Stay tuned and join data excellence mailing list
https://bit.ly/data_excellence

# Welcome!

## Data Excellence:
## Better Data for Better AI

Google Research

data is the compass for AI - AI advances where there is data

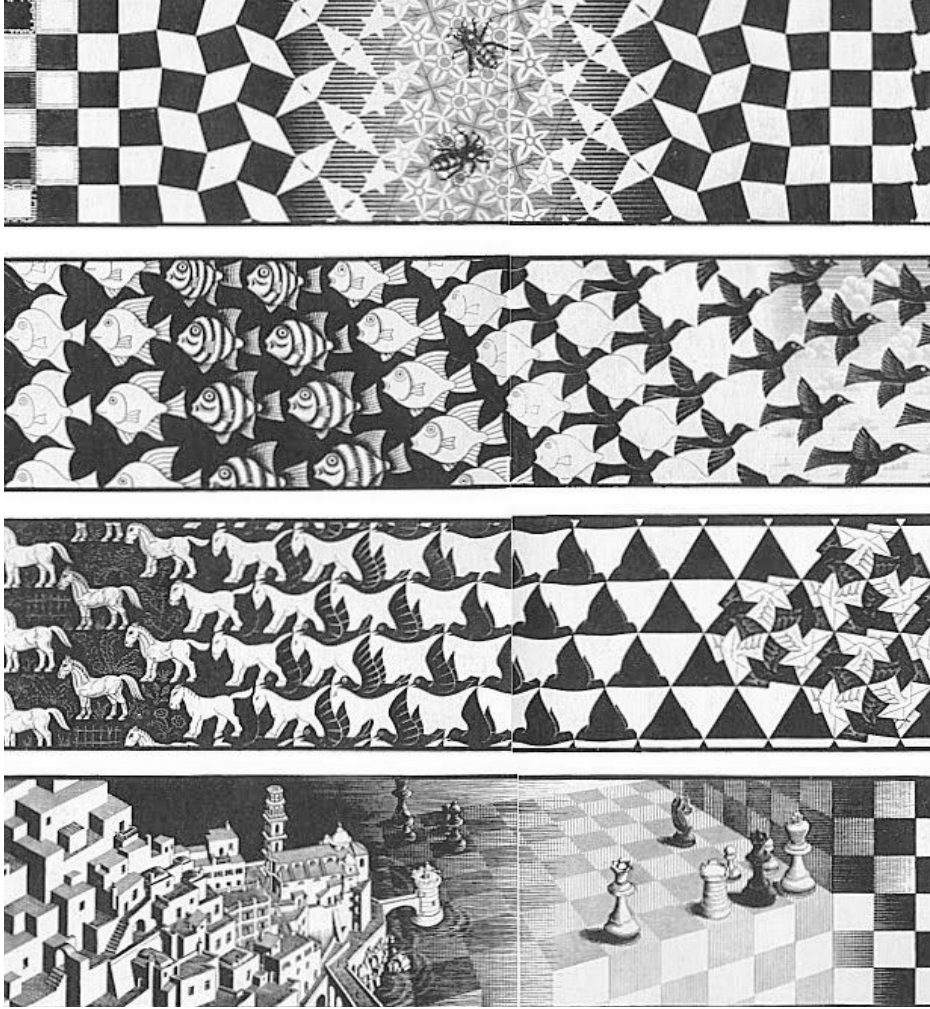
data quality must be addressed in AI practices especially in the way we evaluate AI

improving evaluation of AI must consider ways to measure variance and capture bias to bring us one step closer to data excellence

to address variance in AI evaluation we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address bias in AI evaluation we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

# TAKE HOME MESSAGE



Google Research

data is the compass for AI - AI advances where there is data

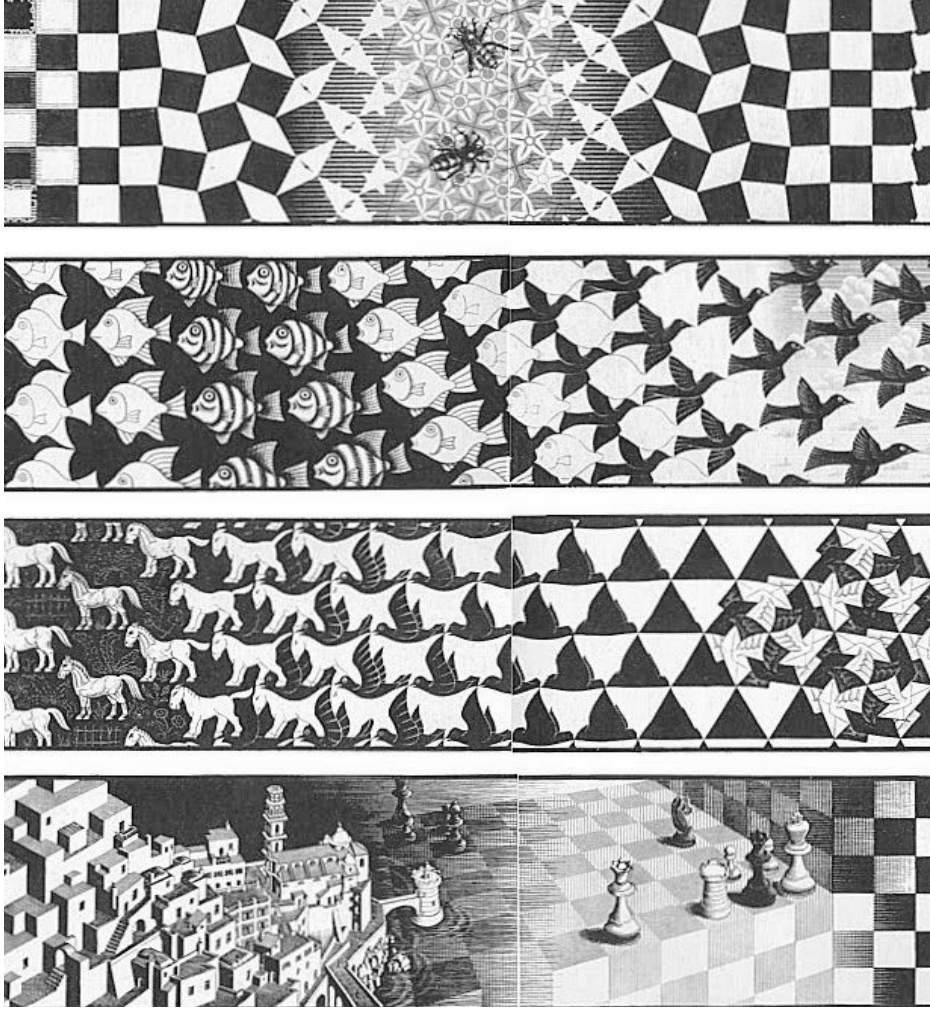
data quality must be addressed in AI practices especially in the way we evaluate AI

improving evaluation of AI must consider ways to measure variance and capture bias to bring us one step closer to data excellence

to address variance in AI evaluation we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address bias in AI evaluation we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

# TAKE HOME MESSAGE



Google Research

data is the compass for AI - AI advances where there is data

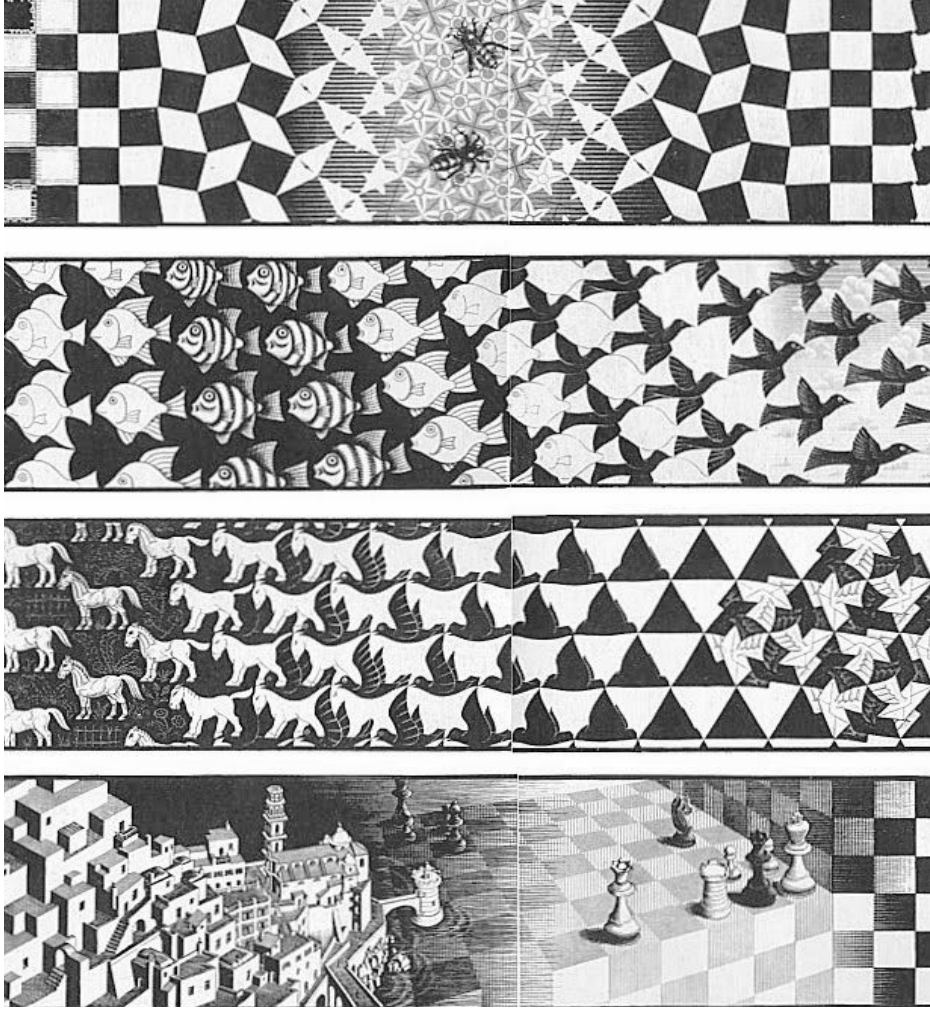
data quality must be addressed in AI practices especially in the way we evaluate AI

improving evaluation of AI must consider ways to measure variance and capture bias to bring us one step closer to data excellence

to address variance in AI evaluation we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address bias in AI evaluation we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

# TAKE HOME MESSAGE



Google Research

data is the compass for AI - AI advances where there is data

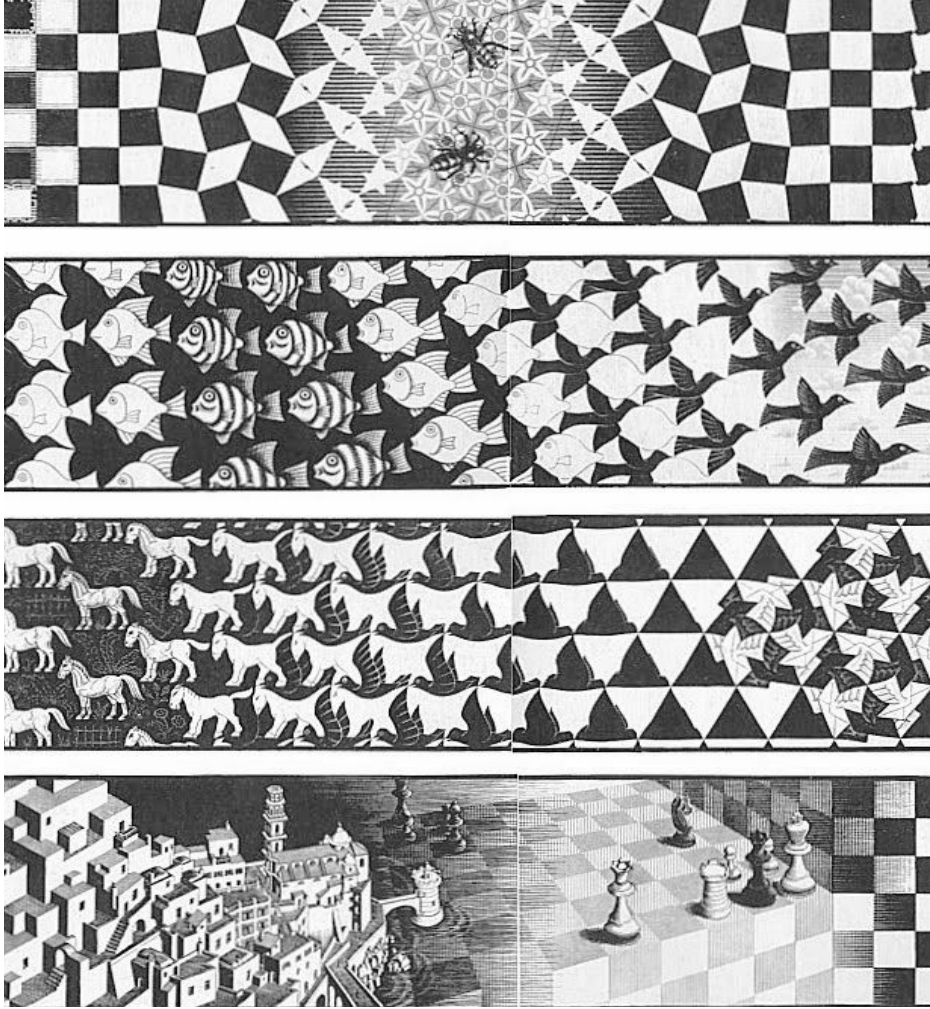
data quality must be addressed in AI practices especially in the way we evaluate AI

improving evaluation of AI must consider ways to measure variance and capture bias to bring us one step closer to data excellence

to address variance in AI evaluation we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address bias in AI evaluation we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

# TAKE HOME MESSAGE



Google Research

# The Rise of the Machines

*lab experiments*

**“AI Winter”**

Expert Systems
small scale
experiments

# The Rise of the Machines

*beat the humans*

"AI Winter" → **"AI Breakthroughs in Games"**

IBM Watson Jeopardy
DeepMind AlphaGo

# The Rise of the Machines

*support the humans*

"AI Winter" → "AI Breakthroughs in Games" → **"Real World Tasks"**

Health diagnostics
Flue prediction
Weather prediction
Text, Image and Video classification
Text Generation
Text Translation
Conversational AI

Google Research

# Mainstream Deployment of AI

**“Real World Tasks” deployed in the wild → Unintended behaviors**

Microsoft Tay bot
IBM Watson Oncology
Amazon Rekognition
Google Photos
Apple Face ID
Facebook chat bots
Various Speech Assistants

# Data is the compass for AI

data quality is essential for guiding AI away from unintended behaviours

getting computers to “see” the diversity of data

Google Research

# The Life of AI Data

*bootstrapping AI with data*

"It exists!"

Caltech101
LabelMe
Berkley-3D

*https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research*

Google Research

# The Life of AI Data

*data hungry AI*

"It exists!" → "It is bigger!"

ImageNet
SIFT10M
OpenImages
COCO
Web 1T 5-Gram

*https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research*

Google Research

# The Life of AI Data

*but before it got better ...*

"It exists!" → "It is bigger!" → "It is better!"

# The Life of AI Data

*but before it got better ...*

"It exists!" → "It is bigger!" → "It is better!"

*it got worse ...*

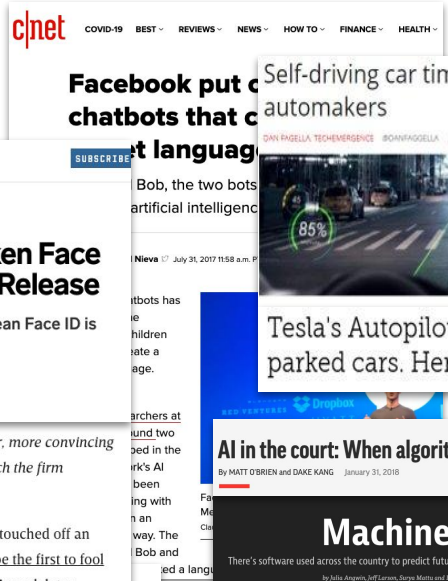
Google Research

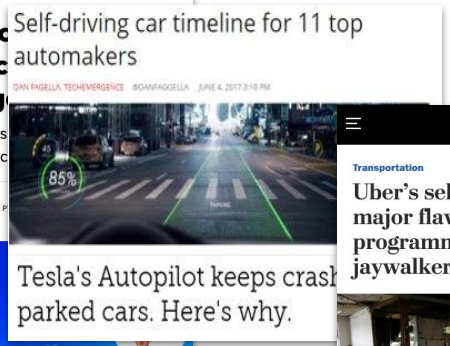
# Unintended Behaviors in AI

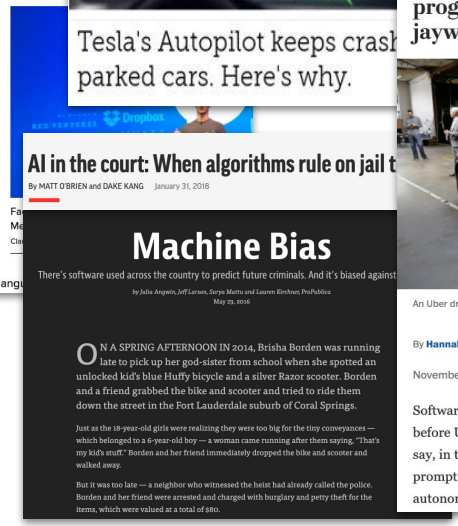


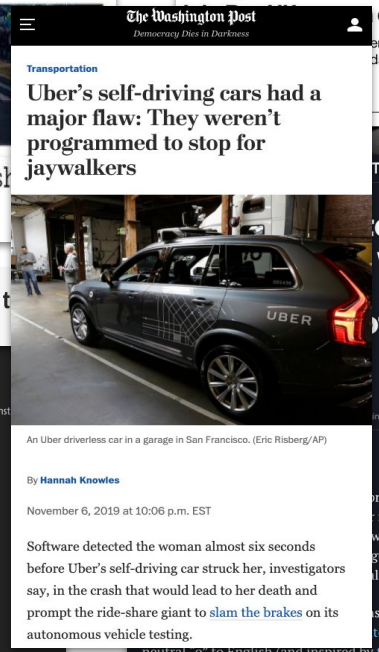




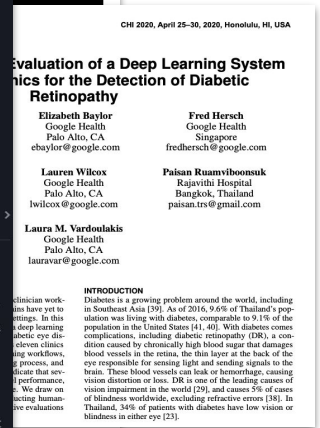










*Adapted from "AI in the Open World: Discovering Blind Spots of AI", SafeAI 2020, Ece Kumar*

Google Research

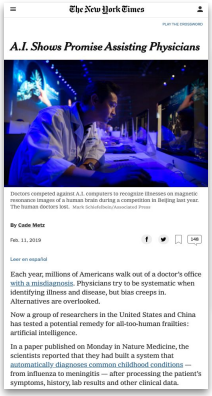
# The Life of AI Data

*but before it got better ...*

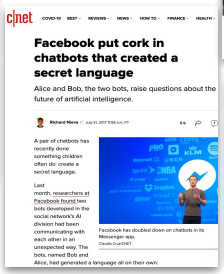
"It exists!" → "It is bigger!" → "It is better!"

*reactive*

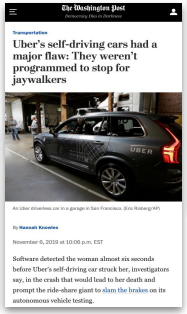
*data improvement*



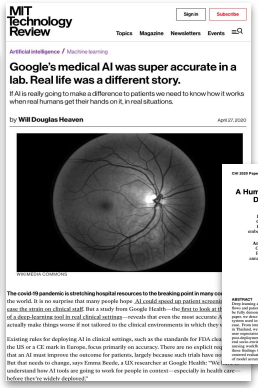












Google Research

Ikart

# The Life of AI Data

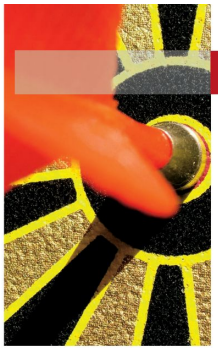
*to reach here*

"It exists!" → "It is bigger!" → "It is better!"

*we need proactive data improvement*

Google Research

# The Life of AI Data



EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

## The Unreasonable Effectiveness of Data

**Alon Halevy, Peter Norvig, and Fernando Pereira,** *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

### Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech rec-

In the decade since then, the research community have done a lot with **quantity**, but **quality has been left behind**

*Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. IEEE Intelligent Systems 24, 2 (2009)*

Google Research

# But ...Data Quality is not easy ...

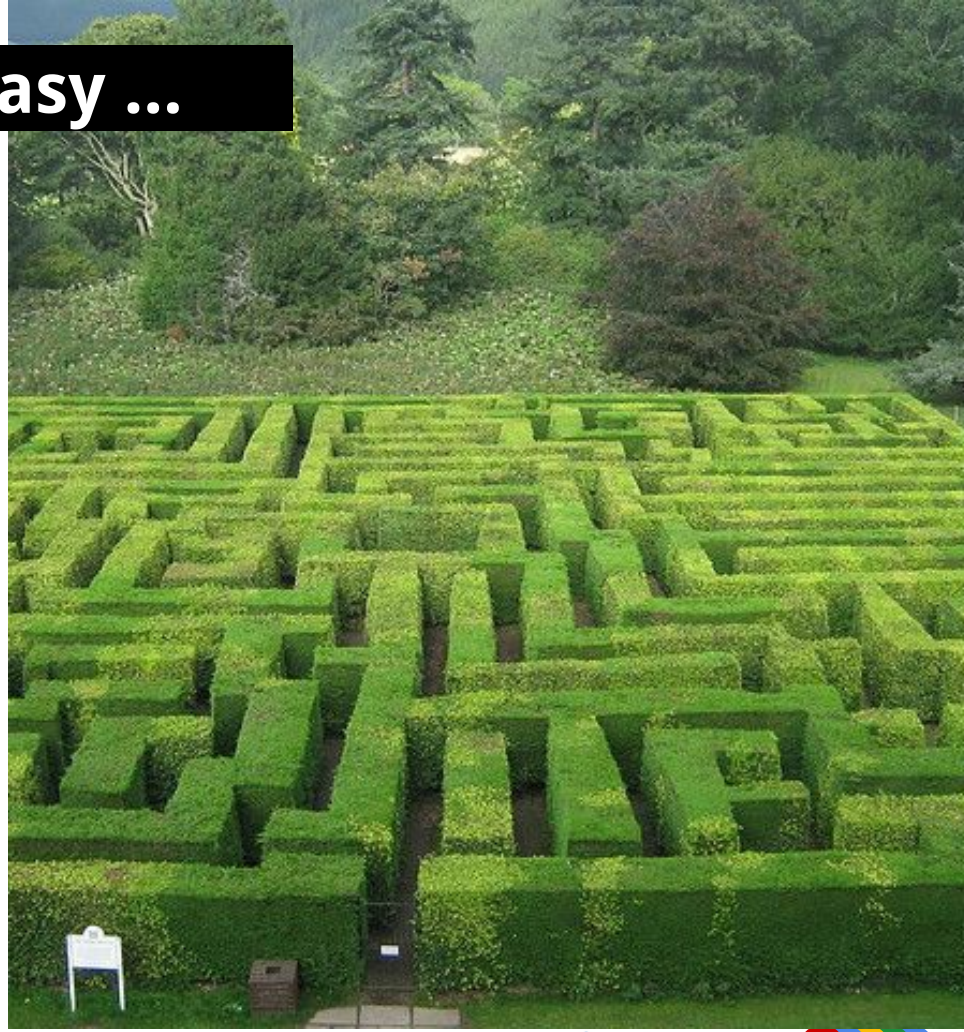
datasets are not easy to debug

**low data quality** is typically **not a result of**:

- software bugs, or
- just human errors

**achieving excellent data** requires change in the way we gather data and use it to evaluate AI systems:

- ensure that the datasets represent the problems they are intended for
- ensure that the human annotation process for these datasets allows to capture the natural variance & bias in the problem
- ensure that the metrics used to measure the AI quality utilize the variance & bias in the human responses



# Data Quality is not only human error

Do these images depict a **GUITAR** ?

**it is not easy to give Y/N answer**
for most of our AI tasks

Google Research

# Data Quality should consider context of use

Do these images depict **NEW ZEALAND** ?

**it is not easy to give Y/N answer**
for most of our AI tasks

the answer typically **depends on the context, on the task, on the usage, etc**

Google Research

# Data Quality should include real world diversity

Do these images depict a **WEDDING** ?

**it is not easy to give Y/N answer** for most of our AI tasks

the answer typically **depends on the context, on the task, on the usage, etc**

**disagreement is signal for natural diversity and variance in human annotations** and should be included in AI training

Google Research

# Data Quality is difficult even with experts

Does the sentence express **TREATS relation** between Chloroquine, Malaria?

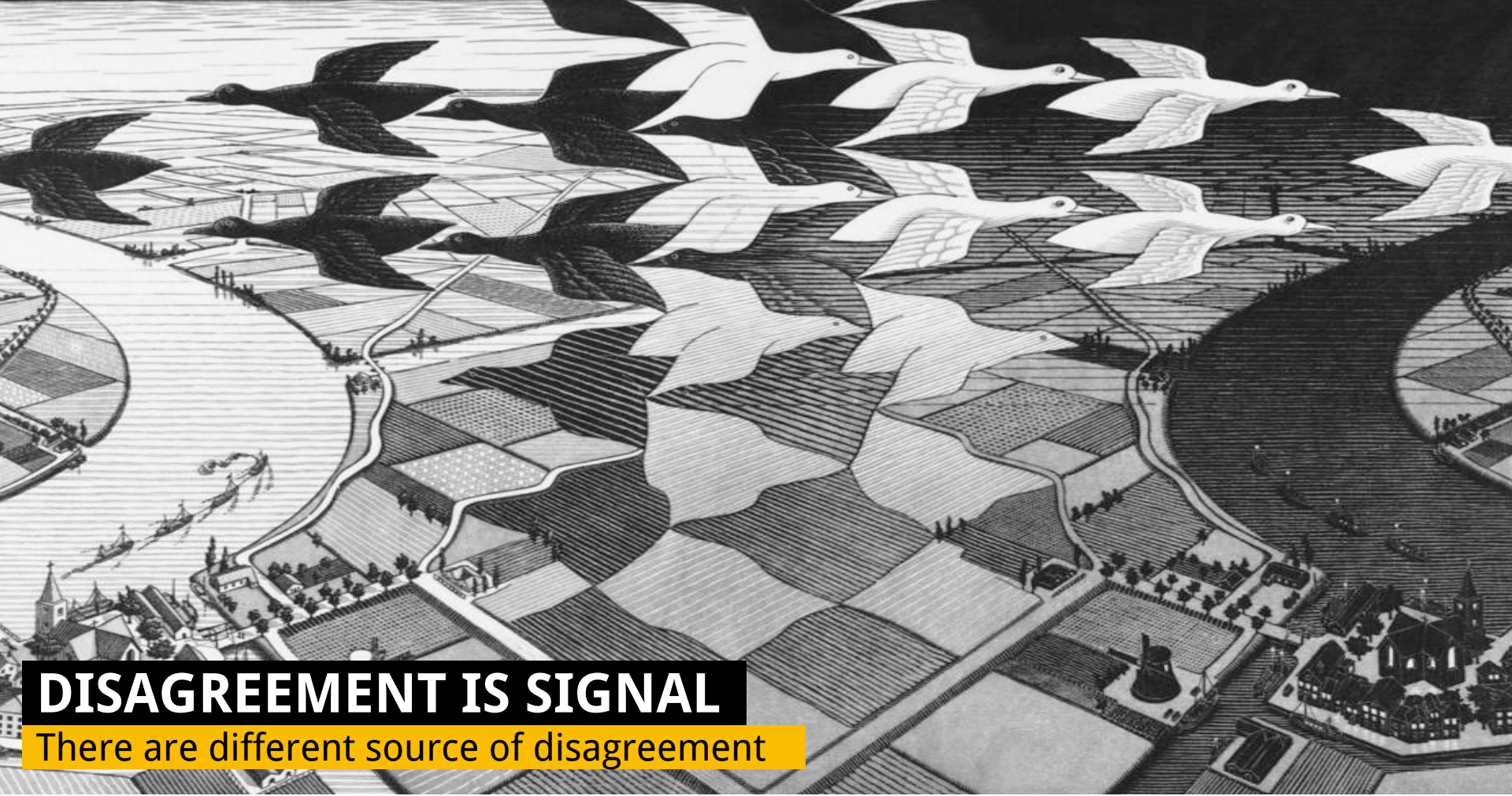
Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

*"Crowdsourcing Ground Truth for Medical Relation Extraction", ACM TiiS Special Issue on Human-Centered Machine Learning, 2018. A. Dumitrache, L. Aroyo, C. Welty*

Google Research



# DISAGREEMENT IS SIGNAL

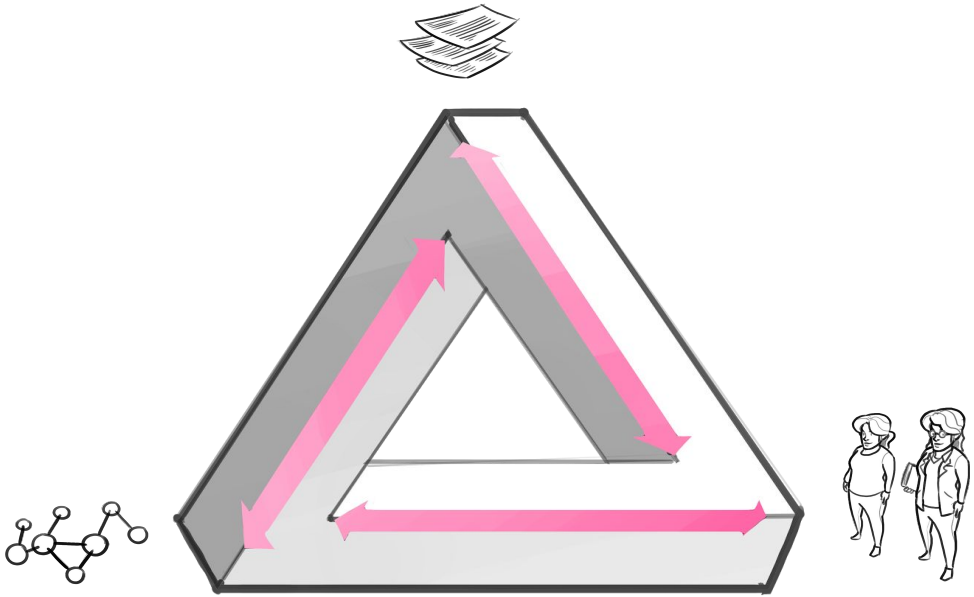
There are different source of disagreement

Google Research

'lɪkərt

# CROWDTRUTH

## Three sides of human interpretation



Disagreement provides **guidance in task analysis:**

- items with **poor semantics**
- items with **salient terms**
- items **difficult to classify**
- items that are **ambiguous**

- users **with/without specific knowledge**
- communities of thought
- spammers

- **subjective** annotations
- **time-sensitive** annotations
- **difficult** annotation tasks
- **mis-translated** annotations

*"Three Sides of CrowdTruth", Human Computation Journal, 2014, L. Aroyo, C. Welty*

Google Research

# … but in current practices disagreement is continuously avoided and the world is forced into a binary setting

## 7 Myths about Human Annotation

**One truth:** knowledge acquisition typically assumes one correct interpretation for every example

**Experts rule:** knowledge is captured from domain experts

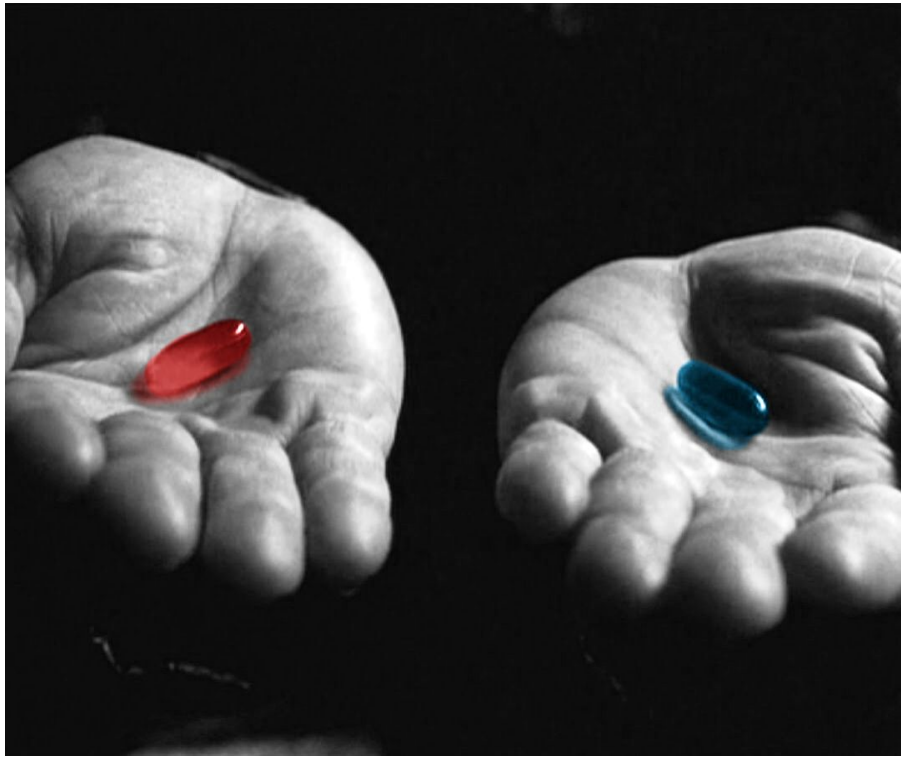
**One is enough:** single expert's knowledge is sufficient

**Disagreement bad:** when people disagree, they must not understand the problem

**Detailed explanations help:** if examples cause disagreement - adding instructions should help

**Once done, forever valid:** knowledge is not updated; new data not aligned with old

**All examples are created equal:** triples are triples, one is not more important than another, they are all either true or false



*"Truth is a Lie: 7 Myths about Human Annotation", AI Magazine 2014, L. Aroyo, C. Welty*

Google Research

data is the compass for AI - AI advances where there is data

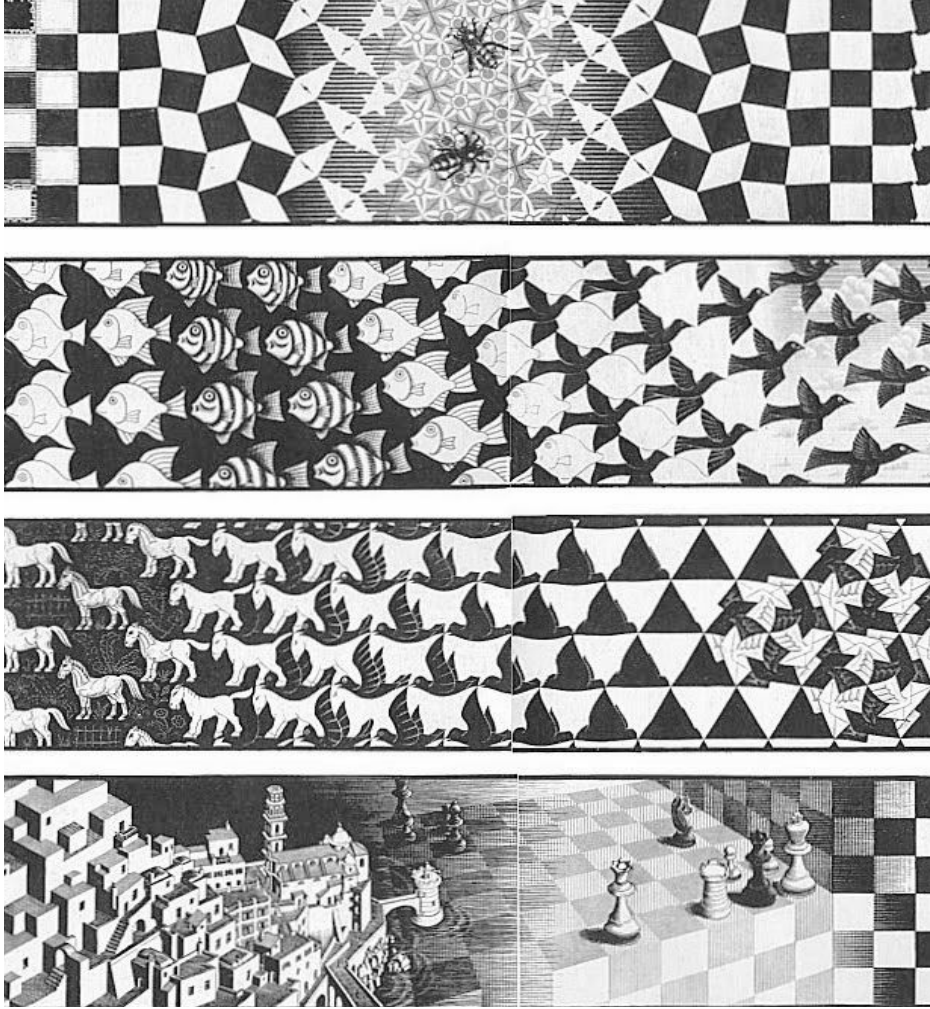
data quality must be addressed in AI practices especially in the way we evaluate AI

improving evaluation of AI must consider ways to measure variance and capture bias to bring us one step closer to data excellence

to address variance in AI evaluation we propose a number of novel metrics for reliability, significance (metrology for AI) and disagreement (CrowdTruth)

to address bias in AI evaluation we propose a novel method for crowdsourcing adverse test sets for ML models (CATS4ML)

# TAKE HOME MESSAGE



Google Research

Your AI model is as good as your **evaluation data** … but is your evaluation data **missing relevant examples**? **How can we find them if they are AI blindspots (i.e. unknown unknowns)?**

Join us tomorrow Dec 9 at 7:30-8:00pm (EST) to hear more about the **CATS4ML challenge (Crowdsourcing Adverse Test Sets for ML):** cats4ml.humancomputation.com


Google Research

Stay tuned and join data excellence mailing list
https://bit.ly/data_excellence

'lɪkərt

Google Research

Praveen Paritosh

Ka Wong

Lora Aroyo

Devi Krishna

# The Team

**Share your voice about ML data!**

We are inviting ML professionals to participate in a short survey to learn about your challenges and needs with data.

Scan to participate now!

People + AI Research

Google Research

# Build for everyone

Fill out our interest form to hear about events and opportunities at **goo.gle/neurips-booth-form**

Google at NeurIPS 2020

Google Research