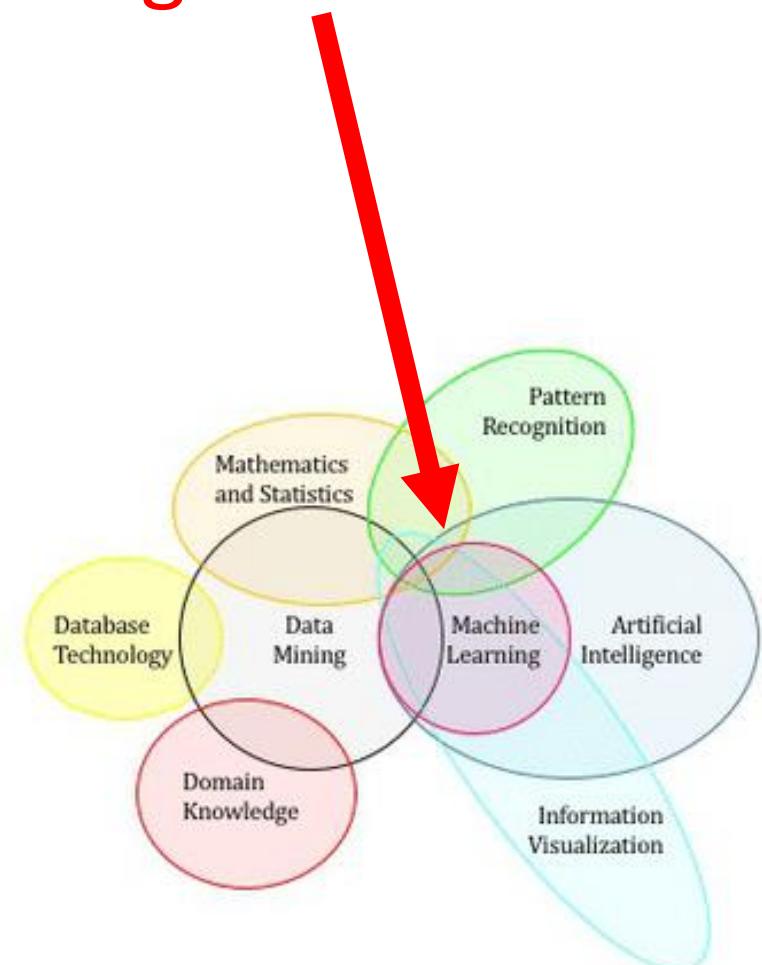


Where is Statistics?

Overlap between
Data Science and ML
/ AI

Machine Learning is
a subset of AI again

Little overlap between Statistics and Machine Learning



What do I think?

AI

Data Science

Machine Learning

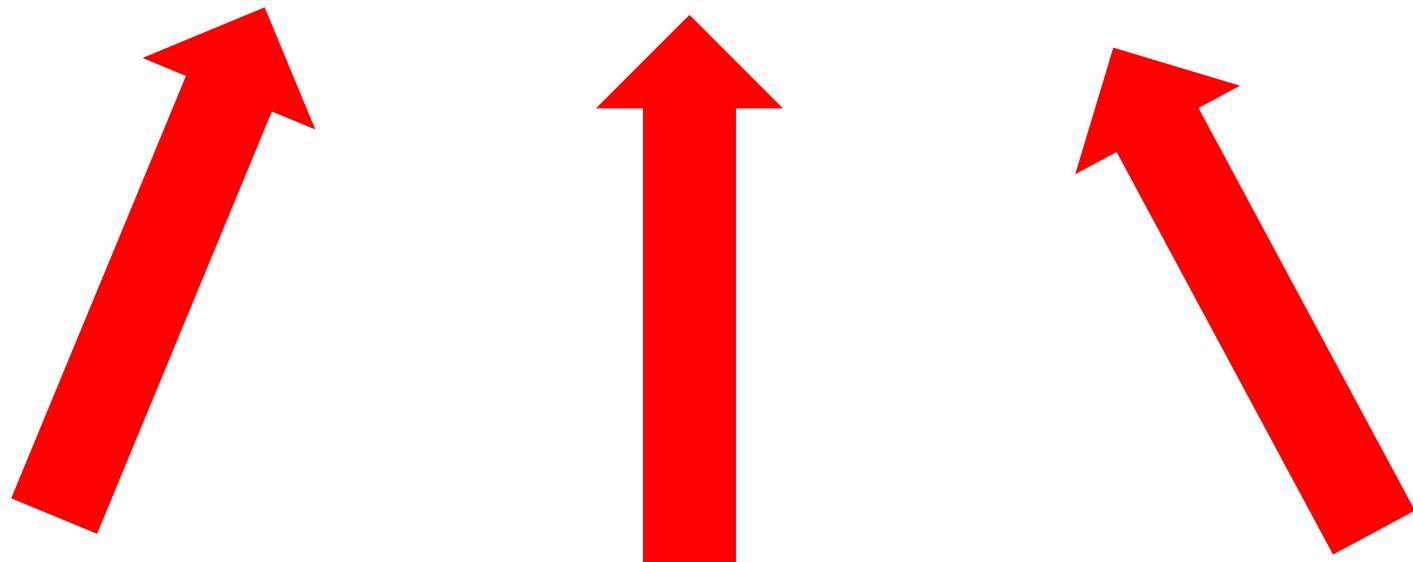
Statistics

Which is best?



Which is best?

It depends on the situation!!!

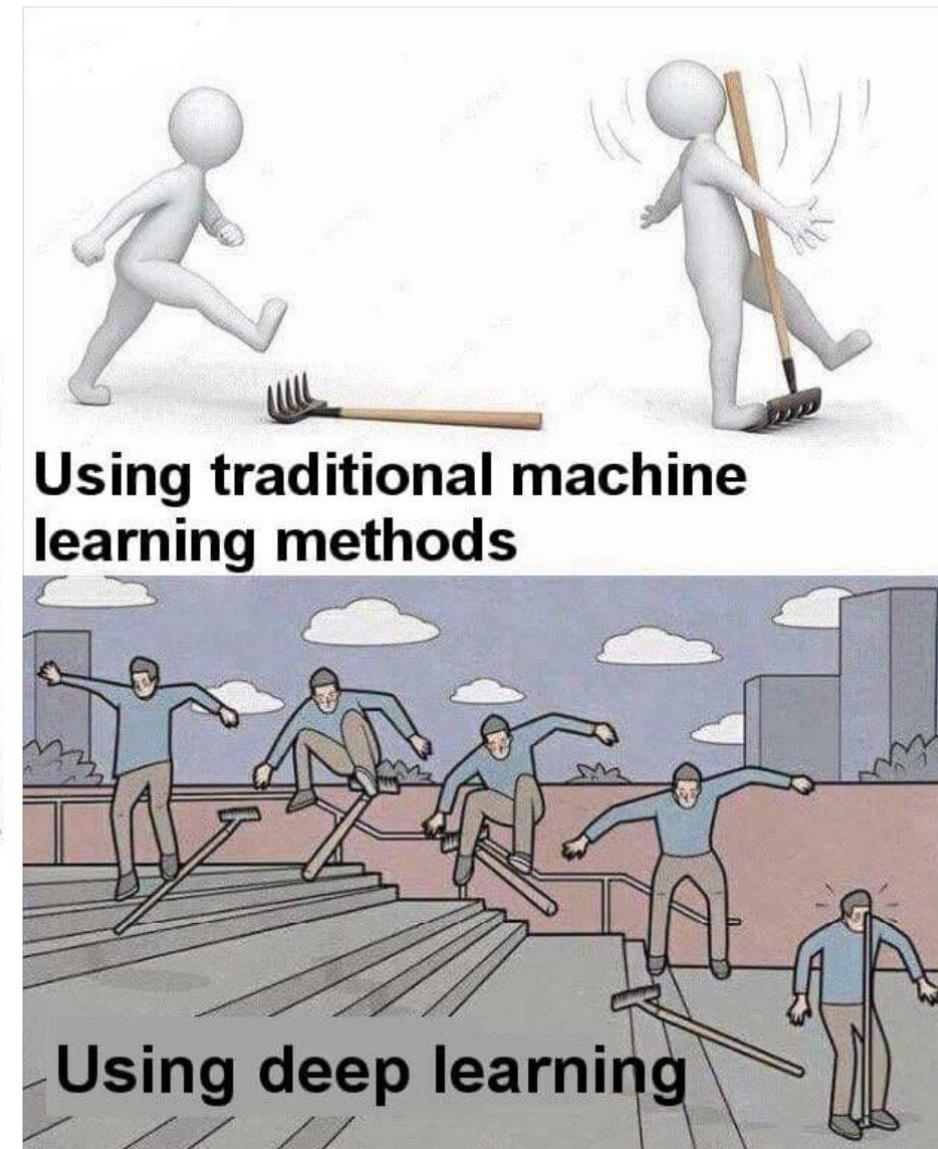
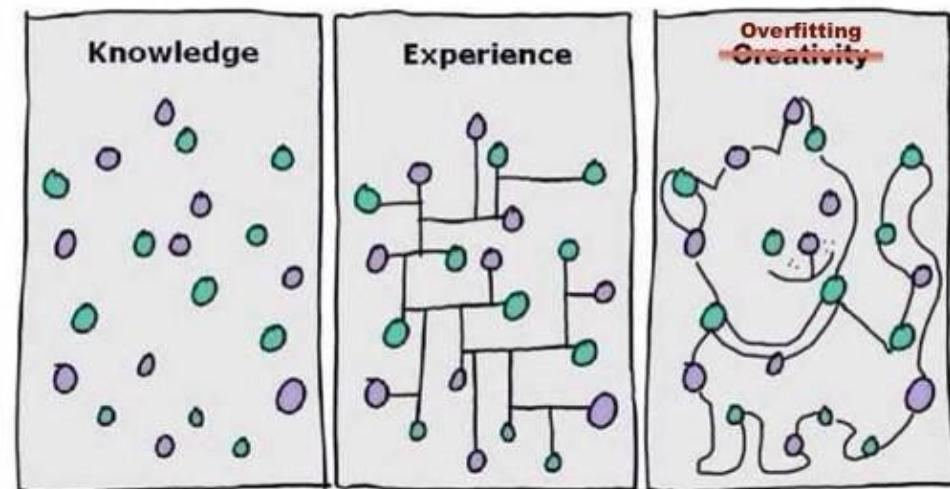


Which is best?

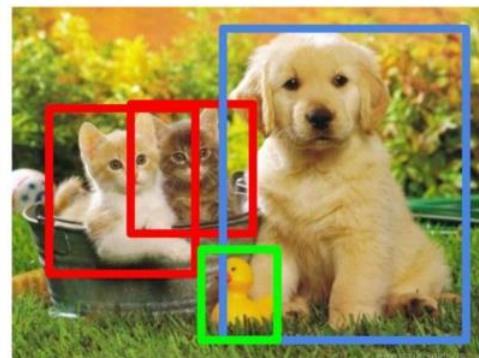
It depends on the situation!!!

- 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?



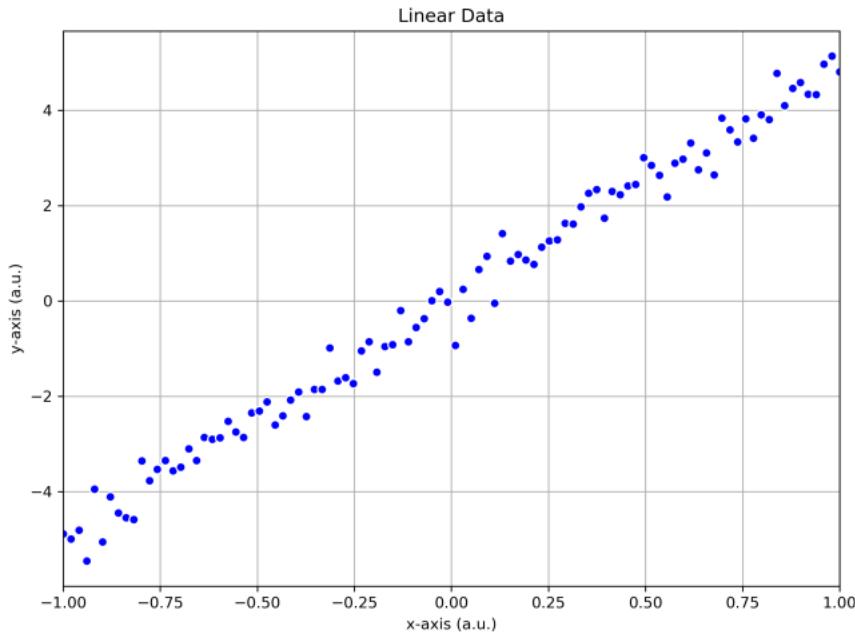
CAT

CAT

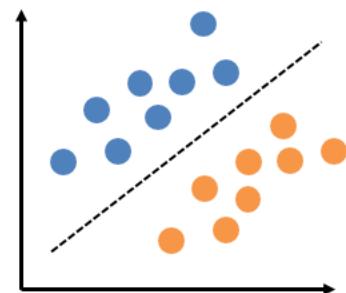
CAT, DOG, DUCK

CAT, DOG, DUCK

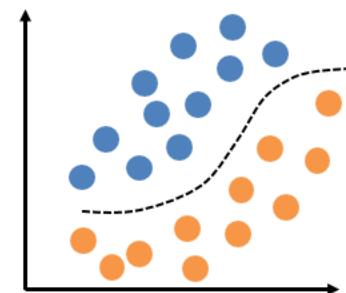
When not to use AI / Deep Learning



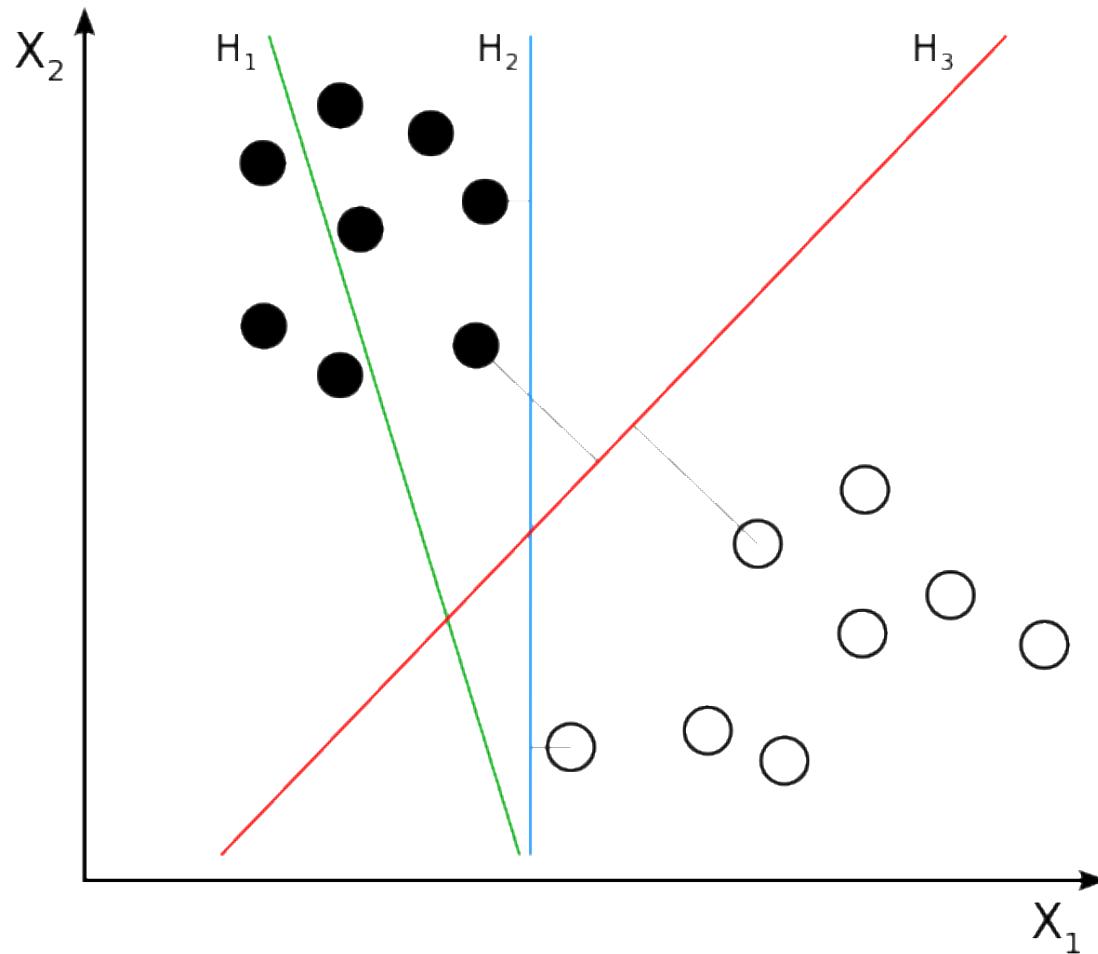
Linear



Nonlinear

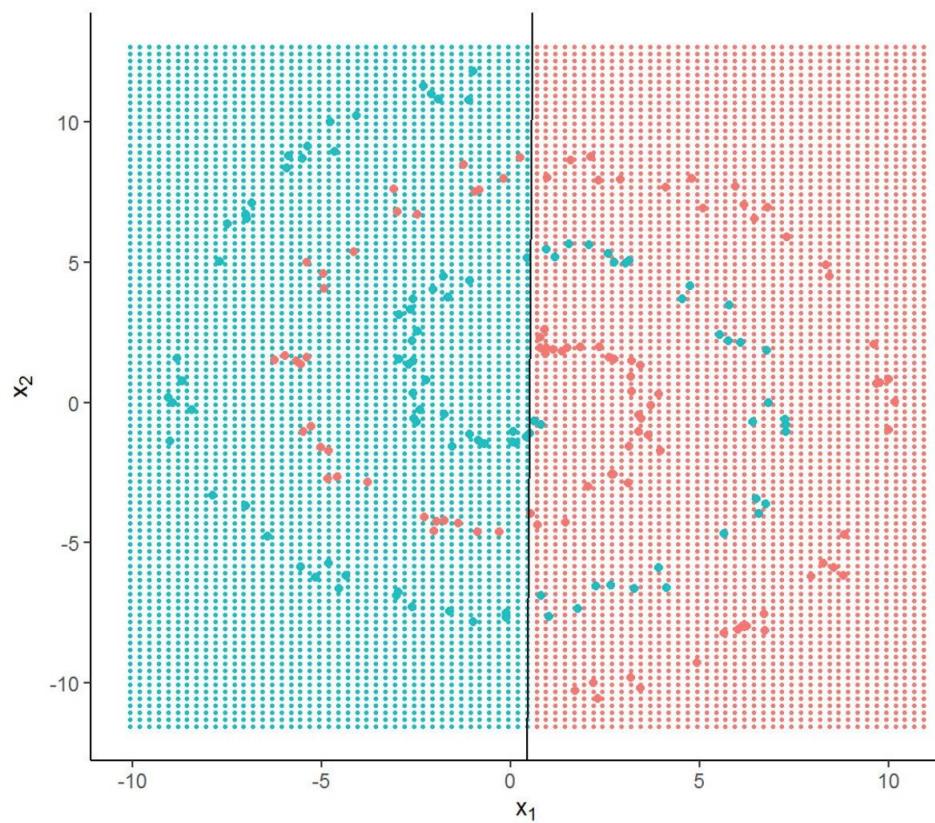


Sometimes Machine Learning does work best

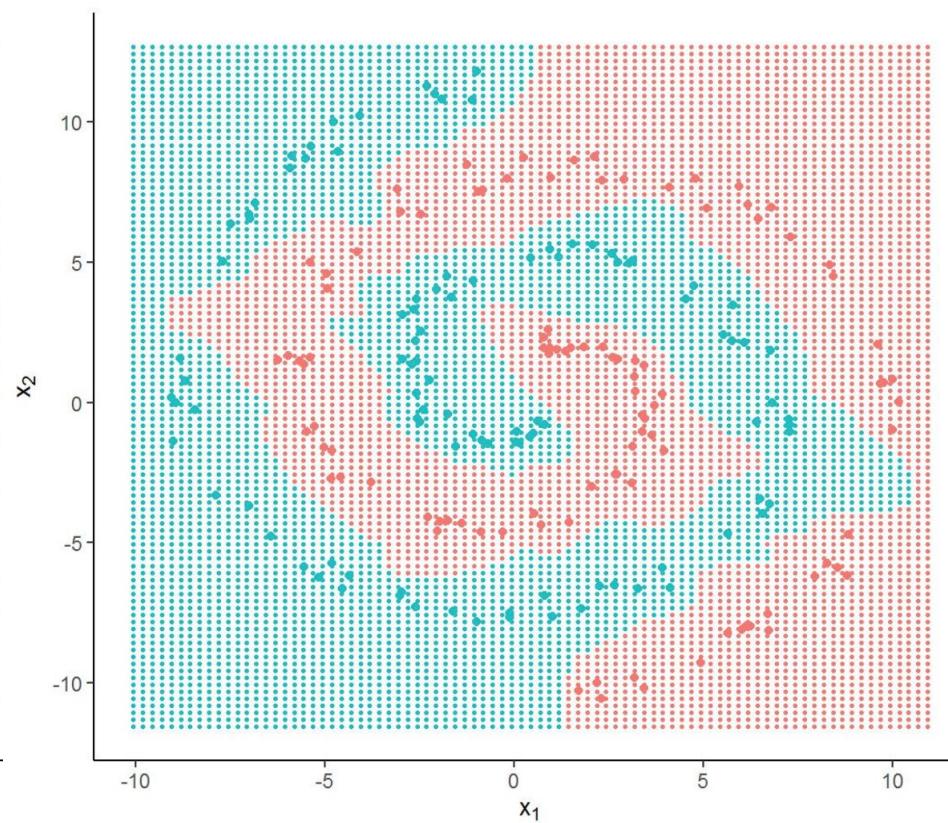


Sometimes Machine Learning does work best

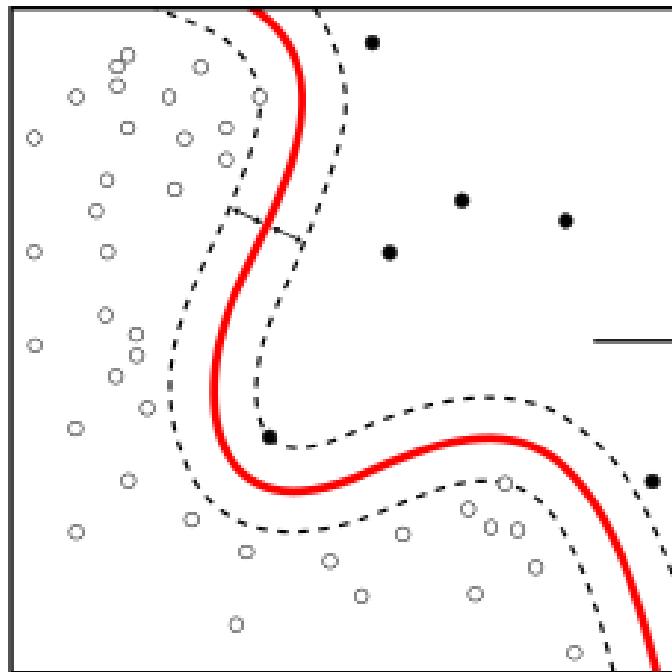
Logistic Regression



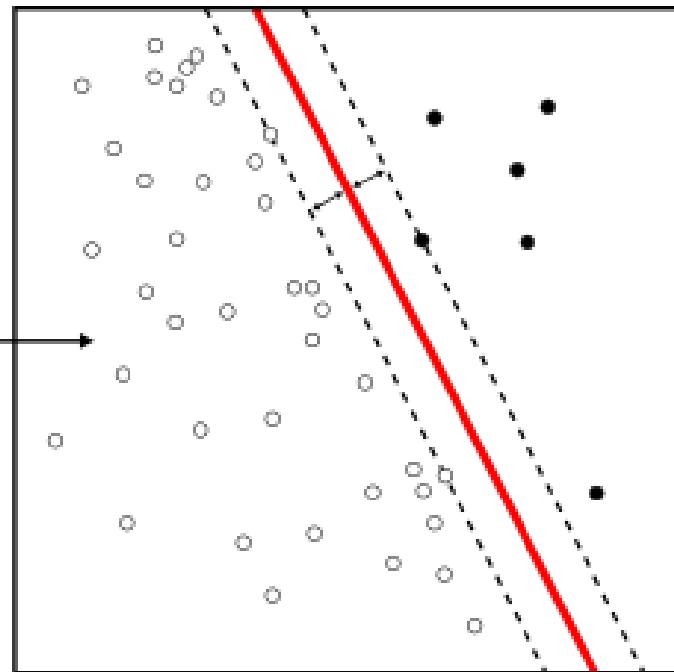
Neural Network



Sometimes a non-linear method is sufficient

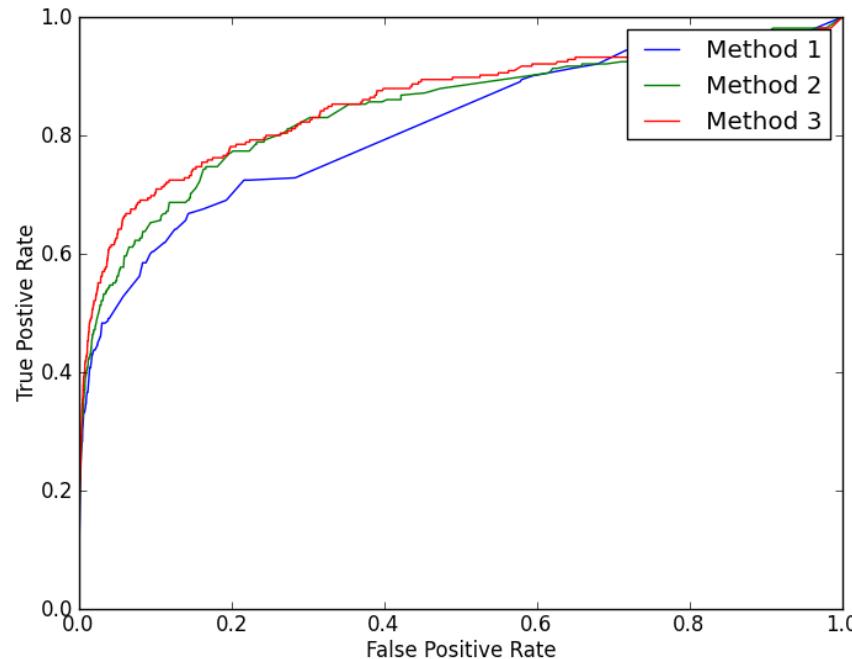


\emptyset



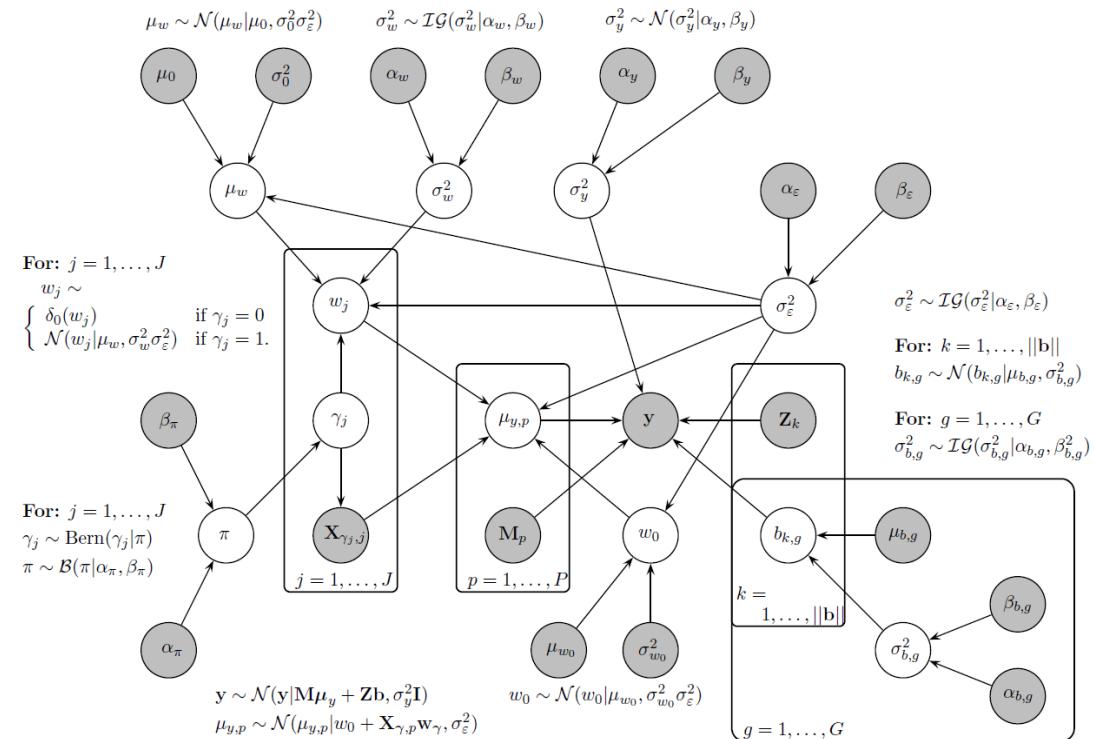
(Always?) Try Statistics First?

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



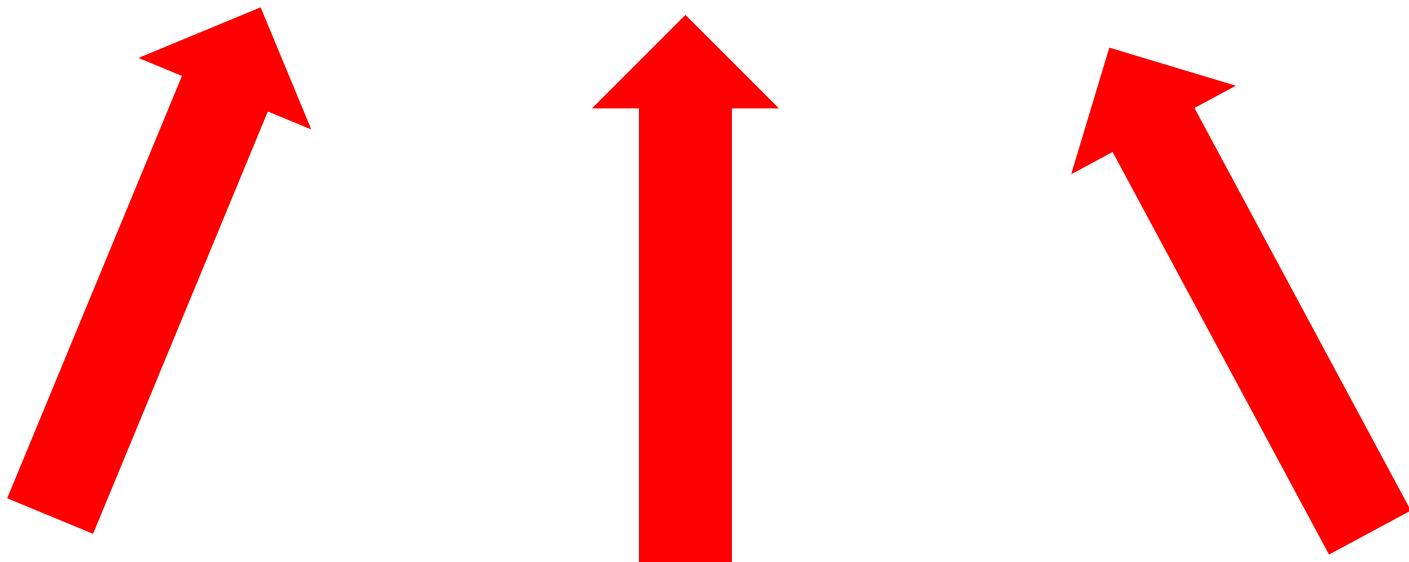
Sometimes Statistics can be complex

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



Still want AI, but don't know how?

Talk to some experts!!!

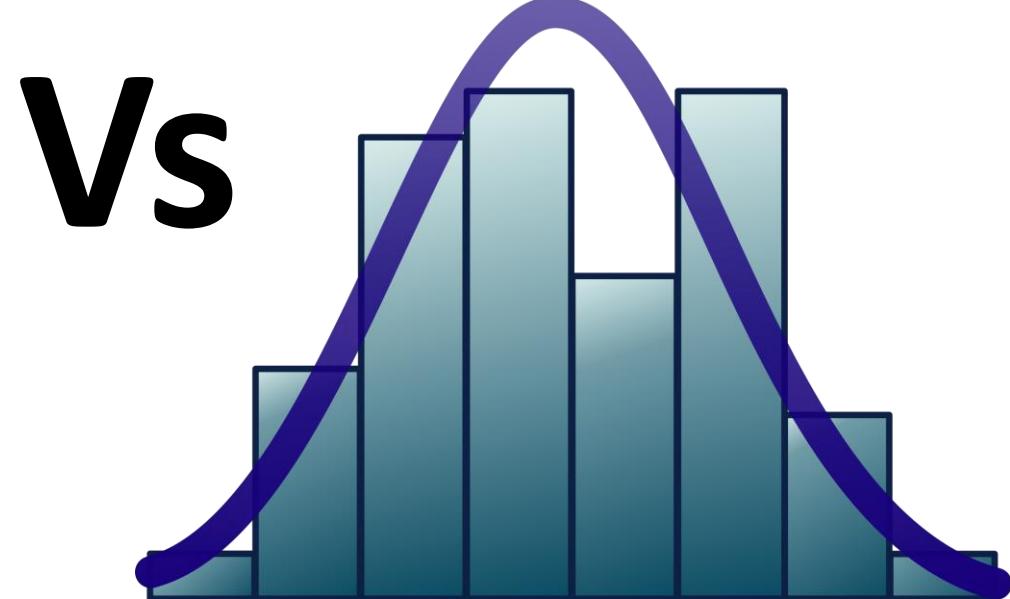
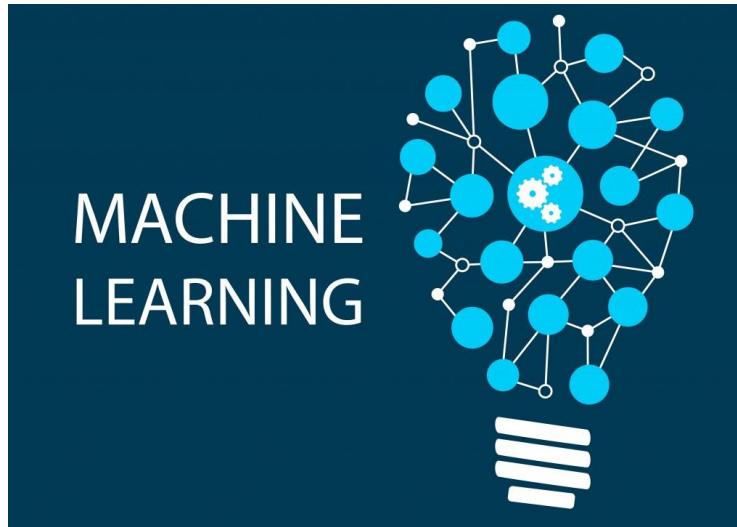


How can Statistics stay relevant?



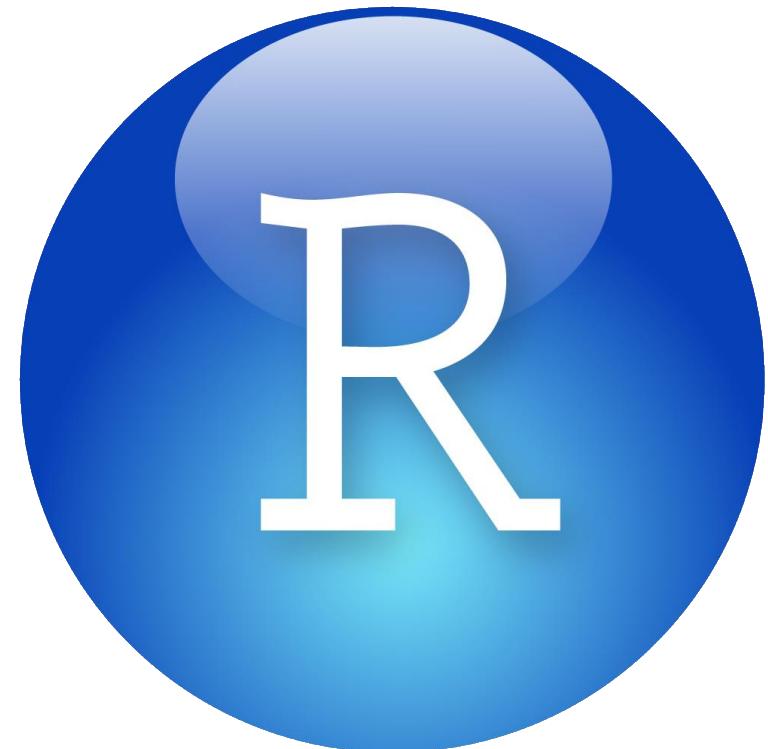
Why Machine Learning, not Statistics?

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



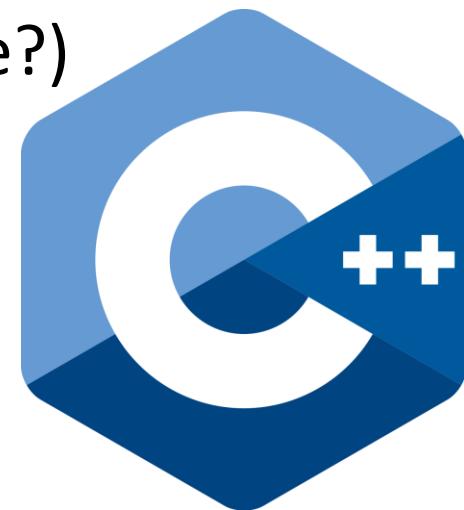
Programming Skills

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



Programming Skills

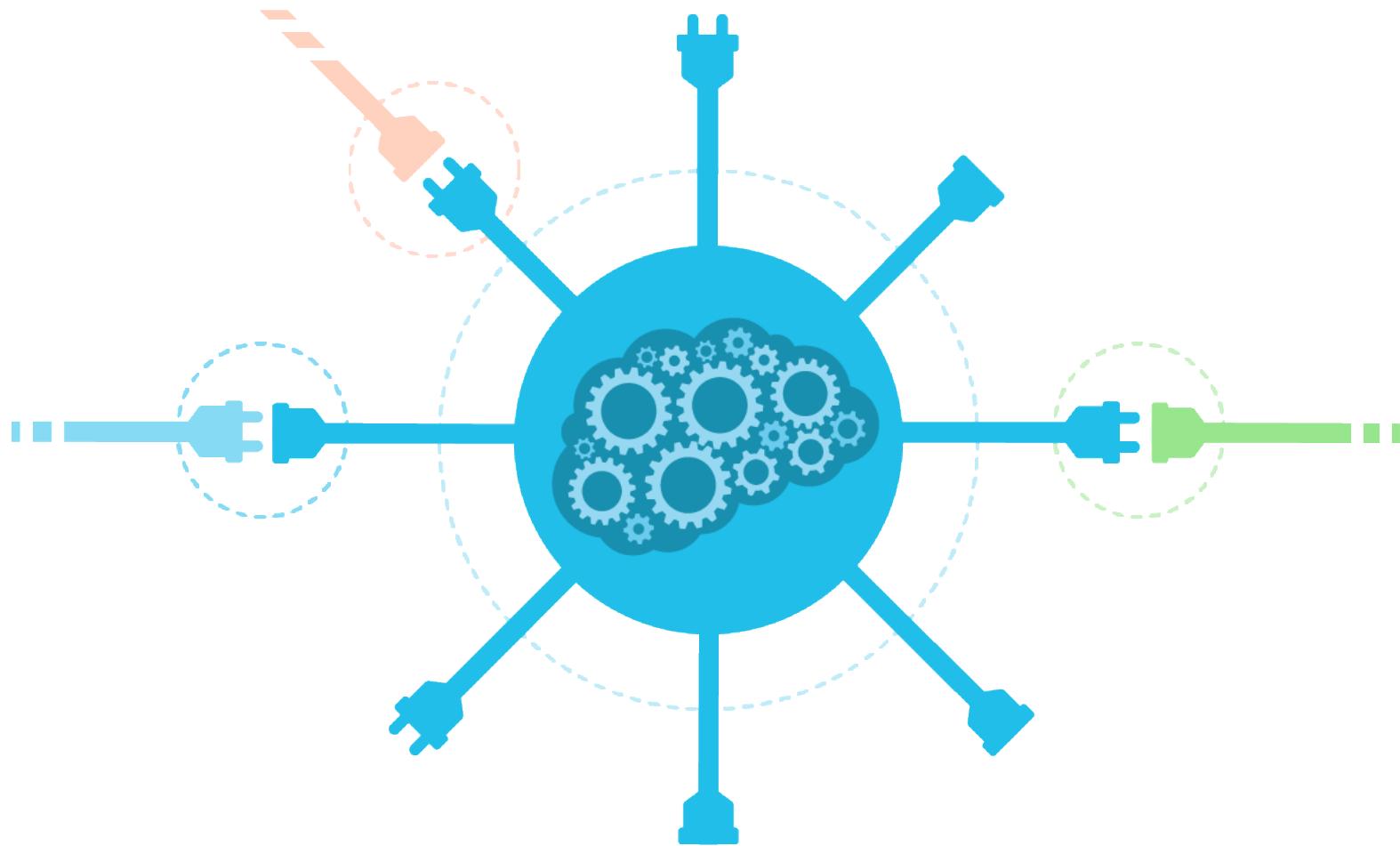
- To keep up with Machine Learners, Statisticians need to get better at programming!
- Statistics courses need to do more programming and at a higher level
 - Teach Python as well as R
 - Teach SQL? (One day extra course?)
 - C++, Java?
 - Good coding practises?



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 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 fit a better model, present the results, then...



Implement our work

- **Example 1: Vinny is daft**
- Two years after my Ph.D. finished and my code isn't even on GitHub
- 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
 2. Help others implement it
 3. Allow time in project to implement the work in practise



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 experience to help do the best work I can!
- I would always use “Statistical Methods” first, “Machine Learning / AI Method” second

Thanks for listening!

Thank you

