Ground: A Data Context Service

Joe Hellerstein, Vikram Sreekanti, Joey Gonzalez, *et al.*

CIDR 2017

https://github.com/ground-context/ground

riselab UC Berkeley awake networks Capital One cloudera

DATAGUISE LinkedIn skyhigh TRIFACTA

ground

Open Source Big Data Community Health

Data Analysis

✓

Data Wrangling

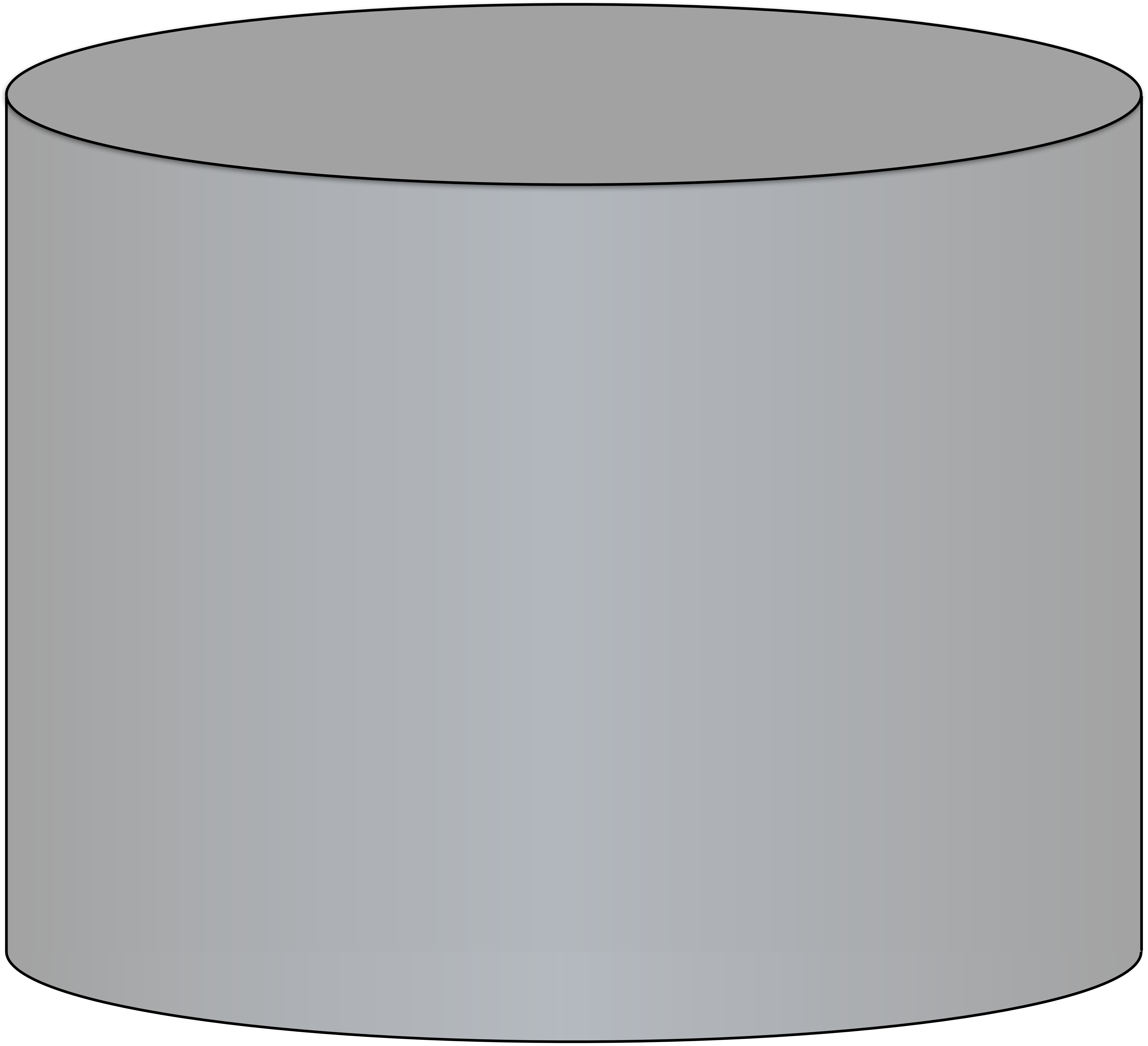
✓

Long-term Data Management

FAIL

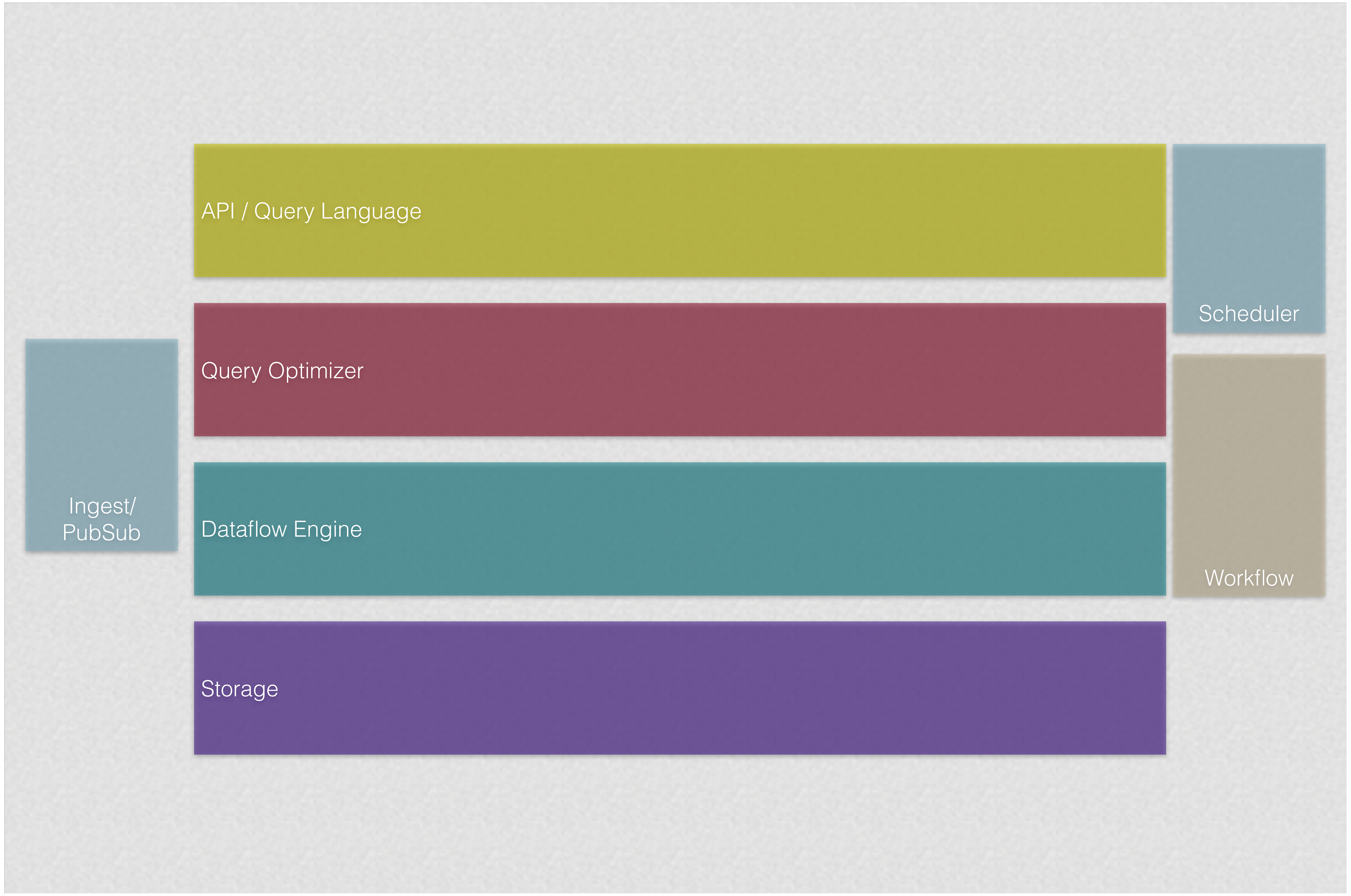
What was the big data revolution really all about?

Database



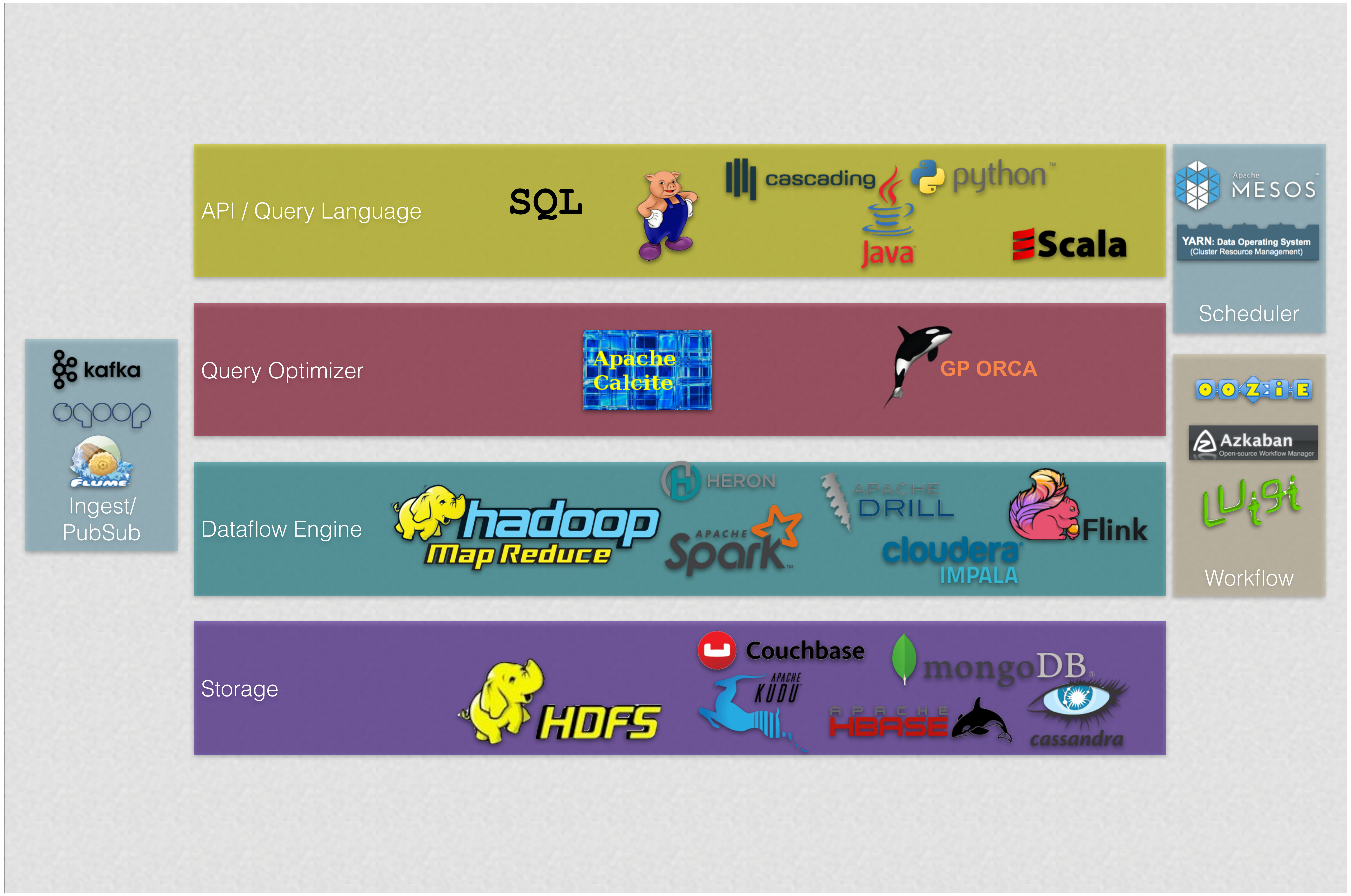
A DECOUPLED STACK

Big Data



A DECOUPLED STACK

The Good: Agility



API / Query Language

Query Optimizer

Dataflow Engine

Storage

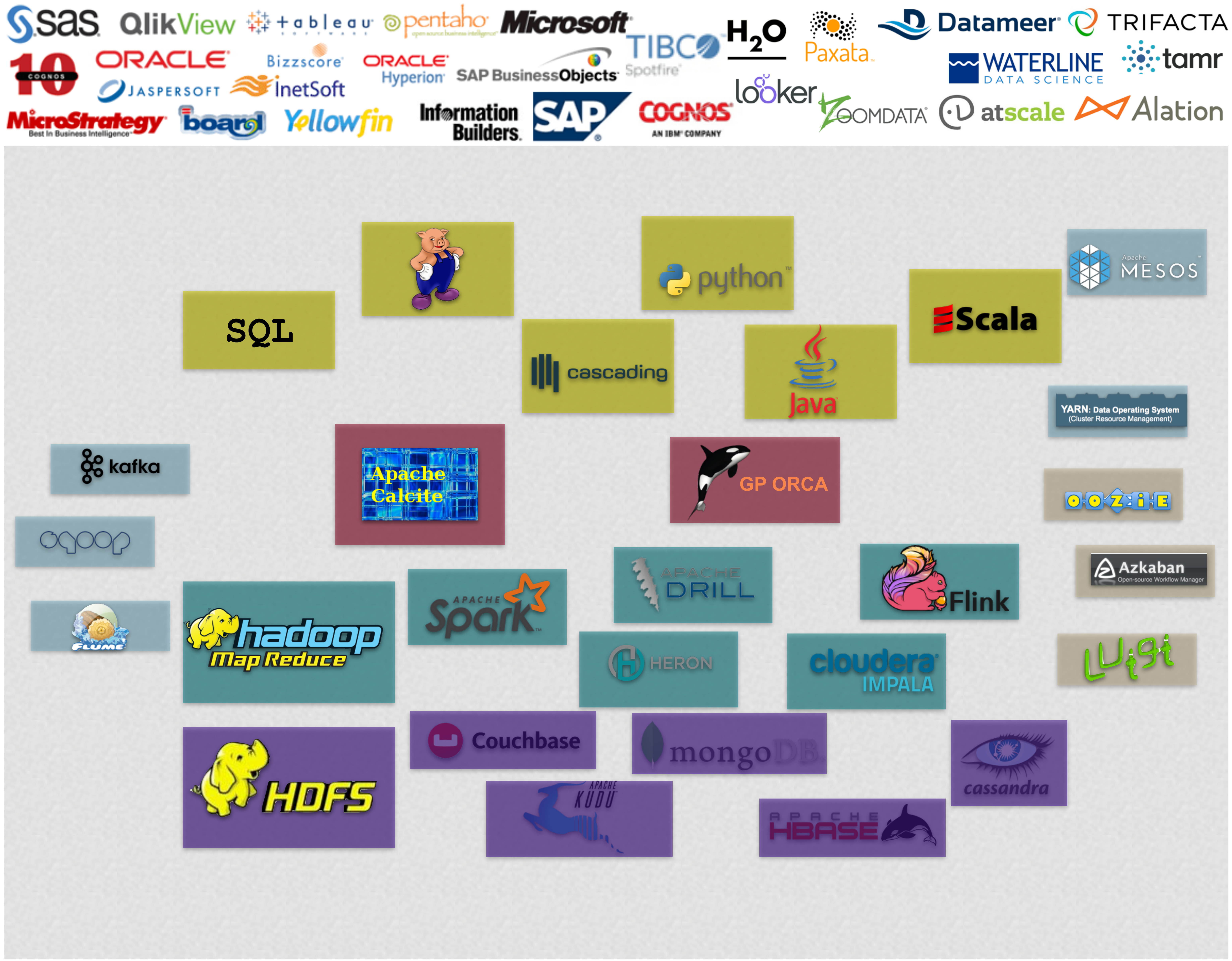
Ingest/ PubSub

Scheduler

Workflow

A DECOUPLED STACK

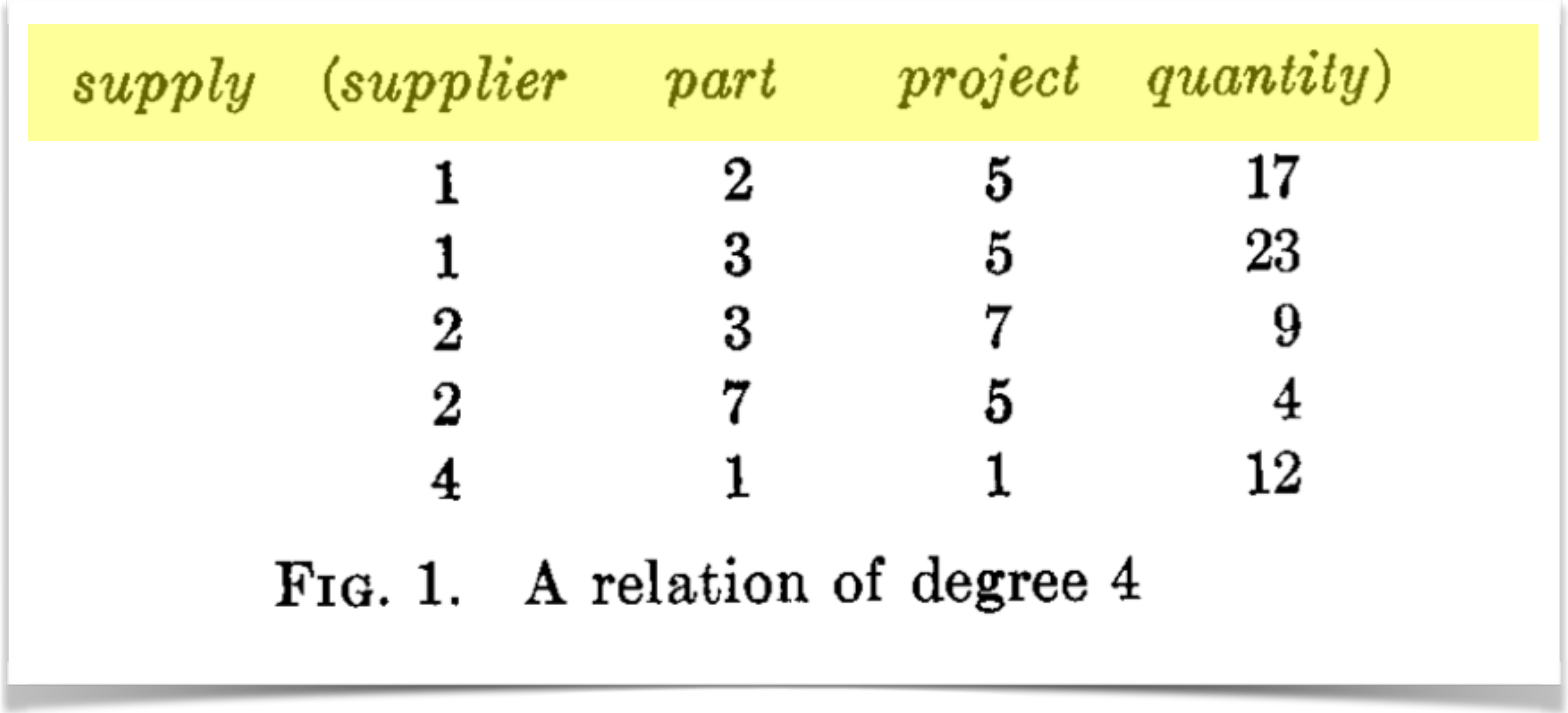
The Bad: Dis-integration.



CRISIS: HOW DO WE SHARE INFORMATION?

METADATA

WHAT IS METADATA?

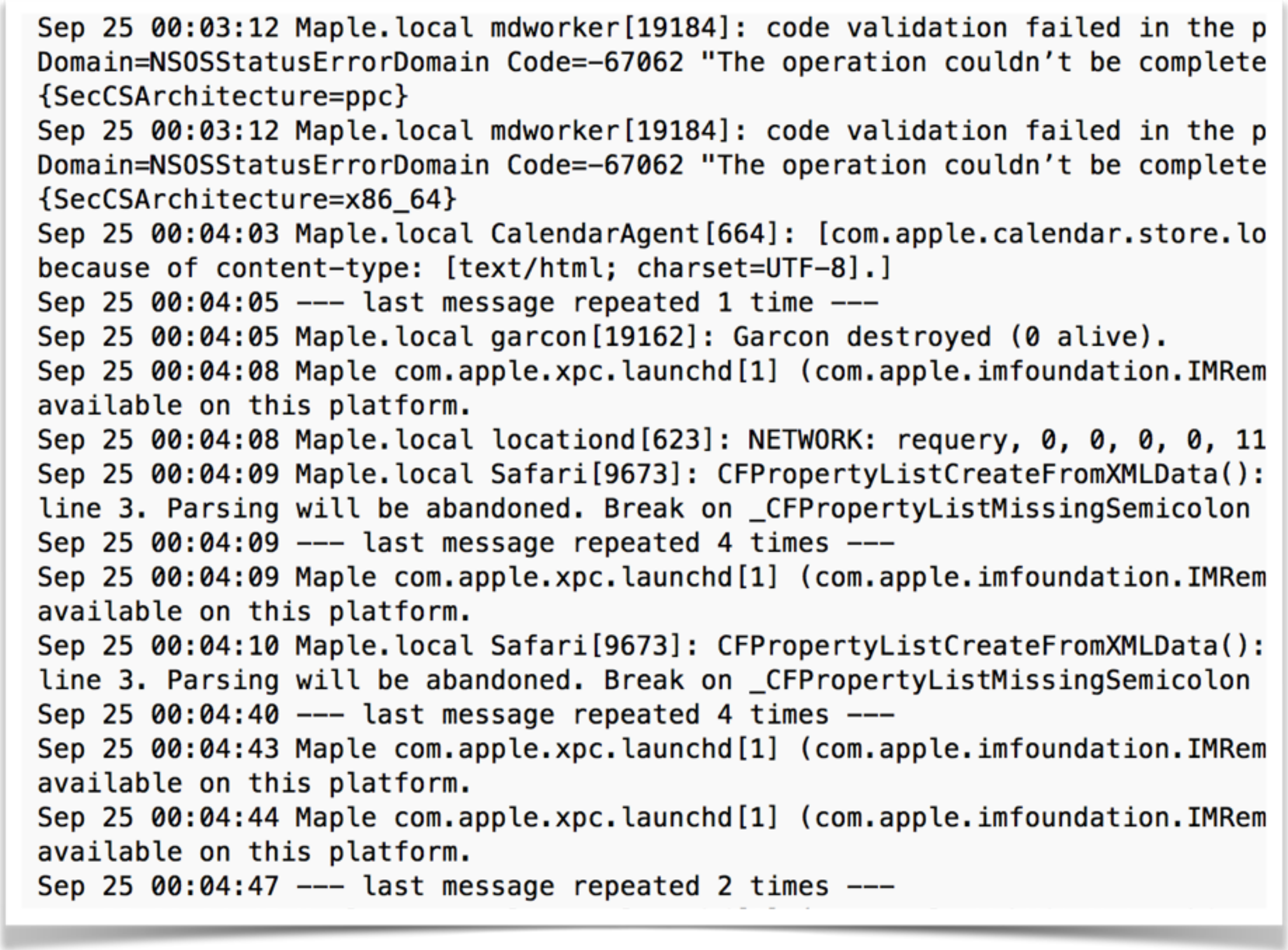


supply	*(supplier*	*part*	*project*	*quantity)*
	1	2	5	17
	1	3	5	23
	2	3	7	9
	2	7	5	4
	4	1	1	12

FIG. 1. A relation of degree 4

WHAT IS METADATA?

- Data about data
  - This used to be so simple!
- But … schema on use
  - One of many changes



```
Sep 25 00:03:12 Maple.local mdworker[19184]: code validation failed in the p
Domain=NSOSStatusErrorDomain Code=-67062 "The operation couldn't be complete
{SecCSArchitecture=ppc}
Sep 25 00:03:12 Maple.local mdworker[19184]: code validation failed in the p
Domain=NSOSStatusErrorDomain Code=-67062 "The operation couldn't be complete
{SecCSArchitecture=x86_64}
Sep 25 00:04:03 Maple.local CalendarAgent[664]: [com.apple.calendar.store.lo
because of content-type: [text/html; charset=UTF-8].]
Sep 25 00:04:05 --- last message repeated 1 time ---
Sep 25 00:04:05 Maple.local garcon[19162]: Garcon destroyed (0 alive).
Sep 25 00:04:08 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:08 Maple.local locationd[623]: NETWORK: requery, 0, 0, 0, 0, 11
Sep 25 00:04:09 Maple.local Safari[9673]: CFPropertyListCreateFromXMLData():
line 3. Parsing will be abandoned. Break on _CFPropertyListMissingSemicolon
Sep 25 00:04:09 --- last message repeated 4 times ---
Sep 25 00:04:09 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:10 Maple.local Safari[9673]: CFPropertyListCreateFromXMLData():
line 3. Parsing will be abandoned. Break on _CFPropertyListMissingSemicolon
Sep 25 00:04:40 --- last message repeated 4 times ---
Sep 25 00:04:43 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:44 Maple com.apple.xpc.launchd[1] (com.apple.imfoundation.IMRem
available on this platform.
Sep 25 00:04:47 --- last message repeated 2 times ---
```

OPPORTUNITY: A BIGGER CONTEXT

Don’t just fill a metadata-sized hole in the big data stack.

Lay the groundwork for rich **data context.**

WHAT IS DATA CONTEXT?

All the information surrounding the use of data.

ground

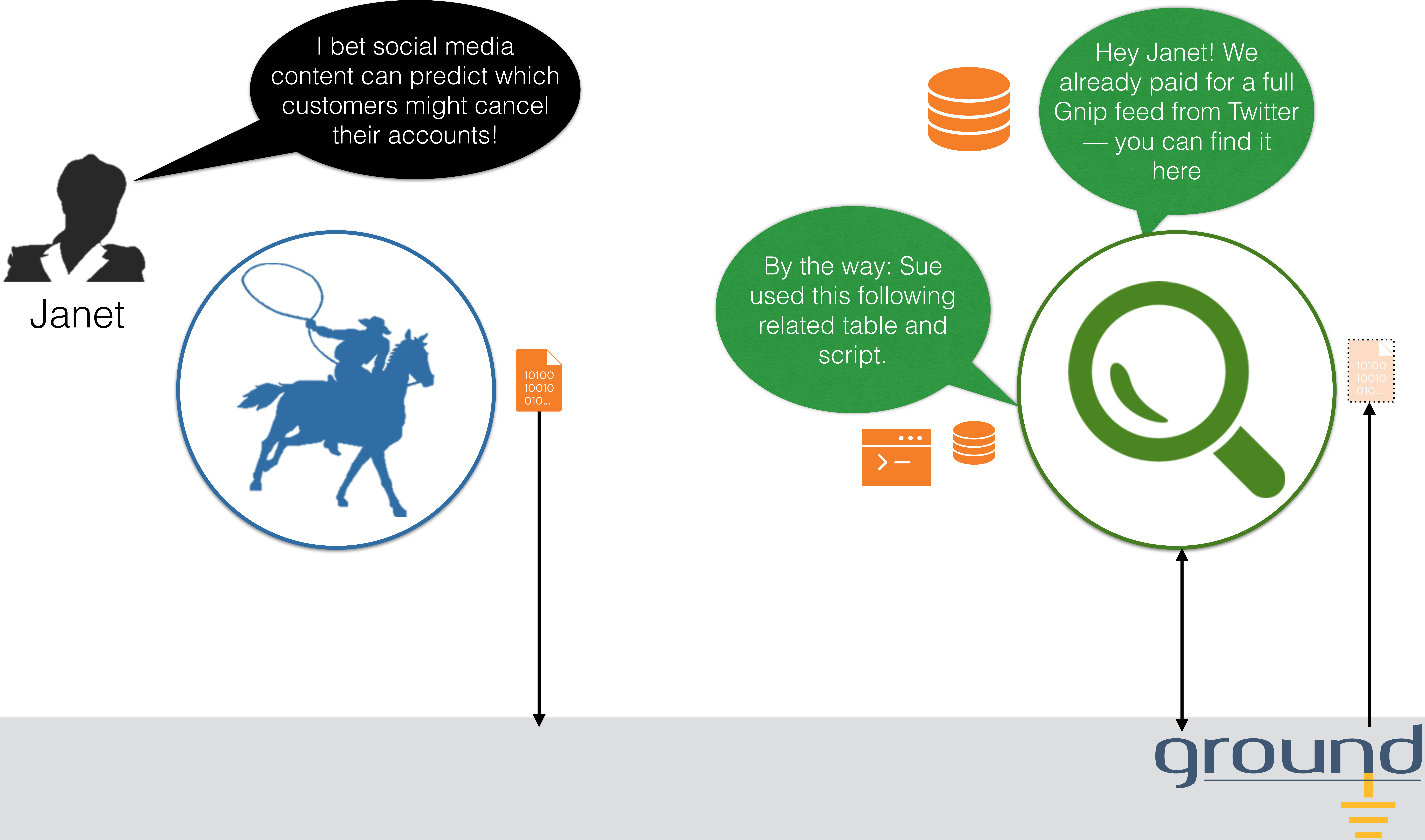
The ABCs of Data Context

Application Context: Views, models, code

Behavioral Context: Data lineage & usage

Change Over Time: Version histories

Generated by—and useful to—many applications and components.



I bet social media content can predict which customers might cancel their accounts!

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.

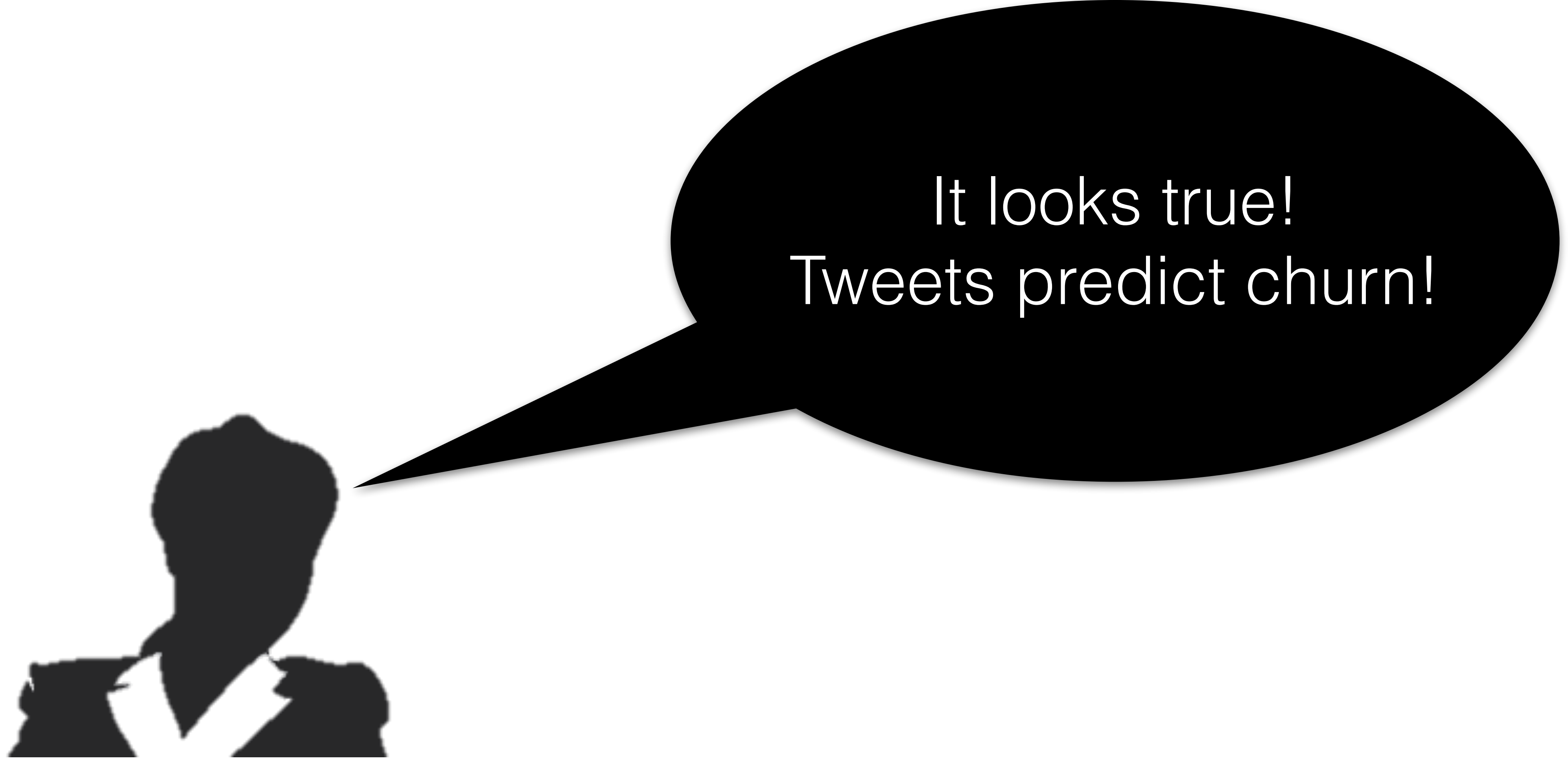
Janet

I bet social media content can predict which customers might cancel their accounts!

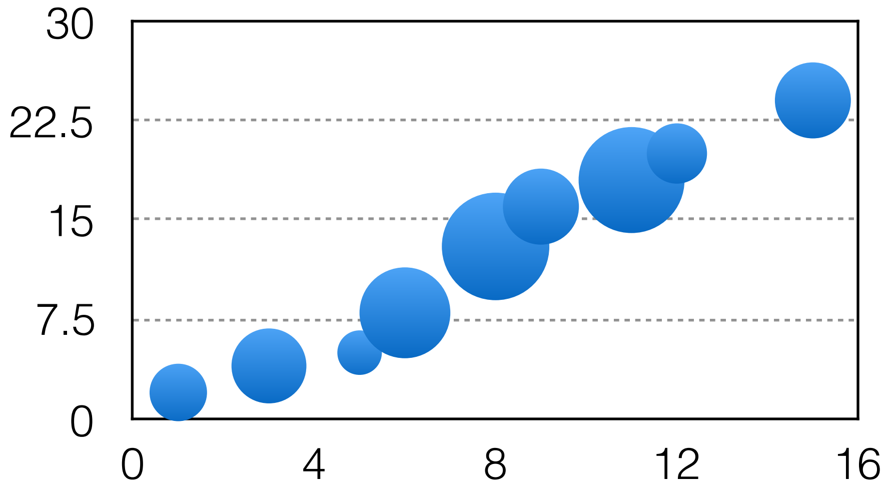
Hey Janet! This looks like Twitter JSON. Many people use this script to turn it into a table.

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?



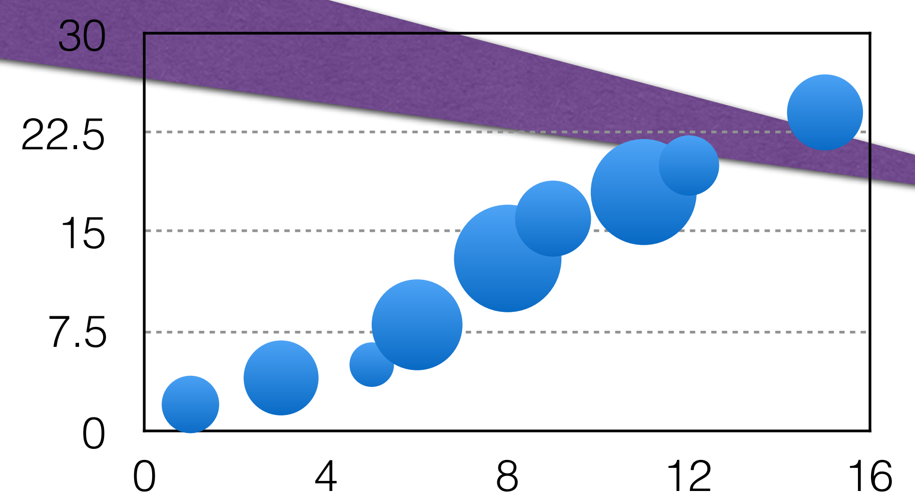
Janet



share

Sue

I wonder if Janet's sentiment analysis will help with my discount targeting pipeline.



Sue

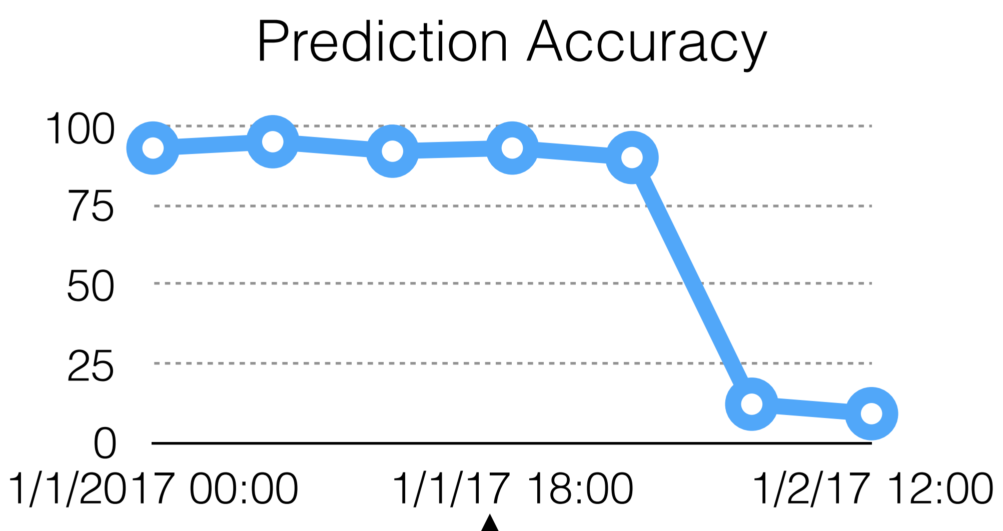
TweetId	Text	neg	pos	neut
47	"sad!"	1	0	0
53	"awesome!"	0	1	0
57	"go packers!"	0	0	1
64	"fleek!"	0	1	0

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
```

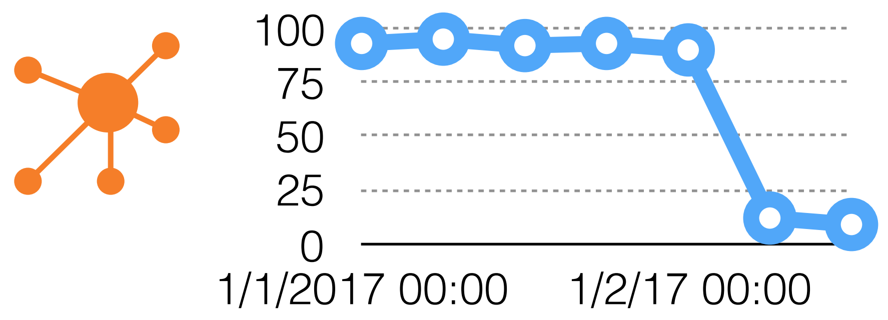
FYI: Janet’s wrangling script changed!

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`

THE BIG CONTEXT

Where are the interesting technical challenges?
All over!

Our goal is **not** to solve all these challenges.
It’s to provide an environment to enable solutions.

ground

COMMON GROUND

ABOVEGROUND API TO APPLICATIONS

METAMODEL

UNDERGROUND API TO SERVICES

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

GOBBLIN
FLUME
git
Datomic
Pachyderm
Solr
elasticsearch
OpenID
OAUTH
OPENAM
OOZIE
Luigi
Chronos

DESIGN REQUIREMENTS

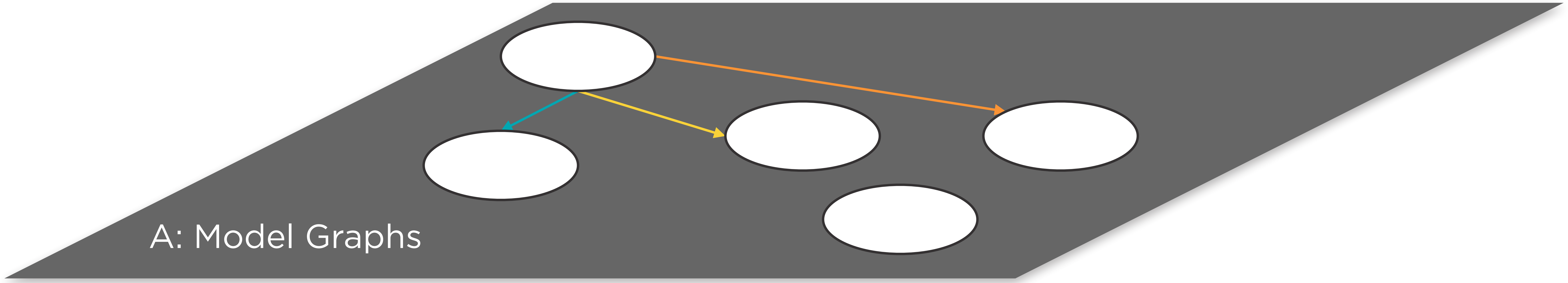
- Model-agnostic
- Immutable
- Scalable
- Politically Neutral

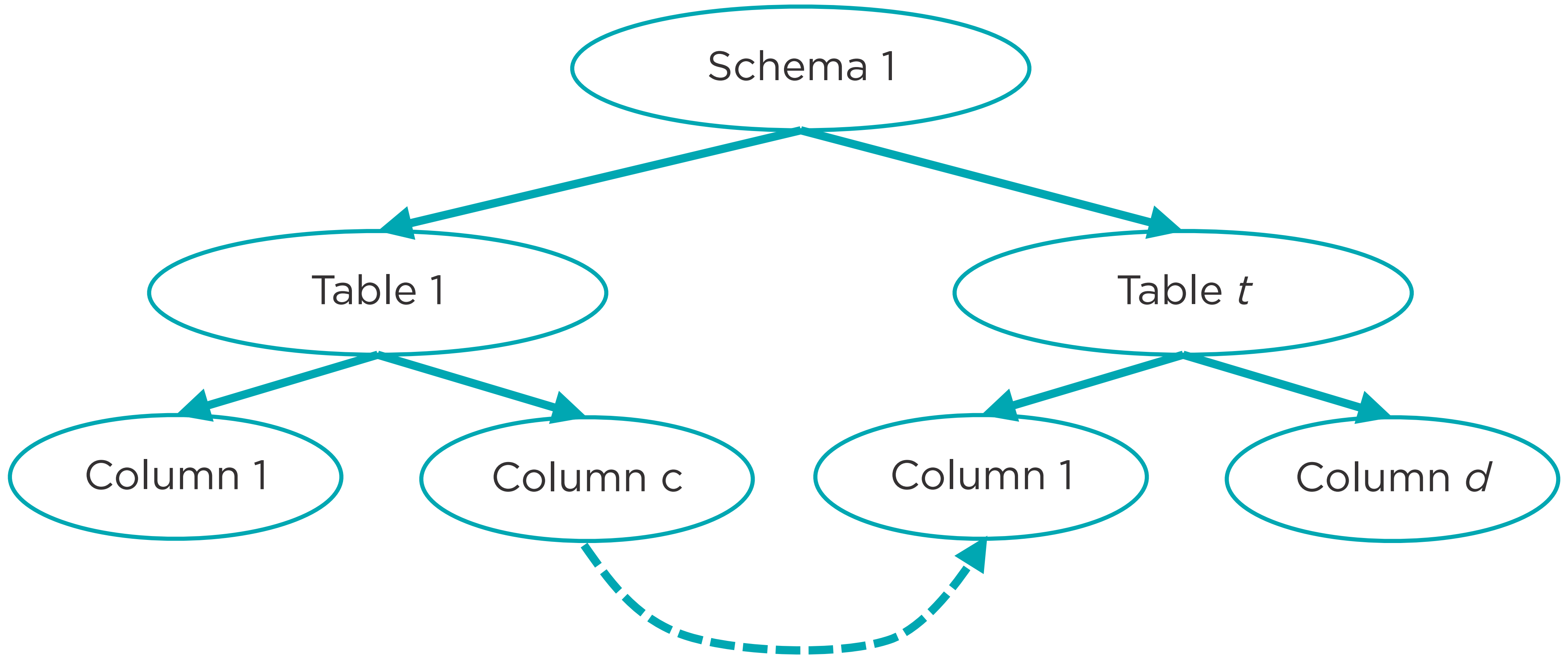
Postel’s Law

“Be conservative in
what you do,
be liberal in what you
accept from others ”

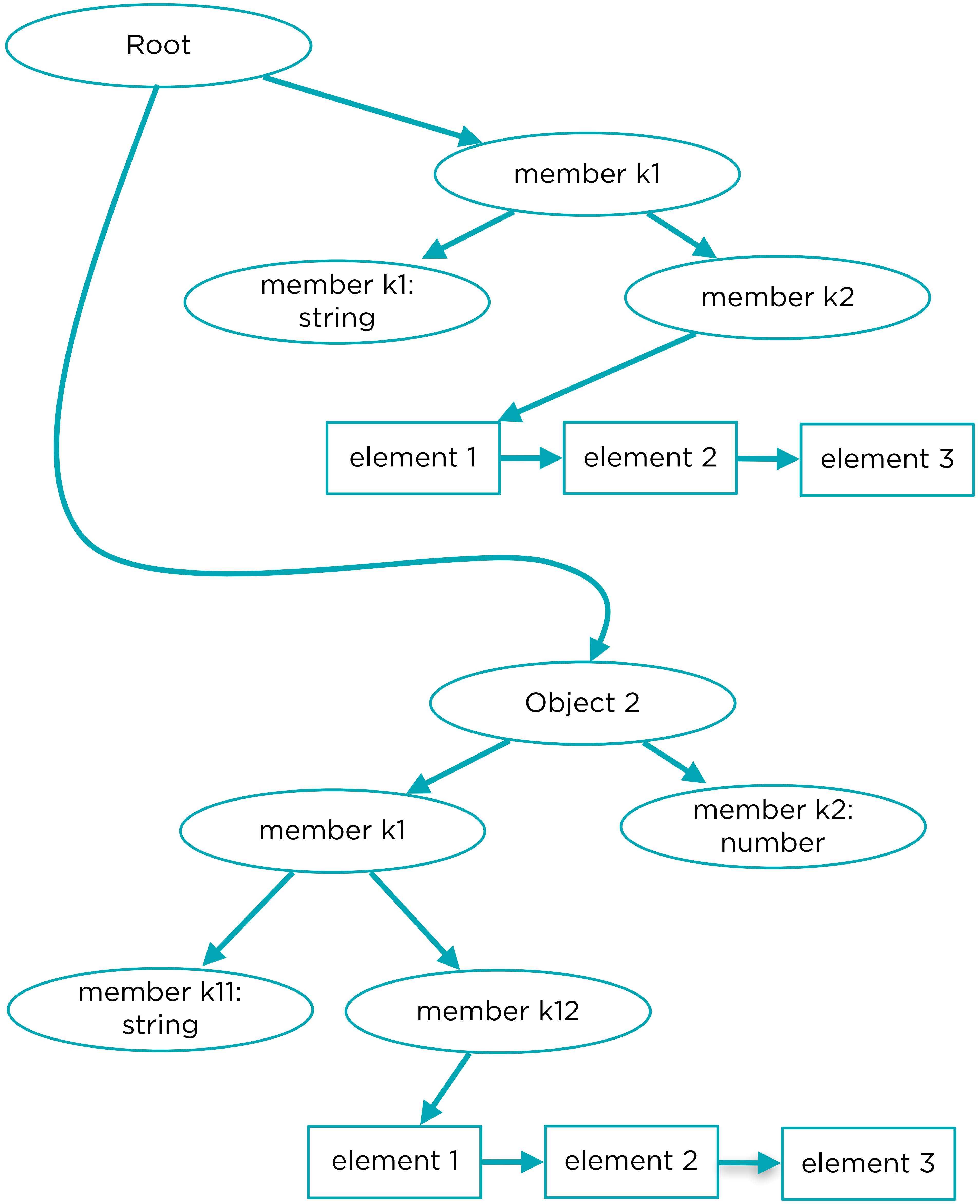
COMMON GROUND

The metamodel





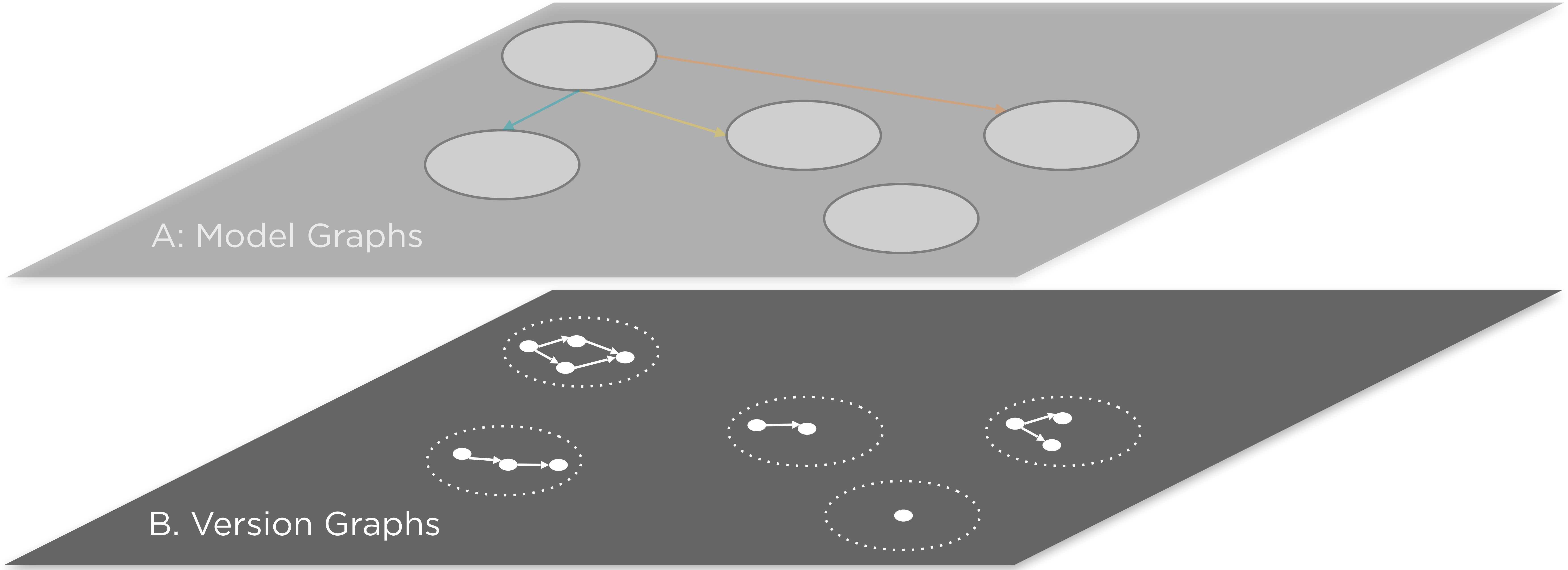
RELATIONAL SCHEMA



JSON DOCUMENT

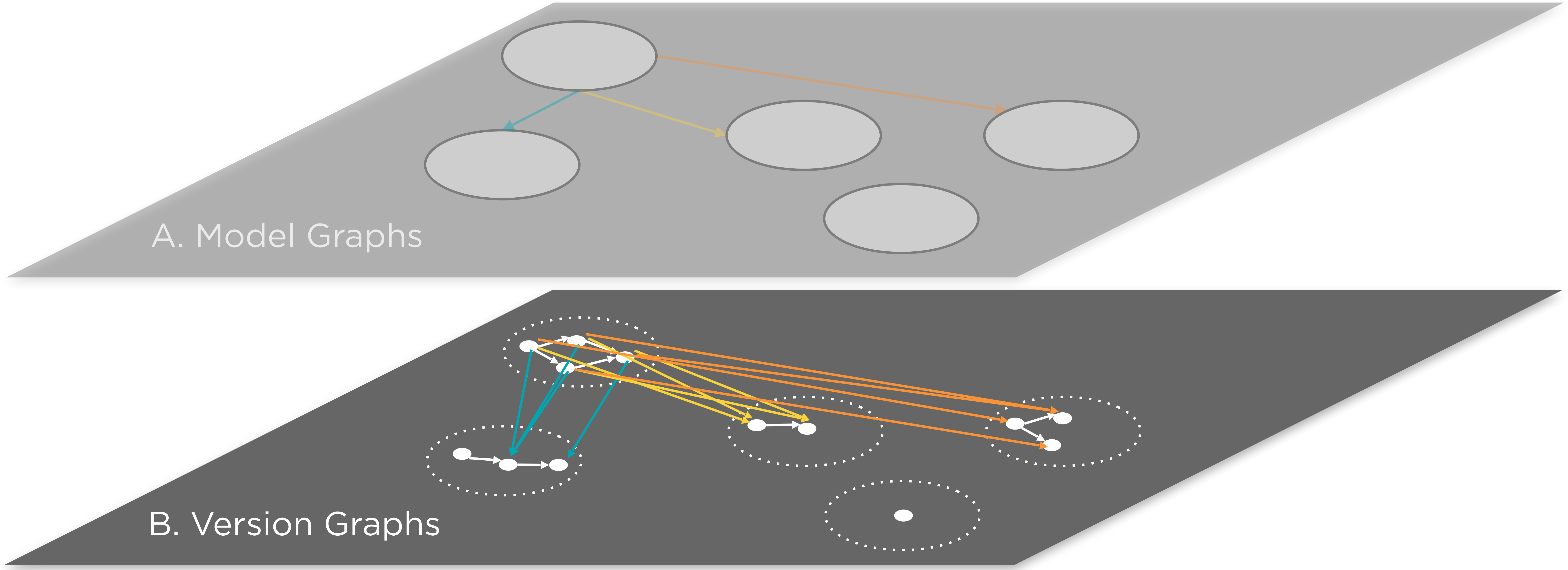
COMMON GROUND

The versioning model



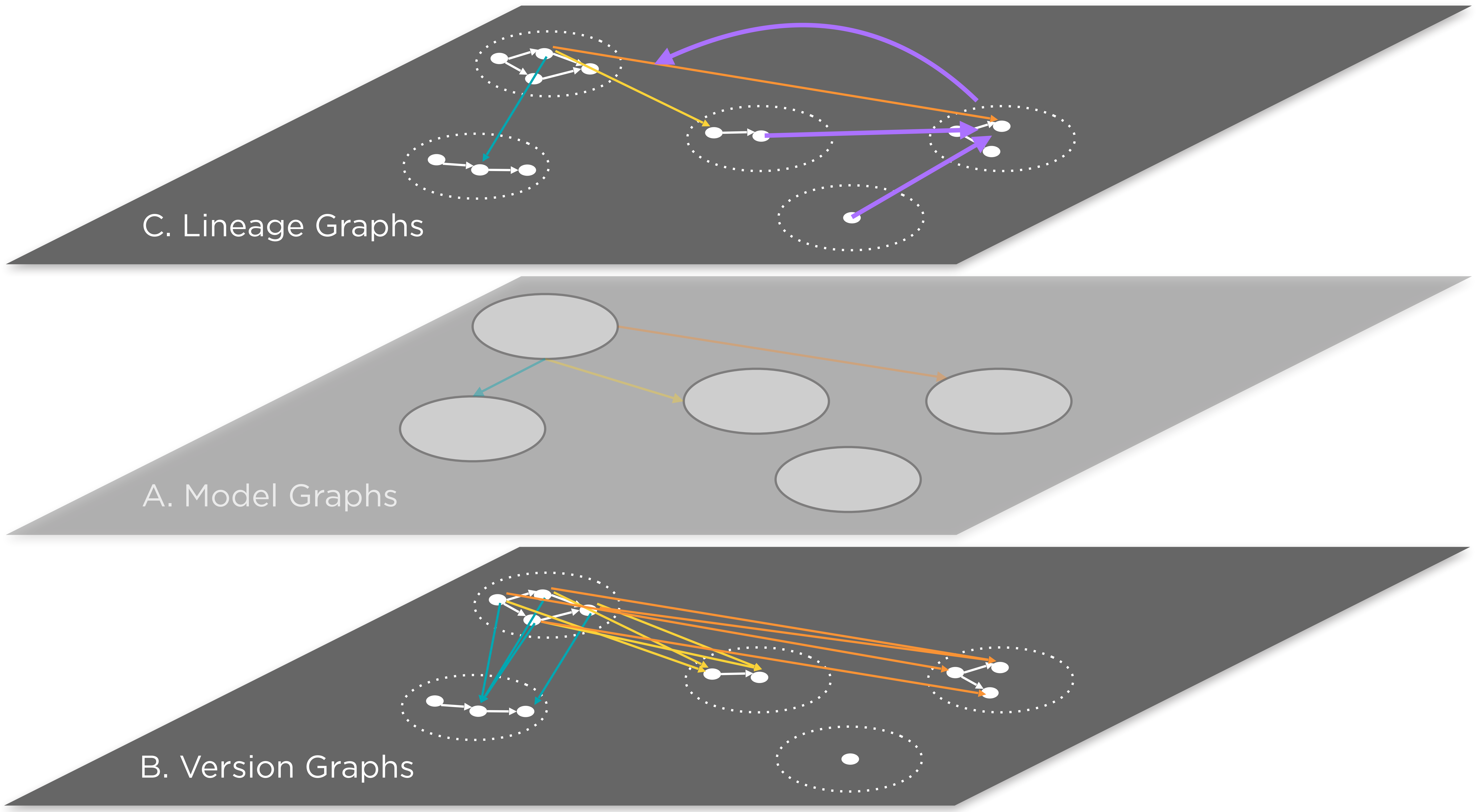
COMMON GROUND

The versioning model



COMMON GROUND

The usage model



SCALABLE, IMMUTABLE BACKEND

Longstanding open problem

Workloads?

- Graph queries for metamodel traversal
- Log analysis queries for usage

Room for improvement

- Goal: compete with in-memory performance ("the McSherry baseline")

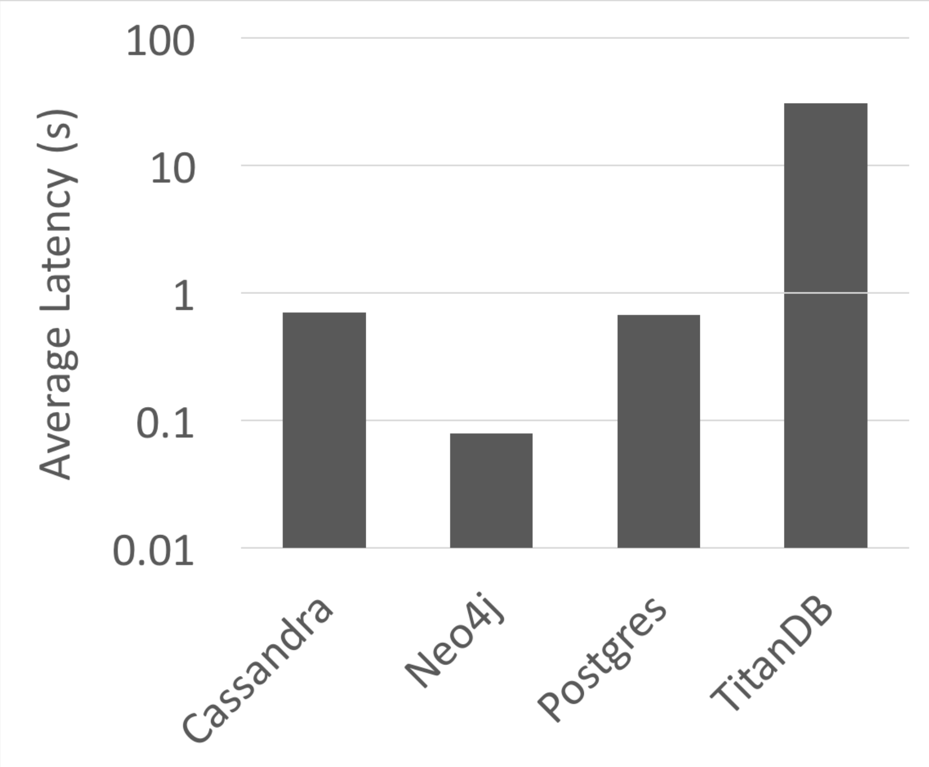


Figure 8: Dwell time analysis.

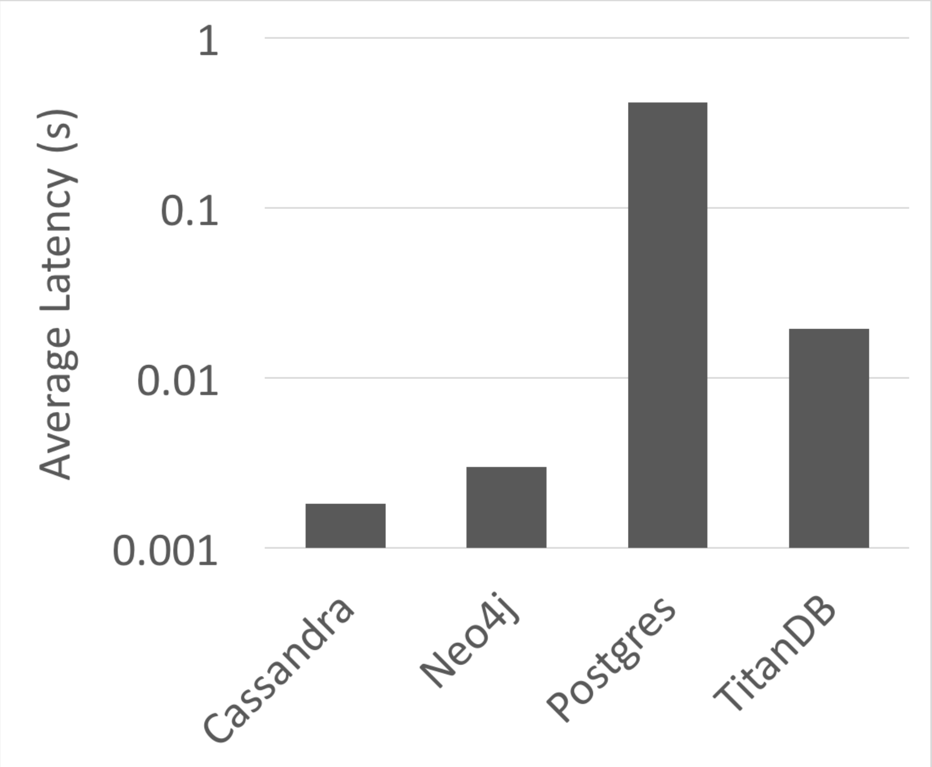


Figure 9: Impact analysis.

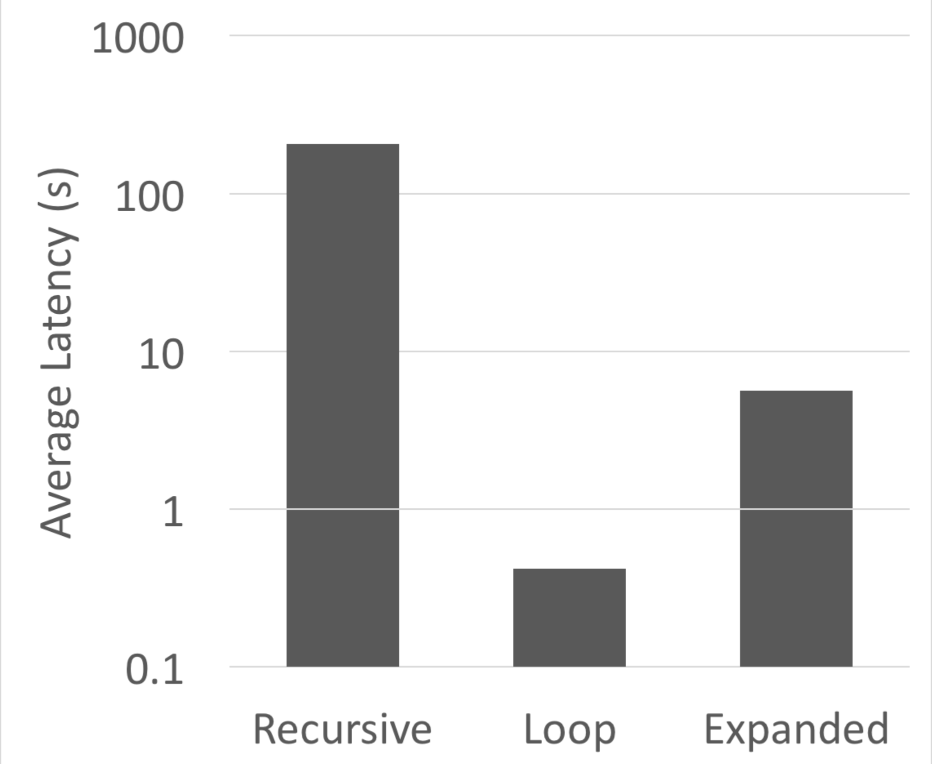


Figure 10: PostgreSQL transitive closure variants.

NEUTRALITY

Reminder:

sas
QlikView
tableau
pentaho
Microsoft
TIBCO
Spotfire
H2O
Paxata
Datameer
TRIFACTA
10 COGNOS
ORACLE
Bizzscore
ORACLE Hyperion
SAP BusinessObjects
WATERLINE DATA SCIENCE
tamr
JASPERSOFT
InetSoft
looker
MicroStrategy
Best In Business Intelligence
board
Yellowfin
Information Builders
SAP
COGNOS
AN IBM COMPANY
ZOOMDATA
atscale
Alation

There will be *k* competing solutions for:

- Data wrangling
- Data cataloging
- Schema extraction
- Feature extraction
- Social network analysis
- Etc.
- This will consolidate somewhat, but only over time

Goal: foster the ecosystem

NEUTRALITY

THE UNIVERSITY OF CALIFORNIA 1868

awake networks

CapitalOne®

cloudera®

DATAGUISE

LinkedIn

skyhigh

TRIFACTA

YOU

MANY 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

CURRENT STATUS

Alpha Release

- Integrated with LinkedIn Gobblin, Kafka, Hive Metastore, Github
- All components have Docker images on DockerHub
- We'd love feedback!

ground

www.ground-context.org