Indexing Validation Datasets
Technical Working Group Meeting, Fall 2023

Agnes Cameron, agnescam@mit.edu
the iiindex

• Now 2 years old!
• A repository of pointers to innovation data, code, and tools
• Builds from a google sheet
• Community-edited, collaboratively annotated, open-source, fully versioned
iiindex 2.0

- Rebuilt to include:
  - Advanced search
  - Relationships between datasets
  - Broader set of datasets, tools and guides
question:

How to adequately credit datasets as contributions to innovation research?
indexing validation data

- Validation datasets are tools in their own right
- In this community, often bespoke and project-specific
- Particularly used for machine learning projects, and in the construction of new datasets
- Expensive to produce!
existing models

• Machine learning community has a culture of counting datasets + validation explicitly as a contribution
• HuggingFace, Kaggle and PapersWithCode provide good models for thinking about this
Huggingface explicitly names tasks that datasets and models are used for.

This structure is helpful for filtering, but also for naming!
what are the tasks?

• Went through the iiindex and classified all ‘Machine Learning’ related projects by task
• CSV and code here
• Almost all text-based tasks!
• Most common ML tasks:
  • semantic analysis
  • matching / reconciliation
• Most common *innovation-specific* tasks:
  • name disambiguation
  • patent similarity
what are the gaps?
filling the gaps

• ‘Naming what’s not yet there’
  -> is there a validation dataset you want to bring into being?

• Crediting and promoting what’s there already!
we would love your help with this!

• contribute to validation tasks sheet!
• Contribute to tagging projects with released validation data!
• Edit the validation guide
• Get in touch -> agnescam@mit.edu