Making Impacts with Data Science

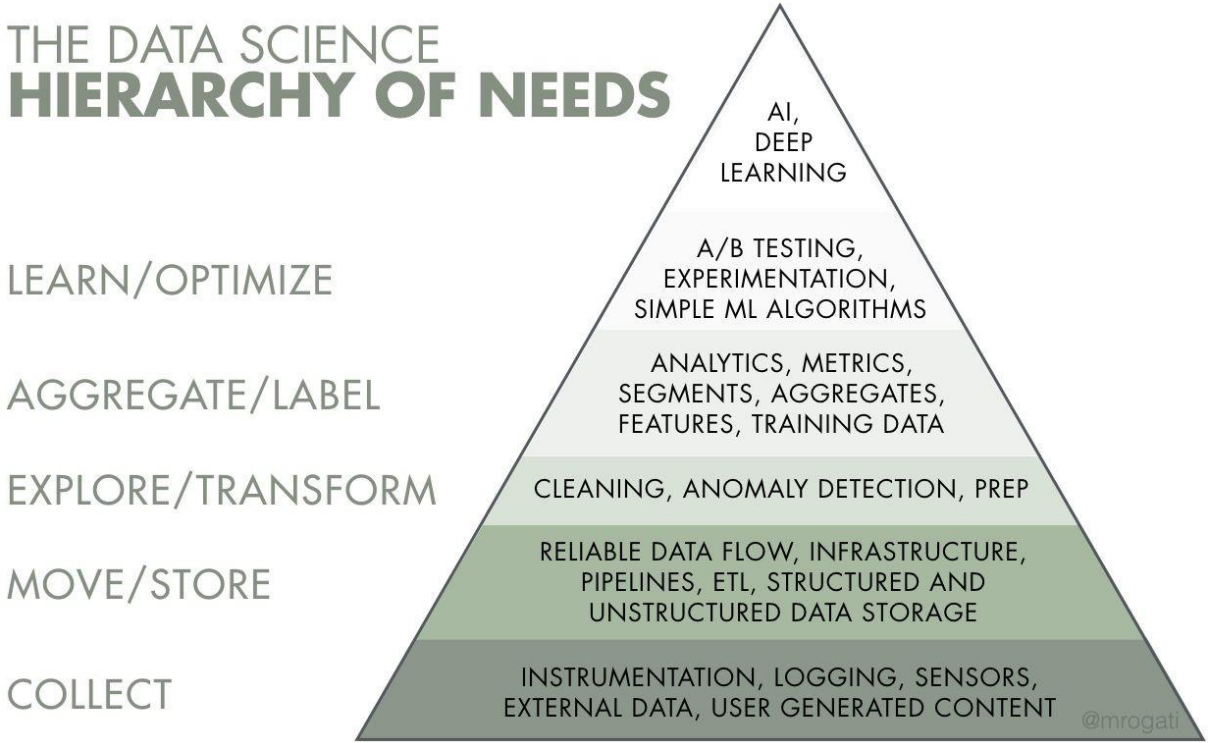
Hari, Tanggal

Data Analytics

Program Zenius Studi Independen Bersertifikat Bersama Kampus Merdeka

1. Data Science Hierarchy of Needs
2. Top Down vs Bottom Up Approach in Data Science
3. Data Products
4. Overcoming Challenges
5. Best Practices: Documentation and Portfolio

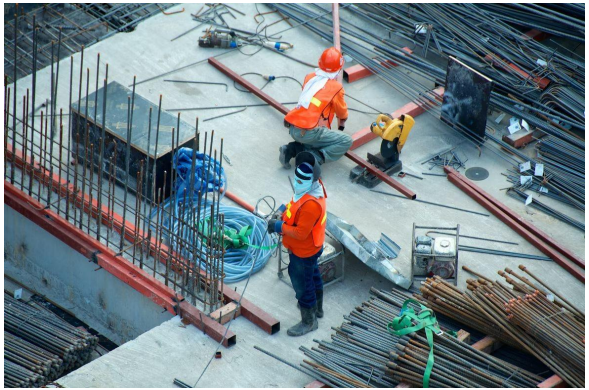
Data Science Hierarchy of Needs



Why do Data Science Job-Desc feels Different in Different Companies?

1. Different companies are at different ‘data ecosystem’ maturity level
2. Mature companies already have…
   a. Data Ingestion
   b. Database
   c. Dashboarding Services
   d. Place to Deploy ML Models
   e. An integrated Cloud Infrastructure

   ..so they have the valid ‘foundations’ to experiment and deploy Machine Learning models.



Top Down vs Bottom Up Approach in Data Science

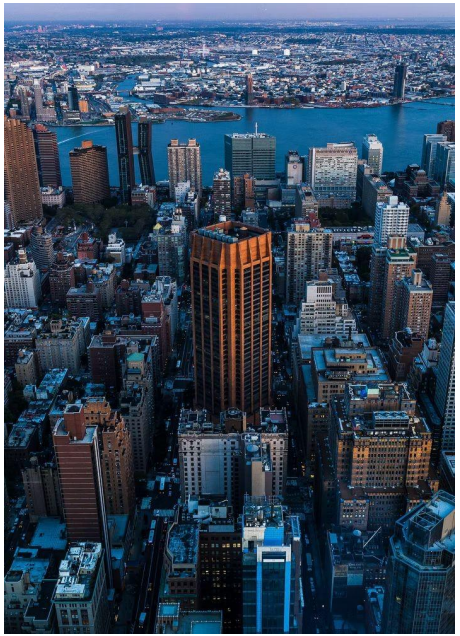
Top Down Approach

People from security / payment team tells us that frauds are harming the company.

Then, to test this hypothesis, we gather data from customers, and analyse the frauds and the non-frauds. We inspect the impact of such cases make to our company's fiscal status.

After validating our hypothesis, we will:

- Estimate the costs of frauds
- Build a Machine Learning model (supervised learning) to classify whether a transaction is fraudulent or not



Bottom Up Approach

We have a collection of transactions data. We initially don't know what to make use of it.

After doing exploratory data analysis, we notice a suspicious pattern.

We bring the data to the security / payment team, and they think it's fraudulent.

We then gather more data, add features, and try to come up with a Machine Learning model to detect frauds.

Top Down vs Bottom Up

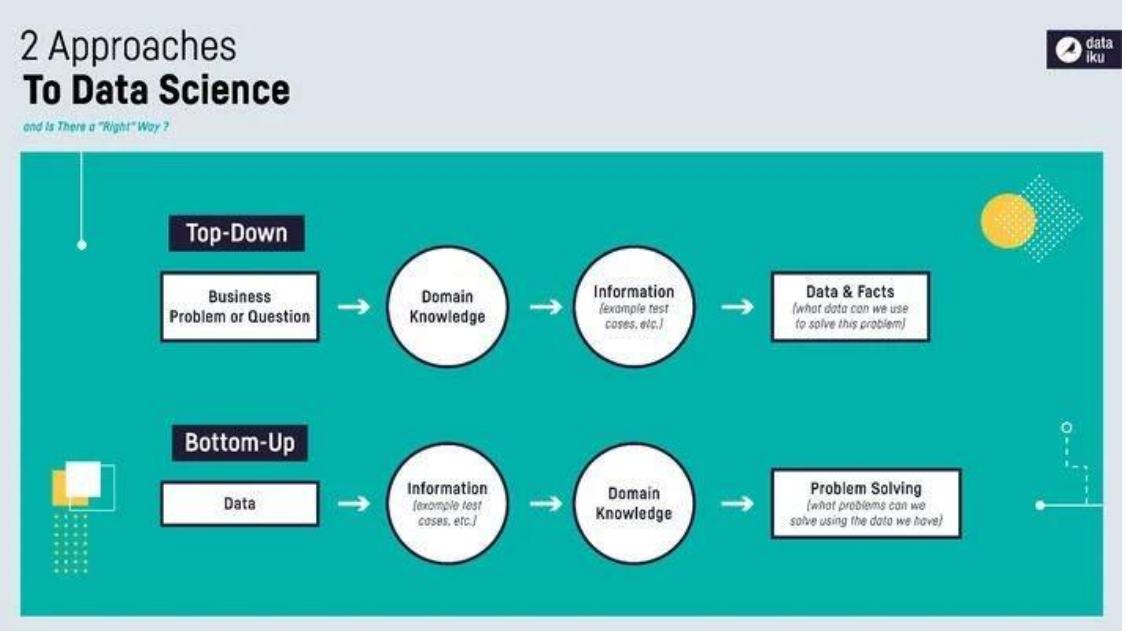


Image Source: © 2013 - 2020 Dataiku. All rights reserved.

Which is Better?

It shouldn’t be a comparison where one approach ‘wins’ or the other ‘loses’.

Both are common approaches and only differ in where the initiative begin - from explorations of the data, or from an expert / leader’s intentions.

Which is Better?

Top Down Approach can lead to Bottom Up Initiatives:

- After the Machine Learning model to detect frauds are completed, data scientists can then explore the data even more to see if there are still potentially 'hidden' frauds that are 'undetected' or 'missed' by the system.

Which is Better?

Bottom Up Approach can lead to Top Down Initiatives:

- After an exploration of data tells us that frauds are happening, we can bring in experts or more seasoned leaders to make a thorough top-down analysis on the issue, and potentially catch a better ‘high-level bigger picture’ of the problem

Data Products

What are Data Products?

In Machine Learning Classes and projects, usually it ends with measuring the accuracy of our model, and saving the report.

In real life, your machine learning model **has to be used** and **applied**.
These are **data products** that you create for your organization.

It's beyond codes and/or jupyter notebooks.

It's an **end to end solution** built on the concepts of **data science.**

Example of Data Products

- Company **performance dashboard** to track **KPI**
- Netflix's **home page** and **recommender** system
- An internal system to automatically: **scan new transaction** for **frauds**, and **automatically flag** them as fraudulent, and **report them**
- "There is a **faster** / alternative **route**!" in Google **Maps**
- Google Mail **autocomplete** / **suggestions**

NOT Example of Data Products

- desmos.com
- Online calculator
- Fonts Selection in Docs / Word
- Alarm Clock / Stopwatch
- Weather Widget

Why? Because they are purely ‘functional’ and does not use ‘data science’ principles.

However, a weather **forecast** - now that’s a data product.

Machine Learning Model as Data Products

It needs to be:

- Deployed
- Accurate
- Scalable
- ‘Easy’ to maintain
- Tracked (to see if the performance deteriorates)

Machine Learning Model as Data Products

When should we retrain the model? That's the 'million-dollar-question'.

1. Training too often: high cost, and high model variability
2. Not training the model: accuracy may worsen over time

Why Machine Learning Models (can) Deteriorate?

Concept Drift:

- When relationship of **dependent** and **independent** variable changes
- Example: a model was trained on a dataset to predict sales numbers. But then…COVID hit. Back then, a more expensive item is sold for less quantity, but now, even with a high price, some items are wanted extremely by the public (mask, oxygen tanks, vitamins)
- The data **stays the same**, but the **relationship between features** changes.

Why Machine Learning Models (can) Deteriorate?

Data Drift:

- When the model now sees input data which it previously did not see during training
- Example: a model is used to predict the monthly electricity consumption of low-income families. Now, the model is planned to be use to predict electricity consumptions of high-income families.
- In this instance, the **data changes**.

Why Machine Learning Models (can) Deteriorate?

How to fix it?

This topic in itself, is another field of study in the world of Machine Learning. Knowing when to re-train your model, and how to re-train it.

Should we train every X months?

Or should we just wait until the accuracy drops?

There is no easy answer - each project, each product, each use case, needs its own solution.

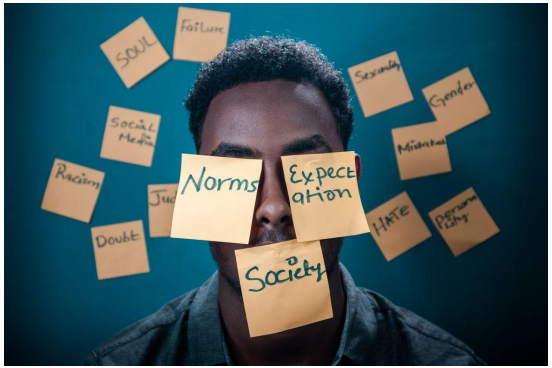
Overcoming Challenges

Unclear Expectations

Data Scientists, mainly working in companies which are **not yet mature in infrastructure**, often faces this.

What **can you do** if you face similar situation(s)?

- Be **proactive**! Ask for data, get a **business understanding** of the usage of that data, and **explore it**! Be creative, and **brainstorm** a **solution** that you can think of.
- Try to **write objectives** as **clear** as possible, and **break it down** to smaller points. **Refer** to that document as your **expectations**.



Project Management

How to juggle your tasks throughout the week?

- Have a **helper system** to **keep track** of your **to-do list**! **Kanban** board, **Jira**, Confluence, or a simple **Google Sheet** should be enough.
- Allocate time to respond to **ad-hoc request**. However, outside that time, focus on your main task!
- Allocate some **focus time** so you can enter 'Zen' mode and stay focused for a prolonged period of time
- Allocate time to take **breaks**!
- **Communication**.

Best Practices: Documentation and Portfolio

Why should we write documentations?

- Document on what we've **developed**
- **Keep track** of progress made
- **Remind** us of past steps
- **Legacy** to be passed to juniors / other teammates (when you quit)
- Other people can **understand** it too by reading it (and **not asking you directly** - which takes your time away)
- Foundation for **future developments**

Writing Readable Documentation

Example:

- https://www.nuclino.com/solutions/project-documentation
- https://github.com/chroline/well_app readme
- https://github.com/anfederico/Clairvoyant readme
- https://huggingface.co/docs

Writing Readable Documentation

Guidelines:

- Start with Why do you create this project
- Start with Who are the stakeholders of this project
- Then, you can state What your project is all about
- Know your audience! Use technical terms if the documentation is to be read by fellow Data Scientists, but don’t do it if it’s for the higher executives
- Give examples! Examples and illustrations help a lot in understanding complex processes
- Use flowchart to show the workflow

Example of Documentation Templates

- https://github.com/dec0dOS/amazing-github-template readme #
- https://www.datascience-pm.com/documentation-best-practices/

Now, let’s talk about Portfolio

What are portfolios?

A curated list of your own personal projects that serve as your personal branding.

Sometimes, we stop at ‘saving the model with 90% accuracy’ and call it a day.

However, in today’s hyper connected and competitive job market, there are things you can do to elevate your personal branding in Data Science.

Now, let’s talk about Portfolio

Example of good Data Science Portfolios:

- https://tdhopper.com/
- https://jameskle.com/data-portfolio
- https://gerinberg.com/

Now, let's talk about Portfolio

Tips on Portfolio creation:

- You don't have to start big. If you can't design a website, that's okay! Use github.io, create a homepage for your github account
- Use minimalistic, clean design, which are eye-friendly. Don't make it too crowded.
- Add your experiences, accomplishments, and how to contact you
- Be a good story teller!

Thank you!

Any Questions?