

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

List of modules

1. Defining Data Science
2. What do data science people do?
3. Data Science in Business
4. Use Cases for Data Science
5. Data Science People

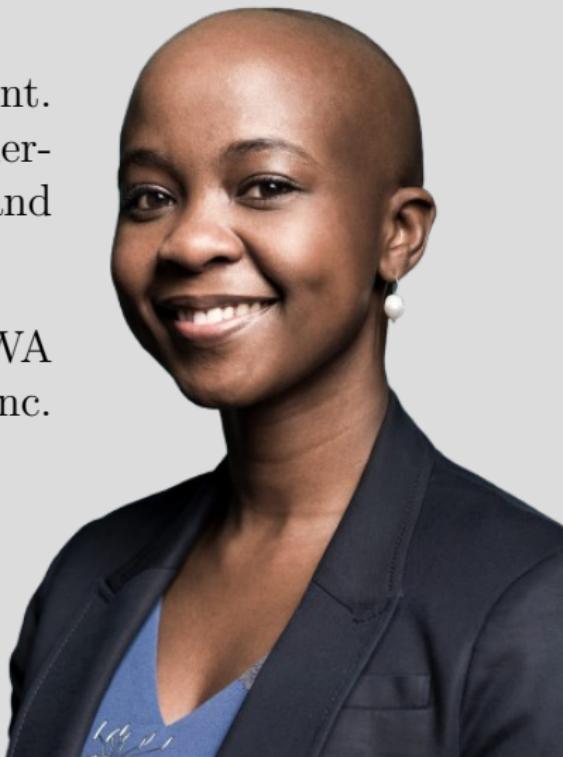
Module 1

Defining 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.



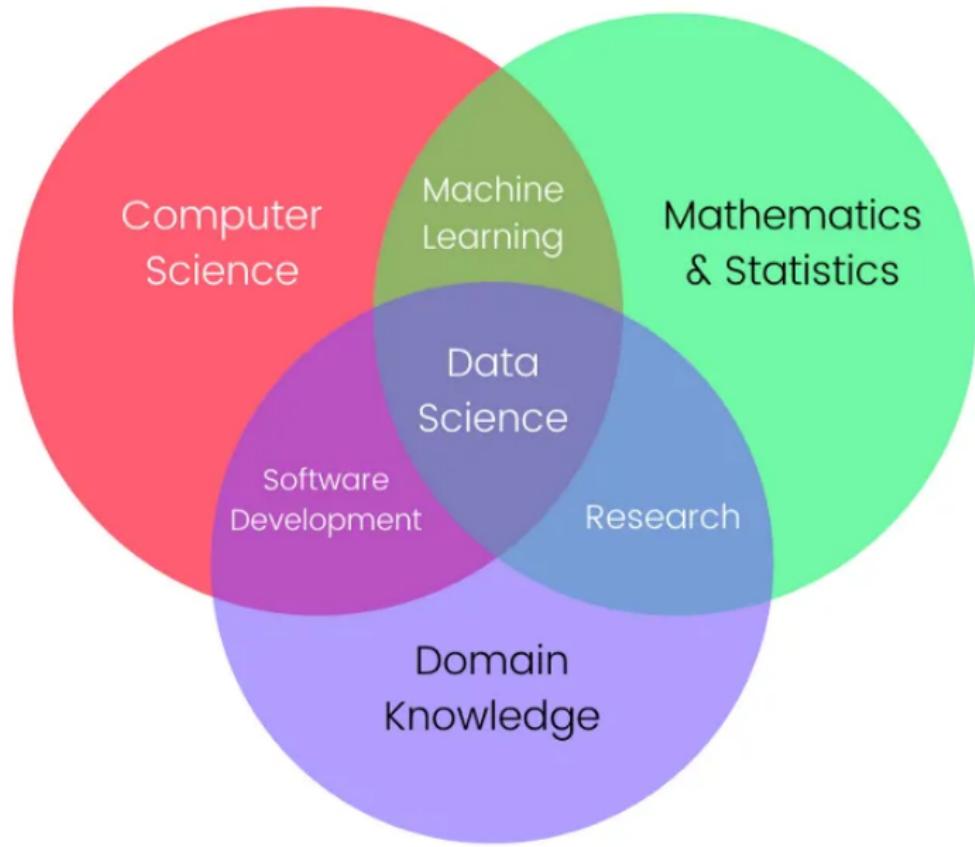
What is data science?

“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

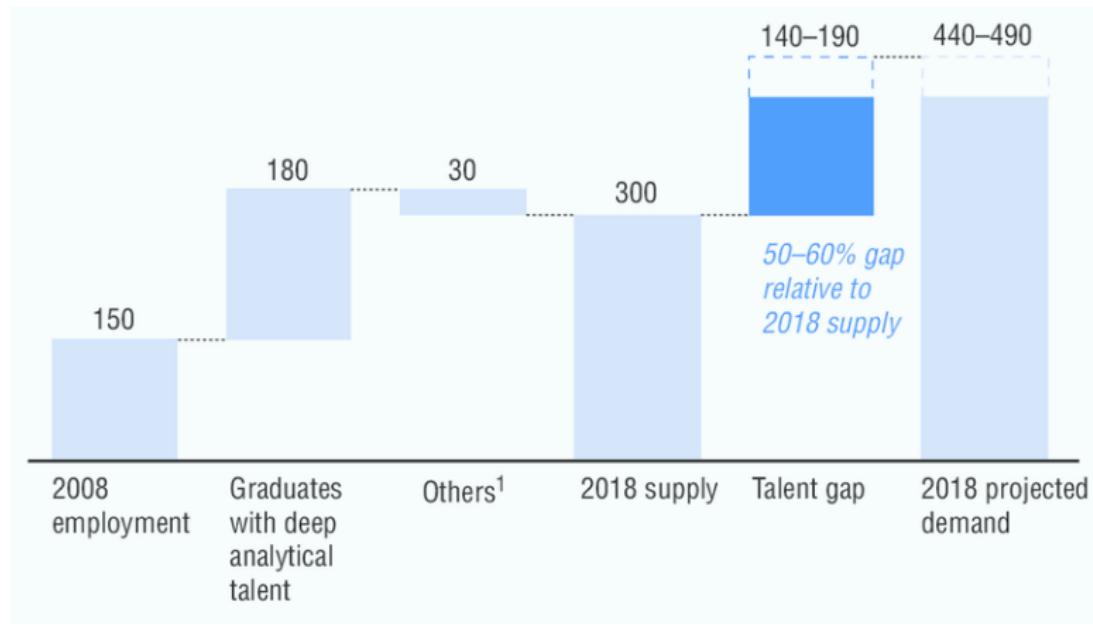


Figure: Demand for data analytics expertise according to the McKinsey Global Institute.

Data Scientist: The Sexiest Job in 21st Century

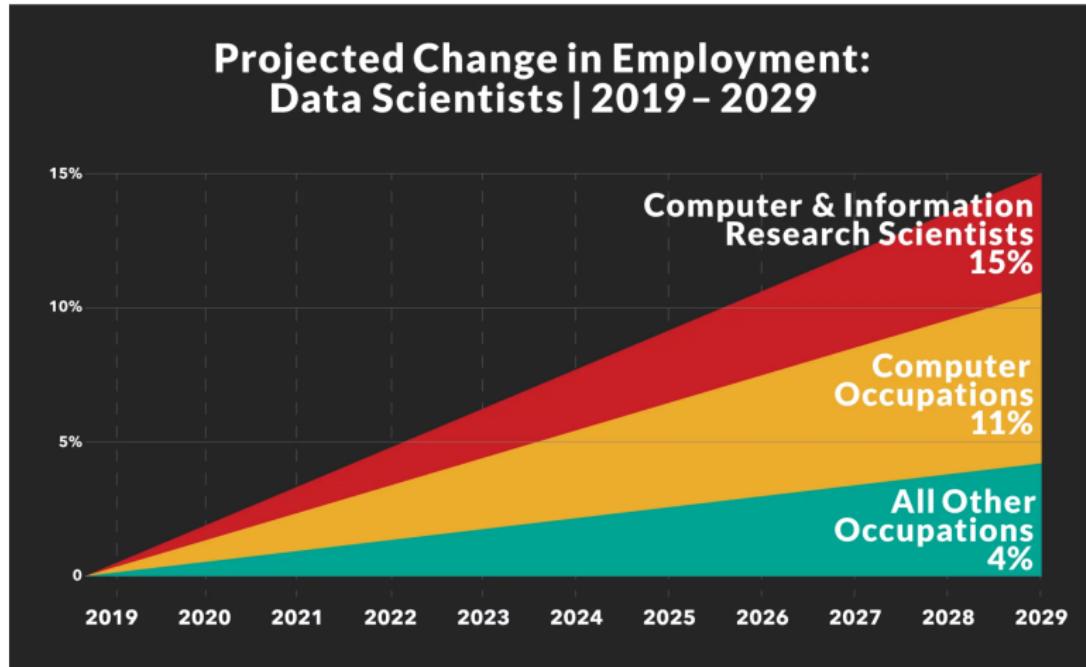


Figure: United States Bureau of Labor Statistics (BLS)
Occupational Outlook Handbook

Module 1.1: Review questions

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?

- 20,000 - 50,000
- 120,000
- 140,000 - 190,000
- 800,000 - 900,000
- 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

Module 2

What do data science people do?

A day in the life of a data science person

Data Scientist



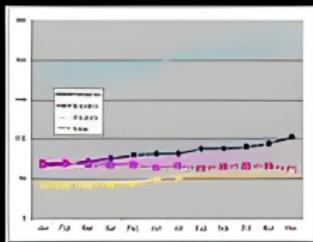
What my friends think I do



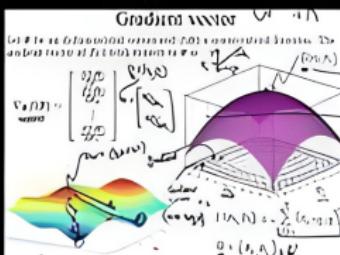
What my mom thinks I do



What society thinks I do



What my boss thinks I do



What I think I do



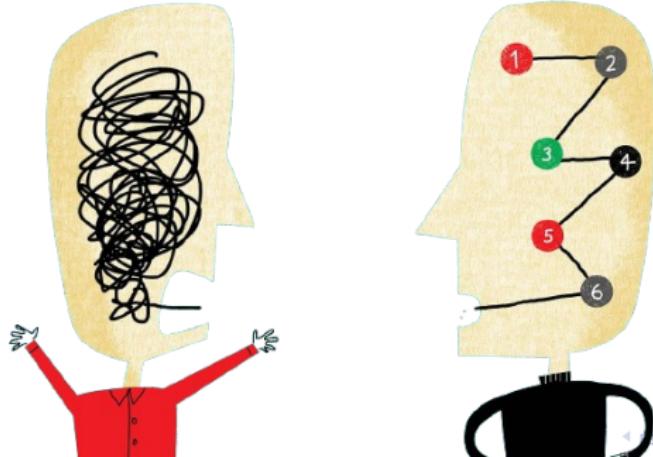
What I actually do

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



Numbers, dates and strings



Estimated 20% of enterprise data (Gartner)



Requires less storage



Easier to manage and protect with legacy solutions



Cannot be displayed in rows, columns and relational databases



Images, audio, video, word processing files, e-mails, spreadsheets



Estimated 80% of enterprise data (Gartner)



Requires more storage

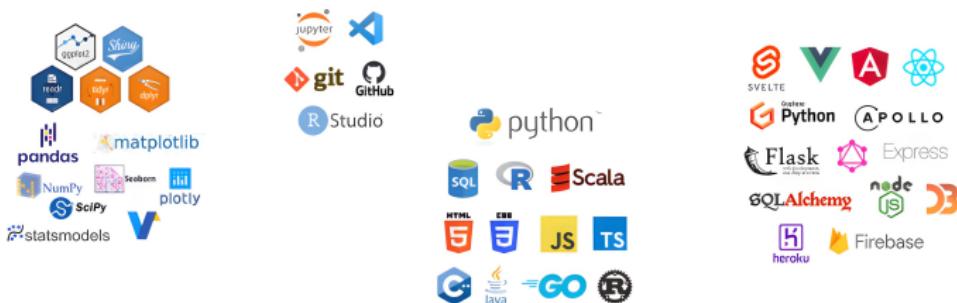


More difficult to manage and protect with legacy solutions

Data science tools and technology



Data science tools and technology



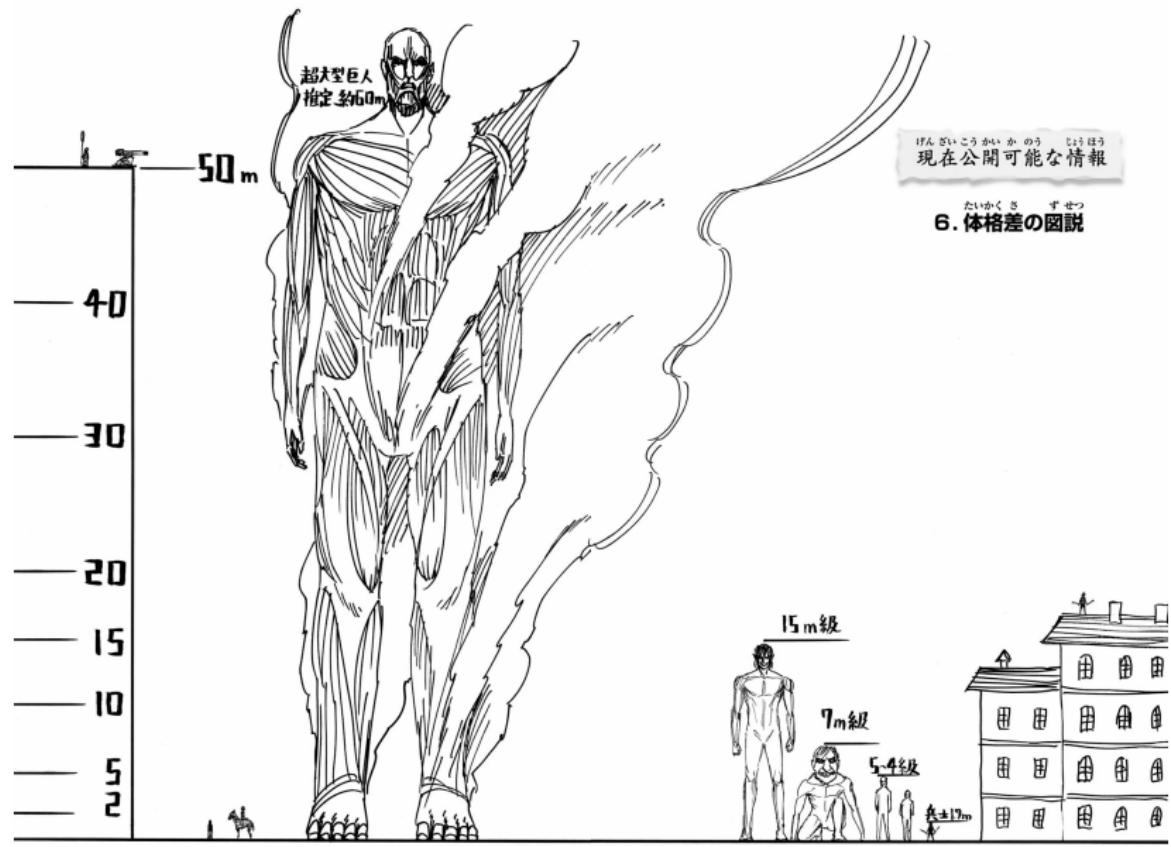
Data science tools and technology



Regression?



Regression?

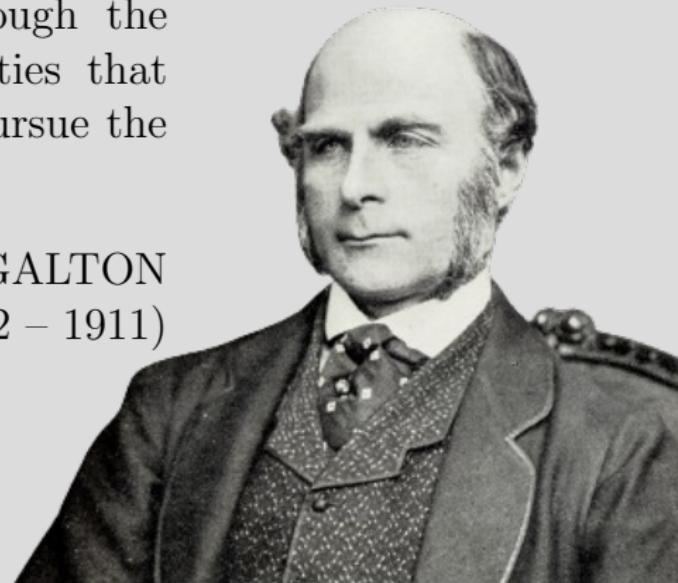


Regression toward the mean

“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)



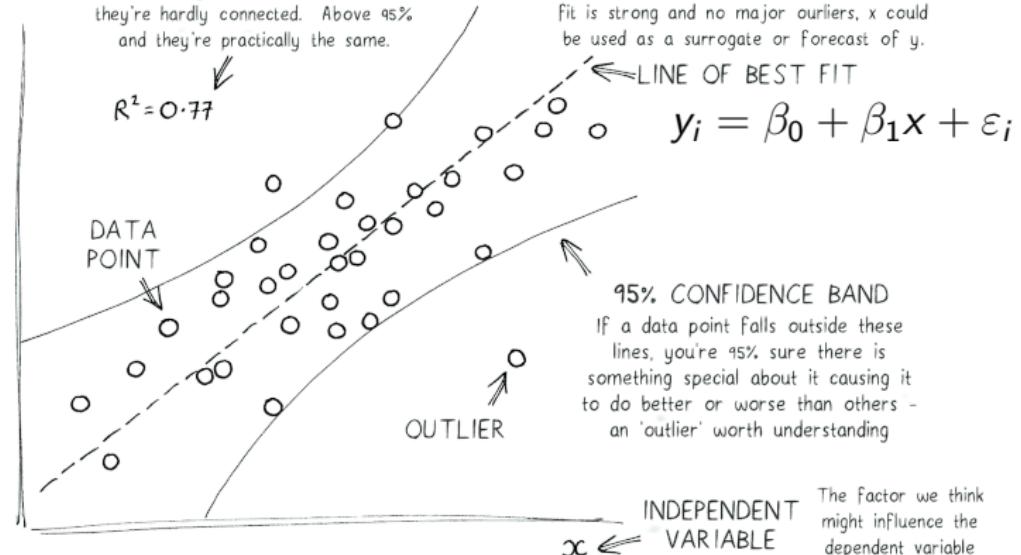
Regression toward the mean

The thing we want
to explain

DEPENDENT
VARIABLE

i.e. 77% of the variance in y is explained by x . Below 0.30% means they're hardly connected. Above 95% and they're practically the same.

$$R^2 = 0.77$$



Case: House Price Prediction

At Use Case

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

Price vs Number of Houses

2. Insights obtained from Lot size:

Lot size provides information on the total size of the land 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 of Houses

At User Case

Forecasting selling price for a house using predictive analysis

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 analyze the features of the house 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.

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 County. The model is further simplified, concerning the distinct features available in the collected data set, to drive more accuracy in predictions.

Forecasting selling price for a house using predictive analysis

on the features we collected by using Data Science tools

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

- Lasso regression.
- Random forest.

In the year 2017 of the Saratoga Country houses with 1 square feet, number of rooms in the house, number of rooms (square feet), etc.

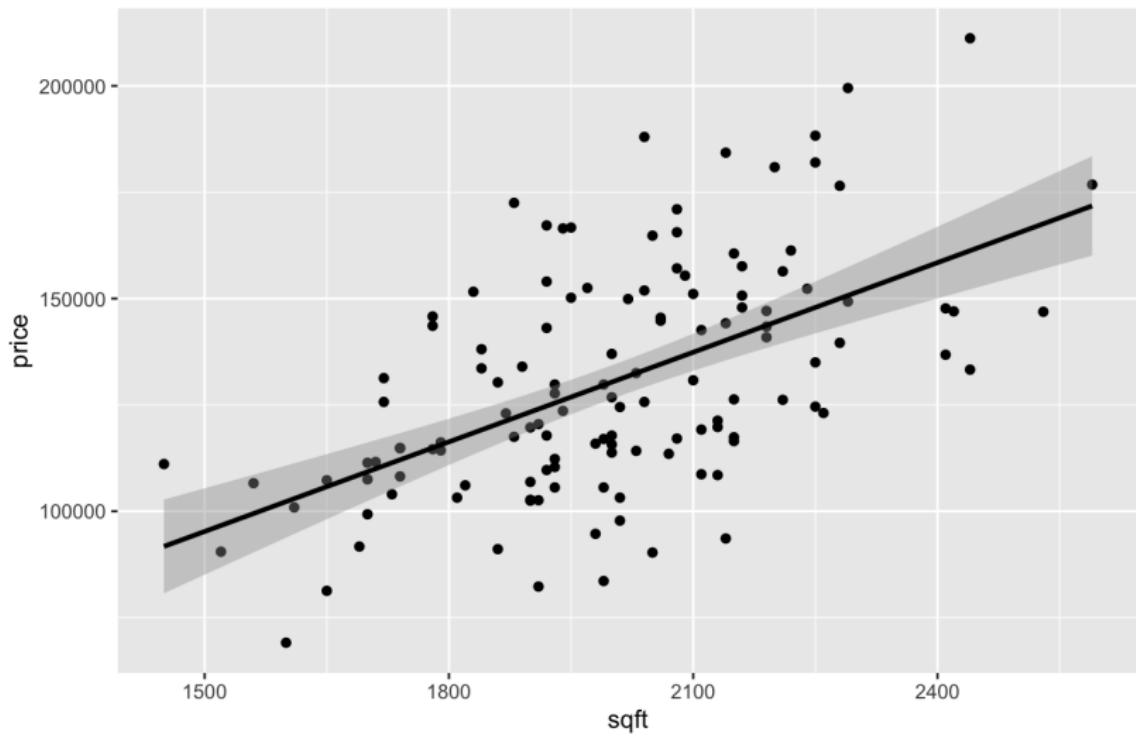
, redundant, and duplicate columns. It was a clean raw data for analysis. We have analyzed the distinct a better predictive model

differences in the price:

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

tribution vs Number of Houses

Case: House Price Prediction



Module 2.1: Review questions

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?:

`base_fair + taxi_rate × travel_distance`

- 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 taxi 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?

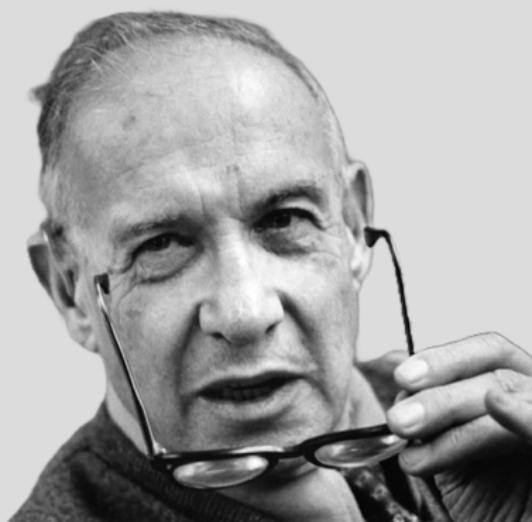
Module 3

Data Science in Business

How should companies get started in Data Science

“If you can’t measure it,
you can’t manage it.”

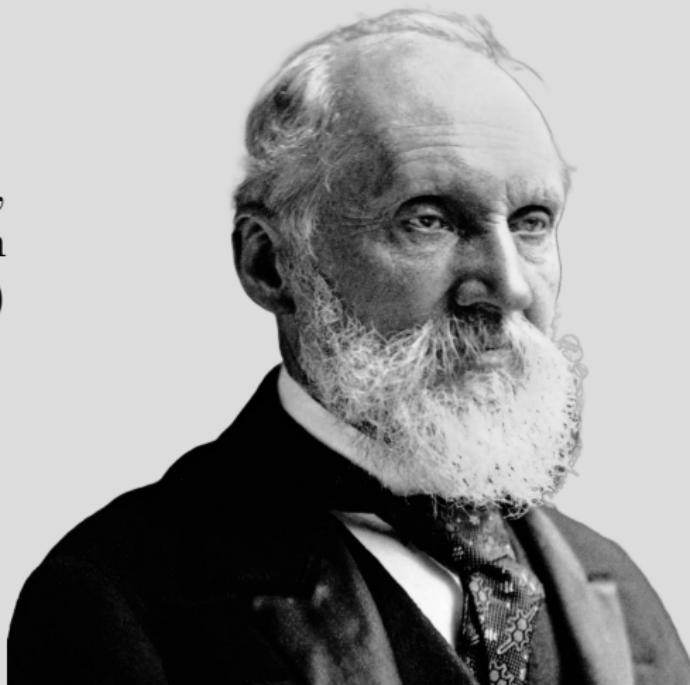
— Peter Drucker
(1909-2005)



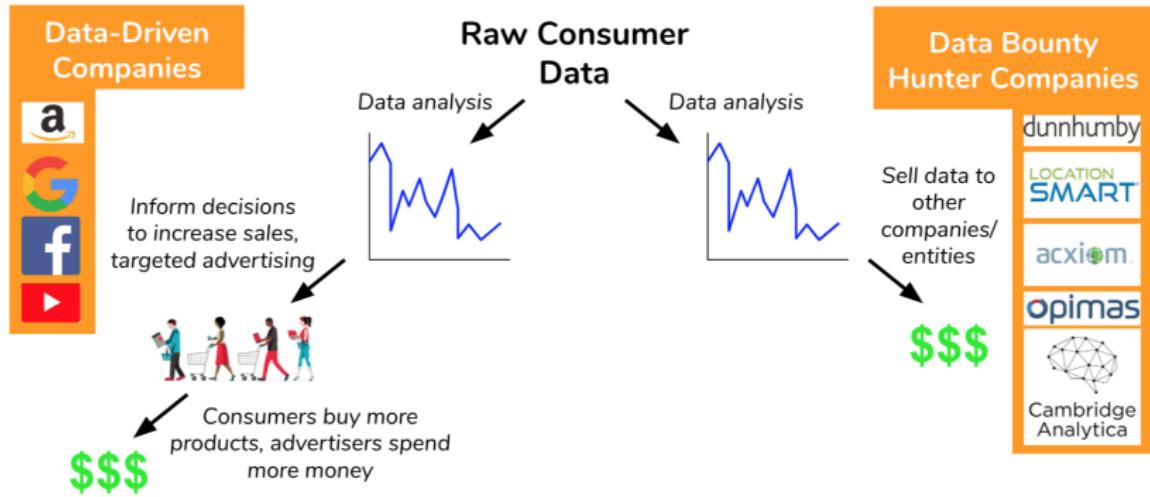
How should companies get started in data science

“If you can’t measure it,
you can’t improve it”

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



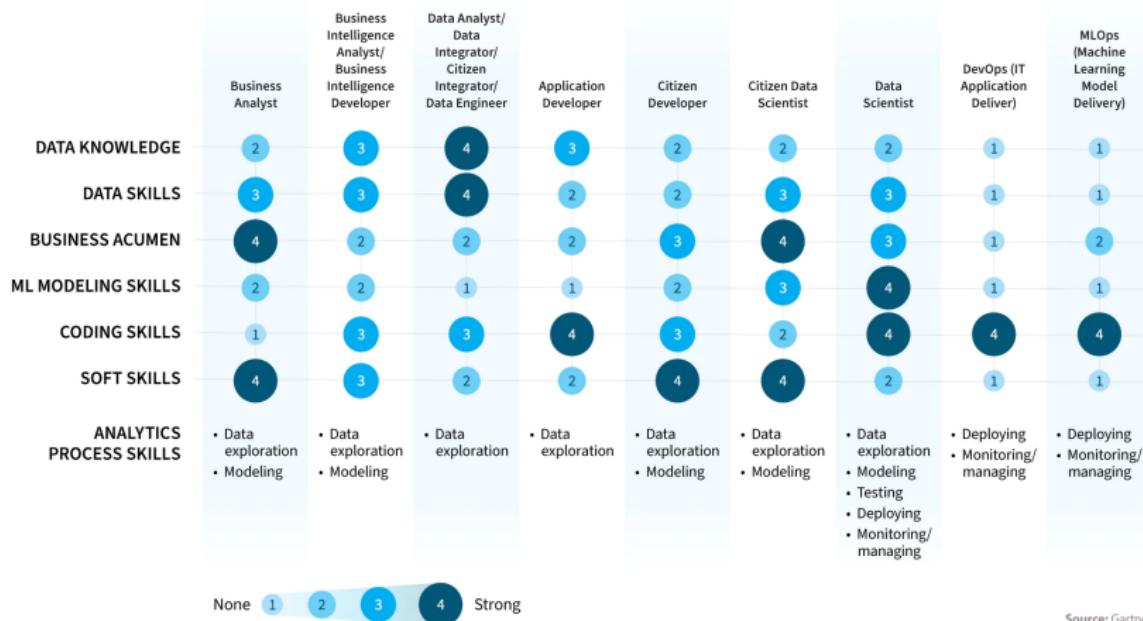
How should companies get started in data science



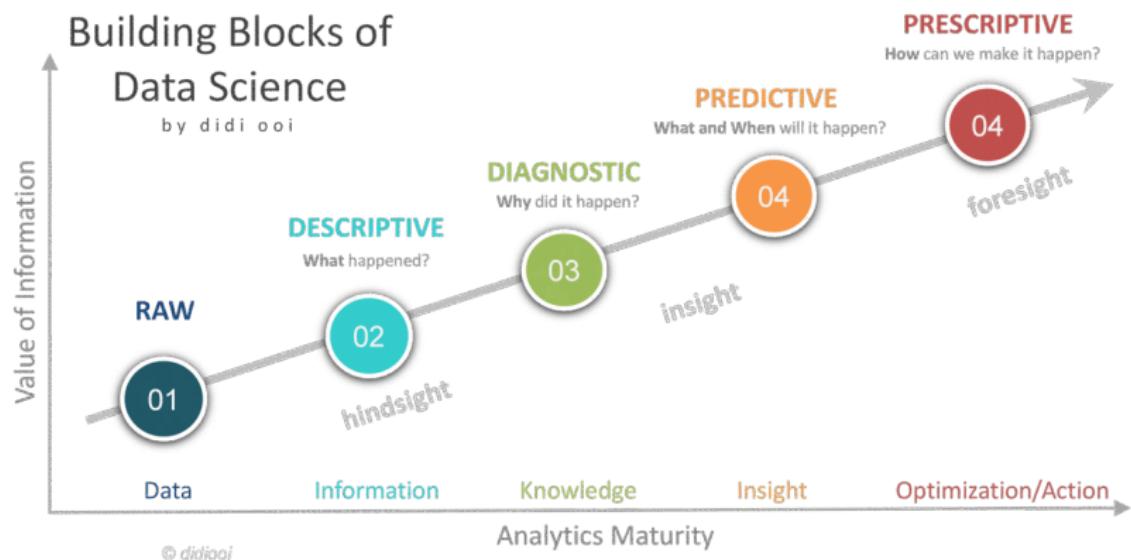
Never delete data, it never get old.

Recruiting for data science

Continuum of Analytics Roles and Skills



The final deliverable



Module 3.1: Review questions

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, ...

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

Review Questions

A business should never:

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

Review Questions

What is the role of the data scientist?

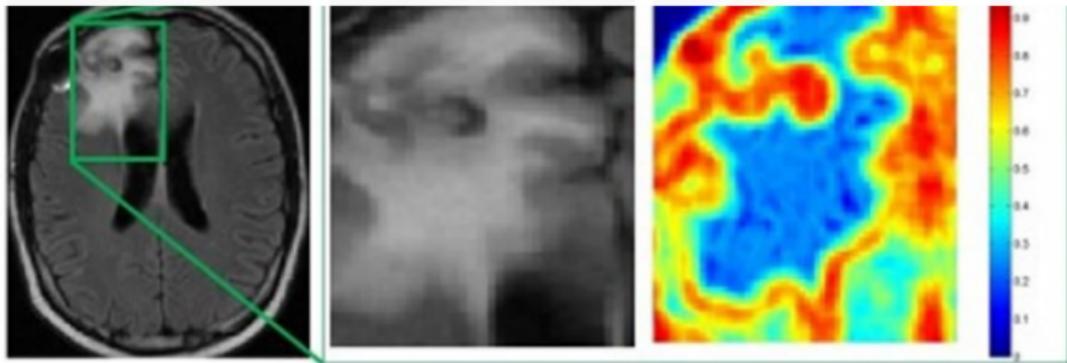
- Email the stakeholders about the analysis
- Manage a team of analysts to create a model
- Develop the strategy to fix the problems in the findings
- Use the insights to build the narrative to communicate the findings
- Use the data to tell the story the CEO wants to tell

Module 4

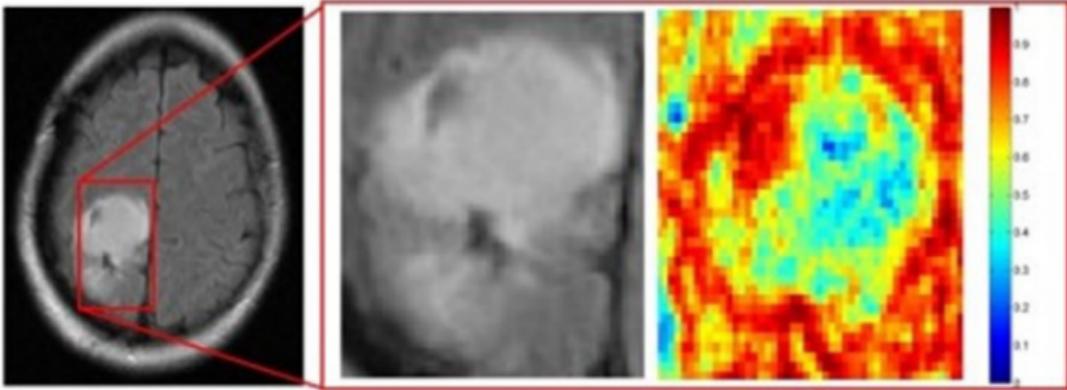
Use Cases for Data Science

Applications: Medical Image Analysis

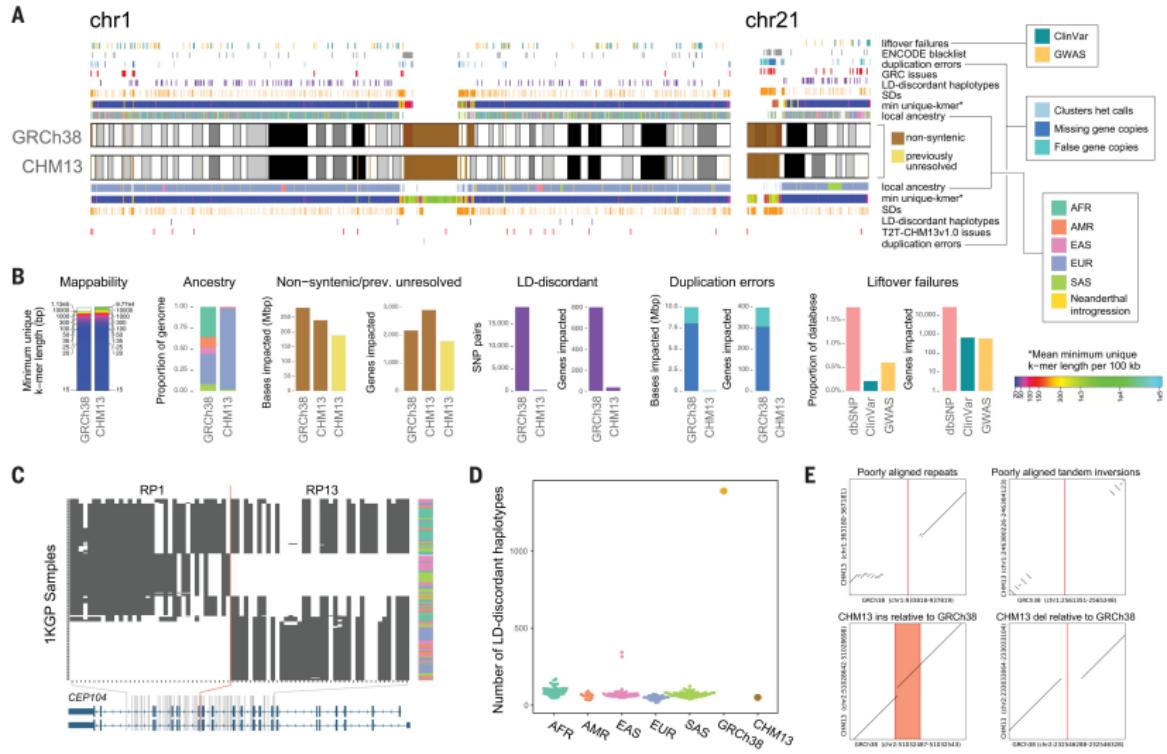
Radiation necrosis



Tumor recurrence



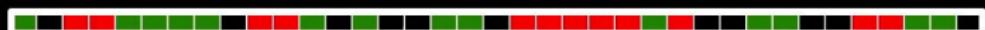
Applications: Genetics & Genomics



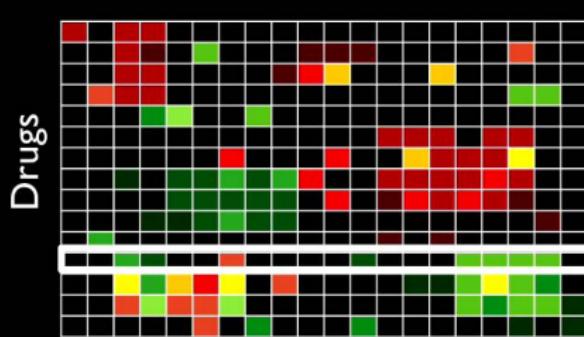
Applications: Drug Development

Computational Pipeline

Disease Gene Expression Signature

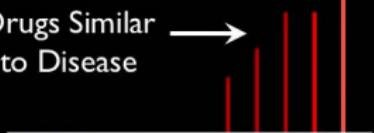


Genes



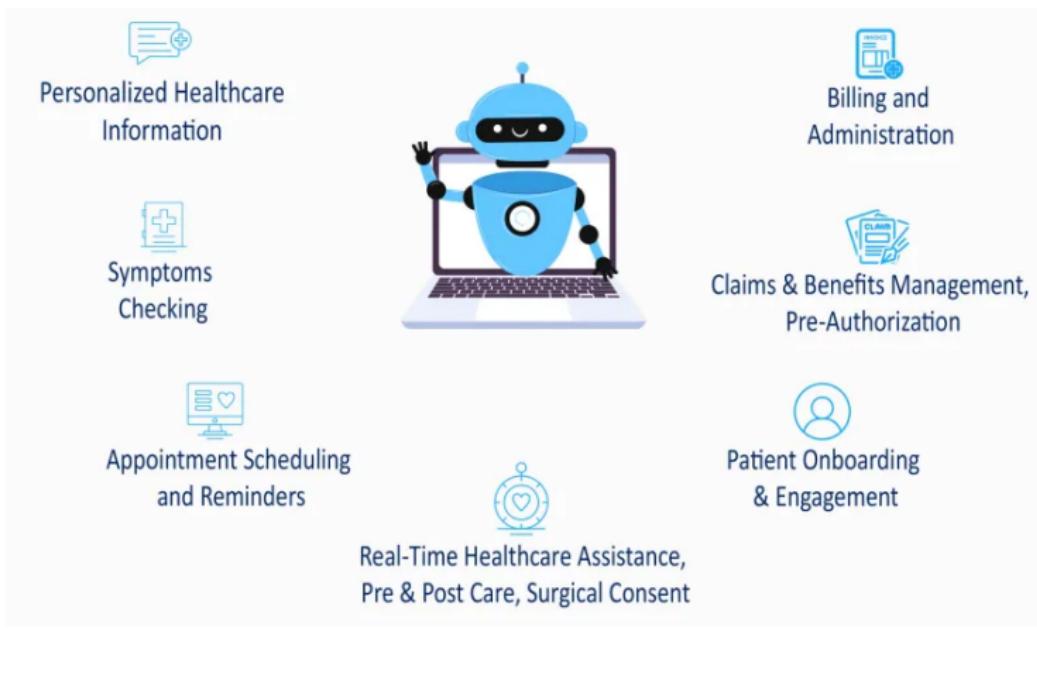
Disease-Drug Scores

Drugs Similar
to Disease →



← Drugs Opposite
to Disease

Applications: Virtual assistance for patients and customer support



Applications: Internet Search

Google search results for "cats".

Search term: cats

Suggested searches:

- cats musical
- cats and dogs
- cats for sale
- cats musical songs
- cats breeds
- cats funny
- cats game
- cats protection
- cats meowing
- cats for adoption

Image cards:

- Cats are so funny you will die laughing - Funny cat compilation (Tiger Productions YouTube - Dec 24, 2016)
- Why Humans Are Obsessed with Cats | Annals of Obsession | The ... (The New Yorker YouTube - Mar 27, 2018)
- How I Trained My Cats (JunsKitchen YouTube - Nov 14, 2017)

Image results:

- Report inappropriate media
- More Images

Image results (large view):

- A large orange tabby cat looking directly at the camera.
- A white cat sitting in green grass.
- A grey and white cat looking towards the camera.
- A small black and tan kitten.
- A close-up of a cat's face.
- A black cat lying down.

Definition card:

Cat

Animal

The cat, often referred to as the domestic cat to distinguish from other felines and tigers, is a small, typically furry, carnivorous mammal. It is often called house cat when kept as indoor pet or feral/feral domestic cat when wild. It is often valued by humans for companionship and for its ability to hunt vermin. [Wikipedia](#)

Information card:

Lifespan: 2 – 16 years (In the wild)
Scientific name: *Felis catus*
Gestation period: 58 – 67 days
Mass: 3.6 – 4.5 kg (Adult)
Daily sleep: 12 – 16 hours
Did you know: The vomeronasal organ in cats can readily detect even the smallest of chemical clues in their environments, which can help them determine the proximity and status of other cats. [fullyfeline.com](#)

Applications: Targeted Advertising

The image displays a collage of various targeted advertising examples from different platforms:

- Netflix:** A banner ad for Netflix with the text "One Month Free" and a "SIGN UP" button.
- Stack Overflow Jobs:** An advertisement for Stack Overflow Jobs with a "Get started" button.
- SHEIN:** A promotional banner for SHEIN featuring a woman in a floral dress, with "CASH ON DELIVERY" and "SHOP NOW >" buttons.
- Prime Video:** An advertisement for Prime Video showing a scene from the TV show "Mirzapur" with the text "#KaleenBhaiya King of Mirzapur" and a "STREAM NOW" button.
- ISB Executive Educators:** An advertisement for ISB Executive Educators' Certificate Programme in BUSINESS ANALYTICS, featuring a ladder leading up to a cloud.
- 7TRADES.com:** An advertisement for 7TRADES.com featuring a profile picture of Ankur I and the text "I earn \$1,000 a month without any knowledge or experience."
- ThoughtWorks Products:** An advertisement for ThoughtWorks Products with a "SUBSCRIBE" button and a thumbnail for a video titled "PIPELINES YOU SHOULD BE INCLUDING".
- Promotional Offers:** A grid of promotional offers including "price drop" items like a red dress and a gold top, and a 40% off offer featuring a woman in a cat-ear top.
- Pizza Hut:** An advertisement for Pizza Hut offering "BUY 2 PIZZAS AT 50% OFF" with a "ORDER NOW" button.

Applications: Website Recommendations

Compare with similar items



This item Bose SoundLink Wireless Around-Ear Headphones with Mic (Black)

Add to Cart



Sennheiser HD 4.40-BT Bluetooth Headphones (Black)

Add to Cart



Bose 741158-0020 SoundLink Wireless Around-Ear Headphones with Mic (White)

Add to Cart

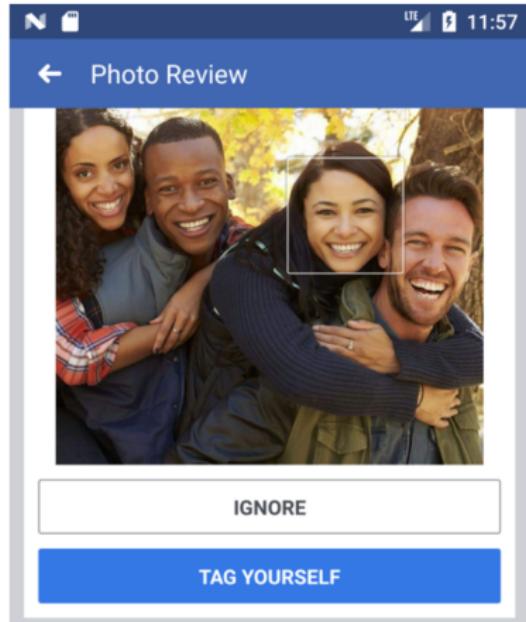
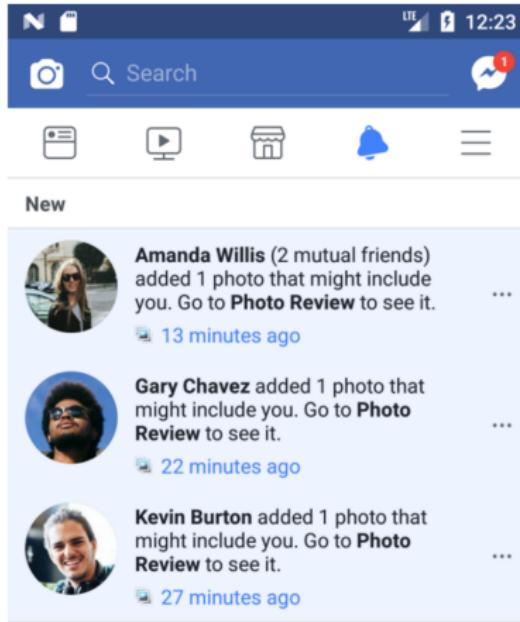


Bose 789564-0030 QuietComfort 35 Wireless Headphone (Blue)-Special Edition

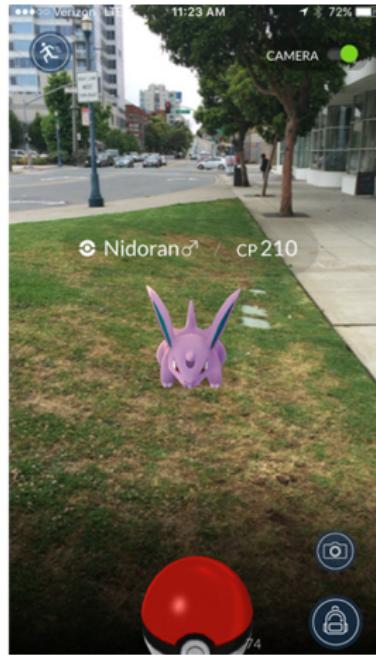
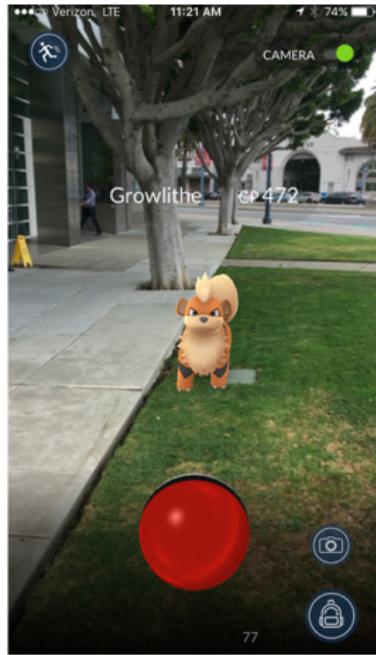
Add to Cart

Customer Rating	 (68)	 (349)	 (22)	 (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			
Colour	Black	Black	White	Blue
Connectivity Technology	Bluetooth wireless	Bluetooth Wireless	Bluetooth Wireless	Bluetooth Wireless

Applications: Image Recognition

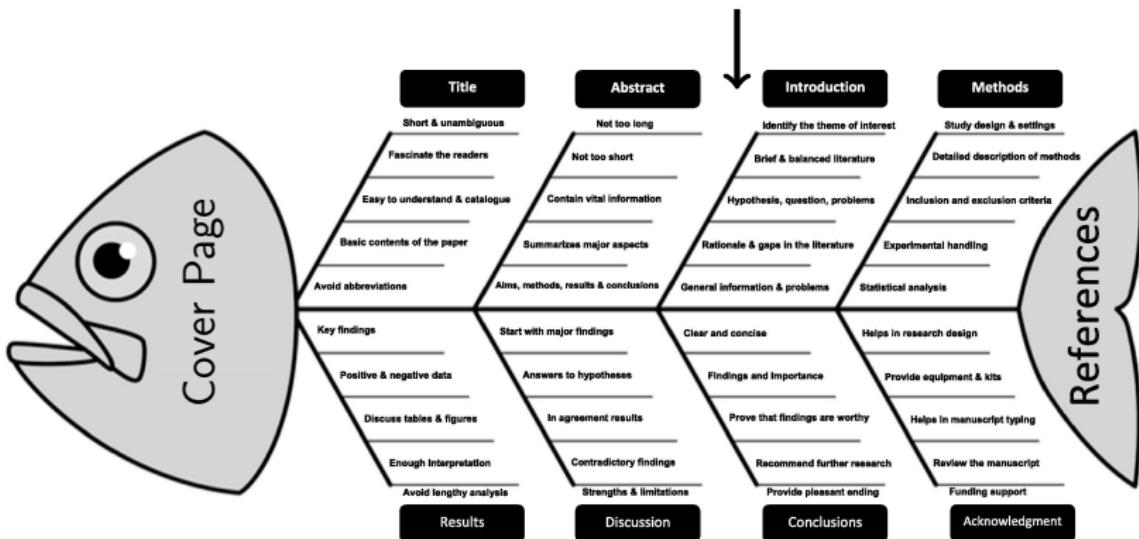


Applications: Gaming



Report Structure

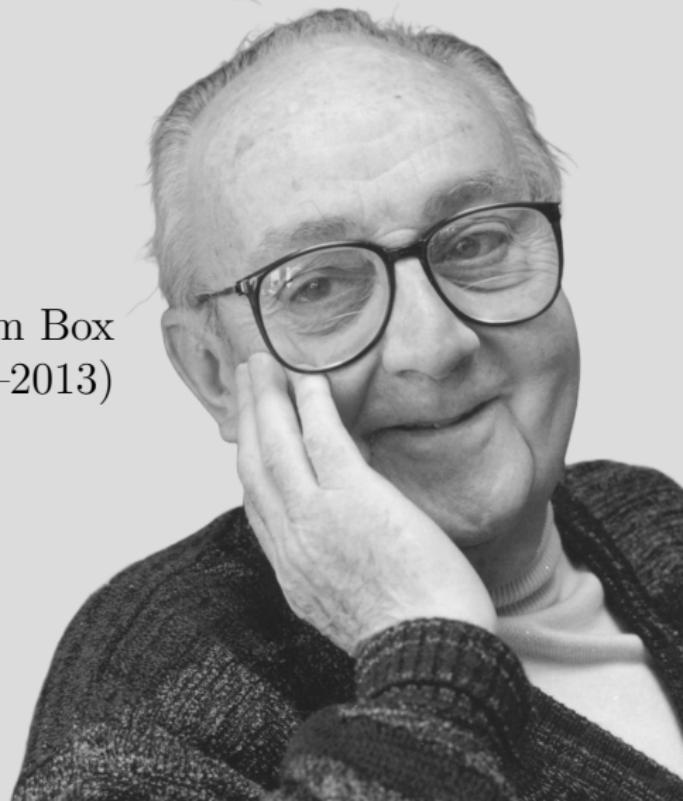
(+ToC)



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?

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

Review Questions

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

- The empirical findings
- R Squared
- The conclusion
- The contributors
- 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

Module 5

Data Science People

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

Module 5.1: Review questions

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 ?

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