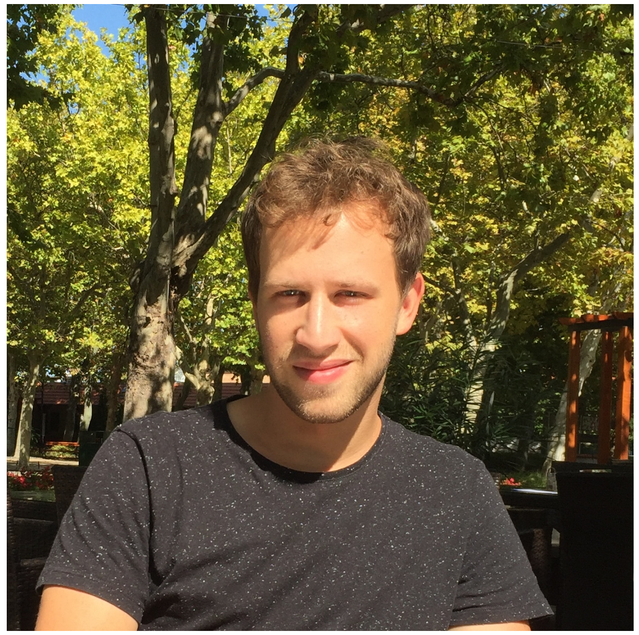
The Proactive Data Catalog

June 2022

stemma™

from the creators of amundsen

Introduction



Bálint Haller

Senior Software Engineer, Stemma

Previously Data Engineering @ Shapr3D

www.linkedin.com/in/balinthaller

Proactive vs reactive

Proactive

acting in anticipation of future problems, needs, or changes

Reactive

readily responsive to a stimulus

Proactive

Anticipate changes

- data freshness
- data structure
- meaning

Reactive

Proactive

Anticipate changes

- data freshness
- data structure
- meaning

Reactive

Observe changes, and act if the need arises

- Fix broken ETL/ELT pipelines
  - calculations
  - meaning

How does this connect to metadata?

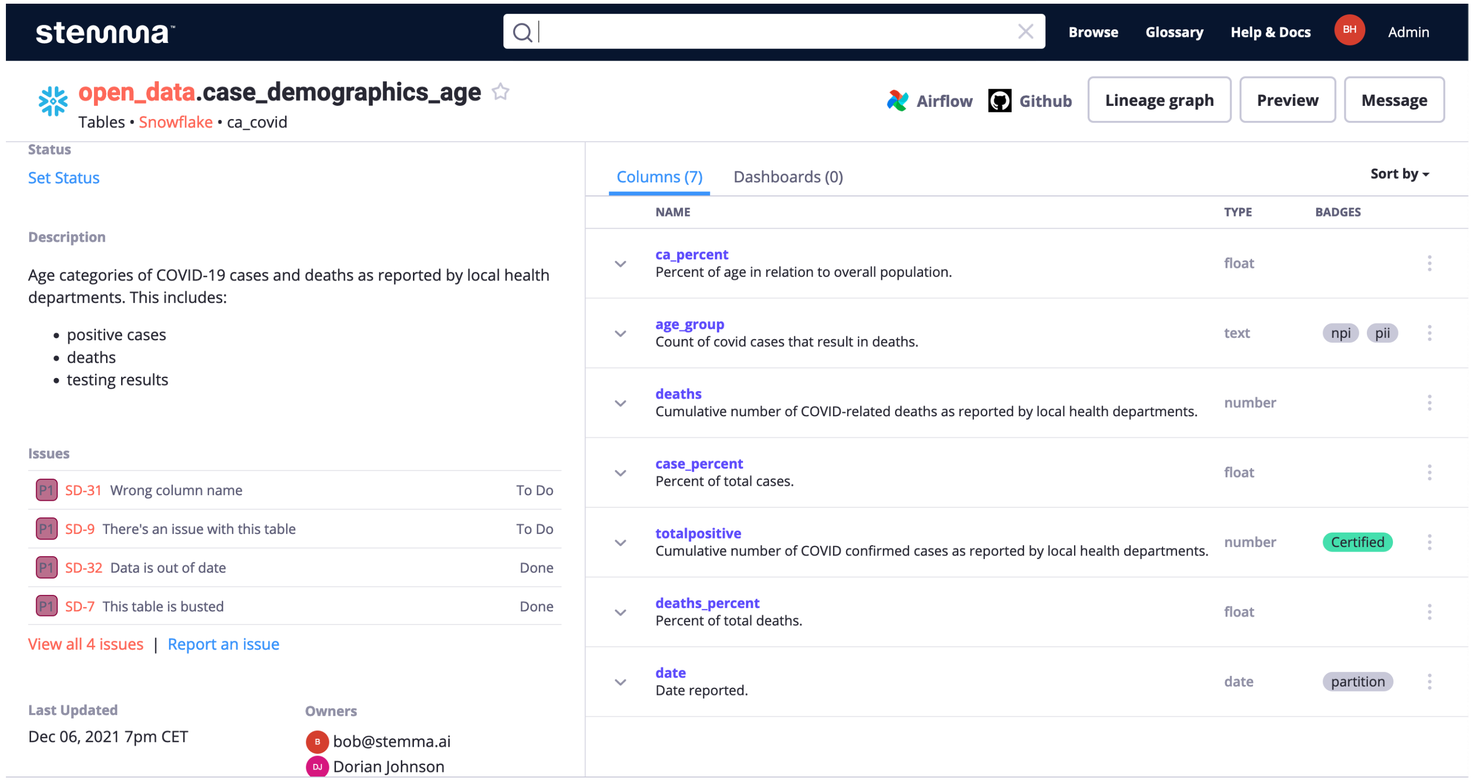
stemma™

Data Catalog as a thing

- An organised inventory of data assets in the organisation
- Helps data professionals collect, organise, access, and enrich metadata to support data discovery and governance

Data Catalog as a thing

- Central part is the **description** of datasets, columns, dashboards
  - And any other data asset potentially
- Slightly different from ‘Data Observability’ tools that are meant to detect anomalies
  - Can be integrated though!



Data Catalog as a thing

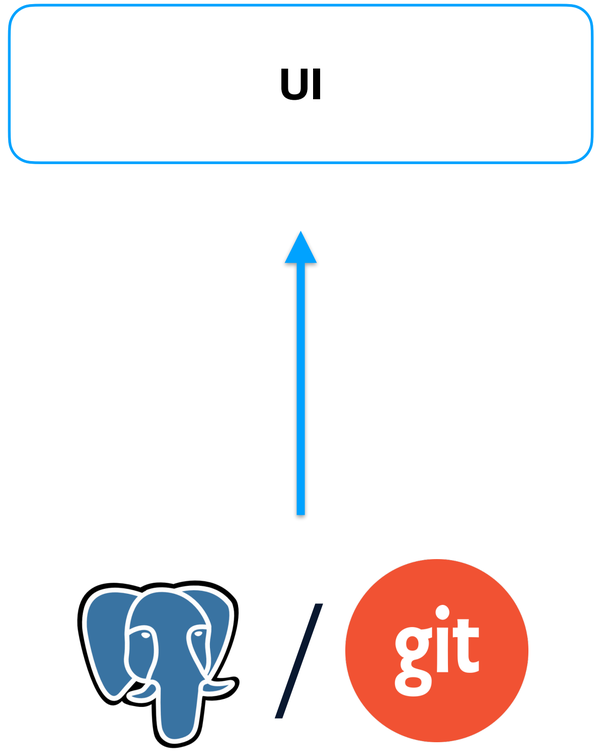
- General shift from people-oriented to automated
  - This is mostly in the **reactive -> proactive** direction
  - Automated **doesn't lose the human element**
  - Dataset **owners maintain** their own descriptions (instead of data stewards)
- Supporting tools in the **modern data stack**
- **connecting things** rather than trying to pull in every feature (SQL editor, dashboards, etc.)

How do you handle descriptions?

stemma™

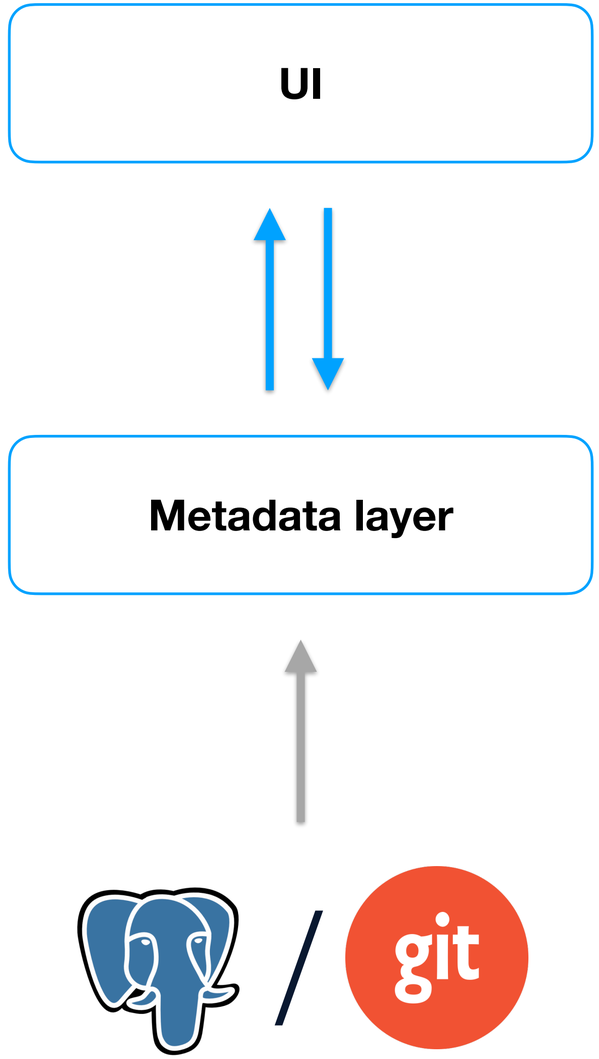
Handling descriptions

- From the DB/event schema
  - The closer to source, the better*
  - Can even get rid of the metadata layer and serve the exact same descriptions



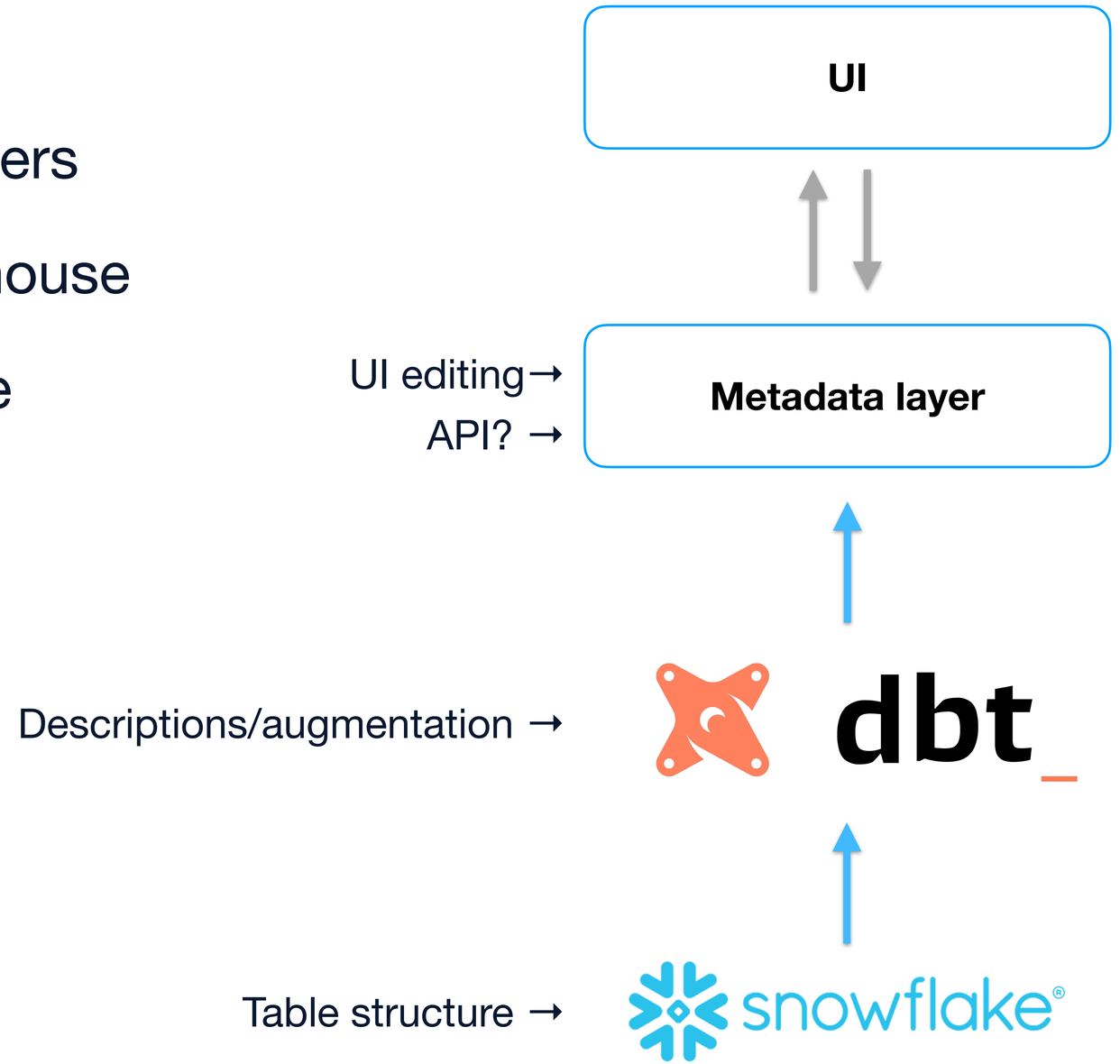
Handling descriptions

- However! You **still need UI editability** for non-technical users
  - Otherwise you run the risk of not having a ‘network effect’
  - Easier to get out of sync
  - Can be very reactive and outdated



Handling descriptions

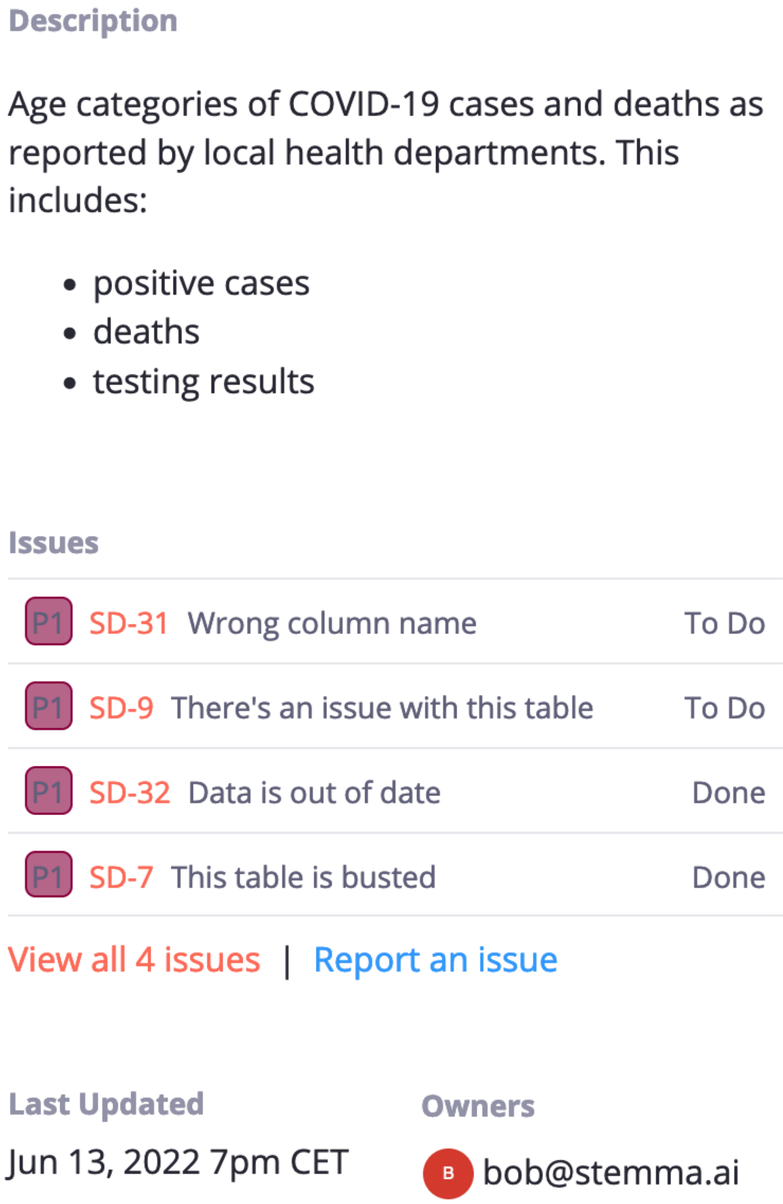
- Some companies have extra layers
- E.g. DBT on top of a data warehouse
- Need to be able to resolve these
- Some catalogs expose an API
  - Ingestion or augmentation
  - Very flexible,
    but needs infra



What to do?

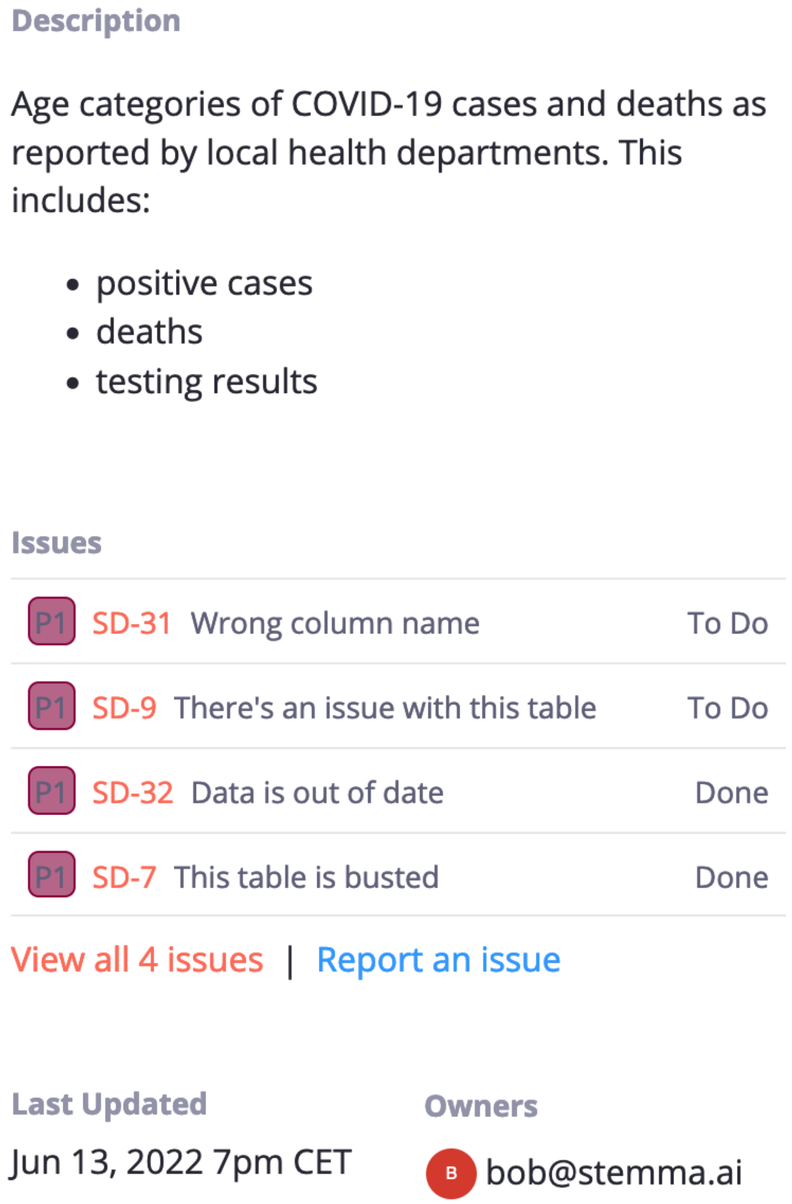
Handling descriptions

1. Display everything
   - More complex
   - Not the right call if you need canonical descriptions
2. Establish a resolution strategy
   1. Static priority list
   2. Based on freshness
   3. Can get fancy here, but it comes at the cost of transparency



Surfacing missing descriptions

- Do you surface missing descriptions?
  - Useful when the proactive approach fails, and you need to augment a significant portion of your descriptions by hand
  - or just to clean up a couple outlier datasets



How do you handle changes?

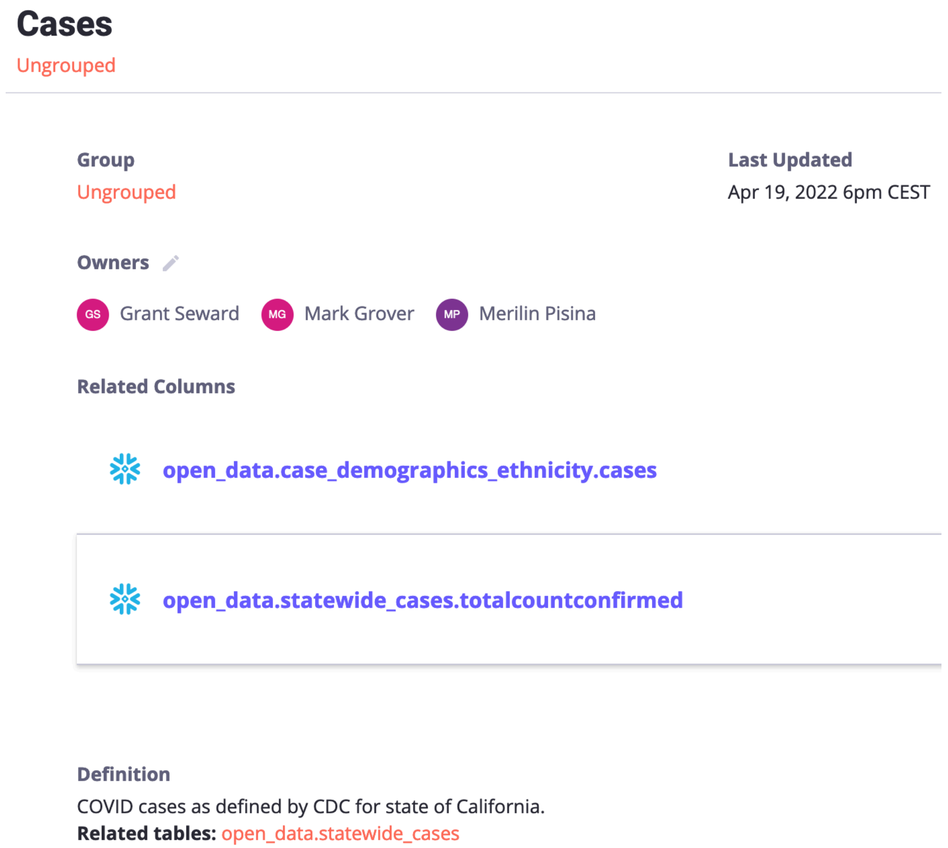
stemma™

How do you handle changes?

- Changes can painful, especially if multiple layers have information about your datasets' structure
- Lineage becomes important
  - Helps with notifying downstream and assessing impact
  - Ability to message downstream dataset owners can help prepare (more automation isn't usually possible)

Handling changes

- Shared term glossary is helpful – Proactive
  - Leaves less room for errors
- Column and table description linking
- Anything to make your descriptions DRY
- Data freshness and size alerts
  - The realm of observability
- Deleting stale data



Finding and fixing errors

Finding and fixing errors

- When the proactive approaches fail
  - unless you have very strong operational checks in place, and you can catch errors on a PR level
  - Helpful to have your data catalog update or DBT update (if the catalog depends on that) built into your deployment pipeline
  - Versioning is super helpful
    - Can even be git

Closing thoughts

- These were just a couple questions through the proactive vs reactive lens
- Reactive has its place, **after the proactive** approach fails
- **Keep all the data** you can, but be conscious about **what you display**
- A lot depends on your organisation structure
  - Do Data Scientists write ETL jobs?
  - Are only Data Engineers responsible for operations?

Thank You

stemma™

from the creators
of amundsen