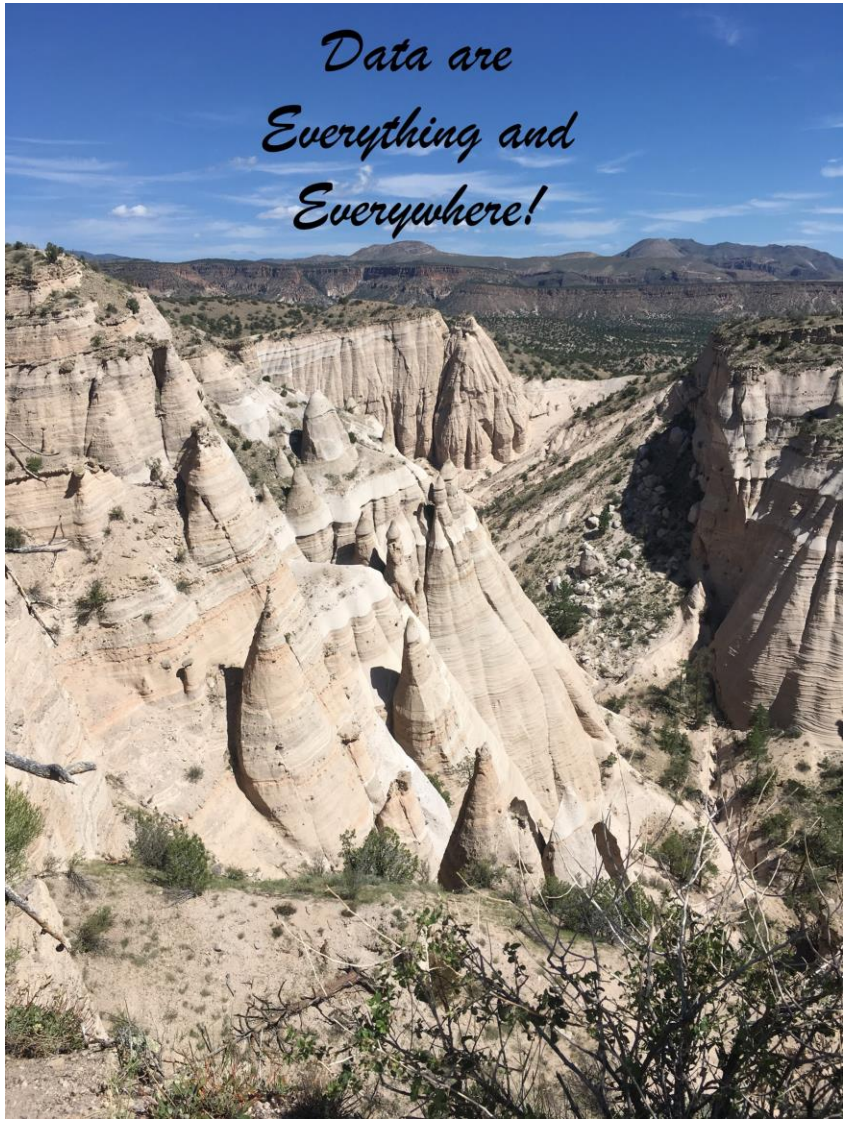
On Democratizing Data:

Diminishing Disparity and Increasing Scientific Productivity

Ophir Frieder
Georgetown University

Data – Everything & Everywhere



- AI models dominate
- Data fuel these models

More data → better models → greater advancements

Data Dominate

- The Economist on Data: **“The World’s Most Valuable Resource”** 2017

- Data democratization – Not new!
  - NIST’s Text REtrieval Conference (**TREC**) (1992 – ongoing)
  - America Online (AOL) **Query Log** Release (2006)
  - Medical Information Mart for Intensive Care (**MIMIC**) collections (1996 – 2023)

Data availability fostered innovations & societal benefits

Foundational Models

- A rose by any other name would smell as sweet (Shakespeare's Romeo and Juliet)
  - Pretrained Language Models (PLM)
  - Large Language Models (LLM)

- Training Data: Larger, larger, and even larger

  - Google's Bidirectional Encoder Representation from Transformers (**BERT**) (2018)
    - BERT: ~ 3.3 B tokens

  - Meta's Large Language Model Meta AI (**LLaMA**) (2023)
    - LLaMA: ~ 1.4 T tokens
    - LLaMA 2: ~ 2.0 T tokens

  - Open AI's Generative Pre-trained Transformer 4 (**GPT-4**) (2023)
    - Undisclosed training data

Encoding Still Demands Data

- Generalist vs. Specialist for human health
  - Primary Care Physician (**Generalist**) – "common cold"
  - Cardiologist or pulmonologist (**Specialist**) – chronic heart or lung ailments

- Generalist vs. Specialist for informational needs
  - BERT – Generalist for general information needs
  - BERT Specialists:
    - **ClinicalBERT** (medical notes),
    - **FinBERT** (financial corpus)
    - **SciBERT** (scientific texts)

Specialization requires relevant data access

Democratized Data Benefit Society

- Taiwan's National Health Insurance Research Database (**NHIRD**)
  - Covering more than **99% of Taiwan's population**
  - Spans **over two decades**
  - Contains **deidentified but some demographic** information

- Clinical advancements derived
  - **Prescription efficacy prediction** via a graph-based deep learning method
  - **Dementia**, **pancreatic cancer**, and **other ailments** related **clinical enhancements**

Availability of NHIRD improved patient care

Don’t Democratize Blindly

- **Some data are sensitive**
  - Include personal identifiable information (PII)
  - Contain proprietary information
- **Some data are biased**
  - Inadvertently
  - By intent
- **Some data are flawed**
  - Procedural or measurement errors
  - Poorly generated “gold standard”

Brighter Tomorrow

- **Democratize data**
  - Fosters (scientific & social) innovation
  - Provides greater participation
  - Improves decision-making by supporting specialization

- **Data democratization**
  - Enhances transparency
  - Exposes bias
  - Enables discovery and repeatability for verification

Safe, accurate, privacy-protecting democratized data opens possibilities yet to be comprehended

With Gratitude

- Arman Cohan
- Nazli Goharian
- Sean MacAvaney
- Eugene Yang
- Hao-Ren Yao
- Andrew Yates

Anonymous Reviewers