Machine Learning - Fairness

DataBite Summer 2021

Dr. Christan Grant

oudatalab.com

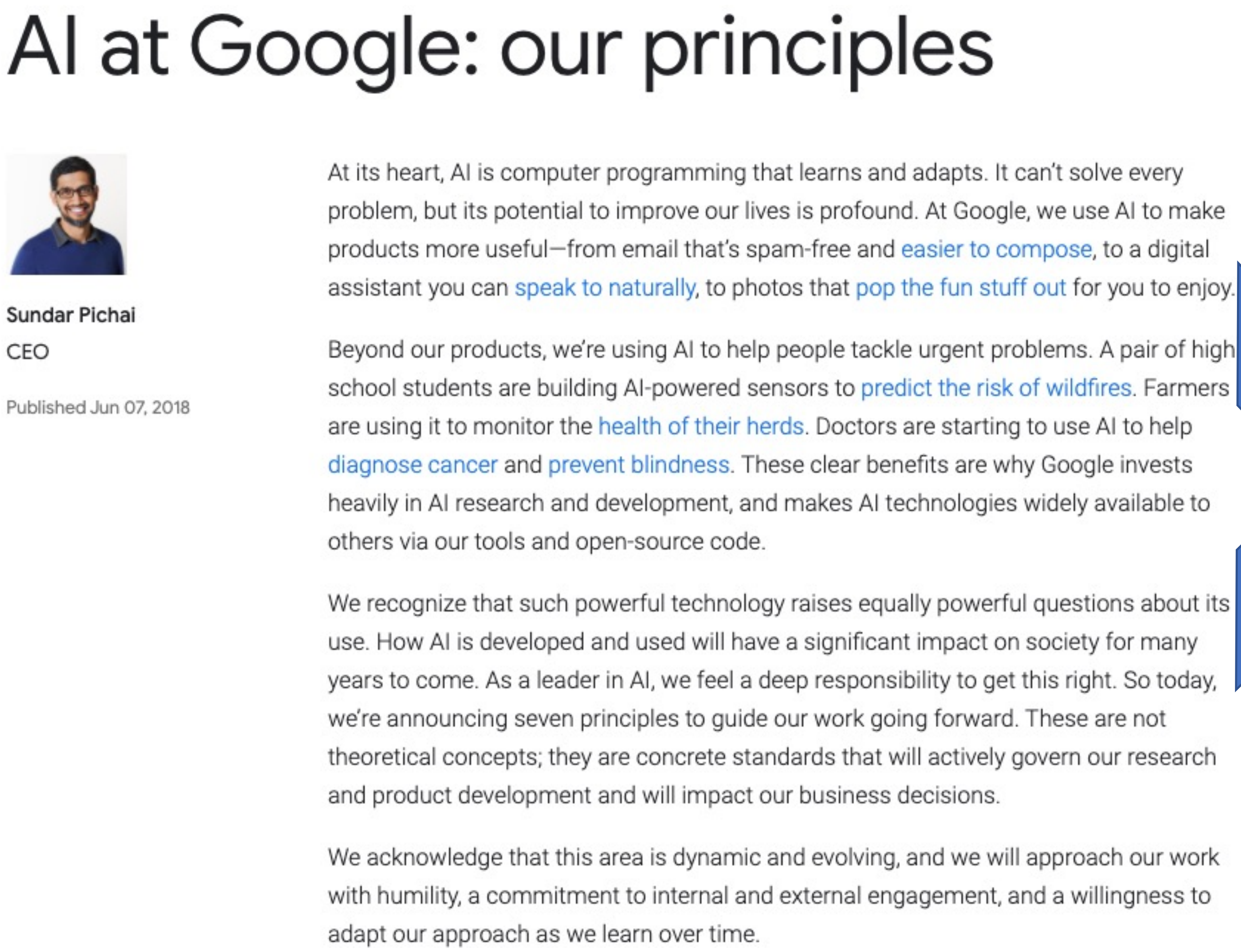
The University of Oklahoma

DataBite Summer 2021 Schedule

Day 1 Welcome	**Day 2** Introduction to Python	**Day 3** Introduction to Probability	**Day 4** Model Olympics
Day 5 Socratic Seminar	**Day 6** Bias and Fairness	**Day 7** Natural Language Processing	**Day 8** Deep Learning

“With great power comes great responsibility.”

Peter Parker Principle (RIP Uncle Ben)



AI at Google: our principles

Sundar Pichai
CEO

Published Jun 07, 2018

At its heart, AI is computer programming that learns and adapts. It can't solve every problem, but its potential to improve our lives is profound. At Google, we use AI to make products more useful—from email that's spam-free and easier to compose, to a digital assistant you can speak to naturally, to photos that pop the fun stuff out for you to enjoy.

Beyond our products, we're using AI to help people tackle urgent problems. A pair of high school students are building AI-powered sensors to predict the risk of wildfires. Farmers are using it to monitor the health of their herds. Doctors are starting to use AI to help diagnose cancer and prevent blindness. These clear benefits are why Google invests heavily in AI research and development, and makes AI technologies widely available to others via our tools and open-source code.

We recognize that such powerful technology raises equally powerful questions about its use. How AI is developed and used will have a significant impact on society for many years to come. As a leader in AI, we feel a deep responsibility to get this right. So today, we're announcing seven principles to guide our work going forward. These are not theoretical concepts; they are concrete standards that will actively govern our research and product development and will impact our business decisions.

We acknowledge that this area is dynamic and evolving, and we will approach our work with humility, a commitment to internal and external engagement, and a willingness to adapt our approach as we learn over time.

https://www.blog.google/technology/ai/ai-principles/

Objectives for AI applications

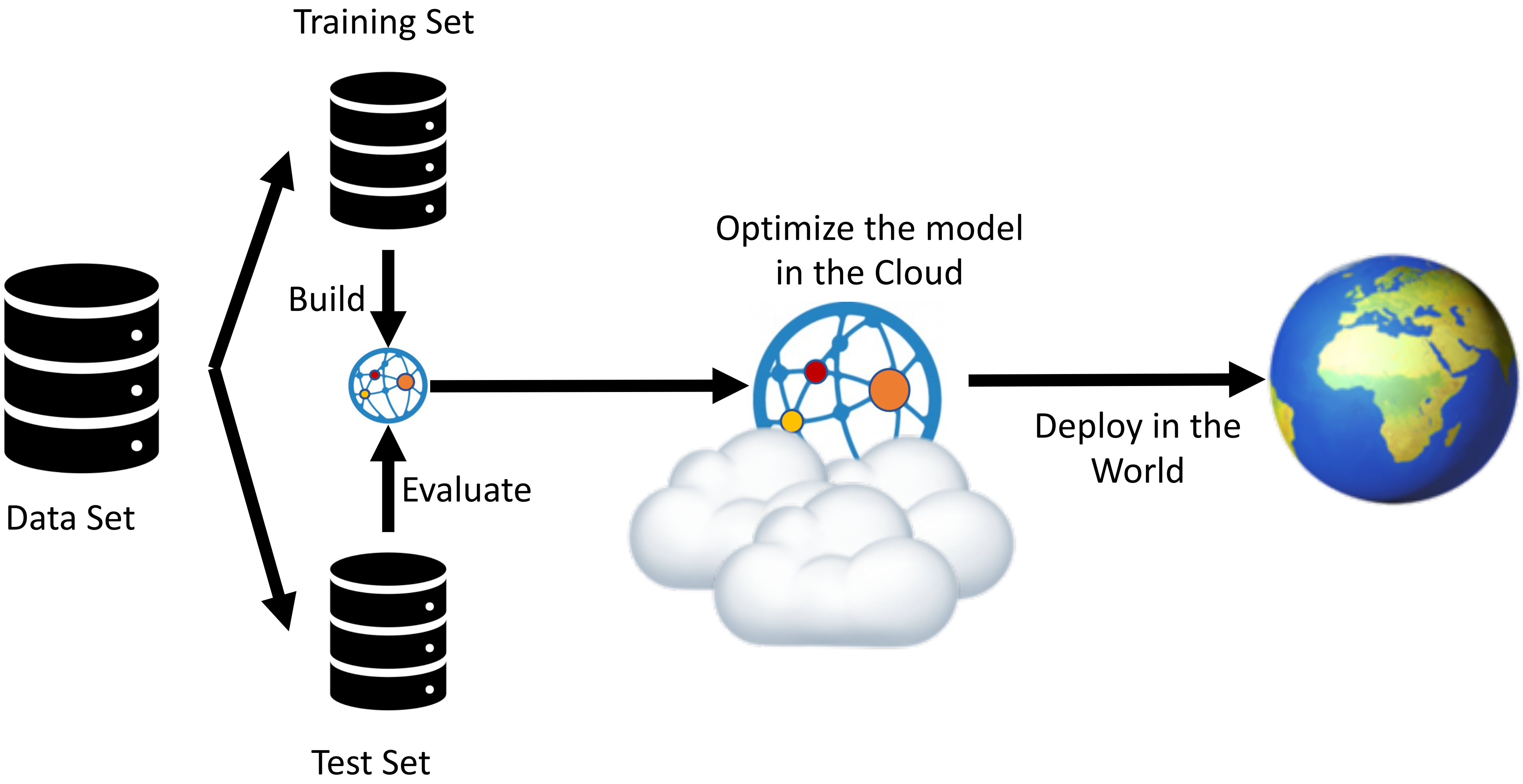
1. **Be socially beneficial.**
2. **Avoid creating or reinforcing unfair bias.**
3. **Be built and tested for safety.**
4. **Be accountable to people.**
5. **Incorporate privacy design principles.**
6. **Uphold high standards of scientific excellence.**
7. **Be made available for uses that accord with these principles.**

AI Applications we won't pursue Google

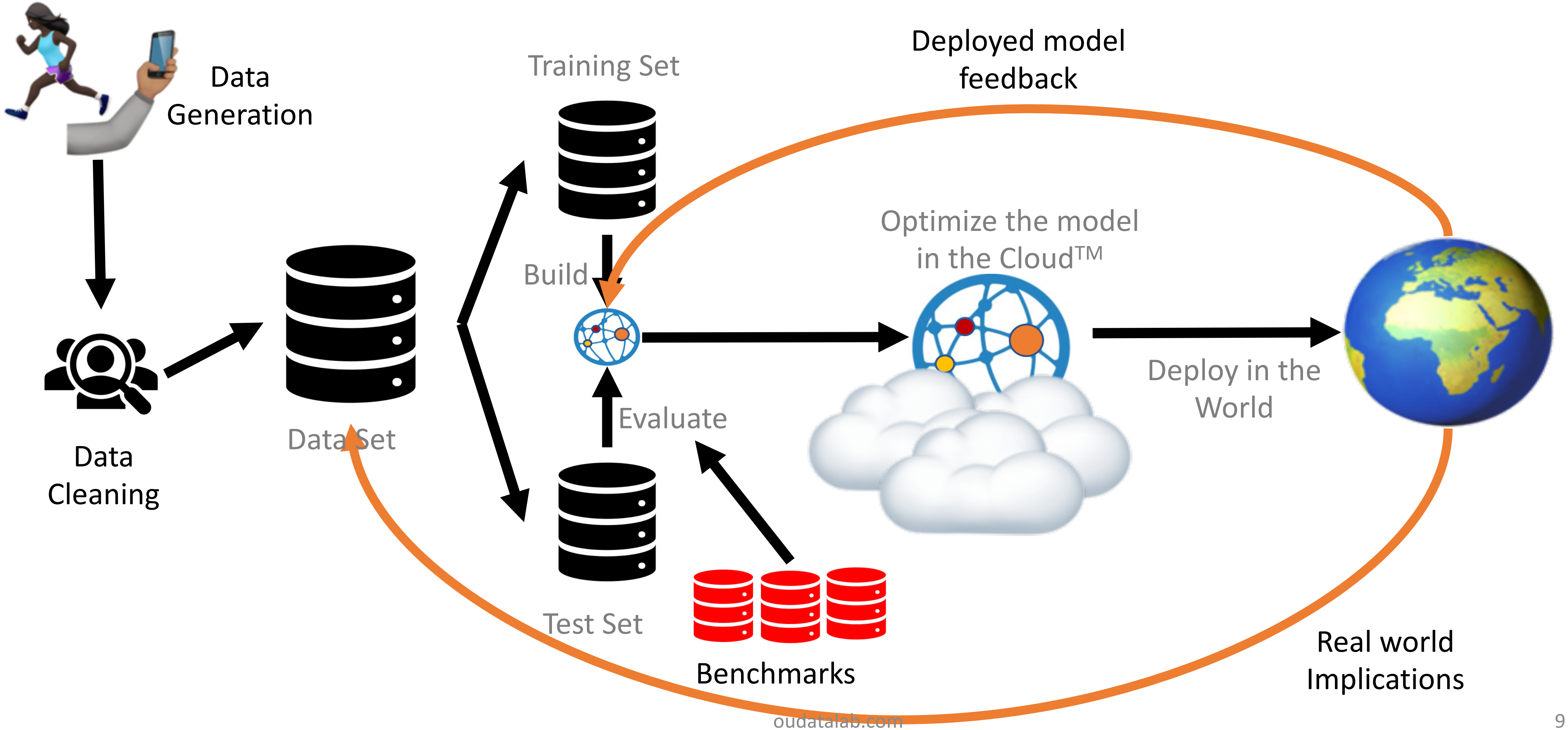
1. **Technologies that cause or are likely to cause harm.** Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints.
2. **Weapons** or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
3. Technologies that gather or use information for **surveillance** violating *internationally* accepted *norms*.
4. Technologies whose purpose contravenes widely accepted principles of *international law* and human rights.

Bias in the AI/ML Pipeline

How we think about the AI/ML “pipeline”.



The (closer) to full picture of the pipeline.



Bias Types

Reporting Bias

occurs when the frequency of events, properties, and/or outcomes captured in a data set does not accurately reflect their real-world frequency. This bias can arise because people tend to focus on documenting circumstances that are unusual or especially memorable

Automation Bias

is a tendency to favor results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each.

Selection bias

occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.

Selection bias

occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.

- **Coverage bias** -- occurs when data is not selected in a representative fashion.

Selection bias

occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.

- **Sampling bias** -- occurs when proper randomization is not used during data collection.

Selection bias

occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.

- **Non-response bias --** occurs when data are unrepresentative due to participation gaps in the data collection process.

Group Attribution Bias

is a tendency to generalize what is true of individuals to an entire group to which they belong.

Group Attribution Bias

is a tendency to generalize what is true of individuals to an entire group to which they belong.

- **In-group bias** -- A preference for members of a group to which you also belong, or for characteristics that you also share.

Group Attribution Bias

is a tendency to generalize what is true of individuals to an entire group to which they belong.

- **Out-group homogeneity bias --** A tendency to stereotype individual members of a group to which you do *not* belong, or to see their characteristics as more uniform.

Confirmation Bias

is where model builders unconsciously process data in ways that affirm preexisting beliefs and hypotheses.

Confirmation Bias

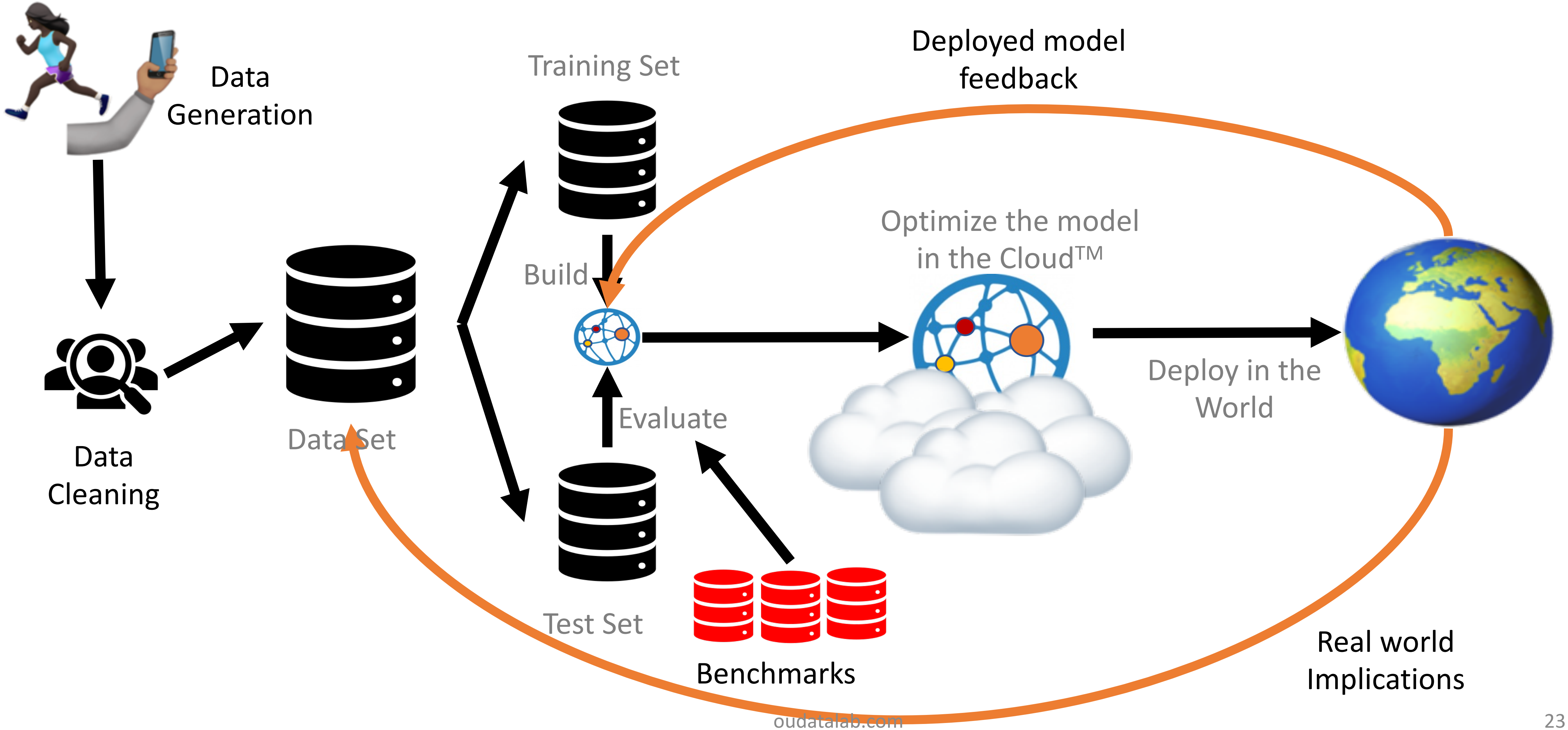
is where model builders unconsciously process data in ways that affirm preexisting beliefs and hypotheses.

- **Experimenter's bias** – a model builder may actually keep training a model until it produces a result that aligns with their original hypothesis.

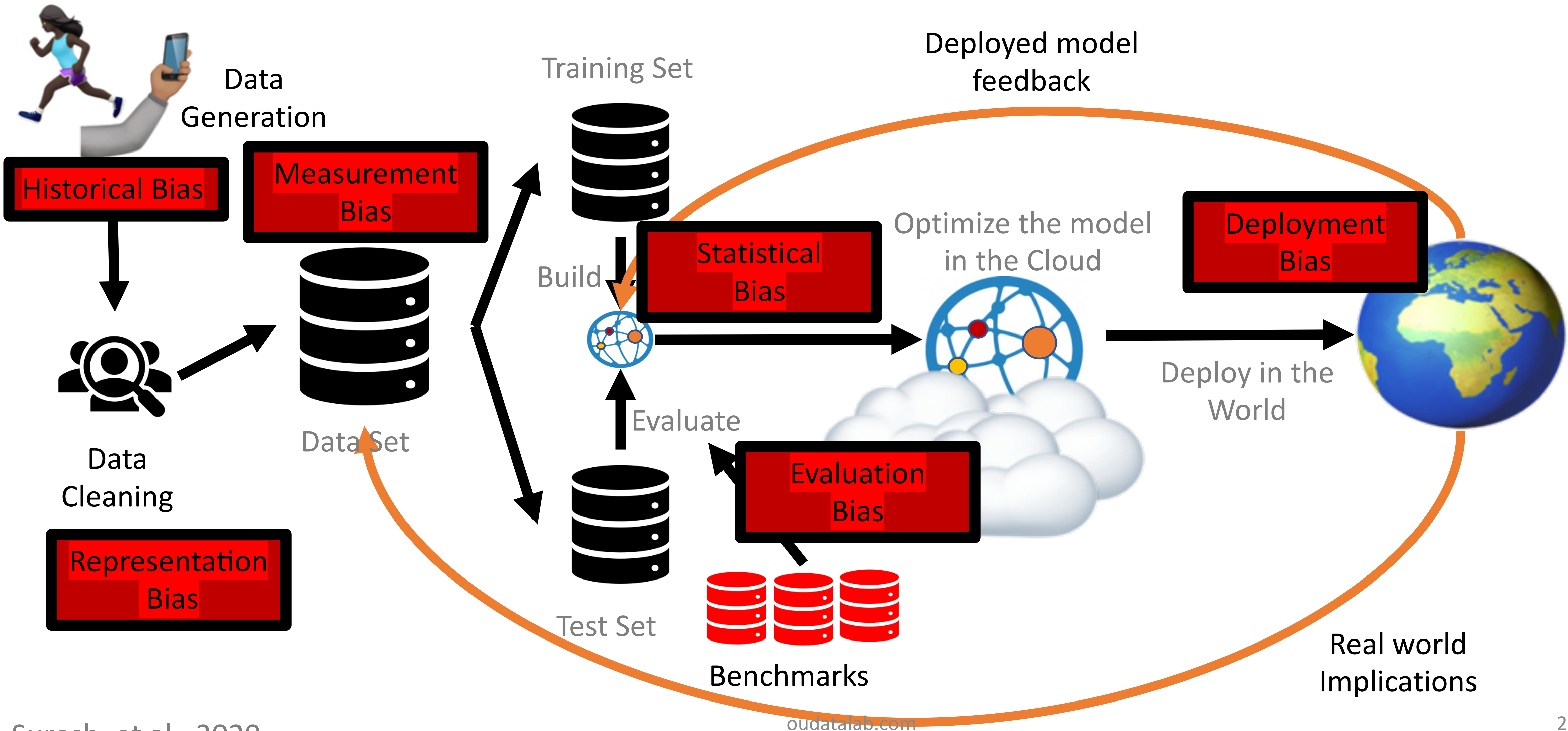
Bias Types

- Reporting
- Automation
- Selection (coverage, non-response, sampling)
- Group attribution (in-group, out-group)
- Implicit (confirmation, experimenters)

The (closer) to full picture of the pipeline.



There is Bias hiding in every step!



Suresh, et al., 2020