Institute of Data

2023

Data Science and AI

Module 0

Succeeding as a data scientist in industry

Agenda: Succeeding in Industry

- Introduction, definitions, purpose and objectives
- What do employers value and what do they complain about?
- Skills required and attitude to succeed in the industry
- Data Science process in industry
- Case study
- Summary, conclusions and call for action

Agenda: Succeeding in Industry

- **Introduction, definitions, purpose and objectives**
- What do employers value and what do they complain about?
- Skills required and attitude to succeed in industry
- Data Science process in industry
- Case study
- Summary, conclusions and call for action

Institute of Data

Ahmed Fattah

- I am an AI Architect, a Full-stack Data Scientist and a Data Science/ AI Lead Trainer with over 20 years of industry experience, primarily at IBM, leading end-to-end large-scale innovative AI, Data Science and analytics solutions.
- I have architected, led and overseen implementation of many large-scale enterprise integrated AI solutions across several industries, especially in the financial, telecommunication and government sectors.
- I have an extensive knowledge and industry experiences in AI, Data Science, Solution and Enterprise Architecture.
- I am an AWS and Open Group Certified Architect. I have contributed many technical papers to journals, blogs and industry conferences.

Working in industry versus research

- Working in the **‘industry’** refers to working or consulting for **commercial entities** in **competitive sectors** such as **financial services**, **telecommunications** or **retail**.
- **Competitive pressures** in these sectors **heighten expectations** from Data Scientists, make it imperative to track **Return on Investment (ROI)** for all projects and **accelerate the pace of work**.
- **Typical** ‘university’ Data Science **education does not prepare graduates** to effectively work in the industry. This is due to **focus on theoretical** topics and the lack of emphasis on **softer skills** such as communication, collaboration and stakeholders management.

Purpose and objectives of this presentation

- The purpose of this presentation is to share my experience in working as a Data Scientist in the industry with the aim to help you **maximise the value of this course**.

- Objectives of the presentation:

  - Describe **what is valued** in the industry

  - **Prioritise the skills** you should focus on

  - Help **you to get hired**

What do employers value?

- Employers value Data Scientists or other data professionals who **use their technical skills and experiences** to:
  - **Asking the 'right' questions**.
  - **Taking initiative** to deliver **business value**.
  - **Manage Stakeholders**' involvement, communication and effective **team work**.
  - Understand **industry**.
  - Participate actively in delivering **solutions in production**.

What do employers complain about?

- Data Scientists care only about **theory**.
- They treat every project as a **6-month 'PhD'.**
- They go down **rabbit holes**.
- They use **confusing language**.
- They **cannot** put solutions into **production**.

What should you do to meet these expectations?

- Focus on **business outcomes.**
- Be agile – effective **communication to stakeholders.**
- Understand the **business value** of projects.
- Use **simple models** and communicate in **business language.**
- Develop a small number of **effective and practical skills** and be prepared to learn on the job.

Agenda: Succeeding in Industry

- Introduction, definitions, purpose and objectives
- What do employers value and what do they complain about?
- **Skills required and attitude to succeed in industry**
- Data Science process in industry
- Case study
- Summary, conclusions and call for action

Skills of various roles in Data Science and AI

- There are a number of roles that are required to deliver Data Science/AI solutions.
- Some roles are closer to business while others are more technical.
- There is a growing demand for Data Scientists to be able to contribute directly to implementing systems in ‘production’.

	Data Engineer	ML/AI Engineer	AI Architect	Data Scientist	Business Analyst
‘Soft Skills’ Data-driven mindset, Communication, Collaboration, Critical Thinking, Creativity	◑	◑	●	●	●
Business Domain Knowledge	◔	◑	◑	◕	●
Software Engineering & Information Management	●	◔	●	◑	○
Programming	●	●	◑	◕	○
Math & statistics Linear Algebra, Calculus, Statistics	◔	●	◑	●	◔

○ Not Important ⟷ ● Very Important

Technical ⟷ Business

Data Science skills for industry

AI Architect/ Full stack Data Scientist

Data Engineer/ Machine Learning Engineer

Data Scientist

Software Engineering
Software Development Lifecycle, cloud computing

Machine Learning
Scikit learn, Keras

Data Analysis
Pandas

Computational Math and statistics
Numpy, Matplotlib

Soft skills
Data-driven mindset, communication, questioning, visualisation, stakeholder's management, *domain knowledge*

Programming

Statistics

Math

Data Science skills

Foundational skills	Core Data Science and AI skills	Applying Data Science in industry
• Programming for Data Science (Python) • Maths and Statistics for Data Science	• Exploratory Data Analysis (EDA) and data wrangling • Data visualisation • Database access • Application Programming Interfaces (APIs) • Supervised learning (Regression and Classification) • Unsupervised learning (Clustering and Dimensionality reduction) • Deep learning • Natural Language Processing (NLP) • Artificial Intelligence • Cloud computing • Machine learning deployment • Data science industry practices	• Applying data science on different data structures and domains • Defining a data science project • Designing a data science project • Delivering data science project • Optimising machine learning model algorithms • Overall end-to-end solution • Presenting to stakeholders and obtaining buy-in • Capstone project

Soft skills

Consulting, Questioning, Critical Thinking, Problem Solving, Documenting, Presenting

Learning how to learn effectively framework

Minimal Viable Learning (MVL), Multimodal learning, Learn-Create cycle

Data Science skills for industry

- **Foundational skills** that are required to learn Data Science:
  - Programming
  - Math, Statistics
  - Basic software engineering
  - Soft skills

Data Science skills for industry

- **Core** Data Science skills
  - Computational math and statistics
  - Data Analysis
  - Machine Learning
- **Complementary** Data Science skills
  - Business domain knowledge
  - Software Engineering
  - Soft skills
    - Data-driven mindset
    - Critical Thinking
    - Communication
    - Curiosity

Programming Data Science in Python

Programming is:

- the **process of creating a set of instructions** that tell a computer how to perform a task
- thinking **systematically and critically**
- breaking a task into steps. Examples include: a recipe, directions to a destination and mathematical problem solving

Python has a very **active community** with a vast selection of **libraries**, especially in scientific computing, data analysis and visualisation which makes it **very suitable for Data Science**.

There are a number of tools available to support the development of Python.

Jupyter notebook has emerged as an effective way to develop and share Data Science projects.

Visual Studio Code (VSC) is an alternative for developing reusable software modules.

Programming (**computational mathematics and statistics**) can be crucial for developing deep mathematical and statistical knowledge and skills.

Why is Statistics important for a Data Scientist?

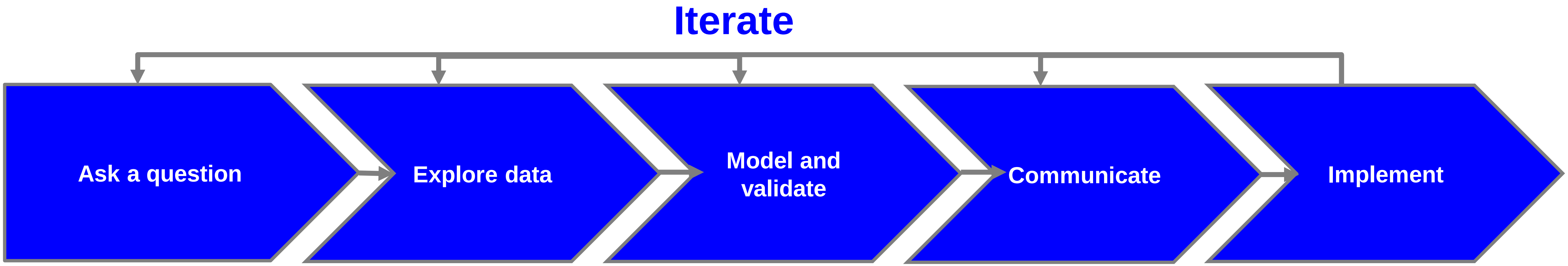
- **Statistical Thinking** is an essential component of a data-driven mindset which is crucial for a Data Scientist
  - Statistical analysis must start with the appropriate **data** (sample)
  - Statistical Inference (reasoning) should start with measurement, ideally, via **controlled experiments**
  - Statistics uses samples (a small subset of the population) and therefore always has a degree of **uncertainty.**
  - Sampling must be **random, and preferably, independent.**

- The best way to learn statistics is by **experimenting with data using Python code and visualisation.**

Agenda: Data Science Industry Experiences

- Introduction, definitions, purpose and objectives
- What do employers value and what do they complain about?
- Skills required and attitude to succeed in the industry
- **Data Science process in industry**
- Case study
- Summary, conclusions and call for action

Data Science Process



Ask a question
- Ask a (business/ data) question
- Define objective, strategy, value and effort budget.
- Translate business question to a data question.

Most important step!

Explore data
- Identify and collect, clean and transform data
- Explore data
- Ascertain quality and ability to answer questions

May consume large proportion of the total effort

Model and validate
- Feature engineering
- Select model, apply and validate

Feature engineering requires deep domain knowledge

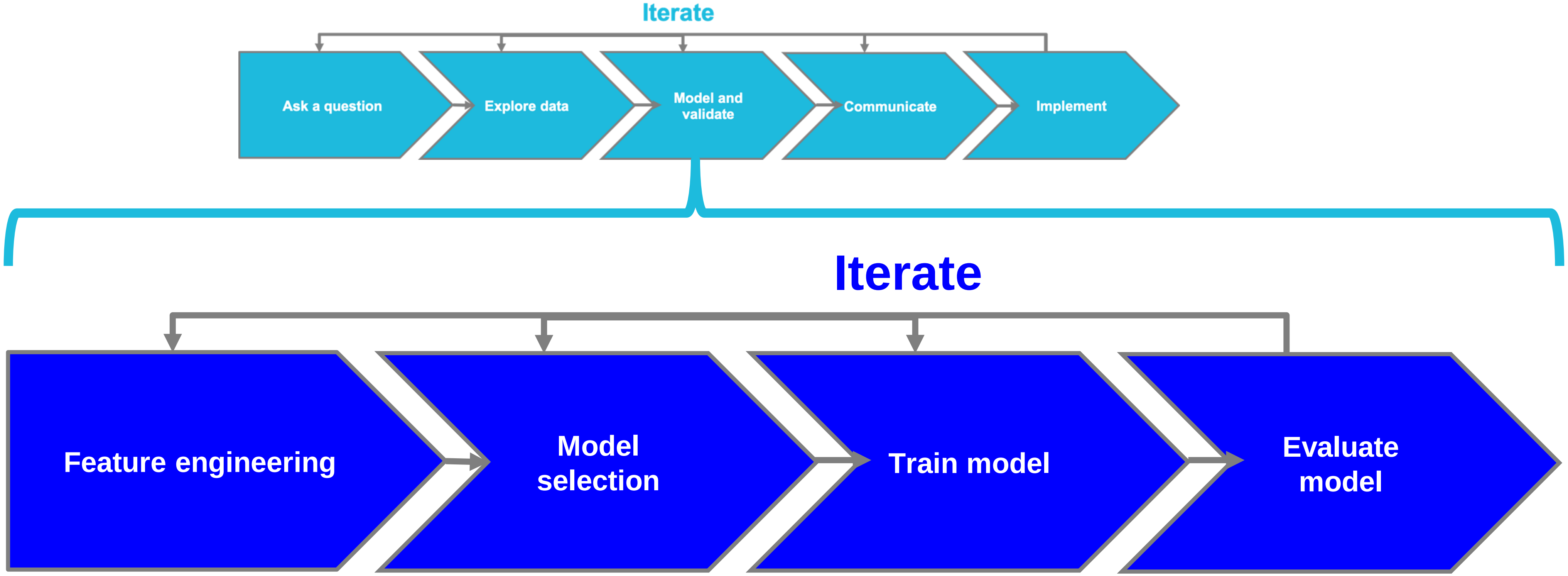
Communicate
- Communicate to stakeholders and obtain buy-in

Perform as early and as often as possible

Implement
- Develop end-to-end solution
- Build, test, deploy and monitor

Consider the total cost of implementation earlier in the process

Modelling Process



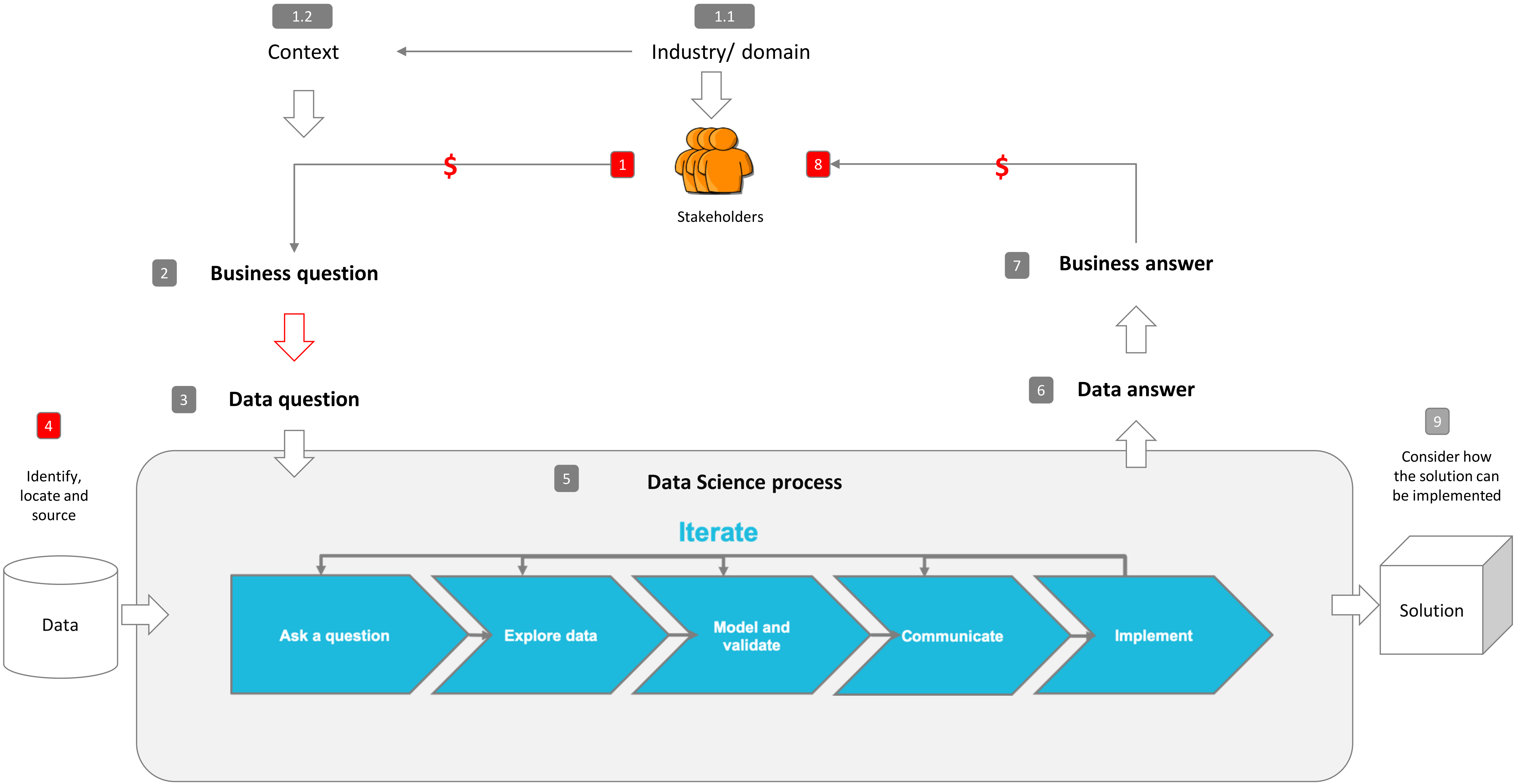
- Dimensionality reduction
- Remove noise
- Feature standardisation
- Categorical feature encoding

- Select appropriate model(s)
- Select hyper-parameters

- Train model

- Accuracy, recall

Applying data science in an industry project



Agenda: Succeeding in Industry

- Introduction, definitions, purpose and objectives
- What do employers value and what do they complain about?
- Skills required and attitude to succeed in the industry
- Data Science process in industry
- **Case study**
- Summary, conclusions and call for action

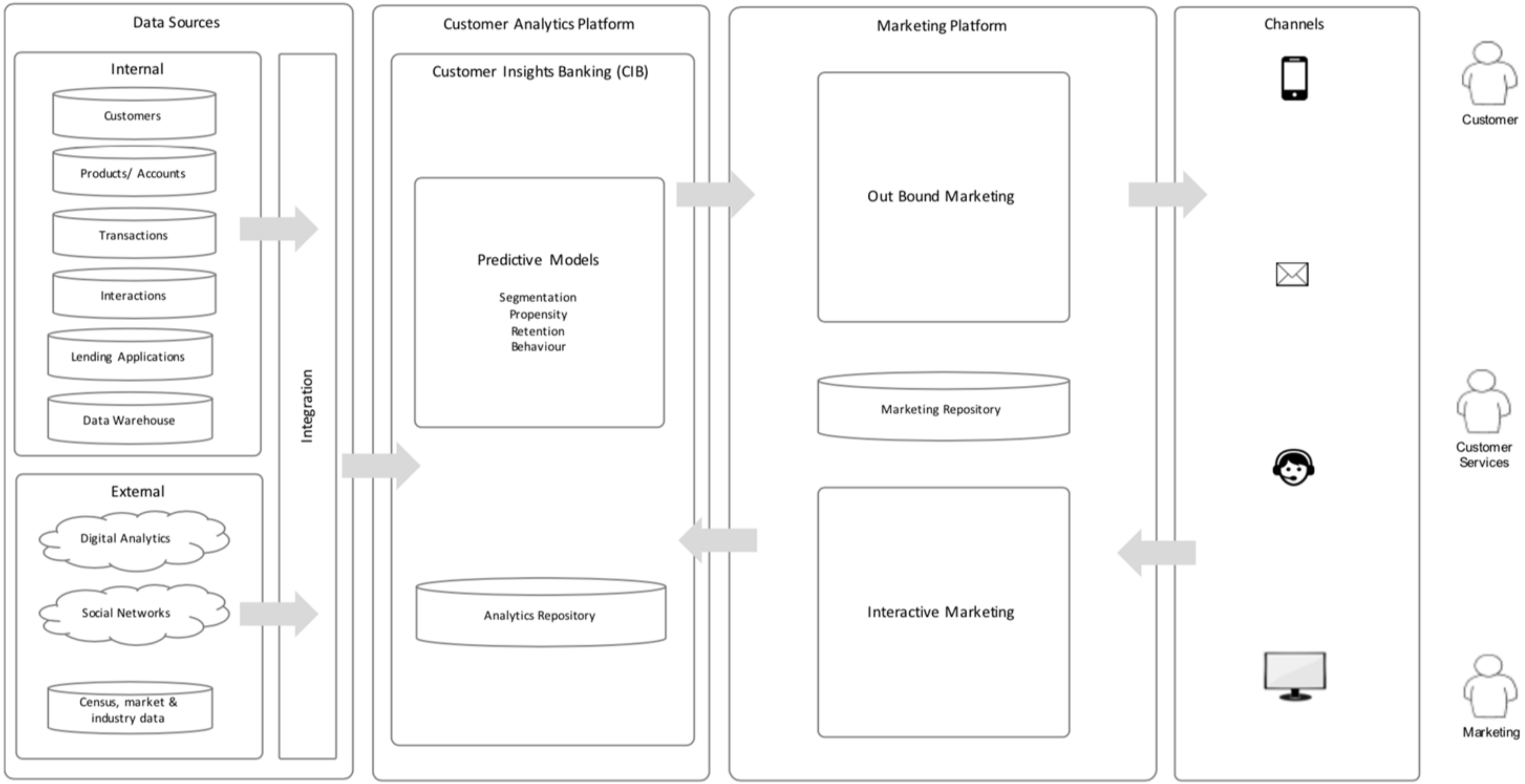
Industry Data Science use cases

- **Marketing, sales and customer services**: customer experience, acquisition, retention and life value.
- **Financial Services**: risk management, fraud detection and loan approval.
- **Telecommunication**: customer churn.

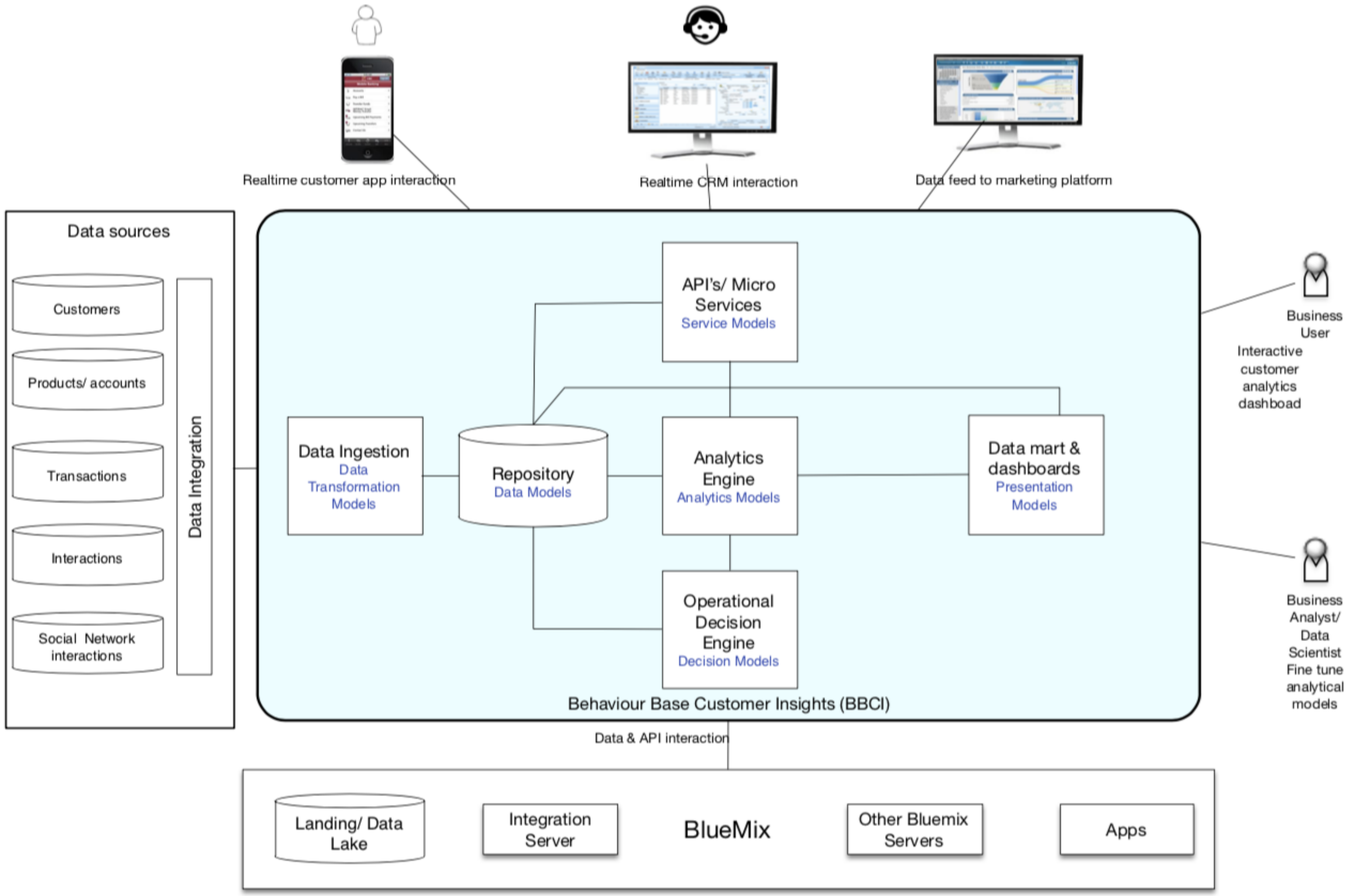
Case study: Home loans marketing

- **Use case:** The primary objective of the project is to develop models to **identify prospective customers** that are likely to take a new **home loan** or re-mortgage their existing loan with the bank within a set time horizon (up to 6 months).
- **Approach:** The models have been created and evaluated based on the **2-year historical data**.
- **Success criteria**: The model was tested on previously “unseen” customer data and successfully predicted customers who did purchase a mortgage.

Case study: Home loans marketing



Case study: Home loans marketing



Case study: Home loans marketing

Results comparison and business case overview

Applying the model for Banking can lead to potential annual **revenue twice as big** as the current model.

Results overview

	Baseline Model	Full Feature Model	Difference
% of identified applicants in top 10%	32%	61%	+29%
Potential Profit	627 x $Y	1,200 x $Y	573 x $Y

Business case overview based on the Final Model

Assumptions:

- Customer Value/year is $1000
- Customer base = 1.4 million
- Top 10% = 140,000 customers
- 2.8% applicants over 2 years ie. 1.4% annually
- 1.4% applicants in top 10% = 1,960
- 61% identified to target = 1,200
- 32% identified to target = 627

Potential profit = 573 x 1000 ≃ $500,000/ year
≃ $1.5m over 3 years

Agenda: Succeeding in Industry

- Introduction, definitions, purpose and objectives
- What do employers value and what do they complain about?
- Skills required and attitude to succeed in industry
- Data Science process in industry
- Case study
- **Summary, conclusions and call for action**

Summary, conclusions and call for action

- **Summary**
  - I have shared with you my industry experience and my views on how you could succeed in the industry by understanding what is required by employers
- **Conclusions**
  - It is **not enough** to develop the Data Science **'technical' skills**, you need soft skills so you can apply these skills to deliver **value**
  - To enable you to effectively work in industry, you need to:
    - Understand which are the most **important skills required in industry**
      - Discover your dream job and research what skills are needed for this job
    - Master a **small number of skills/tools** and
    - Decide on your **focus areas** including **domain (industry)**

Summary, conclusions and **call for action**

- **Call for action**
  - ***Start now!***
    - Identify/ refine your focus areas and skills,
    - **learn,**
    - **create,**
    - Identify gaps,
    - Iterate.

Questions?

Appendices

Data Scientist’s responsibilities

Data scientist's responsibilities

- Identify business needs
- Analyse data
- Develop machine learning models and solutions
- Present insights
- Manage data science projects

Data scientist’s responsibilities

- **Identify business needs**
  - Work with stakeholders to define business and information needs
  - Support the translation of business needs into data questions that can be addressed by available data
  - Defining what data is needed to answer the business question
- **Analyse data**
  - Collect, extract, query, clean, and aggregate data for advanced analytics purposes
  - Clean data to remove duplicate, outdated or irrelevant information
  - Perform statistical and visual analysis on data
  - Perform data validation and quality control checks
  - Mine data to identify trends, patterns and correlations

Data scientist's responsibilities

- **Develop machine learning models and solutions**
  - Build, implement, and evaluate advanced analytics problems solving using appropriate machine learning models and algorithms
  - Apply data mining techniques to investigate leads, identify patterns and regularities in data
  - Implement automated pipelines to create reproducible, scalable models
  - Identify areas of improvement of current analytics processes, products/services or models
- **Present insights**
  - Use data visualisation tools to communicate findings
  - Create clear and concise presentations reports for stakeholders
  - Design data reports and visualisation tools to facilitate data understanding
  - Assist with the development of actionable recommendations
  - Develop compelling, logically structured presentations, including story-telling of research and/or analytics findings
  - Guide stakeholders on how to act on findings
  - Use business consulting skills and frameworks in data science to assist managers and stakeholders understand the application of AI technology

37

Data scientist’s responsibilities

- **Manage data science projects**
  - Assists in the conceptualisation of data science projects
  - Maintain project plans and status of the project
  - Provide feedback to stakeholders throughout the whole analytics lifecycle
  - Prepare documentation to outline data sources, models and algorithms used and developed

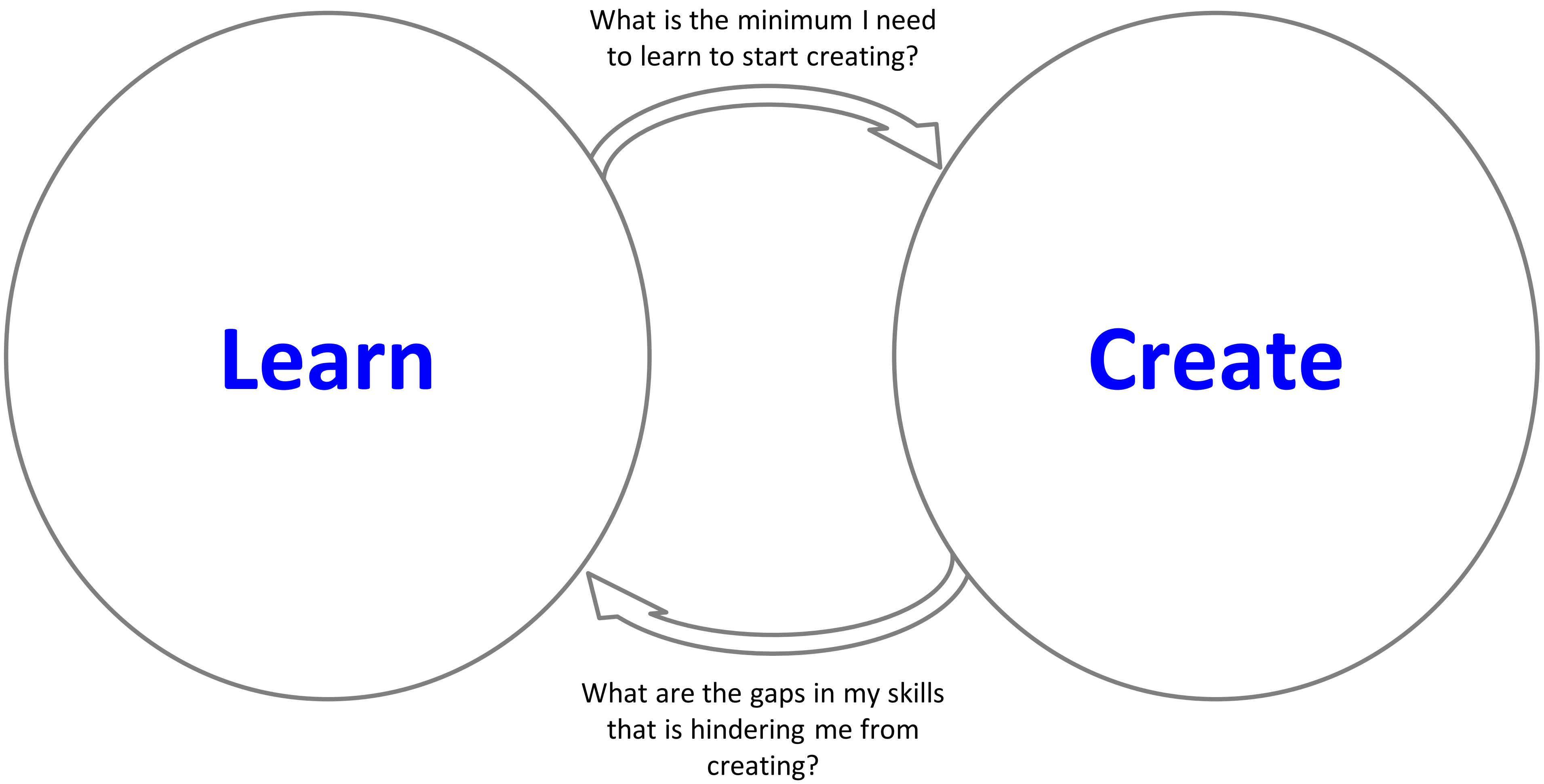
Mapping responsibilities to skills

Responsibility	Skills
Identify business needs	Applying data science in industry: • Define projects Soft skills: • Consulting, questioning and documenting projects
Analyse data	Core data science skills: • Exploratory Data Analysis (EDA) and data wrangling • Visualisation • Unsupervised machine learning
Develop machine learning models and solutions	Core data science skills: • Visualisation • Supervised machine learning (regression and classification) • Unsupervised machine learning Applying data science in industry • Design projects • Deliver project

Responsibility	Skills
Present insights	Soft skills: • Presenting Core data science skills: • Visualisation • Supervised machine learning (regression and classification) • Unsupervised machine learning
Manage data science projects	Applying data science in industry • Define projects • Design projects • Deliver project

Minimal Viable Learning for data science

The Minimal Viable Learning (MVL) concept helps you move quickly from learning to creating. As you create you identify possible gaps in your skills and go back to learning to fill these gaps



Mapping data scientist responsibilities to skills

Area	Minimal Must be mastered	Reasonable Expected to know	Nice to have Learn as required
EDA, data wrangling and visualisation	Pandas (see Pandas cheat sheet) Matplotlib Seaborn		Tableau
Supervised machine learning	**Sk-Learn (see sk-learn cheat sheet)** Linear Regression Logistic Regression K-NN	Ridge Regression Decision Trees Random Forest	SVM XGBoost Naive Bayes
Unsupervised machine learning	**Sk-Learn (see sk-learn cheat sheet)** K-means	Principal Component Analysis (PCA)	Locally Linear Embedding (LLE)

Minimal Viable Learning for data science

Area	Minimal Must be mastered	Reasonable Expected to know	Nice to have Learn as required
NLP	**Spacy (see Spacy cheat sheet)** Text classification	Named Entity Recognition Sentiment analysis Word2Vec	Latent Dirichlet Allocation (LDA)
Deep Learning	**Keras & TensorFlow (see Keras Cheat Sheet)** Convolutional Neural Networks (CNN)	Recurrent Neural Networks (RNN) Long-Short Term Memory (LSTM)	
Applying ML on different data types and applications	Tabular cross-sectional data Tabular longitudinal data (time series) Text	Images Time series forecast	Techniques for processing large datasets

Minimal Viable Learning for data science

Area	Minimal Must be mastered	Reasonable Expected to know	Nice to have Learn as required
Data access	SQL APIs		Web scraping Nosql
Programming development environments	Jupyter Notebook Anaconda Github	Google Collab	Jupyter Labs Visual Studio Code (VSC) Other cloud environments (e.g. AWS Sage maker)
Software engineering	Developing reusable, reproducible Python function		Developing reusable, reproducible Python classes

Strategies for learning Data Science & AI

Strategies for learning Data Science & AI

- The aim of course is to make you an **effective** Data Scientist in **industry**.
- Information about DS and AI are readily available for anyone to learn but **few** are able to develop their learning and achieve a level where they can **effectively perform** the role of a Data Scientist.
- Using the following **strategies,** you can do it:
  - Develop a **data-driven mindset**
    - Look at every topic or a question as a mini **Data Science project**
    - Ask yourself what **data** can answer this question and what the data is telling me
    - Develop your **Statistical Thinking**
  - Accumulate your learning in **Cookbook** Notebook(s)
  - Understand **your strengths** (including your previous experiences). Be flexible but avoid a Data Scientist **generalist** stance
  - Build your own **portfolio**
  - **Learn how to learn**

Strategies for learning Data Science & AI

- Learning Data Science (in a traditional way) can be very hard
  - It's very broad, covering business, programming, math, visualisation, etc
  - It's new, therefore has not developed a stable body on learning
  - It's very active. There is a huge level of research and new methods coming everyday
  - It's full of hype
- To make your job of learning data science easier, do the following:
  - **Focus!**

    Decide if you prefer to be a **generalist** or a **specialist**.

    Select an **industry** to apply your technical skill
  - Decide on your **focus** but develop an **appreciation** of the entire process and different levels of abstraction
  - Make up your mind **for now** on the above points but be open to change your mind

Level of abstraction, understanding and execution in Data Sconce

Business

- Business value (ROI)
- Value chain
- Key concepts
- Stakeholders consulting and communication

Data

- Key entities and relationships
- Data sources and structure
- Data cleaning, munging, analysis, visualisation

Modelling

- What features to use
- Which model to use
- How to evaluate models
- Interpreting output of models

Algorithms

- How to tune model hyperparameters
- Understand limitations of algorithms
- How to optimise algorithm performance

Math/ Statistics

- Intuition on how algorithms work
- How to modify, enhance or extend functions of algorithms
- Understanding limits of algorithms

Learning to learn (better) framework

Learning how to learn better

Multi-modal learning

- **Multi-modal** learning means, simply, that you use multiple modes of learning alternately and repeatedly.
- Modes of learning include:
  - Reading
  - Writing
  - Observing
  - Researching
  - Programming
  - Experimenting
  - Visualising
  - Communicating
  - Collaborating
  - Questioning

Learning how to learn better

Some learning modes are more effective

More Effective

Less Effective

Questioning
Why are doing this?

Collaborating

Communicating

Experimenting
What if?

Visualising

Programming
Make the process reproduceable

Writing

Researching

Critiquing

Reading

Watching

Listening

This is not the **sequence** of how you learn. It is an iterative process.
You should cycle through **all** these learning modes to produce an effective outcome.

Once you identify your questions, go back and collect the data to answer the questions

Passive

Active

End of Presentation!