

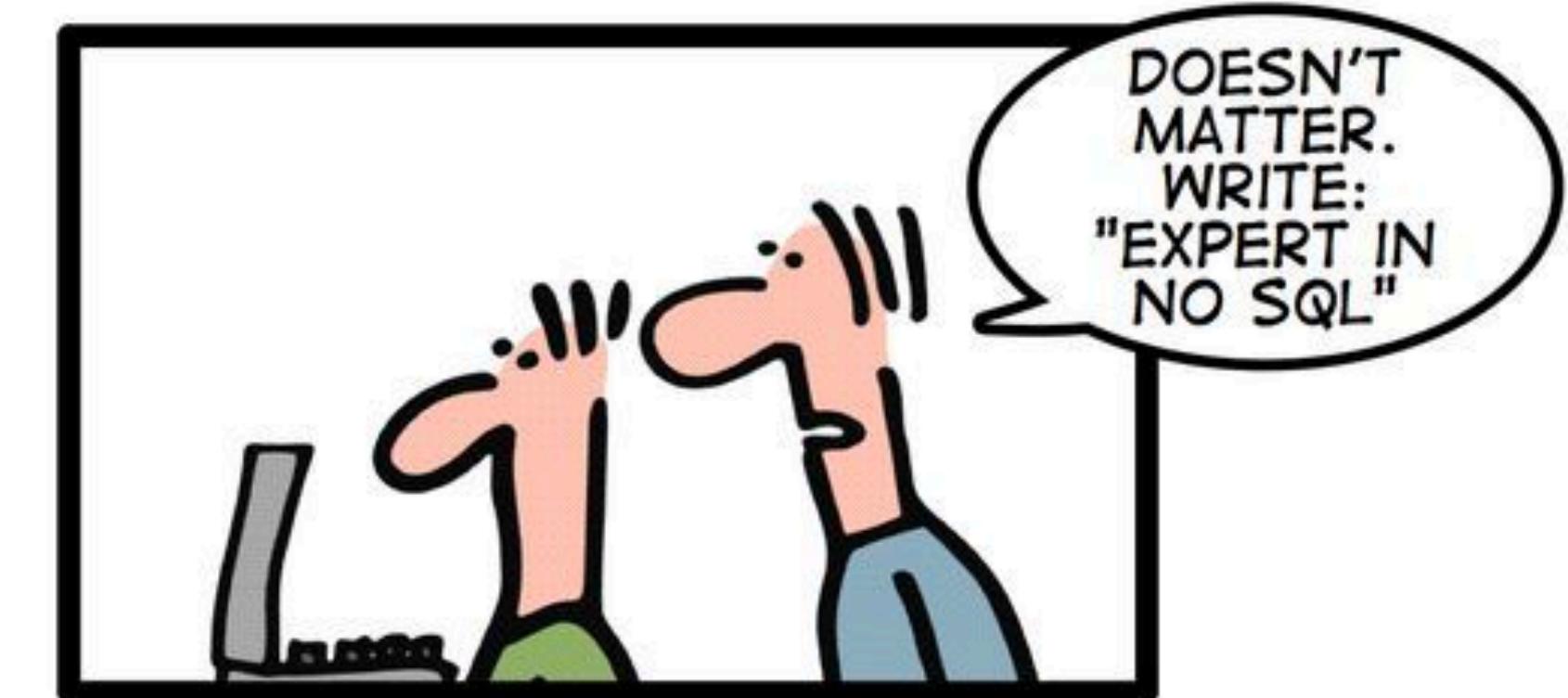
Data Engineering

2019 and Beyond

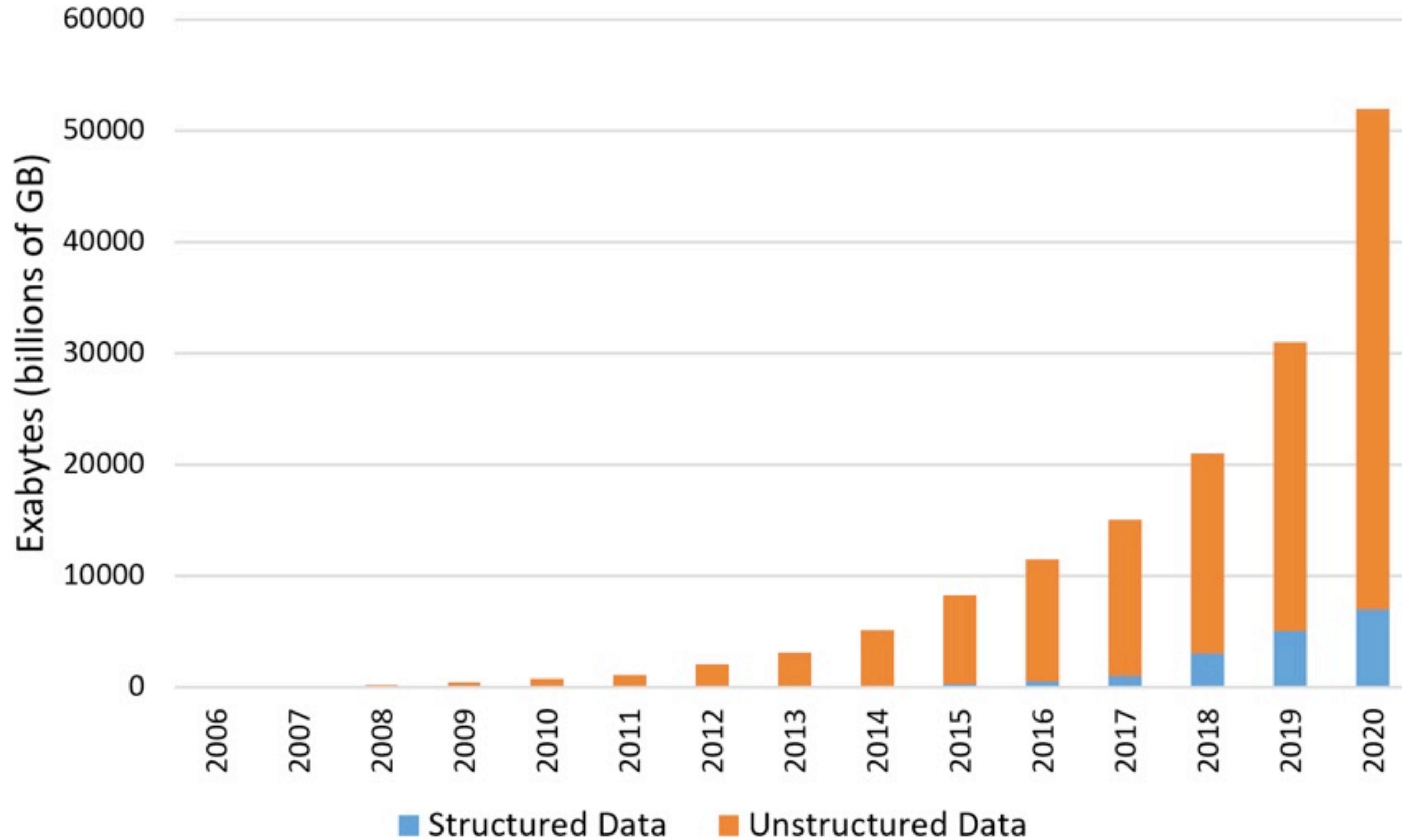


ITZIK FELDMAN | DE PRODUCT MANAGER | @ITZIKFELDMAN

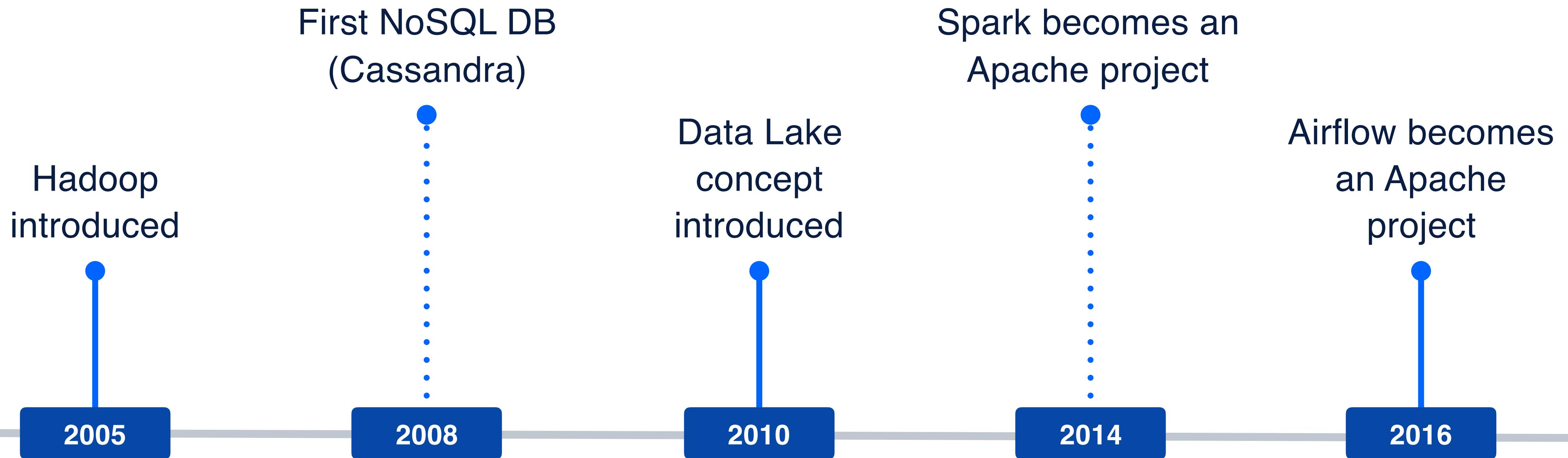
HOW TO WRITE A CV



The Cambrian Explosion...of Data



(BRIEF) BIG DATA HISTORY



By 2020



Growth

Number of users will double and business value per user will double



Simplicity

50% of analytical queries will be generated via search, NLP, voice, or will be automatically generated



Governance

Organisations with a curated data catalog will derive twice as much business value from analytics investments

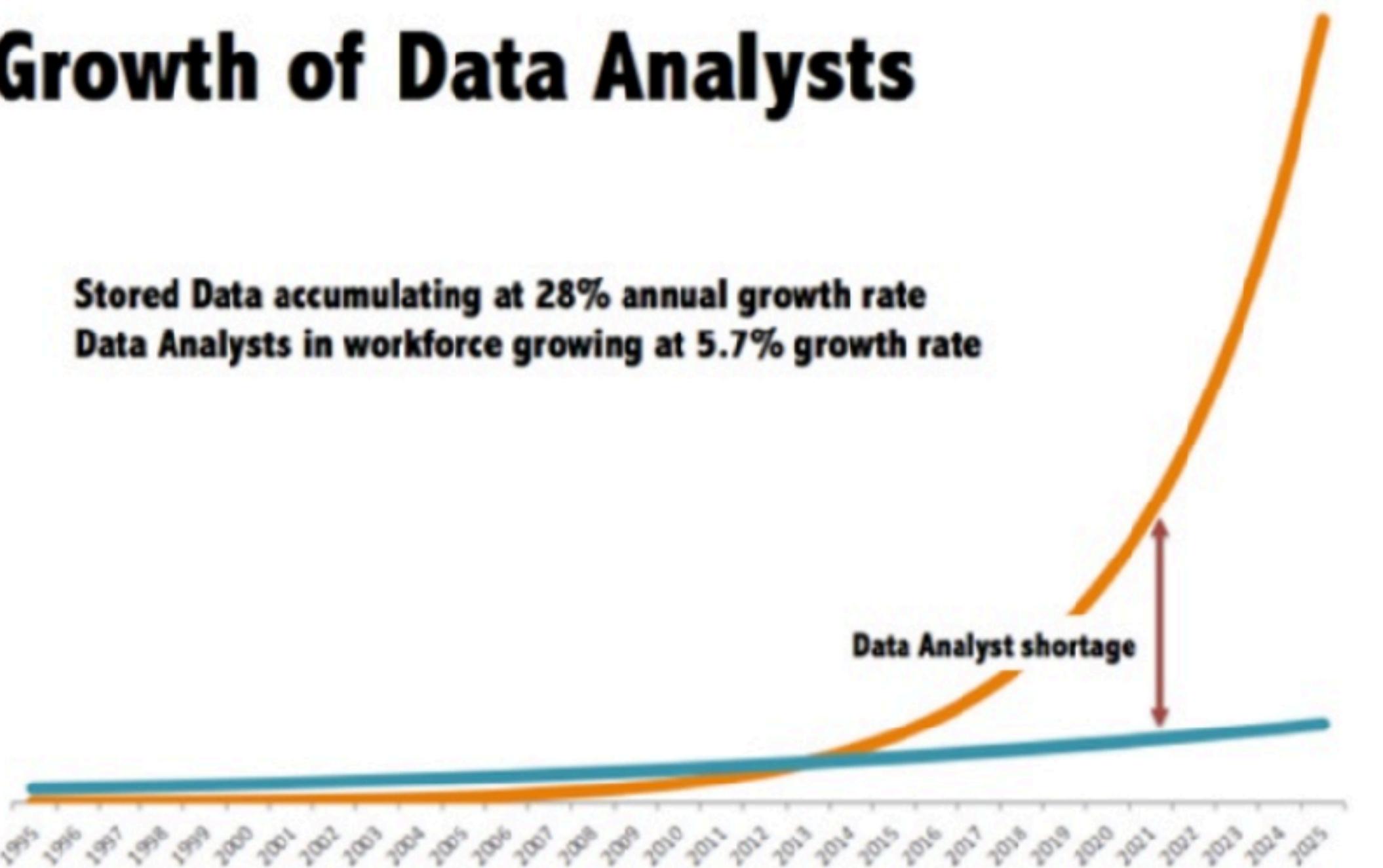


Democratisatio n

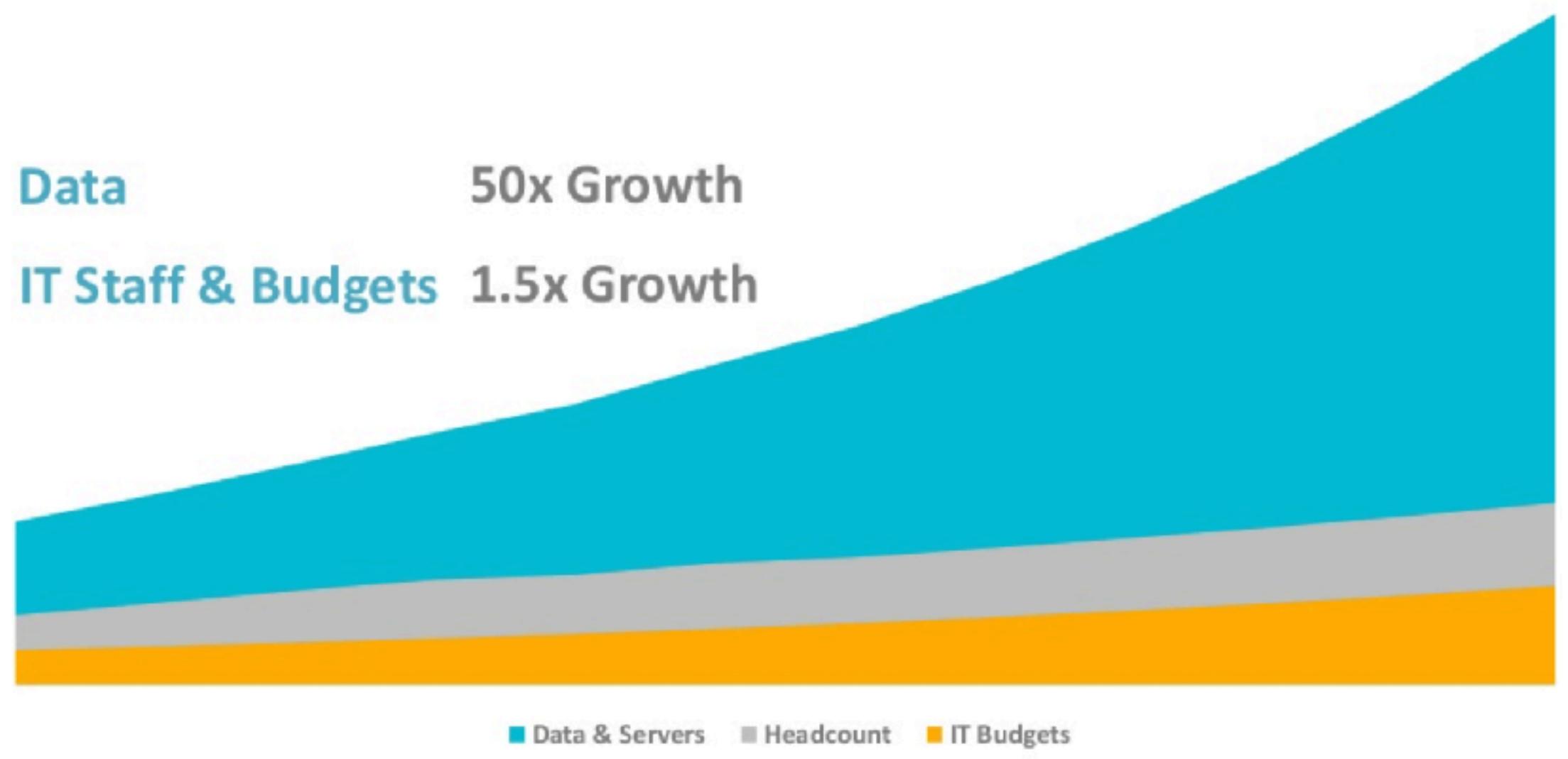
The number of citizen data scientists will grow five times faster than the number of expert data scientists

Challenges

Growth of Data vs. Growth of Data Analysts



We don't
have
enough
analysts to
handle the
data

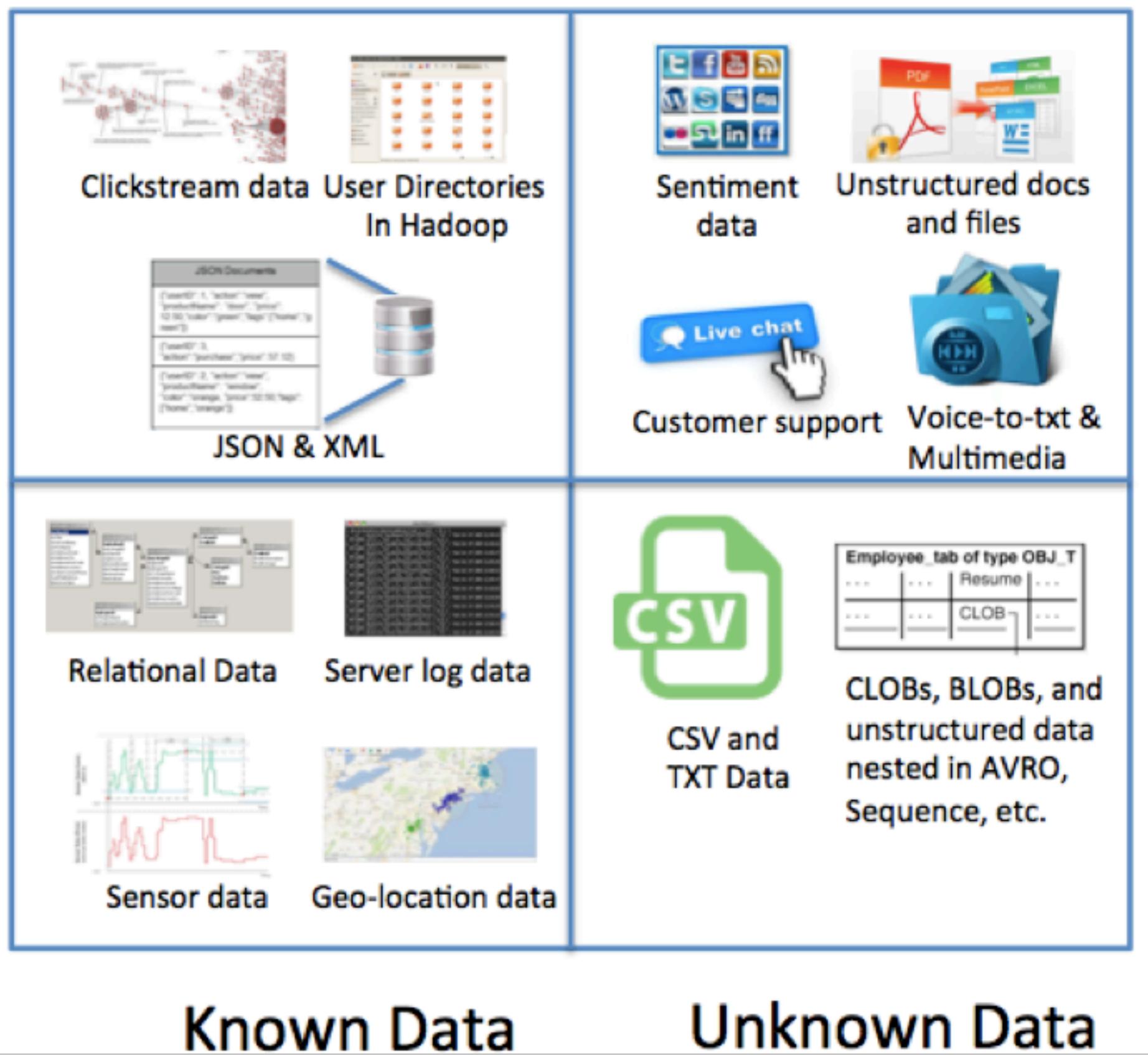


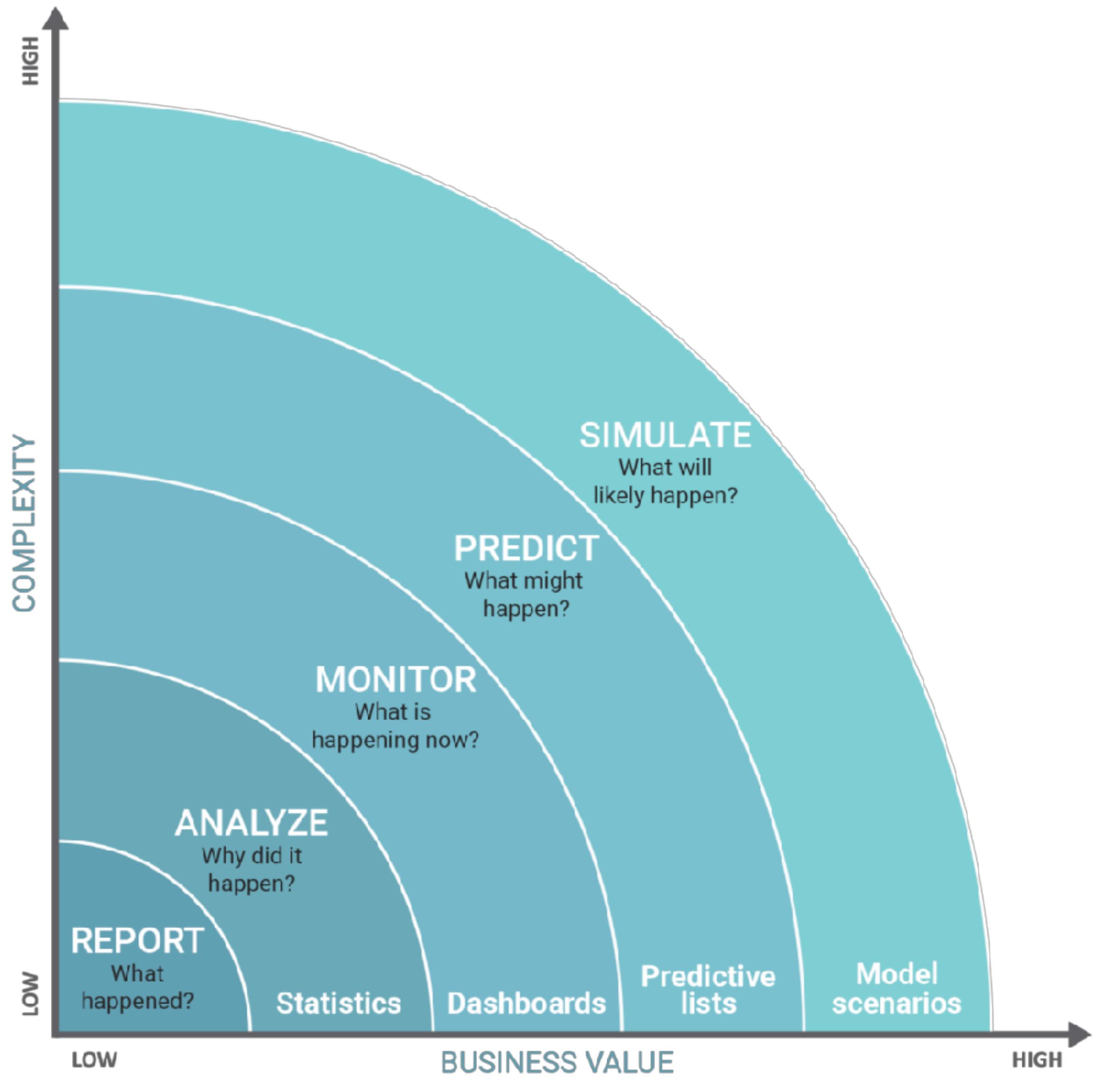
IT budgets
are not
growing as
fast as data

The rise of the unknown

Unclear
Meta-Data

Clear
Meta-Data





To derive business value we need to raise the complexity

Trends

REAL TIME PREDICTIVE ANALYTICS

When predictions are needed in



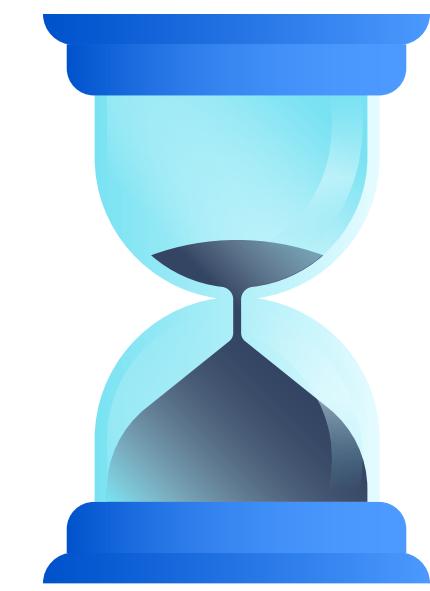
Seconds

- Dynamic Product Behaviour
- Fraudulent transactions



Minutes

Customer Service



Hours

Usage Drop
Predictive Churn

Realtime brings challenges





Pioneer Tax:

- Conventional ETL is batch
- Training ML models on streaming data is a new ground



Pioneer Tax:

- Conventional ETL is batch
- Training ML models on streaming data is a new ground

Demanding SLOs

- Batch failures have to be addressed urgently, streaming failures have to be addressed immediately



Pioneer Tax:

- Conventional ETL is batch
- Training ML models on streaming data is a new ground

Demanding SLOs

- Batch failures have to be addressed urgently, streaming failures have to be addressed immediately

Fault-tolerant infrastructure

- Monitoring, alerts
- Rolling Deployments

A close-up photograph of a baby elephant's front legs and trunk as it kicks a soccer ball. The elephant's skin is wrinkled and brown. The soccer ball is white with black pentagonal panels. The background is a lush green jungle with sunlight filtering through the leaves.

But there are
wins

Other Wins



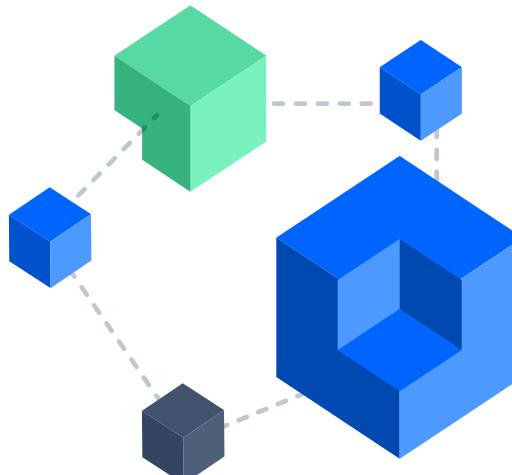
Faster Turnaround

Less impact for the business



Realtime Auditing

Of key metrics



Easy Integration

With other realtime systems



Cost Savings

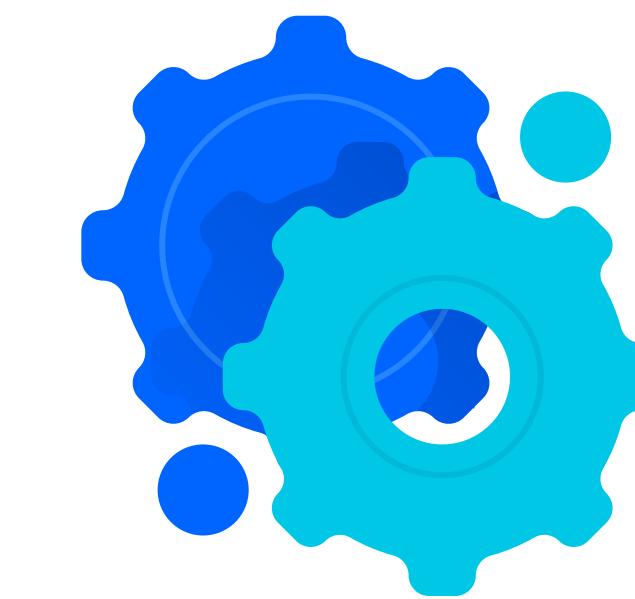
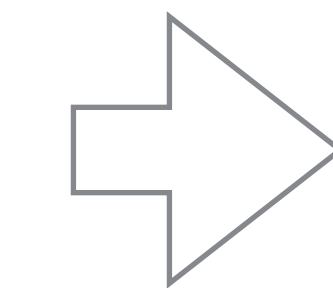
No need for one big cluster

AUTO ML

Methods and processes to
make Machine Learning
available for non-Machine
Learning experts

Workflow Example

User	Event	Time	Properties	Paid?
USER1	Main evaluation Viewed	1/1/2018 12:10:00	Channel = Campaign111	Yes
USER1	JSW/Con. Bundle Viewed	1/1/2018 12:11:00	Source = Main evaluation	No
USER2	Main evaluation Viewed	1/1/2018 12:10:00	Channel = atlassian.com	Yes
USER2	JSW/JSD Bundle Viewed	1/1/2018 12:11:00	Source = Main evaluation	No

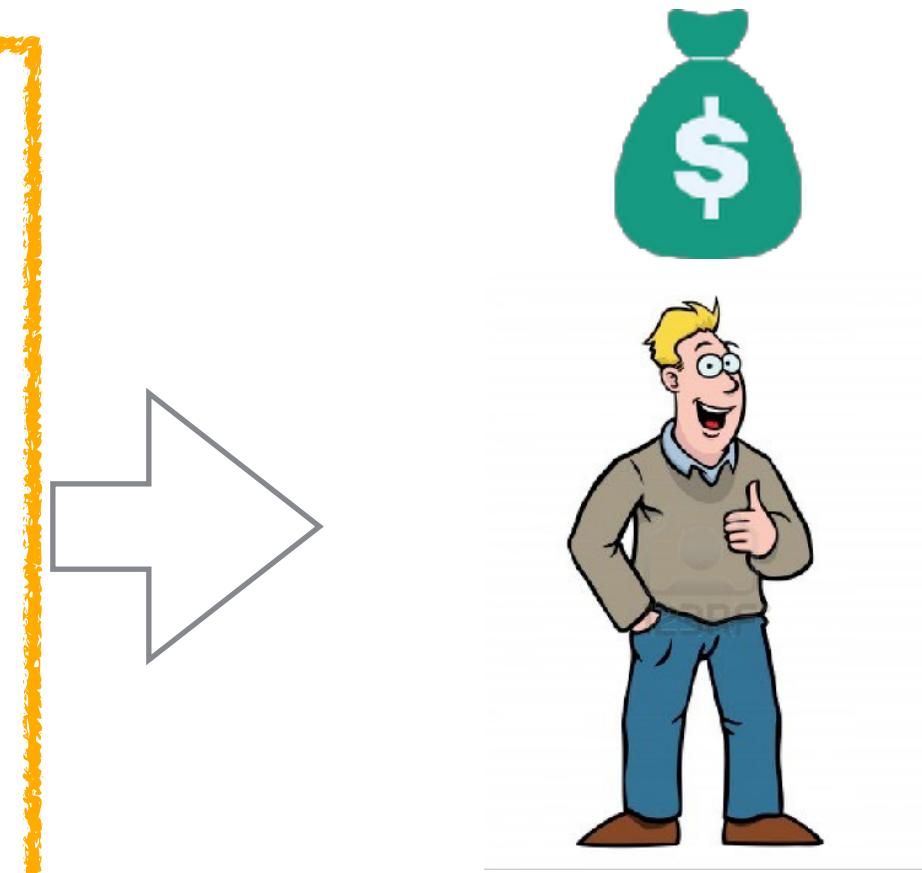


Auto ML

Workflow Example



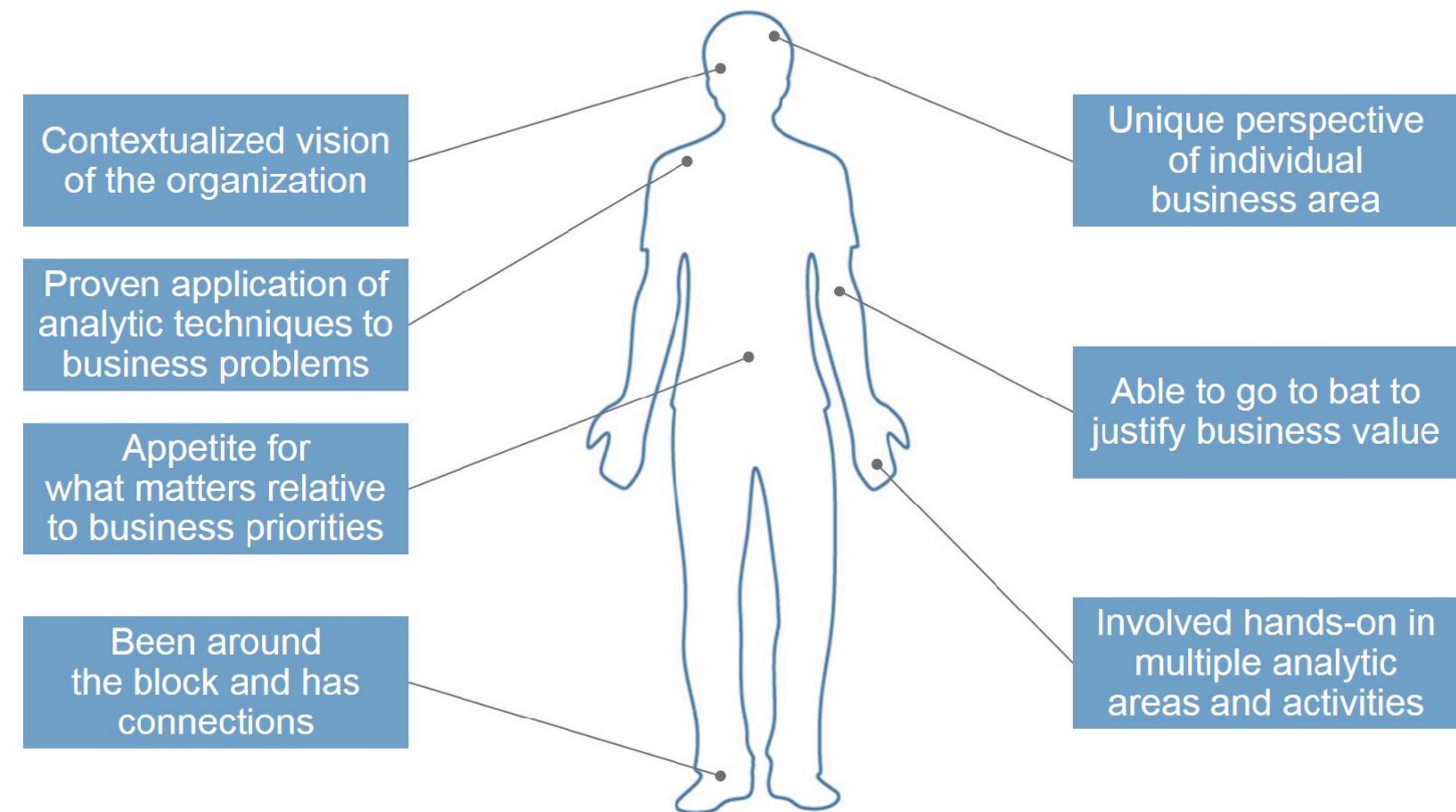
If (user selected JSW/JSD bundle)
AND (user selected confluence) within 1 week
AND (channel = atlassian.com)
Then True
Else False



1/1/2019 - Tried JSW/JSD
from atlassian.com

3/1/2019 - Tried Confluence

Traits of a Citizen Data Scientist



Begin a project by dragging a dataset here

or

simply import from:

ODBC

URL

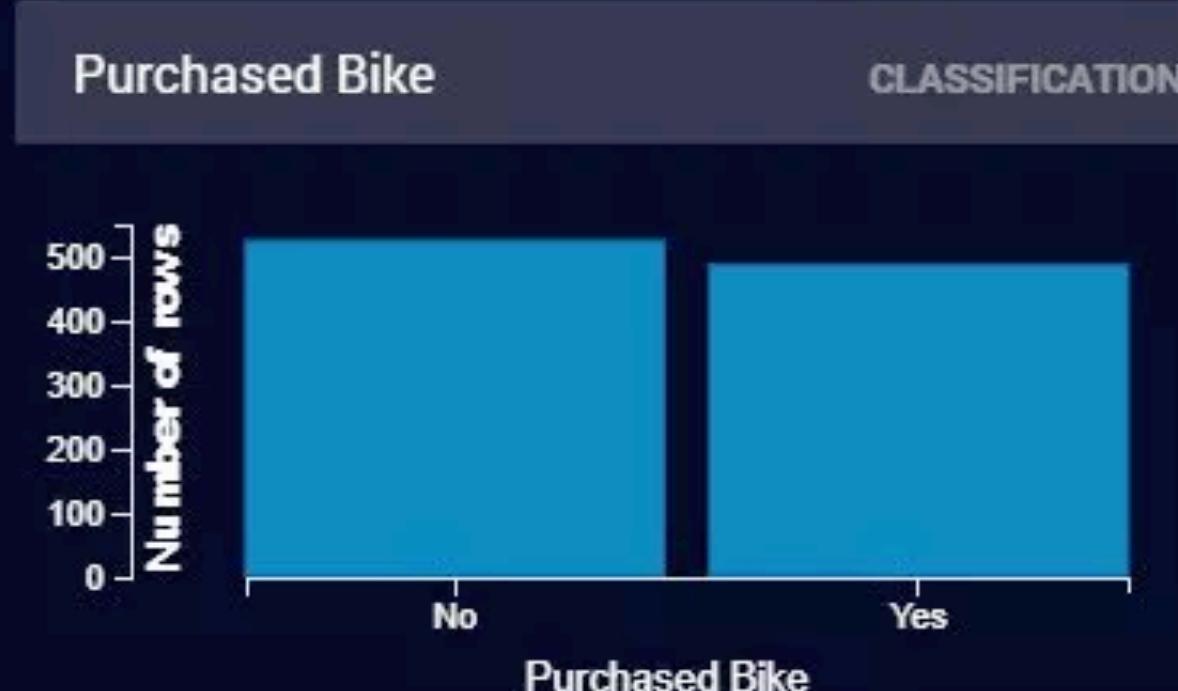
HDFS

Local File

We currently accept .csv, .tsv, .dsv, .xls, .xlsx, .sas7bdat, .bz2, .gz, .zip, .tar, .tgz

What would you like to predict?

Purchased Bike CLASSIFICATION



Purchased Bike	Number of rows
No	~450
Yes	~450

Start Modeling Mode: Autopilot ▾

Feature list: Informative Features
Optimization Metric: LogLoss

Show Advanced Options

Explore BikeBuyers.csv

- Workers: 00
1. Uploading Data (0.359 sec.)
 2. Reading raw data (Quick) (12.658 sec.)
 3. Exploratory Data Analysis : 13 / 13 Features (1.919 sec.)

Single Row Lookup

Column: H2O Frame Row # Value:

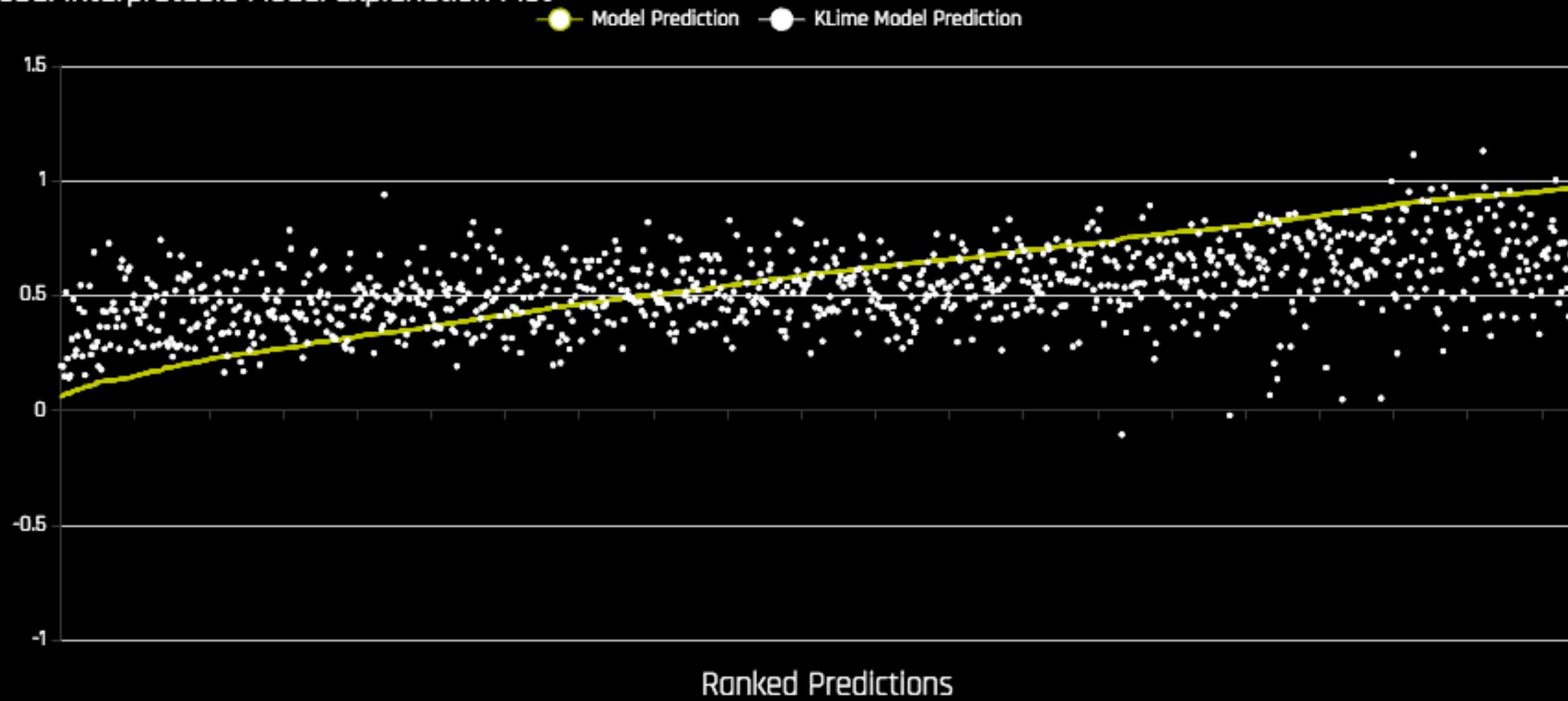
SEARCH

Plot: Global

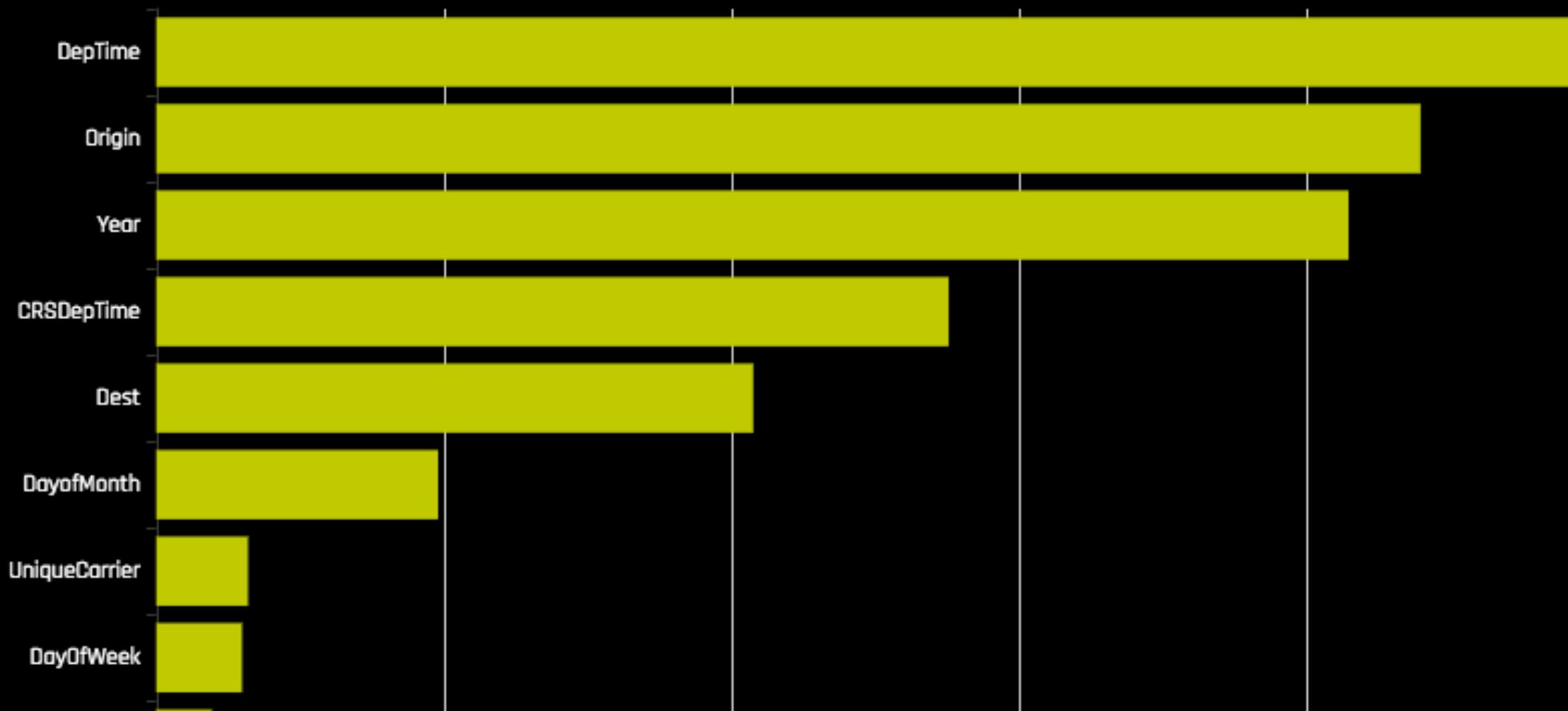
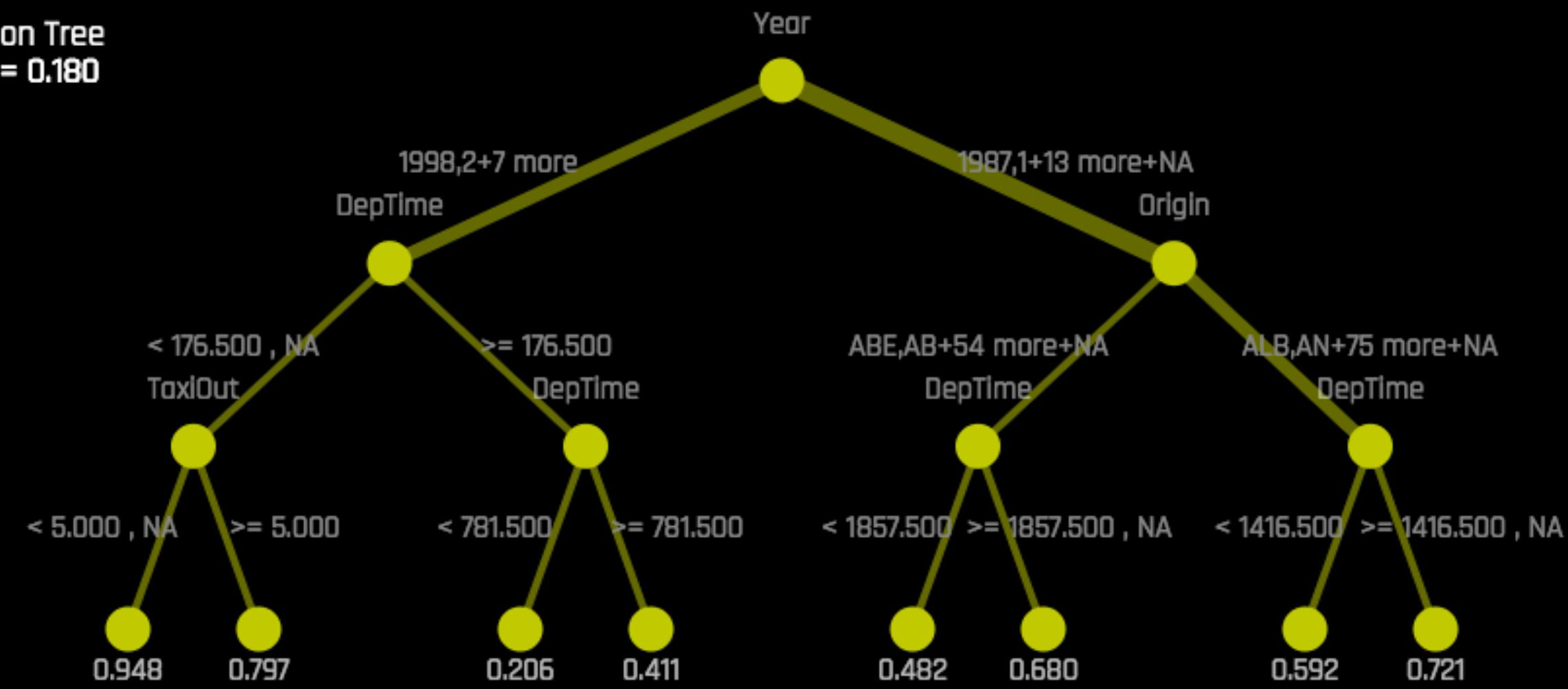
EXPERIMENT

EXPLANATIONS

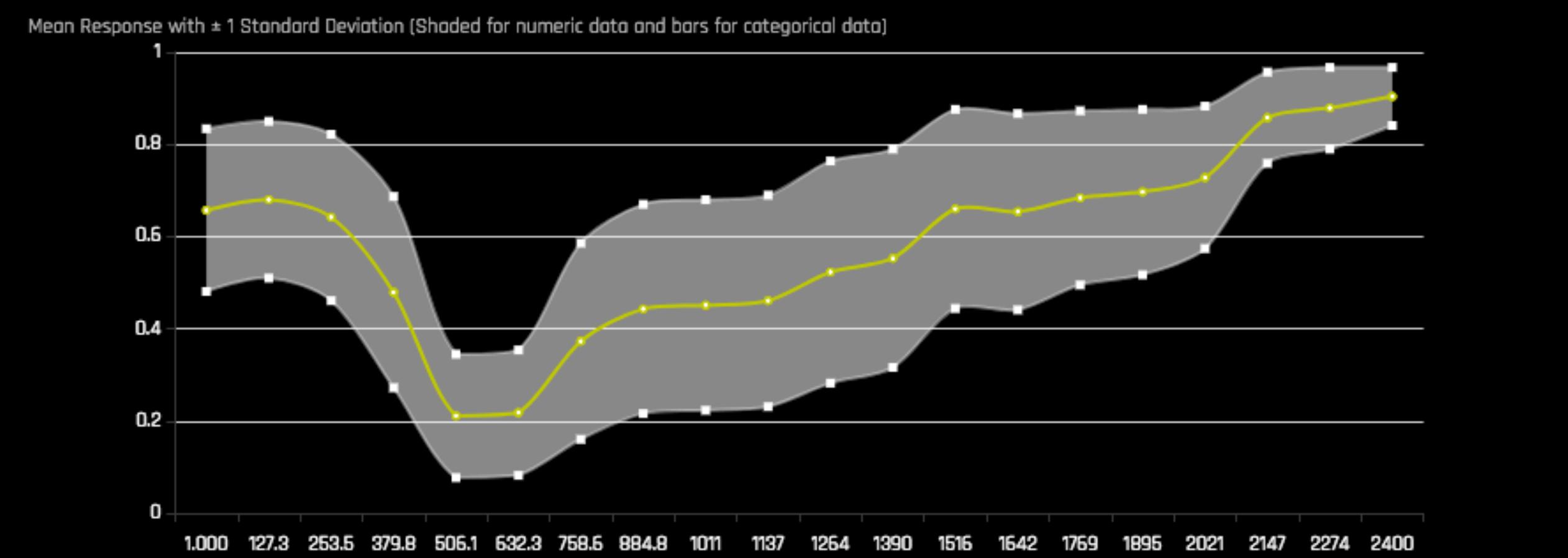
Global Interpretable Model Explanation Plot



Variable Importance

Decision Tree
RMSE= 0.180

Partial Dependence



SMART DATA DISCOVERY

“

**Google, what are
the top three
things I can do to
improve my close
rate today?**



A business user with no technical
knowledge

Smart Discovery Types

Logical Expressions

Search

Chatbots

What is it?

Using human readable interface for an SQL like interaction

The screenshot shows a user interface for defining a cohort of users based on their activity in Jira. The title is "Bento cohort". The interface includes a "Jira" logo and a "J" icon.

The main query is:

```
The Users who ..performed issue viewed (client)  
where attributes.designVersion = v2-bento  
with count >= 1 time  
any time during Last 30 days
```

Below the query, there is a "And also" section with four options:

- ..performed Event
- ..had been Active
- ..had been New
- ..had Property

Smart Discovery Types

Logical Expressions

Search

Chatbots

What is it?

Using NLP to construct the query and return results

The screenshot shows the Tableau TC18 - Ask Data interface. At the top, there's a navigation bar with icons for + a b | e a u, TC18 - Ask Data, Content, Users, Groups, Schedules, Tasks, Status, Settings, and a search icon.

The main area displays a data source named "wines" (DATA SOURCE · By Samantha Kwok · 0 views · 0 stars). Below the title, there are tabs for Ask Data, Connections (1), Connected Workbooks (0), and Details. The Ask Data tab is active, showing a text input field with the placeholder "Ask a question about wines". A user has typed "how many wineries are ther" into this field, and the system has suggested "distinct count of Winery, filter Designation to T".

On the left side, there are two sections: "Dimensions" and "Measures". The Dimensions section lists various categories like Country, County, Description, Designation, Province, Taster Name, Taster Twitter Handle, Title, Variety, Vintage, and Winery. The Measures section lists Number of Records, Points, and Price.

Below the input field, there are "Tips for success:" and a numbered list:

1. At left, hover over dimensions and measures to see data available to you.
2. Start with simple questions and then expand them.
3. In the drop-down list of suggestions, choose the best match.
(If you don't see a match, try a rephrased, simple question.)
4. Learn more in the Ask Data section of the [2019.1 Beta](#).

On the right side, there's a "Try asking these questions:" section with several examples:

- by County
- sort County in alphabetical order
- top County by sum of Number of Records
- sum of Points
- Points at least 80

Smart Discovery Types

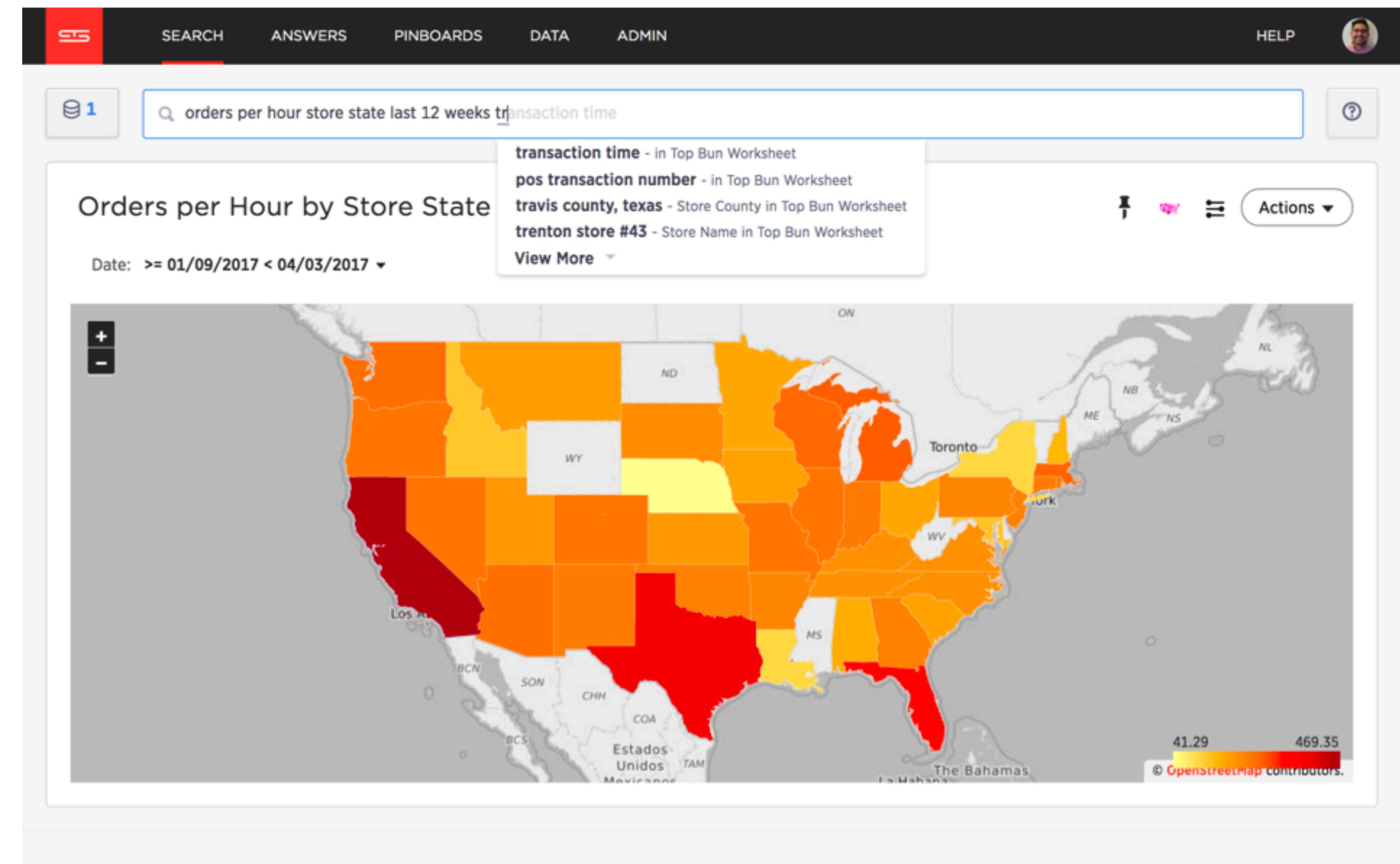
Logical Expressions

Search

Chatbots

What is it?

Using NLP to construct the query and return results



Smart Discovery Types

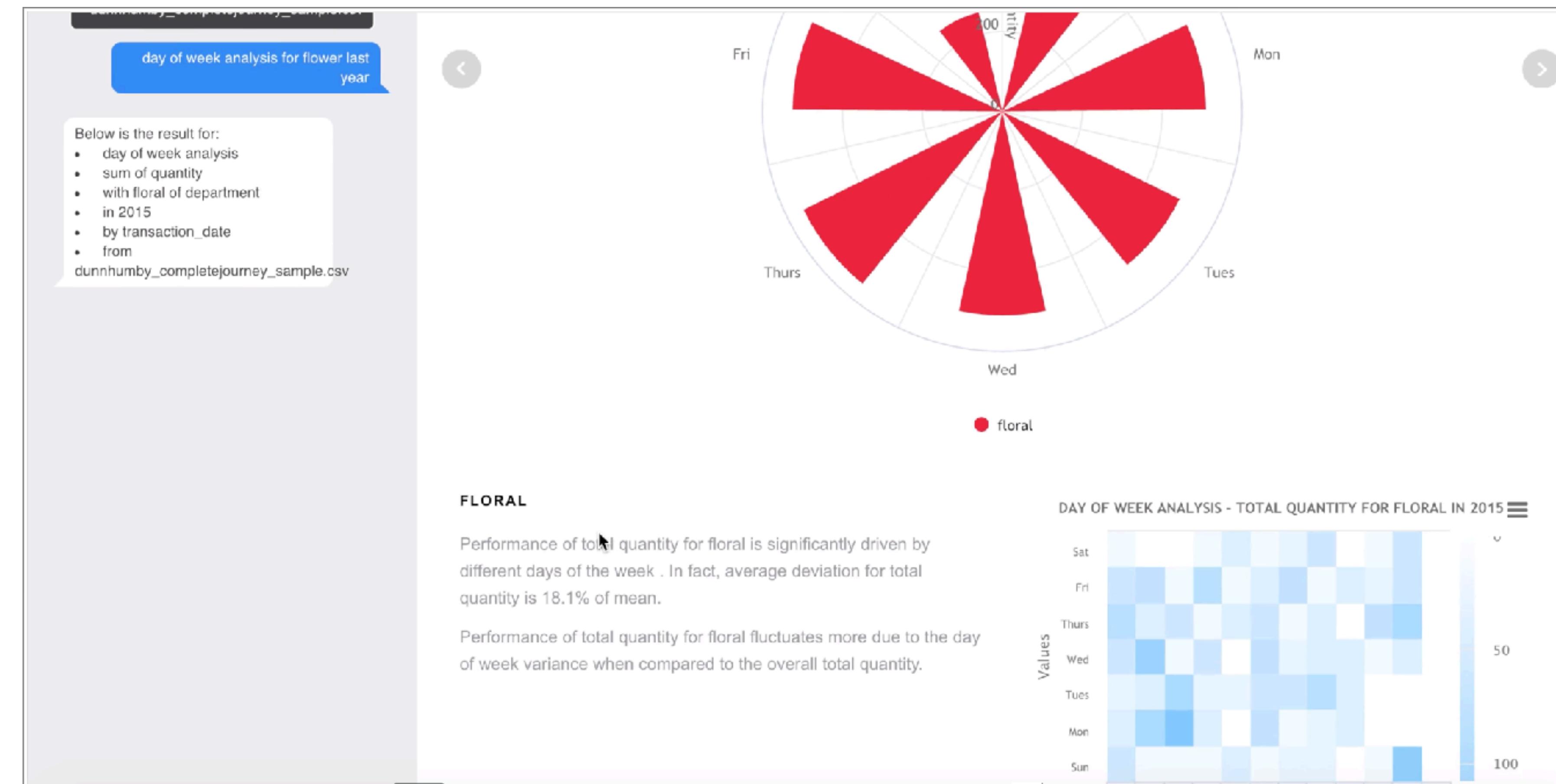
Logical Expressions

Search

Chatbots

What is it?

Human-like interaction to retrieve data



BIG DATA STEWARDSHIP





ML examples in stewardship solutions



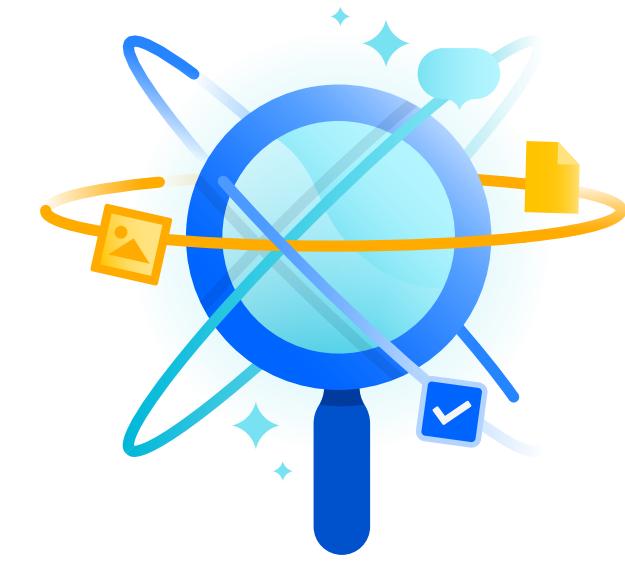
Trust

Suggest which data can be trusted



Consistency

Automatic metadata curation



Search

Find relations in the data

Search Alation

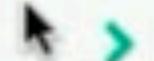


Data

[MySQL] Analytics



Alation DB



[Hive] Finance and Research ...



File Systems

Cataloging Flag Files Example



[S3] Data Lake FS: Bucket X



S3 RightNow Data



Queries

California Bank Customers

select * from bank.customer wh...
by Paul Walker · Endorsed 0 times

Articles

Applications

Building a Taxonomy

Business Areas

Conversations

what was our q2 revenue?

Join Vince Kuhn in this conversation

where is the medicaid table?

Business Intelligence

IBM Cognos Demo



MicroStrategy Demo



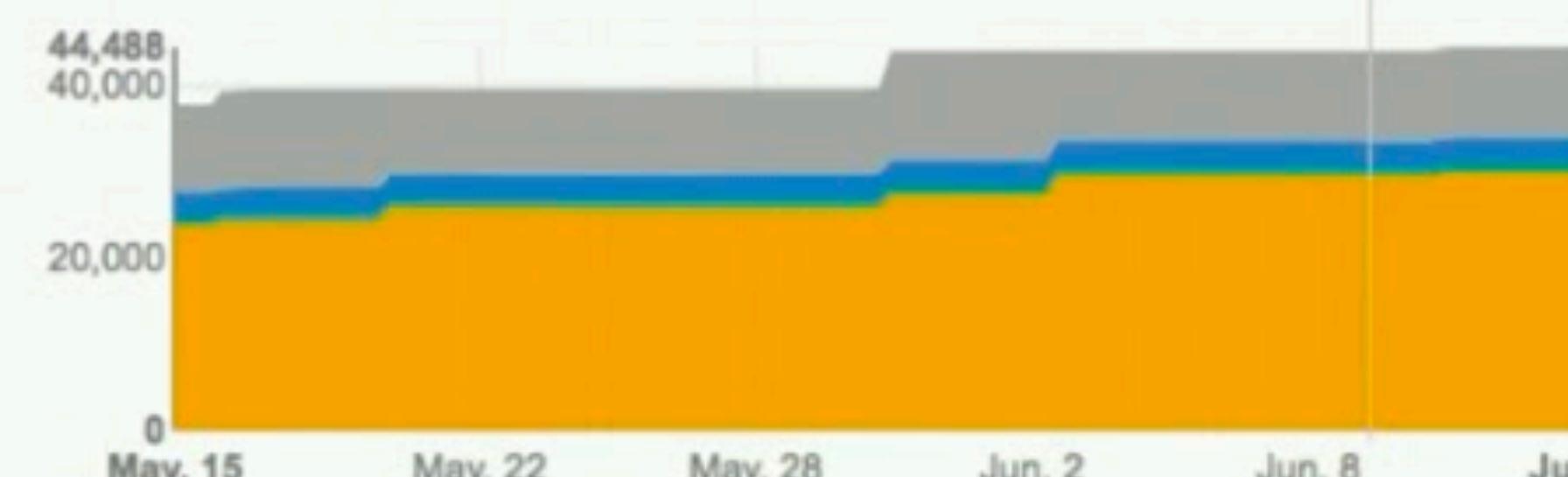
Microstrategy Old Data Model



Titles of data objects in All data sources

over the last 1 month

Today



Guessed and Unconfirmed	29964
Guessed and Confirmed	774
Hand-Populated	3255
Blank	10495
Total	44488

Getting Started in the Data Catalog

In order to get started with the data catalog, you can begin navigating the various pages.

You will find:

- Data sources like [MySQL] Healthcare
 - including summ_top_drg
- BI Servers like Tableau - Analytics Environment
- Glossaries like Business Glossary

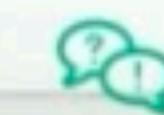
And more..

Quick Contacts / Links

In cases of data documentation questions, you can reach out to anyone on the Governance Team

To better delegate and motivate data curation and stewardship, see Alation Analytics

For definitions, see the Business Glossary



grid order Columns

	Name	Title	Type
1	id	ID	INT
2	ordr_tp	Order Type	INT
3	ordr_dt	Order Date	DATETIME
4	byr_id	Buyer ID	INT
5	item_id	Item ID	INT
6	qty	Quantity	
7	shp_dt	Ship Date	
8	dlvry_dt	Delivery Date	
9	shpg_addr_id	Shipping Address	
10	dscnt_cd	Discount Code	

Alation has guessed that shp_dt means "Ship Date". Is that right?

Confirm

Reject

Allie, the Alation robot, will learn from your response.

SMART DATA PREPARATION



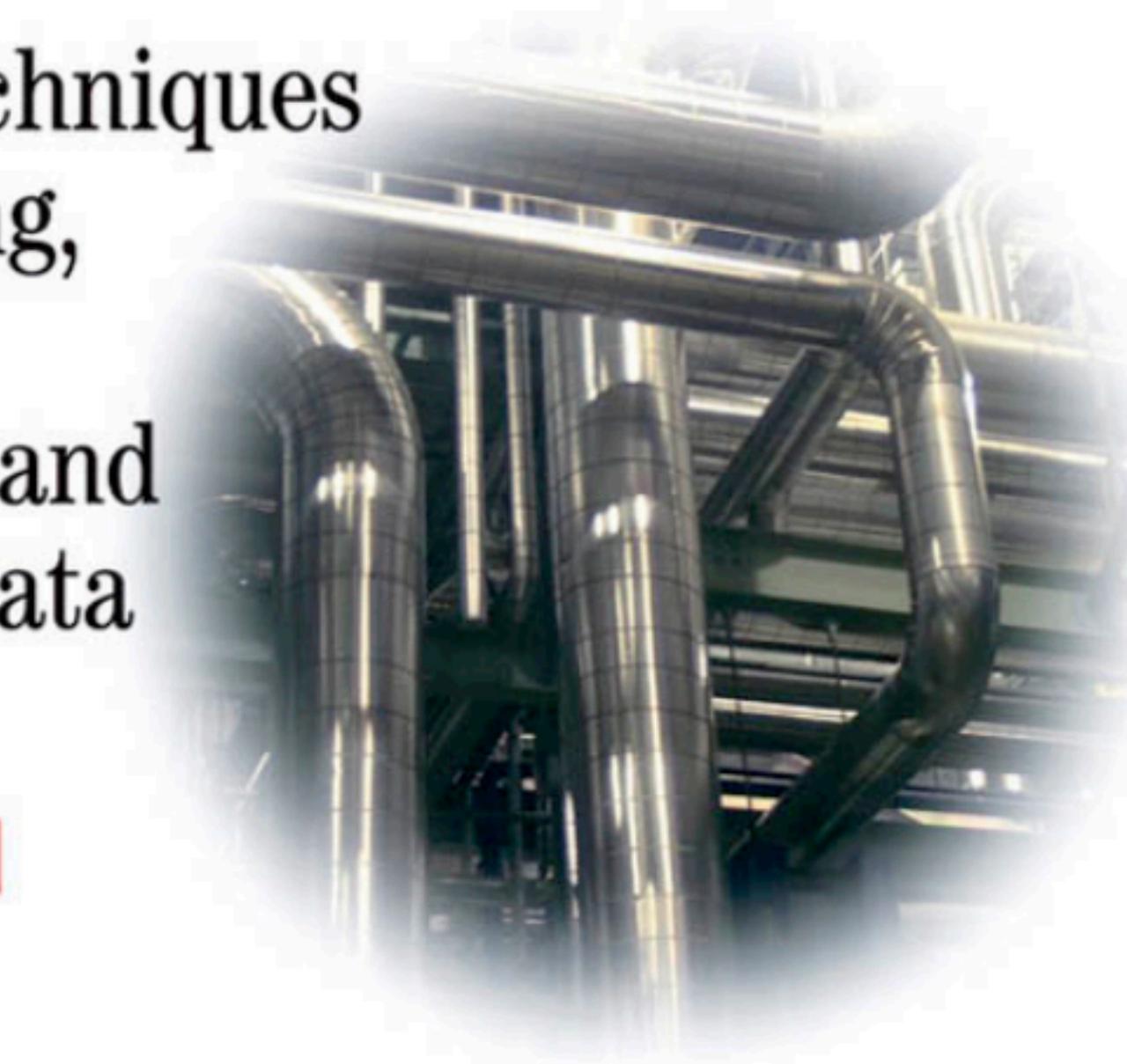
TIMELY. PRACTICAL. RELIABLE.

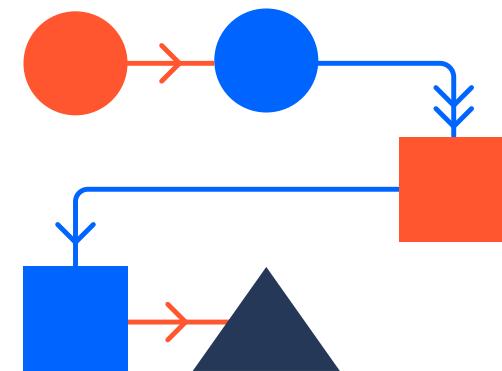
The Data Warehouse ETL Toolkit

Practical Techniques
for Extracting,
Cleaning,
Conforming, and
Delivering Data

Ralph Kimball

Joe Caserta





Drag & Drop ETL

Enable non technical users to prepare data

The screenshot shows a data preparation interface with a visual workflow builder and a detailed data preview section.

Workflow Overview:

```

graph LR
    O1[Orders (East)] --> A1[All Orders]
    O2[Orders (West)] --> A1
    O3[Orders (Central)] --> A1
    O4[Orders (South)] --> A1
    A1[All Orders] --> OR[Orders + Returns]
    A1 --> SC[Split Customer]
    A1 --> CAA[Create 'All Orders' ...]
    OR[Orders + Returns] --> A2[Aggregate]
    SC[Split Customer] --> A2
    A2[Aggregate] --> Q1[Pivot Quotes]
    Q1[Pivot Quotes] --> Q2[Quota]
    Q2[Quota] --> DO[Quota + Orders]
    DO[Quota + Orders] --> CA[Create 'Annual ...']
  
```

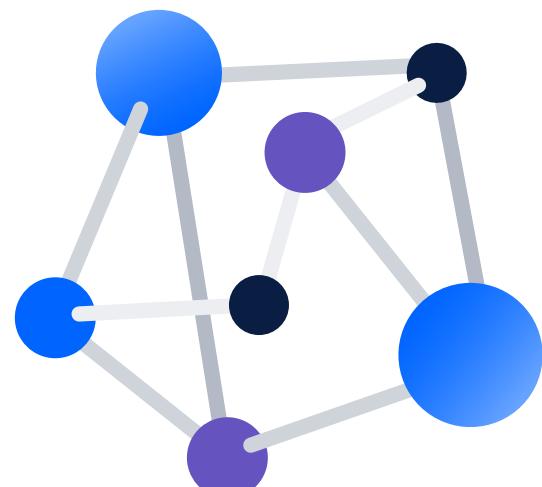
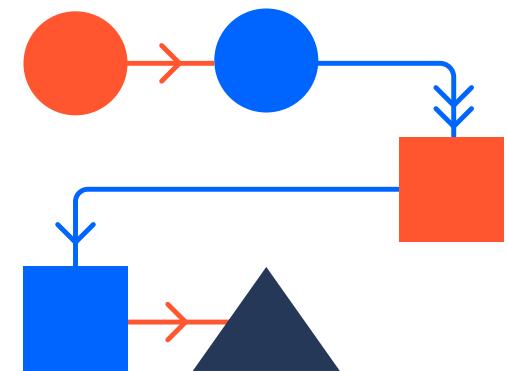
Data Preview Section:

The preview area displays several data tables and charts:

- Changes (2):** A table showing state changes with a bar chart of Row ID counts.
- Order ID 922:** A table showing order details with a dropdown menu highlighting the Segment column.
- Customer ID 512:** A table showing customer names and their corresponding customer IDs.
- Ship Mode 4:** A table showing ship mode categories.
- Order Date 604:** A table showing order dates with a histogram.

Table Data Preview:

Sales	Quantity	Profit	Discount	Region	State	Row ID	Order ID	Segment	Customer ID	Customer Name	Ship Mode	Order Date	Ship Date
18.648	7	-12.432	0.7	South	North Carolina	231	US-2015-156216	Corporate	EA-14035	Erin Ashbrook	Standard Class	09/13/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
178.384	2	22.298	0.2	South	Florida	315	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
15.552	3	5.4432	0.2	South	Florida	316	CA-2015-167850	Corporate	AG-10525	Andy Gerbode	Standard Class	08/09/2015, 12:00:00 AM	08/16/2015, 12:00:00 AM
39.072	6	9.768	0.2	South	North Carolina	404	CA-2015-155208	Corporate	SP-20650	Stephanie Phelps	Standard Class	04/16/2015, 12:00:00 AM	04/20/2015, 12:00:00 AM
10.368	2	3.6288	0.2	South	North Carolina	705	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
166.84	5	18.7695	0.2	South	North Carolina	706	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
15.216	1	2.2824	0.2	South	North Carolina	707	CA-2015-138527	Corporate	BN-11470	Brad Norvell	Standard Class	09/12/2015, 12:00:00 AM	09/17/2015, 12:00:00 AM
11.36	2	5.3392	0	South	Louisiana	764	CA-2015-162775	Corporate	CS-12250	Chris Selesnick	Second Class	01/13/2015, 12:00:00 AM	01/15/2015, 12:00:00 AM



Data Blends

Connecting different datasets automatically

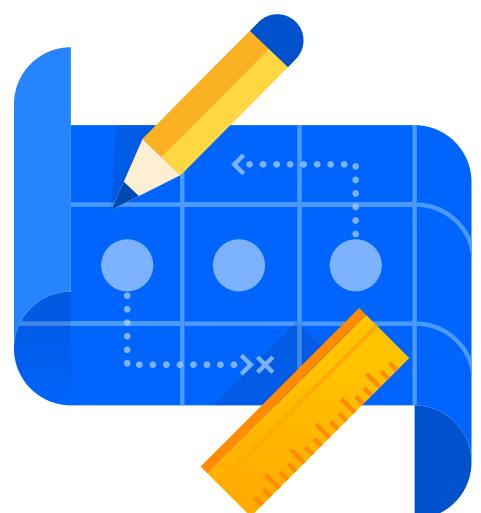
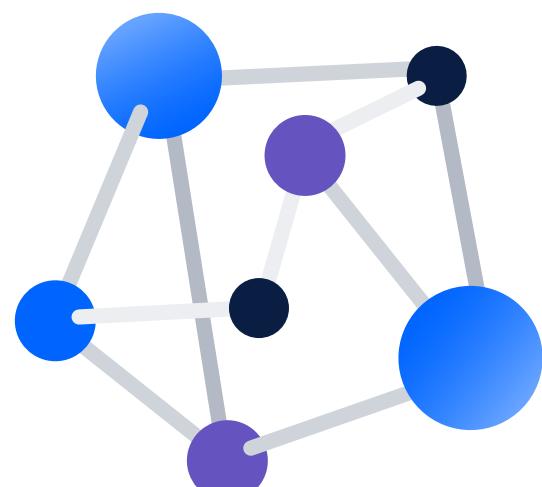
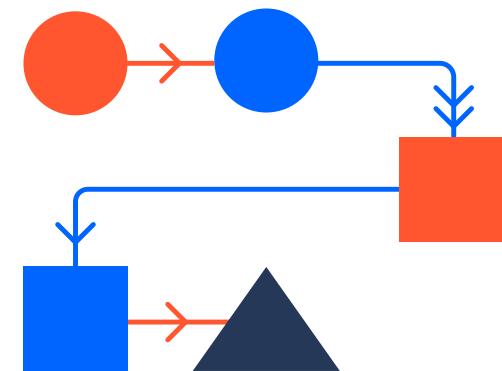
The screenshot shows a data blending interface with the following components:

- Left Panel (Tools):** A sidebar with various tools: steps, versions, highlight, attach, columns, compute, remove, sampling, shape, 123, auto #, and preview.
- Top Bar:** Shows "Steps" and buttons for "Save" and "Cancel".
- Current Step:** "Lookup from SFDC Leads" (highlighted in pink).
- Next Step:** "Append with Web Campaigns (FY15 Q4)" (highlighted in orange).
- Start Step:** "Start with 2015 Web Campaigns".
- Right Panel (Lookup using):**
 - Header:** "Lookup using Full Sampled SFDC Leads".
 - Text:** "For each row in the current dataset, find the first match from SFDC Leads with automatic match. (change options)".
 - Section Connections:**
 - Current:** Shows "First" and "Last" with a 91% match rate.
 - SFDC Leads:** Shows "Full Name" with a 91% match rate.
 - Suggested Connections:**

	Match Rate	Fields
Email	92%	E-mail
Last, First	91%	Full Name
Direct Number, Last, First	53%	Phone, Full Name
Direct Number, Company	33%	Phone, Company Name
Position, Direct Number	19%	Title, Phone
- Bottom Panel (Table):**

Position	LOOKUP (LEFT)	LOOKUP (LEFT)	LOOKUP (RIGHT)	Full Name	E-mail
Supervisor, Business Applications	Heinaman	James	Heinaman, James	Heinaman, James	jheinaman@fanniemae.c
	Muppipi	Brent	Muppipi, Brent	Muppipi, Brent	bmuppipi@markelcorp.cc
	Davidson	David	Davidson, David	Davidson, David	ddavidson@centerplate.c
Systems Analyst	Taylor	Mark	Taylor, Mark	Taylor, Mark	mtaylor@premera.com
	McClellen	Dan	McClellen, Dan	McClellen, Dan	dmcclellen@weyerhaeus
Desktop Support Technician	Cotty	Karen	Cotty, Karen	Cotty, Karen	kcott@canadiantire.ca
	Brotsky	Joshua	Brotsky, Joshua	Brotsky, Joshua	jbrotsky@kingsfamily.cc
	Hanley	Karsten	Hanley, Karsten	Hanley, Karsten	khanley@testamericainc.
Analyst, Enterprise Application En...	Lecosia	Gary	Lecosia, Gary	Lecosia, Gary	glecosia@mcdonalds.con

Bottom Right: Source: Paxata



Data cleansing

ML for manipulating the data

Group and Replace Done

County 68

Stop Date 1H

01/01/2012

01/01/2017

County Stop Date Violation Stop Outcome ID

County	Stop Date	Violation	Stop Outcome	ID
Barron	04/16/2012	Other (non-mapped)	Arrest	WI-2012-060620
Barron	07/11/2012	Other (non-mapped)	Arrest	WI-2012-114876

County 68 Group Replace Done

Barron 3 members

- Baron
- Barron
- Barronn
- Adam
- Adams
- Ashland
- Bayfield
- Brown
- Buffalo
- Burnett
- Calumet
- Chippewa
- Clark
- Columbia
- Crawford

Baron
91 rows
91 (100%) highlighted

THE PRIVACY WAVE

Bigger Responsibility, Bigger Repercussions

Fines of up to 4% of turnover
Organizations in breach of GDPR can be fined up to 4% of annual global turnover or €20 Million.



Increased territorial scope
Applies to any company processing personal data of EU citizens, regardless of location.



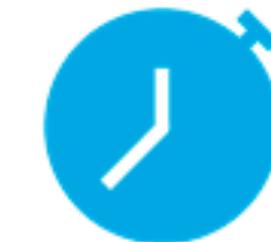
Consent matters
Explicit consent must be provided in an intelligible and easily accessible form.



Right to access and portability
Users can inquire whether and how their personal data is being processed.



Breach notification within 72 hrs
Breaches must be reported within 72 hours of first having become aware of the breach.



Privacy by design
Data protection from the onset of the designing of systems, rather than a retrospective addition.

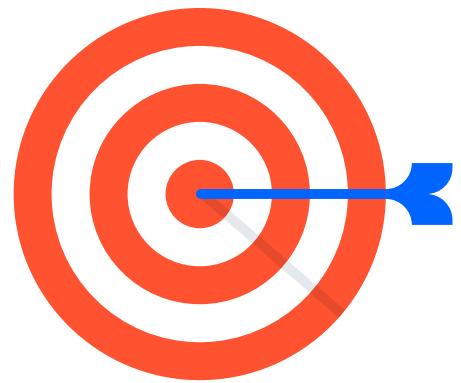


Right to be forgotten
Entitles the data subject to have the data controller erase his/ her personal data (and potentially third parties, too).



Mandatory data protection officers
Appointed in certain cases, to facilitate the company's need to demonstrate GDPR compliance.





Direct Marketing

can only be used when the intended recipient has given consent



Electronic Communication

Privacy of individuals as it relates to the confidentiality of electronic communications



CCPA

Passed as a law on June-2018. Goes into effect in January 2020



Consent Act:

The US version of GDPR (not a law yet)

Summary

Summary



Vision

- Growth
- Simplicity
- Governance
- Democratisation



Challenges

- Data growth vs. Analysts growth
- IT budgets
- The rise of the unknowns
- Complexity vs. Business value



Trends

- Realtime Predictive Analytics
- Auto ML
- Smart Data Discovery
- Big Data Stewardship
- Smart Data Preparation
- The Privacy Wave



Thank you