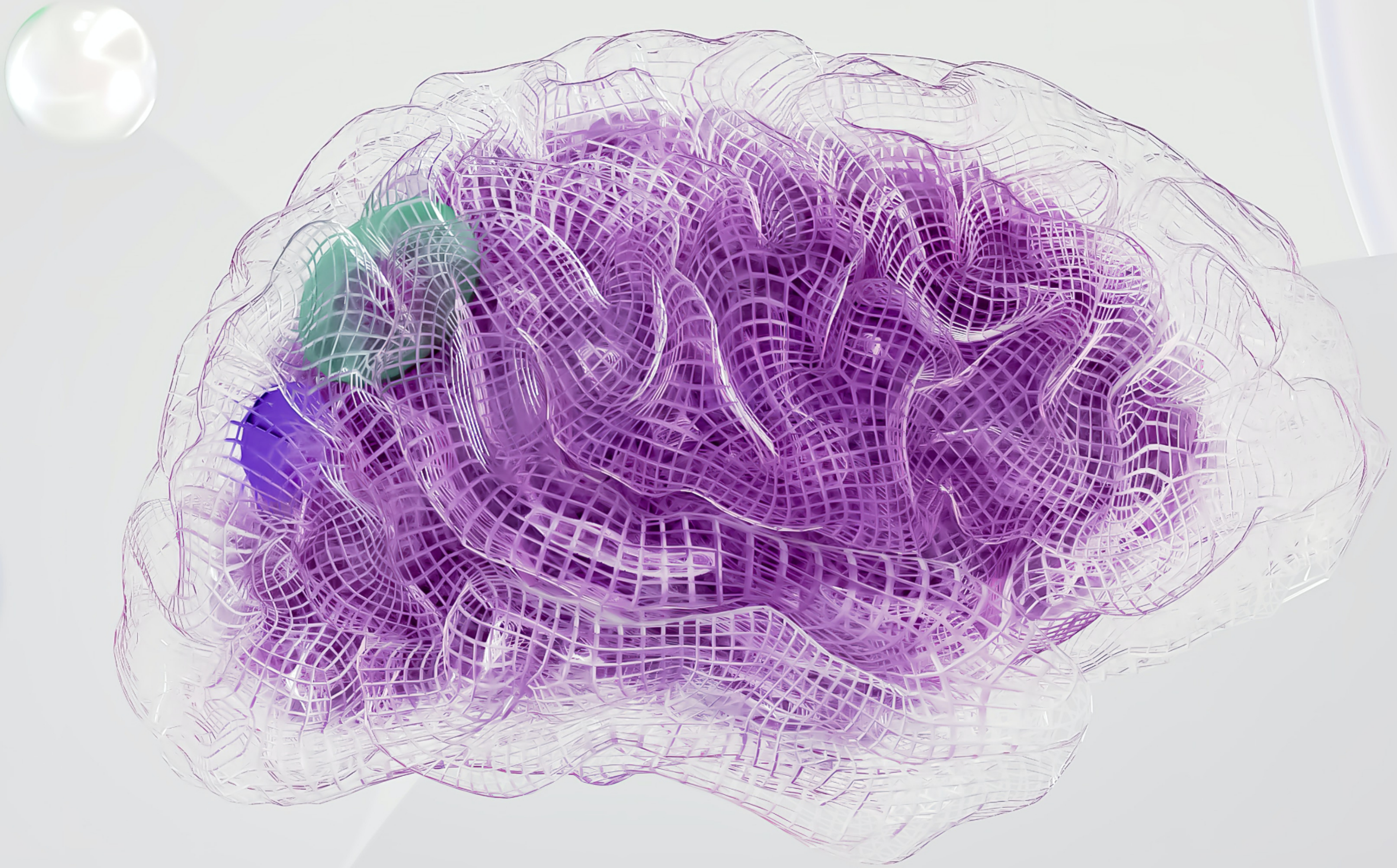
censius Data Council

Evolving AI Laws and the Imperative to Build Safe, Compliant, and Risk-proof AI



Ayush Patel

Founder @ CensiusAI

About Me

- I'm the co-founder of Twelvefold (https://twelvefold.ai), an AI start-up studio, where I manage a portfolio of MLOps and Generative AI companies with entrepreneurs. I'm also the CEO of Censius, an AI Observability platform that helps to optimize AI models' real-world performance.

- I have closely worked with customers across industry verticals, AI teams, and research projects to build reliable and compliant AI solutions to solve everyday business problems and scale models at production

- Connect with me on Twitter and LinkedIn @ayushpatelxyz

Agenda

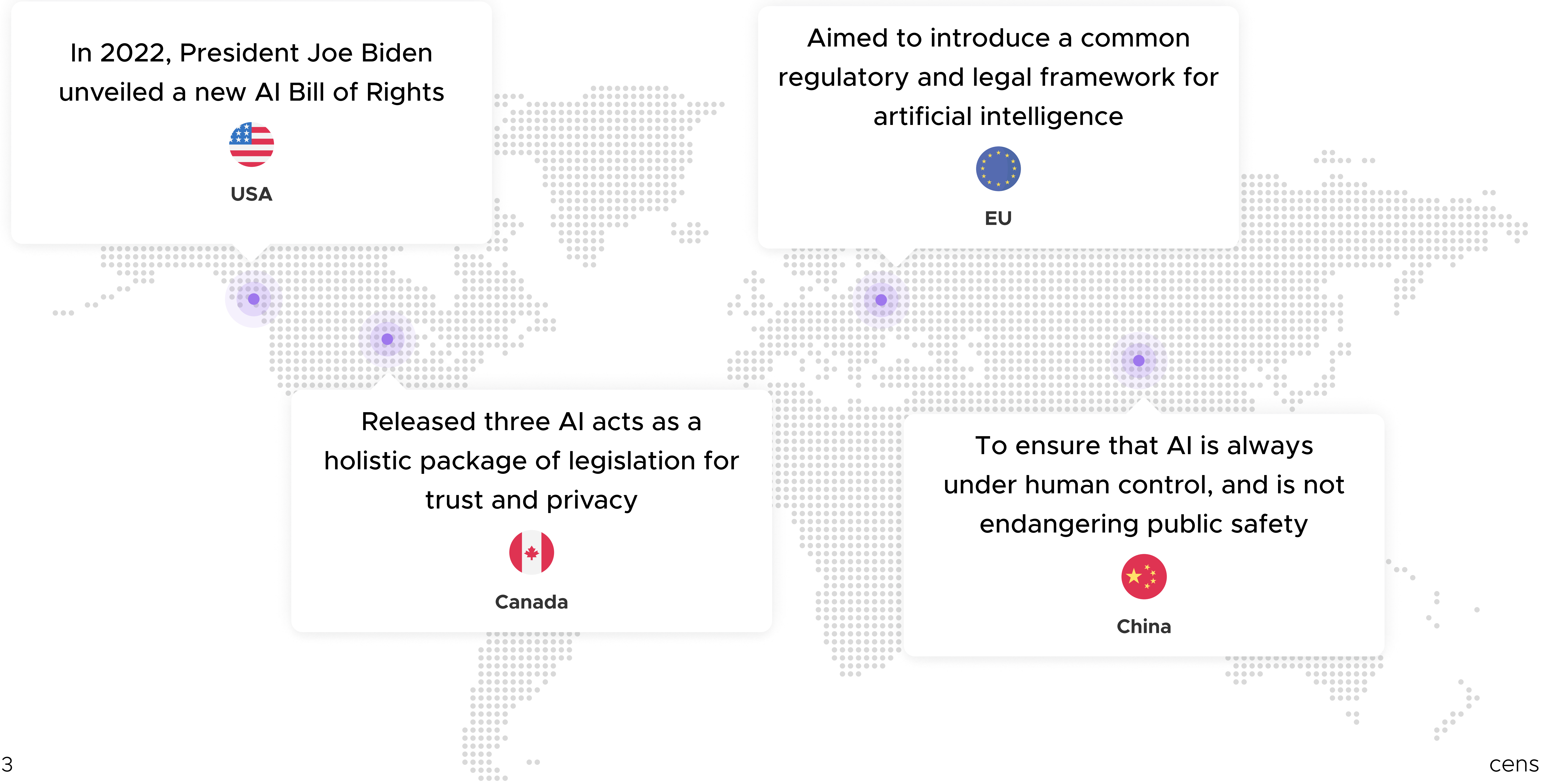
01 What's in the Blueprint of USA's AI Bill of Rights and its implications

02 Key issues in the opacity of ML models

03 Strategies to stay on par with the evolving AI laws and establish Responsible AI (RAI)

04 About Censius

Evolution of AI regulations at a global scale



In 2022, President Joe Biden unveiled a new AI Bill of Rights

USA

Aimed to introduce a common regulatory and legal framework for artificial intelligence

EU

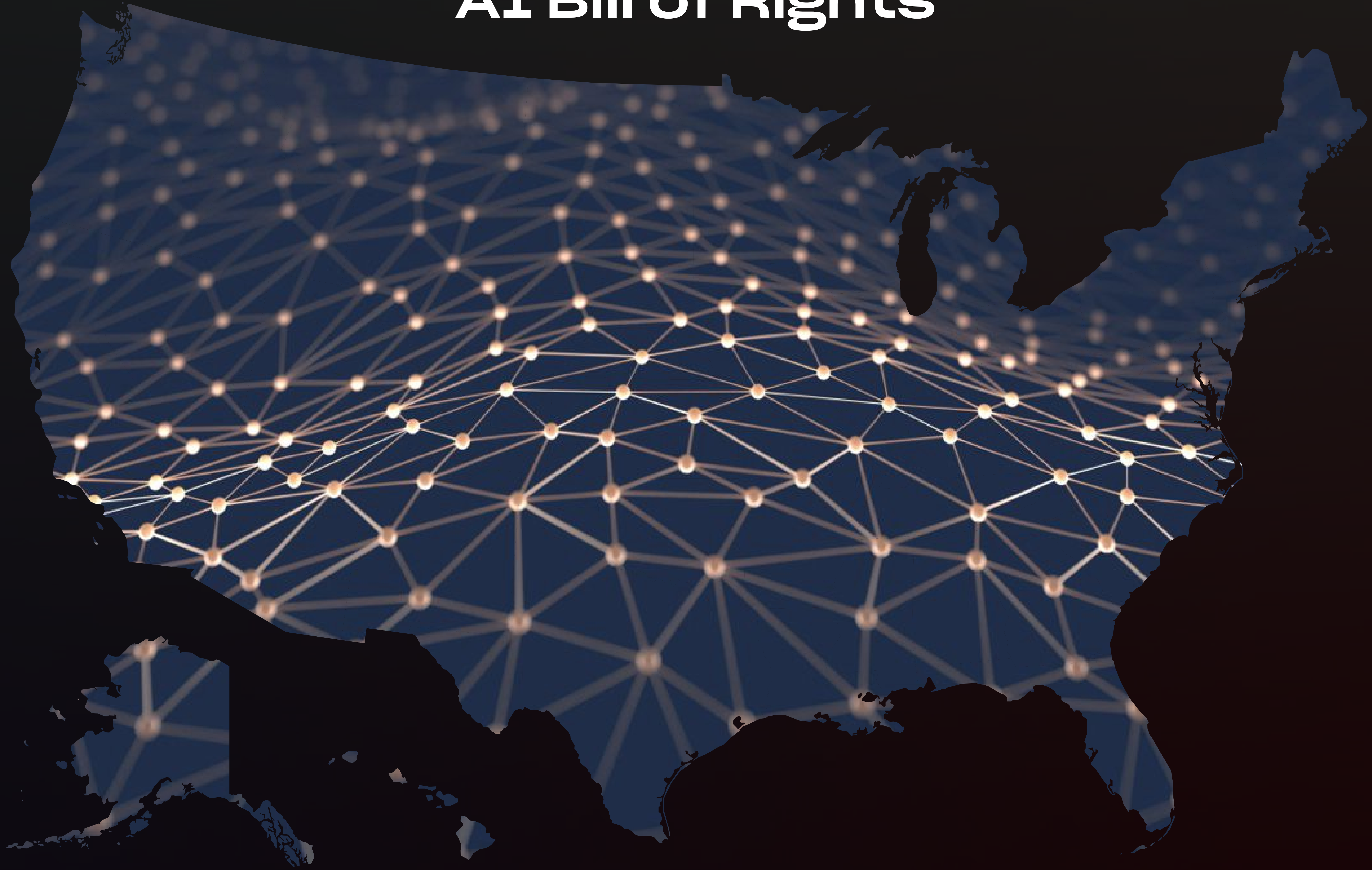
Released three AI acts as a holistic package of legislation for trust and privacy

Canada

To ensure that AI is always under human control, and is not endangering public safety

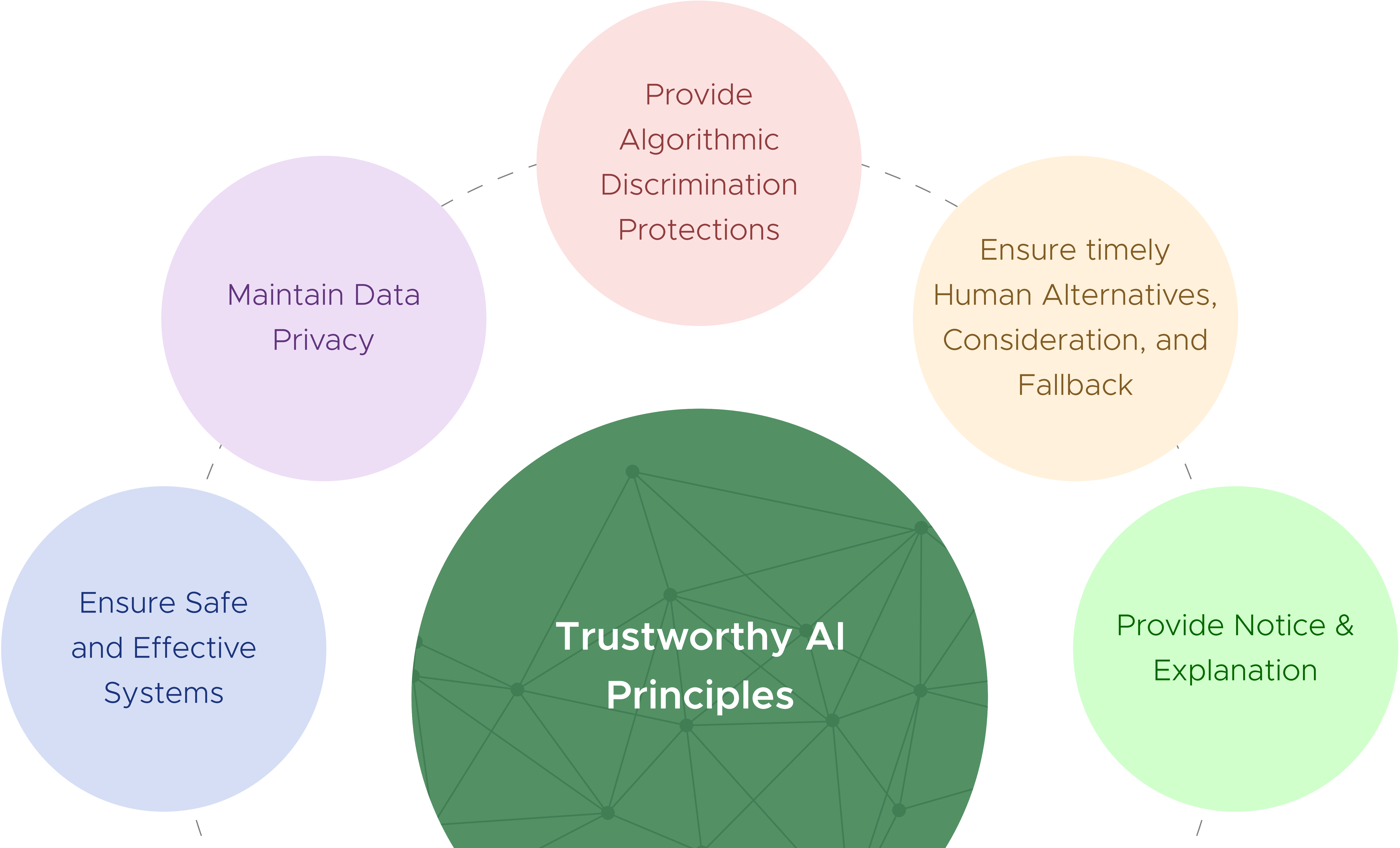
China

Under the lens: The Blueprint of USA's AI Bill of Rights



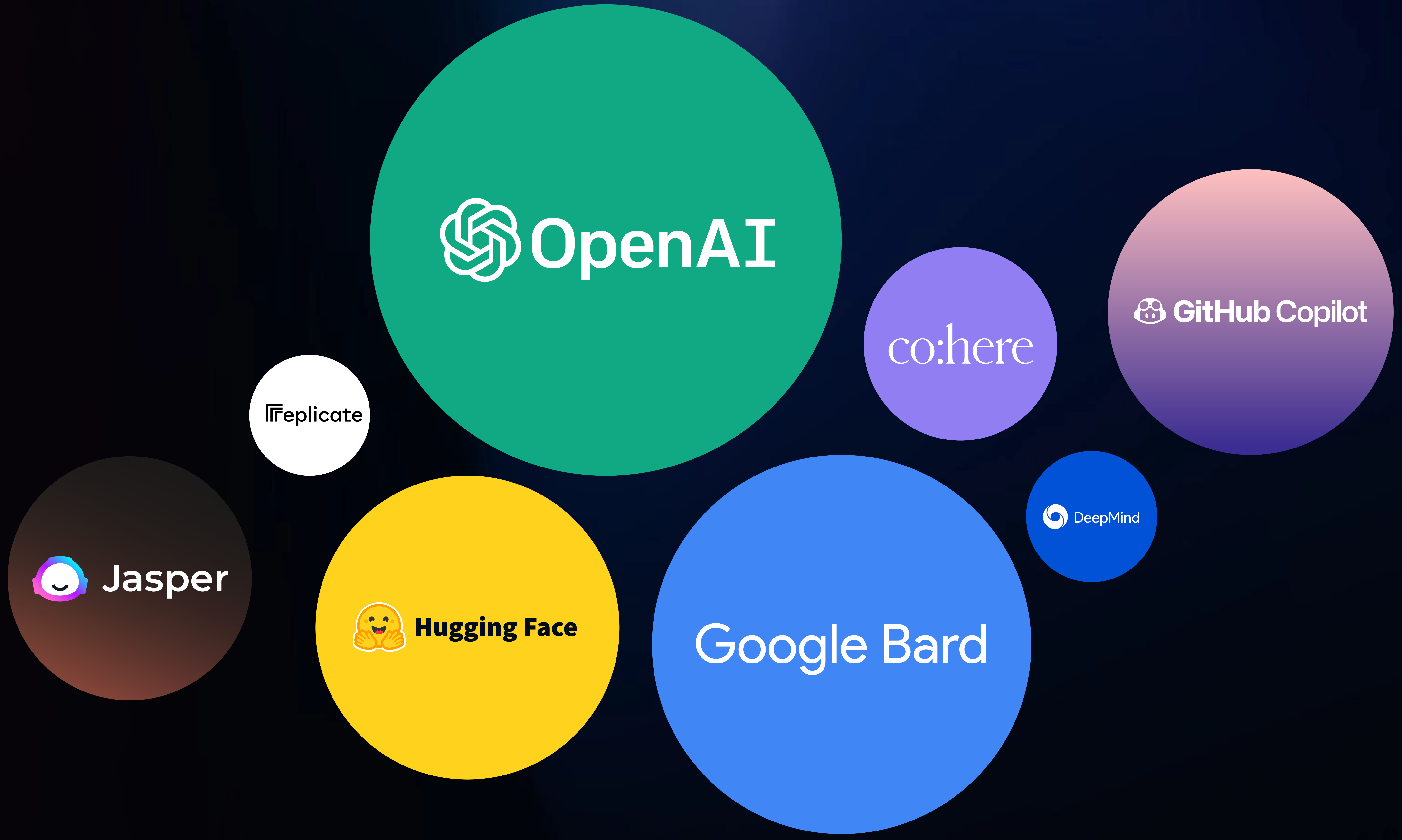
Overview of The Blueprint of USA's AI Bill of Rights

Covers five guiding principles for automated systems to meaningfully impact the public's rights, opportunities, or access to critical needs; is currently a non-binding set of guidelines


Provide Algorithmic Discrimination Protections
Maintain Data Privacy
Ensure timely Human Alternatives, Consideration, and Fallback
Ensure Safe and Effective Systems
Trustworthy AI Principles
Provide Notice & Explanation

Why now?
Case in point: Boom of Generative AI and LLMs

The exponential growth of LLMs and Generative AI tools comes with a cost



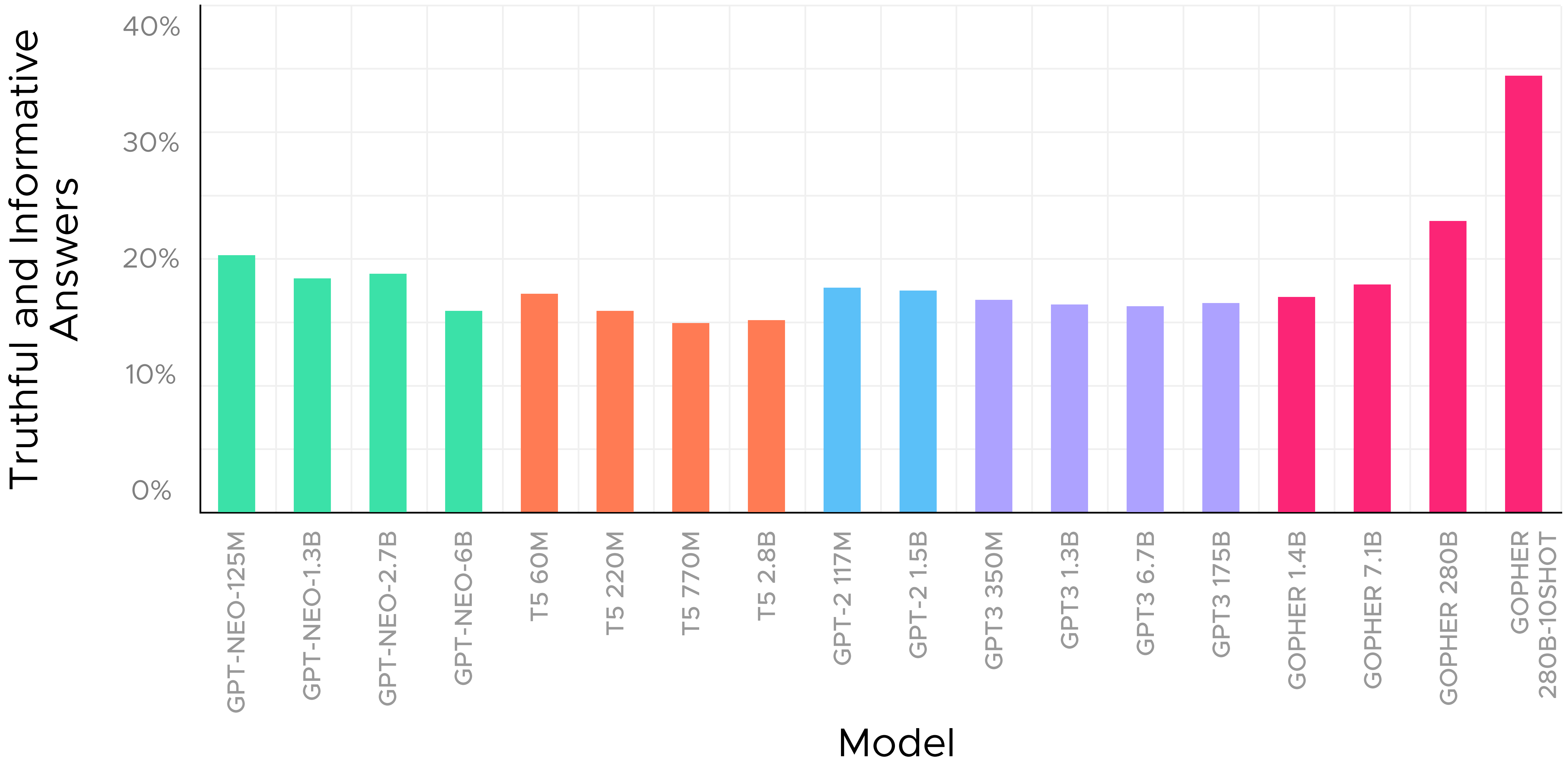
Why now?
Case in point: Boom of Generative AI and LLMs

Risk of inaccurate of fabricated details

On average, most generative models are truthful only **25%** of the time, according to the TruthfulQA benchmark test

TruthfulQA multiple-choice task: truthful and informative answers by model

(Source: Stanford University Artificial Intelligence Index Report 2022)



Why now?
Case in point: Boom of Generative AI and LLMs

Ethical concerns

Last year, OpenAI's DALL-E was re-launched to fix some issues related to gender and racial bias

Increased risk of generating false, biased, and unrealistic outputs

55% of technology leaders experienced AI incidents due to biased or incorrect outputs that resulted in financial losses, measurable loss of brand value, and customer attrition

[Source: McKinsey]

How the new AI Bill of Rights is designed to protect end-users

Aims to protect the rights, opportunities, and access of:

CIVIL RIGHTS, CIVIL LIBERTIES, AND PRIVACY

including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy

EQUAL OPPORTUNITIES

including equitable access to education, housing, credit, employment, and other programs

ACCESS TO CRITICAL RESOURCES OR SERVICES

Such as healthcare, financial services, safety, social services, non-deceptive information about goods and services, and government benefits

Key challenges in the opacity of ML models

Garbage in, garbage out

Model bias

Models are prone to drift

Zero traceability

Garbage in, garbage out

Data can come with a lot of baggage such as:

- Stale, inaccurate, incomplete, or incorrect data
- Sampling error
- Model drifts
- Limited visibility into model performance health
- Legacy software dependency leading to siloed collaboration and delayed error resolution

IMPLICATION

- Faulty input data can lead to incorrect output, driving misinformed and possibly harmful decisions
- The more data you receive, the more it costs to manage and clean it

Model bias

Data bias

Historical bias, sampling bias, prejudicial bias, labeling bias

Human bias

Cognitive error. information processing, preconception, model mechan

Algorithmic bias

Training data, focus, processing, context transfer, intepretation

IMPLICATION

- Revenue loss
- Lost customer trust
- Negative publicity
- Risk of gender prejudice, racial bias, age discrimination, and recruiting inequality

Models are prone to drift

Data drift

A shift observed due to changes in the statistical properties of the independent variables, such as feature distributions

Concept drift

The law underlying the data changes, assumptions made by the model on past data need to be revised based on current data

Upstream data changes

These are operational data changes in data pipelines like changes in measurement systems such as miles to kilometers

AI IN THE REAL-WORLD

Spam detection model

A model trained in 2020 to classify spam emails might underperform in 2022 as spammers also upgrade day by day. But according to IBM, **68%** of organizations are not tracking performance variations and model drift.

Zero traceability

Challenges in the absence of model traceability

- No visibility into the model’s performance across datasets
- No way to perform root cause analysis causing performance degradation
- No way to pinpoint specific feature values or cohorts of data where the model is performing poorly
- Difficulty in explaining black box AI decisions to stakeholders

Framework for assessing AI risk

Outline

your goals and objectives for building Responsible AI

Measure & compare

Perform quantitative analysis of outcomes and disparities across different user segments

Optimize

Model to improve performance and establish fairness with proactive re-training

Monitor & build explainable systems

to flag and proactively resolve AI issues

Strategies to stay on par with evolving AI regulations

Strategy for building Responsible AI (RAI)

Implement data governance

Lean towards interpretability and explainability tools

Maintain human oversight

Ensure provision of information to users

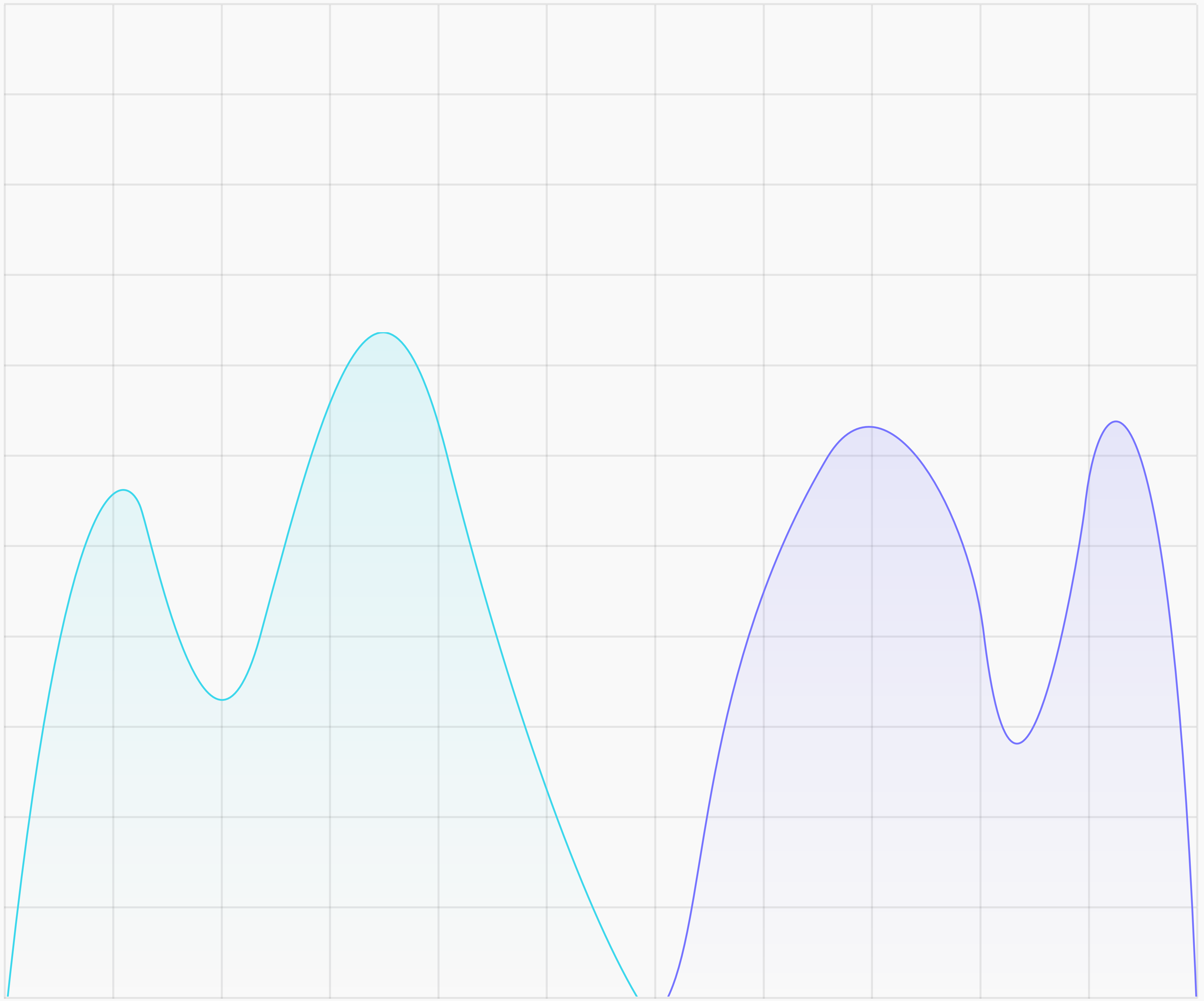
Maintain accuracy, robustness, and security

Optimize documentation and reporting activities

Implement data governance

Utilize and enable granular level monitoring of data segments

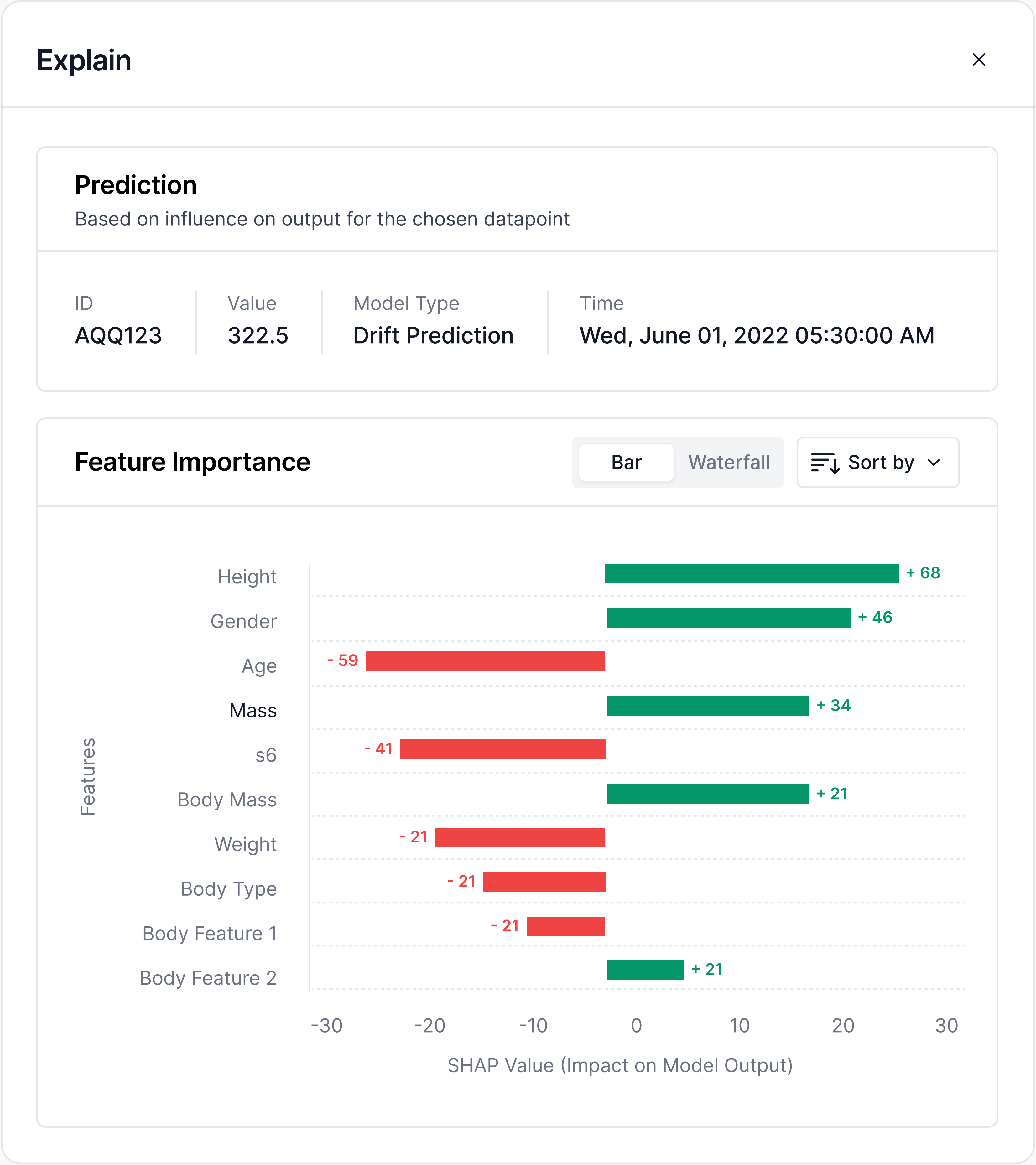
- Consistently log and supervise model inputs and outputs to detect performance dips
- Ensure models are running correctly, even without the ground truth
- Detect and flag specific data segments where the performance dipped below the threshold value
- Leverage monitoring tools to enable real-time alerts for violations



Lean towards AI explainability

Identify blind spots in your model's decision-making for proactive correction by training your models

- Explain Black Box AI decisions and perform root cause analysis
- Improve performance for specific cohorts
- Visualize and observe changing trends or anomalous patterns in feature distributions



Explain ×

Prediction

Based on influence on output for the chosen datapoint

ID	Value	Model Type	Time
AQQ123	322.5	Drift Prediction	Wed, June 01, 2022 05:30:00 AM

Feature Importance

Maintain human oversight

Build and maintain a button that stops operations if overseer detects risk

Monitor for signs of anomalies, dysfunctions, & unexpected performance

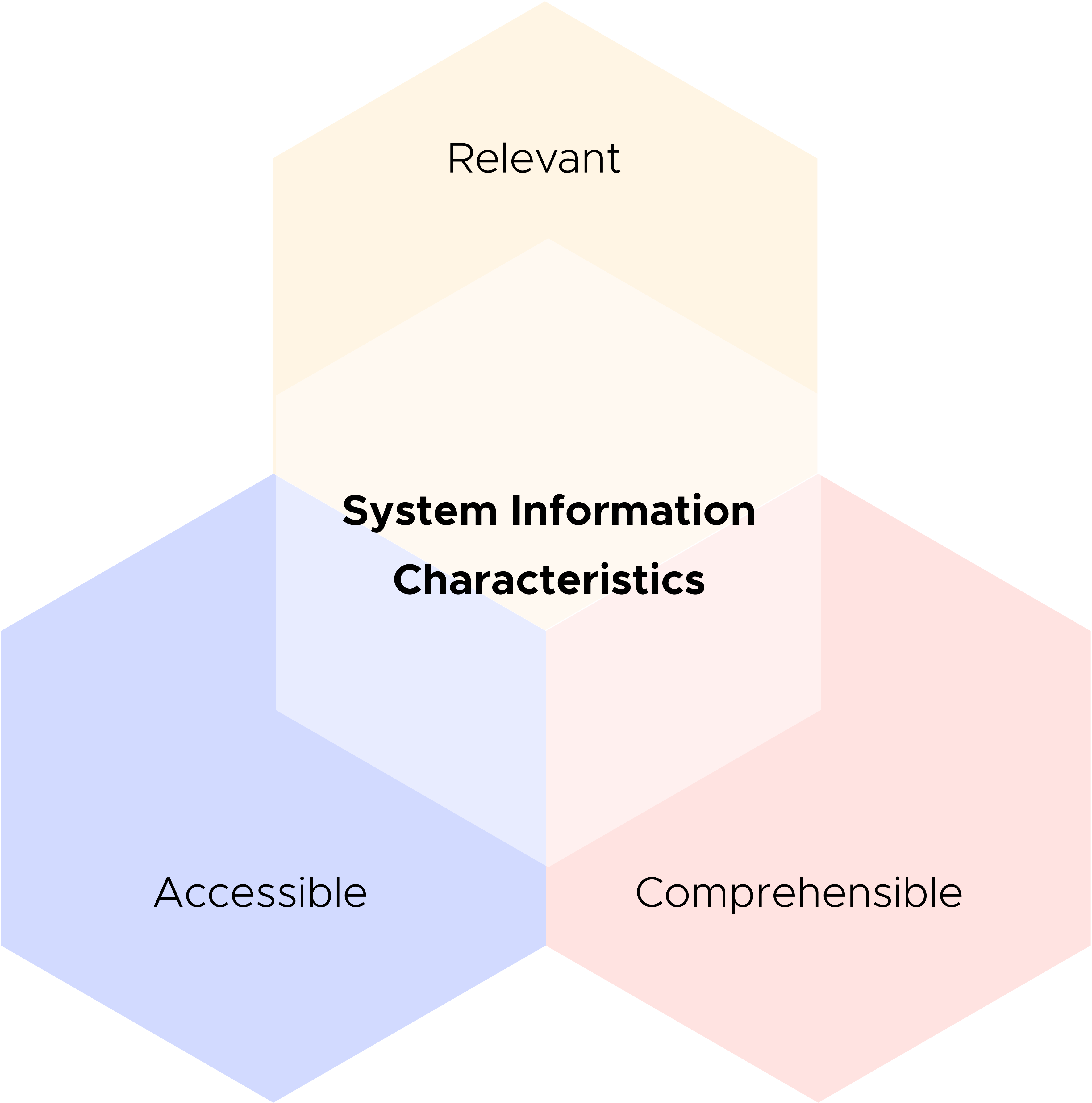
Adopt Explainability tools that offer root cause analysis widgets and provides reports for stakeholder alignment

Ensure provision of information to users

The system should be accompanied by Instructions for use that follow the

4Cs

CONCISE, COMPLETE, CORRECT, CLEAR



Relevant

System Information Characteristics

Accessible

Comprehensible

Maintain security and compliance

- Spot and fix suspicious model patterns with root cause analysis
- Establish robust compliance standards to stay on par with new AI regulations
- Ensure secure flow and transition of data across parties involved, such as developers, overseers, stakeholders, and end-users

Optimize documentation and reporting activities

Dedicated Resources

Assign dedicated resources or experts to maintain documentation

Automated Logging Capabilities

Leverage automated logging capabilities to log metadata from every endpoint of your ML cycle

Stay on top of model health reports

Leverage monitoring and explainability tools to get continuous reports on model health

It's a wrap.
Checklist to ensure adherence to new AI laws

Logging

Bird's eye view of the model

↓

Documentation and reporting

Performance Analysis

Is something going wrong?

↓

Monitor

Root cause analysis

What is going wrong?

↓

Explain

Troubleshoot

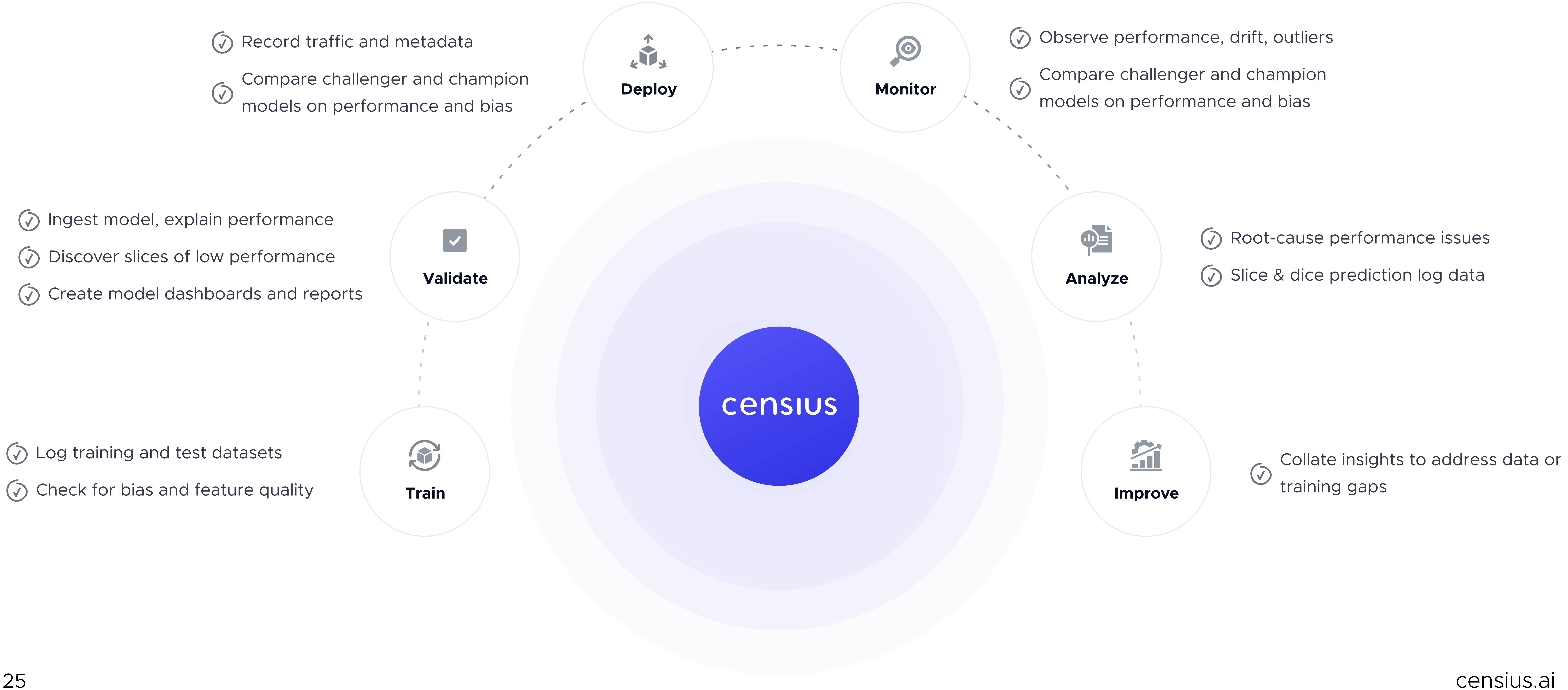
Why it went wrong?

↓

Debug

About Censius

A single platform for delivering enterprise level observability at scale.



Train

- Log training and test datasets
- Check for bias and feature quality

Validate

- Ingest model, explain performance
- Discover slices of low performance
- Create model dashboards and reports

Deploy

- Record traffic and metadata
- Compare challenger and champion models on performance and bias

Monitor

- Observe performance, drift, outliers
- Compare challenger and champion models on performance and bias

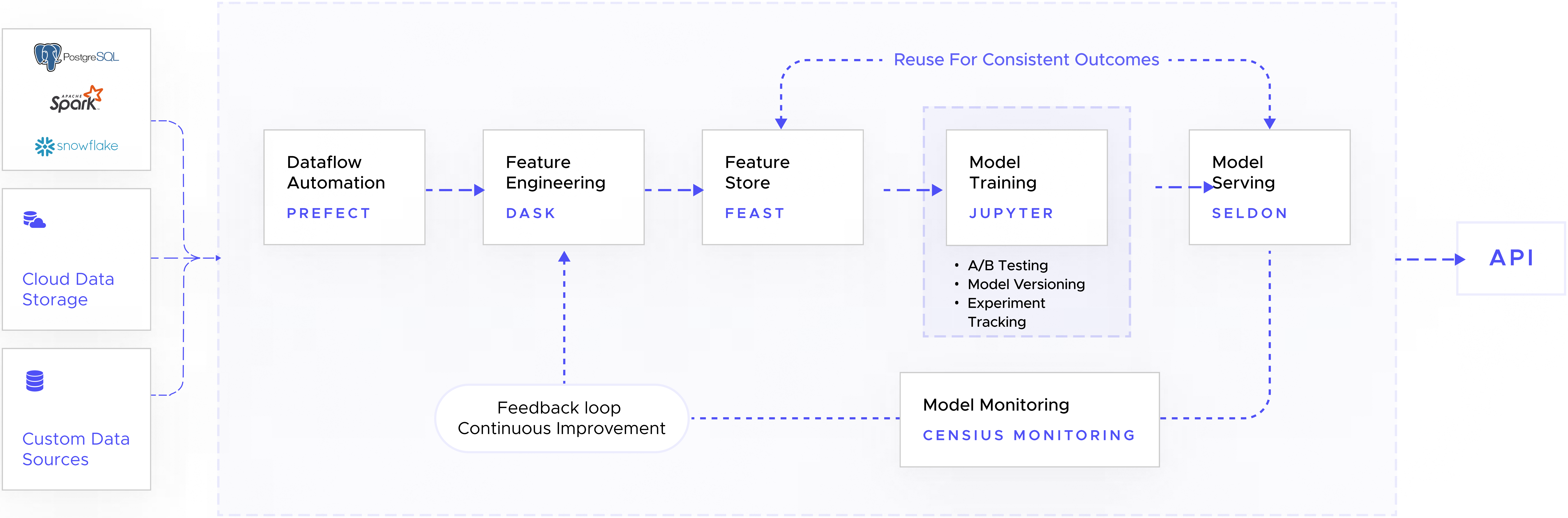
Analyze

- Root-cause performance issues
- Slice & dice prediction log data

Improve

- Collate insights to address data or training gaps

Censius Tech Stack



PostgreSQL

Apache Spark

snowflake

Cloud Data Storage

Custom Data Sources

Dataflow Automation

PREFECT

Feature Engineering

DASK

Feature Store

FEAST

Model Training

JUPYTER

- A/B Testing
- Model Versioning
- Experiment Tracking

Model Serving

SELDON

Reuse For Consistent Outcomes

API

Feedback loop
Continuous Improvement

Model Monitoring

CENSIUS MONITORING

Thanks

Get started with
Census AI Observability

censius.ai

Learn more about our
AI Startup Studio

twelvefold.ai

Connect with me

ayush@twelvefold.ai