Data Scientist

Prepared by Dr. Salimah Mokhtar

If you are asking questions and using data to find answers, you are a __________.

Learning Objectives

1. To describe data scientist as a new profession.
2. To explain the path and pillars to becoming a data scientist.
3. To discuss the 12 tasks that data scientist do.
4. To summarize the skills and training for a data scientist.
5. To plan a data science portfolio.

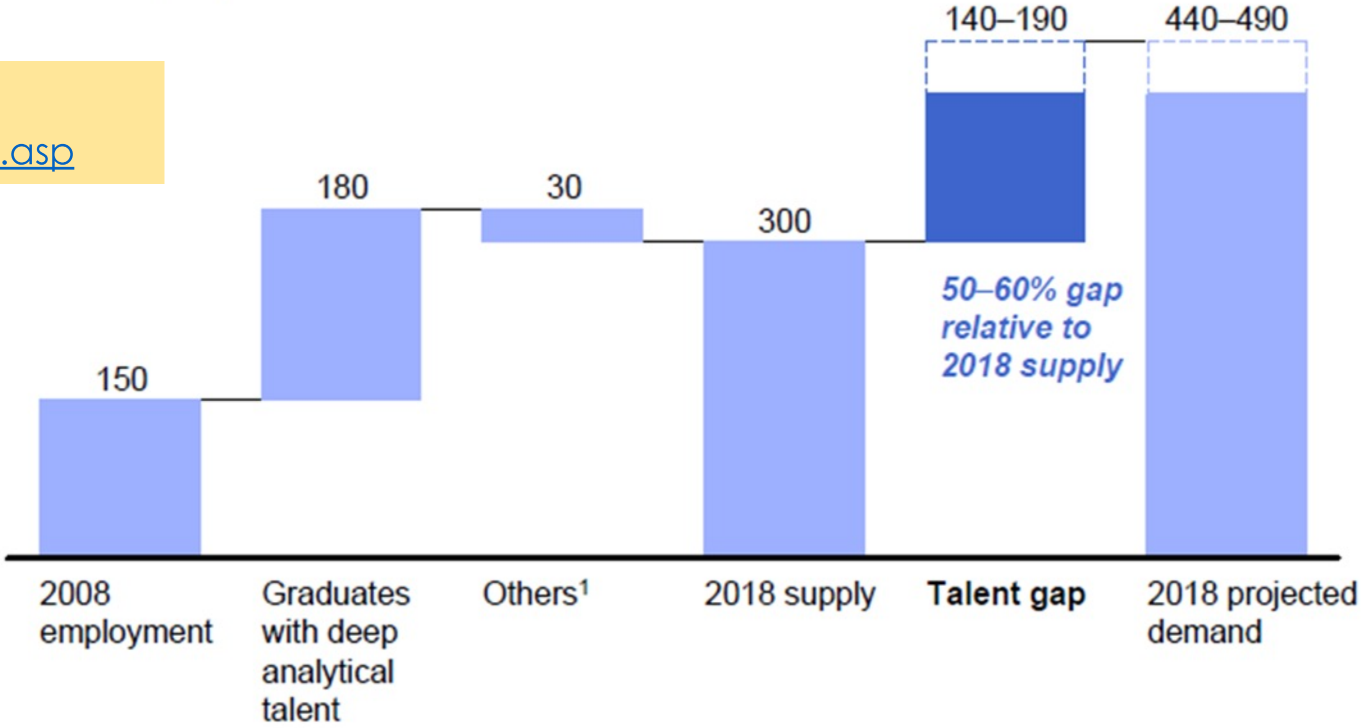
What REALLY is Data Science? Told by a Data Scientist

https://youtu.be/xC-c7E5PK0Y

Data Scientist: New Profession and Opportunities

McKinsey Institute on Big Data Jobs (2011)
http://www.mckinsey.com/mgi/publications/big_data/index.asp

There is **a shortage of talent** necessary for organizations to take advantage of Big Data.
By 2018, the United States alone could face a **shortage** of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.



Malaysia Digital Economy Corp (MDEC) targets to have **20,000** big data professionals by **2020**.

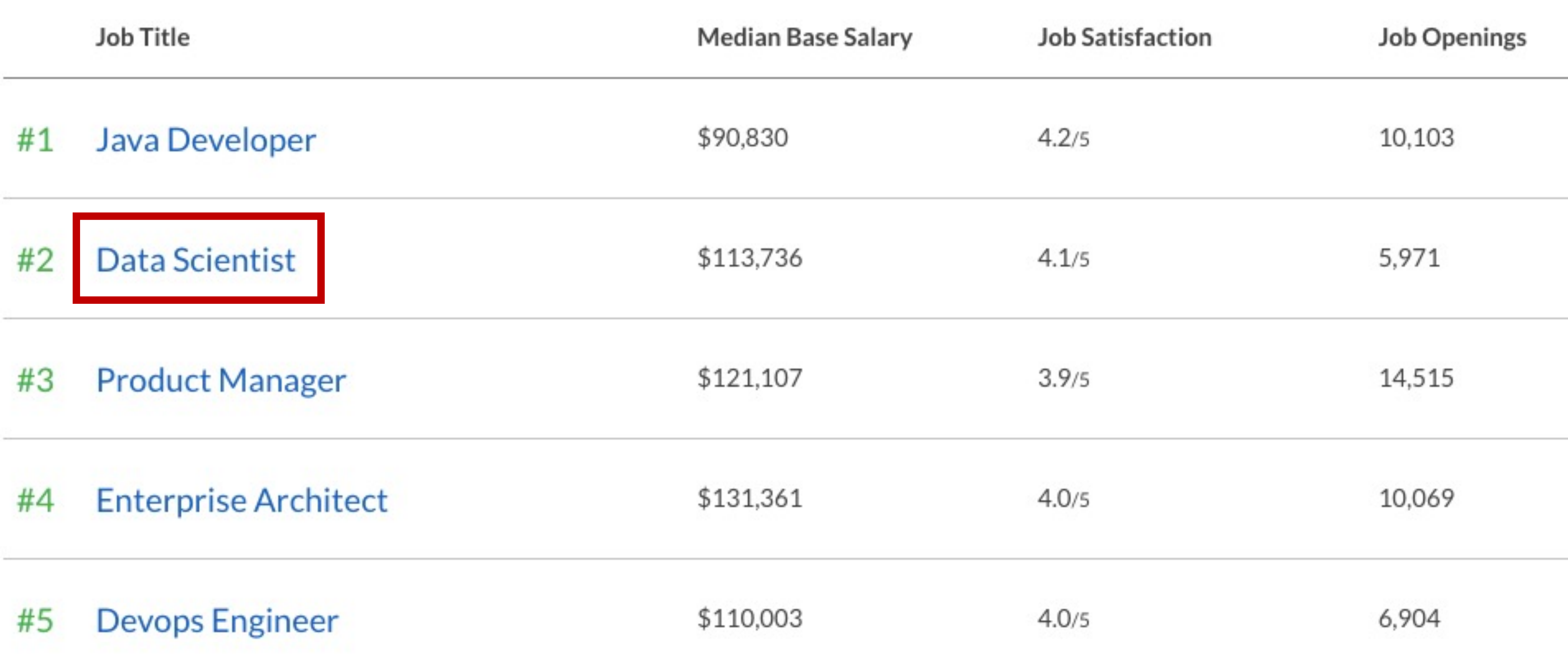
MALAYSIA CONTINUES TO GROOM DATA PROFESSIONALS

During Big Data Week Kuala Lumpur 2018 (4th Oct), it was announced that Malaysia currently has around **8,000** professionals in big data, and it is on track in achieving the target of 20,000 data professionals by year 2020.

- In the effort to embrace the 4th industrial revolution, Malaysia now needs to prepare its talent for AI.
  - **Data professionals** need to be equipped with the know-how of Artificial Intelligence.

Glassdoor ranks **data scientist** as the best job in the US from **2016 – 2019**.

Glassdoor - Best Jobs in America for 2021 (Top 5)



	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Java Developer	$90,830	4.2/5	10,103
#2	Data Scientist	$113,736	4.1/5	5,971
#3	Product Manager	$121,107	3.9/5	14,515
#4	Enterprise Architect	$131,361	4.0/5	10,069
#5	Devops Engineer	$110,003	4.0/5	6,904

Some tasks that might reasonably be labeled as part of doing data science.

What constitutes data science?

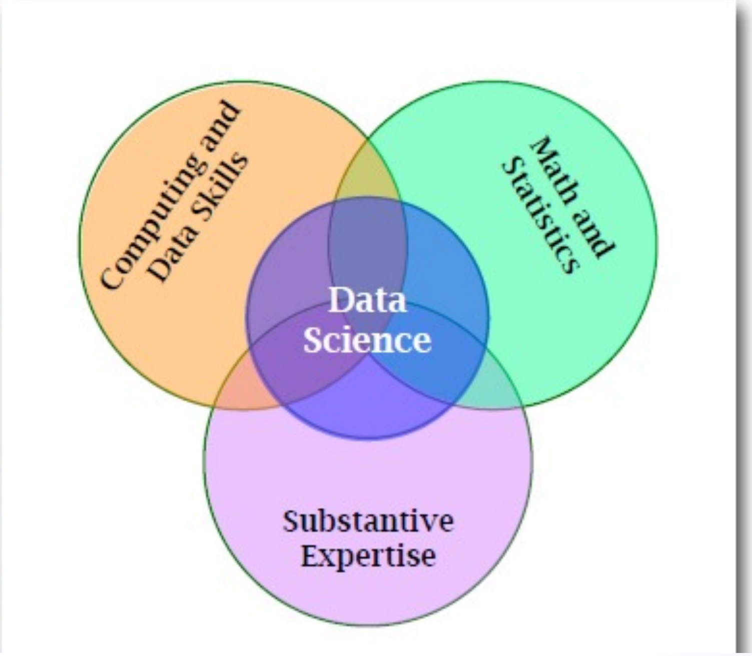
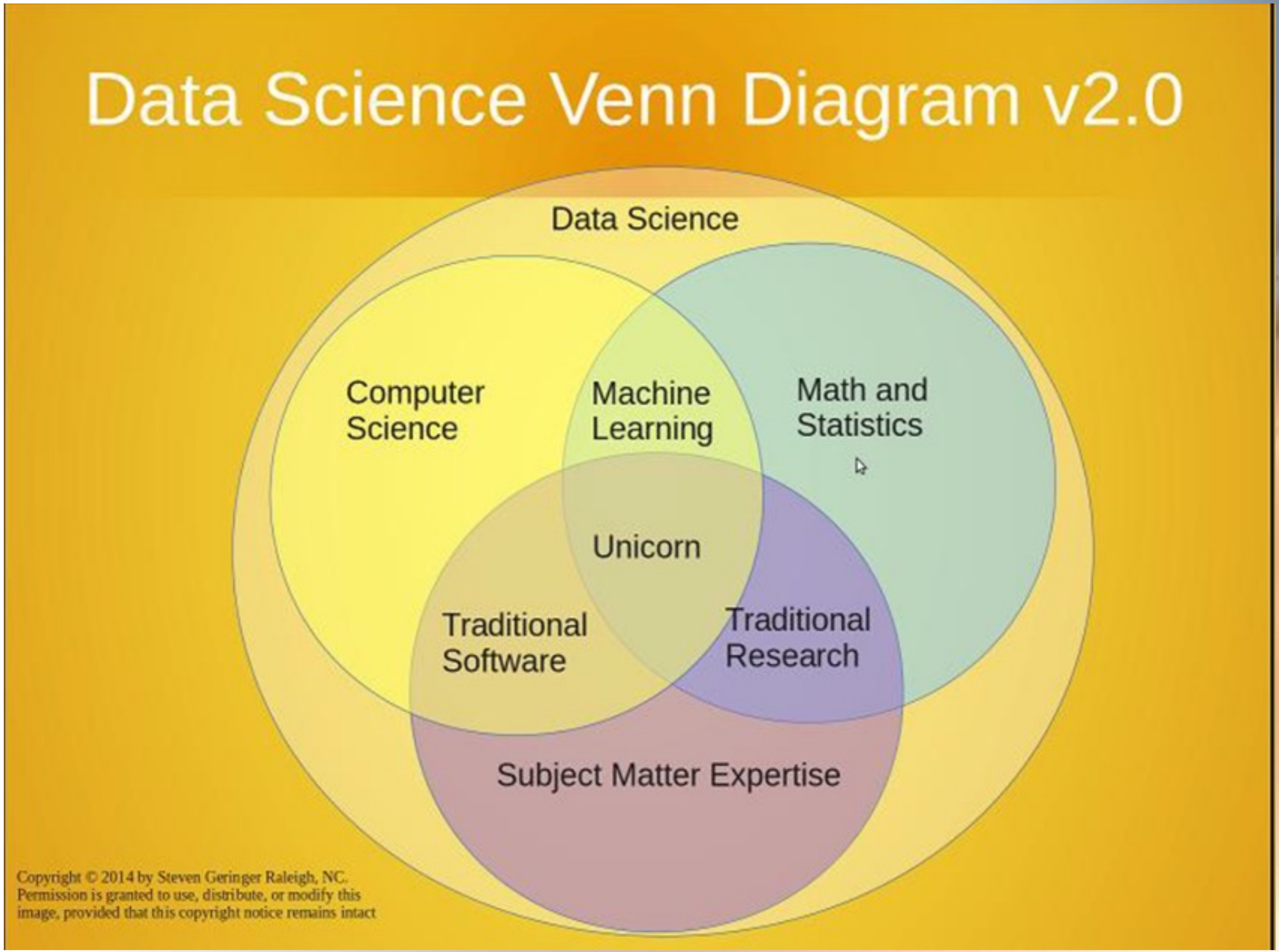


Fig 1. Data Science Venn diagram adapted from Conway (2010)

Table 1. Representative data science tasks and the portion of the Data Science Venn diagram from which they primarily draw. (CD = Computing and Data Skills, MS = Math and Statistics, SE = Substantive Expertise)

Task	Primary Expertise
Set up a server as a repository of data streaming in real time from a large array of geographically distributed sensors.	?
Explain the origin of outliers in a particular data set.	?
Decide to what extent the conclusions drawn from analysis can be generalized.	?
Design a data visualization suitable for publication in an article for non-experts.	?
Decide what data should be gathered.	?
Decide whether certain disparate data sets can be meaningfully merged.	?
Automate the merger of multiple data sets.	?
Detect the signal within the noise.	?
Reduce the number of variables that need to be considered for a particular analysis.	?
Set up a version management system for data that will be gathered over a number of years.	?

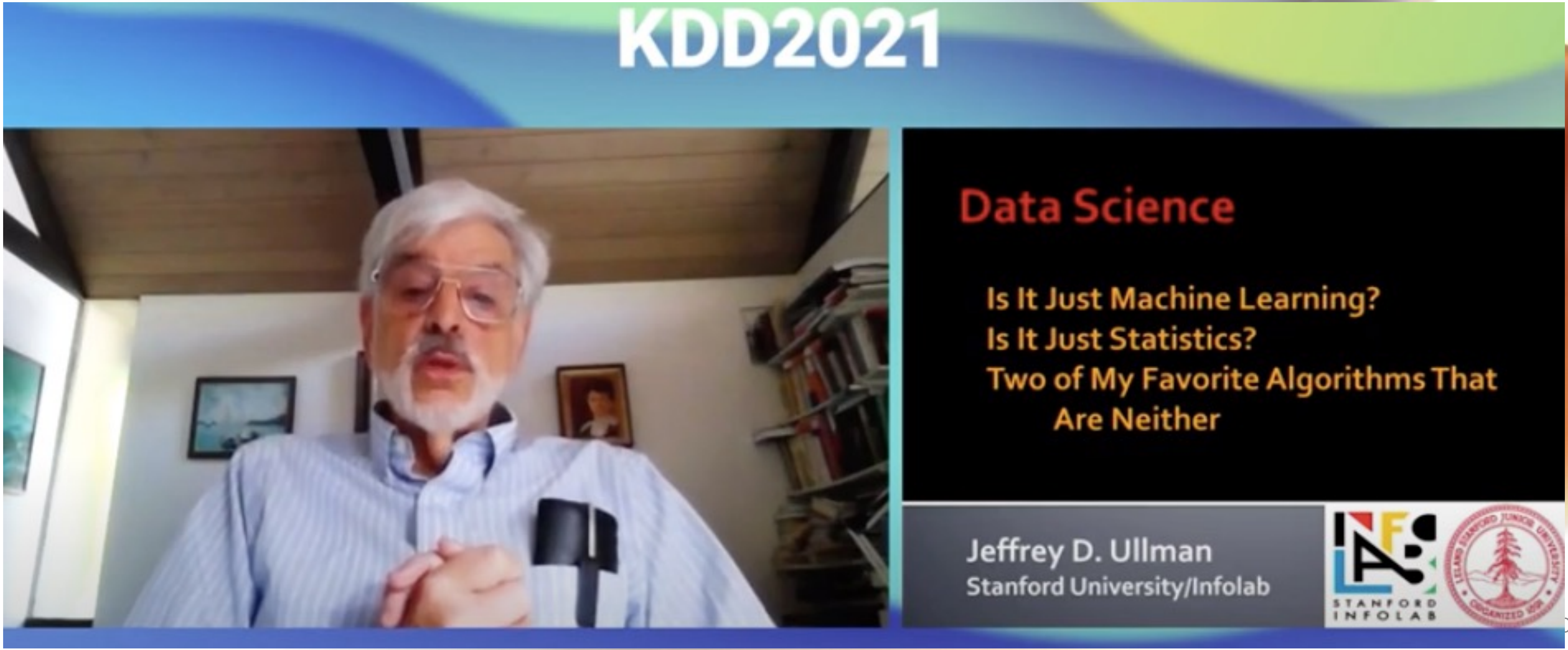
Conway, D. (2010). The data science venn diagram. Dataists [Web page]. Retrieved September 9, 2016, from http://www.dataists.com/2010/09/the-data-science-venn-diagram/



Data Science Venn Diagram v2.0

KDD 2021 Keynote Talk--On the Nature of Data--Jeffrey D Ullman

https://youtu.be/Kkx--T5NUy4



KDD2021

Data Science

Is It Just Machine Learning?

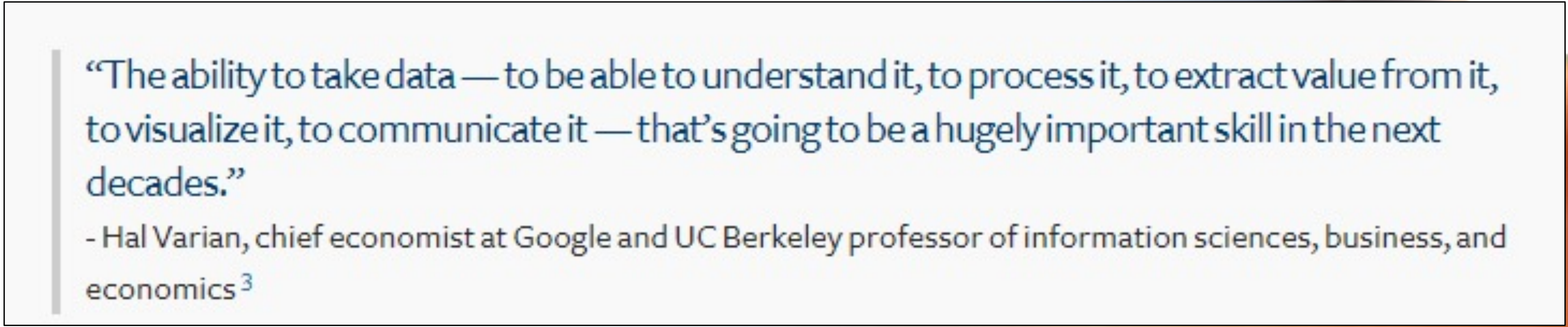
Is It Just Statistics?

Two of My Favorite Algorithms That Are Neither

Jeffrey D. Ullman

Stanford University/Infolab

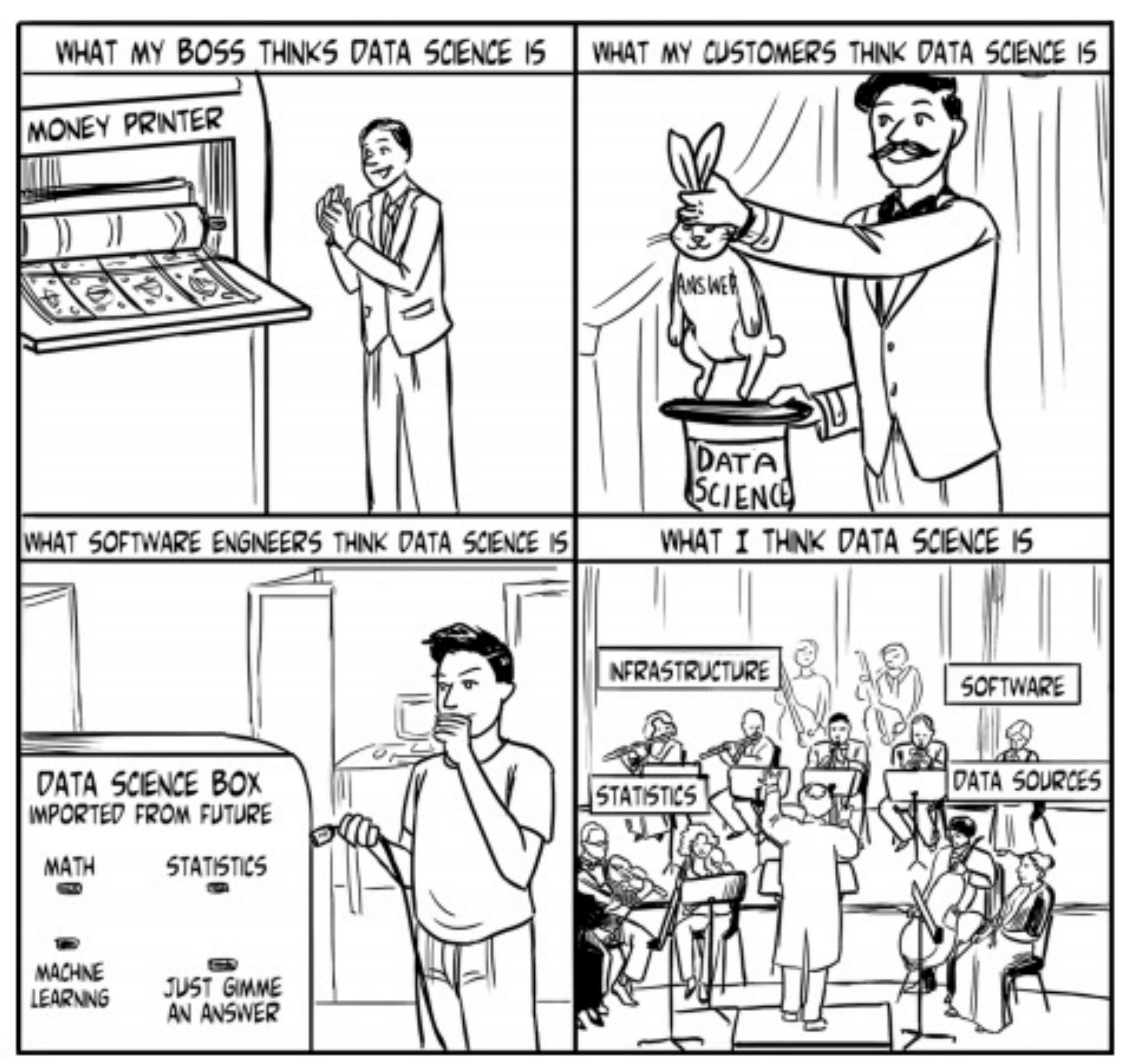
Data science is using **data** to answer questions.

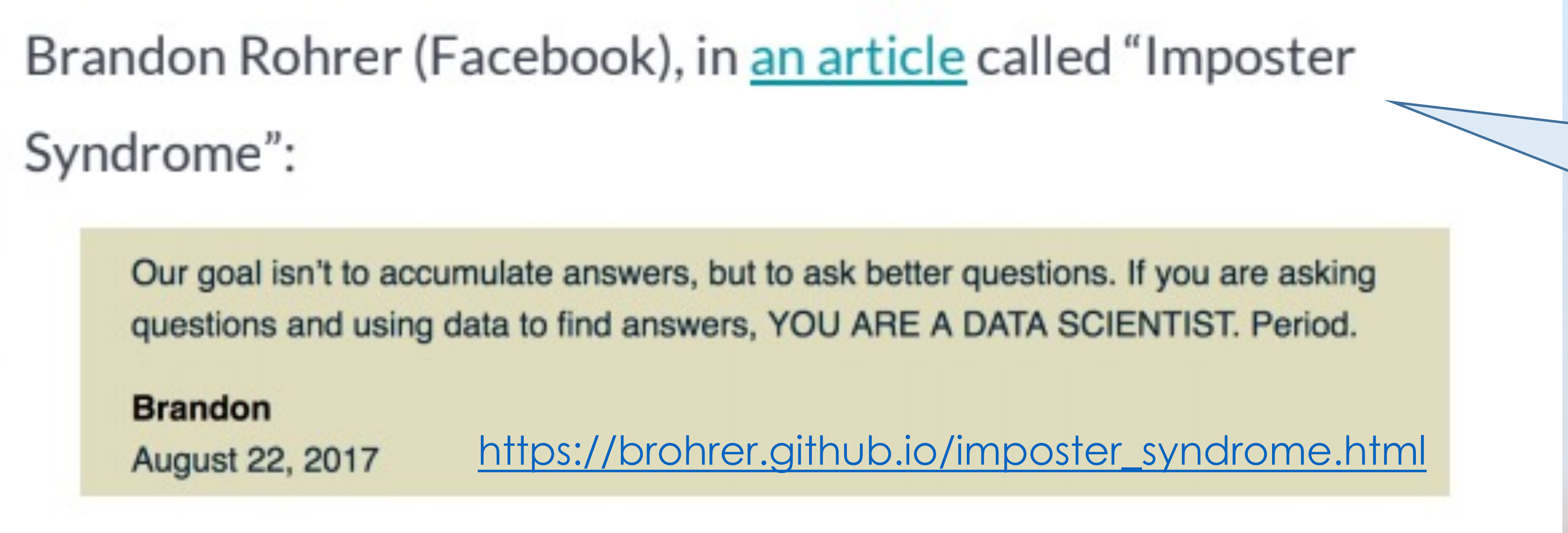


> “The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”
>
> - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics[3]

Data science is the science of analyzing raw data using **statistics** and **machine learning** techniques with the purpose of drawing **conclusions** about that information.

Some stereotypical perspectives on data science





Brandon Rohrer (Facebook), in an article called "Imposter Syndrome":

> Our goal isn't to accumulate answers, but to ask better questions. If you are asking questions and using data to find answers, YOU ARE A DATA SCIENTIST. Period.
>
> **Brandon**
> August 22, 2017

https://brohrer.github.io/imposter_syndrome.html

Everyone else in DS field is more qualified than me!!!

There are two paths forward: Generalist and Specialist.

A good generalist

is superficially familiar with every part of data science, recognizes all the jargon and technical terms, has a good notion of what tools and expertise are needed to solve a given problem, and asks insightful questions in technical reviews.

A good specialist

understands one area deeply, can explain their area of expertise to non-experts, understands the tradeoffs between different approaches, is up to date on current research and new tools, and can use their tools quickly to produce high-quality results.

Navigating the Sprawling Terrain of Data Science

The reason data science feels so big is because it’s no longer a single field (multi-disciplinary). There are 3 **separate pillars**.

Data Analysis - turning raw information into knowledge that can be acted on.

Data Modeling - using the data we have to estimate the data we wish we had.

Data Engineering - making everything work faster, more robustly, and at greater scale.

Reference: https://youtu.be/_2g7yqwgAVg

Data Analysis

There are several aspects of data analysis, the process of taking data and turning it into information that we can use to make a decision.

- **Domain knowledge** - translate a business need to a question, it involves making trade-offs between how **accurate the answer** should be and how much **time and money** you want to spend finding it (accuracy-cost trade-offs).
- **Research** - translating information into actions: gather the data, design and conduct experiments.
- **Interpretation** – Given a large collection of data, we can summarize, aggregate, visualize (turning numbers into a picture), and apply statistical tools for statistical summaries.

Data Modeling

Data modeling also has several major subcategories, often referred to as machine learning.

Creating a simplified description of your data that you can use to make estimates for data that you have not measured.

Supervised learning - uses labelled examples to discover patterns e.g., classification (if labels are categorical), regression (if labels are numerical), anomaly detection (to determine whether examples are in line with previously seen patterns).

Unsupervised learning - the data does not come with labels e.g., clustering (to discover patterns), dimensionality reduction (groups of variables that tend to behave similarly).

Custom algorithm development - feature engineering, numerical optimization

Data Engineering

Data engineering is the third pillar of data science. It has a few major areas.

Data management - is the storing, moving, and handling of data. Include data base management, pipeline construction, data collection.

Production – taking code that works well in a prototype and making it ready for operation in the wider world.
e.g. automation, system integration, robustification.

Software engineering - ensure maintainability, scaling, collaborative development.

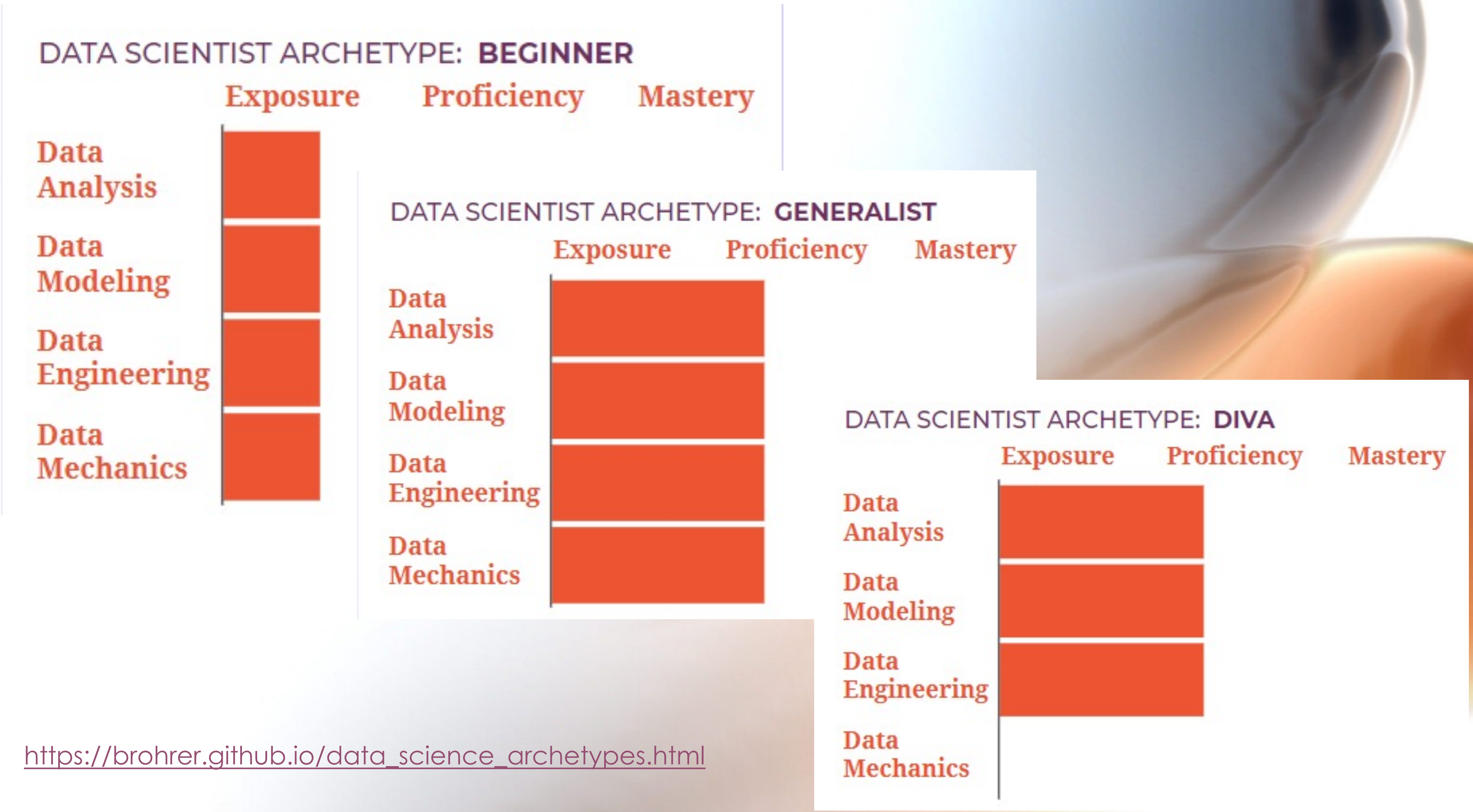
Data Mechanics

This is the dirty work that everyone needs to do but nobody likes to talk about.

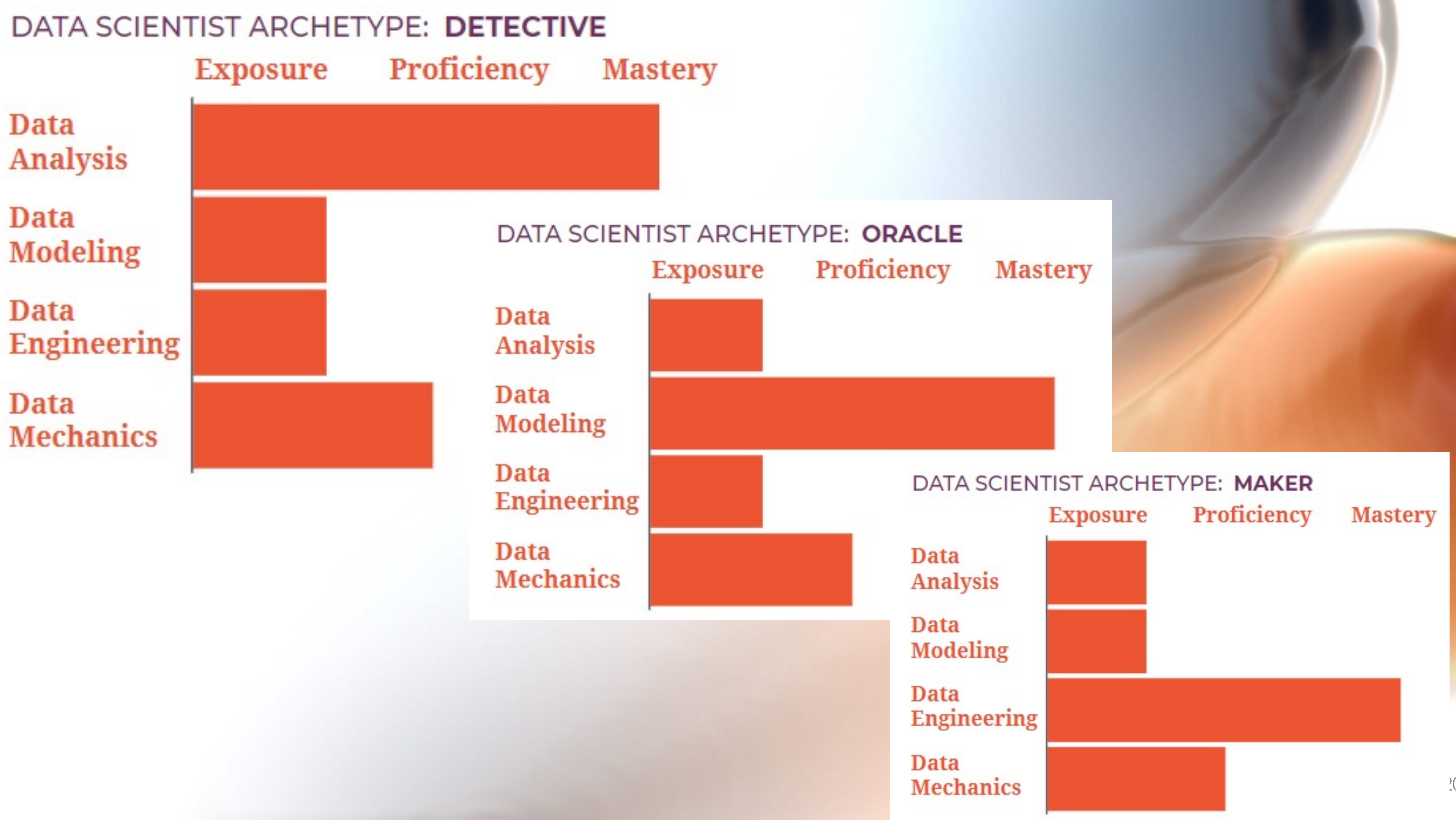
Data formatting - type conversion, string manipulation, fixing errors

Value interpretation - handling dates and times in an interpretable and consistent way, responding correctly to missing values, making sure units of measurement are consistent and documented

Data handling - querying, slicing, joining.
This lets you pull the subset of data that you want for an operation from the great lake of what you have available.

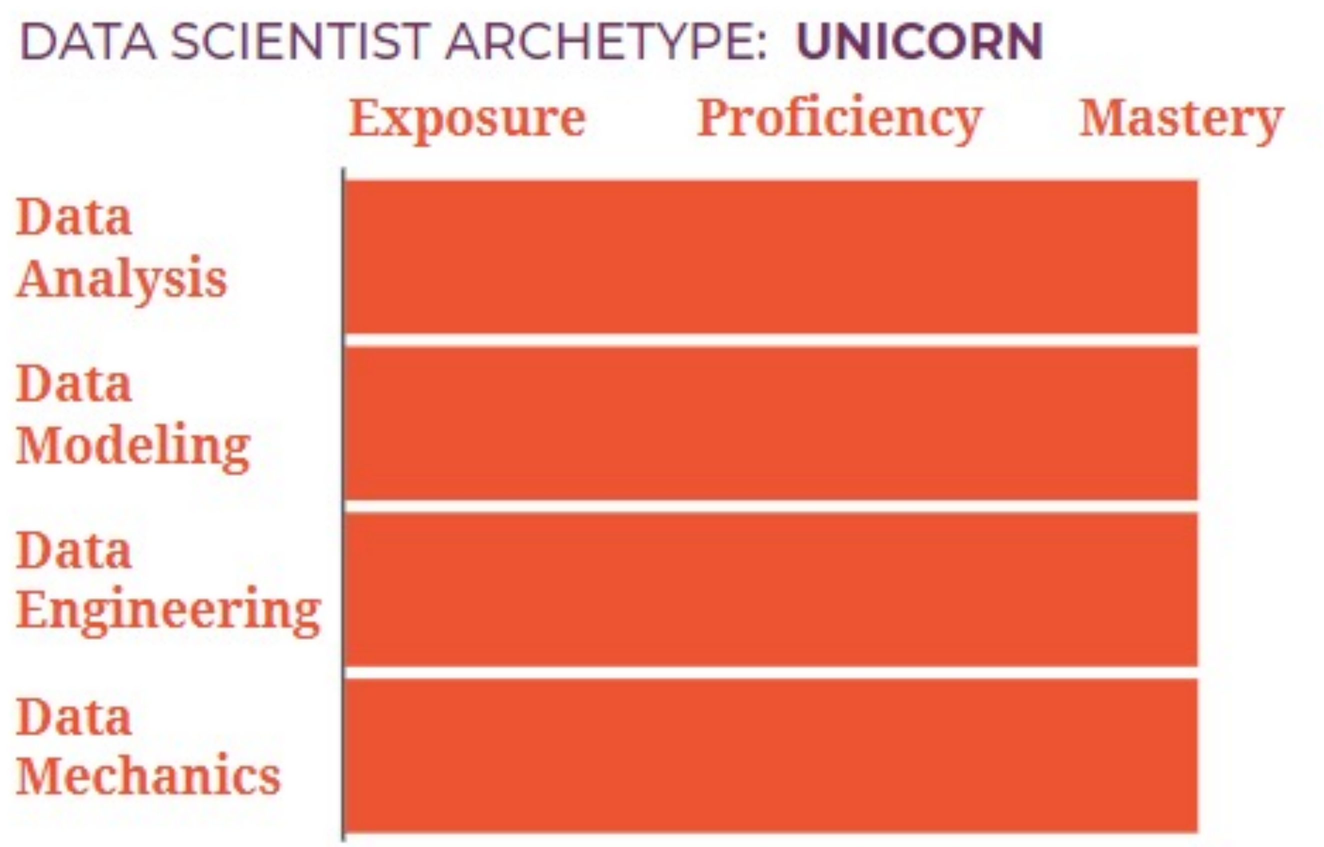


https://brohrer.github.io/data_science_archetypes.html

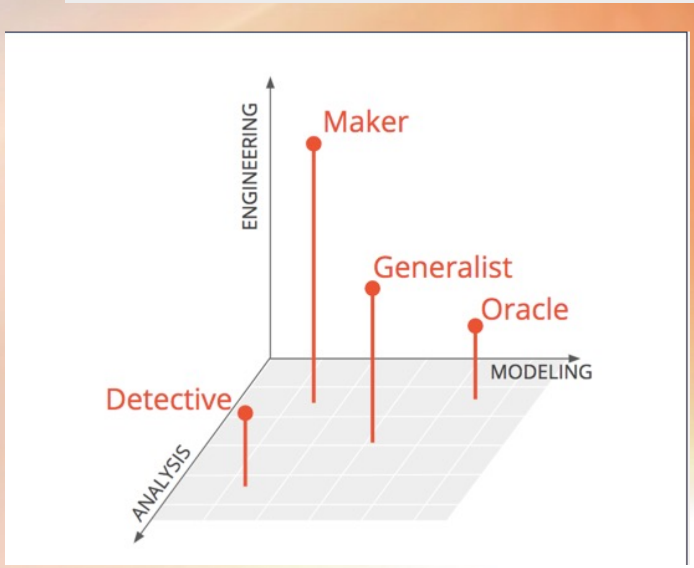

DATA SCIENTIST ARCHETYPE: **DETECTIVE**

DATA SCIENTIST ARCHETYPE: **ORACLE**

DATA SCIENTIST ARCHETYPE: **MAKER**



“Unicorn Data Scientists (upgraded from “sexy data scientists”) are hard to find …” - Gil Press, Forbes, 2015

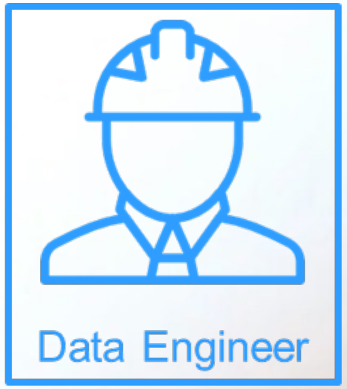


The **archetypes** that are most useful for **career planning**, then, are

- The Generalist, proficient at everything,
- The Detective, a master of analysis,
- The Oracle, a master of modelling, and
- The Maker, a master of engineering.

Solution to Replace Unicorn Data Scientist

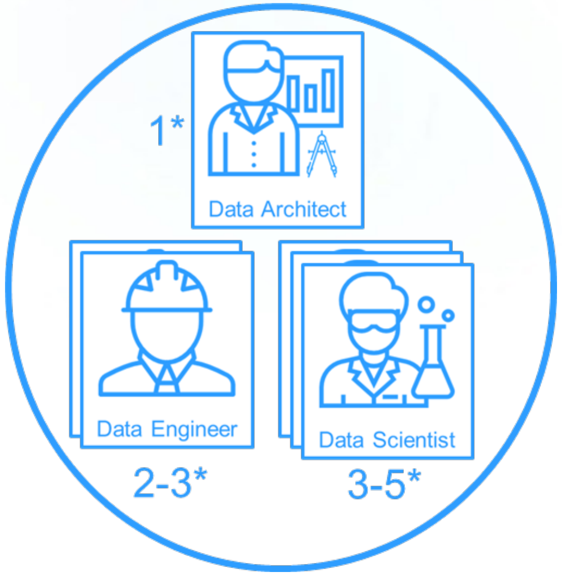
Data Science Team



Build out **teams** to reflect a variety of **backgrounds** and **experience.**

A **data science team** can have various roles like data engineer, data analyst, business intelligence, research scientist, decision scientist, and of-course data scientist.

Ideal Scenario



1 Data Architect/Data team manager
2-3 Data engineer to support the development and go-lives
3-5 Data Scientist for fast prototyping and complex models/analysis

Reality

1. Learn how to use the whole data toolbox
   He must be able to face the majority of IT challenges by yourself, from bash to Dockerfile.

2. Learn how to explain your work
   A good story with average results is better than a boring story with good results (stunning results always win).

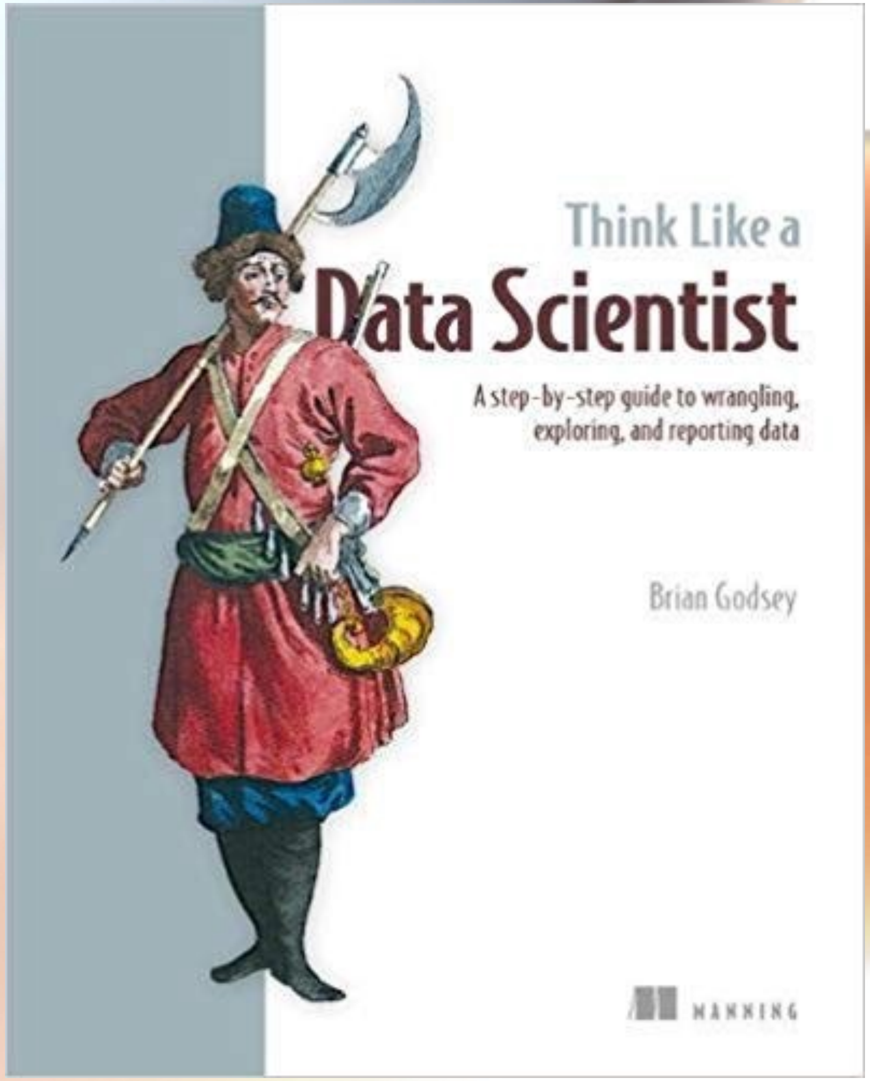
3. Learn how to write "good code"
   Working like the person you are going to show your code one day is a psychotic killer.

4. Learn how to understand business needs
   Well posed questions are very rare.

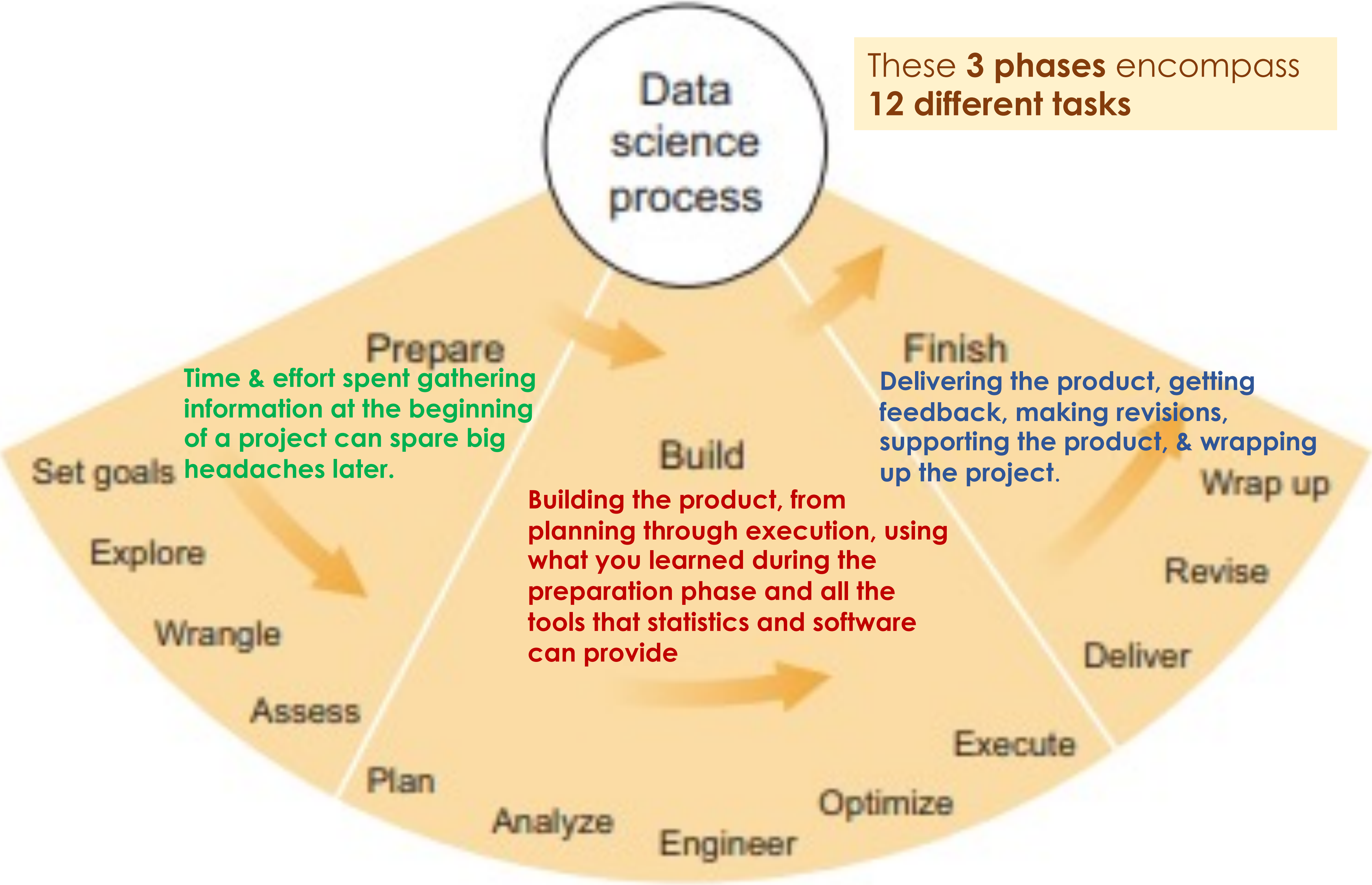
5. Learn how to make models work
   Data Superman needs to be comfortable with mathematics and statistics.

How to Think Like a Data Scientist in 12 Steps

https://medium.com/cracking-the-data-science-interview/how-to-think-like-a-data-scientist-in-12-steps-157ea8ad5da8


Think Like a
Data Scientist
A step-by-step guide to wrangling, exploring, and reporting data
Brian Godsey
MANNING

These **3 phases** encompass **12 different tasks**



Time & effort spent gathering information at the beginning of a project can spare big headaches later.

Building the product, from planning through execution, using what you learned during the preparation phase and all the tools that statistics and software can provide

Delivering the product, getting feedback, making revisions, supporting the product, & wrapping up the project.

Phase 1 - Preparation

The process of data science begins with **preparation**.

- Need to **establish**
  - what you know,
  - what you have,
  - what you can get,
  - where you are, and
  - where you would like to be.
- A project in data science needs to have a **purpose** and corresponding **goals**.
- Only when you have well-defined goals can you begin to survey the **available resources** and all the possibilities for moving toward those goals.

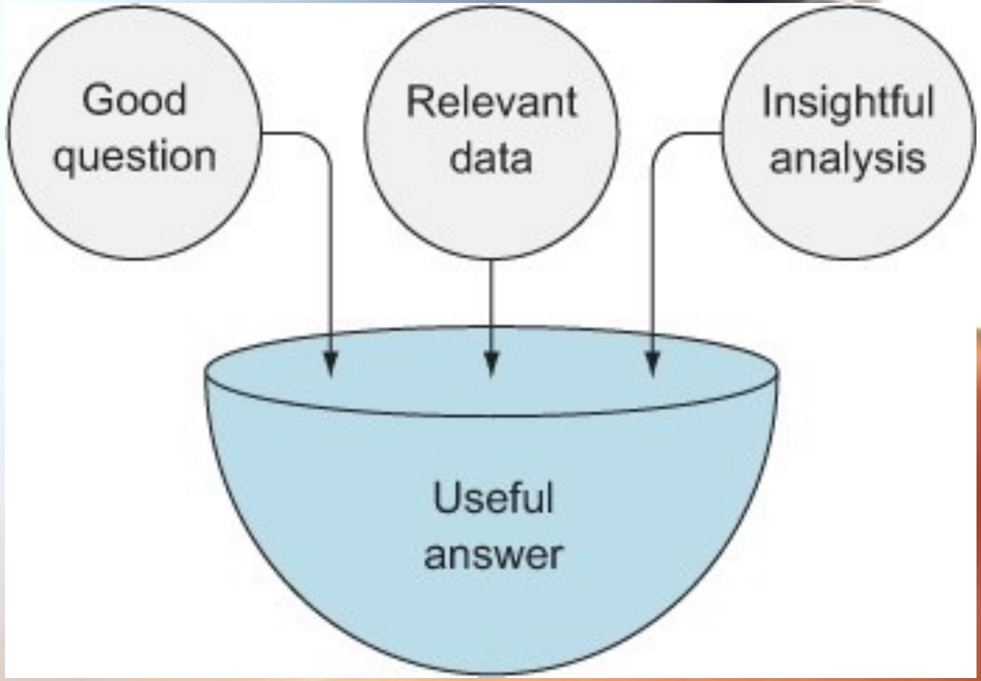
Step 1 – Setting Goals

- Every project in data science has a **customer** - with some **expectations**.
- Need to **ask good questions** about their data.
- Asking questions that lead to **informative answers** and subsequently **improved results** is an important and nuanced challenge.
- **Good questions** are concrete in their assumptions, and good answers are measurable success without too much cost.

The recipe for a useful answer in a data science project

Each **goal** should be put through a **pragmatic filter** based on data science.
This filter includes asking these questions:

(1) What is possible?
(2) What is valuable?
(3) What is efficient?



If you have a **good question** BUT **irrelevant data**, an *answer will be difficult to find*.

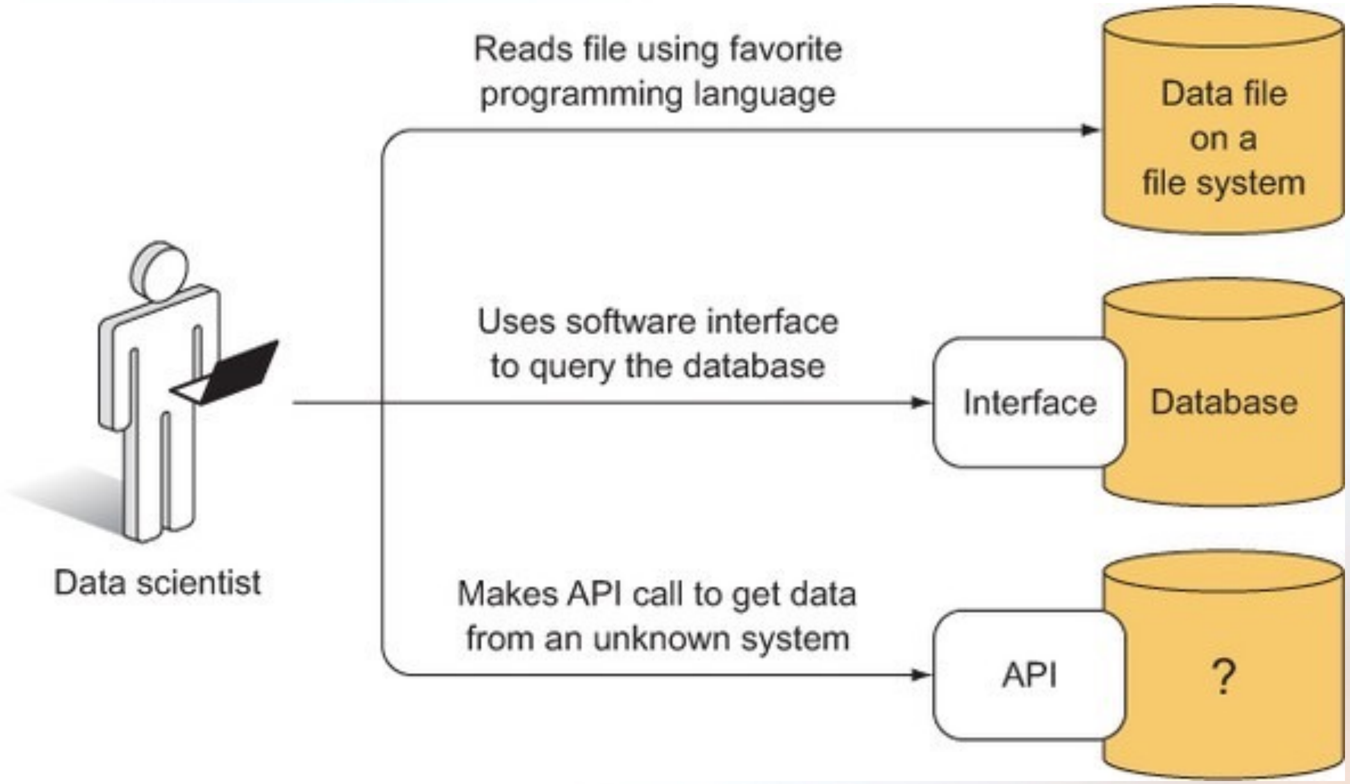
Applying this filter to all putative goals within the context of the good questions, possible answers, available data, and foreseen obstacles can help you arrive at a **solid set of project goals** that are, well, possible, valuable, and efficient to achieve.

To get real truth and useful answers from data

We must use the **scientific method**, or in our case, the **data scientific method**:

1. Ask a question.
2. State a hypothesis about the answer to the question.
3. Make a testable prediction that would provide evidence in favor of the hypothesis if correct.
4. Test the prediction via an experiment involving data.
5. Draw the appropriate conclusions through analyses of experimental results.

Step 2 – Exploring Data


Reads file using favorite programming language
Data file on a file system
Uses software interface to query the database
Interface
Database
Data scientist
Makes API call to get data from an unknown system
API
?

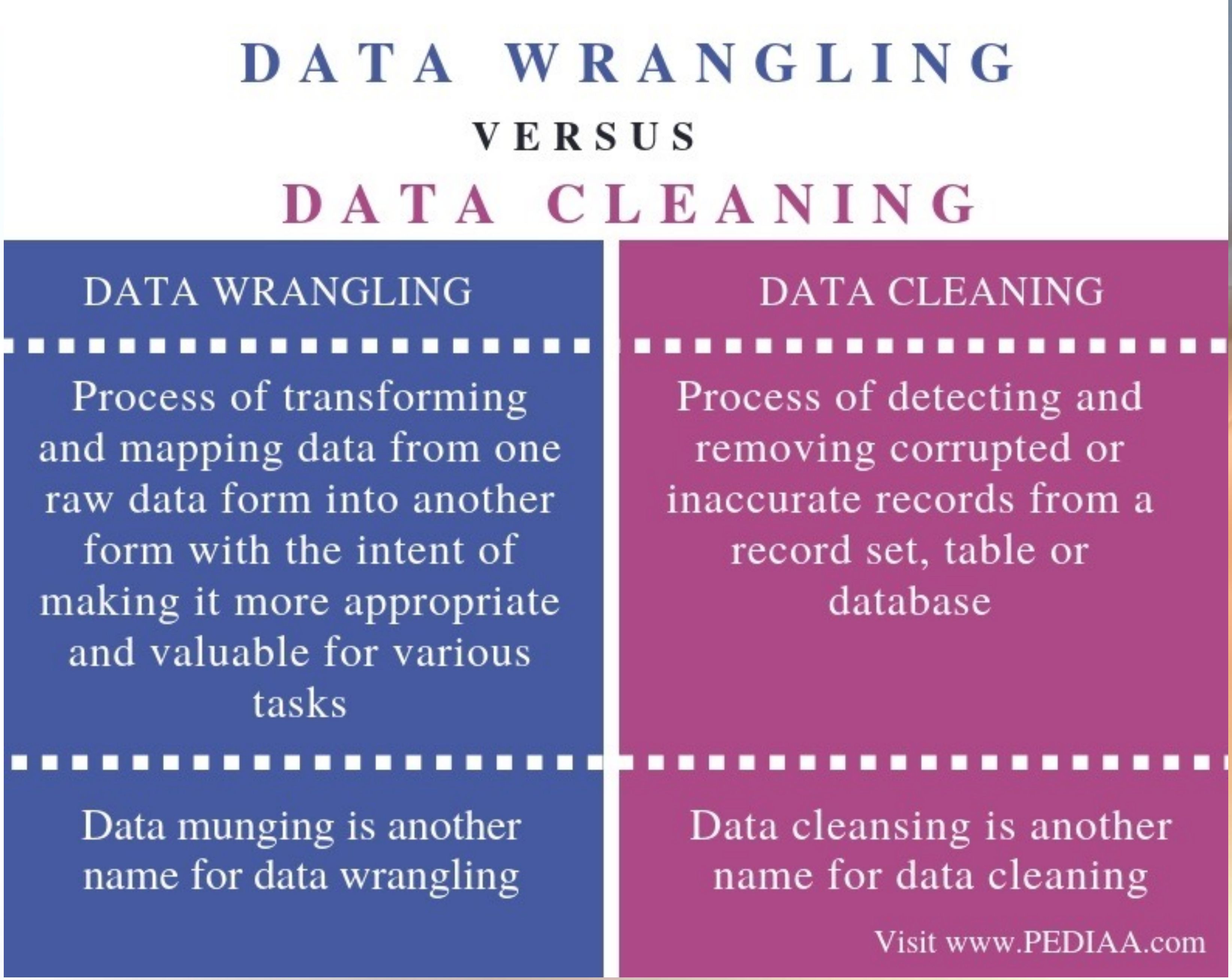
The data comes in a certain format, and you have to deal with it.

Common forms of data includes:
Flat Files (csv, tsv),
HTML, XML, JSON,
Relational Databases,
Non-Relational
Databases, APIs.

Two important things that a **web scraper** must do well are visit lots of URLs programmatically and capture the right information from the pages.

Step 3 – Wrangling Data

- A.K.A as **data munging** - the process of taking data and information in difficult, unstructured, or otherwise arbitrary formats and converting it into something that conventional software can use.
- Making messy data clean.
- Data wrangling is an uncertain thing that requires specific tools in specific circumstances to get the job done.
- Can try using file format converters or proprietary data wranglers and writing a script to wrangle data.



DATA WRANGLING

VERSUS

DATA CLEANING

DATA WRANGLING	DATA CLEANING
Process of transforming and mapping data from one raw data form into another form with the intent of making it more appropriate and valuable for various tasks	Process of detecting and removing corrupted or inaccurate records from a record set, table or database
Data munging is another name for data wrangling	Data cleansing is another name for data cleaning

Visit www.PEDIAA.com

Step 4 - Assessing Data

- To get to know the data better and to avoid problems with outliers, biases, precision, specificity, or any number of other inherent aspects of the data.
- **Descriptive statistics** is the discipline of quantitatively describing the main features of a collection of information, or the quantitative description itself.
- Think description, max, min, average values, summaries of the dataset.
- **Inferential statistics** is the practice of using the data you have to deduce — or infer — knowledge or quantities of which you don’t have direct measurements or data.
- With respect to a data set, you can say the following:
    - **Descriptive statistics** asks, “*What do I have?*”
    - **Inferential statistics** asks, “*What can I conclude?*”

Phase 2 - Building

- You learned something in Phase 1, and now you may already be able to answer some of the questions that you posed at the beginning of the project.

Step 5 – Developing Plans

- Plans and goals can change at any moment, given new information or new constraints or for any other reason.
- **Focus on what the customer cares about**: progress has been made, and the current expected, achievable goals are X, Y, and Z.
- Consider **communicating** your basic plan to the customer.

Step 6 – Analyzing Data

Involves statistical analysis of data.

- **Statistical modeling** is the general practice of describing a system using statistical constructs and then using that model to aid in analysis and interpretation of data related to the system.
- **Machine learning and black box methods** and other complex statistical methods can be good tools for accomplishing the otherwise impossible.

Step 7 – Engineering Product

A wide range of software is available for doing data science, from spreadsheets to mid-level statistical tools to statistical programming languages and libraries

Involves building statistical software.

- **Spreadsheets and GUI-based applications** - Common software tools here are Excel, SPSS, Stata, SAS, and Minitab.
- **MATLAB** is a proprietary software environment and programming language that's good at working with matrices.
- **R** or **Python** or **Java**.
  - iPython for Python and RStudio for R
- Frameworks for building web applications in data-friendly languages:
  - □ Flask for Python
  - □ Shiny for R
  - □ Node.js for JavaScript, plus D3.js for awesome data-driven graphics
- **Natural Language Toolkit (NLTK)** - most popular and most robust tool for natural language processing (NLP)

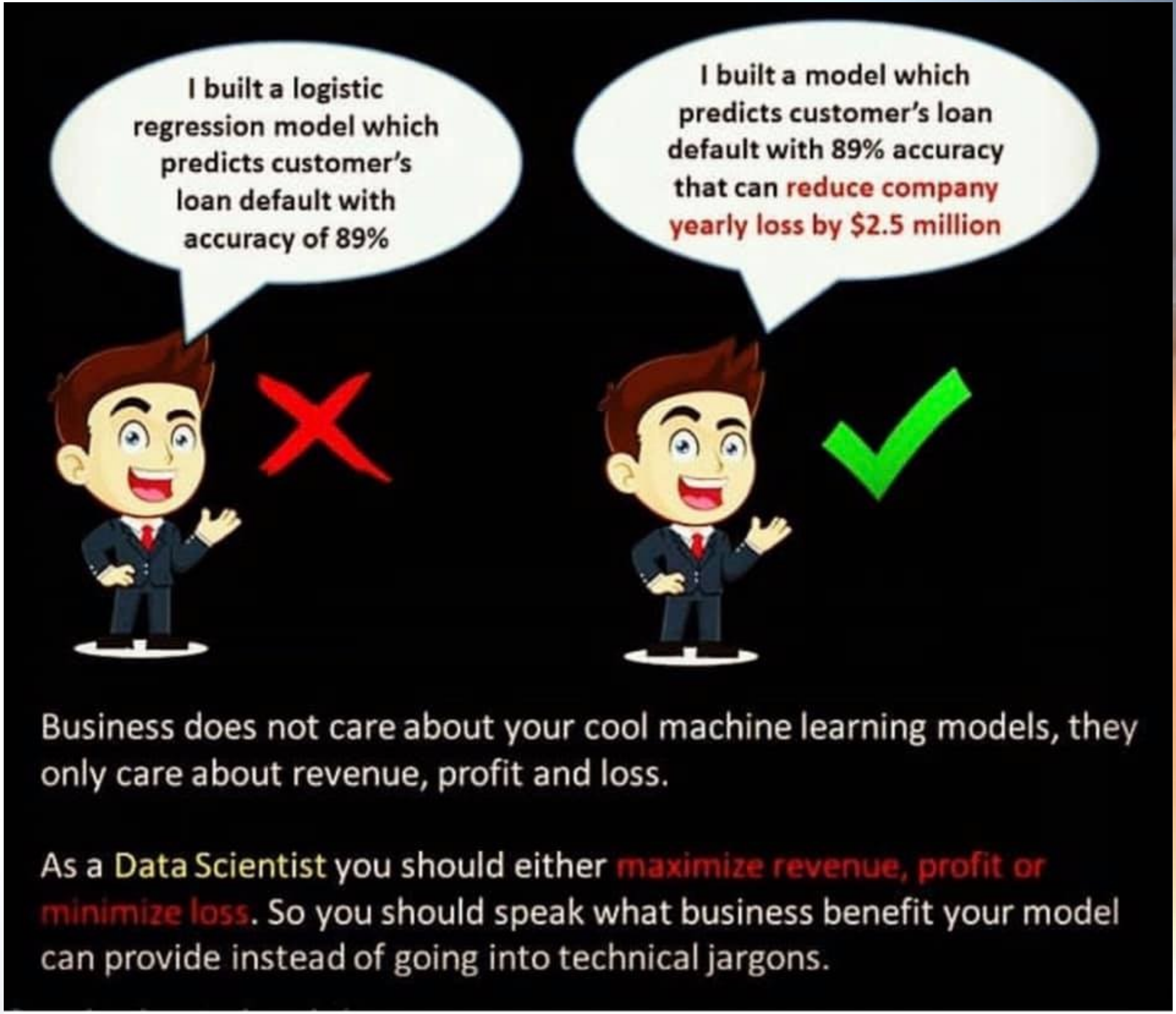
Step 8 – Optimizing Data

- Know some of the most popular and most beneficial software for making your life and work as a data scientist easier.
- **Databases** - The 2 most common types are relational (SQL) and document-oriented (NoSQL, ElasticSearch).
- **High-performance computing (HPC)** – for cases where there's a lot of computing to do and you want to do it as fast as possible.
  - ***Supercomputer*** (which is millions of times faster than a personal computer),
  - ***Computer clusters*** (a bunch of computers that are connected with each other, usually over a local network, and configured to work well with each other in performing computing tasks), or
  - ***Graphics Processing Units*** (which are great at performing highly parallelizable calculations).
- **Cloud services** – Google, Amazon, Microsoft
- **Big data technologies**: Hadoop, HBase, and Hive — among others.

Step 9 – Executing Plan

A data science project involving statistics, expectations are based either on a notion of **statistical significance** or on some other concept of the **practical usefulness** or **applicability of those results or both.**

- No matter how good a plan is, there’s always a chance that it should be revised as the project progresses.
- A project plan can unfold in a number of ways; maintaining an awareness of outcomes as they occur can mitigate risk and problems.
- If you’re a software engineer, be careful with statistics.
- If you’re a statistician, be careful with software.
- If you’re a member of a team, do your part to make a plan and track its progress.
- Modifying a plan in progress is an option when new, external information becomes available, but make modifications deliberately and with care.
- Good project results are good because they’re useful in some way, and **statistical significance** might be a part of that.


I built a logistic regression model which predicts customer's loan default with accuracy of 89%
I built a model which predicts customer's loan default with 89% accuracy that can reduce company yearly loss by $2.5 million

Business does not care about your cool machine learning models, they only care about revenue, profit and loss.

As a Data Scientist you should either maximize revenue, profit or minimize loss. So you should speak what business benefit your model can provide instead of going into technical jargons.

Phase 3 - Finishing

Step 10 - Delivering Product

In order to create an effective product that you can deliver to the customer:

1. DS must understand the customer perspective.
2. DS need to choose the best media for the project and for the customer.
3. DS must choose what information and results to include in the product and what to leave out.

The thing that DS create and deliver to customers—the **product**—can take many

forms.

Passive product - a **report** or **white paper**; the customer can find in this only the answers that are in the text, tables, and figures present in the document.

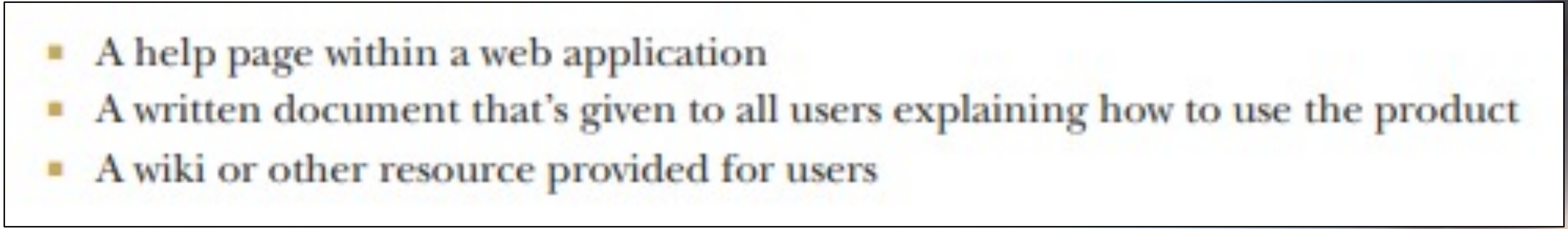
Active product - **application** that allows customers to interact with data and analysis in order to answer some questions on their own.

Step 11 – Making Revisions

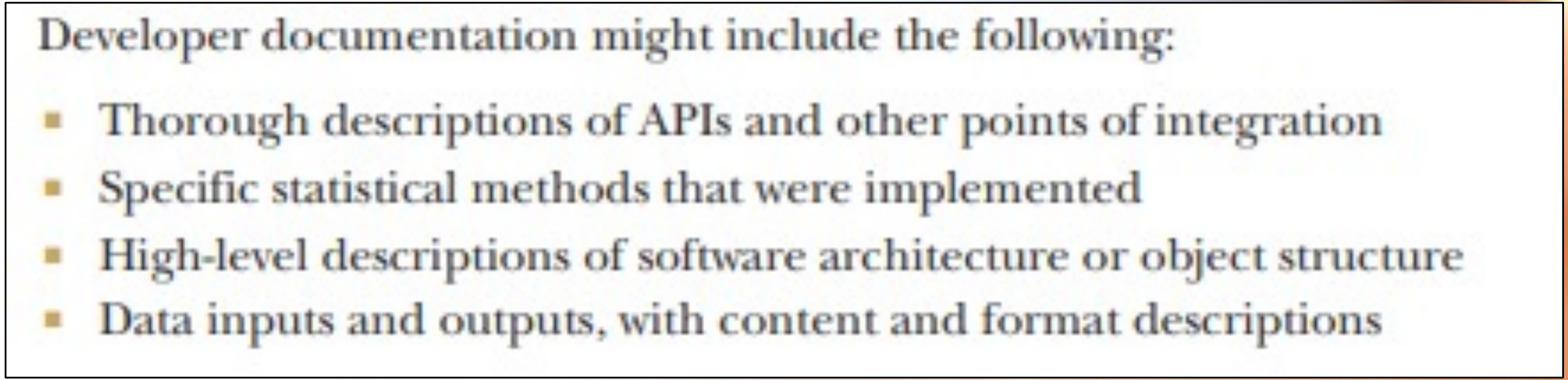
- The process of recognizing, diagnosing, and fixing problems in the product should be undertaken deliberately and carefully.
- Once the customer begins using the product, there's the potential for a whole new set of problems and issues to pop up.
- Getting feedback is helpful, but it shouldn't be taken at face value.
- Product revisions should be designed and engineered with the same level of care (or more) as when you designed and built the product itself.
- Not every problem needs fixing.

Step 12 – Wrapping up project

- Organizing project materials and **storing them** in a reliable place can spare you from headaches later if you have to revive the project for any reason.
- It's important to **document** the software and the methods so that you and your colleagues can understand every aspect of the project and work with it in the future.
- **Documentation** is an exercise in empathy; you have to imagine what you and others might not understand in the future and write explanations accordingly.
- Conducting a formal **project postmortem** can reveal many lessons that may not have been obvious otherwise.
- Every project offers many **lessons to be learned**, and many of them can be generalized to apply to almost any future data science project.
- Data science is mostly about recognizing when something unexpected might occur, and awareness of such uncertainties can make the difference between success and failure in future projects.

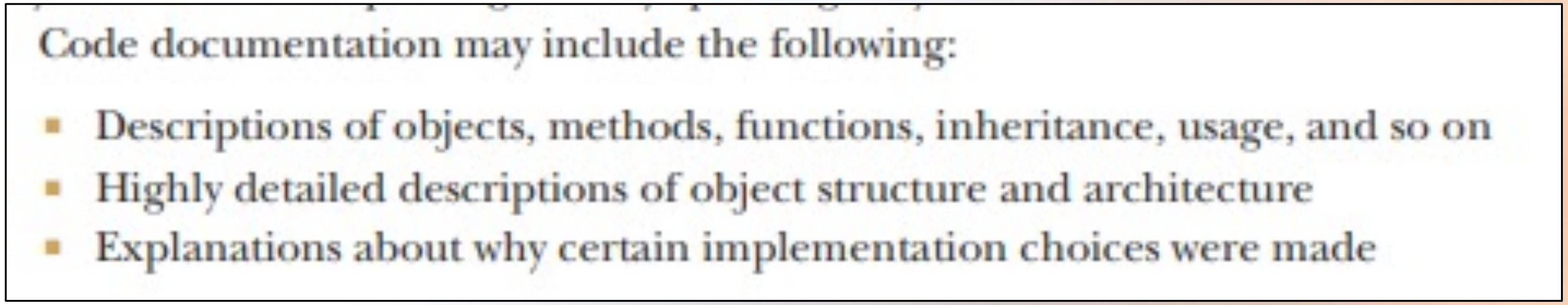


- A help page within a web application
- A written document that's given to all users explaining how to use the product
- A wiki or other resource provided for users



Developer documentation might include the following:

- Thorough descriptions of APIs and other points of integration
- Specific statistical methods that were implemented
- High-level descriptions of software architecture or object structure
- Data inputs and outputs, with content and format descriptions



Code documentation may include the following:

- Descriptions of objects, methods, functions, inheritance, usage, and so on
- Highly detailed descriptions of object structure and architecture
- Explanations about why certain implementation choices were made

What Skills Make Data Scientists Exceptional?

Beyond technical proficiency, there are several skills that a data scientist must have to excel.

1. Communication

- In their **writing**, **presentations** and **emails** strong data scientists are **clear**.
- They focus on the audience, considering what they already know, what they need to know and what they care about.
- They can explain their methods and results to a non-technical audience at the appropriate level of technical depth.
- Data scientists that are lacking in this area fail to convey their work or persuade teammates and leaders of its importance.

2. Breadth

- Strong data scientists are not afraid to move between roles, say, migrating between data analysis, data engineering, modeling, and back, over the course of a project.
- This breadth provides a huge benefit. For example, doing data analysis while keeping the limitations of modeling in mind produces results that are more accurate, more useful and more timely.
- Data scientists that are lacking in this area might say "I'm a modeler. Data cleaning is a job for someone else." Overspecialization leads to blind spots, such as neglecting code health or neglecting statistical rigor.

3. Readiness to learn new tools, skills and domains

- There's no practical way to learn all the tools you will need before you need them. The only way to be prepared for this is get **comfortable** with the **process of learning**.
- The set of tools a data scientist comes with doesn't matter as much as their ability to embrace new ones.
- Data scientists that are lacking in this area will be **limited in what they can contribute**. Most project work will be frustrating. (Their teammates will be frustrated too.) The solution is to adopt a willingness to feel dumb, also known as "a beginner's mindset". This helps navigate the uncomfortable start-up period when every step of working with a tool is unfamiliar. The beginner's mindset manifests itself as a curiosity about the field, the company, the products, and the customers.

Success Patterns

1. Learning through practice

- Have a broad understanding of all the roles a data scientist can play and have deep skills in at least one.
- These data scientists are the ones who have worked on realistic data science problems in several domains.
- Concrete examples with rich context and ambiguity are powerful teachers. Applying the same skill in several different domains bestows a facility on the learner that is hard to get any other way.

2. Mentoring / cross-mentoring / community contribution

- Share your work with others in the form of teaching activities with those less experienced, such providing advice, tutorials, or explanations.
- It can also manifest between peers in such varied ways as publishing project summaries, asking advice, cooperative coding, and creating cheat sheet references for a new tool. These can take place in person or on-line.
- Every major social network has its own data science community, each with its own flavor.

3 Education options - A Career as a Data Scientist

1. **Degrees and graduate certificates** provide structure, internships, networking and recognized academic qualifications for your résumé. They will also cost you significant time and money.
2. **MOOCs and self-guided learning courses** are free/cheap, short and targeted. They allow you to complete projects on your own time – but they require you to structure your own academic path.
3. **Bootcamps** are intense and faster to complete than traditional degrees. They may be taught by practicing data scientists, but they won’t give you degree initials after your name.

Training for Data Science

A few universities and other organizations have started to offer data science degrees, training, and certificates.

- University of Washington (Seattle, WA)
- Northwestern University (Evanston, IL)
- UC Berkeley (Berkeley, CA)
- CUNY (New York, NY)
- New York University (New York, NY)
- Columbia University (New York, NY)
- Stanford University (Palo Alto, CA)

Training for Data Science

Corporate and Association Training Programs

A few other types of organizations — both private companies and professional organizations — offer certifications or training.

- INFORMS (Operations Research Society): Analytics certificate
- Digital Analytics Association: certificate
- TDWI (The Data Warehousing Institute): Courses; focus is on database architecture
- American Statistical Association: Chartered statistician certificate
- Data Science Central: Data science apprenticeship
- International Institute for Analytics: More like a think tank. Founded by the famous Tom Davenport (visiting Harvard Professor and one of the fathers of data science).
- Statistics.com: Statistics courses

Training for Data Science

Free Training Programs

- Coursera.com
    - Machine Learning (Stanford University)
    - Web Intelligence and Big Data (Indian Institute of Technology)
    - Introduction to Data Science (University of Washington)
    - Maps and Geospatial Revolution (Penn State)
    - Introduction to Databases (Stanford University, self-study)
    - Computing for Data Analysis (Johns Hopkins University)
    - Statistics One (Princeton University)
- Data Science Central
    - Training basics
    - Tutorials
    - Data sets
    - Real-life projects
    - Sample source code

Real World Experience

To develop a "real world experience" in data science is by working through lots of small projects.

Follow the process of

a) finding a data set you are interested,

b) asking and trying to answer questions about it,

c) write up the results, and

d) rinse and repeat to develop "real world experience".

https://www.datascienceweekly.org/articles/the-simple-process-to-get-real-world-data-science-experience

Data Science Portfolio

4 components:

i. Candidate’s particulars, like in CV.

ii. Candidate’s DS knowledge & skills & attitude

iii. Candidate’s real world experience doing DS work

iv. Candidate’s contribution on DS, like participating in competition, bootcamp, talk, etc.

Ideas on Data Science Project

- http://www.data.gov/applications.
- http://www.datakind.org/projects
- https://theartandscienceofdata.wordpress.com/2017/06/01/the-billionaire-clusters/
- http://demos.datasciencedojo.com/

DATA
WITH INSIGHT
LEADS YOU TO
DECISIONS
WITH CLARITY.

THANK YOU
FOR
YOUR
ATTENTION!
ANY QUESTIONS?