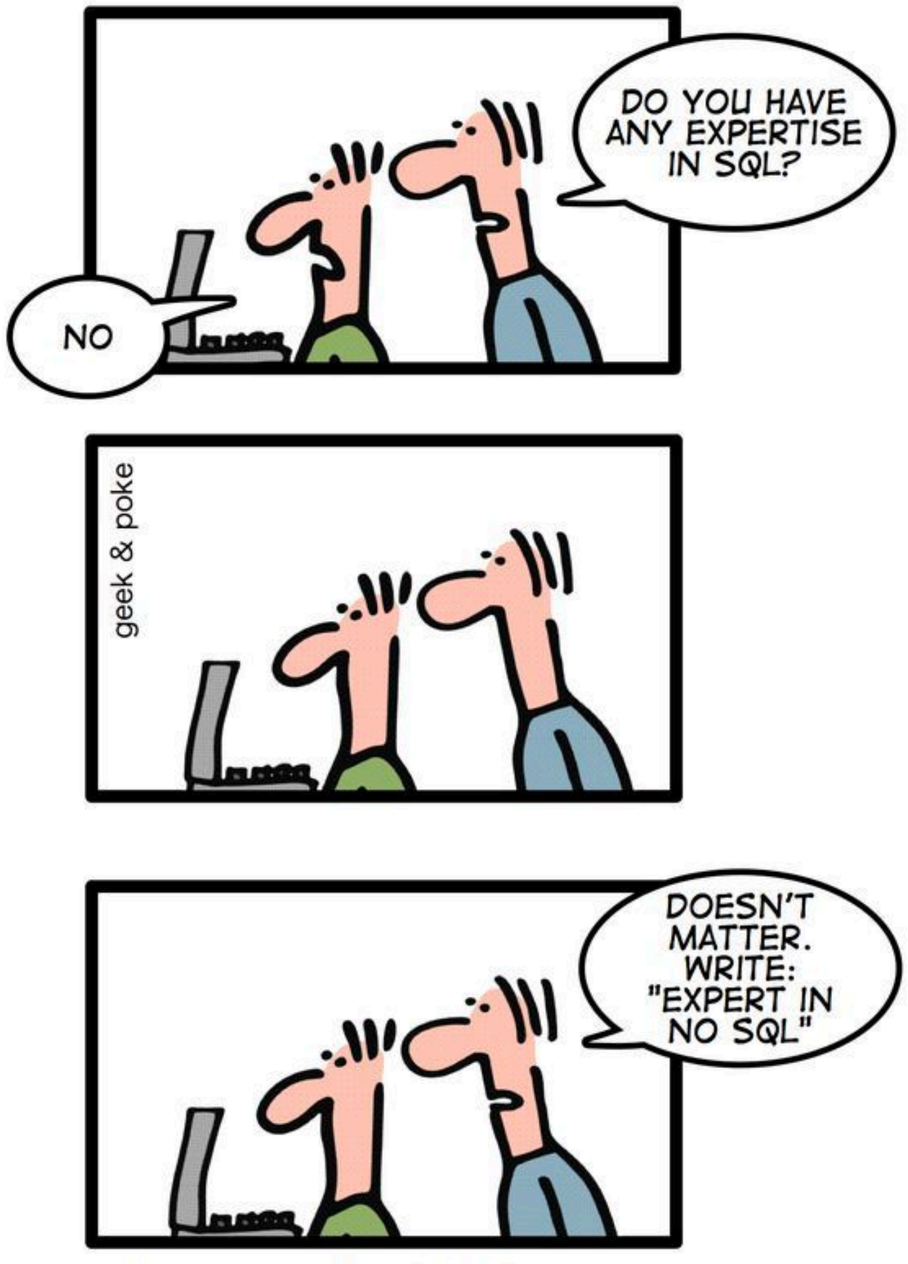
Data Engineering

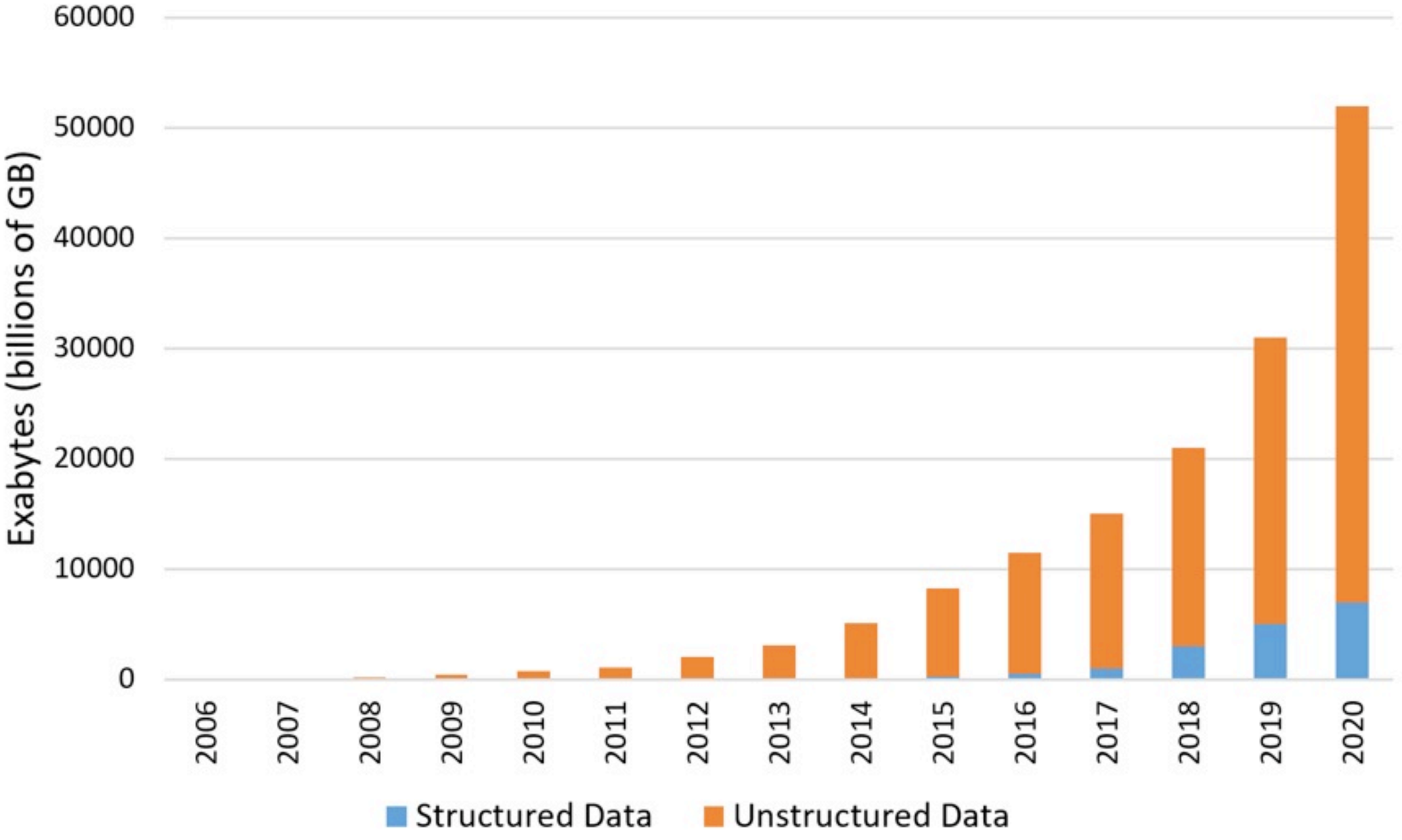
2019 and Beyond

ITZIK FELDMAN | DE PRODUCT MANAGER | @ITZIKFELDMAN

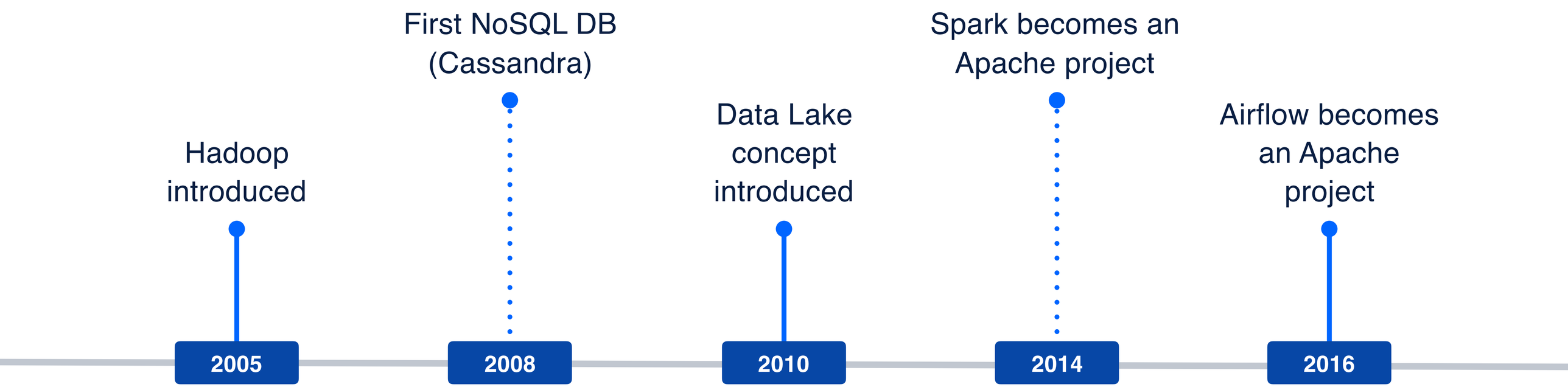
HOW TO WRITE A CV



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

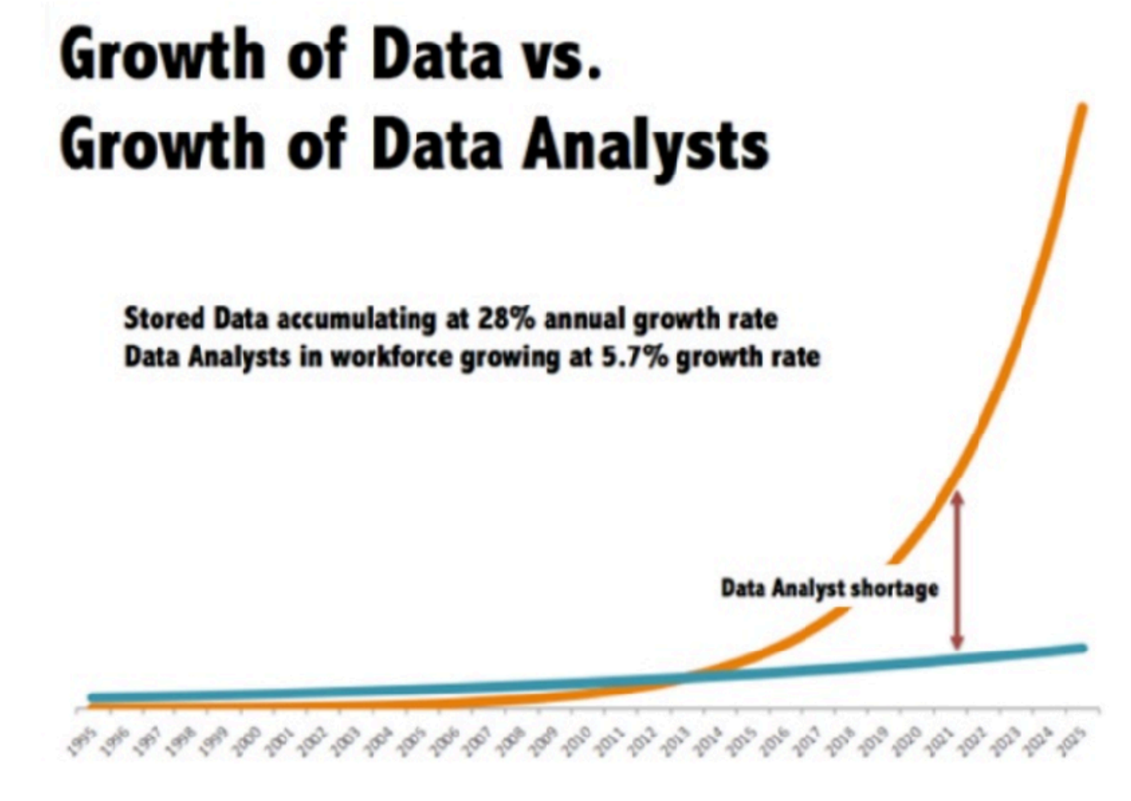
Governance

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

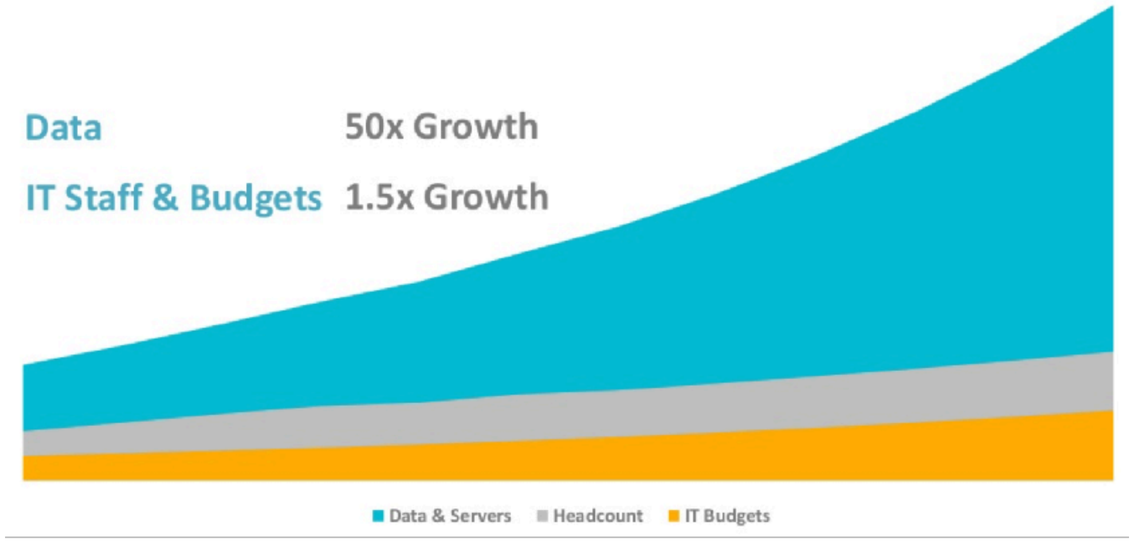
Democratisation

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

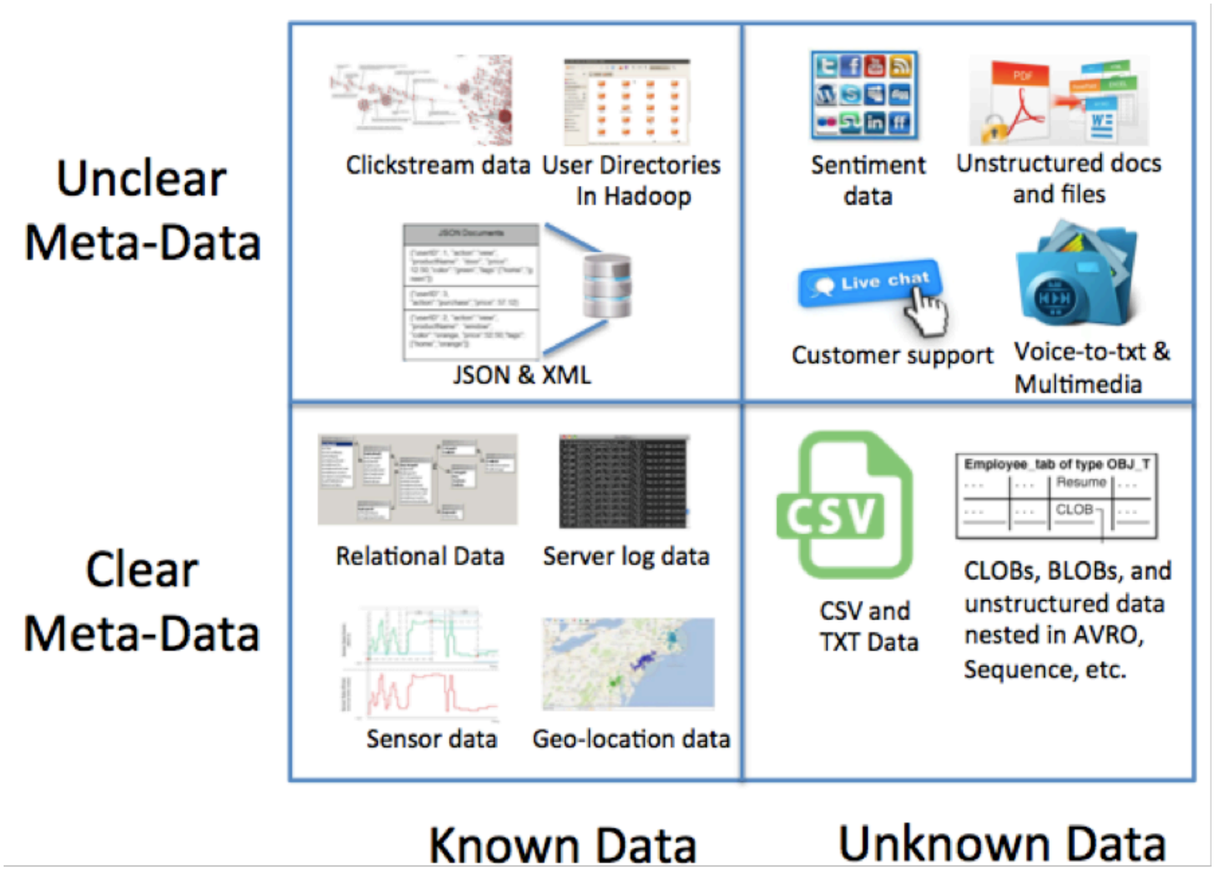
Challenges



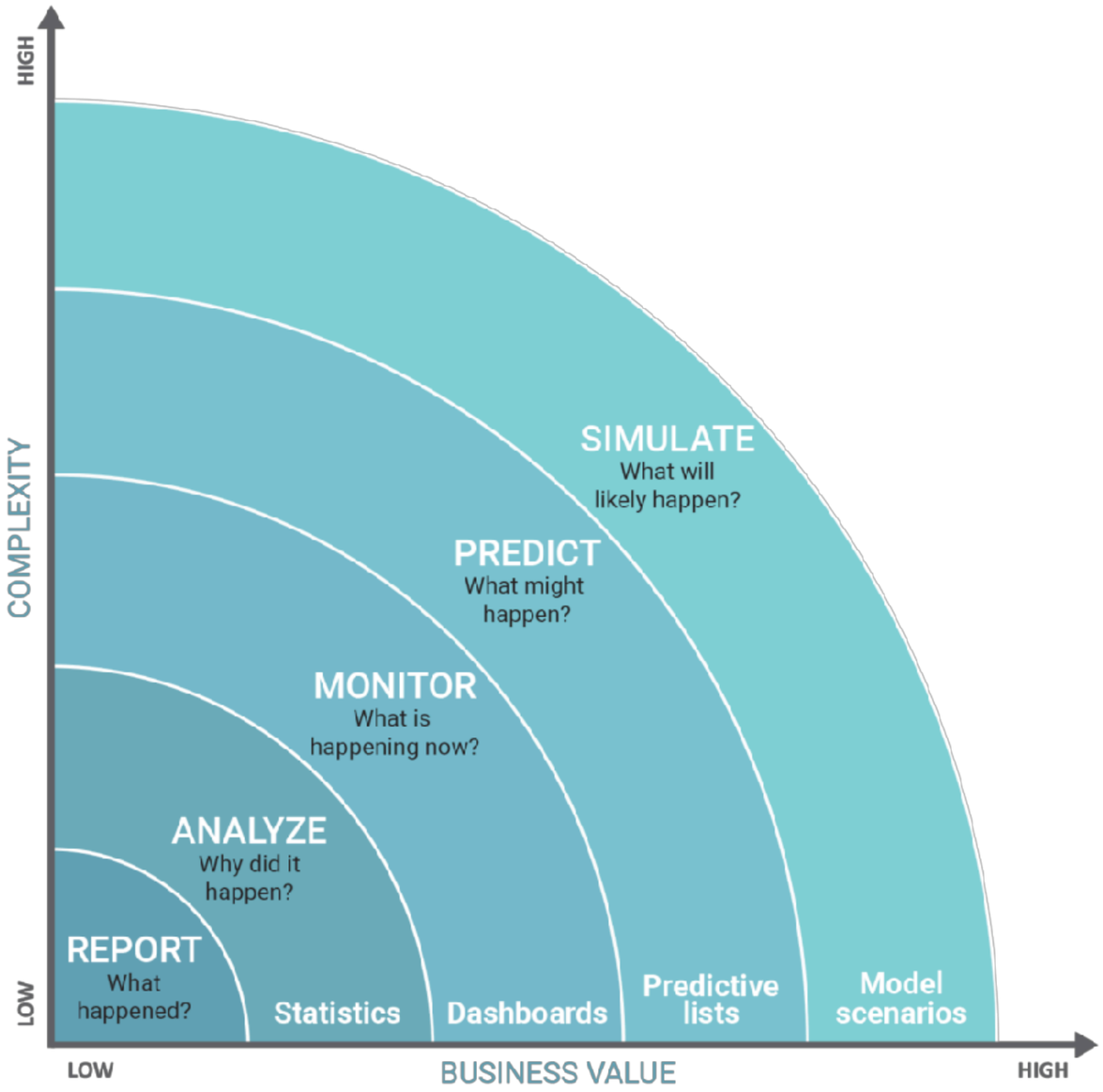
We don’t have enough analysts to handle the data



IT budgets are not growing as fast as data



The rise of the unknown



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

But there are wins

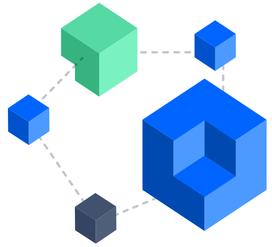
Other Wins

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

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

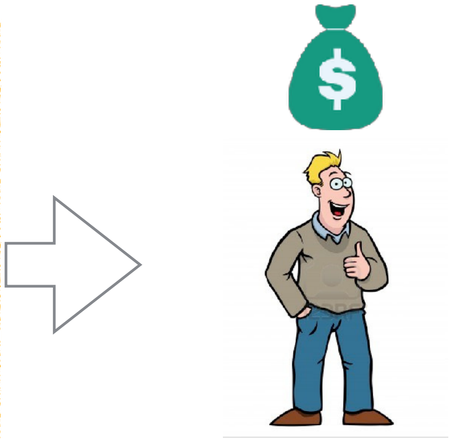
If (user selected JSW/JSD bundle)

AND (user selected confluence) within 1 week

AND (channel = atlassian.com)

Then True

Else False

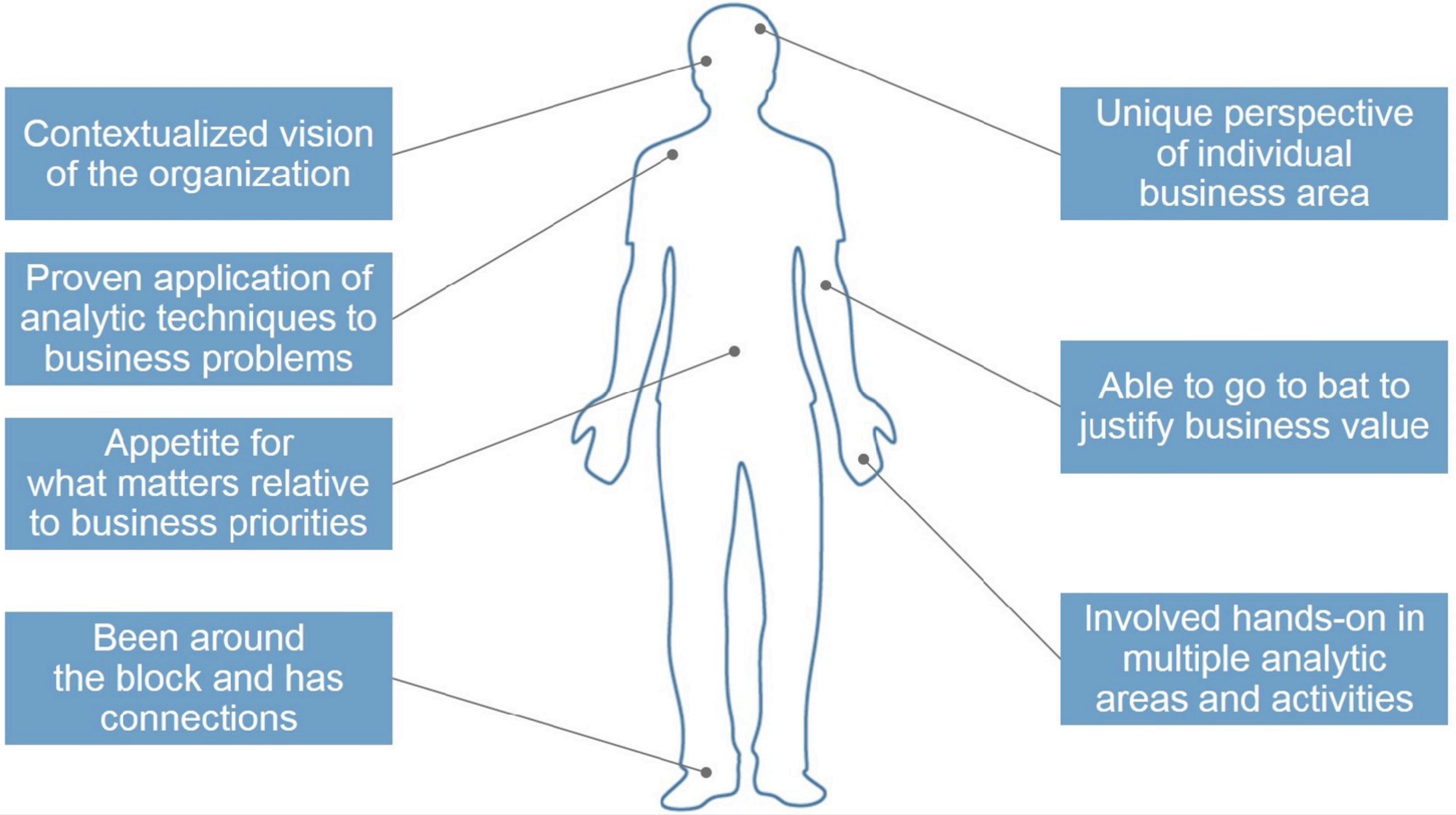


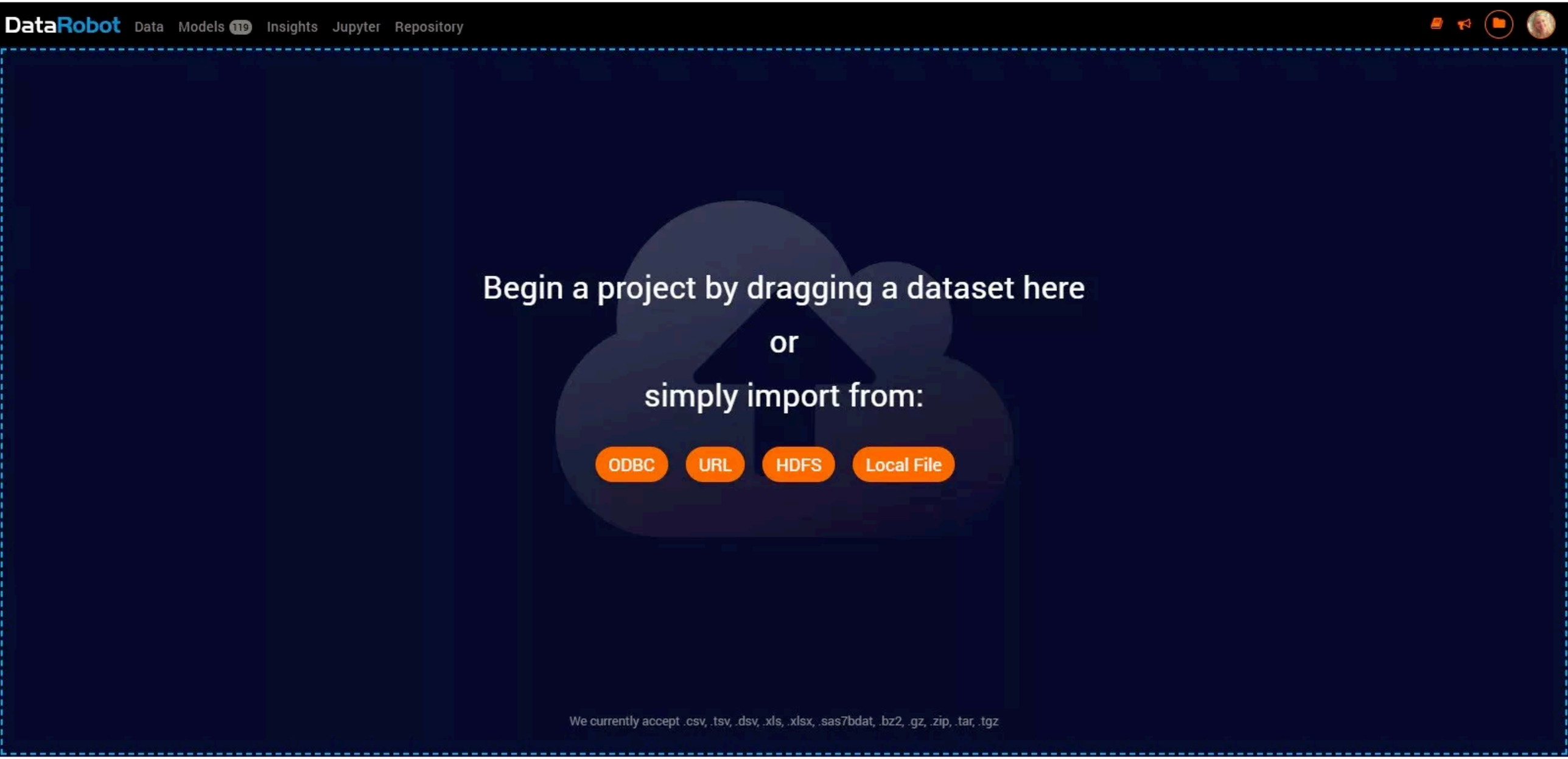
1/1/2019 - Trialed JSW/JSD

from atlassian.com

3/1/2019 - Trialed Confluence

Traits of a Citizen Data Scientist




DataRobot
Data
Models 119
Insights
Jupyter
Repository
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

DataRobot Data Models Insights Jupyter Repository

BikeBuyers.csv

What would you like to predict?

Purchased Bike CLASSIFICATION

Start

Modeling Mode: Autopilot

Feature list: Informative Features

Optimization Metric: LogLoss

Time-Aware Modeling

Time-aware modeling is not available for this project because DataRobot did not detect any time features in the dataset.

Set up time-aware modeling

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.)

Show Advanced Options

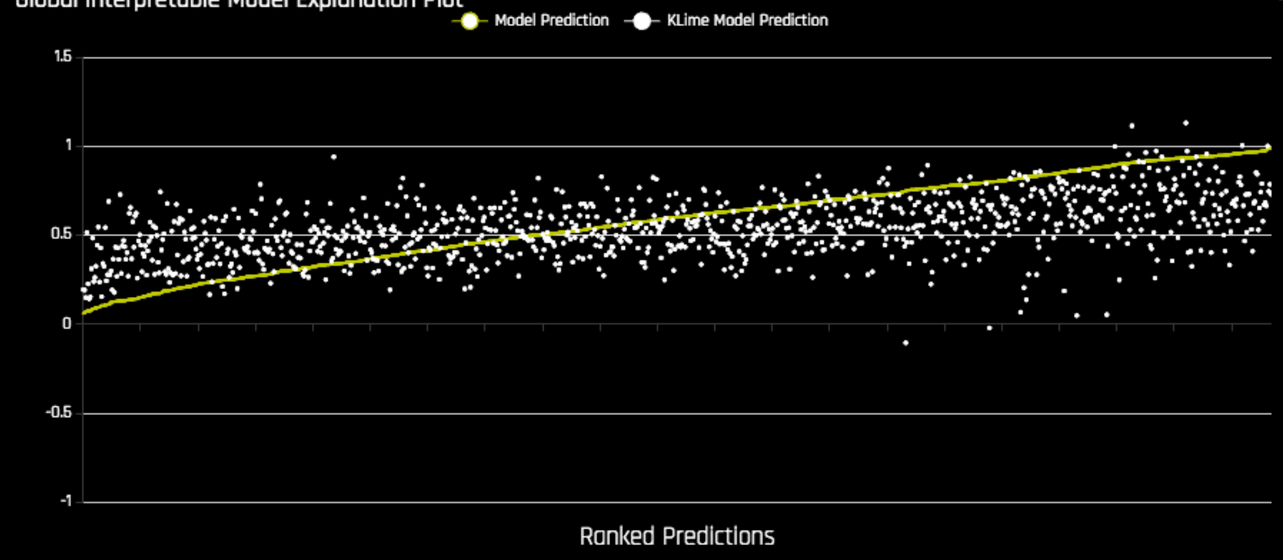
Explore BikeBuyers.csv

Single Row Lookup

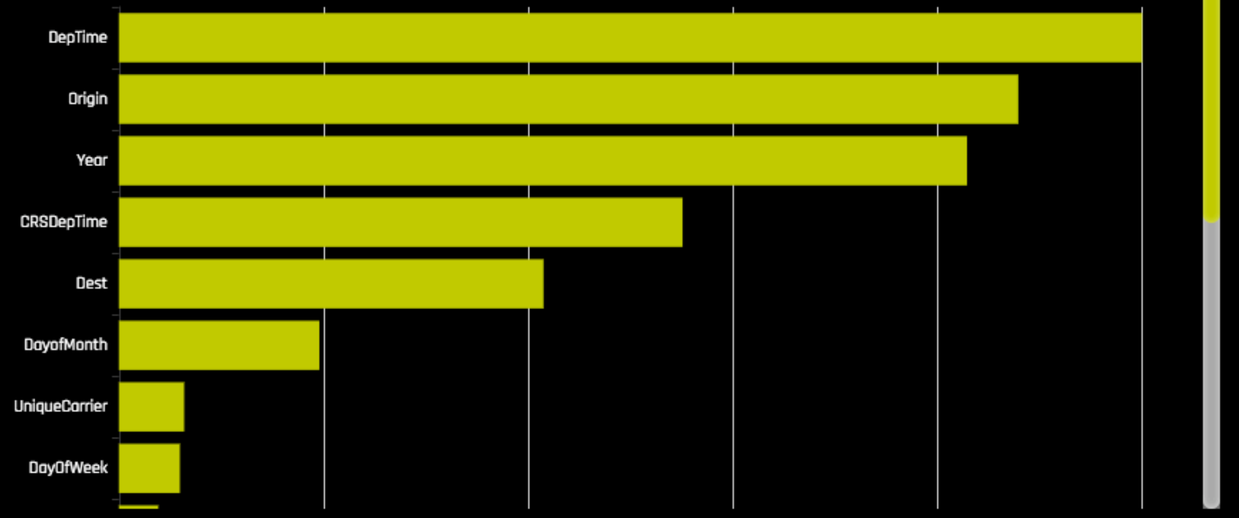
Column: H2O Frame Row # Value: SEARCH Plot: Global

EXPERIMENT EXPLANATIONS

Global Interpretable Model Explanation Plot

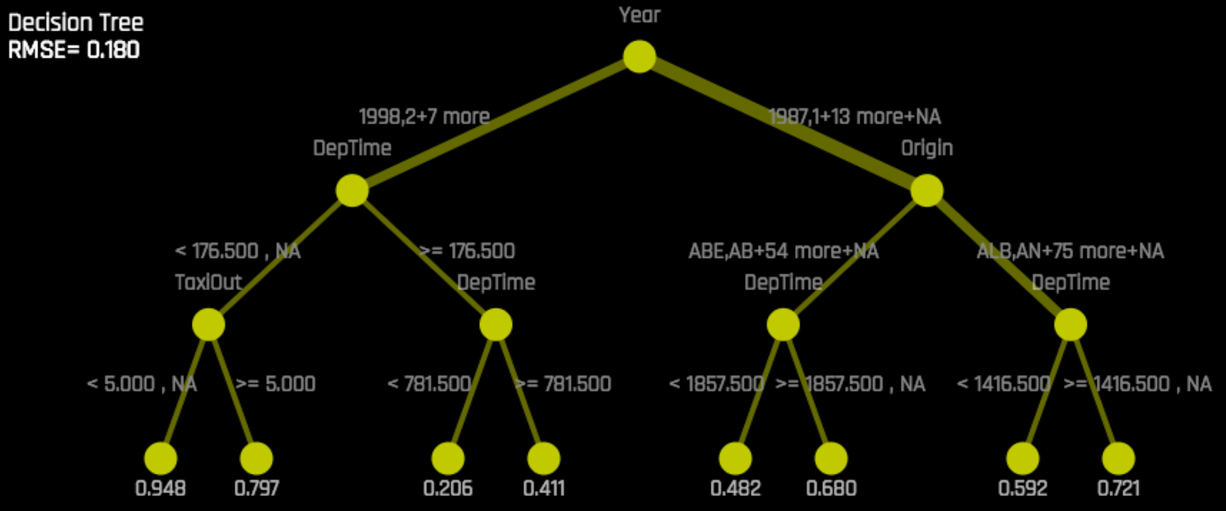


Variable Importance



Decision Tree

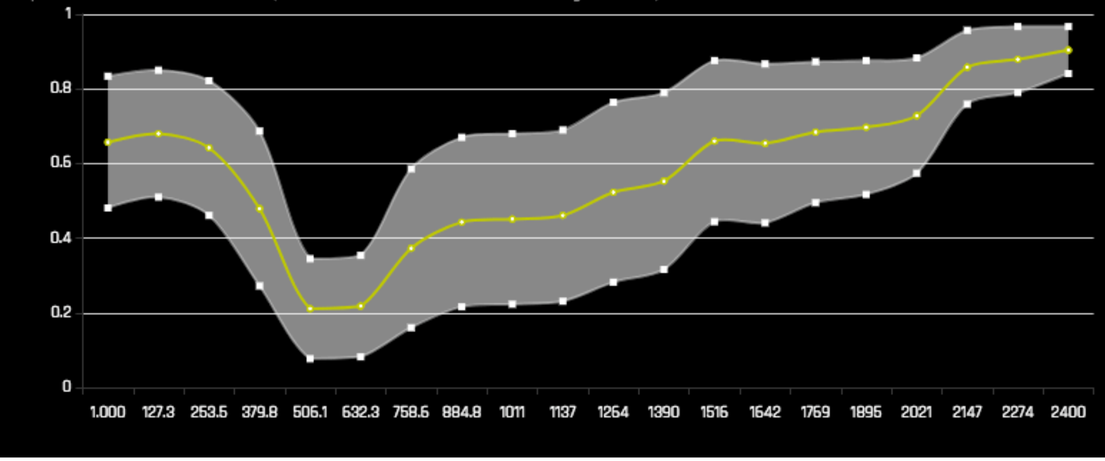
RMSE= 0.180



Partial Dependence

DepTime

Mean Response with ± 1 Standard Deviation (Shaded for numeric data and bars for categorical data)



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

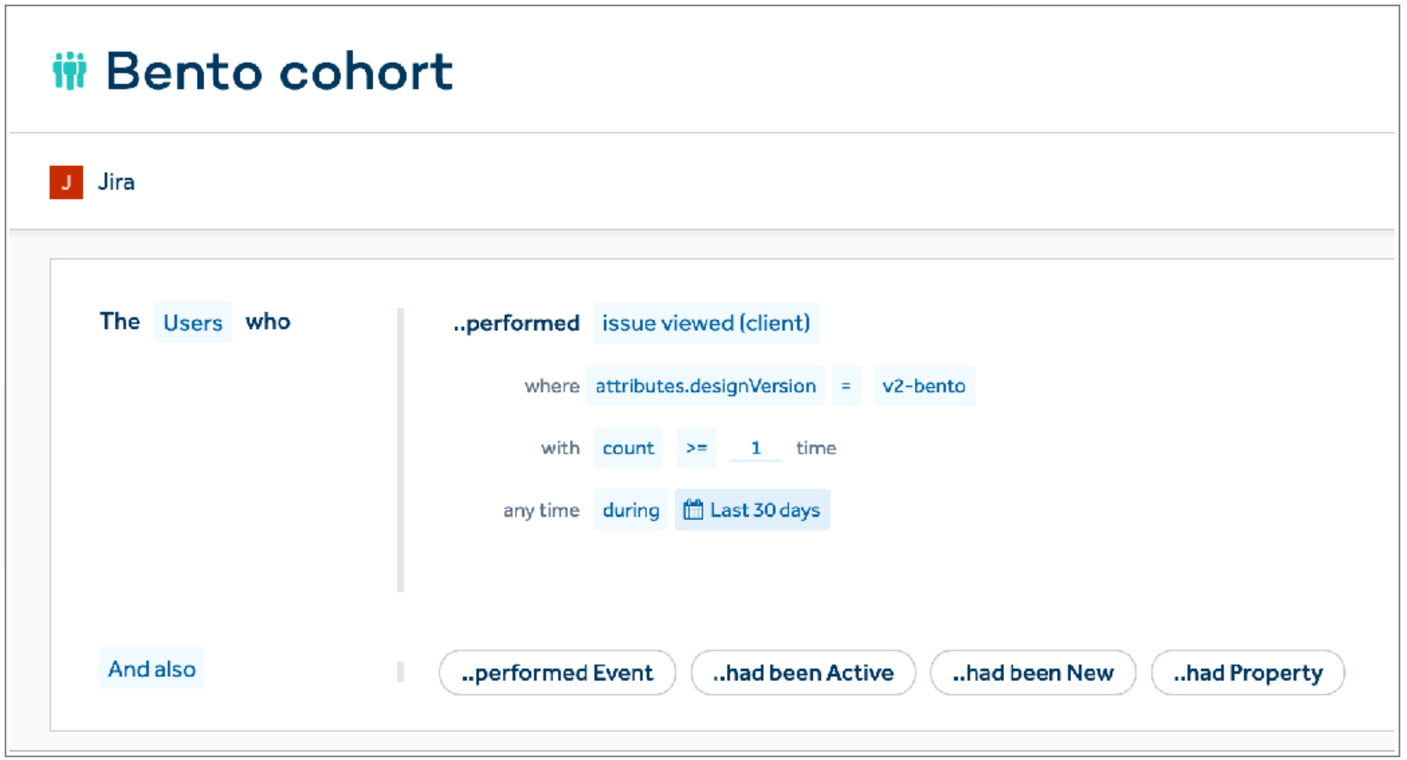
Logical Expressions

Search

Chatbots

What is it?

Using human readable interface for an SQL like interaction



Smart Discovery Types

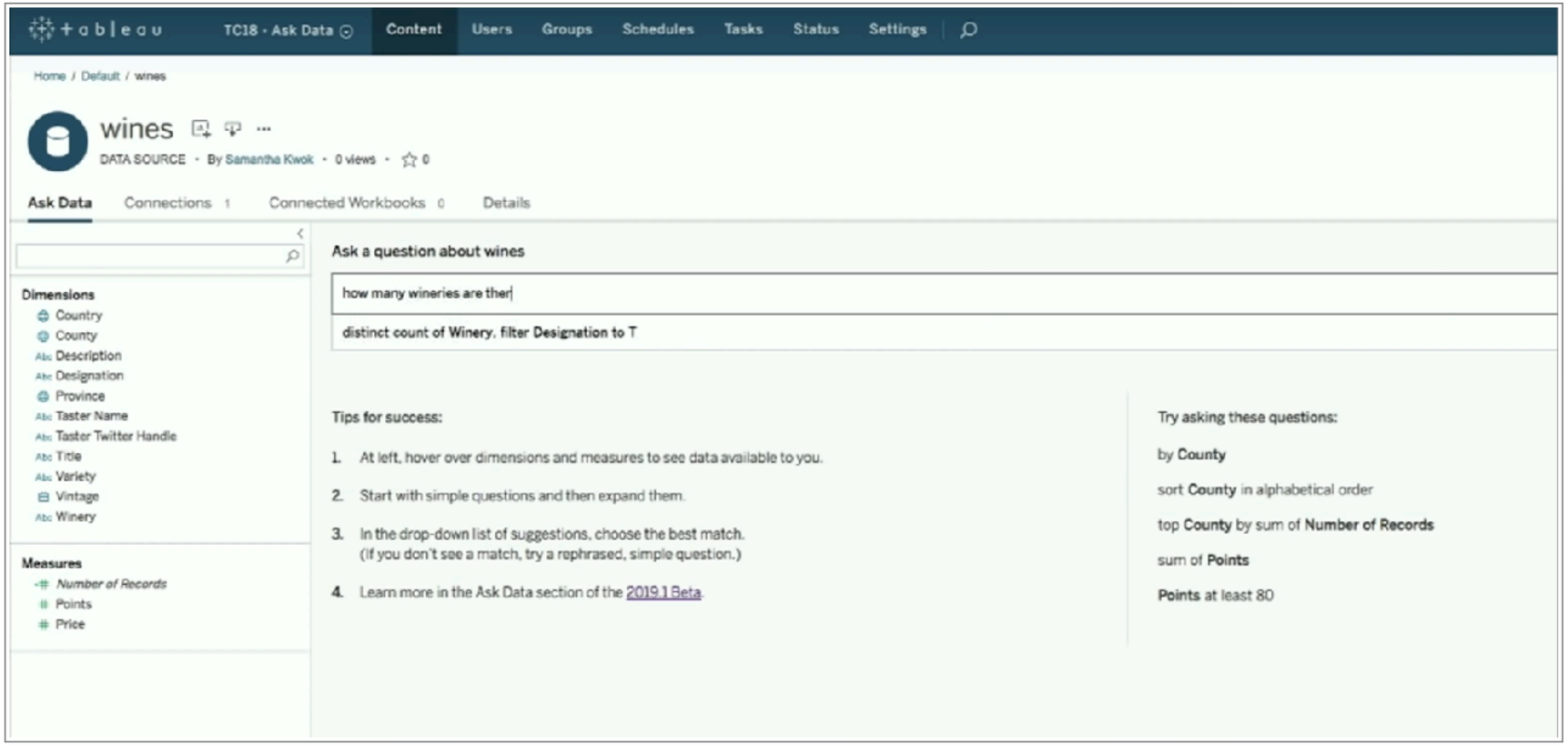
Logical Expressions

Search

Chatbots

What is it?

Using NLP to construct the query and return results



Smart Discovery Types

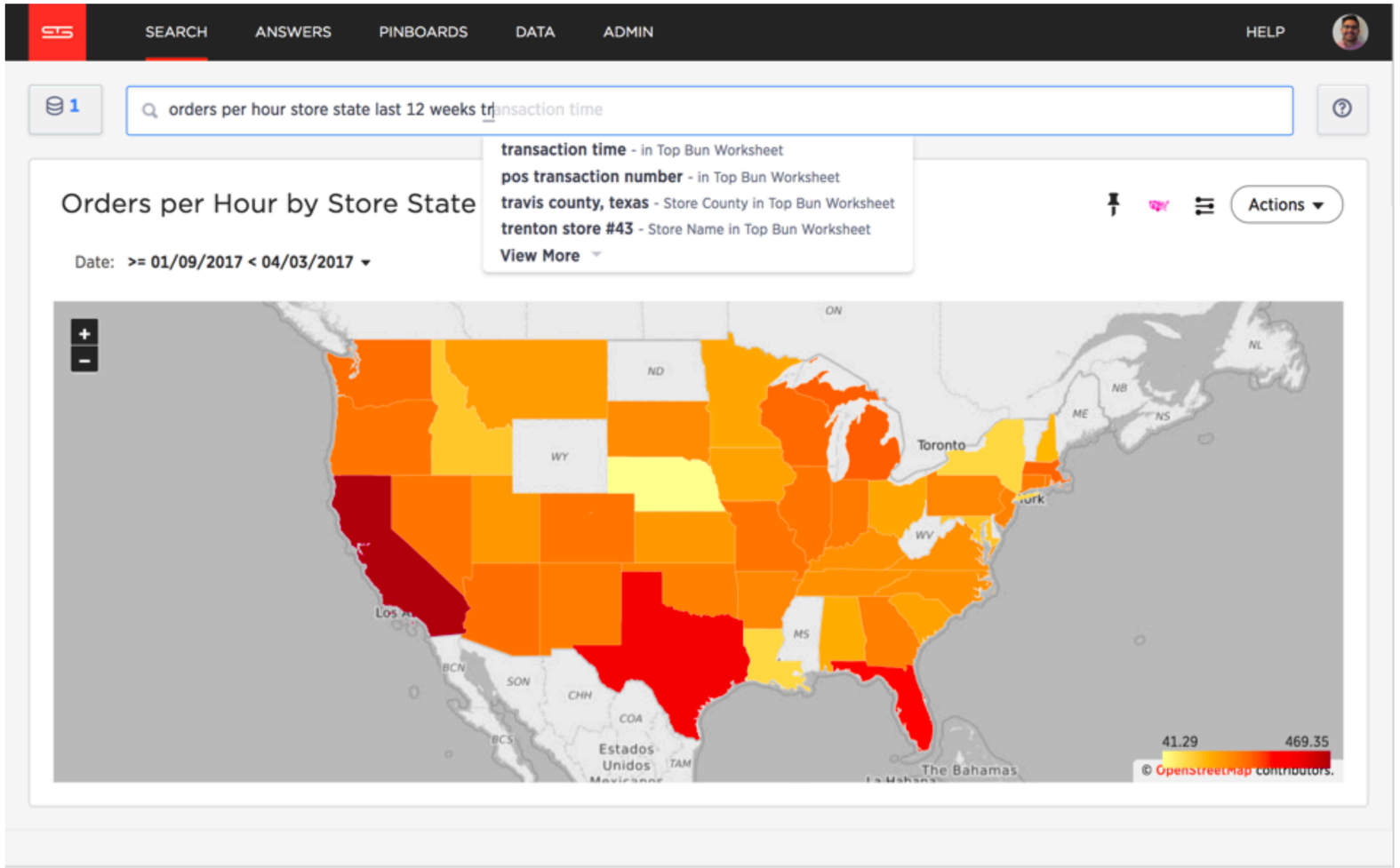
Logical Expressions

Search

Chatbots

What is it?

Using NLP to construct the query and return results



Smart Discovery Types

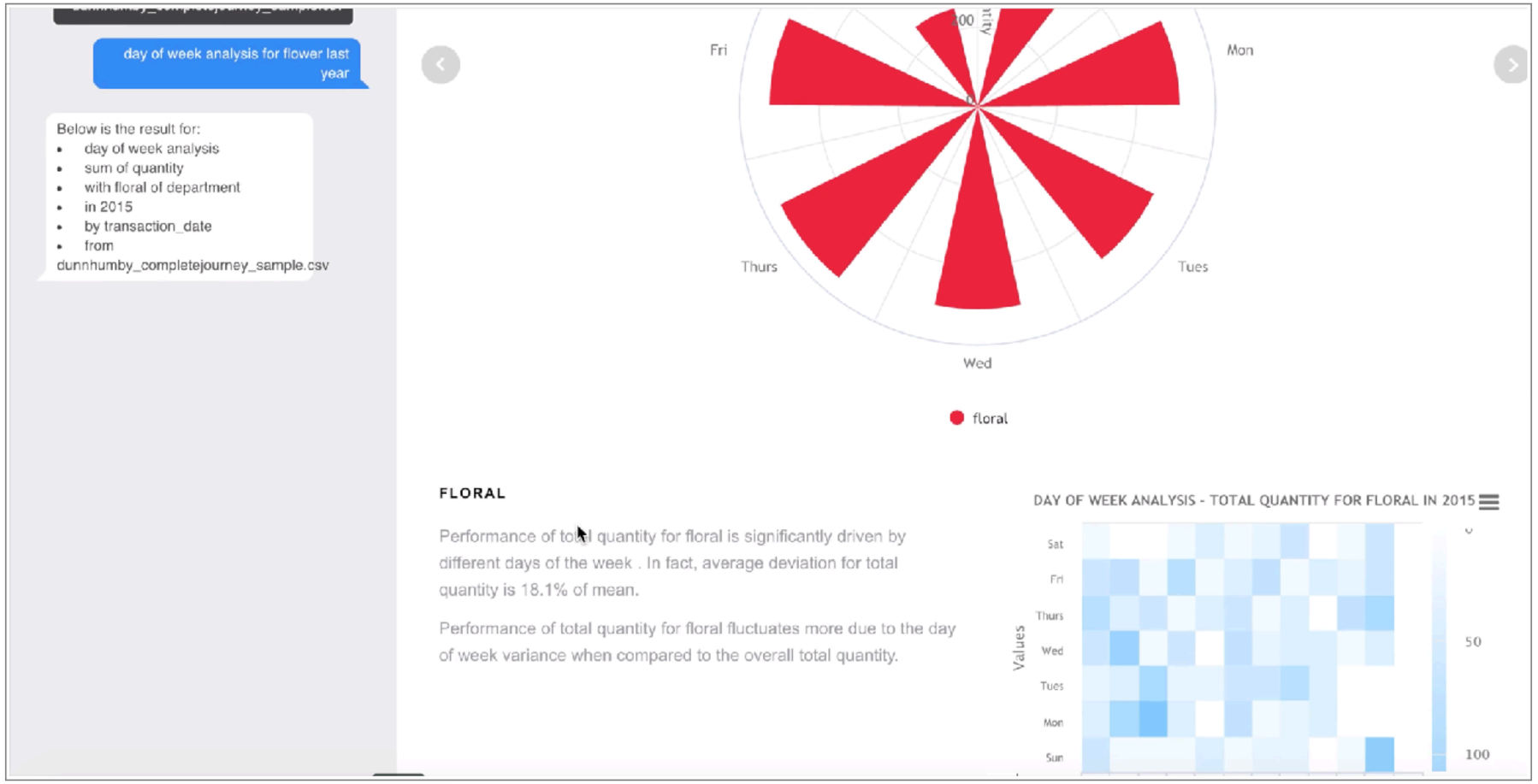
Logical Expressions

Search

Chatbots

What is it?

Human-like interaction to retrieve data



BIG DATA STEWARDSHIP


DS


DS


CROWD
SOURCING

ML examples in stewardship solutions

Trust

Suggest which data can be trusted

Consistency

Automatic metadata curation

Search

Find relations in the data

Secure | https://demo-sales-v.alationdata.com

Alation | Data Catalog | Glossaries | Track | Search | Admin | Help

Search Alation

Data
- [MySQL] Analytics
- Alation DB
- [Hive] Finance and Research ...

File Systems
- Cataloging Flag Files Example
- [S3] Data Lake FS: Bucket X
- S3 Flightaware Data

Queries
- California Bank Customers
  `select * from bank.customer wh…`
  by **Paul Walker** · Endorsed **0** times

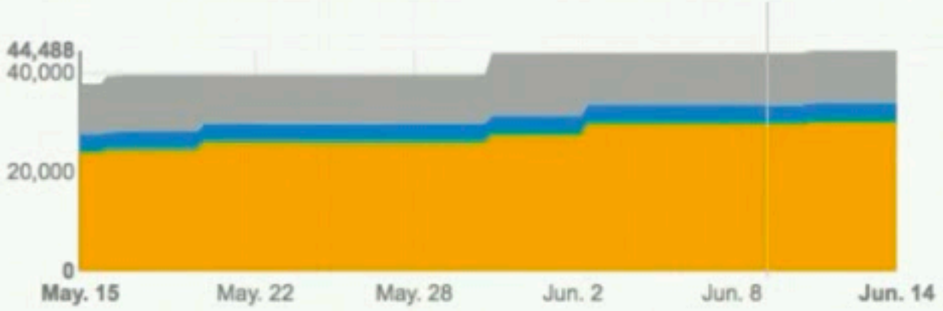
Articles
- Applications
- Building a Taxonomy
- Business Areas

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

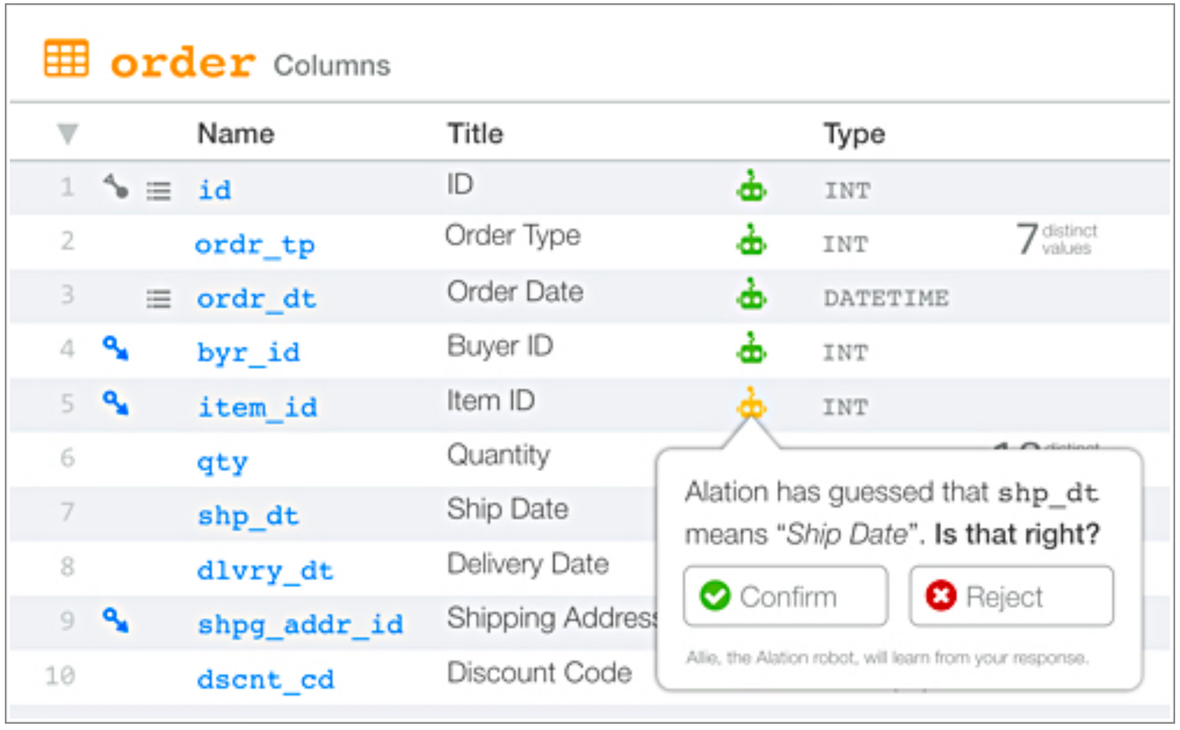
And more..

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


order Columns
Name
Title
Type
1 id ID INT
2 ordr_tp Order Type INT 7 distinct values
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 Addres
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.

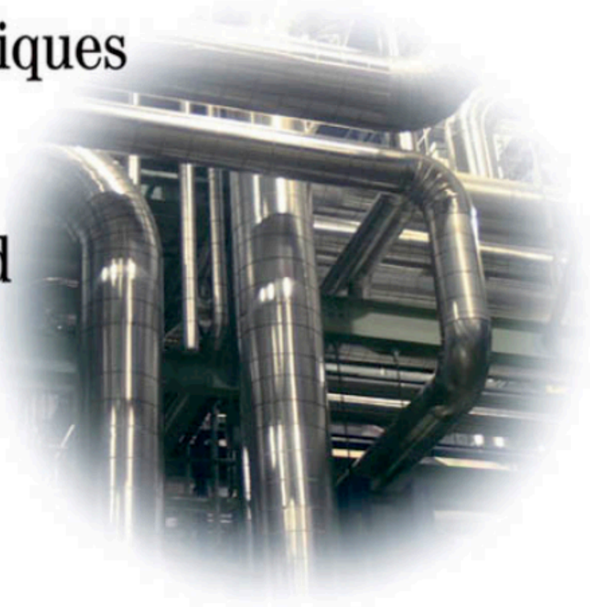
SMART DATA PREPARATION

WILEY

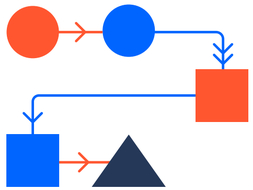
TIMELY. PRACTICAL. RELIABLE.

The Data Warehouse ETL Toolkit

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

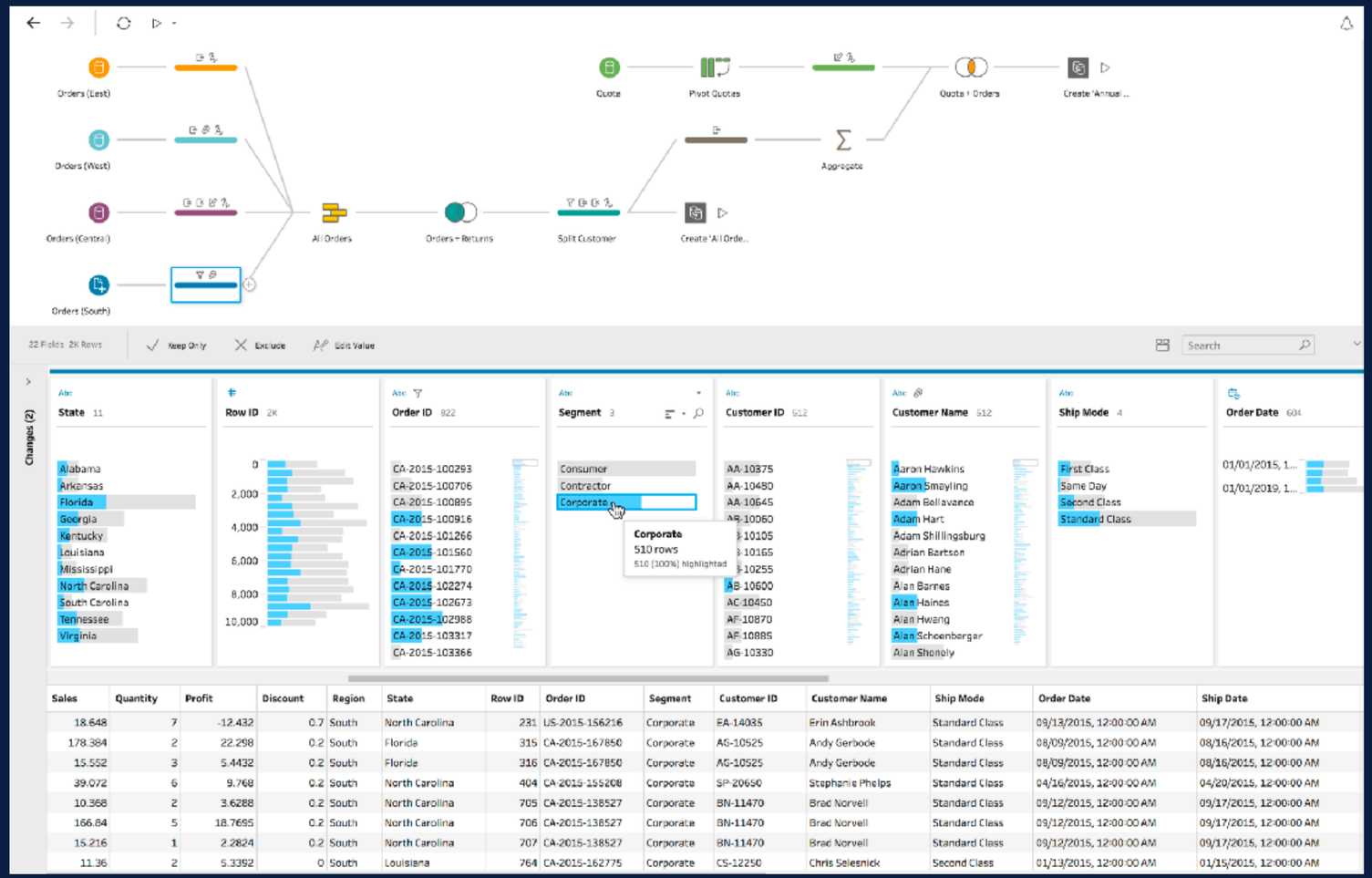


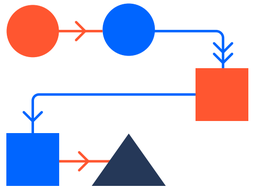
Ralph Kimball
Joe Caserta

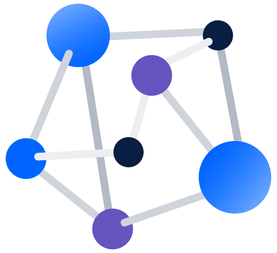


Drag & Drop ETL

Enable non technical users to prepare data

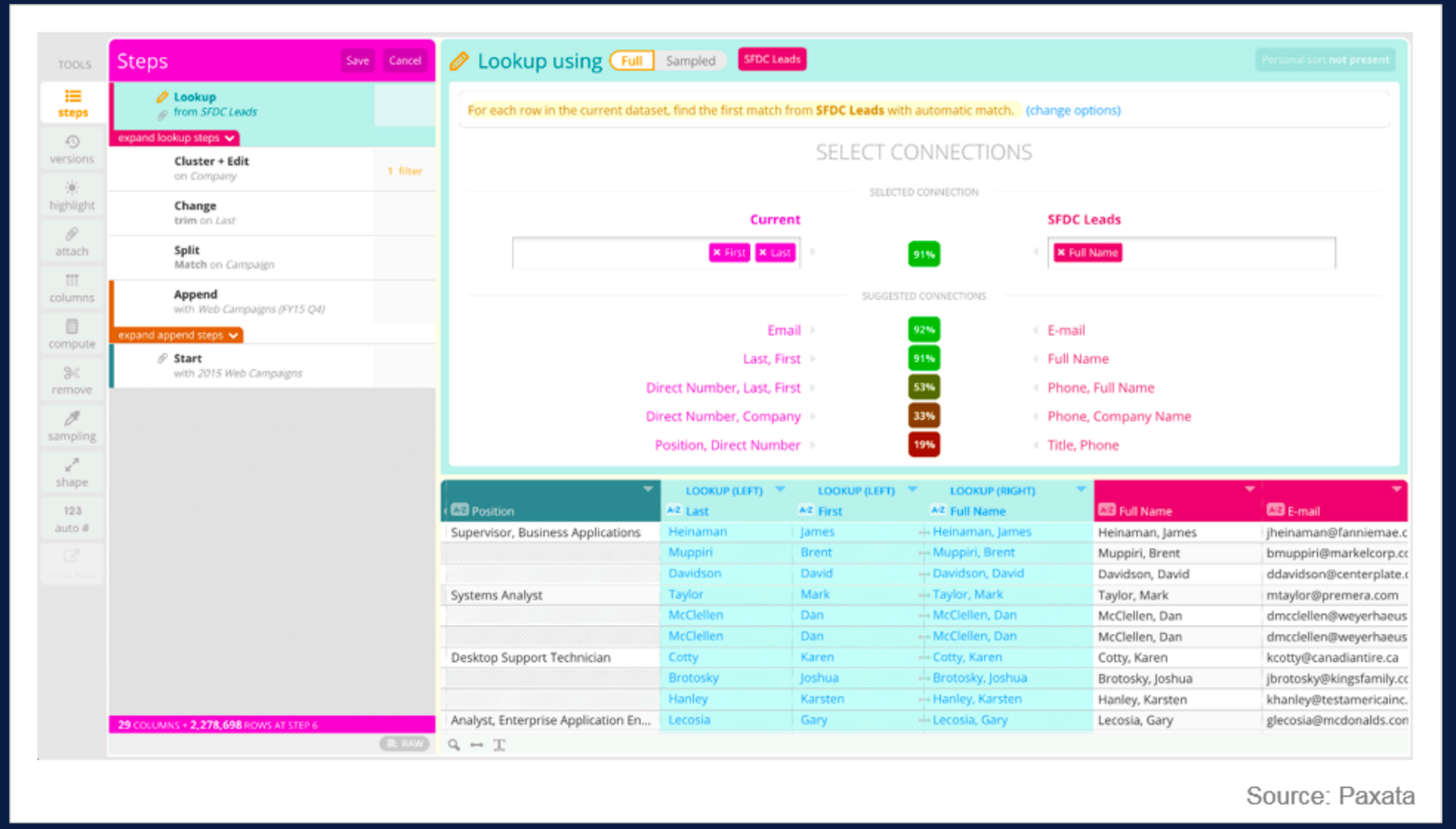




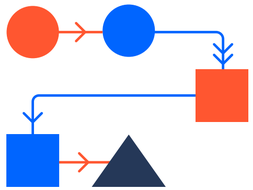


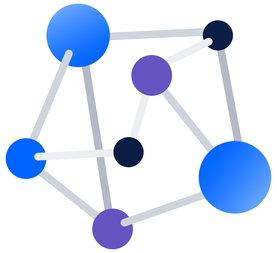
Data Blends

Connecting different datasets automatically



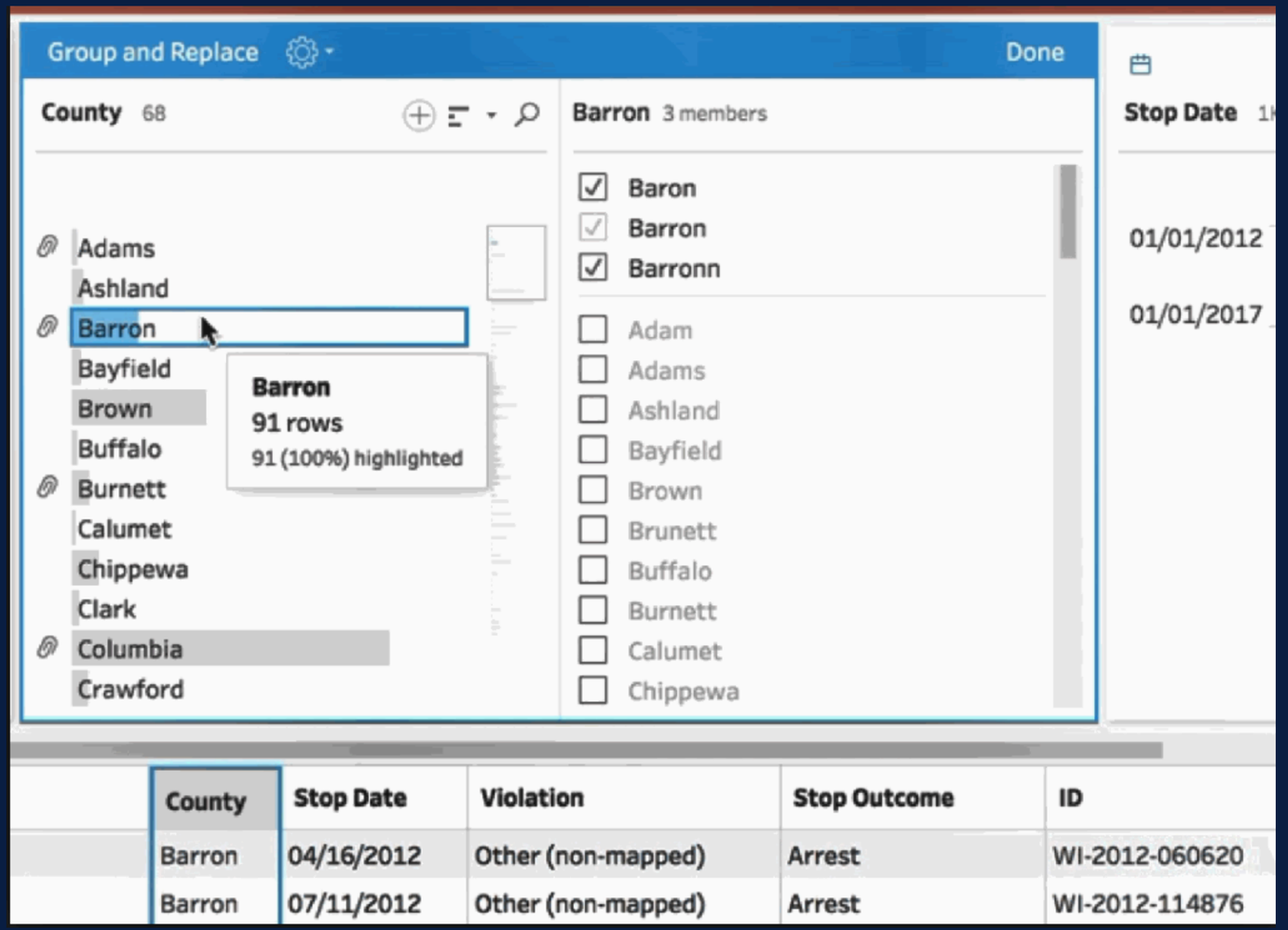
Source: Paxata





Data cleansing

ML for manipulating the data



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.

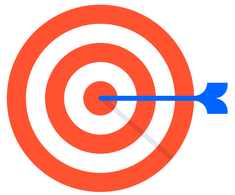
Breach notification within 72 hrs
Breaches must be reported within 72 hours of first having become aware of the breach.

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.

ePrivacy REGULATION > GDPR The General Data Protection Regulation



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