HOW TO DEMOCRATIZE DATA AND SCALE AUGMENTED ANALYTICS

ATSCALE

Today’s Speaker Panel

Jay Schuren

SVP of Customer Success & Enablement, DataRobot

Jay is an industry veteran who has been with DataRobot since 2017 as a Data Scientist and General Manager prior to his current role. Prior to joining DataRobot, Jay held research and leadership roles at Nutonian, the Air Force Research Laboratory, and Bettis Atomic Power Laboratory.

Jay holds a Ph.D. in Mechanical Engineering from Cornell University.

@DataRobot

Kirk Borne

Chief Science Officer DataPrime, Inc.

Kirk has been an influential globally recognized leader in the data science space for 20 years.

His areas of passion and focus include Big Data & Data Science, Artificial Intelligence (AI), and Astrophysics. Kirk is also the co-creator of the field of Astroinformatics.

@dataprime_ai

Gal Barnea

Database Engineering Lead, Redshift

Gal leads the Database Engineering team at Amazon Redshift. In this capacity, Gal works closely with Redshift’s most strategic customers word-wide to optimize and maximize the business value they gain from their data warehouse.

During his career, Gal built & lead engineering teams, oversaw large scale data and analysis initiatives, and worked with some of the world’s largest brands. In his spare time, Gal enjoys cycling in the Bay Area hills and following way too much sports.

@awscloud

Daniel Gray

VP, Solutions Engineering,AtScale

Daniel brings rich experience in technical solutions engineering as well as software engineering to his work with global enterprise organizations.

Prior to joining AtScale to lead the Solutions Engineering team, Daniel spent many years in the analytics space including Hewlett-Packard’s Advanced Technology Center, Vertica, and Domino Data Lab.

@atscale

Data & Analytics Maturity Model

Capabilities	Level 0 Initial	Level 1 Procedural	Level 2 Proactive	Level 3 Leading
Data	Siloed	Centralized	Enhanced	Shared
Access	Extracts	ETL/ELT	Virtualization	DataOps
Model	Object	Tabular	Logical	Dimensional
Analyze	Analyst	BI User/Data Scientist	Citizen Data Scientist	Everyone
Consume	Query	Dashboards	Self-service Analysis	Data as Code
Insights	Descriptive	Diagnostic	Predictive/Augmented	Prescriptive

Smarter Decisions

The Challenge of Bridging Silos

Data

Data Warehouses, Data Lakes, Apps, Feature Stores, 3rd party Data, etc…

Teams

BI, Data Science, IT, Line of Business, etc…

Technologies

BI Platforms v. AutoML Platforms, SQL v. Python, etc…

Priorities

Where to invest limited resources? Budget, resources, cycles, attention, etc…

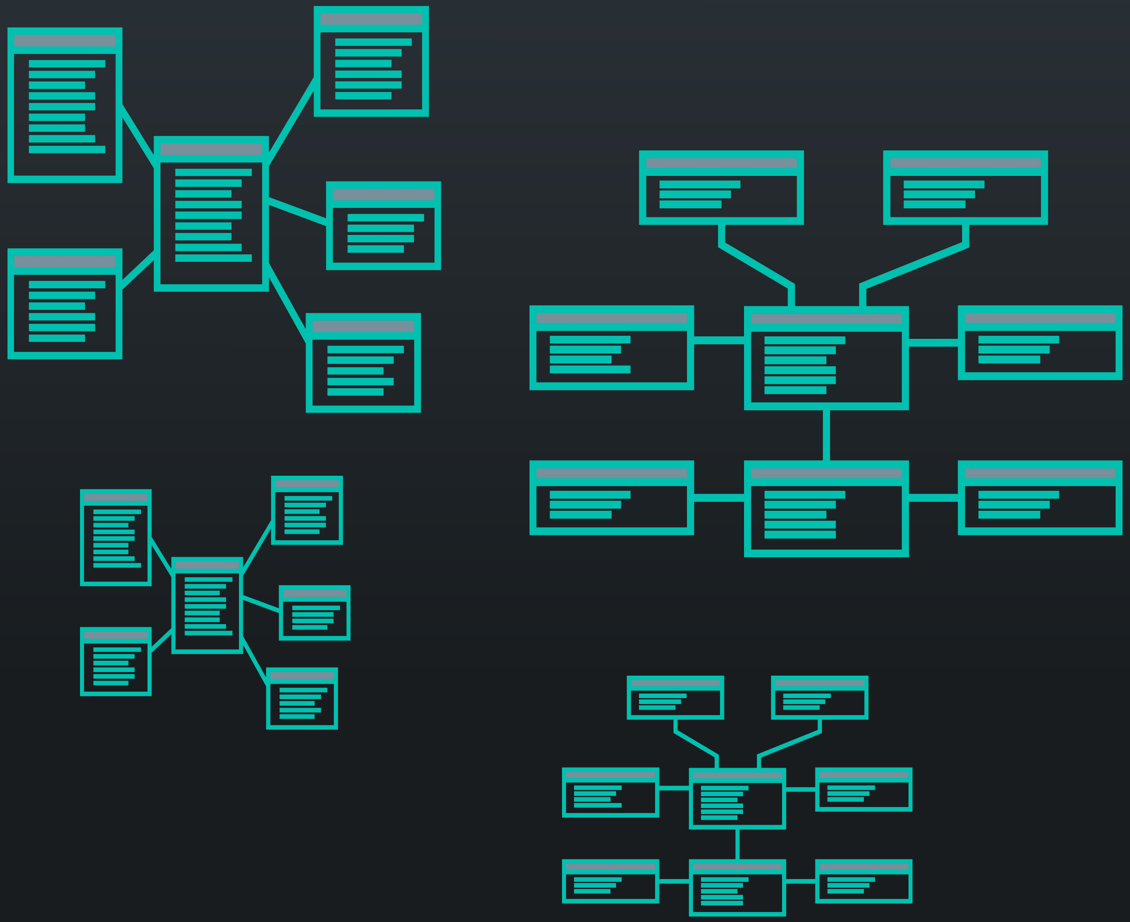
Cultures

IT Pros v. Business Operators v. Data Scientists, etc…

Semantics

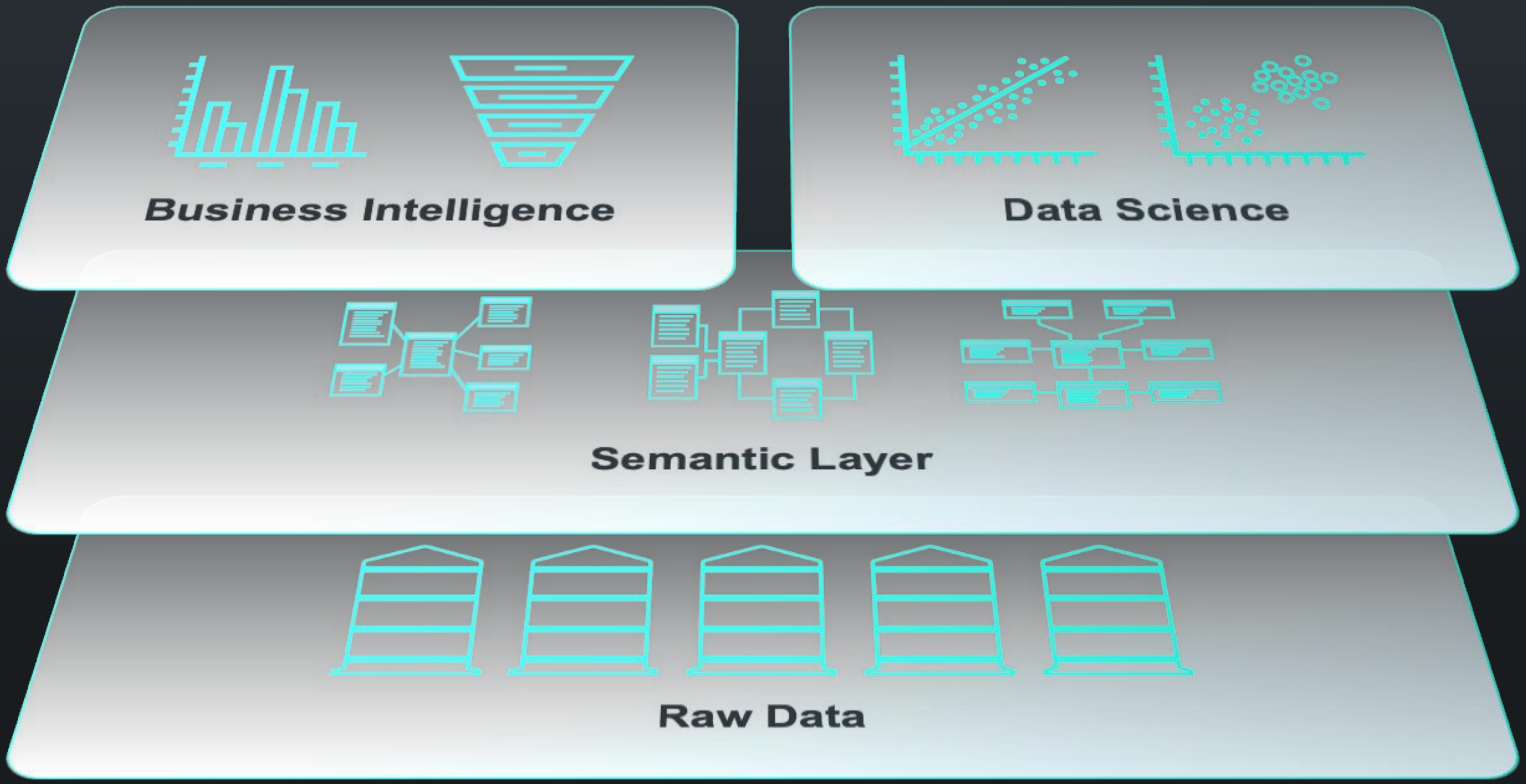
Descriptive v. Predictive v. Prescriptive, Dimensions v. Features, etc…

The Power of Semantics



- Define and Enforce Logical Data Relationships Across Tables and Data Sources
- Business Metric Definitions
- Conformed Dimensions and Reusability
- Centralized Metric Repository
- Data Model Library for Common Data Sets

Bridging Raw Data to BI and Data Science with a Semantic Layer

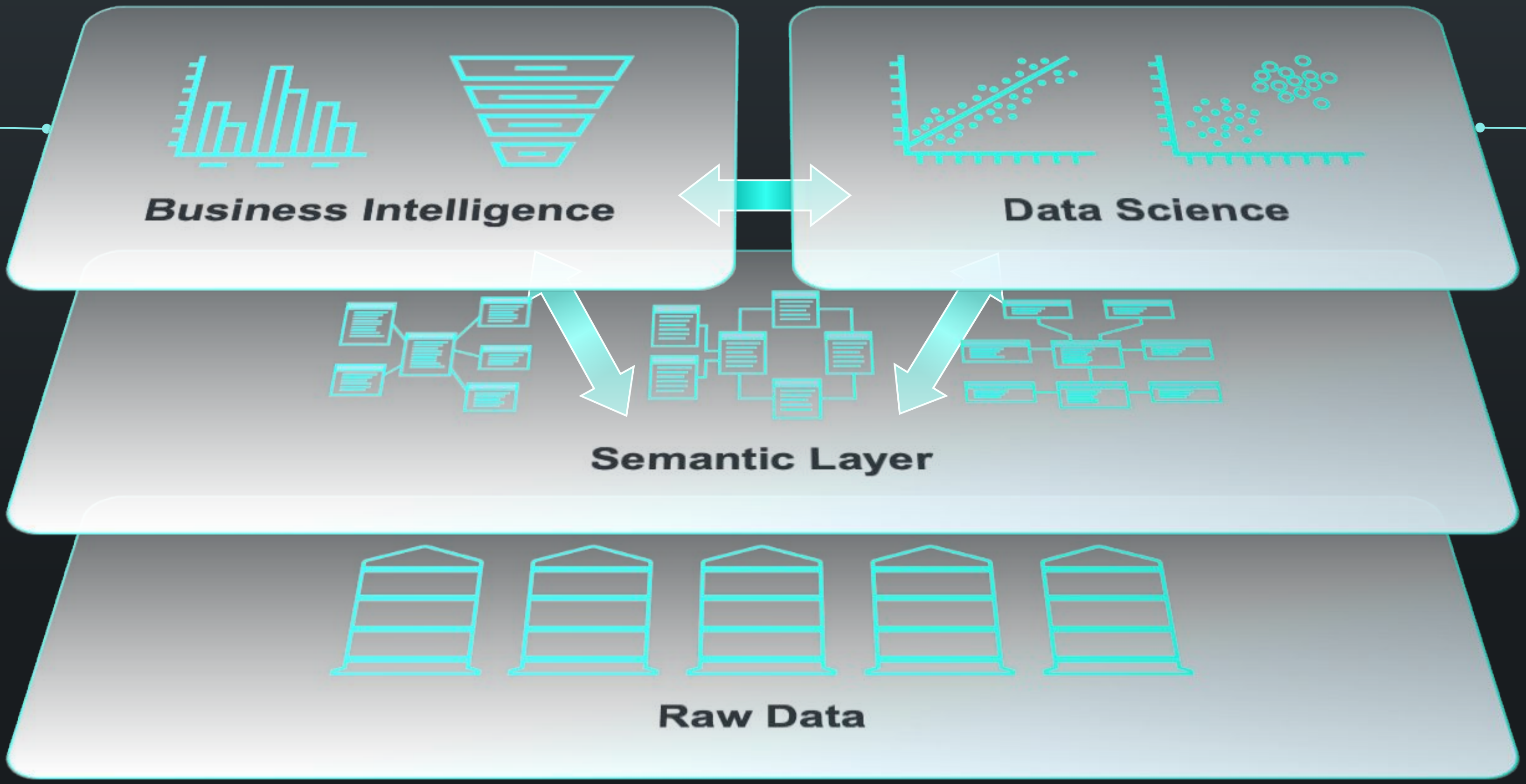


Creating a Structured Analysis Feedback Loop

KPI Definition

Dimensionality

Hierarchical Definition



Feature Creation

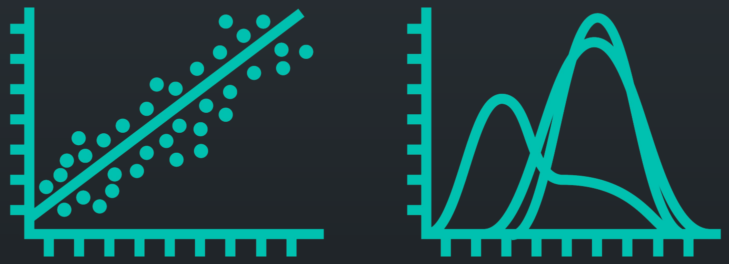
Prediction Modeling

Explain-ability

Scoring/Drift

Exploratory vs Structured Data Analysis

EXPLORATORY ANALYSIS



- Informal modeling
- Investigative
- Data mining
- Pattern, Anomaly, Hypothesis testing

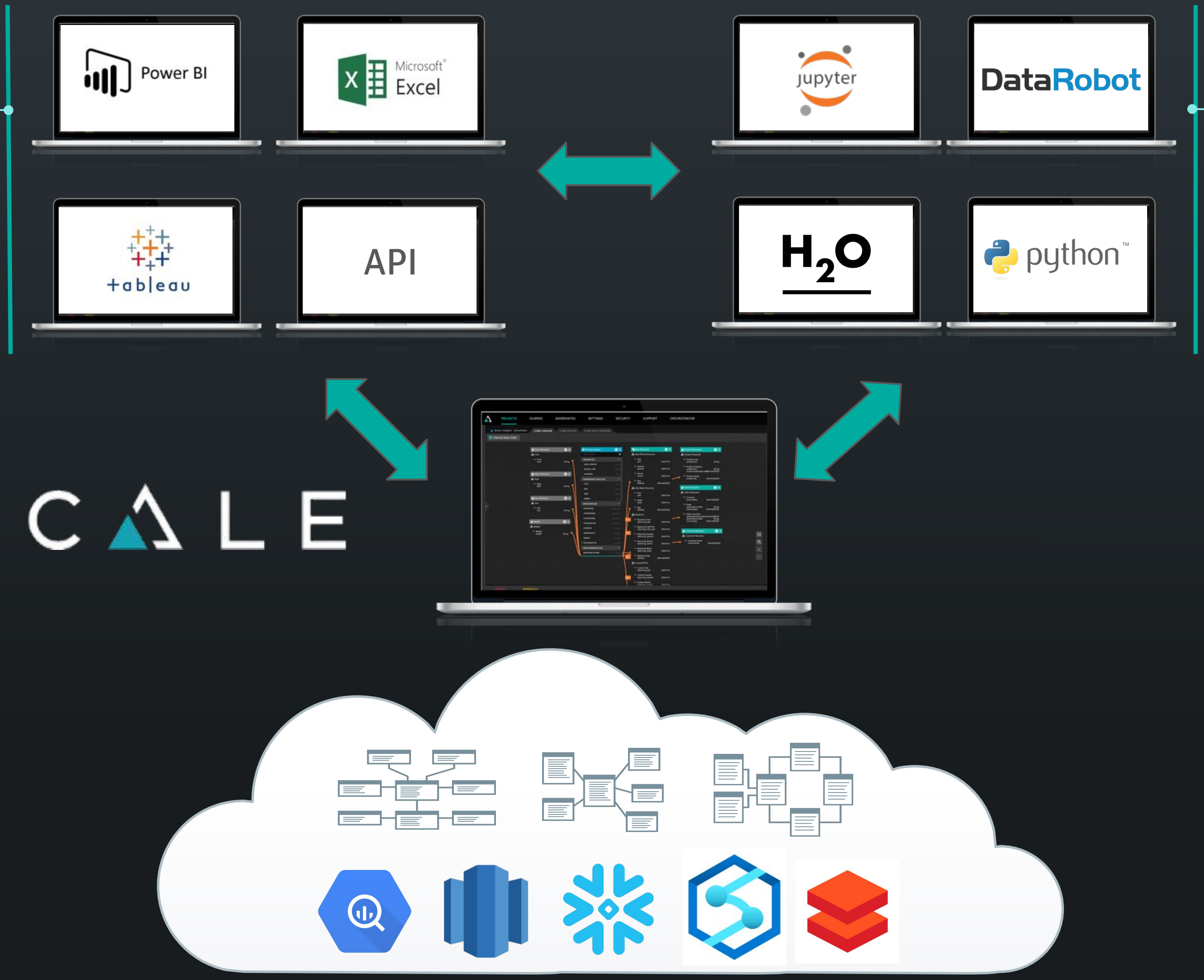
STRUCTURED ANALYSIS

- Importance of the Data Model
- Consistency and Reuse
- Domain vetted
- Governance and Reliability
- Production - Run your business on

Bridging Raw Data to BI and Data Science with a Semantic Layer

BI Teams

- Define KPIs used by the business
- Data dimensionality (e.g. time, geography, product, customer, etc.)
- Hierarchical definition (i.e. time series analytics, drill into data for granular analysis)



ATSCALE

Data Science Teams

- Develop domain specific features
- Build predictive models based on features
- Time series predictions
- Ability to explain predictive model outcomes
- Score models and understand model drift

Questions for Daniel

- Daniel, I know you are on the front lines with our customers so you see a lot of trends before anyone else does. Is the merging of BI and AI top of mind with the forward leading customers out there?

Modern data platform requirements

Use: Support exploratory data analysis and ML

- Embedded analytics
- Notebook automation
- Predictive analytics
- Operational analytics
- One-time query

Manage: Data discovery, search, and collaboration

- Catalog and search
- Governance share data
- Master data management
- Data quality
- Security and management

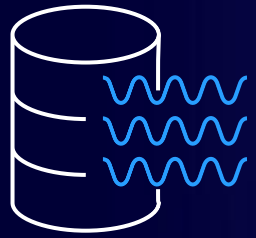
Run: Data processing and platform frameworks

- Data ingestion
- Data transformation
- Databases and storage
- Code and infrastructure automation

Amazon Redshift

THE MOST WIDELY USED CLOUD DATA WAREHOUSE, WITH TENS OF THOUSANDS OF CUSTOMERS

ANALYZE ALL YOUR DATA



Take a **lake house approach** by analyzing all your data across your data warehouse, your Amazon S3 data lake, and operational databases with consistent security and governance policies

PERFORMANCE AT ANY SCALE

Get up to **3x better price performance** than other cloud data warehouses with a **self-tuning** system, and boost queries up to **10x with AQUA**

LOWER YOUR COSTS

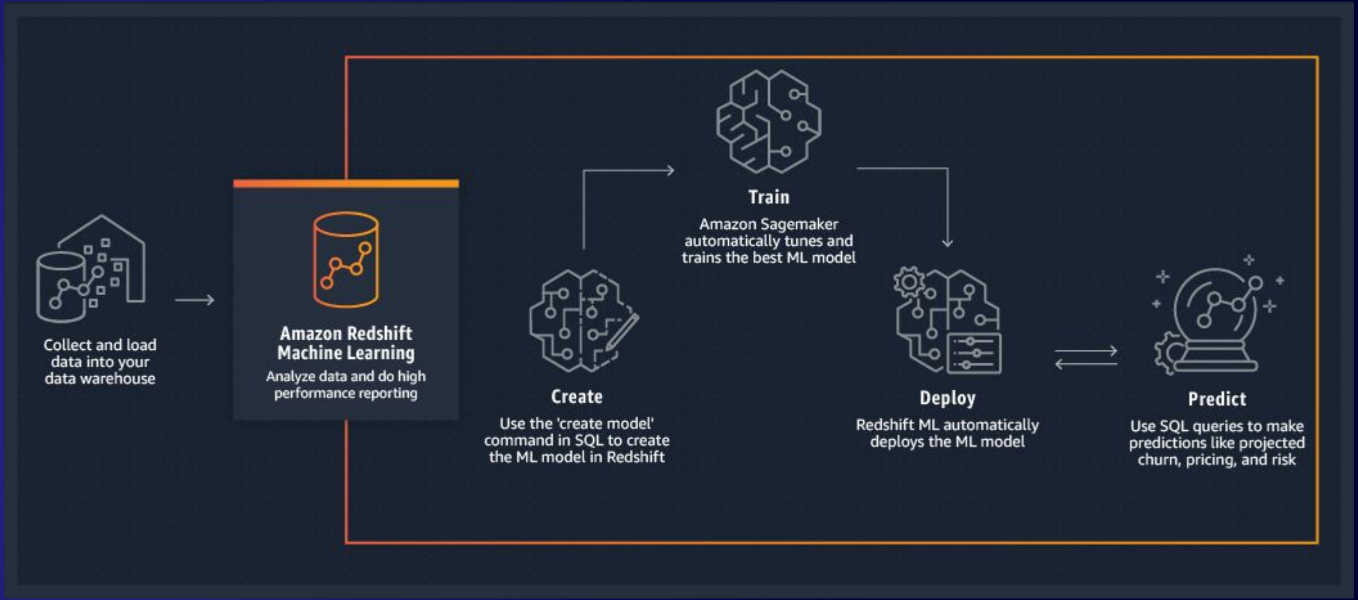
Start small and pay only for what you use with **predictable** monthly costs, Amazon Redshift is at least **50% less expensive** than other cloud data warehouses

Amazon Redshift ML

NEW

EASILY CREATE AND TRAIN ML MODELS USING SQL QUERIES WITH AMAZON SAGEMAKER

- Use case: Product recommendations, fraud prevention, reduce customer churn
- Train and apply ML models using SQL
- From fully automated training to partially or fully guided training
- Automatic pre-processing, creation, training, deployment of your model



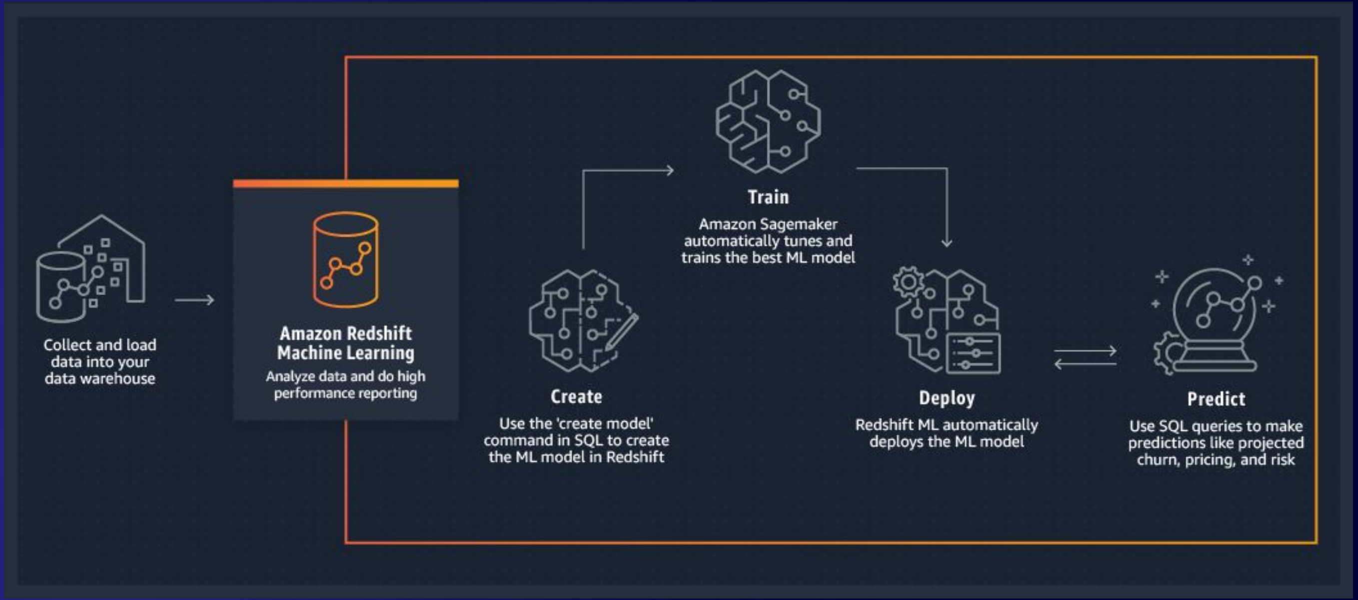

```
CREATE MODEL customer_churn
FROM (SELECT c.age, c.zip, c.monthly_spend,
c.monthly_cases, c.active FROM customer_info_table c)
TARGET c.active
FUNCTION predict_customer_churn
…;
```

Amazon Redshift ML

EASILY CREATE AND TRAIN ML MODELS USING SQL QUERIES WITH AMAZON SAGEMAKER

NEW

- Deploy inference models locally in Amazon Redshift
- Run an inference as invoking a user-defined function as part of SQL statements

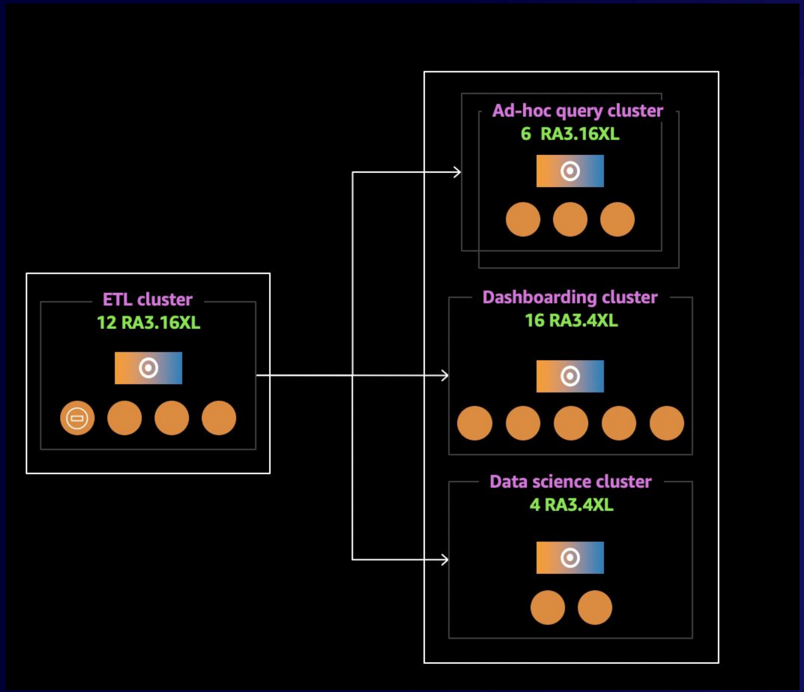


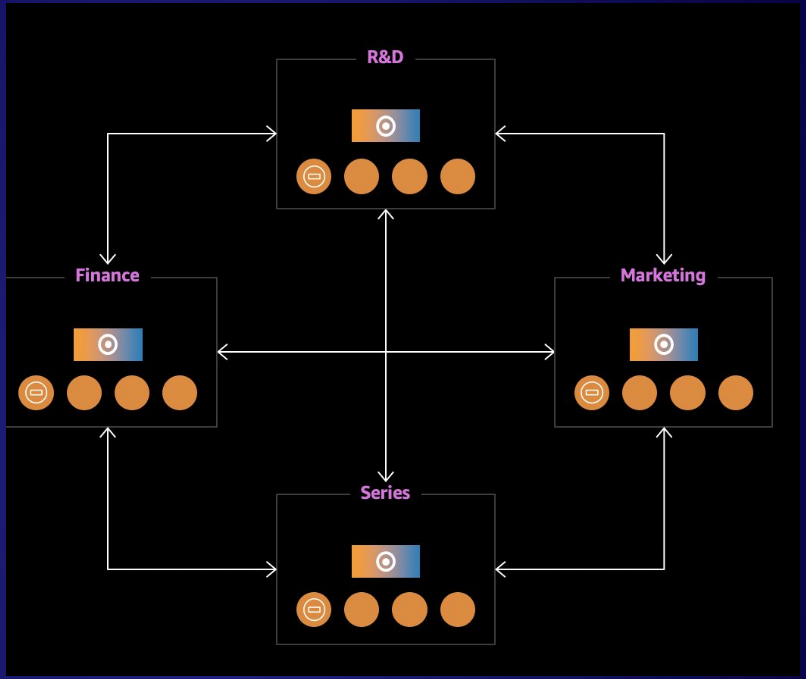

```
SELECT n.id, n.firstName, n.lastName,
predict_customer_churn(n.age,c.zip,..)
AS activity_prediction
FROM new_customers n
WHERE n.marital_status = ‘single’
…;
```

NEW

Data sharing

- A SECURE AND EASY WAY TO SHARE DATA ACROSS AMAZON REDSHIFT CLUSTERS





- Instant, granular, high-performance data access without data copies / movement
- Live and consistently updating views of data across all consumers
- Secure and governed collaboration within and across organizations and with external parties

- Workloads accessing shared data are isolated from each other
- Use cases: Cross-group collaboration and sharing, workload isolation and chargeability, data as a service

"Data sharing feature seamlessly allows multiple Amazon Redshift clusters to query data located in our RA3 clusters and their managed storage. This eliminates our concerns with delays in making data available for our teams, reduces the amount of data duplication and associated backfill headache. We now can concentrate even more of our time making use of our data in Amazon Redshift and enable better collaboration instead of data orchestration."

Steven Moy, Yelp

The Data-Driven Organization

“An organization that harnesses data as an asset, to drive sustained innovation and create actionable insights to supercharge the experience for their customers so they demand more.

85% of businesses want to be data driven

37% have been successful

Source: Forbes Online; New Vantage Partners—Big Data Executive Survey

Questions for Gal

- We hear a lot about adopting a Lakehouse architecture instead of a Data Warehouse architecture. It seems like Redshift is offering the best of both worlds. Why would customers want to leave their data in the data lake (i.e. S3) rather than loading it into the Redshift data warehouse?

Data Science Perspectives on Augmented Analytics

by Kirk Borne
Chief Science Officer, DataPrime.ai

“How many cows are in Texas?”

1) **Standard search – type this question into a search box, which then finds documents containing the keywords “how”, “many”, “cows”, “Texas”**

2) **Voice-assisted search – verbally ask the question to a voice-enabled search bot, which then finds documents containing the keywords “how”, “many”, “cows”, “Texas”**

3) **Database search – build a data model to answer all possible questions from your humans, load the database with all possible data, and then:**
select animal.counts where animal.type=‘cow’ and animal.location=‘Texas’

4) **Augmented Analytics search – ask a natural language question to an AI-powered search bot with a semantic data layer (between the input layer and the data layer). The semantic layer then interprets the question, which enables the analytic engine to provide the correct answer.**

The Power of Semantics in Augmented Analytics

a) Makes your data “smarter” = Augments the data with “meaningful” metadata (labels, annotations).

b) Enables integration of diverse data sources by representing (and making searchable) the knowledge content and the context of the data, not simply storing each dataset’s source-specific syntax and labels.

c) Differentiates business analytics use cases between old school (querying the database for factoids) and new (querying the data for knowledge).

d) Powers smarter analytics-driven decision-making.

e) Delivers business answers at the speed of business questions!

f) How? …with graph (linked) data, ontologies, taxonomies, folksonomies.

5 Dimensions of Analytics-Driven Outcomes & Decisions

1) Descriptive Analytics

- **Hindsight** (What happened?)
- Asks the required questions.

2) Diagnostic Analytics

- **Oversight** (Real-time / What is happening? Why did it happen?)

3) Predictive Analytics

- **Foresight** (What will happen?)

4) Prescriptive Analytics

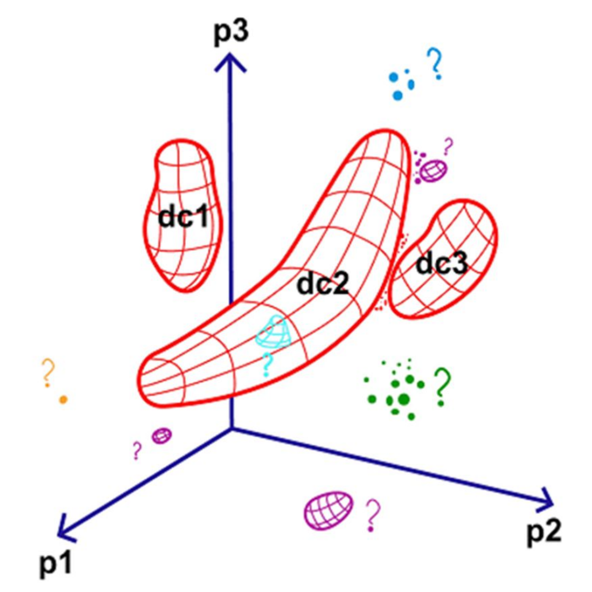
- **Insight** (How can we optimize what happens?) (Search high-dimensional datasets for causal factors and responses.)

5) Cognitive Analytics

- **Right Sight** (the right decision and the right action right now, in the right context)
- Moves beyond simply providing answers, to **generating new questions from the data.**

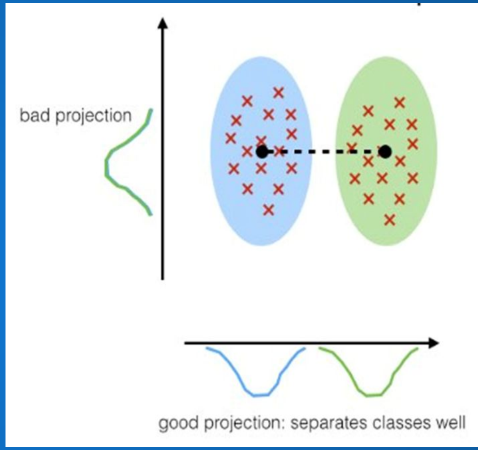
Big Benefits of High Dimensions in Augmented Analytics

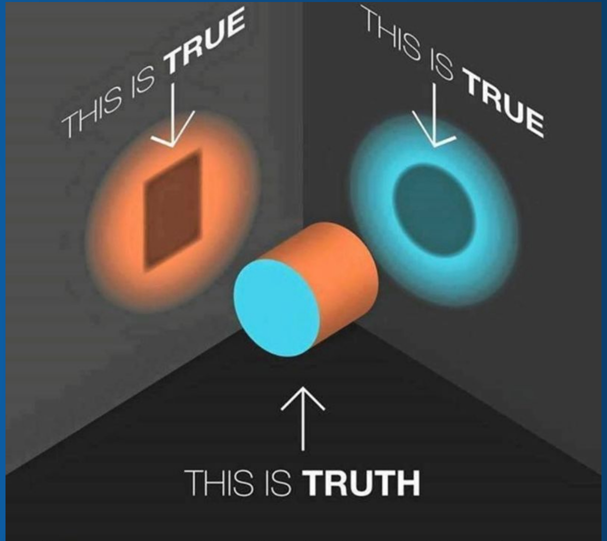
PRESCRIPTIVE (CAUSAL) ANALYSIS



- Moves beyond correlation discovery (for prescriptive analytics)
- Explores inter-dependencies among additional (contextual) data features to reveal governing principles, causal relations, behavioral patterns

BIAS-BUSTING

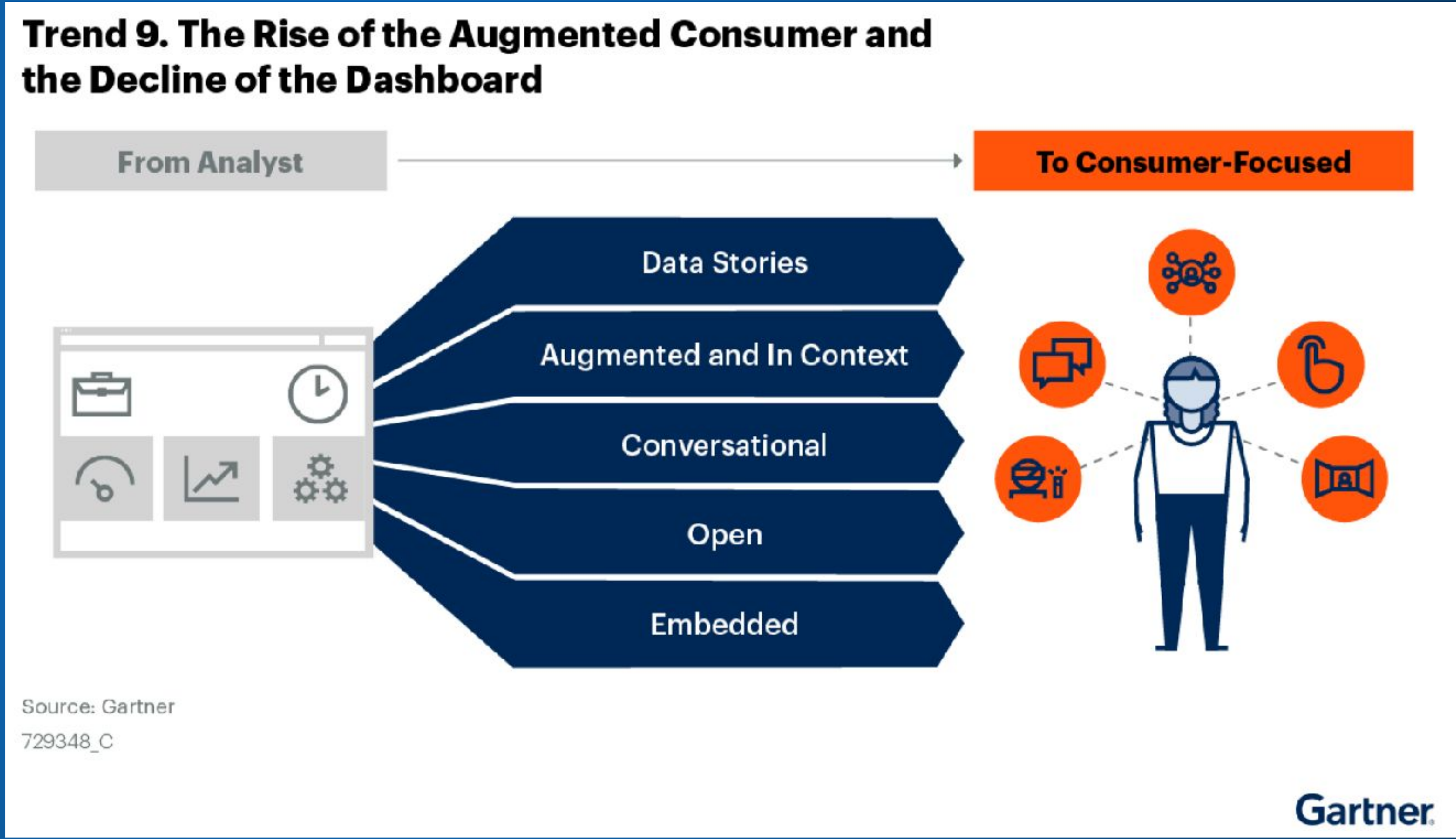




http://www.transformativeinsights.co.nz/blog/new-perspective-on-conflict

- Decreased model bias (underfitting)
- Discovery of new classes (superposed in low-D)
- Improved separation of overlapping classes
- Entity Disambiguation & Entity Deduplication

The Rise of the Augmented Analytics Consumer

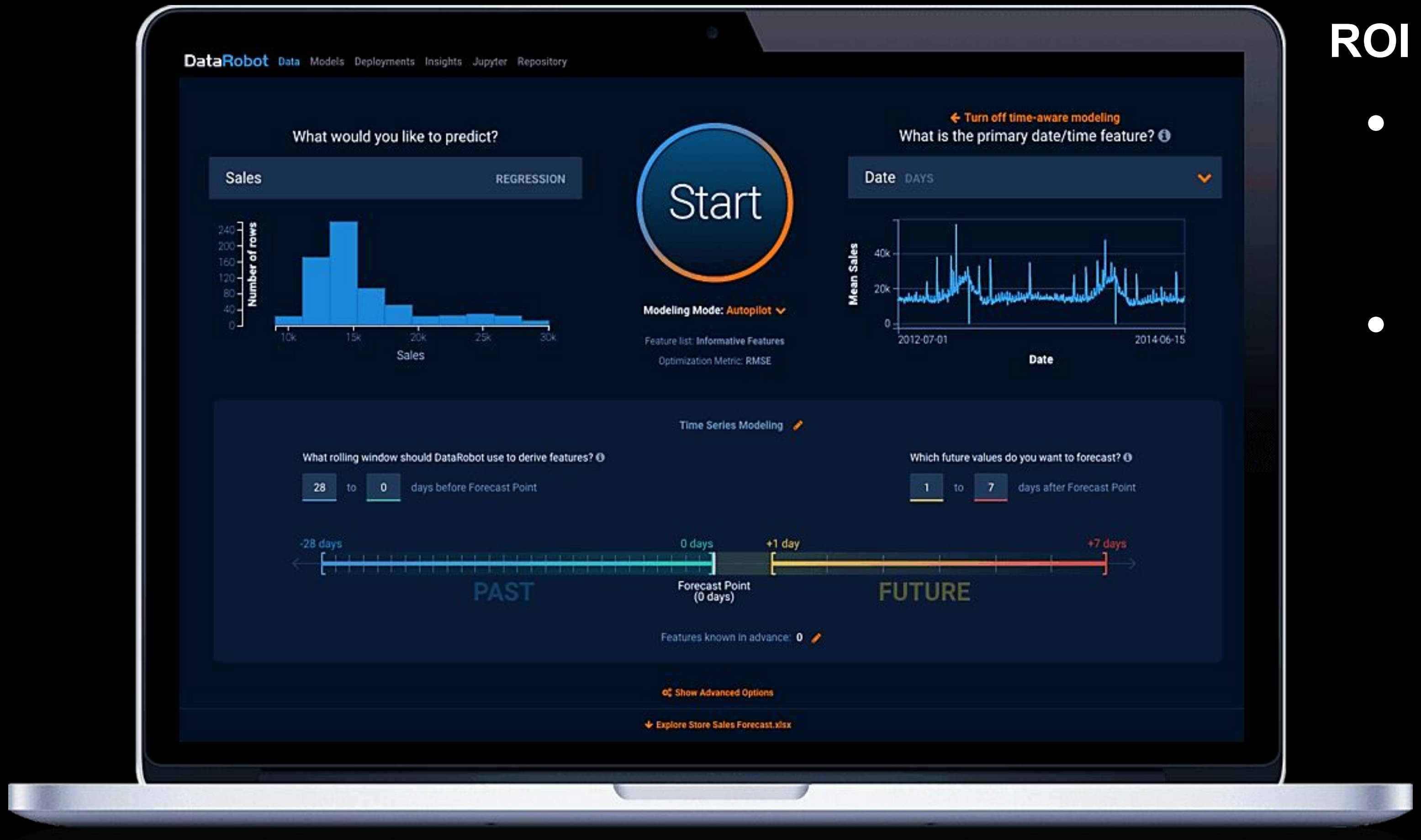


https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021/

Questions for Kirk

- I love your last slide with Gartner's predictions, especially the prediction that Data & Analytics becomes a core business function. We are already seeing that in our most mature customers. For Data & Analytics to become a core function, the business definitely has to move beyond the dashboard as you discussed. Given that, if it's not the dashboard, what do you predict will become the predominant analytics "interface" for consumers?

DataRobot is an End-to-End Enterprise AI Platform

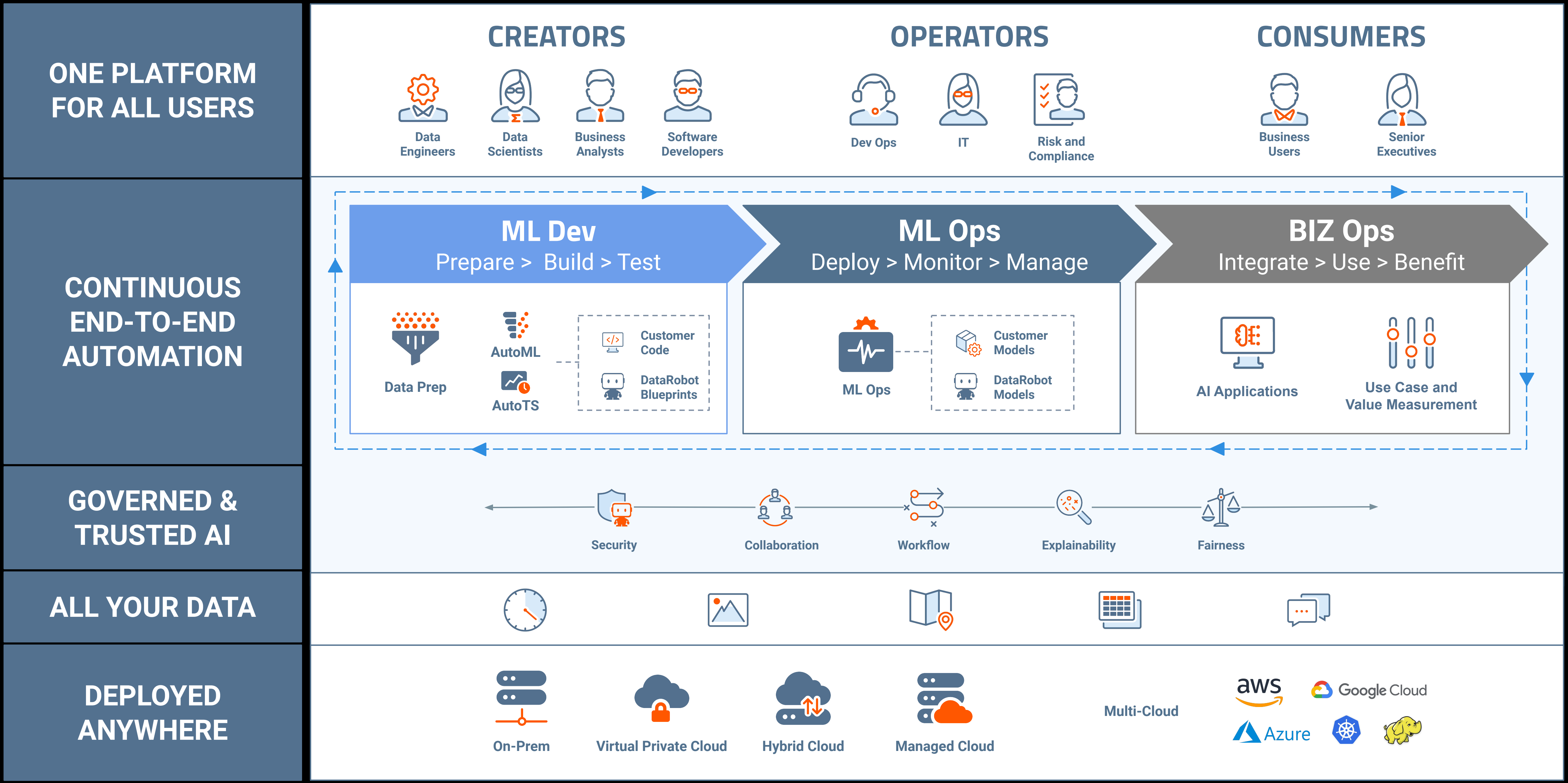


ROI of AI: Forrester Total Economic Impact

- Recognized value through increased revenue, cost savings, avoid over-hiring employees, faster delivery time from data science team, improved opportunity conversion, and reduced customer churn.
- Average ROI of **514%** over four companies, with a **payback in under 3 months**

$18M	$200M	$400M
Healthcare system reduces patient stays and gets more efficient nurse staffing.	Realized value by retailer for multiple SKU-level demand forecasting use cases	Value to top 5 global bank across multiple use cases and models in production

DataRobot Enterprise AI Platform



Scaling AI through Cohort Based Enablement

Establish a program to upskill roles across entire orgs., focused on high-value delivery

Identify

Across a massive workforce, find and recruit the capable and motivated

Enable

Form cross-functional cohorts to build AI solutions or broad fluency for the entire organization

Prioritize

Build culture that can identify and prioritize AI opportunities

Execute

Put theory to practice with hands on experience building AI solutions

AI Awareness Sessions

Cohort Learning and Delivery

Broad AI Fluency

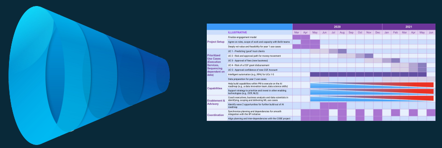
Role-Specific Training

Instructor-led sessions tailored to each specific background to identify and deliver AI solutions

Leaders | Strats/Quants | Data Scientists | Analysts | Engineers

AI Roadmapping

Leverage best practice approaches to identify, evaluate, and prioritize high-value and feasible AI solutions



Build and Deliver AI Solutions

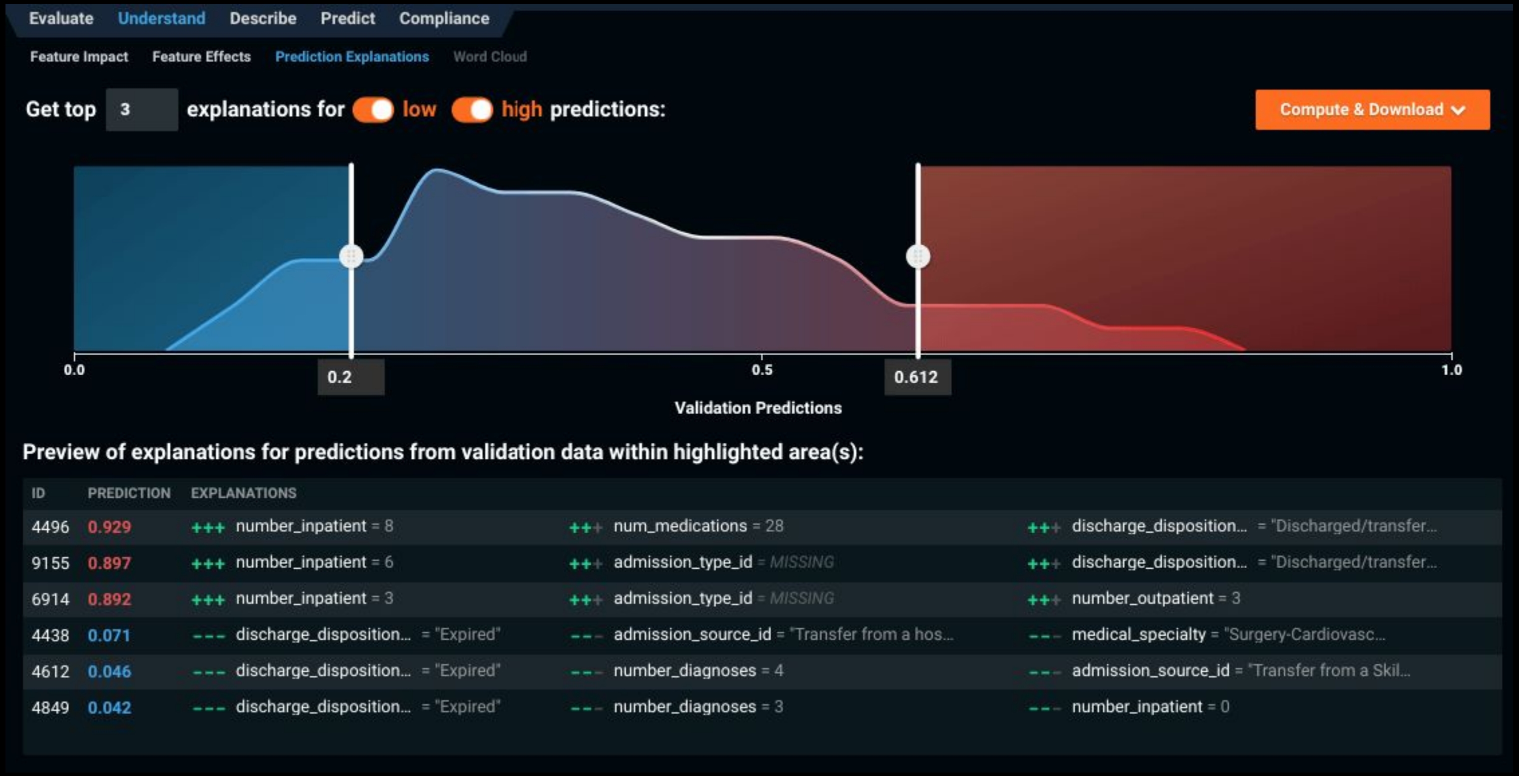
Cross functional teams formed to deliver prioritized AI solutions

Monthly Sessions and Forums

Sessions to build an AI culture broadly across the organization

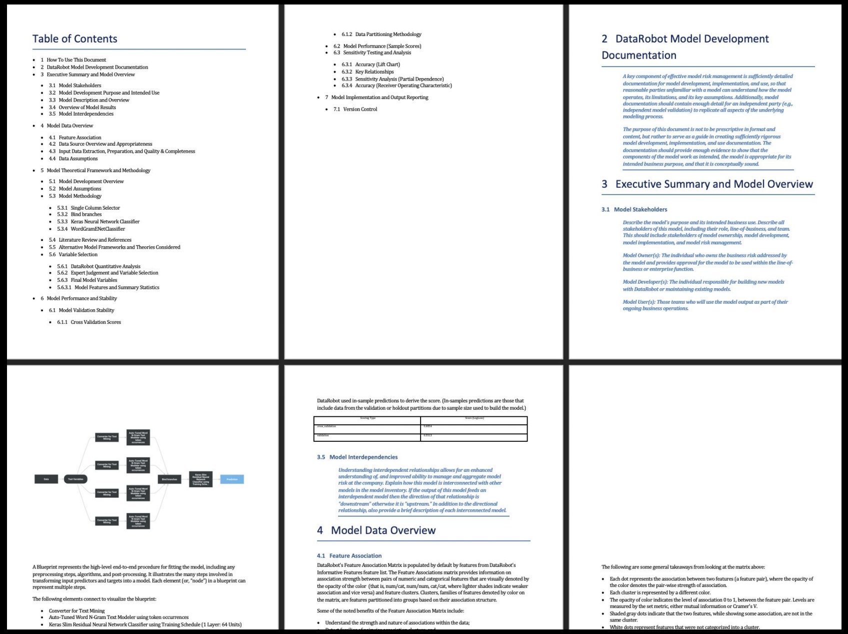
Communicate clearly with stakeholders.

All stakeholders need a shared foundation of knowledge to develop trust with a system.



Individual predictions have clear explanations supporting the AI recommendation.

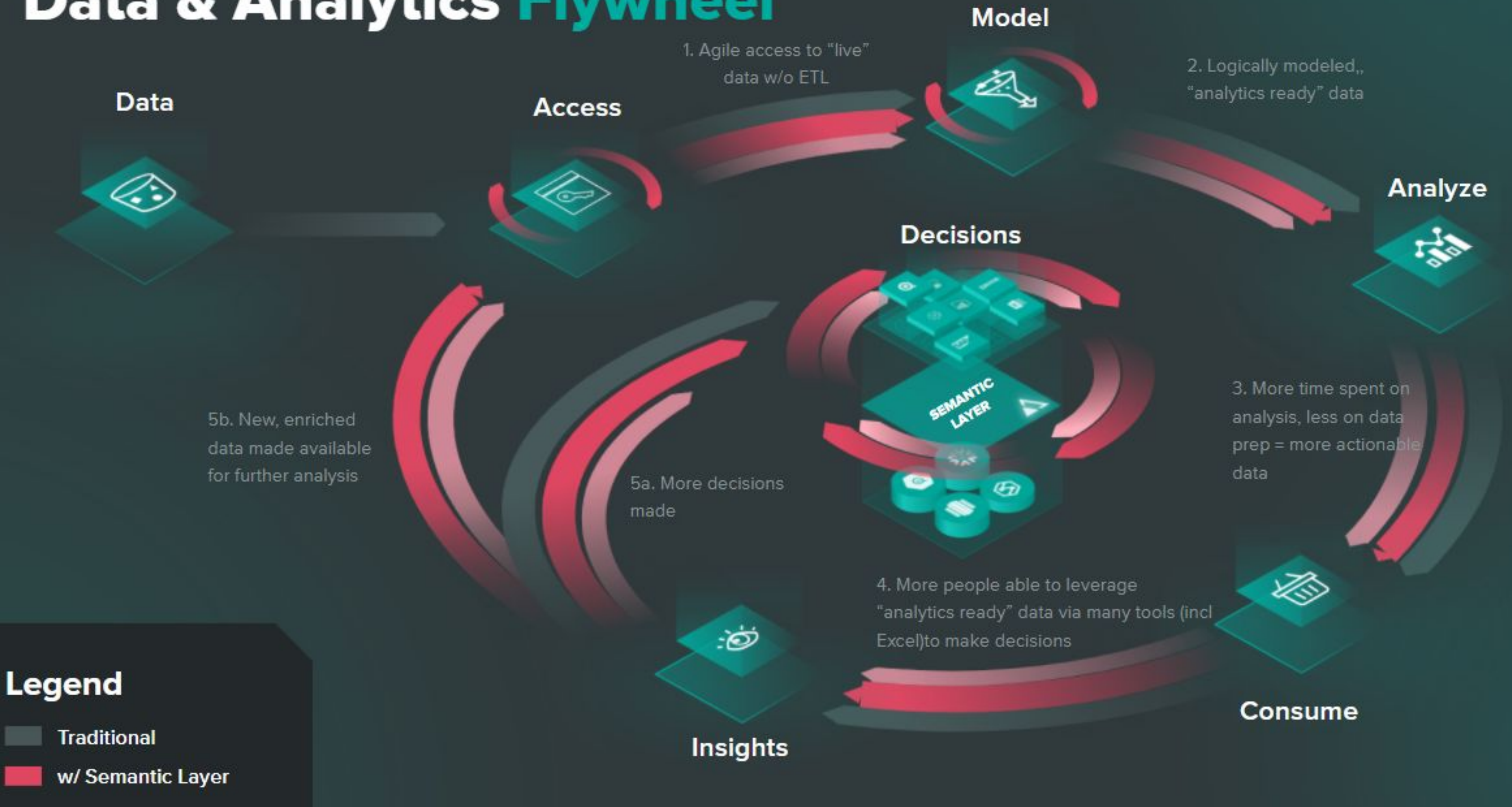
Automated Compliance Documentation accelerates the path to production



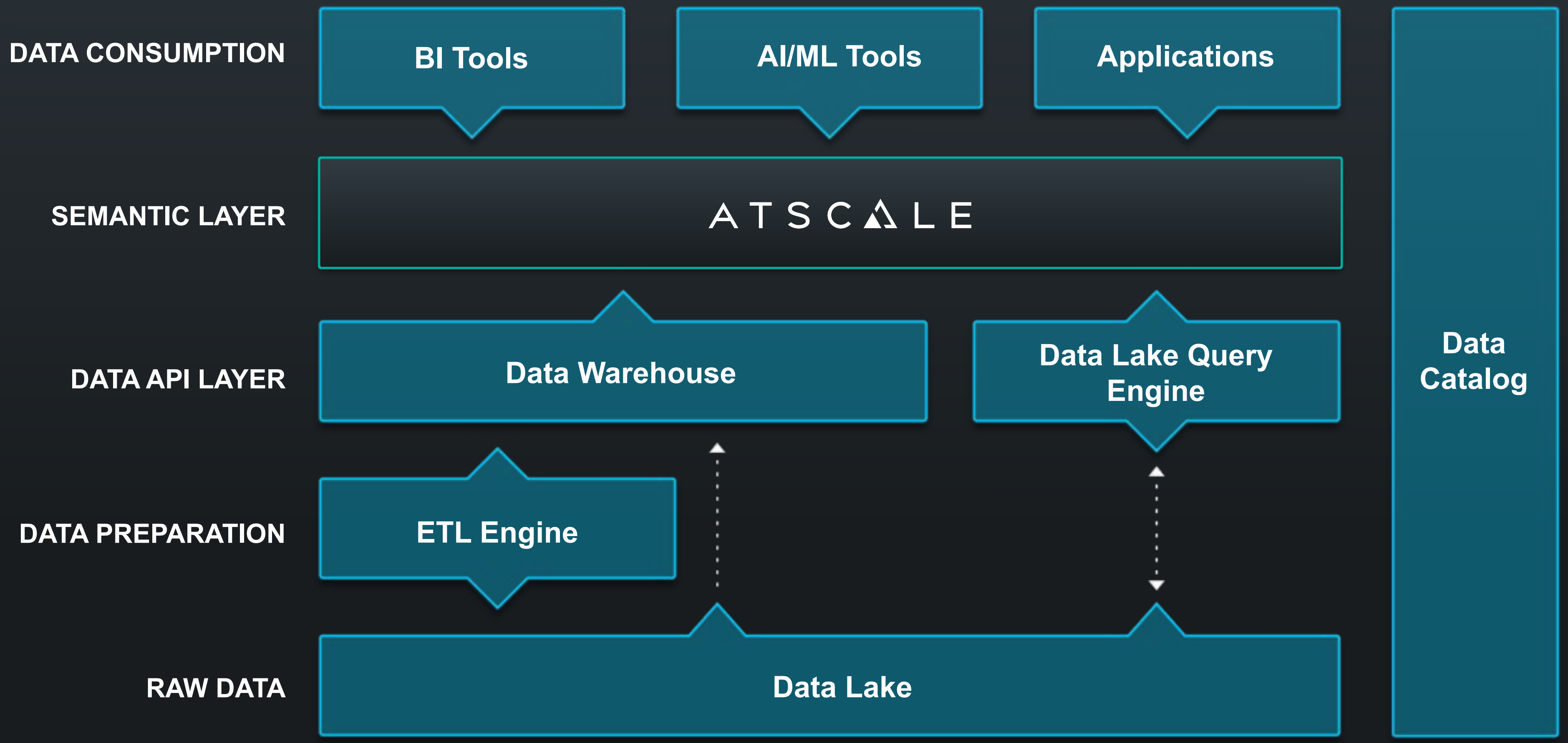
Questions for Jay

- It looks like DataRobot is to machine learning what Tableau was to ad hoc analytics - meaning, you've made machine learning more approachable to a much wider audience within the enterprise just like Tableau did for BI. For your customers, how do you see the predictions coming out of DataRobot making their way into decision making by business users?

Data & Analytics Flywheel



AtScale: *Where we fit.*



ATSCALE

www.atscale.com

400 S El Camino Real, Ste 800, San Mateo, CA 94402