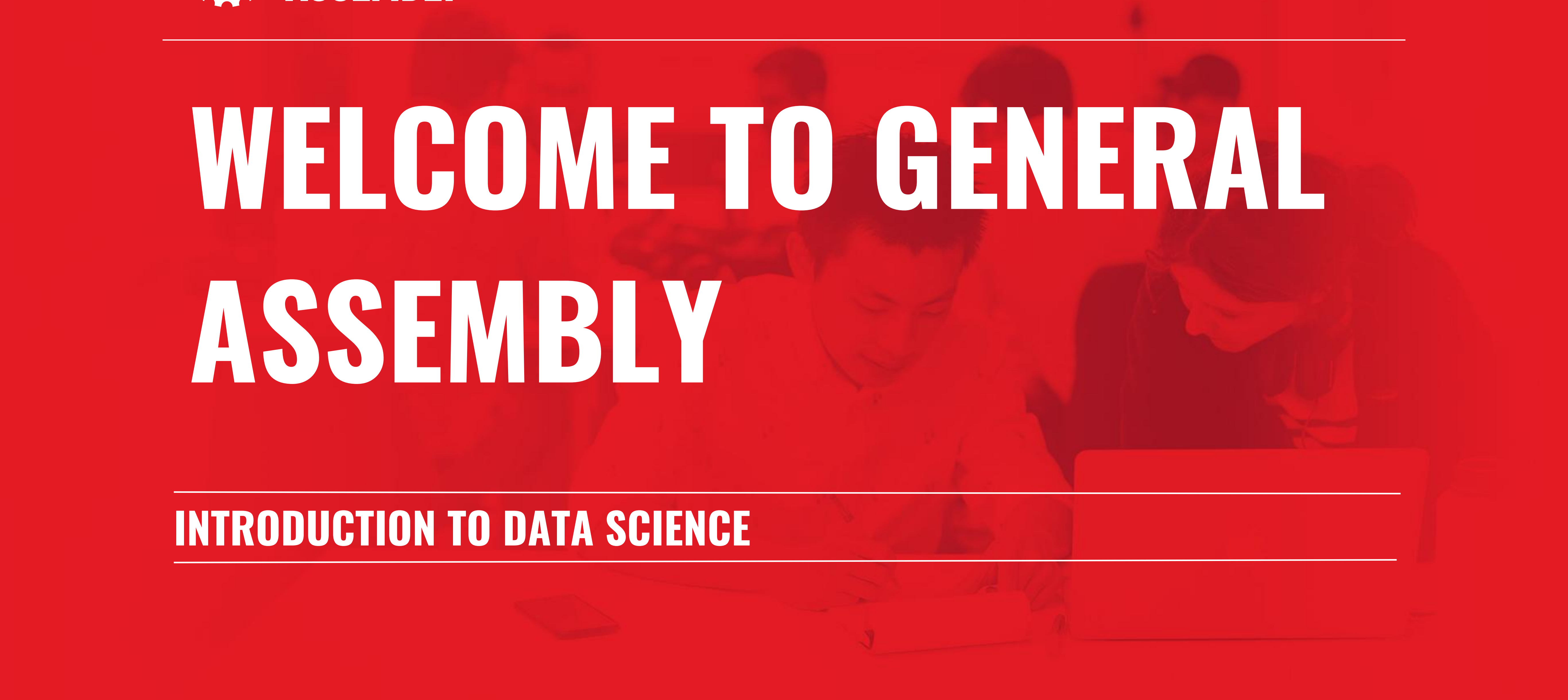
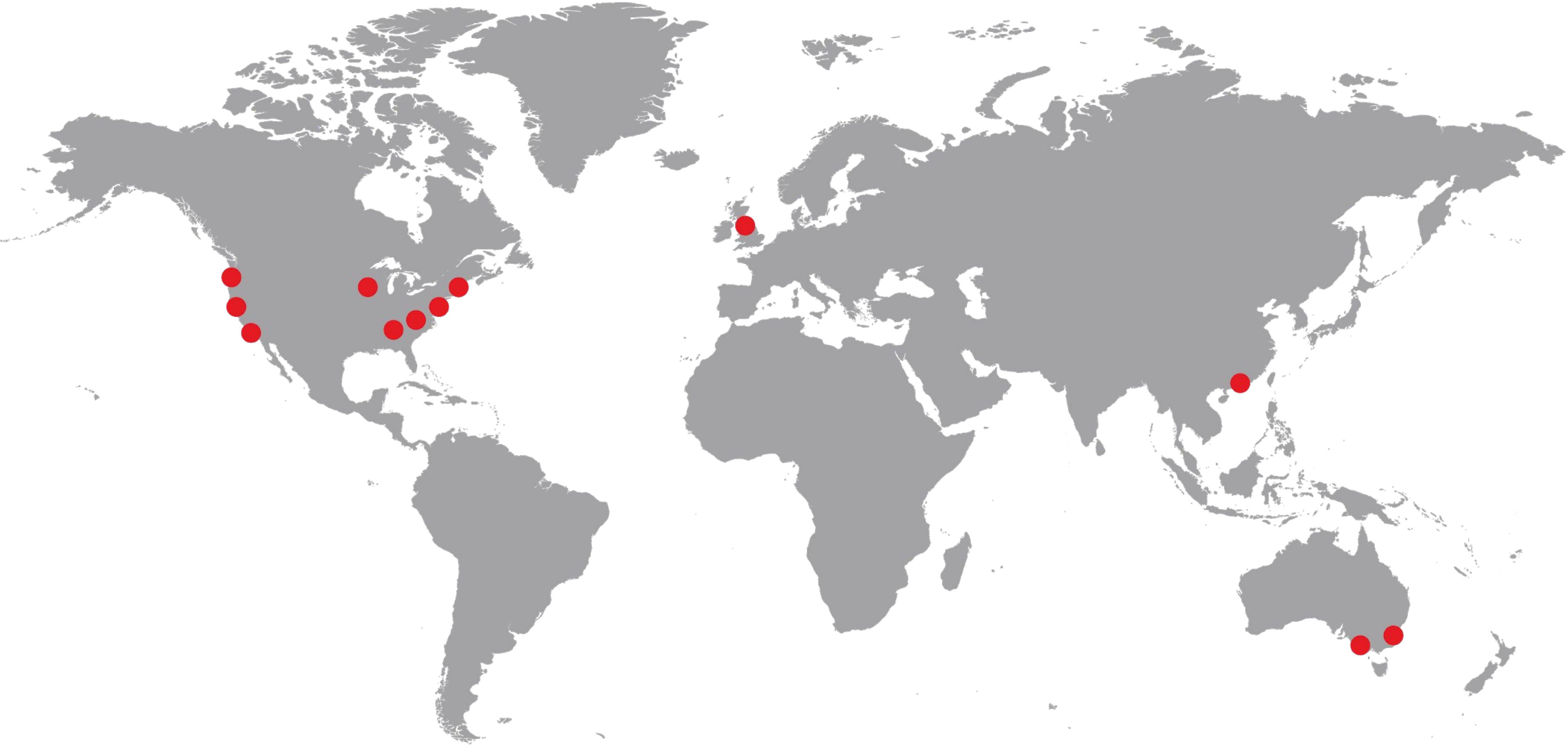
GENERAL ASSEMBLY



WELCOME TO GENERAL ASSEMBLY

INTRODUCTION TO DATA SCIENCE



Let’s get started...

INTRODUCTION TO DATA SCIENCE

AGENDA

- What Is Data Science?
- How Is Data Science Used?
- The Data Science Workflow
- Practical Data Science Example
- Questions and Next Steps

INTRODUCTION TO DATA SCIENCE

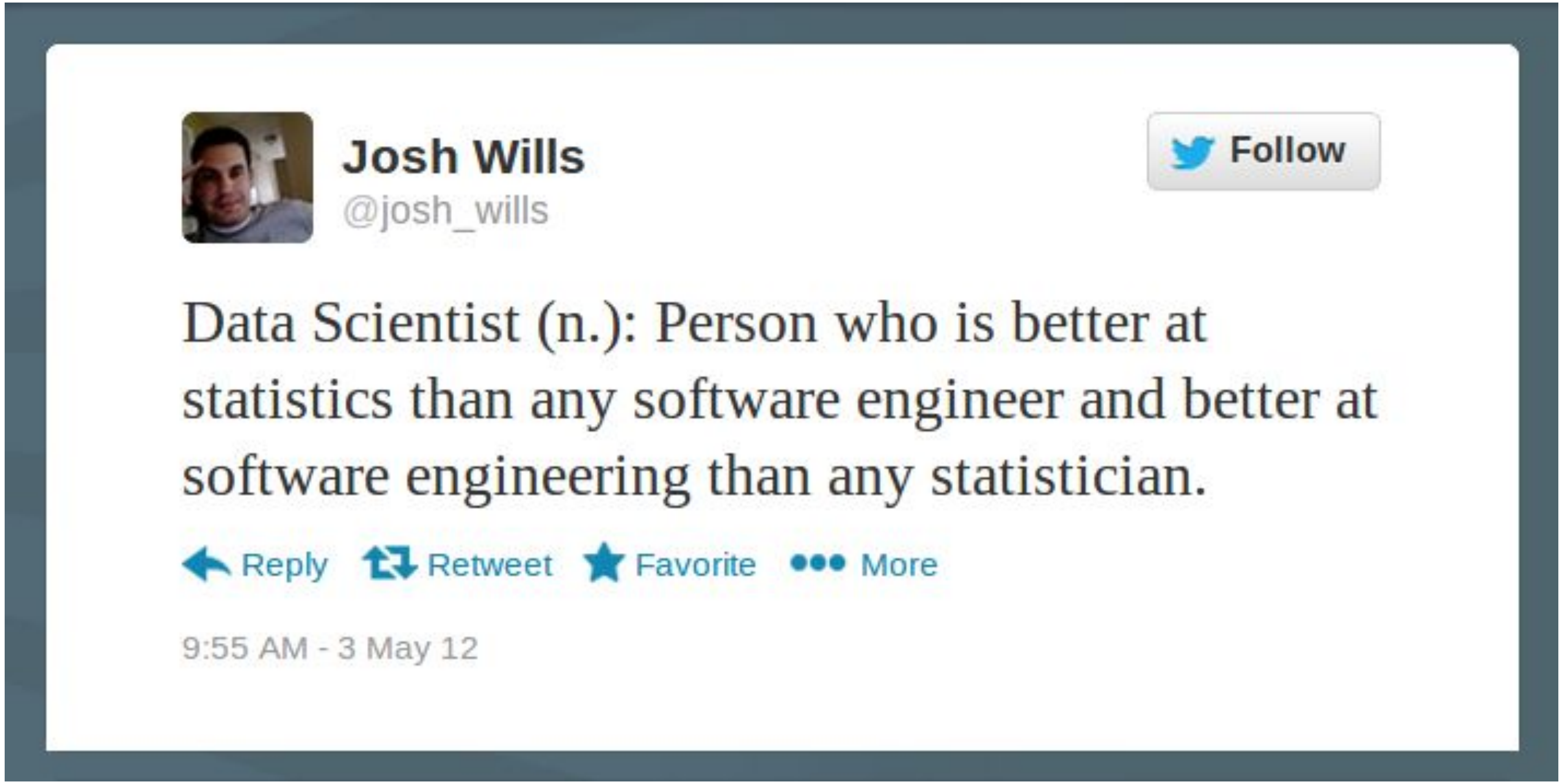
WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

> “A data scientist is a ***statistician*** who lives in ***San Francisco.***”

@cdixon

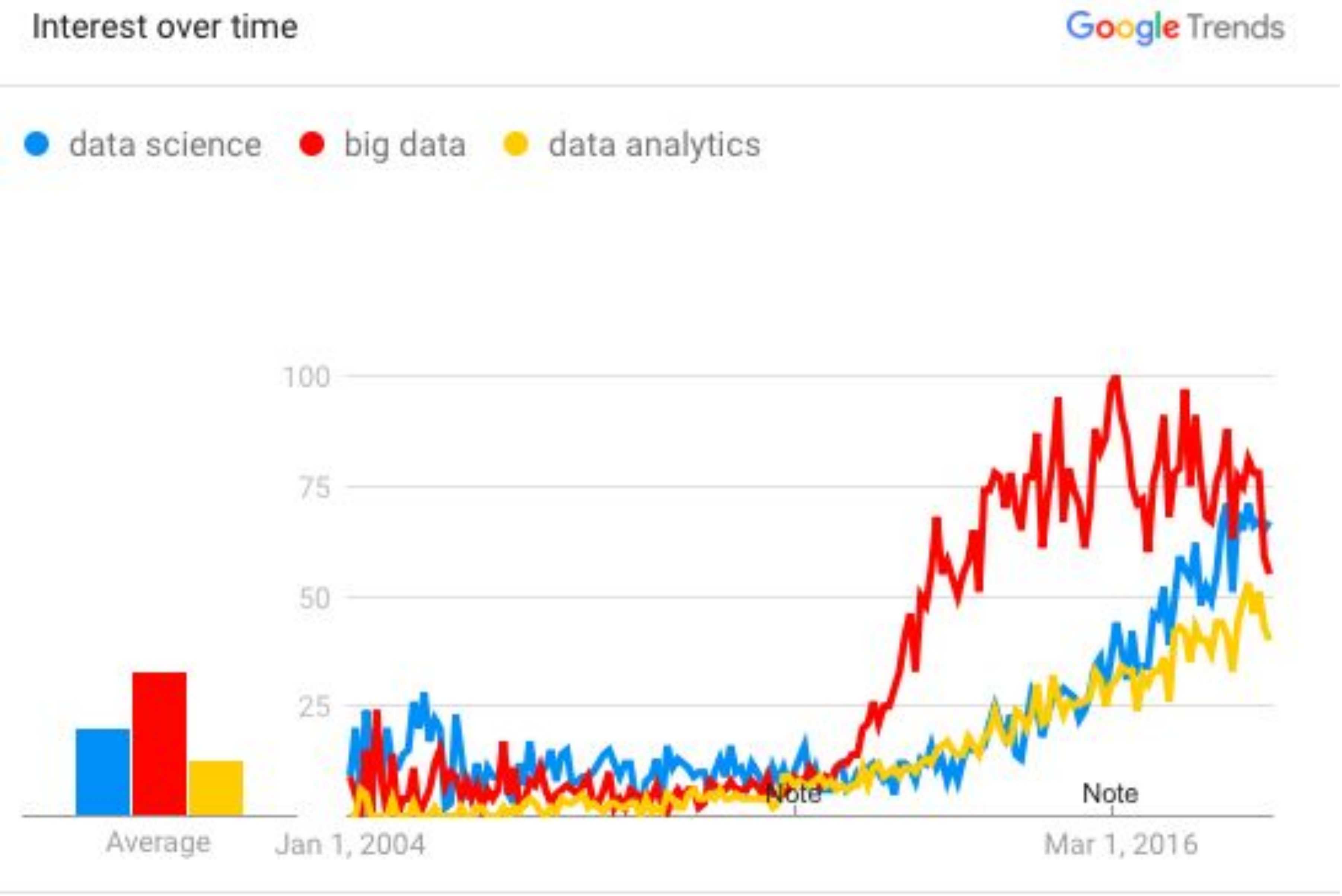
WHAT IS DATA SCIENCE?



Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

9:55 AM - 3 May 12

WHAT IS DATA SCIENCE?



WHAT IS DATA SCIENCE?

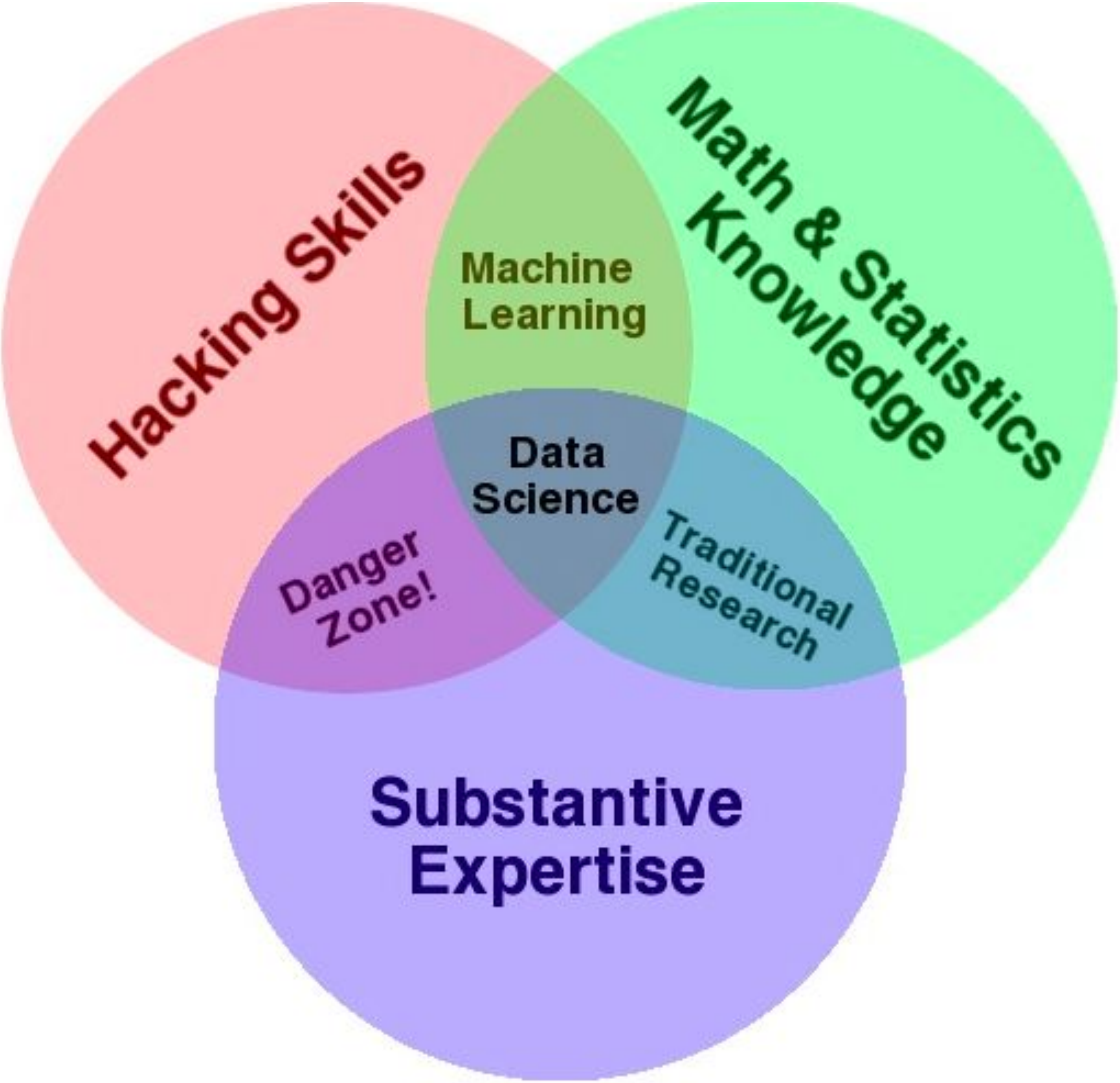


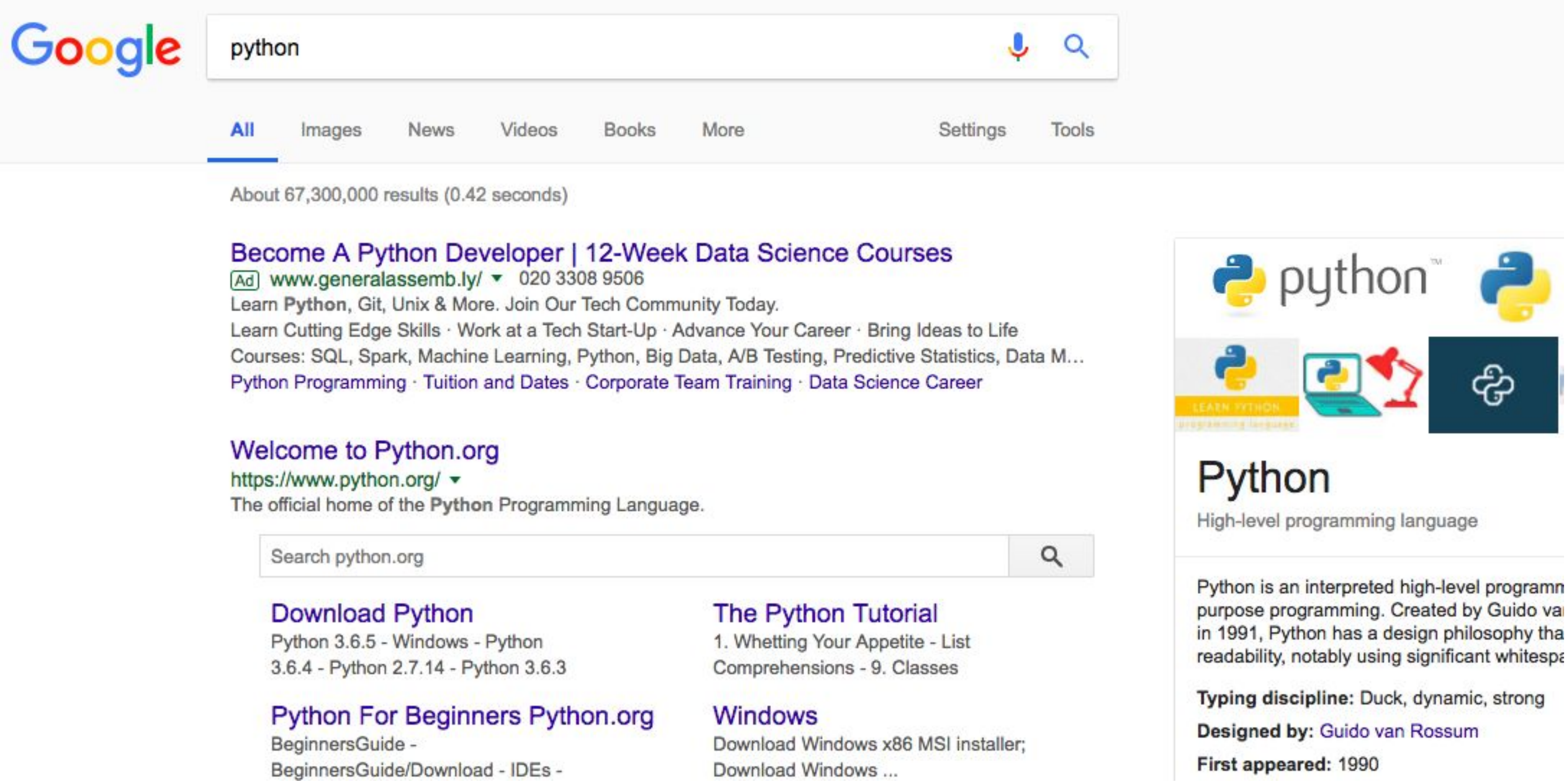
Figure 1-1. Drew Conway's Venn diagram of data science

WHAT IS DATA SCIENCE?

- A set of **tools and techniques** used to extract **useful information** from data
- **Interdisciplinary**, but domain-centric
- **Evidence-based problem solving** and decision-making
- The application of **scientific techniques** to practical problems

HOW IS DATA SCIENCE USED?

HOW IS DATA SCIENCE USED?



About 67,300,000 results (0.42 seconds)

Become A Python Developer | 12-Week Data Science Courses
Ad www.generalassemb.ly/ ▾ 020 3308 9506
Learn **Python**, Git, Unix & More. Join Our Tech Community Today.
Learn Cutting Edge Skills · Work at a Tech Start-Up · Advance Your Career · Bring Ideas to Life
Courses: SQL, Spark, Machine Learning, Python, Big Data, A/B Testing, Predictive Statistics, Data M...
Python Programming · Tuition and Dates · Corporate Team Training · Data Science Career

Welcome to Python.org
https://www.python.org/ ▾
The official home of the **Python** Programming Language.

Download Python
Python 3.6.5 - Windows - Python 3.6.4 - Python 2.7.14 - Python 3.6.3

The Python Tutorial
1. Whetting Your Appetite - List Comprehensions - 9. Classes

Python For Beginners Python.org
BeginnersGuide - BeginnersGuide/Download - IDEs -

Windows
Download Windows x86 MSI installer; Download Windows ...

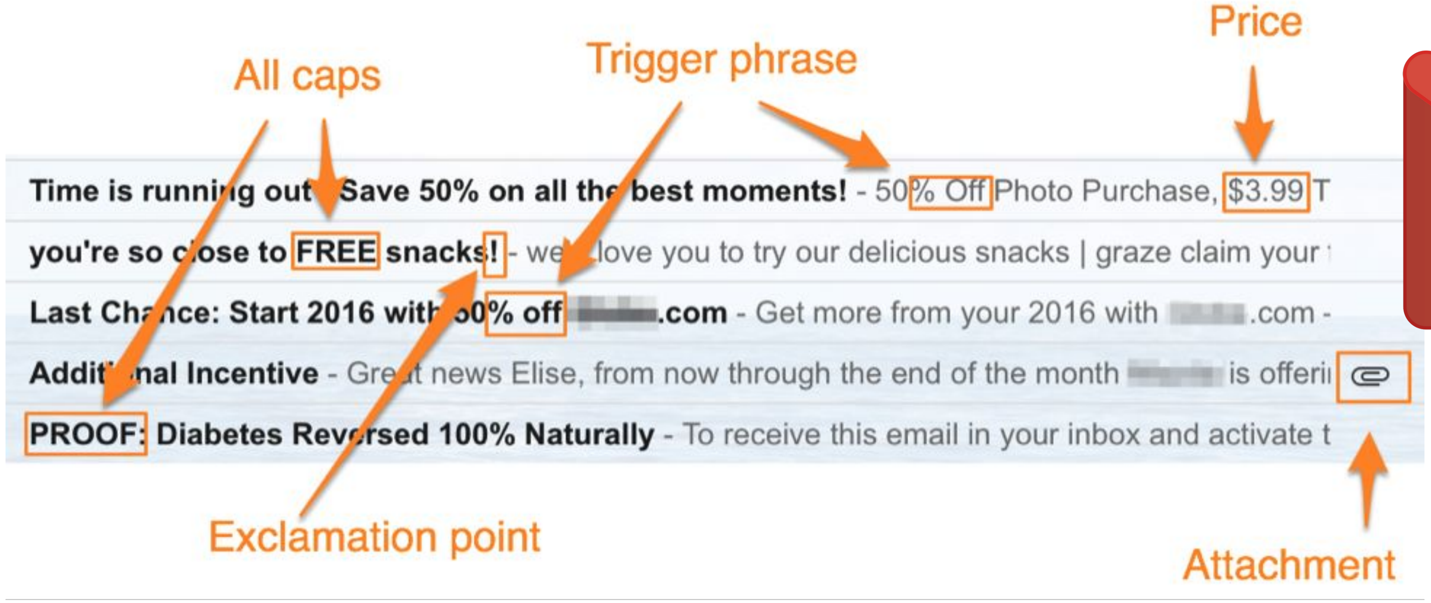
Python
High-level programming language

Python is an interpreted high-level programm
purpose programming. Created by Guido va
in 1991, Python has a design philosophy tha
readability, notably using significant whitesp

Typing discipline: Duck, dynamic, strong
Designed by: Guido van Rossum
First appeared: 1990

HOW IS DATA SCIENCE USED?

NETFLIX



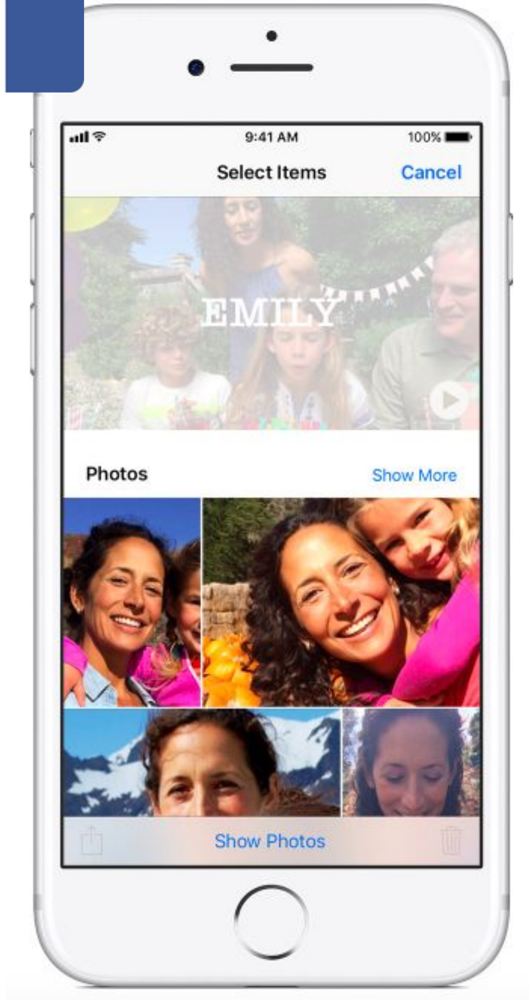
Time is running out! Save 50% on all the best moments! - 50% Off Photo Purchase, $3.99 T

you're so close to FREE snacks! - we'd love you to try our delicious snacks | graze claim your

Last Chance: Start 2016 with 50% off [redacted].com - Get more from your 2016 with [redacted].com -

Additional Incentive - Great news Elise, from now through the end of the month [redacted] is offeri

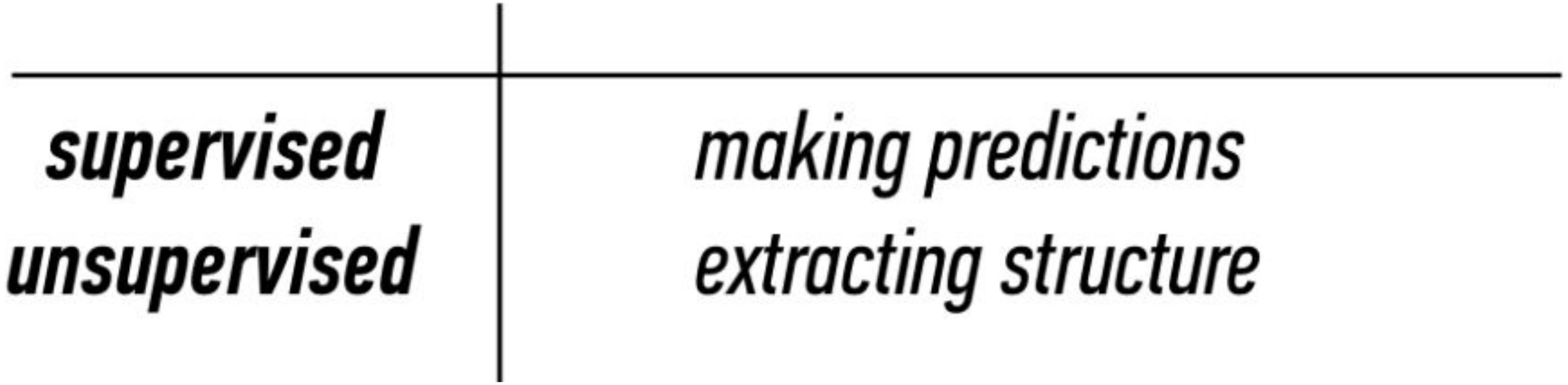
PROOF: Diabetes Reversed 100% Naturally - To receive this email in your inbox and activate t



DIFFERENT TYPES OF MACHINE LEARNING

HOW IS DATA SCIENCE USED?

SUPERVISED VS. UNSUPERVISED

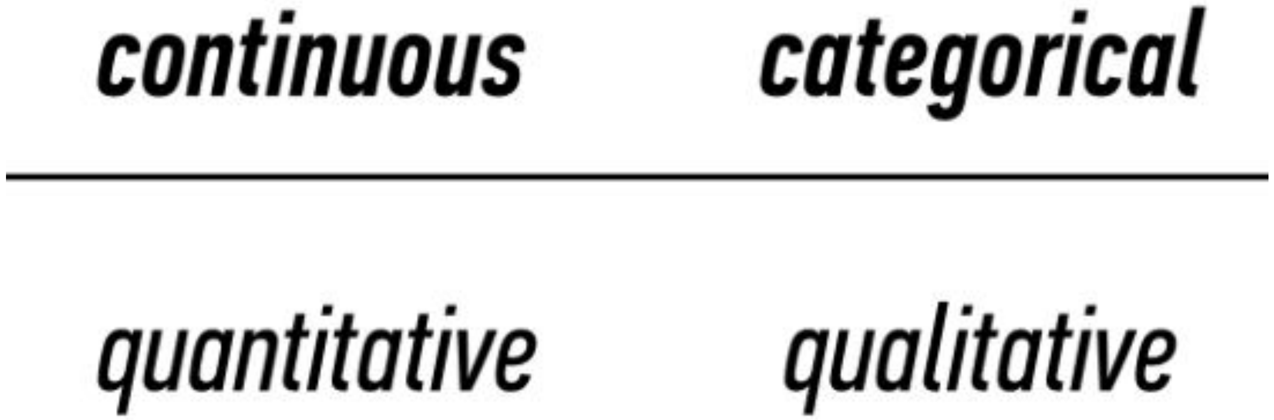


supervised	*making predictions*
unsupervised	*extracting structure*

- Supervised: **Generalization**
- **Unsupervised: Representation**

HOW IS DATA SCIENCE USED?

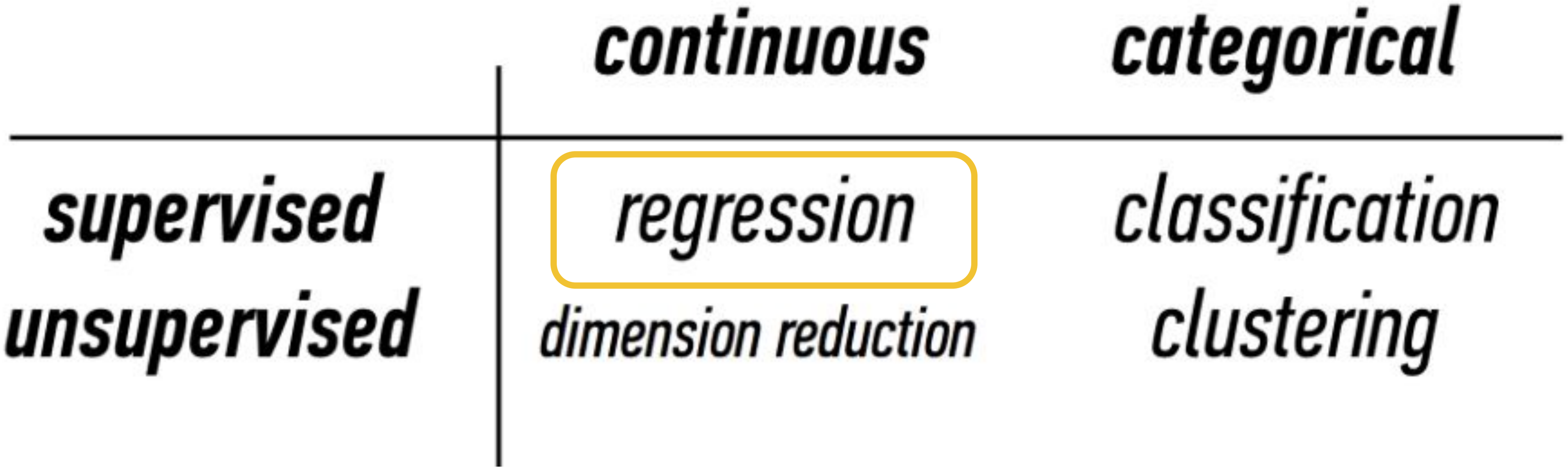
CONTINUOUS VS. CATEGORICAL



continuous	categorical
quantitative	qualitative

HOW IS DATA SCIENCE USED?

MACHINE LEARNING PROBLEMS



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

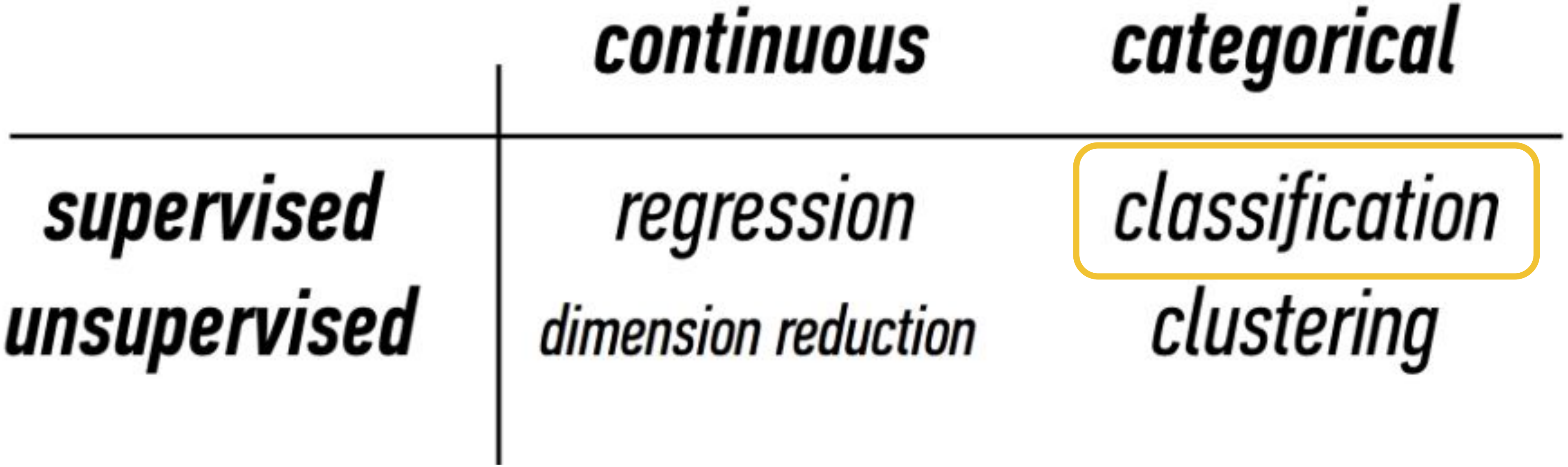
HOW IS DATA SCIENCE USED?

REGRESSION EXAMPLE
(CONTINUOUS, SUPERVISED)

- If a company want to expand into a new city, can they predict their sales based on other locations?
- Target is continuous - can use a regression algorithm
- GDP, city population, growth, average salaries, average rent

HOW IS DATA SCIENCE USED?

MACHINE LEARNING PROBLEMS



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

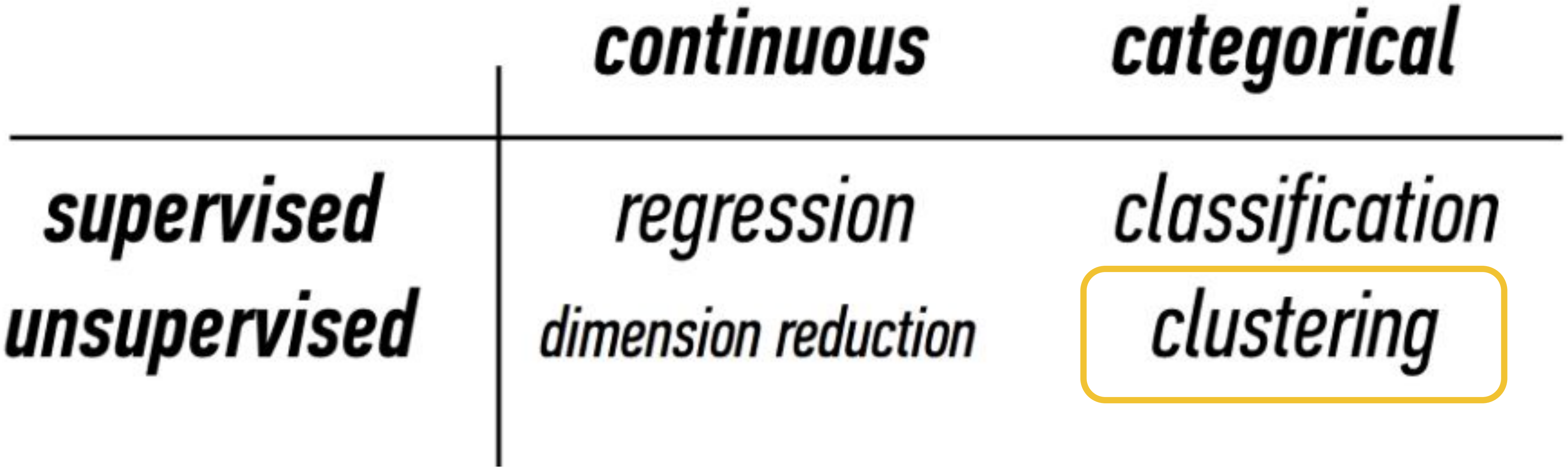
HOW IS DATA SCIENCE USED?

CLASSIFICATION
(CATEGORICAL, SUPERVISED)

- Can we identify when a transaction is fraudulent
- Binary response - can use a classification algorithm
- Past buying behaviours of customer
- Average amount, time since last transaction, average time since last transaction, average number of merchants, currency, time of day

HOW IS DATA SCIENCE USED?

MACHINE LEARNING PROBLEMS



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

HOW IS DATA SCIENCE USED?

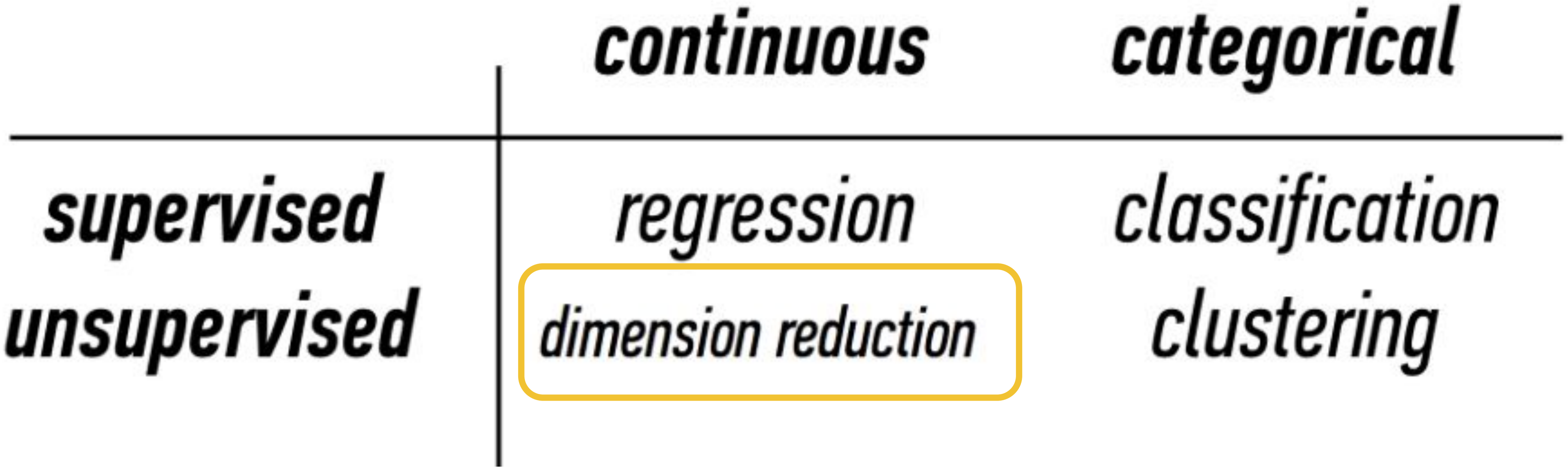
CLUSTERING

(CATEGORIAL, UNSUPERVISED)

- Can we identify groups of users, through looking for patterns in behaviour as they move through the site
- Looking for (unknown) structure in the data
- Data not labelled - unsupervised
- Looking for natural groupings in the data

HOW IS DATA SCIENCE USED?

MACHINE LEARNING PROBLEMS



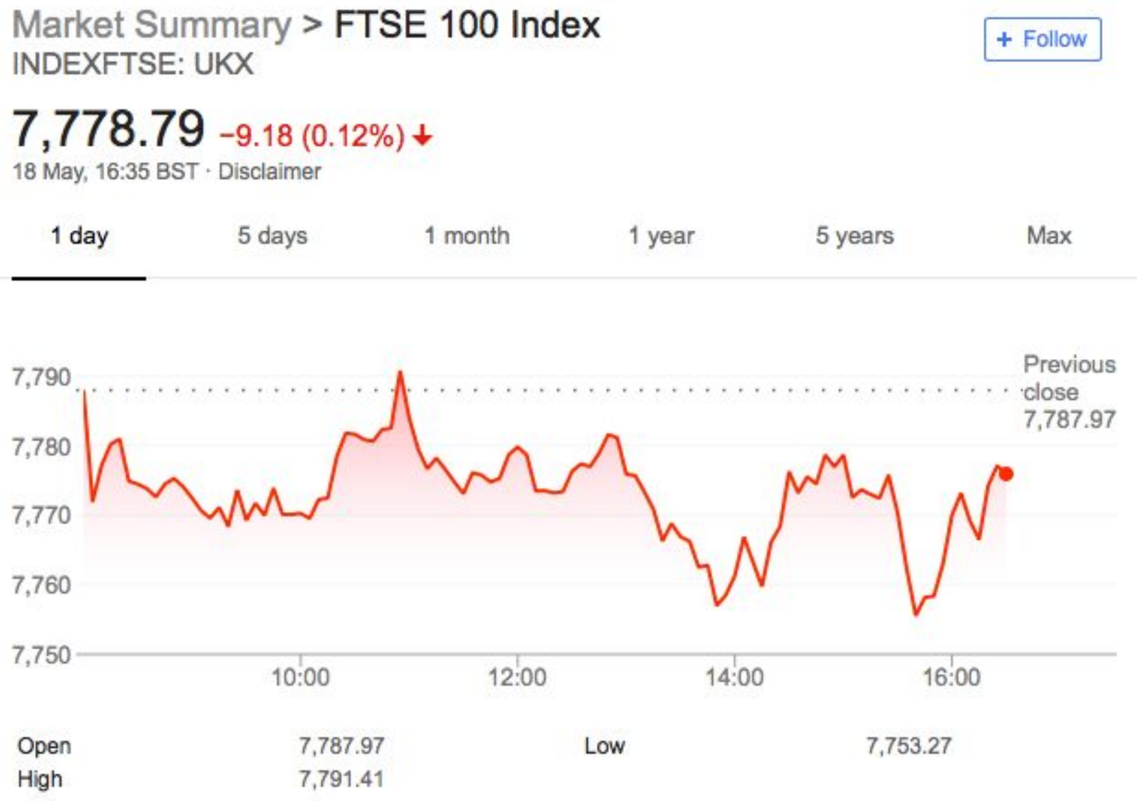
	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

HOW IS DATA SCIENCE USED?

DIMENSIONALITY REDUCTION

(CONTINUOUS, UNSUPERVISED)

- Every new feature adds a <u>dimension</u>
- Can we reduce the noise to find a signal?
- Want to understand the prosperity of UK businesses
- Could look at share price of each business individually or we could look at a weighted average like the FTSE 100.



Market Summary > FTSE 100 Index
INDEXFTSE: UKX

7,778.79 −9.18 (0.12%) ↓

18 May, 16:35 BST · Disclaimer

Open	7,787.97	Low	7,753.27
High	7,791.41		

HOW IS DATA SCIENCE USED?

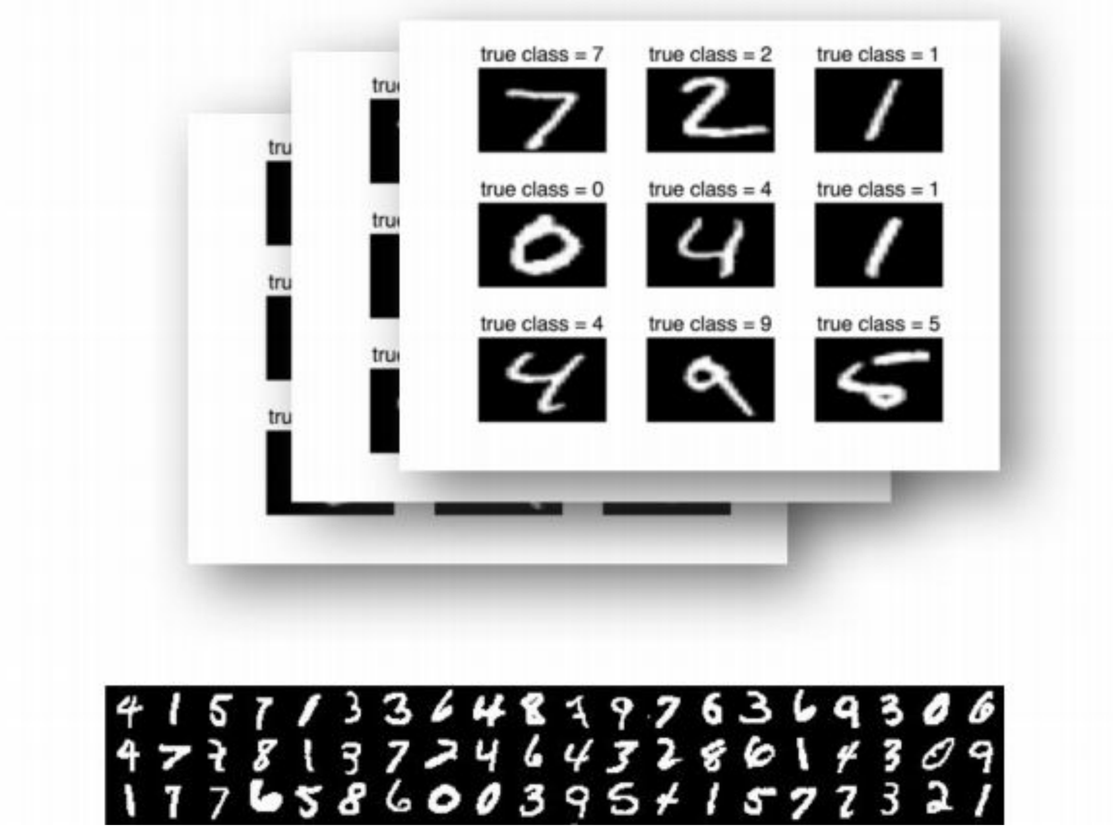
WHAT TYPE OF PROBLEM? – HOUSE PRICES

HOW IS DATA SCIENCE USED?

WHAT TYPE OF PROBLEM? – LAND USE

HOW IS DATA SCIENCE USED?

WHAT TYPE OF PROBLEM? – OCR



HOW IS DATA SCIENCE USED?

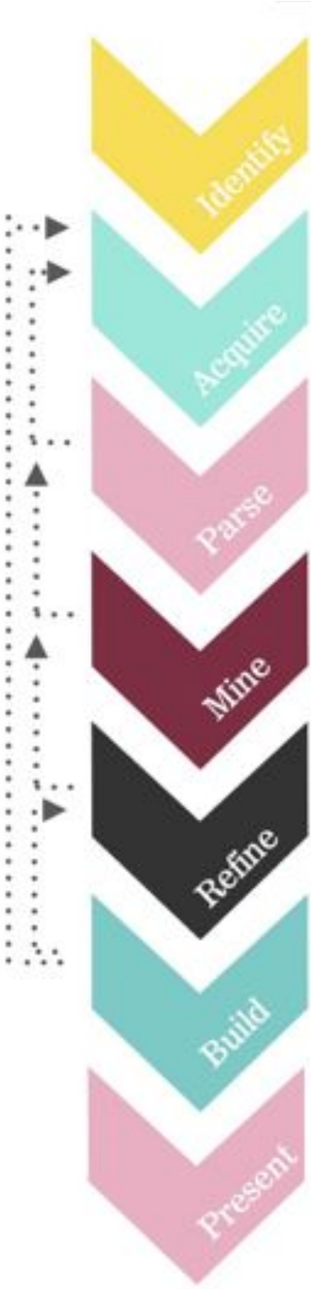
WHAT TYPE OF PROBLEM? – MORTGAGE APPLICATION

THE DATA SCIENCE WORKFLOW

WHAT IS DATA SCIENCE?

DATA SCIENCE WORKFLOW

1. Identify the problem
2. Obtain the data
3. Explore the data available
4. Mine the data
5. Build a model
6. Present the results
7. Deploy the model



GETTING STARTED WITH DATA SCIENCE

https://github.com/hapenfold/intro_to_data_science

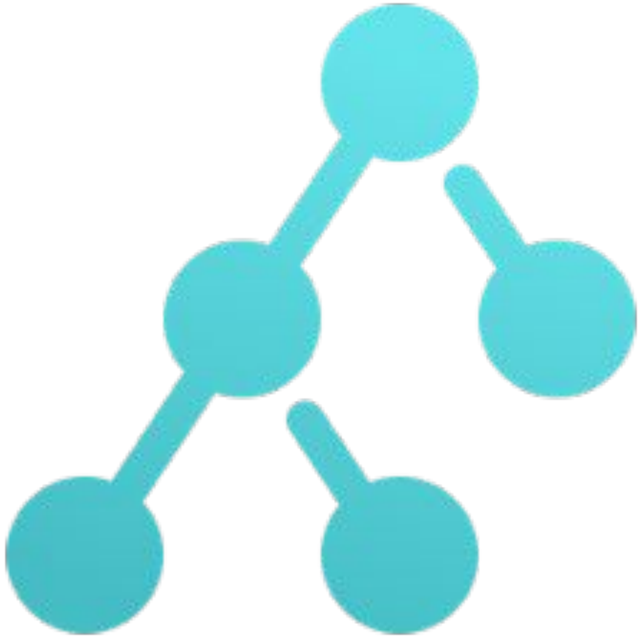
MODELLING MISCONCEPTIONS

Most well-executed Data Science projects don’t..

- Use complicated tools
- Fit complicated models

Instead, they…

- Focus on solving the problem
- Use appropriate data
- Use relatively standard models



MODELLING MISCONCEPTIONS

80-20 rule of modelling

- The first reasonable thing you can do gets you 80% of the way
- Everything after that is the remaining 20%

Often at a significant additional cost

INTRODUCTION TO DATA SCIENCE

WRAP-UP

Data Ethics

- Predictive modelling often reinforces traditional stereotypes
- Data is incredibly powerful and is making huge changes to society, therefore we need to make sure that the change that is coming, is the one that we **all** want to see.

Next-Steps

- Meetups - get a chance to meet people and learn more about the field. Go to meetup.com and find one that looks interesting
- Kaggle competitions
- Hackathons
- Free online courses - start learning to code!
- Data science bootcamp
- Khan Academy
- Statistics - “Elements of statistical learning” and “An introduction to statistical learning in R”
- Read white papers and blogs

We appreciate your feedback!

Please take 60 seconds to complete our survey.

www.ga.co/survey

Questions?