
# INTRODUCTION TO DATA SCIENCE

# Agenda

1. **Data Science**
   - The data flow
   - Who is a data scientist?
   - Industry use cases

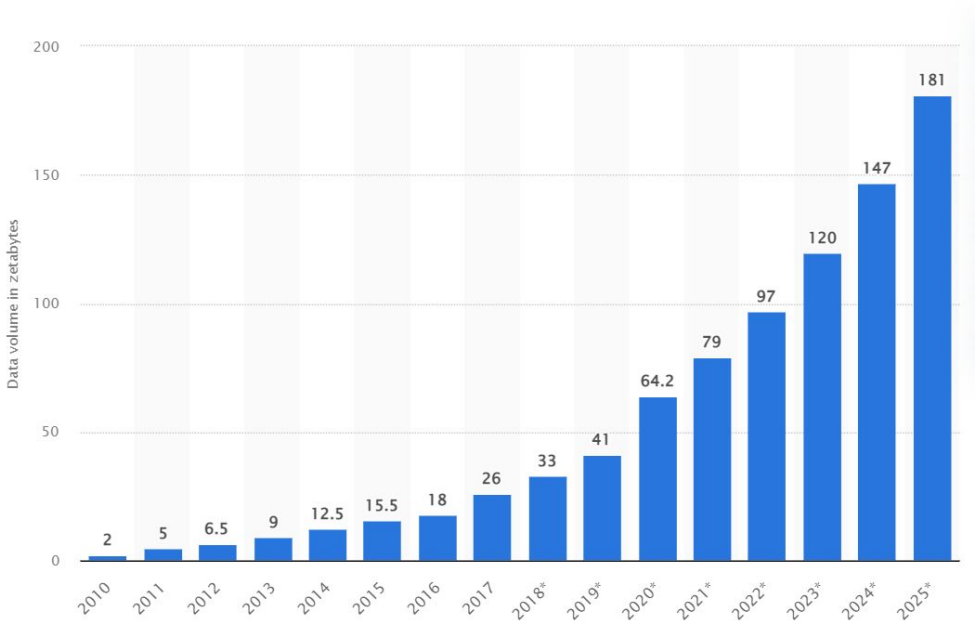
2. **Definitions and Process Flows**
   - AI or ML or DS or DL

3. **Problems**
   - Case Studies
   - Most common Machine Learning problems
   - ML techniques used to solve business problems

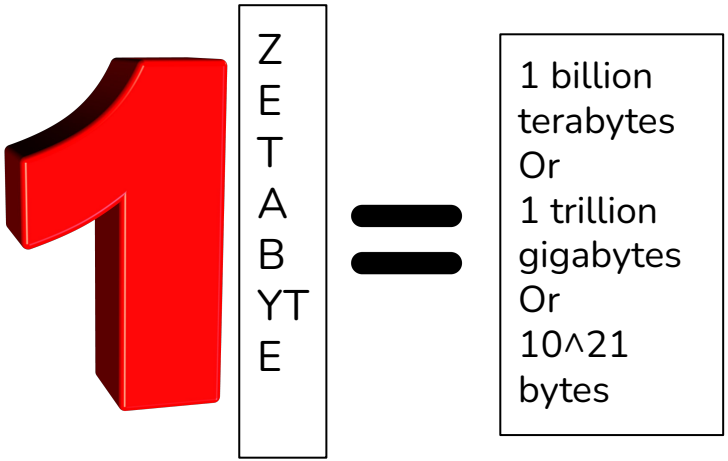
4. **Learning Focus**
   - Program Outcome
   - Next Steps!

# Why are we talking Data Science?

## DATA GROWTH




Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025



# What is Data Science?

Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

- IBM

# Who is a Data Scientist?

## Data Scientist

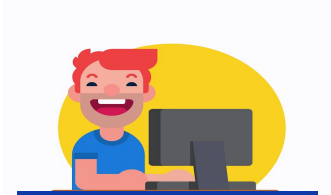
A data scientist uses machine learning and statistical concepts to make predictions and answer key questions. They may need to work with different data types.

Skill sets - Machine Learning and Deep Learning, data mining, statistical analytical skills, data optimization and soft skills

## Data Analyst

A data analyst role is associated with analyzing the trends and patterns on existing data and helping stakeholders make better decisions

Skill sets - SQL and database knowledge, reporting and data visualization, data warehousing



## Data Engineer

A data engineer develops, constructs, tests and maintains database architectures, deploys models and ensures data integrity

Skill sets - Data warehousing and ETL, data architecture, pipeline, scripting

# Skills required for a Data Scientist

## Math and Statistics

- Experimental Design
- Linear Algebra & Calculus
- Machine Learning
- Supervised Learning
- Unsupervised Learning
- Model Optimization
- Deep Learning

## Programming Skills

- Knowledge of Programming languages like R & Python
- A basic understanding of Database Languages like SQL

## Communication and Visualisation

- Ability to convey insights to senior leadership
- Art of storytelling
- Effective visualization skills
- Ability to collaborate with other team members
- Data to decisions

## Domain Knowledge

- Operational settings, functions and technologies specific to an industry
- Business acumen for that domain

# Industry - Use Cases

# Various Business Use Cases

## Consumer Goods

- Demand Forecasting
- Marketing Personalization
- Cross-sell and upsell
- Pricing and Promo Effectiveness

## Banking / Finance

- Credit Risk Assessment
- Fraud Detection
- Insurance / Loan Underwriting
- Churn and Retention Modeling

## Telecom

- Next Best Offer
- Churn and Retention Modeling
- Lifetime Value Prediction
- Product Optimization

## Energy

- Logistics Optimization
- Predictive Maintenance
- Outage Detection
- Project Risk Management

# Consumer Goods Use Cases

**Consumer Goods**

| | |
|---|---|
| **Demand Forecasting** | • Inventory Planning<br>• Production Planning |
| **Marketing Personalization** | • Personalized Advertising<br>• Personalized Offers / Discounts |
| **Cross-sell & Upsell** | • Selling a long term membership plan, get 1 month at $9 and yearly membership at $99<br>• Amazon: Customer who bought these also bought these (Buying a phone cover with mobile phone) |
| **Pricing & Promo Effectiveness** | • Marketing budget allocation to get maximum profits<br>• Holiday or special day offers |

# Banking / Finance Use Cases

**Banking / Finance**

| | |
|---|---|
| **Credit Risk Assessment** | • Generating customer credit score indicating risk based on profile and past financial activity |
| **Fraud Detection** | • Automatic Credit card fraud detection<br>• Preventing money laundering |
| **Insurance/ Loan Underwriting** | • Tailoring insurance premium as per customer's risk profile which is calculated using data from different sources |
| **Churn & Retention Modeling** | • Identifying customers who are at risk of churning so they can acted upon<br>• Keeping old & disengaged customers loyal using targeted communication |

# Telecom Use Cases

**Telecom**

| | |
|---|---|
| **Next Best Offer** | • Next Best Offer (NBO) automatically evaluates and recommends one or more customized offer from a diverse offer catalogue that will most likely satisfy a customer |
| **Churn & Retention Modeling** | • Identifying customers who are at risk of churning so they can acted upon<br>• Keeping old & disengaged customers loyal using targeted communication |
| **Lifetime Value Prediction** | • Segment and target customers based on their value - present value of all future cash flow generated during his/her business with organization |
| **Product optimization** | • Providing best product to customers on the basis of usage, feedback etc |

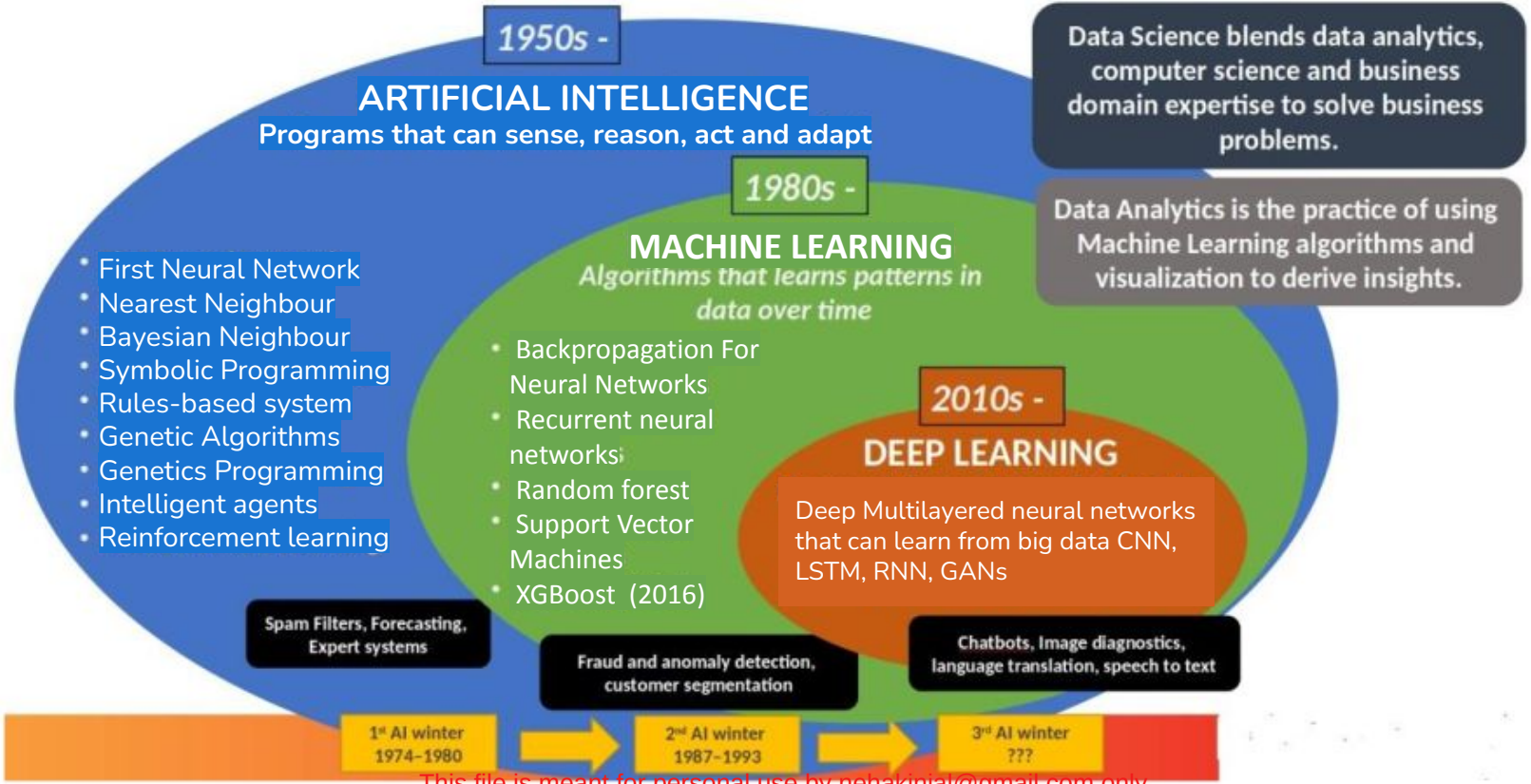
# Energy Use Cases

Energy

| | |
|---|---|
| Logistics Optimization | • Route optimization<br>• Demand predictions |
| Predictive Maintenance | • predicting remaining useful life time<br>• predicting failure in given time period |
| Outage Detection | • Anomaly detection |
| Project Risk Management | • Risk estimation |

# Definitions - AI / ML / DL

# The Evolution of AI



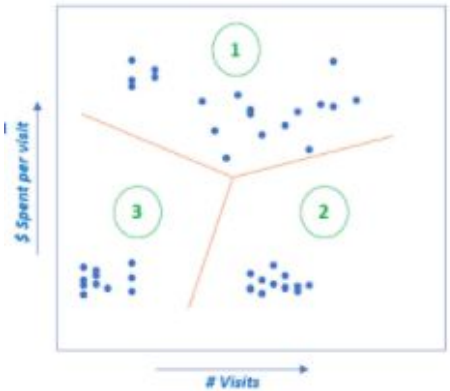
# Data Science - Algorithms

# Types of Machine Learning Problems

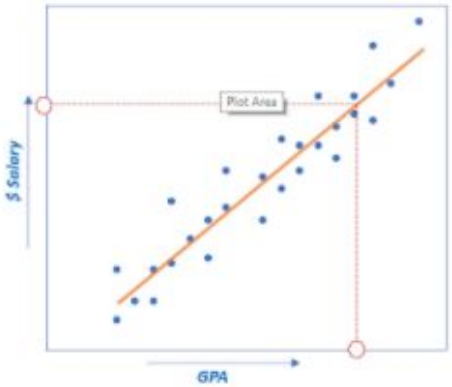
HYPOTHESIS TESTING

Is A > B ?

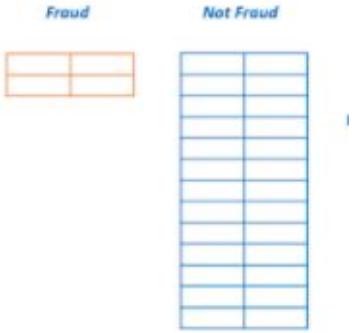
CLUSTERING




REGRESSION




CLASSIFICATION



# ML Techniques for Business Problems

## Clustering

Marketing Personalization

## Regression - Prediction

Demand Forecasting

Lifetime Value Prediction

Predictive Maintenance

Insurance / Loan Underwriting

Workforce Planning

Logistics Optimization

## Regression - Drivers

Pricing and Promo Effectiveness

## Classification

Credit Risk Assessment

Fraud Detection

Product Optimization

Predictive Maintenance

Churn & Retention Modeling

Next Best Offer

Cross-sell and Upsell

Outage Detection

Attrition and Retention Modeling

Training Planning

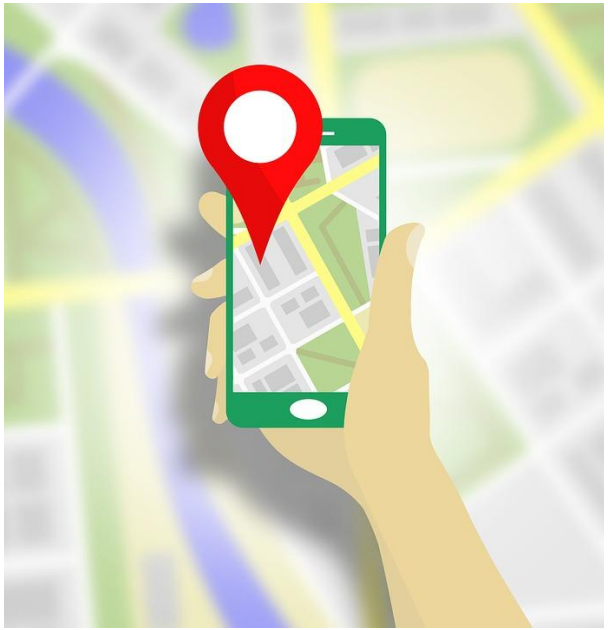
Project Risk Management

# Case Study: How Google uses Data Science

# Google Maps - Forecasting and Recommendations

**Google's biggest asset is data**, and lots of it! It collects data every day from a multitude of reliable sources, including your phone. If you have your location services on, Google will be fed with a stream of anonymous bits that speak of location, relative velocity and itinerary.

With the help of  data collected from drivers, passengers, and pedestrians, Google's Machine Learning algorithms are able to predict traffic jams, help you find your optimal route and even determine which areas should be avoided due to road works or accidents. Enormous amounts of historical data and cutting-edge algorithms from Network-based Data Science give Google the power to predict traffic patterns in a highly accurate fashion.

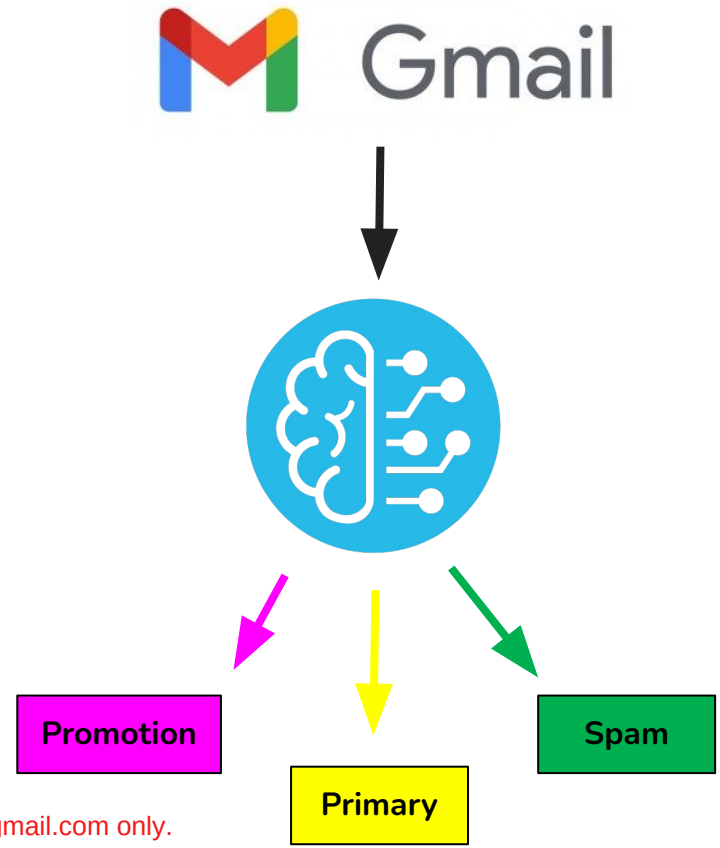


# Gmail - A Classification Problem

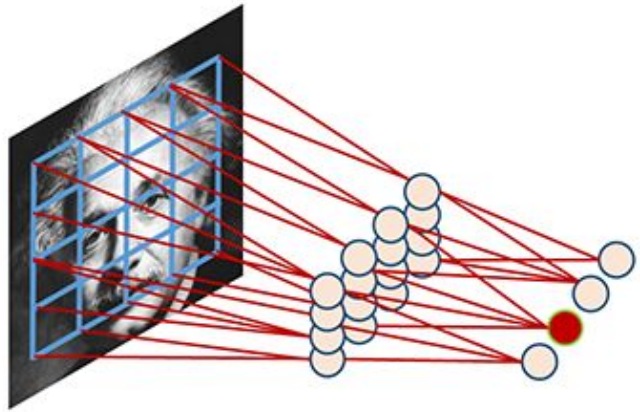
Do you know how the mails are categorized as social, promotion or primary are **automatically labelled**?

Gmail is a good example of how Data Science and Machine Learning is being used to understand the text data in emails and classify them in an automated way.

Smart Reply and Smart Compose are also examples of using Machine Learning for classification to automatically generate replies to emails or predict the next words you wish to type while composing emails.



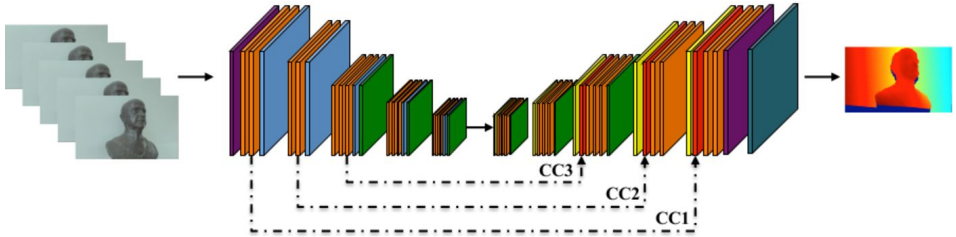
# Deep Dream - A Computer Vision Application



**DeepDream** is a Computer Vision oriented Deep Learning program created by **Google** engineer Alexander Mordvintsev.

It uses Convolutional Neural Networks to find random patterns in various images and amplifies them in different artistic ways.

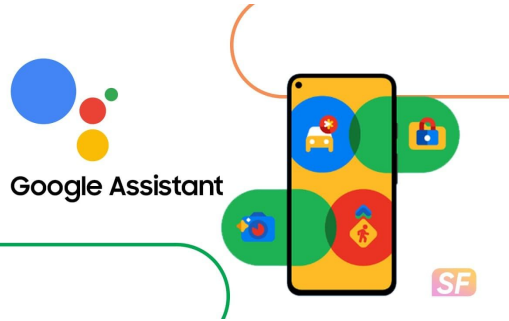
These images can be tweaked in any possible manner using the input data and various parameters so that the results obtained can be funny or weird.



# Google Assistant - Natural Language Processing

Google Assistant is built based on ***Natural Language Processing***, the Data Science domain that deals with machine understanding of text and audio data.

1. A local device that has Google Assistant enabled **first receives your speech**.
2. Google breaks down what you said into individual sounds. It then consults a database containing various words' pronunciations to **find which words most closely correspond to the combination of individual sounds**.
3. It then **identifies key words to make sense of the tasks** and carry out corresponding functions. For example, if Google Assistant notices words like *"weather"* or *"temperature"*, it would open the weather app.
4. **Google's servers send this information back to your device and the Google Assistant will respond**. If the Google Assistant needs to say anything back to us, it would go through the same process described above, but in reverse order.



# Case Study: How Airbnb uses Data Science

# Airbnb analyzes customer feedback text with Natural Language Processing

- Airbnb witnessed a 43,000 % hyper-growth in as little as five years, using several business solutions primarily based on Data Science and Machine Learning.

- One of the ways Data Science is being used is in the analysis of customer feedback and text using NLP. This text is analyzed to understand customer sentiment about a property, and the platform boosts listings with a higher percentage of positive reviews to maximize the likelihood of gaining customers.

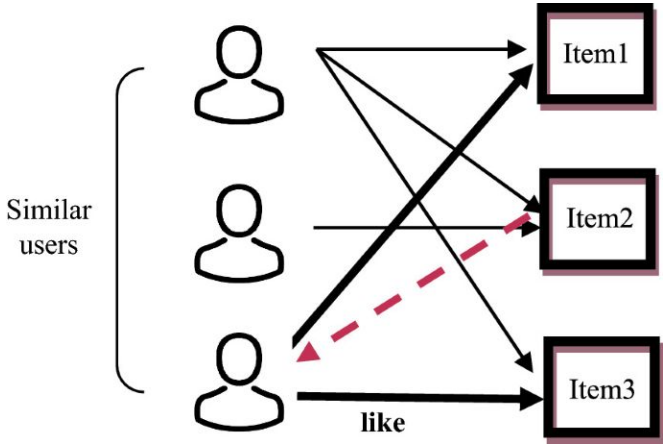
# Image Recognition and Analysis at Airbnb

On Airbnb, guests are likely to make a decision on going for a particular listing based on **whether it attracts their eye.**

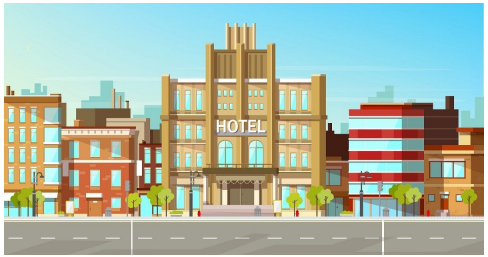
Airbnb in fact has algorithms that automatically suggest homes and accommodation that **visually look similar** to previous listings you have browsed or booked, in order to boost engagement and utilization. This is accomplished with Computer Vision & Deep Learning techniques.

# Recommendation Systems at Airbnb

Airbnb utilizes an algorithm called Collaborative Filtering, which is one of the basic approaches in Recommender Systems. This aims to give **good and reliable recommendations to a user based on what has worked for other similar users.**




Using collaborative filtering, the users (hosts) and the items (trips) data can be used to understand user preferences for items by combining historical ratings through statistical learning.



# Learning Focus

# Program Outcomes

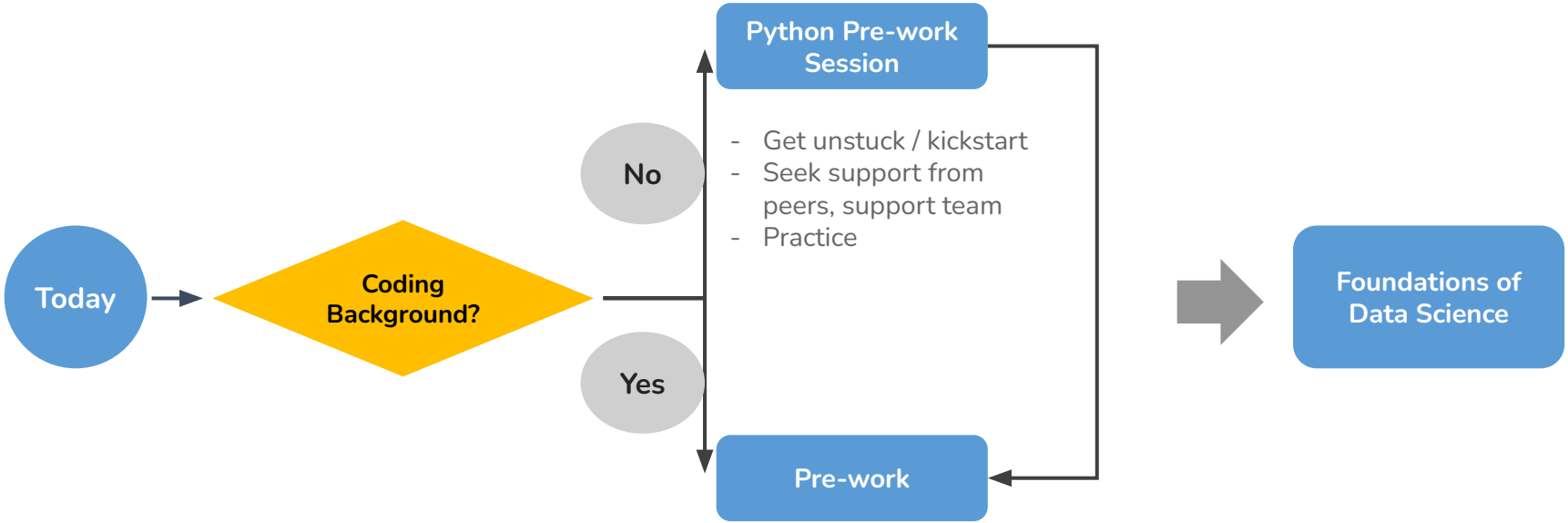
Ability to solve complex business problems

Applied knowledge of Data Science techniques – what, when & how

Ability to tell stories from data

Ability to work with complex data to generate business insights and deliver recommendations

# Next Steps



greatlearning
*Power Ahead*

# Happy Learning !