AI-Ready Data

The Census Track to Machine Understandable Data

Kenneth Haase

Senior Computer Scientist for Artificial Intelligence Applications

US Census Bureau

The views expressed in this presentation are those of the presenter and not the US Census Bureau.

Plan

- AI, GenAI, LLMs, Oh my!
- What’s really changed
- Thinking Bigger
- Machine Understandable Data
- Rethinking Data

What is AI?

- Patrick Winston’s definition: Systems which can do things which people would say requires “intelligence”
- Alan Turing’s definition: Systems which an convince people they’re human (Imitation Game)
- Marvin Minsky’s definition: Systems which adaptively pursue goals and can learn new kinds of adaptations
- Ken’s definition: Systems which can acquire and invent new models and ways of thinking

Generative AI

Systems which generate artifacts – narratives, images, videos, podcasts, etc. – for which a human would be willing to take credit

- Amazing progress in the past four years
- Moves the focus from how it works to what it produces
- Typically based on large arithmetic models with billions of parameters
- Requires significant compute to use and massive compute (and data) for training (parameter identification)
- Opaque, fantasy-prone, difficult to debug or correct

What’s really changed

- We now have technology to create software components whose function is derived from training on very large data sets
- We can embed those components in larger systems including a variety of components and business logic
- Creating these composite systems requires expertise, iteration, experimentation, and “real understanding”
- As data-centric organizations, that requires making our data “machine understandable” and not just “machine readable”

Machine Understandable Data

Data accompanied by **metadata** to support **complex processing**

- Level 1: Machine Readable
    - Digital formats with separation into records and fields
- Level 2: Table level metadata (including for discovery)
    - Topics, source, provenance, licensing, etc.
- Level 2: Field/variable (for ‘scalar’ values)
    - Labels, types, tanguage, precision, accuracy, etc.
- Level 3: Identity and constraints
    - Primary/secondary keys, valid ranges, business logic, etc.

Understanding More

- Application-level logic pushed into the database
    - First,Middle,Last <=> Personal Name
    - Unpacking adaptive design logic
- Data interdependency and independence
    - Important for combining variables
    - Important for realistic imputation
- Distributional expectations
- Non-scalar and complex compound data
    - Ontological variables
    - Flexible value schemas
    - Varieties of missingness

Thinking outside of the phone

Reimagining the

World

Data

Model

Insight

Action

World

pipeline

United States® Census Bureau

Seeds of Possibility

- Accessibility & Democratization
  - Question answering and explanation for non-expert end users
  - Pedagogical UX providing “framing” with facts
- Hybrid Data Franken-products
  - Using rich metadata to responsibly combine data sources
- Knowledge-driven imputation
  - Using both general knowledge and complex models
- Complex ‘non-scalar’ data processing
  - Ontological, ambiguous, and multi-scale data values
- Interactive Survey Instruments
  - Multiple-choice -> Free text -> Conversations