

DR J ROGEL-SALAZAR

# INTRODUCTION TO DATA SCIENCE


A PRACTICAL POINT OF VIEW

ODSC Europe, London  
Sep 22nd, 2018

ODSC

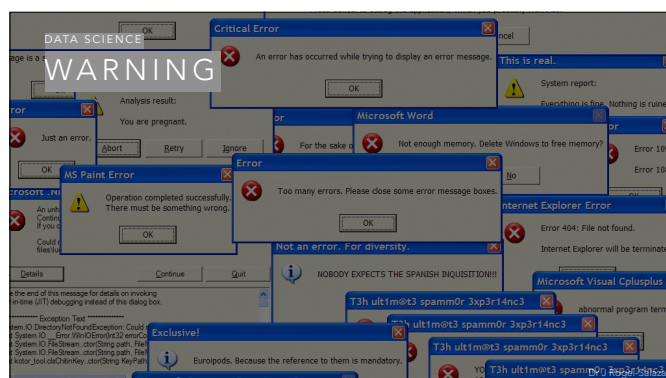
jrogel.datascience@icloud.com  
@quantum\_tunnel / @datascience

## SOCIAL MEDIA



- Use the hashtags
- #ODSC2018
- #IntroDataScience

Dr J Rogel-Salazar



## TODAY

- We will not be covering "everything data science"
- This is not a technical presentation.
- We will cover some general aspects about working in the field
- Discussion points - hopefully useful to all

### INSTALL.SH


```
#!/bin/bash

pip install "$1" &
easy_install "$1" &
brew install "$1" &
npm install "$1" &
yum install "$1" &
dnf install "$1" &
docker run "$1" &
pkg install "$1" &
apt-get install "$1" &
sudo apt-get install "$1" &
steamcmd +app_update "$1" validate &
git clone https://github.com/"$1"/"$1" &
cd "$1";./configure;make;make install &
curl "$1" | bash &
```

Dr J Rogel-Salazar

## AGENDA


- Data Science - A working definition
- Skills, competencies and more
- The data science workflow
- Tools and methodologies
- Challenges and opportunities



Dr J Rogel-Salazar

## DISCUSSION

- Who are you? What brings you to ODSC?
- What comes to mind when thinking of "data science"? Can you define it?
- What would you want to know or do if you could get access to ALL the transactional data (not just a small sample); if, for example, you had all data from the last 10 or 20 years?
- Thinking about your current rôle, what is the most important skill: programming, business knowledge, maths & stats, communication, other?
- What tech and tools do you use in your work?



Dr J Rogel-Salazar

DATA SCIENCE

WHO ARE YOU?  
WHAT BRINGS YOU TO ODSC?

Dr J Rogel-Salazar

A LITTLE BIT ABOUT me...

Physics

Imperial College London

DOW JONES

PRUDENTIAL

AKQA

DR J ROGEL-SALAZAR

DATA SCIENCE AND ANALYTICS WITH PYTHON

ESSENTIAL MATLAB AND OCTAVE

Dr J Rogel-Salazar

DATA SCIENCE

WHAT COMES TO MIND WHEN  
THINKING OF "DATA SCIENCE"?  
CAN YOU DEFINE IT?

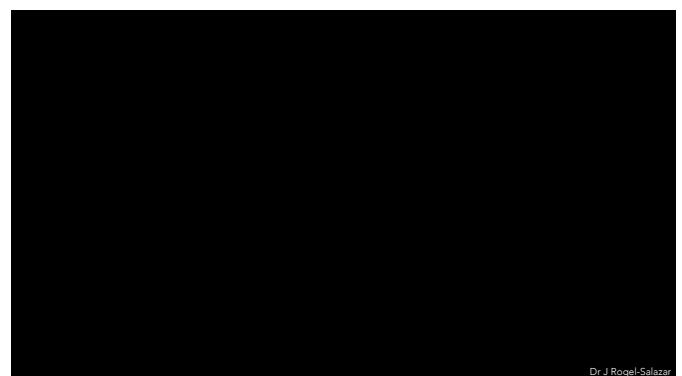
Dr J Rogel-Salazar



DATASCIENCE

A BRIEF HISTORY

Dr J Rogel-Salazar





Dr J Rogel-Salazar

## DATA SCIENCE

- Use of data (and metrics) to drive and inform key business decisions
- Employ data science and engineering to enhance product performance

Dr J Rogel-Salazar

“And, what about other buzz words used in the area?”

Dr J Rogel-Salazar

## Computer Science

Artificial Intelligence	Machine Learning	Deep Learning
Intelligence displayed by machines	A pillar of artificial intelligence	A subset of machine learning

Dr J Rogel-Salazar

# MACHINE LEARNING

DATA

Pattern recognition in the data

And make predictions...

Dr J Rogel-Salazar

## Tasks in Machine Learning

Techniques to find patterns are called **unsupervised tasks**

UNSUPERVISED MACHINE LEARNING

Those that make predictions are **supervised tasks**

SUPERVISED MACHINE LEARNING

Dr J Rogel-Salazar



## A generic term

Does not fit in memory, or  
Does not fit in a machine...

Dr J Rogel-Salazar

## DATA... SCIENTISTS...

- There are many definitions....
- "Data science is the combination of analytics and the development of new algorithms... You may have to invent something, but it's okay if you can answer a question just by counting. The key is making the effort to ask the questions."

Hillary Mason



Dr J Rogel-Salazar

## DATA SCIENCE

WHAT WOULD YOU WANT TO KNOW OR DO IF  
YOU COULD GET ACCESS TO ALL THE  
TRANSACTIONAL DATA...?

Dr J Rogel-Salazar

## DATA SCIENCE

WHERE? WHAT? WHO?

Dr J Rogel-Salazar

## INDUSTRIES



Dr J Rogel-Salazar

## SOME USE CASES



Customer  
Retention



Marketing



Security



Life Event  
Prediction



Product  
Recommendation

Dr J Rogel-Salazar

DATA SCIENCE

CONSIDER THE FOLLOWING...

Dr J Rogel-Salazar

DATA SCIENCE

DIGITAL WORLD


The world's largest taxi company...

owns no cars



The world's most popular media content...

creates no content



The world's most valuable retailer...

owns no stores



The world's largest accommodation provider...

owns no state



The world's largest

...




Dr J Rogel-Salazar

DATA SCIENCE

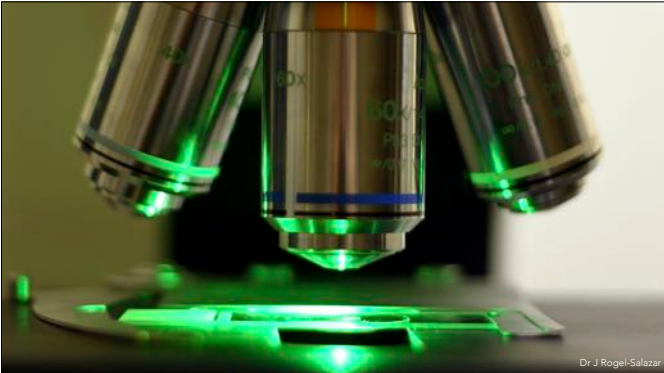
WHAT IS THE MOST IMPORTANT SKILL:  
PROGRAMMING, BUSINESS KNOWLEDGE,  
MATHS & STATS, COMMUNICATION,  
OTHER?

Dr J Rogel-Salazar

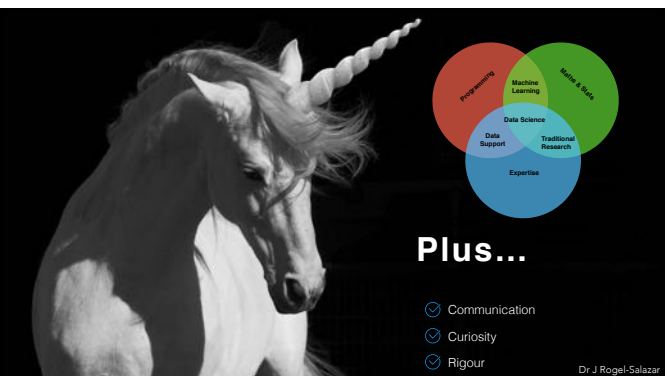
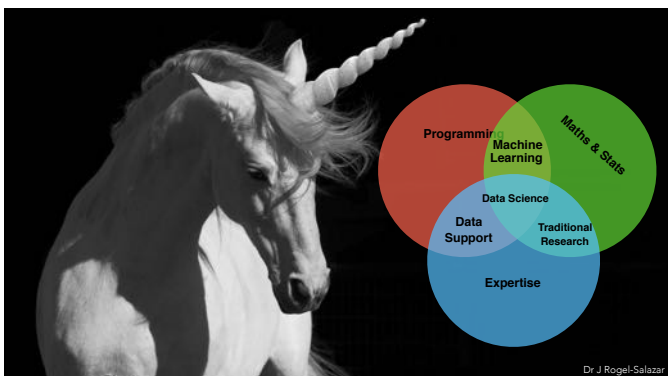


AND AFTER BIG DATA  
SOLVES ALL OUR PROBLEMS,  
WE'LL RIDE AWAY ON  
MAGIC FLYING UNICORNS.

Dr J Rogel-Salazar









# JACKALOPE DATA SCIENTISTS

- Or... you can be more realistic and become a Jackalope data scientist
- This can help you build an effective team

Dr J Rogel-Salazar

# JACKALOPE?

- A mythical animal of North American folklore:
- A Jackrabbit with antelope horns
- Douglas Herrick (1930s) popularised the jackalope by making one
- Mythical but with a hint of reality...
- the cottontail rabbit papilloma virus (CRPV) infects certain leporids, causing keratinous carcinomas resembling horns

Dr J Rogel-Salazar

# DATA SCIENCE

## THE WORKFLOW

Dr J Rogel-Salazar

# EXPLOITING VALUE IN DATA

Understand what's happening

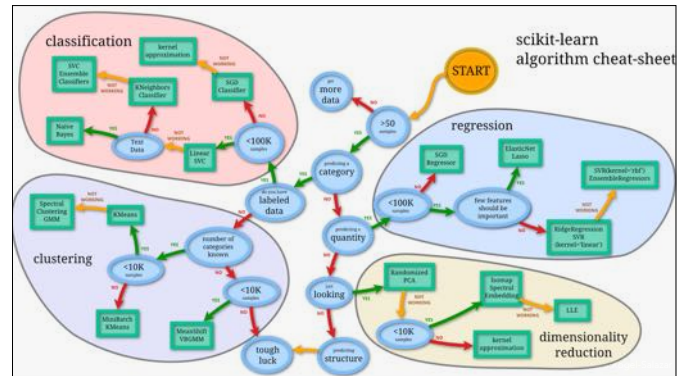
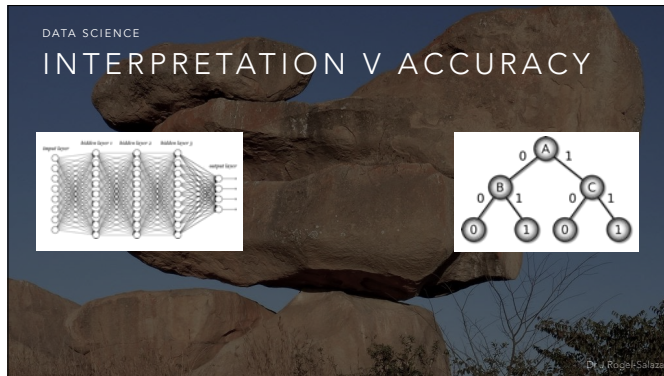
Mine content

Accelerate decisions

Optimise processes

Find patterns & Derive Insights

Dr J Rogel-Salazar



## REQUIREMENTS

- Understand what we need from the model
- Have the clearest idea on what you want. Consider the technological setup
- Dedicate enough communication and mutual understanding between parts
- Having a clear requirement helps with the main ingredient: data!

Dr J Rogel-Salazar

## DATA

- Data is the raw material for the model
- Good quality data is better
- Data preparation is almost always necessary.
- You may not need machine learning!!

Dr J Rogel-Salazar

## MODEL

- Start simple
- Tinkering with new techniques is fun...
- But, clients need a problem fixed
- Agree on the metrics
- Beware the asymptotic chase

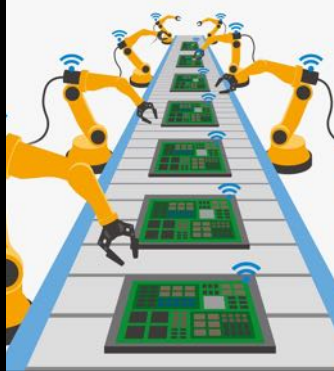
Dr J Rogel-Salazar



## PRODUCTION

- Production-ready model:
  - Integration with existing setup - Decouple!
  - Updating the model - Is retraining needed?
- Both are not mandatory, but...
- It all depends!

Dr J Rogel-Salazar



## GDPR

- GDPR has been adopted and is applicable as of 2018
- It also has international reach – applying to any organisation that processes data of EU data subjects
- Fines for non-compliance will increase substantially up to a maximum fine of €20 million or 4% of global annual sales, whichever is higher
- GDPR has fundamentally change the way companies must manage their data
- Any handling of Personal Data throughout its entire life cycle, from collection to deletion, is considered “processing”. Even remote access is considered “processing”

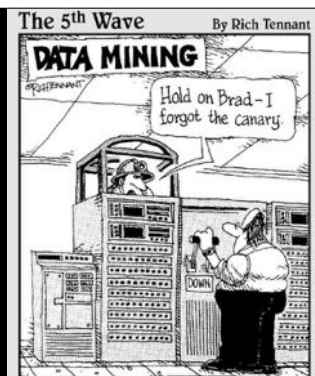
Dr J Rogel-Salazar



## DATA SCIENCE

## TOOLS AND METHODOLOGIES

Dr J Rogel-Salazar

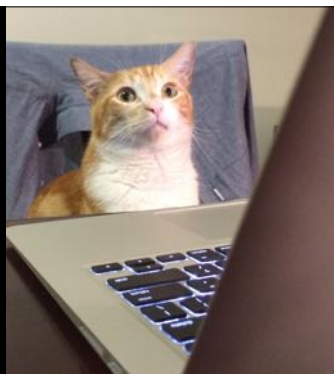


Dr J Rogel-Salazar

## PROGRAMMING

- Obtaining data from DBs, APIs
- Processing data
- Reproducible
- Automation

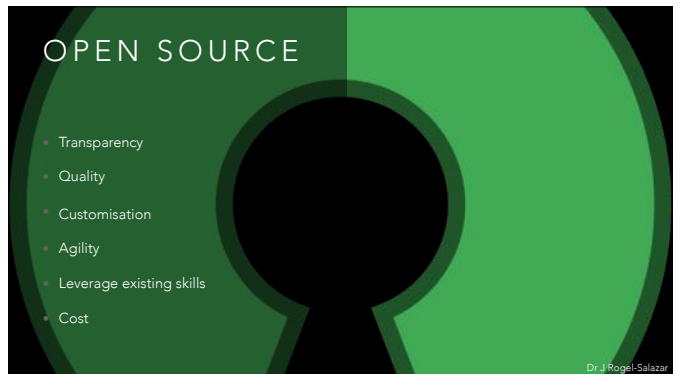
Dr J Rogel-Salazar



## OPEN SOURCE

- Transparency
- Quality
- Customisation
- Agility
- Leverage existing skills
- Cost

Dr J Rogel-Salazar



## R

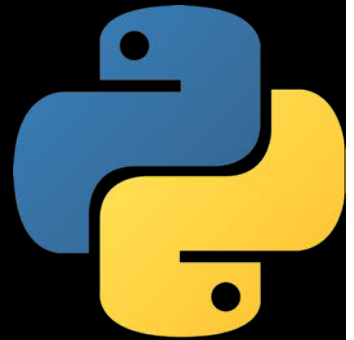
- A statistical programming language
- R started out as an implementation of S
- Developed by statisticians for statisticians
- Cutting-edge algos available
- ggplot2



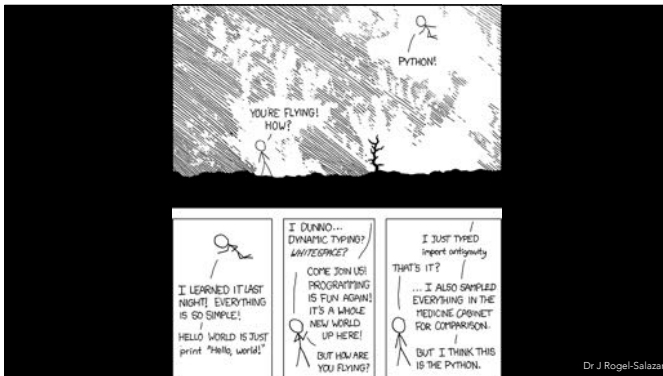
Dr J Rogel-Salazar

## PYTHON

- A very useful general-purpose language
- Rapidly growing since 2000s
- Libraries to access DBs, APIs, machine learning, plotting, network analysis, NLP, web dev, etc
- For reference: it is interpreted and dynamically typed



Dr J Rogel-Salazar



Dr J Rogel-Salazar

## DATABASES - SQL AND NOSQL

- Relational databases
- SQL - Structured query database
- The main DB tech for many a year
- What has changed?



Dr J Rogel-Salazar

## NOSQL

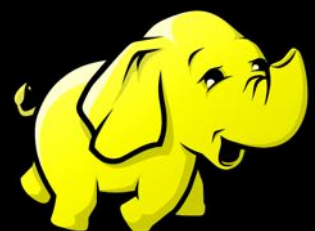
- Emergence of web scale data
- Distributed, large scale, non structured data



Dr J Rogel-Salazar

## HADOOP

- Comes from an effort to create an open source search engine
- Yahoo! was an important contributor and a large user
- It is not a database, but a data storage and management system
- Data extracted and manipulated via MapReduce



Dr J Rogel-Salazar

## SPARK

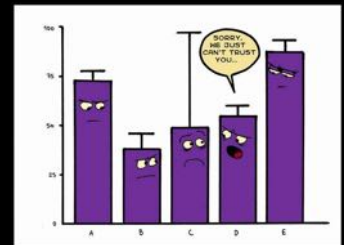
- Fast, in-memory data processing engine
- Real-time stream processing
- Integration - Hadoop for example
- Core APIs - Scala, Java, R, Python, SQL



Dr J Rogel-Salazar

## STATISTICS

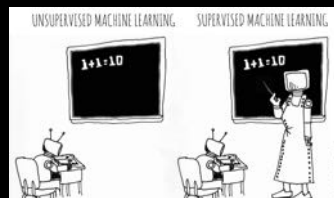
- A science of data
- Predates computers
- Emphasises formal statistical inference (in low dimensionality)
  - Confidence intervals
  - Hypothesis tests



Dr J Rogel-Salazar

## MACHINE LEARNING

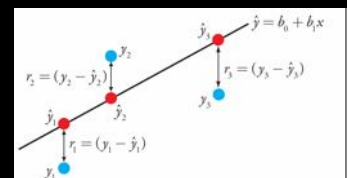
- Close to stats
- A computer science discipline
- There is an algorithmic component
- Many concepts are similar to stats, but with a different name



Dr J Rogel-Salazar

## STATS - REGRESSION

- The "workhorse of data science"
- Allows us to characterise relationships between variables
- Can be used to build a model to predict values of  $y$  given observations of  $x$



Dr J Rogel-Salazar

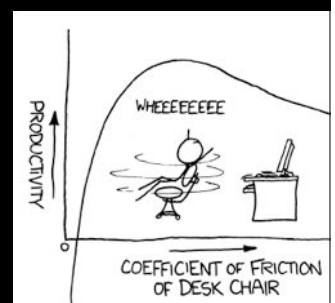
DATA SCIENCE

## CHALLENGES AND OPPORTUNITIES

Dr J Rogel-Salazar

## YOUR JOB


- The organisation you work for probably generates more data than you imagine
- Is the value from each data source exhausted?
- Is data from multiple sources? Combined?
- Are users aware of possibilities offered by SNA, NLP, others?



Dr J Rogel-Salazar

## KAGGLE

- Crowd-sourcing analytics problems
- Thousands compete by using whatever methods to produce best prediction
- Cash prizes




problem data crowd tools models

Dr J Rogel-Salazar

## YOUR STARTUP

- Software development becomes commoditised
- Many not very technical ideas only need a WordPress install
- Many new companies differentiate themselves through their use of data



Dr J Rogel-Salazar

## NEW JOBS

- In April 2012 McKinsey predicted 1.5 million shortage of data scientists
- More and more companies are looking for people to unlock the value in their data
- Rise in available positions




Dr J Rogel-Salazar

## DATA SCIENCE IN SUMMARY

Dr J Rogel-Salazar

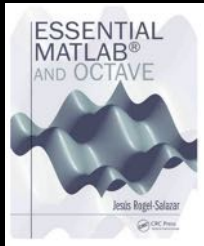
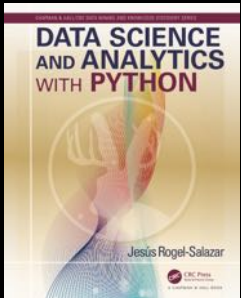
Data science is a product of our time

- Being a data scientist requires people and technical skills
- Data science is a team sport
- Methodological and rigorous approach in a business context



Dr J Rogel-Salazar

## BOOKS - KEEP AN EYE

Dr J Rogel-Salazar



DR J ROGEL-SALAZAR

THANKS A LOT

[jrogel.datascience@cloud.com](mailto:jrogel.datascience@cloud.com)  
[@quantum\\_tunnel](#) / [@dat\\_science](#)

ODSC Europe, London  
Sep 22nd, 2018

