

INTRODUCTION TO DATA SCIENCE

Tan Kwan Chong

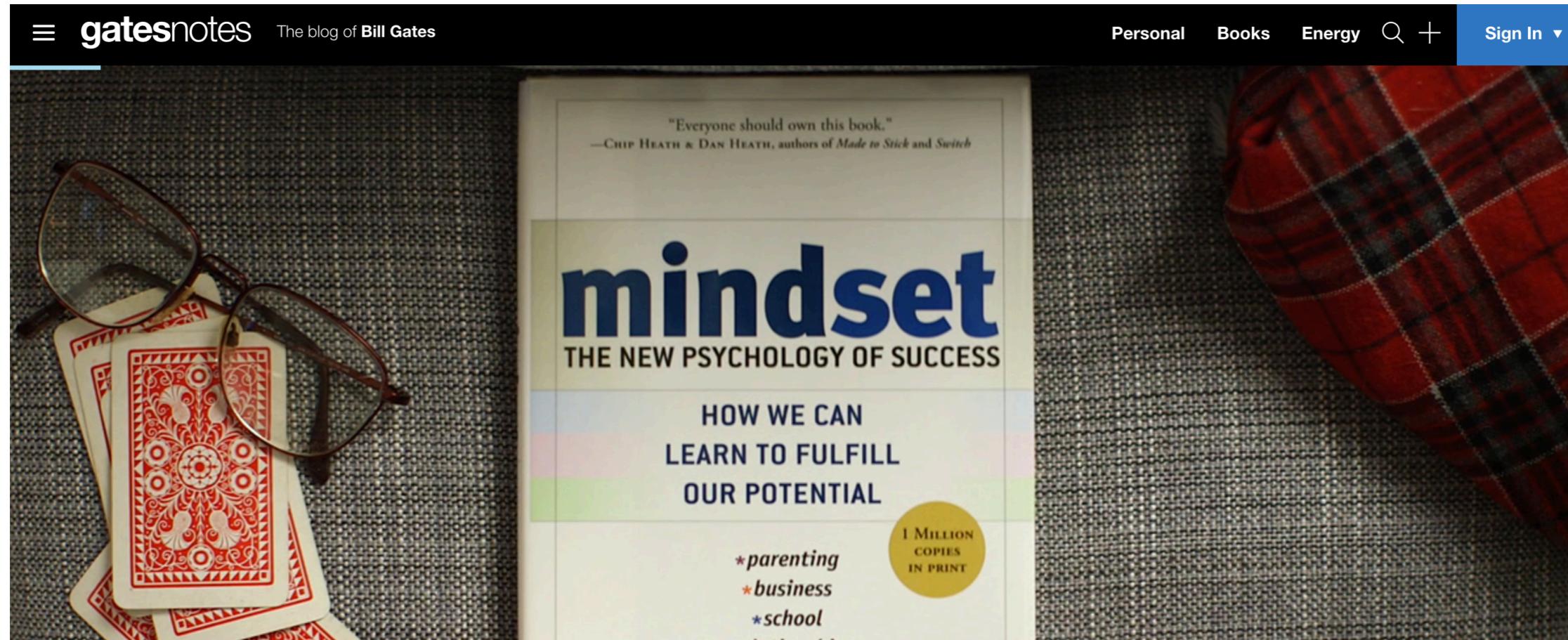
Chief Data Scientist, Booz Allen Hamilton

SELF INTRODUCTIONS

Hello
my name is

- Name
- Background
- What do you want to gain out of this course?
- Experience in programming and data analytics?

GROWTH VS FIXED MINDSET



Mindset Over Matter

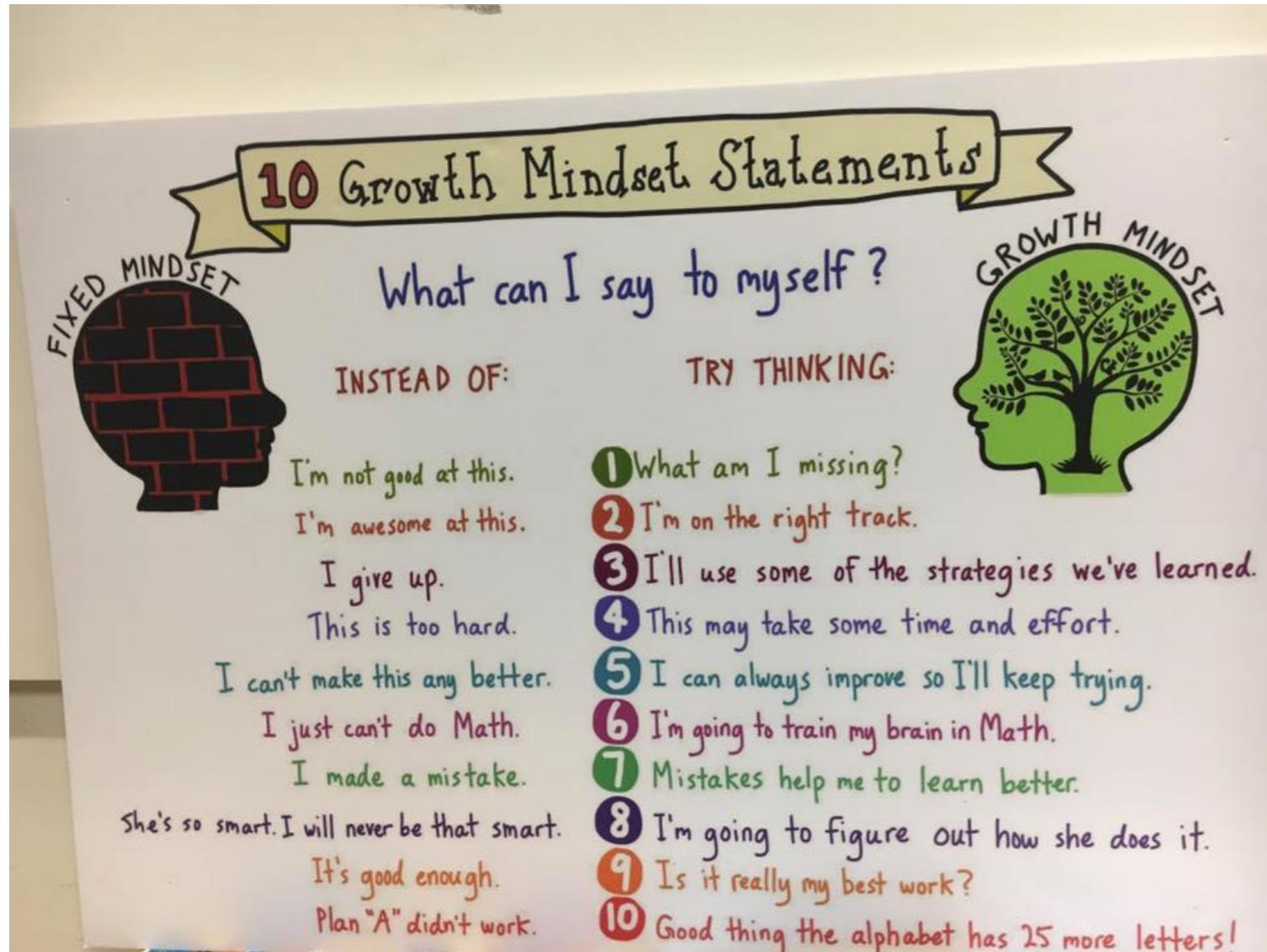


1789 SHARES

97 SHARES

What You Believe Affects What You Achieve

GROWTH VS FIXED MINDSET



Fixed Mindset –
Intelligence and talents are fixed at birth, capabilities derive from DNA and destiny

Growth Mindset –
Basic qualities including intelligence can be strengthened like muscles with practice and perseverance

COURSE LEARNING OBJECTIVES

- Understand the applications and business use cases for data science
- Acquire, parse, clean and apply various modeling techniques to make predictions from your data using Python
- Communicate findings to both a non-technical and technical audience in both written and verbal formats

COURSE OUTLINE

UNITS

UNIT 1: RESEARCH DESIGN AND EXPLORATORY DATA ANALYSIS

- › What is Data Science Lesson 1
 - › Research Design and Pandas Lesson 2
 - › Statistics Fundamentals I Lesson 3
 - › Statistics Fundamentals II Lesson 4
 - › Flexible Class Session Lesson 5
-

UNIT 2: FOUNDATIONS OF DATA MODELING

- › Introduction to Regression Lesson 6
 - › Evaluating Model Fit Lesson 7
 - › Introduction to Classification Lesson 8
 - › Introduction to Logistic Regression Lesson 9
 - › Communicating Logistic Regression Results Lesson 10
 - › Flexible Class Session Lesson 11
-

UNIT 3: DATA SCIENCE IN THE REAL WORLD

- › Decision Trees and Random Forests Lesson 12
- › Natural Language Processing Lesson 13
- › Dimensionality Reduction Lesson 14
- › Time Series Data I Lesson 15
- › Time Series Data II Lesson 16
- › Database Technologies Lesson 17
- › Where to Go Next Lesson 18
- › Flexible Class Session Lesson 19
- › Final Project Presentations Lesson 20

COURSE PROJECTS

Research Design and EDA		Foundation of Data Modeling				Data Science in the Real World					
1	2	3	4	5	6	7	8	9	10		
		PROJECT 1	PROJECT 2		PROJECT 3	PROJECT 4					
				FINAL PROJECT DELIVERABLE 1			FINAL PROJECT DELIVERABLE 2	FINAL PROJECT DELIVERABLE 3	FINAL PROJECT DELIVERABLE 4	FINAL PROJECT DELIVERABLE 5	

WHAT IS DATA SCIENCE

Tan Kwan Chong

Chief Data Scientist, Booz Allen Hamilton

WELCOME TO DATA SCIENCE

LEARNING OBJECTIVES

- Define data science and the data science workflow
- Setup your development environment and review command line, git, and python basics

INTRODUCTION

WHAT IS DATA SCIENCE?

BACKGROUND AND CONTEXT

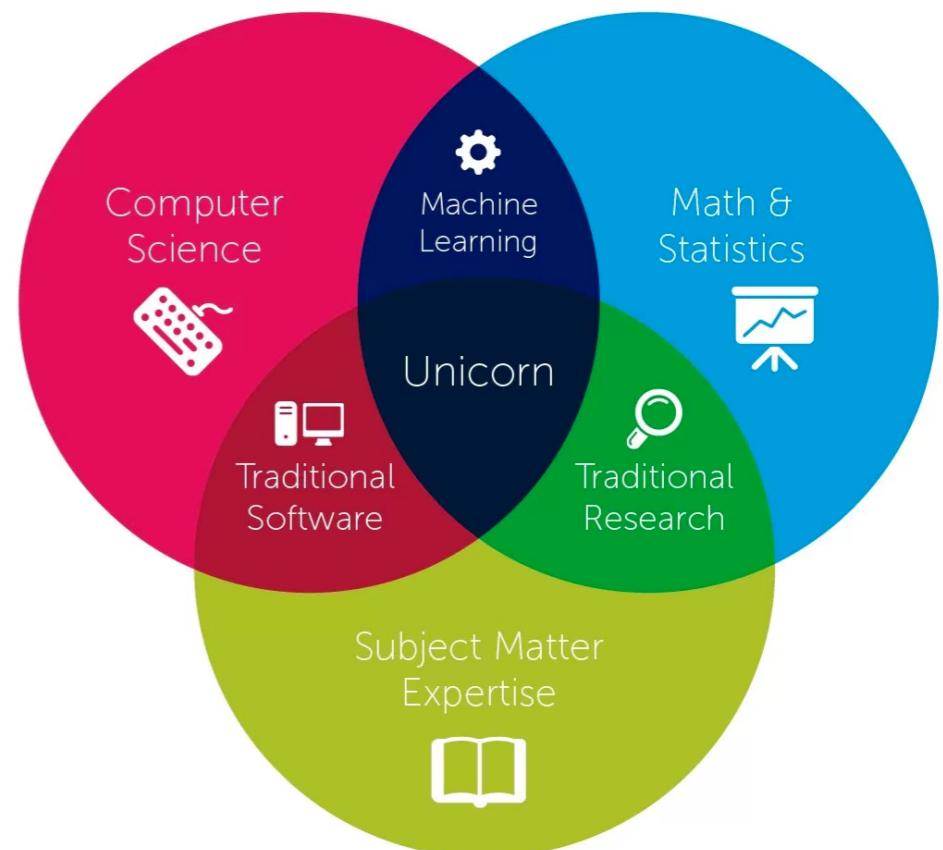
- 1 Data is increasingly cheap and ubiquitous
- 2 New technologies are emerging to organize and make sense of this avalanche of data
- 3 Organizations need to harness internal and external data effectively to attract and retain their customers



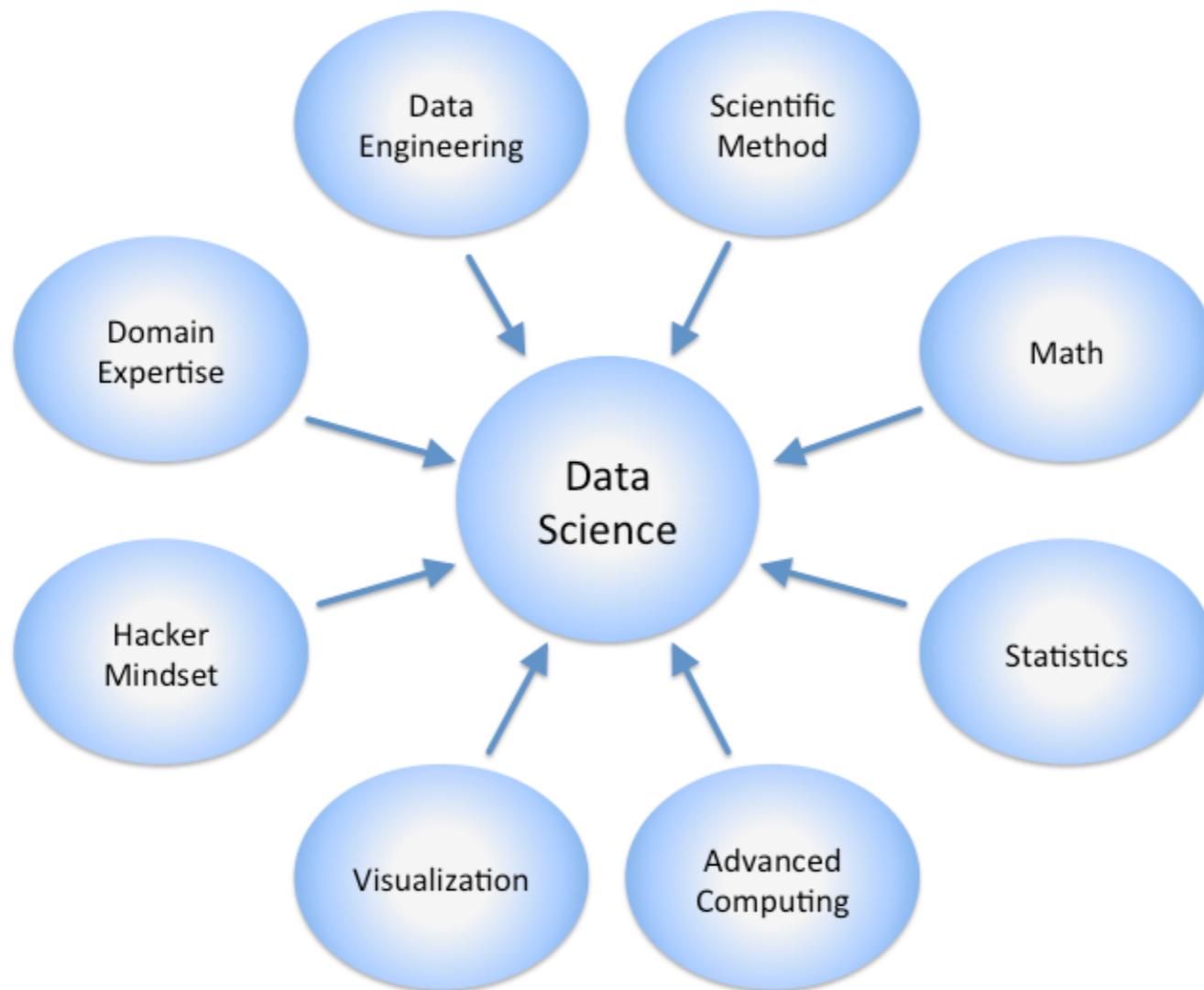
WHAT IS DATA SCIENCE?

- A set of tools and techniques for data
- Interdisciplinary problem-solving
- Application of scientific techniques to practical problems

Data Science



WHAT IS DATA SCIENCE?



The scientific approach to knowledge extraction from data

WHO USES DATA SCIENCE?

NETFLIX

amazon.com®



Google

Grab

 **FiveThirtyEight**

 **DBS**



**GOVTECH
SINGAPORE**

LAZADA
Effortless Shopping

SINGAPORE EXAMPLES - IS THIS DATA SCIENCE?



datagovsg [Follow](#)

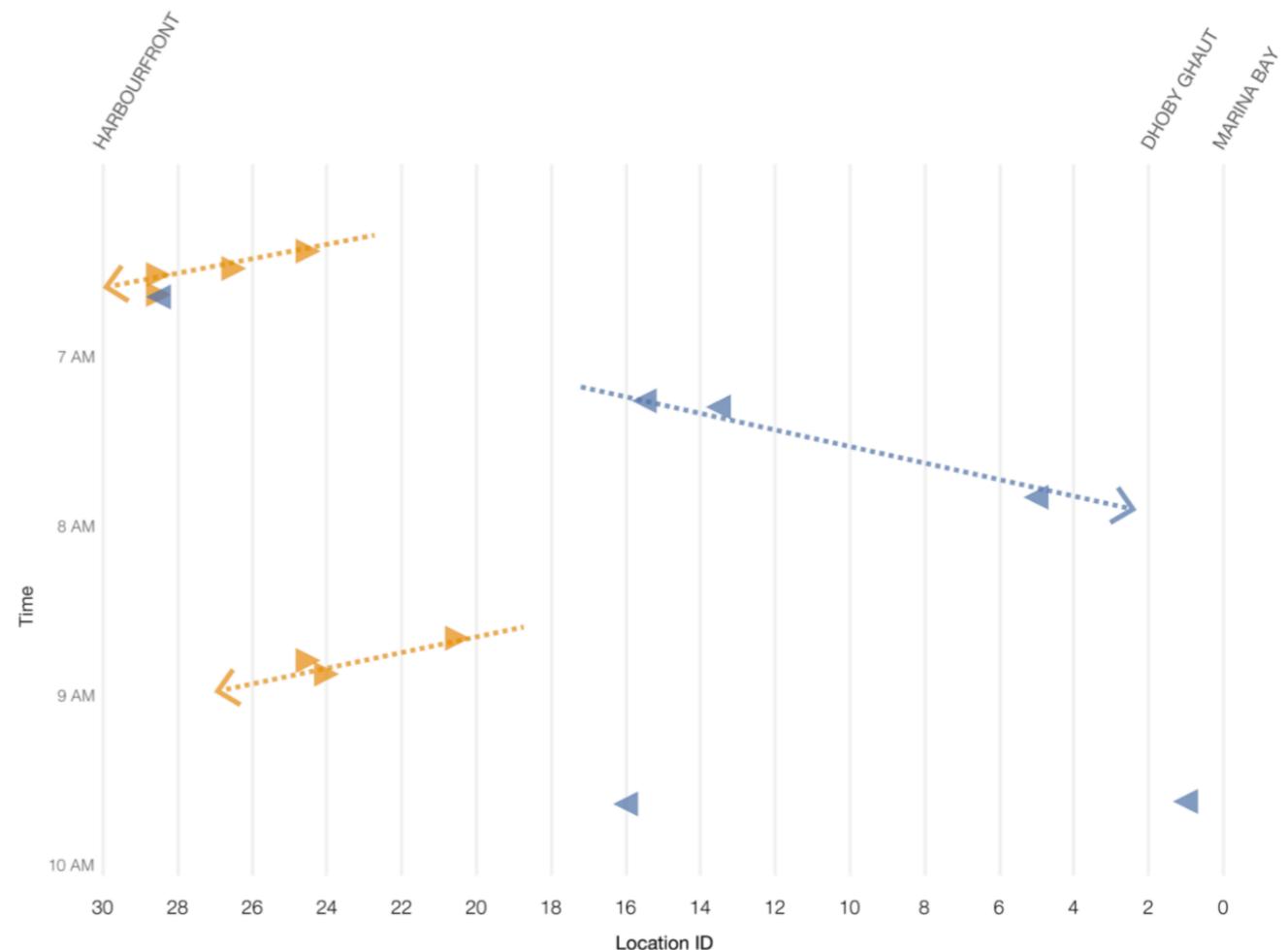
Official Medium account for <https://data.gov.sg>, Singapore's open data portal.

Dec 1, 2016 · 8 min read

How the Circle Line rogue train was caught with data

Text: Daniel Sim | Analysis: Lee Shangqian, Daniel Sim & Clarence Ng

Singapore's MRT Circle Line was hit by a spate of mysterious disruptions in recent months, causing much confusion and distress to thousands of commuters.



<https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a>

SINGAPORE EXAMPLES - IS THIS DATA SCIENCE?

Yellow taxis have fewer accidents than blue ones, says study



Researchers from NUS and the Chinese University of Hong Kong say that yellow taxis are more noticeable than blue taxis in both daylight and under street lighting. PHOTO: BLOOMBERG NEWS

PUBLISHED MAR 8, 2017, 5:00 AM SGT | UPDATED MAR 8, 2017, 11:01 AM



Tay Hong Yi

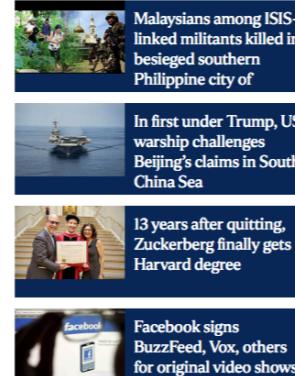
The next time you hail a taxi, take note of its colour, for it might mean a safer ride.

Researchers at the National University of Singapore (NUS) found that taxis painted yellow, a colour that stands out, were involved in significantly fewer traffic accidents than taxis painted blue.

Their results were based on analysing three years' worth of detailed taxi, driver and accident data from a large fleet of over 4,000 yellow taxis and 12,500 blue taxis locally.



ST VIDEOS ▶



Recommended by @utbrain

SPONSORED CONTENT

Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue

Teck-Hua Ho^{a,b,1}, Juin Kuan Chong^c, and Xiaoyu Xia^d

^aOffice of the Deputy President (Research & Technology), National University of Singapore, Singapore 119077; ^bHaas School of Business, University of California, Berkeley, CA 94720; ^cNational University of Singapore Business School, National University of Singapore, Singapore 119245; and ^dDepartment of Decision Sciences and Managerial Economics, Chinese University of Hong Kong Business School, Chinese University of Hong Kong, Shatin, NT, Hong Kong

Edited by George A. Akerlof, University of California, Berkeley, CA, and approved January 31, 2017 (received for review August 3, 2016)

Is there a link between the color of a taxi and how many accidents it has? An analysis of 36 mo of detailed taxi, driver, and accident data (comprising millions of data points) from the largest taxi company in Singapore suggests that there is an explicit link. Yellow taxis had 6.1 fewer accidents per 1,000 taxis per month than blue taxis, a 9% reduction in accident probability. We rule out driver difference as an explanatory variable and empirically show that because yellow taxis are more noticeable than blue taxis—especially when in front of another vehicle, and in street lighting—other drivers can better avoid hitting them, directly reducing the accident rate. This finding can play a significant role when choosing colors for public transportation and may save lives as well as millions of dollars.

car color | road safety | data science | transportation science | sensory perception

Accidents involving public transport are common and cause significant economic losses as well as loss of human life. Applying statistical analysis to a unique and comprehensive dataset we establish that a change in color can avert a significant number of taxi accidents, leading to a reduction in economic losses. Specifically, analysis of a complete set of accident records from the largest taxi operator in Singapore, which uses yellow and blue taxis, shows that yellow is safer than blue because yellow is more noticeable, with the result that potential accidents are avoided by other drivers' timely responses.

Yellow has been a popular color for taxis since 1907, when the Chicago Yellow Cab Company chose the color based on a survey conducted at the University of Chicago. The survey showed that yellow was the most noticeable color, which would make it easy for potential passengers to spot a yellow taxi in the sea of mass-

demographic characteristics. These two datasets include millions of observations on the company's drivers and taxis, and accidents involving these taxis. The data from both datasets have been anonymized and are available in Datasets S1–S6.

The company uses yellow or blue for all its regular taxis (approximate colors are shown in Fig. 1).[†] The colors are the remnants of a 2002 merger that took place between two taxi companies, one of which used yellow and the other, blue. The company owns ~16,700 taxis in a ratio of one yellow to three blue (1y:3b), which translates to 4,175 yellow taxis and 12,525 blue ones. These account for 60% of the ~27,800 taxis in Singapore.[‡]

To control for the difference in the number of taxis used by the company (1y:3b), we calculated a normalized accident rate using the average number of accidents that occurred per 1,000 taxis

Significance

This paper examines the phenomenon that yellow taxis have fewer accidents than blue taxis. Statistical analysis of a unique and comprehensive dataset suggests that the higher visibility of the color yellow makes it easier for other drivers to avoid getting into accidents with yellow taxis, leading to a lower accident rate. This suggests that color visibility should play a major role in determining the colors used for public transport vehicles.

Author contributions: T.-H.H. and J.K.C. designed research; T.-H.H. and J.K.C. performed research; T.-H.H., J.K.C., and X.X. contributed new reagents/analytic tools; T.-H.H., J.K.C., and X.X. analyzed data; and T.-H.H., J.K.C., and X.X. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

SINGAPORE EXAMPLES - IS THIS DATA SCIENCE?

How Lazada ranks products

to improve customer experience and conversion

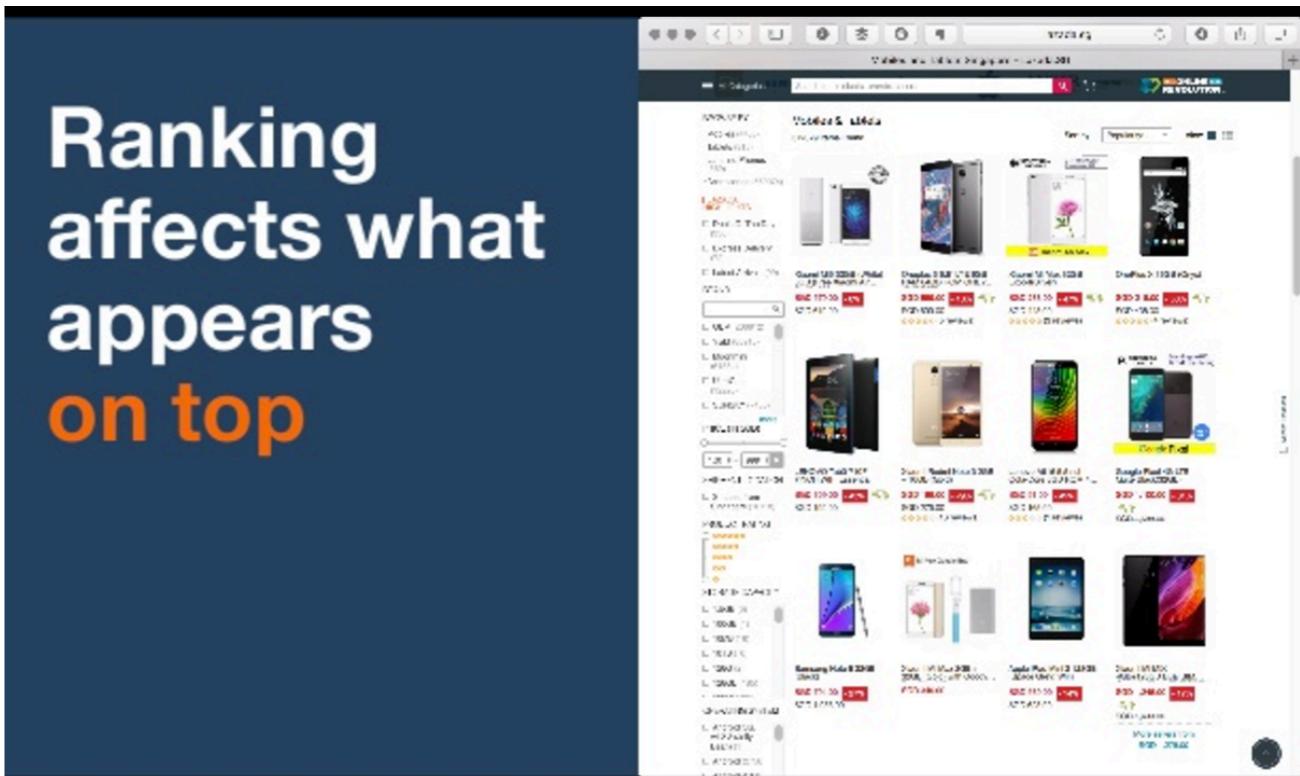
Strata Hadoop Singapore 2016

Overall results

Better ranking improved conversion and revenue per session

Introducing new products improved new product engagement

Emphasizing product quality had neutral to positive outcomes



COMMON QUESTIONS ASKED IN DATA SCIENCE

How much? How many?

- What will the stock price of Microsoft be next week?
- What will my fourth quarter sales be?
- How many kilowatts of electricity will be demanded between 7-8pm?
- How many new Twitter followers will I get next week?

Regression

- Predict a continuous outcome
- K-Nearest Neighbors
 - Linear Regression
 - Regression Trees

COMMON QUESTIONS ASKED IN DATA SCIENCE

Is this A, B or C?

- Is this a fraudulent transaction?
- Is this an image of a man, a cat, or a dog?
- Will this customer click on the advertisement?
- Is the sentiment of the review positive or negative?

Classification

- Predict a discrete outcome
- K-Nearest Neighbors
 - Logistic Regression
 - Classification Trees

COMMON QUESTIONS ASKED IN DATA SCIENCE

How is this Data Organized?

- What are the different types of coffee drinkers?
- Which viewers like the same kind of movies?
- Are there common clusters of cable channels that customers tend to purchase together?
- What is a natural way to break these documents into five topics?

Clustering

What are the “categories” within the data?

WHAT IS A DATA SCIENTIST?



Michael E. Driscoll

@medriscoll

Follow

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu

[@peteskomoroch](#)

4:57 AM - 18 Jul 2012



Data Science and Moneyball: A Profile of Pet...

Pete Skomoroch is a Principal Data Scientist at LinkedIn, where he leads a team that builds features like LinkedIn Skills. He was previously the [metamarkets.com](#)

47 45

i



Scott Vokes

@silentbicycle

Follow

"What is a 'Data Scientist'? An analyst who lives in California." -
[@edmundjackson](#) [#clojure_conj](#)

10:00 PM - 16 Nov 2012

20 9

i

“A Data Scientist should have a wide breadth of abilities: **academic curiosity, storytelling, product sense, engineering experience** and just a catch-all I call **cleverness**. But he or she should also have deep domain expertise in **Statistical and Machine Learning Knowledge**” – Thomson Nguyen

WHAT ARE THE ROLES IN DATA SCIENCE?

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning



HIRED BY
Google Microsoft Adobe

DATA SCIENTIST 'AS RARE AS UNICORNS'

Role

Cleans, massages and organizes (big) data

Mindset

Curious data wizard

DATA ANALYST 'DATA DETECTIVE'

Role

Collects, processes and performs statistical data analyses

Mindset

Intuitive data junkie with high "figure-it-out" quotient



HIRED BY
IBM hp DHL

Languages
R, Python, HTML, Javascript, C/C++, SQL

- ✓ Spreadsheet tools (e.g. Excel)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Communication & visualization
- ✓ Math, Stats, Machine Learning

Languages
SQL, XML, Hive, Pig, Spark

- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development



HIRED BY
VISA Coca-Cola logitech

DATA ARCHITECT 'THE CONTEMPORARY DATA MODELLER'

Role:

Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:

Inquiring ninja with a love for data architecture design patterns

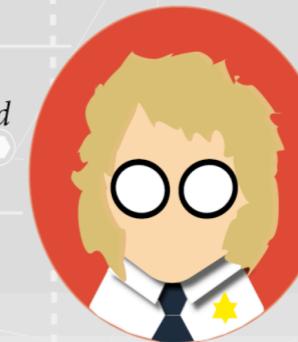
DATA AND ANALYTICS MANAGER 'DATA SCIENCE TEAM LEADER'

Role

Manages a team of analysts and data scientists

Mindset

Data Wizards' Cheerleader



HIRED BY
coursera slack **MOTOROLA** SOLUTIONS

Languages
SQL, R, SAS, Python, Matlab, Java

- ✓ Database systems (SQL and NO SQL based)
- ✓ Leadership & project management
- ✓ Interpersonal communication
- ✓ Data mining & predictive modeling

WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of roles, not just one.

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepreneur

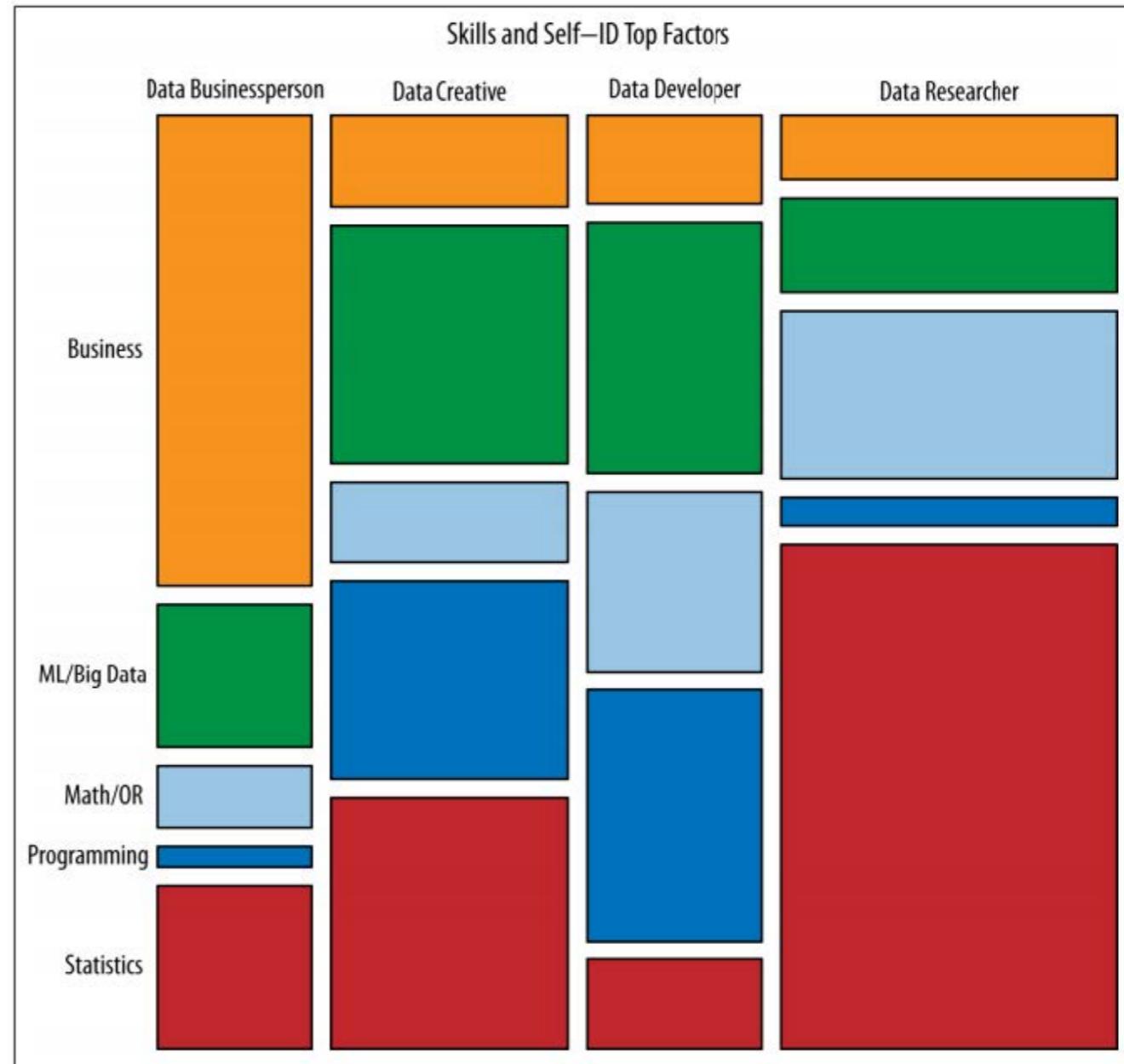
WHAT ARE THE ROLES IN DATA SCIENCE?

- Data Science involves a variety of skill sets, not just one.

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Development Business	Unstructured Data Structured Data Machine Learning Big and Distributed Data	Optimization Math Graphical Models Bayesian / Monte Carlo Statistics Algorithms Simulation	Systems Administration Back End Programming Front End Programming	Visualization Temporal Statistics Surveys and Marketing Spatial Statistics Science Data Manipulation Classical Statistics

WHAT ARE THE ROLES IN DATA SCIENCE?

- These roles prioritize different skill sets.
- However, all roles involve some part of each skillset.
- Where are your strengths and weaknesses?



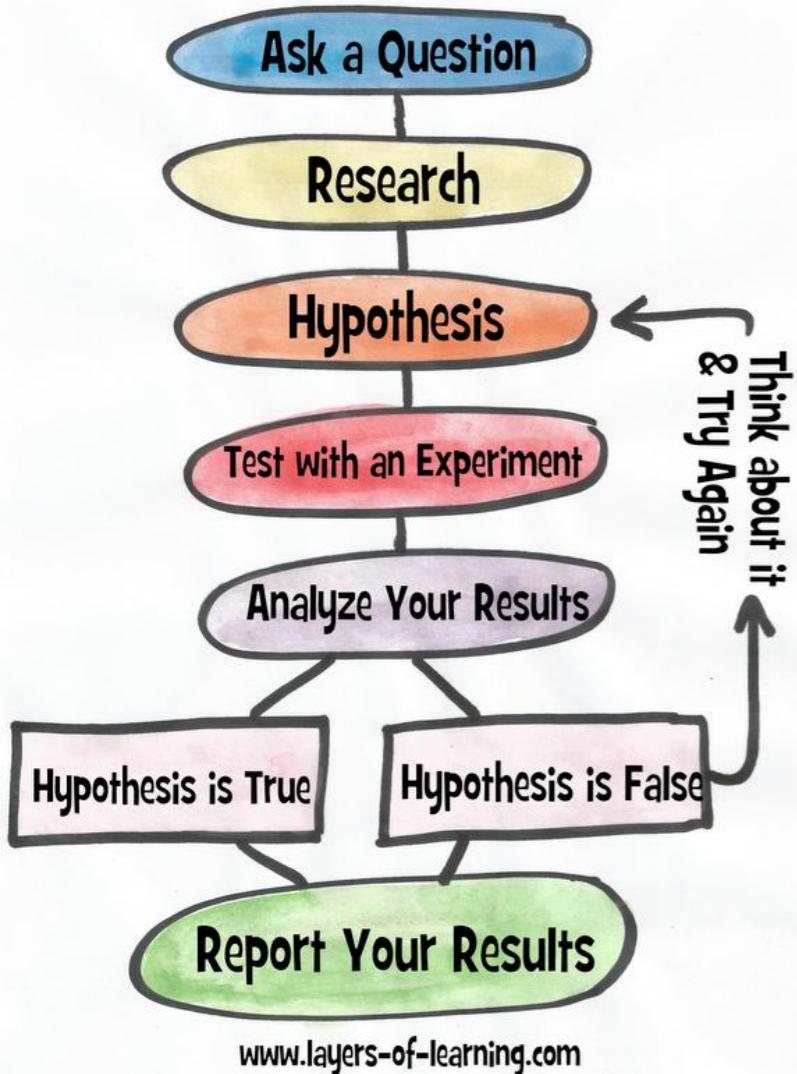
INTRODUCTION

THE DATA SCIENCE WORKFLOW

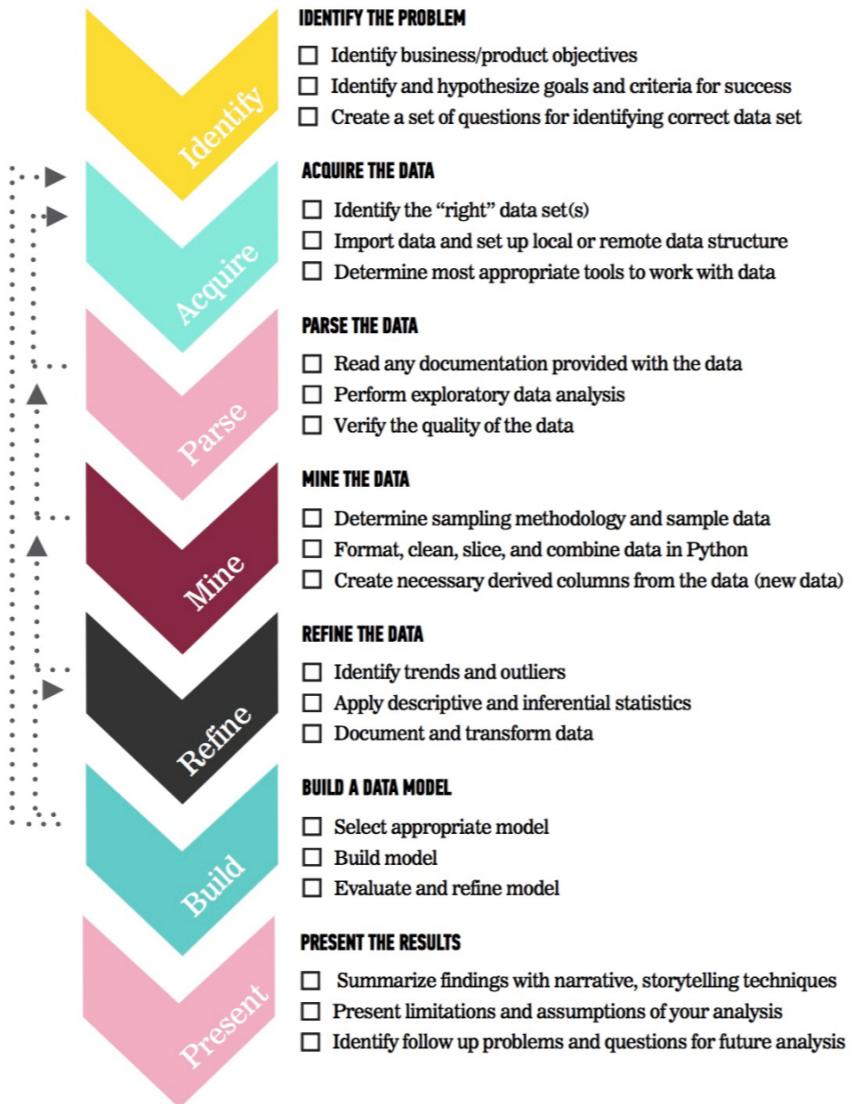
OVERVIEW OF THE DATA SCIENCE WORKFLOW

- A methodology for doing Data Science
- Similar to the scientific method
- Helps produce *reliable* and *reproducible* results
 - *Reliable*: Accurate findings
 - *Reproducible*: Others can follow your steps and get the same results

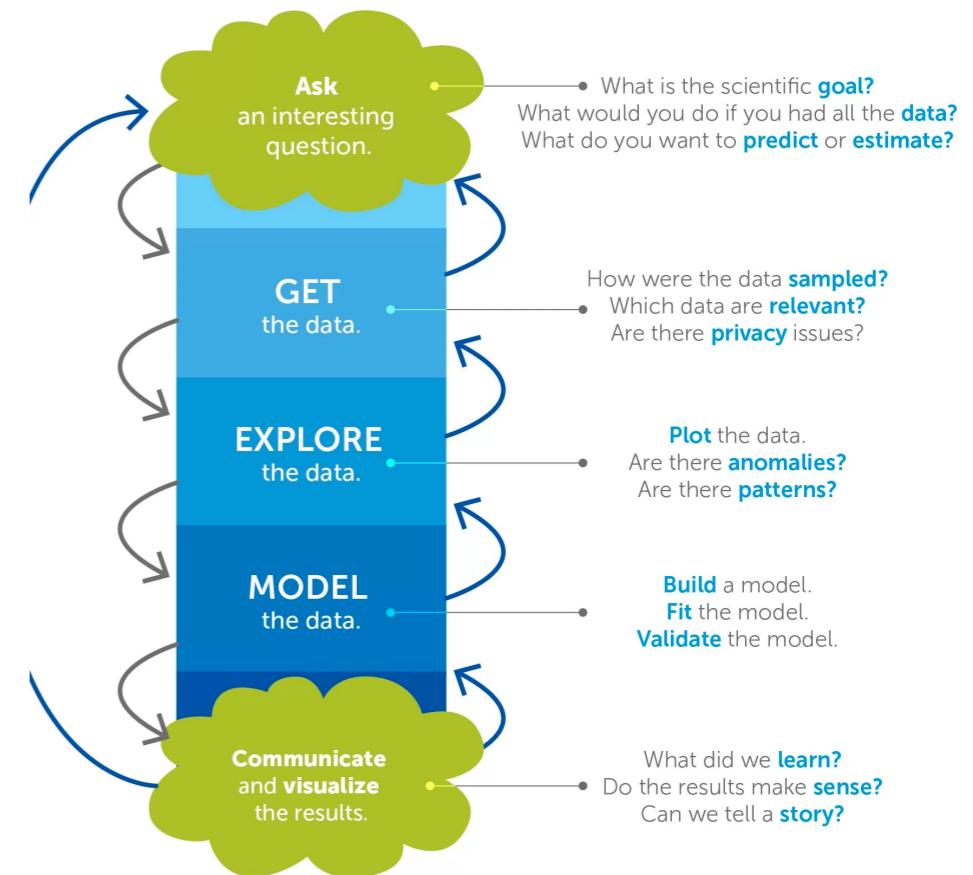
OVERVIEW OF THE DATA SCIENCE WORKFLOW



DATA SCIENCE WORKFLOW

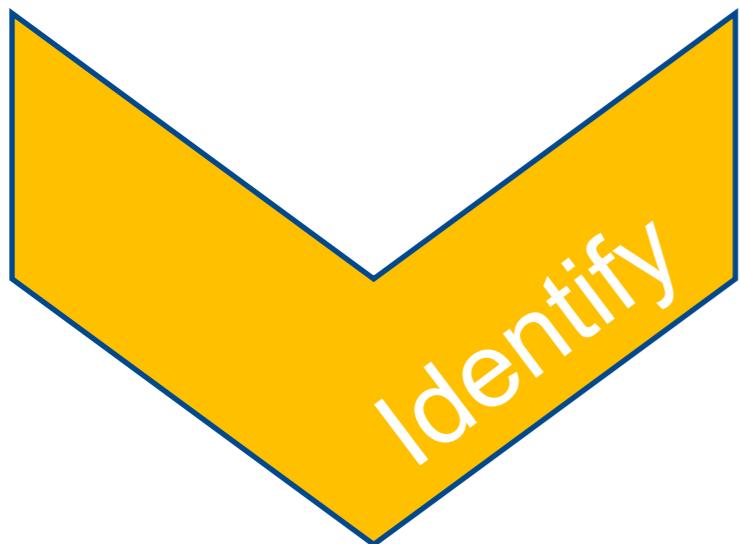


The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course <http://cs109.org/>.

IDENTIFY THE PROBLEM



IDENTIFY THE PROBLEM

- Identify business / product objectives
- Identify and hypothesize goals and criteria for success
- Create a set of questions for identifying the correct data set

IDENTIFY THE PROBLEM

- Begin by identifying the objectives and desired outcomes of the analysis
- Typical scenarios:
 - Given dataset(s) and tasked to explore and uncover insights
 - Specific business problem and need to source the relevant data

Example Problem Statement: What will the expected resale value of my HDB flat be in 10 years time?



ACQUIRE THE DATA

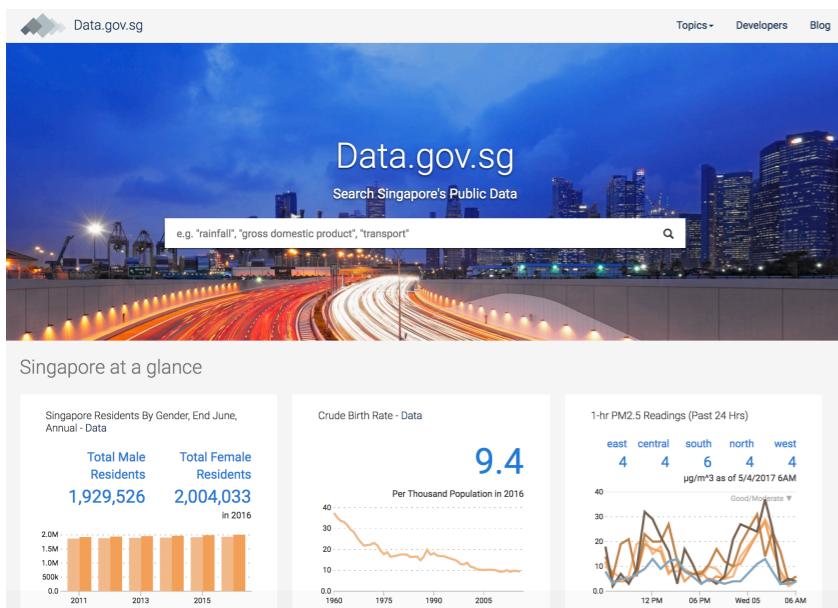


ACQUIRE THE DATA

- Identify the “right” data set(s)
- Import data and set up local or remote data structure
- Determine most appropriate tools to work with data

ACQUIRE THE DATA

- Internal data – spreadsheets, csv files, databases, system logs
- External data – csv files, API requests, web scraping



The screenshot shows the Welcome to Kaggle Datasets page. It features three main sections: "Discover" (with a magnifying glass icon), "Explore" (with a brain and test tubes icon), and "Create a Dataset" (with a signal icon). Below these are "Learn More" and "New Dataset" buttons. A dismissible notification at the bottom says "510 featured datasets".

The screenshot shows the Facebook for Developers Graph API documentation page. The left sidebar includes links for "All Docs", "Graph API", "Overview", "Using the Graph API", "Reference", "Common Scenarios", "Other APIs", "Webhooks", "Advanced", and "Changelog". The main content area is titled "The Graph API" and discusses the primary way for apps to read and write to the Facebook social graph.

ACQUIRE THE DATA

Data.gov.sg

Topics ▾ Developers Blog

Resale Flat Prices (Based on Registration Date), From March 2012 Onwards

Display 10 records Search: Filter

Month	Town	Flat Type	Block	Street Name	Storey Range	Floor Area (Sqm)	Flat Model	Lease Commence Date	Resale Price (\$\$)
2017-03	WOODLANDS	5 ROOM	742	WOODLANDS CIRCLE	10 TO 12	122	Improved	1997	428,000
2017-03	WOODLANDS	5 ROOM	749	WOODLANDS CIRCLE	07 TO 09	122	Improved	1998	398,888
2017-03	WOODLANDS	5 ROOM	787C	WOODLANDS CRES	01 TO 03	123	Improved	1997	380,000
2017-03	WOODLANDS	5 ROOM	502A	WOODLANDS DR 14	07 TO 09	123	Improved	1998	443,000
2017-03	WOODLANDS	5 ROOM	526	WOODLANDS DR 14	10 TO 12	126	Premium Apartment	2000	430,000
2017-03	WOODLANDS	5 ROOM	503	WOODLANDS DR 14	01 TO 03	122	Improved	1998	400,000
2017-03	WOODLANDS	5 ROOM	504	WOODLANDS DR 14	10 TO 12	120	Improved	1999	412,000

Showing 1 to 10 of 192512 records (Only last 2000 records shown)

« 1 2 3 4 5 ... 200 »

Resale Flat Prices (Based on Registration Date), From March 2012 Onwards Embed View

Resale Flat Prices

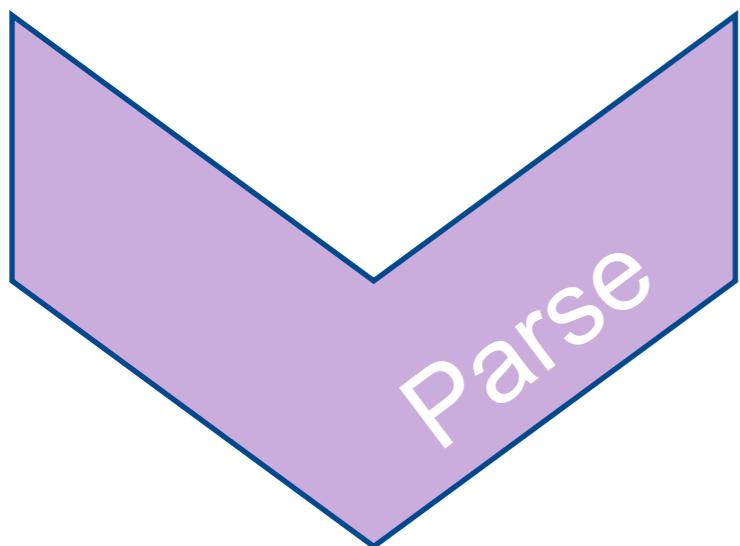
Managed by Housing and Development Board

Resale transacted prices. Prior to March 2012, data is based on date of approval for the resale transactions. For March 2012 onwards, the data is based on date of registration for the resale transactions.

Download

- Resale flat prices data available for download from Data.gov.sg website

PARSE THE DATA

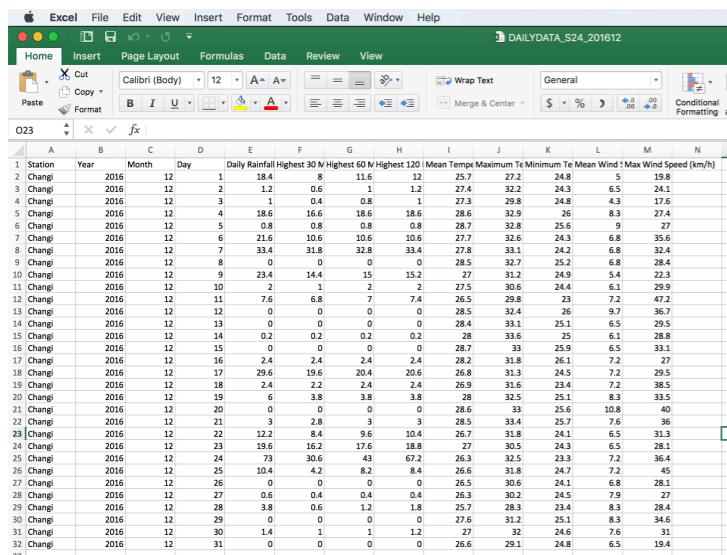


PARSE THE DATA

- Read any documentation provided with the data
- Perform exploratory data analysis
- Verify the quality of the data

PARSE THE DATA

- Structured data – well formatted, data schema defined, generally easy to parse
- Unstructured data – may require additional processing or programming tools to extract relevant metadata e.g. sentiments, word count, entities
- Semi-structured data – may require use of regex or programming logic to extract relevant fields e.g. logs, JSON, XML



Station	Year	Month	Day	Daily Rainfall	Highest 30 M	Highest 60 M	Highest 120 M	Mean Temp	Maximum Temp	Minimum Temp	Mean Wind	Max Wind Speed (km/h)
2 Changi	2016	12	1	18.4	8	11.6	12	25.7	24.8	5	19.8	
3 Changi	2016	12	2	1.2	0.6	1	1.2	27.4	32.2	24.3	6.5	24.1
4 Changi	2016	12	3	1	0.4	0.8	1	27.3	29.8	24.8	4.3	17.6
5 Changi	2016	12	4	18.6	16.6	18.6	18.6	28.6	32.9	26	8.3	27.4
6 Changi	2016	12	5	0.8	0.8	0.8	0.8	28.7	32.8	25.6	9	27
7 Changi	2016	12	6	21.6	10.6	10.6	10.6	27.7	32.6	24.3	6.8	35.6
8 Changi	2016	12	7	33.4	31.8	32.8	33.4	27.8	33.1	24.2	6.8	32.4
9 Changi	2016	12	8	0	0	0	0	28.5	32.7	25.3	6.8	28.4
10 Changi	2016	12	9	23.4	14.4	15	15.2	27	31.2	24.9	5.4	22.3
11 Changi	2016	12	10	2	3	2	2	27.5	30.6	24.4	6.1	29.9
12 Changi	2016	12	11	7.6	6.8	7	7.4	26.5	29.8	23	7.2	47.2
13 Changi	2016	12	12	0	0	0	0	28.5	32.4	26	9.7	36.7
14 Changi	2016	12	13	0	0	0	0	28.4	33.1	25.1	6.5	29.5
15 Changi	2016	12	14	0.2	0.2	0.2	0.2	28	33.6	25	6.1	28.8
16 Changi	2016	12	15	0	0	0	0	28.7	33	25.9	6.5	33.1
17 Changi	2016	12	16	2.4	2.4	2.4	2.4	28.2	31.8	26.1	7.2	27
18 Changi	2016	12	17	29.6	19.6	20.4	20.6	26.8	31.3	24.5	7.2	29.5
19 Changi	2016	12	18	2.4	2.2	2.4	2.4	26.9	31.6	23.4	7.2	38.5
20 Changi	2016	12	19	6	3.8	3.8	3.8	28	32.5	25.1	8.3	33.5
21 Changi	2016	12	20	0	0	0	0	28.6	33	25.6	10.8	40
22 Changi	2016	12	21	3	2.8	3	3	28.5	33.4	25.7	7.6	36
23 Changi	2016	12	22	12.2	8.4	9.6	10.4	28.7	31.8	24.1	6.5	31.3
24 Changi	2016	12	23	19.6	15.2	17.6	18.8	27	30.5	24.3	6.5	28.1
25 Changi	2016	12	24	73	30.6	43	67.2	26.3	32.5	23.3	7.2	36
26 Changi	2016	12	25	10.4	4.2	8.2	8.4	26.6	31.8	24.7	7.2	45
27 Changi	2016	12	26	0	0	0	0	26.5	30.6	24.1	6.8	28.1
28 Changi	2016	12	27	0.6	0.4	0.4	0.4	26.3	30.2	24.5	7.9	27
29 Changi	2016	12	28	3.8	0.6	1.2	1.8	25.7	28.3	23.4	8.3	28.4
30 Changi	2016	12	29	0	0	0	0	27.6	31.2	25.1	8.3	34.6
31 Changi	2016	12	30	1.4	5	1	1.2	27	32	24.6	7.6	31
32 Changi	2016	12	31	0	0	0	0	26.6	29.1	24.8	6.5	19.4

Structured



ORAL ANSWERS TO QUESTIONS
FEATURES FOR MERGED SKILLSFUTURE AND JOBS BANK PORTALS

1 Ms Sun Xueling asked the Minister for Education (Higher Education and Skills) (a) whether the merger of the SkillsFuture and Jobs Bank portals will incorporate a review function for training providers to ensure outcome-based training provisions; (b) whether the merged portal can include internship opportunities for young Singaporeans; (c) whether the merged portal can include personality assessments that gauge the fit of the user to jobs as opposed to just skills; and (d) whether jobs listed in the Jobs Bank are regularly updated to ensure relevance.

The Parliamentary Secretary to the Ministers for Education (Assoc Prof Dr Muhammad Faishal Ibrahim) (for the Minister for Education (Higher Education and Skills)) : Madam, the Individual Learning Portfolio (ILP) is designed to be a one-stop online portal which empowers individuals to make informed learning and career choices. There will be platforms for individuals to review and provide feedback on the training programmes available on the ILP, and information on training outcomes will be published on the portal.

```
893252015.307 14 <client-ip> TCP_HIT/200 227 GET
http://images.go2net.com/metacrawler/images/transparent.gif - NONE/- image/gif
893252015.312 23 <client-ip> TCP_HIT/200 4170 GET
http://images.go2net.com/metacrawler/images/head.gif - NONE/- image/gif
893252015.318 38 <client-ip> TCP_HIT/200 406 GET
http://images.go2net.com/metacrawler/images/bg2.gif - NONE/- image/gif
893252015.636 800 <client-ip> TCP_REFRESH_MISS/200 8872 GET
http://www.metacrawler.com/ - DIRECT/www.metacrawler.com text/html
893252015.728 355 <client-ip> TCP_HIT/200 5691 GET
http://images.go2net.com/metacrawler/images/market2.gif - NONE/- image/gif
893252016.138 465 <client-ip> TCP_HIT/200 219 GET
http://images.go2net.com/metacrawler/templates/tips/.../images/pixel.gif -
NONE/- image/gif
893252016.430 757 <client-ip> TCP_REFRESH_HIT/200 2106 GET
http://images.go2net.com/metacrawler/templates/tips/.../images/ultimate.jpg -
DIRECT/images.go2net.com image/jpeg
```

```
1 - {
  "odata.metadata": "http://datamall2.mytransport.sg/ltaodataservice/$metadata#CarParkAvailability",
  "value": [
    {
      "CarParkID": 1,
      "Area": "Marina",
      "Development": "Suntec City",
      "Latitude": 1.39375,
      "Longitude": 103.85718,
      "Lots": 1460
    },
    {
      "CarParkID": 2,
      "Area": "Marina",
      "Development": "Marina Square",
      "Latitude": 1.39115,
      "Longitude": 103.85728,
      "Lots": 1789
    },
    {
      "CarParkID": 4,
      "Area": "Marina",
      "Development": "The Esplanade",
      "Latitude": 1.39011,
      "Longitude": 103.85728
    }
  ]
}
```

Unstructured

Semi-structured

PARSE THE DATA

```
# Metadata for Resale Flat Prices
Identifier: '7a339d20-3c57-4b11-a695-9348adfd7614'
Name: 'resale-flat-prices'
Title: 'Resale Flat Prices'
Description:
- 'Resale transacted prices.'
- 'Prior to March 2012, data is based on date of approval for the resale transactions.'
- 'For March 2012 onwards, the data is based on date of registration for the resale transactions.'
Topics:
- 'Infrastructure'
Keywords:
- 'Cost of Living'
- 'HDB'
- 'Housing'
- 'Property'
- 'Public Housing'
- 'Resale Flats'
Publisher:
Name: 'Housing and Development Board'
Admin 1:
Name: 'Michelle Tay'
Department: 'CDG'
Email: 'Michelle_MB_TAY@hdb.gov.sg'
Sources:
- 'Housing and Development Board'
License: 'https://data.gov.sg/open-data-licence'
Frequency: 'Monthly'
Coverage: '1990-01-01 to 2017-04-30'
Last Updated: '2017-05-15T07:03:31.090642'
Resources:
- Identifier: '83b2fc37-ce8c-4df4-968b-370fd818138b'
Title: 'Resale Flat Prices (Based on Registration Date), From March 2012 Onwards'
Url: 'https://storage.data.gov.sg/resale-flat-prices/resources/resale-flat-prices-based-on-regis'
Format: 'CSV'
Coverage: '2012-03-01 to 2017-04-30'
Last Updated: '2017-05-15T07:03:30.565947'
Schema:
- Name: 'month'
Title: 'Month'
Type: 'datetime'
Sub Type: 'month'
Format: 'YYYY-MM'
```

```
In [156]: resale_prices.head()
Out[156]:
```

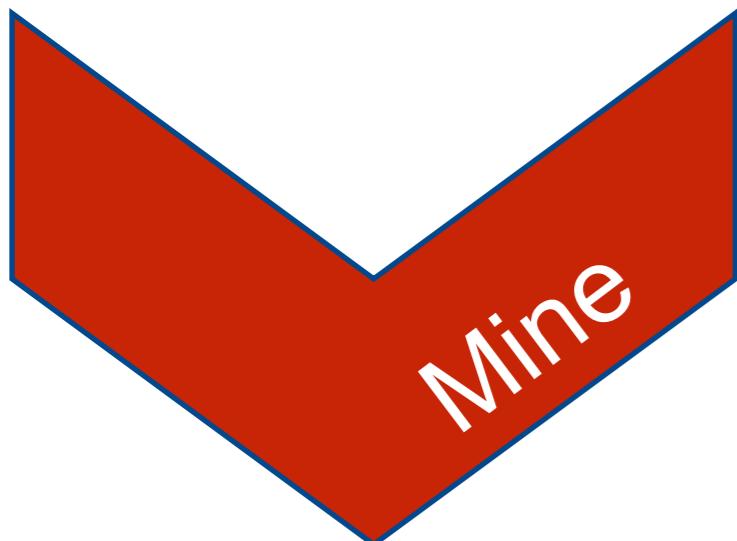
	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price
0	2012-03	ANG MO KIO	2 ROOM	172	ANG MO KIO AVE 4	06 TO 10	45.0	Improved	1986	250000.0
1	2012-03	ANG MO KIO	2 ROOM	510	ANG MO KIO AVE 8	01 TO 05	44.0	Improved	1980	265000.0
2	2012-03	ANG MO KIO	3 ROOM	610	ANG MO KIO AVE 4	06 TO 10	68.0	New Generation	1980	315000.0
3	2012-03	ANG MO KIO	3 ROOM	474	ANG MO KIO AVE 10	01 TO 05	67.0	New Generation	1984	320000.0
4	2012-03	ANG MO KIO	3 ROOM	604	ANG MO KIO AVE 5	06 TO 10	67.0	New Generation	1980	321000.0

```
In [169]: resale_prices.isnull().values.any()
Out[169]: False
```

```
In [160]: resale_prices.count()
Out[160]:
```

month	96631
town	96631
flat_type	96631
block	96631
street_name	96631
storey_range	96631
floor_area_sqm	96631
flat_model	96631
lease_commence_date	96631
resale_price	96631
dtype:	int64

MINE THE DATA



MINE THE DATA

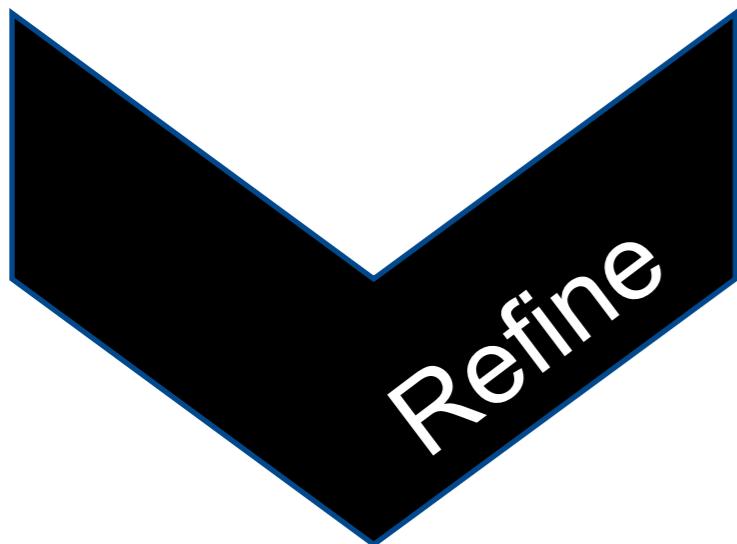
- Determine sampling methodology and sample data
- Format, clean, slice, and combine data in Python
- Create necessary derived columns from the data (new data)

MINE THE DATA

```
In [171]: resale_prices = resale_prices.rename(columns={'month': 'year-month'})  
  
In [181]: resale_prices['year'] = resale_prices['year-month'].apply(lambda x: int(x.split("-")[0]))  
  
In [182]: resale_prices['month'] = resale_prices['year-month'].apply(lambda x: int(x.split("-")[1]))  
  
In [174]: resale_prices['lower_storey_range'] = resale_prices['storey_range'].apply(lambda x: int(x.split()[0]))  
  
In [175]: resale_prices['upper_storey_range'] = resale_prices['storey_range'].apply(lambda x: int(x.split()[2]))  
  
In [184]: resale_prices['flat_age'] = resale_prices['year'] - resale_prices['lease_commence_date']  
  
In [185]: resale_prices.head()
```

Out[185]:	id	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price	year	month	lower_storey_range	upper_storey_range	flat_age
	06 TO 10	45.0	Improved	1986	250000.0	2012	3	6	6	10	26
	01 TO 05	44.0	Improved	1980	265000.0	2012	3	1	5	5	32
	06 TO 10	68.0	New Generation	1980	315000.0	2012	3	6	6	10	32
	01 TO 05	67.0	New Generation	1984	320000.0	2012	3	1	5	5	28

REFINE THE DATA



REFINE THE DATA

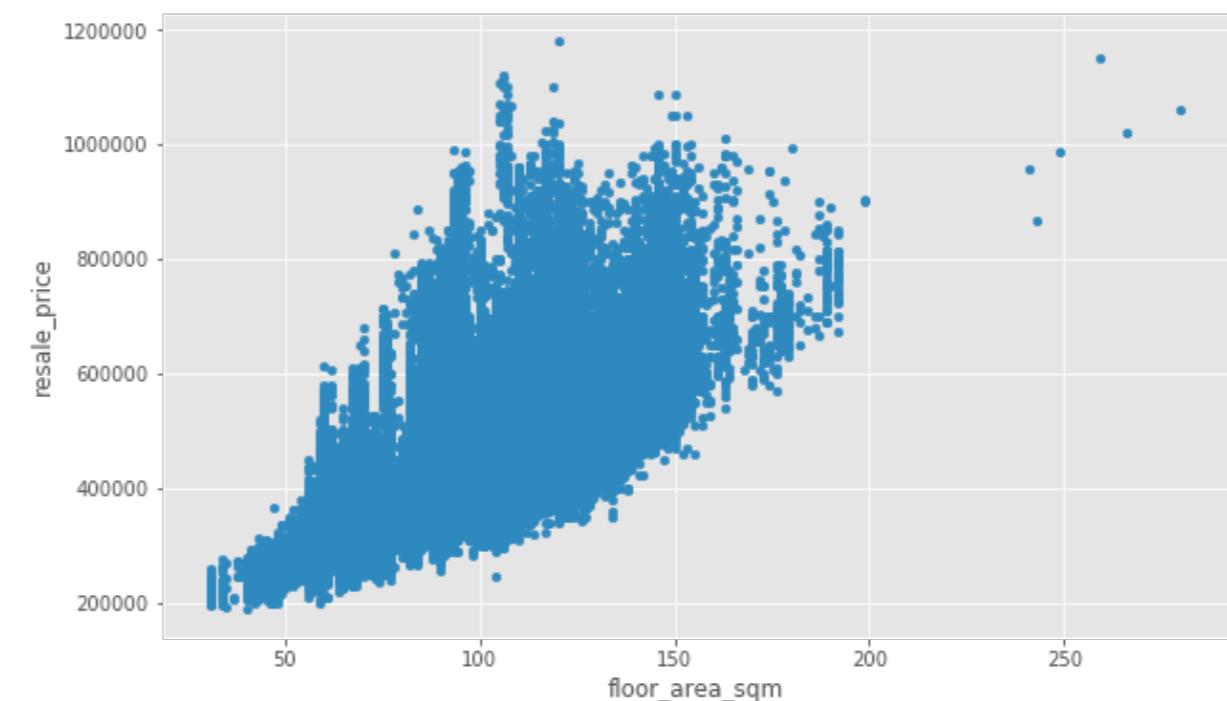
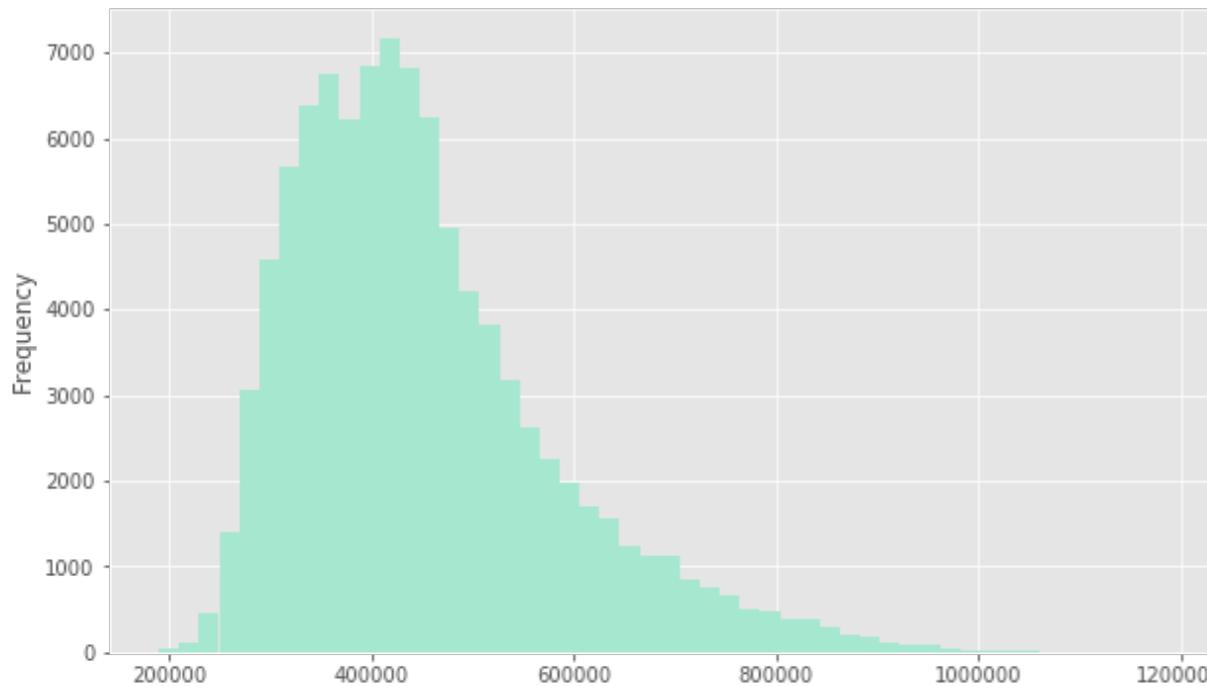
- Identify trends and outliers
- Apply descriptive and inferential statistics
- Document and transform data

REFINE THE DATA

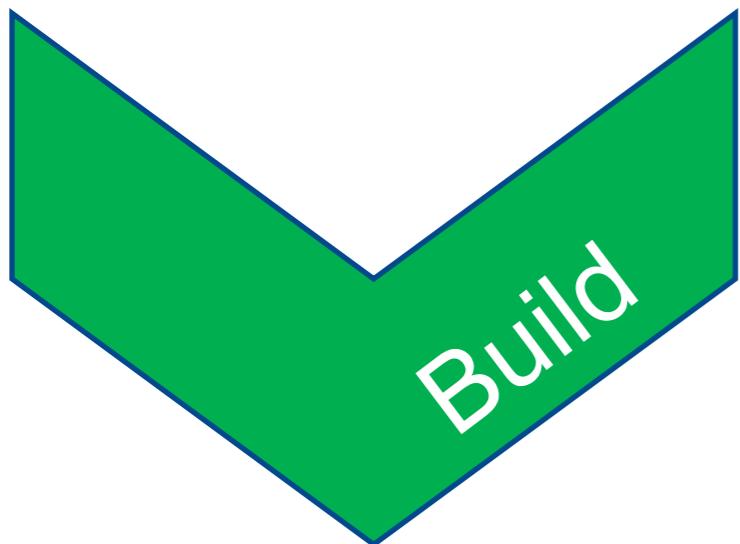
```
In [186]: resale_prices.describe()
```

Out[186]:

	floor_area_sqm	lease_commence_date	resale_price	year	month	lower_storey_range	upper_storey_range	flat_age
count	96631.000000	96631.000000	9.663100e+04	96631.000000	96631.000000	96631.000000	96631.000000	96631.000000
mean	96.570929	1990.219039	4.502107e+05	2014.204003	6.382331	6.843156	8.984684	23.984963
std	24.615748	10.549772	1.298110e+05	1.585023	3.324086	5.149889	5.143670	10.554514
min	31.000000	1966.000000	1.900000e+05	2012.000000	1.000000	1.000000	3.000000	1.000000
25%	74.000000	1983.000000	3.550000e+05	2013.000000	4.000000	4.000000	6.000000	15.000000
50%	95.000000	1988.000000	4.250000e+05	2014.000000	6.000000	7.000000	9.000000	26.000000
75%	111.000000	1999.000000	5.150000e+05	2016.000000	9.000000	10.000000	12.000000	32.000000
max	280.000000	2013.000000	1.180000e+06	2017.000000	12.000000	49.000000	51.000000	51.000000



BUILD A DATA MODEL



BUILD A DATA MODEL

- Select appropriate model
- Build model
- Evaluate and refine model

BUILD A DATA MODEL

- Predicting a continuous variable -> regression model
- Target variable -> resale_price
- Iteratively add / remove / modify input variables, validate errors

```
from sklearn import linear_model

reg = linear_model.LinearRegression()

reg.fit(resale_prices[["floor_area_sqm", "upper_storey_range", "flat_age"]], resale_prices["resale_price"])

LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)

reg.predict([[50,10,10], [67,5,28]])

array([ 305218.45070752,  315782.87716889])

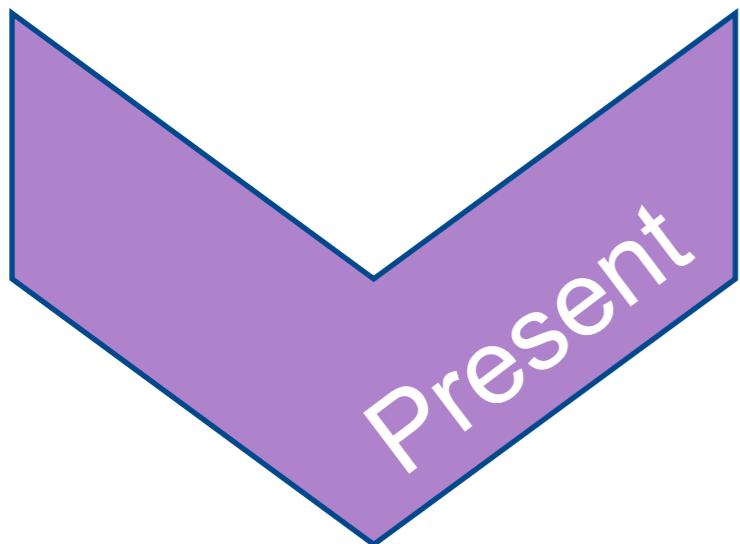
from sklearn.metrics import mean_squared_error

rmse = mean_squared_error(resale_prices["resale_price"],
                           reg.predict(resale_prices[["floor_area_sqm", "upper_storey_range", "flat_age"]]))**0.5

rmse

84658.318517472144
```

OVERVIEW OF THE DATA SCIENCE WORKFLOW



PRESENT THE RESULTS

- Summarize findings with narrative, storytelling techniques
- Present limitations and assumptions of your analysis
- Identify follow up problems and questions for future analysis

PRESENT THE RESULTS

- Key factors of a good presentation include:
 - Summarize findings with narrative and storytelling techniques
 - Refine your visualizations for broader comprehension
 - Present both limitations and assumptions
 - Determine the integrity of your analyses
 - Consider the degree of disclosure for various stakeholders
 - Test and evaluate the effectiveness of your presentation beforehand

GUIDED PRACTICE

DATA SCIENCE WORK FLOW

ACTIVITY: DATA SCIENCE WORKFLOW

DIRECTIONS (25 minutes)

EXERCISE

1. Divide into groups of 2-3 people
2. **IDENTIFY:** Each group should develop a data science problem statements on topics of personal or work interest. Create a hypothesis to your question. (5 minutes)
3. **ACQUIRE:** Discuss and list down the required data sources and their formats and fields required to address the problem statement. Do a search to see if the data is publicly available (10 minutes)
4. **PRESENT:** Present on your problem statements and approach to acquiring the data to the class (10 minutes)

DEMO

ENVIRONMENT SETUP

GETTING HELP

1. Google & Stack Overflow are your friend

sort pandas dataframe column value

All Videos Images News More Settings Tools

About 99,700 results (0.56 seconds)

pandas.DataFrame.sort_values — pandas 0.20.1 documentation

https://pandas.pydata.org/pandas-docs/stable/.../pandas.DataFrame.sort_values.html ▾
Sort by the values along either axis ... Specify list for multiple sort orders. ... For DataFrames, this option is only applied when sorting on a single column or label.

pandas.DataFrame.sort — pandas 0.18.1 documentation

<https://pandas.pydata.org/pandas-docs/version/0.18.1/.../pandas.DataFrame.sort.html> ▾
DataFrame.sort(columns=None, axis=0, ascending=True, inplace=False, ... Sort DataFrame either by labels (along either axis) or by the values in column(s) ...

How to sort pandas data frame using values from several columns?

<https://stackoverflow.com/.../how-to-sort-pandas-data-frame-using-values-from-sever...> ▾
Jul 12, 2013 - I have the following data frame: df = pandas.DataFrame([{'c1':3 ... Your code works for me. >>> import pandas >>> df = pandas.DataFrame([{'c1':3 ...

sorting - python, sort descending dataframe with pandas - Stack ...

<https://stackoverflow.com/questions/.../python-sort-descending-dataframe-with-pandas> ▾
Jul 28, 2014 - [False] , being a nonempty list, is not the same as False . You should write: ... Sort pandas DataFrame with function over column values.

python - Sort Pandas DataFrame by value - Stack Overflow

<https://stackoverflow.com/questions/37287938/sort-pandas-dataframe-by-value> ▾
May 17, 2016 - If I'm understanding you correctly, you're trying to sort that df by 'retweets'? use: ... I Know this question has a lot of answers, for example: How to sort pandas data frame using values from several columns? I tried the solutions ...

stackoverflow Questions Jobs Documentation BETA Tags Users Search... ? Log In Sign Up

python, sort descending dataframe with pandas Ask Question

I'm trying to sort a dataframe by descending. I put 'False' in the ascending argument, but my order is still ascending.
My code is:

```
from pandas import DataFrame
import pandas as pd

d = {'one':[2,3,1,4,5],
     'two':[5,4,3,2,1],
     'letter':['a','a','b','b','c']}

df = DataFrame(d)

test = df.sort(['one'], ascending=[False])
```

but the output is

	letter	one	two
2	b	1	3
0	a	2	5
1	a	3	4
3	b	4	2
4	c	5	1

python sorting pandas

share improve this question asked Jul 28 '14 at 5:25 user3636476 364 ● 1 ● 3 ● 14

1 Your code actually gives the desired results on pandas version 0.14.1, so you may want to upgrade if possible. – Marius Jul 28 '14 at 6:12

Jobs near you

Expert node.js Developer for Singapore Zuhlike Engineering Ltd Singapore javascript node.js

Work on SaaS products sold to the world's Fortune 500 Lucep Pte Ltd Singapore \$36K - \$60K javascript html5

Full-Stack Engineer (Web) Grab Singapore RELOCATION javascript reactjs

Software Engineer, Backend (Incentives) Grab Singapore RELOCATION python javascript

More jobs near Singapore...

GETTING HELP

2. Check the documentation

The screenshot shows a web browser displaying the pandas 0.20.1 documentation for the `pandas.DataFrame.sort_values` method. The URL in the address bar is https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_values.html. The page has a green header bar with the text "pandas 0.20.1 documentation » API Reference » pandas.DataFrame ». The main content area has a blue header "pandas.DataFrame.sort_values". Below it, the method signature is shown: `DataFrame.sort_values(by, axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')` with a "[source]" link. A note states "Sort by the values along either axis". A "New in version 0.17.0." badge is present. To the left is a "Table Of Contents" sidebar with many links to various pandas documentation pages. The right side contains detailed parameter descriptions for `by`, `axis`, `ascending`, `inplace`, `kind`, and `na_position`, along with a "Returns" section describing the `sorted_obj`.

`DataFrame.sort_values(by, axis=0, ascending=True, inplace=False, kind='quicksort', na_position='last')` [source]

Sort by the values along either axis

New in version 0.17.0.

Parameters:

- by : str or list of str**
Name or list of names which refer to the axis items.
- axis : {0 or 'index', 1 or 'columns'}, default 0**
Axis to direct sorting
- ascending : bool or list of bool, default True**
Sort ascending vs. descending. Specify list for multiple sort orders. If this is a list of bools, must match the length of the by.
- inplace : bool, default False**
if True, perform operation in-place
- kind : {'quicksort', 'mergesort', 'heapsort'}, default 'quicksort'**
Choice of sorting algorithm. See also `ndarray.np.sort` for more information. `mergesort` is the only stable algorithm. For DataFrames, this option is only applied when sorting on a single column or label.
- na_position : {'first', 'last'}, default 'last'**
`first` puts NaNs at the beginning, `last` puts NaNs at the end

Returns: `sorted_obj : DataFrame`

GETTING HELP

3. Review the error logs

1. Acquire the data

- Identify the "right" data set(s)
- Import data and set up local or remote data structure
- Determine most appropriate tools to work with data

Data downloaded from <https://data.gov.sg/dataset/resale-flat-prices>

```
: resale_prices = pd.read_cs('data/resale-flat-prices-based-on-registration-date-from-march-2012-onwards.csv')

-----
AttributeError                               Traceback (most recent call last)
<ipython-input-19-de8346ce4883> in <module>()
----> 1 resale_prices = pd.read_cs('data/resale-flat-prices-based-on-registration-date-from-march-2012-onwards.csv')

AttributeError: 'module' object has no attribute 'read_cs'
```

```
resale_price.head()
```

```
-----
NameError                                Traceback (most recent call last)
<ipython-input-21-e451aa5364ef> in <module>()
----> 1 resale_price.head()

NameError: name 'resale_price' is not defined
```

GETTING HELP

4. If you can't resolve the problem yourself within half an hour, reach out for assistance via our class Slack channel or on Stack Overflow

Provide details about the steps you have taken, code you are running and errors to facilitate troubleshooting

DEV ENVIRONMENT SETUP

- Brief intro of tools
- Environment setup
 - Create a Github account
 - Install Python 2.7 and Anaconda
 - Practice Python syntax, Terminal and Git commands
- Jupyter Notebook test and Python review

CONCLUSION

REVIEW

CONCLUSION

- You should now be able to answer the following questions:
 - What is Data Science?
 - What is the Data Science workflow?

WELCOME TO DATA SCIENCE

Q & A

WELCOME TO DATA SCIENCE

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET