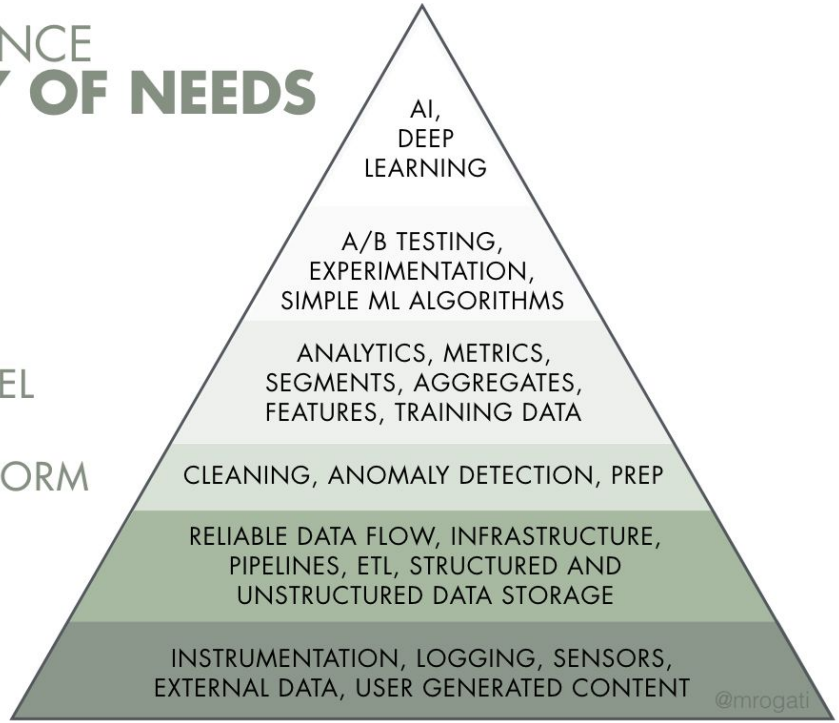
Data quality monitoring

Budapest Data Forum 2020

Data quality issues the hard way

- Started as a machine learning engineer
- Risk analytics for insurance / banking
- Many 3rd party dependencies
- Mistake: focussing on AI not on data pipelines

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

AGGREGATE/LABEL

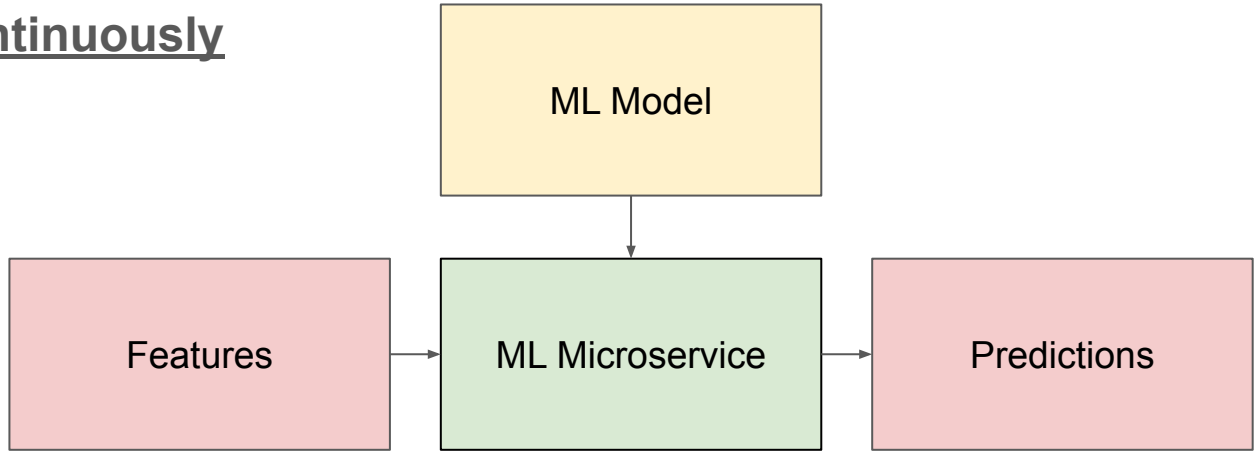
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

Machine learning services change infrequently

- Microservice changes **rarely**
- ML model changes **occasionally**
- Features change **continuously**



Dependency management is unsolved for data

- 3rd party vendors
- Customer data dumps
- Website tracking code (avo.app)
- Internal microservices
- ETL pipelines and batch transforms
- Database owned by another team
- Contracts are often undefined or broken

The data warehouse has become the hub

Co-founded Hubble

Hubble runs tests on your data warehouse so you can identify issues with data quality. You can test for things like missing values, uniqueness of data or how frequently data is added/updated.

Extract & Load are getting commoditised

Fivetran / Stitch / Matillion (others available)

-> EL is managed but not free of problems

Tests

- Schema changes
- Data volume
- Freshness

SQL is the new Transform (but not always)

Running transforms in the warehouse is commonplace.

But often spark / custom scripting too

Tests

- Business logic checks
- Business metrics
- Foreign key checks
- Uniqueness
- String formats

3 types of tests

Hard tests: the answer is known

e.g. User ID should not be null

Soft tests: the answer is a grey area

e.g. approximately 1,000 new records should be created each day

Metadata tests: the answer is relevant for governance

e.g. schema changes, permissions, audit logs