The Boston Area SAS® Users Group

Gathering SAS Users Since 1983

Trustworthy AI

Sterlina Smith is a Senior Manager in the SAS Data Ethics Practice (DEP), a cornerstone of the company's Responsible Innovation efforts. Prior to joining the DEP, she led the SAS Corporate Social Innovation and Brand division's data for good programs and has served in numerous other roles over the course of her 22 years at SAS. She is also a licensed attorney in the state of North Carolina.

The DEP works internally and externally to guide SAS and its customers in developing and deploying technologies that promote human well-being, agency and equity; establish SAS as a leading contributor to the growth of responsible innovation in AI and analytics; ensure SAS' approach to responsible innovation is coordinated globally; and develop best practices to quickly respond to regulatory changes.

A Comprehensive Approach to Trustworthy AI

Sterlina Smith

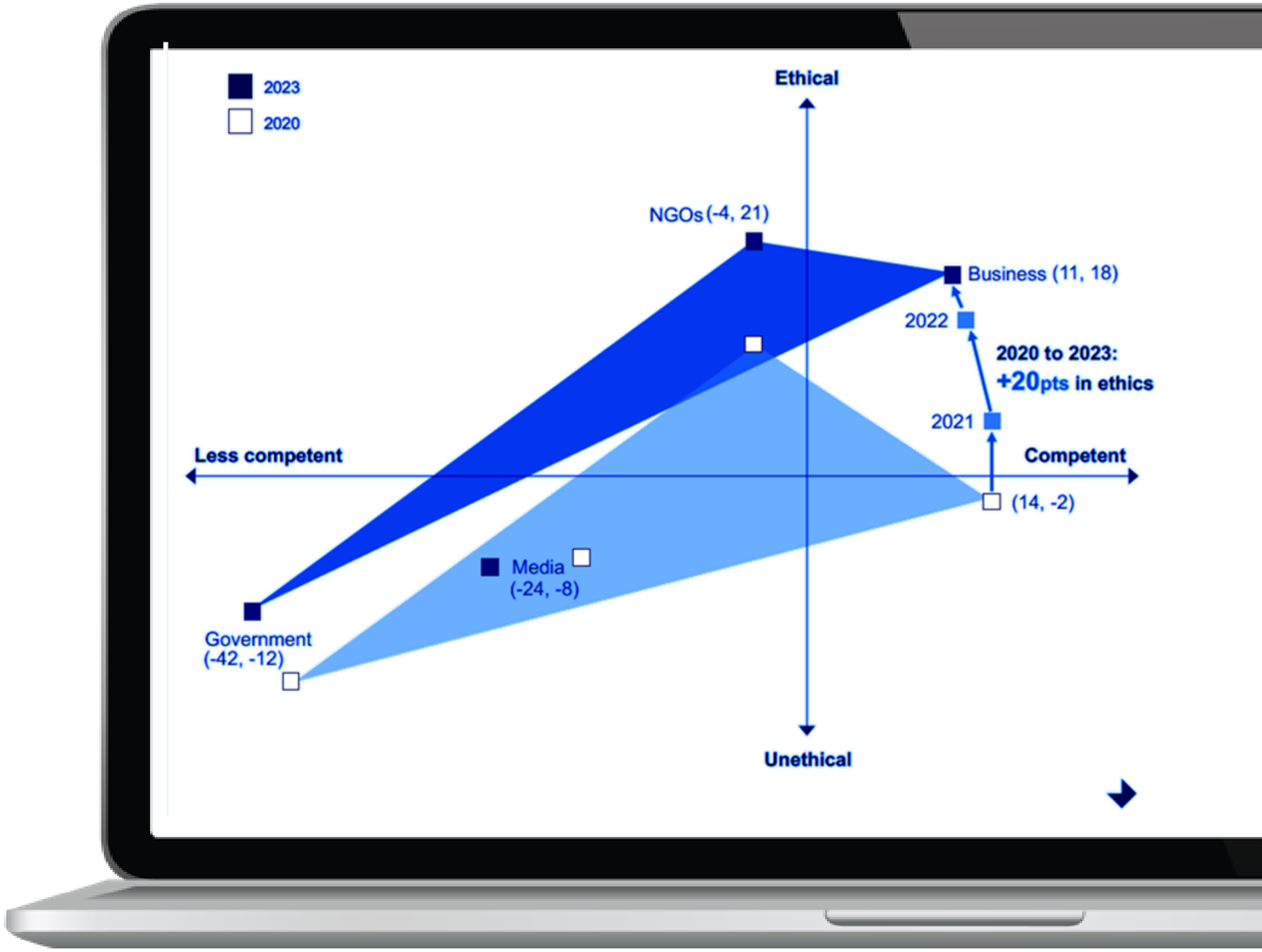
TRUST

2023 Edelman Trust Barometer

Only **Business** is Competent and Ethical, Sustains Rise in Ethics for Third Year

(Competence score, net ethical score)

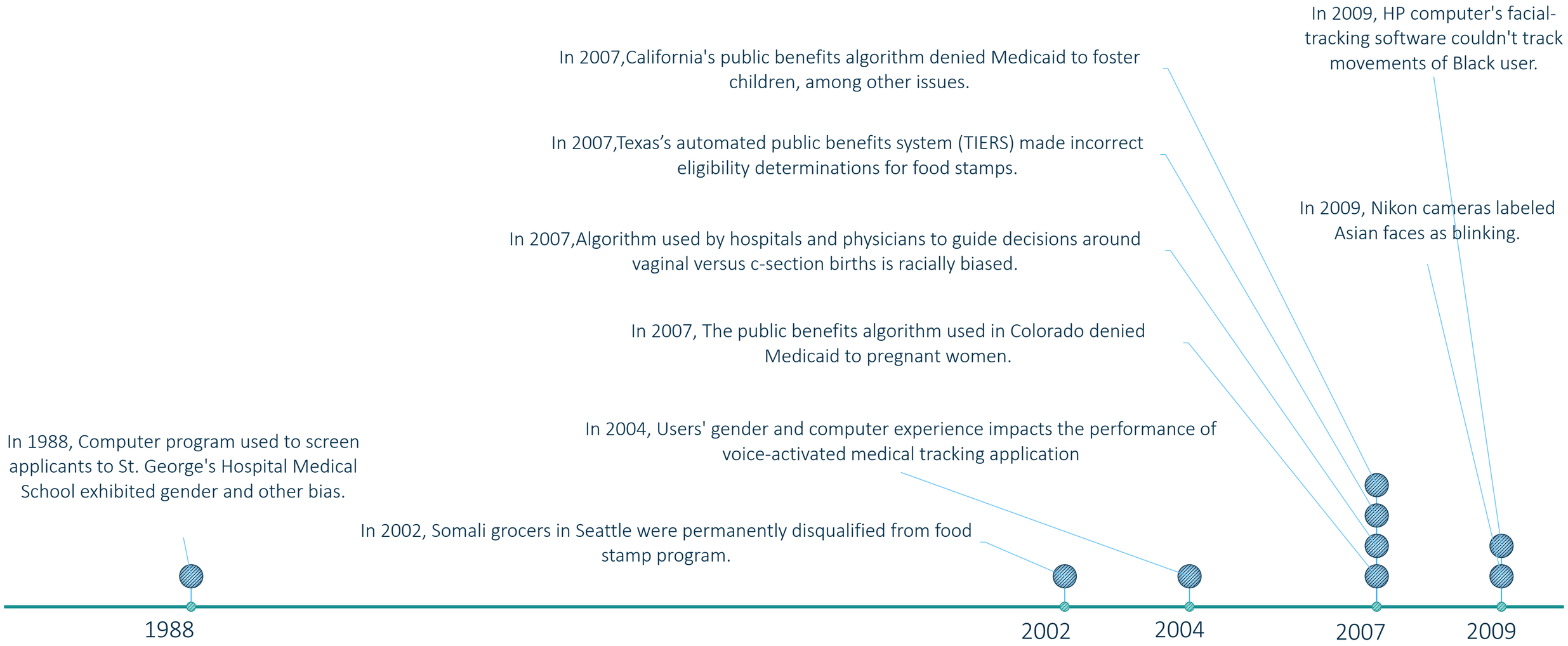
GLOBAL 24 Excludes China and Thailand



2023 Edelman Trust Barometer. The ethical scores are averages of nets based on [INS]_PER_DIM/1-4. Government and Media were only asked of half of the sample. The competence score is a net based on TRU_3D_[INS]/1. Government and Media were only asked of half of the sample. General population, 24-mkt avg. Data not collected in China and Thailand. For full details regarding how this data was calculated and plotted, please see the Technical Appendix.

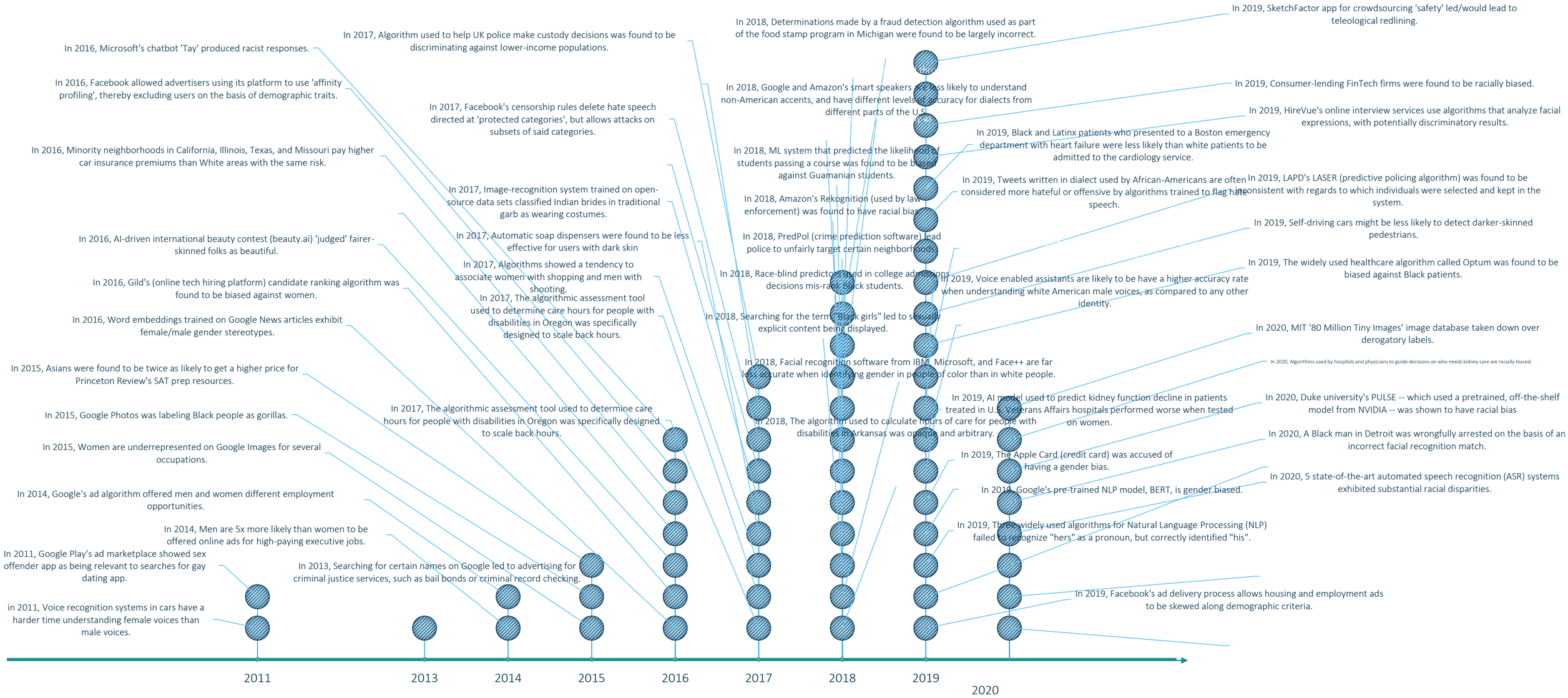
Why might people not trust AI?

Unintended Harms Occur…



BERKELEY HAAS CENTER FOR EQUITY, GENDER & LEADERSHIP Examples of Bias in Artificial Intelligence

...Now at **MASSIVE** Scale



But AI is everywhere, right?

AI and Advanced Analytics

Analytical modeling is an effective and reliable technique for turning data into information you can use to make decisions.

Financial Services

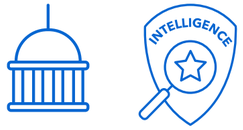
Fraud Detection
Credit Analysis
Automate Financial Advisors

Health and Life Sciences

Predictive Diagnostics
Biomedical Imaging
Health Monitor

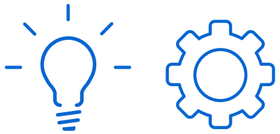
Communications

Conversational Chat Bots
Contextual Marketing
Network Analytics



Government

Smart Cities
Sensor Fusion
Facial Recognition



Manufacturing and Energy

Supply Chain Optimization
Automated Defect Detection
Energy Forecasting

AI Impacts Every Industry

Financial Resources	Insurance & Retail	Criminal Justice	Healthcare & Life Sciences	Human Resources
 Who gets approved for a loan?	Who gets the best prices?	 Who gets a harsher sentence?	 Who gets seen faster at the doctor's office?	 Who gets hired?

AI: Positive or Negative?

AI: Positive or Negative?

Positive things that AI can accomplish:

- Diagnosing diseases
- Finding and fixing mistakes
- Rooting out fraudsters
- Improving agricultural and industrial operational efficiency
- Finding the shortest path, saving time and energy
- Improving motor vehicle safety, saving lives
- Quickly identifying and locating school shooters
- Assessing drug and vaccine efficacy
- Entertainment
- Gaining supply chain efficiency and security
- Reducing emergency responder response times
- Intelligent traffic signals reducing gridlock

But AI can also be associated with negative actions:

- Deception, such as spreading misinformation
- Invasion of privacy
- Fostering device addiction
- Blackmailing or embarrassing people with private information
- Providing biased results
- Lending credibility to bogus and inaccurate results
- Heavy electric power usage
- Use of resources in limited supply (e.g., rare earth elements)
- Creating revenge porn
- Creating fraud
- Lack of transparency
- “Stealing” jobs from humans

Bias Examples

USER INTERFACE

DISCLOSE DATA
A person, process, or system creates, and publishes/ shares data

ACQUIRE
Ingest data from sensors, systems, or humans, recording its provenance and consent for use wherever possible

STORE
Record data to a trusted location that is both secure and easily accessible for further manipulation

STAGE AND PREPARE DATA
A person, process, or system transforms, moves, or analyzes data.

PRE- PROCESS
Combine disparate datasets to create a larger dataset that is greater than the sum of its parts

MODEL BUILD
Examine and transform data with the purpose of extracting information and discovering new insights

CONSUME DATA
A person, process, or system benefits from manipulated data

MODEL DEPLOY
Apply the insights gained from data analysis towards making decisions, affecting change, or delivering a product or service

SHARE/ SELL
Provide Access to datasets or data insights to new sets of data manipulators or consumers

DISPOSE
Remove data from servers to prevent future release or use

Availability Bias
Recall Bias

Exclusion Bias
Pre-processing Bias
Measurement Bias
Time-interval bias
Historical Bias

Sample Selection:
Selection Bias
Attrition Bias

Confirmation Bias
Cause/ Effect Bias
Confounding Bias
Collider Bias
Prediction Bias
Performance Bias
Hindsight Bias
Chronological Bias
Funding Bias
Proxy Bias
Aggregation Bias
Survivorship Bias

Automation bias
Deployment Bias
Drift Bias

Reporting Bias

COMMON BIASES

sas

Exploring Bias:
Facial Recognition Example

If your training data set looks like this…

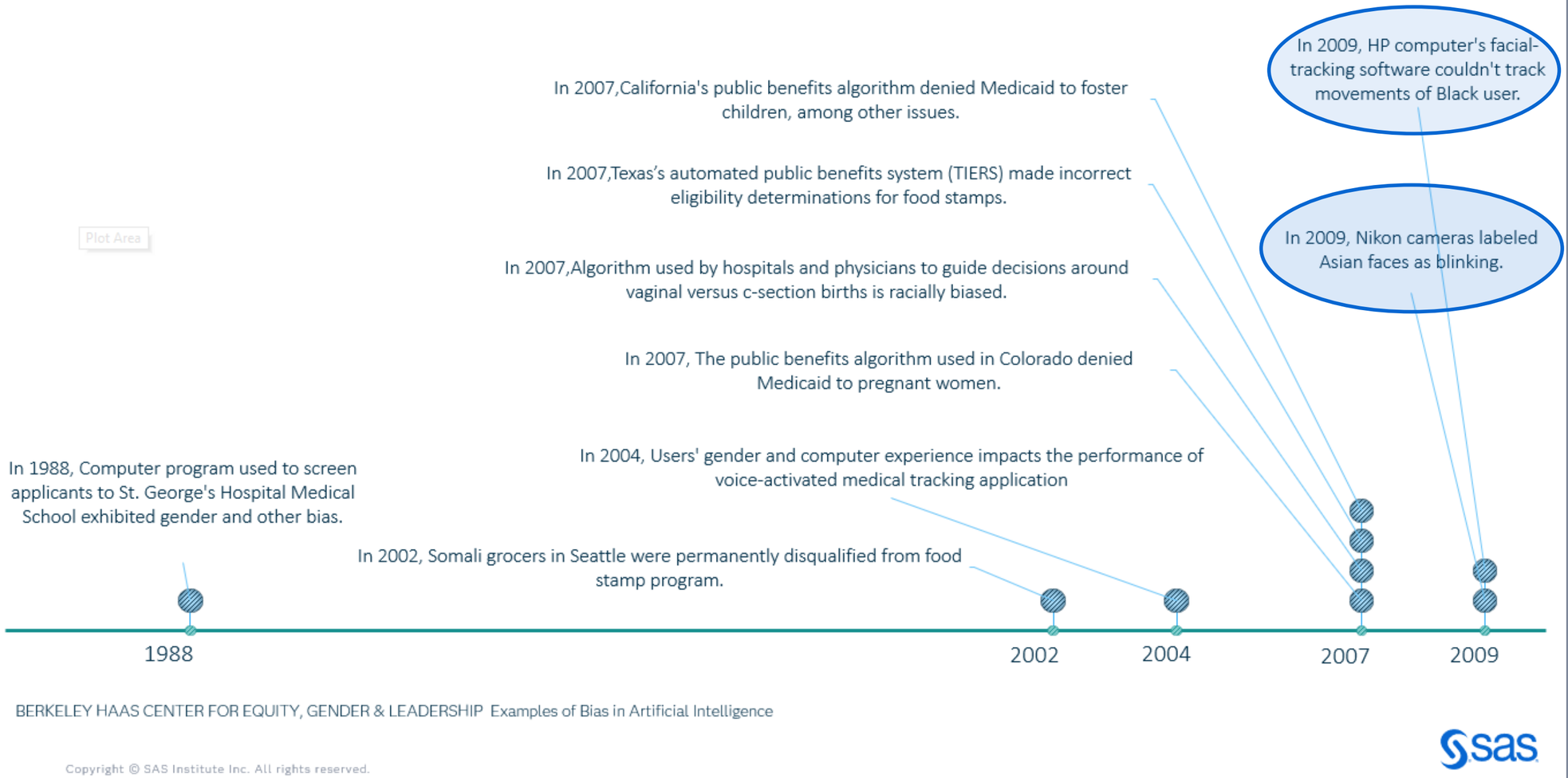
It won’t extrapolate to a population that includes

or

or

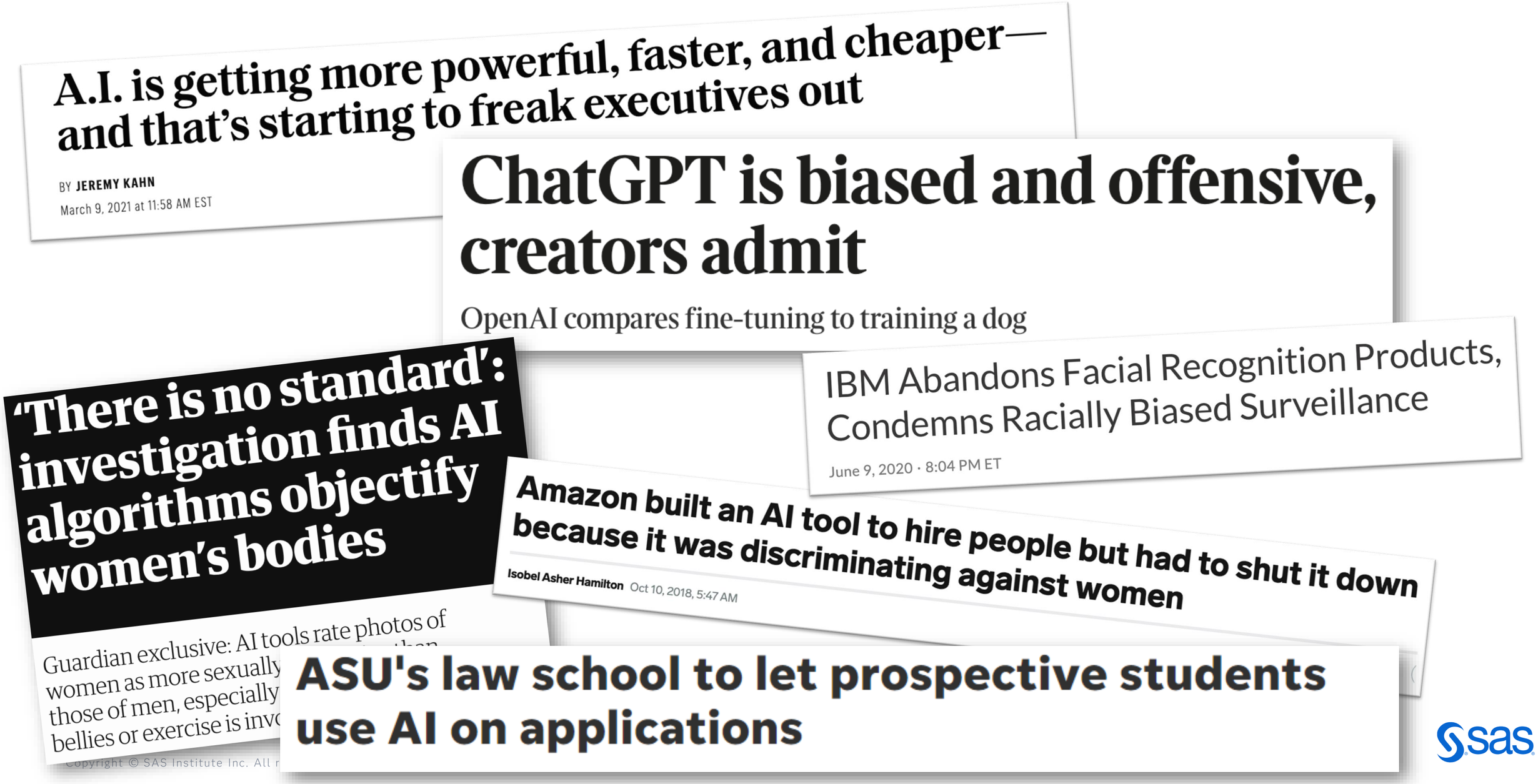
… and it’s not just theoretical

Unintended Harms Occur…



BERKELEY HAAS CENTER FOR EQUITY, GENDER & LEADERSHIP Examples of Bias in Artificial Intelligence

AI in the News



A.I. is getting more powerful, faster, and cheaper—and that's starting to freak executives out

BY JEREMY KAHN

March 9, 2021 at 11:58 AM EST

ChatGPT is biased and offensive, creators admit

OpenAI compares fine-tuning to training a dog

'There is no standard': investigation finds AI algorithms objectify women's bodies

Guardian exclusive: AI tools rate photos of women as more sexually ... those of men, especially ... bellies or exercise is inv...

IBM Abandons Facial Recognition Products, Condemns Racially Biased Surveillance

June 9, 2020 · 8:04 PM ET

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct 10, 2018, 5:47 AM

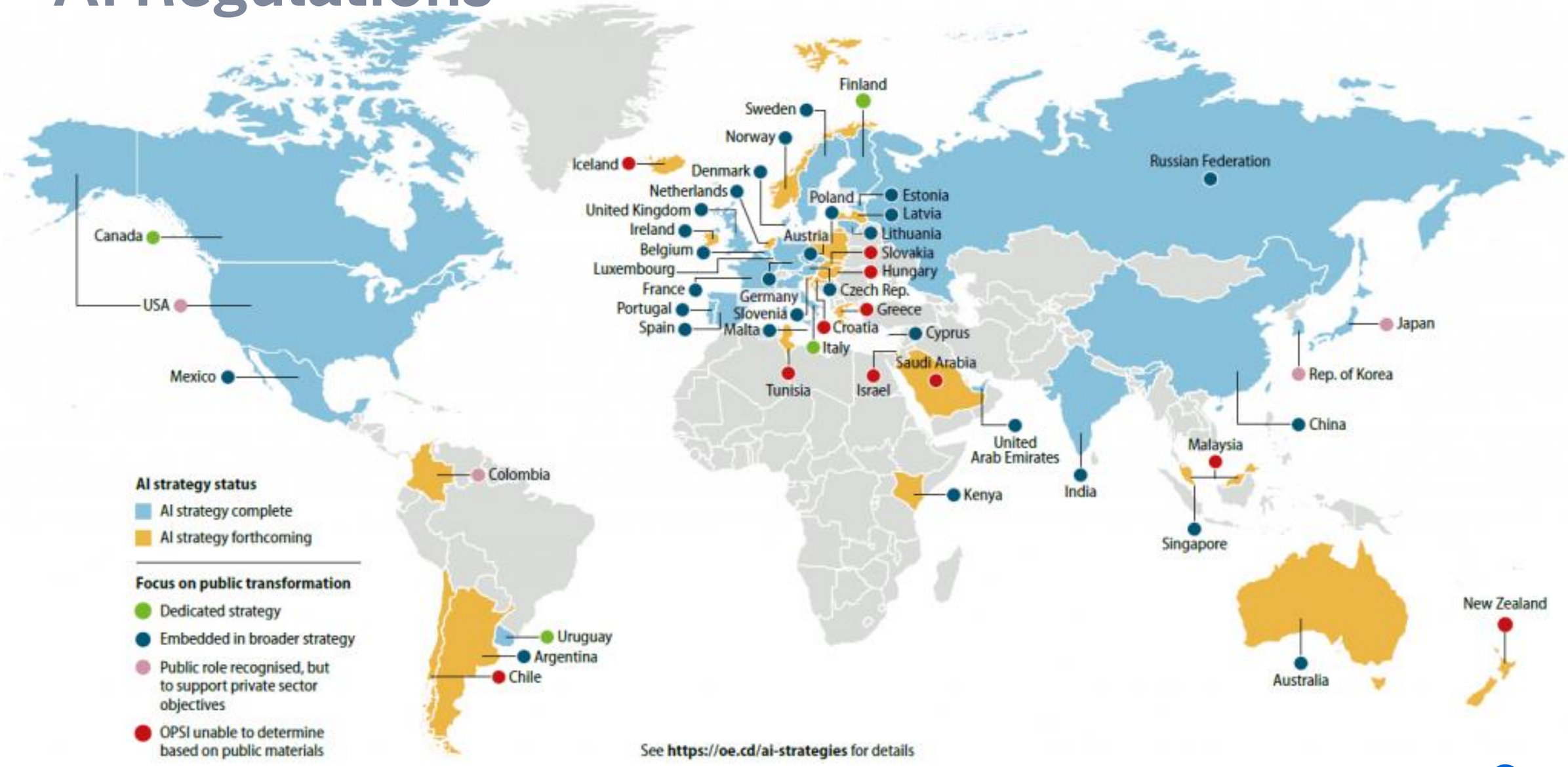
ASU's law school to let prospective students use AI on applications

AI in the Future

- Personal assistants
- Self-driving cars
- Healthcare diagnostics and treatment
- Business process optimization
- Cybersecurity threat detection
- Customized education and tutoring
- Immersive responsive entertainment
- Supply chain and factory automation
- AI writing and speech enhancement
- Art/music/film generation
- AR/VR experiences
- Autonomous drones
- Smart fitness tracking clothing
- Robotic companions
- Fake media identification
- Etc

65% of today's students will be employed in jobs that don't yet exist.

AI Regulations



Why Should I Care?

Moral Imperative

Reputational and Financial Risk

Compliance

Who has heard of Trustworthy AI?

What is Trustworthy AI?

Ethics by Design

Ethical AI

Trustworthy AI

AI for Good

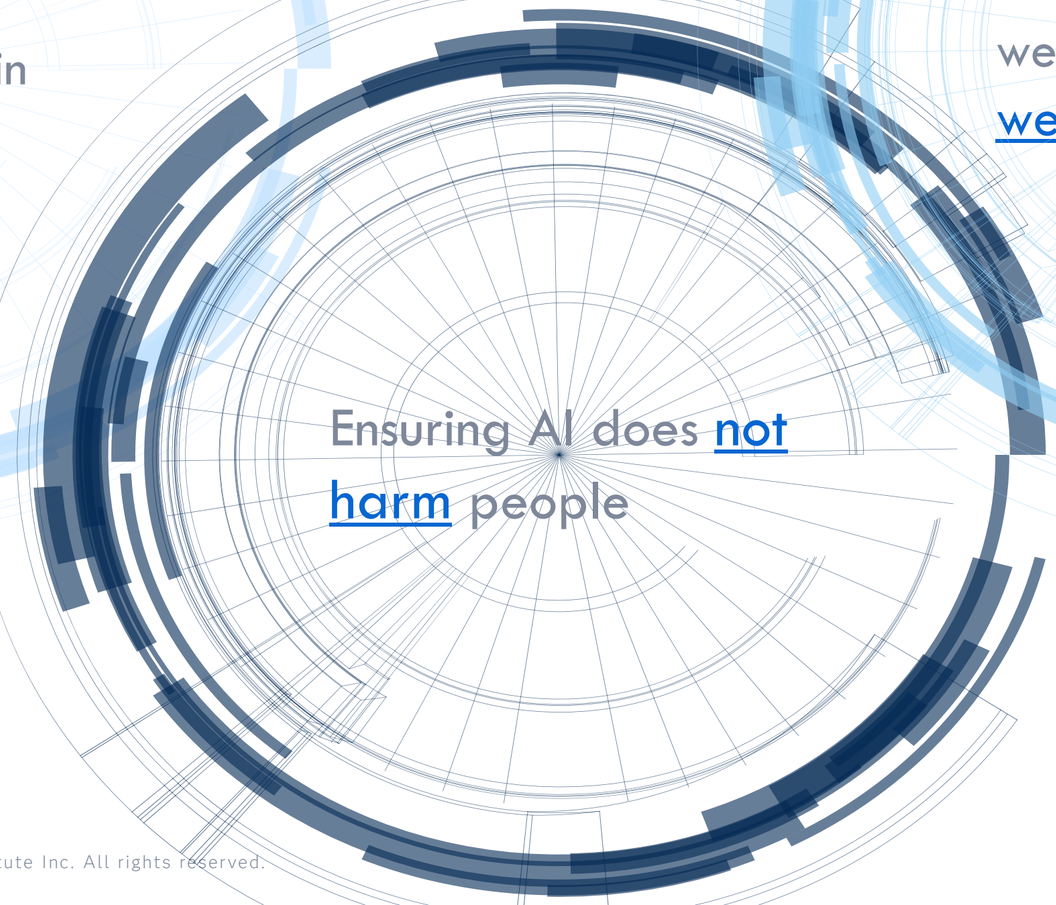
Responsible AI

AI Ethics

What is Trustworthy AI?

Developing and using AI technologies in an ethical manner

Asking not just, “Could we?” but also, “Should we?”



Ensuring AI does not harm people

Building AI that reflects our values as a society

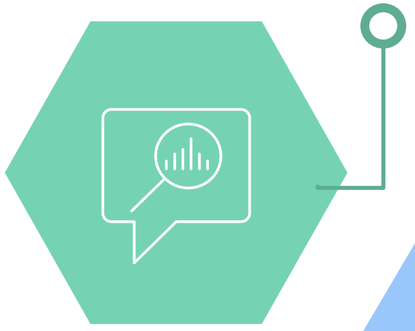
Our Principles

Human-Centricity

Promote human **well-being**, human **agency** and **equity**.

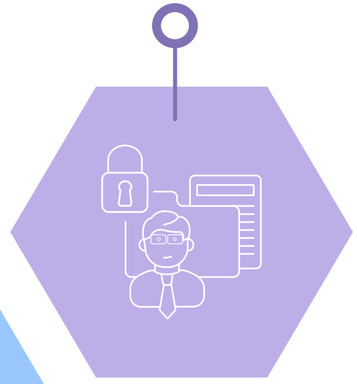
Transparency

Explain and **instruct** on usage openly, including potential risks and how decisions are made.



Privacy & Security

Respect the privacy of data subjects.



Inclusivity

Ensure **accessibility** and include **diverse perspectives** and **experiences**.

Accountability

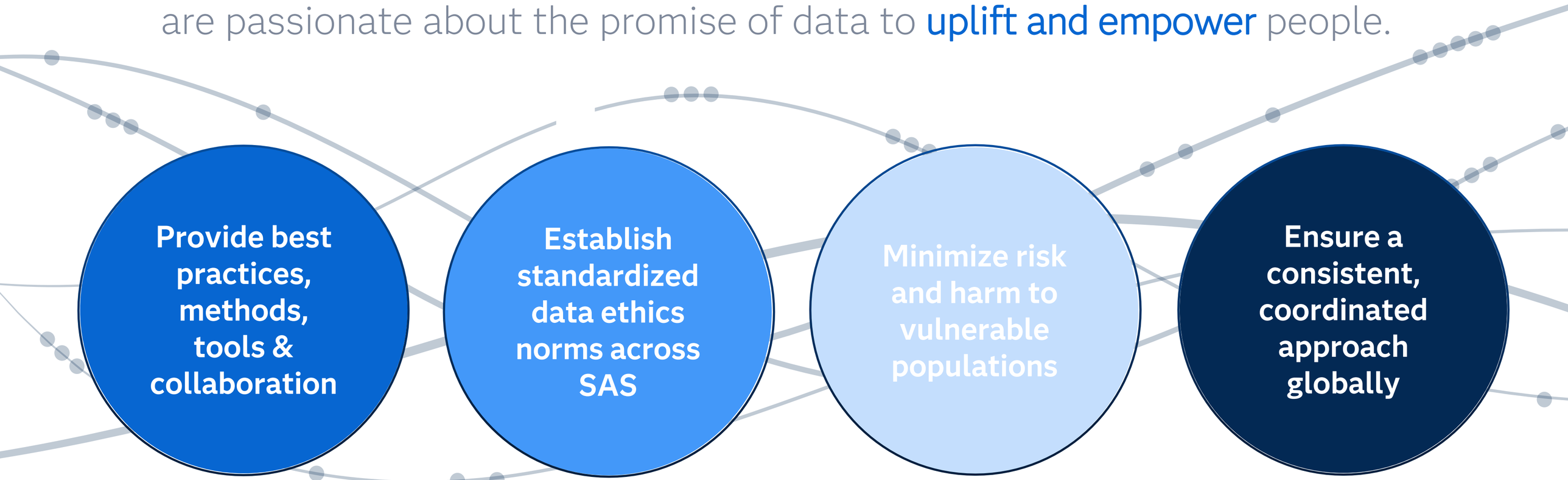
Proactively identify and mitigate adverse impacts.

Robustness

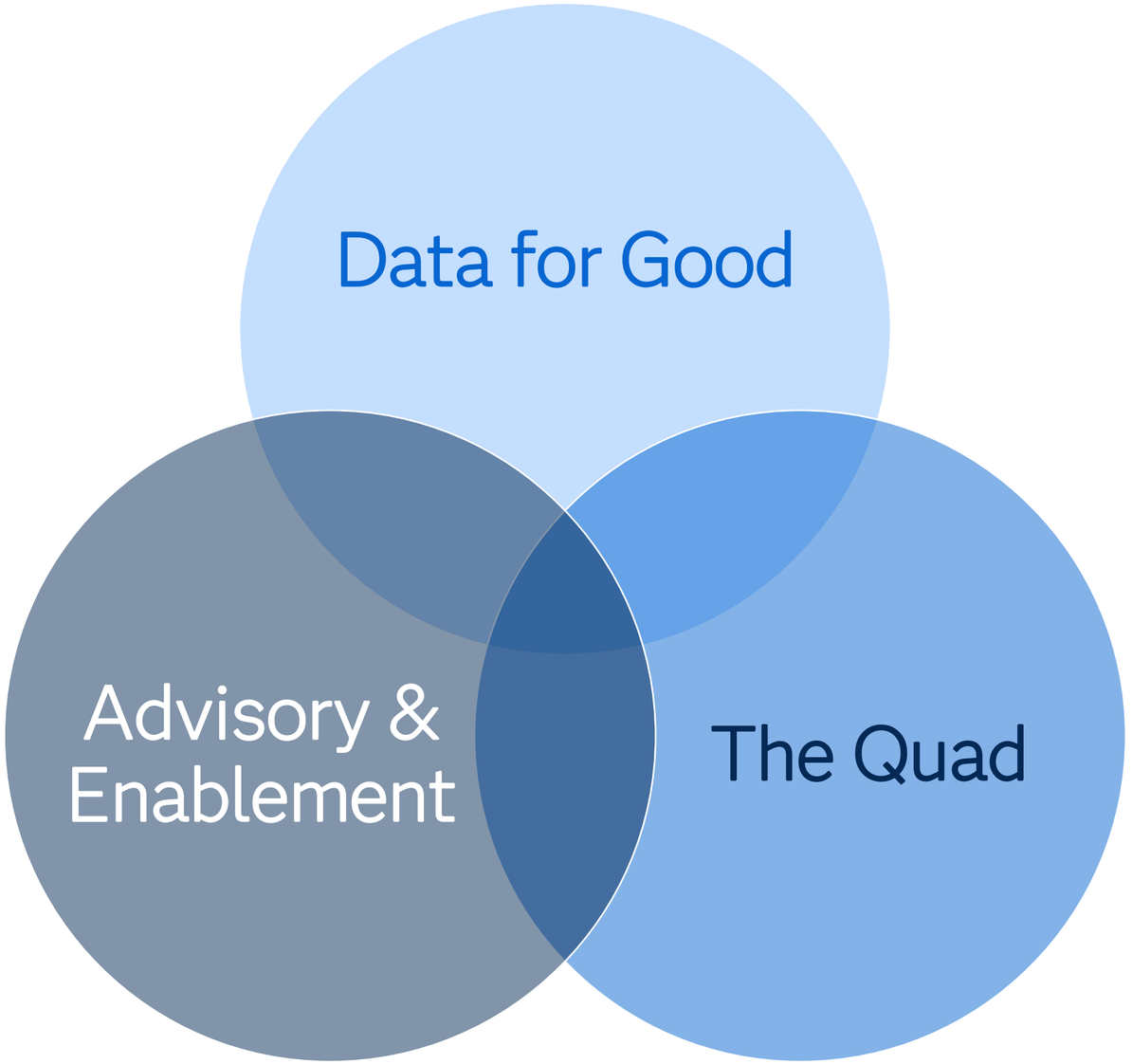
Operate **reliably** and **safely**, while enabling mechanisms that assess and manage potential risks throughout a system's lifecycle.

Data Ethics Practice

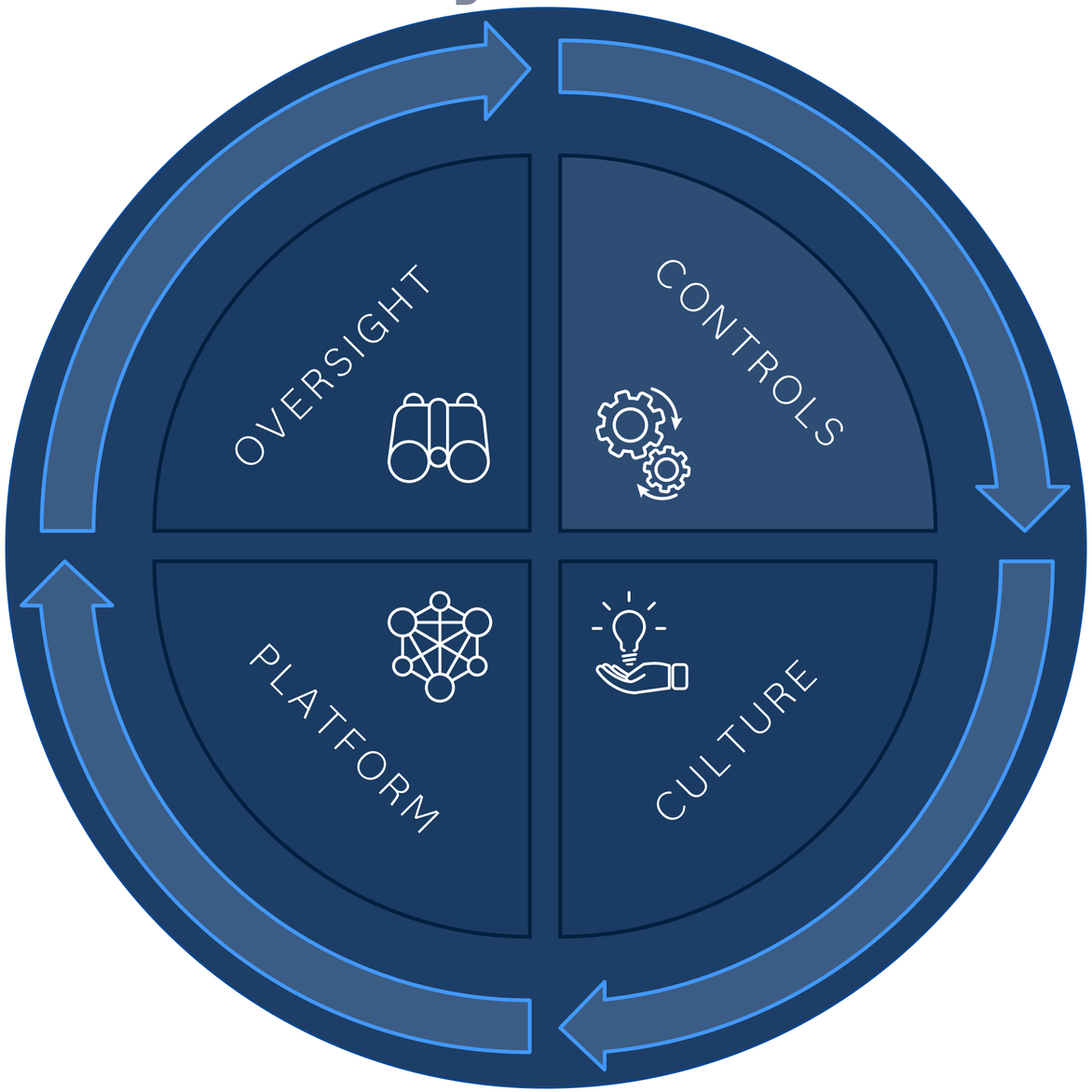
A team of experienced data scientists, developers and others at SAS that are passionate about the promise of data to uplift and empower people.



Data Ethics Practice Pillars



Governing Trustworthy AI



Activation

Oversight
AI Governance, Strategy and Enforcement

Compliance
Performance and Risk Management

Operations
Operating Procedures with Infrastructure

Culture
Systems, Norms and Practices

Technology

Data Management
- Data Quality
- Variable Metadata
- Data Preparation
- Data Asset Catalog

Explanation
- Natural Language Explanation
- Explainable ML
- Counterfactual Explanation
- Surrogate Model Interpretation
- Causal Inference

Detection
- Bias Detection
- Fairness Assessment

Privacy & Security
- Privacy Preservation
- Model Security
- Autonomy Preservation
- Consent & Control

Mitigation
- Bias Mitigation
- Bias Prevention
- Synthetic Data Generation

ModelOps
- Model Cards
- Decisioning
- Lifecycle Management
- Metric Monitoring
- Model Robustness

Solutions (industry & domain)

Accessible and Intuitive UI

Logging & Auditability

Interoperability

*It **is essential** to ensure that these models are safe, reliable and do not harm patients. **And that's exactly where I found a partner in SAS**. We both understand that **the true challenge** is not about developing AI models but is about implementing analytics at the bedside in a responsible fashion.*

Dr. Michel van Genderen
Erasmus Medical Center

Erasmus MC

Copyright © SAS Institute Inc. All rights reserved.

TRUST