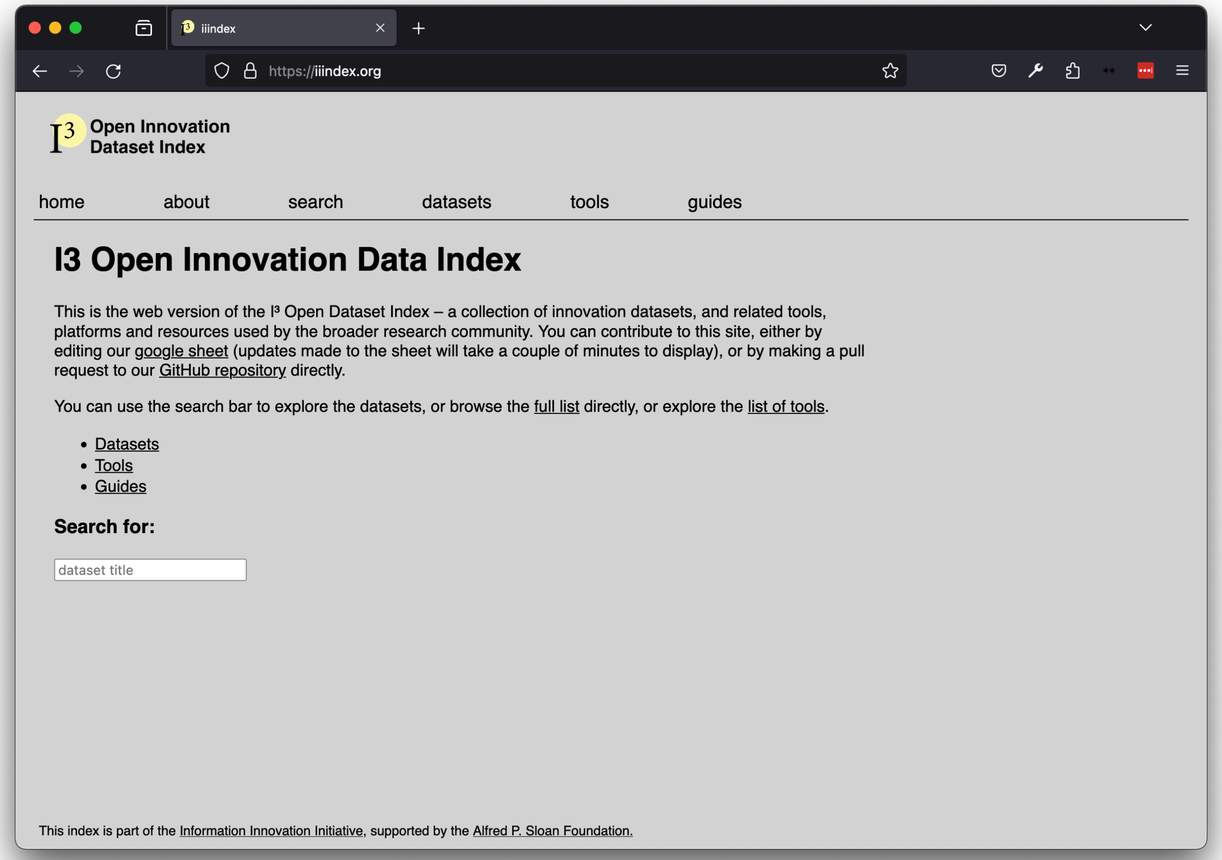
Indexing Validation Datasets

Technical Working Group Meeting, Fall 2023

Agnes Cameron, agnescam@mit.edu

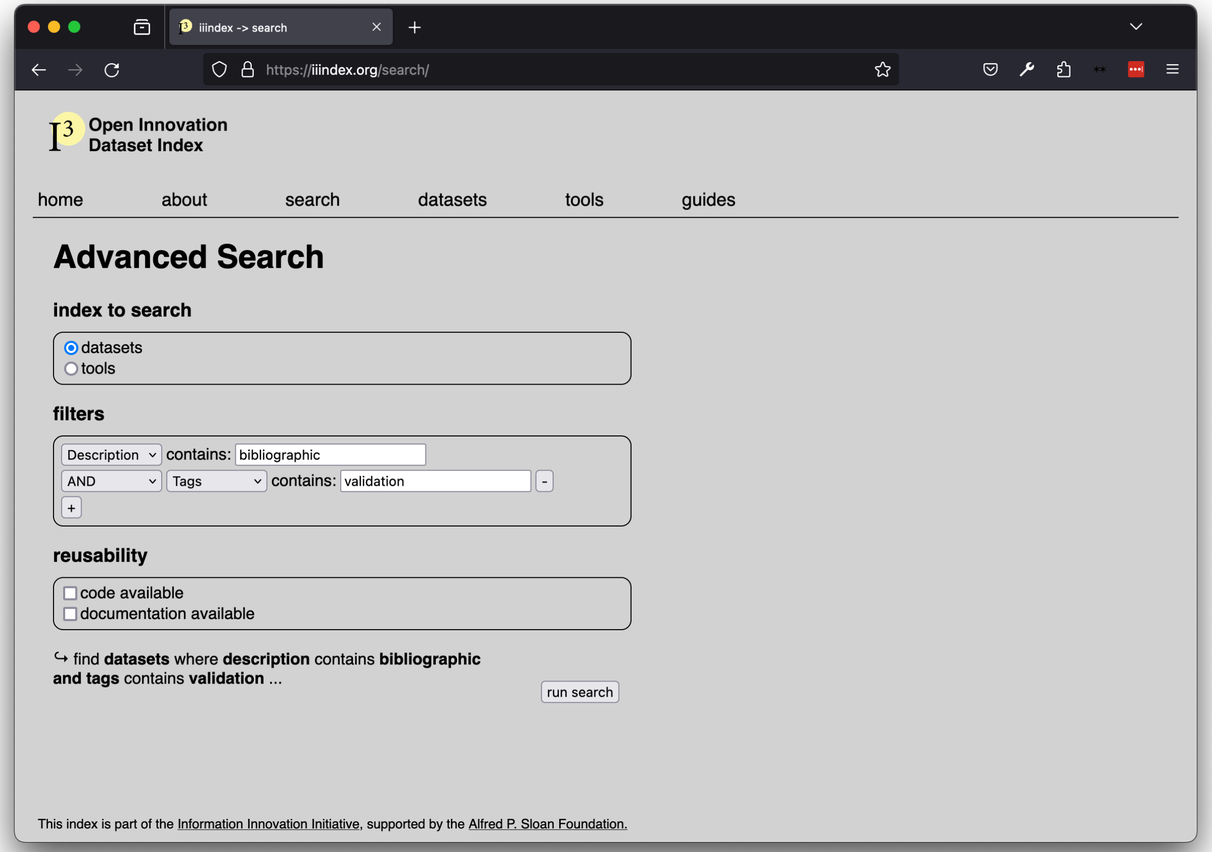
the iiindex

- 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

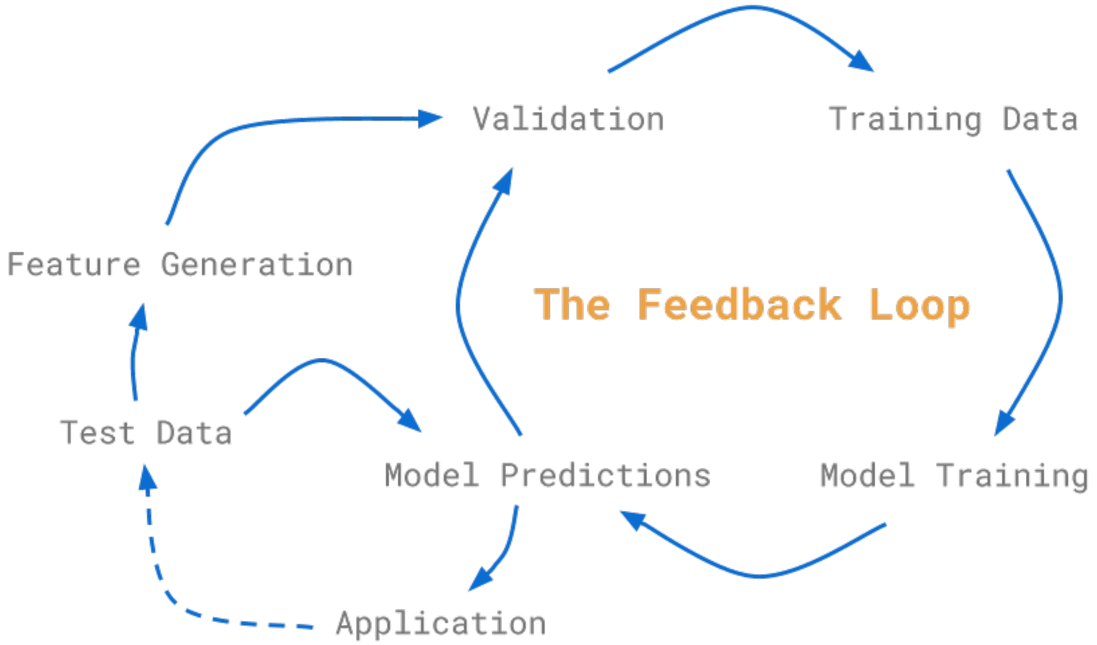


question:

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

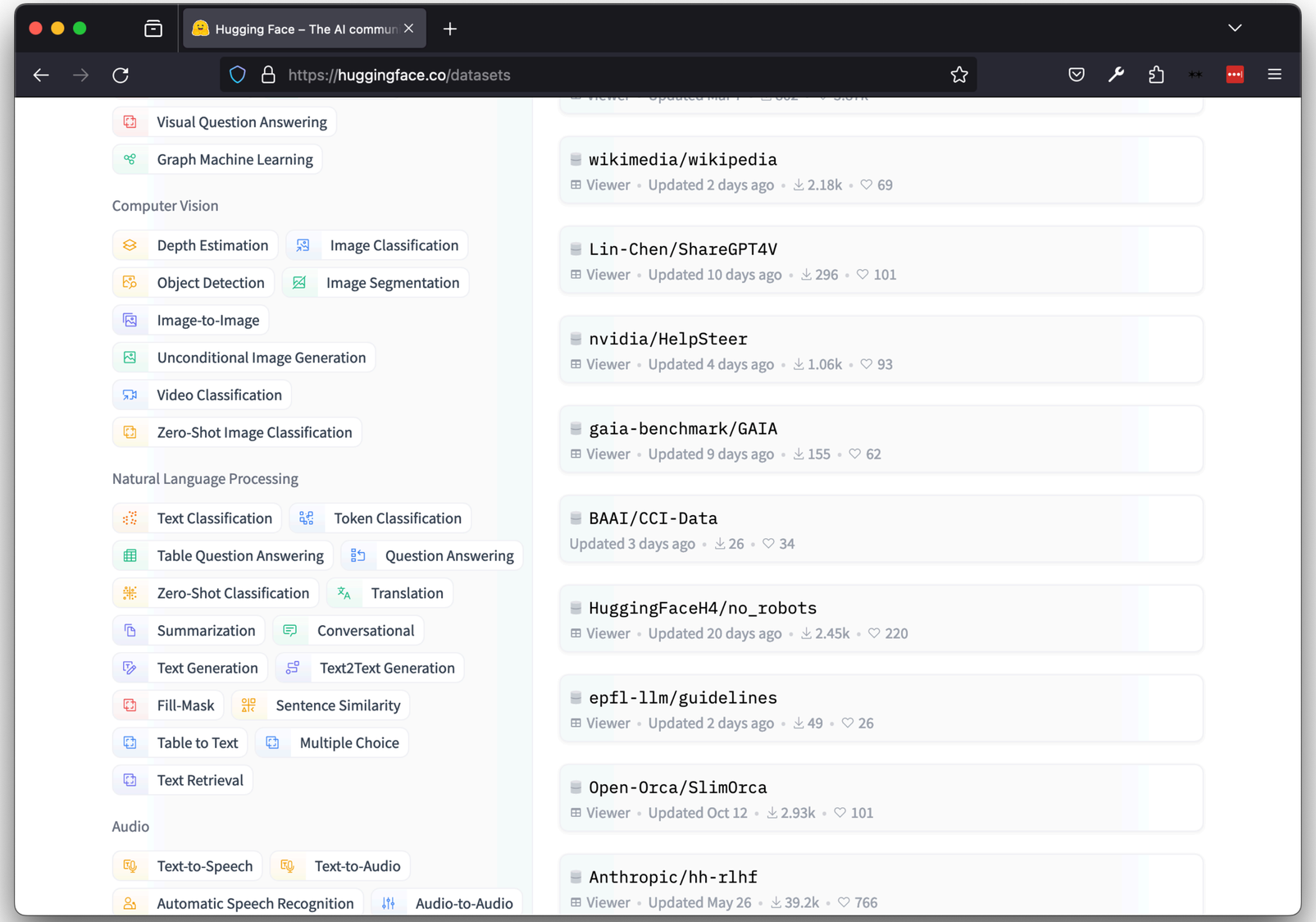
Hugging Face

kaggle

Papers With Code

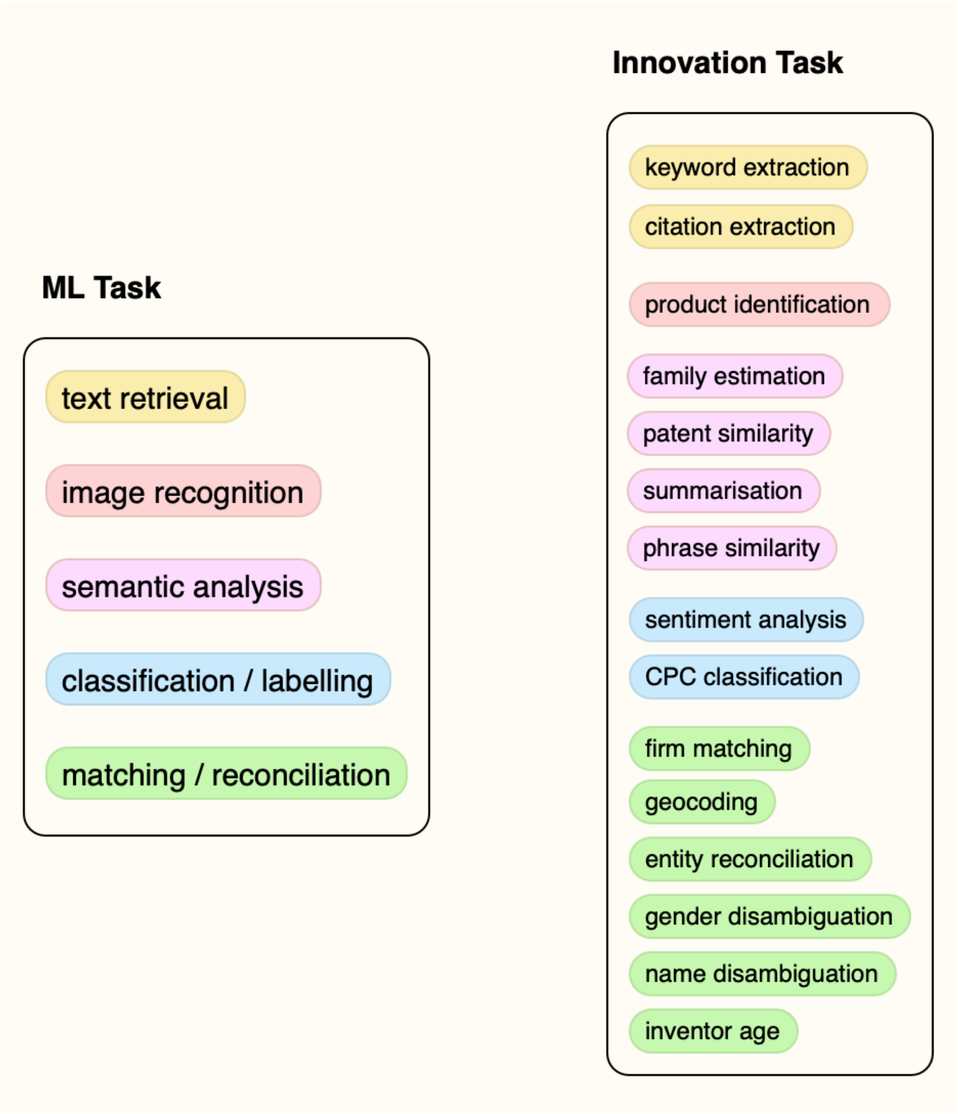
naming!

- Huggingface explicitly names tasks that datasets and models are used for
- This structure is helpful for filtering, but also for naming!

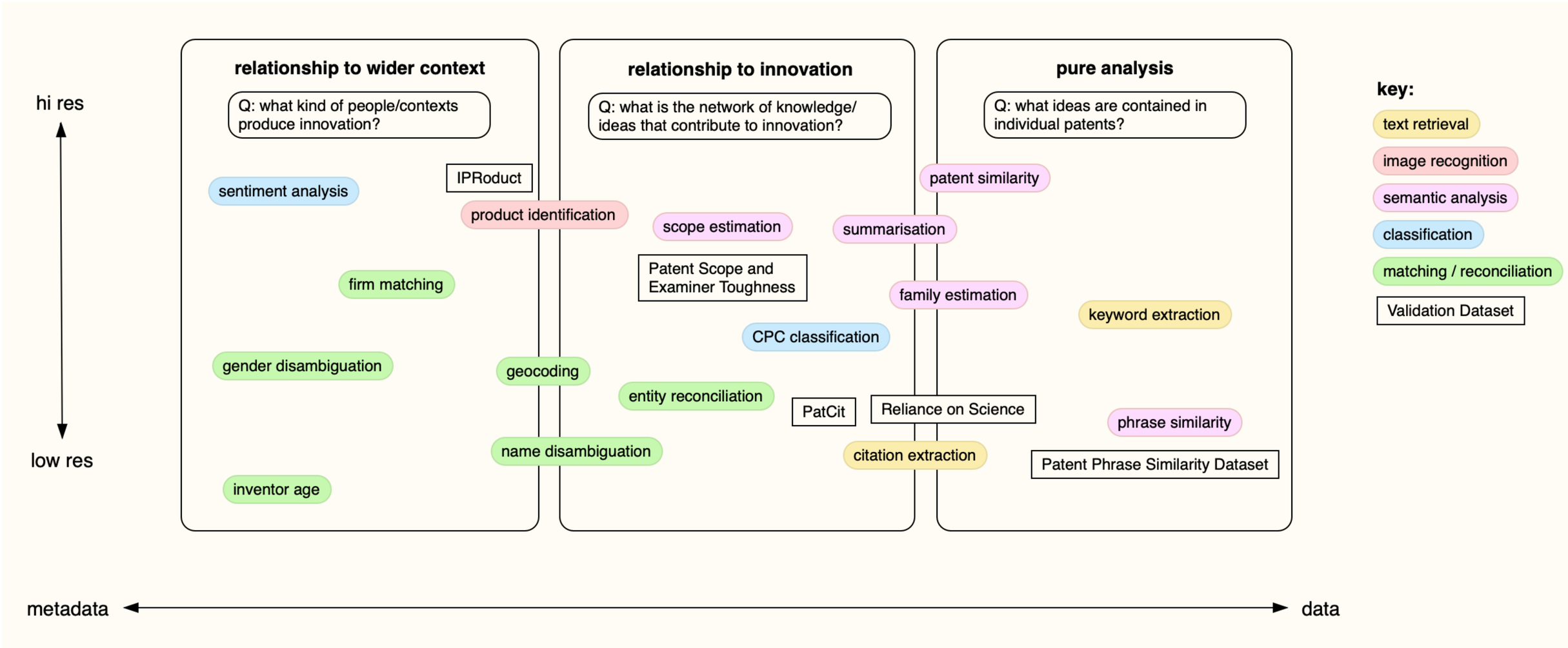


what are the tasks?

- 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 anaylsis
  - matching / reconciliation
- Most common *innovation-specific* tasks:
  - name disambiguation
  - patent similarity

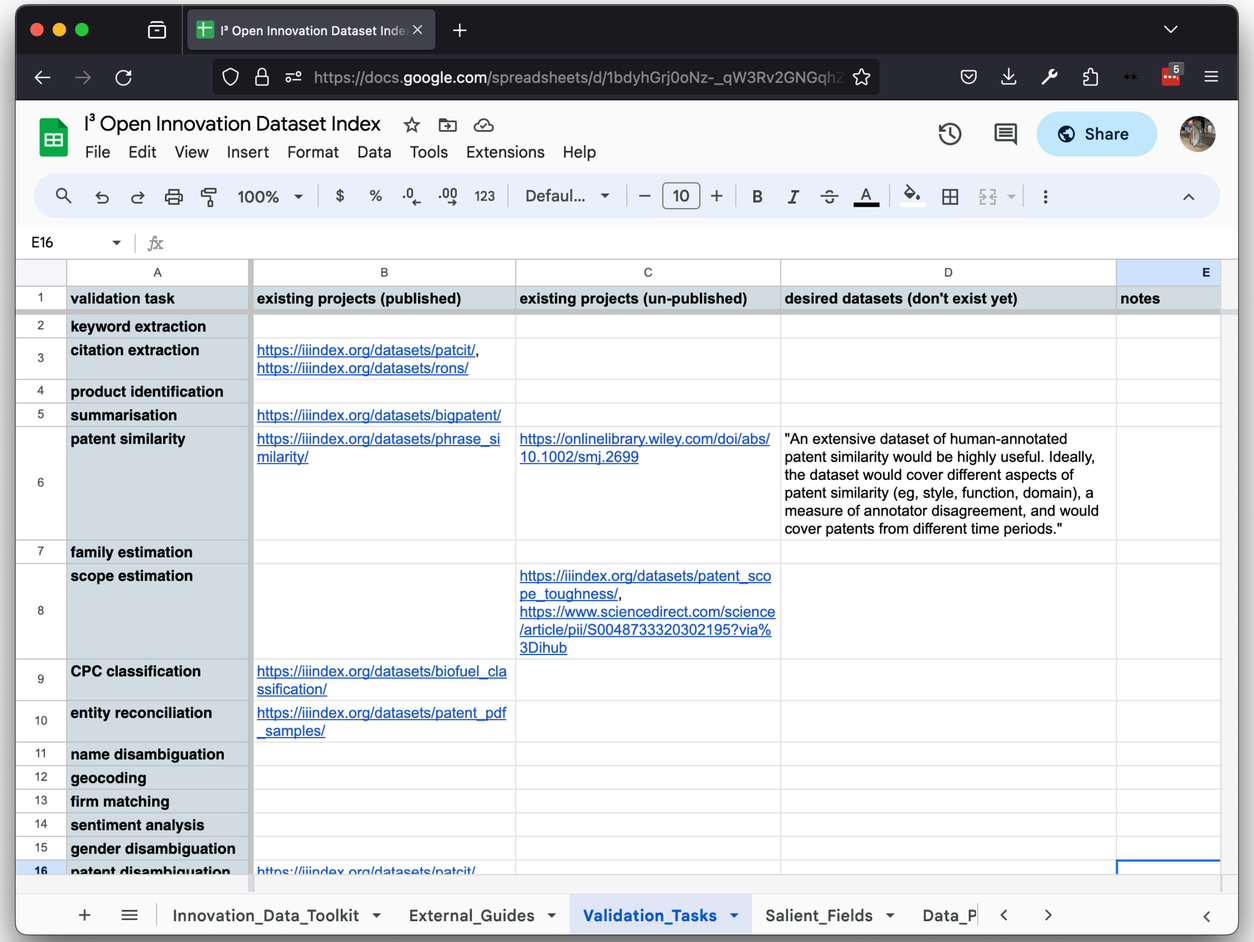


what are the gaps?



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!



I³ Open Innovation Dataset Index

	A	B	C	D	E
1	validation task	existing projects (published)	existing projects (un-published)	desired datasets (don't exist yet)	notes
2	keyword extraction				
3	citation extraction	https://iiindex.org/datasets/patcit/, https://iiindex.org/datasets/rons/			
4	product identification				
5	summarisation	https://iiindex.org/datasets/bigpatent/			
6	patent similarity	https://iiindex.org/datasets/phrase_similarity/	https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.2699	"An extensive dataset of human-annotated patent similarity would be highly useful. Ideally, the dataset would cover different aspects of patent similarity (eg, style, function, domain), a measure of annotator disagreement, and would cover patents from different time periods."	
7	family estimation				
8	scope estimation		https://iiindex.org/datasets/patent_scope_toughness/, https://www.sciencedirect.com/science/article/pii/S0048733320302195?via%3Dihub		
9	CPC classification	https://iiindex.org/datasets/biofuel_classification/			
10	entity reconciliation	https://iiindex.org/datasets/patent_pdf_samples/			
11	name disambiguation				
12	geocoding				
13	firm matching				
14	sentiment analysis				
15	gender disambiguation				
16	patent disambiguation	https://iiindex.org/datasets/patcit/			

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