

What is Data Science?

Leen-Kiat Soh

Computer Science & Engineering
University of Nebraska, Lincoln, NE

Note on Perspectives

- First part is based on the ACM Data Science Task Force's Computing Competencies for the **Undergraduate Data Science Curricula** Draft 2 (December 2019)
- Second part is based on diagrams found online from both the **industry and academia**

From ACM Data Science Task Force

Brief Background

- In 2009, Turing award winner Jim Gray spoke of **data science as a fourth paradigm of science** (empirical, theoretical, computational and data-driven) arising from and capitalizing on the huge amount of data that is now available for investigation
- For CS, the emergence of data science offers both tremendous opportunity and something of a conundrum, as once again the emergence of a new and closely related computing practice or field raises inevitable questions about **whether and how it fits into current postsecondary computer science curricula**

Purpose of Report

- This document represents an effort by the ACM Education Board through the work of the Data Science Task Force to answer this question
- It is an effort to put our own data science house in order
- This document is ***not***, however, an effort to claim ownership or even primacy in data science
 - To do so would be to negate the powerful **interdisciplinarity** that data science makes possible

Charter

- At the August 2017 ACM Education Council meeting, a task force was formed to explore a process to add to the broad, interdisciplinary conversation on data science, with an articulation of the role of computing discipline-specific contributions to this emerging field
- There is a shared desire to form **a broad interdisciplinary definition of data science** and to develop **curriculum guidance for degree programs in data science**
- There is a need to acknowledge the definition and description of the individual contributions to this interdisciplinary field
 - E.g., those interested in the business context for these concepts generally use the term “**analytics**”
 - E.g., the abbreviation DSA appears, meaning Data Science and Analytics

Motivating the Study of Data Science

- Those who study Data Science have to **develop a mind set with a strong focus on data**
 - the collection of data and, through analyzing it appropriately, using this to bring about beneficial insights and changes
- Students of Data Science need to be imbued with the **‘joy of data,’ seeing data as the ‘currency or fuel of our time’**
- They also need to be imbued with a **strong sense of professional and ethical responsibility**

Current View of Data Science

- Data Science is an **inherently interdisciplinary** field
- Data Science is the field that brings together **domain data**, **computer science**, and the **statistical** tools for interrogating the data and extracting useful information
 - the domain that provides the data, expectation, understanding, needs for tools and techniques;
 - statistics for analysis, modeling, and inference; and
 - computer science for data access, management, protection, as well as effective processing in modern computer architectures

Current View of Data Science2

- Early programs in Data Science will often work with a group of existing courses from the participating disciplines
 - practical and easy to bring a new program into existence
 - But ... it is difficult to make the essential connections so that all the parts work together to support discovery and decision making in the domain

Prior Work that Informed the Report

- The EDISON Data Science Framework (2018)
- The National Academies of Science, Engineering, and Medicine Report on Data Science for Undergraduates (2018)
- The Park City Report (2017)
 - Includes an outline of the Data Science Major
 - ACM Task Force report builds on this work with **a heavy orientation toward CS**
- The Business Higher Education Framework (BHEF) Data Science and Analytics (DSA) Competency Map (2016)
 - Provides a 4-tier competency map
 - Tier-2 is most relevant (post-secondary)
- Business Analytics Curriculum for Undergraduate Majors (2015)
- Initial workshops related to this ACM Data Science Curriculum effort (2015)

Prior Work that Informed the Report: Summary

- **Task Force's position:** any Data Science program will have to reflect competencies in math, statistics, and CS, possibly with different emphases
 - consistent with the view of the National Academies report
- Important to capture in a single volume the **contributions that computing makes to data science**
 - **Computing opens up many avenues for data collection**
 - Internet of Things, sophisticated sensors, face recognition and voice recognition, automation, etc.,
 - Computing can play a vital role as a **custodian of information** with great attention being paid to **maintenance** but crucially also to **security** and **confidentiality** matters
 - Then the **analysis of large amounts** of information and utilization of that for the purposes of **machine learning** or **augmented intelligence** in its various roles can bring significant benefit

Body of Knowledge: **Knowledge Areas**

- Analysis and Presentation (AP)
- Computing and Computer Fundamentals (CCF)
- Programming, Data Structures, and Algorithms (PDA)
- *Artificial Intelligence (AI)*
- *Big Data Systems (BDS)*
- *Data Mining (DM)*
- *Machine Learning (ML)*
- Data Acquisition, Management, and Governance (DG)
- Data Privacy, Security, Integrity, and Analysis for Security (DP)
- Professionalism (PR)
- Software Development and Maintenance (SDM)

Body of Knowledge: Knowledge Areas with Sub-Domains (computing oriented)

Question: Where's statistics?

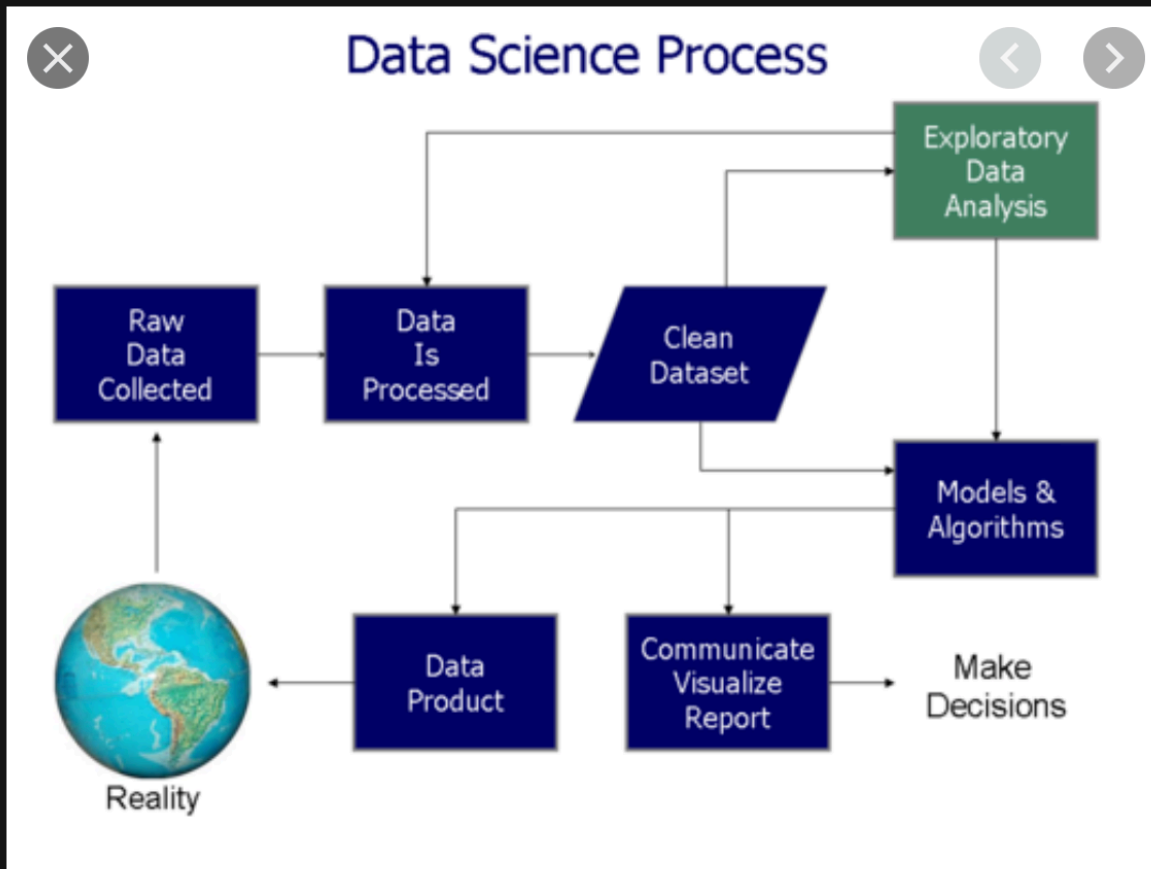
- Embedded into various sub-domains

Analysis and Presentation <ul style="list-style-type: none">• Foundational considerations• Visualization• User-centered design• Interaction design• Interface design and development Artificial Intelligence <ul style="list-style-type: none">• General• Knowledge representation and reasoning<ul style="list-style-type: none">– logic based• Knowledge representation and reasoning<ul style="list-style-type: none">– probability based• Planning and search strategies Big Data Systems <ul style="list-style-type: none">• Problems of scale• Big data computing architectures• Parallel computing frameworks• Distributed data storage• Parallel programming• Techniques for Big Data applications• Cloud computing• Complexity theory• Software support for Big Data applications Computing and Computer Fundamentals <ul style="list-style-type: none">• Basic computer architecture• Storage systems fundamentals• Operating system basics• File systems• Networks• The web and web programming• Compilers and interpreters Data Acquisition, Management, and Governance <ul style="list-style-type: none">• Data acquisition• Information extraction• Working with various types of data• Data integration• Data reduction and compression• Data transformation• Data cleaning• Data privacy and security	Data Mining <ul style="list-style-type: none">• Proximity measurement• Data preparation• Information extraction• Cluster analysis• Classification and regression• Pattern mining• Outlier detection• Time series data• Mining web data• Information retrieval Data Privacy, Security, Integrity, and Analysis for Security <ul style="list-style-type: none">• Data privacy• Data security• Data integrity• Analysis for security Machine learning <ul style="list-style-type: none">• General• Supervised learning• Unsupervised learning• Mixed methods• Deep learning Professionalism <ul style="list-style-type: none">• Continuing professional development• Communication• Teamwork• Economic considerations• Privacy and confidentiality• Ethical considerations• Legal considerations• Intellectual property• On automation Programming, data structures and algorithms <ul style="list-style-type: none">• Algorithmic thinking and problem solving• Programming• Data structures• Algorithms• Basic complexity analysis• Numerical computing Software development and maintenance <ul style="list-style-type: none">• Software design and development• Software testing
--	---

Figure 3-2 The (Computing) Data Science Knowledge Areas (with sub-domains)

From the Industry & Academia

(a very small sample ...)

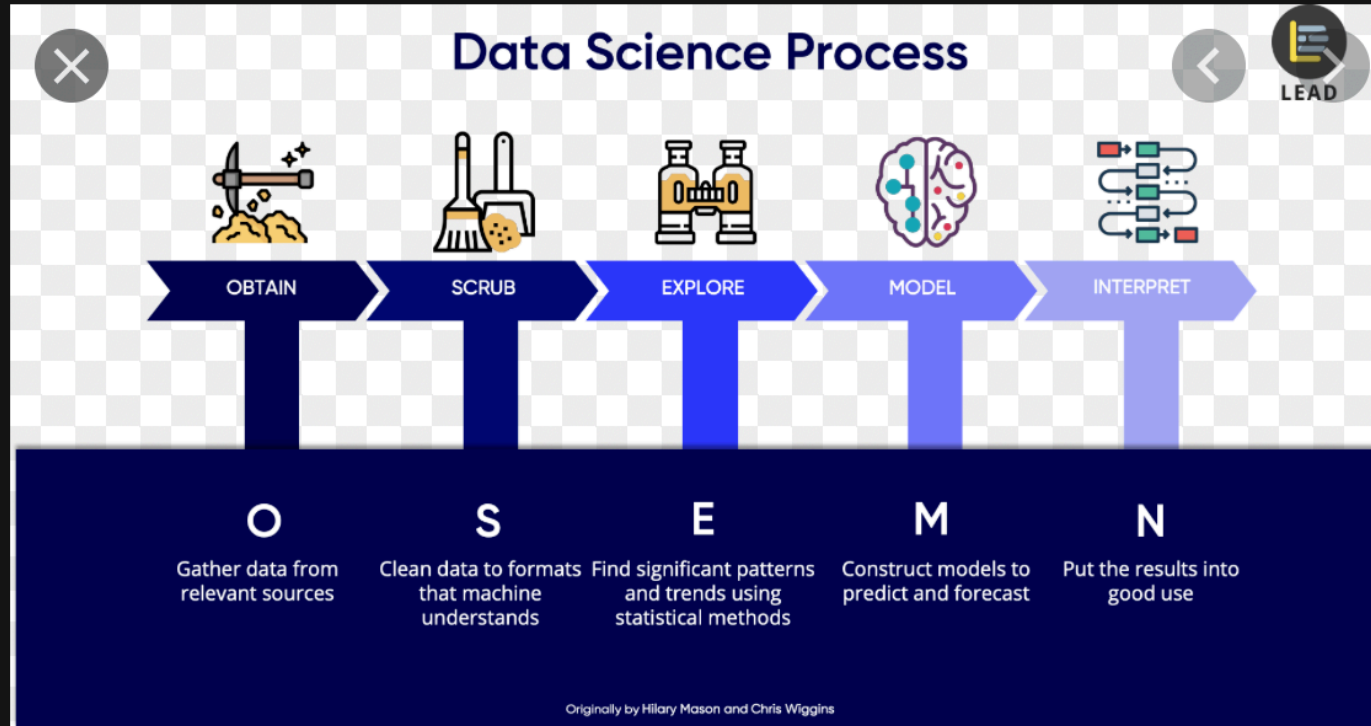


KDnuggets



The Data Science Process

Images may be subject to copyright. [Learn More](#)

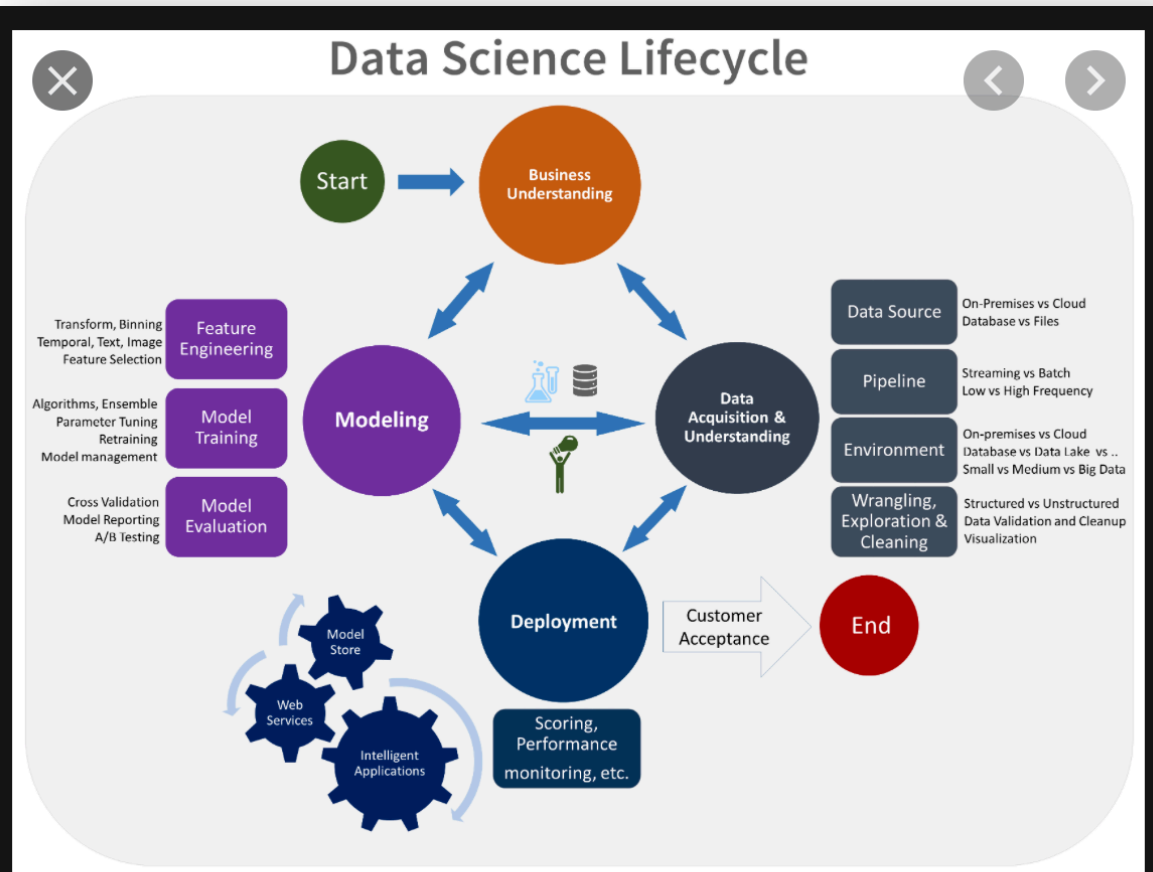



Towards Data Science

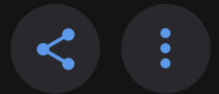


5 Steps of a Data Science Project Lifecycle | by Dr. Cher Han Lau ...

Images may be subject to copyright. [Learn More](#)

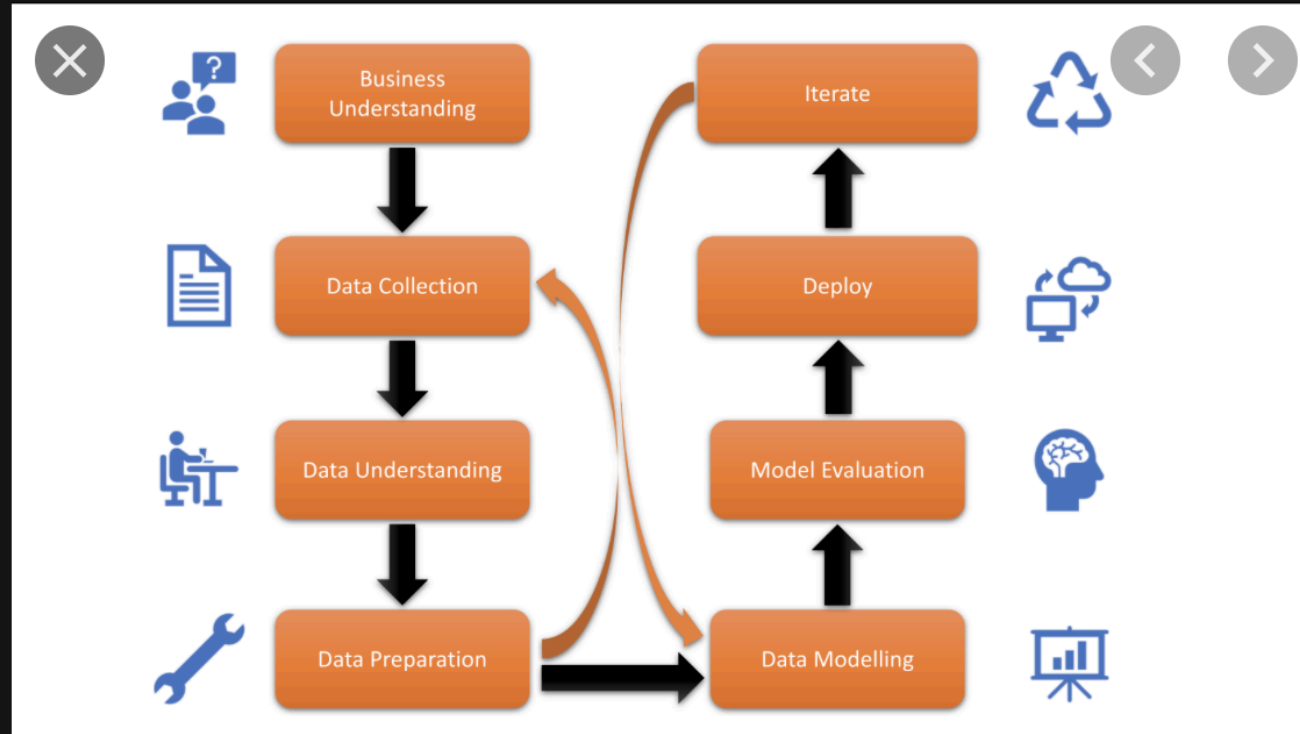


 Microsoft Docs



The Team Data Science Process lifecycle |
Microsoft Docs

Images may be subject to copyright. [Learn More](#)

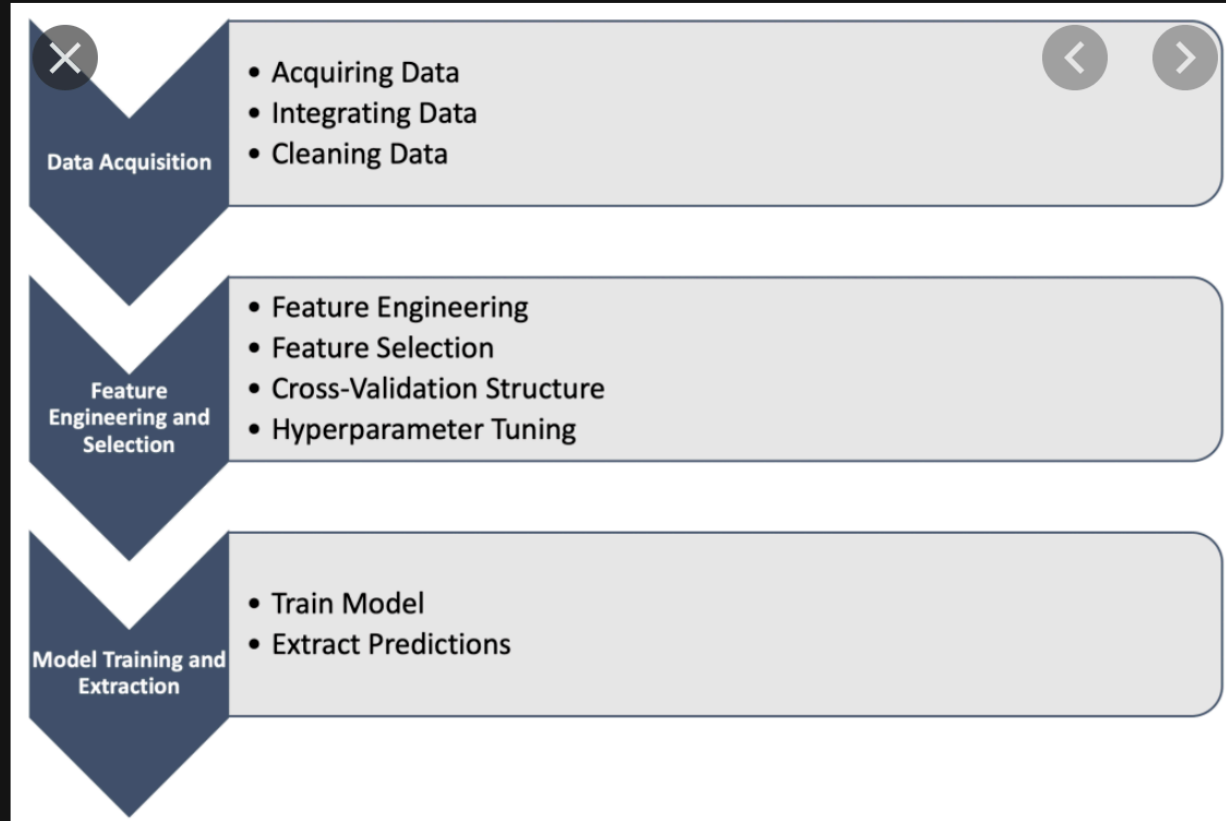


Towards Data Science



Life Cycle of a Data Science Project | by Rishi Sidhu | Towards ...

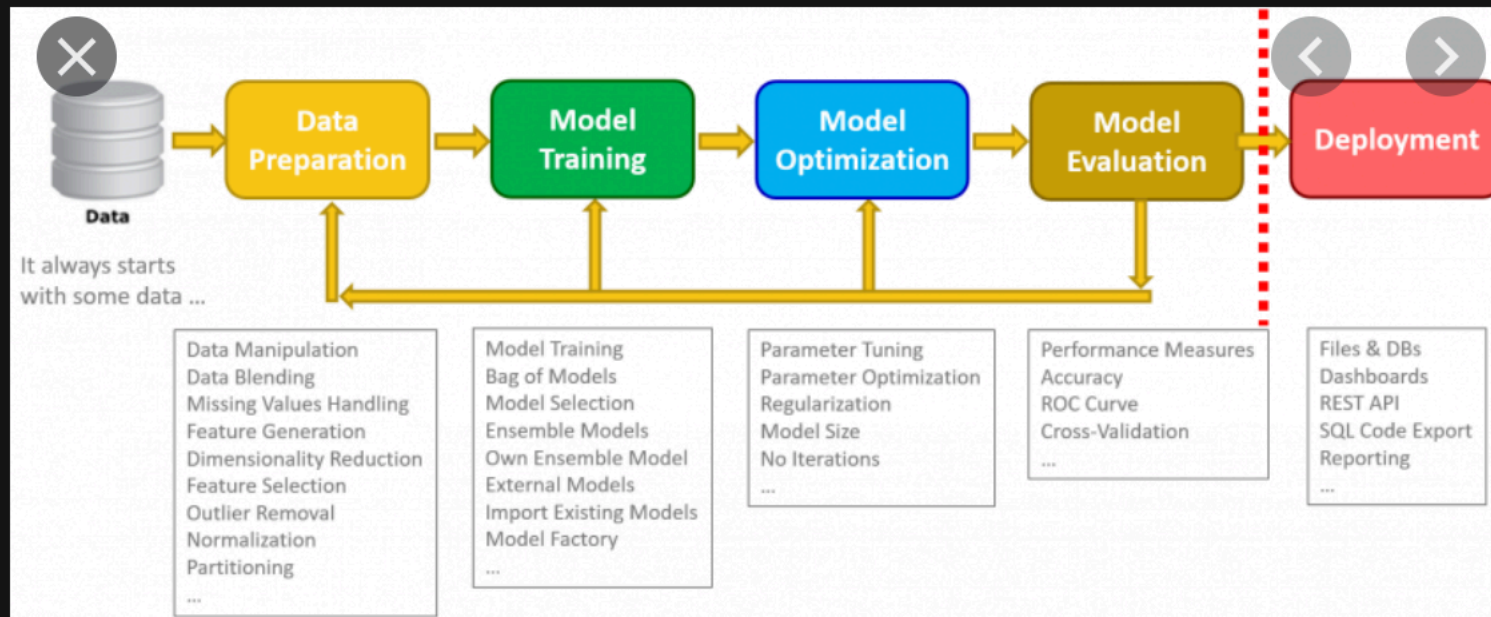
Images may be subject to copyright. [Learn More](#)



Cloudvane.net



The Data Science Process - Cloudvane Data Science

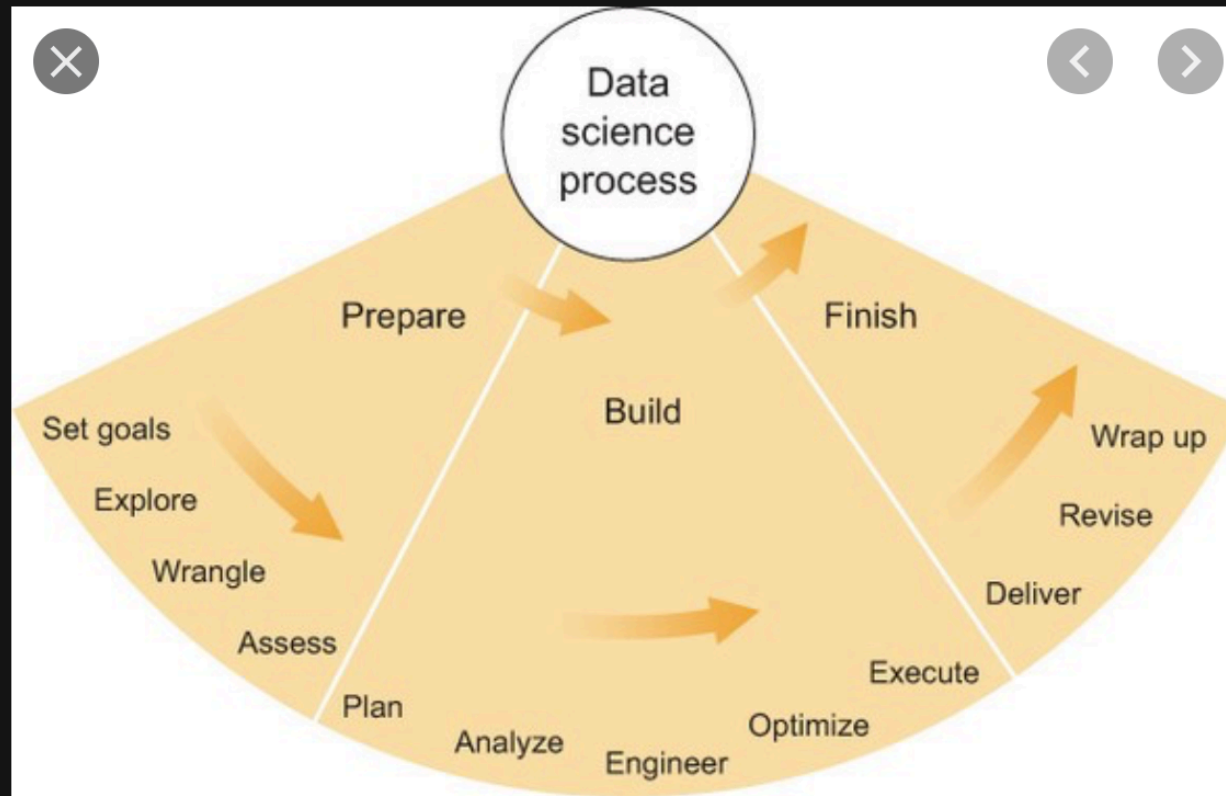


Towards Data Science



Best Practice for Data Science Projects | by Murat Yalcin ...

Images may be subject to copyright. [Learn More](#)



Medium



How to Think Like a Data Scientist in 12 Steps | by James Le ...

Images may be subject to copyright. [Learn More](#)

Gain Data Science Skills in 10 steps with the Microsoft Data Science Track



Do you want to know more? Visit www.md2c.nl