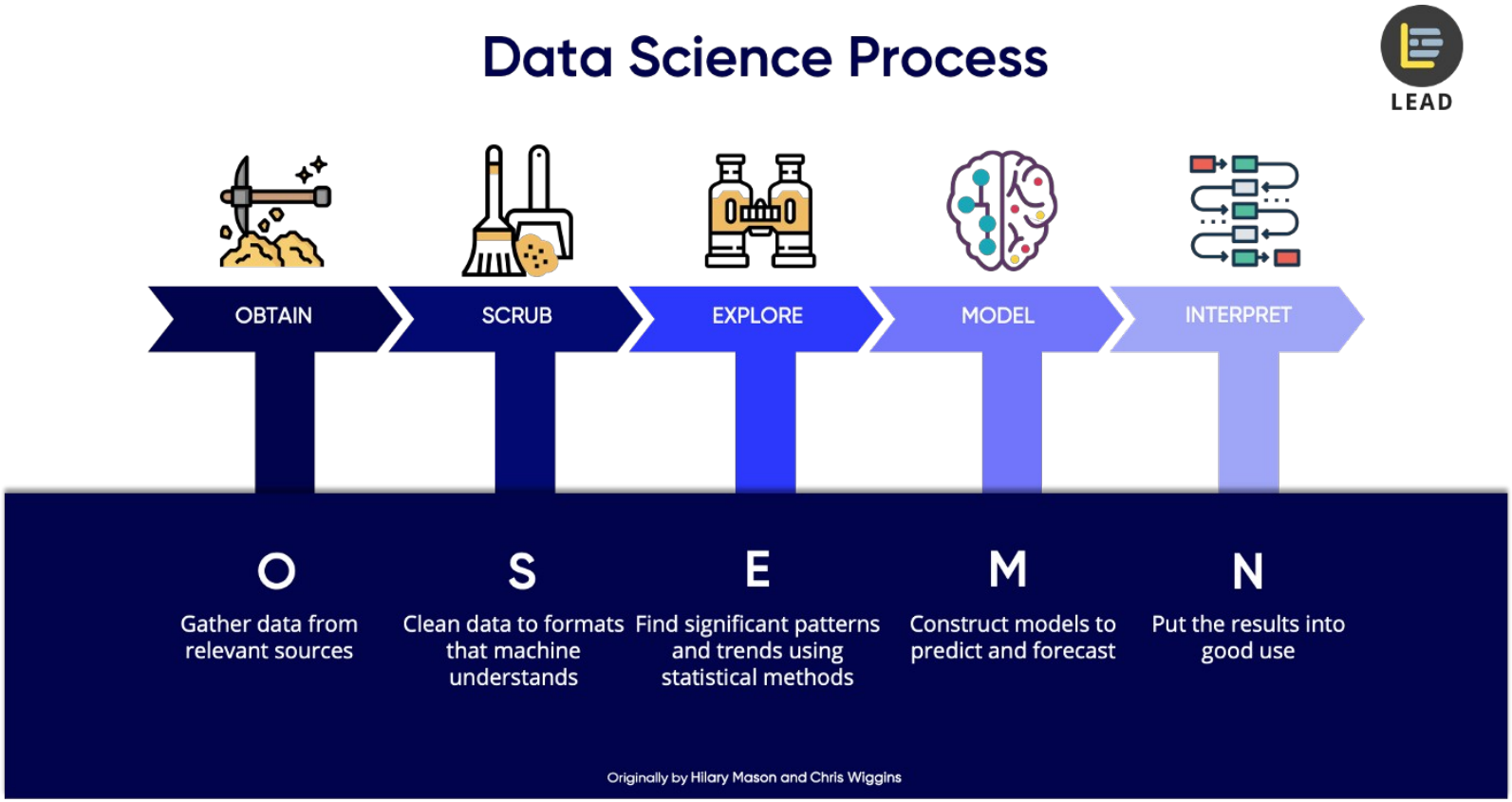
Data Science Life Cycle

Tushar B. Kute,
http://tusharkute.com

Data Science Life Cycle

- Often, when we talk about data science projects, nobody seems to be able to come up with a solid explanation of how the entire process goes.
- From gathering the data, all the way up to the analysis and presenting the results
- We break down the data science framework, taking you through each step of the project lifecycle, while discussing what the key skills and requirements are.

Data Science Life Cycle



Obtain Data

- The very first step of a data science project is straightforward. We obtain the data that we need from available data sources.
- In this step, you will need to query databases, using technical skills like MySQL to process the data.
- You may also receive data in file formats like Microsoft Excel.
- If you are using Python or R, they have specific packages that can read data from these data sources directly into your data science programs.

Obtain Data

- The different type of databases you may encounter are like PostgreSQL, Oracle, or even non-relational databases (NoSQL) like MongoDB.
- Another way to obtain data is to scrape from the websites using web scraping tools such as Beautiful Soup.
- Another popular option to gather data is connecting to Web APIs.
- Websites such as Facebook and Twitter allows users to connect to their web servers and access their data.
- All you need to do is to use their Web API to crawl their data.

Obtain Data

- And of course, the most traditional way of obtaining data is directly from files, such as downloading it from Kaggle or existing corporate data which are stored in CSV (Comma Separated Value) or TSV (Tab Separated Values) format.
- These files are flat text files. You will need to use special Parser format, as a regular programming language like Python does not natively understand it.

Obtain Data

- Skills required
  - To perform the tasks above, you will need certain technical skills. For example, for Database management, you will need to know how to use MySQL, PostgreSQL or MongoDB (if you are using a non-structured set of data).
  - If you are looking to work on projects on a much bigger data sets, or big data, then you need to learn how to access using distributed storage like Apache Hadoop, Spark or Flink.

Scrub Data

- After obtaining data, the next immediate thing to do is scrubbing data.
- This process is for us to “clean” and to filter the data. Remember the “garbage in, garbage out” philosophy, if the data is unfiltered and irrelevant, the results of the analysis will not mean anything.

Scrub Data

- In this process, you need to convert the data from one format to another and consolidate everything into one standardized format across all data.
- For example, if your data is stored in multiple CSV files, then you will consolidate these CSV data into a single repository, so that you can process and analyze it.
- In some situations, we will also need to filter the lines if you are handling locked files.
- Locked files refer to web locked files where you get to understand data such as the demographics of the users, time of entrance into your websites etc.

Scrub Data

- On top of that, scrubbing data also includes the task of extracting and replacing values. If you realise there are missing data sets or they could appear to be non-values, this is the time to replace them accordingly.
- Lastly, you will also need to split, merge and extract columns. For example, for the place of origin, you may have both “City” and “State”. Depending on your requirements, you might need to either merge or split these data.
- Think of this process as organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardising the format across all the data collected.

Scrub Data

- Skills Required
  - You will need scripting tools like Python or R to help you to scrub the data. Otherwise, you may use an open-sourced tool like OpenRefine or purchase enterprise software like SAS Enterprise Miner to help you ease through this process.
  - For handling bigger data sets require you are required to have skills in Hadoop, Map Reduce or Spark. These tools can help you scrub the data by scripting.

Explore Data

- Once your data is ready to be used, and right before you jump into AI and Machine Learning, you will have to examine the data.
- Usually, in a corporate or business environment, your boss will just throw you a set of data and it is up to you to make sense of it.
- So it will be up to you to help them figure out the business question and transform them into a data science question.

Explore Data

- To achieve that, we will need to explore the data. First of all, you will need to inspect the data and its properties.
- Different data types like numerical data, categorical data, ordinal and nominal data etc. require different treatments.
- Then, the next step is to compute descriptive statistics to extract features and test significant variables. Testing significant variables often is done with correlation.
- For example, exploring the risk of someone getting high blood pressure in relations to their height and weight. Do note that some variables are correlated, but they do not always imply causation.

Explore Data

- The term “Feature” used in Machine Learning or Modelling, is the data features that help us to identify the characteristics that represent the data.
- For example, “Name”, “Age”, “Gender” are typical features of members or employees dataset.
- Lastly, we will utilise data visualisation to help us to identify significant patterns and trends in our data.
- We can gain a better picture through simple charts like line charts or bar charts to help us to understand the importance of the data.

Explore Data

- Skills Required
  - If you are using Python, you will need to know how to use Numpy, Matplotlib, Pandas or Scipy; if you are using R, you will need to use GGplot2 or the data exploration swiss knife Dplyr.
  - On top of that, you need to have knowledge and skills in inferential statistics and data visualization.
  - As much as you do not need a Masters or Ph.D. to do data science, these technical skills are crucial in order to conduct an experimental design, so you are able to reproduce the results.

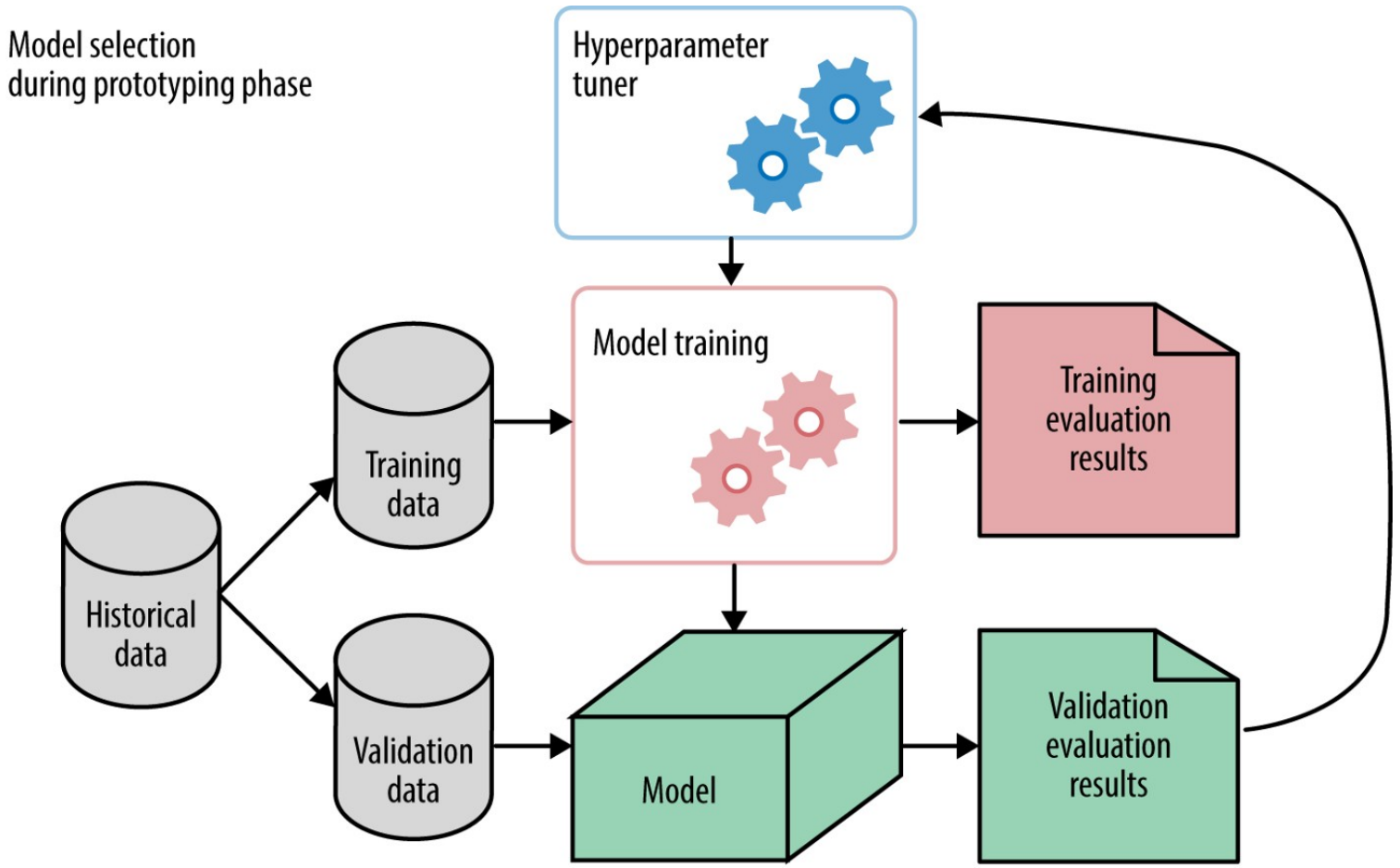
Explore Data

- Additional Tips:
  - Be curious. This can help you develop your spidey senses to spot weird patterns and trends.
  - Focus on your audience, and understand their background and lingo. So that you are able to present the data in a way that makes sense to them.

Model Data

- This is the stage where most people consider interesting. As many people call it “where the magic happens”.
- Once again, before reaching this stage, bear in mind that the scrubbing and exploring stage are equally crucial to building useful models. So take your time on those stages instead of jumping right to this process.
- One of the first things you need to do in modelling data is to reduce the dimensionality of your data set.
- Not all your features or values are essential to predicting your model. What you need to do is to select the relevant ones that contribute to the prediction of results.

Model Data


Model selection
during prototyping phase
Hyperparameter
tuner
Model training
Training
evaluation
results
Training
data
Historical
data
Validation
data
Model
Validation
evaluation
results

Model Data

- There are a few tasks we can perform in modelling. We can also train models to perform classification to differentiating the emails you received as “Inbox” and “Spam” using logistic regressions.
- We can also forecast values using linear regressions. We can also use modelling to group data to understand the logic behind those clusters.
- For example, we group our e-commerce customers to understand their behaviour on your website.
- This requires us to identify groups of data points with clustering algorithms like k-means or hierarchical clustering.

Model Data

- Skills Required
  - In Machine Learning, the skills you will need is both supervised and unsupervised algorithms. For libraries, if you are using Python, you will need to know how to use Sci-kit Learn; and if you are using R, you will need to use CARET.
  - After the modelling process, you will need to be able to calculate evaluation scores such as precision, recall and F1 score for classification.
  - For regressions, you need to be familiar with R^2 to measure goodness-of-fit, and using error scores like MAE (Mean Average Error), or RMSE (Root Mean Square Error) to measure the distance between the predicted and observed data points.

Interpreting Data

- We are at the final and most crucial step of a data science project, interpreting models and data.
- The predictive power of a model lies in its ability to generalise. How do we explain a model depends on its ability to generalise unseen future data.
- Interpreting data refers to the presentation of your data to a non-technical layman.
- We deliver the results in to answer the business questions we asked when we first started the project, together with the actionable insights that we found through the data science process.

Interpreting Data

- Actionable insight is a key outcome that we show how data science can bring about predictive analytics and later on prescriptive analytics.
- In which, we learn how to repeat a positive result, or prevent a negative outcome.
- On top of that, you will need to visualise your findings accordingly, keeping it driven by your business questions.
- It is essential to present your findings in such a way that is useful to the organisation, or else it would be pointless to your stakeholders.

Interpreting Data

- In this process, technical skills only are not sufficient.
- One essential skill you need is to be able to tell a clear and actionable story. If your presentation does not trigger actions in your audience, it means that your communication was not efficient.
- Remember that you will be presenting to an audience with no technical background, so the way you communicate the message is key.

Interpreting Data

- Skills Required
  - In this process, the key skills to have is beyond technical skills. You will need strong business domain knowledge to present your findings in a way that can answer the business questions you set out to answer, and translate them into actionable steps.
  - Apart from tools needed for data visualization like Matplotlib, ggplot, Seaborn, Tableau, d3js etc., you will need soft skills like presentation and communication skills, paired with a flair for reporting and writing skills will definitely help you in this stage of the project lifecycle.

Summary

- We have presented here are the steps that data scientists follow chronologically in a typical data science project.
- If it is a brand new project, we usually spend about 60–70% of our time just on gathering and cleaning the data. Since it is a framework, you may use it as a guideline with your favorite tools.
- The true north is always that business questions we defined, before even started the data science project.
- Always remember that solid business questions, clean and well-distributed data always beat fancy models.

Open Exit: Reaching the End of the Data Lifecycle

- Scientific research data often have a longer lifespan than the project that creates them.
- This is particularly true when good research data management practice is put into place.
- Good data management throughout the data lifecycle is essential for successful long-term preservation and sharing, ensuring a long lifespan of use for research data.
- In addition to good data management, many of us would agree that the importance, impact, and relevance of one's research data often influences the potential long-term value of it—that is that relevance, value assessment, and retention are all closely linked.

Open Exit: Reaching the End of the Data Lifecycle

- Yet it remains uncertain whether or not the retention of data increases their inherent value. More fundamentally, do data in the lifecycle smoothly progress from one stage to another without a gap or an exit?
- Should review, assessment, and evaluation functions for scientific records and data be included at every stage prior to reaching the end of the data lifecycle?
- These questions and similar inquiries about the lifespan (and ‘death’ of data1) have motivated us to investigate a variety of actions involved in curation decisions for data retention or deletion.

Open Exit: Reaching the End of the Data Lifecycle

- In order to advance our understanding of the actions and decisions that adequately safeguard data for future use, we examine a variety of technical, legal, and institutional responses, controls, and resources that influence actions (and the actors involved in these actions) to retain or not retain the data.
- Three areas provide context for discussion: university records & information management, library collections management, and data curation.

Open Exit: Reaching the End of the Data Lifecycle

- As we approach disposition and end-of-lifecycle issues from the three perspectives (university records & information management, library collections, and data curation), we focus on the following five areas:
  - 1) Terminology (usage and interpretation);
  - 2) Scope (types, formats, and uses of objects);
  - 3) Authority (actors and directives);
  - 4) Appraisal Criteria (actions and factors that influence those actions); and,
  - 5) Resources (human, financial, and physical space)

Open Exit: Reaching the End of the Data Lifecycle

Table 1: Comparison in End of Lifecycle Terminology

	University records & information management	**Library Collection management**	**Data Curation**
Terminology	Official record; Active/inactive records; Disposition: retention or destruction	Collection; Maintenance; Weeding; Deaccessioning; "Data-driven" deaccession; Deselection	Digital content; Retention; Appraisal/Re-appraisal; Selection/Acquisition; Data transfer/migration; Disposition; Destruction

Scope

	University records & information management	Library Collection management	Data Curation
Scope	Theoretically embraces all information created by organization; includes any information created in support of the organization's mission or in fulfillment of its legal obligations. Disposition of paper records is often more effective than electronic records management.	Everything the library or archive subscribes to or collects, including provisions for gift and legacy materials: books, journals, digital resources, media, hardware, software, etc., There is a difference between discarding the object and discarding the metadata.	Theoretically everything that researchers generate out of research projects—recorded factual material commonly accepted in the scientific community as necessary to validate research findings.[9] Decisions are often influenced by types of data, state of data (e.g., raw, primary, analyzed, published) and the sensitivity of data.

Authority

Table 3: Comparison in Authority

	University records & information management	**Library Collection management**	**Data Curation**
Authority	Records schedule;	Collection	Data steward;
(actors and directives)	Retention schedule; Records manager or records coordinator; Legal directives.	Developer/Manager; Librarians; State-level directives; Consortial policies.	Data producer; Repository collection policies; Institutional regulations; Funder directives.

Apraisal Criteria

Table 4: Comparison in Appraisal Criteria

	University records & information management	**Library Collection management**	**Data Curation**
Appraisal Criteria	Criteria include: liability; administrative need; superseded, obsolete, rescinded; time period after event/action.	Criteria include: space; currency; subject; coverage; usage/cost-per-use; duplication (in format or consortial location).	Criteria include: funder ROI; compliance; (repository) collection alignment; scientific/historical/continuing value of data (in terms of re-usability); quality; integrity.

Apraisal Criteria

Table 5: Comparison in Resources

	University records & information management	**Library Collection management**	**Data Curation**
Resources	Costs of staffing and centralized records management program; costs of staff time and resources for records management tasks within records-creating units	Costs: budget and subscription/purchase models, staffing resources, space resources (physical); digital space counted by numbers of titles/items than by storage size.	Storage/backup costs; Preservation costs; Cost of creating and managing preservation metadata (to ensure discoverability).

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com