PyLadies Amsterdam

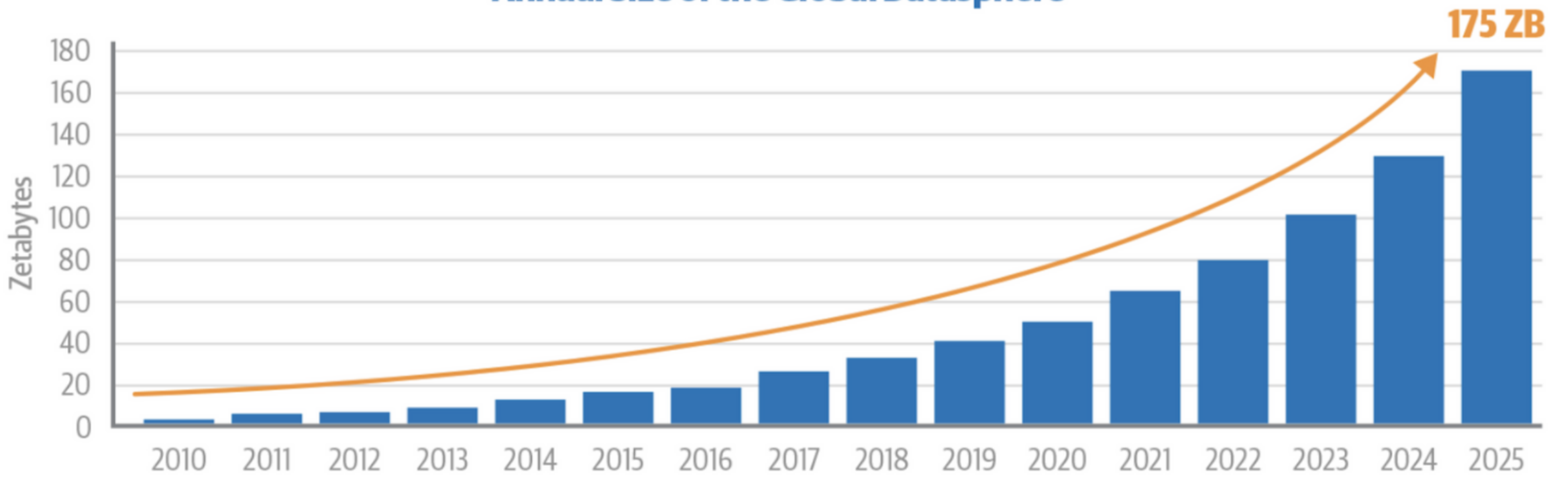
Let's talk about Data Engineering

Carolina Londoño
Data engineer @ felyx

Why Data Engineering?

- Increase of users on the internet
- Increase of time spent online
- Increase of services online
- Exponential data growth

Annual Size of the Global Datasphere



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Why Data engineering?

Definition of a new role

Data scientist + BI analyst + Data Analyst +
Software Developers

Many Data scientist are now data engineers

What does a DE do?

- Create and maintain the infrastructure for the data
- Model the data
- Keep the data up to date
- Privacy and security
- Support the data team

What does a DE do?

- Create and maintain the infrastructure for the data
  - On Premise
  - Cloud
  - Hybrid

Create and maintain the infra

On Premise

- Physical administration for network and config
- Licenses
- Common mostly for financial or healthcare institutions

Create and maintain the infra

Cloud

- Delegation of physical admins
- Interact with it remotely with APIs, or the web console.
- Provides services for common use cases

Create and maintain the infra

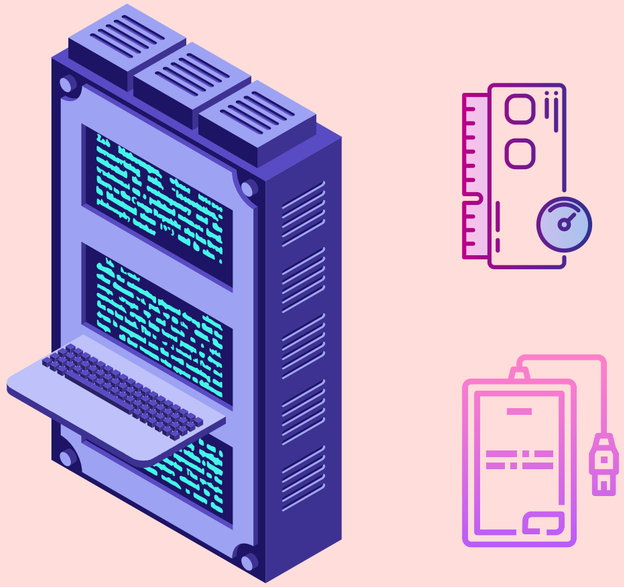
Hybrid

- Mix of both

Scaling

Scale Vertically

Or Scale up



Scale Horizontally

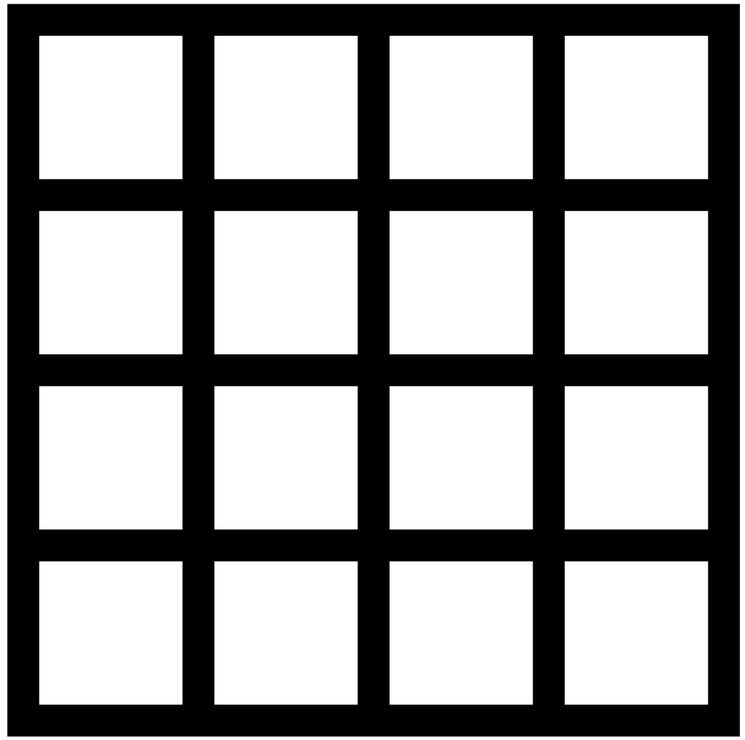
Or Scale out

What does a DE do?

- Model the data
  - How is the data going to be stored
  - How is the data going to be accessed

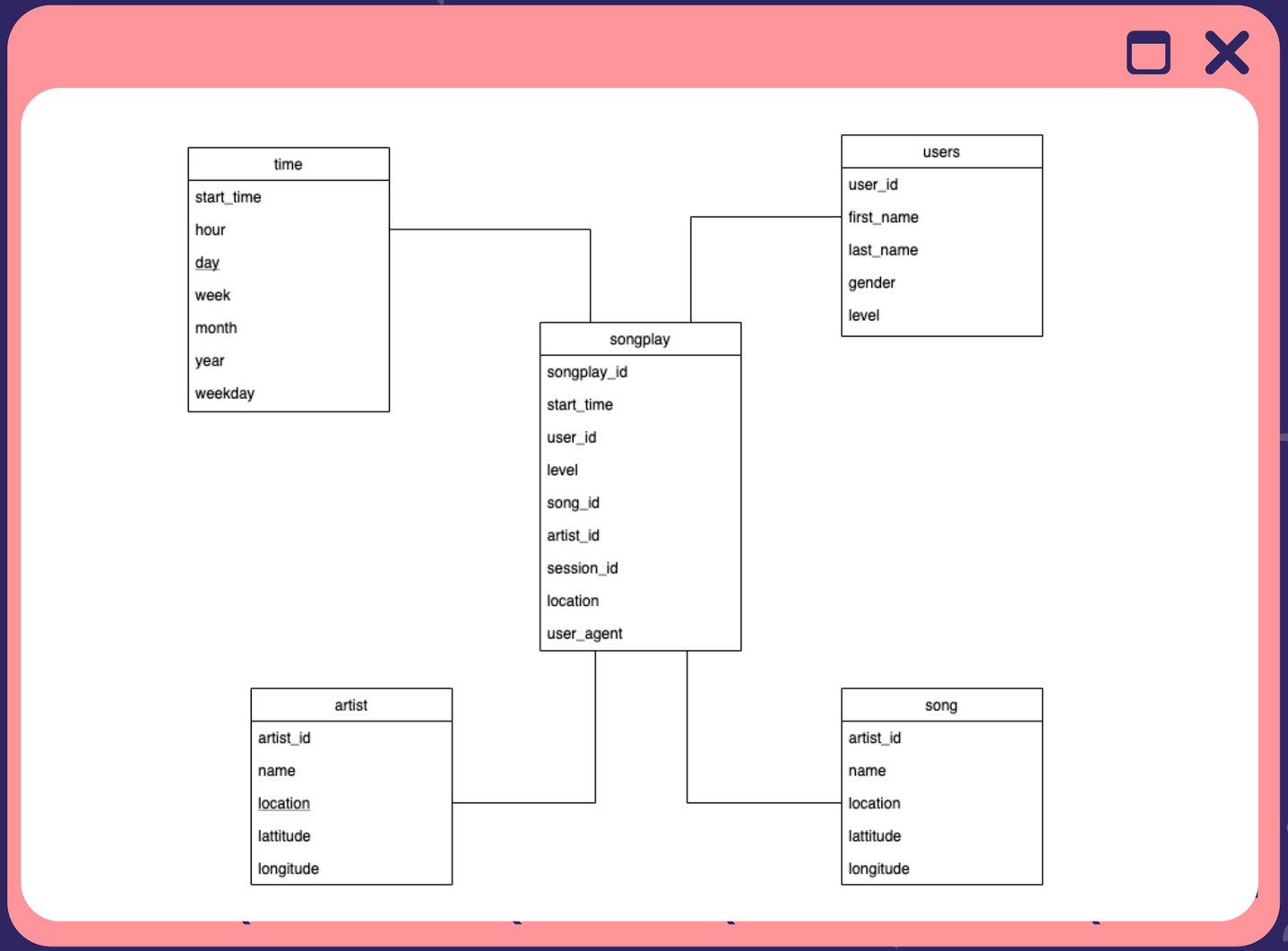
Data modeling

Relational



Data modeling

Relational



Data modeling

Relational databases

OLTP

- Online transactional database
- Usually on the operations side, apps, websites

OLAP

- Online Analytical processing DBs
- Usual for analysing/processing data of 1 month, 1 year, etc.
- Used on the business side
- Usually Data Warehouses

Data modeling

non-relational
JSON example

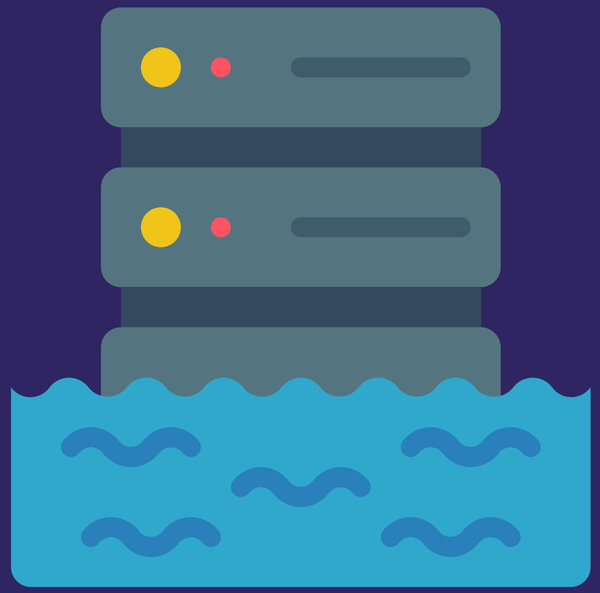
```
{
    "person" : {
        "firstName": "John",
        "lastName": "Smith",

        "address": {
            "streetAddress": "21 2nd Street",
            "city": "New York",
            "state": "NY",
            "postalCode": 10021
        },

        "phoneNumbers": [
            "212 555-1234",
            "646 555-4567"
        ]
    }
}
```

Data modeling

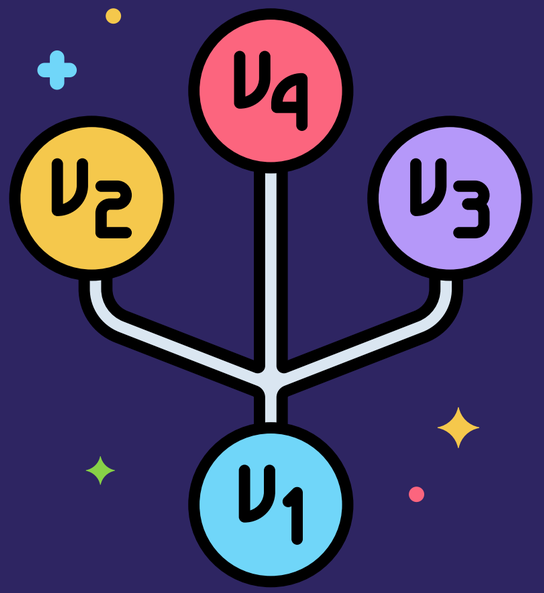
Non-relational



Datalakes

- Hold multiple types of data
  - JSON, especially deeply nested
  - Columnar
  - Documents

Data modeling



Data model evolution

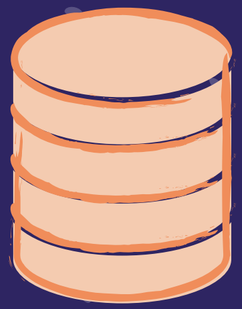
- What happens when a column is added, removed, or changed its type
- We should be able to time travel to see the data at x point in time -> define policies
- Version the database
- GDPR compliant

What does a DE do?

- Keep the data up to date
  - Batch
  - Streaming

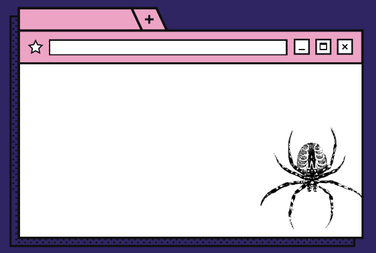
Keeping the data up to date

BATCH



Extract data from other DBs

API responses
Usually from 3rd parties



Web Scrapping

Files

Keeping the data up to date

STREAMING

Sensors

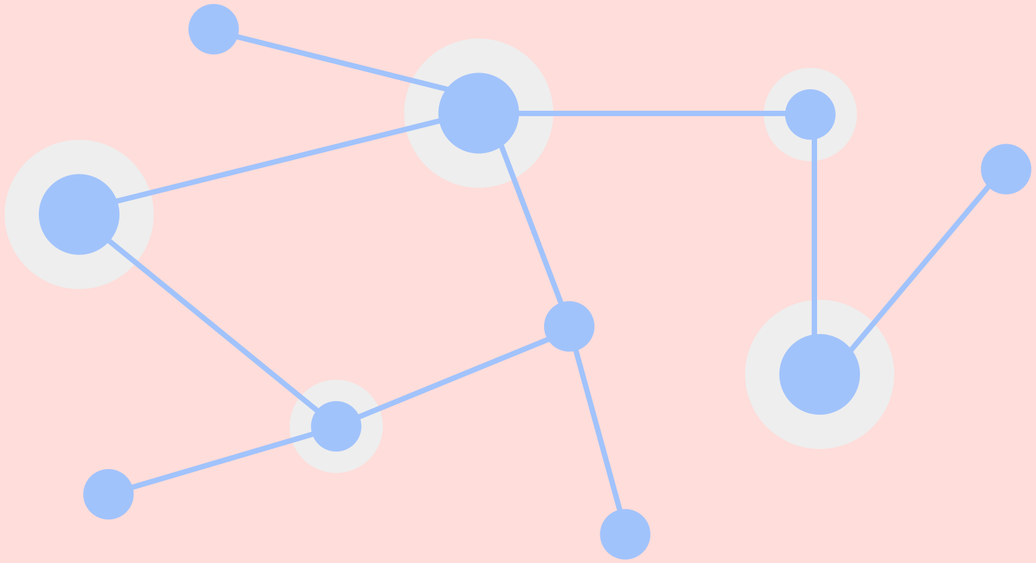
API stream response

What does a DE do?

- Privacy and security

Maintain the infra

Networking



- Is it going to be publicly accessible on the internet?
- Is every user who has access going to be able to read all the data, or just a portion?

Support the team

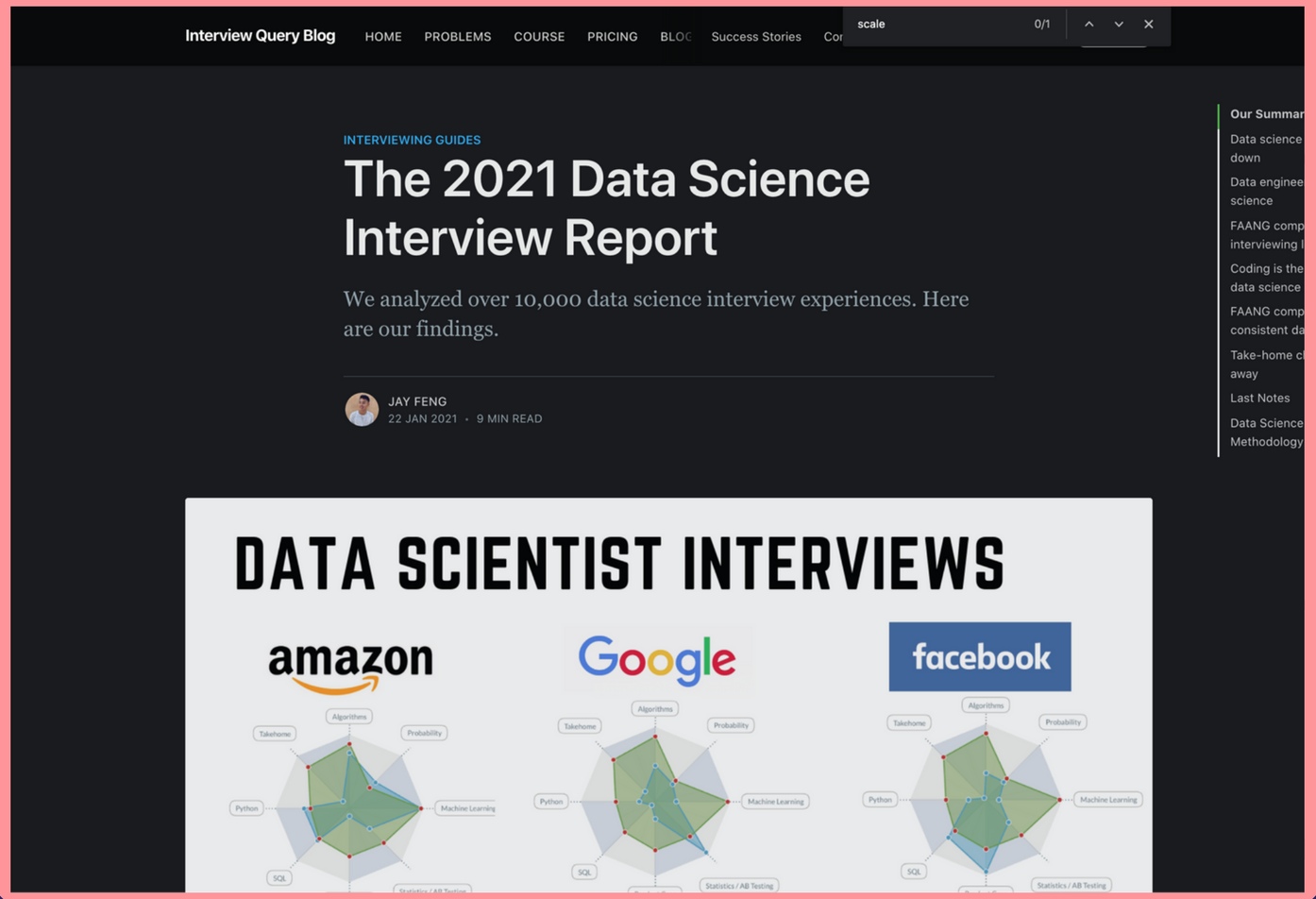
- How can I add new data?
- The data source is wrong
- Investigate new tools to improve the stack

Data engineering - Job market

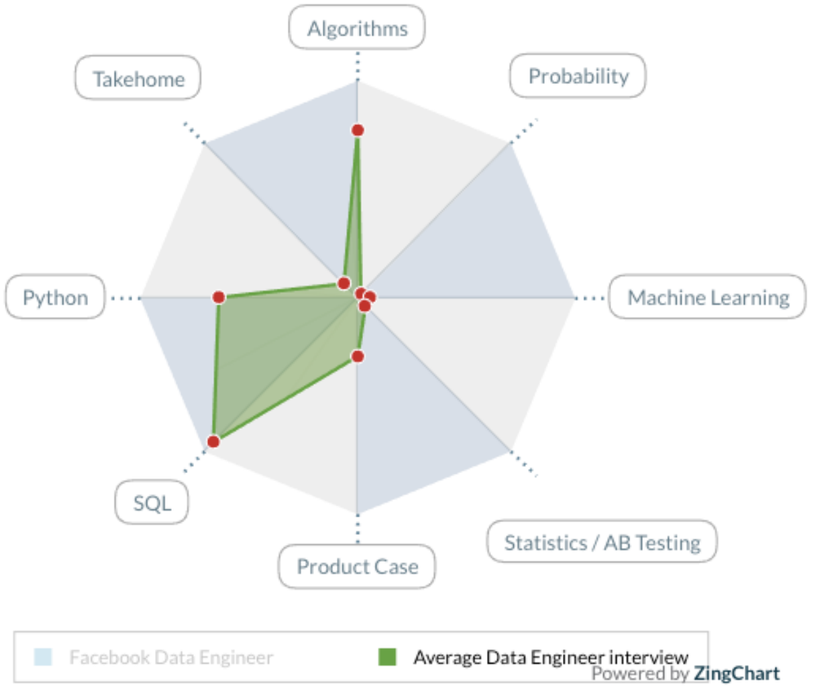
The top-15 emerging jobs in the Netherlands.

#1	Data Protection Officer	**#6**	Data Scientist	**#11**	Cyber Security Specialist
#2	Growth Hacker	**#7**	Data Engineer	**#12**	Salesforce Consultant
#3	Privacy Officer	**#8**	Customer Success Specialist	**#13**	Key Account Management Specialist
#4	Robotics Engineer	**#9**	Human Resources Administrative Officer	**#14**	Analytics Consultant
#5	Artificial Intelligence Specialist	**#10**	Cloud Engineer	**#15**	Full Stack Engineer

Data engineering - Job market



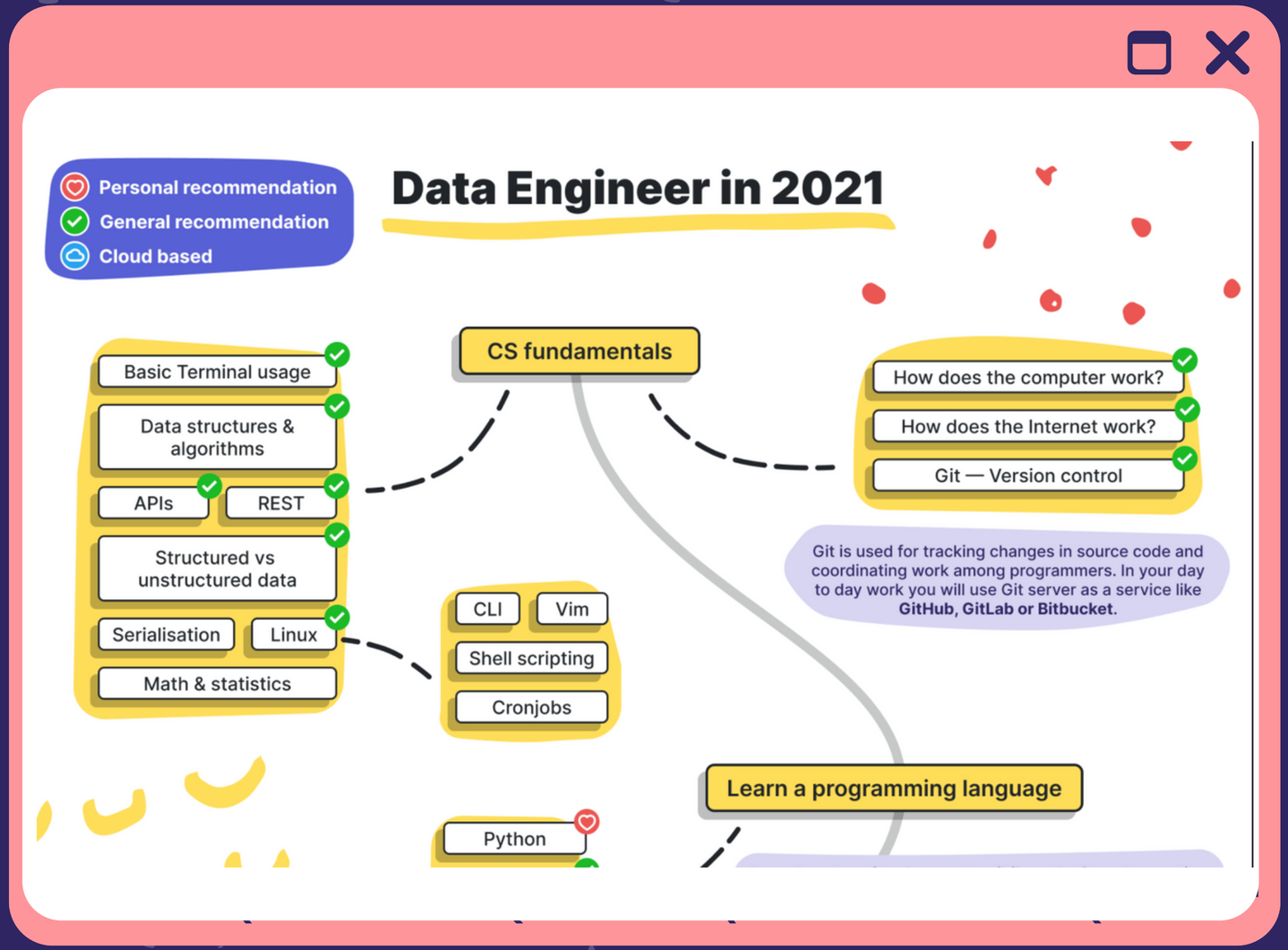
Data engineering - Job market



Where to get started

- Infrastructure
  - Free tiers of the cloud
  - Containers - local development environment
- Keeping the data up to date
  - APIs
  - Frameworks (Airflow, Kafka, spark)
  - Cleaning the data

Where to get started

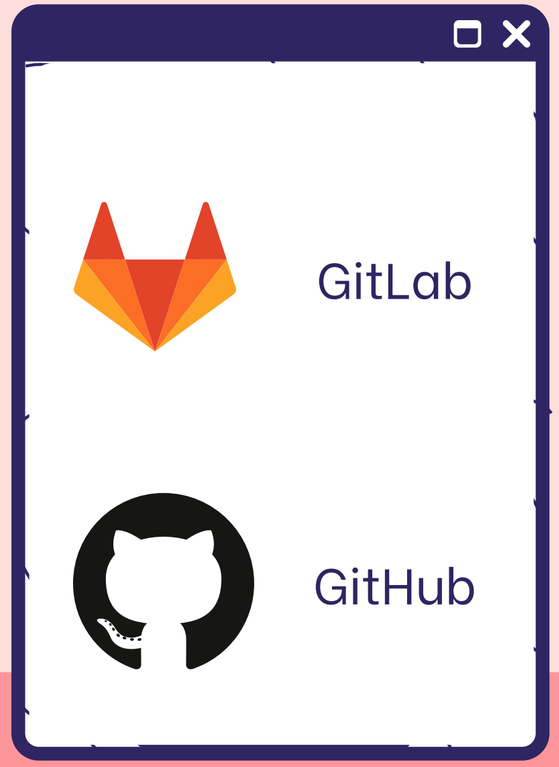


Build projects

- Abstract Ideas and concepts need practice.
- Work on different sets of problems.
- Learn to deal with frustration and unwanted emotions.
- Identify your weakest skill so you work on it in your next iteration.

We remember the things that we have built more than the things that we have seen being done.

Portfolio



Share your work, and get familiar
with others' work and Open-source

Communities!

We become similar to the people around us

Get close with people with similar goals

Get close to people who believe and support you

Communities!

Time for questions!

Thank you!

PyLadies Amsterdam

Carolina

@Kittylon