Chad Sanderson

I've worked as an Experimentation Product Manager at some of the world's biggest companies, both in and outside of technology.

- 2x Junior Olympian
- Co-produced a reality TV show
- Eppo Advisor!
- https://www.linkedin.com/in/chad-sanderson/

CONVOY Microsoft

SUBWAY SEPHORA

CONVOY

Experimentation @ Scale

Convoy

Founded in 2015

World's first digital freight brokerage

$400M Series D

~1000 employees

~50,000 active carriers per month

Experimentation is a core part of business - Founded by Amazon employees!

Conducted an evaluation of multiple 3rd party experimentation tools in 2018

Existing Experimentation Platforms cater to a very specific type of customer: Extremely high-traffic websites using exclusively clickstream events.

~~How can we run A/B Tests?~~ ***How can we measure important things?***

Experimentation Challenges
Use Cases

Convoy is a ML based B2B Marketplace with Small customer volume on both sides.

Problem:

Convoy wants to be able to measure changes inside and outside the product

What we needed:

The ability to assign various entities into the experiment and analyze impacts:

- *Modifications to pricing algorithm*
- *Randomizing on geographies*
- *Measuring Ops Efficiency*

Experimentation Challenges
Use Cases

CONVOY

Small sample size, two-sided B2B marketplace (~50K Carriers, ~1k shippers)

ML-centric. Many changes to pricing models targeting non-user based entities (Shipments)

Ops Efficiency is a major improvement vector. Requires offline analysis and manual intervention

Microsoft

Many product surfaces (Office, Bing, Teams, Xbox, Store) with huge sample size

ML-centric in some cases (Bing) Product in others (Xbox) and Marketing in others (Store). Wide variety of entity types

Safety was a core priority! Experimentation was used to determine if things were breaking

SUBWAY

Small sample size online, massive sample size in brick-and-mortar locations

No Machine Learning at all. 100% driven by marketing use cases: promotions, upsells, and loyalty

The main focus was finding the most effective selling messaging for deals and optimizing in-store behavior

Experimentation Challenges
Metrics

Convoy is a ML based B2B Marketplace with Small customer volume on both sides.

Problem:

Convoy’s primary success metrics are financial and growth based.

What we needed:

The ability to create metrics based on Data Warehouse queries:

- *Margin*
- *Variable Cost per Shipment*
- *Price relative to the market*

Experimentation Challenges
Metrics

CONVOY

Carrier Experience: Bid Intents, Batching Frequency, On Time Pick Up, On Time Delivery, Total Moving Minutes

Shipments: Total Margin, Variable Cost per shipment, Completed shipments, Layover time, Detention

Shipper Experience: Inbound Emails, Shipper Quotes, Inbound Calls, Escalations

Microsoft

Bing: Ads Clicked, Result Relevance, Ad Revenue Generated, Search Result Latency

Office: Documents Saved, Documents Created, Documented Continued, Application Crashes, Edit Frequency

Store: Number of Web Conversions, Website Registration, Purchase Volume

SUBWAY

Marketing: App/Website Accounts Created, Revenue, Purchase Frequency, Loyalty Sign-Up

In-Store: Menu Mix per Store, Revenue by Region, Foot traffic, Variable cost per store

Experimentation Challenges
Organization

Convoy is a ML based B2B Marketplace with Small customer volume on both sides.

Problem:

Product Strategy and Career Growth depends on experiment outcomes, which necessitates high trustworthiness.

What we needed:

Self-serve experiment deployment and analysis at scale, which validated methods for all product and non-product teams.

- *Product Teams*
- *Ops Organizations*
- *Marketing and Sales*

Experimentation Challenges
Organization

CONVOY

Convoy has an extremely high data-specialist to software engineer ratio due to the analytical complexity of freight

Data scientists needed the tools to move quickly, independently, and uniformly to run experiments

Quarterly business results were driven by experimentation outcomes. This required trustworthy results

Microsoft

Microsoft has many product organizations at different stages of science maturity

A central experimentation team was required to educate, onboard, and serve as a customer-success division for less mature organizations

Some teams wholly adopted experimentation, but others saw it as a barrier to product development

SUBWAY

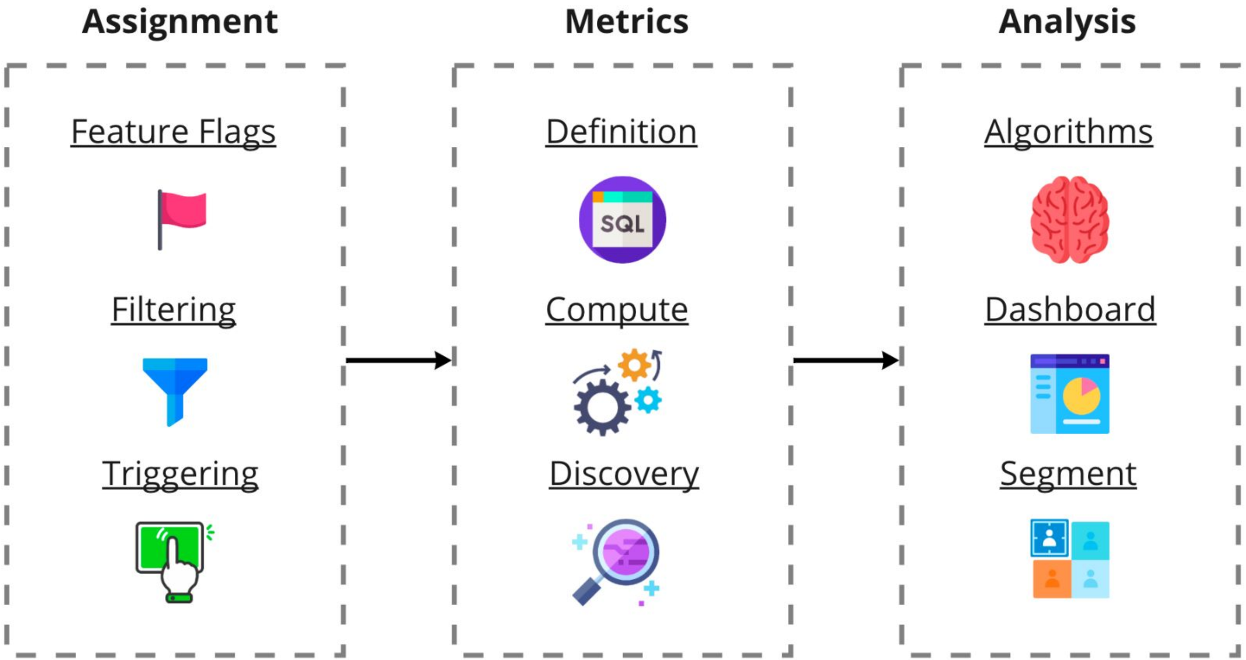
Experimentation was introduced to Subway as a marketing strategy and was highly leveraged by the web/app teams

Had almost no data scientists working in product, most concentrated on in-store analytics

A central CRO team designed, developed, and analyzed all experiments.

Product Needs

An Ideal Stack



"Serving the right experience at the right time, to the right entity."

"Creating and monitoring the experiment success KPIs"

"Determining the outcome of experiments"

Product Needs

An Ideal Stack

Assignment

Carriers

Shippers

Trucks

RFPs

Shipments

Lanes

Facilities

...

"Serving the right experience at the right time, to the right entity."

Metrics

Click Events

Margin

Live Ops Cost

Shipment Duration

Service Failure Cost

Ops Touches

On-Time Rate

"Creating and monitoring the experiment success KPIs"

Analysis

T-Test (Freq)

T-Test (Bayes)

Diff-in-diff

Causal Impact

Synthetic Control

Downstream Impact

...

"Determining the outcome of experiments"

Product Needs

What Current 3rd Party Tools Provide (Clickstream Events & User Randomization)

Assignment

- Carriers
- Shippers
- ~~Trucks~~
- ~~RFPs~~
- ~~Shipments~~
- ~~Lanes~~
- ~~Facilities~~
- ...

"Serving the right experience at the right time, to the right entity."

Metrics

- Click Events
- ~~Margin~~
- ~~Live Ops Cost~~
- ~~Shipment Duration~~
- ~~Service Failure Cost~~
- ~~Ops Touches~~
- ~~On-Time Rate~~

"Creating and monitoring the experiment success KPIs"

Analysis

- T-Test (Freq)
- T-Test (Bayes)
- ~~Diff-in-diff~~
- ~~Causal Impact~~
- ~~Synthetic Control~~
- ~~Downstream Impact~~
- ...

"Determining the outcome of experiments"

Problem:

What we observed was only the tip of the ‘use case iceberg.’ Existing platforms would not satisfy these requirements.

What we needed:

A flexible platform that let us randomize on any entity, use any query as a metric, and perform any analysis we wanted

The Solution

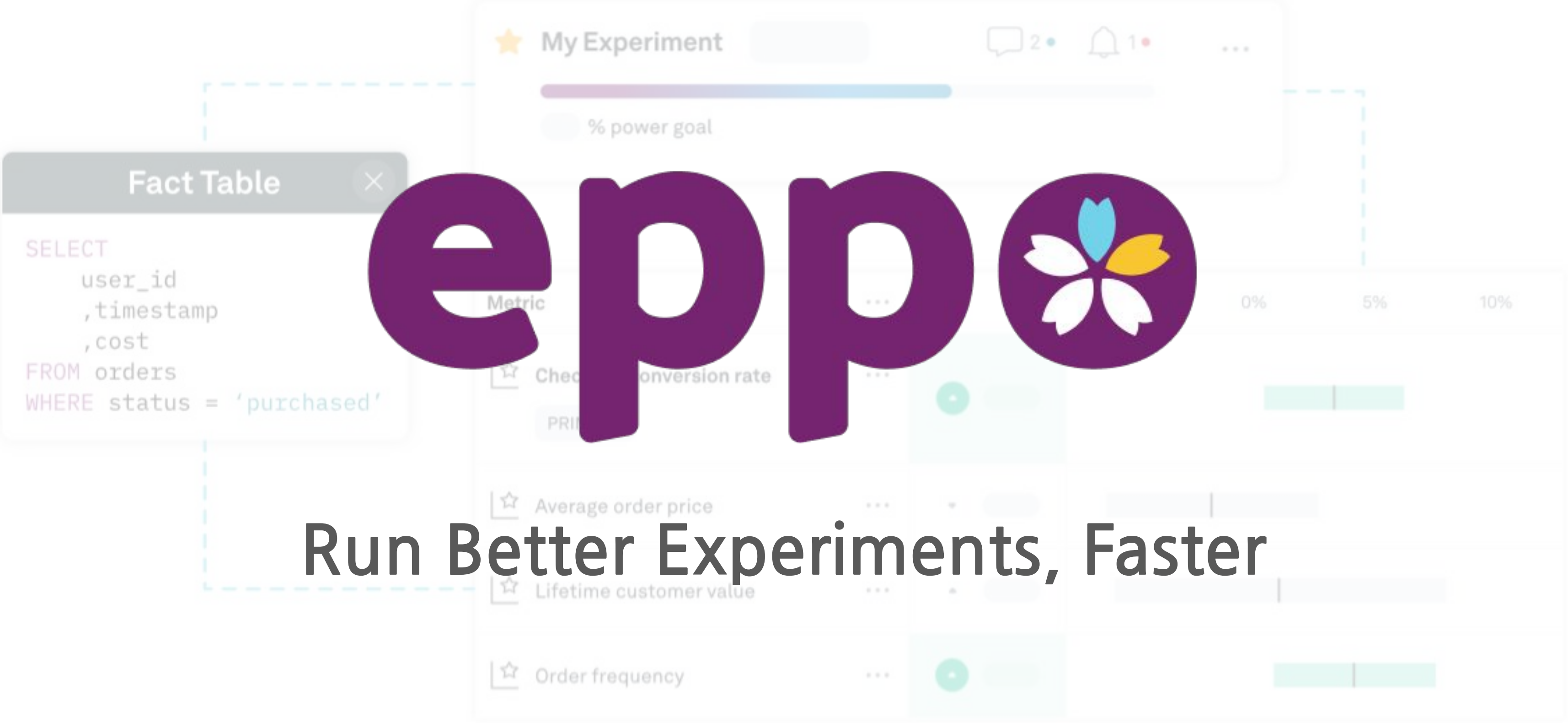
Time to build it! *(4 Engineers x Years of features and hard learnings)*

What we needed:

A flexible platform that let us randomize on any entity, use any query as a metric, and perform any analysis we wanted

Structure your experimentation system around ***use cases and business needs.***

Don't just follow the crowd! Limiting your use cases by selecting the wrong tool can severely impact the value and adoption of experimentation.



eppo

Run Better Experiments, Faster

01 TEAM
Leadership

Chetan Sharma
Chief Executive Officer

- 4th data scientist at Airbnb
- Data scientist at Webflow, Next Trucking
- Consulted at many growth stage companies

Carlin Eng
Head of Data Engineering

- Sales Engineer at Snowflake
- Head of Data Engineering at Strava
- Stanford Alumnus (BA Economics; MS Statistics)

02 EXPERIMENTATION
Then & Now

Experimentation used to be only at mega tech companies

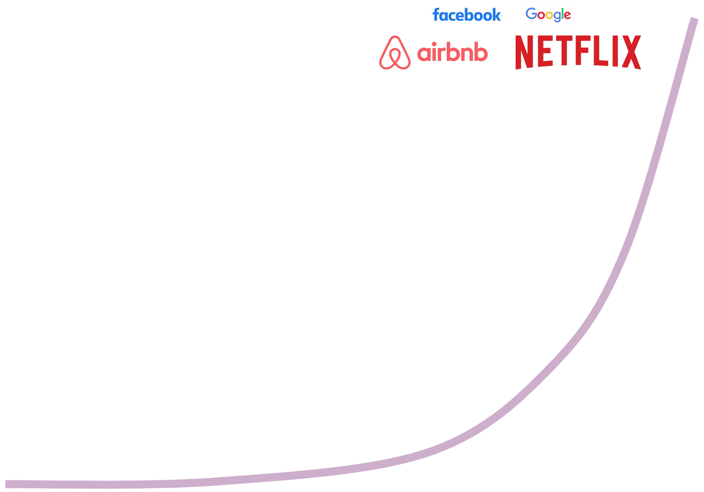
Technology

hadoop

Apache Pig

HIVE

Companies



02 EXPERIMENTATION
Then & Now

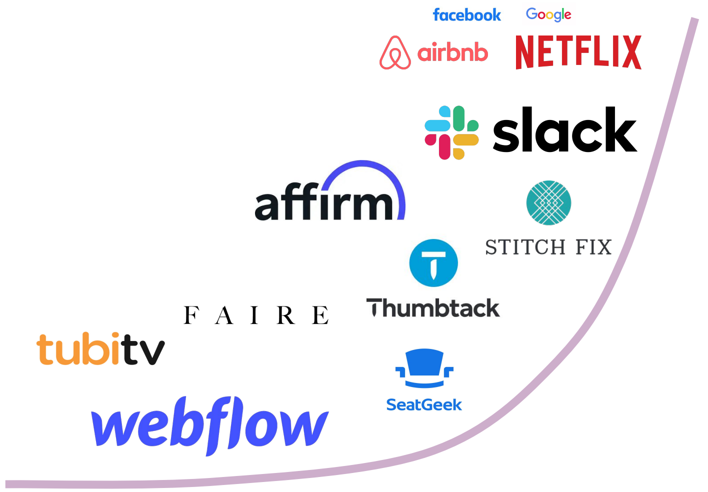
Now, companies run experiments early and often

Technology

dbt
Apache Airflow

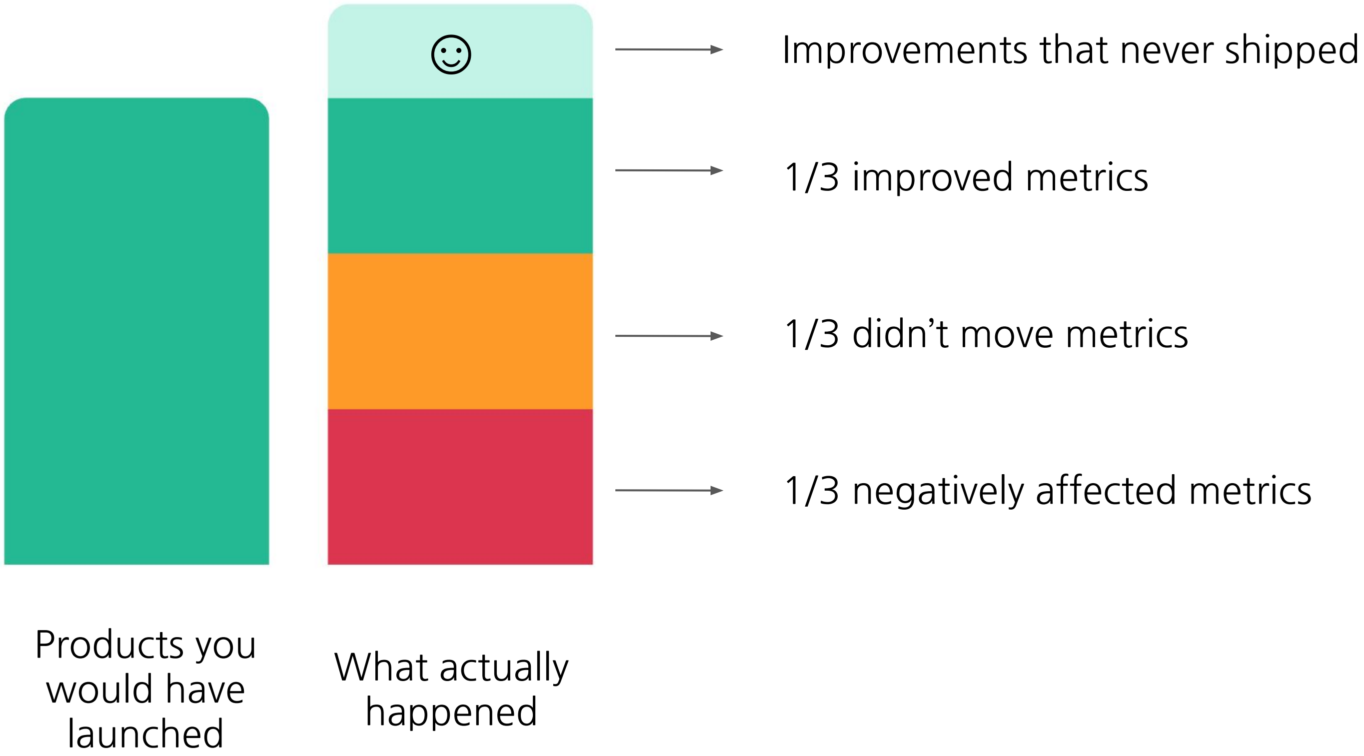
snowflake
amazon REDSHIFT
BigQuery

Companies



02 **EXPERIMENTATION**

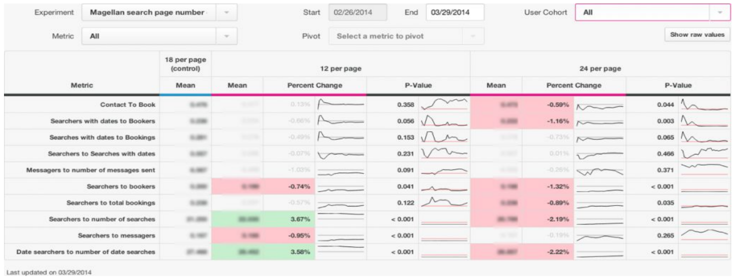
Why Conduct Experiments?



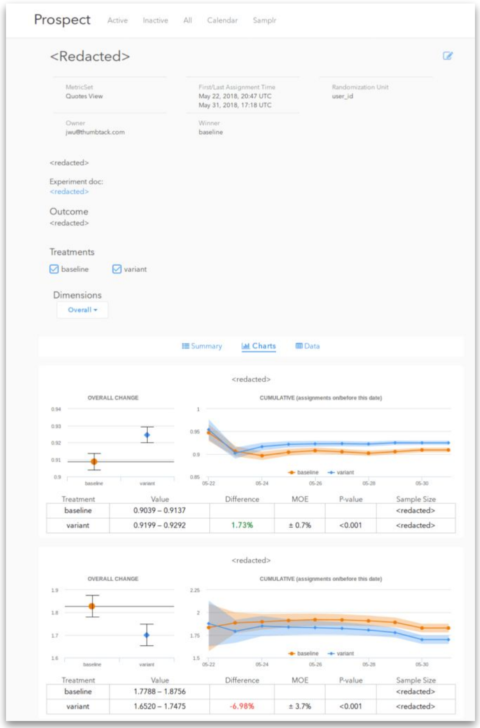
02 EXPERIMENTATION
Experimentation Culture

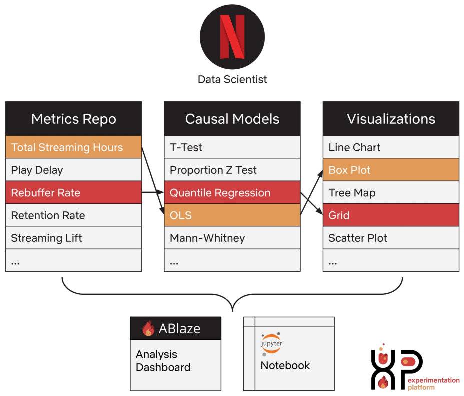
Experiment culture requires dedicated applications.

airbnb



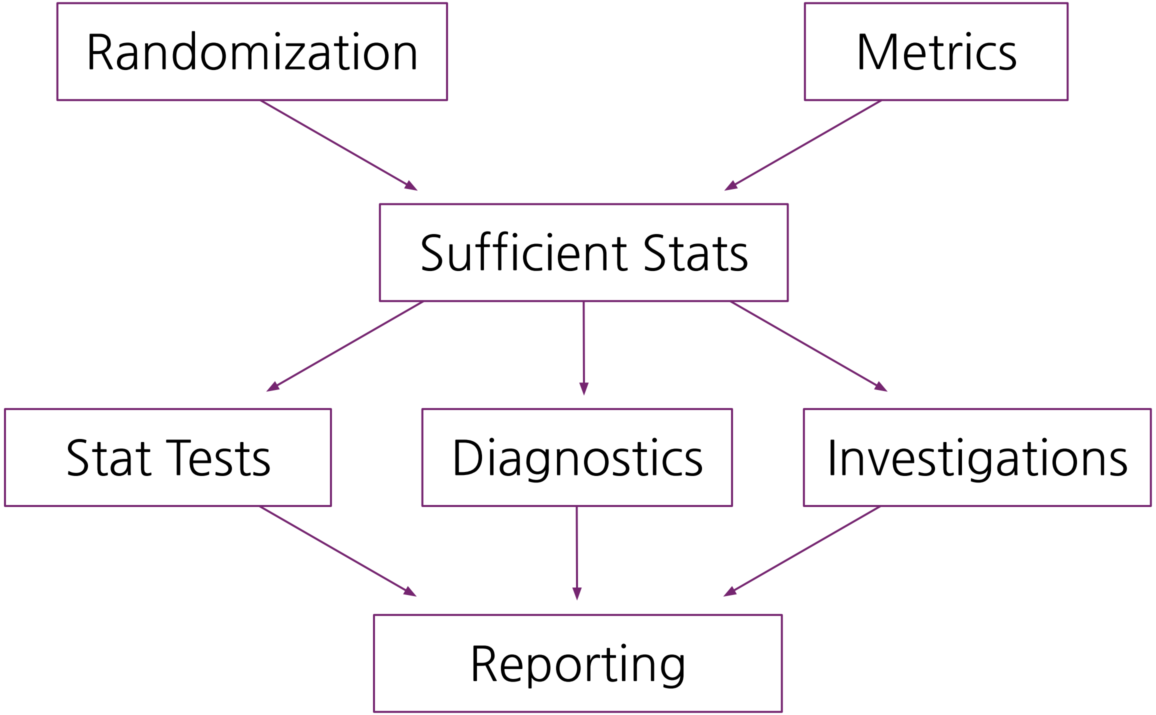
Thumbtack





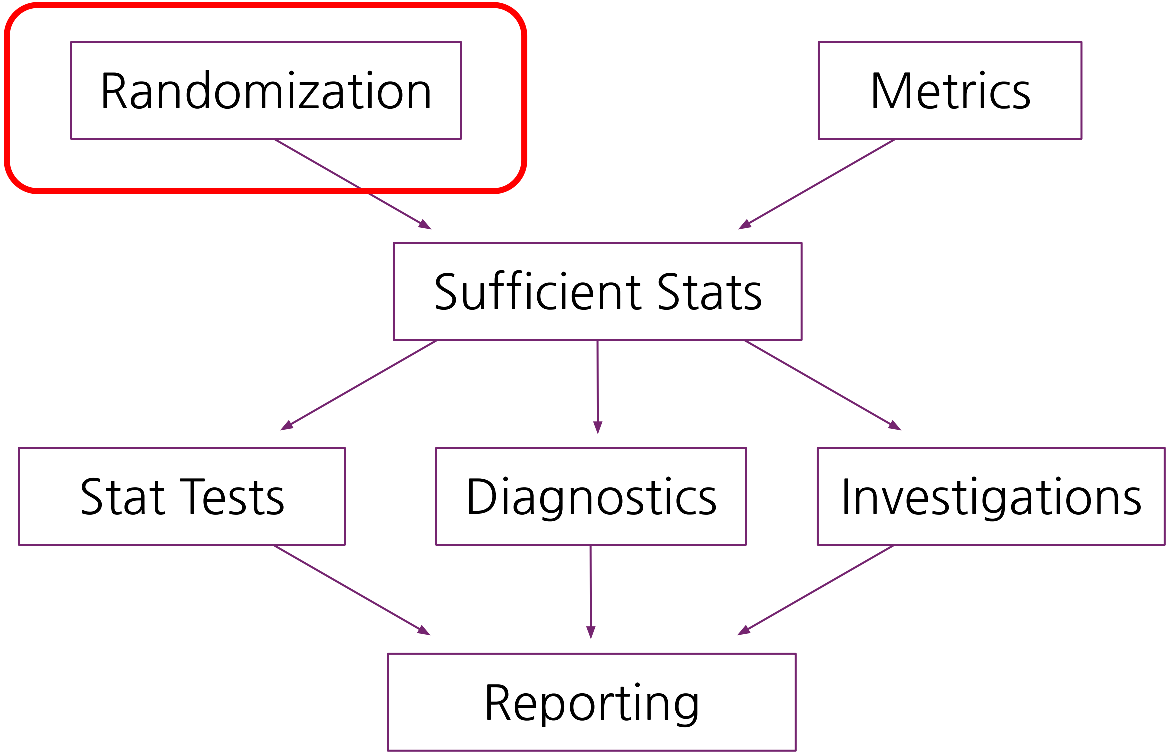
03 ARCHITECTURE
Building Blocks of Experimentation

Every experimentation system has the same architecture.



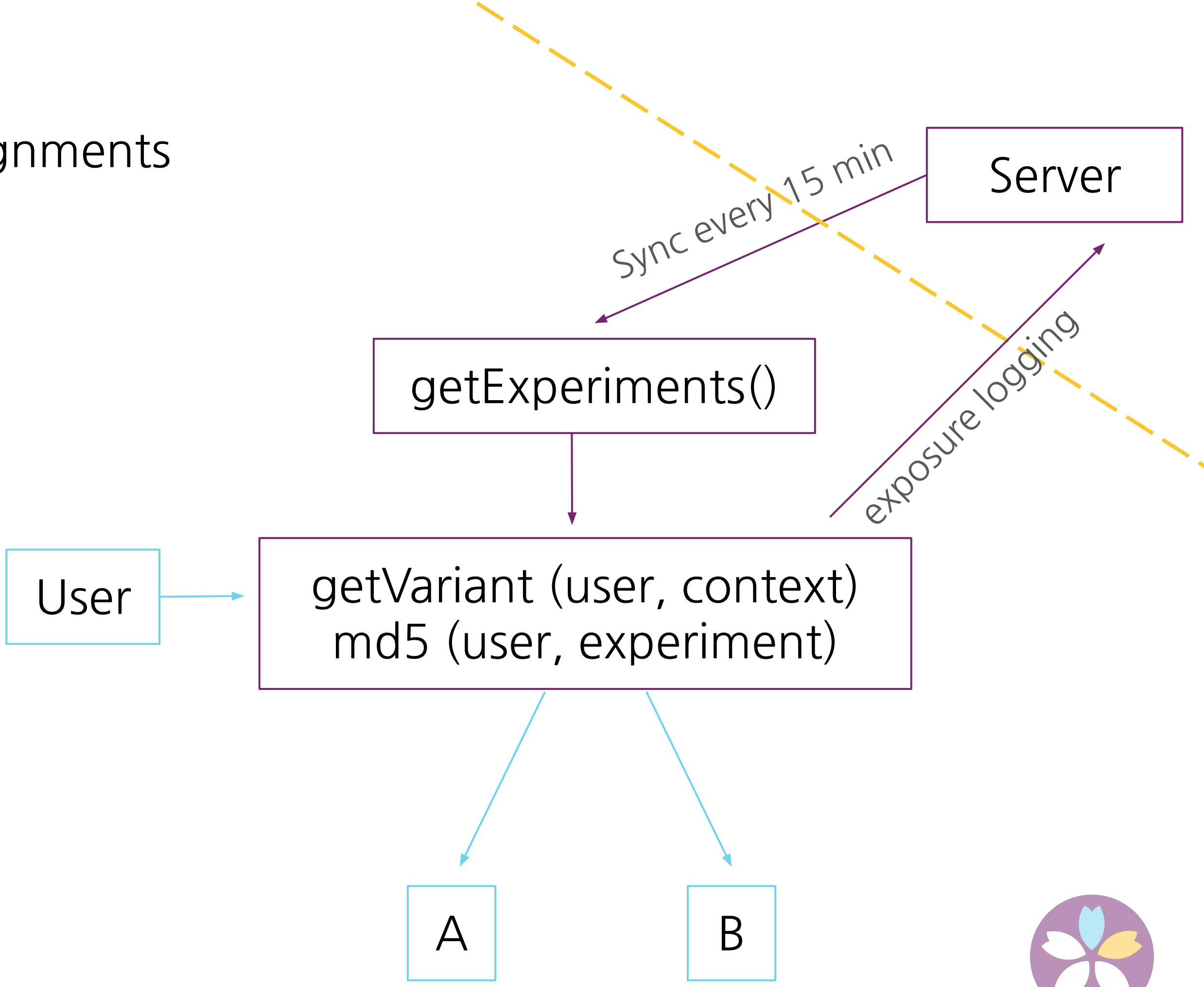
03 **ARCHITECTURE**

Randomization / Assignments

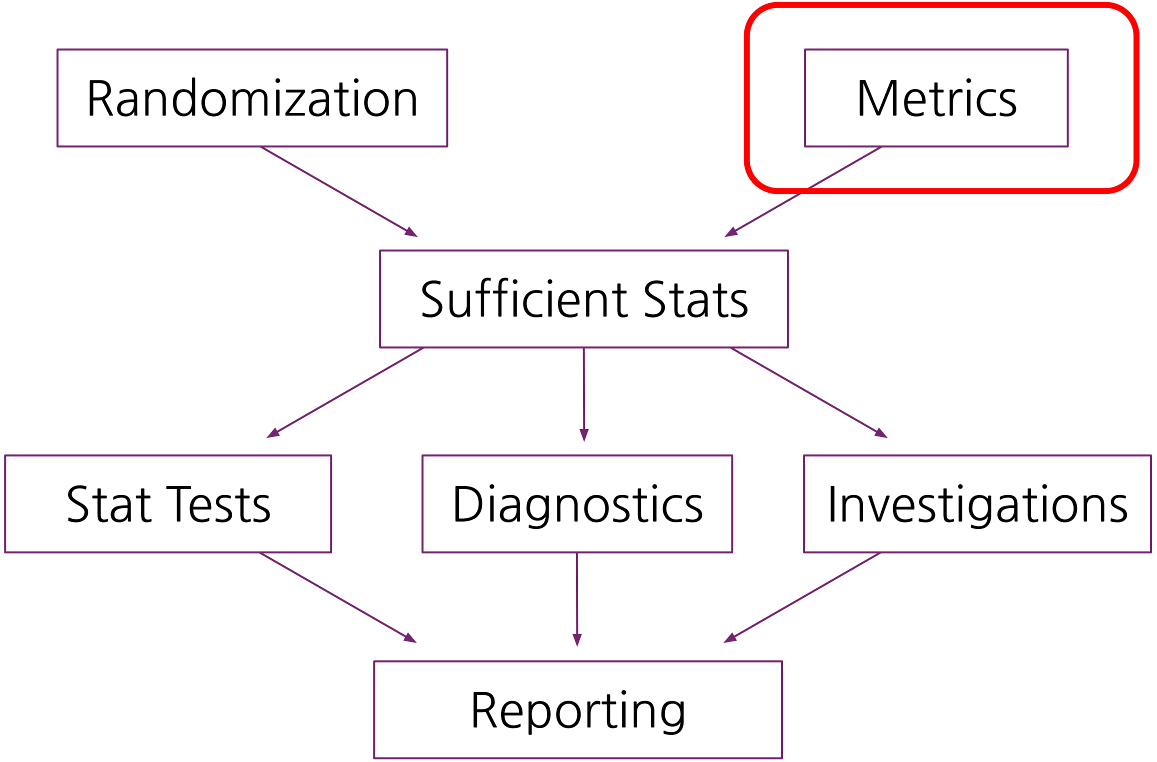


Pro Tip #1:

Use md5() hashes for assignment



03 ARCHITECTURE
Metrics



03 **ARCHITECTURE**
Metrics

Use metrics that matter!

The biggest gap between Airbnb / Netflix/et al. and commercial tools is how easily you can use **business metrics**.

Business Metrics

- Revenue, Activation, Purchases
- What the CFO reads
- From databases, Stripe, multiple POS

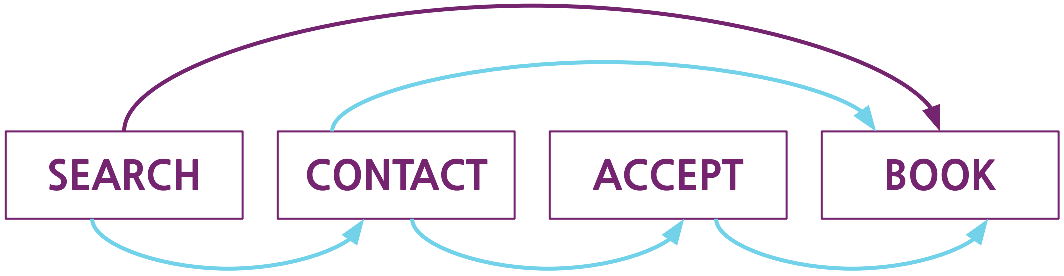
Shallow Metrics

- Signups, “conversions”
- “Directionally accurate”
- From event streams

03 ARCHITECTURE
Metrics

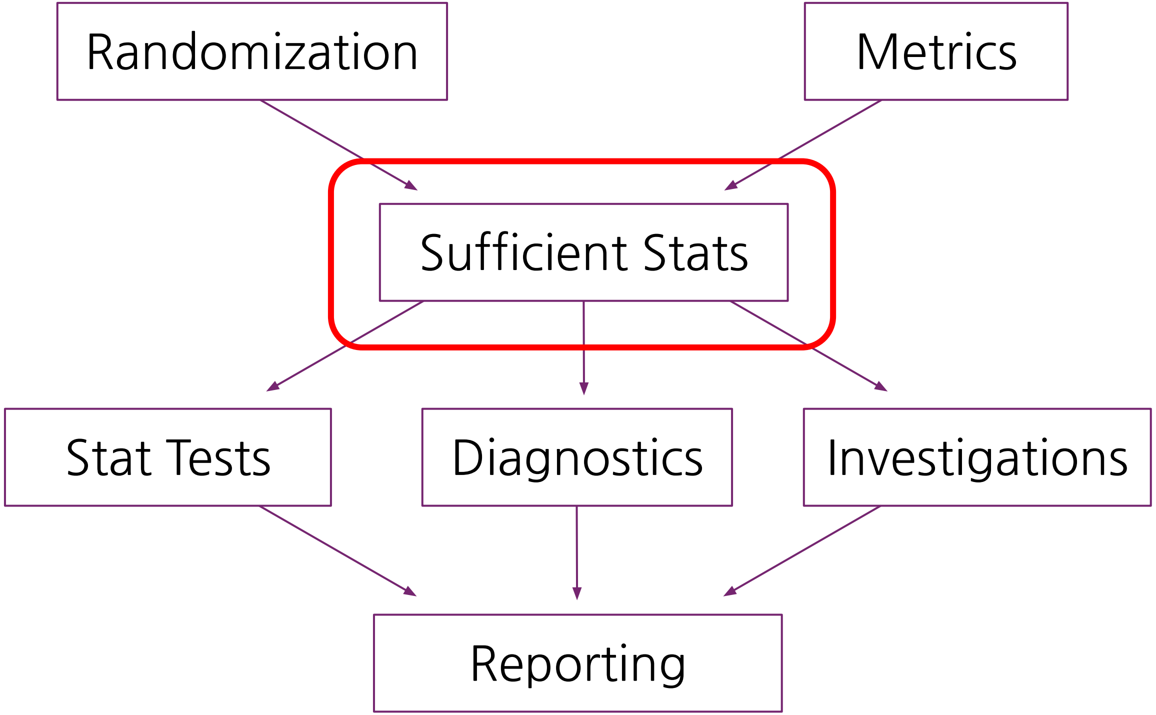
It's common to inflect one part of the funnel, move another down

Metric	Δ	p
Search to Book	-0.31%	0.37
Search to Contact	-1.29%	0.04
Contact to Book	0.99%	0.06
Contact to Accept	1.58%	0.00
Accept to Book	-0.58%	0.11



Sufficient Statistics

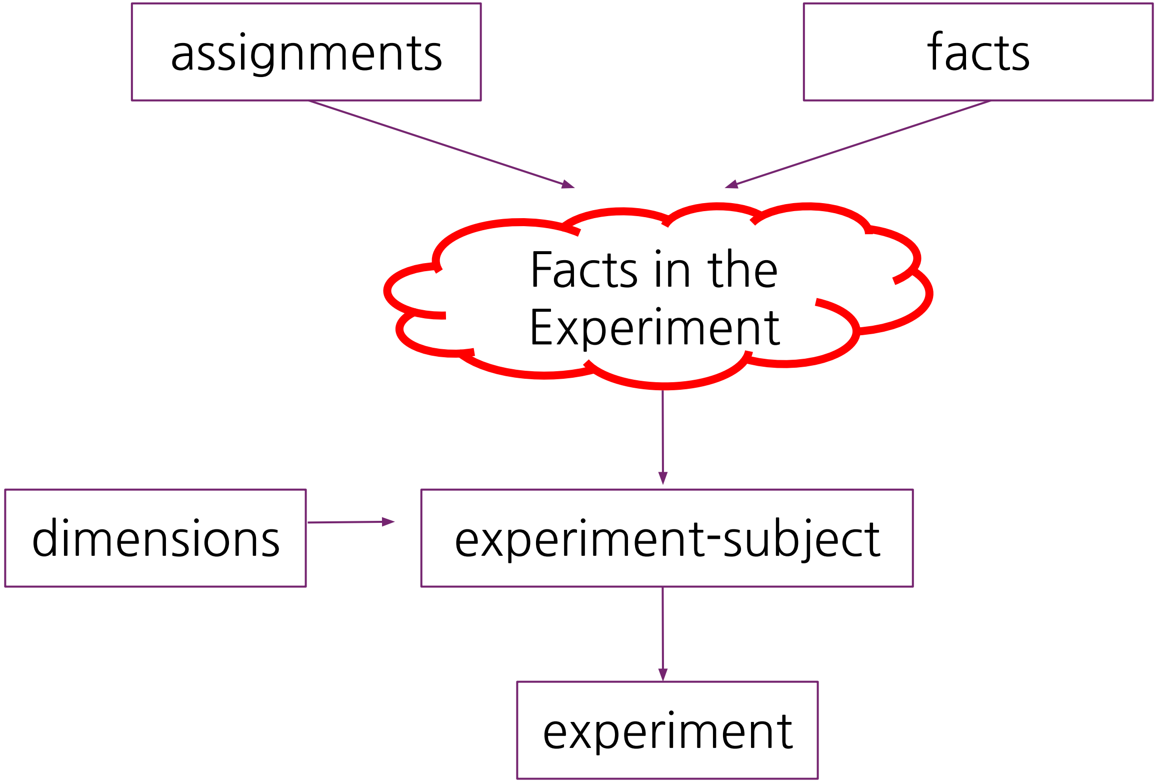
Aka the data aggregations that feed into the stats



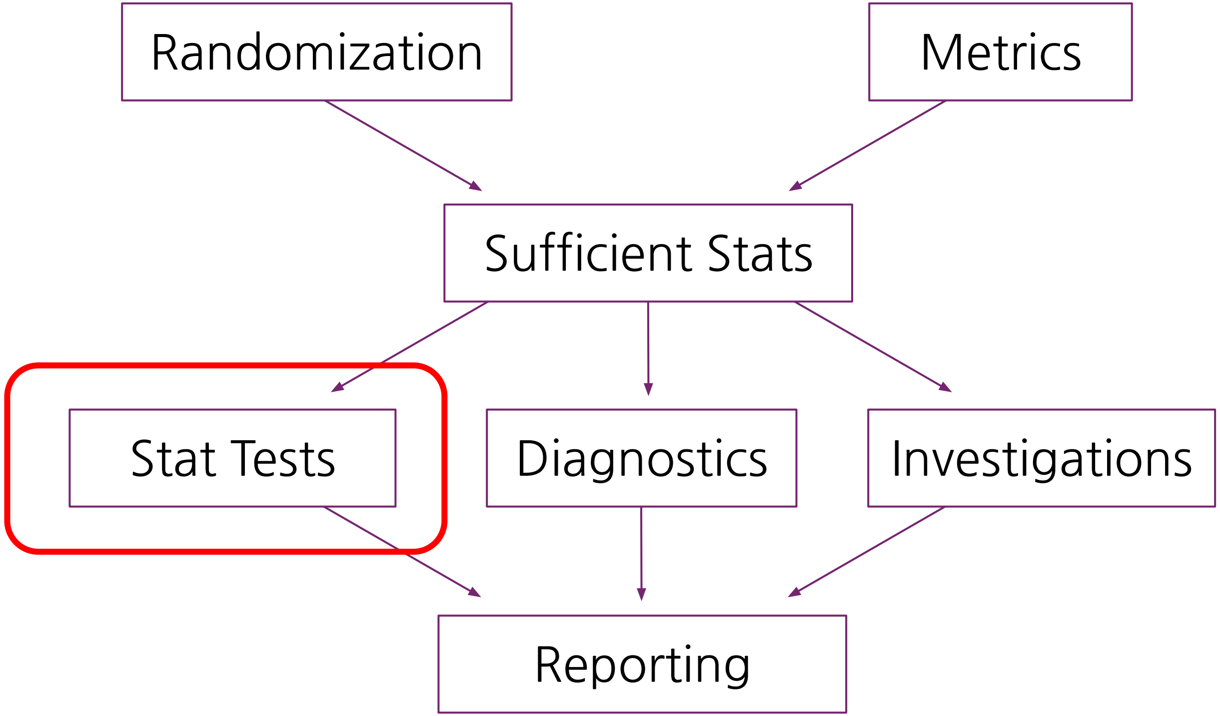
03 **ARCHITECTURE**

Sufficient Statistics

Experiment Data Pipelines: One big JOIN and some GROUP BYs



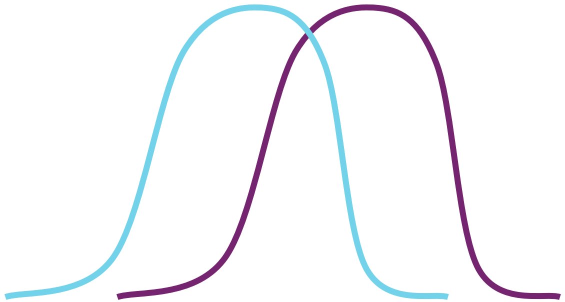
03 ARCHITECTURE
Statistical Tests



03 ARCHITECTURE
Statistical Tests

Just run a t- test,
……right?

statistics



$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma_{1,2}}$$

03 ARCHITECTURE
Statistical Tests

Simple statistical tests stress the organization.

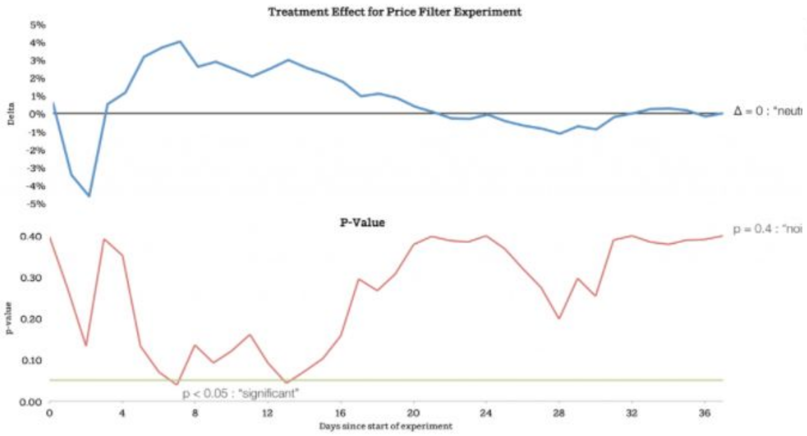
When using t-tests, you need to:

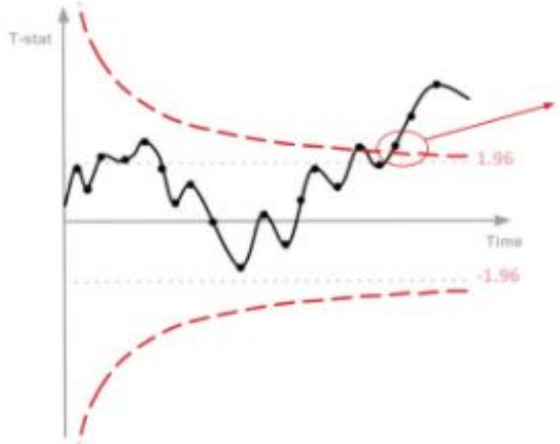
- 🚫 Not look at the results until it's done
- 👪 Not test multiple variants without a statistical correction
- ⚡ Not have outliers / power laws
- ➕ Only use sum(), count()
- ➗ Not use ratios, time_to()

03 **ARCHITECTURE**

Statistical Tests

Sequential testing prevents people from cheating





03 ARCHITECTURE
Statistical Tests

CUPED speeds up experiments

Goal:

Improve # happy meals purchased

Insight:

We can predict whether someone will purchase a happy meal

...are they a family?

...have they purchased a happy meal recently?

CUPED is just OLS, controlling for prior history

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Y = # of happy meals in experiment window

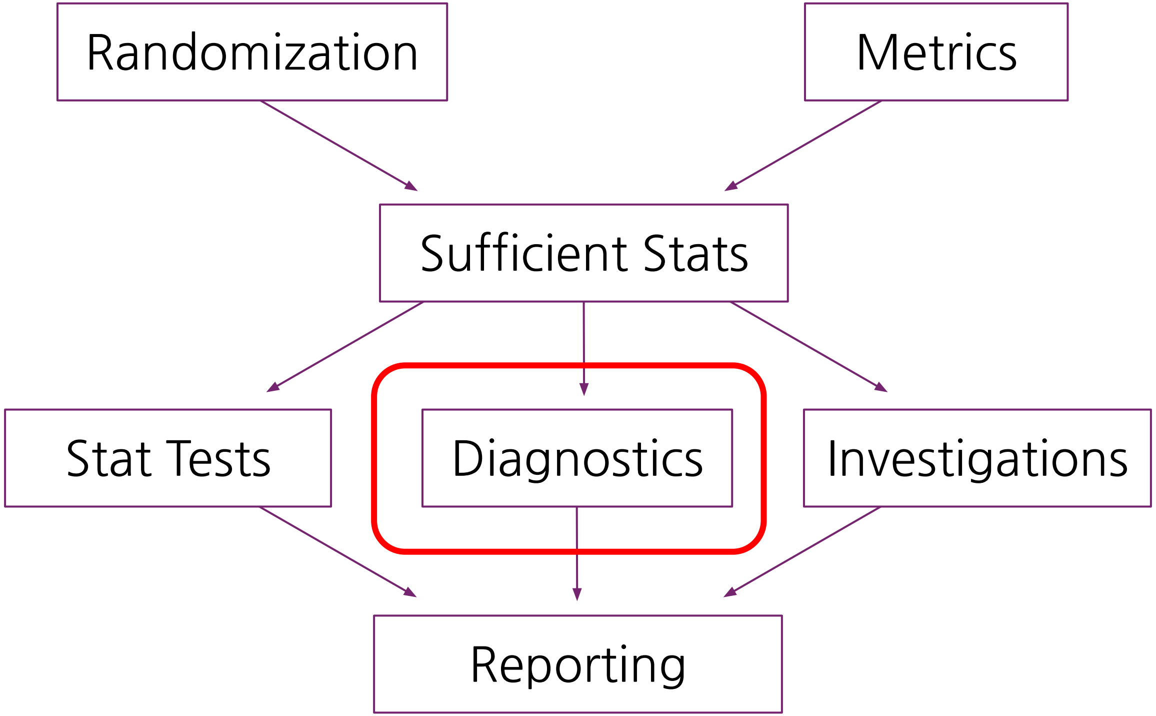
Beta1 = # of children in group

Beta2 = # of happy meals <u>in 60 days prior to experiment</u>

…

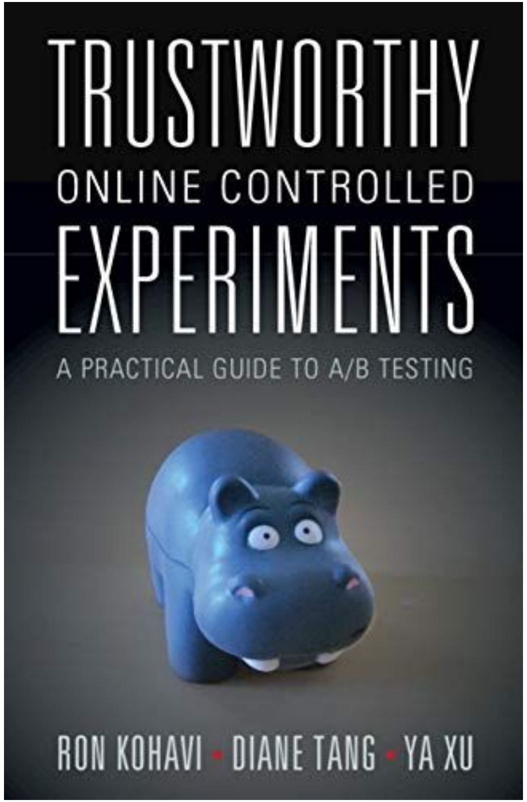
BetaN = Indicator for treatment group

03 ARCHITECTURE
Diagnostics



The first principle of experimentation:

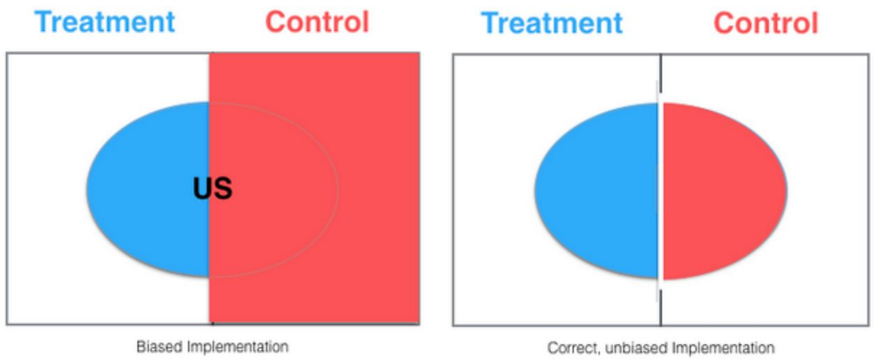
TRUST



03 ARCHITECTURE
Diagnostics

Make sure you have balanced groups!

These issues are usually due to:
- Latency of experiment delivery
- Bad implementation

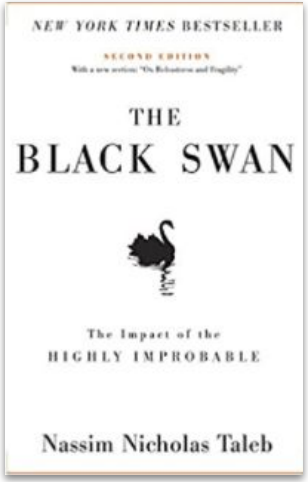


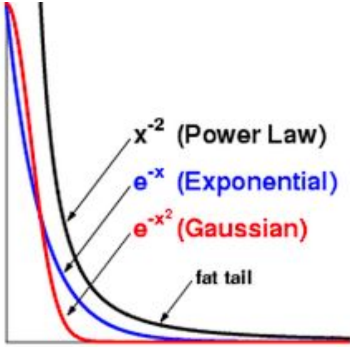
Solution: Sample ratio mismatch test (SRM)

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

03 ARCHITECTURE
Diagnostics

Watch for outliers!



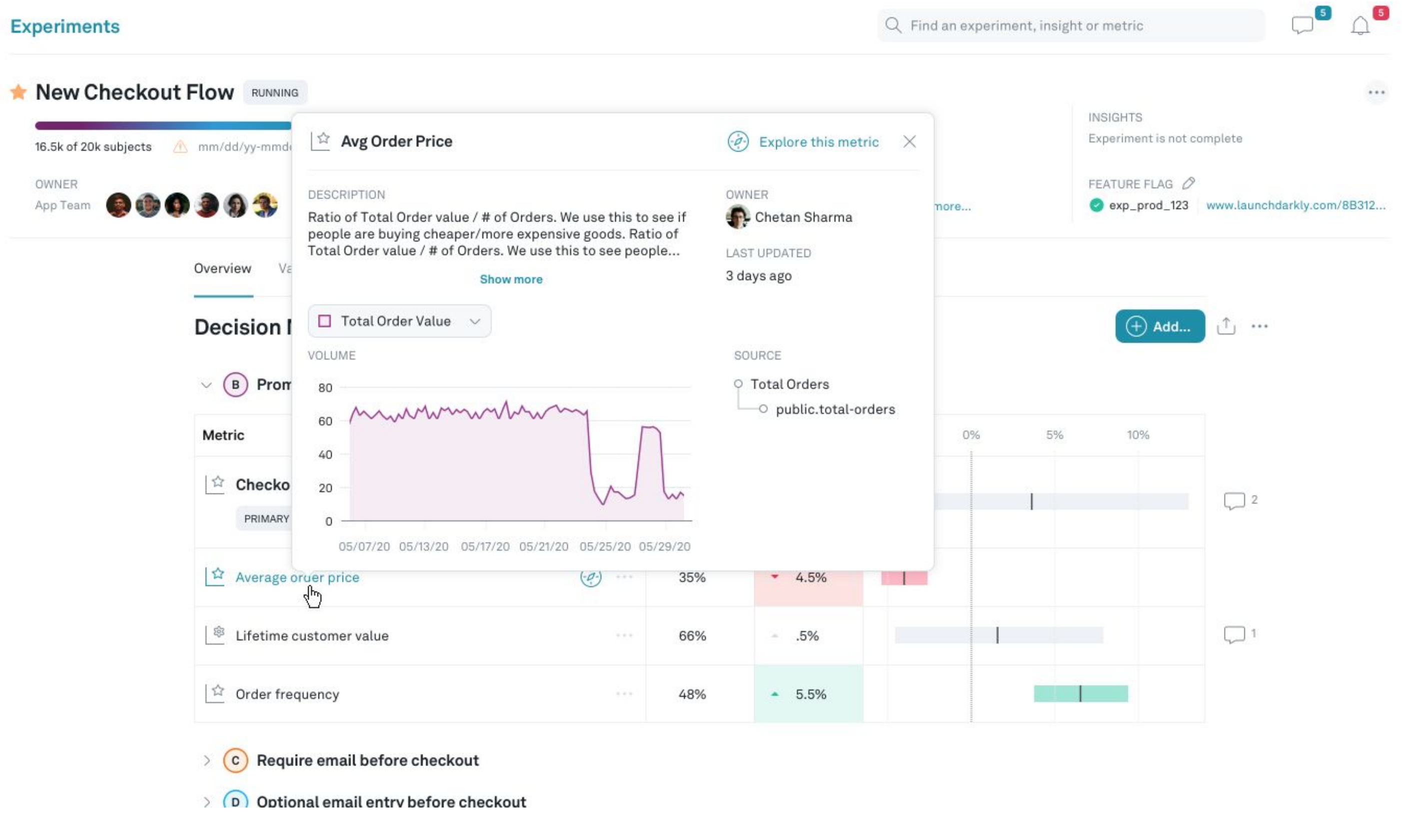


Solutions:

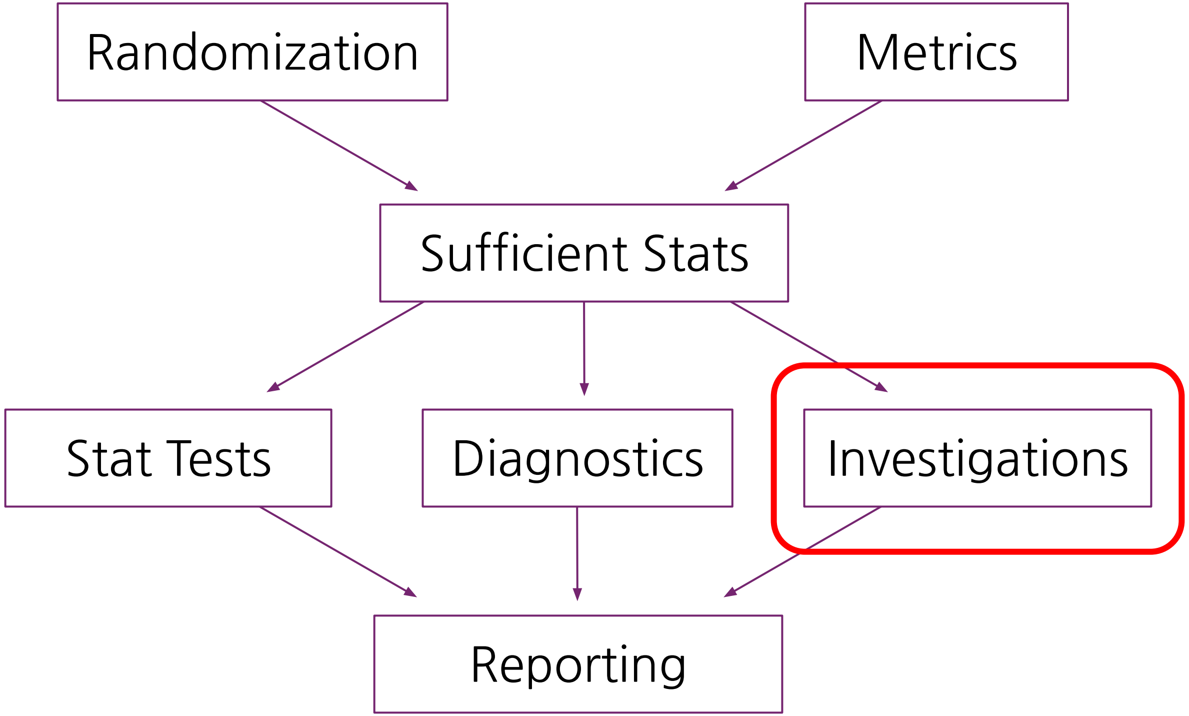
- Winsorization: cap values at 99th percentile
- CUPED

03 ARCHITECTURE
Diagnostics

Bad data becomes invisible



03 ARCHITECTURE
Investigations



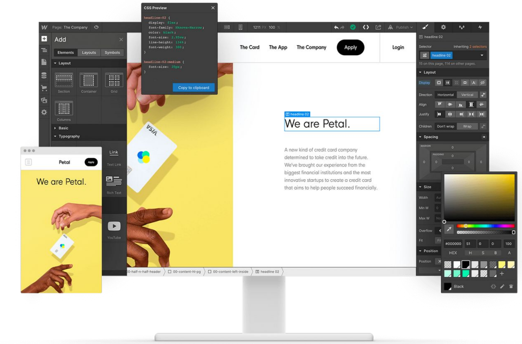
Investigations help you learn

“First you must learn to test.
Then you learn to learn.
Then you learn to win”

— *Elena Verna*
Reforge EIR, previous SVP Growth @ Survey Monkey

03 ARCHITECTURE
Investigations

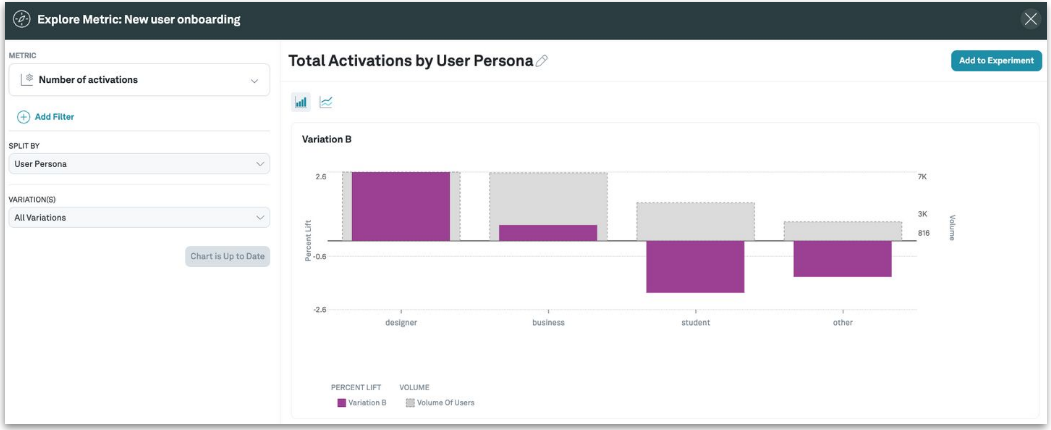
Some users might particularly hate your experiment

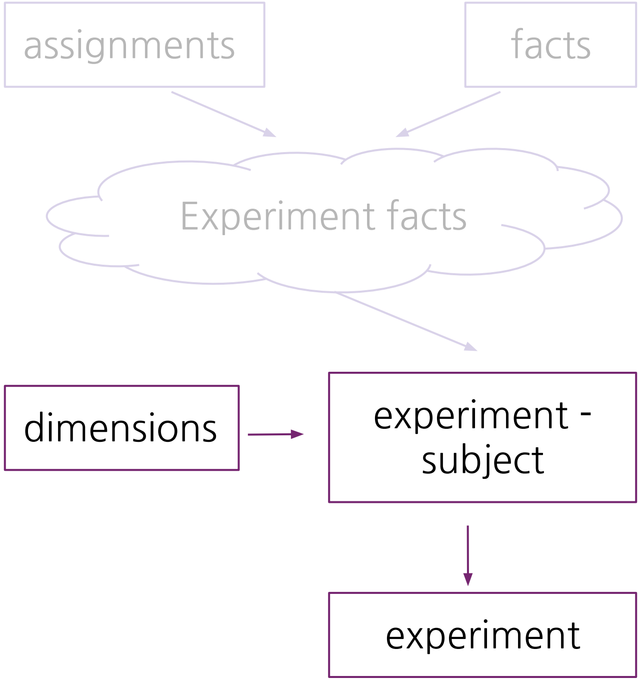


Browser	Δ	p
All	-0.27%	0.29
Chrome	2.07%	0.01
Firefox	2.81%	0.00
IE	-3.66%	0.00
Safari	0.86%	0.26
Rest	-0.74%	0.33

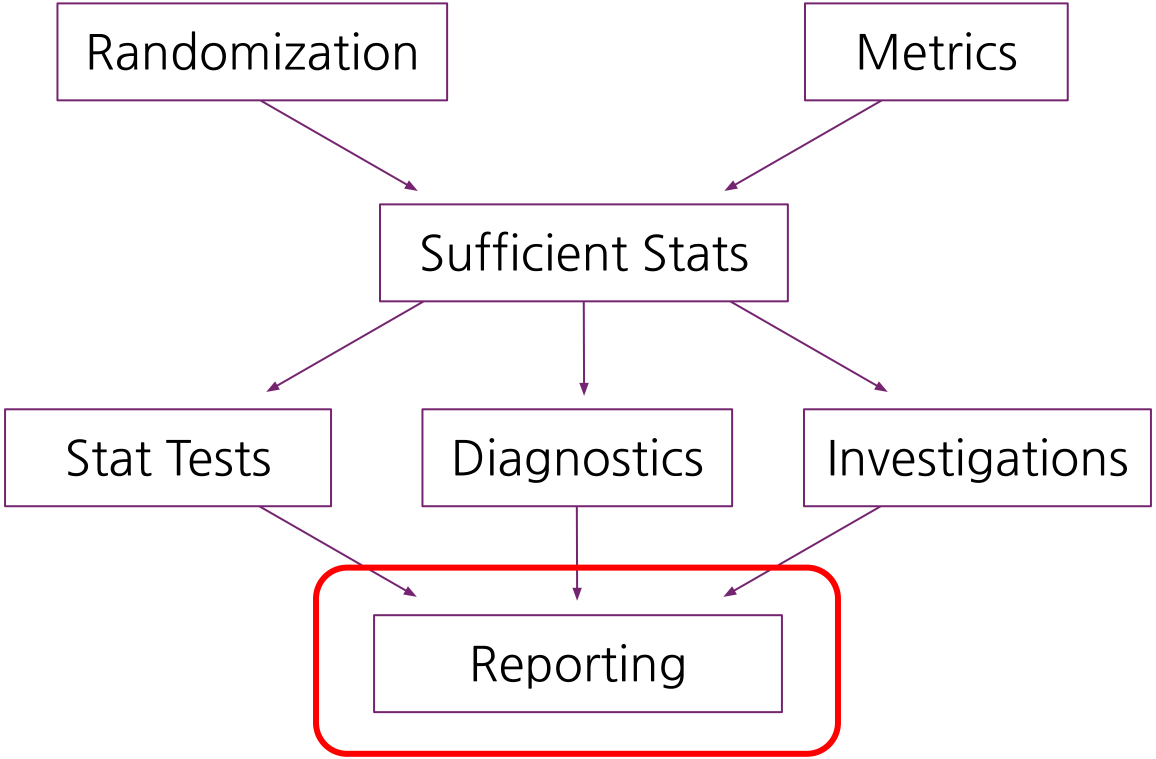
03 ARCHITECTURE
Investigations

Make slice-dice investigations easy





03 ARCHITECTURE
Reporting



03 ARCHITECTURE
Reporting

Bad reporting will undo all of your math, engineering

SET UP: PRE TEST

Sample Size

Please note that specifying the sample size should have been done **prior** to running the test.

If you prefer to use a website to calculate the sample size, I would highly recommend this one (the numbers are going to be very very close to what you see below).

If you would like more details about each of the variables used in calculating the sample size and how they impact it, please see the original A/B Testing Playbook here.

~ Set up Variables ~

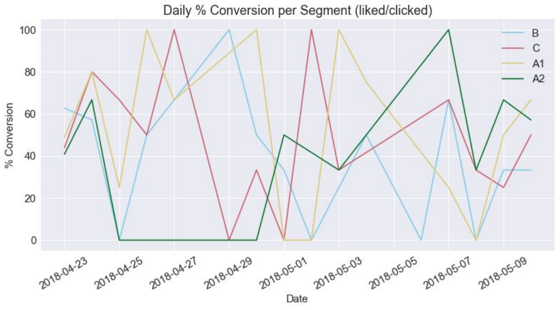
```
# Variables needed to determine the required sample size for test
baseline_probability = 0.10
beta = 0.2
alpha = 0.05
effectSize = 0.01

# Specify if you would like to run a one-sided test or two-sided test
one_sided = True
```

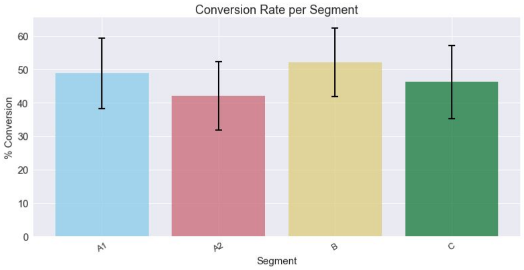
Based on the above metrics...

```
alpha_Zscore = stats.norm.ppf(1-alpha*(1-0.5*(1-one_sided)))
beta_Zscore = stats.norm.ppf(1-beta)

print('Baseline Conversion (pr) = ', baseline_probability*100, '%')
print('Confidence Level = ', (1-alpha)*100, '%')
print('Alpha Z-score (za) = ', round(alpha_Zscore,2))
print('Power = ', (1-beta)*100, '%')
print('Beta Z-score (zb) = ', round(beta_Zscore,2))
print('Effect Size (E) = ', effectSize*100, '%')
```

```
Baseline Conversion (pr) =  10.0 %
Confidence Level =  95.0 %
Alpha Z-score (za) =  1.64
```

```
# Chart Labels and Font Size
plt.xlabel('Date', fontsize=15)
plt.ylabel('% Conversion', fontsize=15)

plt.xticks(fontsize=15, rotation=30)
plt.yticks(fontsize=15)
chart_title = 'Daily % Conversion per Segment ({}/{})'.format(conversion_label, traffic_label)
plt.title(chart_title, fontsize=18)
plt.legend(fontsize=15)
plt.show()
```





IMPORTANT: Please do not make any conclusions from the above graph alone. The formulas below call out segments that are statis

FILTERS Users A B Gender is "f"

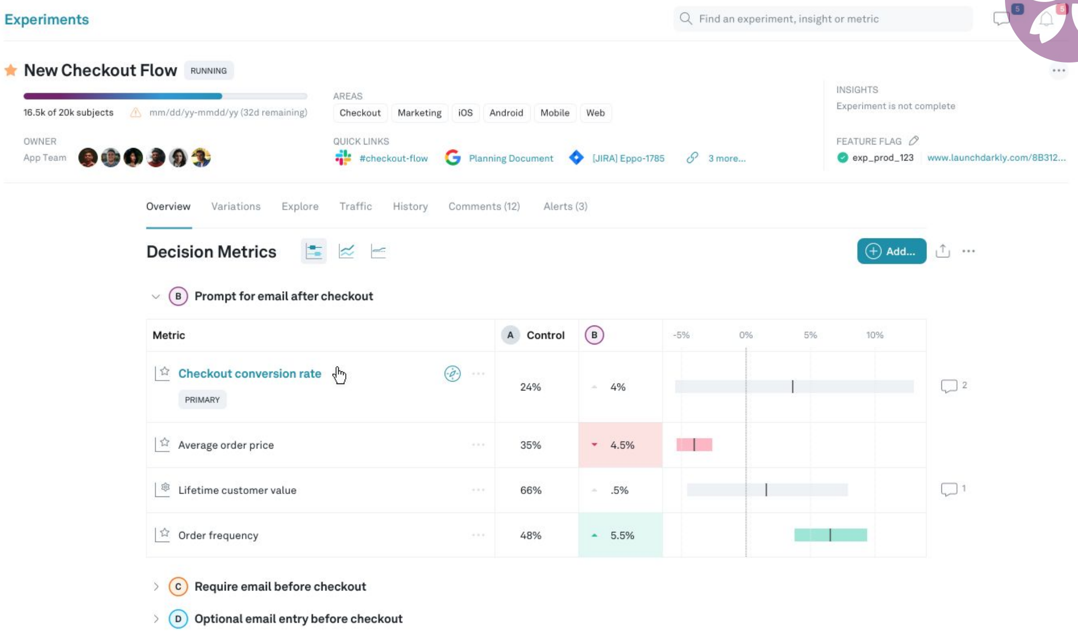
VISUALIZATION

DATA TABLE SQL Calculations Row Limit 500 Totals

	Users Age Tier	Users Count A	Users Count B	Users Average Lifetime Orders A	Users Average Lifetime Orders B	Users T Score	Users Significance
1	[10,20)	398	395	2.59	2.25	2.13	(3) .025 sig. level
2	[20,30)	823	888	2.78	2.34	3.79	(7) .0005 sig. level
3	[30,40)	862	834	2.72	2.32	3.65	(7) .0005 sig. level
4	[40,50)	846	867	2.77	2.39	3.19	(6) .001 sig. level
5	[50,60)	820	808	2.67	2.24	3.93	(7) .0005 sig. level
6	[60,70)	604	601	2.63	2.24	2.90	(5) .005 sig. level
7	[70,80)	360	330	2.61	2.23	2.38	(4) .01 sig. level
8	[80,inf)	231	242	2.99	2.44	2.50	(4) .01 sig. level

03 ARCHITECTURE
Reporting

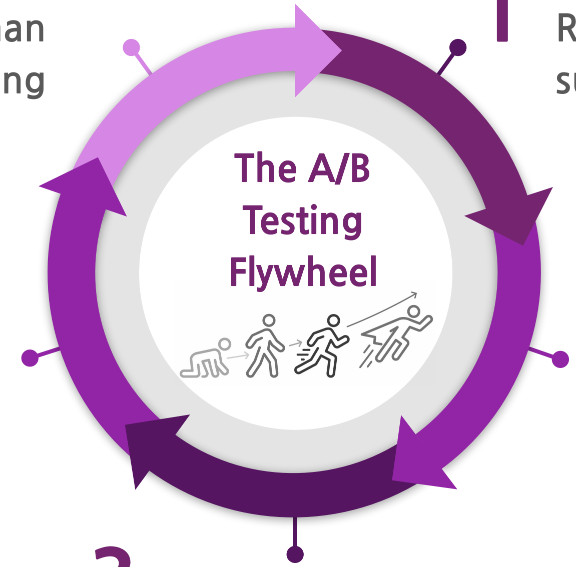
Good reporting assumes no statistics, infrastructure knowledge



- Don’t try to teach p-values, stat tests
- Don’t list 100 numbers without guidance
- Be opinionated, consistent with choice of numbers

04 A/B Testing Flywheel

Crawl, Walk, Run, Fly Progression



1. Running more A/B tests to support more decisions
2. Measuring value to decision making
3. Increasing interest in A/B testing
4. Investing in A/B testing infrastructure and data quality
5. Lowering human cost of A/B testing

We'd love to hear from you!

Reach out to see how Eppo can help grow experimentation culture in your company.

WEBSITE

www.geteppo.com

EMAIL ADDRESS

che@geteppo.com

TWITTER

@chesharma87

Beyond Linearity

Building reactive notebooks for data

Caitlin Colgrove, CTO @ Hex

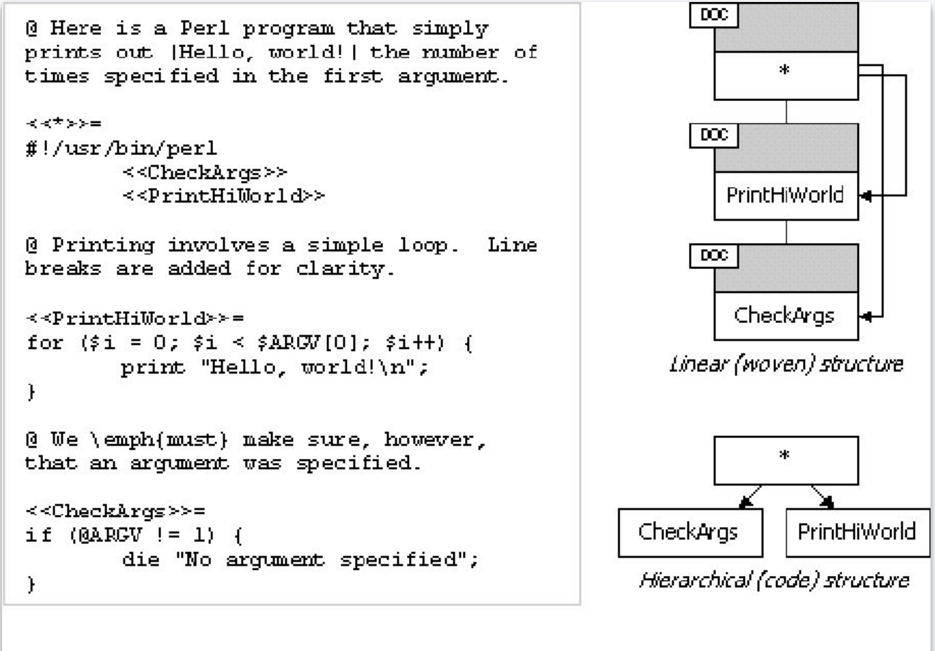
HEX

Poll: how do code notebooks make you *feel*?

A. I use notebooks for everything! Analysis, text editing, email... all notebooks!
B. They’re useful sometimes but they have their drawbacks.
C. I will literally quit my job if they make me use a notebook.
D. You mean, like... to write in?

Historical background: literate programming

In 1984, Donald Knuth introduced the concept of "literate programming", a way of developing that mixes code, explanation, and outputs together in a way that's meant to be more interpretable by humans.

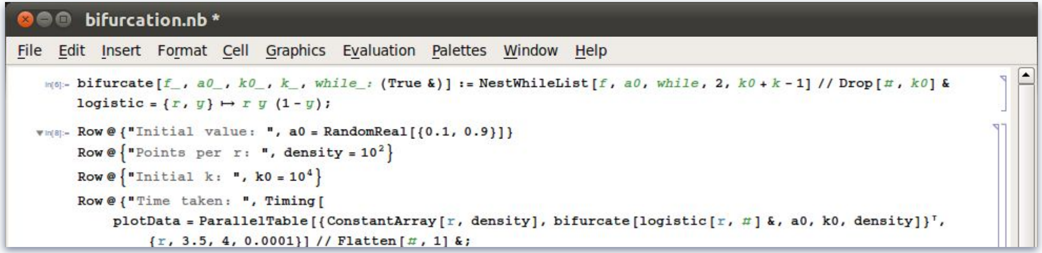


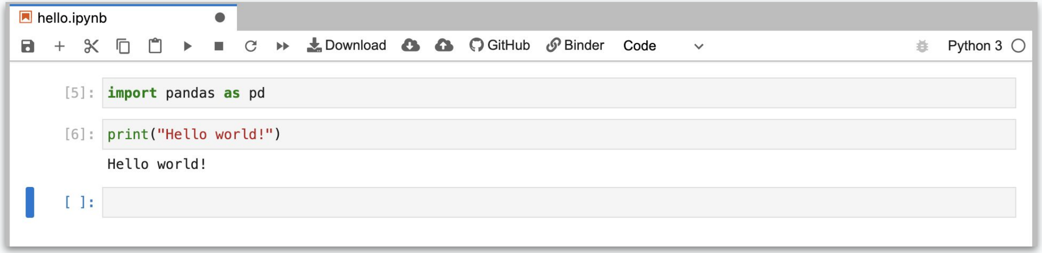
(a) Literate source.

(b) Linear and hierarchical views.

Fast forward to 2022

Notebooks are the most widely-used example of literate programming in practice.

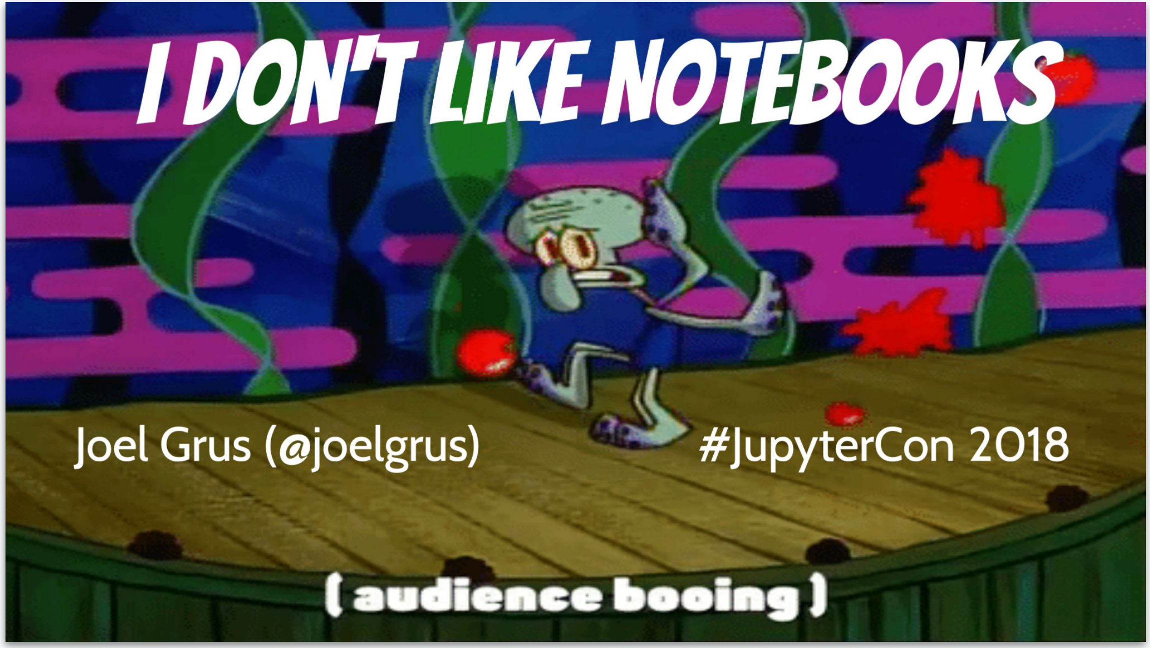




Why notebooks?

- Mix code and outputs together
- Great for iterating on smaller chunks of code; well-suited to exploration
- Linear, narrative layout that is great for storytelling

But notebooks have... issues



The State Problem

```
a = 1
```

```
a = 2
```

```
print(a)
```

What does this print?

imperative programming

a programming paradigm that uses statements that change a program's state.

Notebook state causes 3 major problems

1. **Interpretability**

   It's hard to reason about what's happening in a notebook, especially someone else's.

2. **Reproducibility**

   Out of order cells make it hard to reproduce work without frequent restart-and-run-alls.

3. **Performance**

   Re-runs are wasteful and time-consuming... especially in Hex :(

Another barrier to entry

This is exactly the kind of thing that scares people off from analytics and data science, and gives code a bad name.

The state of state


DATA DEPARTMENT
HAVE YOU TRIED RESTARTING
AND RUNNING FROM SCRATCH
imgflip.com

Re-thinking state

reactive programming

a programming paradigm oriented around data flows and the propagation of change.

In practice, this means that reactive objects maintain references to their dependencies and update automatically when their dependencies change.

Why reactive programming?

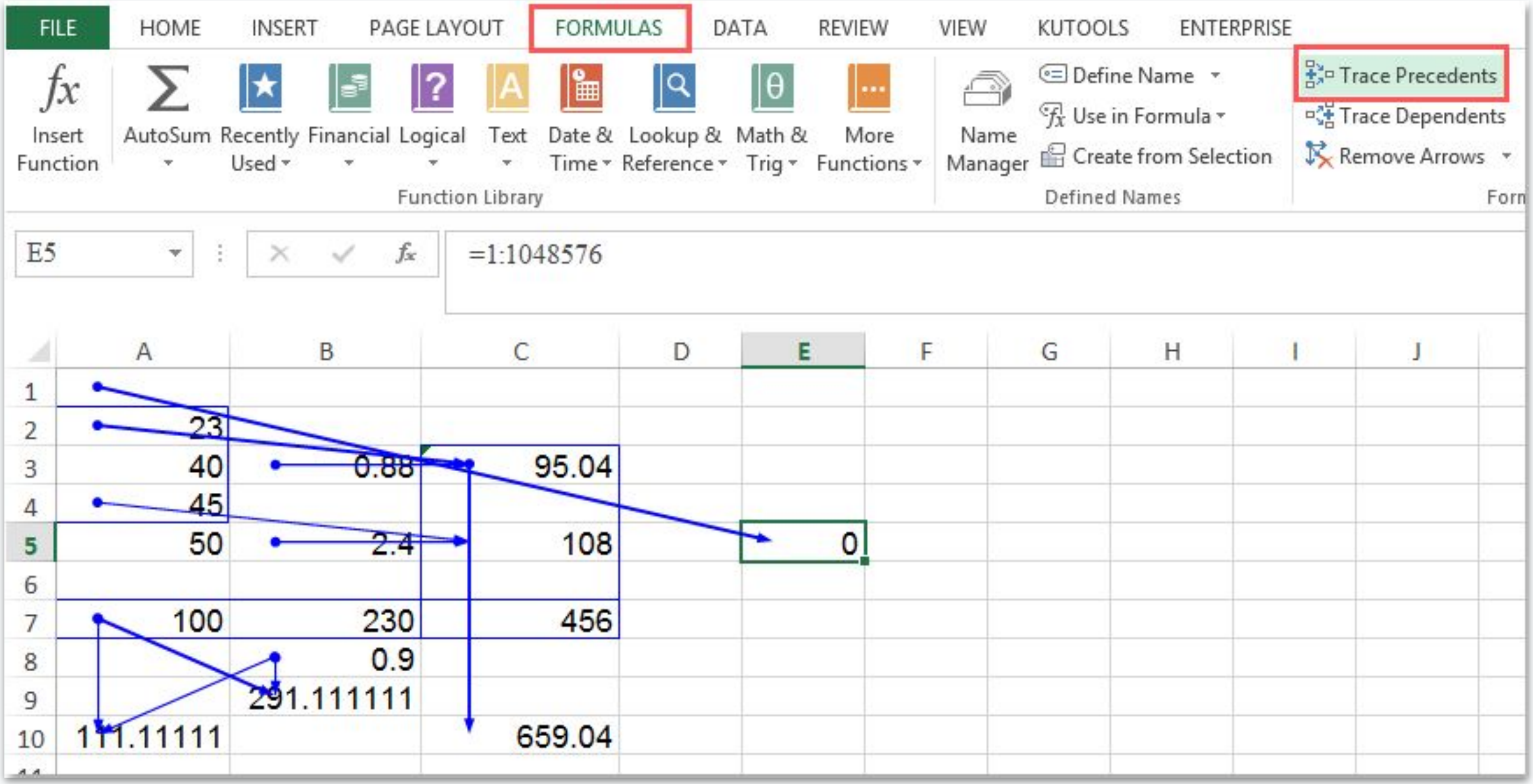
- State consistency
- Performance
- Nice abstractions for async and concurrent data flows

Imperative

```
>> a = 4
>> b = 10
>> c = a + b
>> c
14
>> a = 25
>> c
14
```

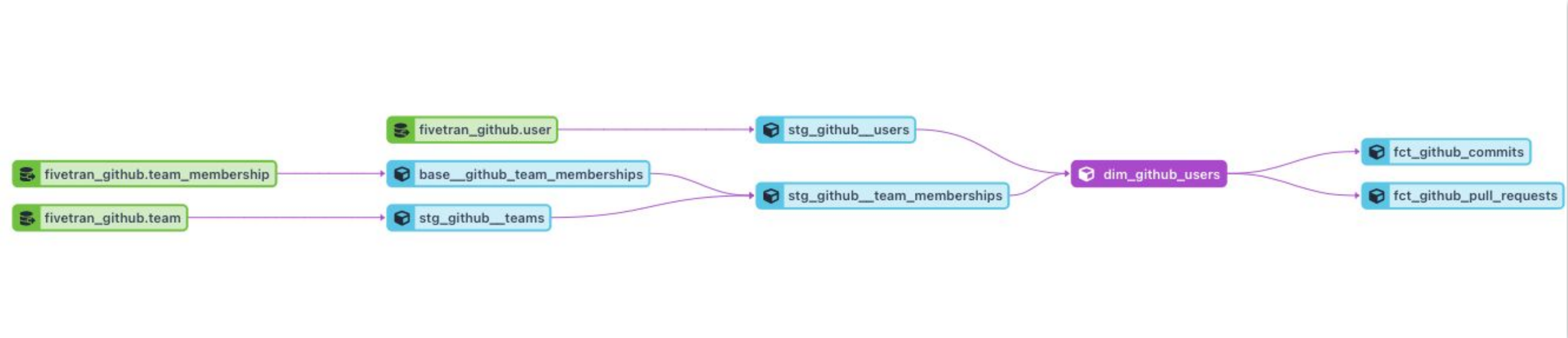
Reactive

```
>> a = 4
>> b = 10
>> c = a + b
>> c
14
>> a = 25
>> c
35
```



Everyone's favorite reactive programming tool

DAGs!

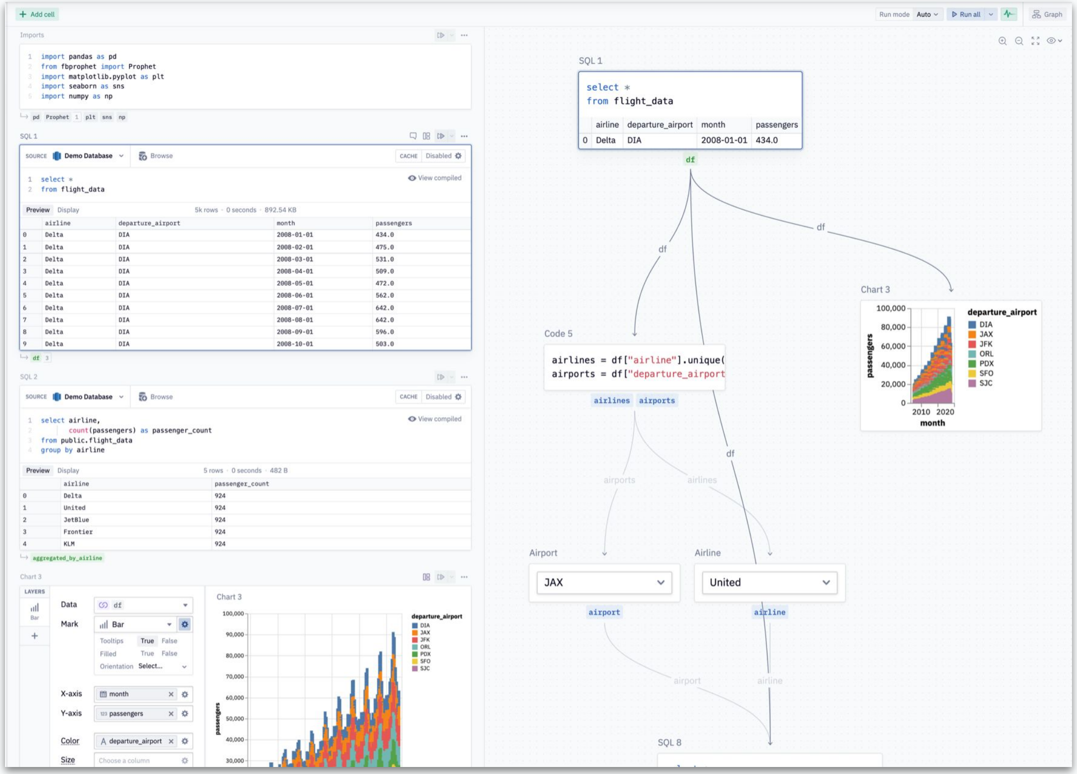


a DAG in dbt

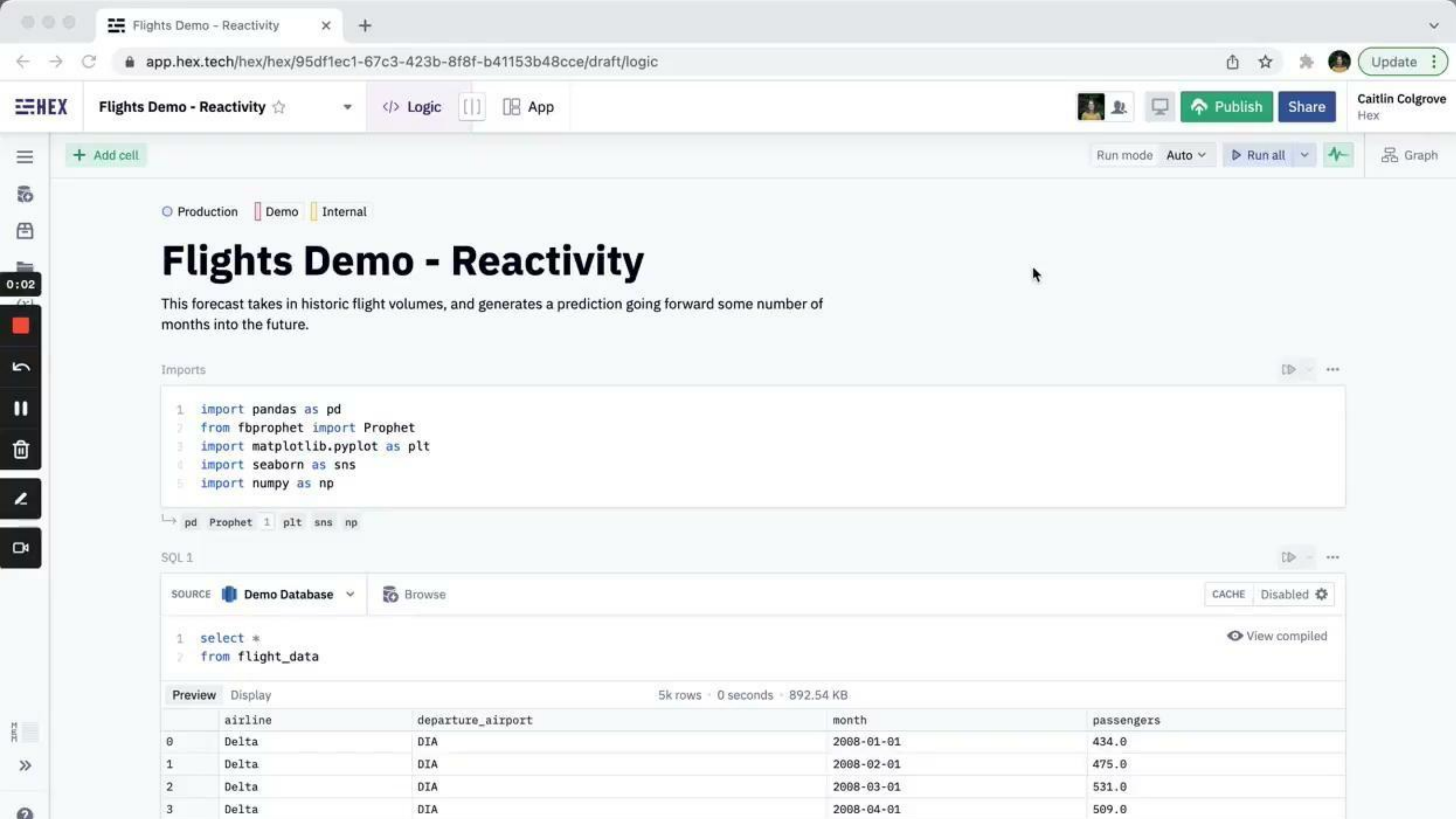
Bringing reactivity and DAGs to notebooks

We introduced a **fully-reactive, DAG-based execution model** in Hex 2.0, which solves for all 3 problems we discussed earlier:

- Interpretability
- Reproducibility
- Performance



Demo



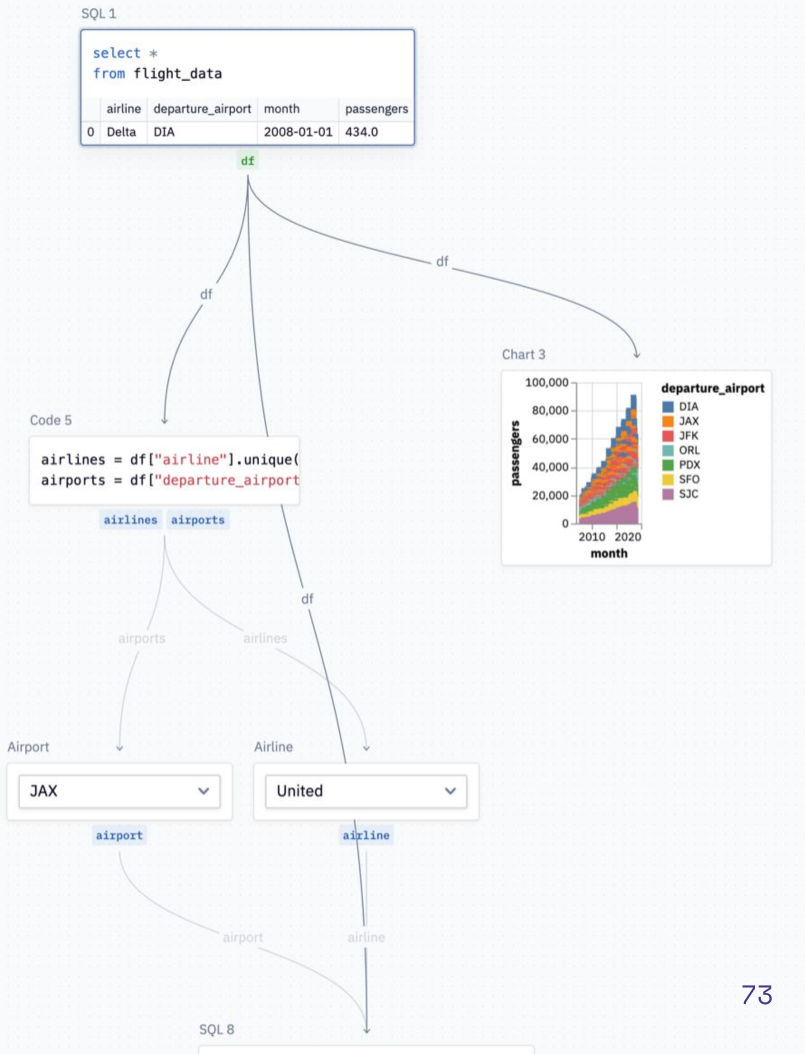
Flights Demo - Reactivity
app.hex.tech/hex/hex/95df1ec1-67c3-423b-8f8f-b41153b48cce/draft/logic
Update
HEX
Flights Demo - Reactivity
Logic
App
Publish
Share
Caitlin Colgrove
Hex
Add cell
Run mode
Auto
Run all
Graph
0:02
Production
Demo
Internal
Flights Demo - Reactivity
This forecast takes in historic flight volumes, and generates a prediction going forward some number of months into the future.
Imports
import pandas as pd
from fbprophet import Prophet
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
pd
Prophet
plt
sns
np
SQL 1
SOURCE
Demo Database
Browse
CACHE
Disabled
select *
from flight_data
View compiled
Preview
Display
5k rows · 0 seconds · 892.54 KB
airline
departure_airport
month
passengers
0 Delta DIA 2008-01-01 434.0
1 Delta DIA 2008-02-01 475.0
2 Delta DIA 2008-03-01 531.0
3 Delta DIA 2008-04-01 509.0

Under the hood: building the DAGs

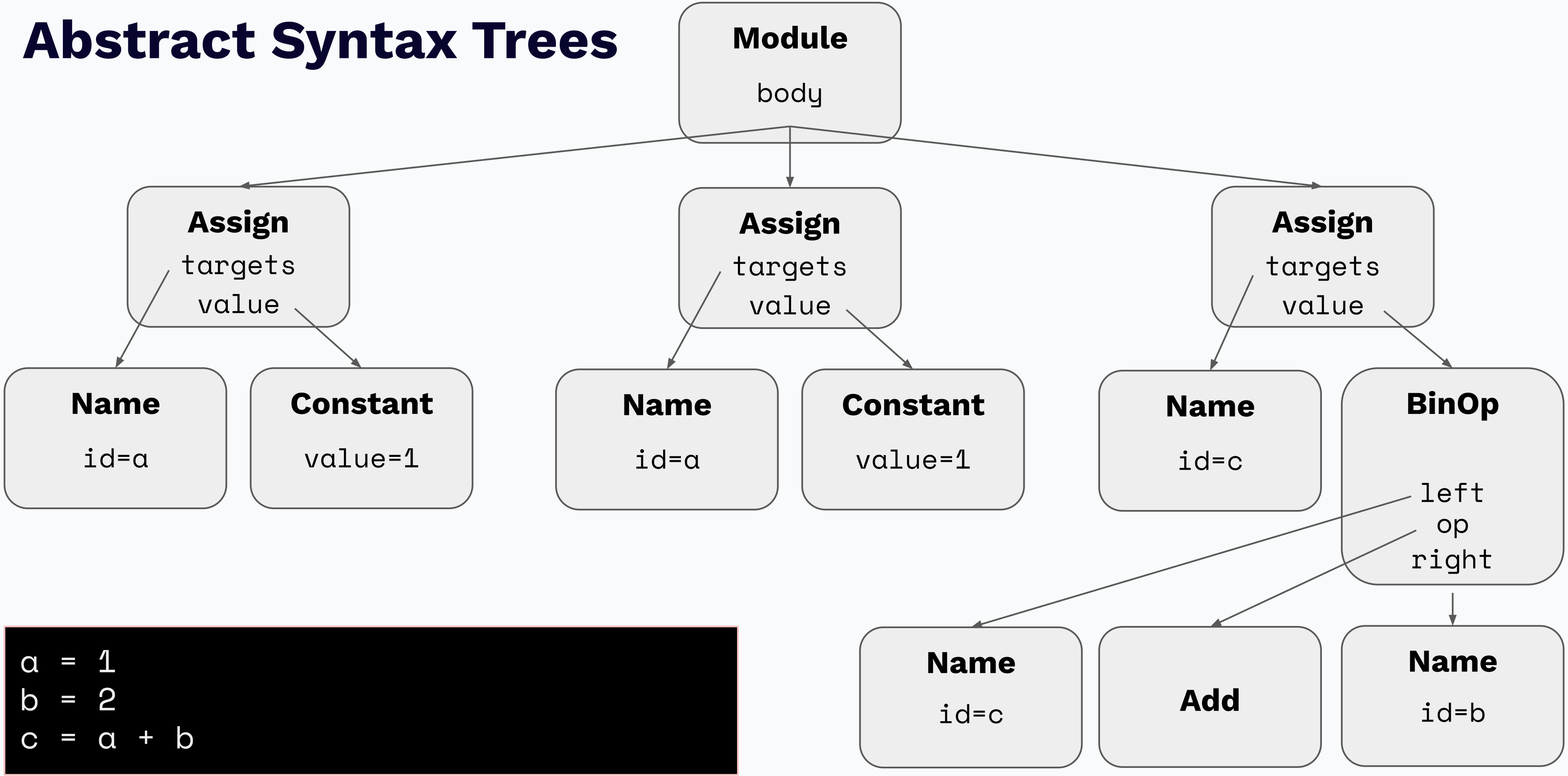
Graphs have **Nodes** and **Edges**:

- Nodes = Cells
- In edges: Variable references
- Out edges: Variable assignments

How do we determine relationships?



Abstract Syntax Trees




```
a = 1
b = 2
c = a + b
```

Issues with this approach

It’s not actually a DAG!

```
a = 1
b = a + 1
```

```
b = 1
a = b + 1
```

The ordering is non-deterministic

```
a = 1
```

```
a = 2
```

```
print(a)
```

Solution: use notebook ordering

```
a = 1
b = a + 1
```

```
a = 1
b = a + 1
```

```
a = 1
```

```
a = 2
```

```
print(a)
```

Pulling it all together: bringing DAGs into Hex notebooks

Determining “staleness”

In order to know which cells to recompute, we track a condition called *staleness*.

A cell is *stale* if:

- It hasn’t been run yet this kernel session
- An upstream cell has been **edited** and it hasn’t been re-run
- An upstream cell has been **run** and it hasn’t been re-run
- An upstream cell has **become stale**

Implementing Reactivity with iPython

On each edit:

- Run each cell through an AST parser to compute inputs and outputs
- Re-compute the cell DAG
- Traverse graph upstream **and** downstream to determine list of cells needed to be run
  - Upstream, filter out cells that are already “up to date”
  - Downstream, mark as “stale”
- Queue all remaining stale cells in notebook order into the kernel
  - Mark cell as “up to date” after successful run

DAG usability cleanup

Code 0

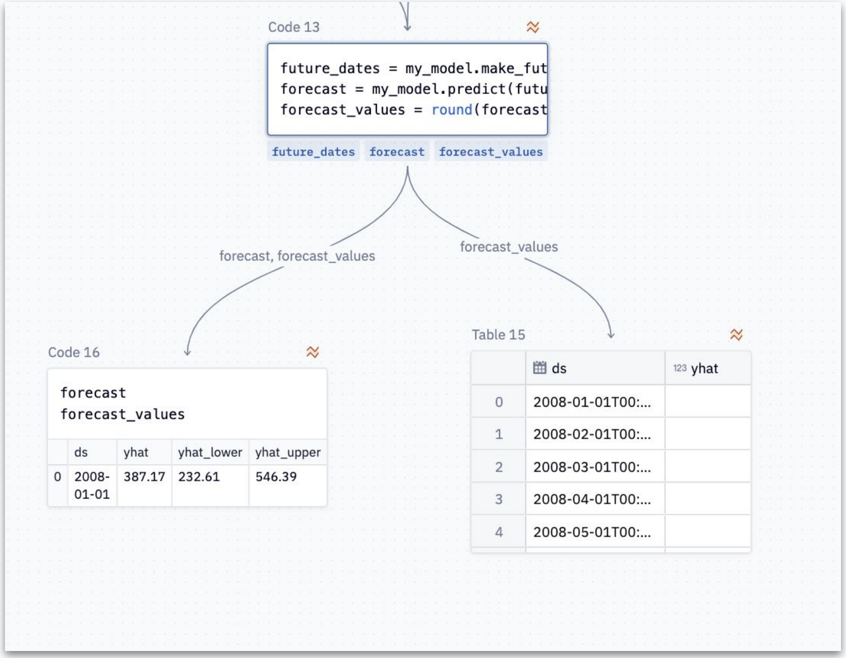
```python
import pandas as pd
from fbprophet import Prophet
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

pd Prophet 1 plt 1 sns

Markdown 1

```
# Flight Traffic Forecast
```

Flight Traffic Forecast



Future exploration

Future exploration

- Lambdas / better isolation
- Cell caching
- Performance & parallelism

Adam Storr
Design Lead

Melissa Carlson
Engineering Lead

Glen Takahashi
Chief Architect

Interested?

Director, Platform Engineering
Backend Engineer
Cloud Engineer
Platform PM
Engineering Lead
... and many more

hex.tech/jobs

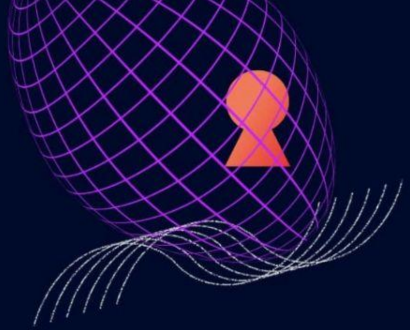
Questions?

The Return of the OLAP Cube

Benn Stancil
Chief analytics officer | Mode

MODE

benn.substack.com



The Return of the OLAP Cube

Benn Stancil, Chief Analytics Officer | Mode

ABOUT THE TALK

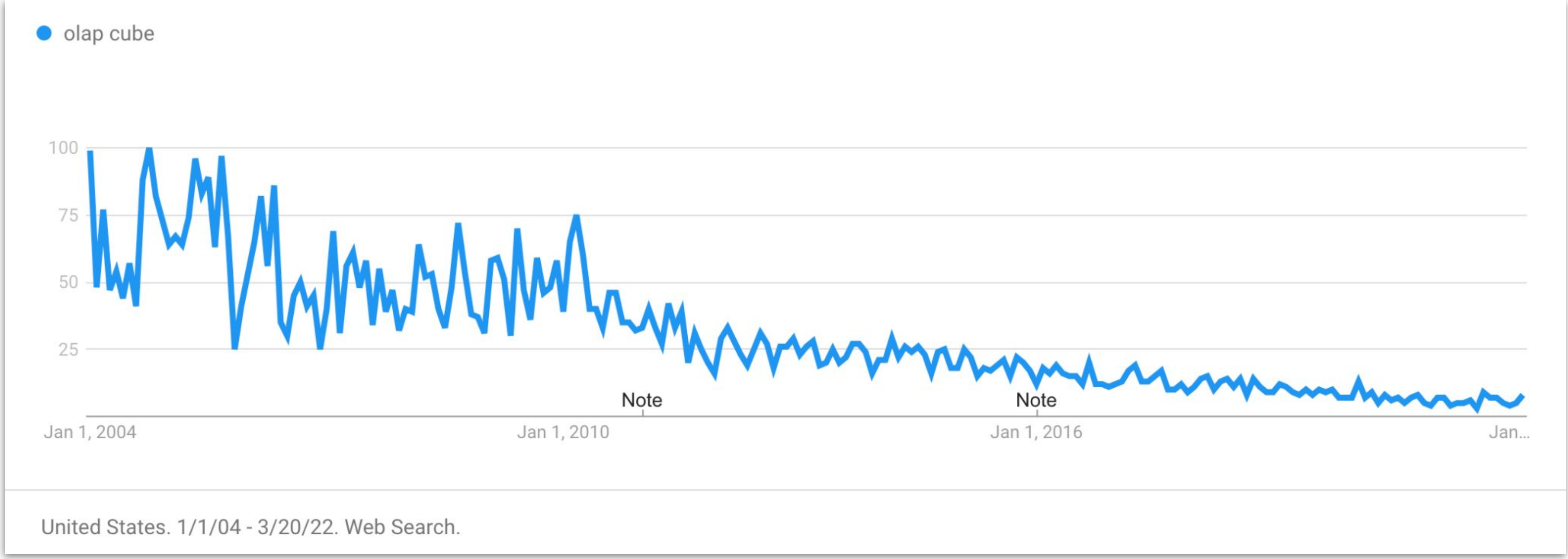
Fifteen years ago, OLAP cubes were a critical part of every analytics and BI stack. In a time when databases were slow and compute was expensive, cubes provided an elegant solution for standardizing multi-dimensional reporting. Over the last decade, however, they've fallen out of favor. As warehouses have gotten bigger, faster, and cheaper, cubes no seem longer necessary. Analysis and reporting is now done directly on top of raw data, no predefined or pre-aggregated cubes required.

Or are they? OLAP cubes are reappearing in the modern data stack—just in a different form and under a different name. Instead of being separate data marts built for reporting and BI, cubes are now synthetic, generalized, and on-demand. In this talk, I'll walk through the history of OLAP cubes and their modern echoes. And I'll explain why this is actually a good thing—and why we should actually be excited about the return of the OLAP cube.

analytics engineers club

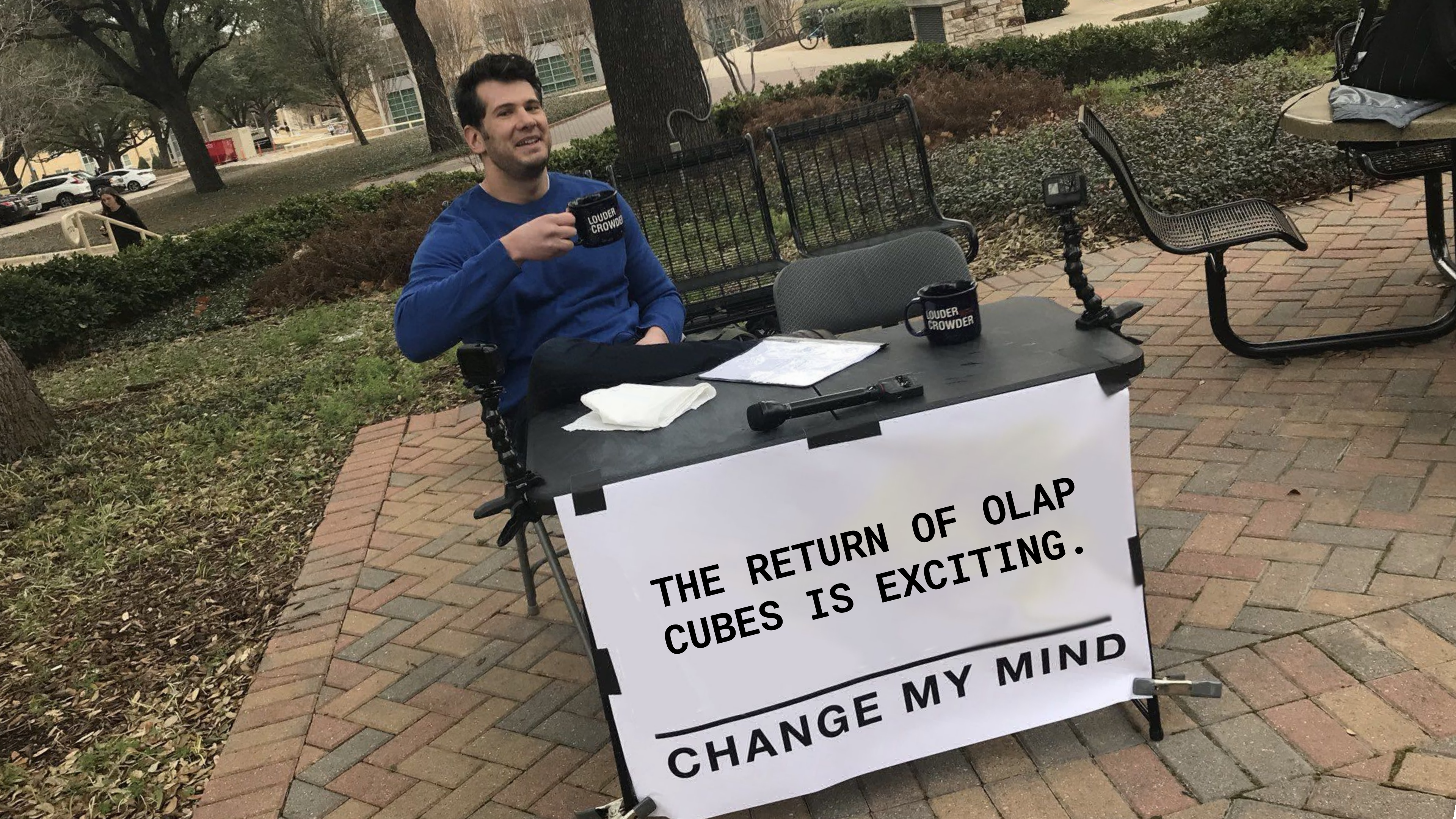
What’s An OLAP Cube? 

By Claire Carroll | August 18, 2021


olap cube
100
75
50
25
Note
Note
Jan 1, 2004
Jan 1, 2010
Jan 1, 2016
Jan...
United States. 1/1/04 - 3/20/22. Web Search.


C
NFL
LIV
SUPER BOWL
BUCCANEERS

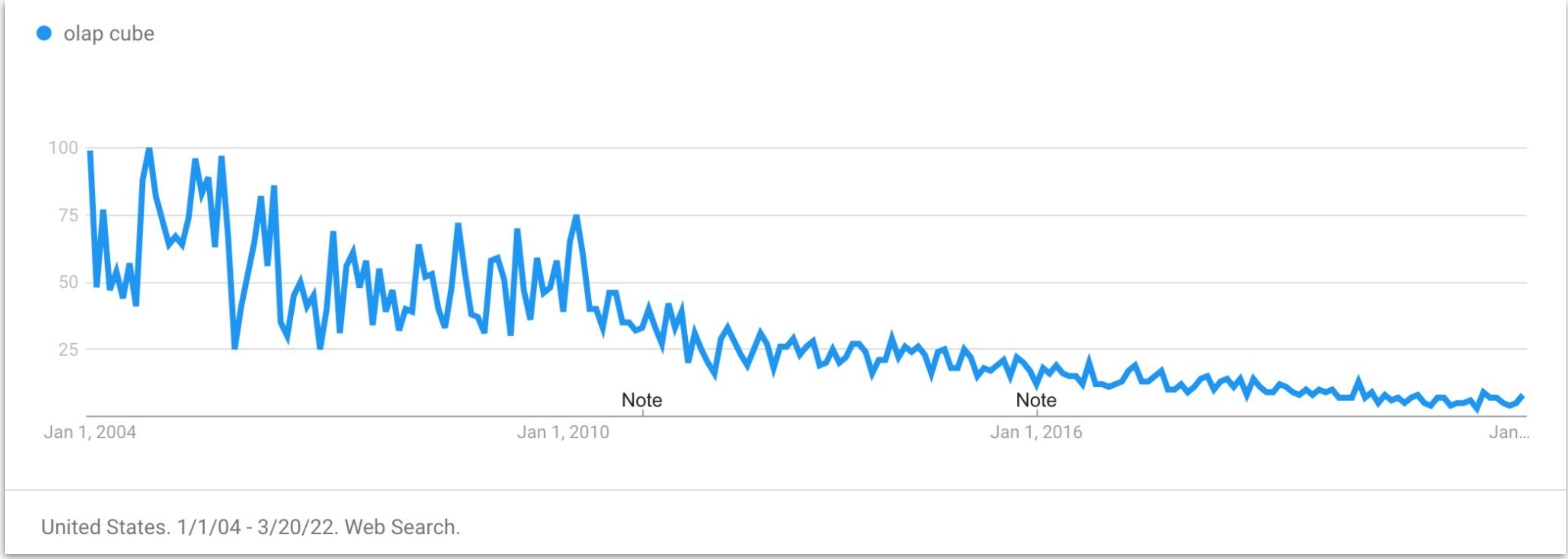

THE WRAP
Reviews Hail Robert Pattinson
Reboot as ‘Best Bat-Movie Yet’


LOUDER CROWDER
LOUDER CROWDER
THE RETURN OF OLAP CUBES IS EXCITING.
CHANGE MY MIND

analytics engineers club

What's An OLAP Cube? 🎲

By Claire Carroll | August 18, 2021


olap cube
100
75
50
25
Note
Note
Jan 1, 2004
Jan 1, 2010
Jan 1, 2016
Jan...
United States. 1/1/04 - 3/20/22. Web Search.


C
NFL
LV
SUPER BOWL
BUCCANEERS

Tilt

The Batman is a Lifeless Reboot

analytics engineers club

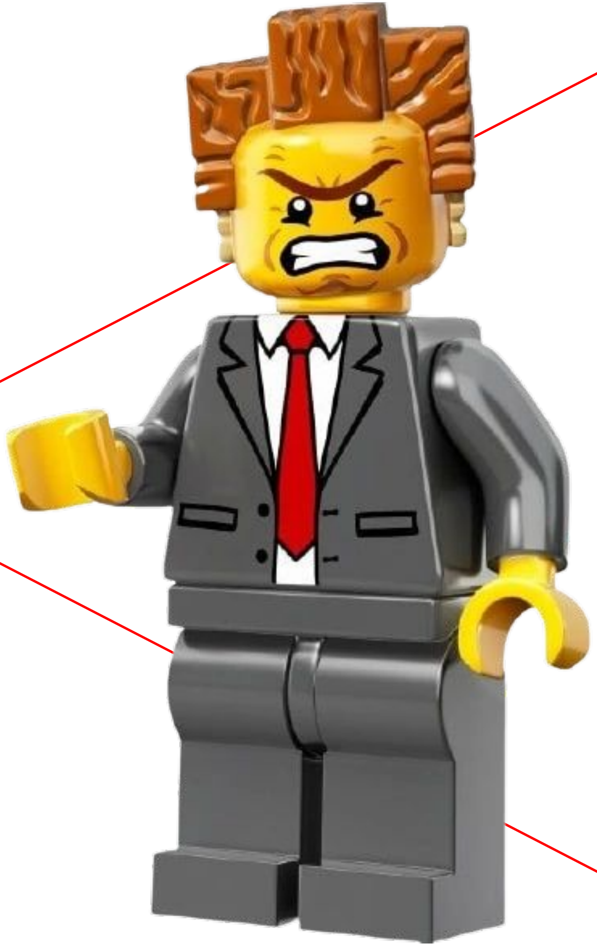
What’s An OLAP Cube? 🎲

By Claire Carroll | August 18, 2021

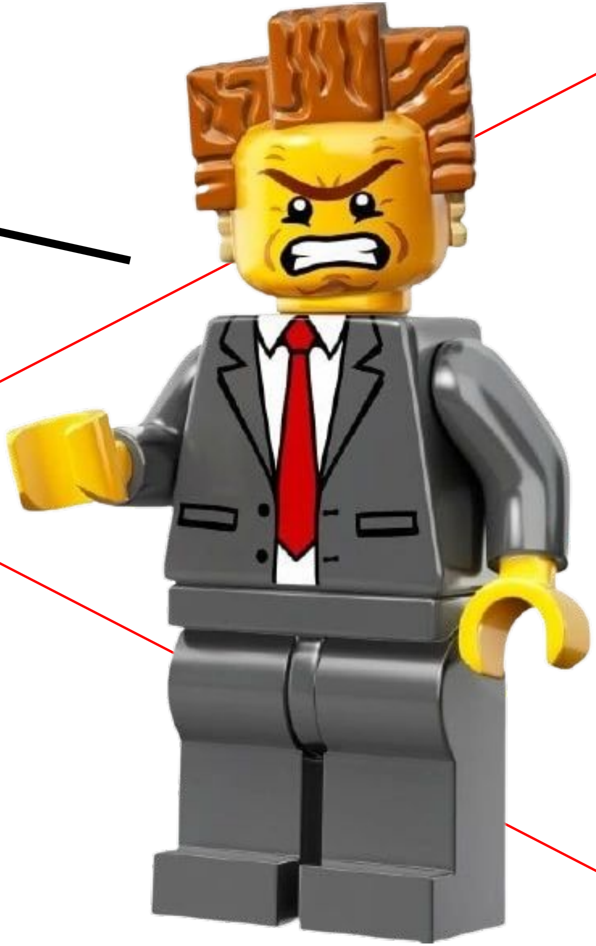
What’s An OLAP Cube? 🎲

By Claire Carroll | August 18, 2021

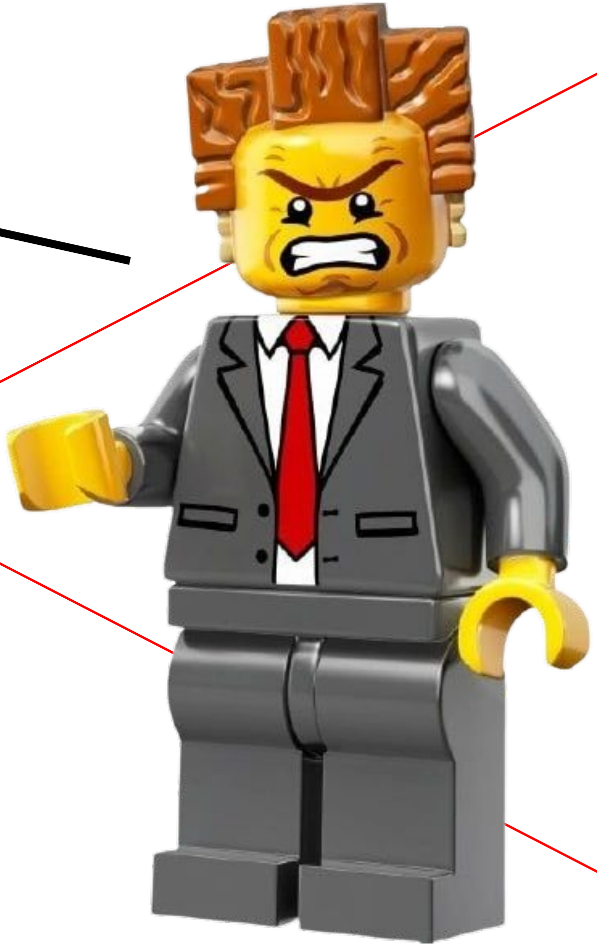
> 👋 Claire here. I’ve been working with data for six years, and always in the context of a “Modern Data Stack” — the first data stack I used included Redshift, Fivetran and Looker! In contrast, many data modeling concepts were coined in an era when analysts used on-prem databases like Oracle and IBM.
>
> As I got further into my career, I came across more terminology that didn’t make sense to me, and I was



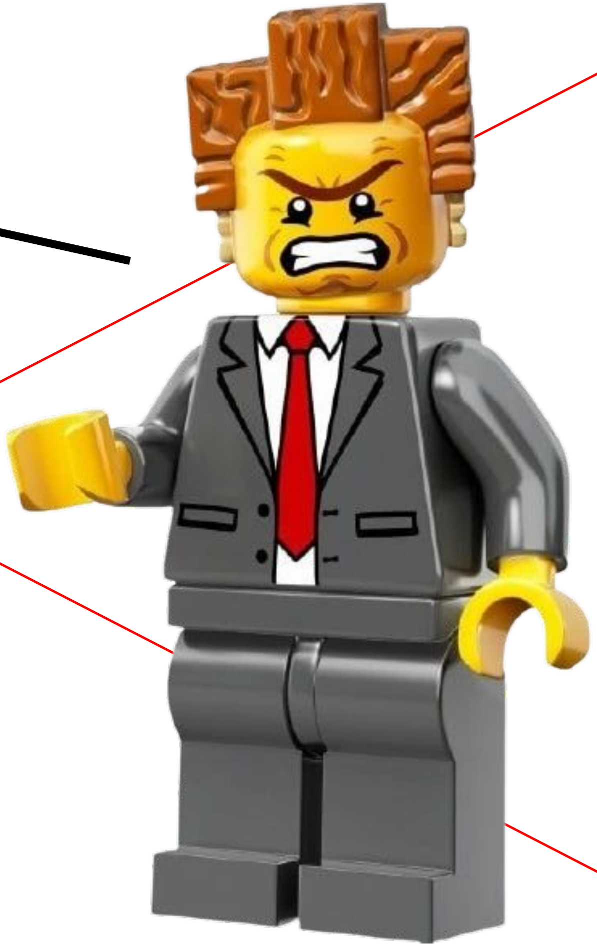
How much money did we make in California?



How much money did we make in Ohio?



How much money did we make in Ohio in January?




LEGO
Technic
8815

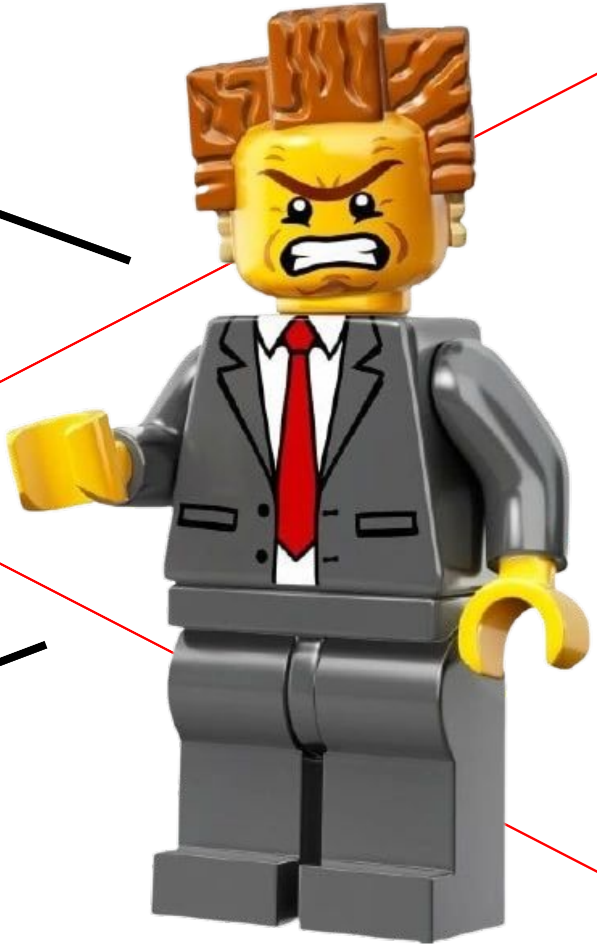
200,000,000

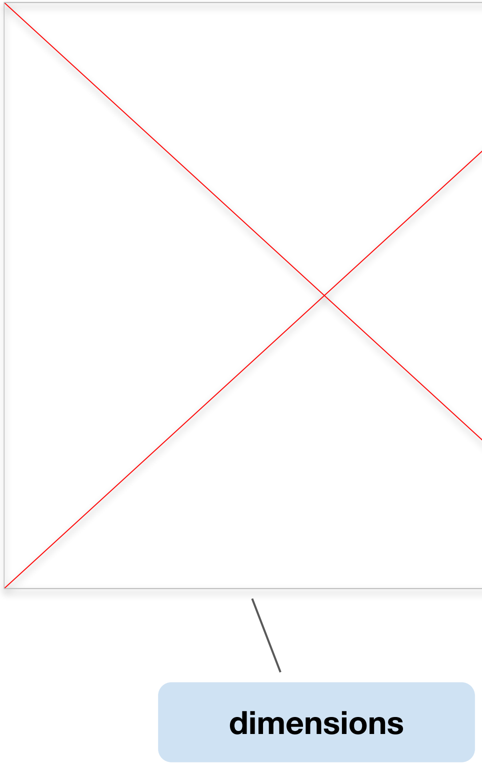
transactions a year

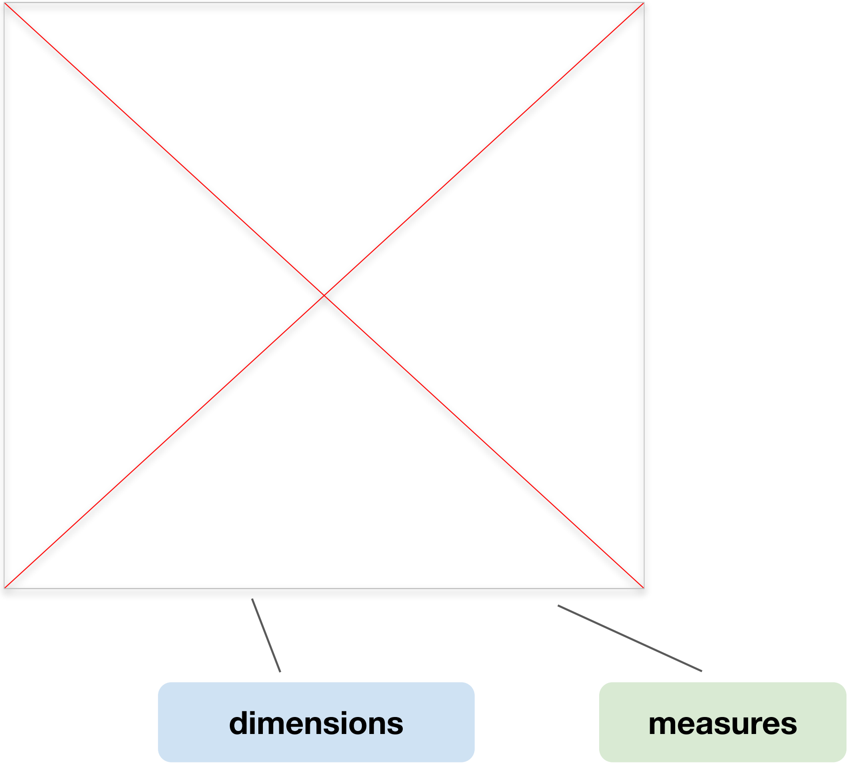
How much money did we make in California?

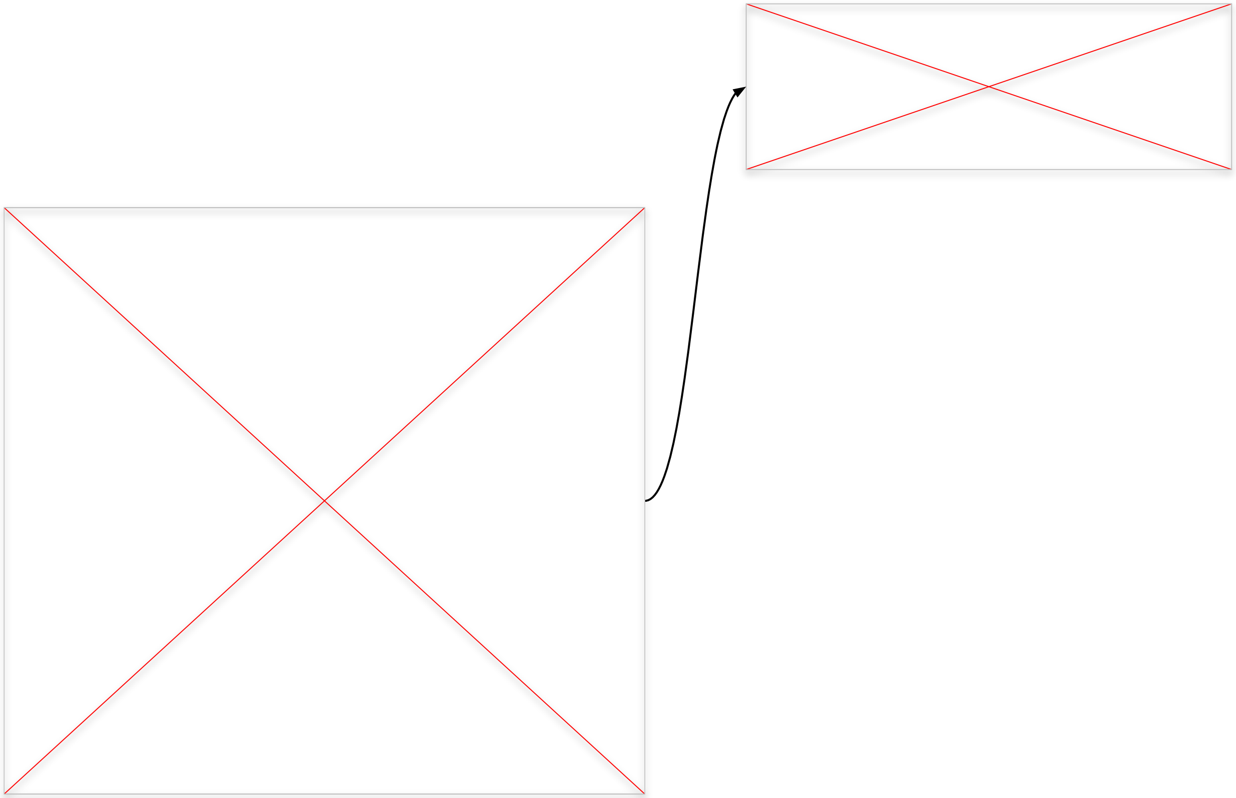
How much money did we make in Ohio?

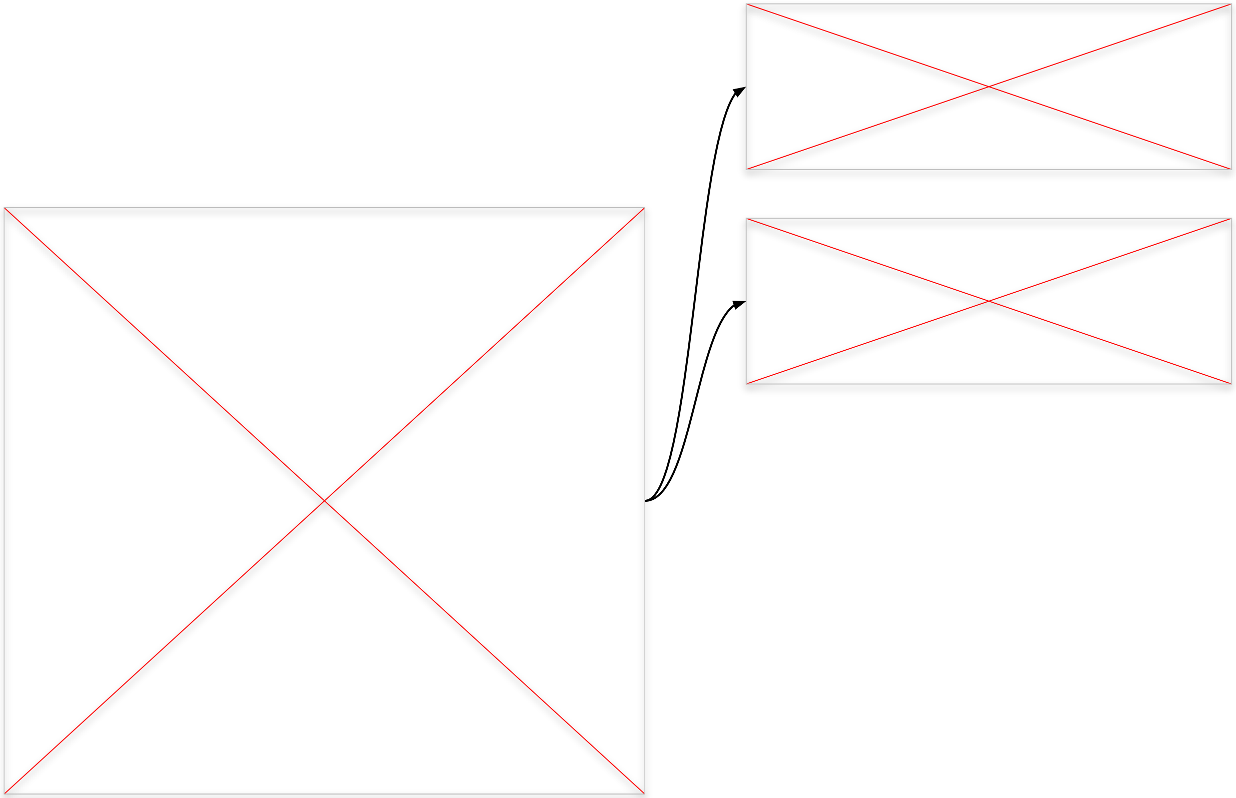
How much money did we make in Ohio in January?

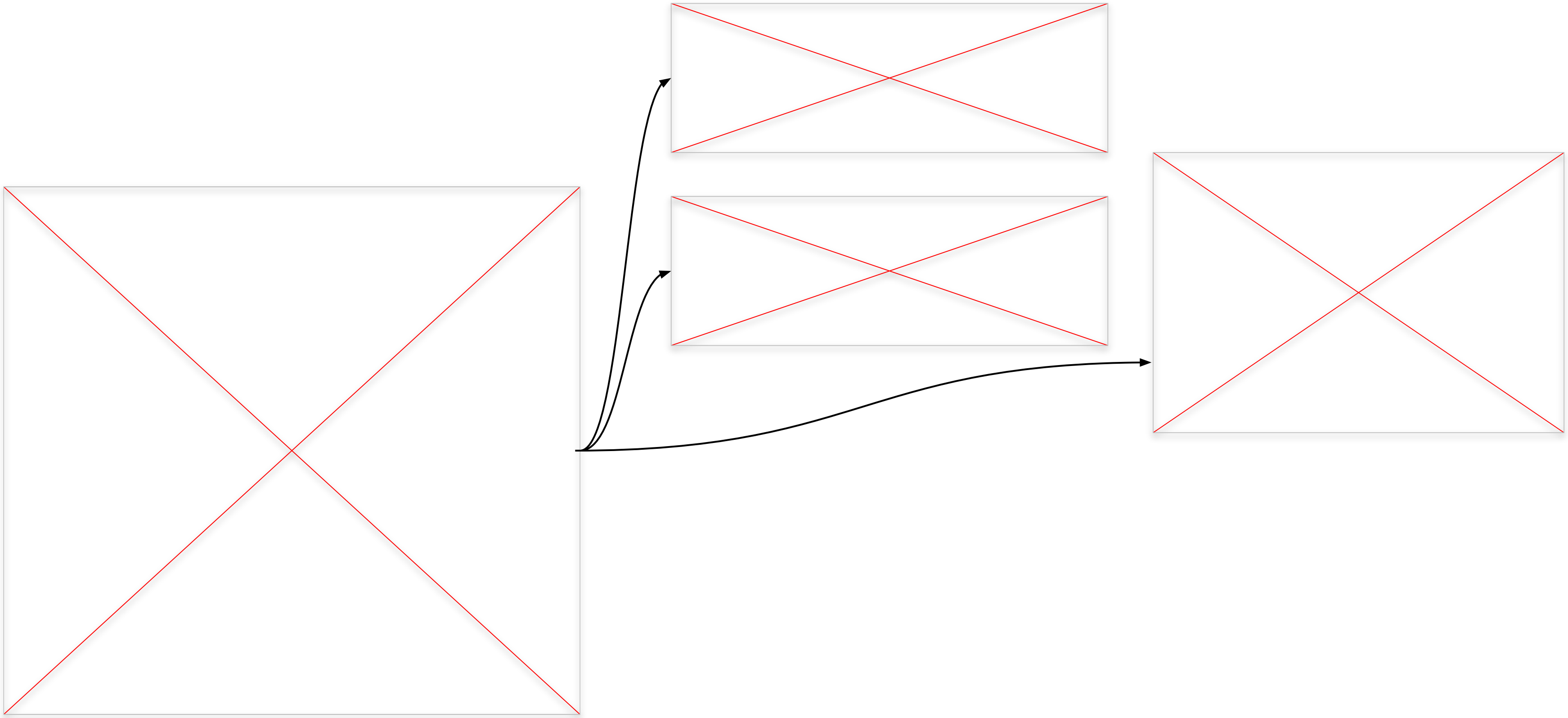


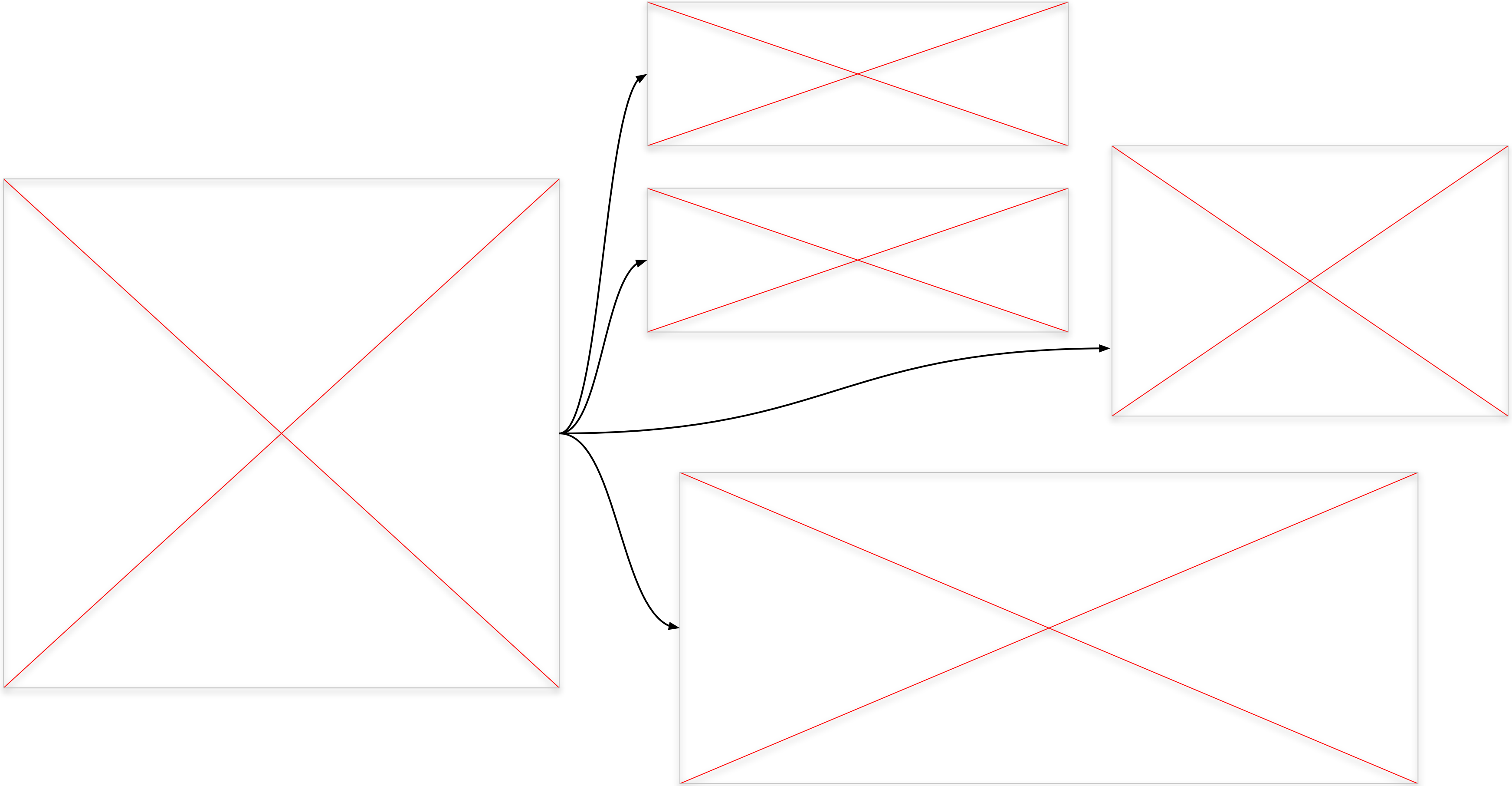

dimensions


dimensions
measures



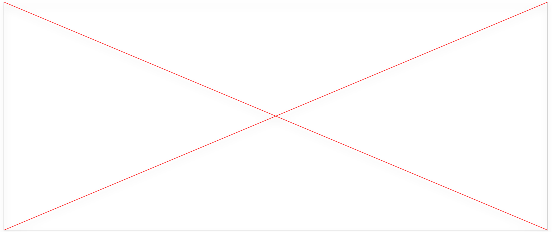




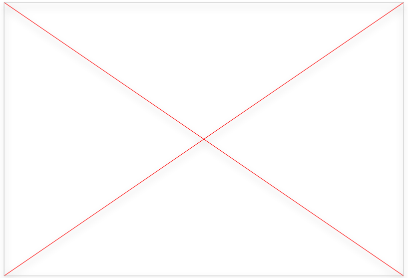


Raw data

Raw data



OLAP cube



Pre-aggregate data an OLAP cube

Raw data

OLAP cube

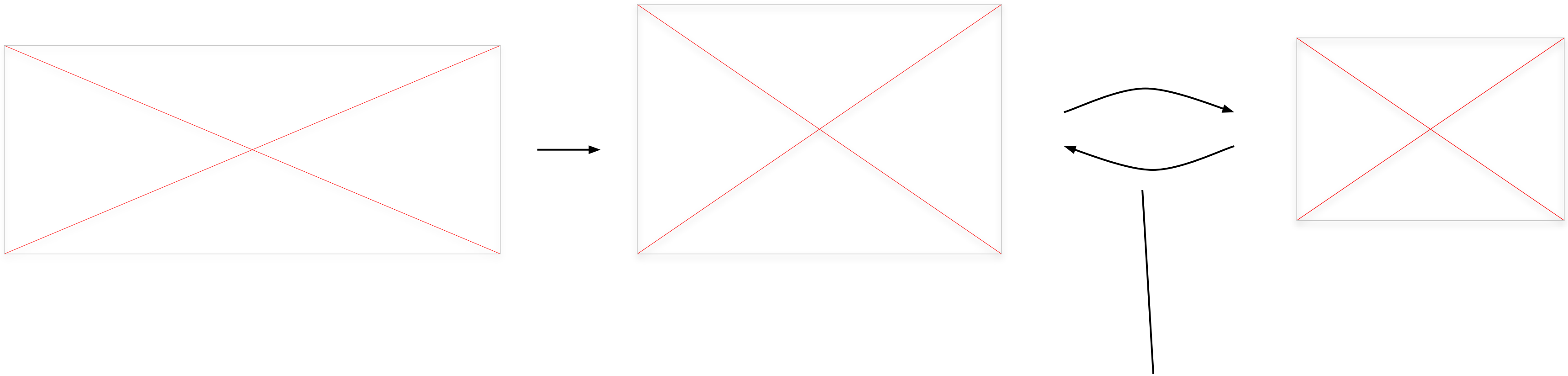
Reporting

Aggregated again to create a report

Raw data

OLAP cube

Reporting

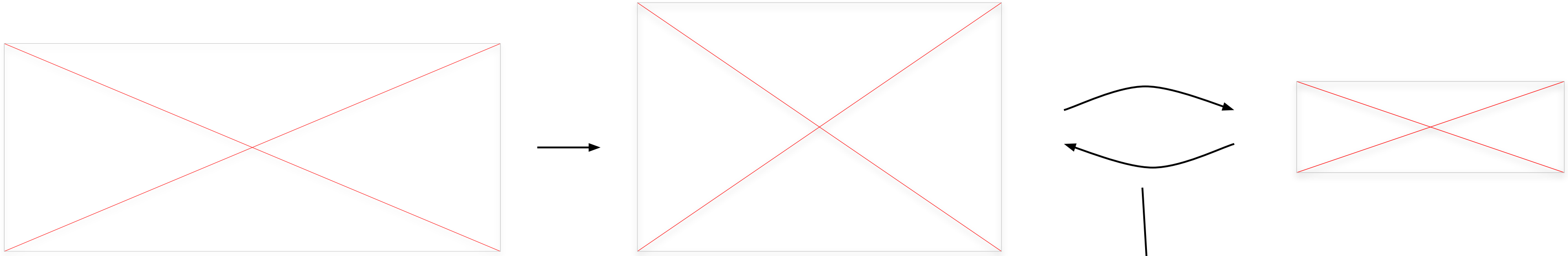


The cube does the computation

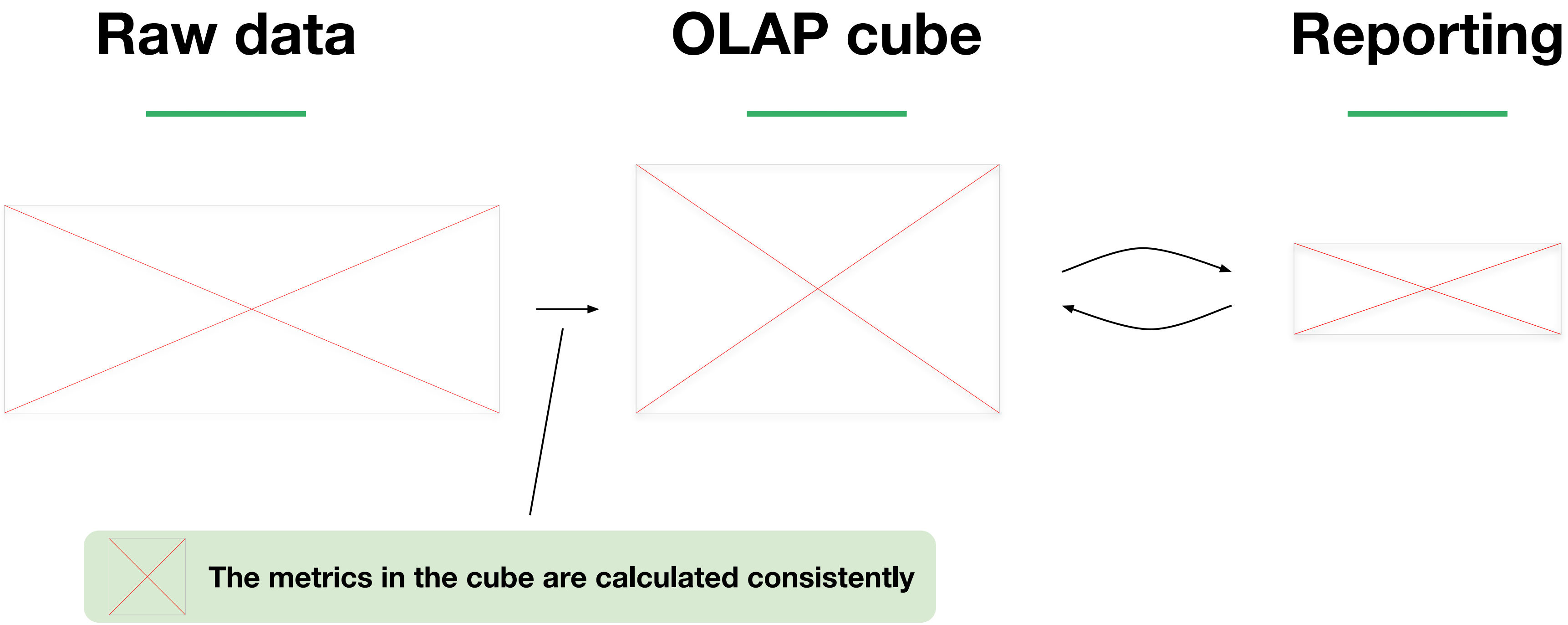
Raw data

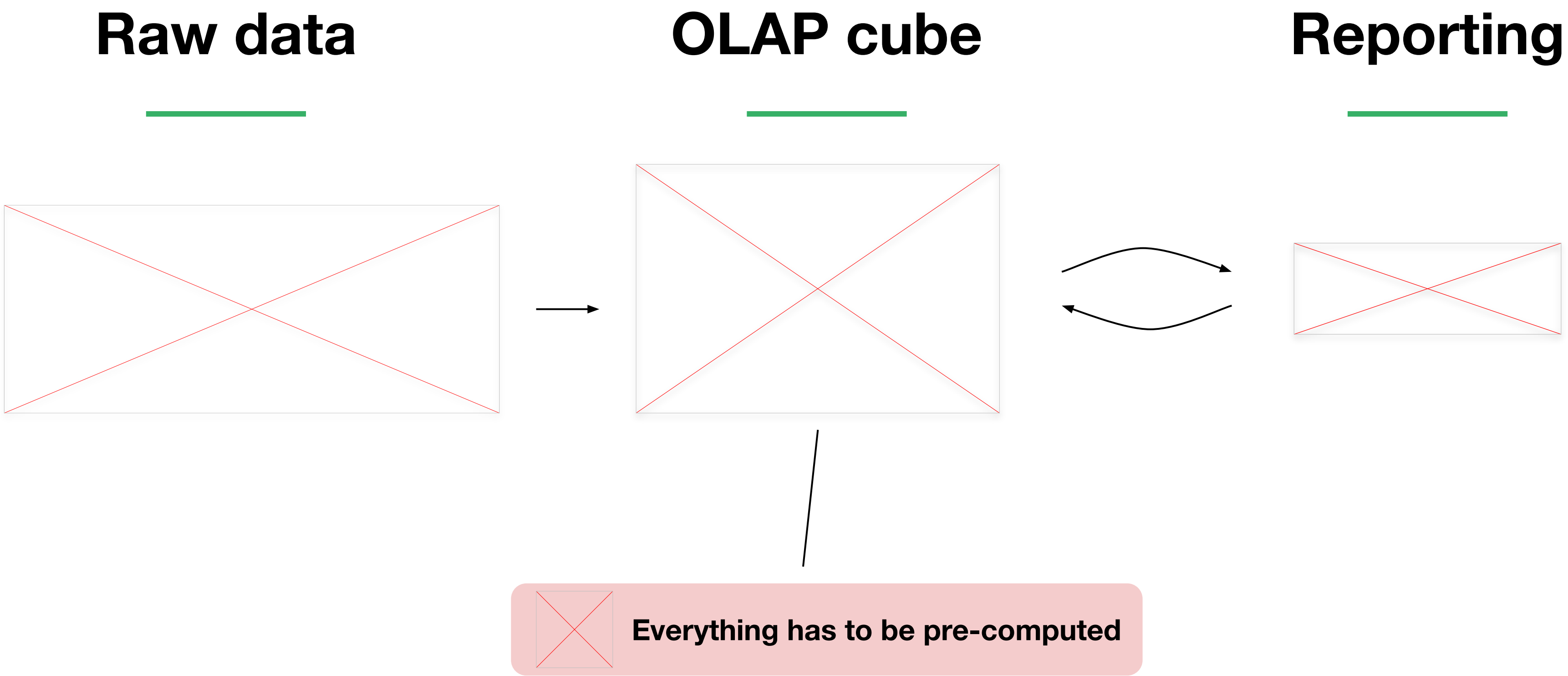
OLAP cube

Reporting



People can ask lots of questions quickly


Raw data
OLAP cube
Reporting
The metrics in the cube are calculated consistently


Raw data
OLAP cube
Reporting
Everything has to be pre-computed

month	state	store	sales	items_sold
January	California	1	$1,205	24
February	California	1	$1,346	11
March	California	1	$1,253	18
April	California	1	$1,184	28
May	California	1	$1,337	17
June	California	1	$1,245	11
January	California	2	$1,426	26
February	California	2	$1,275	26
March	California	2	$1,036	30
April	California	2	$1,357	22
May	California	2	$1,246	17
June	California	2	$1,074	23
January	California	3	$1,070	12
February	California	3	$1,480	29
March	California	3	$1,374	20
April	California	3	$1,105	26
May	California	3	$1,425	18
June	California	3	$1,205	25
January	Ohio	52	$390	8
February	Ohio	52	$461	3
March	Ohio	52	$428	7
April	Ohio	52	$420	13
May	Ohio	52	$425	14
June	Ohio	52	$435	8
January	Ohio	84	$381	3
February	Ohio	84	$487	5
March	Ohio	84	$421	5
April	Ohio	84	$528	12

In 1999…

In 1999…

4,597

sets

In 1999...

4,597 x 50

sets states

In 1999...

4,597	x	**50**	x	**52**
sets		states		weeks

In 1999…

11,952,200

combinations

People can answer questions quickly

People can answer questions quickly

Metrics are computed consistently

- People can answer questions quickly
- Metrics are computed consistently
- Everything has to be pre-computed

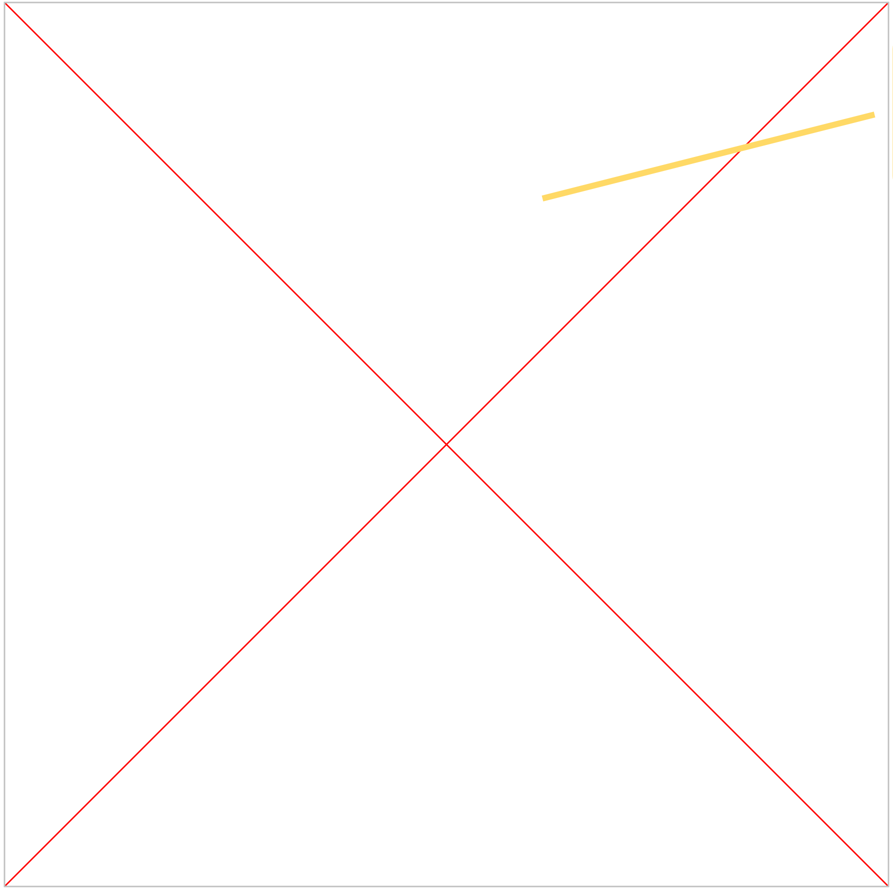
People can answer questions quickly

Metrics are computed consistently

Everything has to be pre-computed

OLAP cubes are unintuitive to use

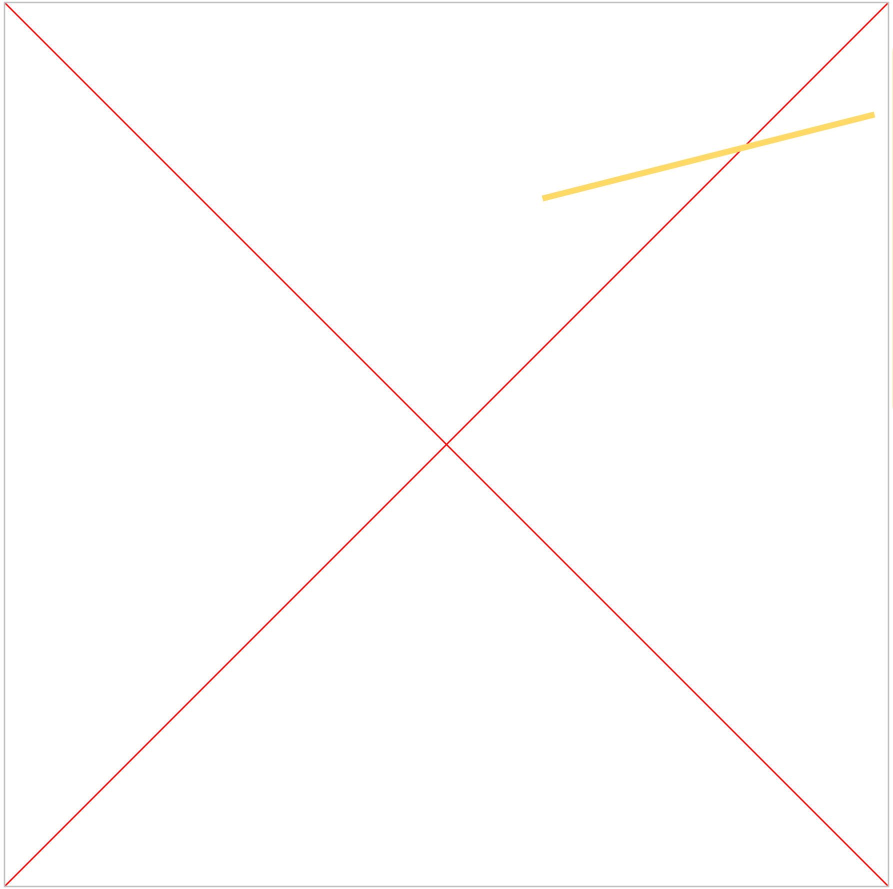
How many bricks did we sell?!?



What is our revenue?!?

What is our revenue…

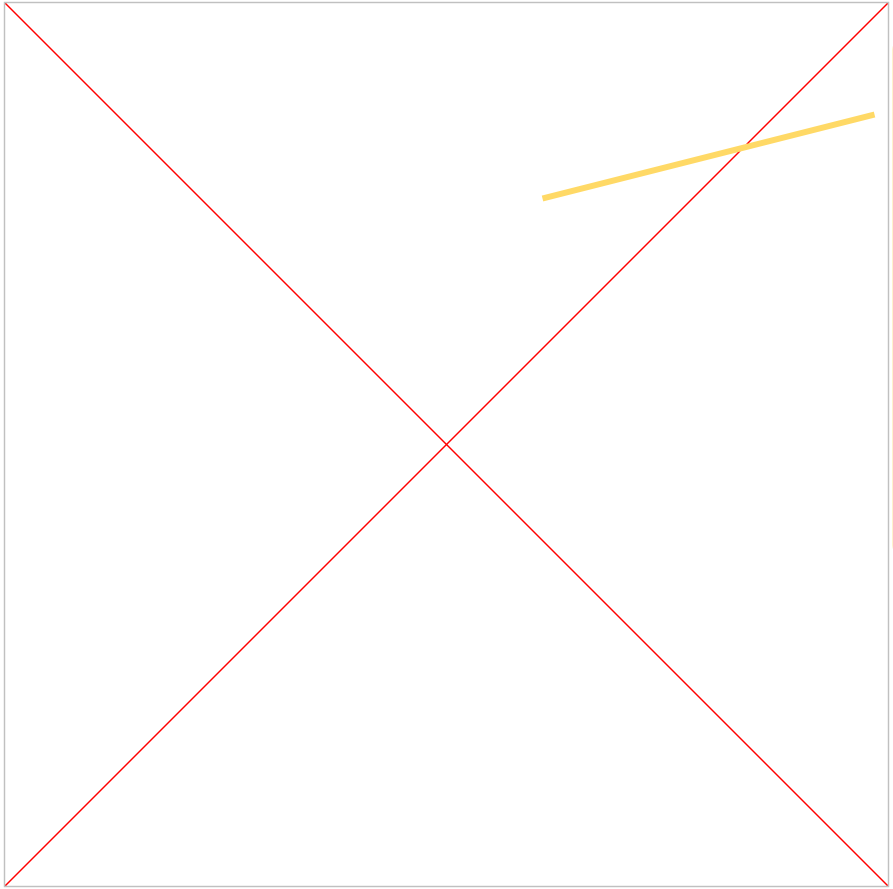
by week?!?



What is our revenue…

by week…

by set theme?!?

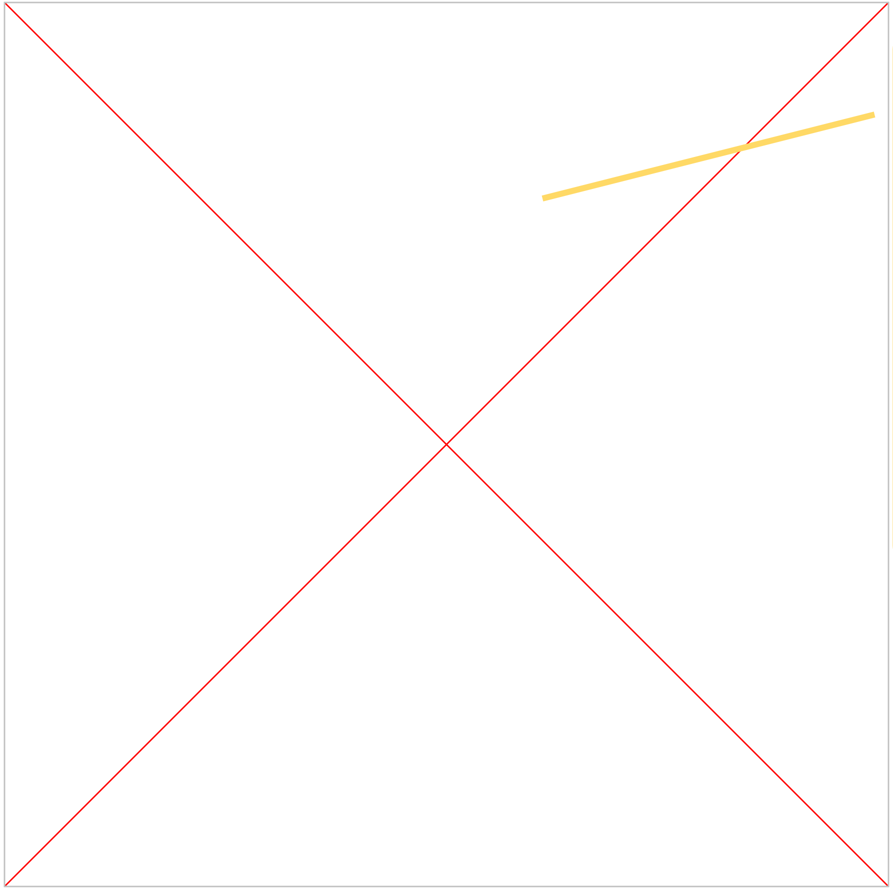


What is our revenue…

by week…

by set theme…

in Europe?!?



Get me a metric!!

Revenue

Stripe logs

Salesforce opportunity

Tax adjustment

Date sales are recognized

CSV from Janice in Accounting

Return policy

Get me a list of stores!

Get me a list of stores!

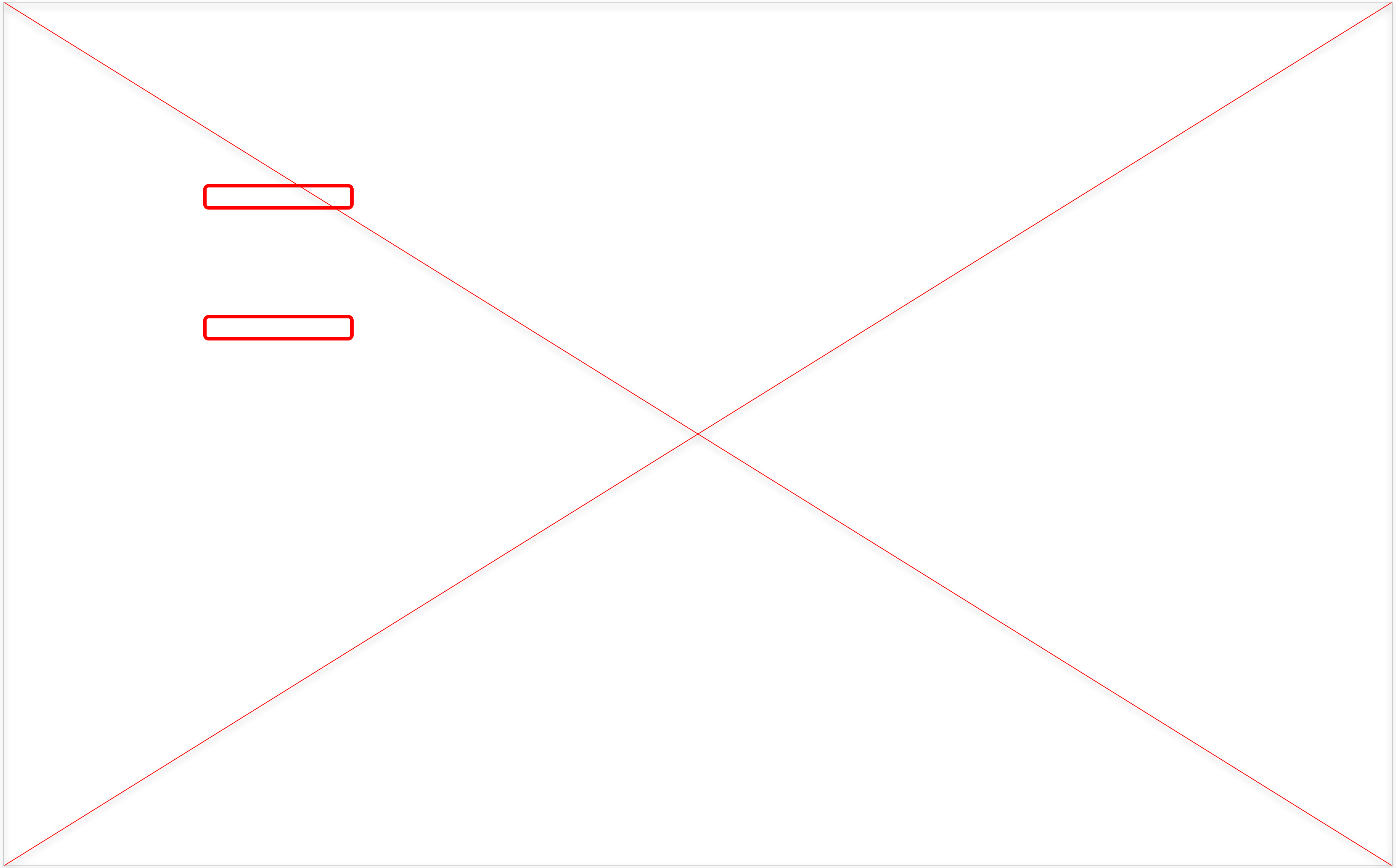
With details on their location and hours!

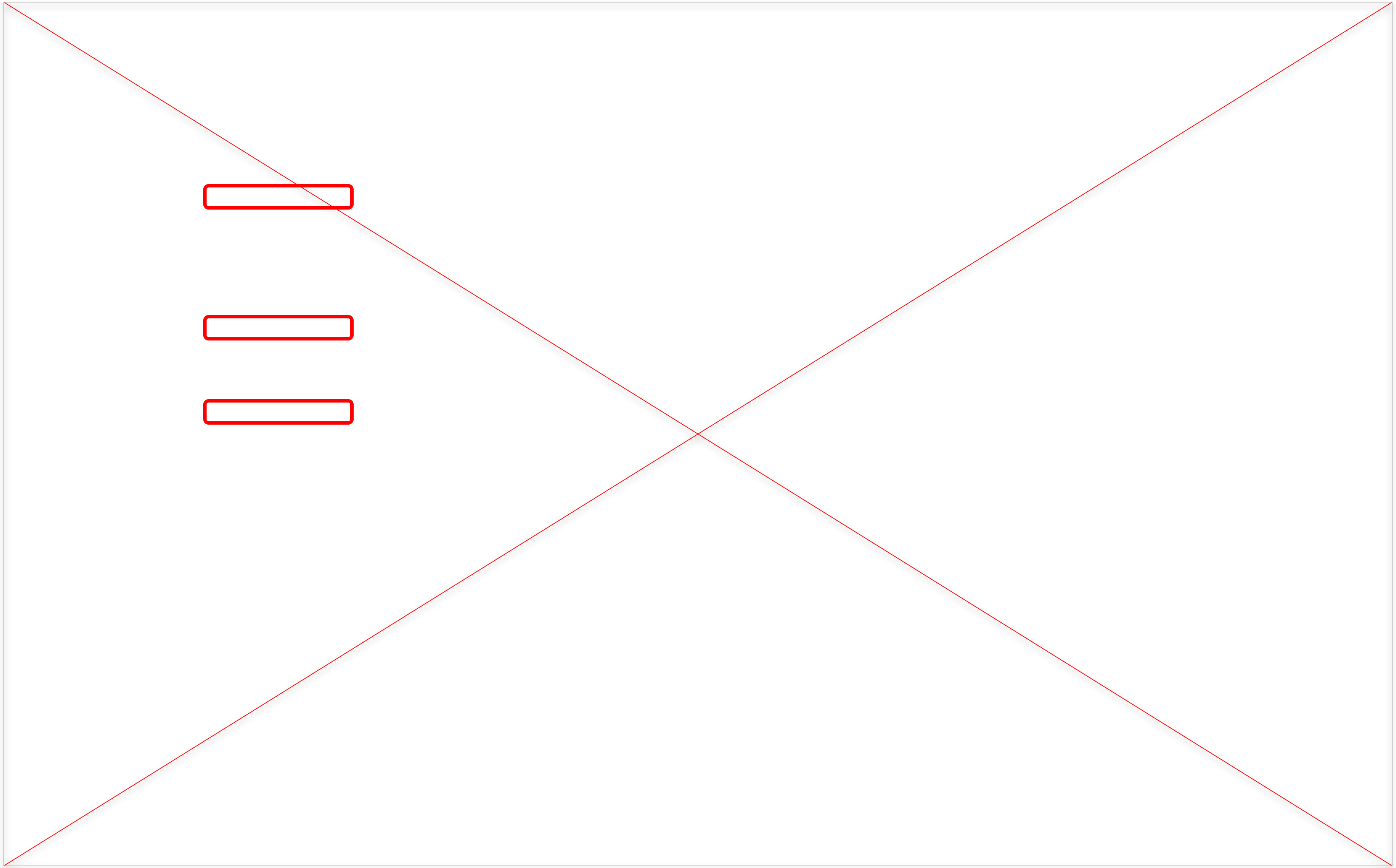
Get me a list of stores!

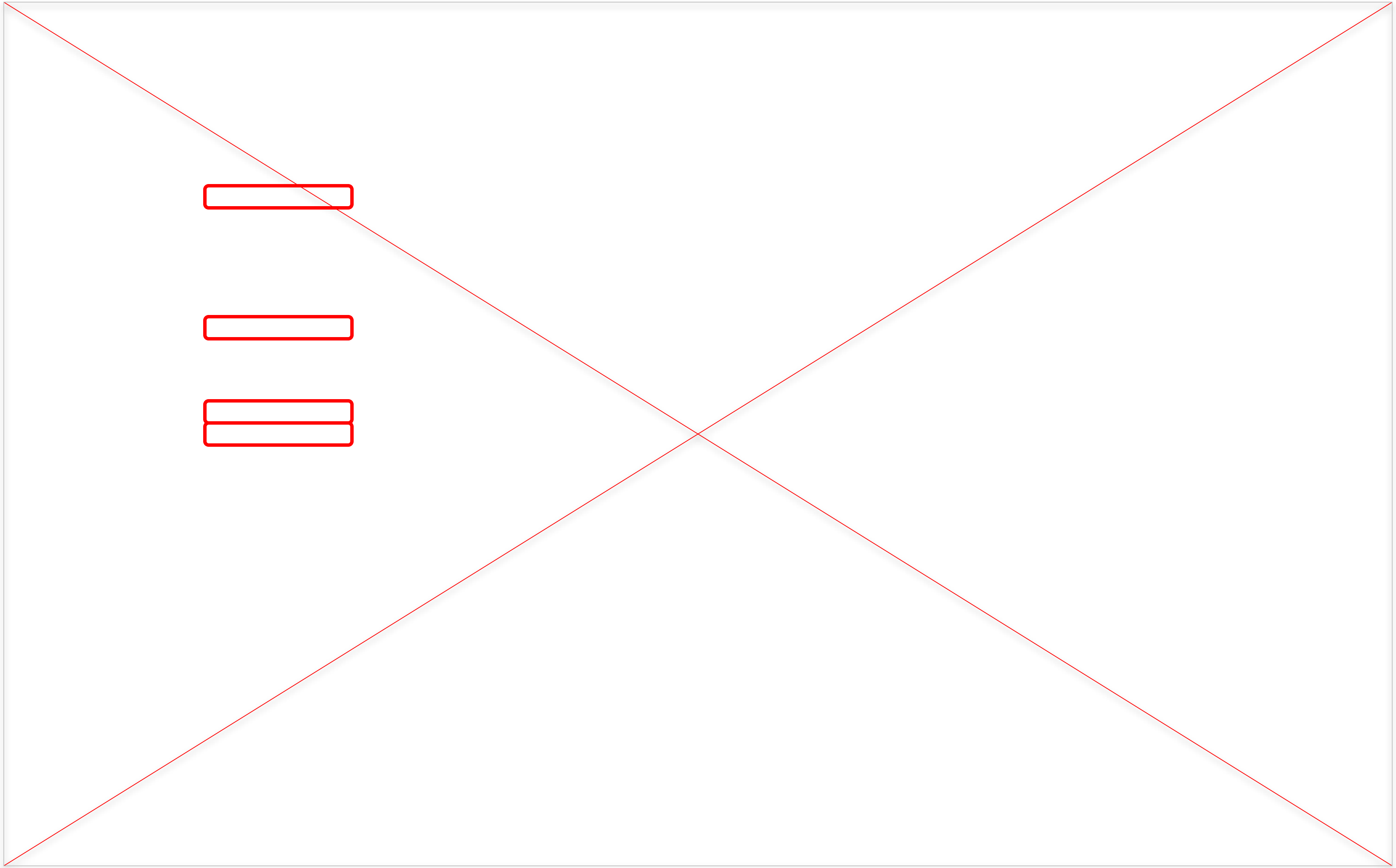
With details on their
location and hours!

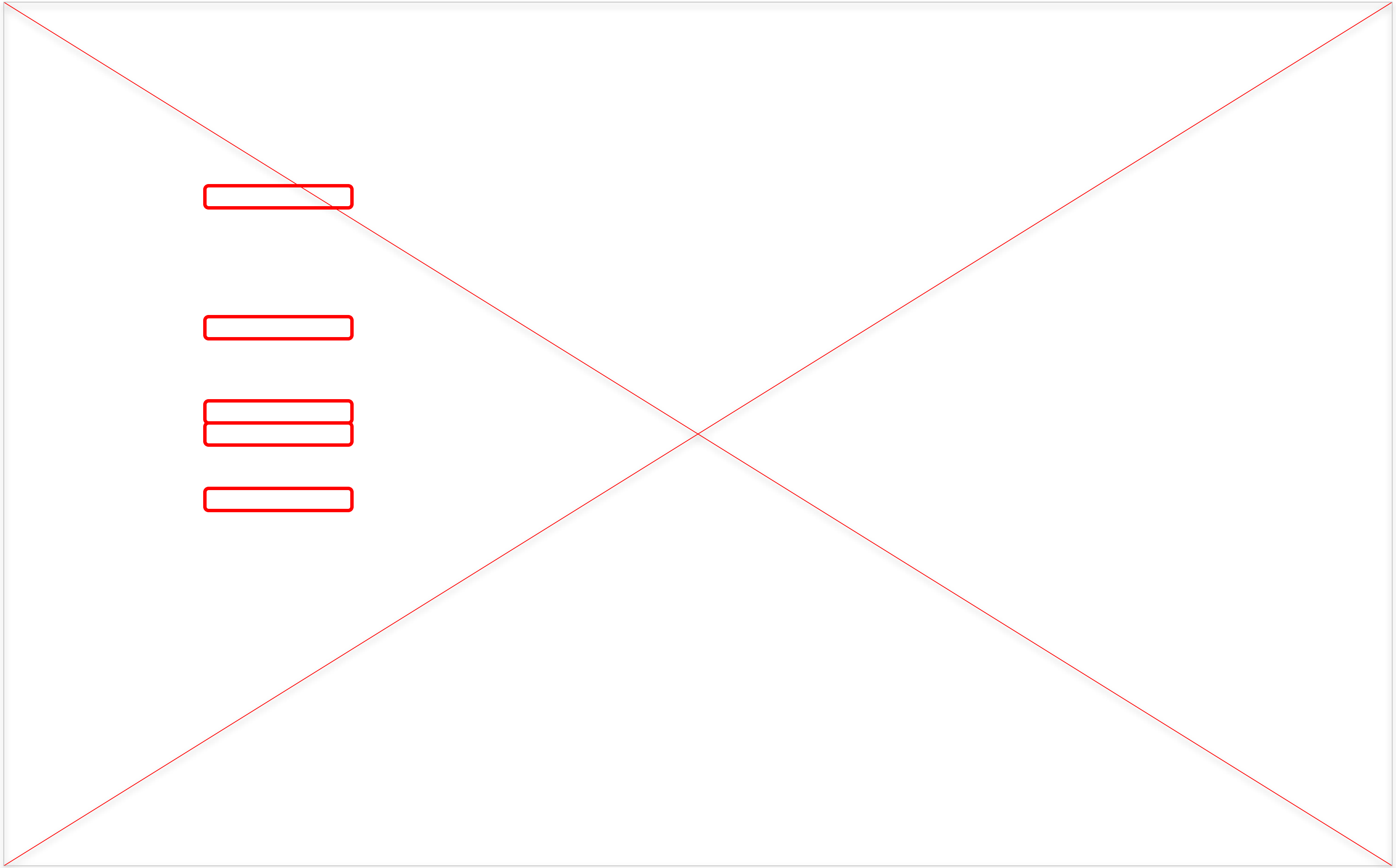
And data about sales
and operating costs!

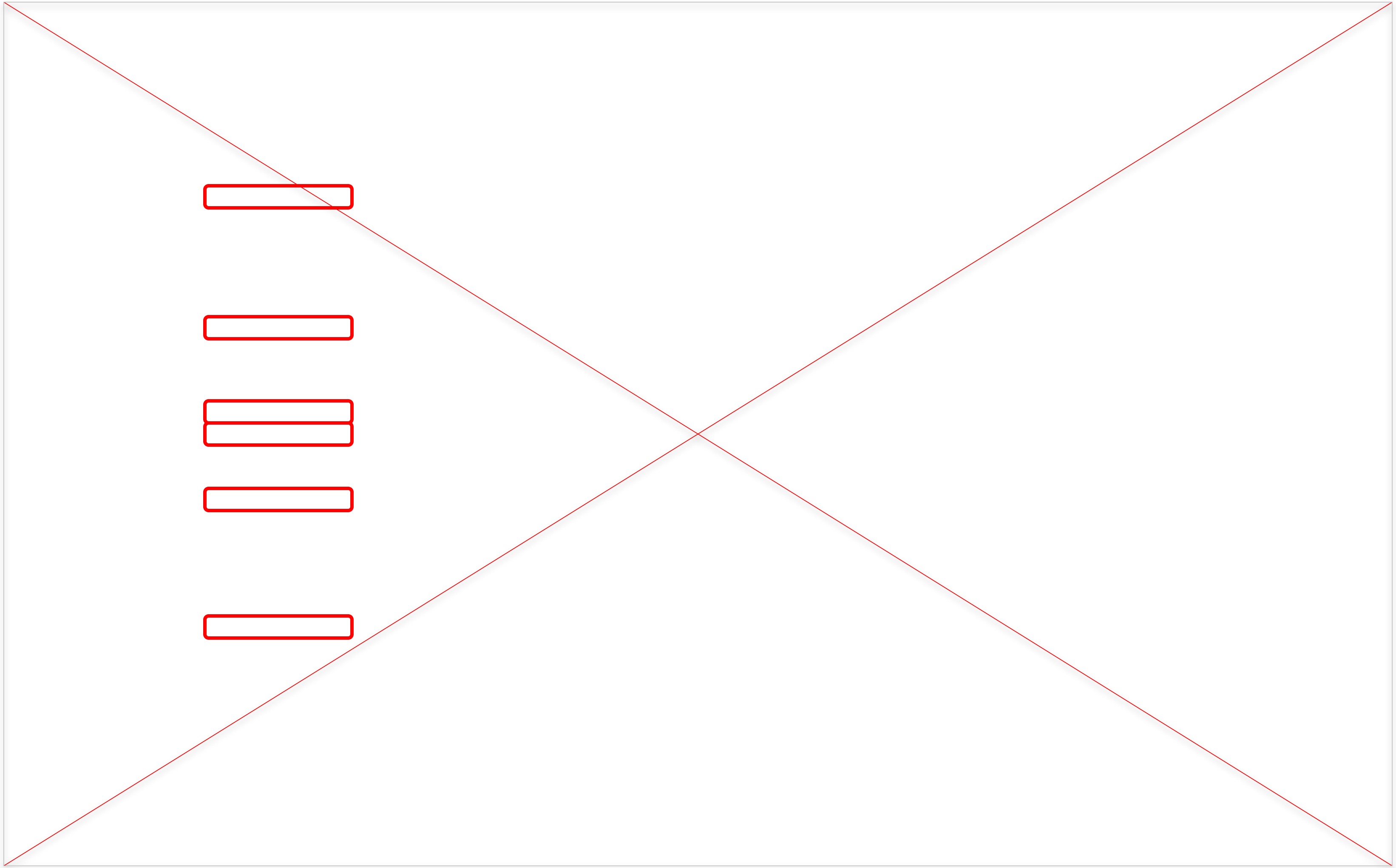
Get me a list of entities!!

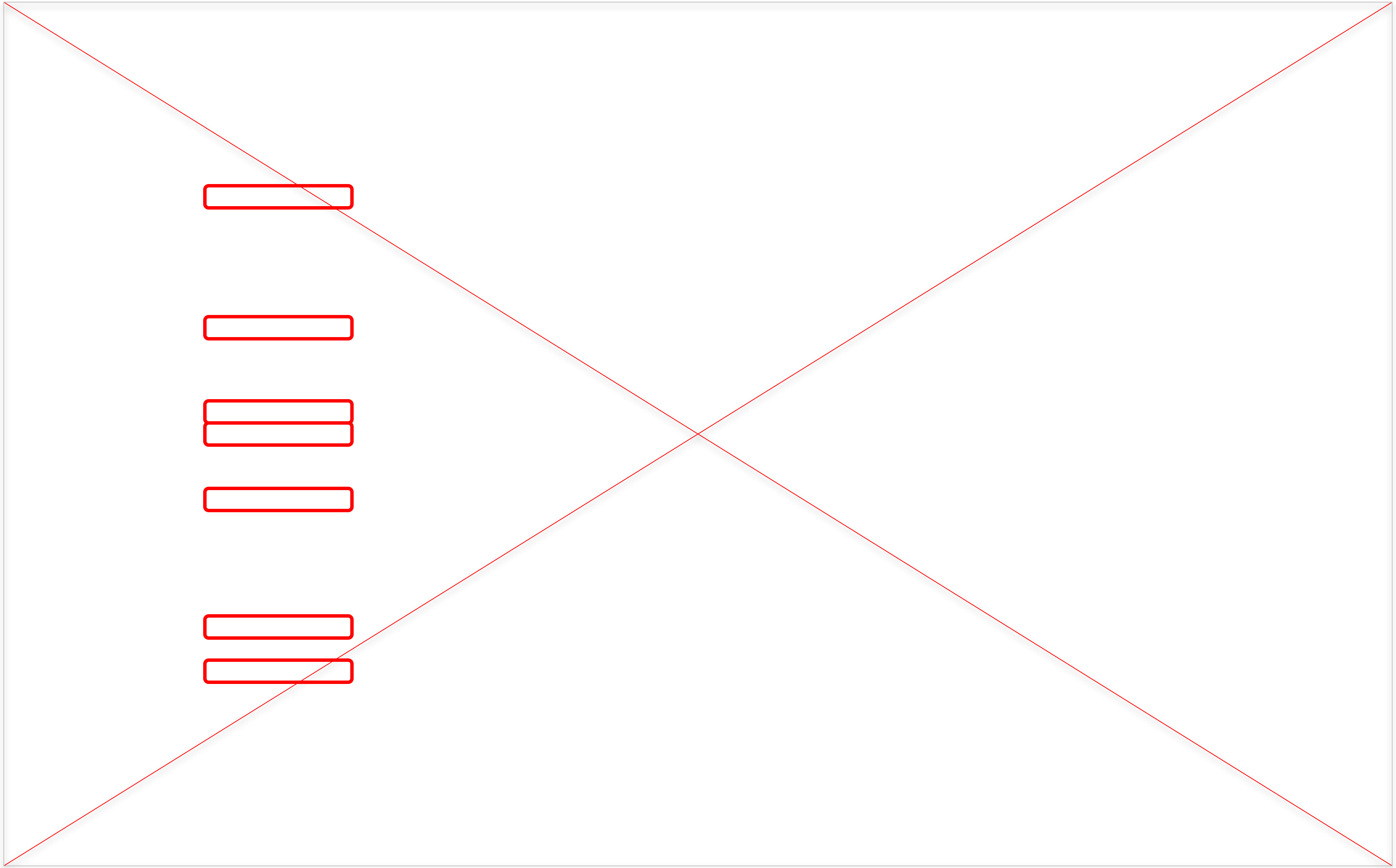


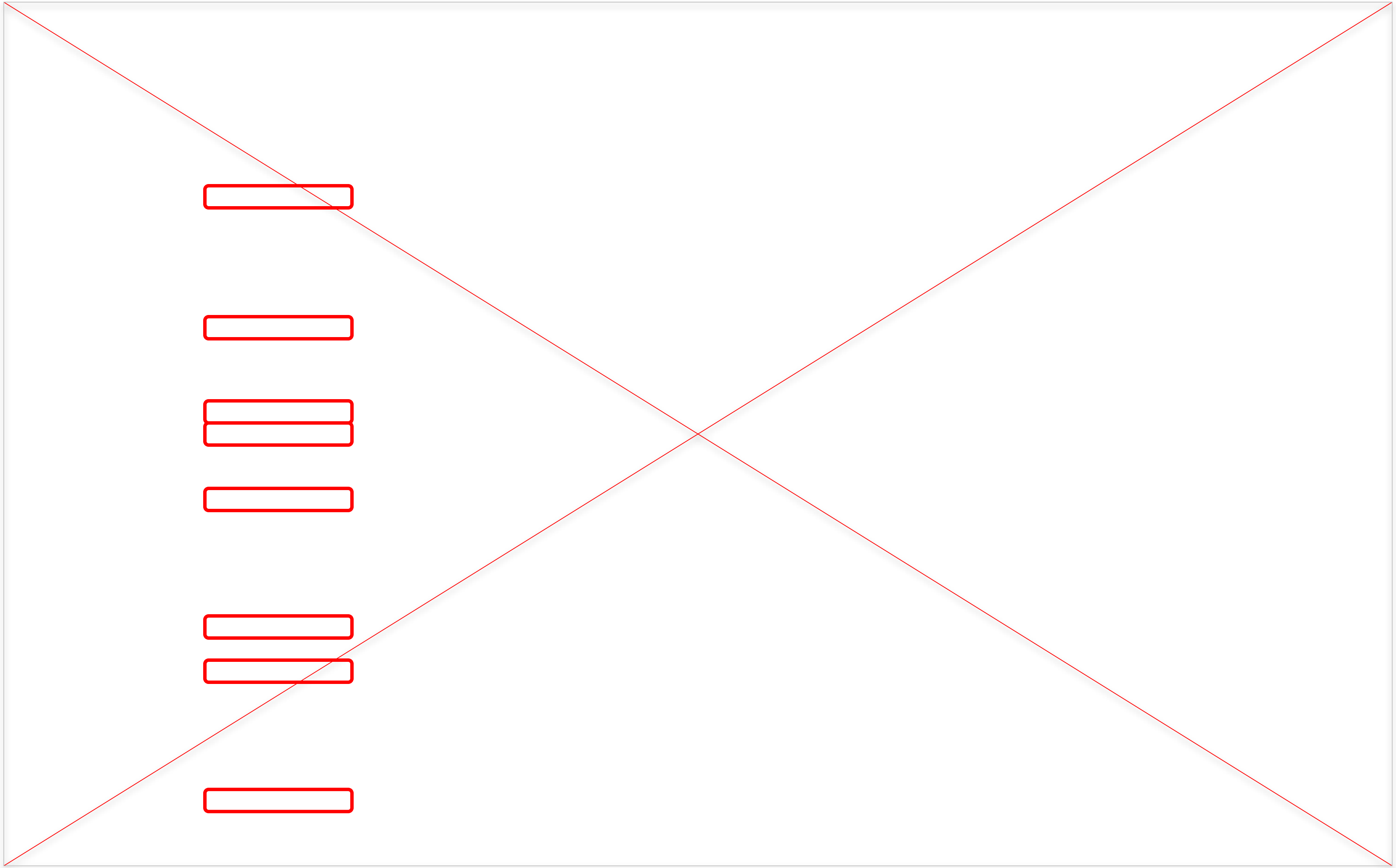










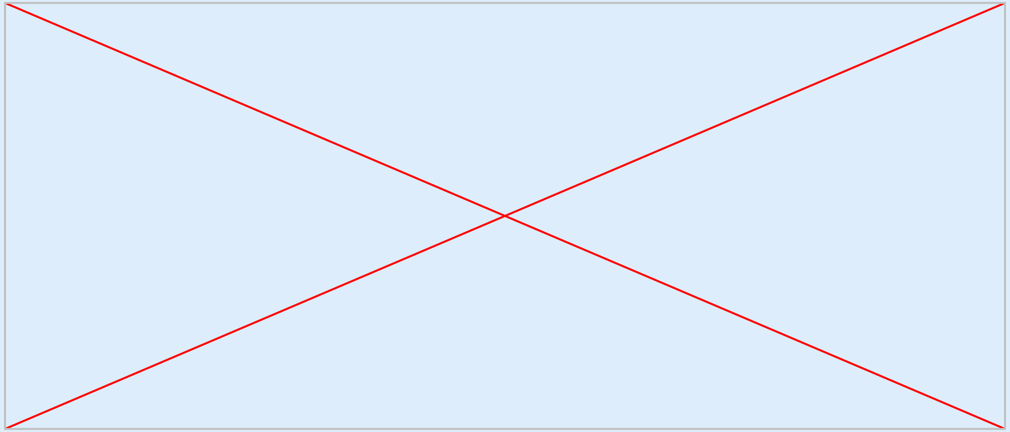


Raw data

OLAP cube

Reporting

OLAP cubes are hard to understand



snowflake®

Today…

11,952,200

combinations

Today…

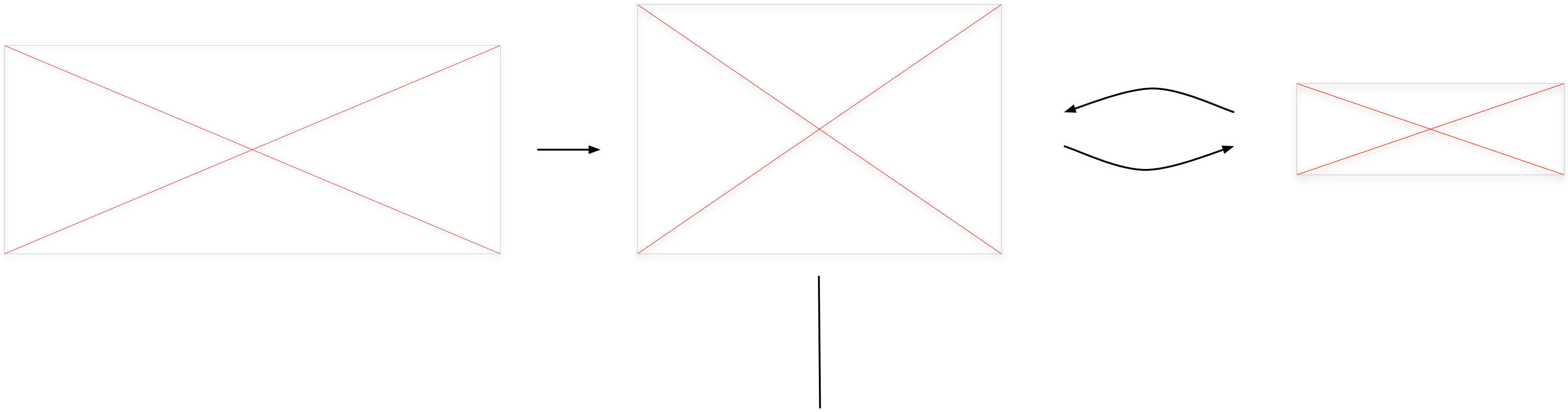
200,000,000

transactions

Raw data

OLAP cube

Reporting

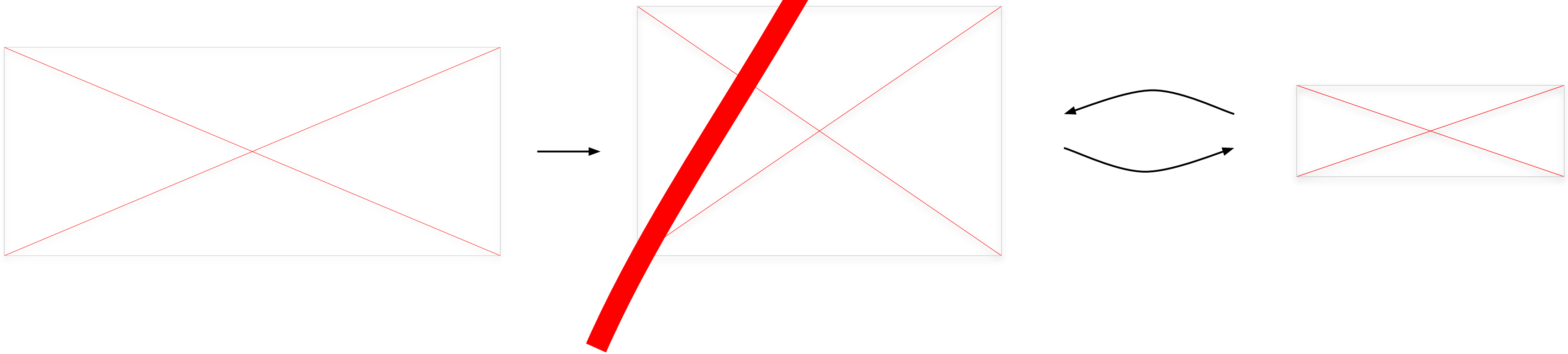


We don’t need to precompute this anymore

Raw data

~~OLAP cube~~

Reporting

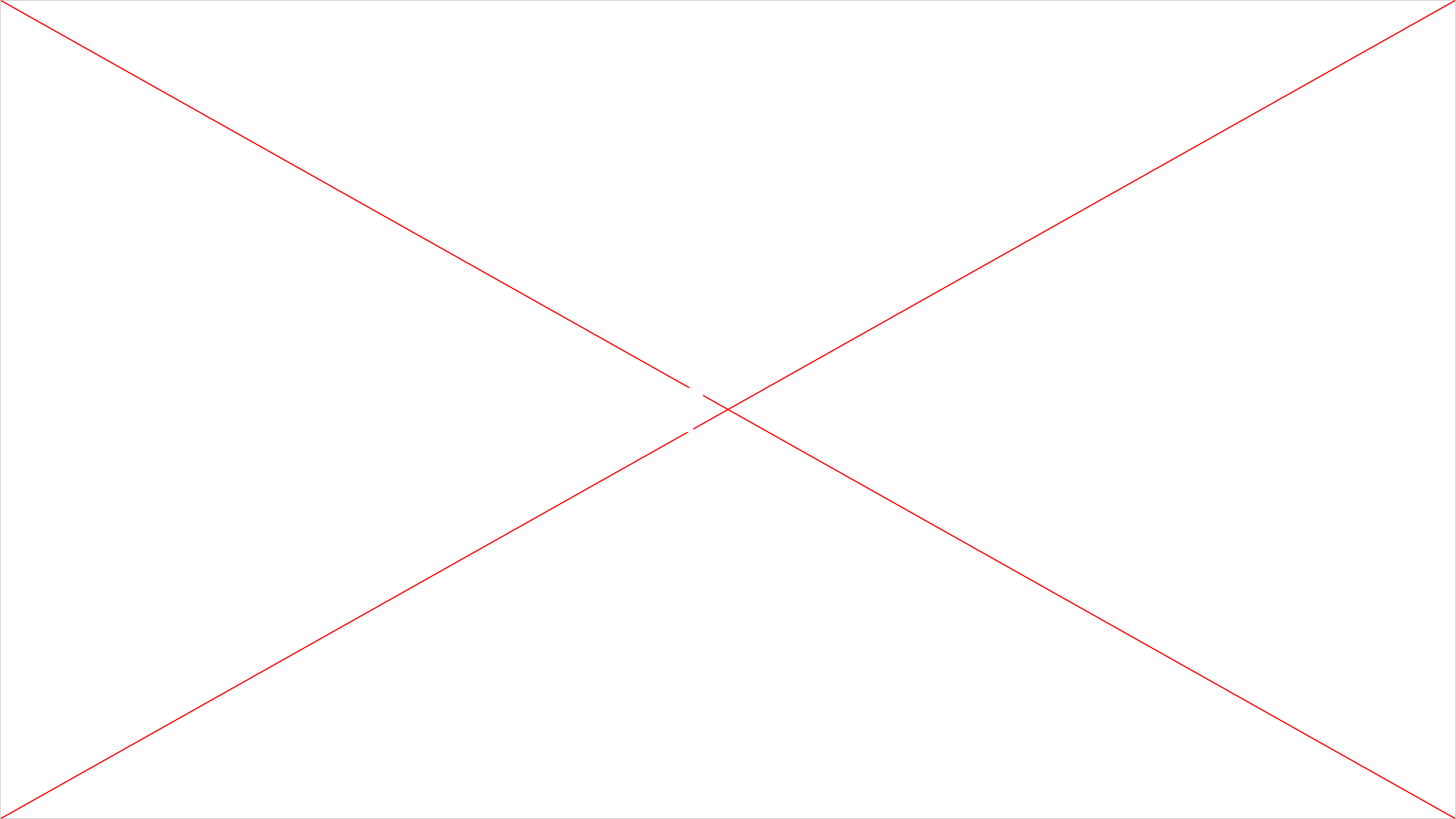


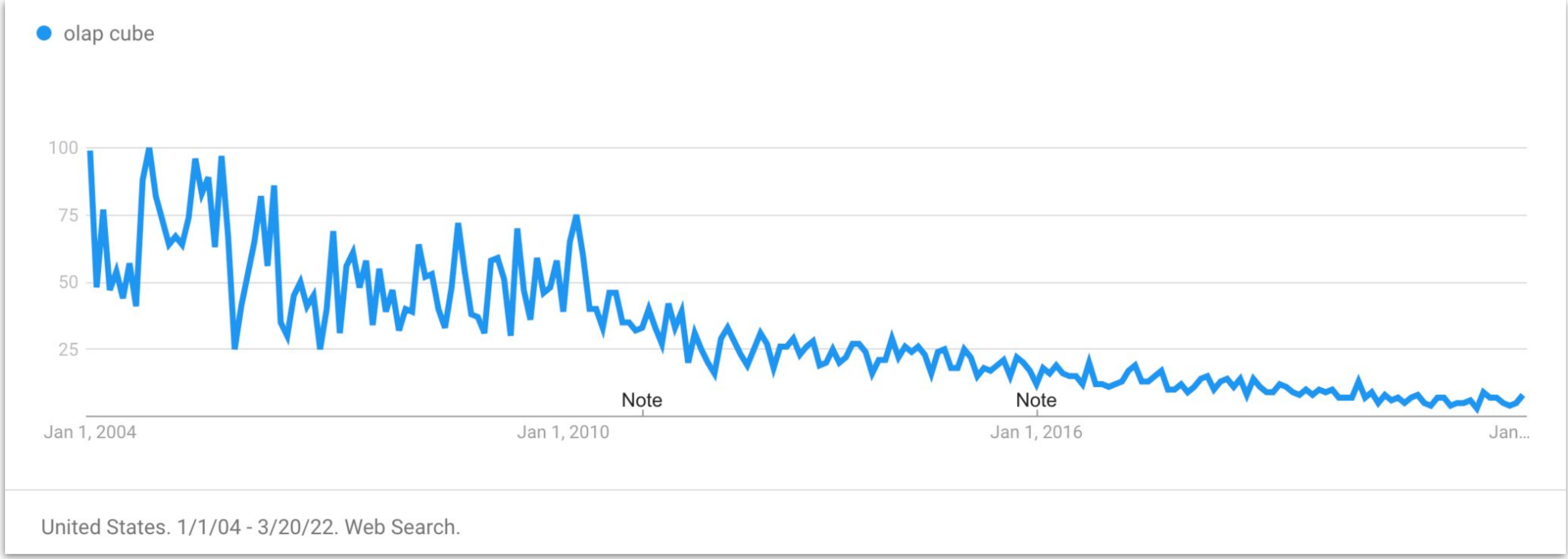
Raw data

~~**OLAP cube**~~

On-the-fly queries, directly against the database

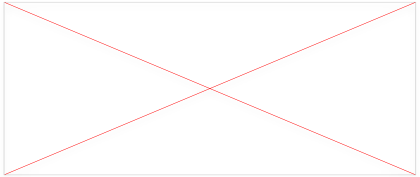
Reporting



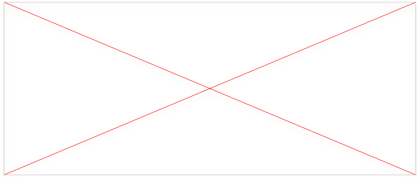

olap cube
100
75
50
25
Note
Note
Jan 1, 2004
Jan 1, 2010
Jan 1, 2016
Jan...
United States. 1/1/04 - 3/20/22. Web Search.

looker

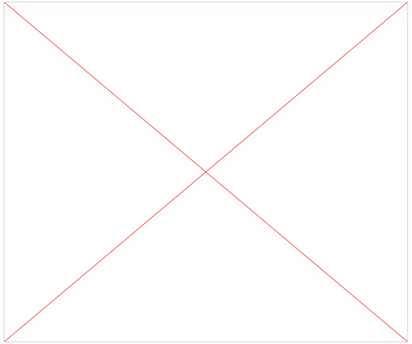
Raw data



Raw data

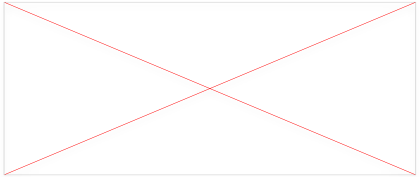


Data model

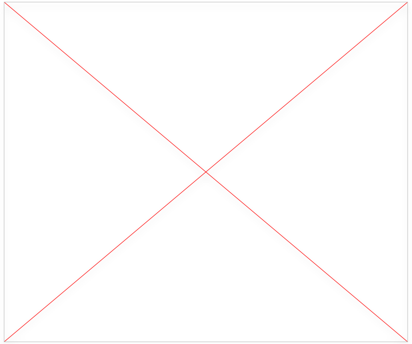


Configure relationships and metrics

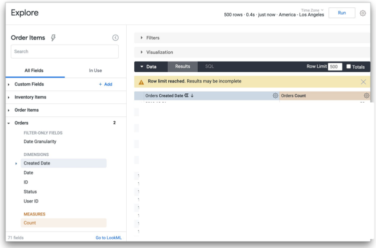
Raw data



Data model



Reporting

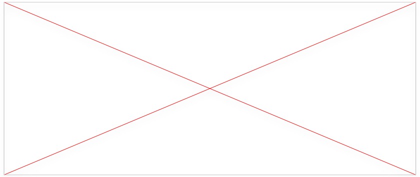


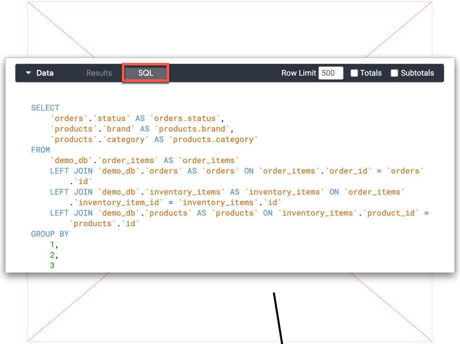
The model creates a UI that shows people what data they can use

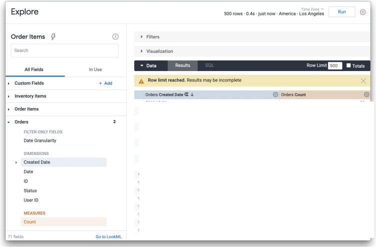
Raw data

Data model

Reporting

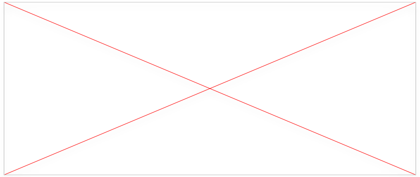




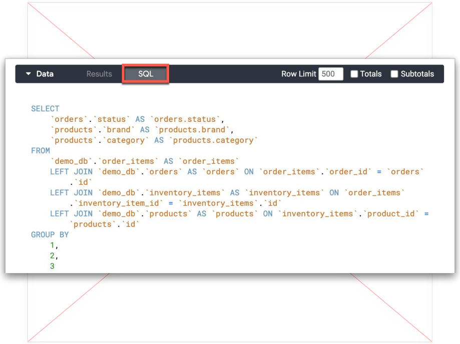


The model transforms requests into SQL queries

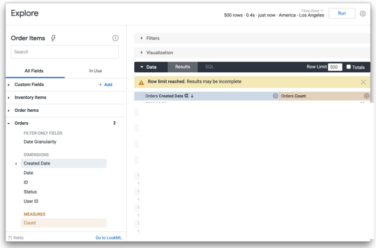
Raw data



Data model

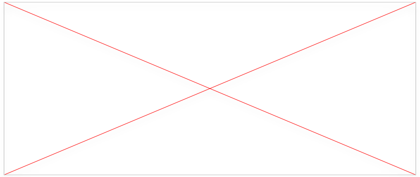


Reporting

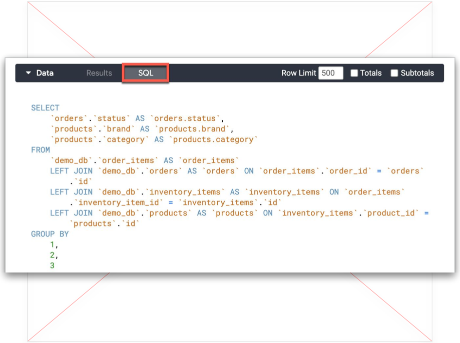


The query runs against raw data

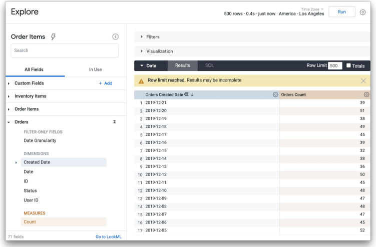
Raw data



Data model

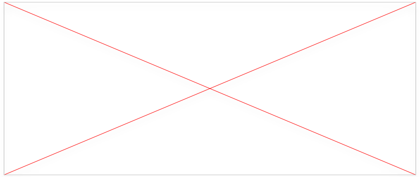


Reporting

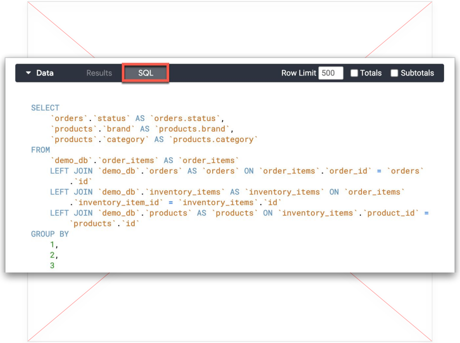


And results get returned

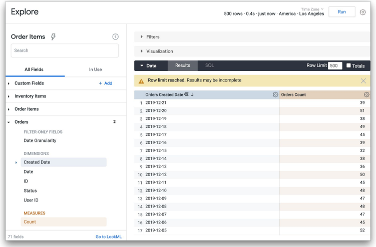
Raw data



Data model



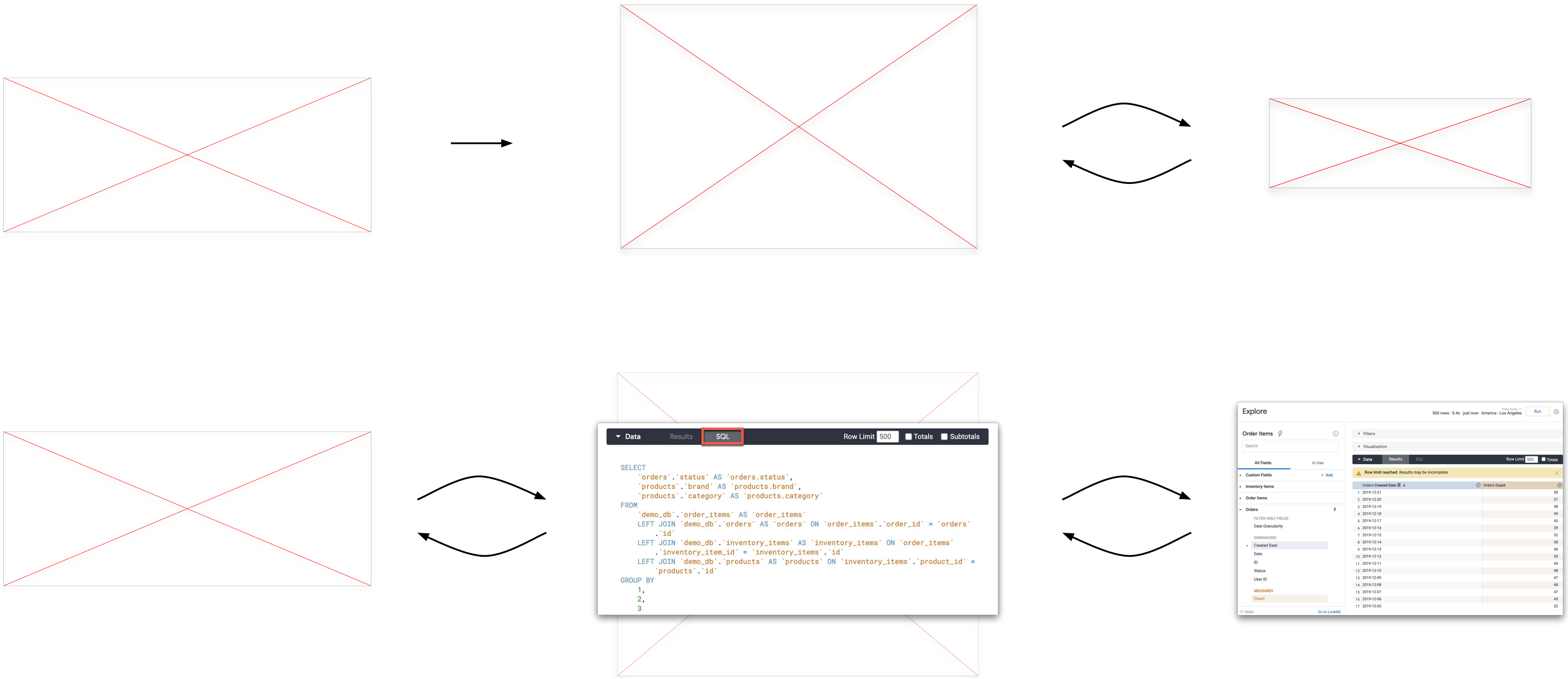
Reporting

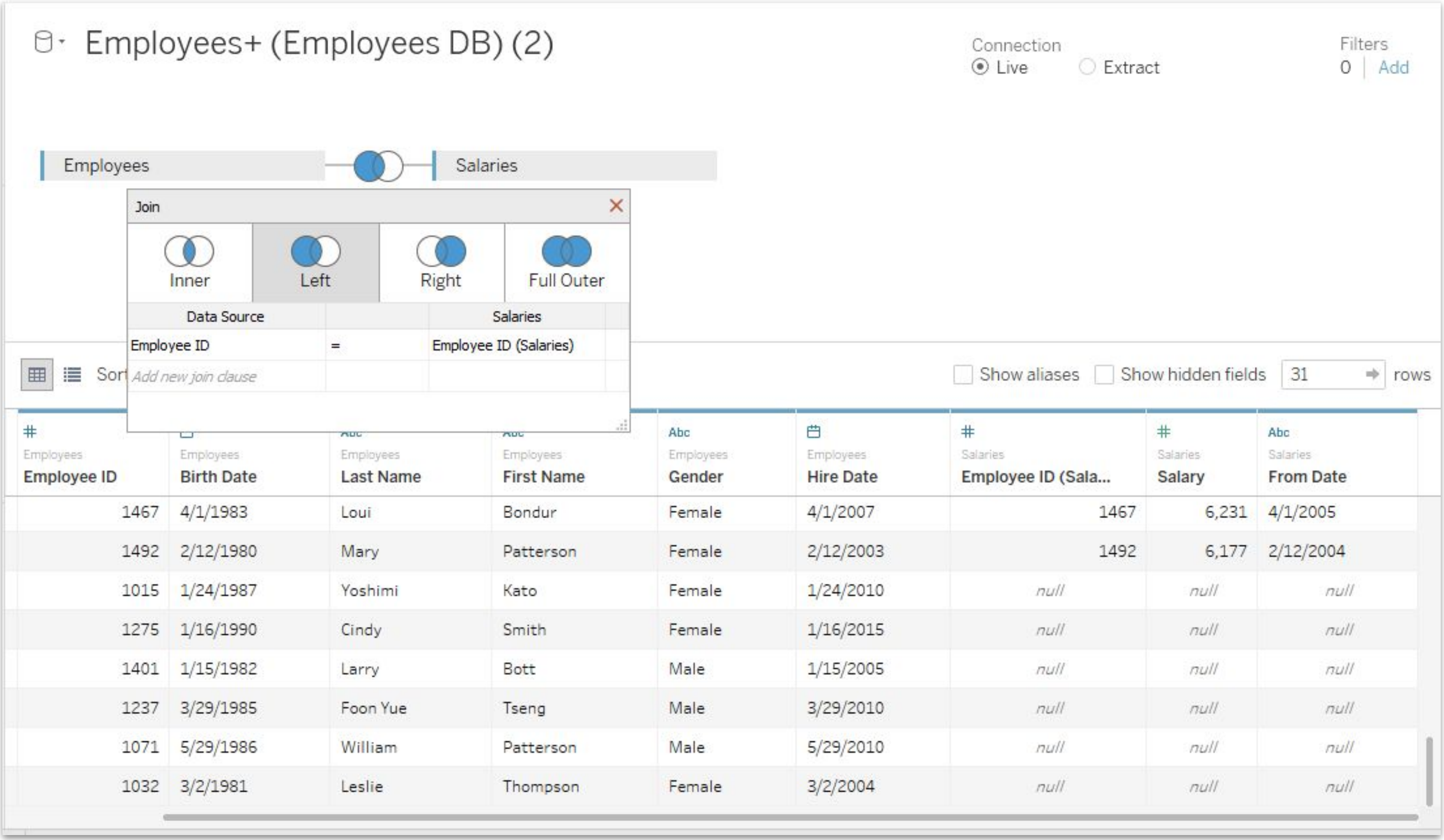


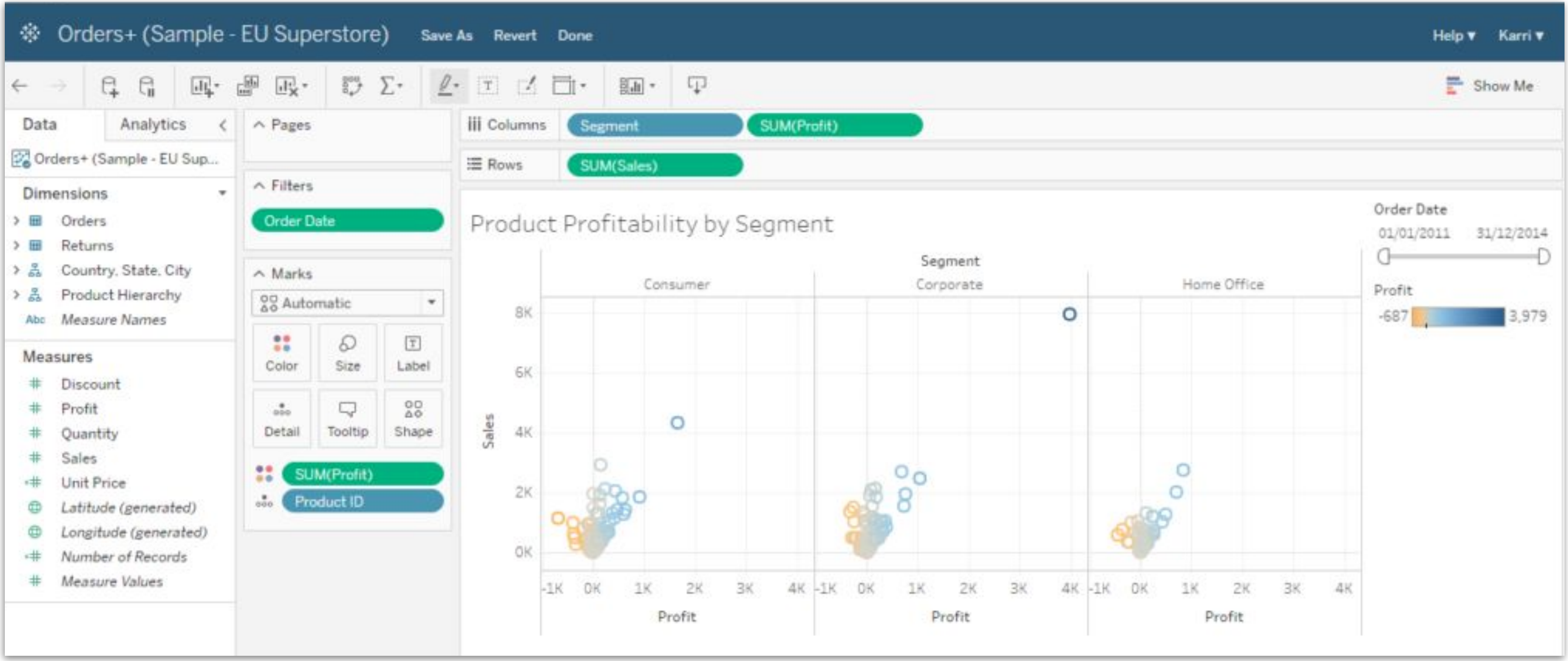
Raw data

OLAP cube?

Reporting




Employees+ (Employees DB) (2)
Connection
Live
Extract
Filters
0
Add
Employees
Salaries
Join
Inner
Left
Right
Full Outer
Data Source
Salaries
Employee ID
=
Employee ID (Salaries)
Add new join clause
Sort
Show aliases
Show hidden fields
31
rows
Employees
Employee ID
Employees
Birth Date
Employees
Last Name
Employees
First Name
Employees
Gender
Employees
Hire Date
Salaries
Employee ID (Sala...
Salaries
Salary
Salaries
From Date
1467 4/1/1983 Loui Bondur Female 4/1/2007 1467 6,231 4/1/2005
1492 2/12/1980 Mary Patterson Female 2/12/2003 1492 6,177 2/12/2004
1015 1/24/1987 Yoshimi Kato Female 1/24/2010 null null null
1275 1/16/1990 Cindy Smith Female 1/16/2015 null null null
1401 1/15/1982 Larry Bott Male 1/15/2005 null null null
1237 3/29/1985 Foon Yue Tseng Male 3/29/2010 null null null
1071 5/29/1986 William Patterson Male 5/29/2010 null null null
1032 3/2/1981 Leslie Thompson Female 3/2/2004 null null null


Orders+ (Sample - EU Superstore)
Save As
Revert
Done
Help
Karri
Show Me
Data
Analytics
Orders+ (Sample - EU Sup...
Dimensions
Orders
Returns
Country, State, City
Product Hierarchy
Measure Names
Measures
Discount
Profit
Quantity
Sales
Unit Price
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Order Date
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
SUM(Profit)
Product ID
Columns
Segment
SUM(Profit)
Rows
SUM(Sales)
Product Profitability by Segment
Segment
Consumer
Corporate
Home Office
8K
6K
4K
2K
0K
Sales
-1K
0K
1K
2K
3K
4K
Profit
Order Date
01/01/2011
31/12/2014
Profit
-687
3,979

Raw data

OLAP cube?

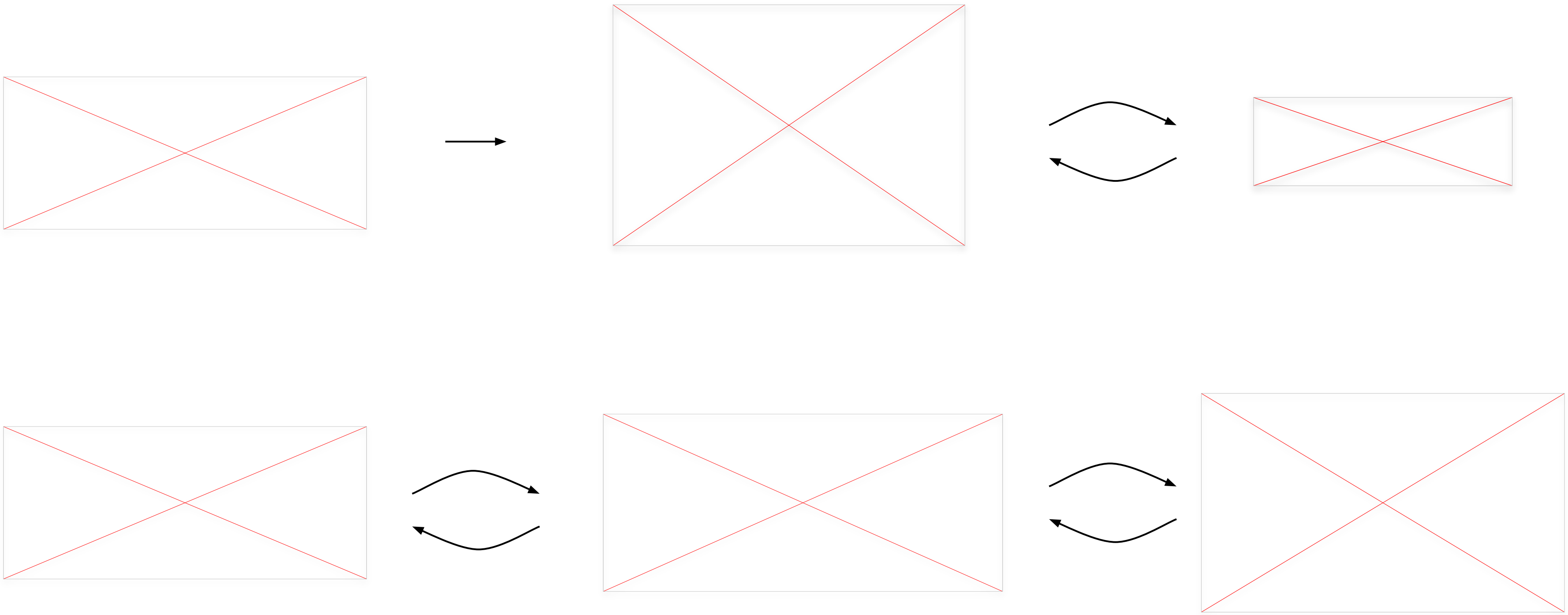
Reporting

MODE

Raw data

OLAP cube?

Reporting



transform

```yaml
name: organizations
description: This datasource is sourced from the demo.orgs table. Each row in this t
             represents an organization. Each user is associated with an organizatio
             can be active with paid or unpaid, has a billing and a usage country, h
             and each organization can be work, social, or education.
owners:
  - support@transformdata.io
sql_table: demo.orgs

identifiers:
  - name: org
    type: primary
    expr: org_id
```

```python
import pandas
from transform import mql


metrics: List[mql.Metric] = mql.list_metrics()

df: pandas.DataFrame = mql.create_query(
  metrics=["rainfall"],
  dimensions=["ds", "country"]
)
```

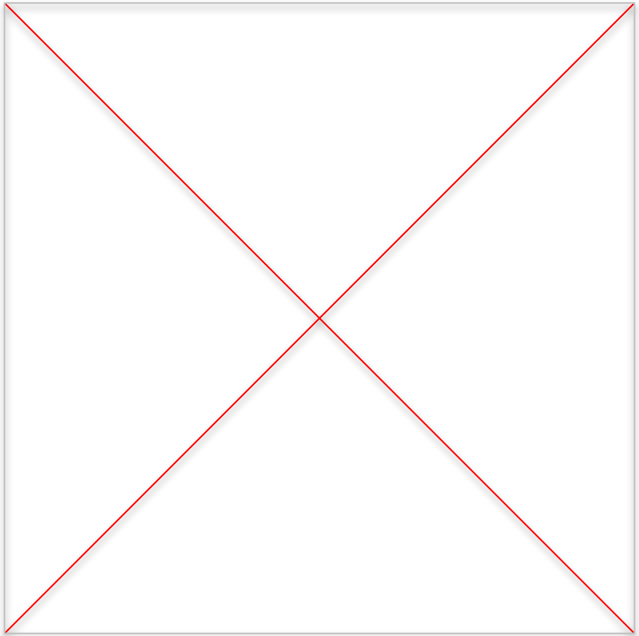
Raw data

OLAP cube?

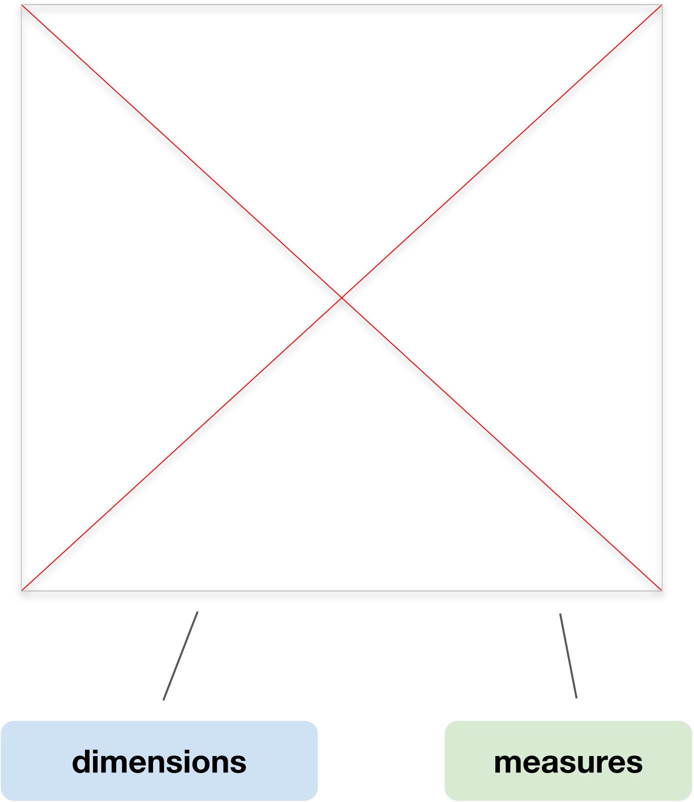
Reporting

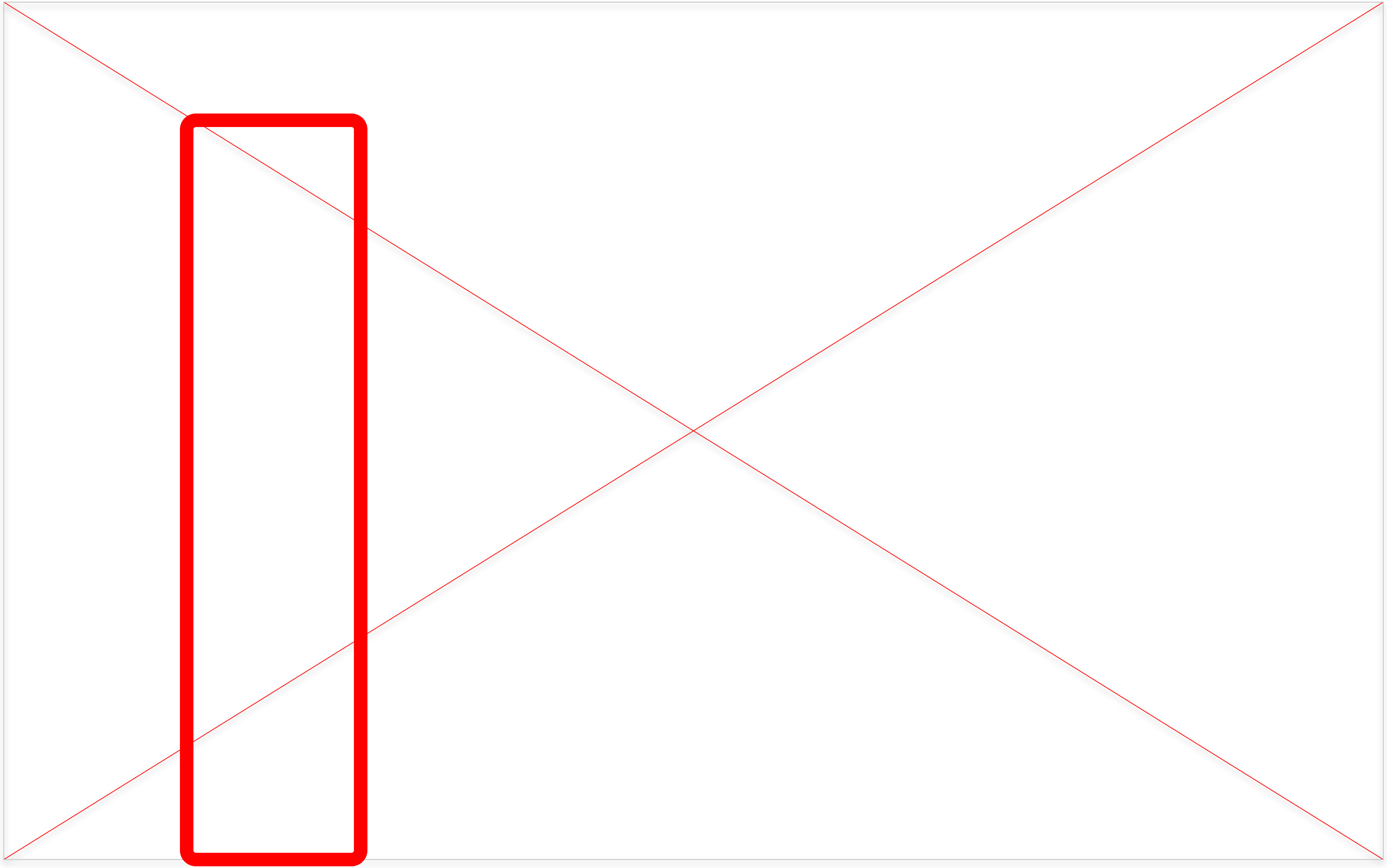
```yaml
name: organizations
description: This datasource is sourced from the demo.orgs table. Each row in this t
             represents an organization. Each user is associated with an organizatio
             can be active with paid or unpaid, has a billing and a usage country, h
             and each organization can be work, social, or education.
owners:
  - support@transformdata.io
sql_table: demo.orgs

identifiers:
  - name: org
    type: primary
    expr: org_id
```

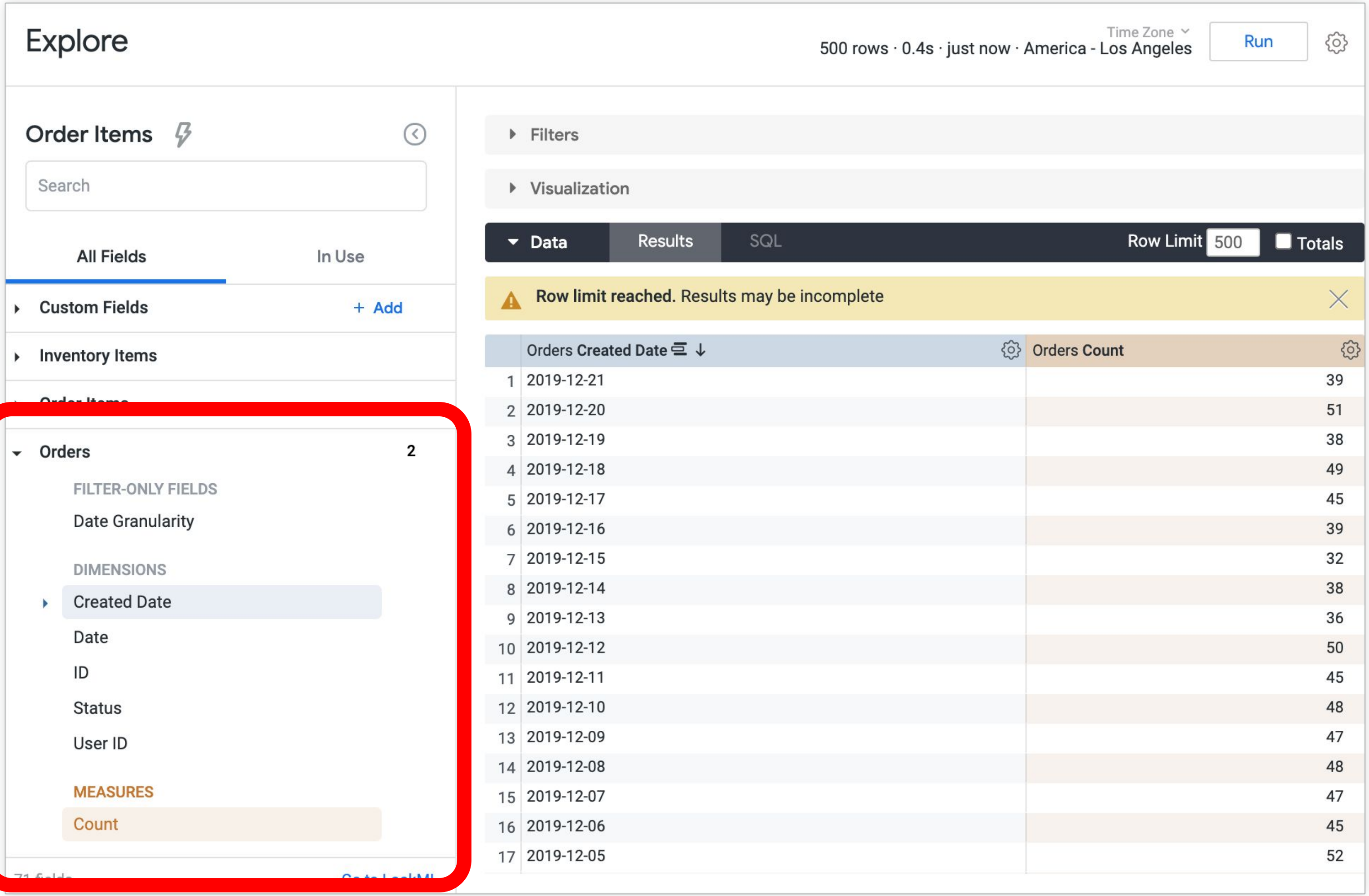
```python
import pandas
from transform import mql


metrics: List[mql.Metric] = mql.list_metrics()

df: pandas.DataFrame = mql.create_query(
  metrics=["rainfall"],
  dimensions=["ds", "country"]
)
```

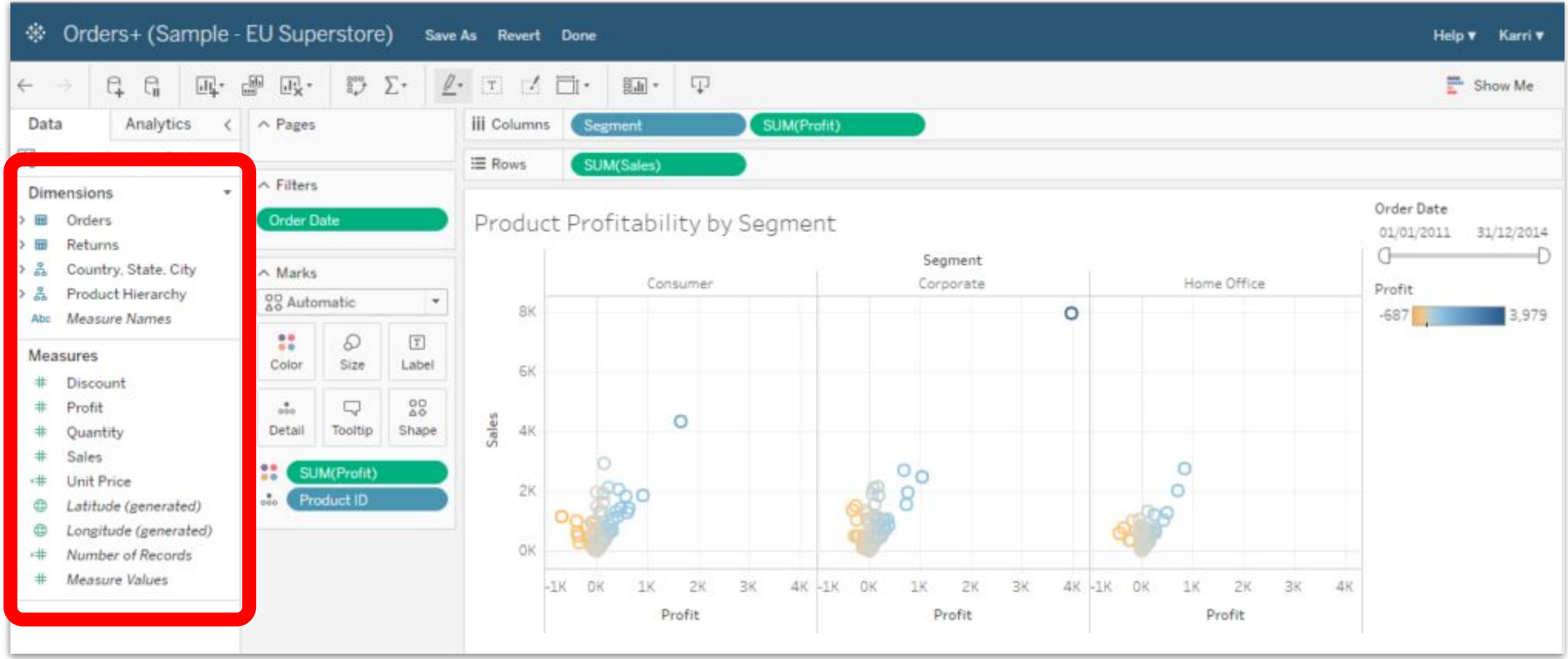


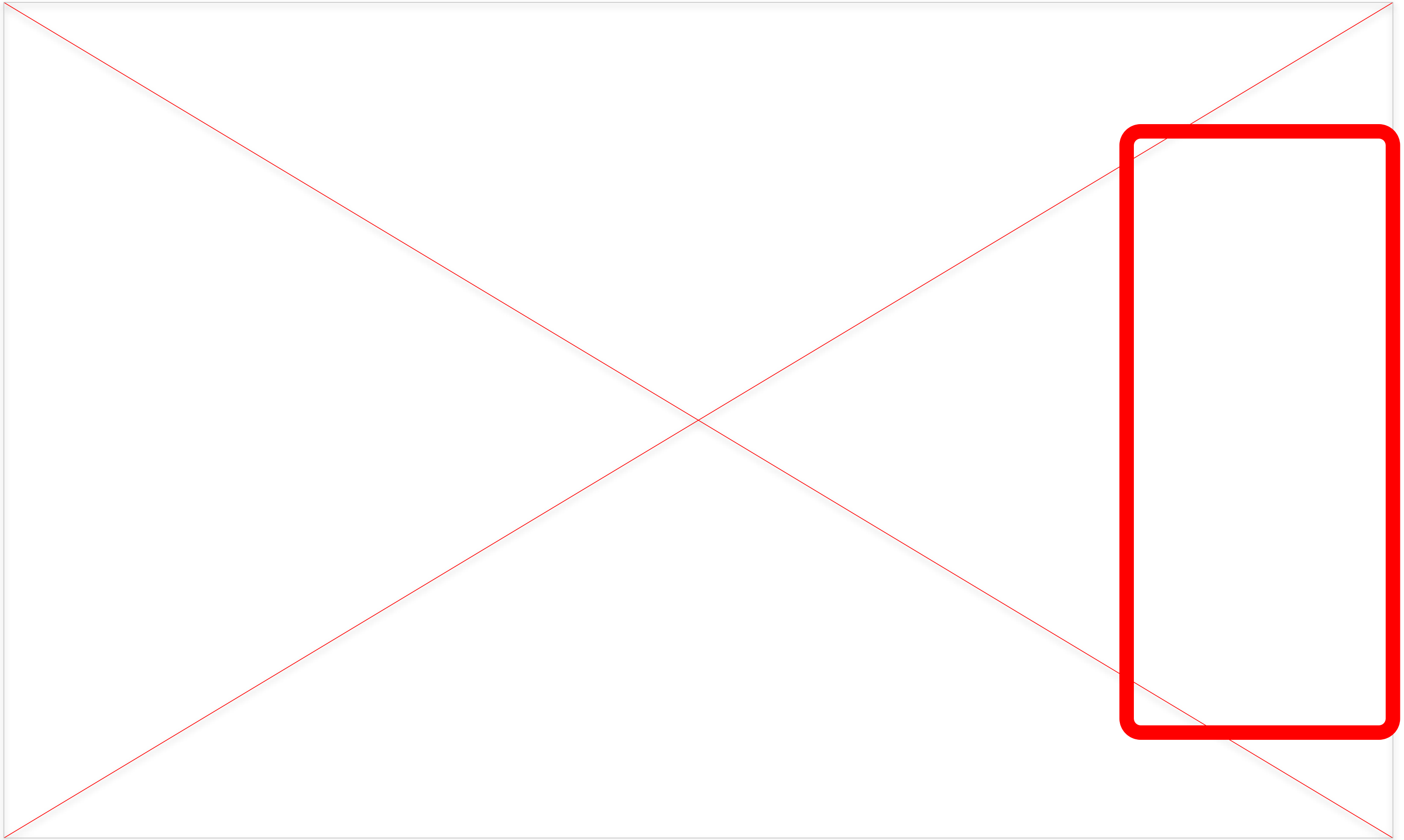
You ship your org chart.


dimensions
measures

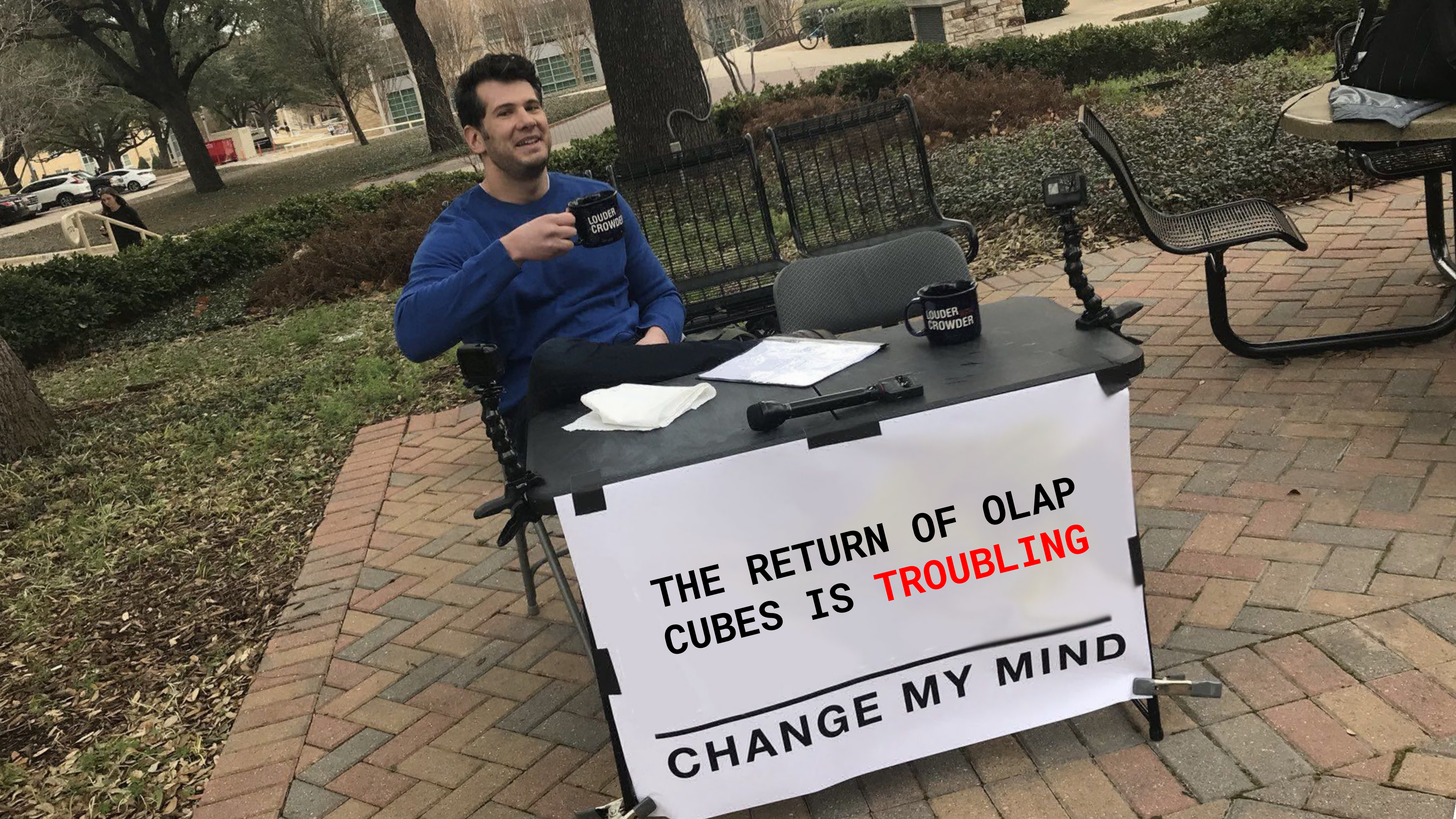



Explore
Time Zone
500 rows · 0.4s · just now · America - Los Angeles
Run
Order Items
Search
All Fields
In Use
Custom Fields
+ Add
Inventory Items
Orders
2
FILTER-ONLY FIELDS
Date Granularity
DIMENSIONS
Created Date
Date
ID
Status
User ID
MEASURES
Count
Filters
Visualization
Data
Results
SQL
Row Limit
500
Totals
Row limit reached. Results may be incomplete
Orders Created Date
Orders Count
1 2019-12-21 39
2 2019-12-20 51
3 2019-12-19 38
4 2019-12-18 49
5 2019-12-17 45
6 2019-12-16 39
7 2019-12-15 32
8 2019-12-14 38
9 2019-12-13 36
10 2019-12-12 50
11 2019-12-11 45
12 2019-12-10 48
13 2019-12-09 47
14 2019-12-08 48
15 2019-12-07 47
16 2019-12-06 45
17 2019-12-05 52


Orders+ (Sample - EU Superstore)
Save As
Revert
Done
Help
Karri
Show Me
Data
Analytics
Pages
Columns
Segment
SUM(Profit)
Rows
SUM(Sales)
Filters
Order Date
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
SUM(Profit)
Product ID
Dimensions
Orders
Returns
Country, State, City
Product Hierarchy
Measure Names
Measures
Discount
Profit
Quantity
Sales
Unit Price
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Product Profitability by Segment
Segment
Consumer
Corporate
Home Office
Sales
Profit
Order Date
01/01/2011
31/12/2014
Profit
-687
3,979



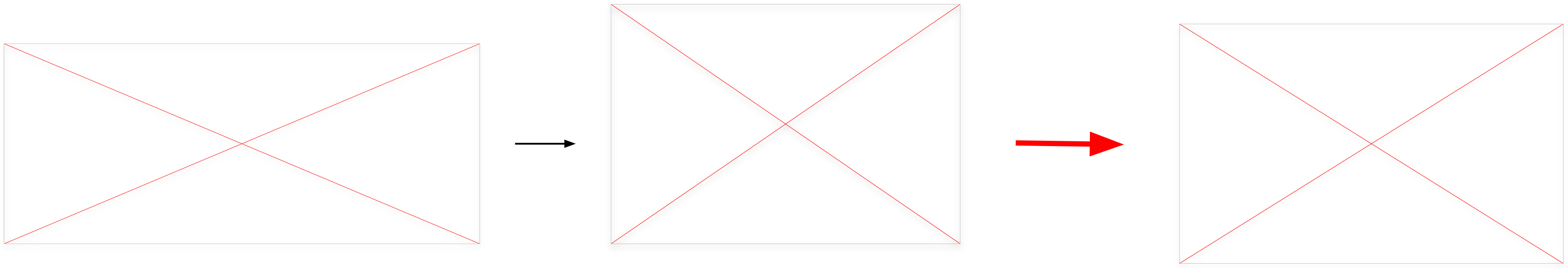
```python
import pandas
from transform import mql


metrics: List[mql.Metric] = mql.list_metrics()

df: pandas.DataFrame = mql.create_query(
  metrics=["rainfall"],
  dimensions=["ds", "country"]
)
```


THE RETURN OF OLAP
CUBES IS TROUBLING
CHANGE MY MIND
LOUDER CROWDER

Raw data

OLAP cube

Reporting



Raw data

OLAP cube

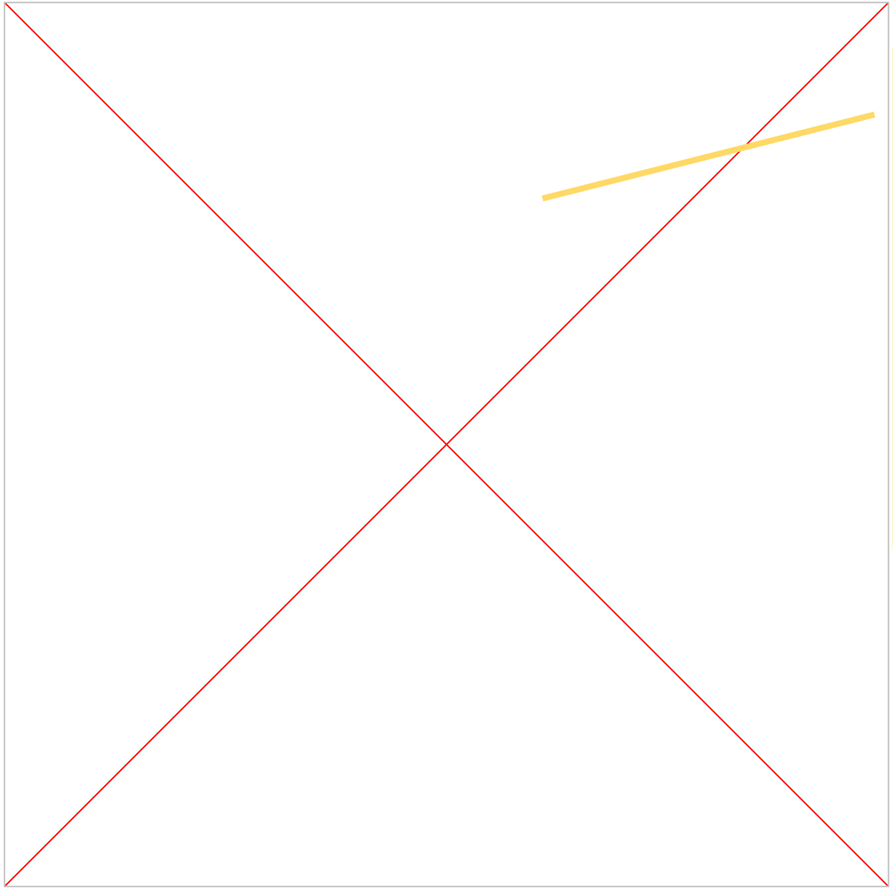
Reporting

Raw data

OLAP cube

Reporting

Get me
a list of
entities!!

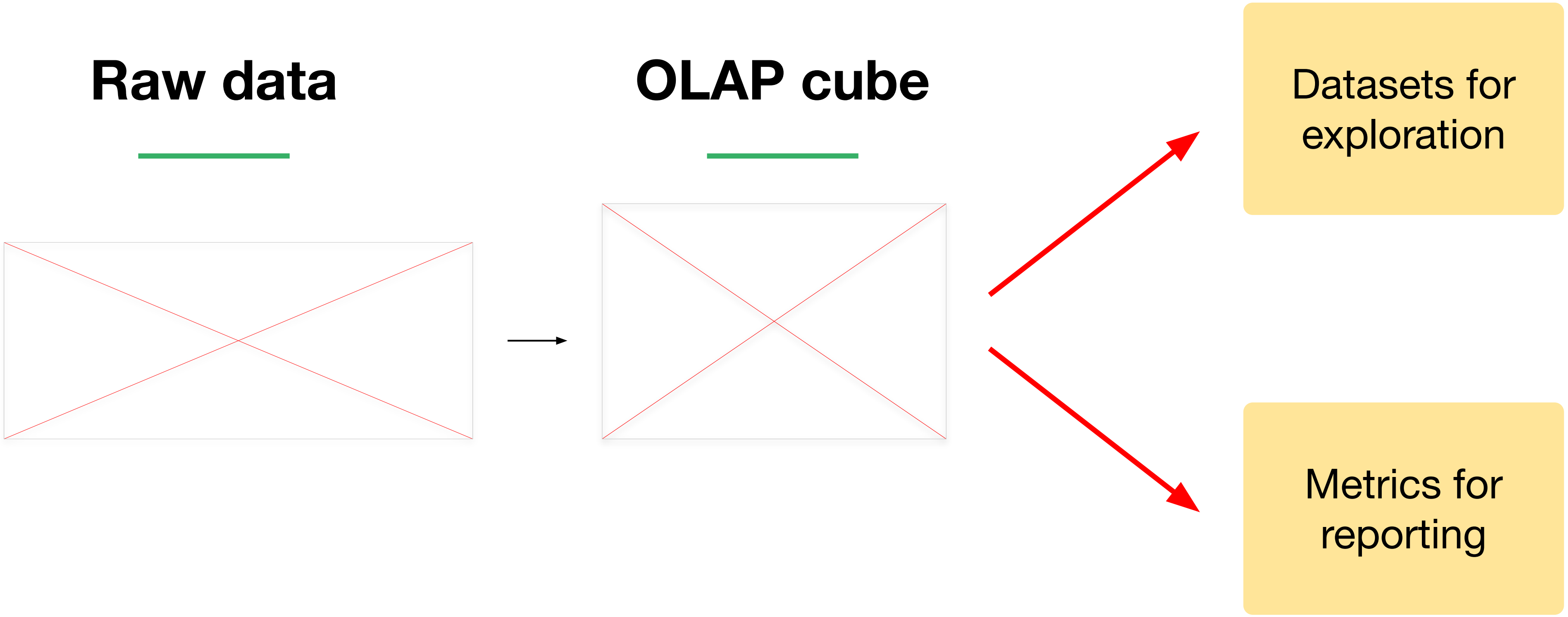


Get me a metric!!

Raw data

OLAP cube

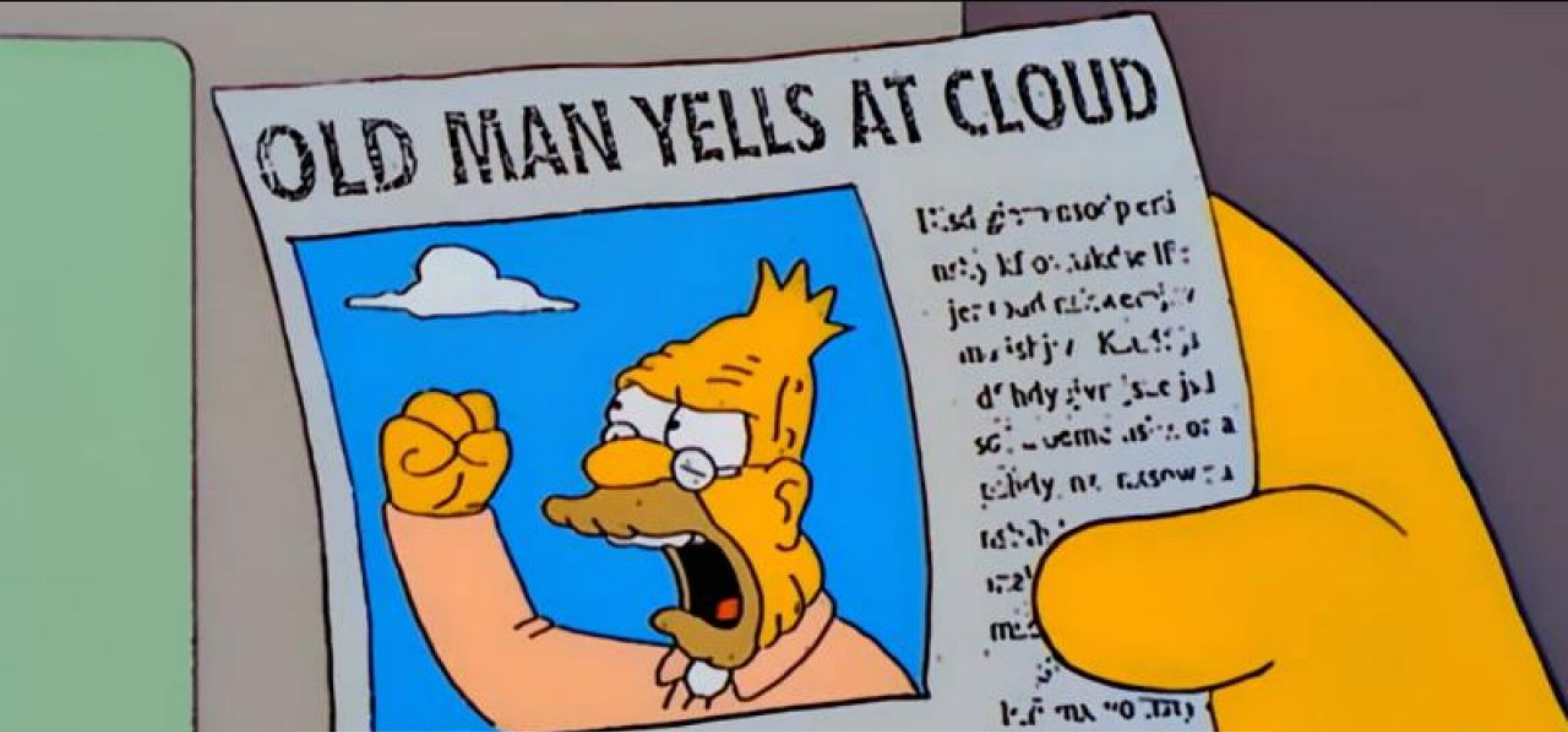
Datasets for exploration


Raw data
OLAP cube
Datasets for exploration
Metrics for reporting

Backup


Lane Open
U-Scan Genesis
Cash In
Change Out
MAGIC BONE

data pulls



benn.substack.com

Self-serve is a feeling

Lots of houses can be made a home.

Jul 9, 2021 ♡ 11

Why is self-serve still a problem?

We're not going to solve it until we define it.

Apr 8, 2021 ♡ 11 8

BI is dead

How an integration between Looker and Tabl

Exploration

Exploration

Reporting


1859
1909
Today
FUSION
Real build. Virtual game.

Get me a list of stores!

Get me a list of stores!

With details on their location and hours!

Get me a list of stores!

With details on their location and hours!

And data about sales and operating costs!

Entity

Get me a list of stores!

With details on their location and hours!

And data about sales and operating costs!

Entity	Get me a list of stores!
Dimensions	With details on their location and hours!
	And data about sales and operating costs!

Entity	Get me a list of stores!
Dimensions	With details on their location and hours!
Metrics	And data about sales and operating costs!

Entity Get me a list of stores!

Dimensions With details on their location and hours!

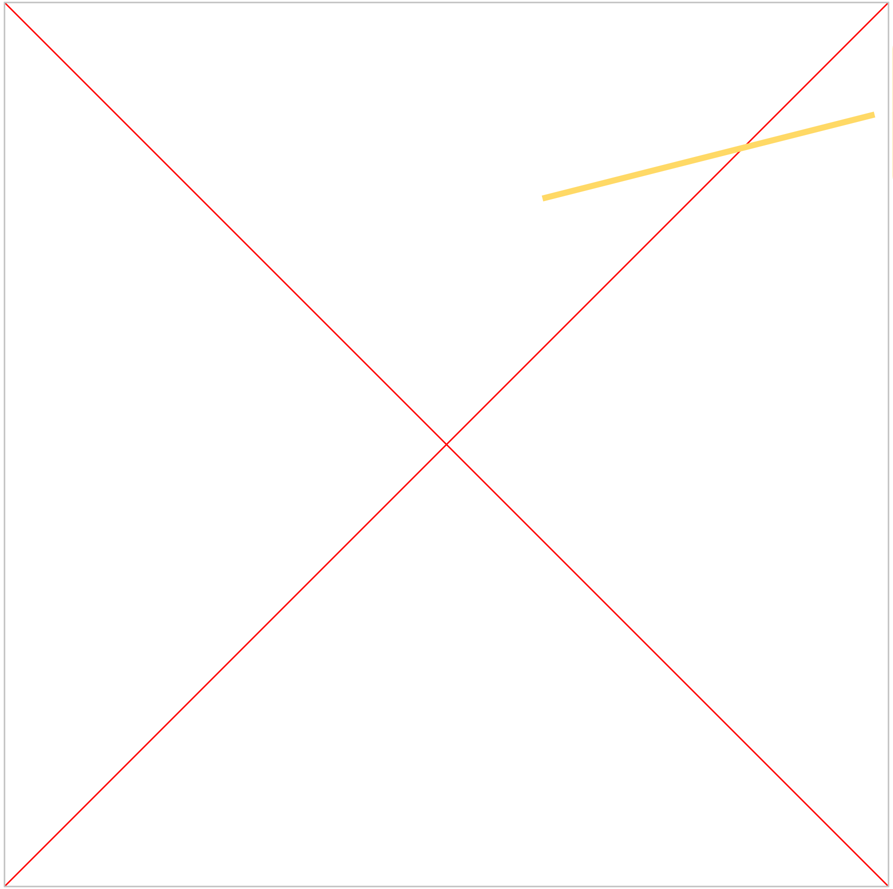
Metrics And data about sales and operating costs!

Entity	Dimensions	Metrics

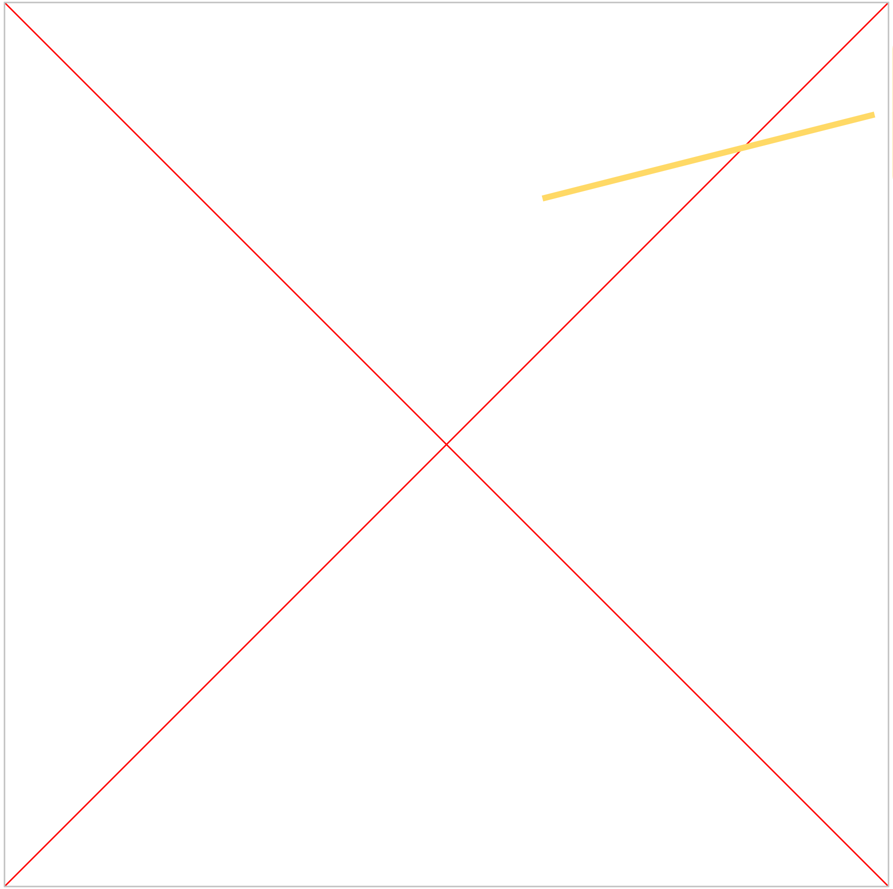
Get me a dataset!!

Exploration

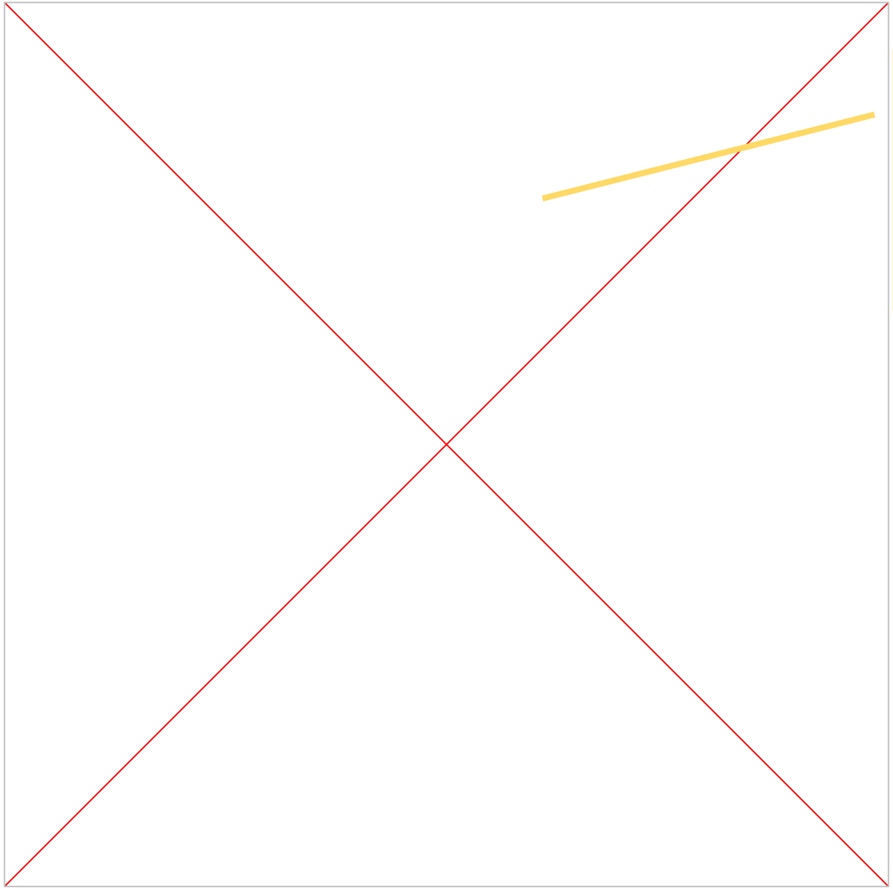
Reporting



How many bricks did we sell?!?

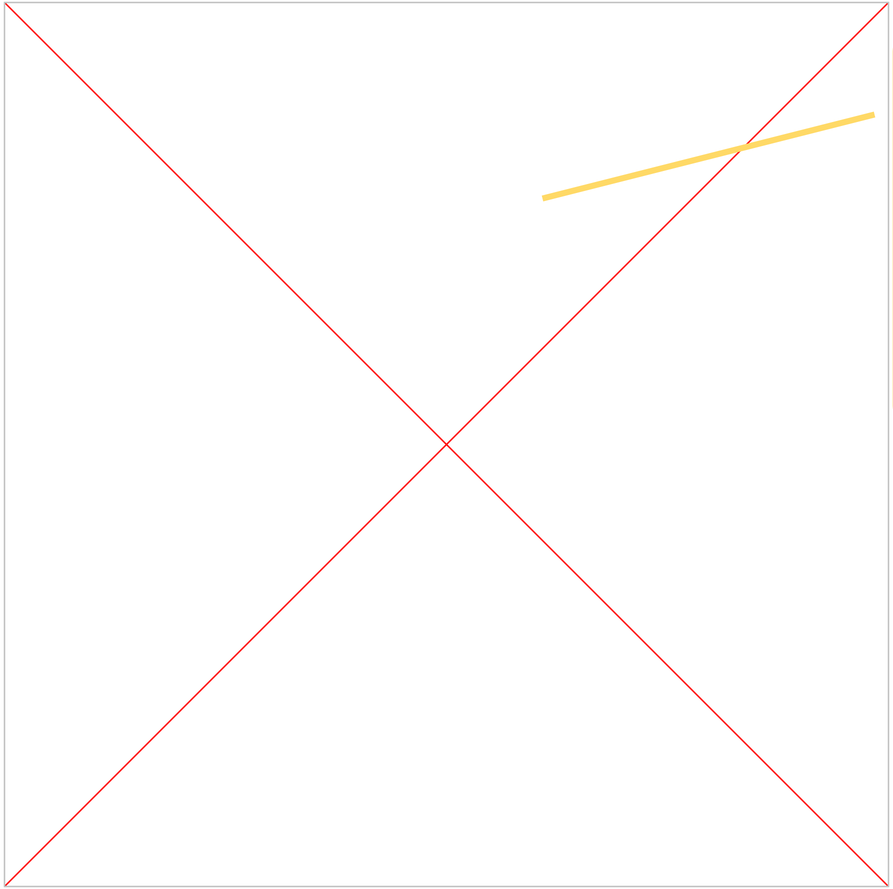


What is our revenue?!?



What is our revenue…

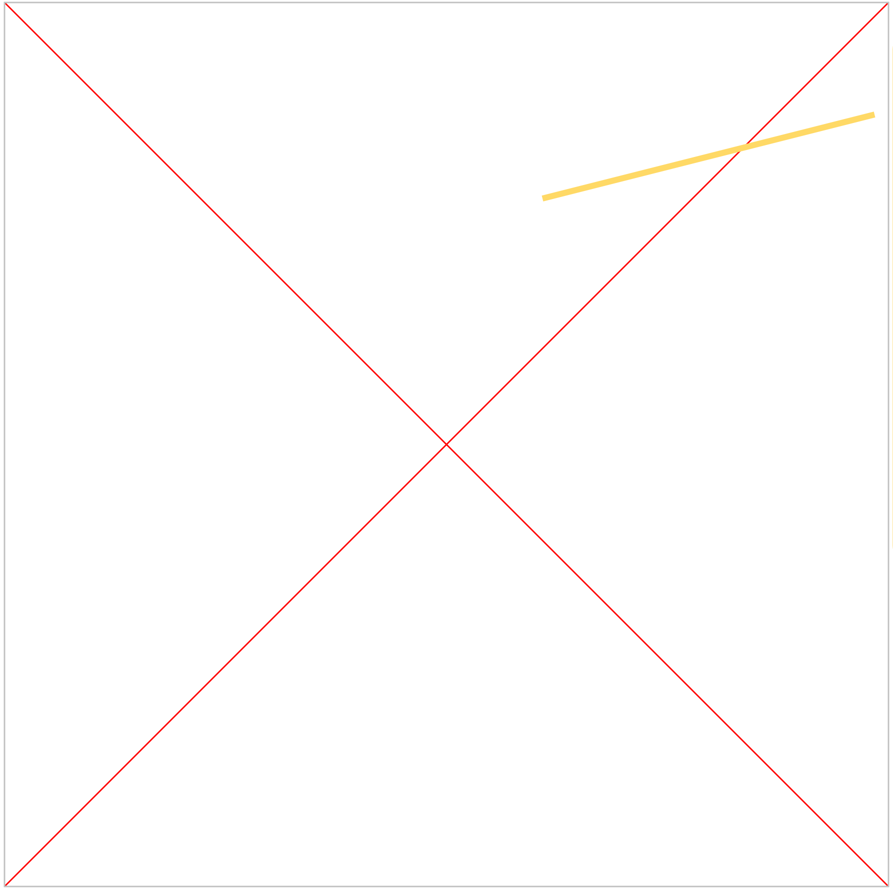
by week?!?



What is our revenue…

by week…

by set theme?!?

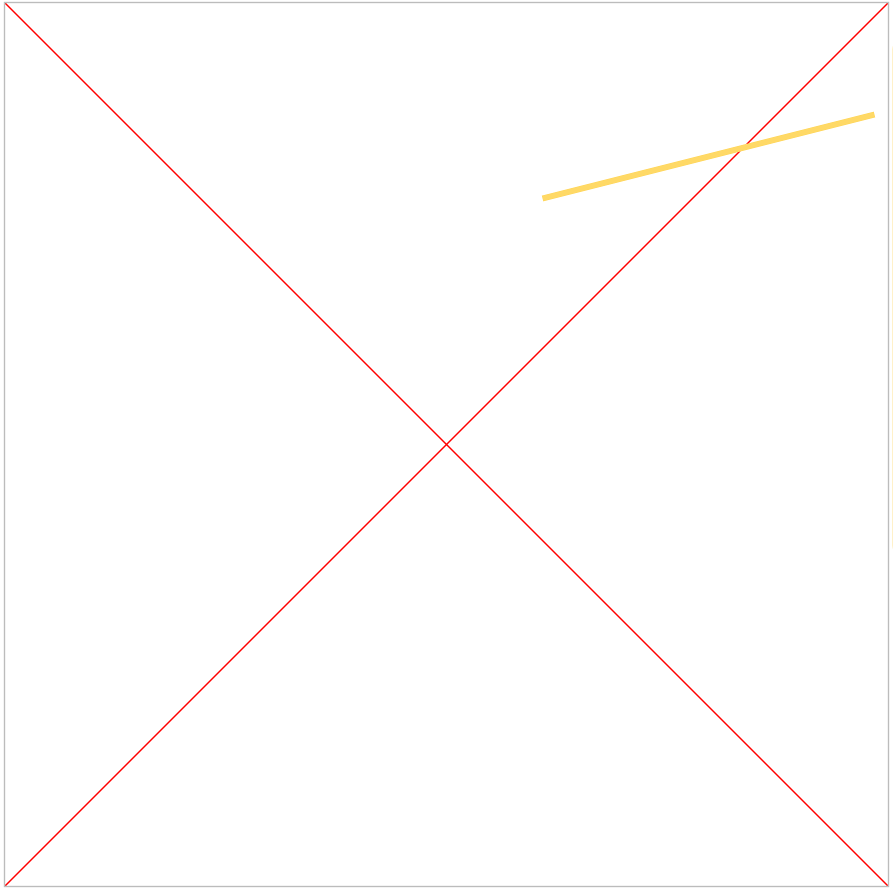


What is our revenue…

by week…

by set theme…

in Europe?!?

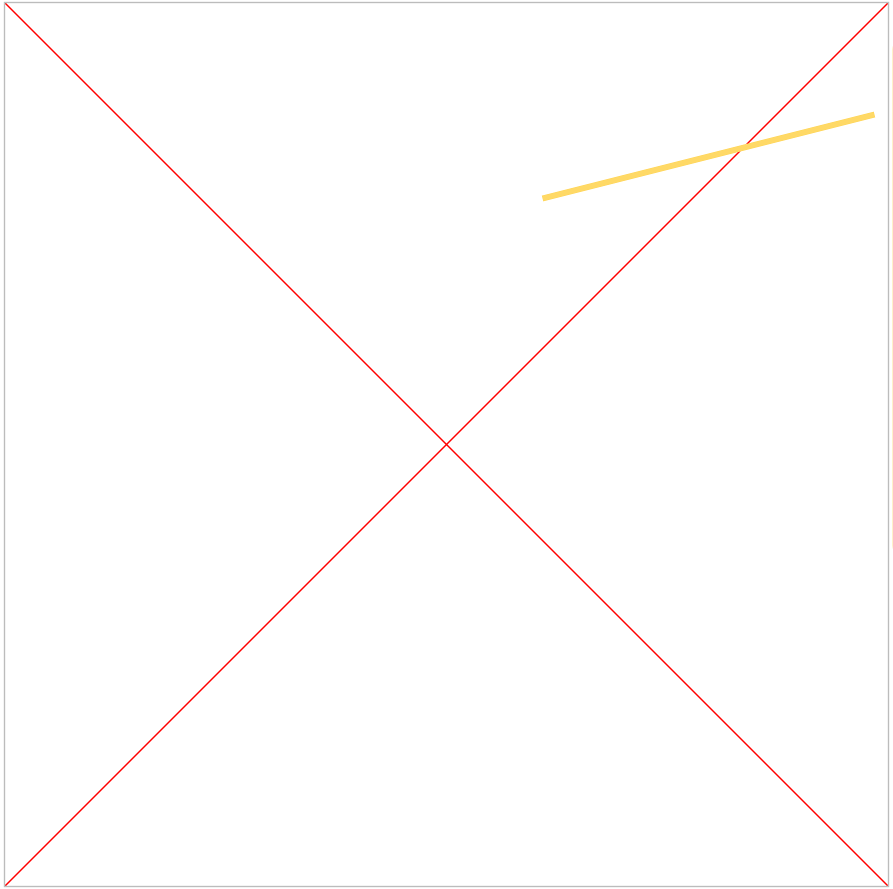


What is our revenue… **Metric**

by week…

by set theme…

in Europe?!?

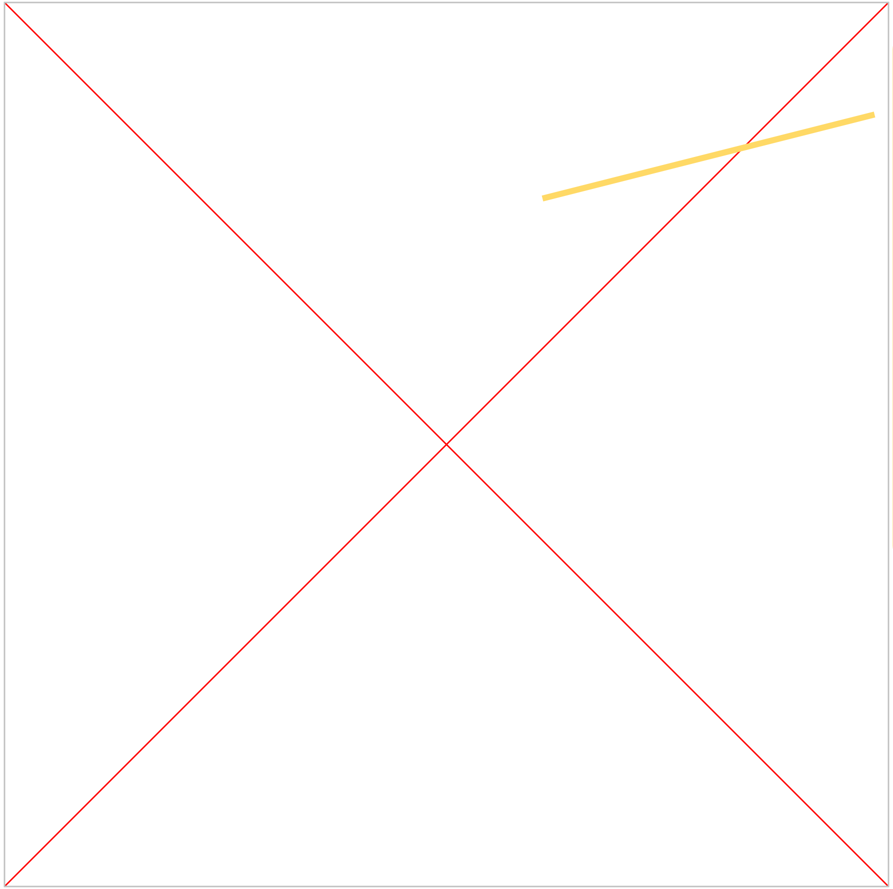


What is our revenue… **Metric**

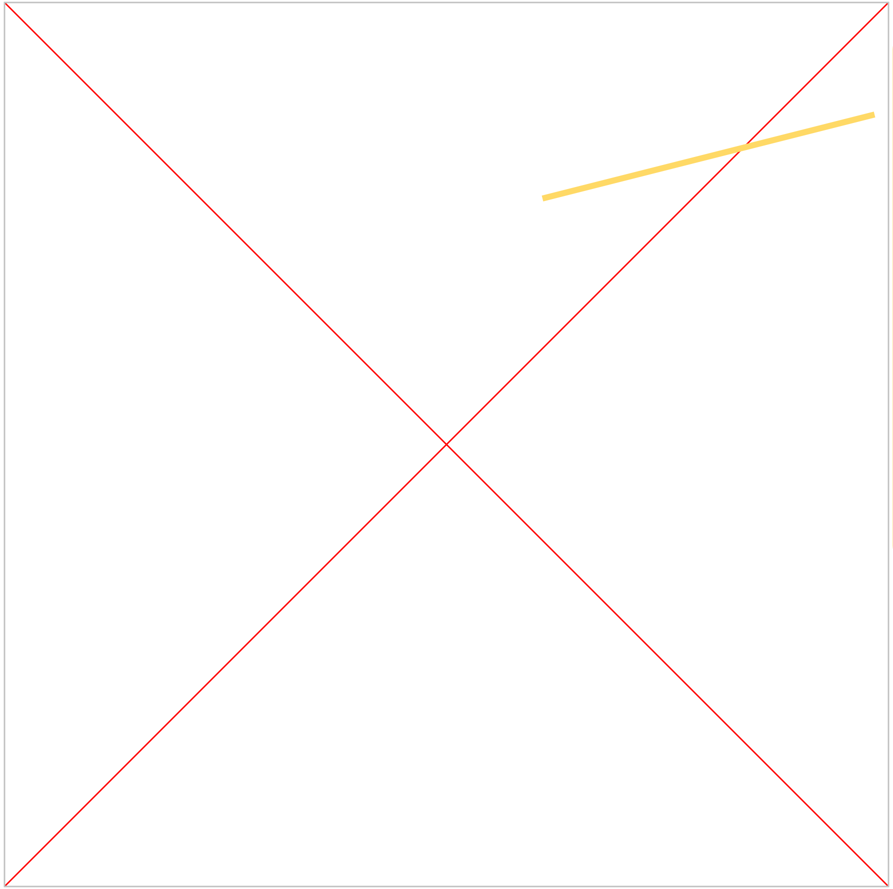
by week… **Time grain**

by set theme…

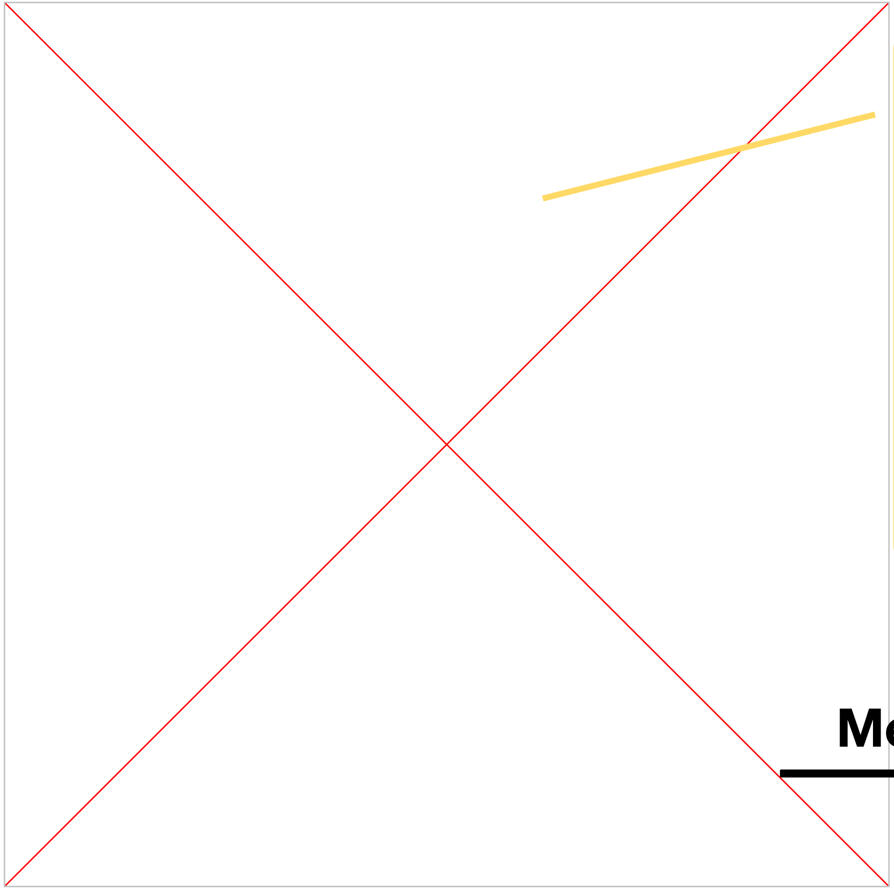
in Europe?!?



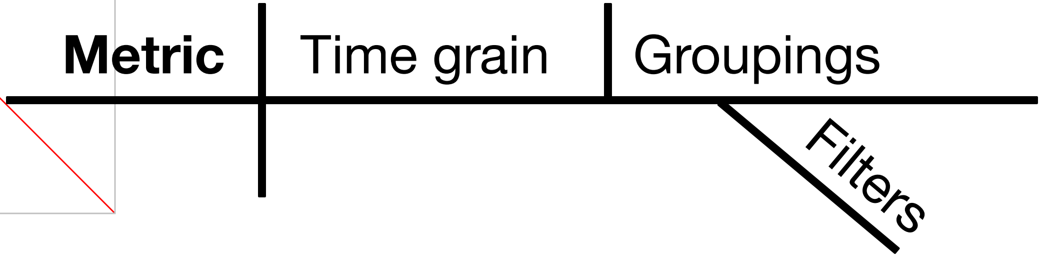
What is our revenue…	**Metric**
by week…	**Time grain**
by set theme…	**Groupings**
in Europe?!?	



What is our revenue…	**Metric**
by week…	**Time grain**
by set theme…	**Groupings**
in Europe?!?	**Filters**



What is our revenue…	**Metric**
by week…	**Time grain**
by set theme…	**Groupings**
in Europe?!?	**Filters**



Revenue

Salesforce opportunity

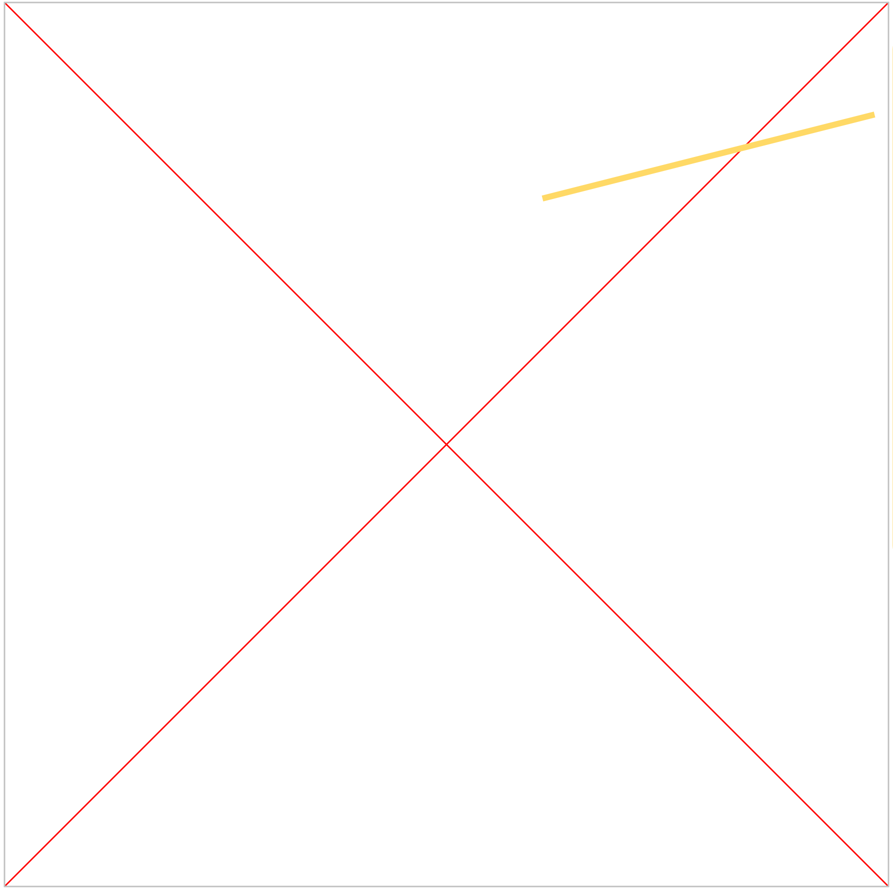
Stripe logs

Tax adjustment

Date sales are recognized

CSV from Janice in Accounting

Return policy



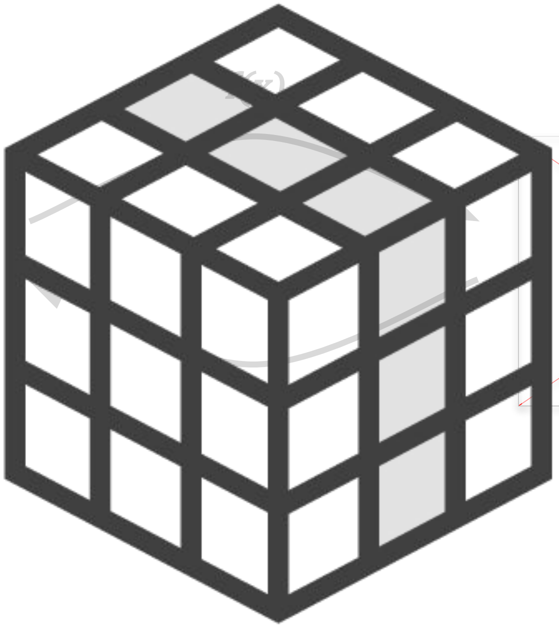
Get me a metric!!


Lane Open
U-Scan Genesis
Cash In
Change Out
MAGIC BONE

Neither? Both?

Exploration

Reporting



Datasets

Metrics

OLAP Cube

month	state	store	sales	items_sold
January	California	1	$1,205	24
February	California	1	$1,346	11
March	California	1	$1,253	18
April	California	1	$1,184	28
May	California	1	$1,337	17
June	California	1	$1,245	11
January	California	2	$1,426	26
February	California	2	$1,275	26
March	California	2	$1,036	30
April	California	2	$1,357	22
May	California	2	$1,246	17
June	California	2	$1,074	23
January	California	3	$1,070	12
February	California	3	$1,480	29
March	California	3	$1,374	20
April	California	3	$1,105	26
May	California	3	$1,425	18
June	California	3	$1,205	25
January	Ohio	52	$390	8
February	Ohio	52	$461	3
March	Ohio	52	$428	7
April	Ohio	52	$420	13
May	Ohio	52	$425	14
June	Ohio	52	$435	8
January	Ohio	84	$381	3
February	Ohio	84	$487	5
March	Ohio	84	$421	5
April	Ohio	84	$528	12

Simple OLAP
What is this?

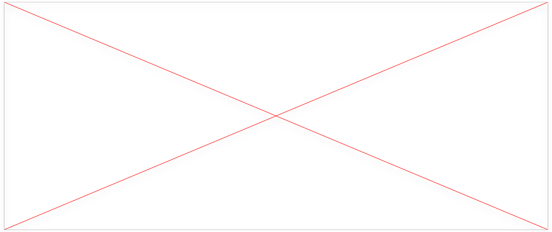
It’s transactions, but not really

And it's metrics, but sorta weirdly decomposed, where you still have to add it up

And that makes them kinda hard to use, because it doesn’t fit the vocabulary

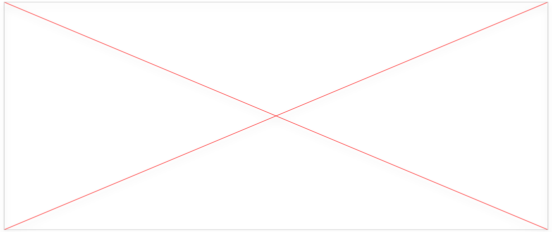

1859
1909
Today
FUSION
FUSION

Exploration



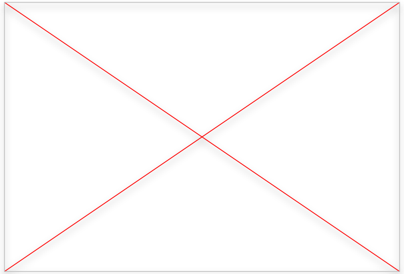
Datasets

Exploration



Datasets

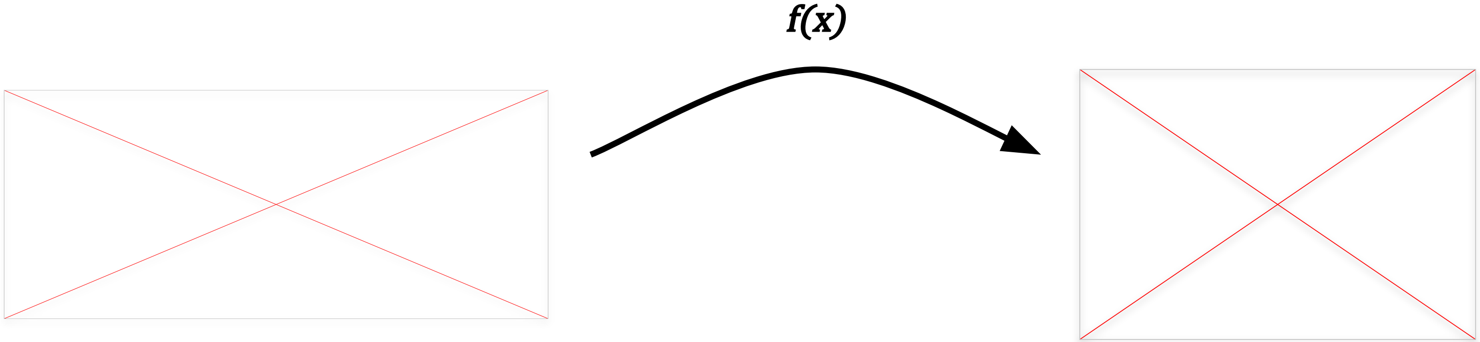
Reporting



Metrics

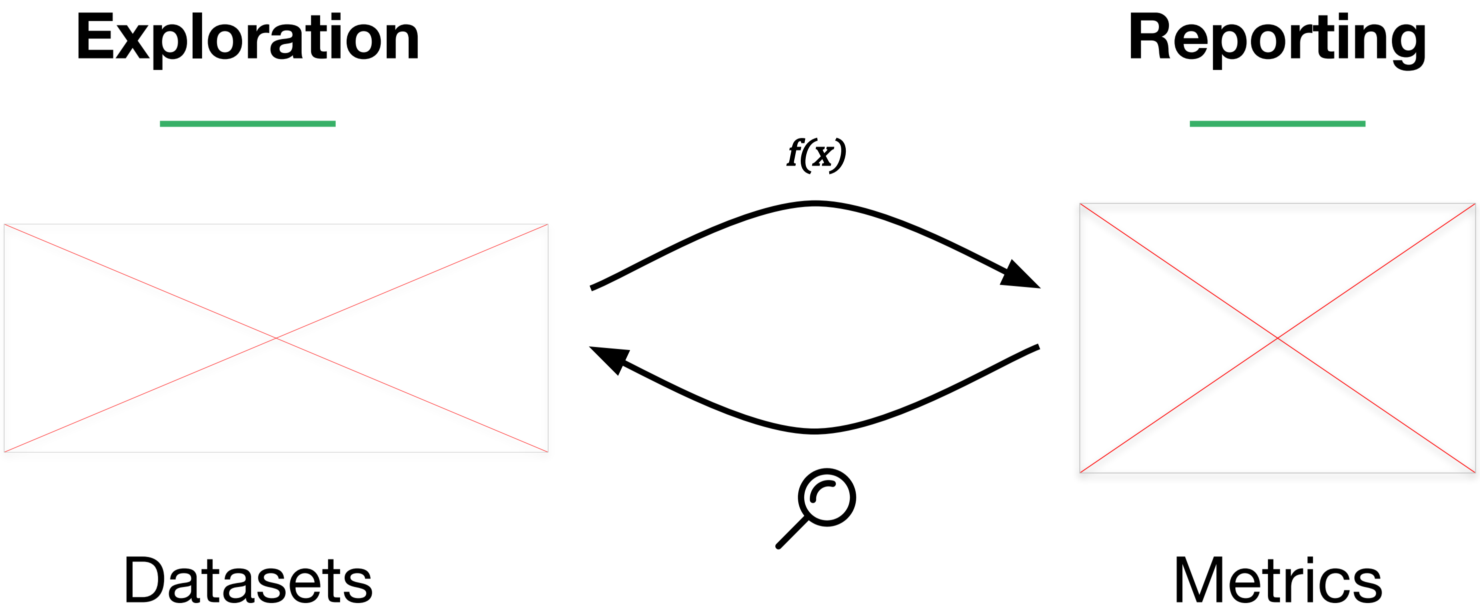
Exploration

Reporting



Datasets

Metrics


Exploration
Reporting
f(x)
Datasets
Metrics

Exploration

Reporting

$f(x)$

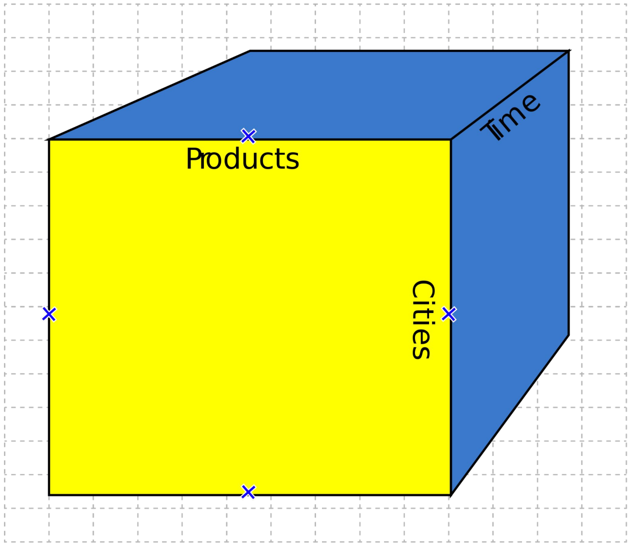
Datasets

Metrics

OLAP cube

From Wikipedia, the free encyclopedia

An **OLAP cube** is a multi-dimensional array of data.


Products
Time
Cities

Adapting to the evolving nature of data through governance

Julie Hollek

Summary

Data Products & Data Science

Consumer and Regulatory Concerns

Case Study: Revenue Data Access Initiative

Background + Acknowledgements

Senior DS + ML manager at Mozilla

- Metrics, Revenue, ML/Data Products, Subscription Services
- Previously: internet health, ad tech

Thank you to the Mozilla Revenue Data Group and Xuan Luo, Arkadiusz Komarzewski

We’re hiring!

careers.mozilla.org

Product Thinking and Data

Data Product

"A product that facilitates an end goal through the use of data"

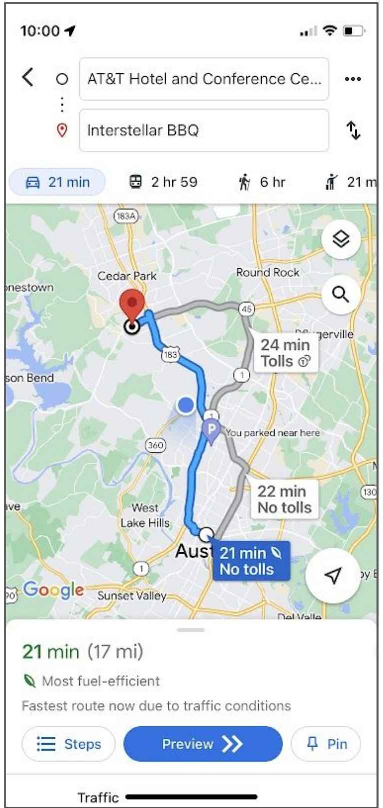
DJ Patil, Data Jujitsu: The Art of Turning Data into Product

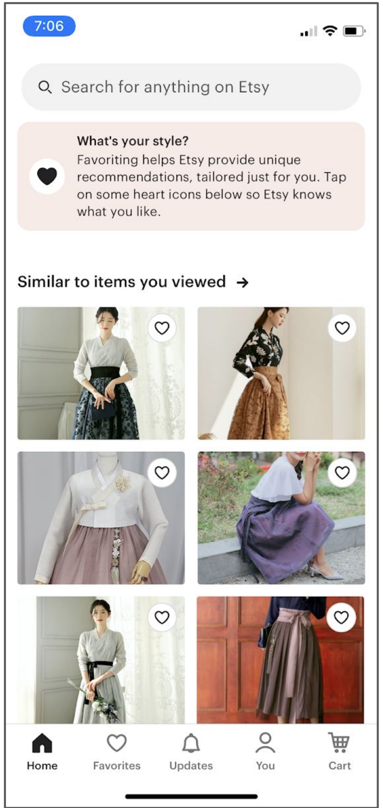
Data as a Product

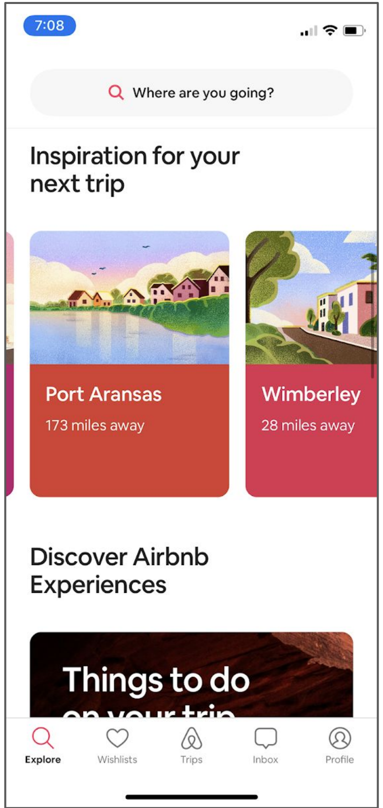
"...data teams must apply product thinking [...] to the datasets that they provide; considering their data assets as their products and the rest of the organization's data scientists, ML and data engineers as their customers."

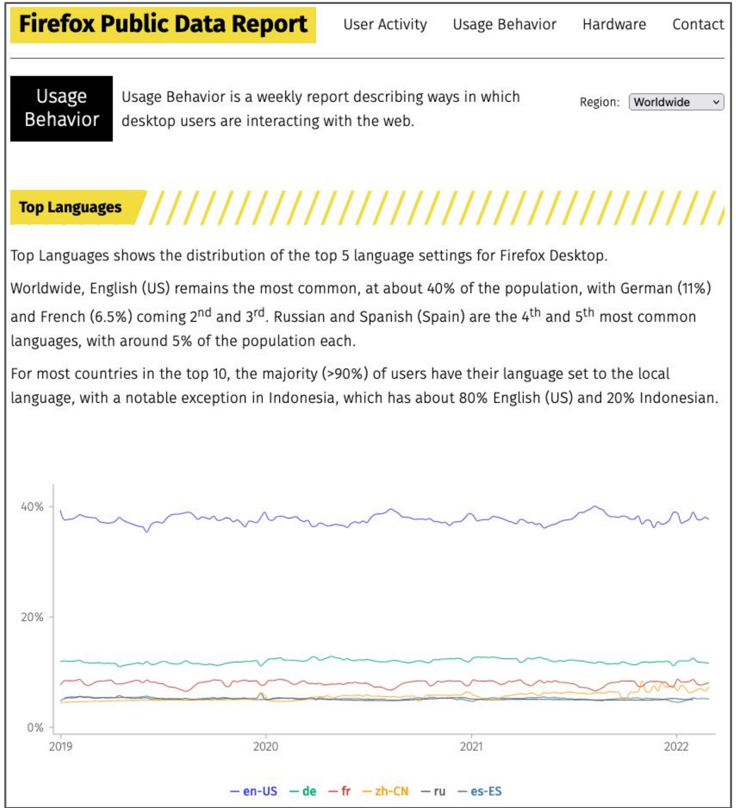
Zhamak Dehghani, How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh

Data Products in the Wild

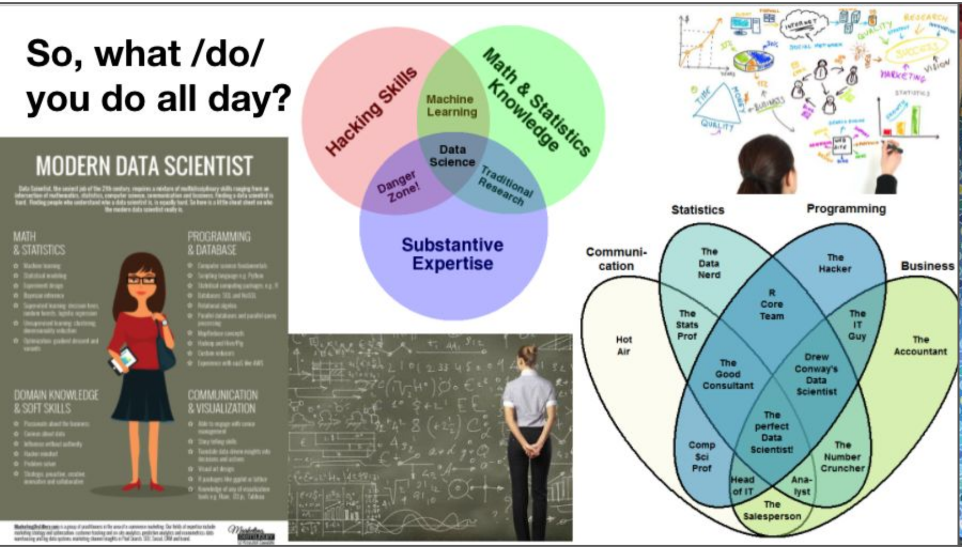




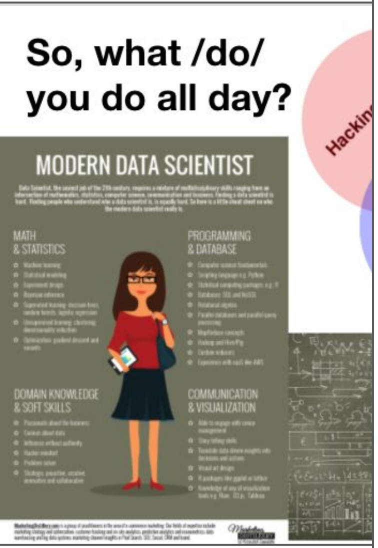


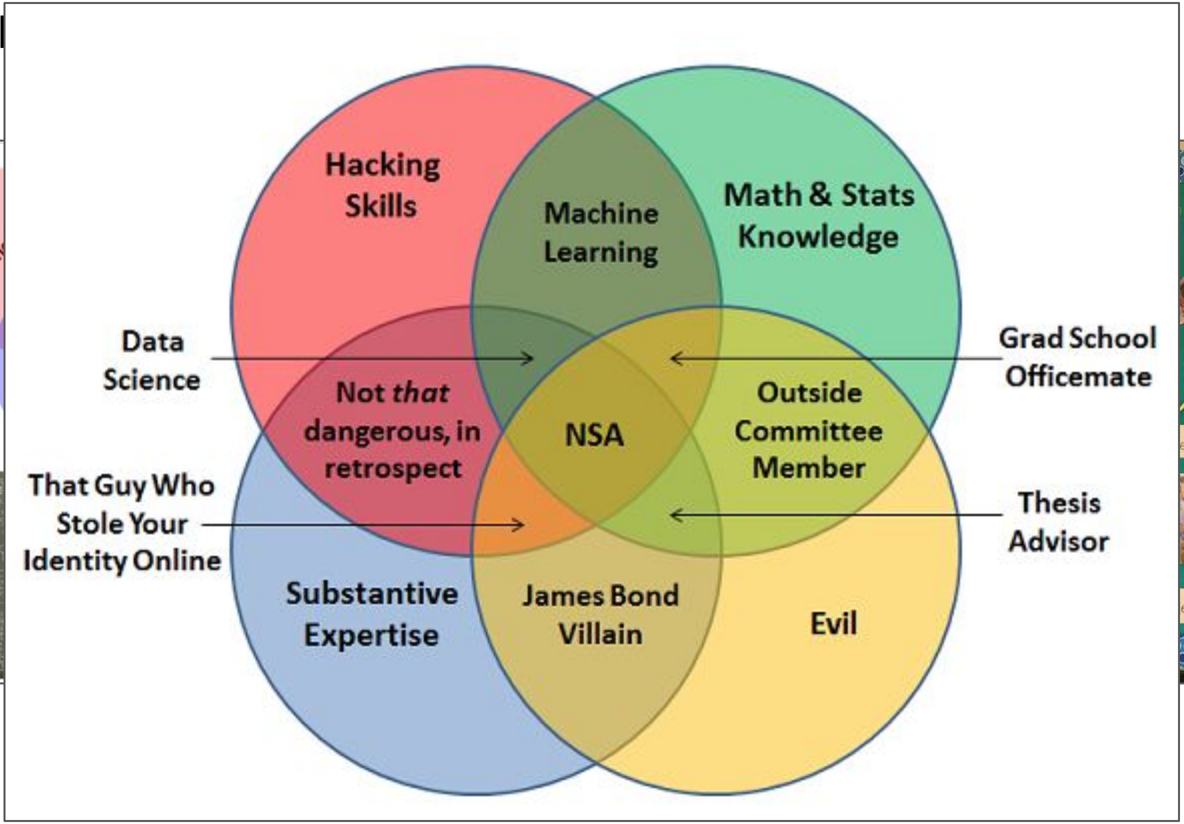


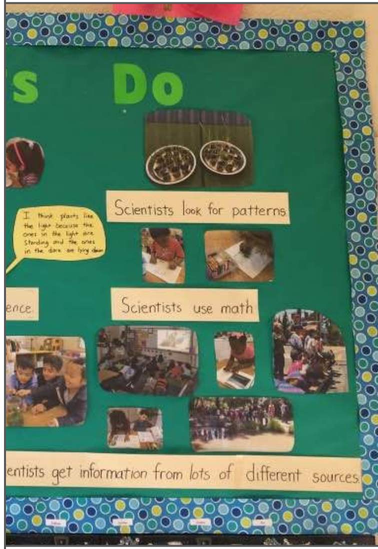
Data Science



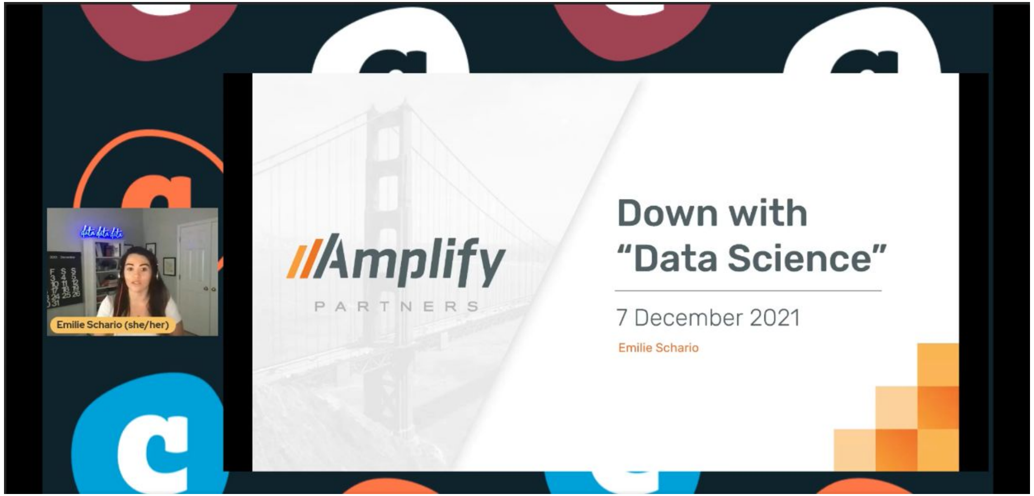
Data Scie


So, what /do/ you do all day?
MODERN DATA SCIENTIST
MATH & STATISTICS
PROGRAMMING & DATABASE
DOMAIN KNOWLEDGE & SOFT SKILLS
COMMUNICATION & VISUALIZATION


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil


Do
Scientists look for patterns
Scientists use math

Beyond the Venn Diagram



- Data Engineers
- Analytics Engineers
- Data Analysts
- Machine Learning Engineers

Emilie Schario - Down with "Data Science"

Analytics

Insight as output

- Play the *objective* voice of your customers
- Metrics + measurements frame how your company views its health
- Looks like
    - Opportunity Sizing/Prototyping
    - Experimentation
    - Impact Analyses

What is the secret of ~~Soylent Green~~ analytics?

(data)

The New York Times

SUBS

Equifax Says Cyberattack May Have Affected 143 Million in the U.S.

1030

By Tara Siegel Bernard, Tiffany Hsu, Nicole Perlroth and Ron Lieber

Sept. 7, 2017

Equifax, one of the three major consumer credit reporting agencies, said on Thursday that hackers had gained access to company data that potentially compromised sensitive information for 143 million American consumers, including Social Security numbers and driver's license numbers.

The attack on the company represents one of the largest risks to personally sensitive information in recent years, and is the third major cybersecurity threat for the agency since 2015.

Equifax, based in Atlanta, is a particularly tempting target for hackers. If identity thieves wanted to hit one place to grab all the data needed to do the most damage, they would go straight to one of the three major credit reporting agencies.

"This is about as bad as it gets," said Pamela Dixon, executive director of the World Privacy Forum, a nonprofit research group. "If you have a credit report, chances are you may be in this breach. The chances are much better than 50 percent."

Facebook appeal over Cambridge Analytica data rejected by Australian court as 'divorced from reality'

Full bench of the federal court confirms earlier ruling that tech giant collects personal information in Australia

Get our free news app; get our morning email briefing

Facebook has been dealt a major blow in its legal fight with the Office of the Australian Information Commissioner over the Cambridge Analytica scandal. Photograph: Artur Widak/NurPhoto/REX/Shutterstock

Facebook has lost a major battle with the Australian regulator over the Cambridge Analytica scandal, after a court dismissed the social media giant's claim that it neither conducts business nor collects personal information in the country.

The Office of the Australian Information Commissioner (OAIC) is suing Facebook, now Meta, for breaching the privacy of more than 300,000 Australian Facebook users in the Cambridge Analytica scandal, exposed more than four years ago by the Guardian.

Throughout the 2010s, consulting firm Cambridge Analytica harvested the personal data of millions of Facebook users without their consent using a

Shortlisted for the FT/McKinsey Business Book of the Year Award 2019

The International Bestseller

THE AGE OF SURVEILLANCE CAPITALISM

THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER

SHOSHANA ZUBOFF

'The true prophet of the information age' *FT*

The **General Data Privacy Regulation** or GDPR is part of the privacy and human rights laws of the EU that set the standards of how companies collect, handle, and protect personal data for EU citizens.

- Users can know how their data are used, what data companies have about them, correct mistakes in the data, have their data deleted, and opt out
- Companies must pay fines for non-compliance such as data breaches or lack of user consent

This is the first of many regulatory standards worldwide. The **California Consumer Privacy Act** is another that is US-based.

Data governance is the set of roles, policies, processes, and technologies that empower an organization to consistently and appropriately handle its data.

Why is this important? It ensures compliance, security, privacy, quality, availability, and usability. It ultimately provides the foundation for an organization's data strategy.

Case Study: Revenue Data Access Initiative

How do we leverage sensitive data for insight and understanding?

Revenue Data Threat Model

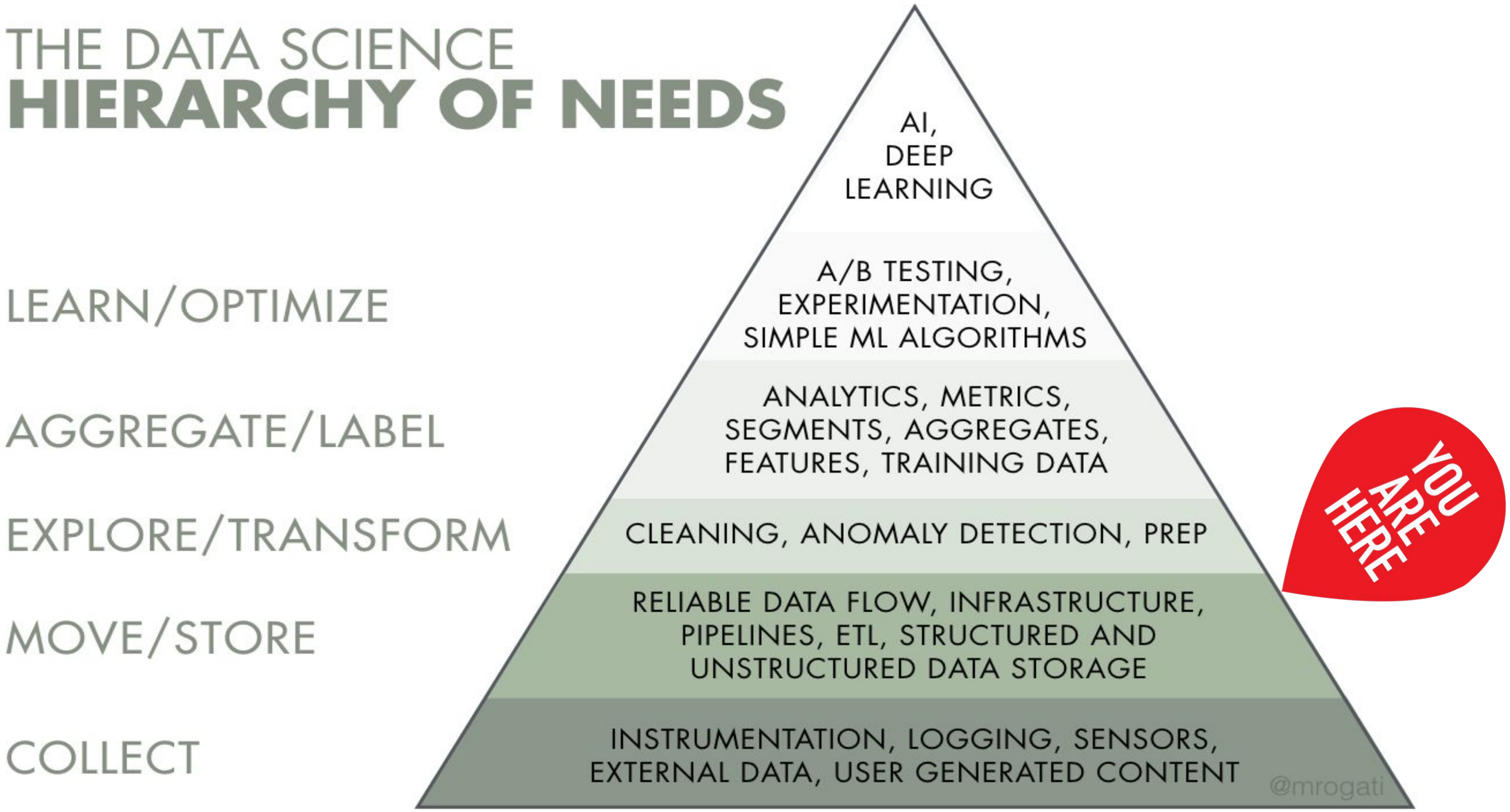
What are we making?
What threats are we concerned about?
What can we do to mitigate these threats?
Do these mitigations work?

Adapted from [Toreon] threat modeling materials

Revenue data are

- important to analyze because they will allow us to make better choices about our business
- sensitive because they contain confidential information that present risk to the business
  - Do not contain personal information
- in need of system designed to
  - grant access to only those who need access to it for processing, evaluation, or decision-making purposes
  - restrict access from the rest of the company

Monica Rogati's Hierarchy of Needs



Revenue Data

- PDFs and Spreadsheets and APIs, oh my!
- Sometimes hand-curated, from non-Mozillians, mostly maintained by hand by non-technologists
- *sensitive*
- All-or-nothing access, but difficult to use
- safeguarded by the CFO herself

Revenue Data Science @ Mozilla

Revenue Forecasting

Product Data Science for our monetizable surface areas

Data Help for Finance and Business Operations Analysts

- Methods
- Data

Given that this is what they do, what do Rev DS look like?

Data Scientist

- Tend to have advanced degrees (Ph.D, MS) in a STEM field
- Advanced skills in SQL and scripting language (usually Python or R)

Finance/Business Analyst

- Subject matter expert
- Simple SQL skills, proficient in Excel
- Straightforward domain-relevant modeling

Revenue Data Access Initiative

- Framework
    - Policy
    - Process
- Technical Infrastructure
    - Differential Access Implementation
    - Data Pipeline Migration & Improvements
- Empowerment
    - Visualization Layer

Policy

Principles-first approach to understand who should get access to sensitive data

Spell out why you need these particular principles

Categories of Data

1	Data that are sensitive but extremely difficult or impossible to calculate sensitive quantities
2	Data that allow someone to back-calculate sensitive quantities
3	Highly sensitive, restricted, and rarely shared data that must be kept confidential

Policy

Principles-first approach to understand who should get access to sensitive data

Role-based Access

- Permanent - you have a job at the company that requires you to deal with these data regularly
- Project - you’re working on a project that requires these data but this is due to the project and not your position

Policy

Principles-first approach to understand who should get access to sensitive data

Compliance

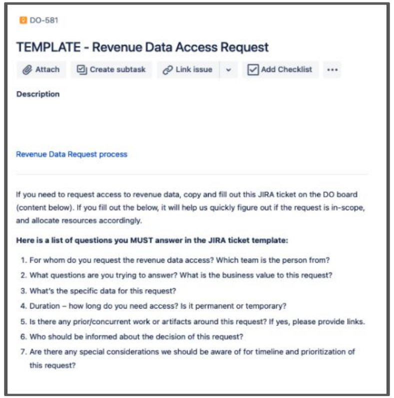
People with access to these data must take a test to demonstrate that they have read and understood the sensitive information training and sign an acknowledgement that they will comply

Process

Practical, standardized workflow to apply our policy

Request and Evaluation Flow

Jira Software



DO-581

TEMPLATE - Revenue Data Access Request

Attach | Create subtask | Link issue | Add Checklist

Description

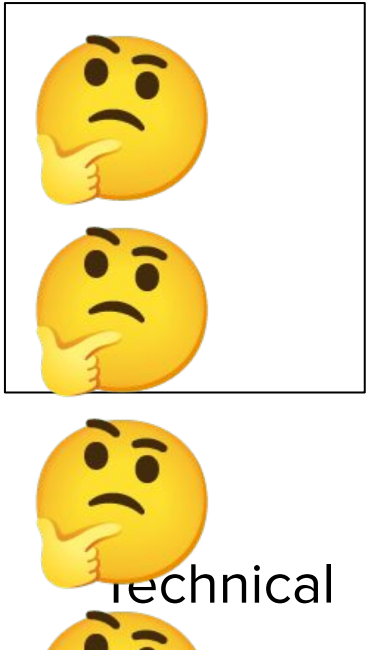
Revenue Data Request process

If you need to request access to revenue data, copy and fill out this JIRA ticket on the DO board (content below). If you fill out the below, it will help us quickly figure out if the request is in-scope, and allocate resources accordingly.

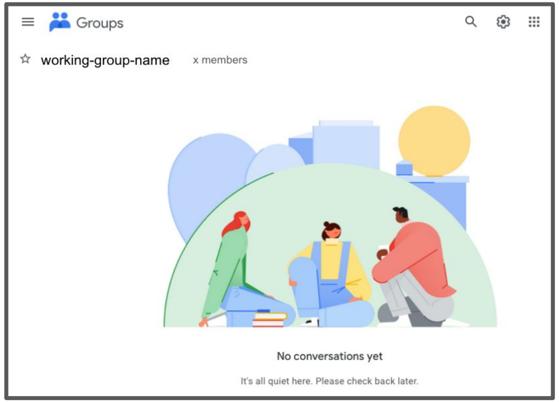
Here is a list of questions you MUST answer in the JIRA ticket template:

1. For whom do you request the revenue data access? Which team is the person from?
2. What questions are you trying to answer? What is the business value to this request?
3. What's the specific data for this request?
4. Duration – how long do you need access? Is it permanent or temporary?
5. Is there any prior/concurrent work or artifacts around this request? If yes, please provide links.
6. Who should be informed about the decision of this request?
7. Are there any special considerations we should be aware of for timeline and prioritization of this request?

Data stewards



Google Groups



Framework

Technical

Empowerment

Process

Practical, standardized workflow to apply our policy

Auditing

- Quarterly audits on permanent access
  - Requires manager and access steward approval
- Extension evaluation for temporary access if needed for project
  - Request is evaluated by access stewards

Technical Infrastructure

Differential Access Implementation

give access to those who need it and restrict access from those who don't

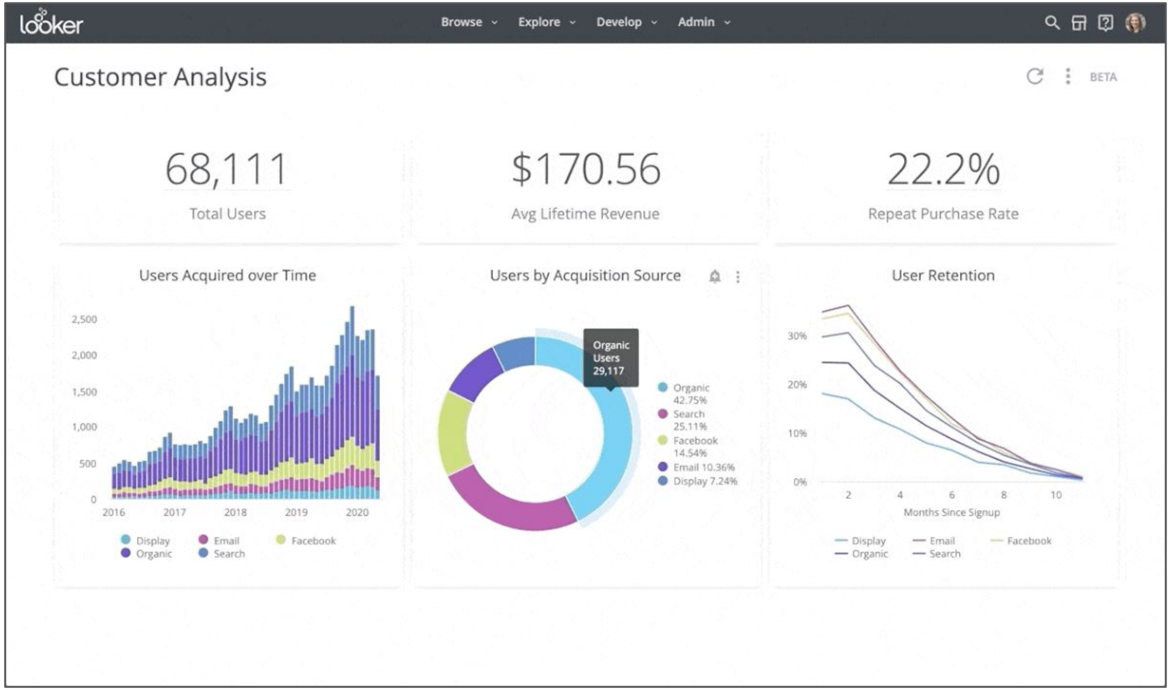
- Leverages BigQuery's authorized views to create differential access based on revenue access policy specifications

Data Pipeline Migration & Improvement

contain all of the revenue data in one place to make easier access configuration, more robust datasets, SRE support

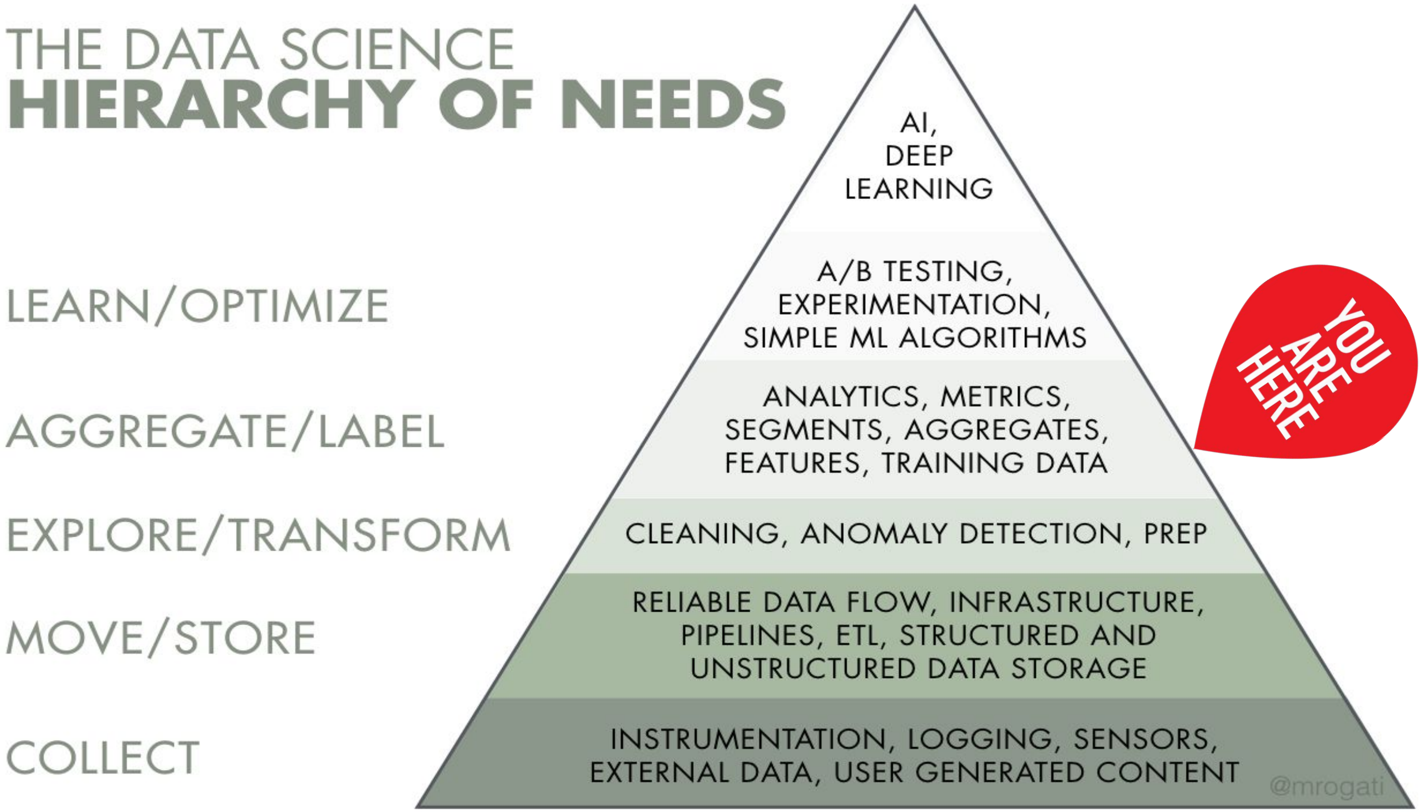
- Syndication of data, new ETL, new connectors that standardize and stabilize pipelines

Visualization Layer



Framework Technical **Empowerment**

Monica Rogati's Hierarchy of Needs

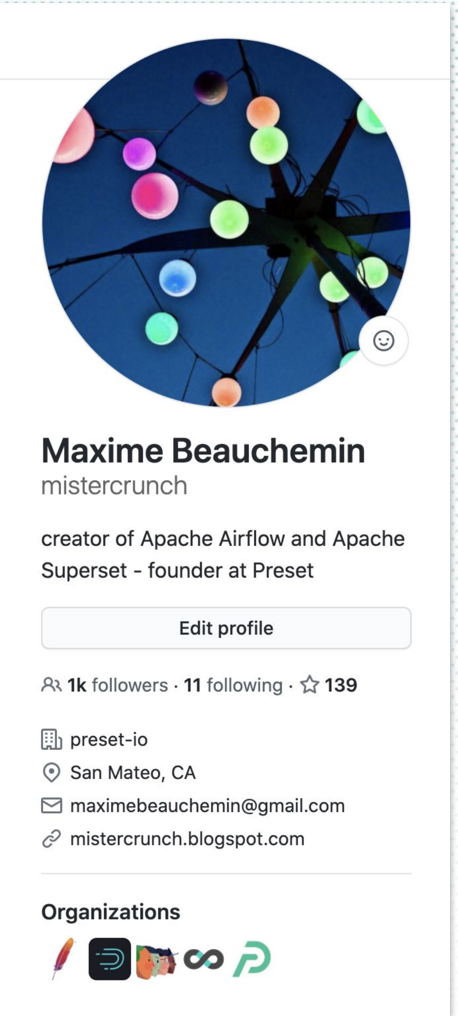


How does this tie back to data privacy?

Domain	Revenue Data	Personal Data
Framework	Policy based on business risk	Policy based on user privacy
Technical Infrastructure	Differential access, data warehouse	Differential access, data warehouse
Empowerment	Insights	Insights, user-facing data product

preset™

The Next Generation of Business Intelligence



- 20+ years swimming in data @ preset
- Started Apache **Airflow** at Airbnb in 2014
- Started Apache **Superset** at Airbnb in 2015
- Started **Preset** - The Apache Superset company in 2019

Apache Superset™

Apache Airflow

preset™

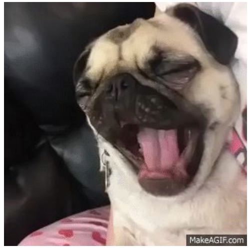
Agenda

- The [accelerated] story of BI
- Enabling analytics everywhere
- Delamination of stack
- Latency over freshness
- Open Source FTW
- Data models
- Still to come

“**Business Intelligence**” - defined

Business intelligence *(**BI**) comprises the strategies and technologies used by enterprises for the data analysis and management of business information.*[1] *Common functions of business intelligence technologies include reporting, online analytical processing, analytics, dashboard* development, *data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics,* and *prescriptive analytics.*



A brief history of BI…

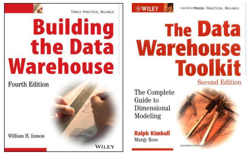
So you thought BI was old…

In **1865**, Richard Millar Devens presented the phrase “Business Intelligence” (BI) in the Cyclopædia of Commercial and Business Anecdotes. He was using it to describe how Sir Henry Furnese, a banker, profited from information by gathering and acting on it before his competition.

More recently, **in 1958**, an article was written by an IBM computer scientist named Hans Peter Luhn, describing the potential of gathering Business Intelligence (BI) through the use of technology.

The contemporary timeline

- 70s - IBM and Siebel enter the market
- 80s - emergence of the data warehouse
- 90s - early vendors appear - highly specialized tooling
- 2000s - self-service and large all-in-one platforms
- 2010s - big data + data goes mainstream
    - explosion of more specialized tools
    - democratization of data
- 2020s!?!?!?!!



Some statements about BI / analytics…

- BI tooling tries to be a solution for ***EVERY* type of data**, every **persona** and every **workflow**. Buyers have been trained to buy a single solution that **SOLVES IT ALL**. This is not realistic.
- Yet most companies have multiple BI tools
- BI is the original 20+Y before no-code **"NO-CODE"** solution!(?)
- BI depends on "the analytics process" and is the last link in an extremely complex and brittle chain
- Yet. **People think data should be easy**, or that the right tool can make it easy. No.

Failed promises

- Solving data for all
- Self-service - making it simple enough for the masses

Analytics Everywhere!

Free analytics from the experts and their specialized tooling!

- In-context analytics > foreign dashboards
- The rise of data literacy = users asking for interactive visualizations
- Every app/SaaS to become a “data app”

[head optional]

Headless software (e.g. "headless Java" or "headless Linux",) is **software capable of working on a device without a graphical user interface**. Such software receives inputs and provides output through other interfaces like network or serial port and is common on servers and embedded devices.

https://en.wikipedia.org › wiki › Headless_software

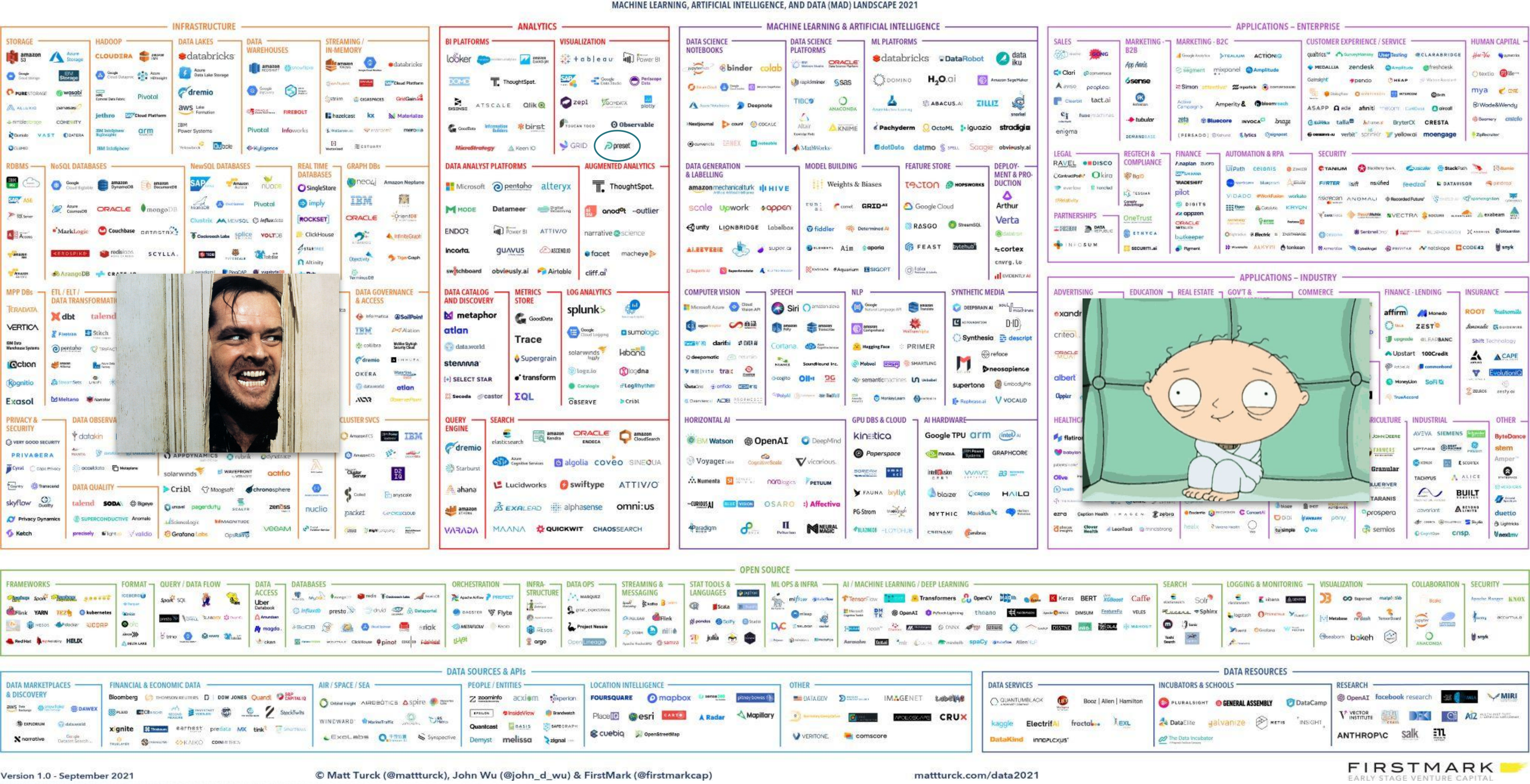
Headless software - Wikipedia

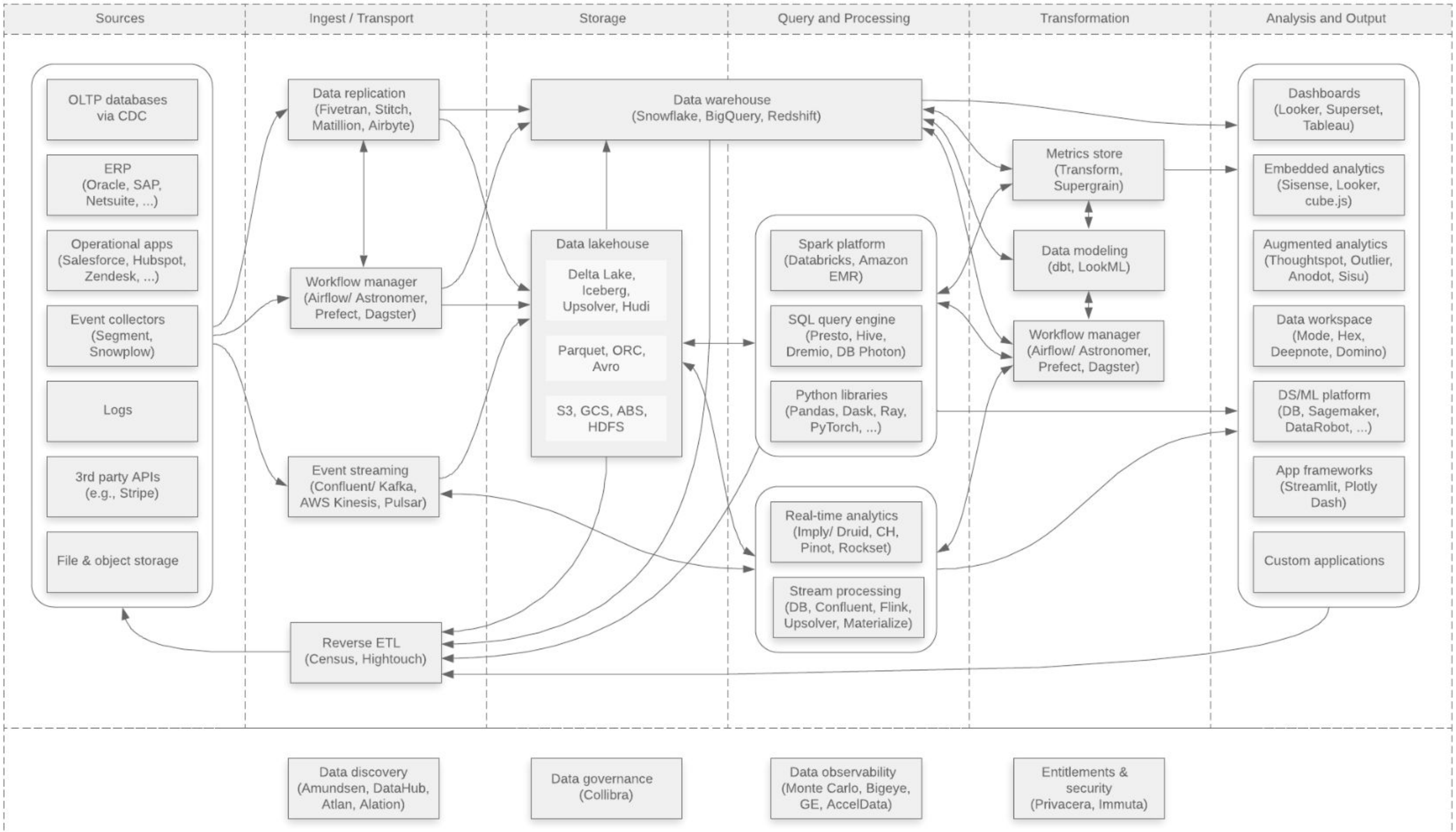
Delamination of the stack

Gartner's BI Magic Quadrant

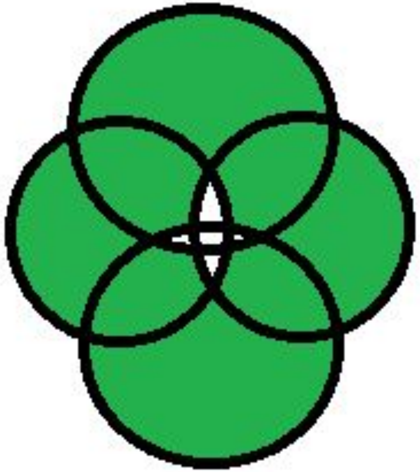
2021

preset


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
STORAGE
HADOOP
DATA LAKES
DATA WAREHOUSES
STREAMING / IN-MEMORY
RDBMS
NoSQL DATABASES
NewSQL DATABASES
REAL TIME DATABASES
GRAPH DBs
MPP DBs
ETL / ELT / DATA TRANSFORMATION
DATA GOVERNANCE & ACCESS
PRIVACY & SECURITY
DATA OBSERVABILITY
DATA QUALITY
CLUSTER SVCS
ANALYTICS
BI PLATFORMS
VISUALIZATION
DATA ANALYST PLATFORMS
AUGMENTED ANALYTICS
DATA CATALOG AND DISCOVERY
METRICS STORE
LOG ANALYTICS
QUERY ENGINE
SEARCH
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
ML PLATFORMS
DATA GENERATION & LABELLING
MODEL BUILDING
FEATURE STORE
DEPLOYMENT & PRODUCTION
COMPUTER VISION
SPEECH
NLP
SYNTHETIC MEDIA
HORIZONTAL AI
GPU DBS & CLOUD
AI HARDWARE
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER EXPERIENCE / SERVICE
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOV'T &
COMMERCE
FINANCE - LENDING
INSURANCE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
FORMAT
QUERY / DATA FLOW
DATA ACCESS
DATABASES
ORCHESTRATION
INFRASTRUCTURE
DATA OPS
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
ML OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH
Version 1.0 - September 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

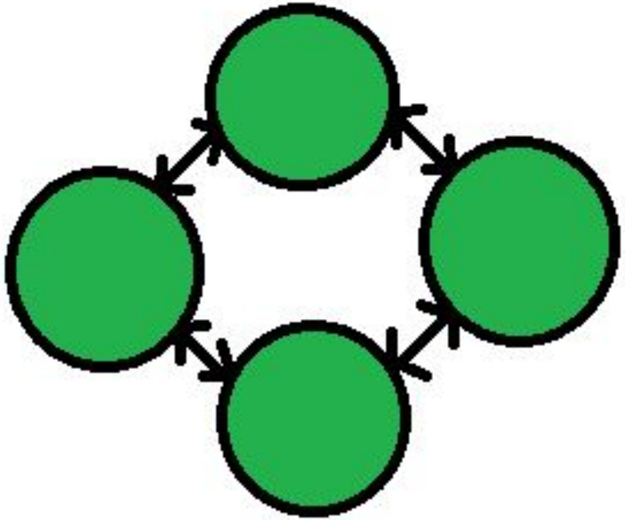


Sources
Ingest / Transport
Storage
Query and Processing
Transformation
Analysis and Output
OLTP databases via CDC
ERP (Oracle, SAP, Netsuite, ...)
Operational apps (Salesforce, Hubspot, Zendesk, ...)
Event collectors (Segment, Snowplow)
Logs
3rd party APIs (e.g., Stripe)
File & object storage
Data replication (Fivetran, Stitch, Matillion, Airbyte)
Workflow manager (Airflow/ Astronomer, Prefect, Dagster)
Event streaming (Confluent/ Kafka, AWS Kinesis, Pulsar)
Reverse ETL (Census, Hightouch)
Data warehouse (Snowflake, BigQuery, Redshift)
Data lakehouse
Delta Lake, Iceberg, Upsolver, Hudi
Parquet, ORC, Avro
S3, GCS, ABS, HDFS
Spark platform (Databricks, Amazon EMR)
SQL query engine (Presto, Hive, Dremio, DB Photon)
Python libraries (Pandas, Dask, Ray, PyTorch, ...)
Real-time analytics (Imply/ Druid, CH, Pinot, Rockset)
Stream processing (DB, Confluent, Flink, Upsolver, Materialize)
Metrics store (Transform, Supergrain)
Data modeling (dbt, LookML)
Workflow manager (Airflow/ Astronomer, Prefect, Dagster)
Dashboards (Looker, Superset, Tableau)
Embedded analytics (Sisense, Looker, cube.js)
Augmented analytics (Thoughtspot, Outlier, Anodot, Sisu)
Data workspace (Mode, Hex, Deepnote, Domino)
DS/ML platform (DB, Sagemaker, DataRobot, ...)
App frameworks (Streamlit, Plotly Dash)
Custom applications
Data discovery (Amundsen, DataHub, Atlan, Alation)
Data governance (Collibra)
Data observability (Monte Carlo, Bigeye, GE, AccelData)
Entitlements & security (Privacera, Immuta)



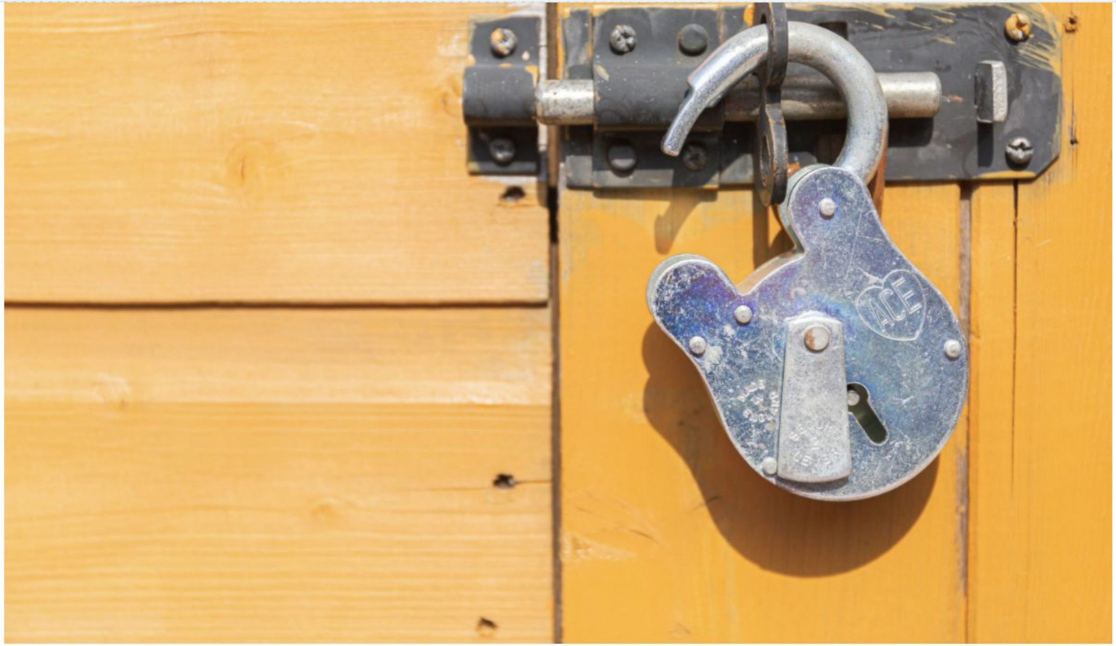
Tight coupling:

1. More Interdependency
2. More coordination
3. More information flow



Loose coupling:

1. Less Interdependency
2. Less coordination
3. Less information flow



COMMUNITY

The Future Of Business Intelligence Is Open Source | Preset

Maxime Beauchemin March 05, 2021

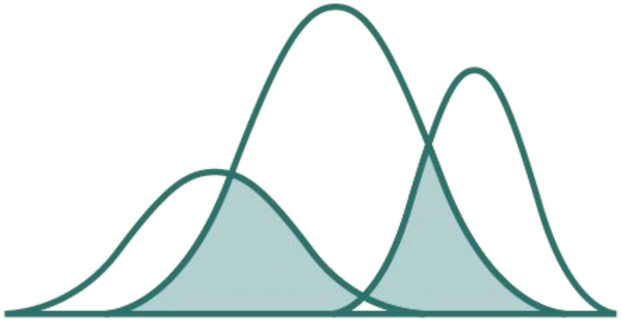
Subscribe

AI Monitoring & Explainability: The Critical Hidden Connection

Anupam Datta

Co-Founder, President, Chief Scientist

TruEra

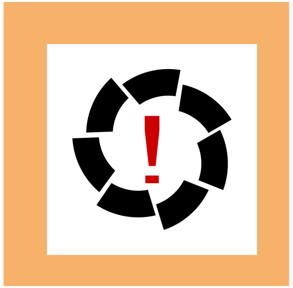


What people think ML Monitoring is like…

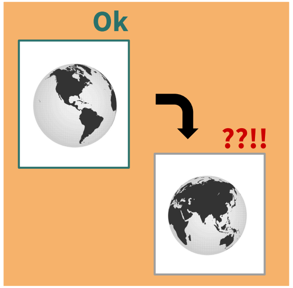
and what it's actually like.

A lot can go wrong.

Data Bugs



Unforeseen Changes



New, untrained use cases

Shifting concepts & behavior

Adversarial attacks

The harsh reality of ML.

The moment you put a model in production, it goes on a wild ride.

So monitoring is key.

Monitoring is not that easy today.
Data Science and ML Ops teams struggle to minimize ML risk.

There's a wild goose chase going on.

How can I better understand how my models are working?

How do I identify real problems with the model?

What is the problem's root cause? How can I debug quickly?

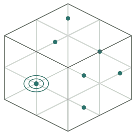
How can I make monitoring work easily in my environment?

Teams struggle with:

- Visibility and observability
- Diagnosis and actionability
- Complex environments and workflow (diverse models, diverse stakeholders)

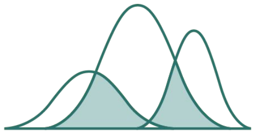
Monitoring Requirements

Fast, precise, and complete.



Broad coverage of model & data quality metrics

Fast, precise debugging

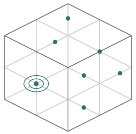


Easy to deploy and scale

AI Monitoring & Explainability:
The Critical Hidden Connection

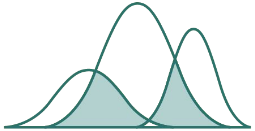
Focus Today: Monitoring Requirements

Fast, precise, and complete.



Model Drift & Performance Metrics

Fast, precise debugging with root cause analysis



Easy to deploy and scale

AI Monitoring & Explainability: The Critical Hidden Connection

Outline

- Overview
  - Why does drift happen?
  - What are different kinds of drift?
  - What is consequential drift?
- How to identify drift?
  - Measures
  - Challenges
- How to mitigate drift?
- Monitoring

Overview of Drift

Overview of Drift

Bikes used to look like this

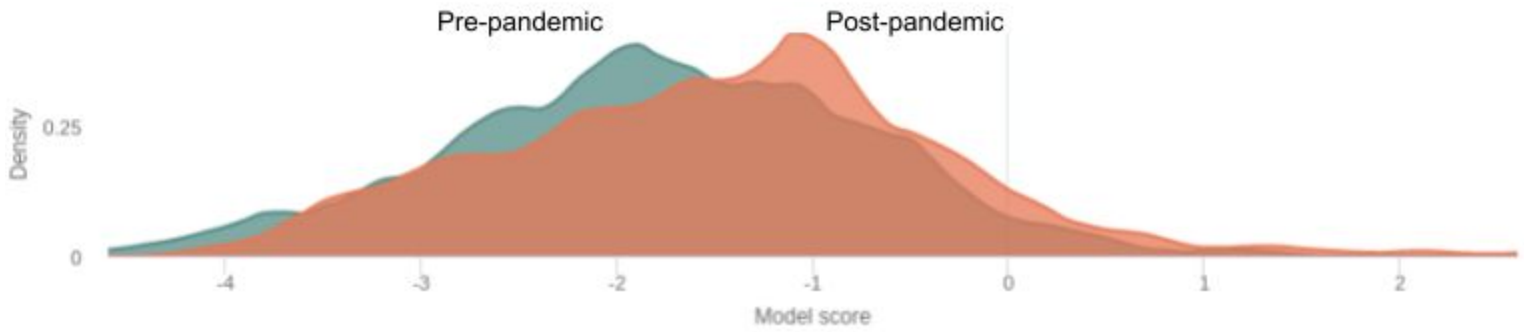
… now they look like this

Will an ML model trained on images like the left continue to work well?

Overview of Drift

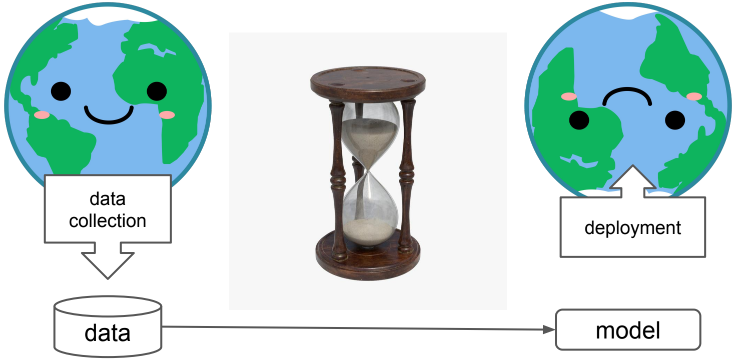
This is similar to what happened to models with Covid.

Example: risk scoring model. Lower model score shows lower risk.



Will an ML model trained on pre-pandemic data continue to work well?

Overview:
Why does drift happen?



Data quality issues

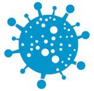
Examples: NaN

- Broken feature pipelines

The External World Has Changed

Examples:

- The pandemic
- Housing market fluctuations



Zillow

Model Applied to a New Context

Example:

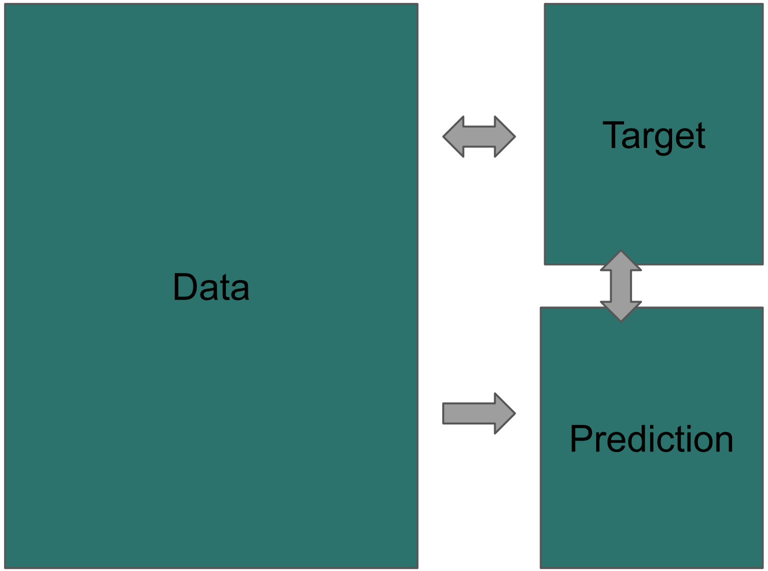
- Model trained on Wikipedia applied to news articles

Collected Training Data Is Different

Example:

- For credit decisions, labels are only available for approved applicants
- Impact of your models on the data

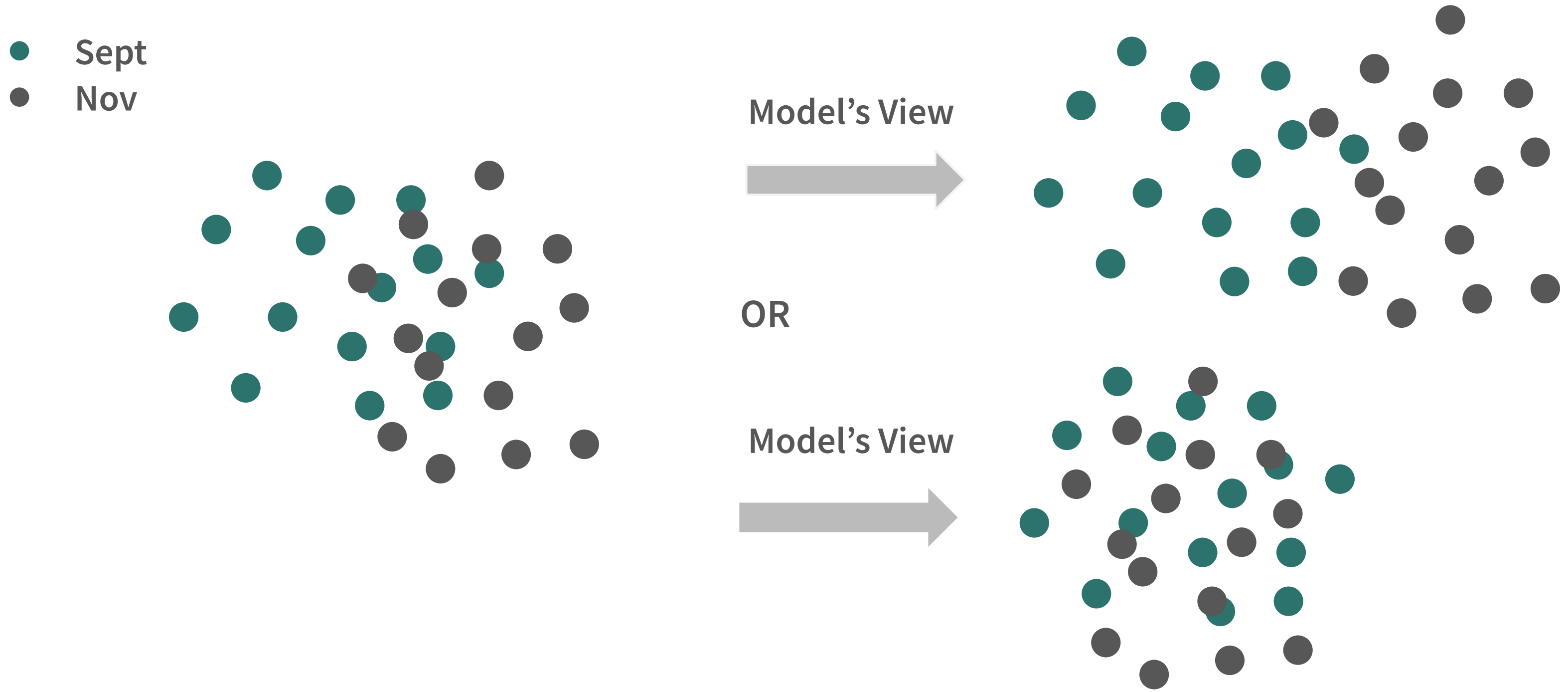
Overview: What are the Different Kinds of Drift?



1. Data drift
   a. Covariate shift -- drift in input features
   b. Concept drift -- drift in relationship between input and target
2. Model decay -- performance loss due to data drift
3. "Prediction shift" -- drift in model predictions

Overview: Which Drifts are Consequential and Why?

truera



High-dimensional data always drifts (curse of dimensionality)

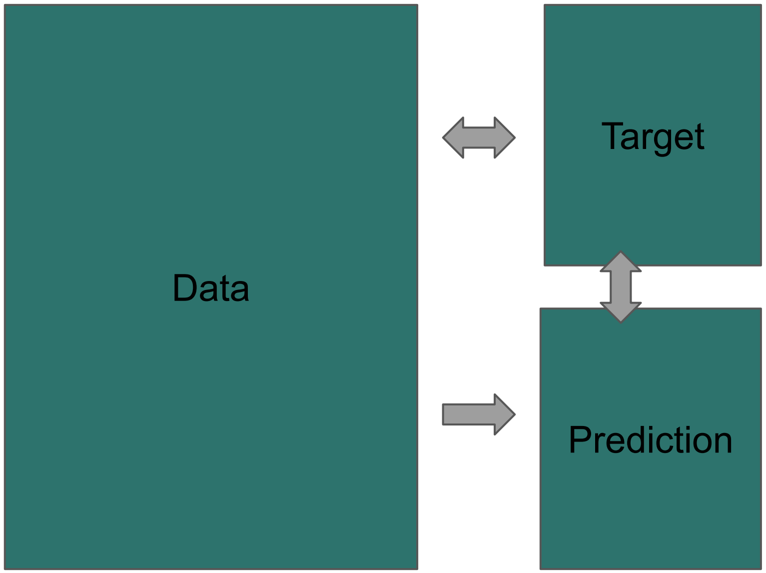
… but not necessarily in ways that affect the model

How to identify drift?

Standard Approaches To Measuring Drift

- Measure model performance in deployment

- Compare distributions of
    - Ground Truth
    - Single input features
    - Prediction
    - Full data sets



Challenges with Standard Approaches

- Measure model performance in deployment
- Compare distributions of
  - Ground Truth
  - Single input features
  - Prediction
  - Full data sets

Don't have ground truth in many cases.

Does a 5% shift in feature 28 matter?

Why is the prediction shifting?

Curse of dimensionality

How to mitigate drift?

example scenarios

Key Takeaway: How to mitigate drift?

Blind model retraining is often not the best answer to counter drift.

Step 1: Understand root causes of drift:

Where is it happening?
When is it happening?
How much is there?
What is causing it?

Monitoring & Explainability – The Critical Hidden Connection!

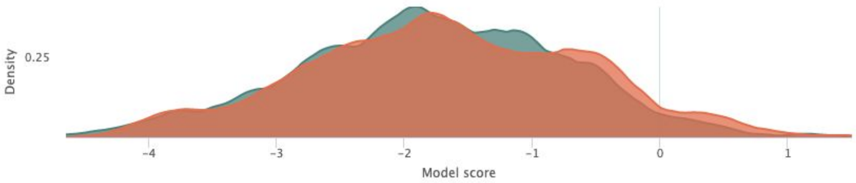
Step 2: Understanding the root cause of drift leads to targeted ways to address drift

How to mitigate drift?

Is the drift caused by an unstable feature?

- Identify and address cause (of prediction drift).
    - Remove a feature without retraining (i.e. replace with mean/mode).
    - Remove a feature and retrain with existing data.

Feature	Importance	Feature value drift ↓	Feature values distribution
annual_inc	1.95% 2.14%	19156.559	
revol_bal	2.65% 2.95%	6702.749	
total_bc_limit	3.68% 3.92%	759.644	



drift in input features → drift in model output

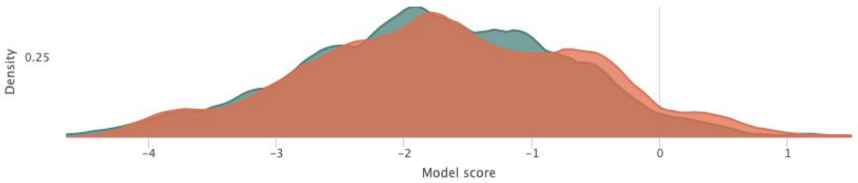
How to mitigate drift?

Is the drift caused by an unstable feature?

Topic for
Q & A

- Explainability technology under the hood
    - Feature importances based on Shapley Values, gradients & more

Feature	Importance	Feature value drift ↓	Feature values distribution
annual_inc	1.95% 2.14%	19156.559	
revol_bal	2.65% 2.95%	6702.749	
total_bc_limit	3.68% 3.92%	759.644	



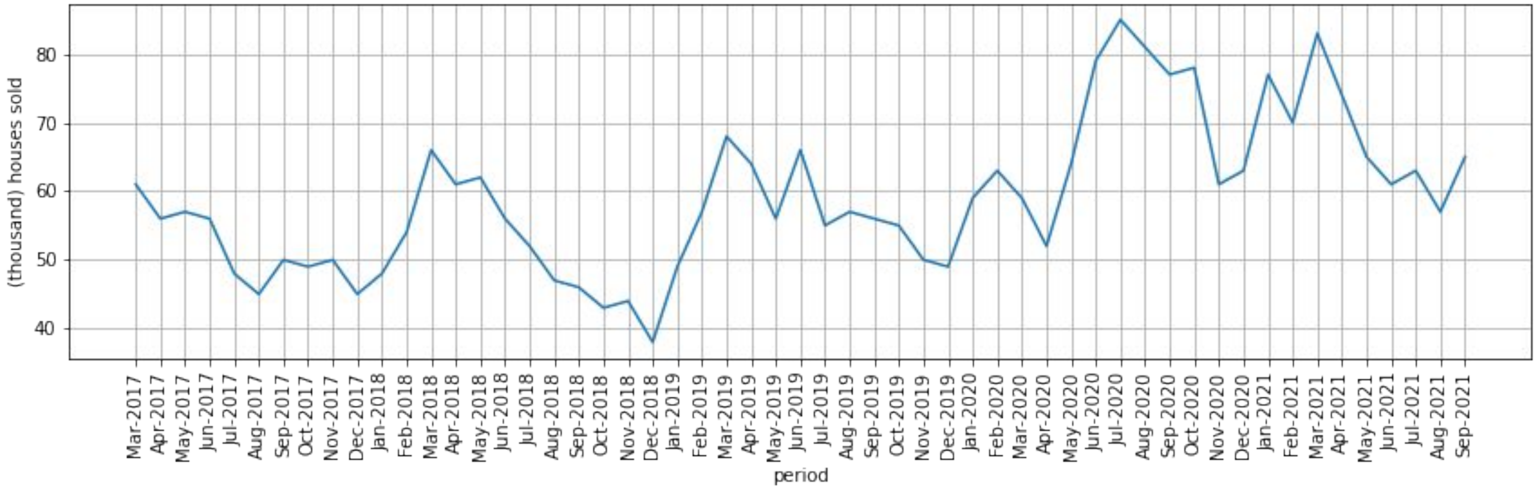
drift in input features → drift in model output

truera

How to mitigate drift?

Is the drift periodic or learnable?

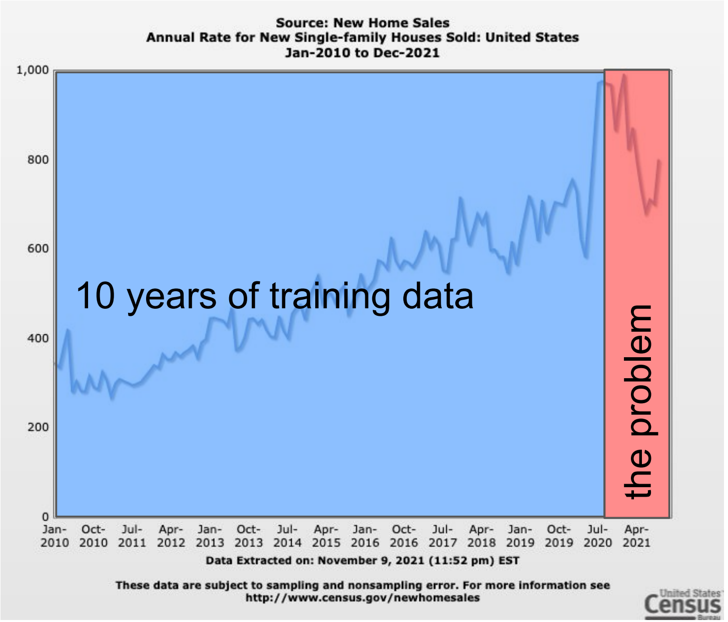
- Concept drift --> Covariate drift with feature engineering
  - Add features to learn periodic change over time.
  - Add indicators of effects of unexpected events ("is-covid" vs "unemployment-rate")
  - Might not need labeling additional data.



How to mitigate drift?

Is the drift sudden relative to training period?

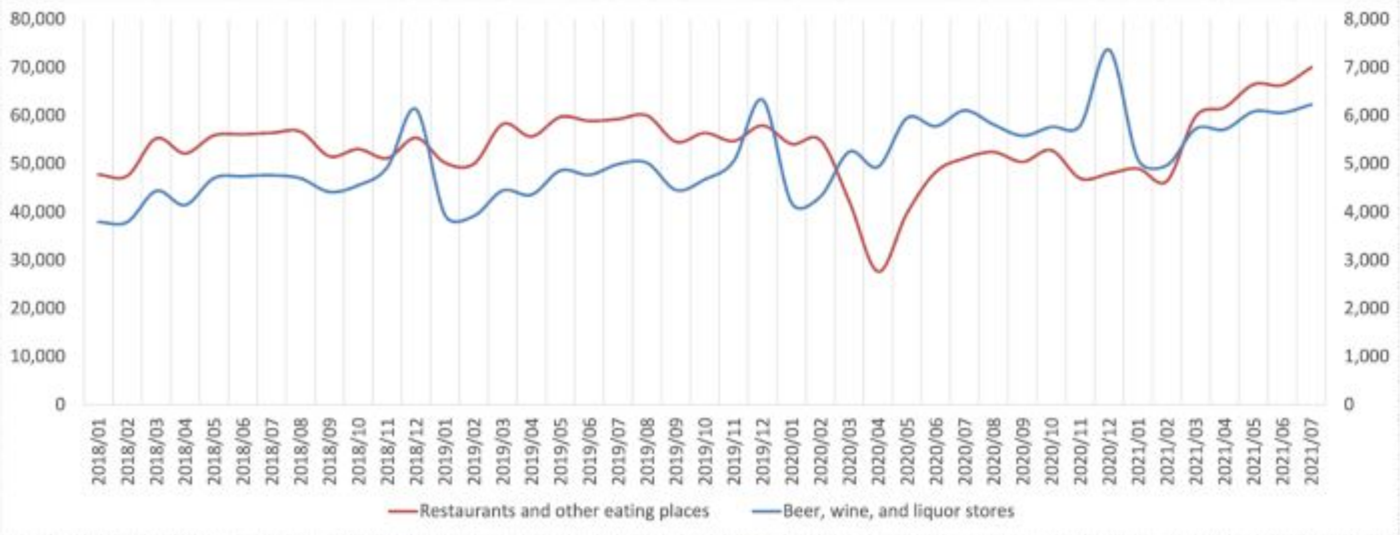
- Drift period may be too insignificant for a retrained model to pick up on it.
- Options:
  - Upweight recent data.
  - Fine tune model with recent data.
  - Identify new features that can help generalize to newer data
    - Example: newer data might be characterized by lower interest rates which might not have been predictive before



What can we do about drift?

Is the drift periodic or learnable?

- Concept drift --> Covariate drift with feature engineering
    - Add features to learn periodic change over time.
    - Add indicators of effects of unexpected events ("is-covid" vs "unemployment-rate")
    - Might not need labeling additional data.



How to mitigate drift?

Is the drift significant enough? Is it affecting model outputs? Is it affecting performance?

- No action may be needed.
    - It might be the case that the model has shifted in a way that is still reasonable.
    - Also needs understanding the root cause of drift.

ML Monitoring

ML Monitoring involves computing drift on data or metrics over time

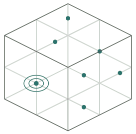
- Track drift over time
    - Basics: Feature Data, Predictions
        - If available: Ground truth, Accuracy
        - Consequences: Influences, MSI, etc
- Set alerts if drift above specific threshold
- Run automated root cause analysis
- Mitigate

Takeaways

- Overview
  - Data drift can happen due to a variety of internal and external causes.
  - Not all drift impacts the model
  - Important to identify consequential drift
- How to identify drift?
  - Different classes of metrics to capture different types of drift: features, ground truth, model output, relationships
  - How to use TruEra to identify root causes of drift
- How to mitigate drift?
  - Not just retrain: Important to understand type and root cause of drift in order to mitigate
  - Retraining, adding features, feature engineering, fixing data quality, and more

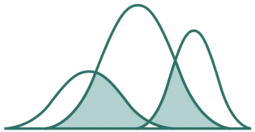
Focus Today: Monitoring Requirements

Fast, precise, and complete.



Model Drift & Performance Metrics

Fast, precise debugging with root cause analysis

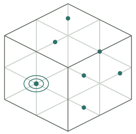


Easy to deploy and scale

AI Monitoring & Explainability:
The Critical Hidden Connection

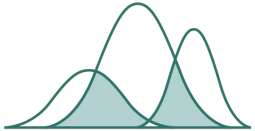
Monitoring Requirements

Fast, precise, and complete.



Broad coverage of model & data quality metrics

Fast, precise debugging



Easy to deploy and scale

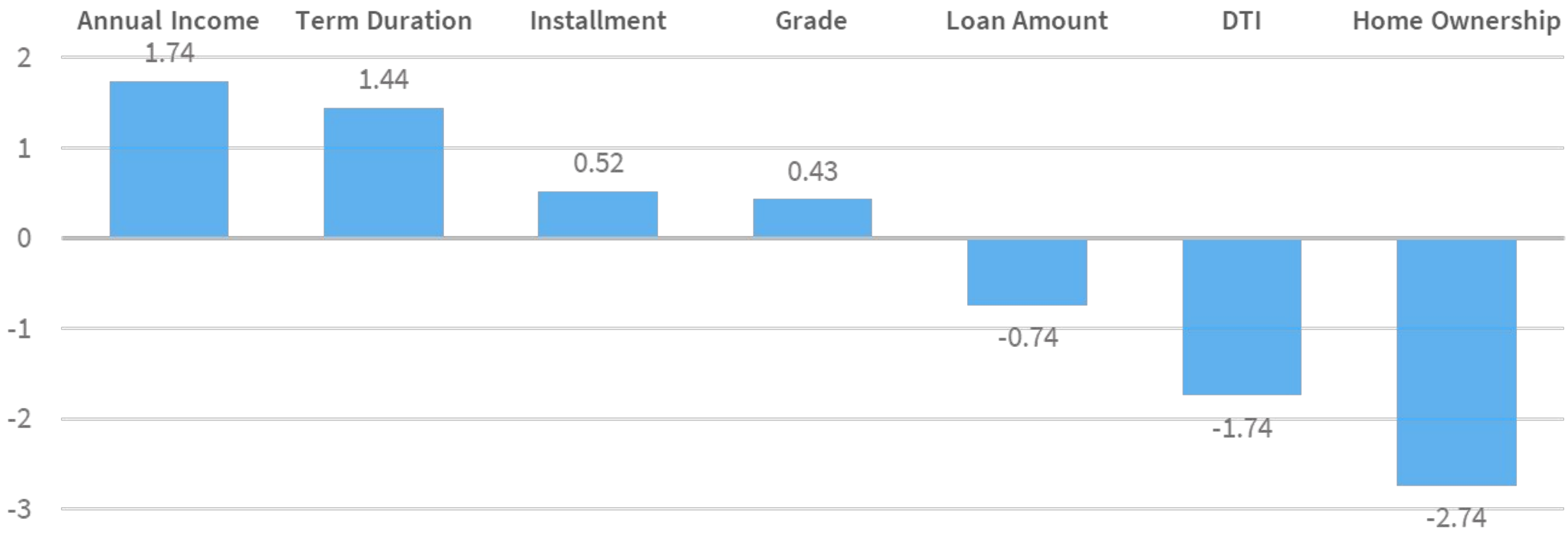
AI Monitoring & Explainability: The Critical Hidden Connection

Thank you!

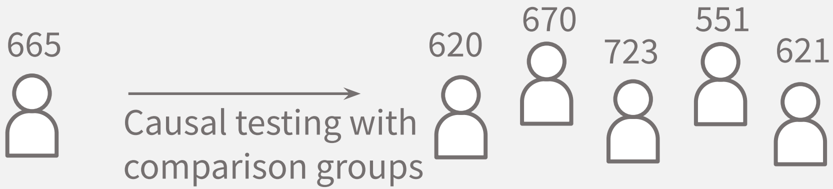
Q&A Time

Appendix: Explainability Methods

Input Feature Importance for a Tree Model



Elements of Explanation Methods

1	QUERY DEFINITION	Why does the model: • have a score of 665 for Jane • have disparate impact • deny Jane
2	OUTPUT COMPARISON	
3	SUMMARIZATION	Of 665, 133 is accounted for by DTI, -45 by income, etc. (Aumann) Shapley accurate estimation

Integrated Gradient

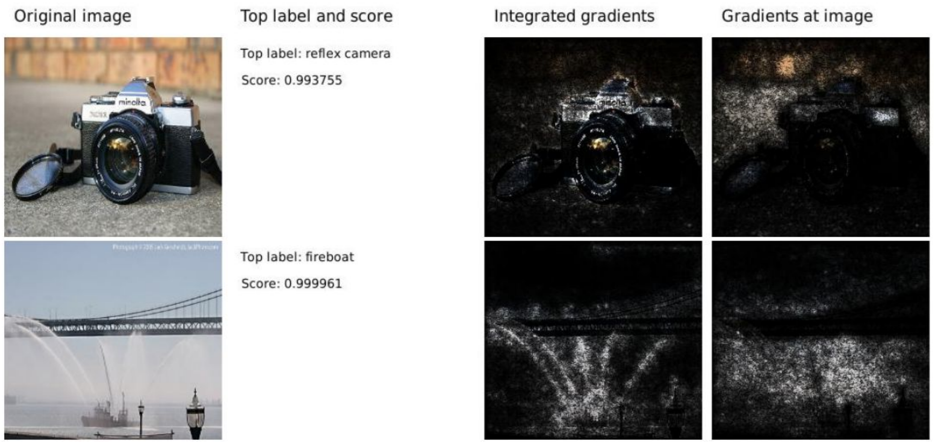
Shapley Value → Aumann Shapley → Integrated Gradient

continues features | differentiable output

[Sundararajan et al. ICML 2017]

Integrated Gradient is the **only** path method that satisfies

- Symmetry
- Dummy
- Efficiency(Completeness)
- Additivity

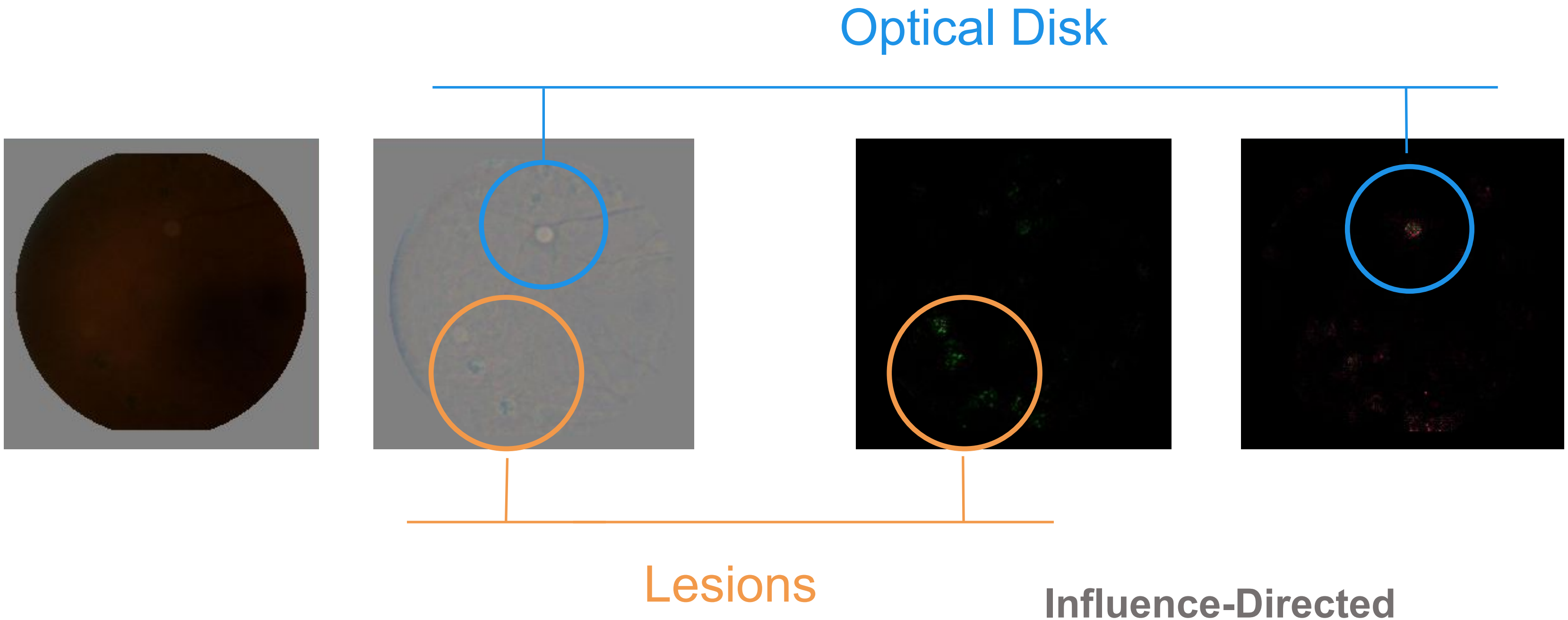


What Makes Orlando Bloom Orlando Bloom?

Internal explanation for a deep network

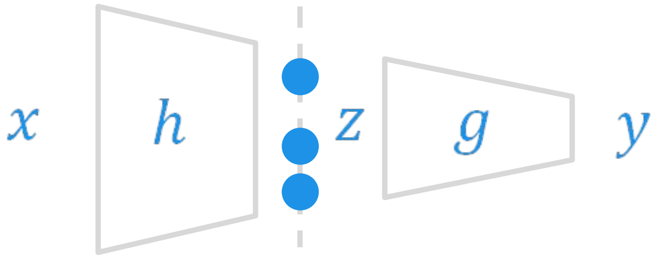
Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li, ITC '18

Detecting Diabetic Retinopathy Stage 5



Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li 2018

Requirements for “Good” Explanations

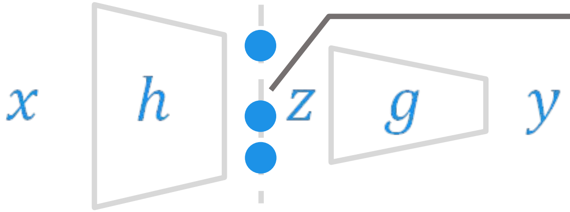


Causal	Succinct	Distributional Faithfulness
Identify features that are causing model predictions	A “few” features explain model predictions	Model is fed “familiar” inputs

Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li, ITC ‘18

Distributional Influence

Influence = average gradient over distribution of interest



$$y = F(x) = g(z), z = h(x)$$

$$\mathrm{I}_j^s(F, P) := \int_{x \in \mathcal{X}} \frac{\partial g(z)}{\partial z_j} P(x) dx$$

z_j

Gradient

For input x
[note z = h(x)]

Weighted by probability
of input x

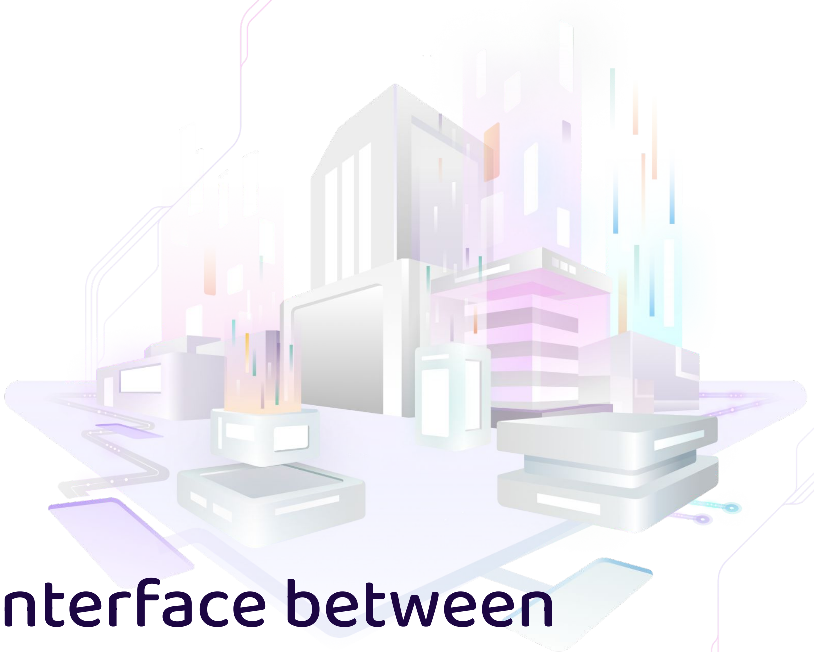
Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li, ITC ‘18

Observability at the long tail:

Why sampling production data doesn't work for rare events

Bernease Herman
Data Scientist, WhyLabs
Data Council Austin
March 23, 2022 in Austin, Texas

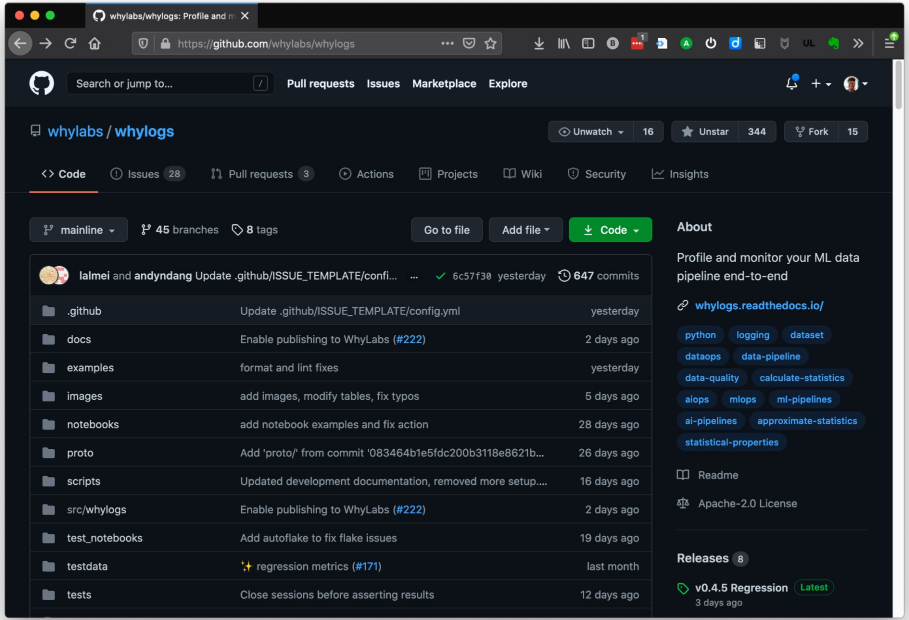
WHYLABS



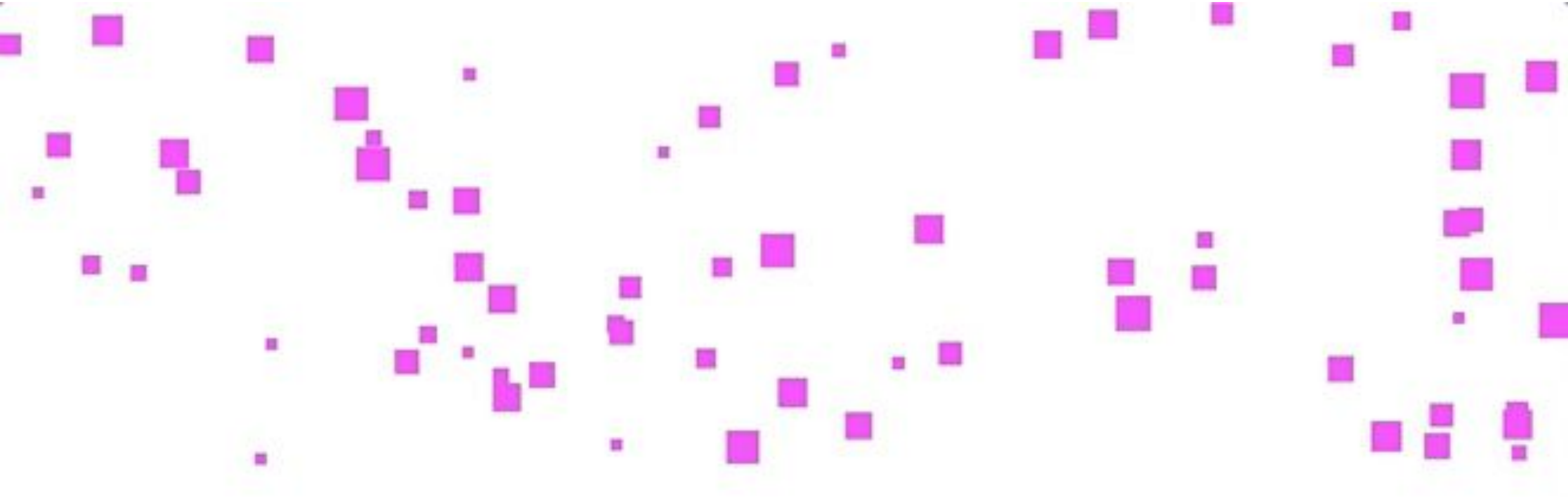
On a mission to build the interface between human operators and AI applications whylogs

bit.ly/whylogs:

Telemetry for the ML stack

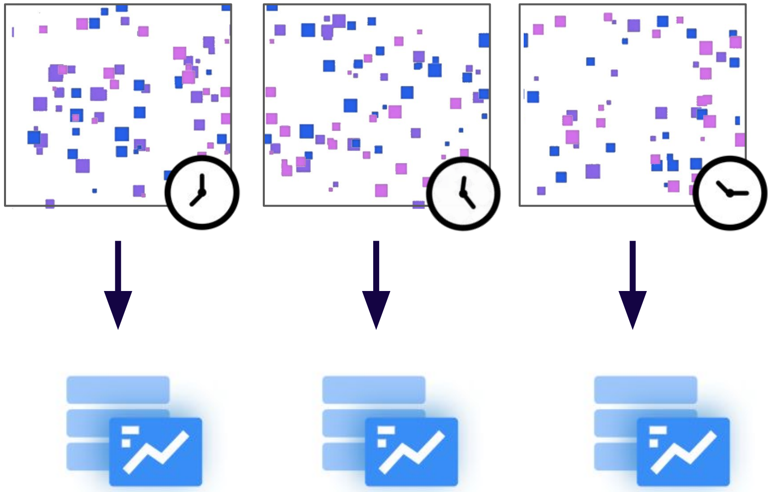


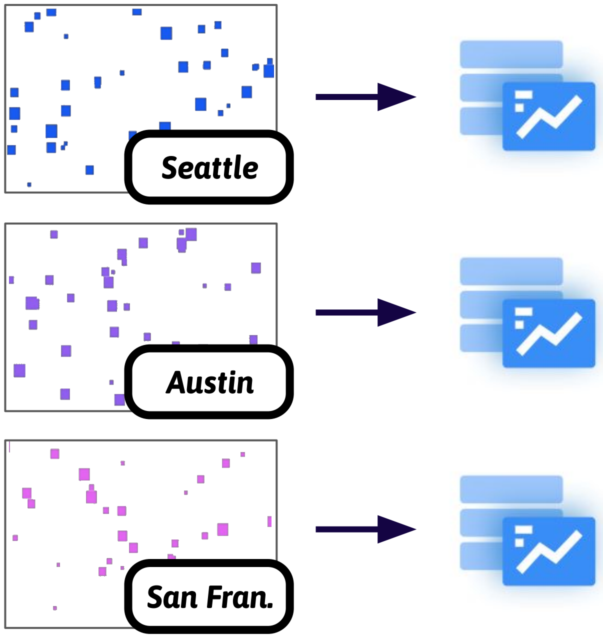
Production ML data is often voluminous, dynamic, and increasingly in the form of streaming data



Complexities of (1) scale and (2) streaming data

Many practitioners try simple sampling techniques; others slice data into segments based on time and other characteristics before conducting analysis





Comparing static windowing, sampling, and profiling

Median and quantile calculation include the following popular approaches:

Static metrics on subsets of data

Predetermine important metrics and store only that information

Random sampling

Store a random sample of the data for further analysis

Data profiling for streaming data

Advanced data structures and algorithms for summarizing data and error

Capturing simple pre-selected metrics for ML data...

```
metrics: {
    mean: 8.0,
    standard_deviation: 1.24,
    quantile_0.25: 5.2,
    ...,
    accuracy: 0.89,
    precision: 0.75,
    recall: 0.92,
}
```

Static metrics approach

Pros:

Fast access to key metrics

Low storage size

Actual metrics on single batch

Cons:

Requires metric pre-selection

Non-mergeable

… isn't enough for root causing production systems!

Using simple pre-selected metrics alone,
you can not answer the following:

Est. value of new metric *x* on prior data?

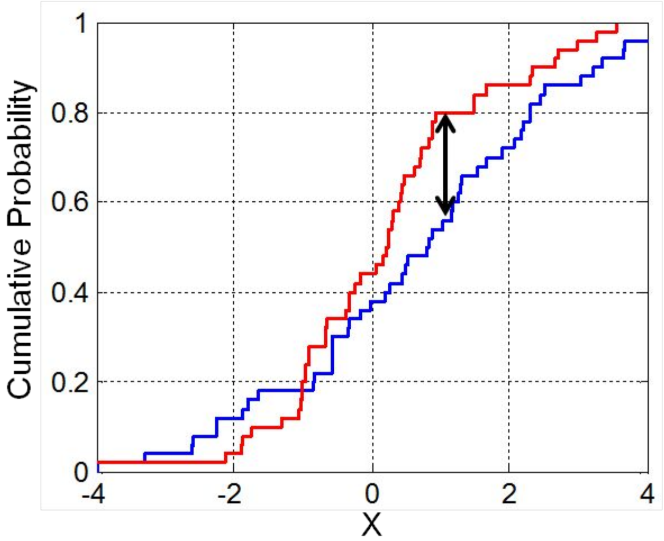
Est, overlap of data with set {*a*, *b*, *c*}?

Relative rank of value *x* on last year's data?

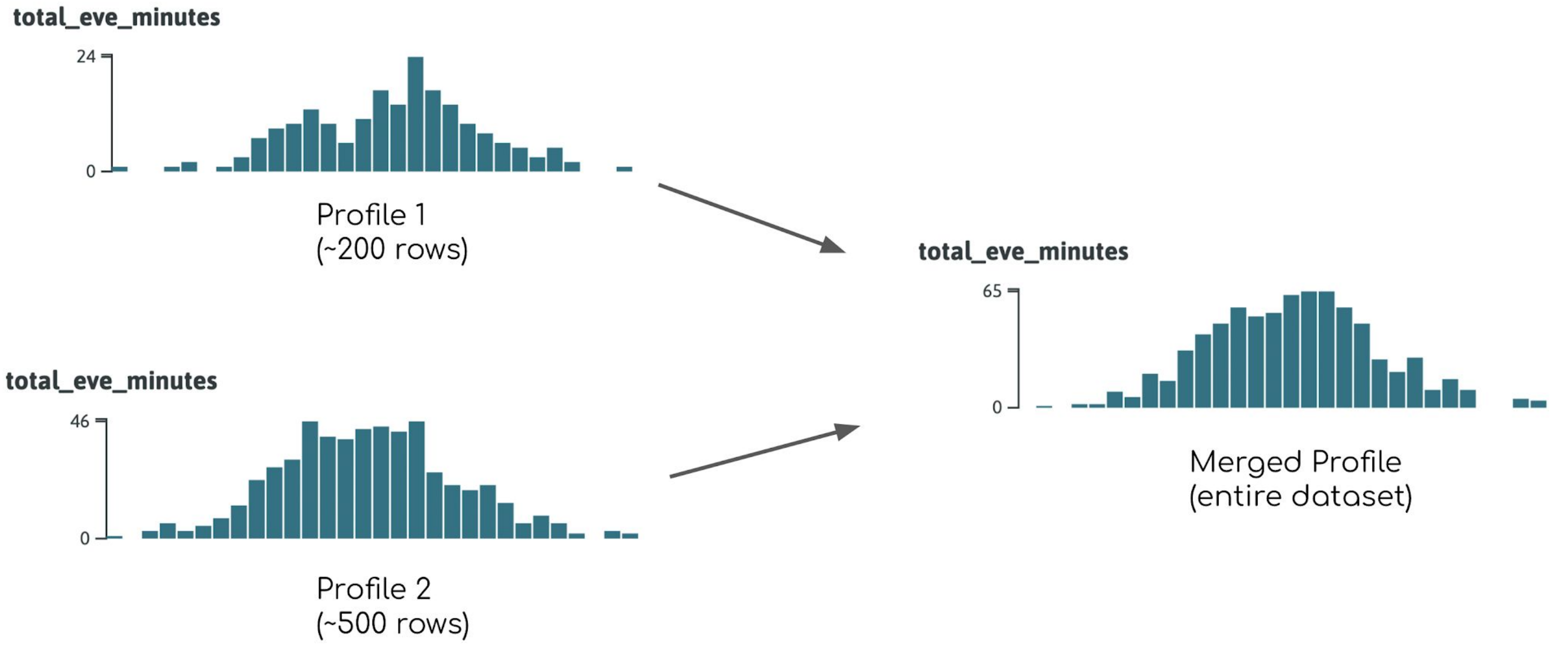
Distribution drift between two datasets?

Error bounds of estimates over the last month of data?

…and many more.



Data mergeability is critical for observing the long tail and rare events



Randomly sampling ML data has issues as well.

```
Sampled rows: 495K                Total rows: 198MM

0 Transaction ID,Customer ID,Quantity,Item Price,
  Total Tax,Total Amount,Store Type,Product
  Category,Product Subcategory,Gender,City Code,
  Age at Transaction Date,Transaction Type,
  Transaction Week,Transaction Batch
1 T24951240379,C267987,12,19.1,24.066000000000003,
  1306.85256,e-Shop,Electronics,Personal
  Appliances,M,9.0,24.0,Purchase,0,2
2 T54251889351,C267740,-3,54.2,17.073,-927.11268
  00000001,MBR,Books,Non-Fiction,M,2.0,36.0,Cancel
  lation,0,2
...
```

Random sampling

Pros:

Same format as original data

High flexibility

Batch or streaming data

Mergeable

Cons:

Poor estimates on tail/outliers

Poor precision (based on %)

High storage size

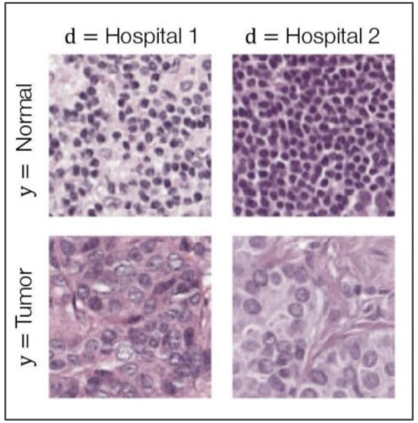
What is data profiling?

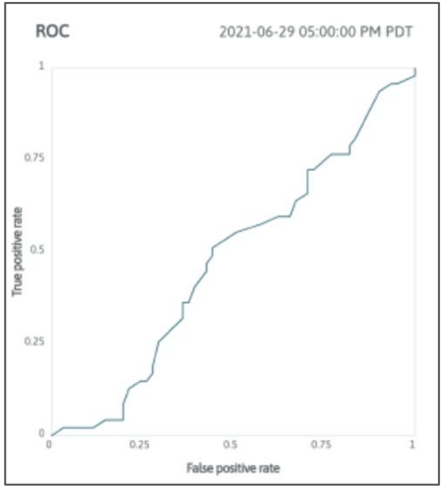
Data profiling is the act of reviewing and analyzing datasets to understand their structure and information. Data profiles can include the following:

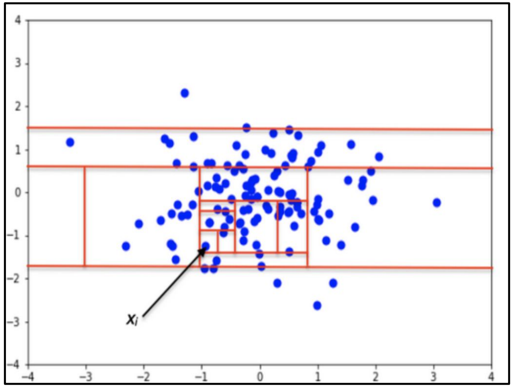
- Collection of descriptive statistics
- Identify different data structures, types, and patterns
- Employ keywords, categorize datasets, and create descriptions
- Conduct data quality examinations
- … and more.

Source: Hanh Truong, “What is Data Profiling?”

Data profiling can include static metrics, but can also contain many more advanced tools needed for analysis

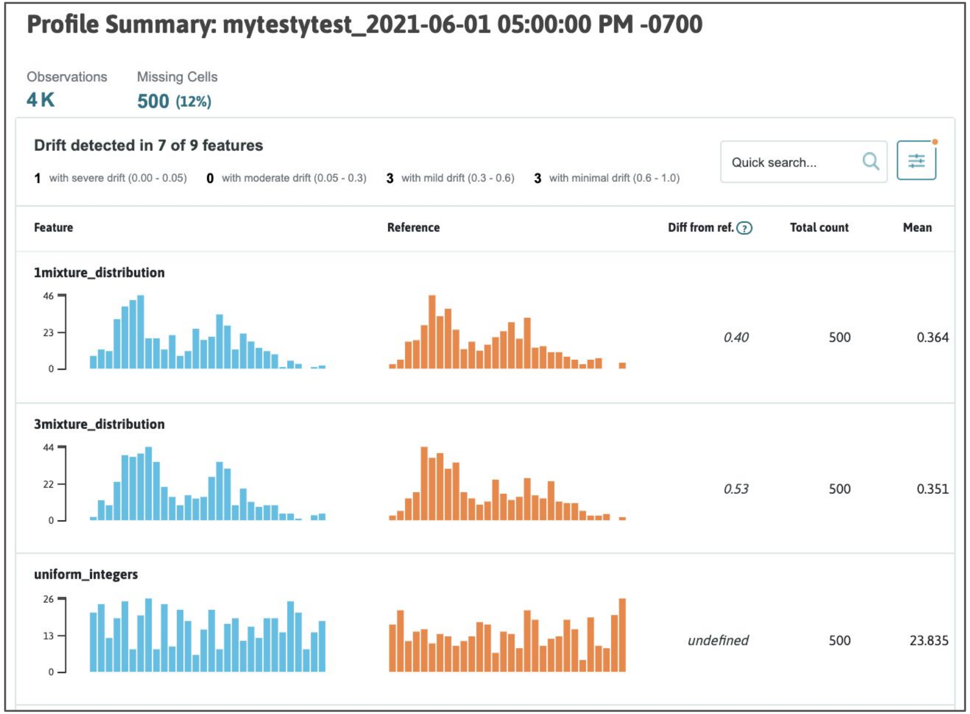






E.g., error bounds for estimates, feature importance, outlier detection, surrogate models.

Sketch-based data profiling for ML data



Data profiling approach

Pros:

Fast access to key metrics

High flexibility

Low memory and storage size

Mergeable

Built on peer-reviewed algos

Cons:

Requires some pre-selection

Underlying algorithm complexity

Building a profiling standard for ML data

Properties of sketch-supported profiling for logging, analysis, and monitoring of ML systems:

- **Lightweight**
- **Configurable**
- **Mergeable**
- **Streaming**
- **Statistically sound**

Powered by:

Apache® DataSketches™

How it works: Notation for median and quantiles

For a stream of numbers $x_1, x_2, \ldots$
with current stream length N:

Rank, $rank(x)$

Number of elements $\leq x$

Relative rank, $r(x)$

Normalized rank, $\dfrac{rank(x)}{N}$

Quantile, $quantile(q)$

Value x s.t. $rank(x) = qN$ or equivalently, $r(x) = q$

Median example

Values: 5 4 1 5 6 2

Sorted: 1 2 4 5 5 6

In this example,

$$rank(4) = 3$$
$$r(4) = \frac{3}{6} = 0.5$$
$$quantile(0.5) = 4$$

Calculating quantiles in P passes over data

Exact calculations

Munro-Paterson proved that the lowest amount of space needed to calculate a quantile in P passes over the data is: $\Omega(N^{1/P})$

You'd need to store N data points to calculate the quantile exactly in streaming setting. Not acceptable!

Approximate calculations

Data sketching techniques allow us to calculate approximate quantiles much more efficiently and in one pass, if desired for streaming.

Numerous algorithms, but KLL (what we use in **whylogs**):

For a single quantile: $(1/\epsilon)loglog^2(1/\epsilon\delta)$

For all quantiles: $(1/\epsilon)loglog^2(1/\delta)$

A brief look at how quantile sketches (KLL) are made

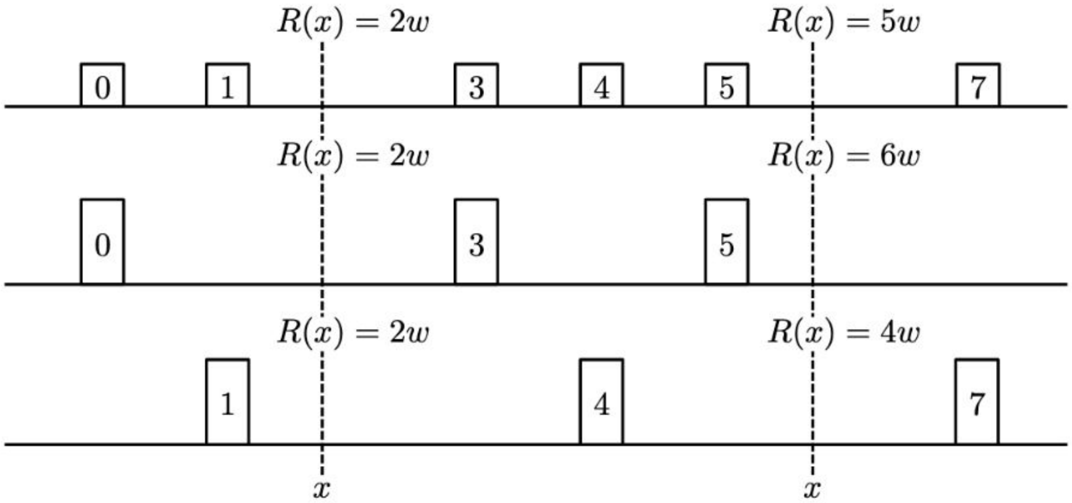


Figure 1: An illustration of a single compactor with 6 items performing a single compaction operation. The rank of a query remains unchanged if its rank with in the compactor is even. If it is odd, its rank is increased or decreased by w with equal probability by the compaction operation.

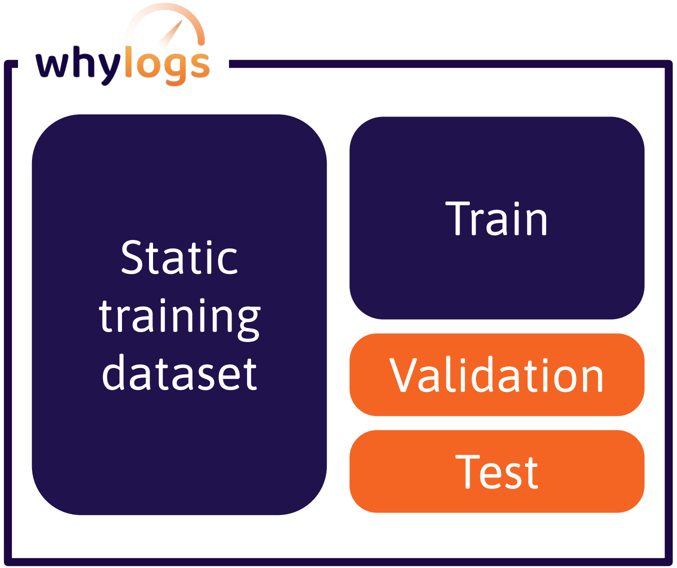
Source: Cardin, Lang and Liberty 2016

Considerations for the whylogs library

Properties of profiling that make whylogs great for logging, analysis, and monitoring ML systems:

- **Lightweight**
- **Mergeable**
- **Configurable**
- **Streaming**
- **Statistically sound**

Profiling training data and other static datasets



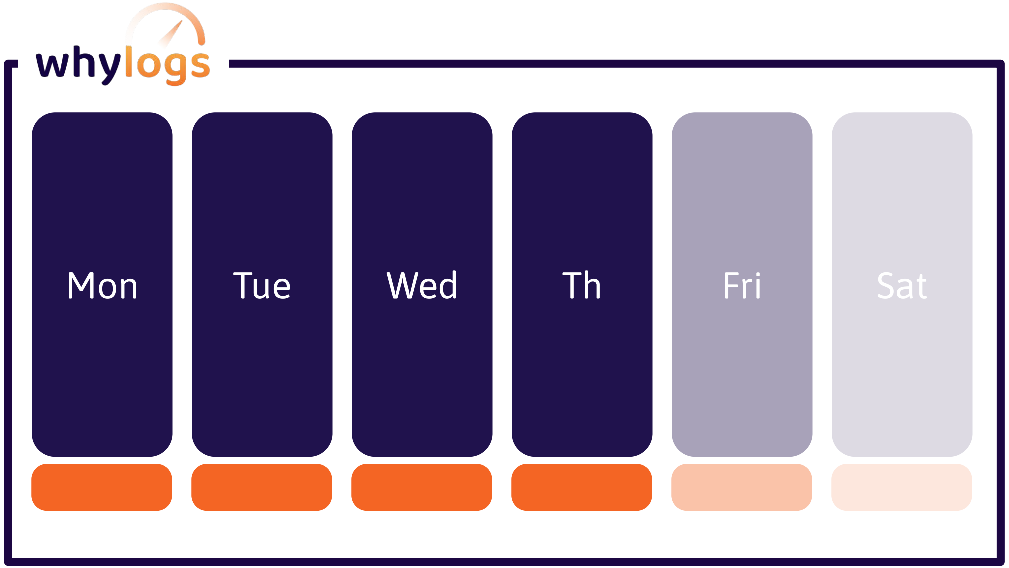
Profile static datasets such as training datasets to store, analyze, and use as a comparison for monitoring.

Uses the same calculations as other profiling, so emphasis on lightweight, speed, and common use cases.

Profiling ongoing production data

Most typical use case, profiling batch or streaming production data.

The underlying data (and perhaps actuals for performance metrics) gets logged regularly while you serve production traffic.



Single profile analysis, but added value for 2+ profiles

whylogs

	Single profile	Two profiles	Three or more
Data documentation	✓	✓	✓
Exploratory data analysis	✓	✓	✓
Data unit testing	✓ NEW!	✓	✓
Ad-hoc comparison to Baseline		✓	✓
Continuous monitoring			✓

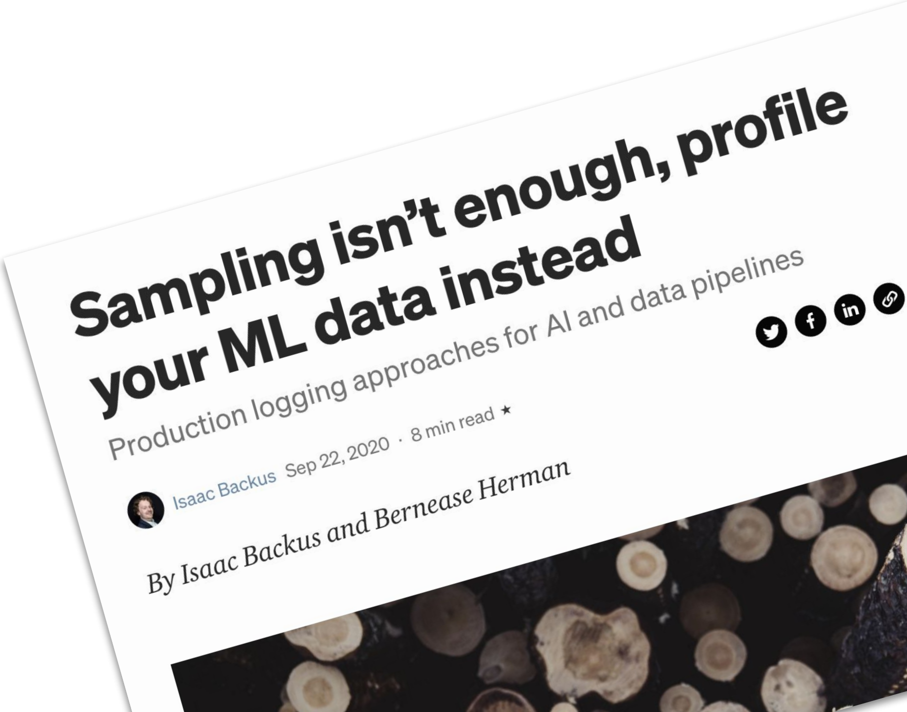
With multiple data profiles, powerful analyses like drift detection, event monitoring, and automated data unit testing become available.

Data sampling versus profiling experiments: Comparing error on common statistical distributions

Experimental procedure:

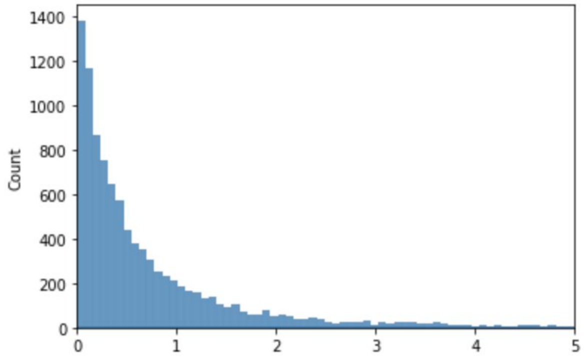
For each statistical distribution:

1. Randomly sample 10^5 records
2. Sample a subset of **`n_sample`** records such that the subset is as many bytes as the profile. This is to compare apples to apples.
3. Compare with exact value on sample
4. Repeat steps 2 through 4 for a total of 24 runs and average the results



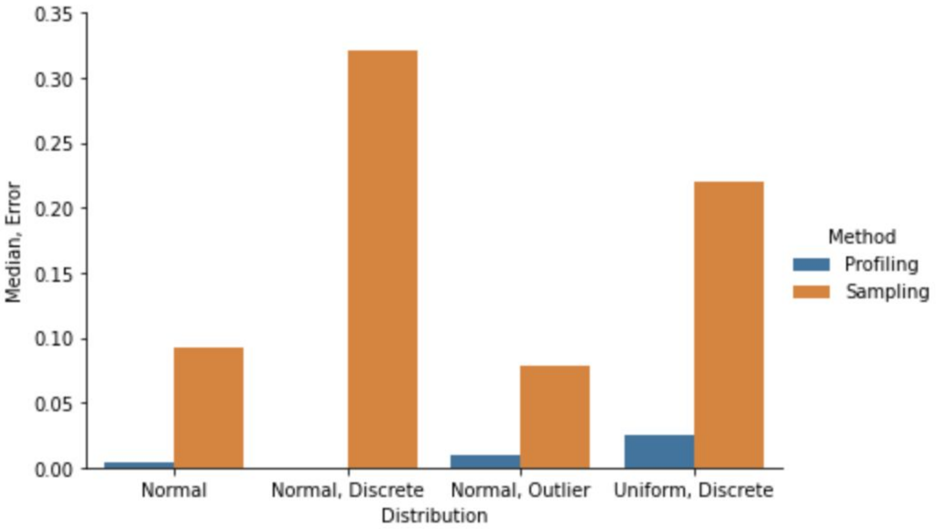
Data sampling versus profiling experiments: Statistical distributions chosen for experiments

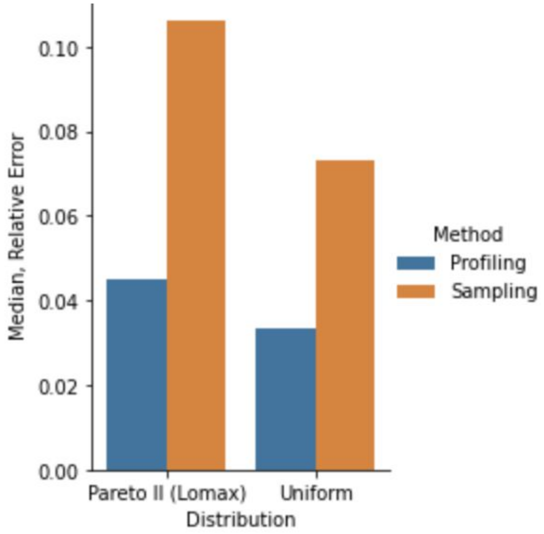
Distribution	Parameters	Purpose
Normal	mu = 0, std dev = 1	A broad class of data. Unskewed, has a tail but is peaked around the center
Uniform	min = 0, max = 1	Data without a tail that is evenly sampled across its domain.
Pareto (type II)	shape = 2, min = 0	A broad class of skewed data with a long tail/outliers.
Discretized normal	mu = 0, std dev = 1 discretized into ~10 categories	Non-uniformly sampled categorical data, occasionally with outliers
Discretized pareto (type II)	shape = 2, min = 0 discretized into ~10 categories	Very non-uniformly sampled categories, with rare events/outliers.
Discrete Uniform	min = 0, max = 1 10 categories	Evenly sampled categorical data



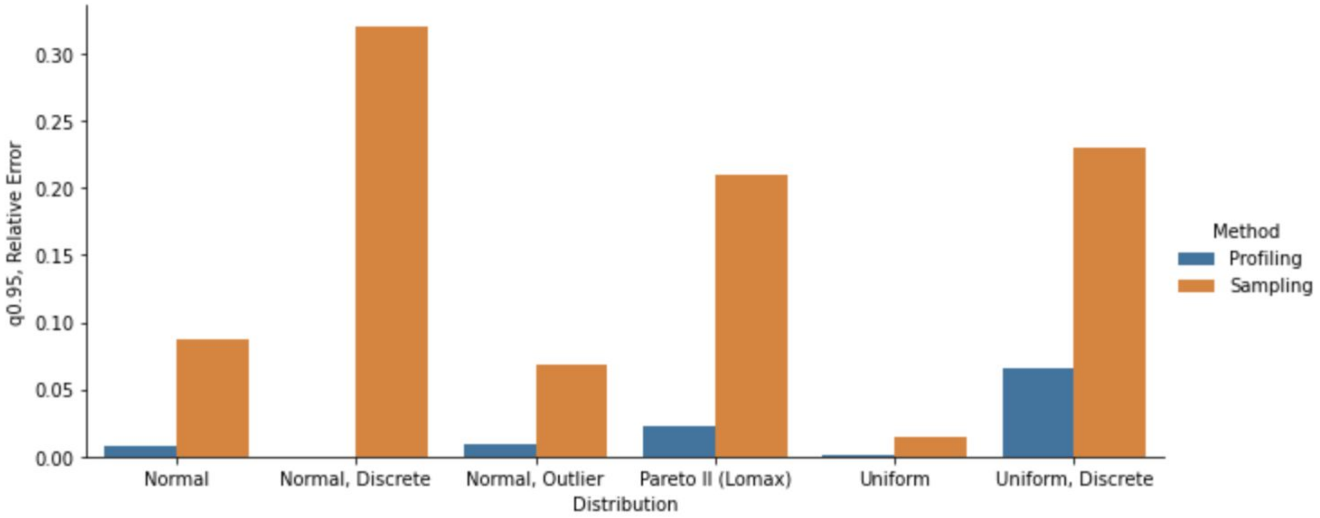
Pareto Type II, or Lomax distribution

Data sampling versus profiling experiments: Comparing error on median across distributions

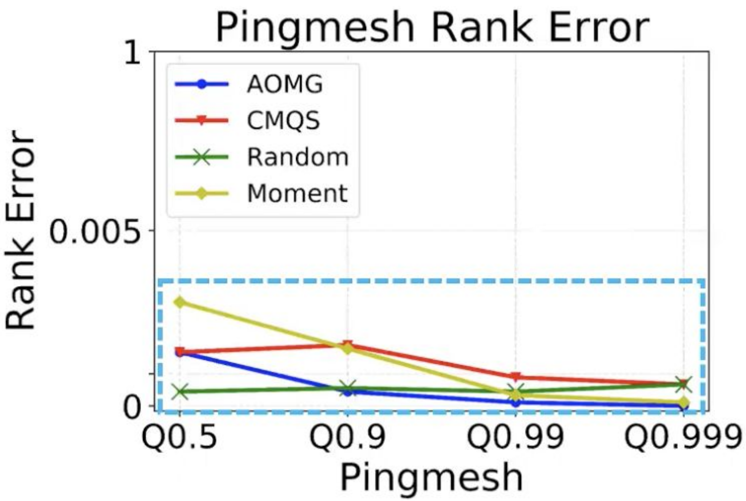


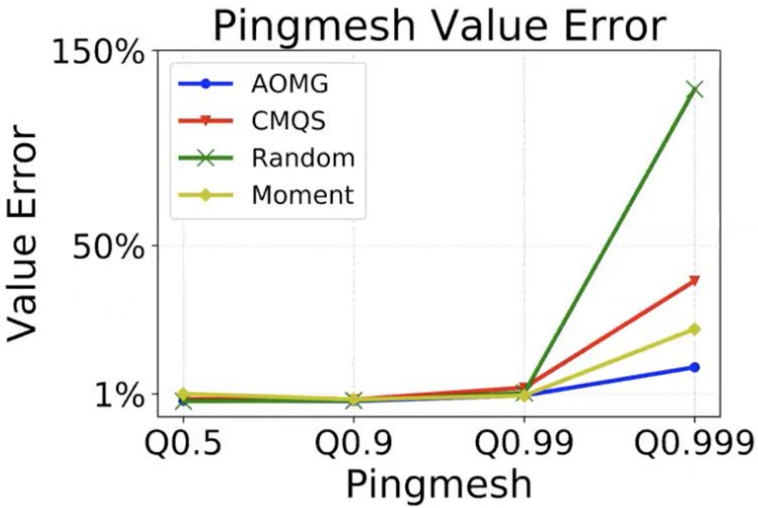


Data sampling versus profiling experiments: Comparing error of across q0.95 across distributions



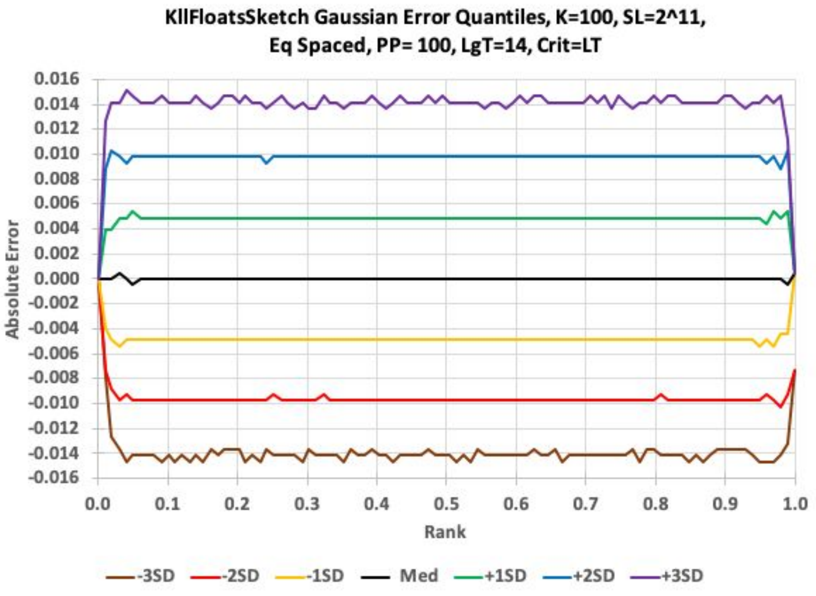
But even low rank error can have a large effect on the tail of the distribution where values may be high

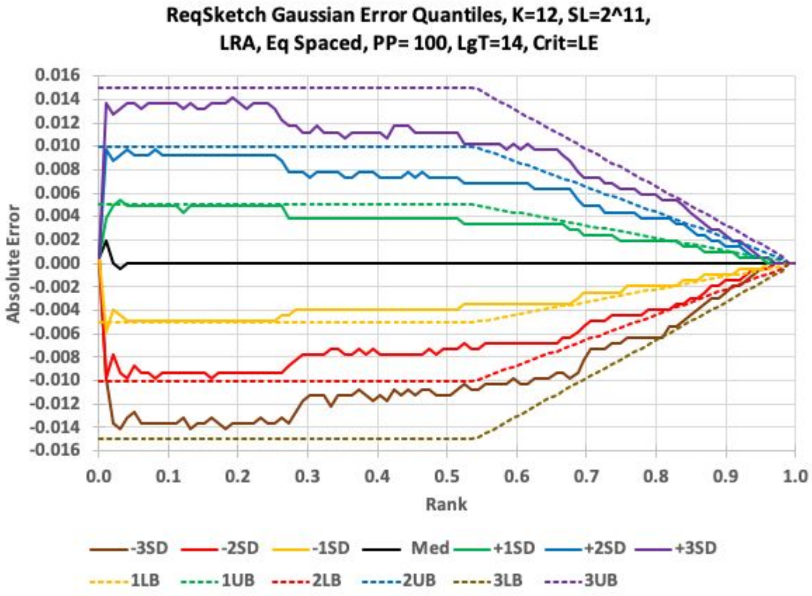




Source: Gangmuk Lim, ICSE 2020 Presentation

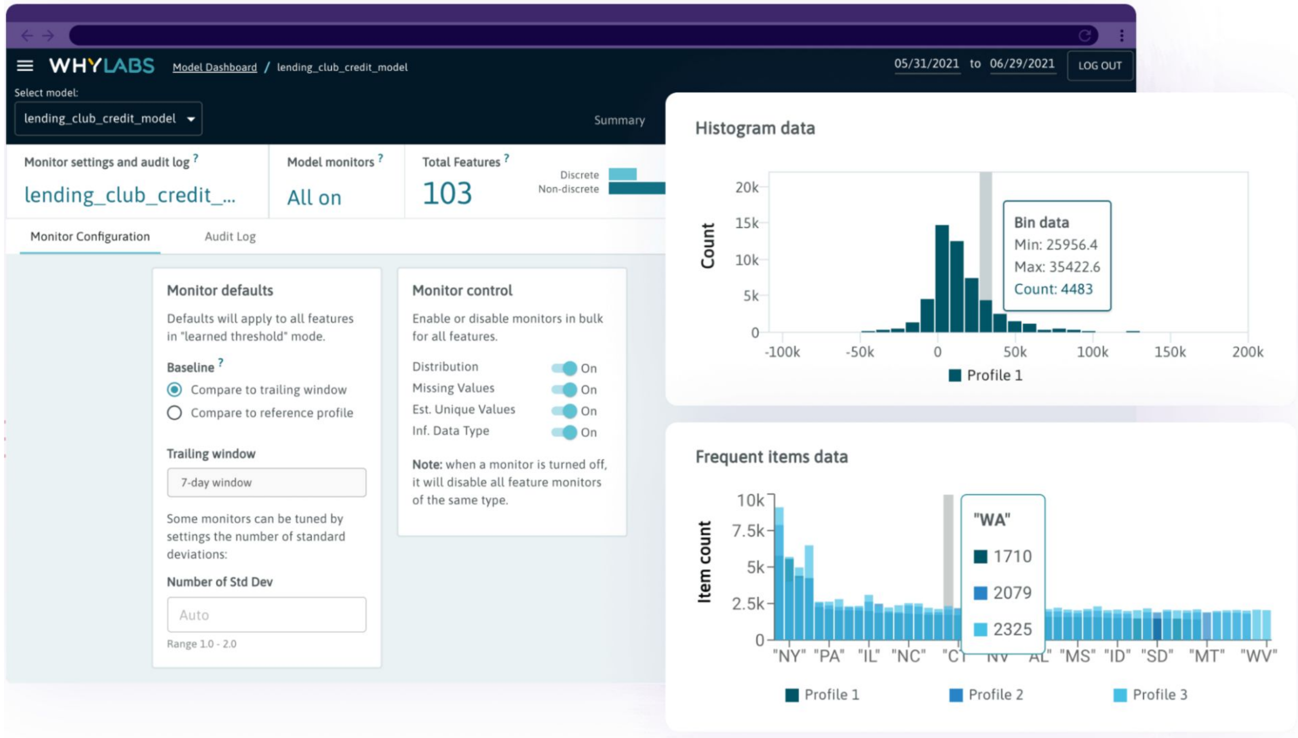
Current sketch treats error evenly across rank, but opportunities to prioritize left or right tail of data





Source: Apache DataSketches, Relative Error Quantiles (REQ)

Want to extend functionality beyond open-source whylogs profiles? Try the WhyLabs SaaS platform



Thank you!
Questions?

Also, help build the open standard for data logging:

github.com/whylabs/whylogs

join.slack.whylabs.ai

Contact me:

In-person at Data Council Austin

Email: **bernease@whylabs.ai**

Social media: **@bernease**

Instructions for getting WhyLabs swag:

- Star the **whylogs** project on Github
- Join our **Community Slack**
- Submit a **form** with relevant info at bit.ly/whylogsswag

A subset of ML issues encountered in production

- Experiment/production environment mismatch
- Wrong model version deployed
- Underprovisioned hardware
- Inappropriate hardware
- Latency/SLA issues
- Data permissions misconfigured
- Untracked changes broke prod
- Traffic sent to the wrong model
- Computational instability
- Customers gaming the model
- PII data exposed
- Expected accuracy doesn't materialize

- Pre-processing mismatch in experiments vs. production
- Retrained on faulty data
- Accuracy improves on one segment, regresses in others
- Outliers predicted incorrectly
- Bias identified
- Correlation with protected features
- Overfitting on training/test
- Surge in missing values
- Surge in duplicates

- Poor performance on outliers
- Data quality issues affect accuracy
- Production data doesn't match test/training
- Accuracy is decaying over time
- Data drift in inputs
- Concept drift in outputs
- Extreme predictions for out of distribution data
- Model not generalizing on new data / new segments
- Major consumer behavior shift

...or it simply doesn't work, and nobody knows why!

Most ML issues are observable from the data itself

- Experiment/production environment mismatch
- Wrong model version deployed
- Underprovisioned hardware
- Inappropriate hardware
- Latency/SLA issues
- Data permissions misconfigured
- Untracked changes broke prod
- Traffic sent to the wrong model
- Computational instability
- Customers gaming the model
- PII data exposed
- Expected accuracy doesn't materialize

- Pre-processing mismatch in experiments vs. production
- Retrained on faulty data
- Accuracy improves on one segment, regresses in others
- Outliers predicted incorrectly
- Bias identified
- Correlation with protected features
- Overfitting on training/test
- Surge in missing values
- Surge in duplicates

- Poor performance on outliers
- Data quality issues affect accuracy
- Production data doesn't match test/training
- Accuracy is decaying over time
- Data drift in inputs
- Concept drift in outputs
- Extreme predictions for out of distribution data
- Model not generalizing on new data / new segments
- Major consumer behavior shift