



#Standards4AI

**'Putting Science Into
Standards' workshop**

Welcome!
We will start soon

CREATING AND DOCUMENTING DATASETS FOR AI

8 June, 15:45-17:15



Panel discussion

CREATING AND DOCUMENTING DATASETS FOR AI



Roundtable speakers

Felix NAUMANN

Hasso-Plattner-
Institut

**Emmanuel
KAHEMBWE**

VDE

Kasia CHMIELINSKI

Dataset Nutrition
label

Flora DELLINGER

Valeo, Confiance.ai

Rapporteurs: Isabelle Hupont Torres (JRC)



Audience interaction



slido.com
#Standards4AI



- ✓ Select the **Dataset for AI** room on Slido
- 💬 Zoom chat - only technical questions to host
- 🚫 Camera and audio OFF



#Standards4AI

**'Putting Science Into
Standards' workshop**

Felix Naumann
Hasso-Plattner-Institut

slido



Please fill in the survey

① Start presenting to display the poll results on this slide.

- ▶ CS PhD in Information Quality and Information Integration
- ▶ Research at Humboldt University, IBM, AT&T, QCRI, SAP
- ▶ Chair for “Information Systems”
at Hasso Plattner Institute and University of Potsdam

- ▶ Data Profiling: Measuring data quality
 - ▶ Dependency discovery
- ▶ Data Cleaning: Improving data quality
 - ▶ Duplicate detection
 - ▶ Data preparation
 - ▶ Change exploration



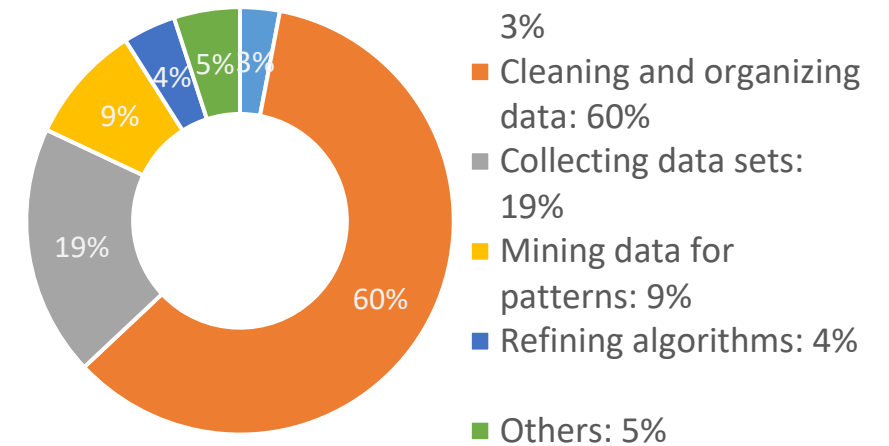
<https://www.vde.com/kitqar>

Challenges Faced & Solutions

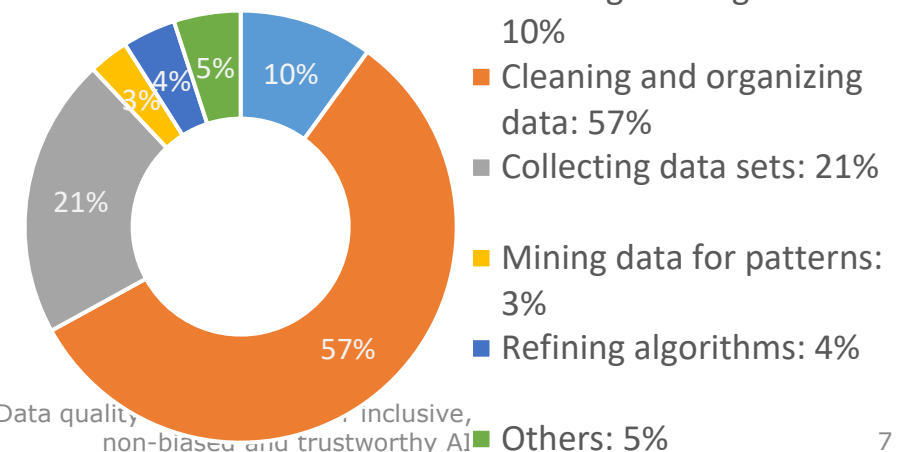
- ▶ Bad files → bad data → bad results
 - ▶ Path of least resistance
- ▶ First challenge: Detect and measure problems
 - ▶ Novel AI-specific quality dimensions
- ▶ Second challenge: Prepare and clean data for AI
- ▶ Third challenge: Transparent documentation
 - ▶ Data quality measures
 - ▶ Data cleaning steps



What data scientists spend the **most time** doing?



What is the **least enjoyable** part of data science?



Way Forward, Next Steps



► Data quality dimensions

► Which established dimensions are relevant?

- Based on learning task, pipeline stage, domain

► Which new dimensions are needed?

- Diversity, privacy, bias, liability, explainability, ...

► Assessment and explanation of data quality

► Which dimensions are (automatically) assessable/testable?

► Can we efficiently measure data quality?

► Can we correlate model errors with data quality problems?





#Standards4AI

**'Putting Science Into
Standards' workshop**

Emmanuel Kahembwe
VDE

Background



- ▶ Education: PhDs in AI & Robotics
- ▶ Research:
 - ▶ Amazon Alexa AI Prize
 - ▶ Multimodal datasets
- ▶ Professional:
 - ▶ CEO @ VDE (UK&I)
 - ▶ Chief AI Architect @ VDE e.V.
 - ▶ Standardization: StandICT EUOS, BSI (ART/1), OECD.AI
- ▶ European Projects:
 - ▶ AI Trust Standard/Label
 - ▶ AI Quality & Testing Hub

Challenges Faced & Solutions



► Web Scrapping:

► Copyright & Provenance

► Multimodality

- Alt-text
- Malignant stereotypes
- Search Engine Bias

► Curation

- RTBF
- Illicit material
- Inclusivity

► Documentation

► Access

- Difficult/expensive to collect, clean and maintain datasets

Challenges Faced & Solutions



- ▶ Documentation
 - ▶ Datasheets (and Model Cards)
- ▶ Auditing
 - ▶ Checklists
 - ▶ AI Trust Label/Standard
- ▶ Algorithmic Tools
 - ▶ Shapley Values
 - ▶ Influence functions
 - ▶ Filtering (and deletion) tools
 - ▶ Labelling tools

Way Forward, Next Steps



►MISSING:

- A clear roadmap or set of standards/guidelines on the large scale collection and curation of web-scraped data.
- How should such data be collected, stored, accessed and used within AI.

►Next Steps:

- Aligning current AI practices with respect to existing laws and regulations.
- A set of standard guidelines and metrics for the collection and curation of web-scraped data.
- A set of standard tools to aid with data curation and documentation



#Standards4AI

'Putting Science Into
Standards' workshop

Thank you!



#Standards4AI

**'Putting Science Into
Standards' workshop**

Kasia Chmielinski

**CO-FOUNDER, DATA NUTRITION PROJECT
SHORENSTEIN CENTER, HARVARD KENNEDY SCHOOL OF
GOVERNMENT**

Professional background



► *Industry: (2007-2021)*

Building algorithmic systems



Digital/McKinsey

► *Research: (2017-current)*

Dataset standards, documentation as a transparency mechanism



Challenges



► Incentives [**Why should I document?**]

Key insight: Internal momentum must be paired with clear expectations and support from leadership, including real power to adjust existing org structures

► Usability / Usefulness [**How does this fit into my existing tasks?**]

Key insight: Friction can be reduced through integrating tools into workflow and focusing on user experience (not just schema)

► Technical Challenges [**How do I document?**]

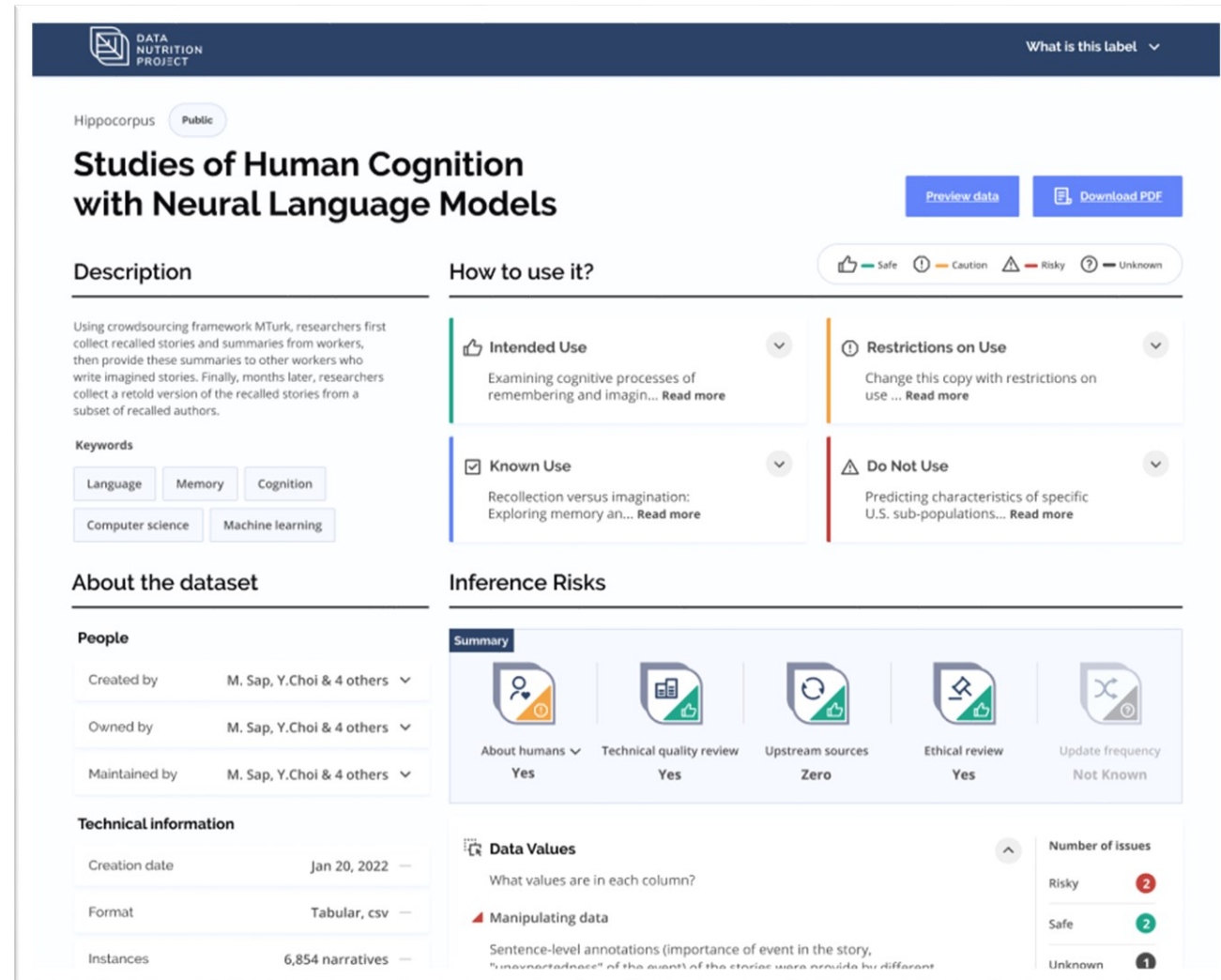
Key insight: Large, unstructured streaming datasets are very challenging to document and will require new tools and approaches. Additionally, there will need to be different standards for different data domains.

► Engagement [**How do we engage communities and create culture?**]

Key insight: This will require balancing values, e.g. community values may be incompatible with open data / science principles

Opportunities

- **Legibility:** Provide standardized, accessible dataset documentation
 - “Nutrition Label” for Datasets (Fall 2022)
 - Streamlined for two pathways: **use** the data and **understand** the data
 - Strong user experience focus for easy reading and comparing
- **Ecosystem:** Leverage existing knowledge structures
 - Impact Assessments, Datasheets, Model Cards, FactSheets, etc
 - Qualitative information prioritized, especially provenance and domain knowledge
- **Impact:** Short and Long Term
 - Labels can be consulted individually, comparatively
 - Future integrations with certification programs
 - Drives consumer expectations of data quality transparency even when Label isn’t present



Hippocorpus Public

Studies of Human Cognition with Neural Language Models

[Preview data](#) [Download PDF](#)

Description

Using crowdsourcing framework MTurk, researchers first collect recalled stories and summaries from workers, then provide these summaries to other workers who write imagined stories. Finally, months later, researchers collect a retold version of the recalled stories from a subset of recalled authors.

Keywords

Language Memory Cognition
Computer science Machine learning

How to use it?

Intended Use
Examining cognitive processes of remembering and imagin... [Read more](#)

Known Use
Recollection versus imagination: Exploring memory an... [Read more](#)

Restrictions on Use
Change this copy with restrictions on use ... [Read more](#)

Do Not Use
Predicting characteristics of specific U.S. sub-populations... [Read more](#)

About the dataset

People

Created by M. Sap, Y.Choi & 4 others
Owned by M. Sap, Y.Choi & 4 others
Maintained by M. Sap, Y.Choi & 4 others

Technical information

Creation date Jan 20, 2022
Format Tabular, csv
Instances 6,854 narratives

Inference Risks

Summary

About humans: Yes
Technical quality review: Yes
Upstream sources: Zero
Ethical review: Yes
Update frequency: Not Known

Data Values

What values are in each column?

Manipulating data

Sentence-level annotations (importance of event in the story, "manipulatedness" of the event) of the stories were provided by different

Number of issues

Risky: 2
Safe: 2
Unknown: 1



#Standards4AI

'Putting Science Into
Standards' workshop

Thank you!



#Standards4AI

**'Putting Science Into
Standards' workshop**

Flora DELLINGER
Valeo, Confiance.ai

Professional background



► **Flora Dellinger** (Valeo, Confiance.AI)

- PhD in computer vision and image processing
- Machine Learning engineer in the industry



Development of AI-based camera perception modules for driving assistance systems.



Confiance.AI: French consortium to design and industrialize trustworthy AI-based critical systems [2021-2024].

Leader of the project **“Trust by data”**: development of methods and tools to obtain trustworthy and relevant datasets for AI.



Behaviour of AI components is difficult to assess !

- Datasets are not representative enough of encountered real-world situations (biases, domain gap, corner cases...).
- Datasets lack quality and integrity over lifecycle.
- Metrics to evaluate AI models are too generic and research oriented.



How to build relevant datasets for a specific use case? How to ensure quality of a dataset?



Create a **methodology for data specification and data collection** activities

- To guide processes
- To ensure compliance with input requirements (safety, functional, operational design domain)
- Inspired by work done by ASAM (OpenODD, OpenScenario)



Propose and develop **metrics for trustworthiness datasets**

- To measure representativity, diversity, traceability, compliance...



Develop **tools to process and analyse datasets**

- To facilitate data processing and get insights on datasets
- Fully integrated in a platform to ensure data integrity.



Way Forward, Next Steps



►Today:

- No standards or methodology for data creation, rely mainly on engineers experience.

►Tomorrow, we need to develop new standards for:

►Data format and label

- To facilitate data exchanges and processes.

►Dataset design and collection

- To ensure relevance of data and to guide data acquisition step.

►Data quality

- To monitor datasets and provide trust in AI components.



#Standards4AI

**'Putting Science Into
Standards' workshop**

Thank you!

Thank you for joining us today

See you tomorrow!

