Welcome to Week 1

We’ll learn about crucial fundamentals that will prepare you for machine learning techniques.

Introduction to Data Science

Why am I learning data science?

Why Data Science?

- Lifelong learning
- Opportunity to be a detective
- Ability to directly impact business strategy
- Relevant across so many industries:
  - Diagnosing cancer
  - Fighting global warming
  - Annoying ads
  - City planning
  - The list goes on….

Why Data Science?

Current demand 10,845 “data scientist” positions on Indeed.com 1/2020

The idea of moving to a “data-driven” or “data-centric” business is a hot topic (with huge revenue potential) and requires data scientists.

Exciting! The amount of data is increasing at a rapid pace, and the tools are constantly being created and to manage this!

As a data scientist I have always felt respected, my opinion valued, and the work fulfilling. Reasonable working conditions/flexibility.

Introduction to Data Science

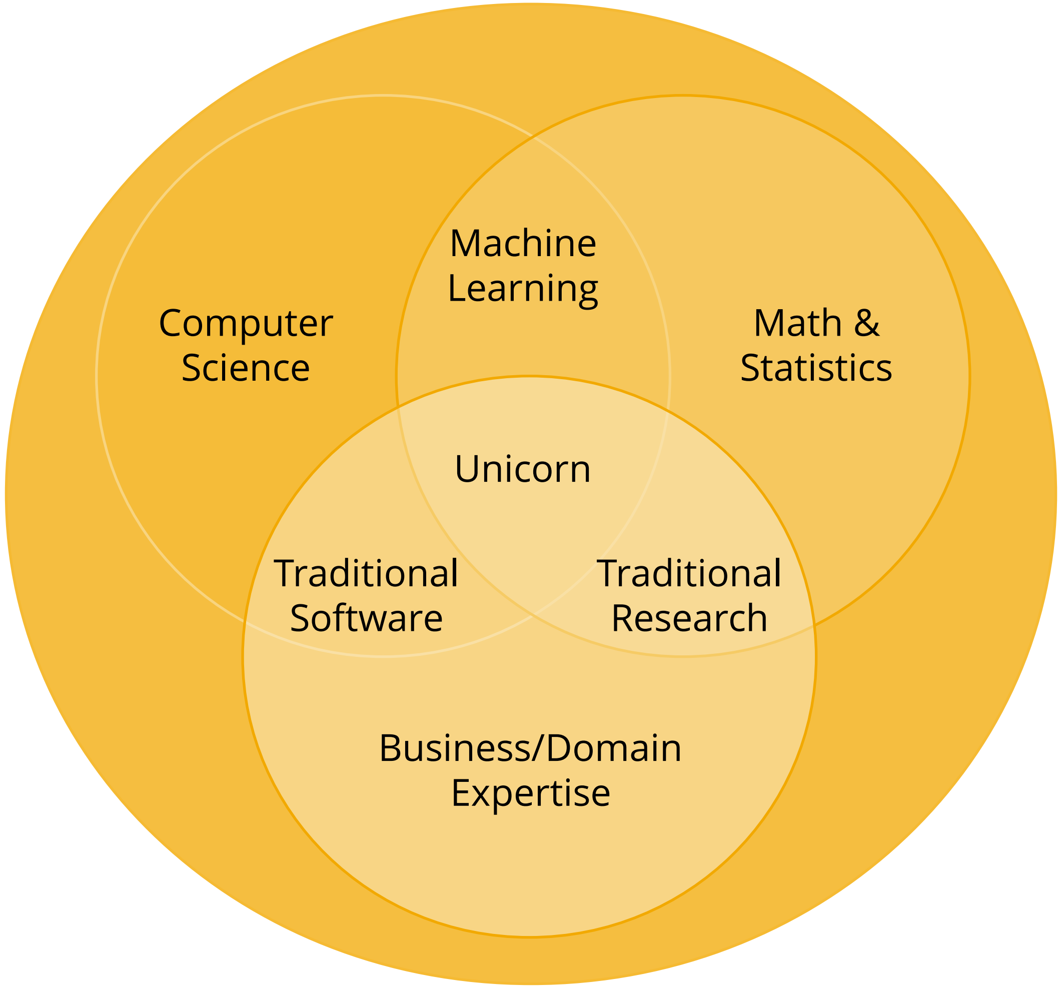
What is data science?

What is Data Science?

Term “data scientist” was coined in 2008.

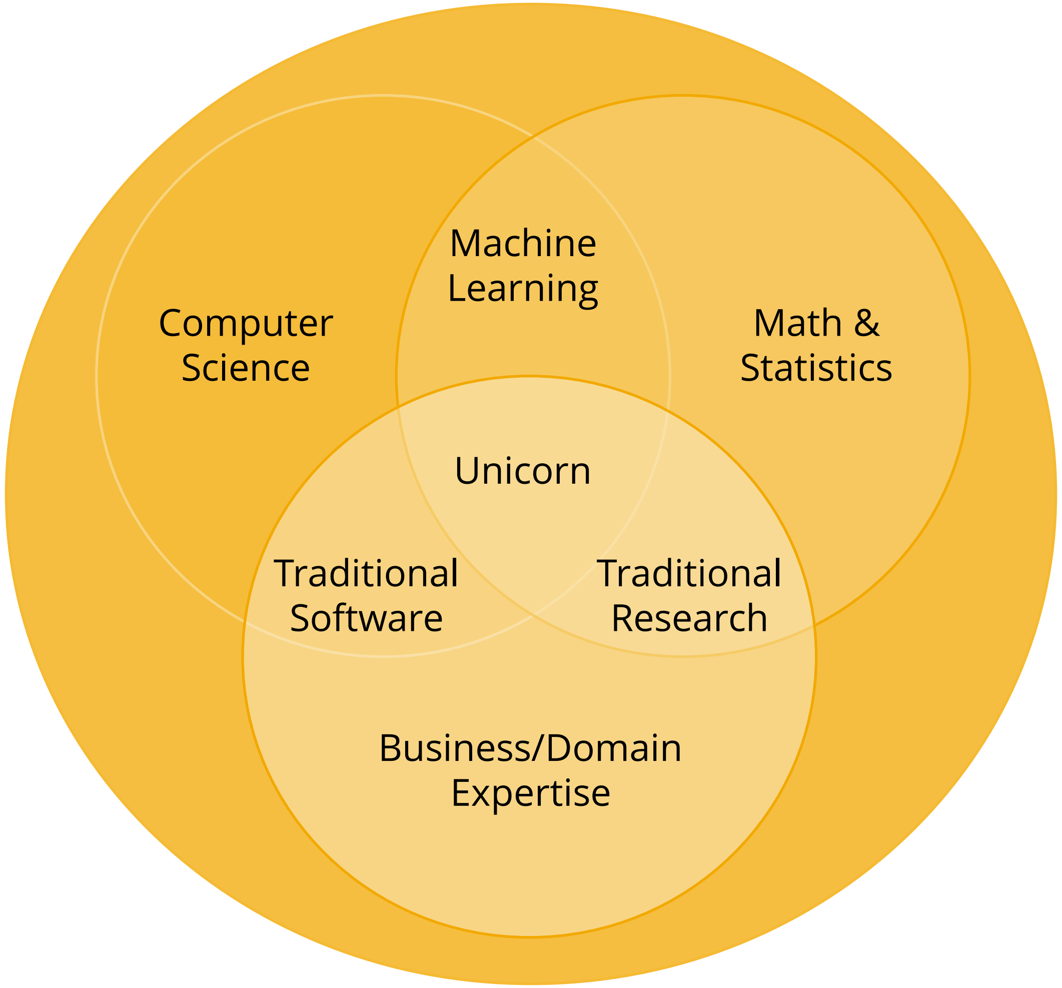
It continues to evolve as tools advance. (8 years ago you might have trained a neural net over the weekend, and that might now take an hour to train, this obviously changes what the day to day job looks like)

What is Data Science?



My definition: It is the understanding and utilization of tools, data and methodologies that enable you to effectively solve problems utilizing data

What is Data Science?



People who have different strengths can all be data scientists. Industry is realizing that "unicorns" do not exist. Data Science is a multidisciplinary field combining math, stats, programming, analysis and business acumen.

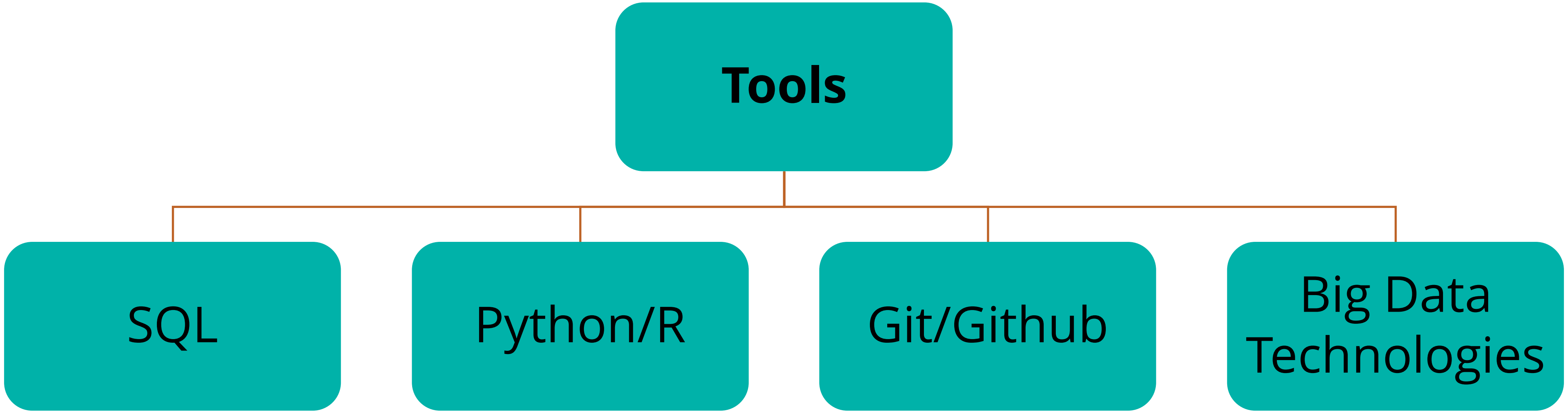
Summary

Data Science is the understanding and utilization of tools, data and methodologies that enable you to effectively solve problems utilizing data

Introduction to Data Science

What are the data scientist’s tools?

Data Scientist’s Toolbox: Tools



Tools

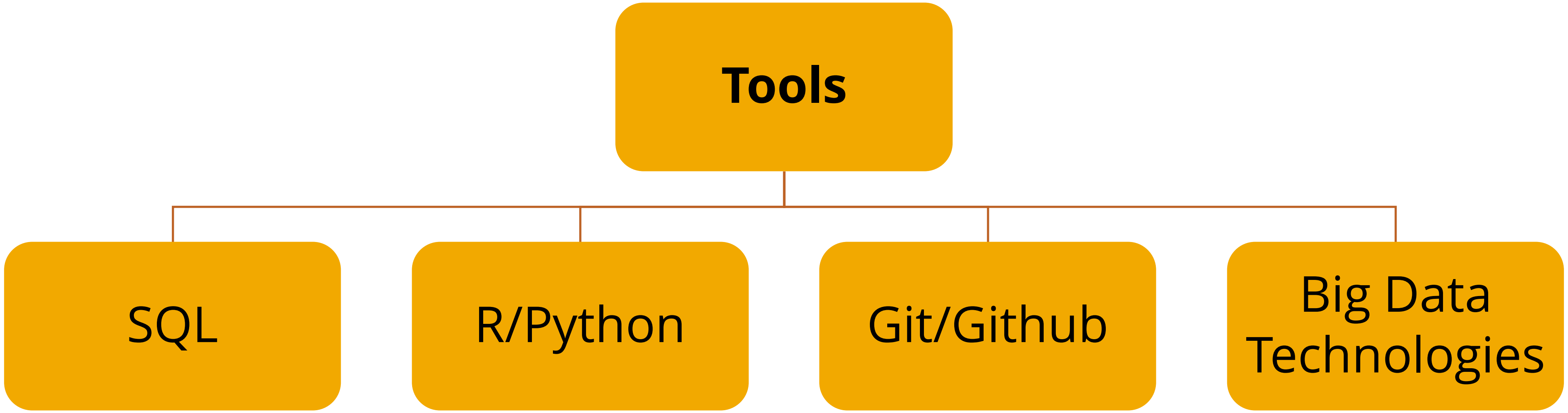
SQL

Python/R

Git/Github

Big Data Technologies

Data Scientist's Toolbox: Tools



Tools

SQL

R/Python

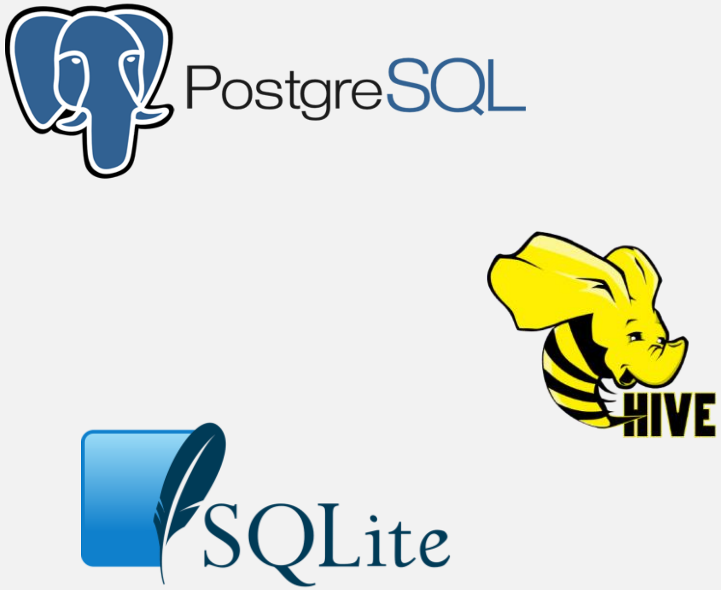
Git/Github

Big Data Technologies

- Tools will change over time
- Data Lakes were cool, now they're considered "Data Swamps"
- Julia and other languages are gaining in popularity as well. It's important that you learn a language. You do not need to learn all of them.

Data Scientist's Toolbox: Tools

SQL | R/Python | Github/Git | Big Data Technologies



- Query language for relational databases
- Oracle, Postgres, MySQL, SQLite, Microsoft SQL Server
- You don't need to learn them all
- Hive is also similar to SQL

Data Scientist's Toolbox: Tools

SQL | R/Python | Github/Git | Big Data Technologies

- Should I learn Python or R?
- The answer is "Yes". You should learn Python or R, or another language with similar capabilities.
- 55% of job description for "Data Scientist" titles list "Python or R" in the description.

Data Scientist’s Toolbox: Tools

SQL

R/Python

Github/Git

Big Data Technologies

- Versioning is imperative in being able to collaborate on code.
- No one wants to see file names with _v45.xlsx on the end of them.
- No more emailing people code or a query, you’ll be sure to have the most recent version in Github.

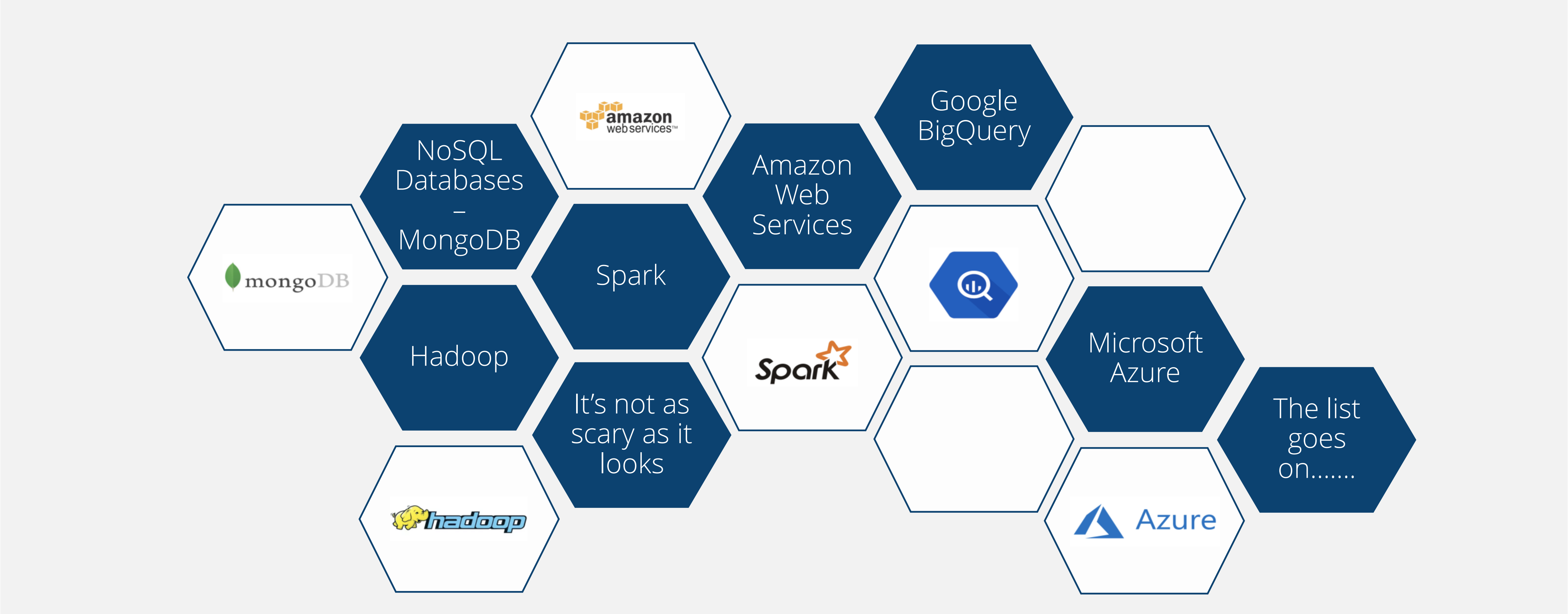
Data Scientist’s Toolbox: Tools

SQL

R/Python

Github/Git

Big Data Technologies



NoSQL Databases – MongoDB

Spark

Amazon Web Services

Google BigQuery

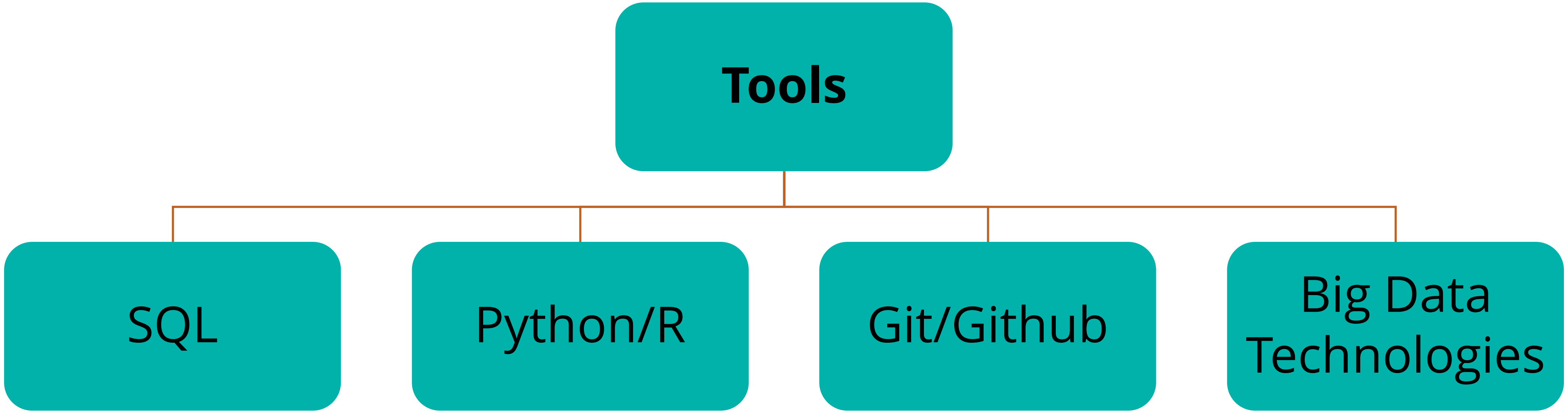
Hadoop

It’s not as scary as it looks

Microsoft Azure

The list goes on…….

Summary



Tools

SQL

Python/R

Git/Github

Big Data Technologies

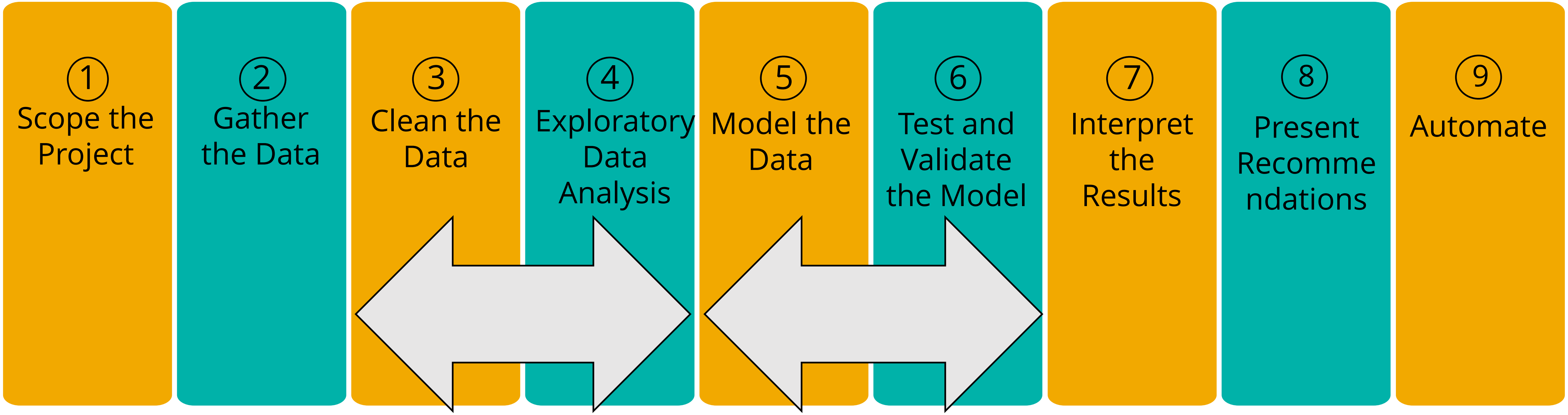
Python for Data Science

Week 1: Introduction to Data Science

Introduction to Data Science

What is the data science lifecycle?

Data Science Pipeline & Project Lifecycle



① Scope the Project

② Gather the Data

③ Clean the Data

④ Exploratory Data Analysis

⑤ Model the Data

⑥ Test and Validate the Model

⑦ Interpret the Results

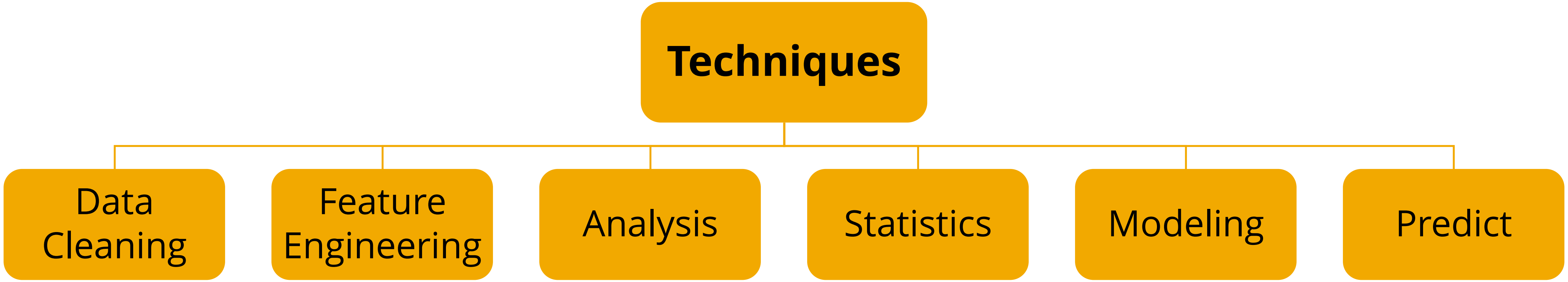
⑧ Present Recommendations

⑨ Automate

Most pipelines in blogs/online leave out the first step and the last two.

Don't let the straight diagram fool you, there is often plenty of back and forth between steps in this process.

Data Scientist’s Toolbox: Techniques



Techniques

Data Cleaning

Feature Engineering

Analysis

Statistics

Modeling

Predict

- The tools are required to implement these different techniques.
- People often ask “how important is statistics?” or “how important is it to know how to code?” No one will be an expert in every area.

Data Scientist’s Toolbox: Techniques

Data Cleaning

Feature Engineering

Analysis

Statistics

Modeling

Predict

- Data cleansing (or data cleaning) and is the first step in data preprocessing (data preprocessing is the term for bringing you from step 1 i.e. data cleaning to a training set that is ready for modeling)
- Validating accuracy – Does the data match the column label? Are there negatives for entries that should be positive? String entries in columns that should be numeric? Etc.
- Are there duplicate rows?
- Handling missing data

Data Scientist’s Toolbox: Techniques

- Data Cleaning
- Feature Engineering
- Analysis
- Statistics
- Modeling
- Predict

- Using domain knowledge to create new columns (called features) that are relevant and useful for machine learning
- Example: You have a customer that has purchased 400 times and another customer that has purchased 20 times. Would I put this directly into the model? I’d probably want to scale it by how long the person has been a customer #purchases/Length tenure. This would be a new feature.
- One hot-encoding or aggregating data in a particular way based on the context. When you aggregate data you lose some information, but if I had a categorical variable with 4,000 categories it may be easier to pick up significance in a model if I grouped them (or some of them) logically.

Data Scientist’s Toolbox: Techniques

- Data Cleaning
- Feature Engineering
- Analysis
- Statistics
- Modeling
- Predict

- Analysis is breaking a complex topic or dataset into smaller parts to make sense of it. Can be really deep or can be a simpler more high-level analysis
- Being able to answer a question effectively leveraging data. Typically, visualizations and charts that tell the story of the data.
- The “A” in EDA ;)
- Finding the root cause of errors that are present in your data

Data Scientist’s Toolbox: Techniques

- Data Cleaning
- Feature Engineering
- Analysis
- Statistics
- Modeling
- Predict

- Data scientists can be the subject matter experts on designing, scoping and analyzing hypothesis tests in their organization (if that is their skillset). Consistency across the organization is important!
- Understanding the appropriate methods for determining statistical significance helps to drive decision making
- Just as important, an understanding of statistics allows you to determine when a method or algorithm is NOT appropriate

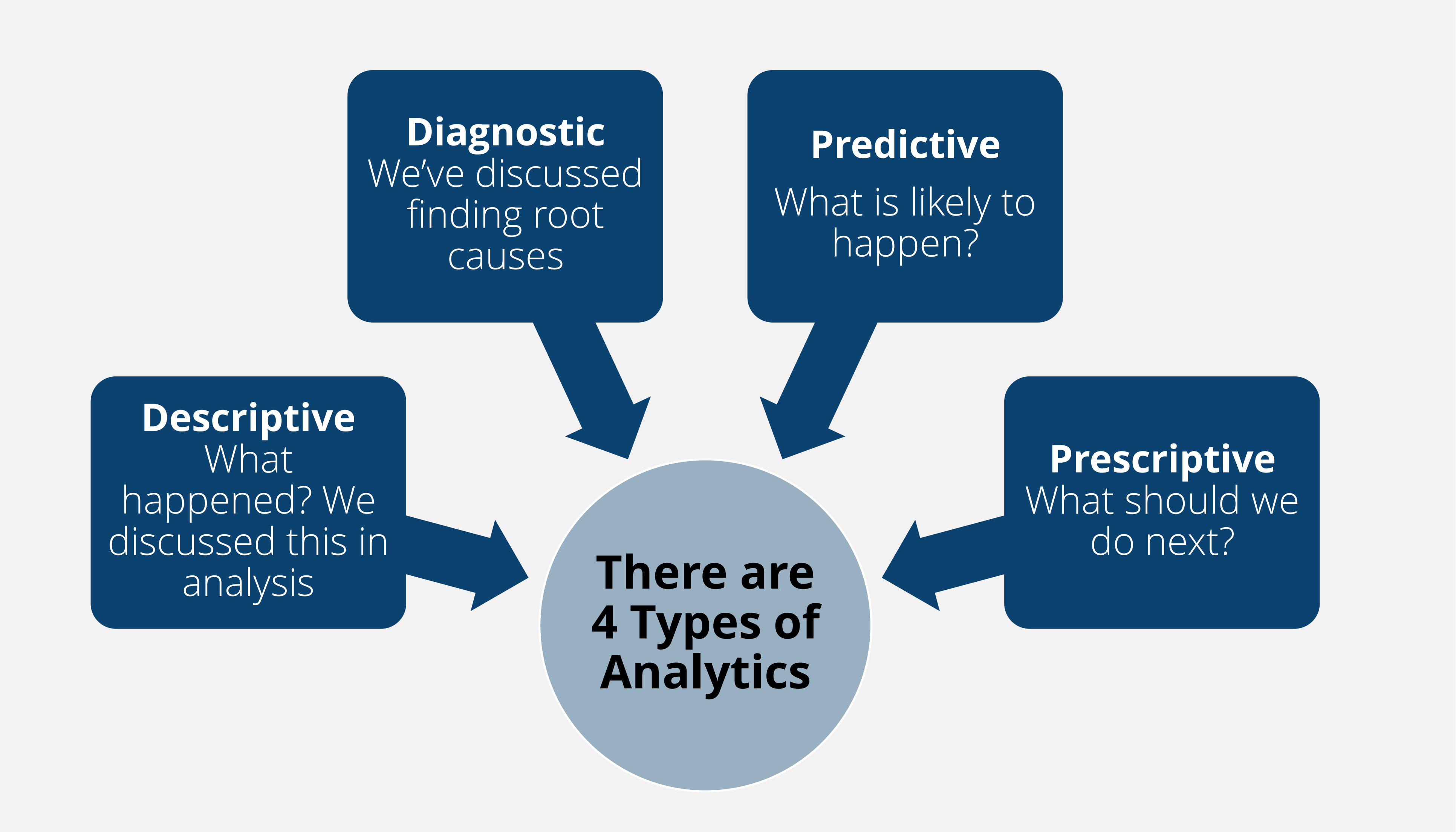
Data Scientist’s Toolbox: Techniques

- Data Cleaning
- Feature Engineering
- Analysis
- Statistics
- Modeling
- Predict

- There is no end to the number of algorithms you could learn. This is a lifelong learning process. You may gain a foundation in modeling, and then get into an industry that goes really deep in one particular area.
- New algorithms are being introduced
- When modeling, simple is preferred over complex if the simple solution performs well and is more interpretable

Data Scientist’s Toolbox: Techniques

- Data Cleaning
- Feature Engineering
- Analysis
- Statistics
- Modeling
- Predict



Descriptive
What happened? We discussed this in analysis

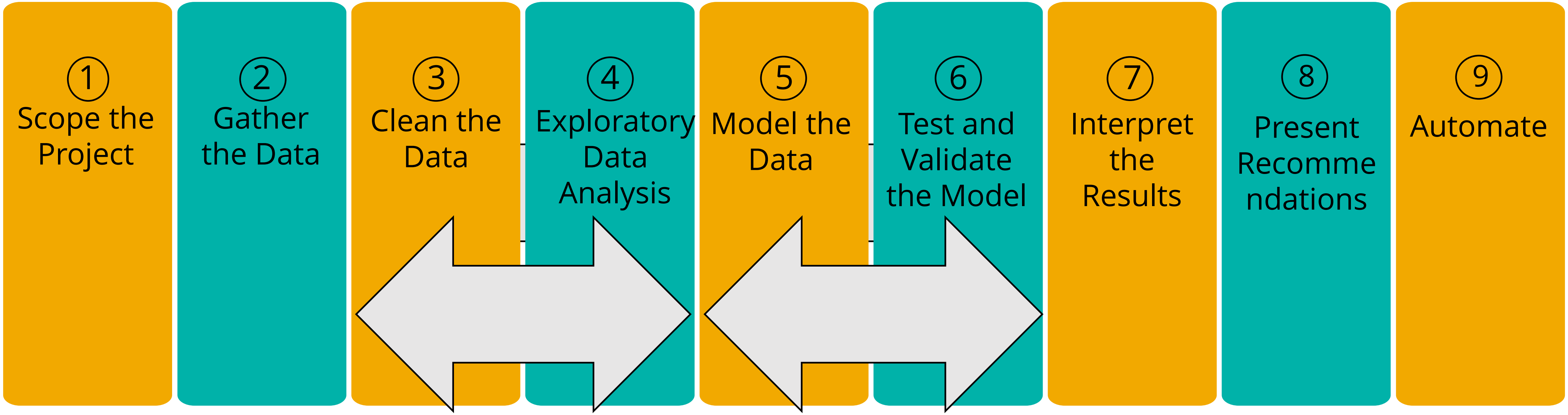
Diagnostic
We’ve discussed finding root causes

Predictive
What is likely to happen?

Prescriptive
What should we do next?

There are 4 Types of Analytics

Summary



① Scope the Project

② Gather the Data

③ Clean the Data

④ Exploratory Data Analysis

⑤ Model the Data

⑥ Test and Validate the Model

⑦ Interpret the Results

⑧ Present Recommendations

⑨ Automate

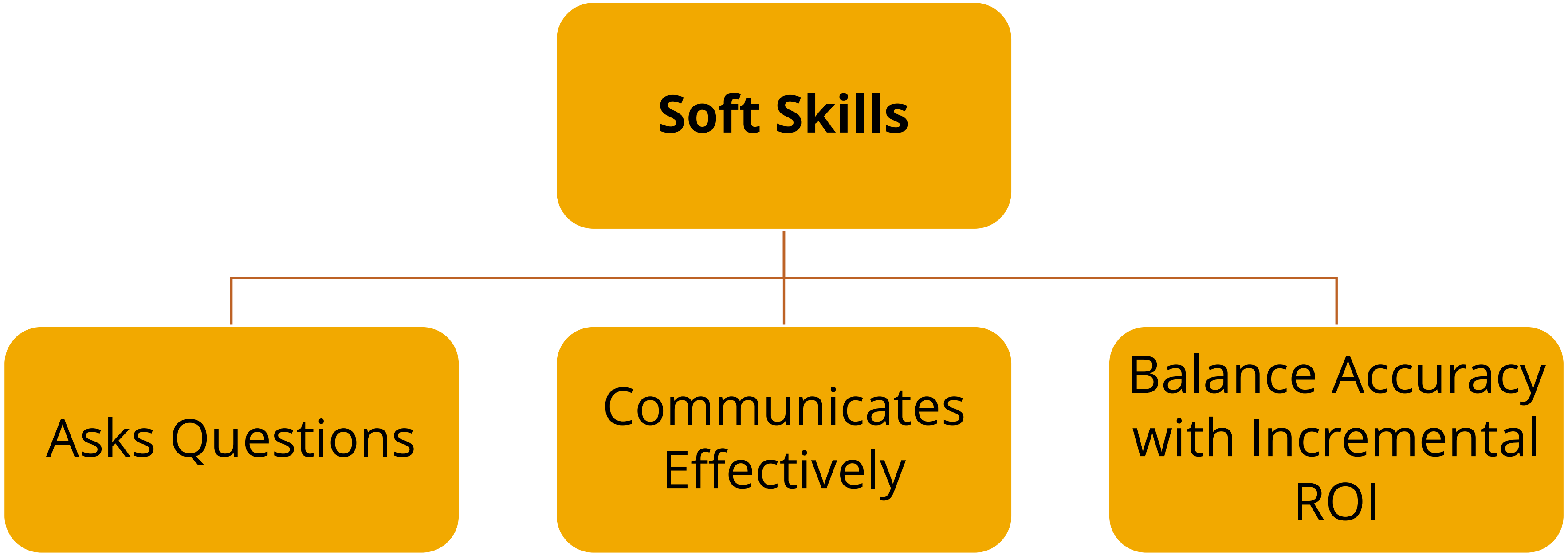
Python for Data Science

Week 1: Introduction to Data Science

Introduction to Data Science

How do I adopt a data science mindset?

Data Science Toolbox: Soft skills



Soft Skills

Asks Questions

Communicates Effectively

Balance Accuracy with Incremental ROI

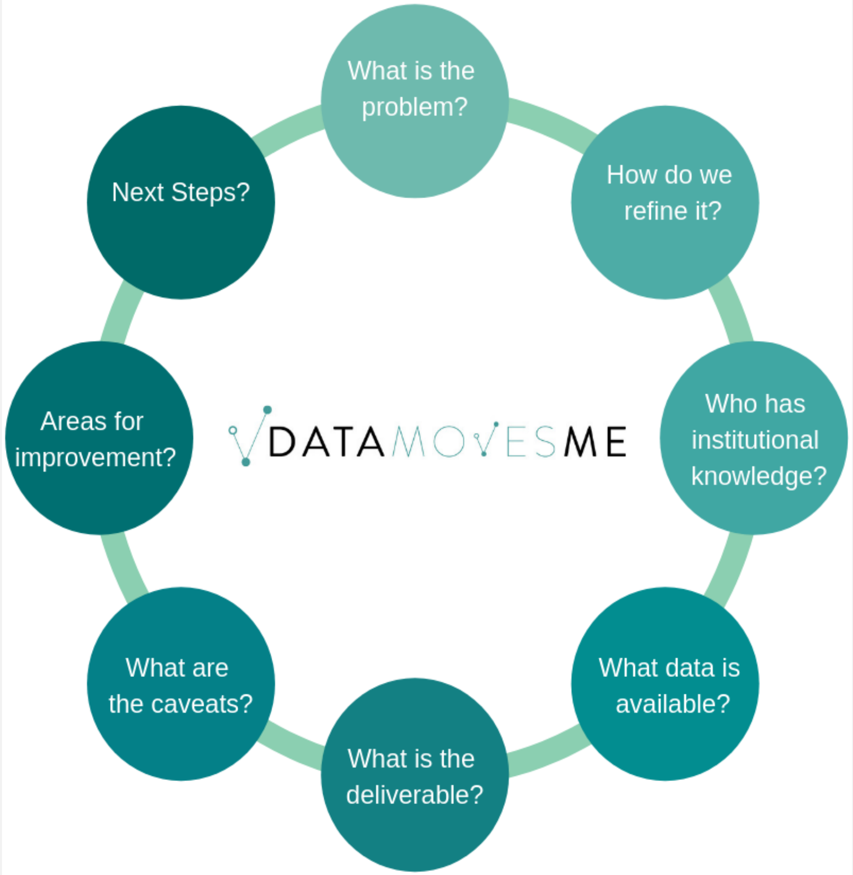
- Do not underestimate the importance of the soft skills.
- Cultural fit and ability to communicate effectively are crucial to both landing a job and then being successful in that role.

Data Science Toolbox: Soft skills

Ask Questions

Communicate Effectively

Balance Accuracy with Incremental ROI



- Data science is a constant collaboration with the business and a series of questions and answers that allow you to deliver the analysis/model/data product that the business has in their head.
- There are questions to be asked at every step in the process.

Data Science Toolbox: Soft skills

Ask Questions

Communicate Effectively

Balance Accuracy with Incremental ROI

- Yes, you’re going to build models. However, the audience you’re presenting to may be non-technical. Being able to communicate results in a way that your business colleagues will understand is imperative in being effective as a data scientist.
- This can truly differentiate you as a data scientist.

Data Science Toolbox: Soft skills

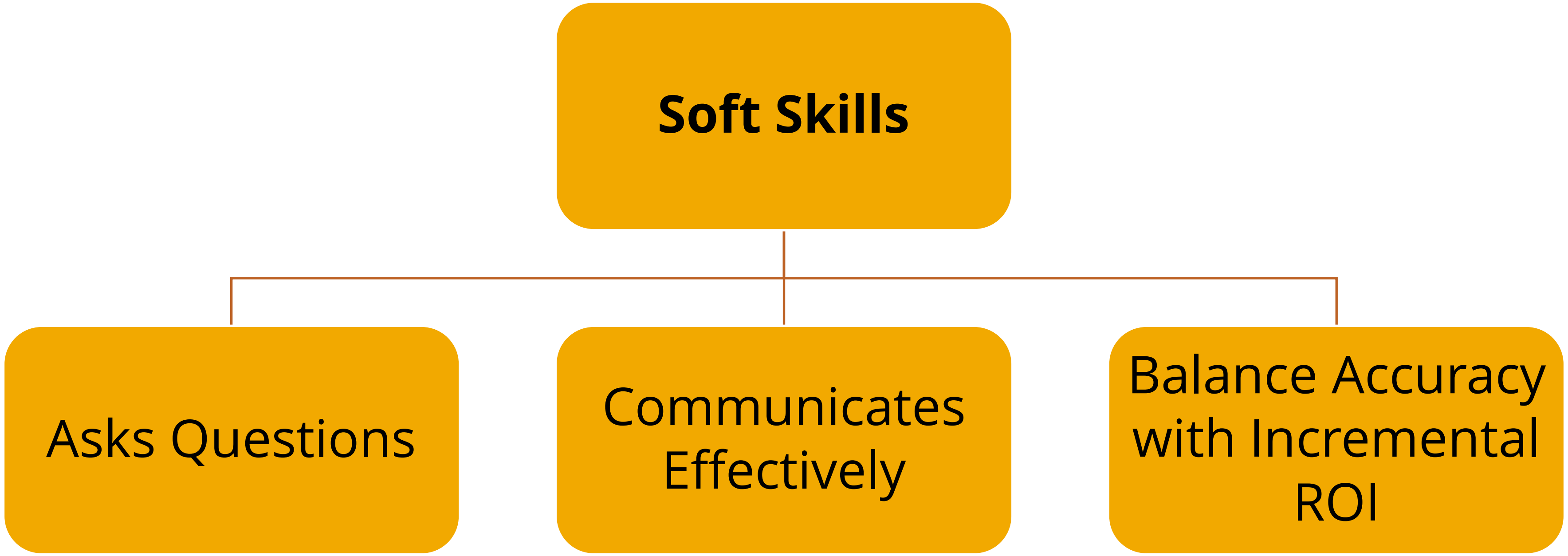
Ask Questions

Communicate Effectively

Balance Accuracy with Incremental ROI

- Knowing when to spend 3 weeks on an analysis or 30 minutes is a skill.
- This can truly differentiate you as a data scientist.

Summary



Soft Skills

Asks Questions

Communicates Effectively

Balance Accuracy with Incremental ROI

Python for Data Science

Week 1: Introduction to Data Science

Introduction to Data Science

What about collaboration, reproducibility and ethics?

Collaboration

Data science is a team effort collaborating with stakeholders and other members of analytics.

Reproducibility

Reproducibility saves time when there are follow up questions that require further analysis.

Allows you to show others that those were your true results. You wouldn’t want to have someone ask how you arrived at that number and your response be “well, I can’t replicate it”

Promotes collaboration and working as a team. If another project becomes higher priority, you’d be able to offload your project to a teammate without as much disruption.

Integrity -> Ethics

Integrity: Your data will always contain some element of bias. However, being fully educated on how to minimize these biases and properly do analysis will help you to be a data scientist with integrity. (if you don’t know how to do something properly, you can easily be led down the wrong path or provide inaccurate results).

Data integrity: Preserving the validity and accuracy of your data through checking, validation, and following up on errors you find in your data.

Summary

Collaboration

Reproducibility

Ethics