Dennis Sun **DATA 401:** Data Science Alexander Dekhtyar

What is Data Science

Definitions

Wikipedia Definition. From Wikipedia[1]:

> **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

Data Science as a Cross-Disciplinary Field definition. Data science is a discipline that lies in the intersection of three fields:

- Statistics
- Computer Science
- Domain expertise

From Statistics, data science borrows hypothesis testing and data analytical techniques. From computer science data science borrows approaches to data manipulation and machine learning algorithms, while domain knowledge directs the combined machinery of statistics and computer science to proper use.

Data Science as a set of skills. Another way to define data science is to outline a set of skills that a pratitioner in the field (i.e., a data scientist) is expected to possess. This approach has its advantages and drawbacks.

Among the **advantages** are:

- **Clear picture.** Skills are easy to evaluate - one either has them or one does not.

[1] https://en.wikipedia.org/wiki/Data_science

- **Reduction from job descriptions.** One can study job descriptions for **data scientist** positions and reverse-engineer the necessary skills[2].

The **disadvantages** of this approach, however, are significant:

- **Such definitions are not stable.** Skills, especially when they are ground in specific technologies, are very present-day. Tomorrow's set of skills necessary for a data scientist may be very different than today's. This means that an effective definition of data science will have to shift with the times. *This is not very convenient.*
- **Such definitions ignore the "science" part of data science.** We do not define what carpentry or medicine is by the sets of tools carpenters or medics have to work with. We define different professions and disciplines by the nature of the work and/or the nature of the product produced. In case of data science a skills-based definition lacks the ability to explain *what these skills are used for.*

As a result:

- Skills-based definitions of data science are very useful in determining the specific portions of the curriculum (what skills to teach, what technologies to present in class).
- But such definitions cannot be all-encompassing.

Data Science as a process. Our final approach to describing what **data science** is, is to observe that productive work with data is a process.

When comparing **what** data scientists do on a day-to-day basis, with what statisticians or computer scientists are taught to do, we can notice significant differences. This leads to an understanding that

> **Data Science** can be thought of as the discipline that studies and the full cycle of work with data and incorporates all stages this work from data acquistion, through data analysis and all the way to presentation of the obtain insight.

Data Science Process

In this class, we will use the term data science process a lot, and we will spend considerable time talking about the specific steps of this process. In a nutshell, the data science process consists of the following steps:

1. Formulation of questions.

[2] In fact, this is, in large part, how we have built the curriculum for the Data Science Minor at Cal Poly.

2. Data acquistion.
3. Data cleaning/pre-processing.
4. Data modelling.
5. Data analysis.
6. Visualisation of results.
7. Presentation of insight/results.

Note 1. This process often needs to be considered as a **cycle**, as shows in Figure 1. Upon the completion of Step 7: Analysis of Results, it is often useful to ask the *"Have we learned everything we need?"* question. The answer to this question is almost always a *"no"*.

Note 2. Data scientists often are not the only people working on parts of this process. Step 1: Formulation of questions is often conducted by other professionals working with data scientists (business analysts, executives, project/product managers, Principal Investigators, and so on). Similarly, the last step of the process, Analysis of results is often performed collaboratively by data scientists and the people who they are working for.

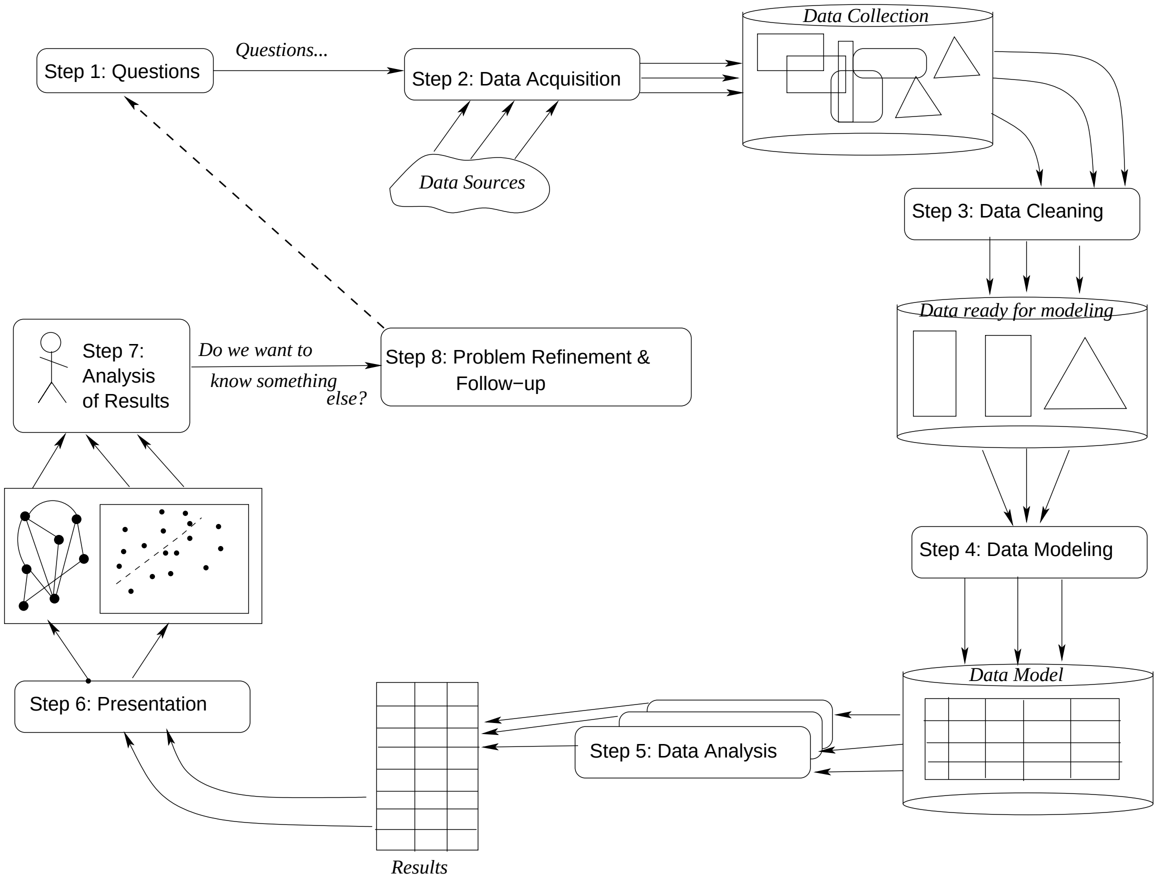

Figure 1: Data Science process as a cycle.

Step by Step

Step 1: Formulation of questions. First step of the data science process. In this step, the specific question that the data scientists must answer is formulated and, if necessary, negotiated.

Step 2: Data Acquisition (Collection). Based on the question given to the data scientists, on this step, the data scientists determine what data is needed to successfully answer it. The data is then collected from a variety of internal (e.g., corporate databases) or external (e.g., world wide web) sources.

Step 3: Data Cleaning. The collected data may contain

(a) data not needed to answer the question(s) studied

(b) data about same objects/events/entities collected from multiple sources

(c) missing data

(d) unreliable, potentially incorrect data.

The data cleaning step of the data science process identifies these categories of data items in the data collection and performs appropriate manipulations with them. For example, unnecessary data is filtered out; data obtained from multiple sources about same objects can be merged together (integrated), unreliable data, if discovered, can be purged, or labelled as such.

Step 4: Data Modeling. The methods used on the Data Analysis step take input in specific format. Also, for these methods to produce high-quality output (this specifically refers to *machine learning* methods), the input data should contain the correct set of features. The data modeling step takes the cleaned up data, transforms it into the formats necessary for the analytical methods. It also, where needed, includes the feature extraction and feature selection procedures to fill in the inputs for the analytical methods with the *right* data.

Step 5: Data Analysis. The step which converts *data* into *knowledge*. A variety of analytical procedures, from data warehouse operations (roll-up, slice-and-dice, pivot, etc), to statistical methods (t-test, linear regression, factor analysis, multivariate analysis), to machine learning methods (classification, clustering, association rule mining, similarity analysis) can be deployed on the collected data on this stage.

Step 6: Visualization and Presentation of results. Often, the output of the methods used in the Data Analysis step is large and hard to immediately understand. During the Visualization and Presentation step, such output is turned into coherent, easy-to-observe representations.

Step 7: Analysis of results. Final step of the linear data science process - on this step the produced results are observed and discussed. Explanations are given to the observed phenomena, and reports are prepared.

Beyond these seven steps, data scientists need to be aware of two more steps of the business process into which their data science/discovery process is embedded.

Step 8: Goal refinement. (See Figure 1). Based on the results obtained in a single cycle of the data science process, data scientists and other professionals they work with ask the *Have we seen enough?* question. The answer to this question is most often a *"no"*. In such cases, based on the information obtained from the most recent analysis (and possibly from some prior stages), new questions are asked, serving as the starting point for the next round of the data science process.

Step 8': Action items. One of the key question following Step 7: Analysis of the results often is *"What do we do with this information?"*. The actual formulation of action items, i.e., things to do based on the completed data analysis, is usually beyond the scope of responibilities of a data scientist. However, the professionals who do make these decisions may trigger a new round of the data science process with the new sets of questions related to their ability to act upon the information produced on the current (and previous) rounds.

Data Science Challenges

Very coarsely the challenges a Data Scientist is facing when working with data can be broken into two categories:

1. Technical challenges. These are challenges related primarily to the data scientist's understanding and command of the specific data analytical methods, and their choice for addressing specific data analysis questions.

2. Logistical challenges. These are challenges related primarily to the data scientist's ability to understand the underlying problem, collect necessary data, perform the required analyses in a timely fashion, and communicate results to the customer/client.

This distinction is imperfect, but in the confines of DATA 401 it is justifiable. The course lectures are split into two groups:

1. Machine Learning Lectures. These lectures are primarily designed to introduce a wider range of data analytical techniques:

   - regression
   - classification
   - clustering
   - collaborative filtering
   - and more...

   These lectures as designed to address the technical challenges of Data Science.

   You will be educated on methodology for analyzing data in specific, sometimes *rather advanced ways*.

2. Data Science Process Lectures. These lectures are primarily designed to address some of the logistical challenges.

Logistical Challenges In Data Science

Below is *a rather incomplete* list of logistical challenges that data scientists face in their work. We plan to discuss these challenges throughout the class.

The list of challenges briefly described below is:

1. Talking to Customers
2. Data Integration
3. Feature Selection/Feature Engineering
4. Measuring the Right Thing
5. Visualization and Reporting
6. Big Data
7. Systems Engineering for Data Science
8. Failure to Discover Insight

Challenge 1: Talking to Customers

Nature of the challenge. Data Scientists rarely are the originators of the initial data analytical problems and questions they need to work on. Typically, outside customers originate such tasks. In such situations a data scientist must communicate with the customer in order to obtain the following information:

- What is the subject matter domain?
- How much subject matter knowledge shall the data scientist acquire?
- What is the data which needs to be analyzed?
- Is the data already collected, or does it need to be collected?
- What are the data-analytical questions of interest to the customer?
- How big the datasets are going to get?

(Note: this is NOT a full list of questions one needs to get answered.)

Why this is a challenge. We observe the following:

- Customers are often domain experts, but are not data scientists themselves.
- Any time people from different fields of study talk to each other, there is a possibility for misunderstanding/miscommunication.

- Working with an outside customer is not something that is routinely taught in college[3].

Notes. This is not dissimilar to the challenges software engineers experience when they elicit software requirements from outside customers. Statisticians[4] face similar challenges when working with outside customers.

Challenge 2: Data Integration

Nature of the challenge. It is a common occurrence in data science to have to build the dataset to be analyzed out of portions of datasets obtained from multiple sources. In such situations, the problem of data integration is often one of the most complex issues to resolve.

Formally, we define data integration as follows[5]:

> Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.

Why this is a challenge. At a somewhat higher granularity, the key issue in data integration is the recognition of portions of different datasets that relate to the same objects/entities.

Recognizing that two different chunks of data belong to the same entity (or are alternative descriptions of the same entity) is notoriously difficult.

There are multiple sources for this difficulty (to be discussed later in the course).

Challenge 3: Feature Selection/Engineering.

Nature of the challenge. In statistics and machine learning feature selection is defined as[6]:

> Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

In addition to feature selection, i.e., the choice of features from already acquired data, this challenge covers an additional problem of *creating new features, relevant for the impending data analysis, from the existing data.*

[3] There are some places in CS and STAT curricula where this happens, but there is no systematic attempt to make this permeate the entire respective curricula.

[4] In this course, we distinguish between Data Scientists who work on all aspects of the data science process, and statisticians, who limit their scope of work to performing only the statistical analysis of data.

[5] http://www.ibm.com/analytics/us/en/technology/data-integration/

[6] https://en.wikipedia.org/wiki/Feature_selection

Why this is a challenge. A lot of data science tasks involve the following:

- large number of raw features brought in (see the data integration challenge above)
- raw features of high granularity/individually representing a very small "slice" of information about the entities
- lack of indication of relative feature importance in the problem statement.

Given p features, there are 2^p feature sets. Additionally, there is no limit to the new features that can be synthesized from the existing ones. Figuring out which ones help analyze the data is difficult if only because of the sheer number of possibilities.

Challenge 4: Measuring the Right Thing

Nature of the challenge. This challenge occurs in a number of data science tasks (although by far, not in all of them) when the insight that the customer wants from the data requires creation of new (i.e., not present in the data at the time of acquisition) measurements.

Often, the customer already has a specific measurement used for analytical purposes.

For example, purchasing activity at a grocery store can be expressed as *dollars spent by all customers each month divided by the square footage of the store.* In this case, the goal of the data scientists is to acquire valid data that allows for this measure to be computed.

In some cases, *however*, the customer's request does not contain a specific measurement. In such cases, *it is the role of the data scientists to come up with a number of alternative ways to produce requested measurements, collect the measurements, compare them, and determine, which measurements provide best insight.*

For example if the customer instead asks the data scientists to simply "measure the purchasing activity at each grocery store and determine which store is the most active," the data scientists must recognize that there is a variety of ways to make such measurements. In addition to the "dollars per square foot", measures of activity may include "number of people shopping per day", "number of items bought per shopping trip", "value of average market basket", "store profit", and many others.

Why this is a challenge. In some cases, the measures are already known and simply need to be computed and compared. In some other cases, it may actually be the job of the data scientist to *invent the measures.*

Challenge 5: Visualization and Reporting

Nature of the challenge. This is actually two challenges wrapped in one:

- Determining the best way to visualize/show(case) the results of the analysis.
- Communicating the results to the customer.

Why this is a challenge. The first of the two challenges is about selecting the appropriate visualization tools, techniques, and tricks to best reflect the insight gained from data. Visualization of information, by itself, is a complex discipline.

The basic reason for the difficulty of information visualization stems from the fact that we must present the visualization in **two** (rarely, **three**) **dimensions**, but the number of dimensions that need to be visualized is often significantly higher.

The second challenge is that of communication (see Challenge 1).

Challenge 6: Big Data

Nature of the challenge. This challenge appears when the amount of data available (and necessary) for analysis exceeds by at least two orders of magnitude the amount of data that can be processed in a "timely fashion"[7] by the analytical methods one would typically select for the required analysis.

The key challenge is determining how the data shall be processed, and what, if any, simplifications and consolations can be made due to the enormous amount of data that needs to be processed.

Why this is a challenge. "Big Data" may mean a variety of things, but let's assume for a second that we are talking about the *Volume* and *Velocity* of data here.

Data mining algorithms are complex and often slow. Their most popular implementations often assume limits on the size of input, such as:

- All input data can fit into main memory of one computer.
- All input data can fit onto hard disk of one computer.
- All input data is static.

[7] "Timely fashion" is in quotes because the notion of what is timely differs from problem to problem. 10 minutes may be very timely for one data analytical task, but may be prohibitively expensive for another.

When these assumptions are violated, data scientists may no longer be able to use off-the-shelf libraries for data analytical tasks.

In order to host the data, the data science team may have to build a hardware/software infrastructure that otherwise would not have been needed.

Challenge 7: Data Science as Systems Engineering

Nature of the challenge. Individual data-analytical problems usually have a straightforward form: obtain a dataset, choose analytical method that takes the dataset as input, collect output (repeat as needed).

However, in practice, a typical data science task consists of numerous subtasks, where each subtask is a standalone data-analytical problem.

In such situations, data scientists must perform some systems engineering tasks in order to assemble a complete solution.

Why this is a challenge. There are two key reasons why building "data science pipelines" is more challenging than solving individual constituent problems in isolation.

- Designing a complex system of n components is more complex than designing n systems each consisting of one component.
- The final accuracy of insight is now subject to compounded error. The design of the system must deal with this issue.

Challenge 8: No Insight is Discovered

Nature of the challenge. Occasionally, the data scientists are able to overcome all other challenges and build a proper data science process for data analysis *only to discover at the end that the analysis reveals no insight.*

Why this is a challenge. Among the reasons are:

- Data science is often performed in situations when *failure is not an option.* (That is, the customer expects answers, and will not accept the "we couldn't find anything" as valid insight, unless the data scientists can provide actual justifications.)
- There may be multiple underlying reasons for lack of insight. Among these are:
  - wrong data
  - insufficient data
  - incorrect data[8]

[8] "Wrong data" refers largely to errors in feature selection/feature engineering, whereas "incorrect data" refers to data that is not reflective of the true state of affairs.

- true lack of sought insight in the data
- wrong technical approach used to data analysis
- more methods can be applied to imrpove results

Determining the reasons for lack of insight thus requires probing all (or most) possible scenarios.