Ground: A Data Context Service

Jiaqi Gao

What have we achieved?

- OLTP vs OLAP
- SQL vs noSQL
- Fast scan, fast index
- Compression
- Transactions
- Version
- ...

A DECOUPLED STACK

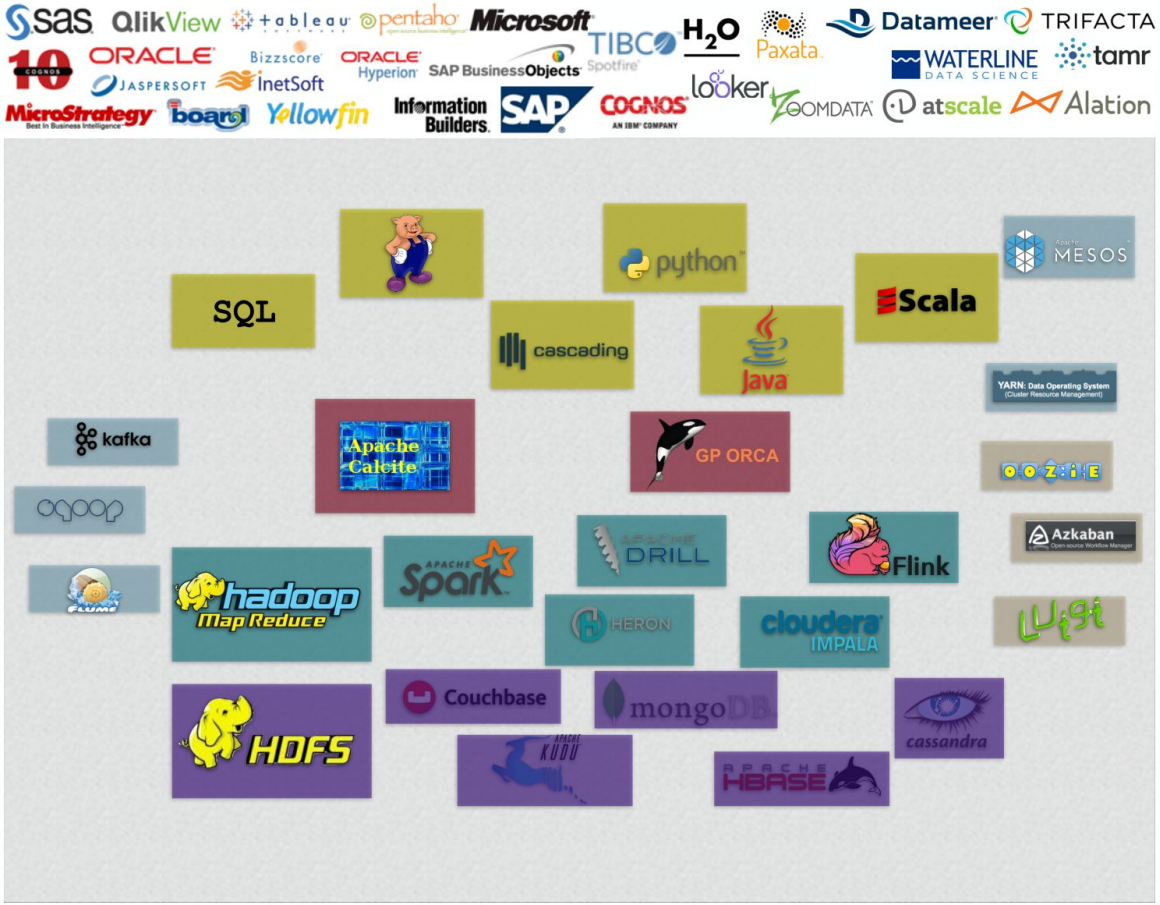
Big Data

A DECOUPLED STACK

The Good: Agility

A DECOUPLED STACK

The Bad: Dis-integration.



Crisis: Big Metadata

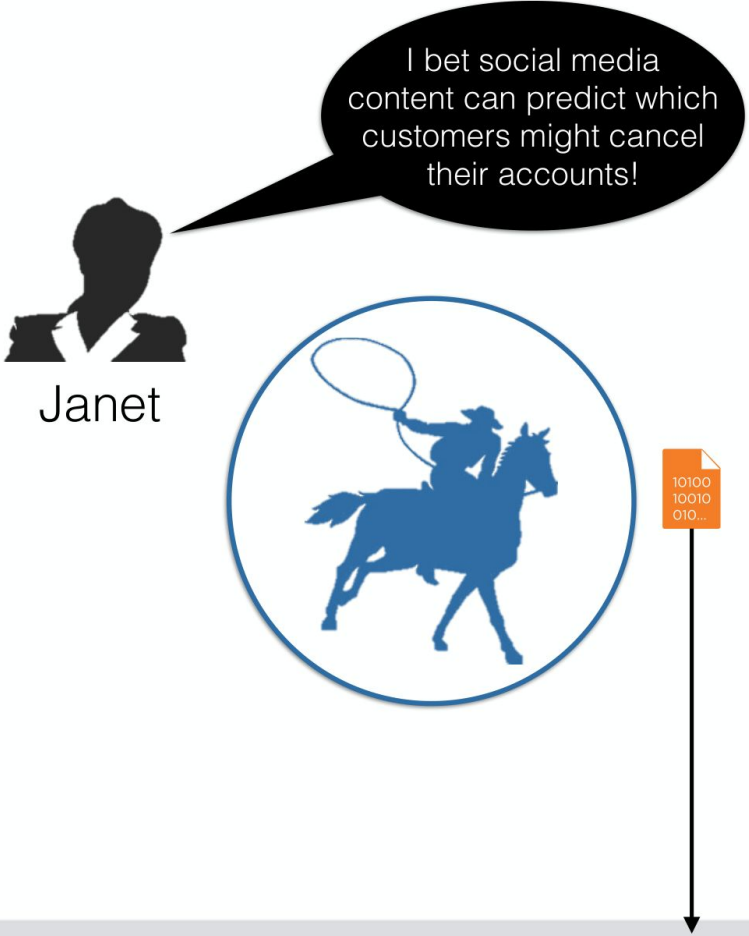
- Lack of standard mechanism to assemble a collective understanding of the origin, scope, usage of the data
- Metadata can be larger than data itself

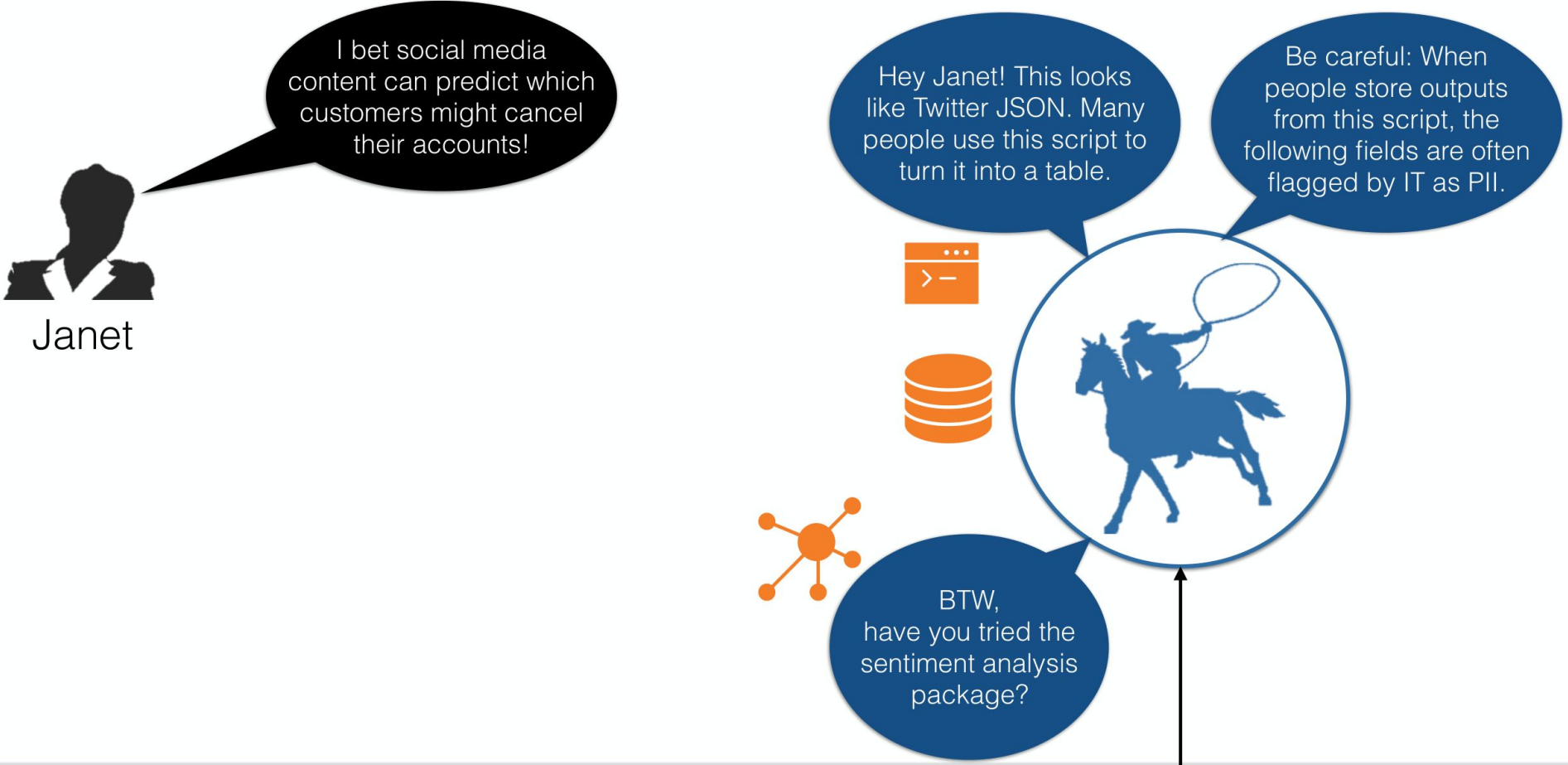
- Poor productivity
  - Hard to discover whether data exists, how it’s been used previously
- Governance risk
  - Hard to know who access the data, what they do with it, where they put it and so on

Solution: Data Context

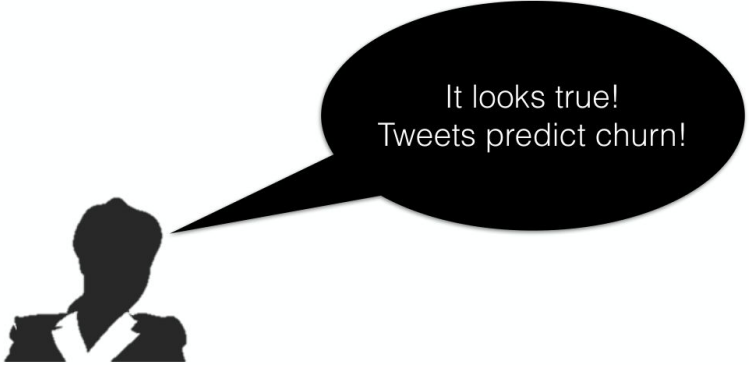
- All the information surrounding the use of data
- ABC of data context
  - Application Context
  - Behavior Context
  - Change Over Time

- *Ground*: open source data context service
  - Provides central model, API and repository for capturing the data context
- A database assistant

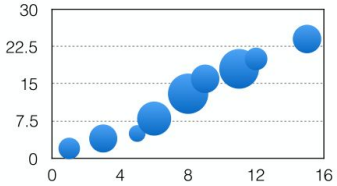

I bet social media content can predict which customers might cancel their accounts!
Janet
10100
10010
010...
Hey Janet! We already paid for a full Gnip feed from Twitter — you can find it here
By the way: Sue used this following related table and script.
10100
10010
010...
ground


I bet social media content can predict which customers might cancel their accounts!
Janet
Hey Janet! This looks like Twitter JSON. Many people use this script to turn it into a table.
Be careful: When people store outputs from this script, the following fields are often flagged by IT as PII.
BTW, have you tried the sentiment analysis package?

ground

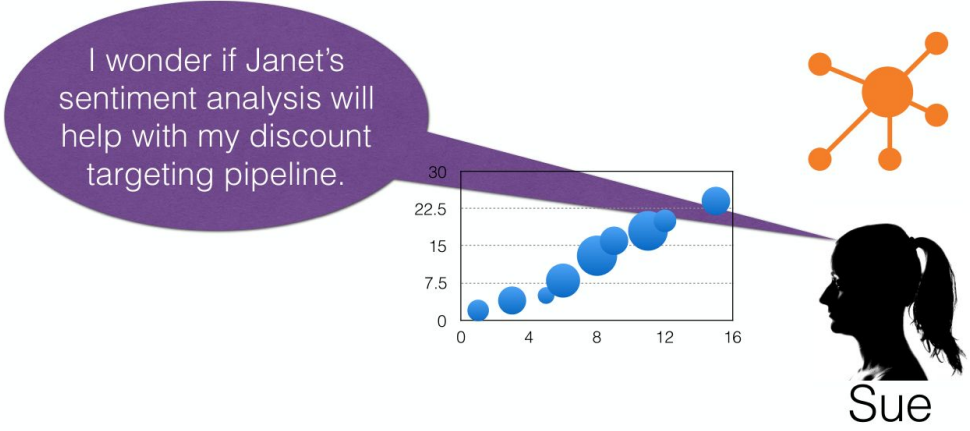


Janet



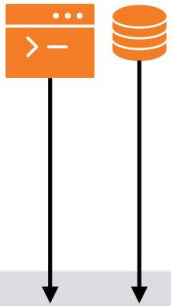
share

Sue



jupyter

TweetId	Text	neg	pos	neut
47	“sad!”	1	0	0
53	“awesome!”	0	1	0
57	“go packers!”	0	0	1
64	“fleek!”	0	1	0



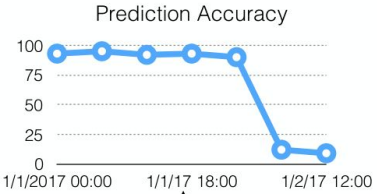
ground

Time passes…

Oh dear. I better call a meeting to introduce better governance on sentiment labeler.

Uh oh, prediction accuracy metrics are down!

Sue



TweetId	Text	Sentiment
47	“sad!”	sadness
53	“awesome!”	elation
57	“go packers!”	sports
64	“fleek!”	trendy

```
VERSION HISTORY
12/31/2016 00:00 -800
hash:
6dda491064bcce14f558bf83867b8c247027c423
user: will
```

FYI: Janet’s wrangling script changed!

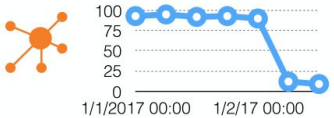
ground

WHAT DID CONTEXT ENABLE?

Self-service catalog, wrangling and analytics.

Collective governance of data.

Fueling our model accuracy monitor.

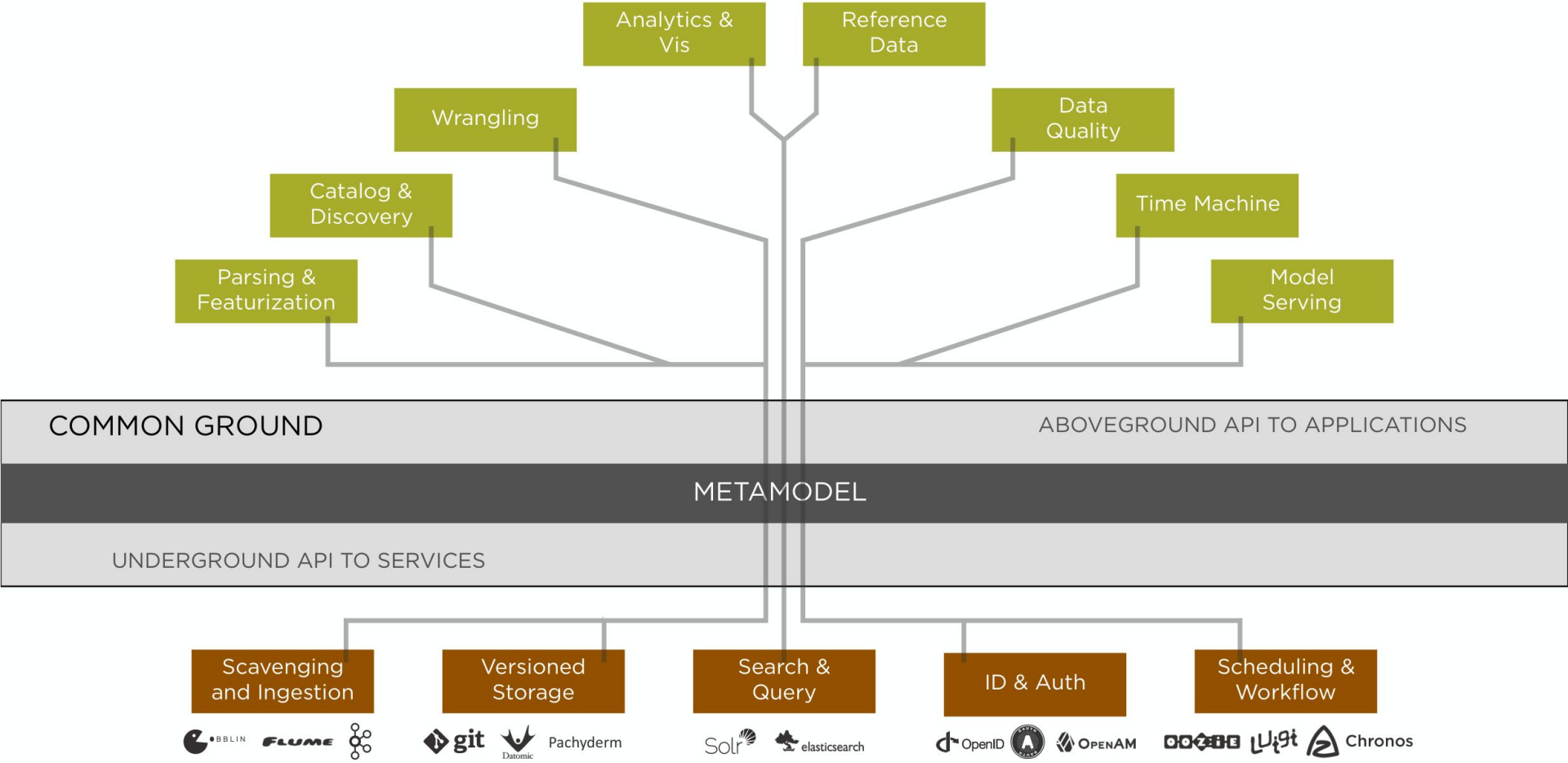


Figuring out which changes introduced the error.

VERSION HISTORY

Determining who made the change to help us resolve the issue.

user: will


Analytics & Vis
Reference Data
Wrangling
Data Quality
Catalog & Discovery
Time Machine
Parsing & Featurization
Model Serving
COMMON GROUND
ABOVEGROUND API TO APPLICATIONS
METAMODEL
UNDERGROUND API TO SERVICES
Scavenging and Ingestion
Versioned Storage
Search & Query
ID & Auth
Scheduling & Workflow
BBLIN
FLUME
git
Datomic
Pachyderm
Solr
elasticsearch
OpenID
OpenAM
Luigi
Chronos

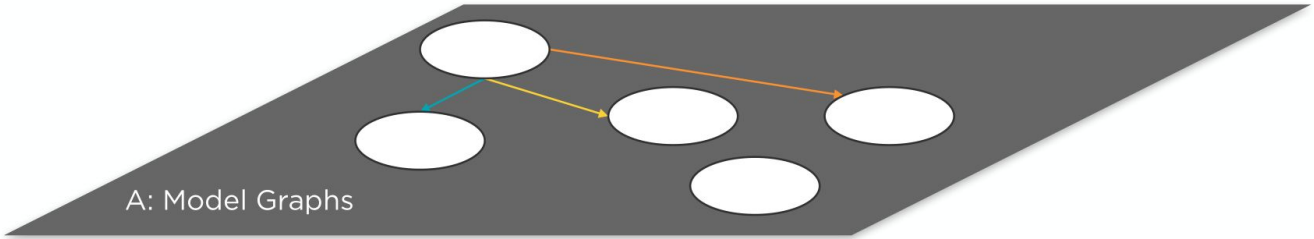
Design Requirements

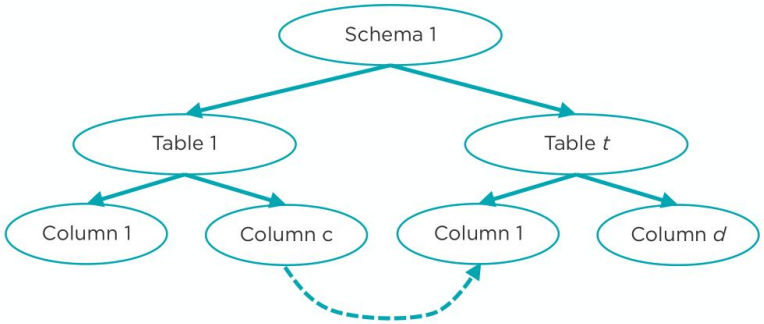
- Model-agnostic
  - Can’t prescribe how metadata is modeled
- Immutable
  - Can’t erase history
- Scalable
  - Log, version history, lineage
- Politically Neutral
  - Interoperate with a wide range of applications

COMMON GROUND

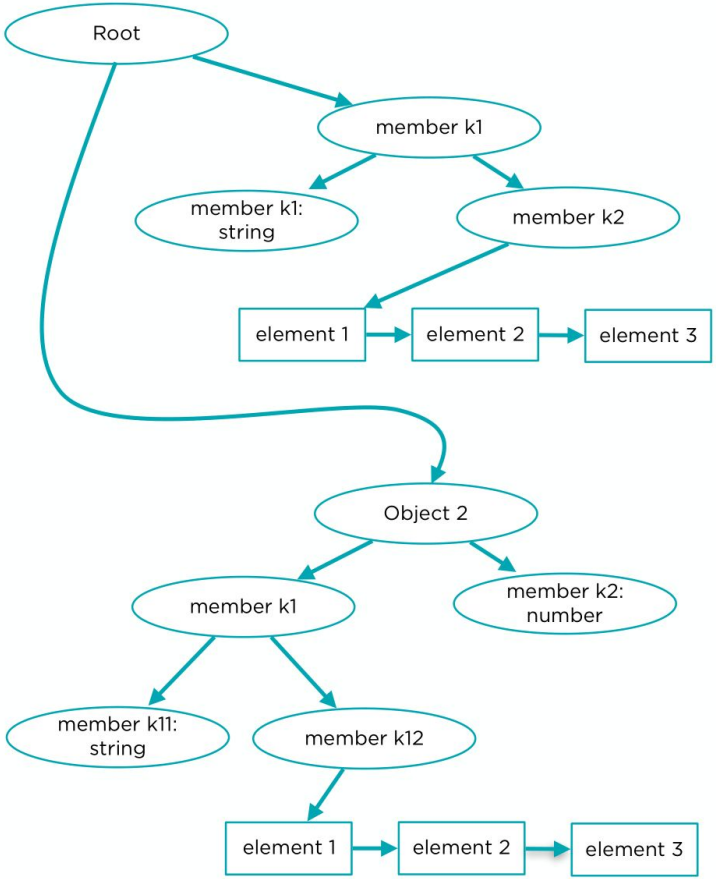
The metamodel

- A DAG shows how metadatas are related





RELATIONAL SCHEMA

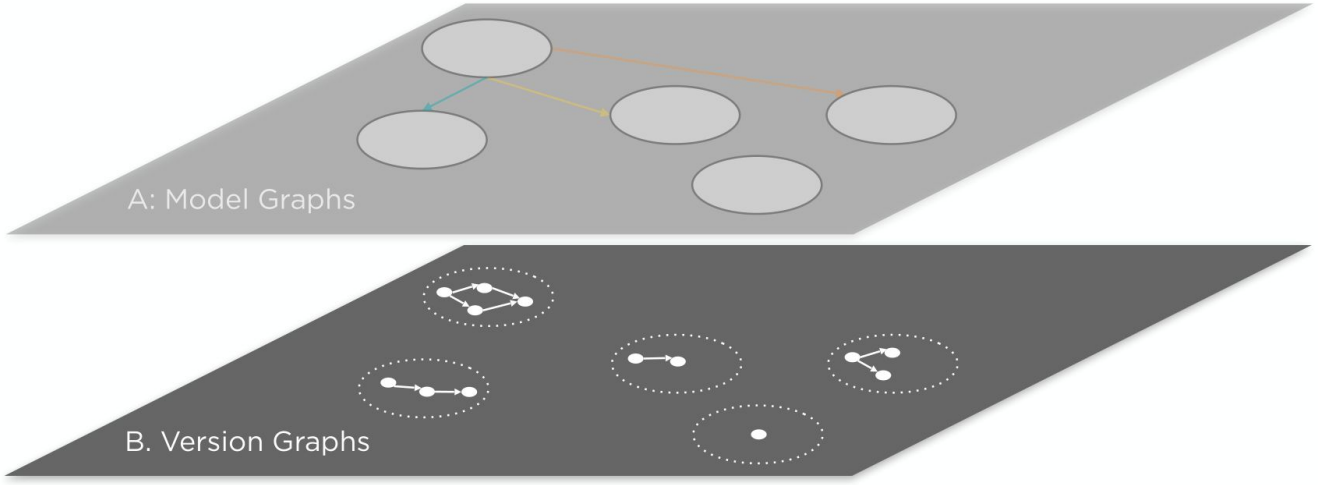


JSON DOCUMENT

COMMON GROUND

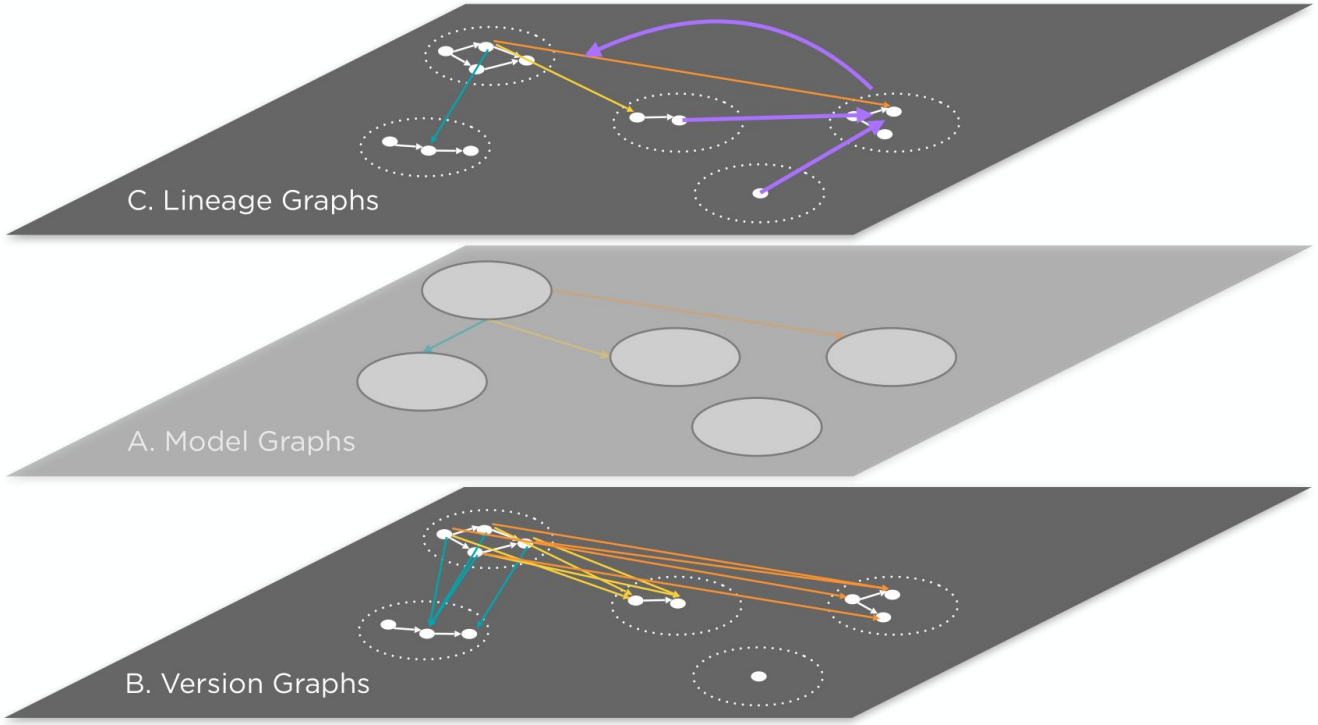
The versioning model

- Shows how metadata changes over time



COMMON GROUND

The usage model



- How used the metadata and how

Open Research Questions

Underground

- Workloads
- Common Ground representations
- No-overwrite versioned DB
- Time travel queries: point and trend Graph queries + log analysis
- Consistency

Aboveground

- Content extraction
- Analytic user exhaust
- Socio-technical networks
- Collective governance
- Reproducibility
- Lifecycle of systems that learn

References

- Ground Context: http://www.ground-context.org/
- CIDR 2017 presentation: http://cidrdb.org/cidr2017/slides/p111-hellerstein-cidr17-slides.pdf