Dr. Andrew L. Tan
Data Science Institute

Data 100

Introduction to Data Science

What is data science?

Data Science

Multi-disciplinary field which aims to derive insights and knowledge from both structured and unstructured data.

Vasant Dhar, 2013

Data Science

Multi-disciplinary field which aims to derive insights and knowledge from both structured and unstructured **data**.

Vasant Dhar, 2013

What id Data?

According to Merriam-Webster :

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available

— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published

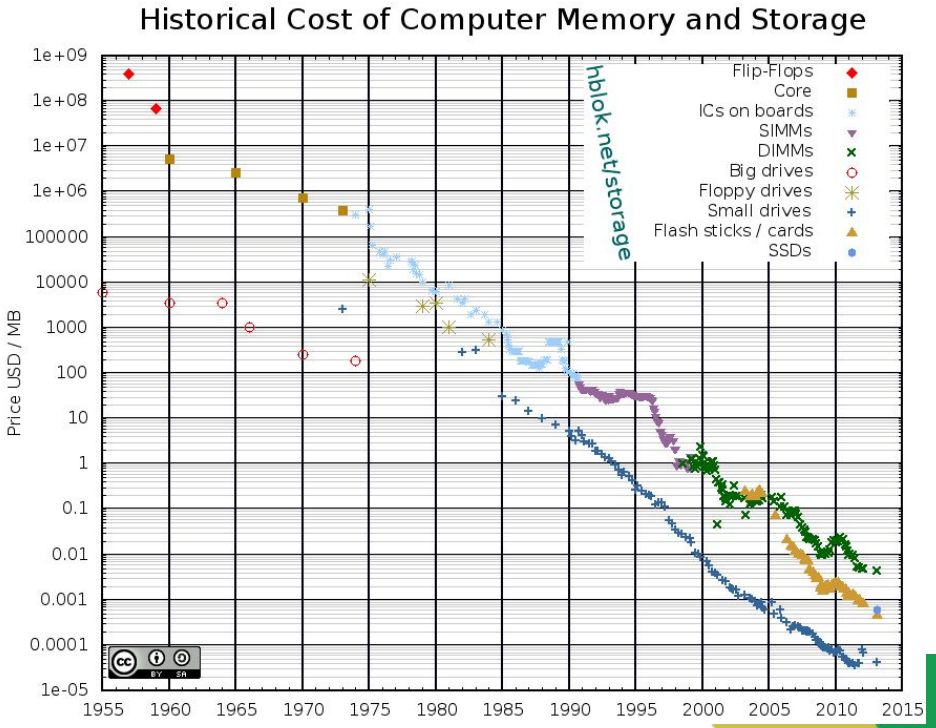
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Why is data more relevant now?

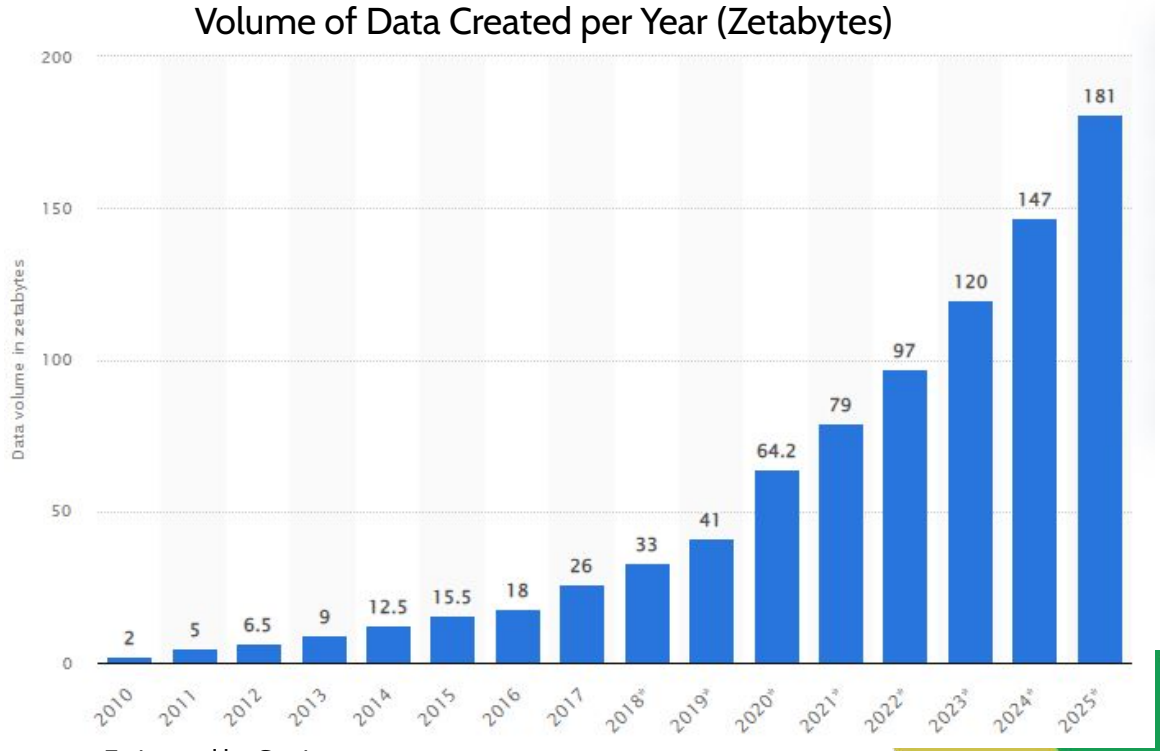
Data storage is cheaper



Why is data more relevant now?

Applications and websites started collecting more data that is :

Expressive
Granular



Estimated by Statista

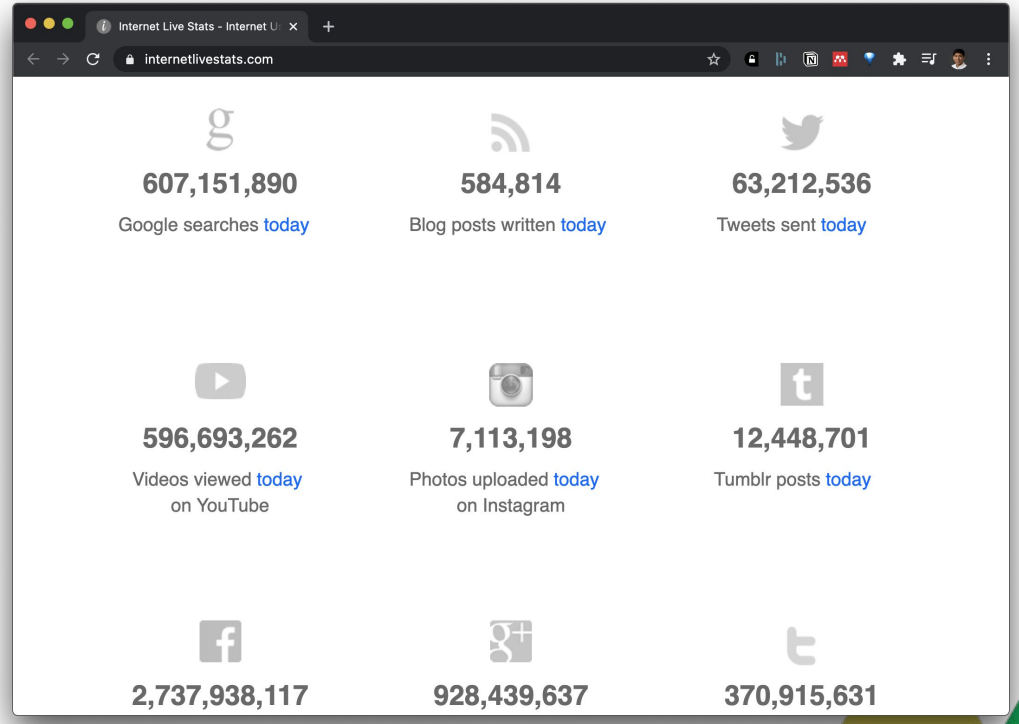
Dr. Andrew L. Tan
Data Science Institute

Why is data more relevant now?

Computing power became easier to access
Setting up a supercomputer that is “pay-per-use”

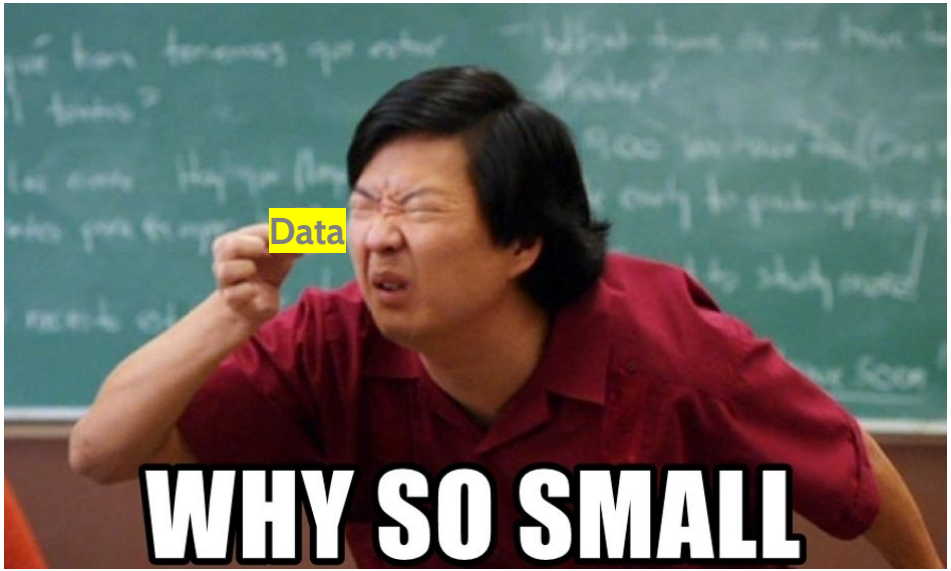
Big Data

Because of internet and connectivity, we generate so much data now!



https://www.internetlivestats.com/

Considerations – When is Data Big?



“If you can open it in excel, it’s not big data”

“If your ram can handle it, it’s not big data”

“Terabytes worth and up”

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

the huge *amount* of data that is generated daily

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

the diversity of *data types* and *data sources*

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

the *speed* at which data is generated, processed and analyzed

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

Sample: Volume, Variety, Velocity

Structural details of data

Uniqueness of Columns and Data Types → **Variety**

Generation of Samples and processing → **Velocity**

Name	Age	Gender	Birthday	Another Col
Person 1	3	M	1/1/2020	
Person 2	25	F	2/1/1995	
Person 3	80	M	3/4/2015	

Volume

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

the authenticity and credibility of the data (and in turn, its quality)

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

V's of Big Data

- Volume
- Variety
- Velocity
- Veracity
- Value

the added value for various stakeholders (company, government, everyone!)

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

Sample: Veracity – Value

Qualitative details

Name	Age	Gender	Birthday
Person 1	.5	M	1/1/2020
Person 2	25	F	2/1/1995
Person 3	80	M	3/4/2050

Veracity

Value

Insight : Majority of the attendees are Male

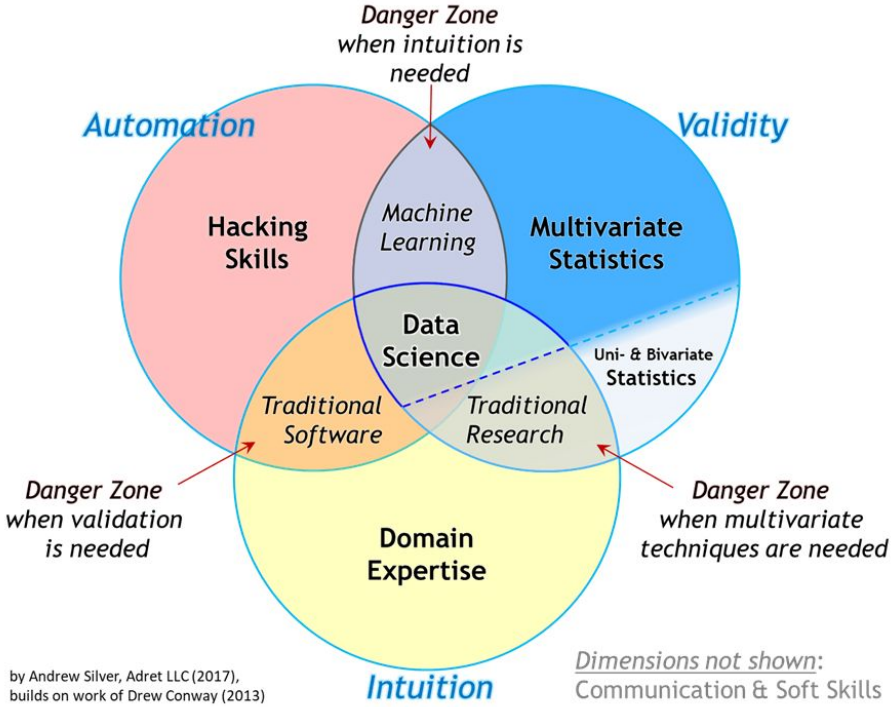
Age and Birthday columns are not aligned

https://blog.unbelievable-machine.com/en/what-is-big-data-definition-five-vs

Data Science

Multi-disciplinary field which aims to derive insights and knowledge from both structured and unstructured data.

Vasant Dhar, 2013



The Essential Data Science Venn Diagram (Silver, A., 2017)

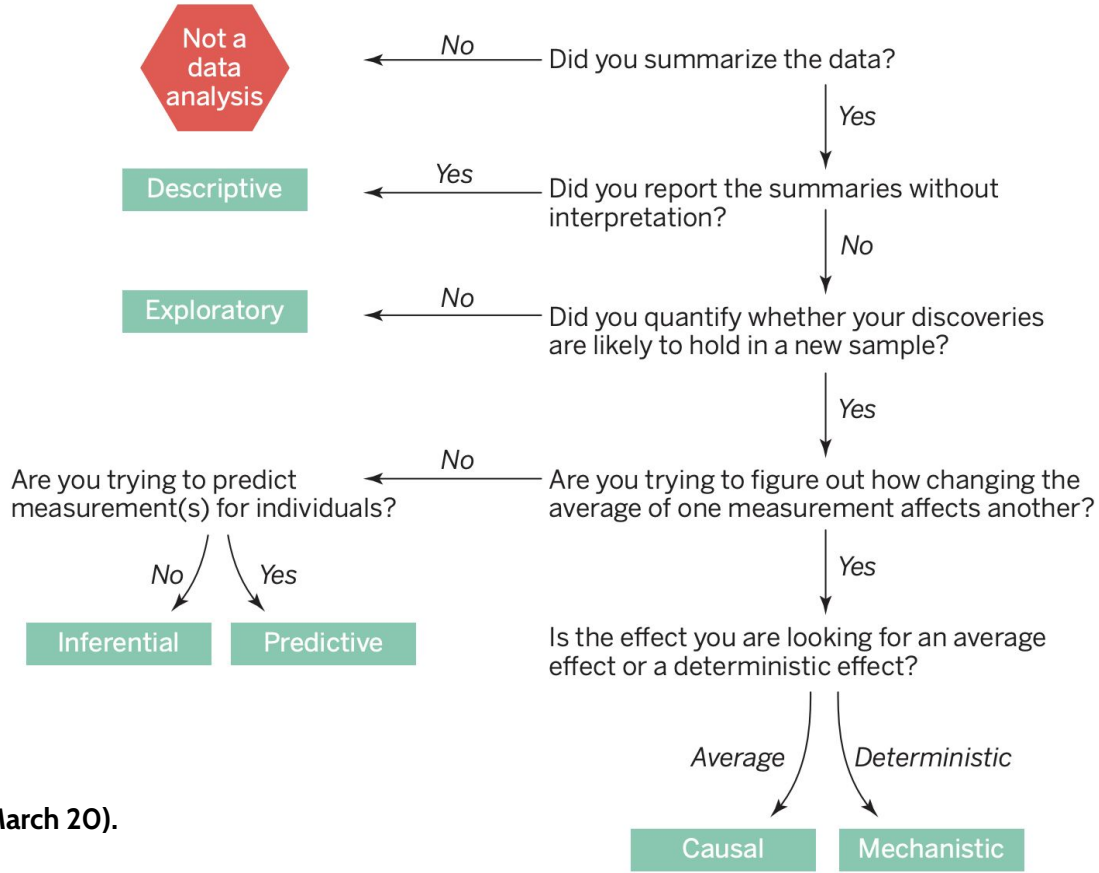
https://towardsdatascience.com/the-essential-data-science-venn-diagram-35800c3bef40

Data Science

Multi-disciplinary field which aims to derive **insights** and **knowledge** from both structured and unstructured data.

Vasant Dhar, 2013

Data analysis flowchart



Leek, J. and Peng, R. (2015, March 20).
What is the Question?
***Science Magazine*, 347(6228), pp. 1314-1315.**

Types of Analysis

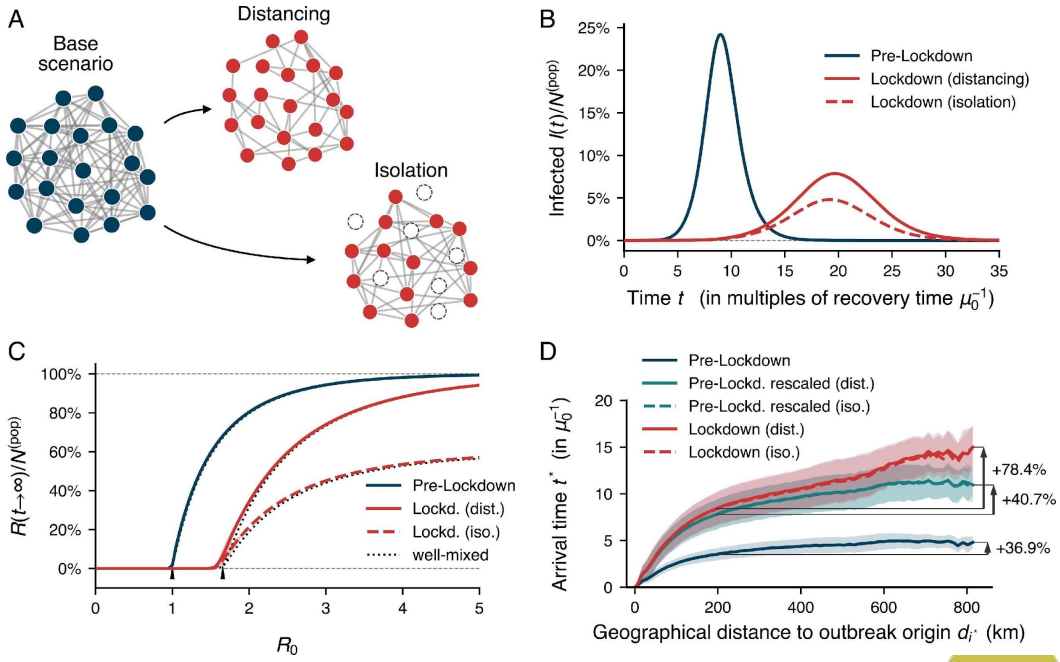
Descriptive	**Descriptive**	Aggregating Totals- What is happening?
Diagnostic	**Exploratory**	Spread and central tendency - What is the average? Are there outliers? Trend?
	Inferential	Difference between two samples (Male vs Female, NCR vs Region4)
Predictive	**Causal**	Relationship and correlation of two variables (What makes our variables tick)
	Predictive	Looking forward to see what will happen in the future
Prescriptive	**Mechanistic**	Effect estimation and simulation of variables (Policy making)

Types of Analysis

Descriptive	**Descriptive**	What areas are high in COVID cases -Philippines?
Diagnostic	**Exploratory**	Is the trend of COVID cases going upwards or downwards?
	Inferential	Is the daily change of cases in NCR the same as the change of cases in Region 4?
Predictive	**Causal**	If mask wearing was not implemented in the Philippines, how will it affect the number of cases in the country?
	Predictive	How many COVID cases will happen in the next month?
Prescriptive	**Mechanistic**	How does mask wearing reduce the transmission of COVID? (Effect estimation and simulation)

Types of Analysis

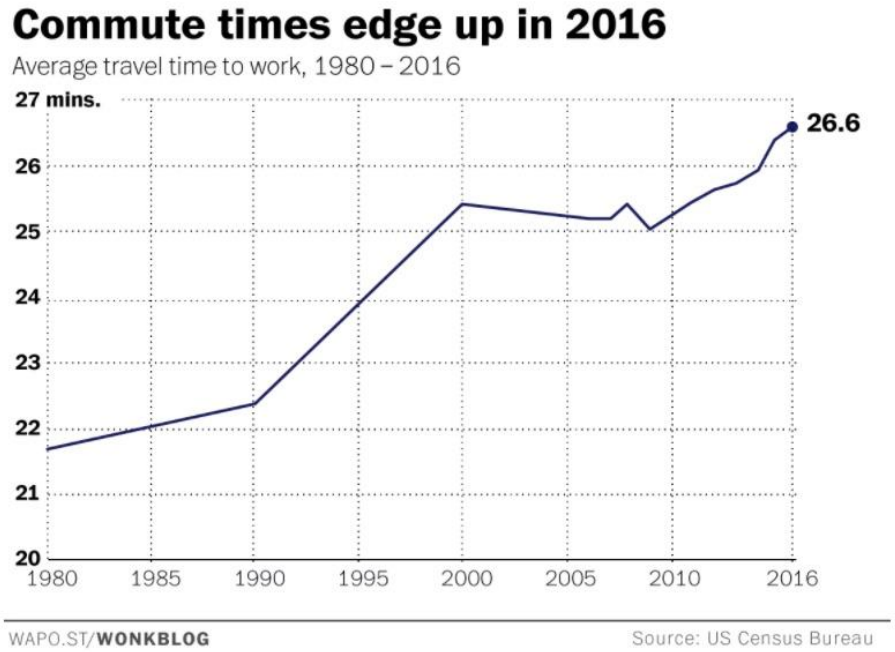
- Descriptive
- Exploratory
- Inferential
- Causal
- Predictive
- **Mechanistic**



Simulating the effects of lockdown with respect to patients that will get infected

Types of Analysis

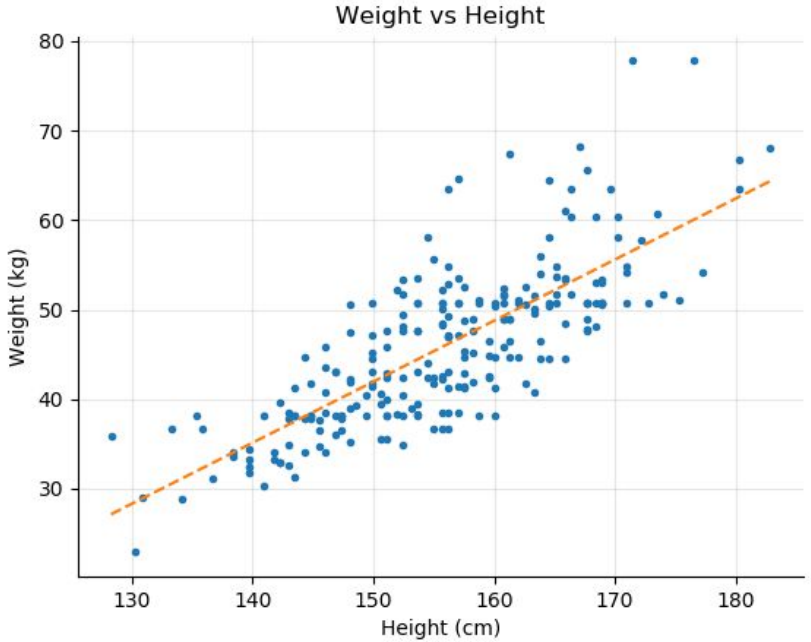
Descriptive
Exploratory
Inferential
Causal
Predictive
Mechanistic



The "trend" of commute time is going up.

Types of Analysis

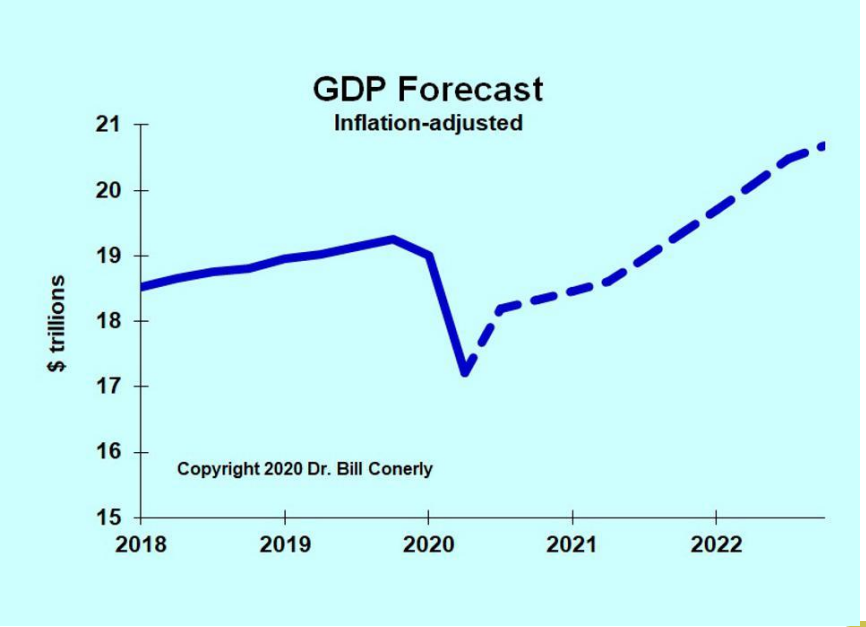
Descriptive
Exploratory
Inferential
Causal
Predictive
Mechanistic



What is the correlation of height and weight? People who are taller are generally observed to be heavier on average.

Types of Analysis

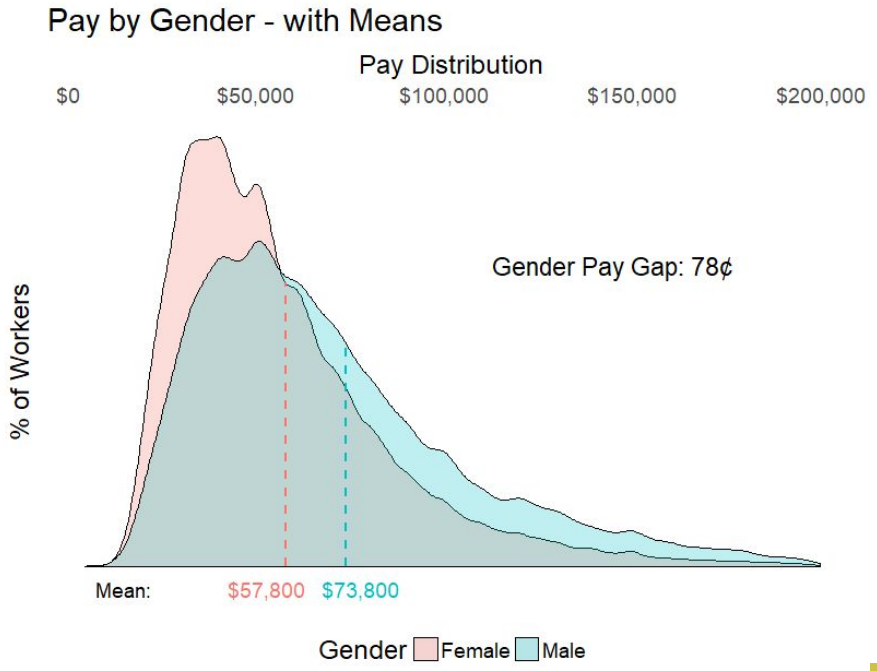
Descriptive
Exploratory
Inferential
Causal
Predictive
Mechanistic



What is the GDP Next year?

Types of Analysis

Descriptive
Exploratory
Inferential
Causal
Predictive
Mechanistic

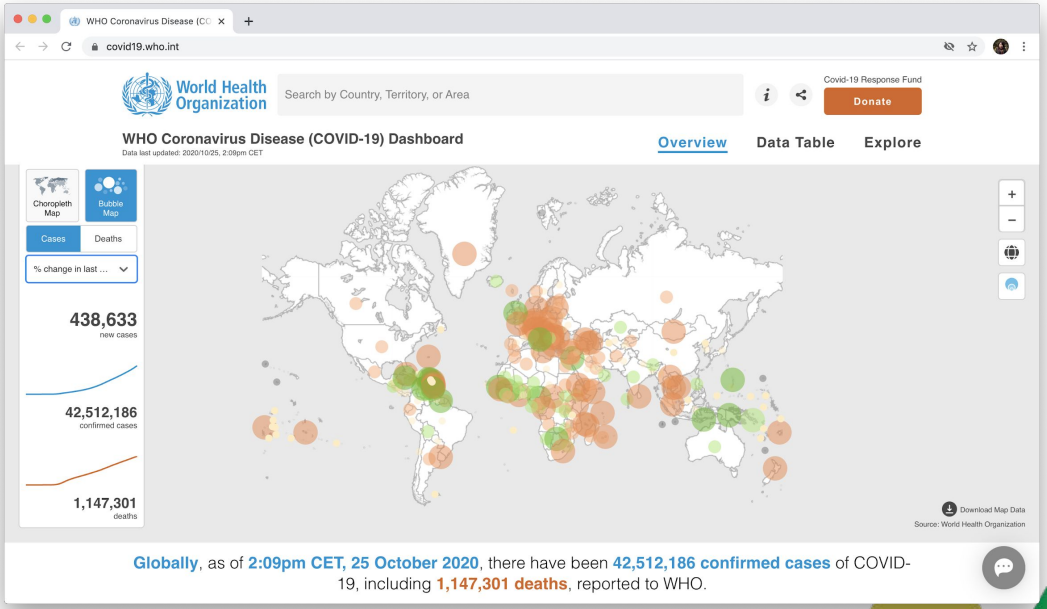


Are Males paid more on average as compared to Females? Is gender gap real? Whats the difference between the two groups?

Dr. Andrew L. Tan Data Science Institute

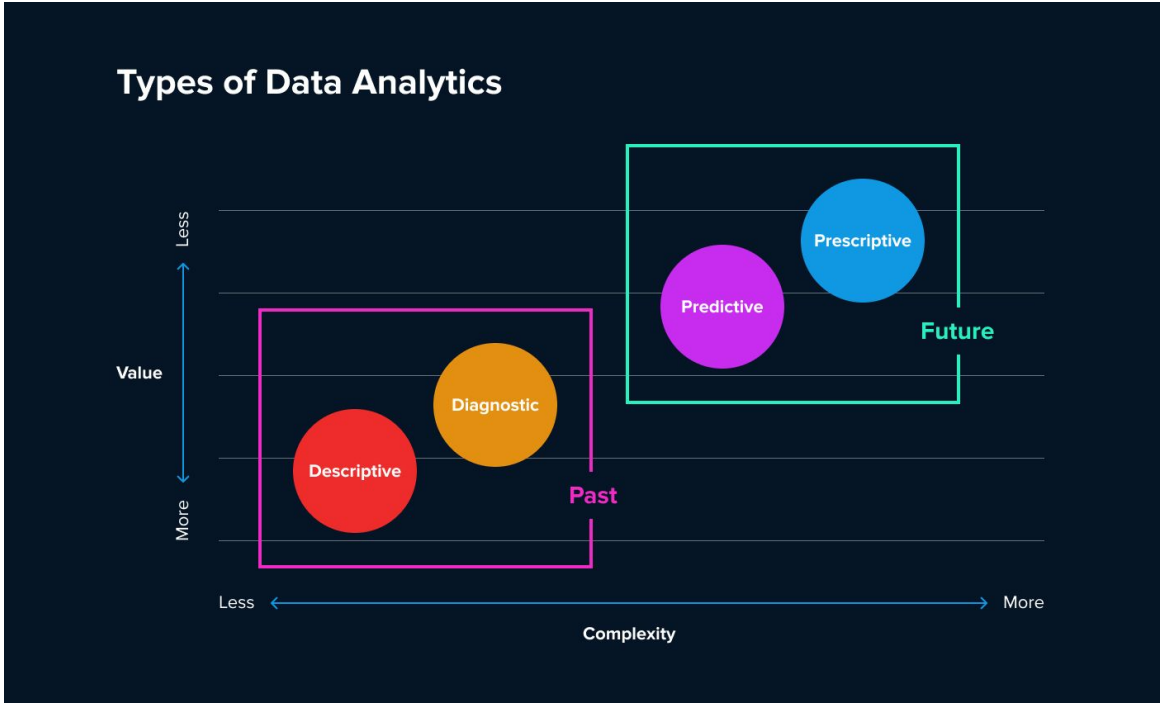
Types of Analysis

Descriptive
Exploratory
Inferential
Causal
Predictive
Mechanistic



What is the total number of covid patients?

Complexity and Types of Analytics



Data Science

Scientific Method

Formulate hypothesis

A question based usually on **observations** that you want to prove is true

Collect data

Collect **data needed** through various means that would help in proving (or disproving) the hypothesis

Test hypothesis

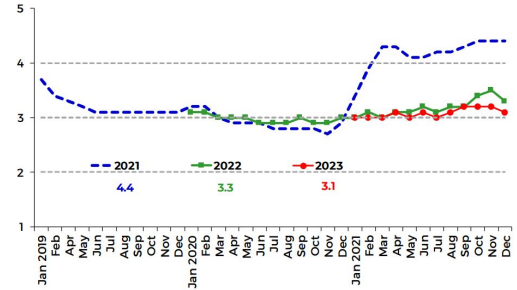
From the data collected, perform analyses and run statistical tests to determine whether or not the hypothesis is **statistically significant** or not

Scientific Method

Formulate hypothesis

Economics
What will happen to inflation in the next quarter?

Chart 3. BSP Private Sector Economists' Survey* (2012=100)
mean forecast for full year; in percent



Collect data

CPI across months and years

Inputs from market xperts (trend insight)

Test hypothesis

Building models

Simple model : Moving average

Advanced model : ARIMA / Neural networks (LSTMs)

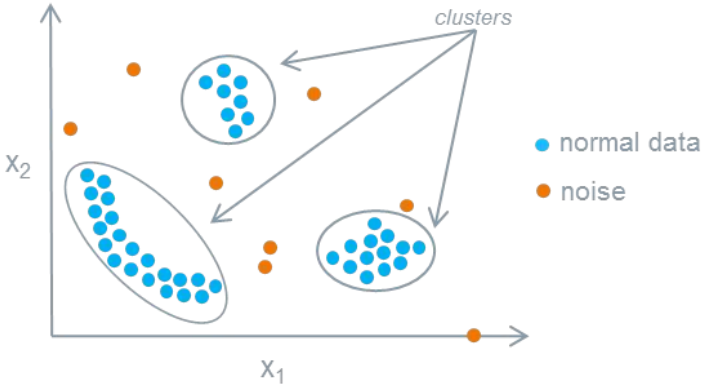
Data Science in Practice

Data Science Pipeline

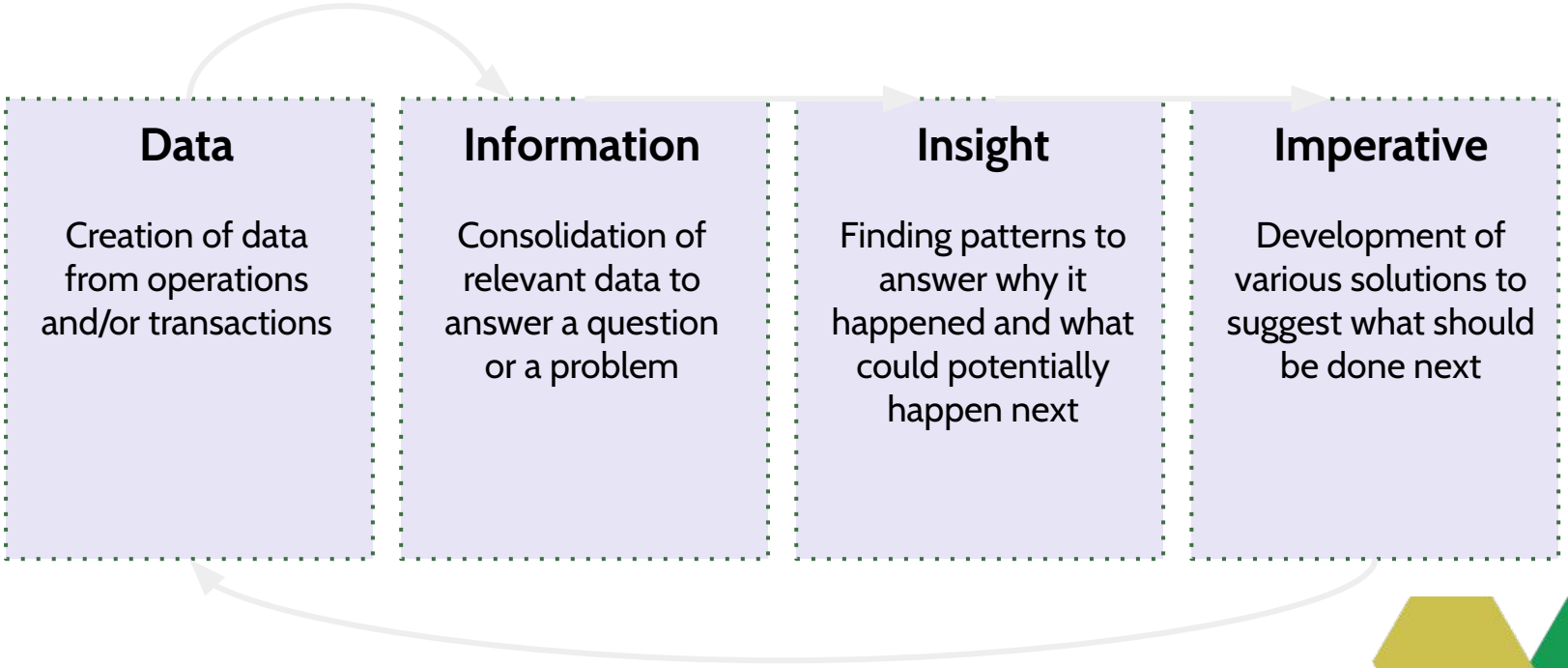
Question forumulation	Data acquisition	Exploratory data analysis	Modeling	Validation & Interpretation
Have a **problem** you want to be solved using **data**	**Retrieve** the data through various sources	**Explore** the datasets **Visualize** to see the patterns **Validate** patterns & importance using statistical tests	**Train** statistical models using (partial) data to predict, classify, etc.	**Validate** models using (remaining) data and measure performance **Organize** results for storytelling

Data Science Pipeline

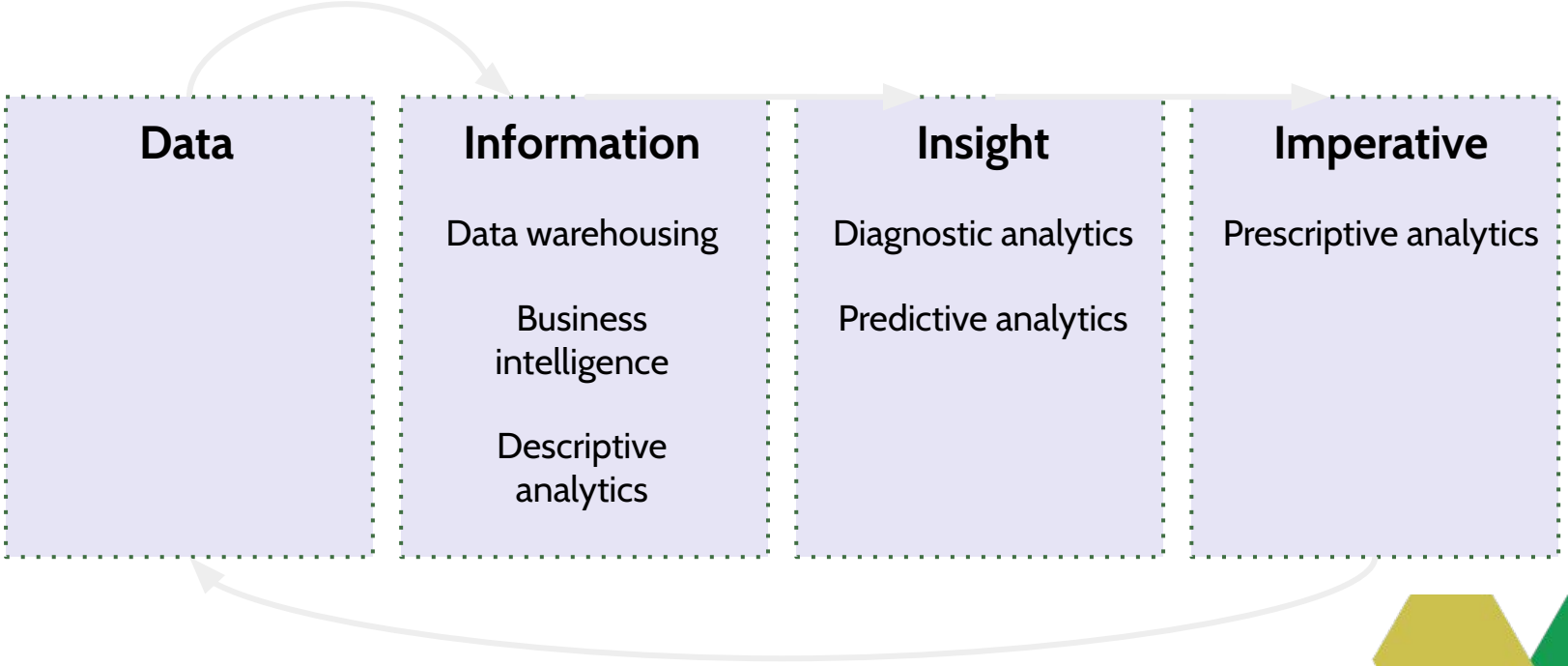
Question forumulation	Data acquisition	Exploratory data analysis	Modeling	Validation & Interpretation
Is anyone fraudulent?	Get transaction of patients and doctors	Visualize, check out pattern	Simple model : Using mean and median Advanced model : Isolation forest	Checking results with HR, Marketing and Operations



General Analytics Pipeline



General Analytics Pipeline



CRISP-ML(Q)

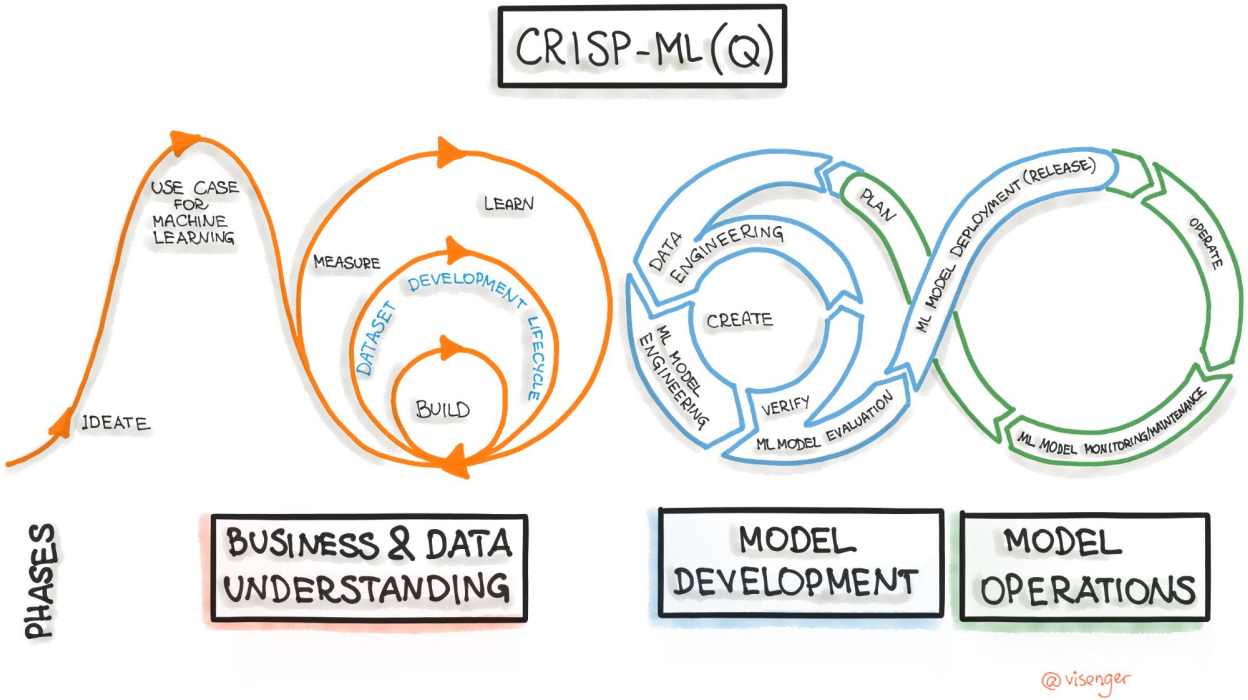


Image from MLOps

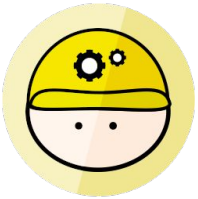
Roles

Job Roles

Data

Data Steward

Information



Data Engineer

Insight

Data Scientist

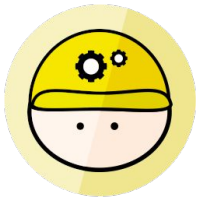
Imperative

Functional Analyst

Analytics Manager

Data Steward

data governance
data privacy
data security
data quality



Data Engineer

data infrastructure
data ETL
(extract, transform, load)
data management
data warehousing

Data Scientist

data analytics
data mining
machine learning
statistical modeling

AAP ANALYTICS
Association of the Philippines

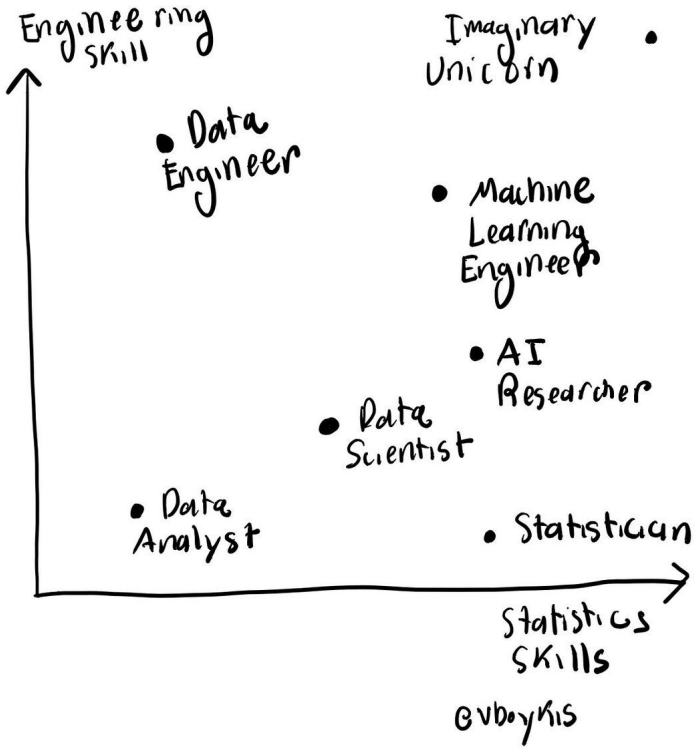
Functional Analyst

domain expert
process/policy expert
data visualization engineer

Analytics Manager

project manager

AAP ANALYTICS
Association of the Philippines



Data Science Roles Scatter Plot (Vicki Boykis, April 2018)

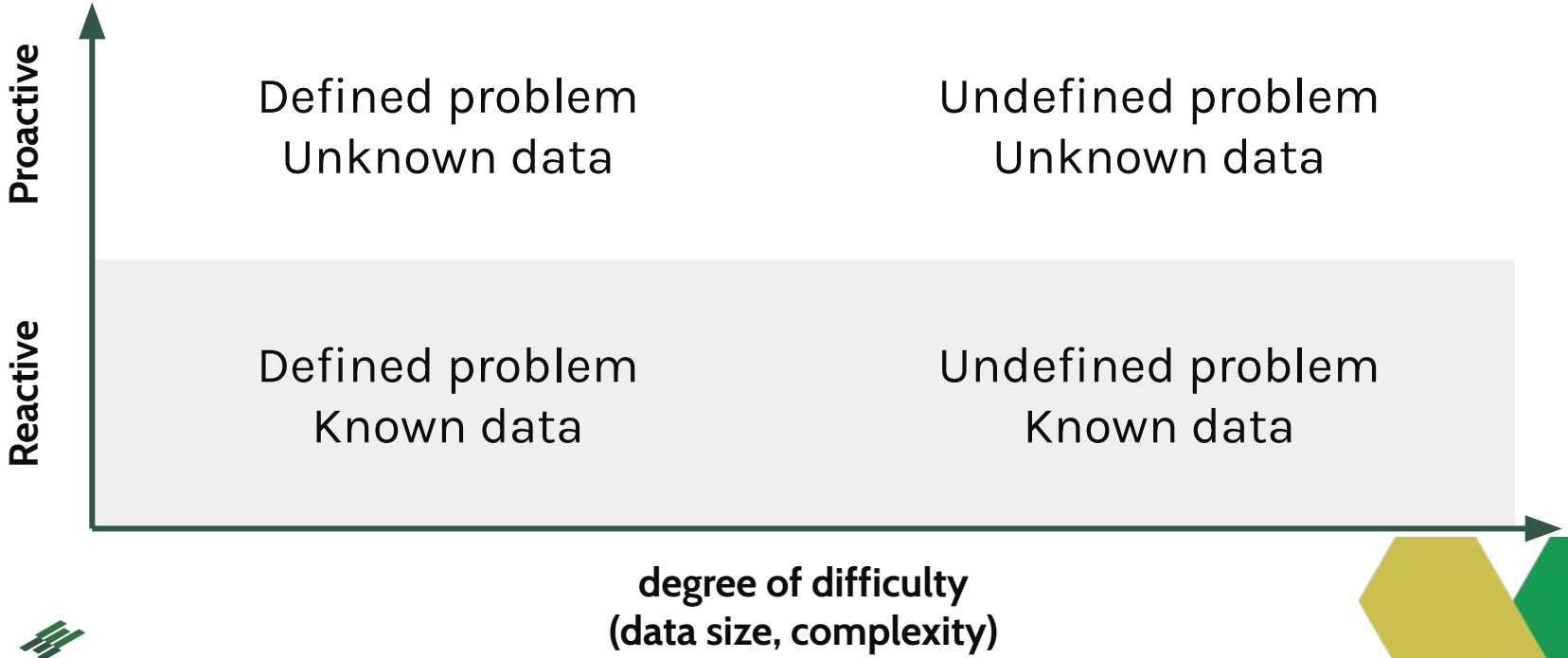
https://twitter.com/vboykis/status/983391619062919170

Which comes first, data or problem?

Data Science Workflow

Problem	Start with a question that you want to answer or a problem you want to solve
Data	Maximize the potential of available data from various sources or collect your own
Explore	Understand the information embedded within the data
Learn	Use statistical models, machine learning and other data mining techniques to find patterns faster
Present	Effectively visualize and communicate the insights and knowledge derived from the data

Type of Data Science Projects



Open Data

Open Data

Different entities now publish openly available data for opportunistic analysis

Dr. Andrew L. Tan Data Science Institute

Open Data Principles

- Open by Default
- Timely & Comprehensive
- Accessible & Usable
- Comparable & Interoperable
- For Improved Governance & Citizen Engagement
- For Inclusive Development & Innovation

Open Data Principles

Open by Default

-For all

-No privacy implications

Timely and Comprenehsive

-Relevant and complete

-Raw and unmodified

Accessible and Usable

-Machine Readable (JSON, CSV, Parquet)

-Free of Charge

Dr. Andrew L. Tan Data Science Institute

https://opendatacharter.net/principles/

Open Data Principles

Comparable and Interoperable

-Follows data standards

-Integration is easy

Improved Governance and Citizen Engagement

-Better sense of whats happening

-"Transparency"

Inclusive Development and Innovation

-Discover new things and ideas

Dr. Andrew L. Tan Data Science Institute

https://opendatacharter.net/principles/

Questions?