Introduction to Data Science

Data Science Fundamental

Ratthaprom PROMKAM, Dr.rer.nat.

Department of Mathematics and Computer Science,
RMUTT

Learning Objectives

Meet people who work in data science

Explore definitions of data science

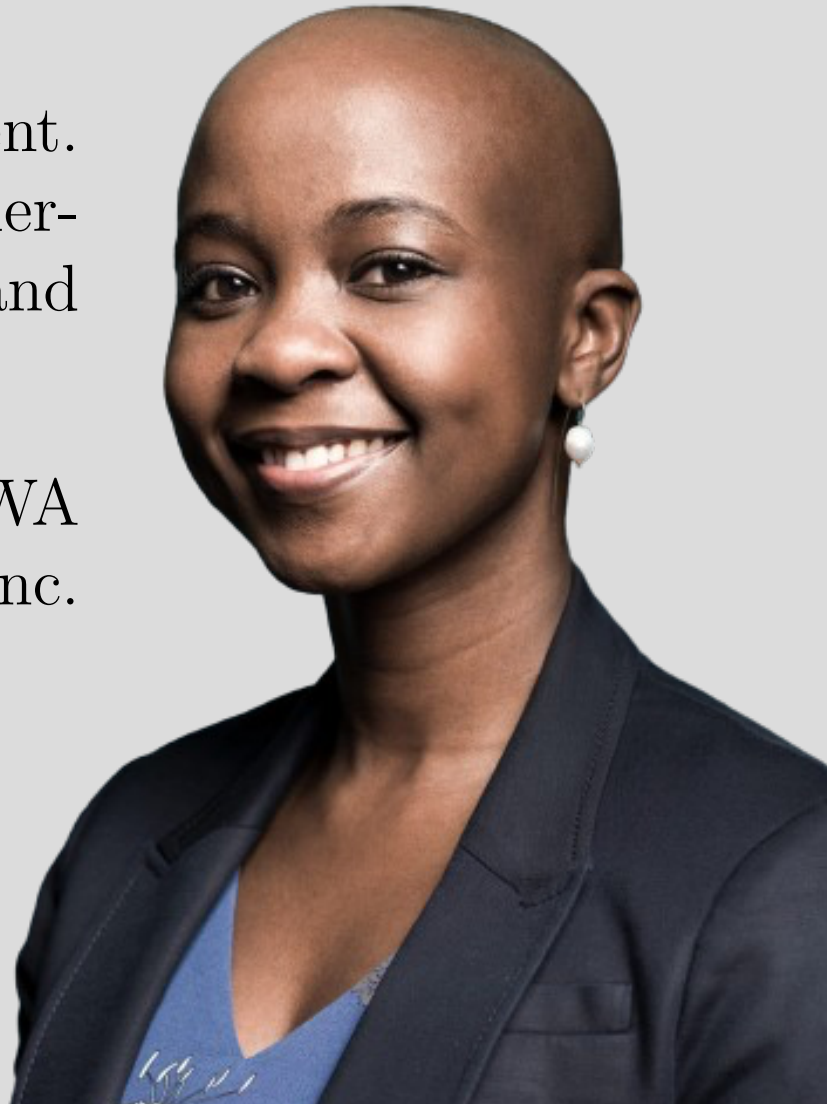
Learn about data science in a business context

Discover some use cases for data science

What is data science?

"Data science is a process, not an event. It is the process of using data to understand different things, to understand the world."

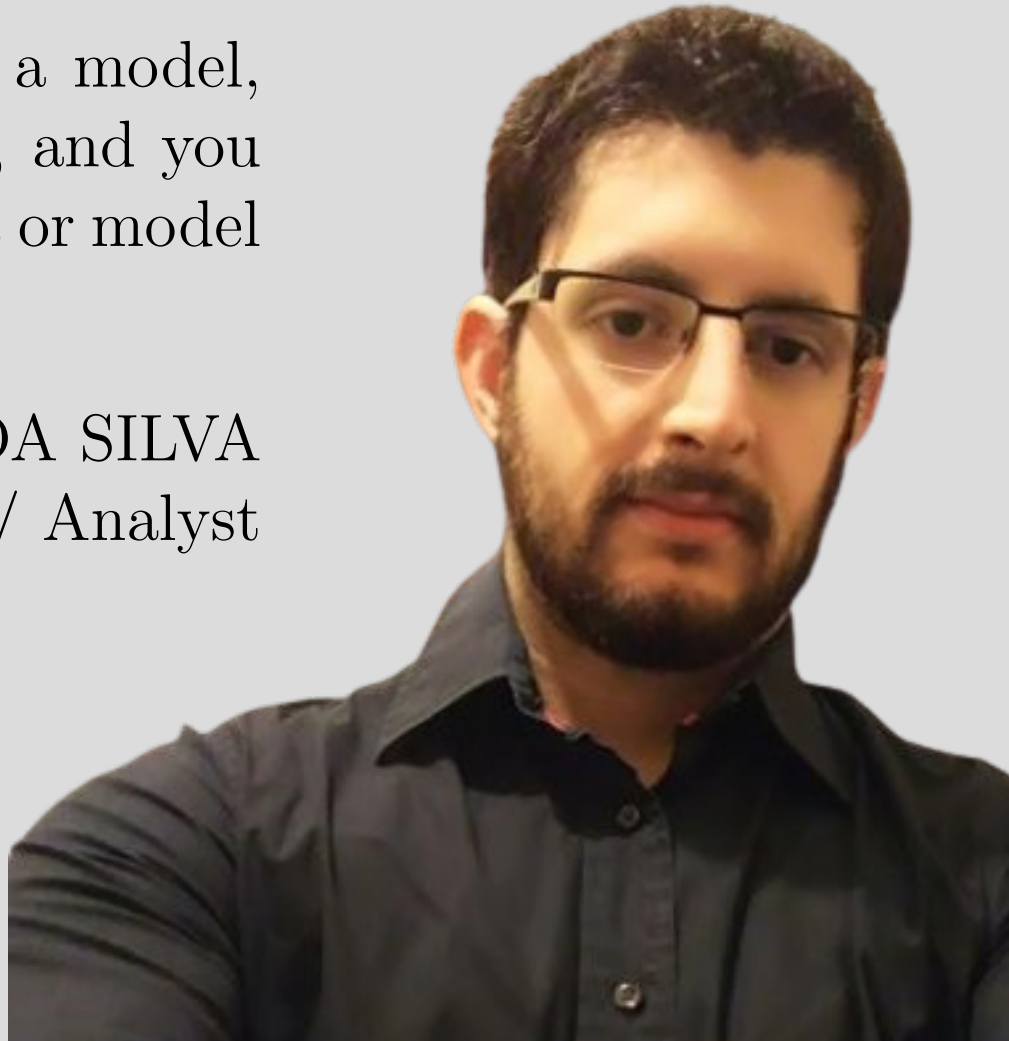
— Shingai MANJENGWA
CEO of Fireside Analytics Inc.



What is data science?

"For me, it's when you have a model, or a hypothesis of a problem, and you try to validate that hypothesis or model with your data."

— Rafael B. DA SILVA
Data Scientist/ Engineer/ Analyst



What is data science?

"Data science is the art of uncovering the insights and trends that are hiding behind data."

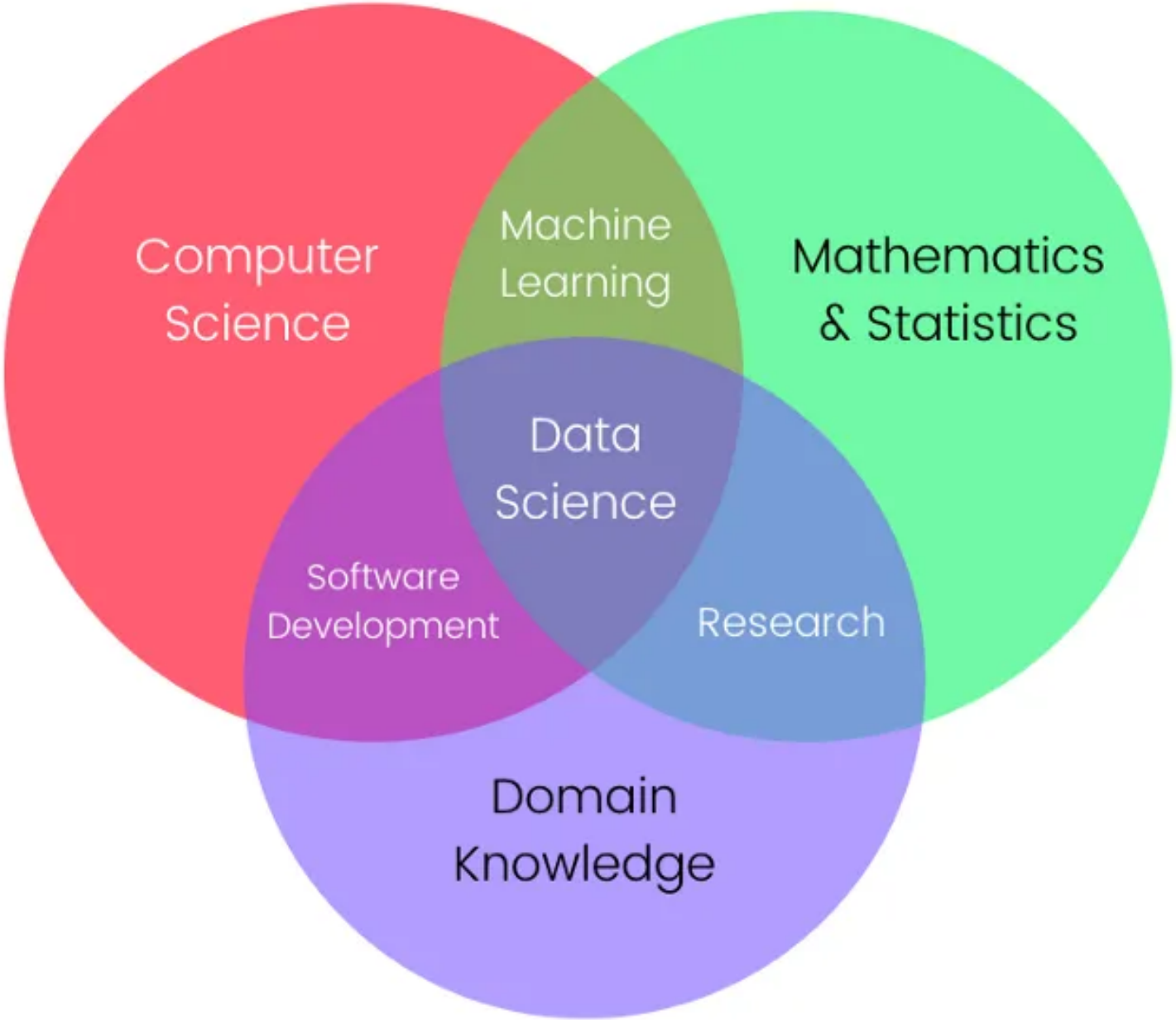
— Diana ZARATE-DIAZ
COO of Sahar Global Summits, Llc.

What is data science?

"I see data science as one's attempt to work with data to find answers to questions that they are exploring.
If you have data, curiosity, and you're going through analyzing data, trying to get some answers from it, is data science. "

— Murtaza HAIDER
Prof. at Toronto Metro. University

What is data science?



Data Scientist: The Sexiest Job in 21st Century

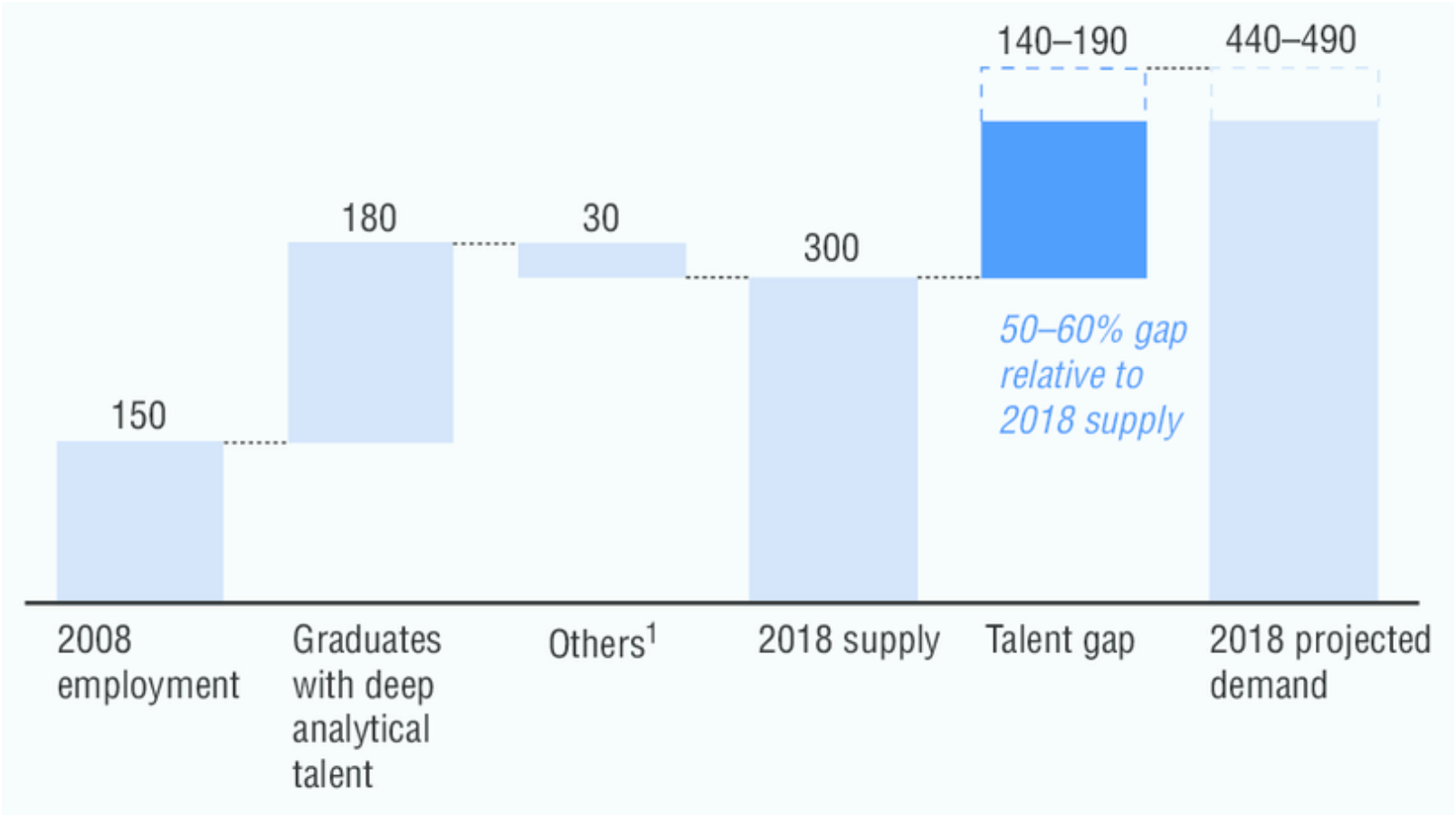


Figure: Demand for data analytics expertise according to the McKinsey Global Institute.

Data Scientist: The Sexiest Job in 21st Century

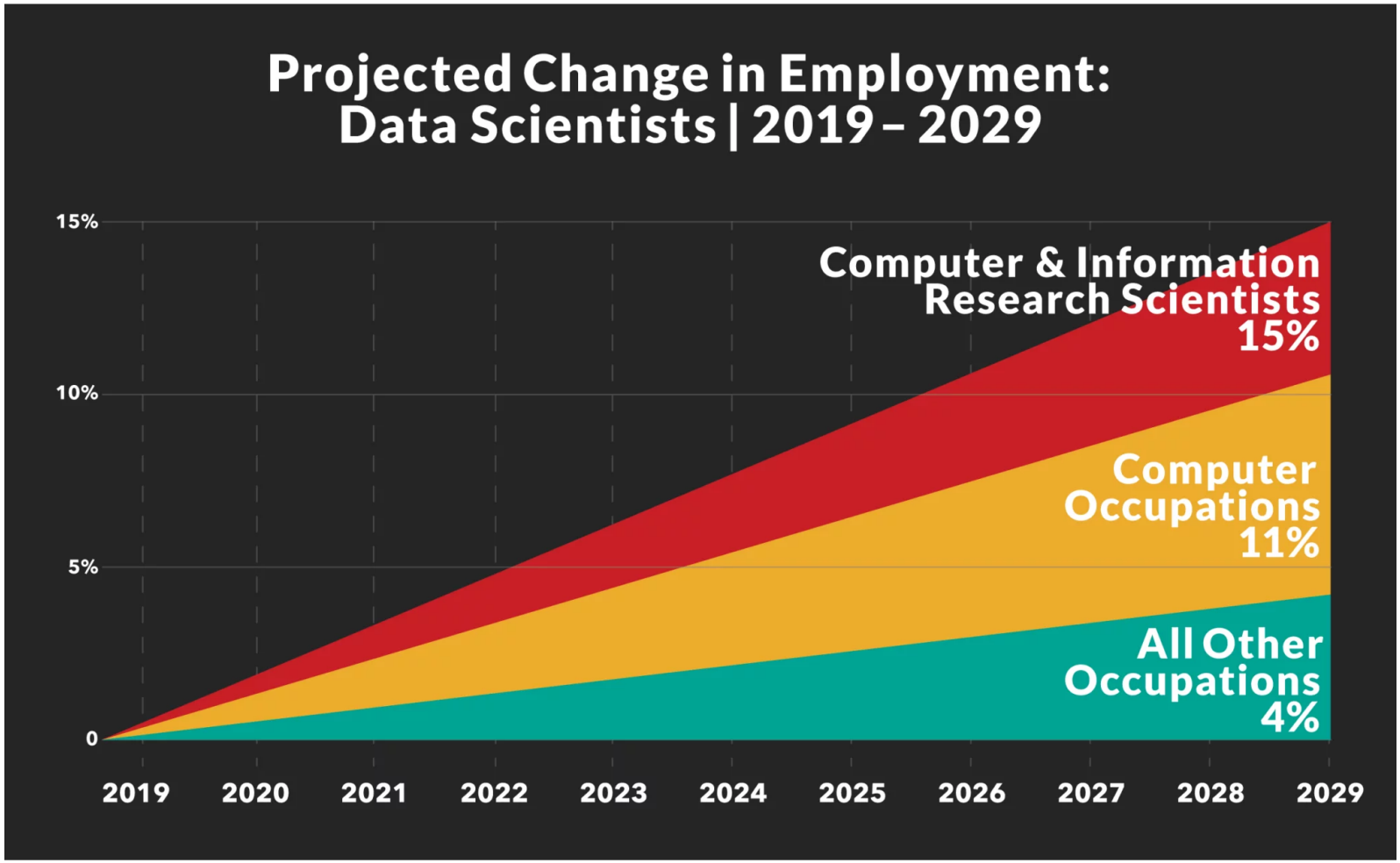


Figure: United States Bureau of Labor Statistics (BLS) Occupational Outlook Handbook

Review Questions:

In the report by the McKinsey Global Institute, by 2018, it is projected that there will be a shortage of people with deep analytical skills in the United States. What is the size of this shortage?

- O 20,000 - 50,000
- O 120,000
- O 140,000 - 190,000
- O 800,000 - 900,000
- O 3,000,000 - 6,000,000

Review Questions:

What has changed from the past to make Data science an in-demand occupation?

- There is now a lack of data
- Laws have changed
- Vast amount of data date being created
- The advent of the free market

Review Questions:

What is the minimum education requirement to become a data scientist?

- You must have a Degree in Computer Science
- You must have a masters degree in Statistics
- You must have a Ph.D. in Machine learning
- The above are all helpful, but they are not necessary to become a data scientist, education backgrounds of data scientists vary

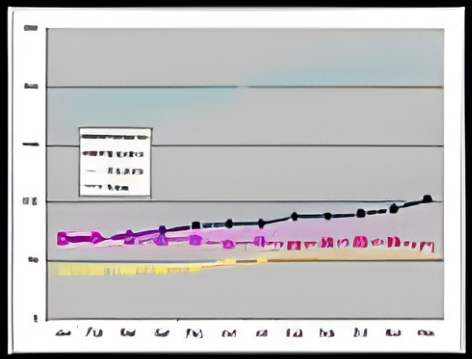
A day in the life of a data science person

Data Scientist

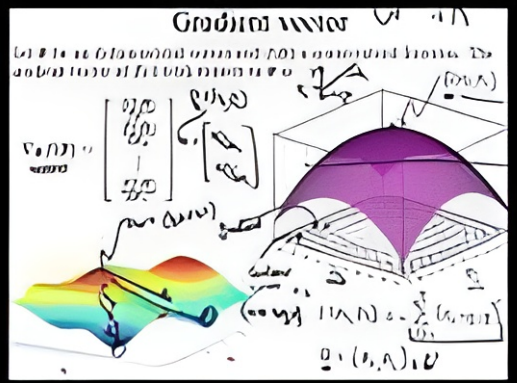
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

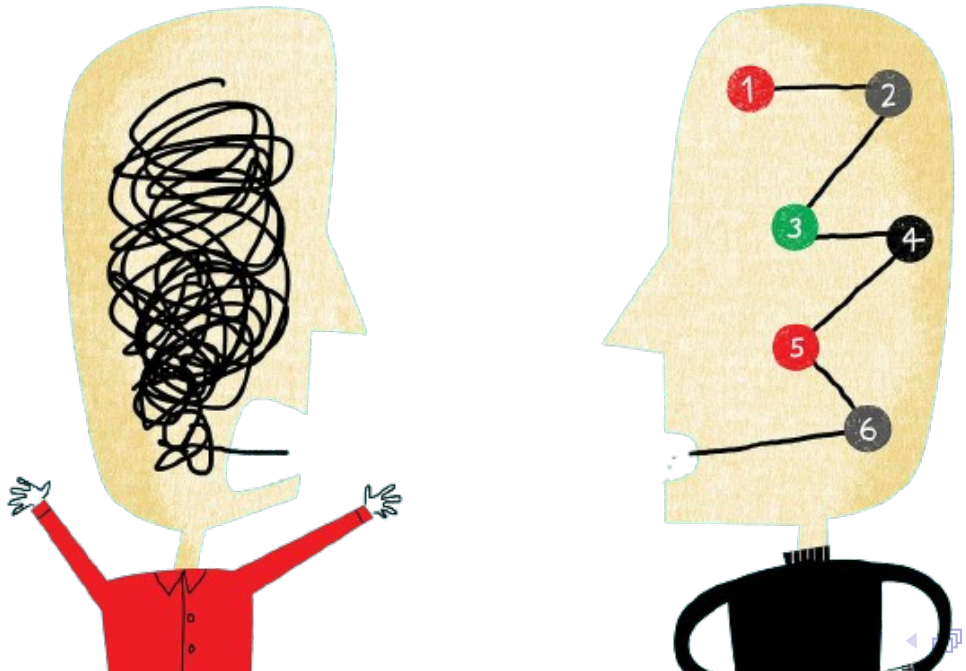
What I actually do

Case: Toronto Transit Commissions

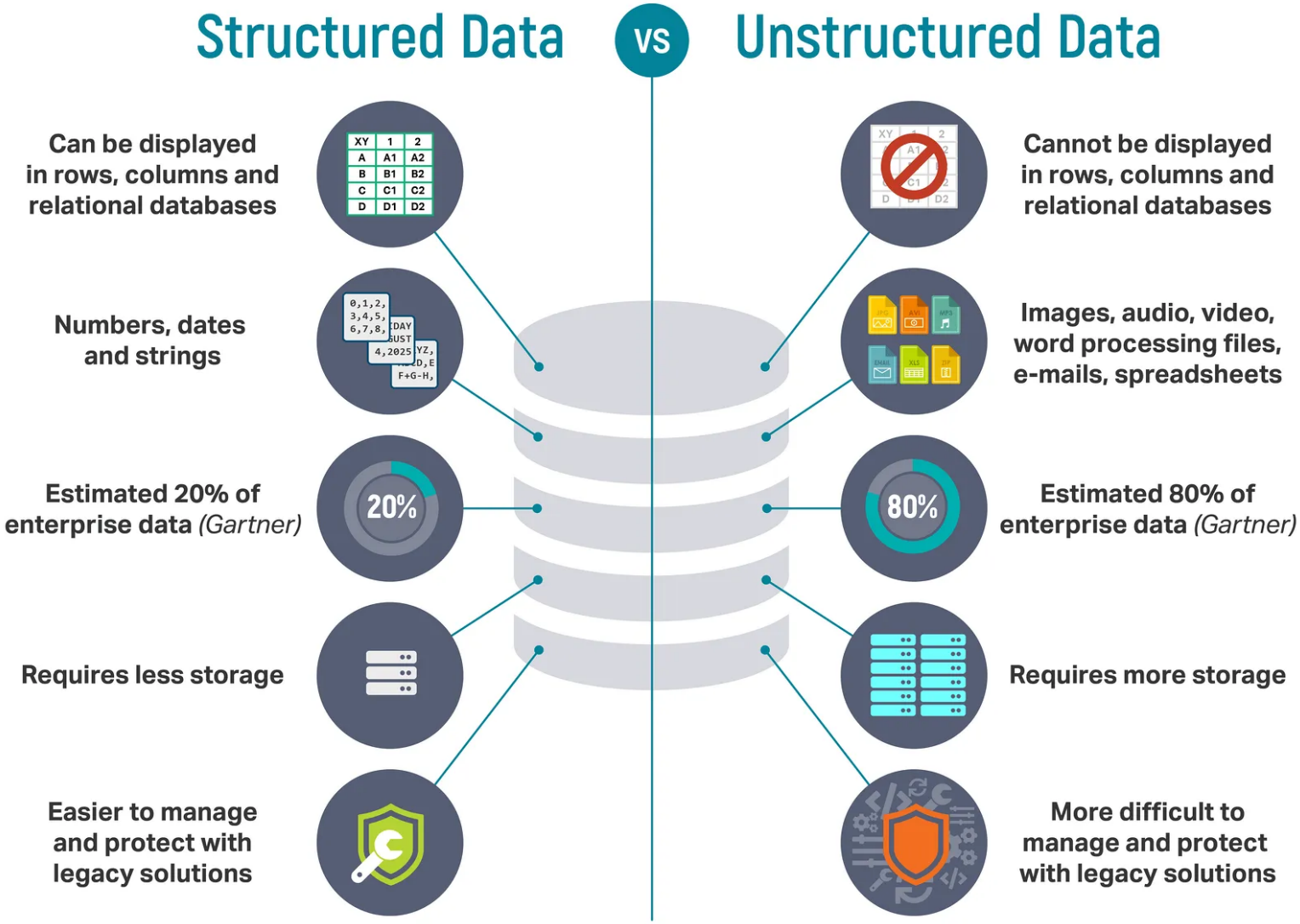
Case: Toronto Transit Commissions

There are **500,000** complaints!

Dates, Names, Types, Status,
Emails and Faxes.



Type of Data


Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data (Gartner)
20%
80%
Estimated 80% of enterprise data (Gartner)
Requires less storage
Requires more storage
Easier to manage and protect with legacy solutions
More difficult to manage and protect with legacy solutions

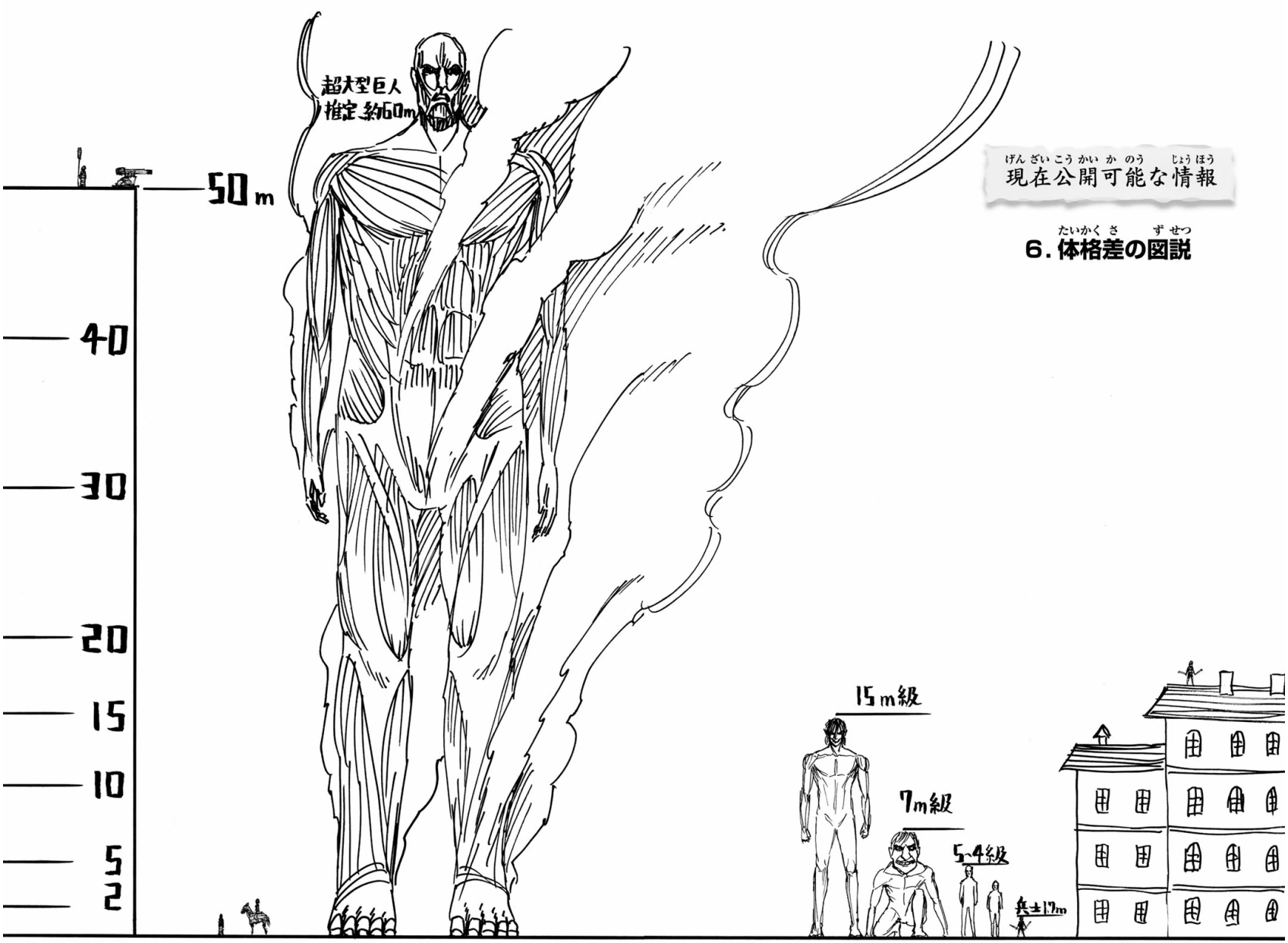
Data science tools and technology

Data science tools and technology

Data science tools and technology

Regression?

Regression?

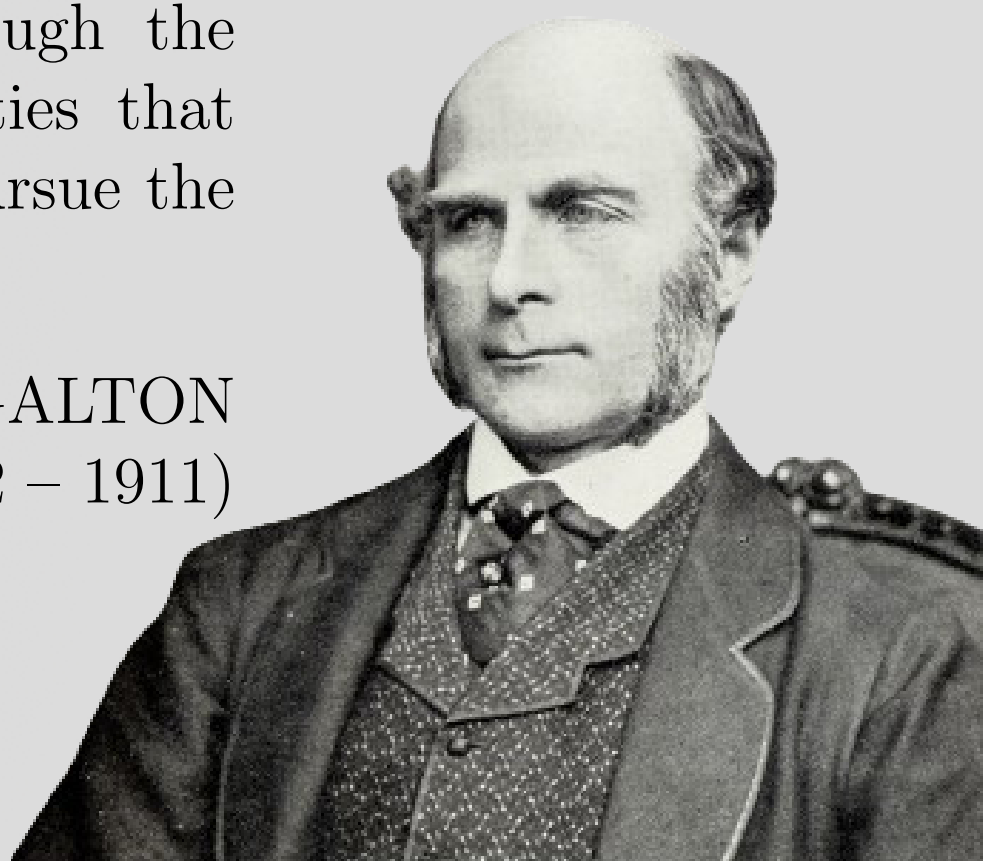

超大型巨人
推定 約60m
50m
40
30
20
15
10
5
2
現在公開可能な情報
6. 体格差の図説
15m級
7m級
5-4級
兵士17m

Regession toward the mean

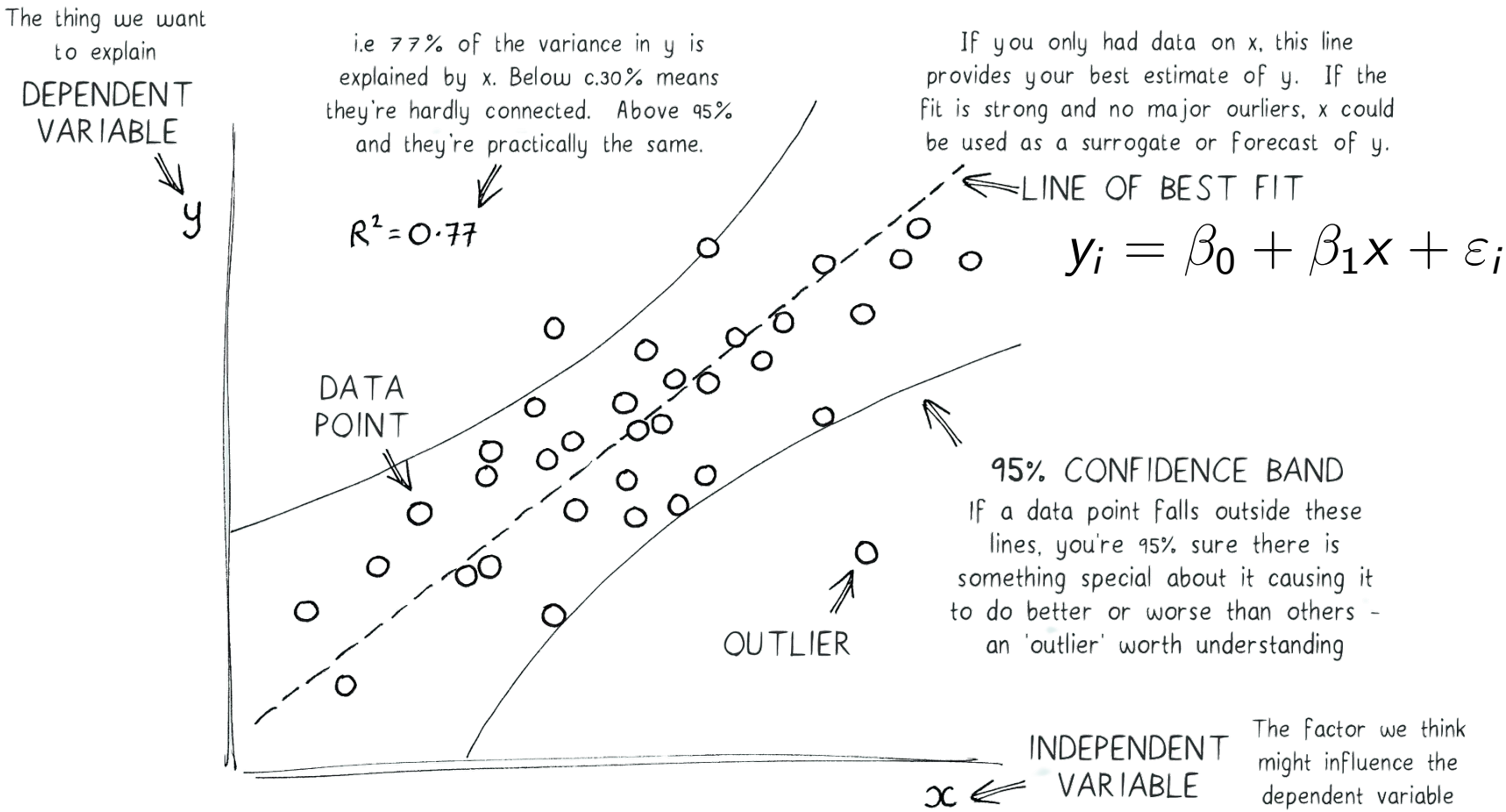
"Whenever you can, count. "

"Statistics are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man. "

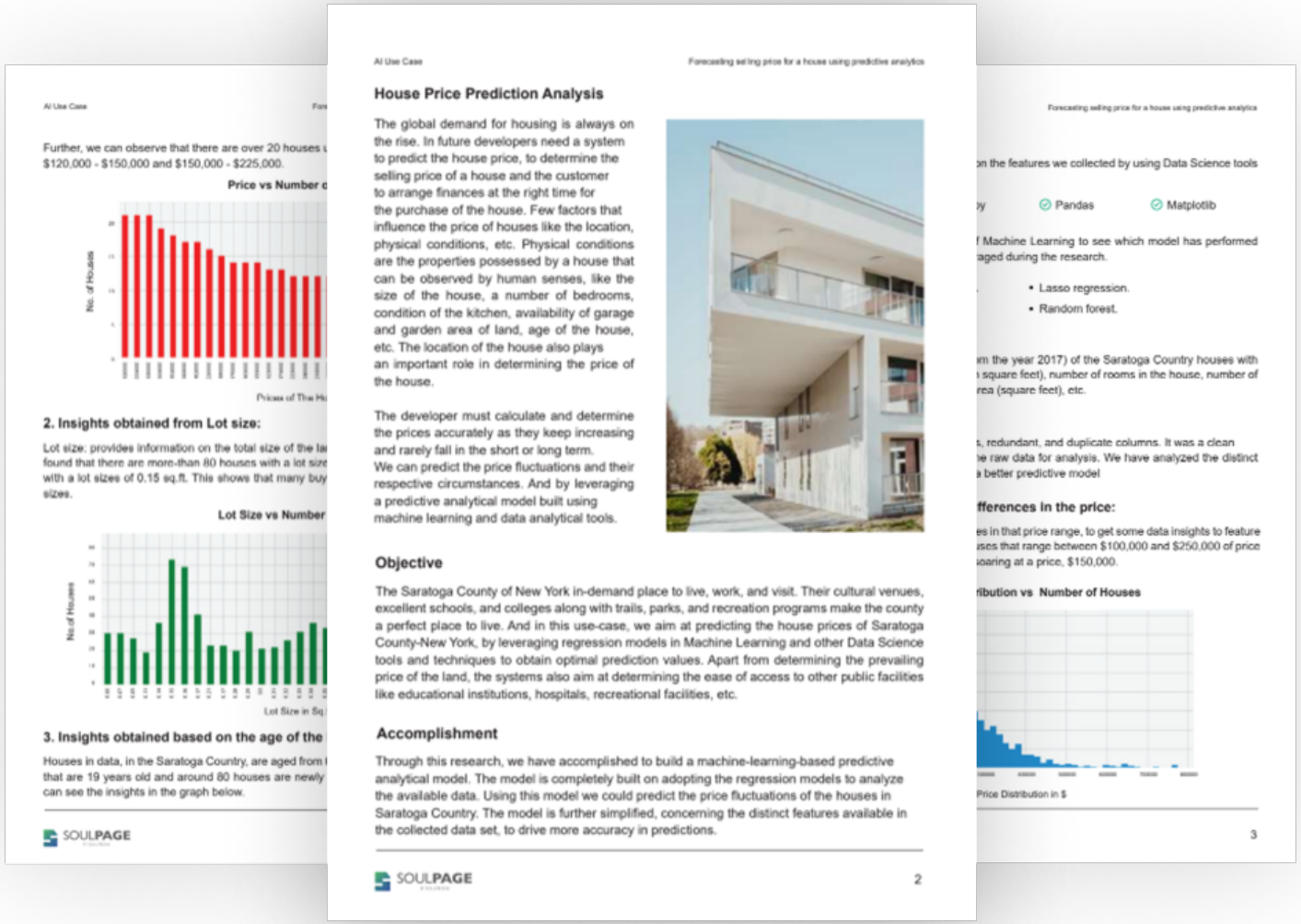
— Sir Francis GALTON
(1812 – 1911)



Regression toward the mean



Case: House Price Prediction



AI Use Case

Forecasting selling price for a house using predictive analytics

House Price Prediction Analysis

The global demand for housing is always on the rise. In future developers need a system to predict the house price, to determine the selling price of a house and the customer to arrange finances at the right time for the purchase of the house. Few factors that influence the price of houses like the location, physical conditions, etc. Physical conditions are the properties possessed by a house that can be observed by human senses, like the size of the house, a number of bedrooms, condition of the kitchen, availability of garage and garden area of land, age of the house, etc. The location of the house also plays an important role in determining the price of the house.

The developer must calculate and determine the prices accurately as they keep increasing and rarely fall in the short or long term.
We can predict the price fluctuations and their respective circumstances. And by leveraging a predictive analytical model built using machine learning and data analytical tools.

Objective

The Saratoga County of New York in-demand place to live, work, and visit. Their cultural venues, excellent schools, and colleges along with trails, parks, and recreation programs make the county a perfect place to live. And in this use-case, we aim at predicting the house prices of Saratoga County-New York, by leveraging regression models in Machine Learning and other Data Science tools and techniques to obtain optimal prediction values. Apart from determining the prevailing price of the land, the systems also aim at determining the ease of access to other public facilities like educational institutions, hospitals, recreational facilities, etc.

Accomplishment

Through this research, we have accomplished to build a machine-learning-based predictive analytical model. The model is completely built on adopting the regression models to analyze the available data. Using this model we could predict the price fluctuations of the houses in Saratoga Country. The model is further simplified, concerning the distinct features available in the collected data set, to drive more accuracy in predictions.

SOULPAGE

2

Further, we can observe that there are over 20 houses
$120,000 - $150,000 and $150,000 - $225,000.

Price vs Number

No. of Houses

Prices of The Ho

2. Insights obtained from Lot size:

Lot size: provides information on the total size of the la
found that there are more-than 80 houses with a lot size
with a lot sizes of 0.15 sq.ft. This shows that many buy
sizes.

Lot Size vs Number

No.of Houses

Lot Size in Sq

3. Insights obtained based on the age of the

Houses in data, in the Saratoga Country, are aged from
that are 19 years old and around 80 houses are newly
can see the insights in the graph below.

SOULPAGE

Forecasting selling price for a house using predictive analytics

on the features we collected by using Data Science tools

Pandas

Matplotlib

f Machine Learning to see which model has performed
aged during the research.

- Lasso regression.
- Random forest.

m the year 2017) of the Saratoga Country houses with
square feet), number of rooms in the house, number of
rea (square feet), etc.

s, redundant, and duplicate columns. It was a clean
e raw data for analysis. We have analyzed the distinct
a better predictive model

fferences in the price:

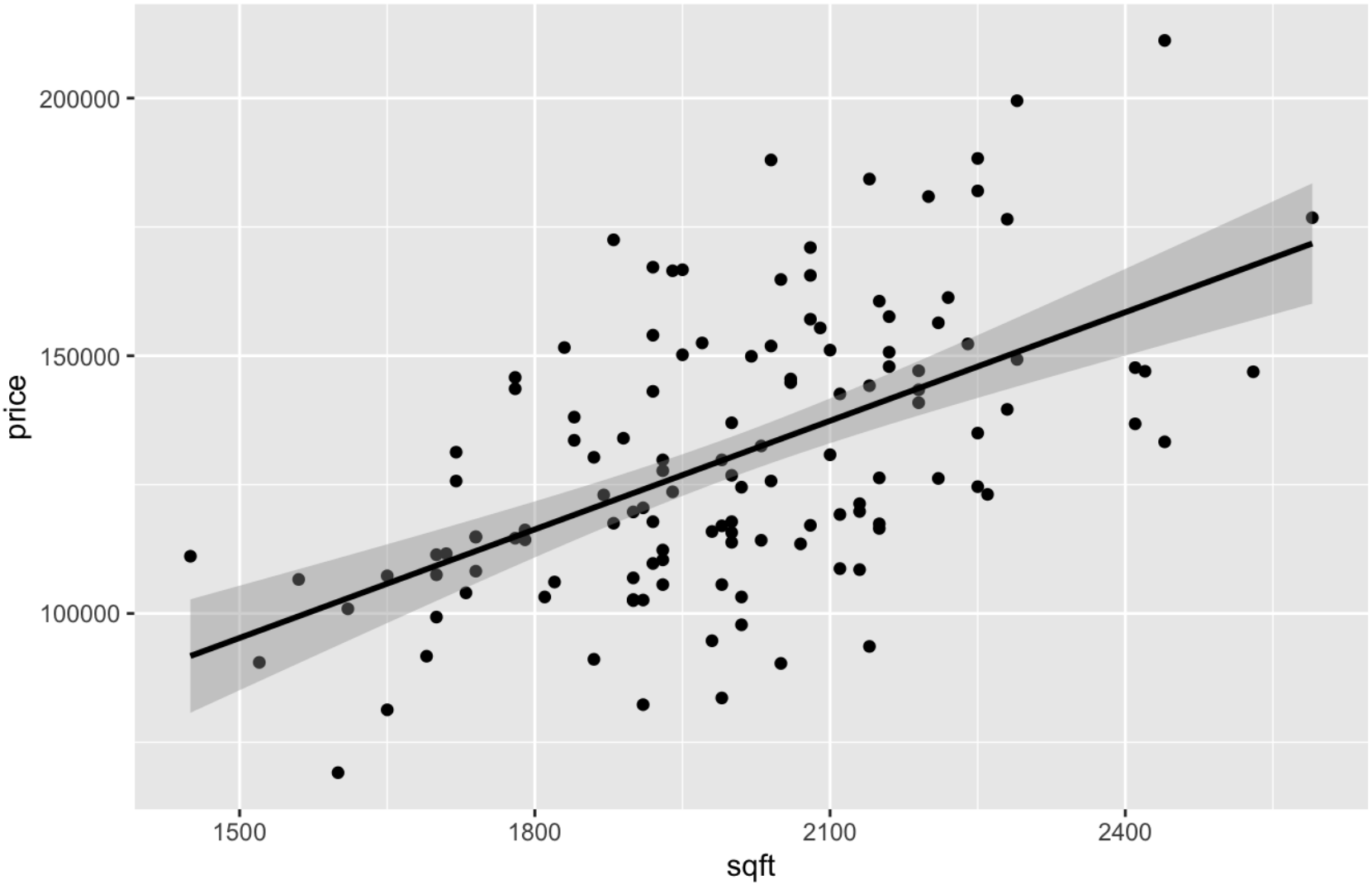
es in that price range, to get some data insights to feature
ses that range between $100,000 and $250,000 of price
oaring at a price, $150,000.

ribution vs Number of Houses

Price Distribution in $

3

Case: House Price Prediction



Review Questions

What is structured data

- Data that can be stored in a database or some tabular form
- Images and video
- Segments of text
- Audio data

Review Questions

What does the following formula represent?:

$$\mathbf{Base\_fair} + \mathbf{Taxi\_rate} \times (\mathbf{Time})$$

- The possible formula used in regression analysis to determine the cost of a taxi ride
- The formula used to build a recommender system for rating a cab service
- A possible formula used in regression analysis used to determine the price of a house
- What is the impact of lot size on housing price?

Review Questions

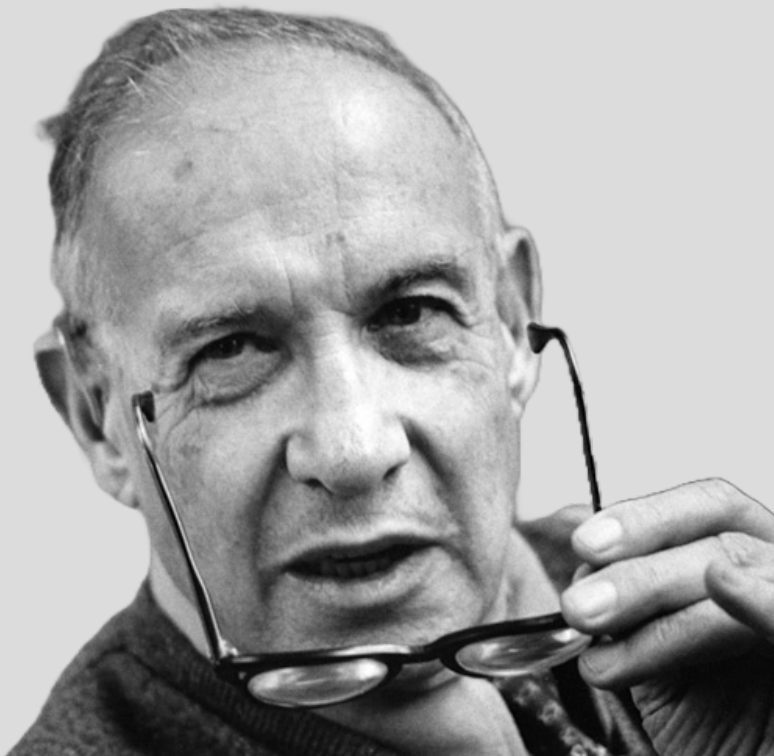
What is an example of a question that can be put to a regression analysis?

- Do homes with brick exterior sell in rural areas?
- What is the impact of lot size on housing price?
- What are typical land taxes in a house sale?
- How much does a finished basement cost?
- How much should a house near a park cost?

How should companies get started in Data Science

"If you can't measure it,
you can't manage it."

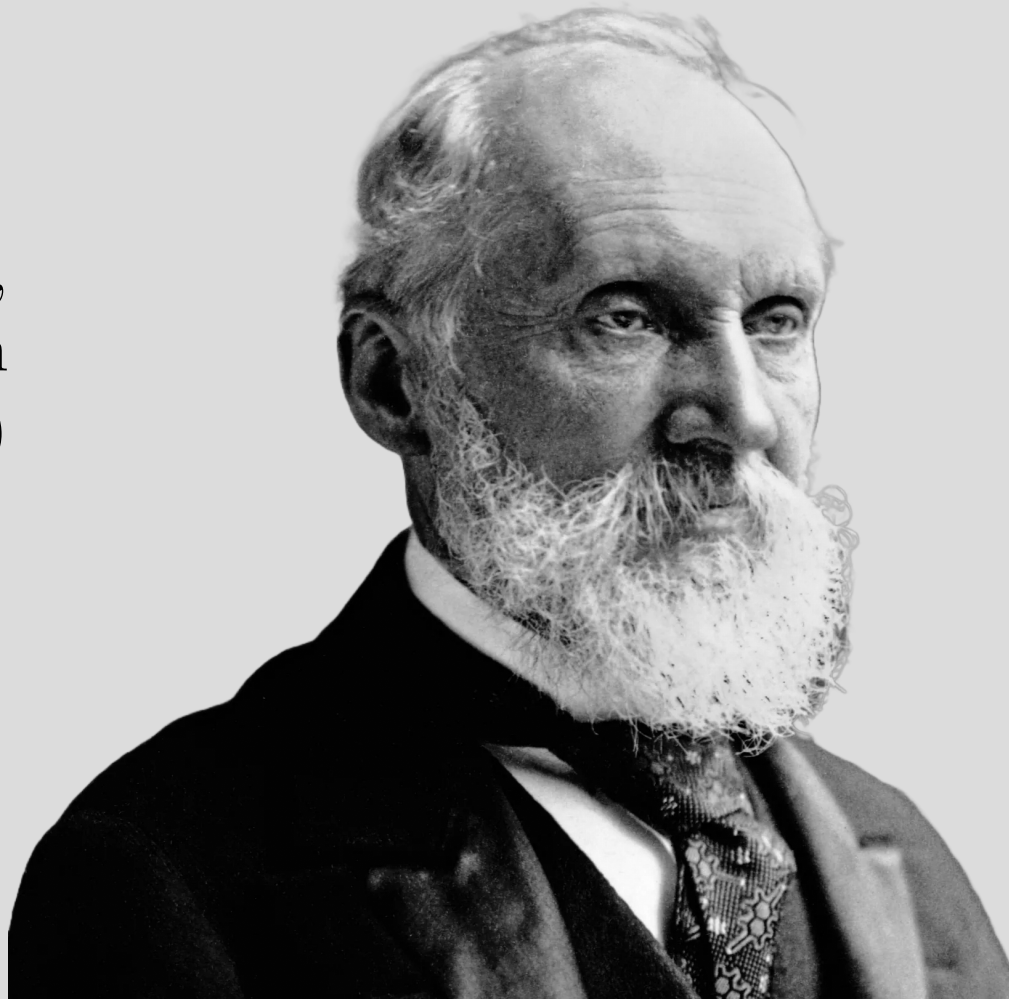
— Peter Drucker
(1909-2005)



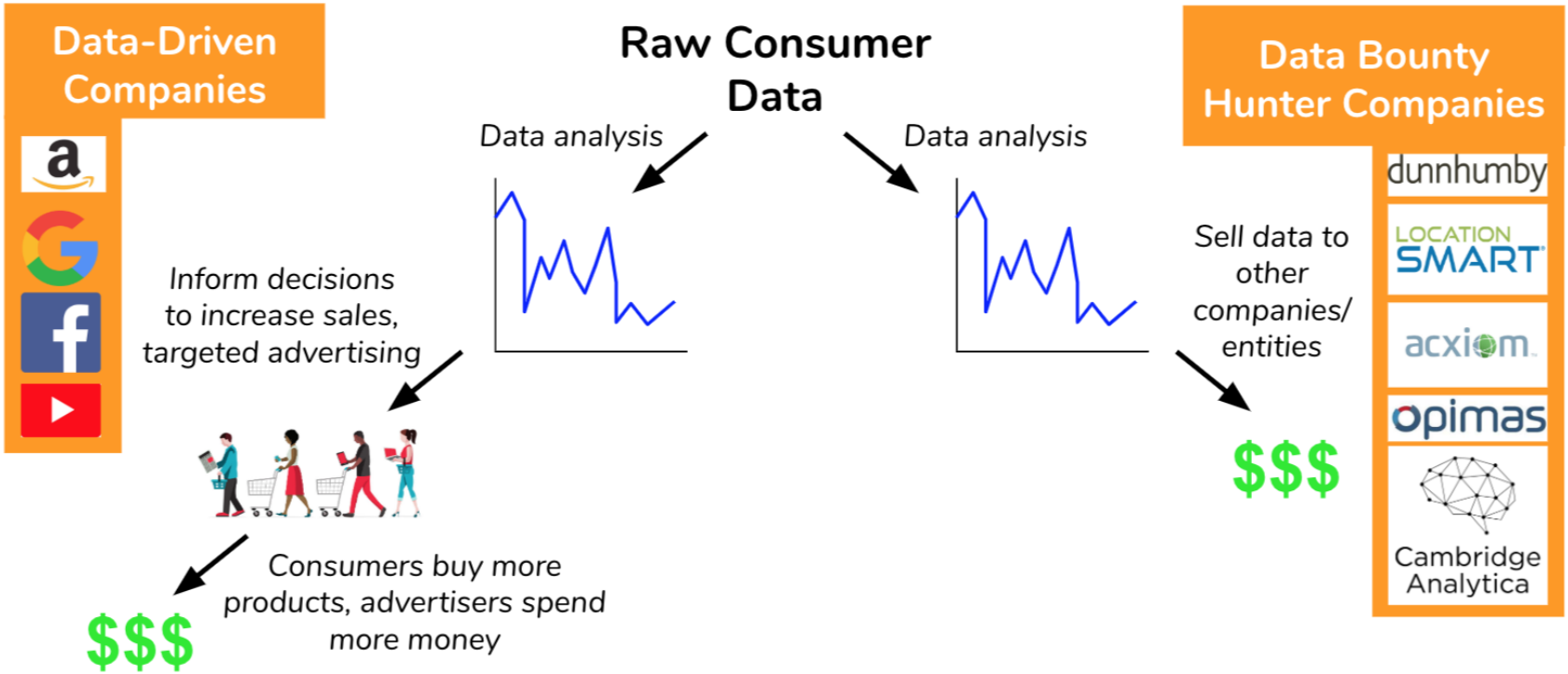
How should companies get started in data science

"If you can't measure it,
you can't improve it"

— William Thomson,
The Lord Kelvin
(1824–1907)

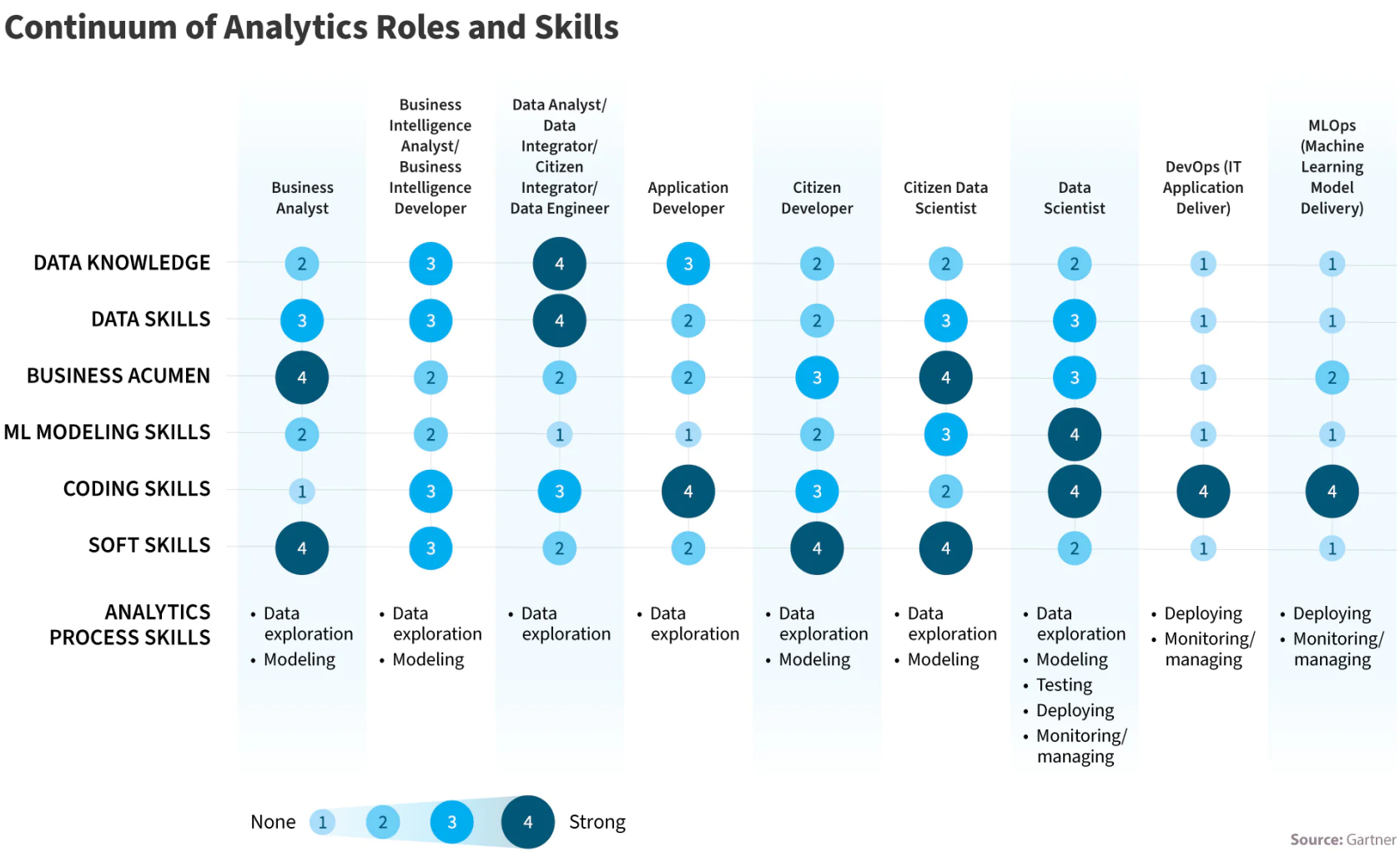


How should companies get started in data science

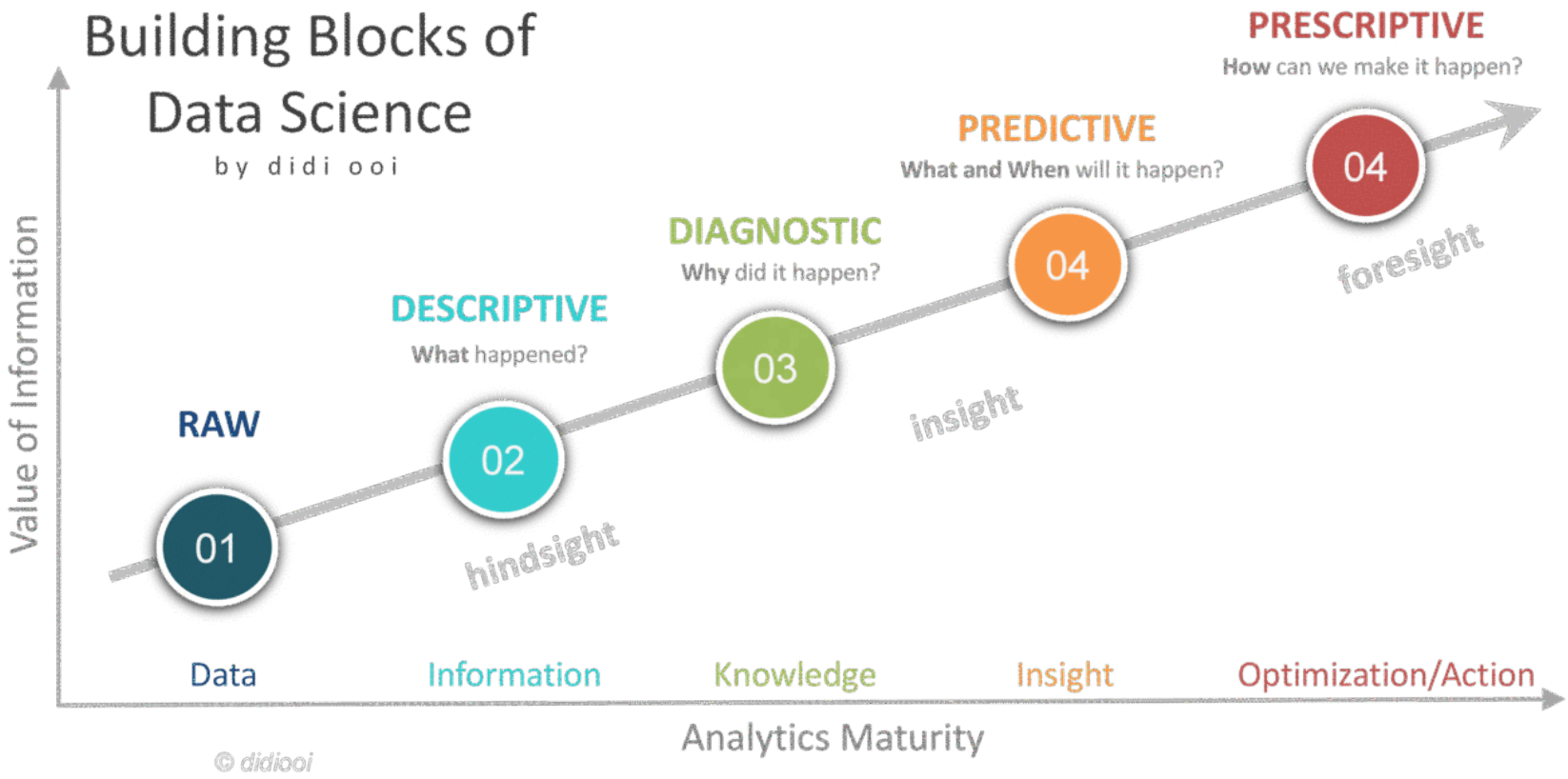


Never delete data, it never get old.

Recruiting for data science



The final deliverable



Review Questions

Complete the following sentence that best explains why business needs to capture data:

At the end of the day, for businesses, they know one thing, that if they are unable to measure something, ...

- O they are unable to graph it
- O they are unable to improve it
- O they are unable to show compliance with tax laws
- O they are unable to facilitate meetings between sales and marketing

Review Questions

A business should never:

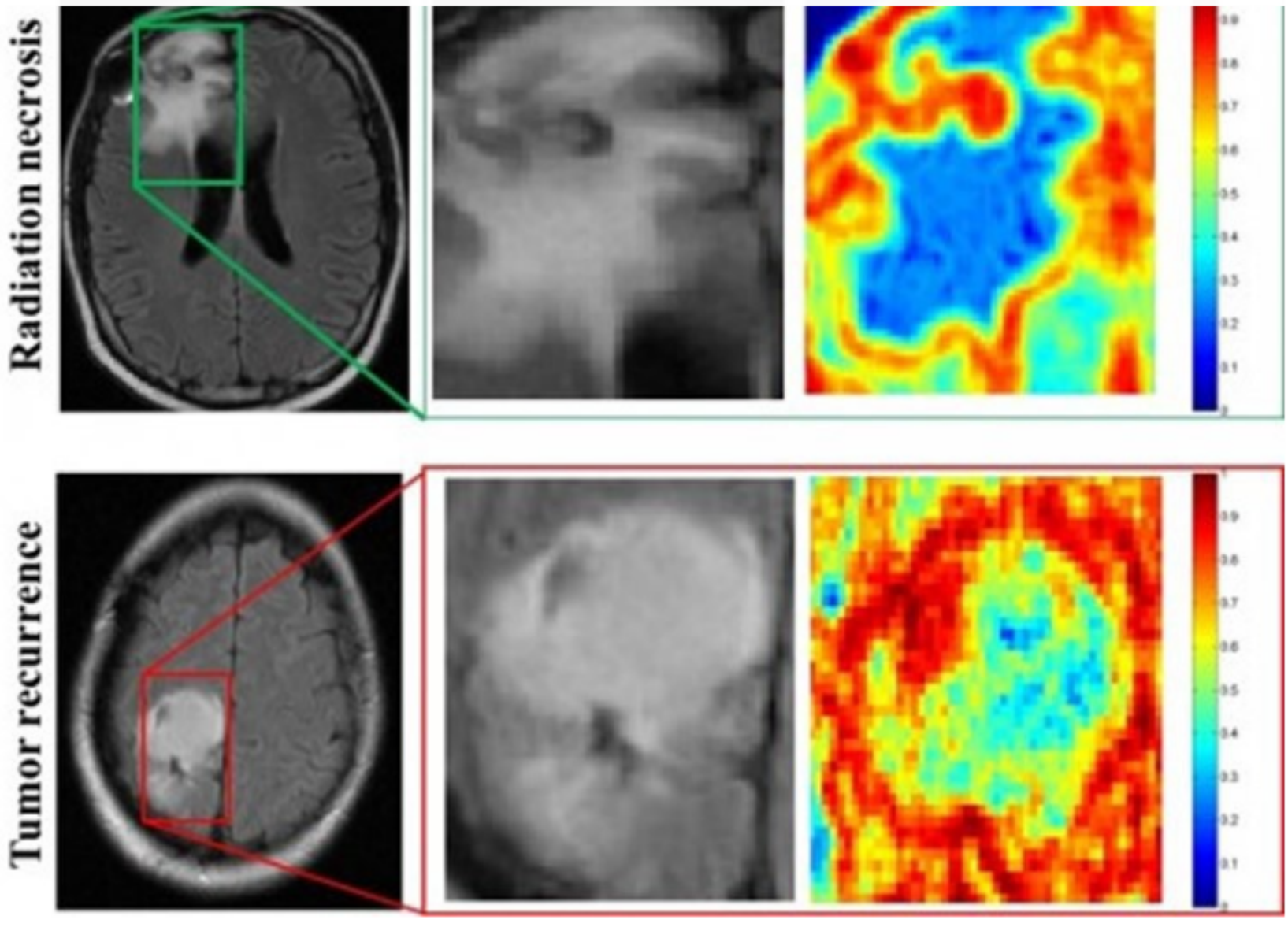
- O delete data
- O use machen learning
- O well document data
- O use PowerPoint to deliver a message

Review Questions

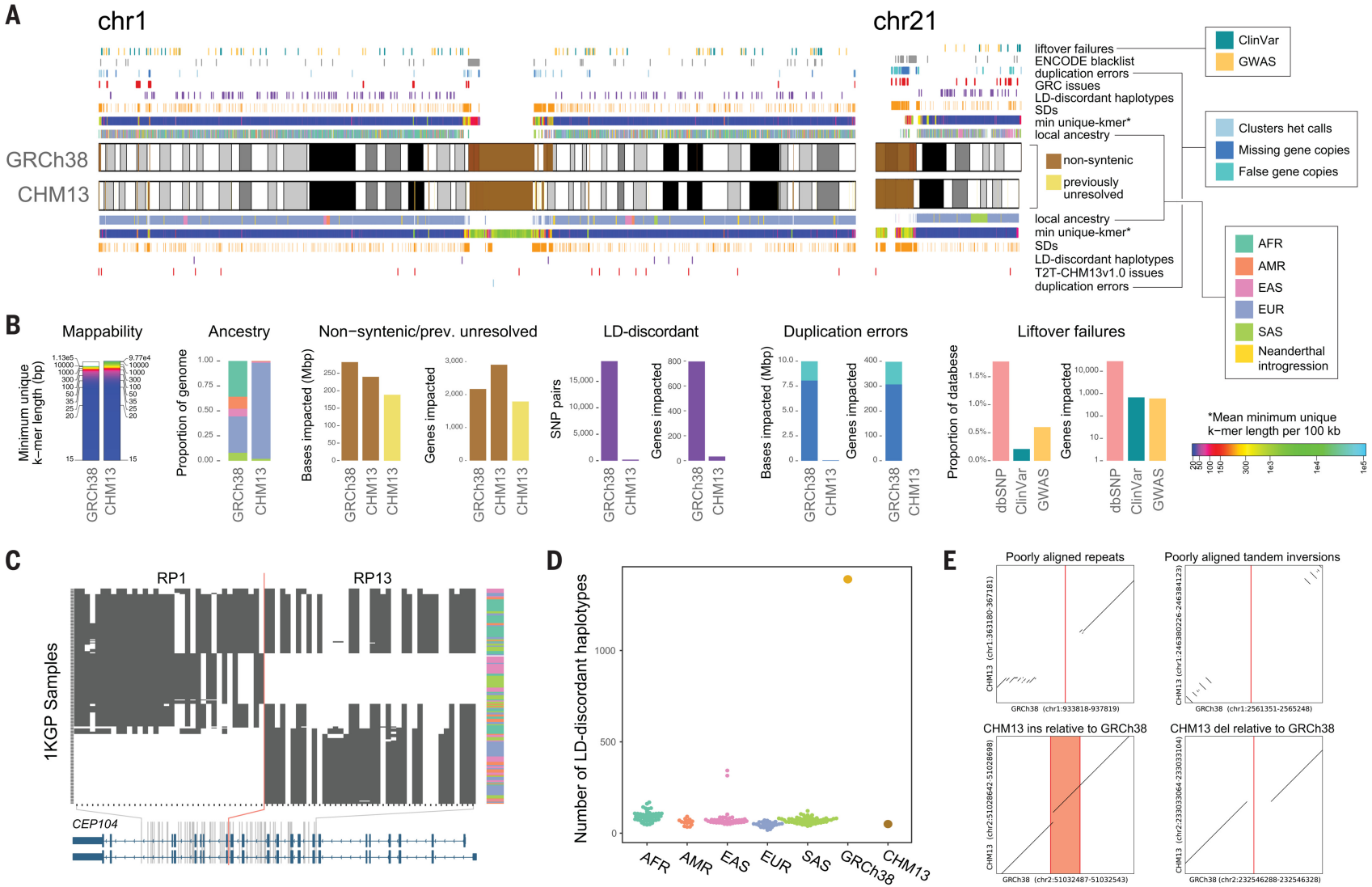
What is the role of the data scientist?

- O Email the stakeholders about the analysis
- O Manage a team of analysts to create a model
- O Develop the strategy to fix the problems in the findings
- O Use the insights to build the narrative to communicate the findings
- O Use the data to tell the story the CEO wants to tell

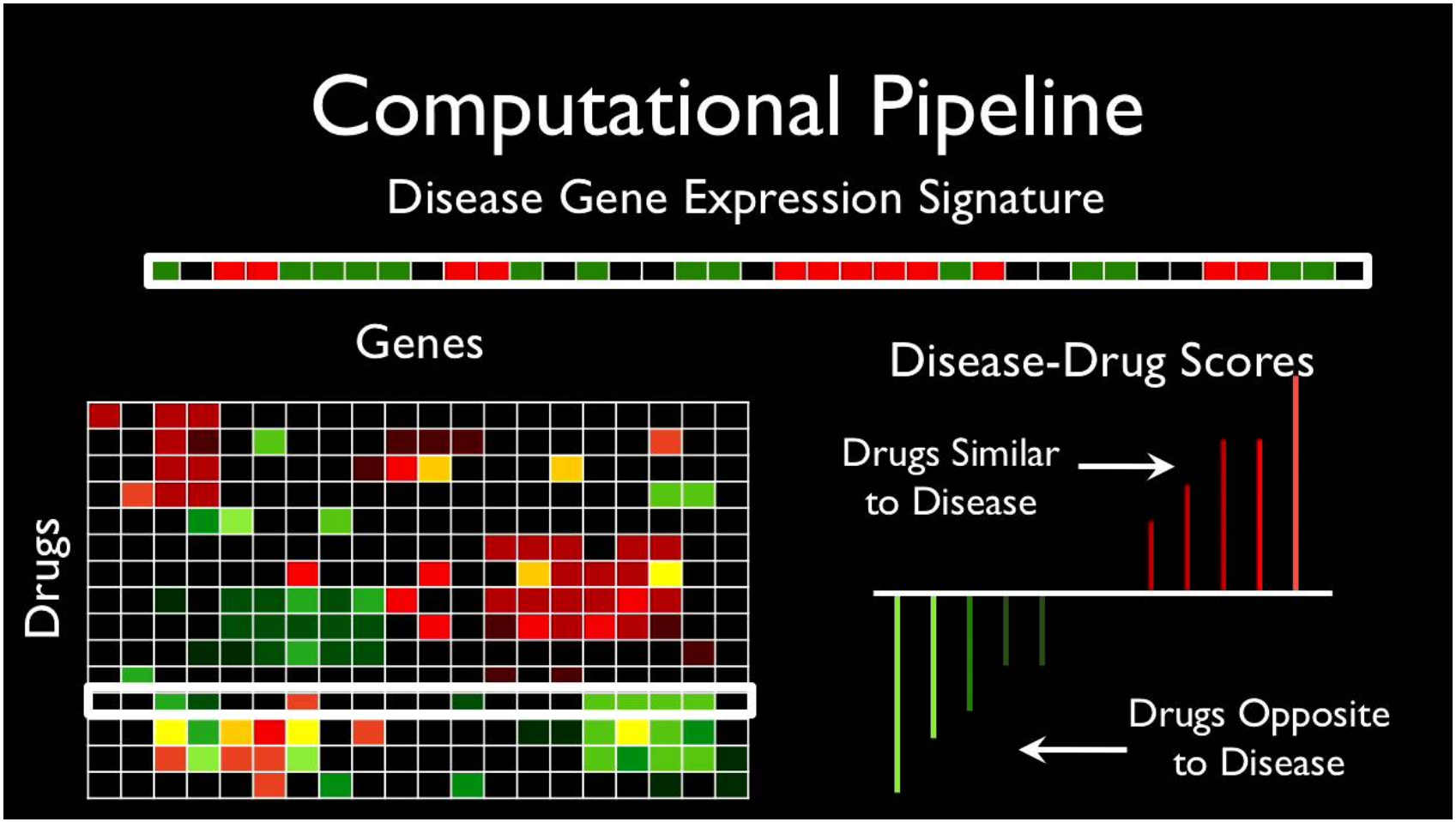
Applications: Medical Image Analysis



Applications: Genetics & Genomics

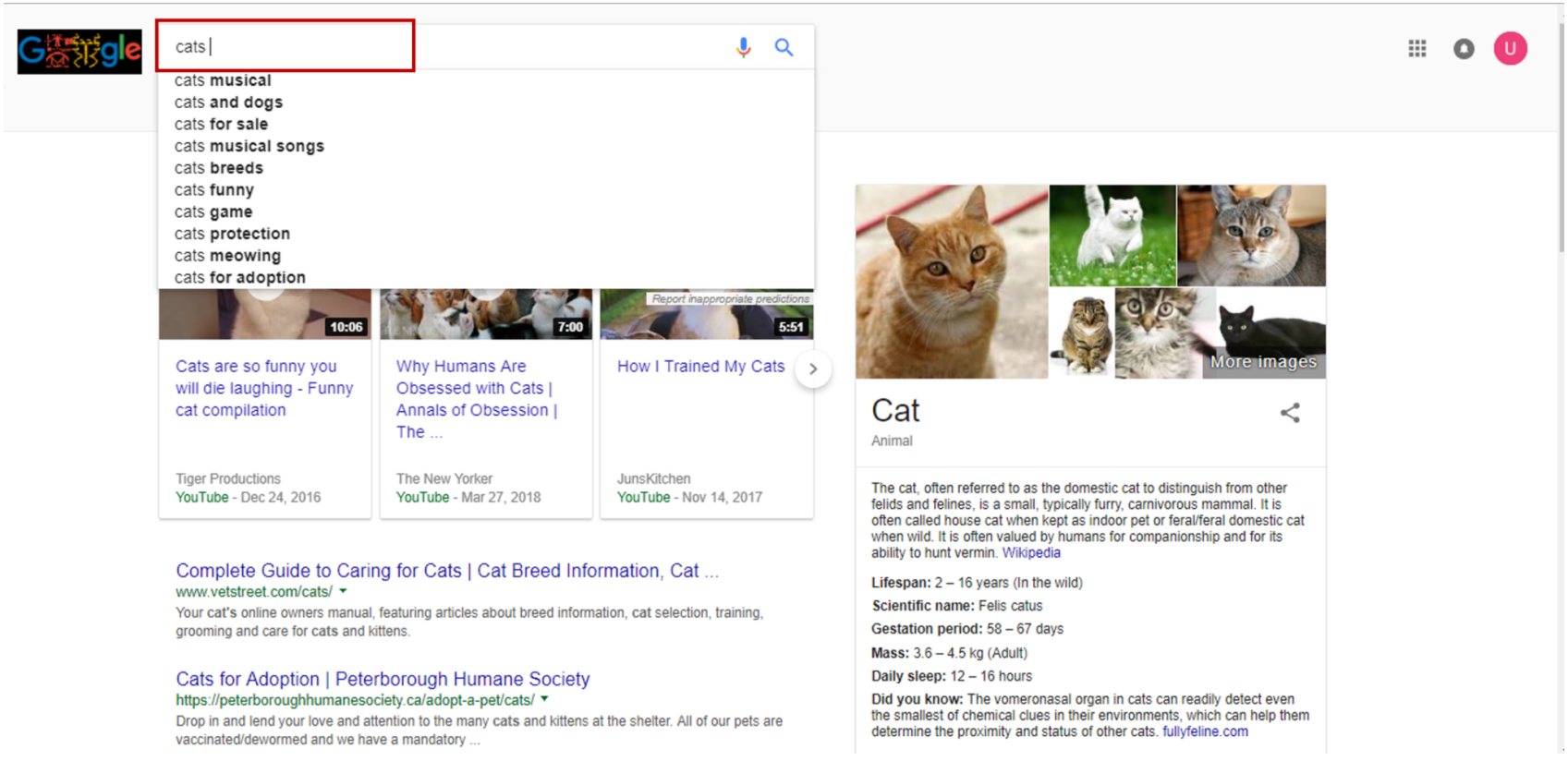


Applications: Drug Development

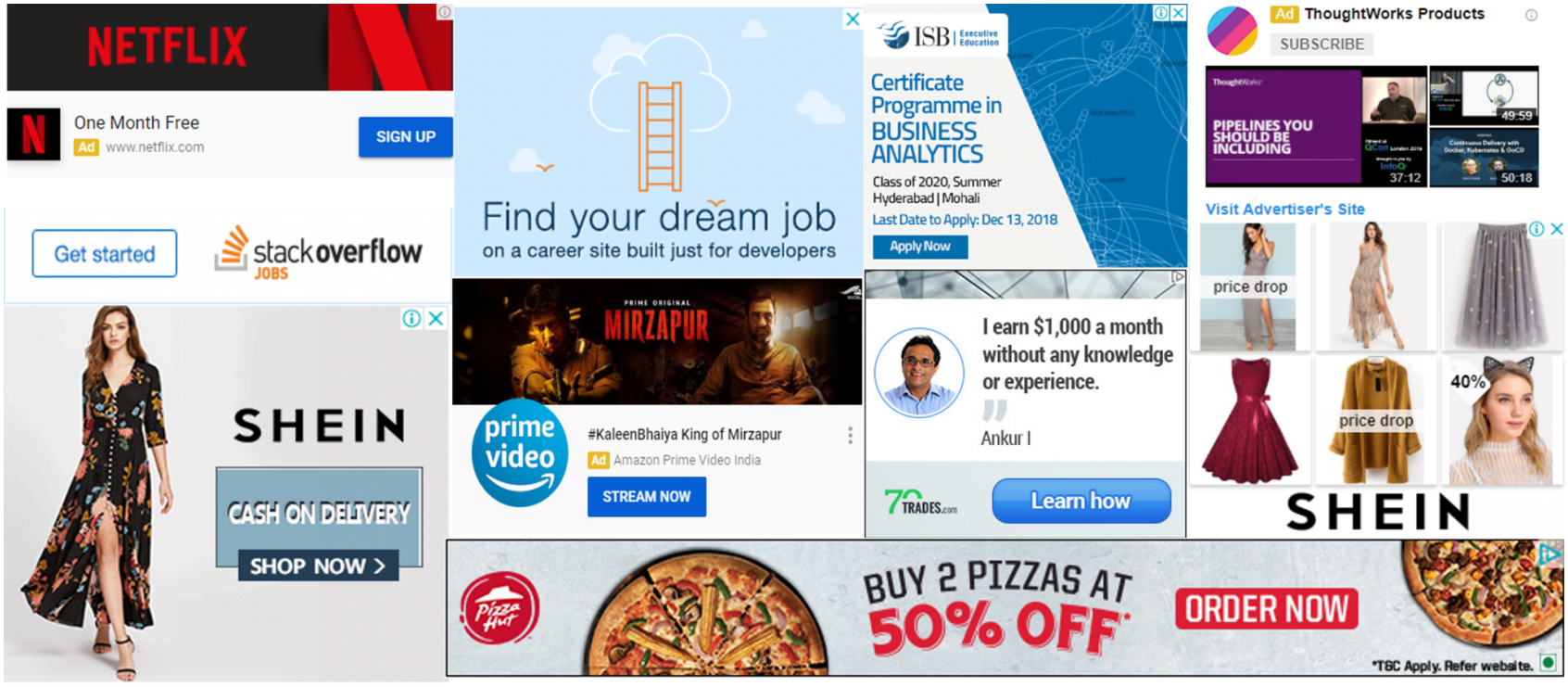


Applications: Virtual assistance for patients and customer support

Applications: Internet Search



Applications: Targeted Advertising

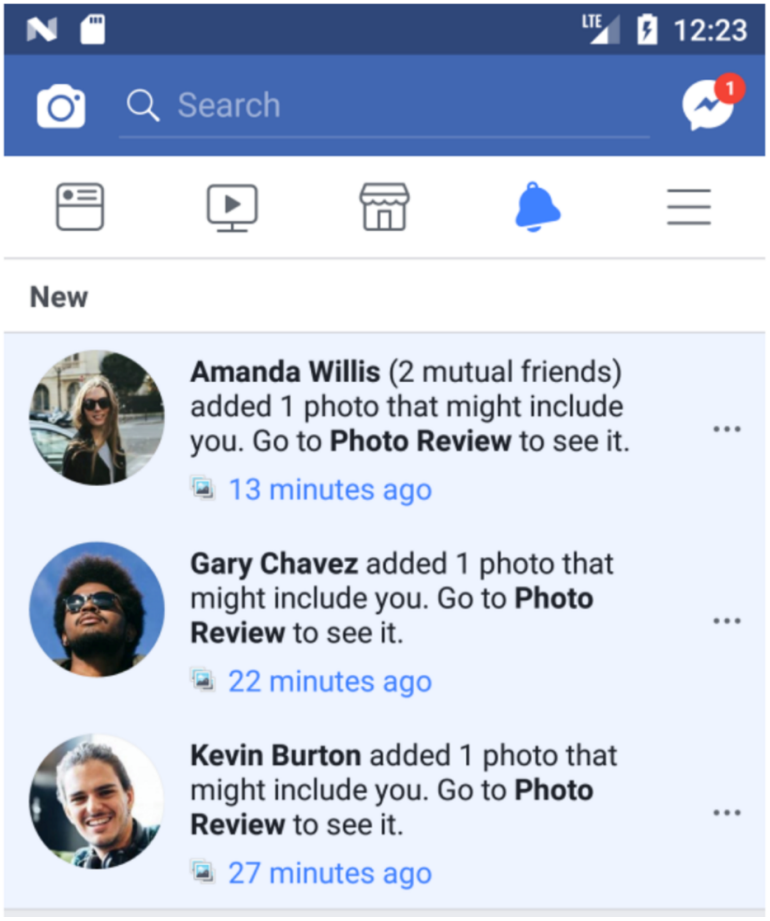


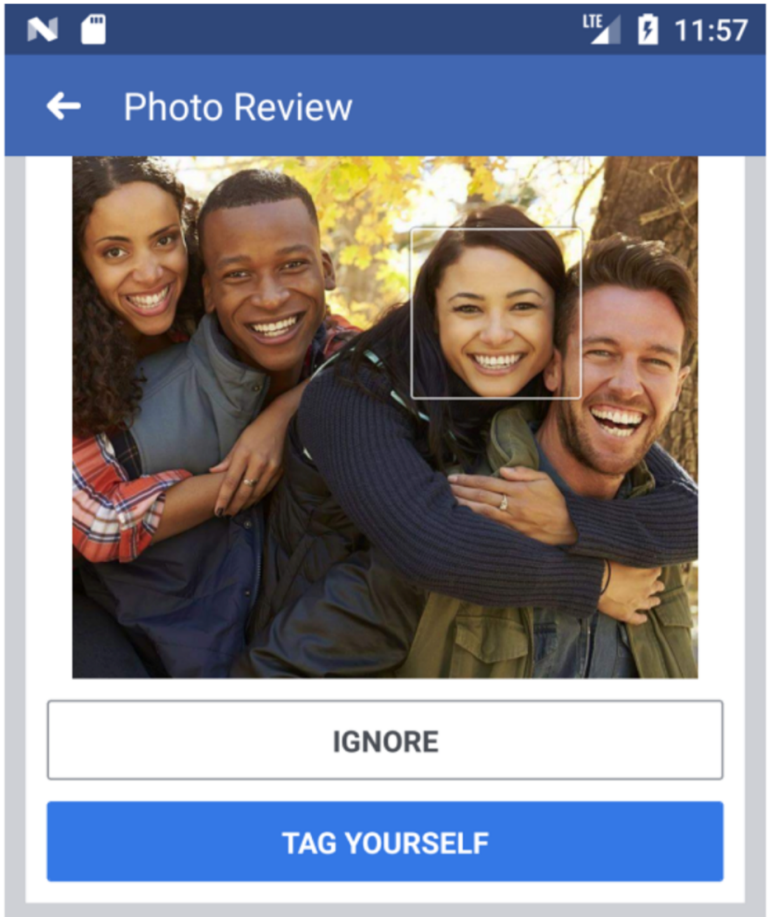
Applications: Website Recommendations

Compare with similar items

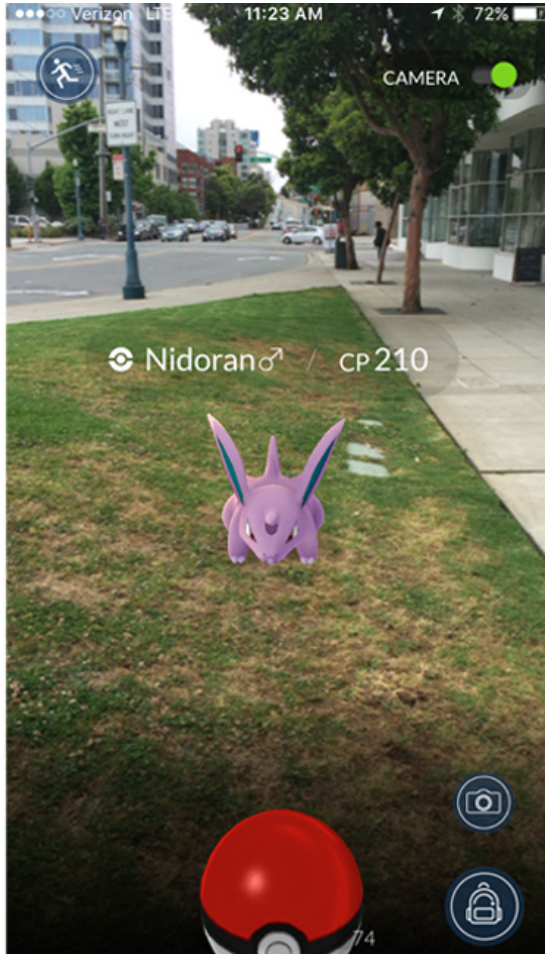
				
	This item Bose SoundLink Wireless Around-Ear Headphones with Mic (Black)	Sennheiser HD 4.40-BT Bluetooth Headphones (Black)	Bose 741158-0020 SoundLink Wireless Around-Ear Headphones with Mic (White)	Bose 789564-0030 Quiet Comfort 35 Wireless Headphone (Blue)-Special Edition
	Add to Cart	Add to Cart	Add to Cart	Add to Cart
Customer Rating	★★★★☆ (68)	★★★★☆ (349)	★★★★½ (22)	★★★★½ (200)
Price	₹ 19,000.00	₹ 7,490.00	₹ 19,000.00	₹ 29,363.00
Shipping	FREE Shipping	FREE Shipping	FREE Shipping	FREE Shipping
Sold By	Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd	Appario Retail Private Ltd
Colour	Black	Black	White	Blue
Connectivity Technology	bluetooth wireless	Bluetooth Wireless	Bluetooth Wireless	Bluetooth Wireless

Applications: Image Recognition

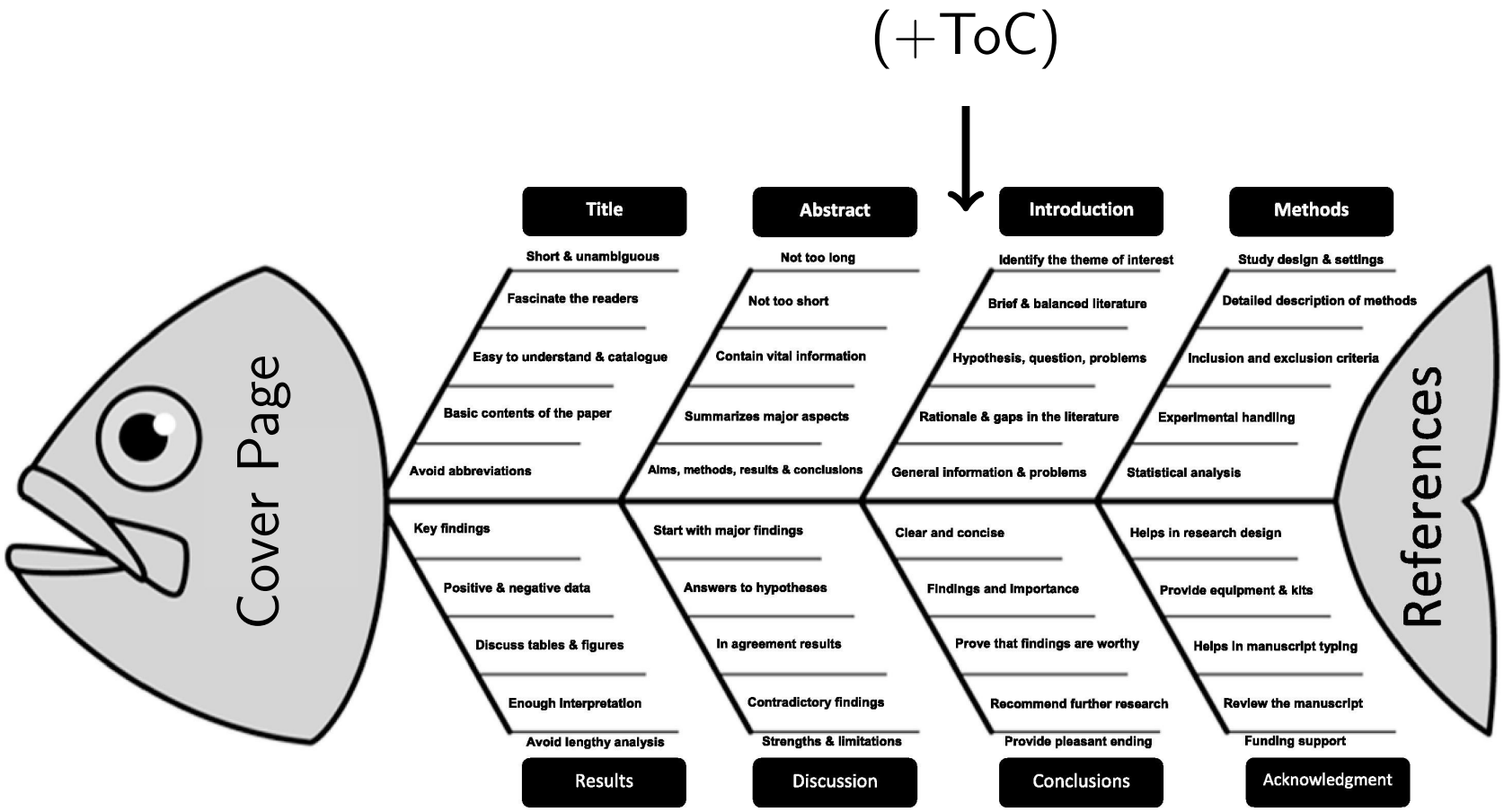




Applications: Gaming



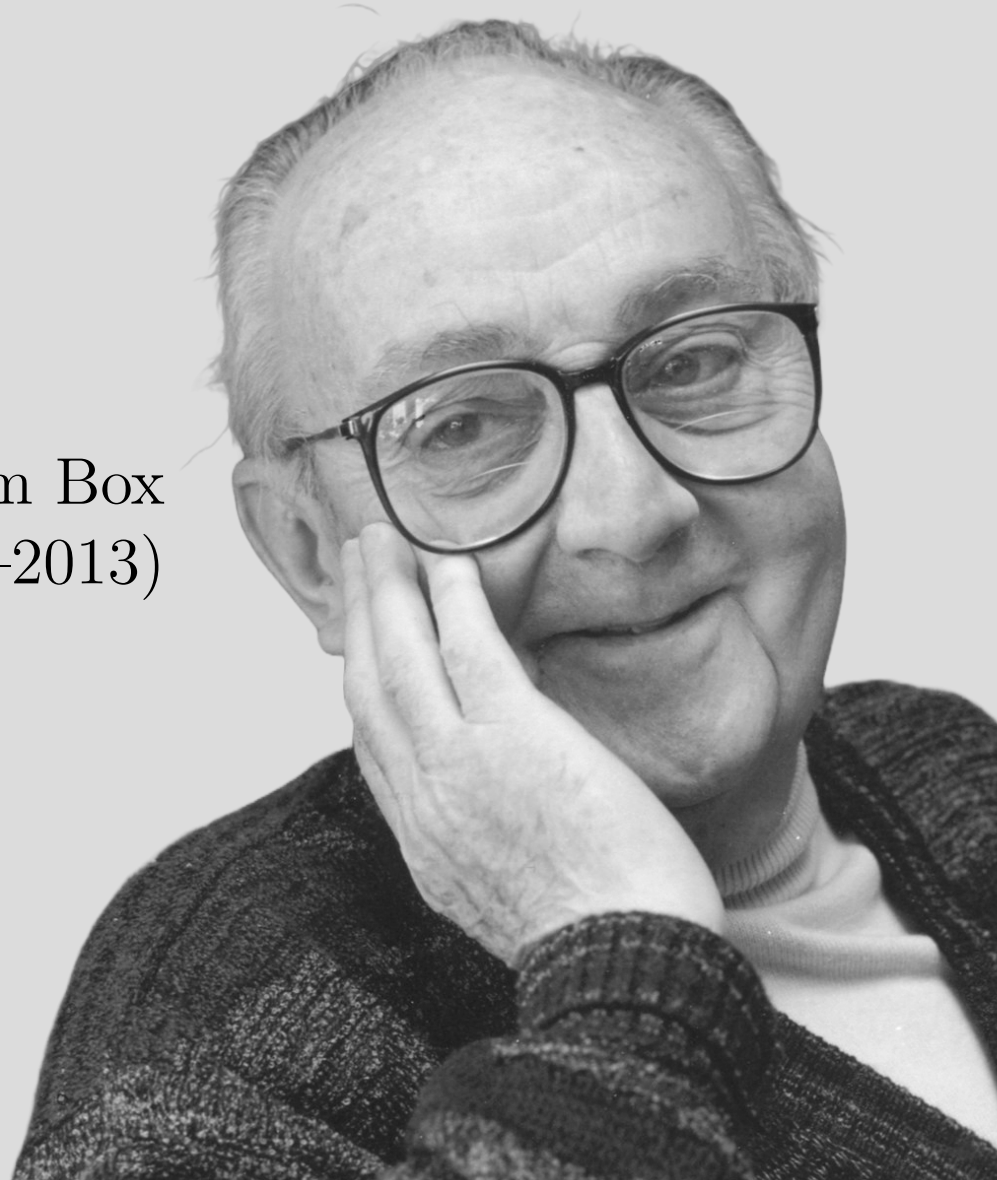
Report Structure



Keep in minds

"All models are wrong,
but some are useful"

— George Edward Pelham Box
(1919–2013)



Review Questions

What popular product is primarily based on data science?

- O smartphone
- O Google search
- O Space X's rockets
- O Tesla's Electric Cars

Review Questions

In a report writing, the results section is where you present:

- O The empirical findings
- O R Squared
- O The conclusion
- O The contributors
- O The methods used

Review Questions

Complete the sentence:

Predictions are useful, ...

- they are always correct
- but you need lots of data
- but they must come from a complicated model
- they are always wrong

Well-known data scientists

Yann LeCun

Chief AI Scientist at Meta

DJ Patil

U.S. Chief Data Scientist

Yoshua Bengio

Godfathers of Deep Learning

Eva Stuler

Chief Data Scientist at IBM

Kamelia Aryafar

Sr. Director of Eng. at Google

Andrew Ng

Co-Founder of Coursera

Review Questions

How does Prof. Haider define data science?

- Data science is way of understanding things, of understanding the world
- Data science is a physical science like physics or chemistry
- Data science is some data and more science
- Data science is what data scientists do
- Data science is the art of uncovering the hidden secrets in data

Review Questions

What is admirable about DJ Patil's definition of a data scientist?

- His definition limits data science to activities involving machine learning
- His definition is only for people who program in Python
- His definition excludes statistics
- His definition is about weaving strong narratives into analytics
- His definition is inclusive of individuals from various academic backgrounds and training

Review Questions

A good data scientist should ?

- O calculate confidence intervals
- O be sceptical
- O use complicated models
- O only use big data