# DSA5101
# Introduction to Big Data for Industry

LX Zhang

Department of Mathematics

National University of Singapore

# Module Information

Teaching Staff

## Lecturers

Dr. Zhang Louxin
matlzx@nus.edu.sg

Dr. Li Xiaoli
matv148@nus.edu.sg

Dr. Zhao Jinyuan
zhaojy@nus.edu.sg

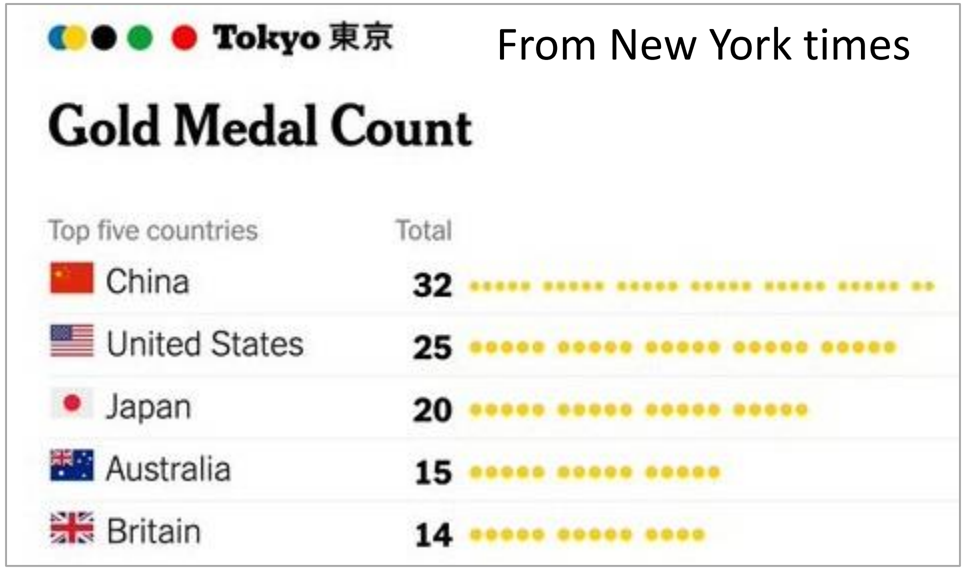
## Teaching Assistants

# Lecture Schedule

| | | |
|---|---|---|
| Lecture 1 (Zhang LX) | 8 Aug<br>15 Aug | What is Data Science? Programming techniques: recursive function and dynamic programming and web development. Python programming Assignment 1 |
| Lecture 2 (Li XL) | 22 Aug | Introduction to ML (clustering and classification methods) and programming. Project 1 |
| Lecture 3 | 29 Aug | |
| Lecture 4 | 5 Sept | |
| Lecture 5 | 12 Sept | |
| Recess Week (17 – 24 Sept) | | |
| Lecture 7 (Zhang LX) | 26 Sept | Selected topics of big data: Theory of visualization, Hadoop, finding similar items, processing data streams, mining social networks, etc. Assignment/Project 2 |
| Lecture 8 | 3 Oct | |
| Lecture 9 | 10 Oct | |
| Lecture 10 (Zhao JY) | 17 Oct | Data analysis and visualization in R (histogram, scatterplot, heatmap, regression.<br>Project 3 |
| ~~Lecture 11~~ | ~~24 Oct~~ | |
| Lecture 12 | 31 Oct | |
| Lecture 13 | 7 Nov | |
| Reading Week (12 – 18 Nov) | | |

# Outcome

- Learn practical issues of  data science:
  - -- data collection
  - -- data manipulation,
  - -- data cleaning
  - -- data analyses
  - -- visualization
- Have Python and R programming skills
- Master algorithms for mining big data, data streams and graph data.



From New York times

Tokyo 東京

## Gold Medal Count

| Top five countries | Total | |
|---|---|---|
| China | 32 | ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● ●● |
| United States | 25 | ●●●●● ●●●●● ●●●●● ●●●●● ●●●●● |
| Japan | 20 | ●●●●● ●●●●● ●●●●● ●●●●● |
| Australia | 15 | ●●●●● ●●●●● ●●●●● |
| Britain | 14 | ●●●●● ●●●●● ●●●● |

# Assessment policy

| | |
|---|---|
| Assignment 1 | 25% |
| (ML) Project 1 | 25% |
| Project 2/Assignment 2 | 25% |
| (R Prog.) Project 3 | 25% |
| **Total** | **100%** |

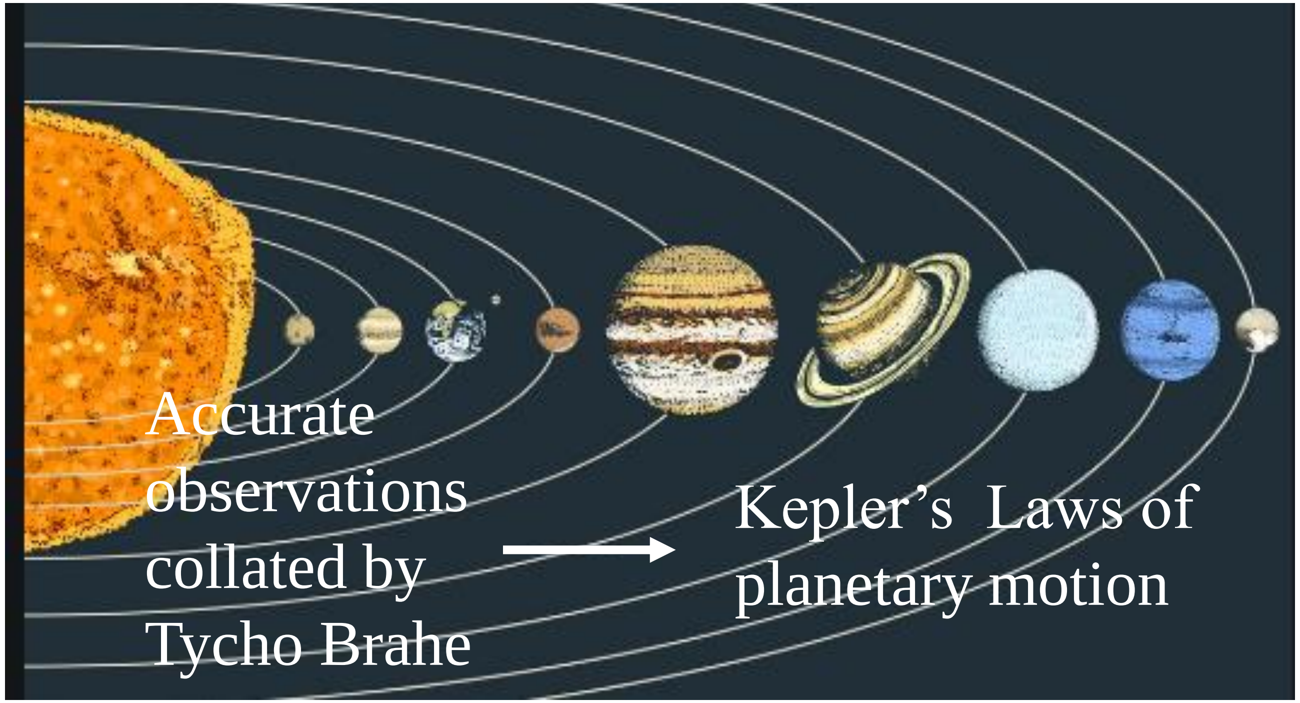
- Penalty for late assignment/project submission will be applied: 3% deduction per day
- Individual/group projects
- Plagiarism: Share ideas but not exact words. Heavy penalty

# Help

- Post the question in the module Chat room and hopefully your peers will answer. Instructors and TAs (if any) monitor and respond frequently the posts.
- Go to online Office Hour (4pm to 5pm) on Monday for the first three weeks, this is the best way to get help.
- For personal matters (regrading appeal, medical leave, etc.) send an email to: **matzlx@nus.edu.sg**

# I. What is Data Science?

- Data science is new, not data science practices have been there for long time



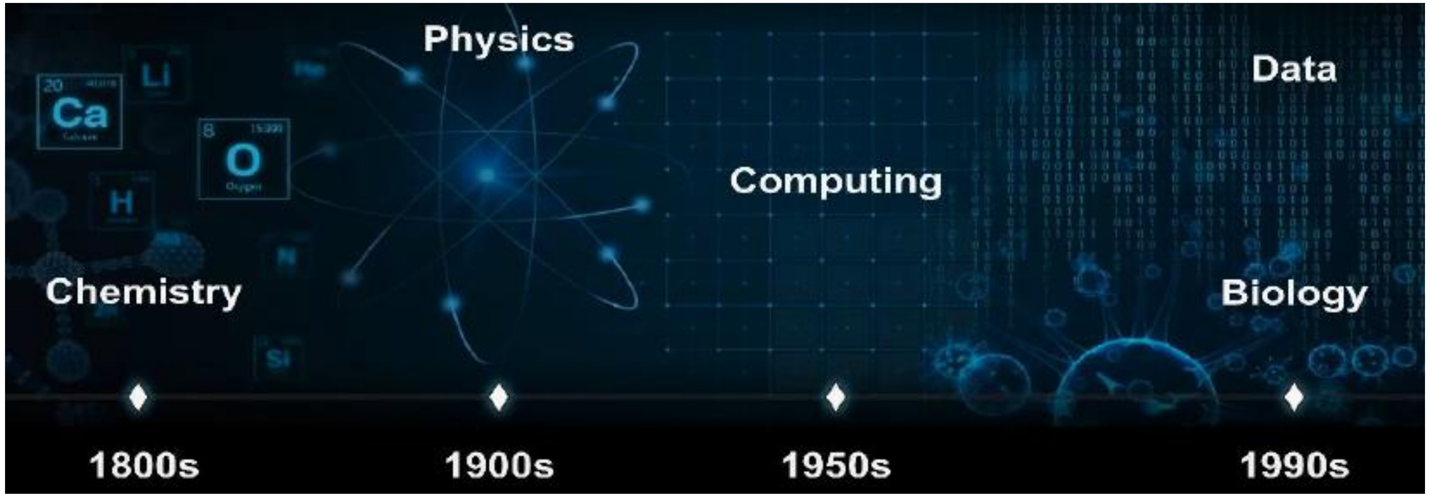
Accurate observations collated by Tycho Brahe → Kepler's Laws of planetary motion

# I. What is Data Science?

- Data science is new, not data science practices have been there for long time
- Data science emerged in the past decade, as

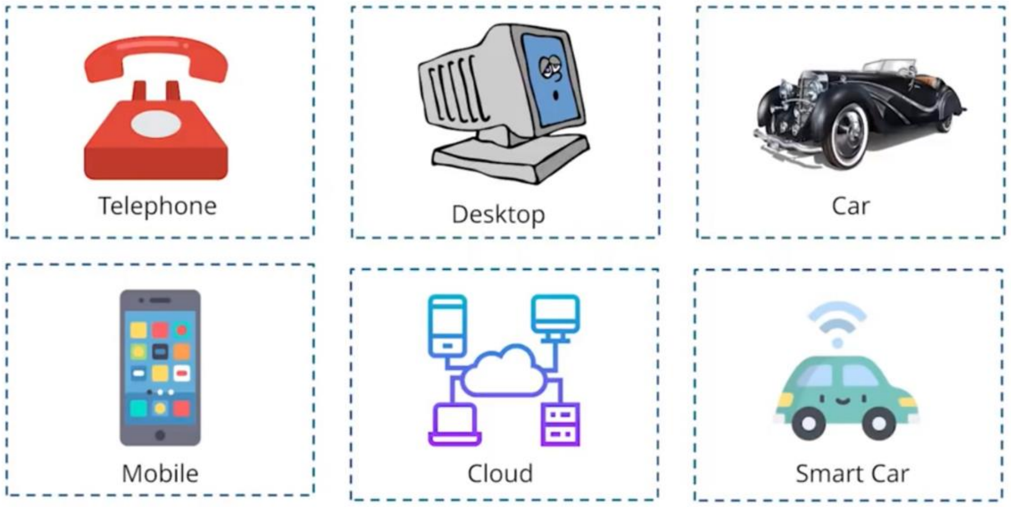
  -- Data are generated in a unstoppable pace

  -- Mobile, social media, and internet of things all produce big data



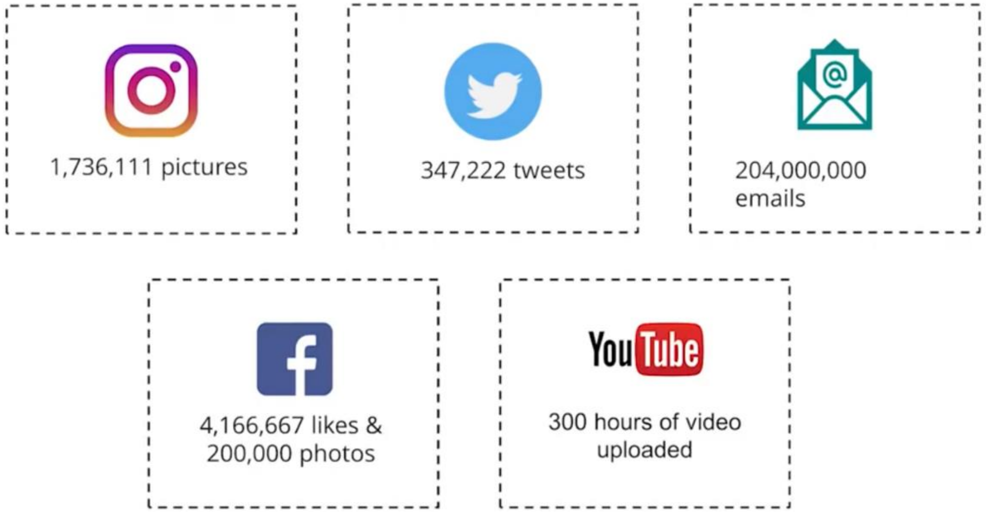
# Data Sources

- Evolution of Tech.



# Data Sources

- Evolution of tech.
- Social media
  --Instagram
  --Facebook
  --Twitter
  --WeChat
  --Tik Tok
  -- YouTube



1,736,111 pictures

347,222 tweets

204,000,000 emails

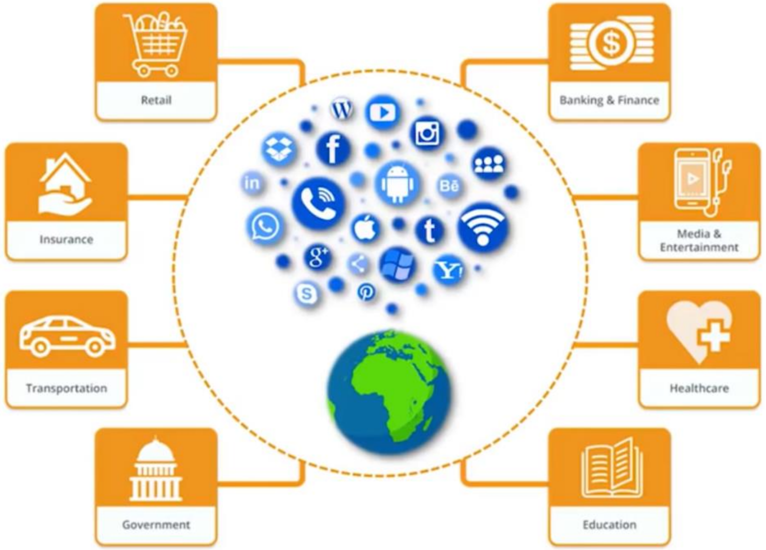
4,166,667 likes & 200,000 photos

300 hours of video uploaded

Data generated per minute

# Data Sources

- Evolution of tech.
- Social media
- Online business
  -- pay bills
  -- online shopping
  -- online education
  -- online healthcare



# Data Sources

- Evolution of tech.
- Social media
- Online business
- Internet of things
  - -- tools and devices that communicate and transfer data via internet
  - -- 500 zetabytes ($10^{21}$) of data per year

# I. What is Data Science?

- Data science is new, not data science practices have been there for long time
- Data science emerged recently, as

  -- Data generated in a unstoppable pace

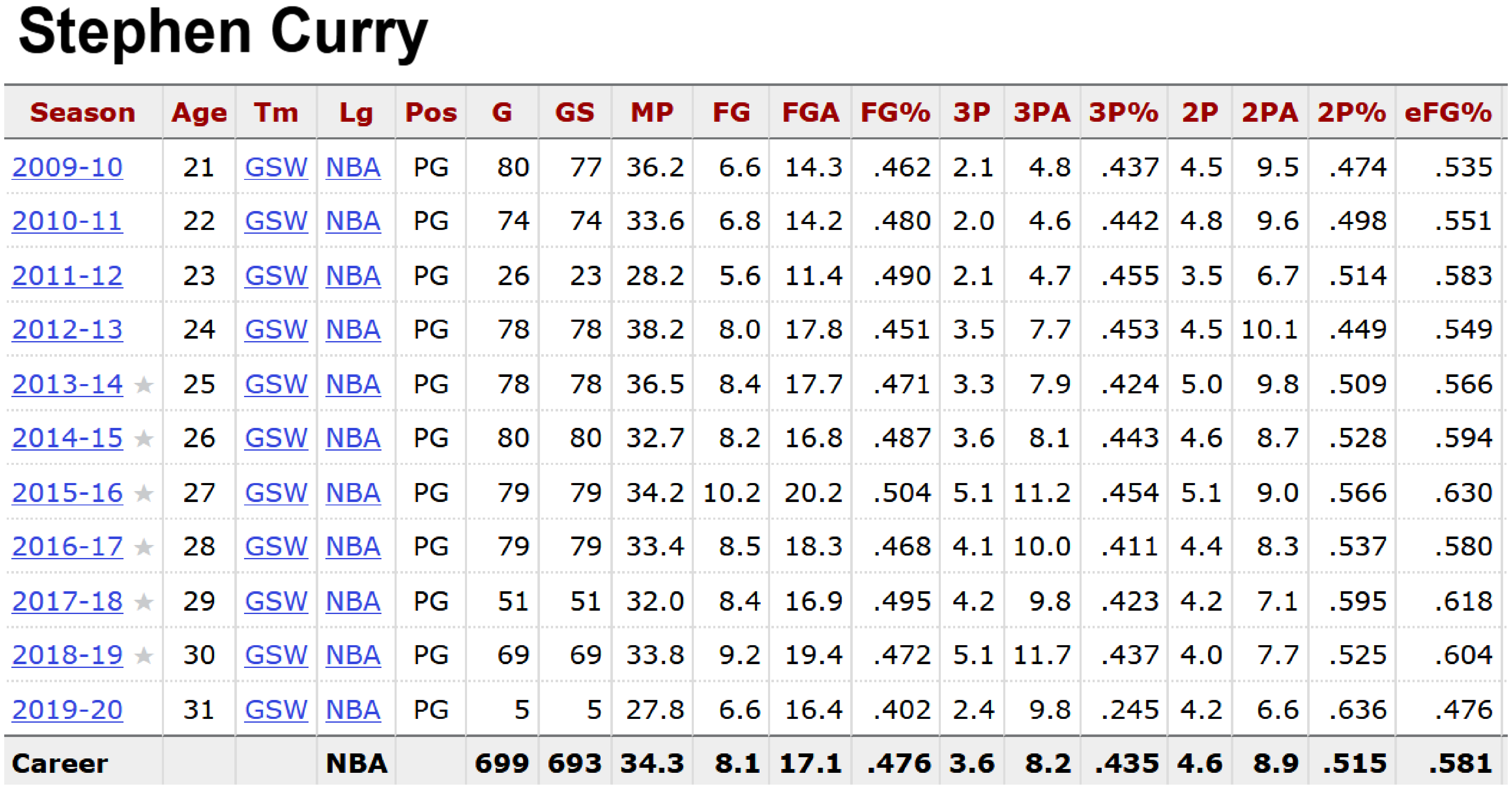
  -- Advances in computing technology allow us to analyze big data to draw useful insights for human beings
- Applications

  -- Classification of news
  -- Retail business empowered by data analyses
  -- Evaluation system in entertainment and sports
  -- Decision making during election
  -- Risk management (like COVID-19)

# Story 1: Walmart use data to improve business

- Walmart has its own data cloud, which is able to process 2.5 petabytes of data every hour.
- Facebook users were crazy about cake pops

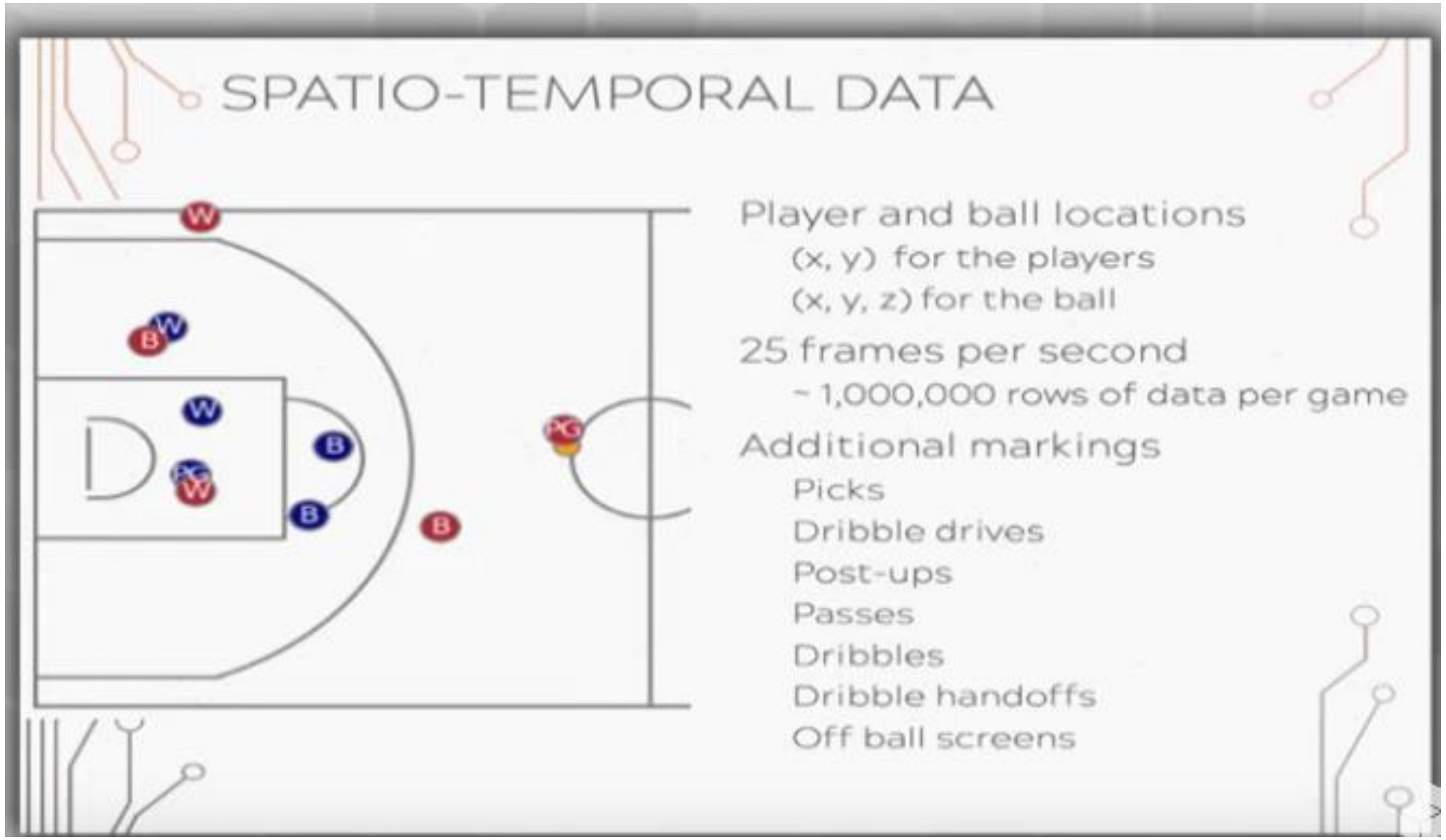
# Story 2: Data analyses in NBA



## Stephen Curry

| Season | Age | Tm | Lg | Pos | G | GS | MP | FG | FGA | FG% | 3P | 3PA | 3P% | 2P | 2PA | 2P% | eFG% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2009-10 | 21 | GSW | NBA | PG | 80 | 77 | 36.2 | 6.6 | 14.3 | .462 | 2.1 | 4.8 | .437 | 4.5 | 9.5 | .474 | .535 |
| 2010-11 | 22 | GSW | NBA | PG | 74 | 74 | 33.6 | 6.8 | 14.2 | .480 | 2.0 | 4.6 | .442 | 4.8 | 9.6 | .498 | .551 |
| 2011-12 | 23 | GSW | NBA | PG | 26 | 23 | 28.2 | 5.6 | 11.4 | .490 | 2.1 | 4.7 | .455 | 3.5 | 6.7 | .514 | .583 |
| 2012-13 | 24 | GSW | NBA | PG | 78 | 78 | 38.2 | 8.0 | 17.8 | .451 | 3.5 | 7.7 | .453 | 4.5 | 10.1 | .449 | .549 |
| 2013-14 ★ | 25 | GSW | NBA | PG | 78 | 78 | 36.5 | 8.4 | 17.7 | .471 | 3.3 | 7.9 | .424 | 5.0 | 9.8 | .509 | .566 |
| 2014-15 ★ | 26 | GSW | NBA | PG | 80 | 80 | 32.7 | 8.2 | 16.8 | .487 | 3.6 | 8.1 | .443 | 4.6 | 8.7 | .528 | .594 |
| 2015-16 ★ | 27 | GSW | NBA | PG | 79 | 79 | 34.2 | 10.2 | 20.2 | .504 | 5.1 | 11.2 | .454 | 5.1 | 9.0 | .566 | .630 |
| 2016-17 ★ | 28 | GSW | NBA | PG | 79 | 79 | 33.4 | 8.5 | 18.3 | .468 | 4.1 | 10.0 | .411 | 4.4 | 8.3 | .537 | .580 |
| 2017-18 ★ | 29 | GSW | NBA | PG | 51 | 51 | 32.0 | 8.4 | 16.9 | .495 | 4.2 | 9.8 | .423 | 4.2 | 7.1 | .595 | .618 |
| 2018-19 ★ | 30 | GSW | NBA | PG | 69 | 69 | 33.8 | 9.2 | 19.4 | .472 | 5.1 | 11.7 | .437 | 4.0 | 7.7 | .525 | .604 |
| 2019-20 | 31 | GSW | NBA | PG | 5 | 5 | 27.8 | 6.6 | 16.4 | .402 | 2.4 | 9.8 | .245 | 4.2 | 6.6 | .636 | .476 |
| **Career** | | | **NBA** | | **699** | **693** | **34.3** | **8.1** | **17.1** | **.476** | **3.6** | **8.2** | **.435** | **4.6** | **8.9** | **.515** | **.581** |

Six camera have been used to catch every movement of players and ball on court in NBA



SPATIO-TEMPORAL DATA

Player and ball locations

- (x, y) for the players
- (x, y, z) for the ball

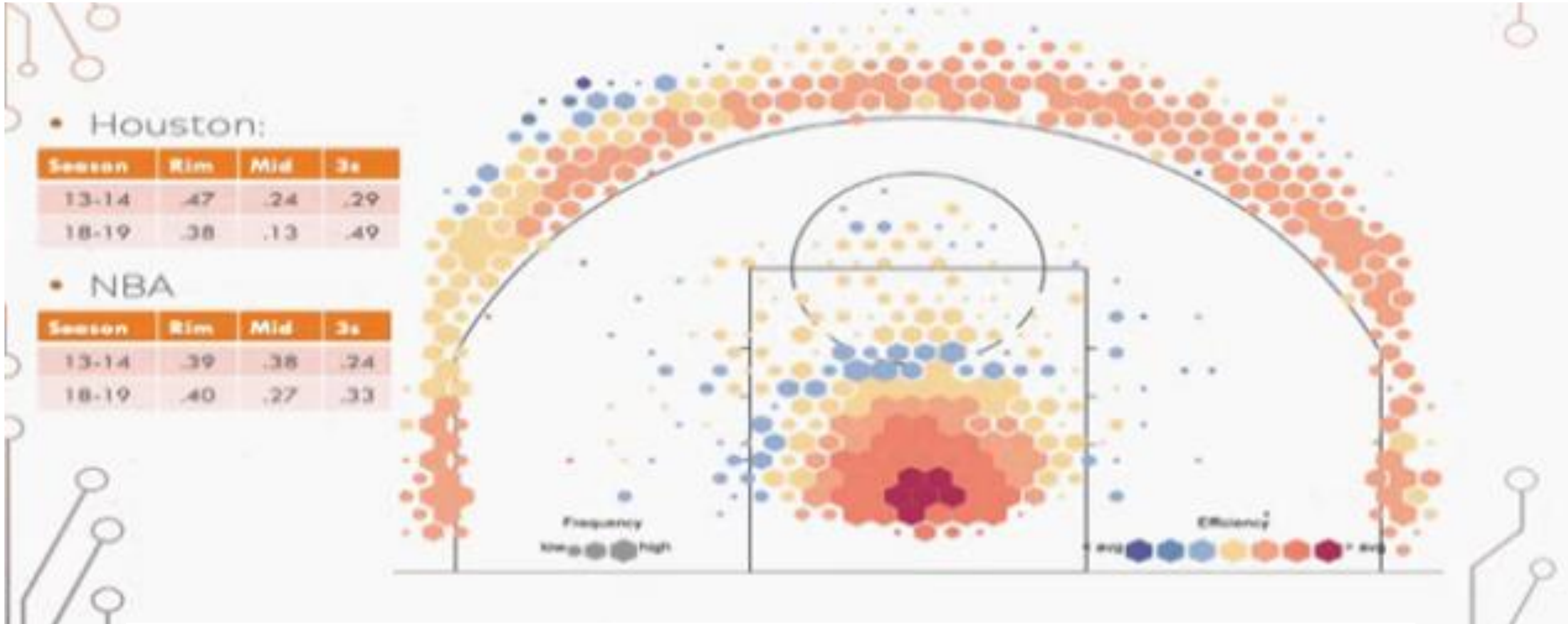
25 frames per second

- ~ 1,000,000 rows of data per game

Additional markings

- Picks
- Dribble drives
- Post-ups
- Passes
- Dribbles
- Dribble handoffs
- Off ball screens

The Economist
INING
EDGE
Subscribe



- Houston:

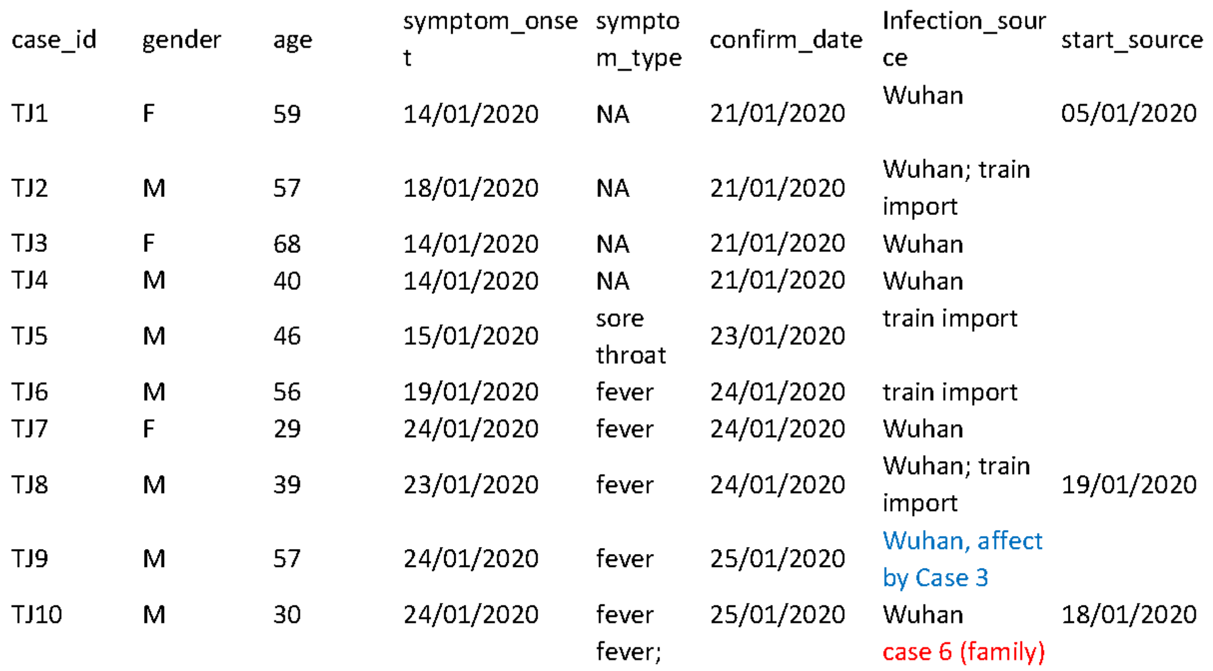
| Season | Rim | Mid | 3s |
|---|---|---|---|
| 13-14 | .47 | .24 | .29 |
| 18-19 | .38 | .13 | .49 |

- NBA

| Season | Rim | Mid | 3s |
|---|---|---|---|
| 13-14 | .39 | .38 | .24 |
| 18-19 | .40 | .27 | .33 |

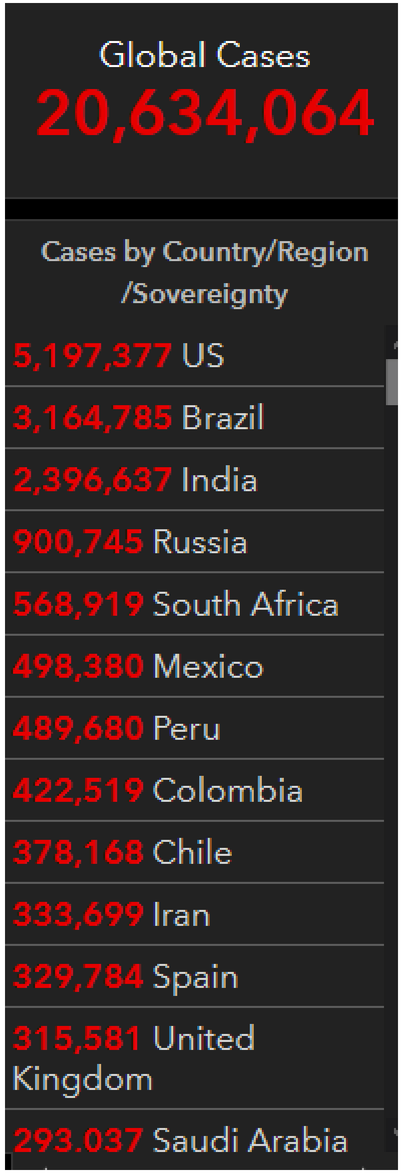
Shooting chart of Houston Rockets

# Story 3: COVID-19 risk management



| case_id | gender | age | symptom_onset | symptom_type | confirm_date | Infection_source | start_source |
|---|---|---|---|---|---|---|---|
| TJ1 | F | 59 | 14/01/2020 | NA | 21/01/2020 | Wuhan | 05/01/2020 |
| TJ2 | M | 57 | 18/01/2020 | NA | 21/01/2020 | Wuhan; train import | |
| TJ3 | F | 68 | 14/01/2020 | NA | 21/01/2020 | Wuhan | |
| TJ4 | M | 40 | 14/01/2020 | NA | 21/01/2020 | Wuhan | |
| TJ5 | M | 46 | 15/01/2020 | sore throat | 23/01/2020 | train import | |
| TJ6 | M | 56 | 19/01/2020 | fever | 24/01/2020 | train import | |
| TJ7 | F | 29 | 24/01/2020 | fever | 24/01/2020 | Wuhan | |
| TJ8 | M | 39 | 23/01/2020 | fever | 24/01/2020 | Wuhan; train import | 19/01/2020 |
| TJ9 | M | 57 | 24/01/2020 | fever | 25/01/2020 | Wuhan, affect by Case 3 | |
| TJ10 | M | 30 | 24/01/2020 | fever | 25/01/2020 | Wuhan | 18/01/2020 |
| | | | | fever; | | case 6 (family) | |

- What is the infection ratio (R0: basic reproduction number)?
- What is the incubation period (time from exposure to visual to the symptom on set)?
- What intervention policy should government adopt to control effectively the Covid-19?

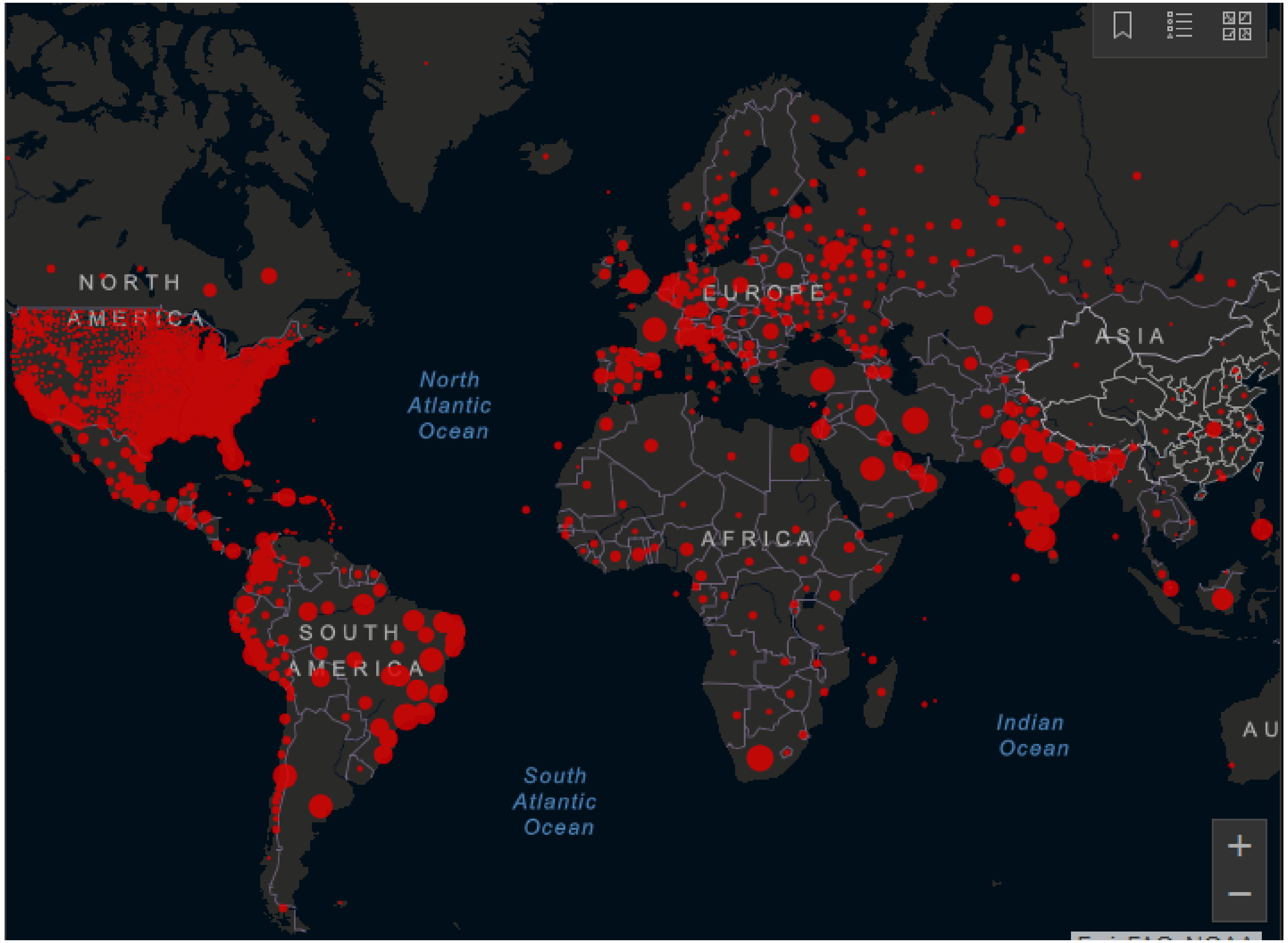


Global Cases

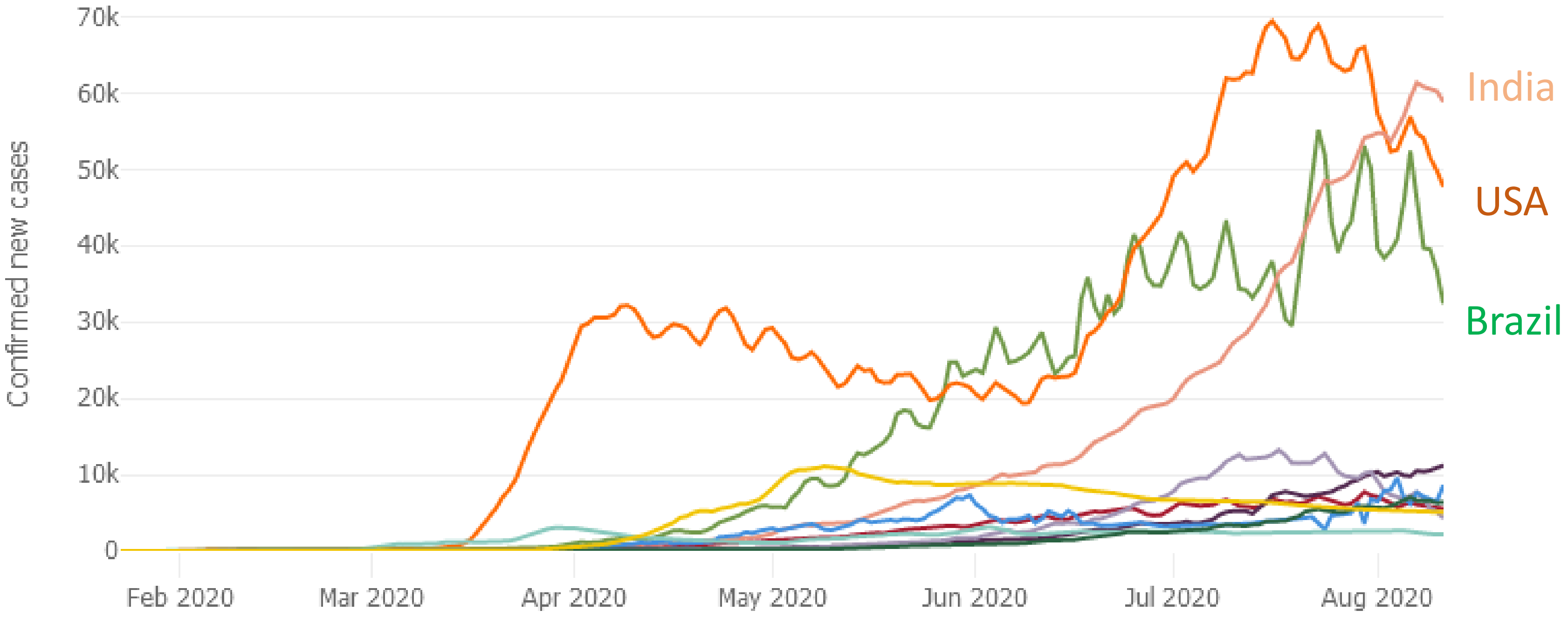
20,634,064

Cases by Country/Region/Sovereignty

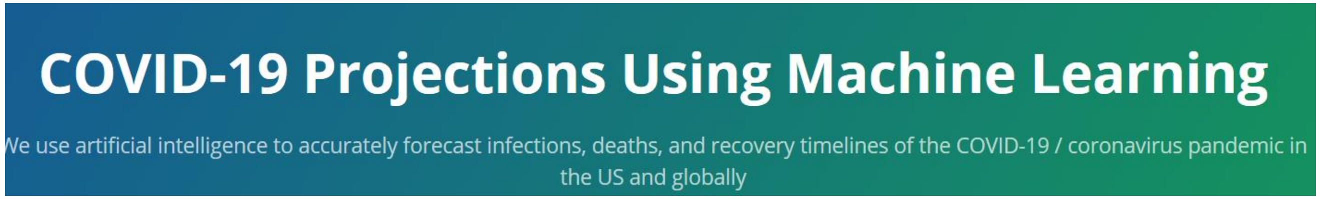
| | |
|---|---|
| 5,197,377 | US |
| 3,164,785 | Brazil |
| 2,396,637 | India |
| 900,745 | Russia |
| 568,919 | South Africa |
| 498,380 | Mexico |
| 489,680 | Peru |
| 422,519 | Colombia |
| 378,168 | Chile |
| 333,699 | Iran |
| 329,784 | Spain |
| 315,581 | United Kingdom |
| 293,037 | Saudi Arabia |



https://coronavirus.jhu.edu/map.html

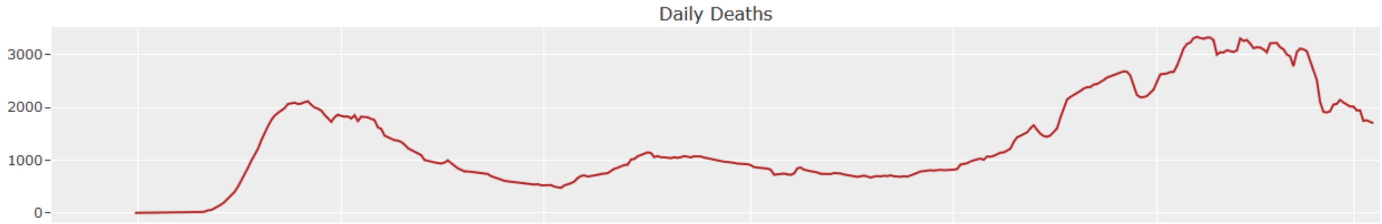


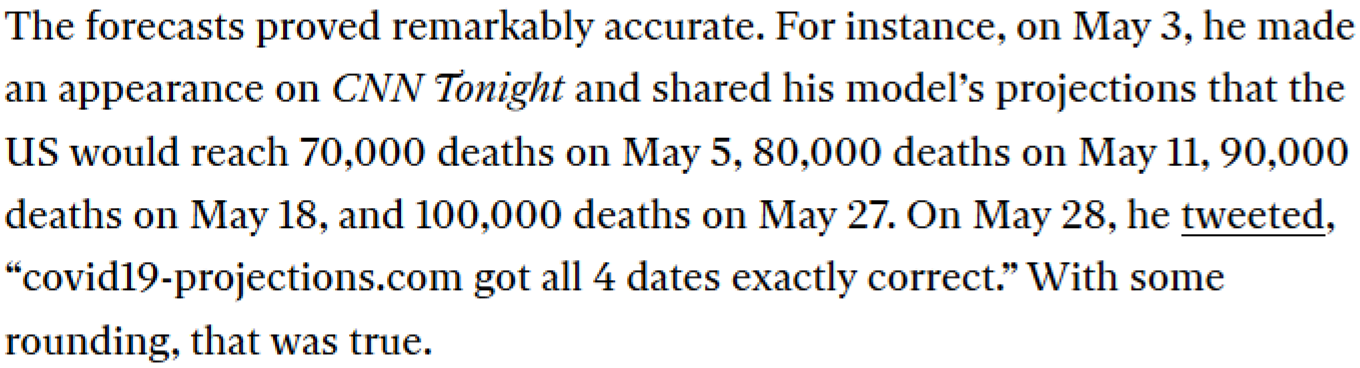
Curves of daily cases



# COVID-19 Projections Using Machine Learning

We use artificial intelligence to accurately forecast infections, deaths, and recovery timelines of the COVID-19 / coronavirus pandemic in the US and globally





The forecasts proved remarkably accurate. For instance, on May 3, he made an appearance on *CNN Tonight* and shared his model's projections that the US would reach 70,000 deaths on May 5, 80,000 deaths on May 11, 90,000 deaths on May 18, and 100,000 deaths on May 27. On May 28, he tweeted, "covid19-projections.com got all 4 dates exactly correct." With some rounding, that was true.

Youyang Gu

https://www.technologyreview.com/2021/04/27/1023657/lessons-from-the-pandemics-superstar-data-scientist-youyang-gu/

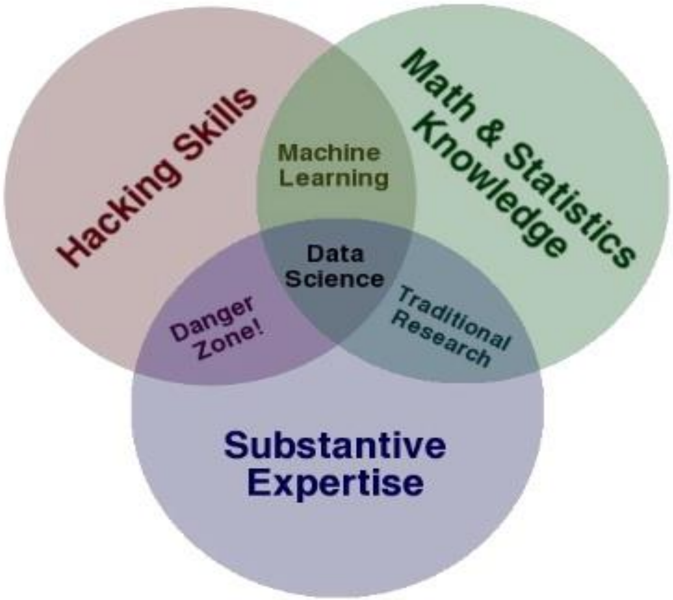
# I. What is Data Science?

- Data science is new, not data science practices have been there for long time
- Data science emerged recently, because

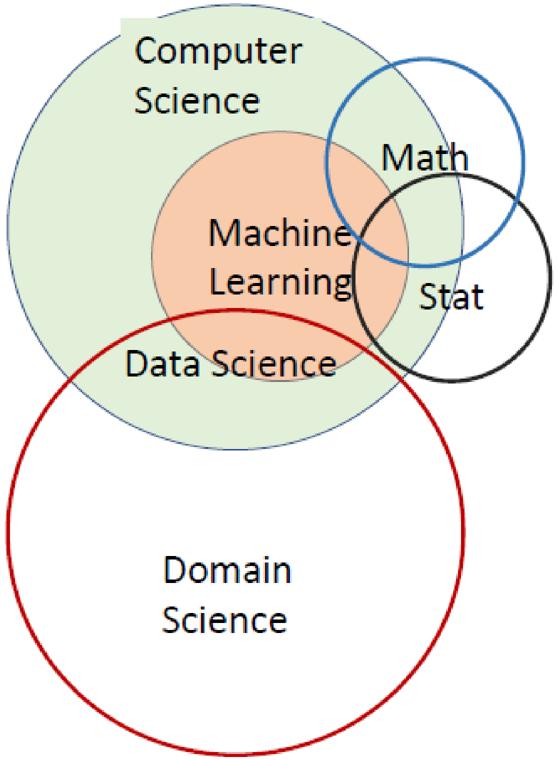
  -- Data generated in a unstoppable pace

  -- Advances in computing technology allow us to analyze huge datasets to draw useful insights for human beings
- Data science is applying scientific methods, algorithms and systems to learn from data and to transfer the data into actionable insights.

# II. Nature of Data Science:  multi-disciplinary practice



http://drewconway.com/



Jeffrey Ullaman’s diagram

# Skills sets for data scientists

- Statistical data analysis and visualization
- Machine learning
- Scalable (cloud and high-performance) computing tech
- Communication skills
- Storytelling skills
- Curiosity

“But I think it’s also important to not just blindly trust science,” he continues. “Scientists aren’t perfect.” It is appropriate, he says, if something doesn’t seem right, to ask questions and find explanations. “It’s important to have different perspectives. If there is anything we’ve learned over the past year, it’s that no one is 100% right all the time.” --- Youyang Gu

"Now I'm old, I'm 30, and I started to realise that all those people who say they know, they actually don't know. Many of them don't know, and especially those who say that they know, don't know, because those who do know say that they don't know."

**Anna Kiesenhofer**

# My experience with Data Science

- Trained as a mathematician & theoret. computer scientist
- Research in computational biology and bioinformatics
- Work on genomic sequence, protein interaction networks, prediction of drug responses
- Interest in data visualization

# Appreciating Data

- Traditionally, mathematicians and computer scientists focus on methods rather than data.
- They validates their methods using random data/simulated data
- But interesting/useful data are a scarce resource

# Reality and virtual world

- Mathematicians and computer scientists build their own clean and organized virtual world
- But, the real world is complicated and messy by nature
- In real world, nothing is completely true or false
- People other than mathematicians and computer scientists are comfortable with errors in data, whereas mathematicians are not.

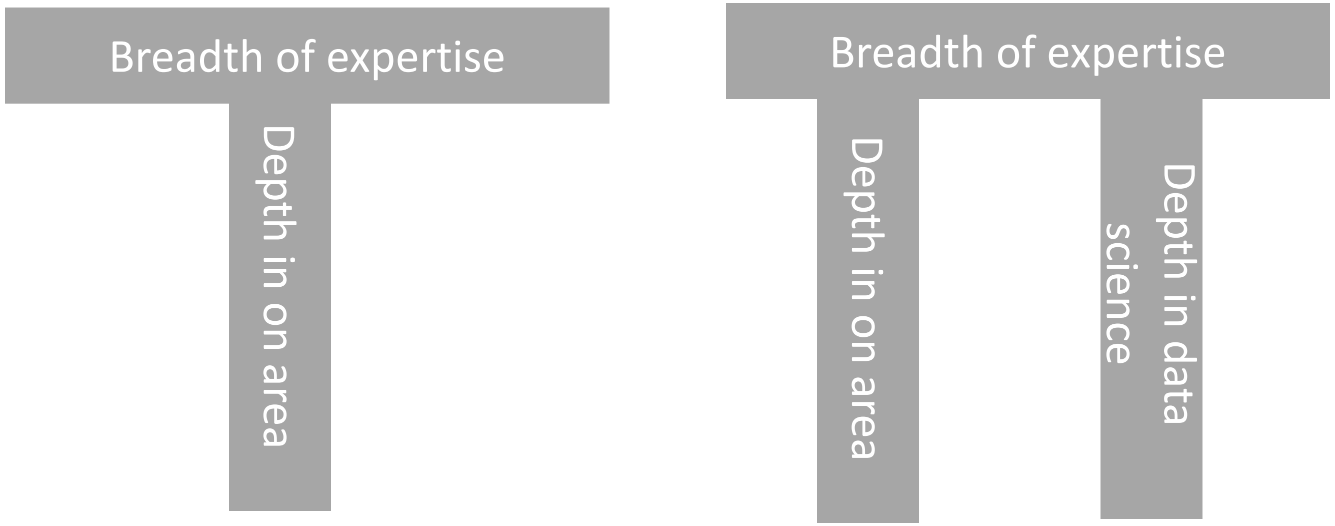
# Wisdom vs Genius

- Genius shows in finding right answers
- Wisdom shows in avoiding the wrong answers
- Data science benefits more from wisdom than from genius

- Wisdom comes from general knowledge
- Wisdom comes from experience: how often you have been wrong and why/how
- Wisdom comes from listening to others.

# T-shaped team vs Pi-shaped team

- Data science is an inherently collaborative art.
- Data science involves teams of people collaborating



Breadth of expertise

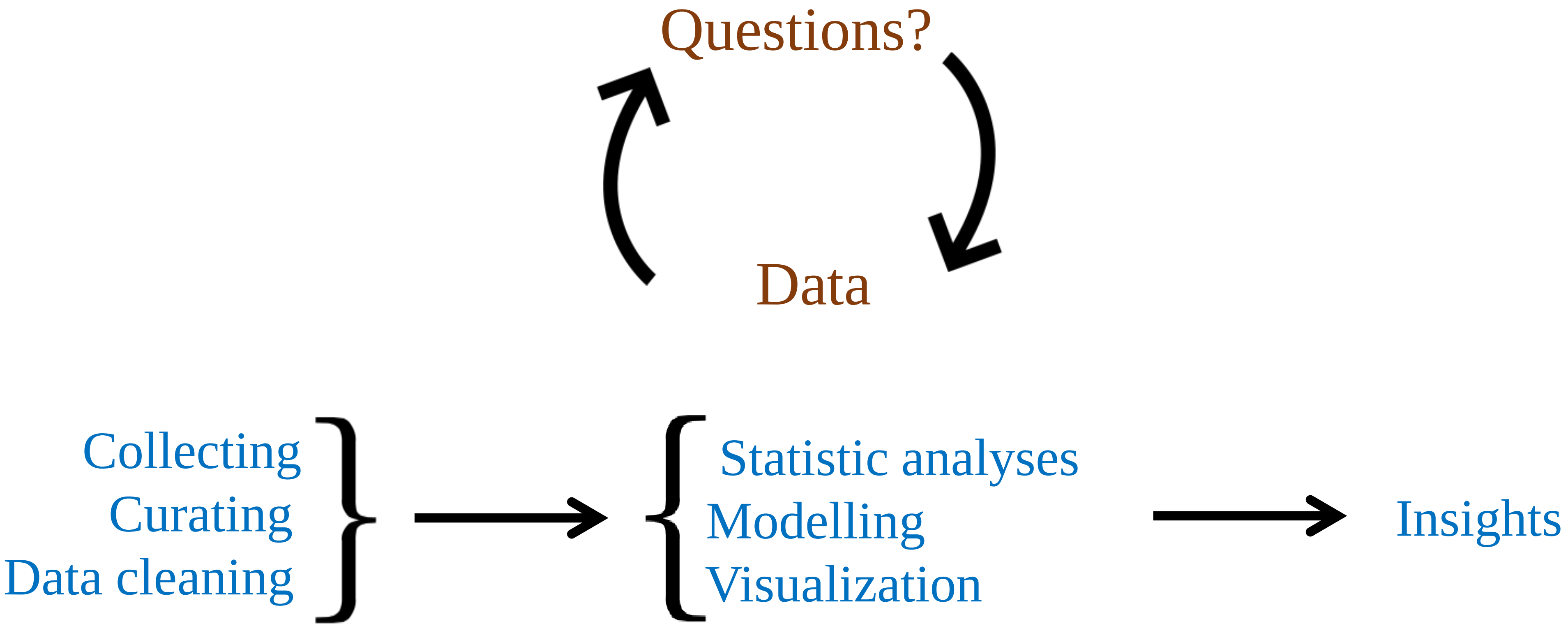
Depth in on area

Breadth of expertise

Depth in on area

Depth in data science

# III. Data Science Process (or Data Life Cycle)



Questions?

Data

Collecting
Curating
Data cleaning

Statistic analyses
Modelling
Visualization

Insights

## Steps

- Collection
- Curating
- Cleaning
- Stat. analysis
- Modelling
- Visualization
- Insights

0. Understand the patterns in the data?
1. Retrieve useful insight?
2. Form hypothesis?
3. Select data features for the machine-learning model
4. Create an accurate model for the purpose
5. Evaluate and test the model.

# Steps

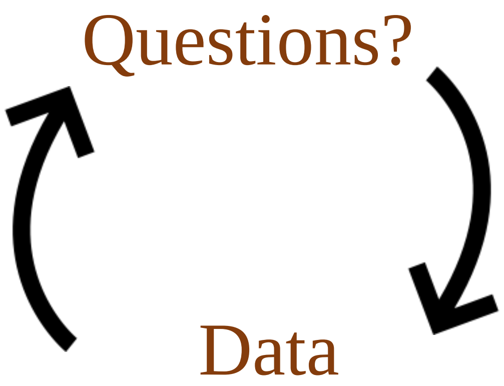
- Collection
- Curating
- Cleaning
- Stat. analysis
- Modelling
- Visualization
- Insights

0. What plots will be used?

1. What are useful insight?

# SUMMARY

- Data science is about how to transform data into information



Questions?

Data

Statistics
Programming languages
Data extraction & processing
Data wrangling & exploration
Machine Learning
Big Data processing frameworks
Data visualisation