THE UNIVERSITY OF WESTERN AUSTRALIA

School of Computer Science and Software Engineering

CITS4009 Introduction to Data Science

SEMESTER 2, 2017: SYLLABUS AND INTRODUCTION

Chapter 1

THE DATA SCIENCE PROCESS

Chapter Objectives

- What is data science?
- Business and scientific applications of data science
- Defining data science project roles
- Understanding the stages of a data science project
- Setting expectations for a new data science project

What is data science?

- Data science is the area of scientific study for extracting knowledge from large volumes of data.
- It is quite often very difficult to understand data in raw form, as numbers for example.
- Moreover, it is very important to interpret key ideas from large volumes of data for various reasons.
- Data science uses a wide repository of tools and techniques for extracting meaningful and useful information from data.

Business Applications

- Businesses collect an enormous amount of data during their day-to-day activities.
- This data can be used for improving insights into the business processes and customer behavior.
- These insights can be used for improving business operations and increase profits.
- There are many different tools that are being used for this purpose, e.g., statistical analysis of business data, data warehousing, machine learning etc.

Scientific applications

- Scientific applications collect enormous amount of data by using different instruments and sensors.
- It is important to extract scientific knowledge from this data as the raw data become overwhelming very fast and become unusable in the long run.
- Some examples are the Square Kilometer Array (SKA) and the Large Hadron Collider (LHC).
- Quite often analyzing scientific data may require domain specific tools as well as tools that are used in analyzing business data.

Data science project

- **Data scientists** are responsible for guiding a data science project from start to finish.
- Success of data science project comes from:
  - quantifiable goals,
  - good methodology
  - cross discipline interactions;
  - repeatable workflow.

The roles in a data science project

Role	Responsibilities
Project sponsor	Represents the business interests; champions the project
Client	Represents end users' interests; domain expert
Data scientist	Sets and executes analytic strategy; communicates with sponsor and client
Data architect	Manages data and data storage; sometimes manages data collection
Operations	Manages infrastructure; deploys final project results

Project Sponsor

- Project sponsor is one of the most important roles in a data science project.
- Usually a sponsor represents the business interest, they are best placed to judge the impact of the project on improving business processes.
- A sponsor is usually responsible for deciding whether a data science project is a success or a failure.
- It is important to keep the sponsor well-informed about the progress of the project.

Project Sponsor

- Project sponsors are usually high ranking persons within a business, and hence they are busy.
- It is important to have very specific quantitative goals to keep the sponsors interested and also to make them approve a project.
- For example, “Identify 90% of the new applicants for a bank loan who will not default in payments, with a false positive rate not more than 20%”.

Client

- The client's role is to ensure the end-users of a data science project are satisfied with the project outcomes.
- The client is more hands-on compared to the sponsor, usually a technical person within the organization.
- The client is responsible for the interface between the business interest (that the sponsor looks after) and the day-to-day business processes within the organization.

Client

- However, the client may not be very well informed mathematically or in their computer/data science knowledge.
- It is the responsibility of the data scientist to convince the client that whatever techniques are used in the project are suitable for the day-to-day functioning of the organization.
- It is very important to keep the client well informed about the progress and the milestones of the project.

Data scientist

- The role of the data scientist is to ensure that all necessary steps are taken for the project to succeed.
- These include the design of the project, the data sources to be used and the techniques to be deployed.
- They must have excellent knowledge in the statistical, machine learning and other techniques in solving data science problems.

Data architect

- The data architect is responsible for all the data to be used by the project during the development process, as well as when the final product from the project is deployed.
- Usually the data architect is someone from within the organization, e.g., data base administrator, and has intimate knowledge how data is procured and stored in the business.
- Data architects have very crucial roles to play in a data science project.

Operations

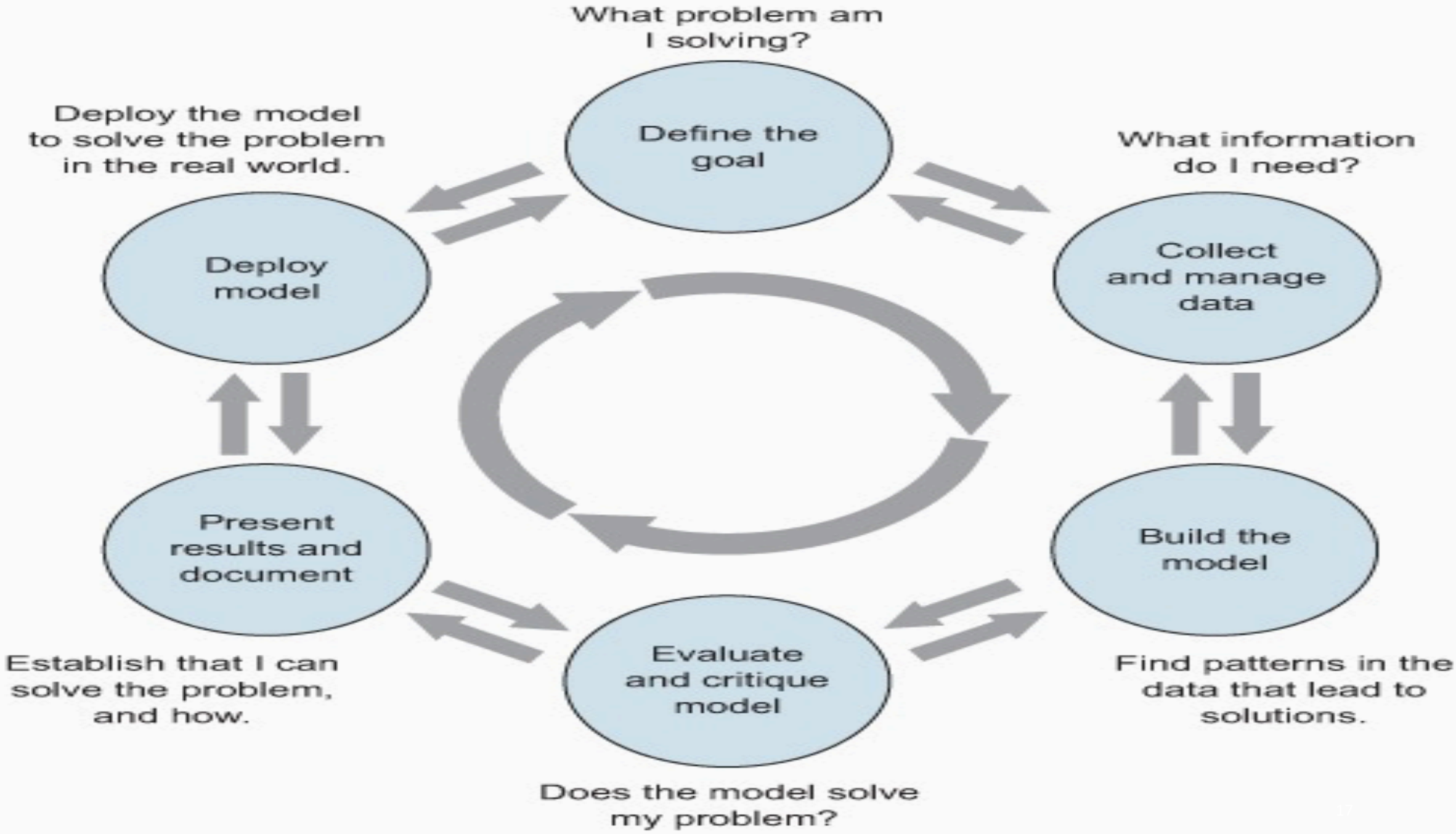
- The operations role is critical both for acquiring data and delivering the final results of the project.
- The operations role is usually filled by someone from the organization, as they have intimate knowledge how their business operates and interfaces with the customers.
- They are also aware of the different constraints, e.g., how frequently the operational data can be uploaded in the data warehouse, how frequently a web interface can be redesigned and deployed etc.

Stages of a data science project

- Ideal data science environment is one that encourages feedback and iteration between the data scientist and all other stakeholders.

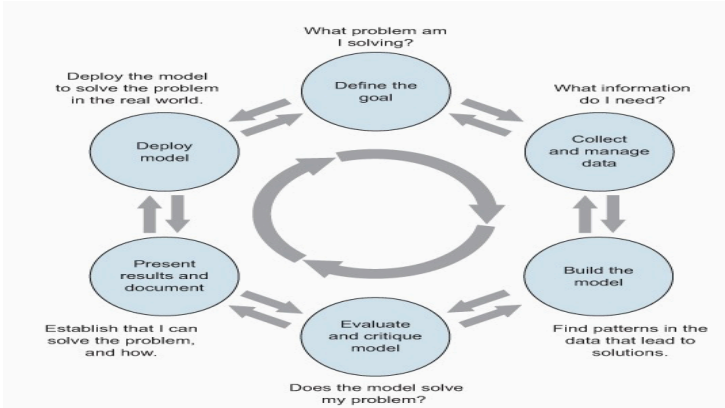
1. Define The goal
2. Collect and manage data
3. Build the model
4. Evaluate and critique model
5. Present results and documents
6. Deploy model

- New issues and questions can arise from seeing that model in action.


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Does the model solve my problem?
Evaluate and critique model
Establish that I can solve the problem, and how.
Present results and document
Deploy the model to solve the problem in the real world.
Deploy model

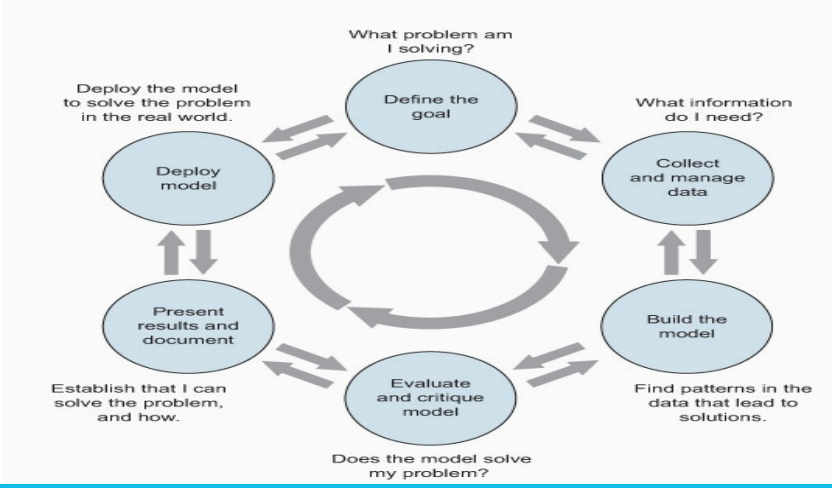
1- Defining the goal

- learn all that you can about the context of your project:
  - Why do the sponsors want the project in the first place? What do they lack, and what do they need?
  - What are they doing to solve the problem now, and why isn't that good enough?
  - What resources will you need: what kind of data and how many staff? Will you have domain experts to collaborate with, and what are the computational resources?
  - How do the project sponsors plan to deploy your results? What are the constraints that have to be met for successful deployment?
- Once you have a good idea of the project's goals, you can focus on collecting data to meet those goals



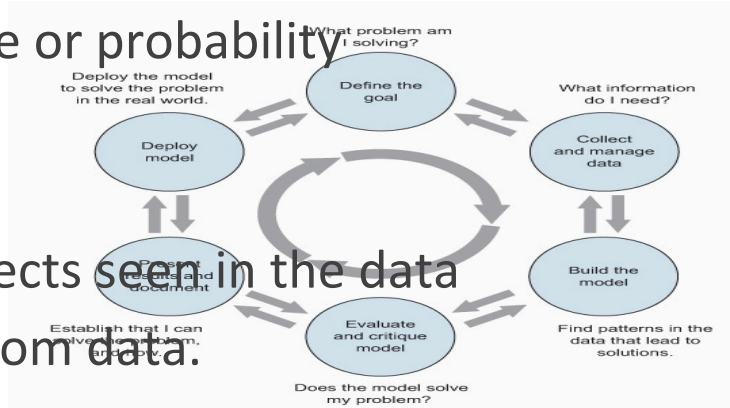
2- Data collection and management

- Identifying the data you need, exploring it, and conditioning it to be suitable for analysis.
- Most time consuming stage and most important:
  1. What data is available to me?
  2. Will it help me solve the problem?
  3. Is it enough?
  4. Is the data quality good enough?
- This stages will be covered in depth in chapter 3 and 4.



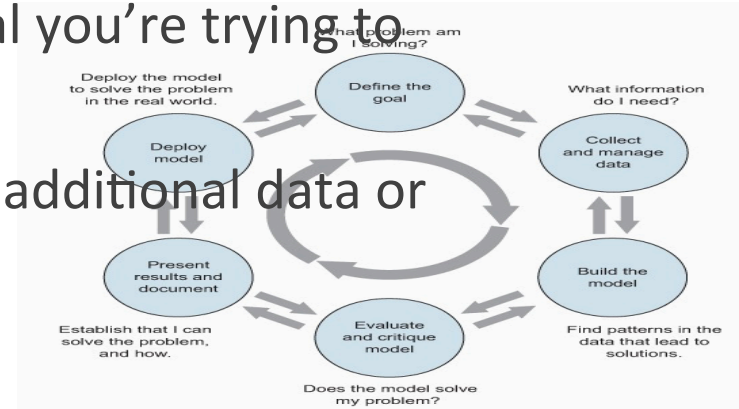
3- Modeling

- This stage is where you try to extract useful insights from the data in order to achieve your goals.
- There will be overlap and back and forth iteration between the modeling stage and the data cleaning stage to try to find the best way to represent the data and the best form in which to model it.
- The most common data science modeling tasks are these:
  - **Classification** — Deciding if something belongs to one category or another.
  - **Scoring**— Predicting or estimating a numeric value, such as a price or probability
  - **Ranking**— Learning to order items by preferences
  - **Clustering**— Grouping items into most-similar groups
  - **Finding relations**— Finding correlations or potential causes of effects seen in the data
  - **Characterization**— Very general plotting and report generation from data.



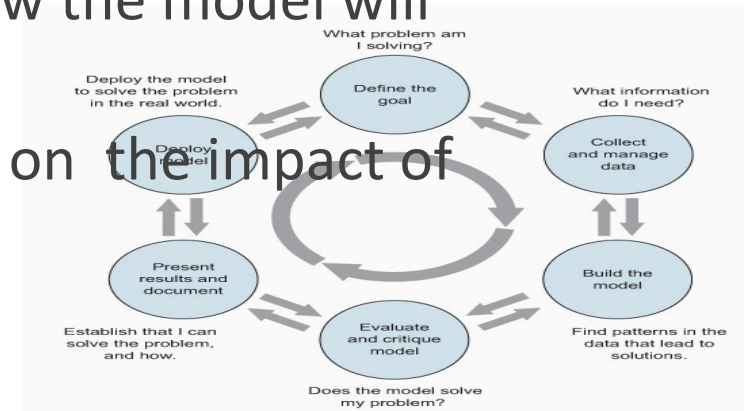
4- Model evaluation and critique

- Once you have a model, you need to determine if it meets your goals:
  - Is it accurate enough for your needs? Does it generalize well?
  - Does it perform better than "the obvious guess"? Better than whatever estimate you currently use?
  - Do the results of the model (coefficients, clusters, rules) make sense in the context of the problem domain?
- If you've answered "no" to any of these questions, it's time to loop back to the modeling step—or decide that the data doesn't support the goal you're trying to achieve.
- This might mean defining more realistic goals or gathering the additional data or other resources that you need to achieve your original goals.
- Chapter 5 covers more details about model evaluation.



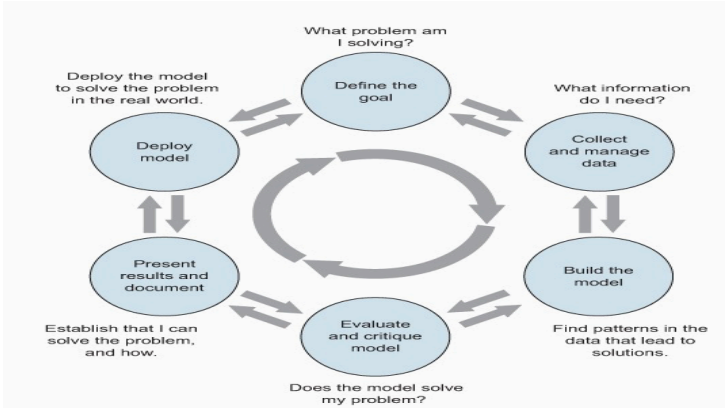
5- Presentation and Documentation

- After meeting the success criteria of the project, you need to present your results to your project sponsor and other stakeholders.
- Documenting the model for those who need to use the model for running, using and maintaining the model once it has been deployed.
- Document for **Business-oriented** audience − understand the impact in terms of business metrics.
- Model Presentation for **end-users** should emphasis on how the model will help them do their jobs better.
- Model Presentation for **operations staff** should emphasis on the impact of the model on the resources they are responsible for.



6- Model deployment and maintenance

- This stage comes after putting the model in operation.
- Ensure the model runs smoothly and won't make disastrous unsupervised decisions.
- Ensure model can be updated as its environment changes.
- More discussion will be in chapter 10.



Setting expectations

- Project sponsor probably already has an idea of the performance required to meet business goals.
- Ways to estimate whether the data you have available is good enough to potentially meet desired accuracy goals:
  1. Determining lower and upper bounds on model performance.
     - The null model: a lower bound on performance
     - The Bayes rate: an upper bound on model performance
     - The Bayes rate gives the best possible accuracy, but the most accurate model doesn't always have the best possible precision or recall (though it may represent the best trade-off of the two).

Takeaway messages

- A successful data science project involves a variety of roles to represent business and client interests, as well as operational concerns.
- Make sure you have a clear, verifiable, quantifiable goal.
- Make sure you've set realistic expectations for all stakeholders.