Curathon

PRIDE repository

05/05/26

Arnaud JUNG - arnaud.jung@etu.unistra.fr

Why should we be re-using data ?

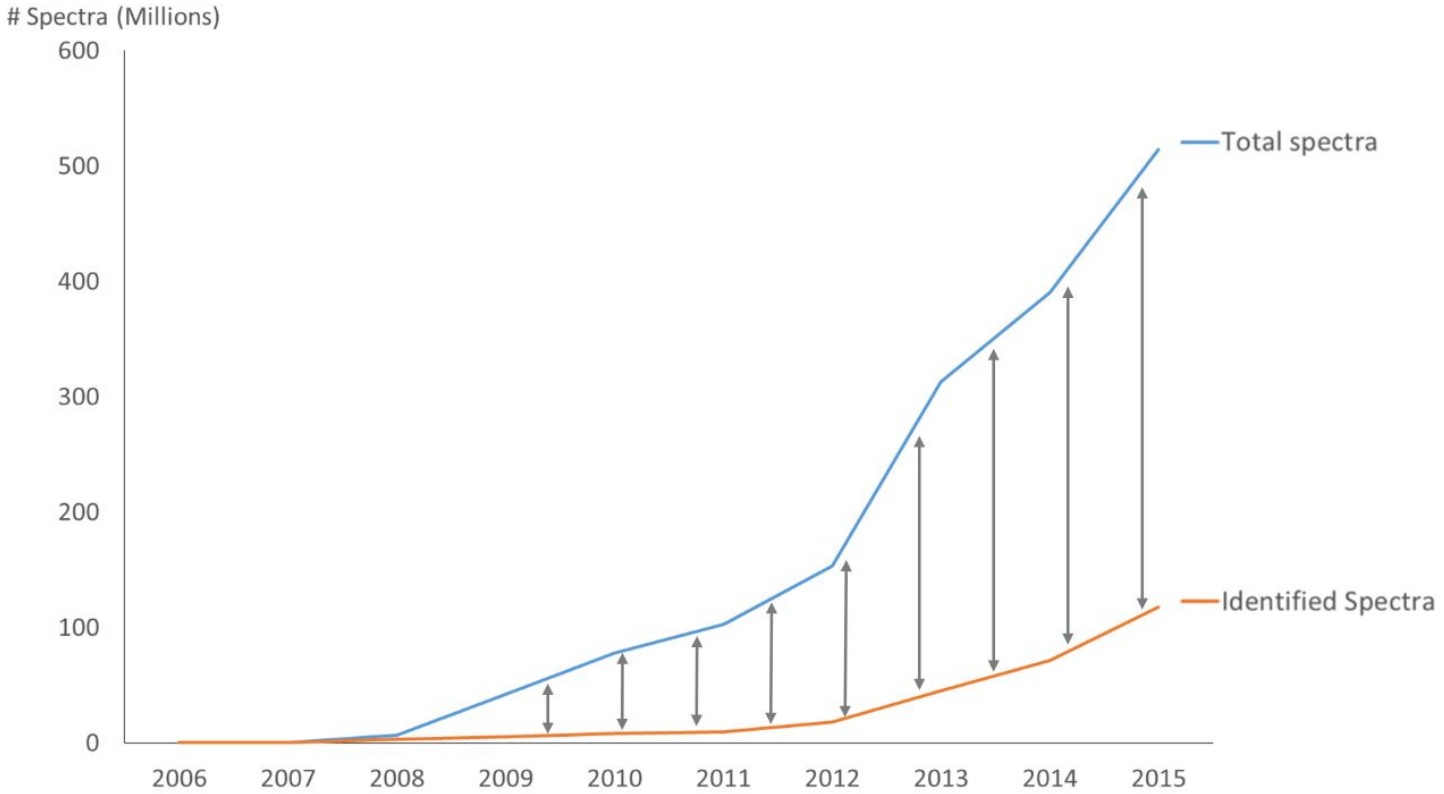
Four types of data re-use

Re-using data to build machine learning models

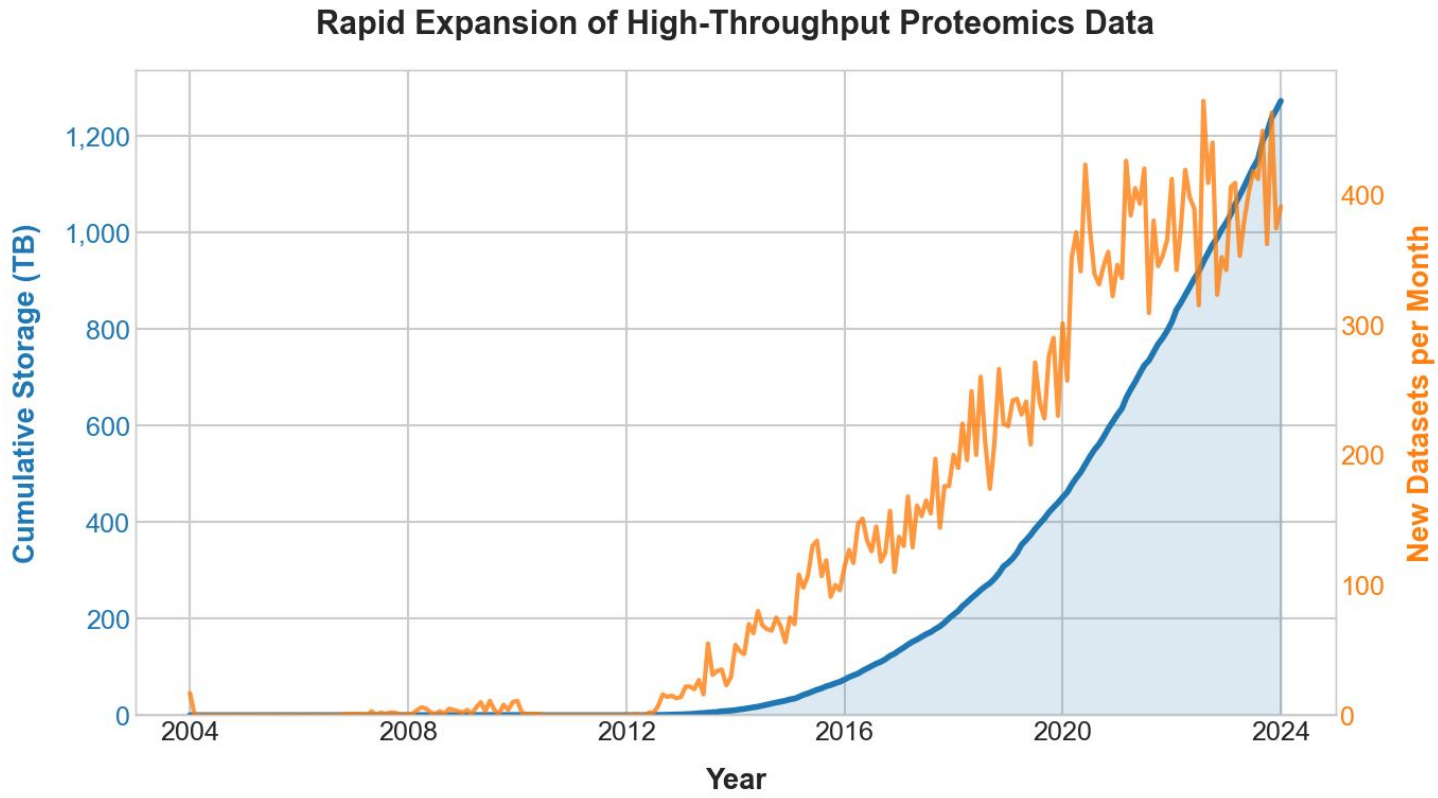
Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

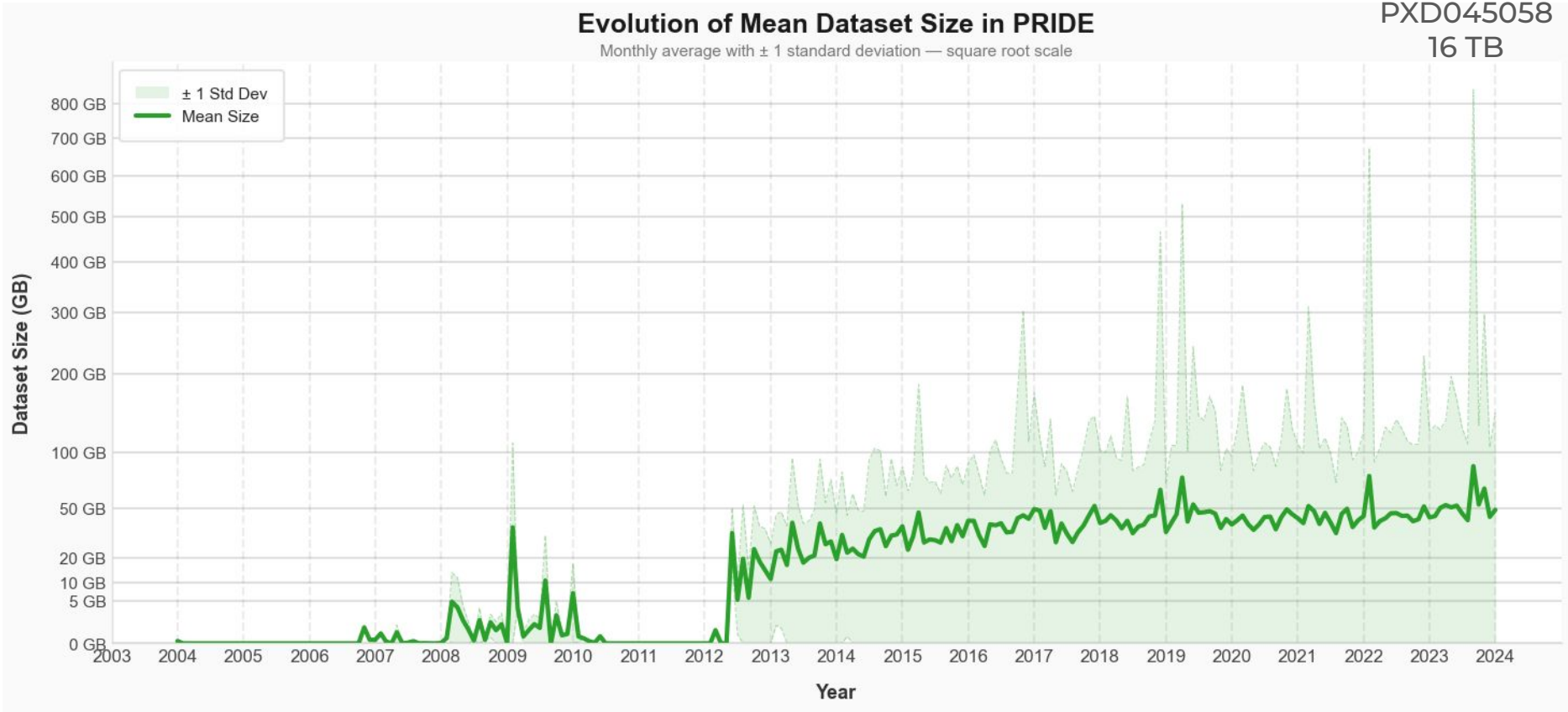
A lot of data is high-content, meaning that much more data is acquired than is used in most papers



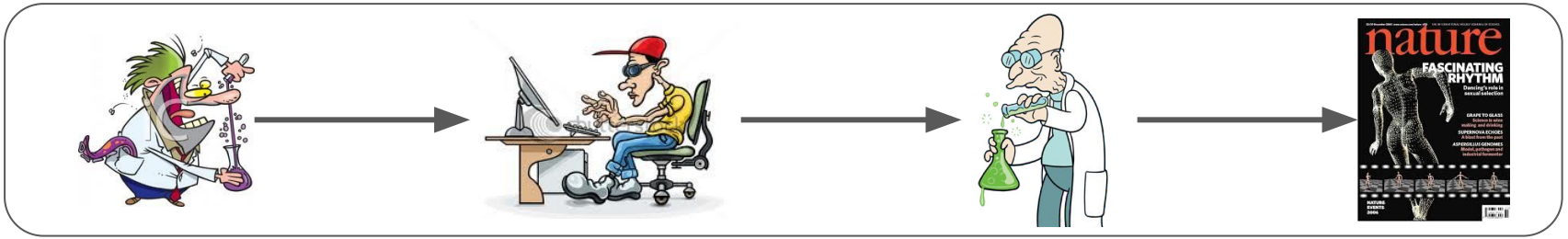
Most of the data is also high throughput, meaning there is lots of data available!

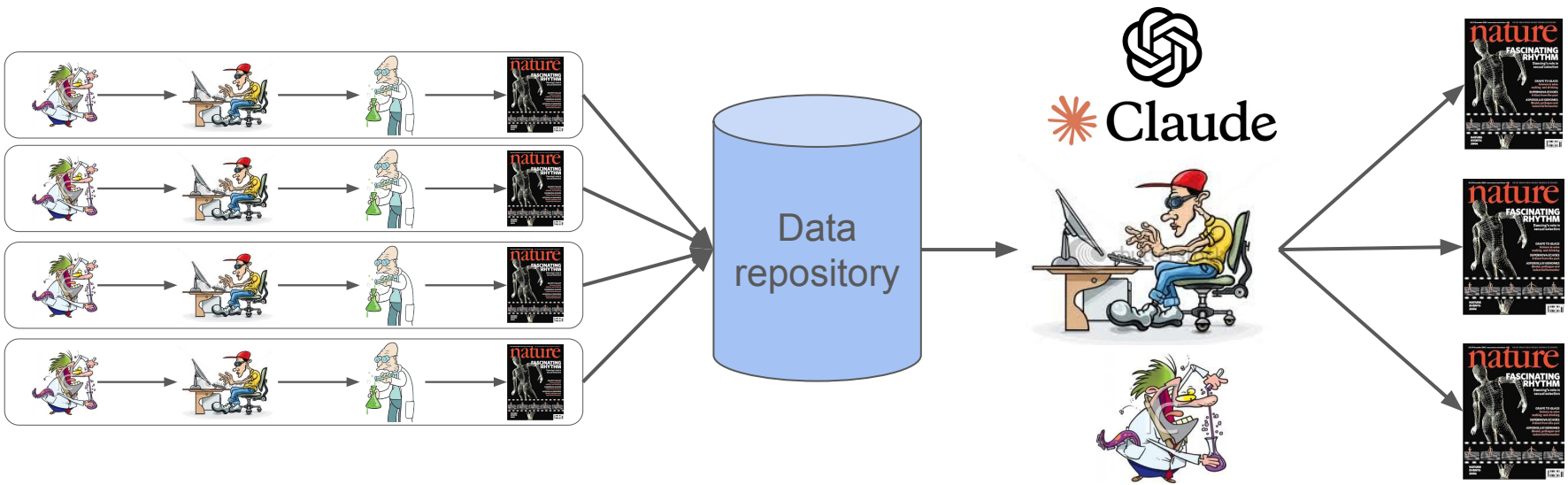


As datasets grow,
so does their informational depth

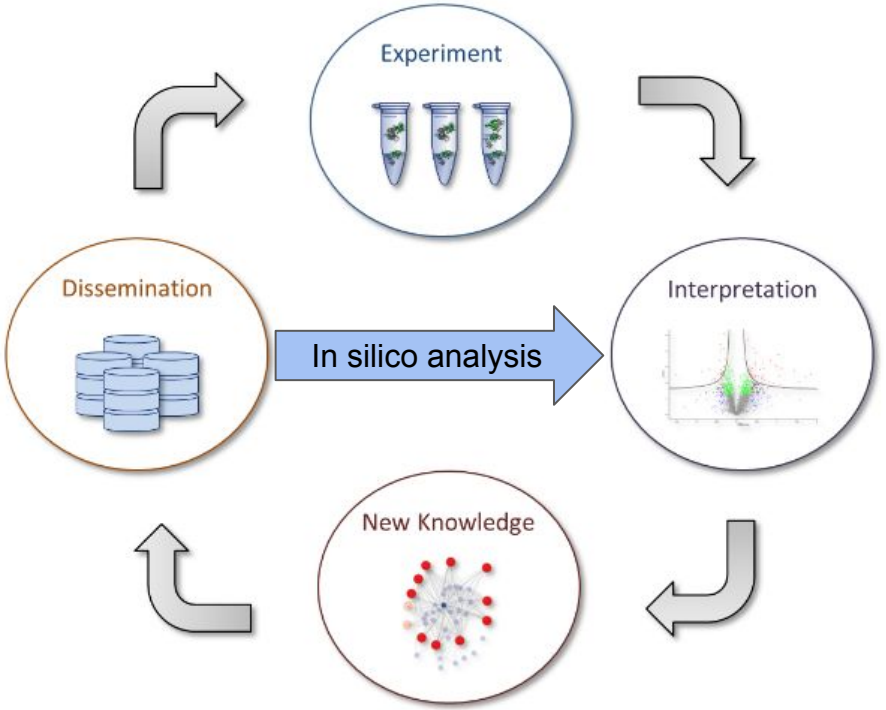


As the volume and content of data increases in a field, the role of informatics in that field changes as well





An open data exchange allows for productive (and completely novel!) data uses



https://doi.org/10.1002/pmic.201500295

Why should we be re-using data ?

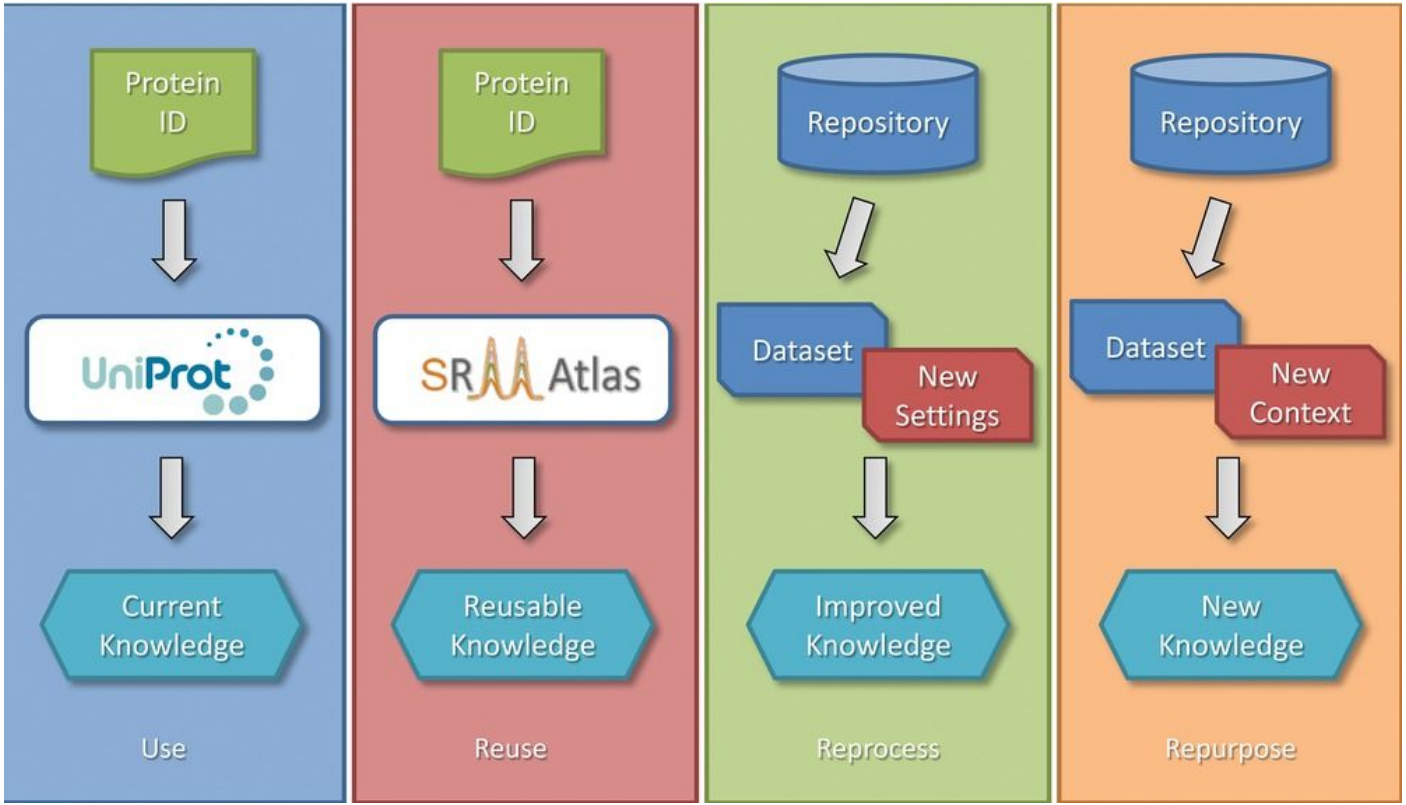
Four types of data re-use

Re-using data to build machine learning models

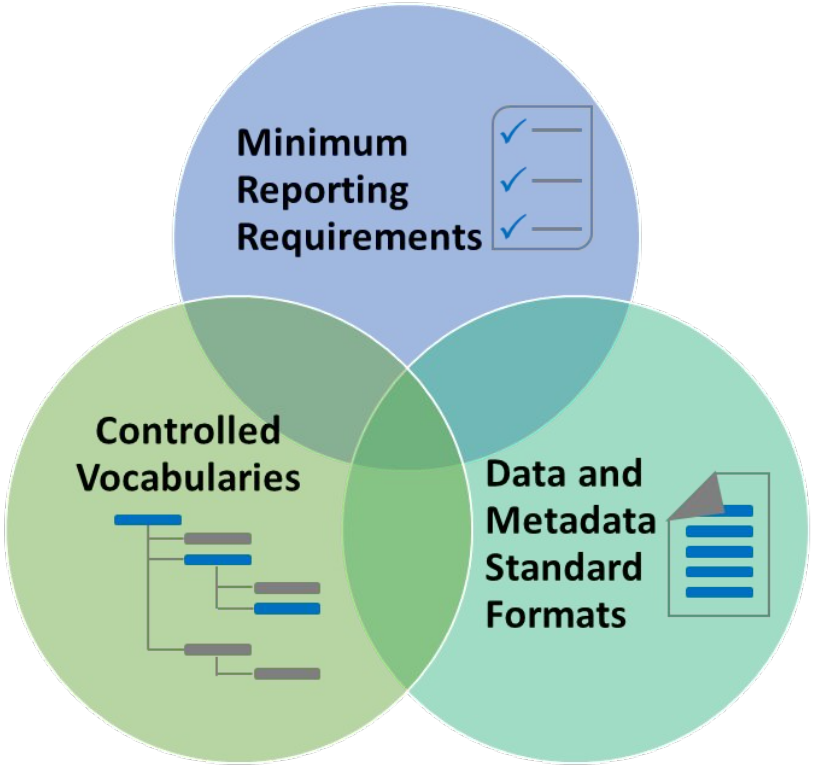
Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

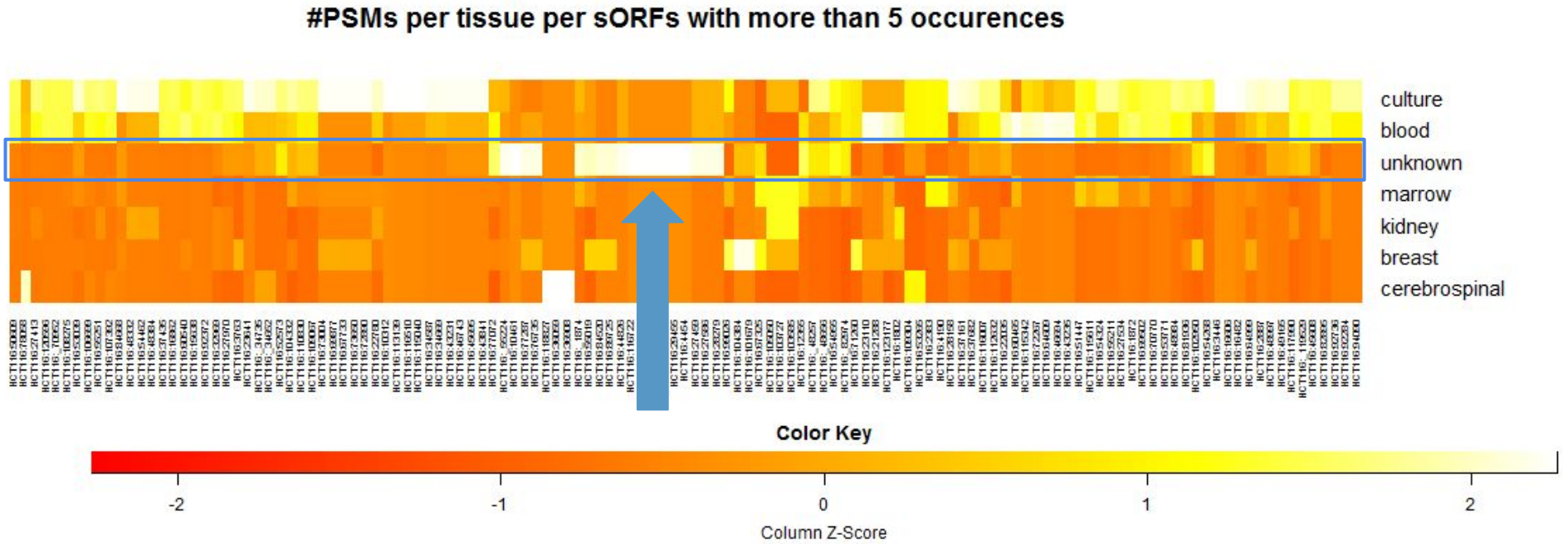
In general, data re-use can take four distinct forms, all of which are somehow applied in our examples



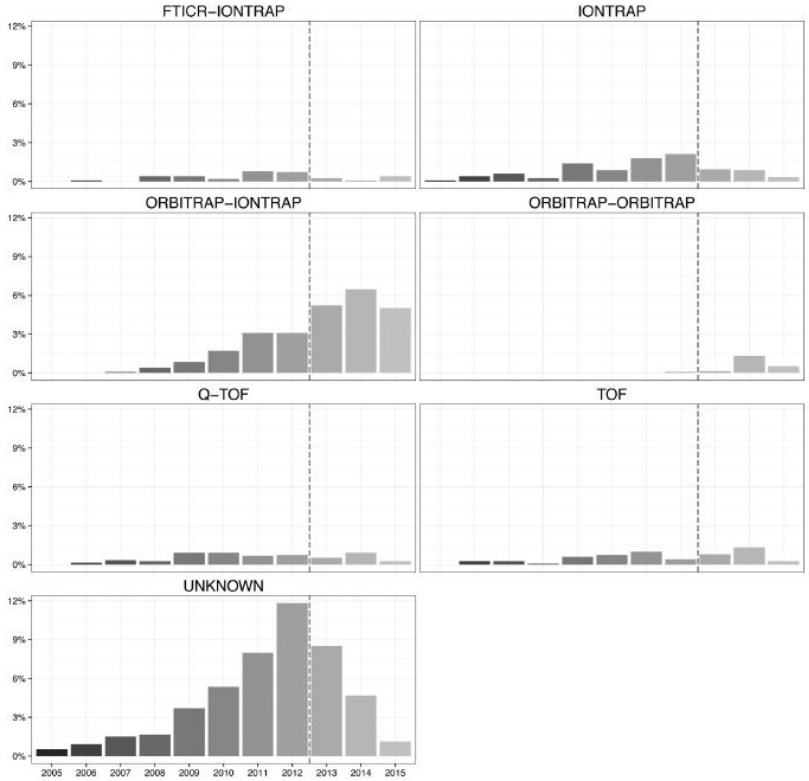
Data sharing requires three complementary building blocks: minimal requirements, controlled vocabularies and formats

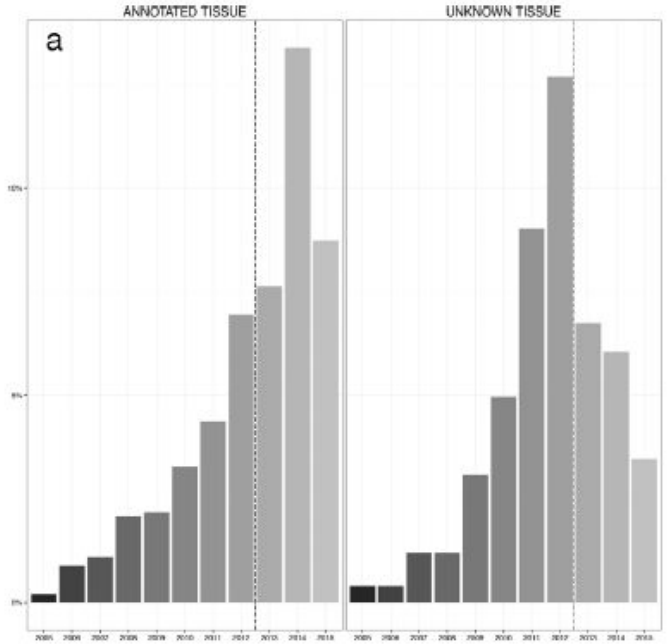


Missing metadata becomes pretty annoying when people successfully re-use your data

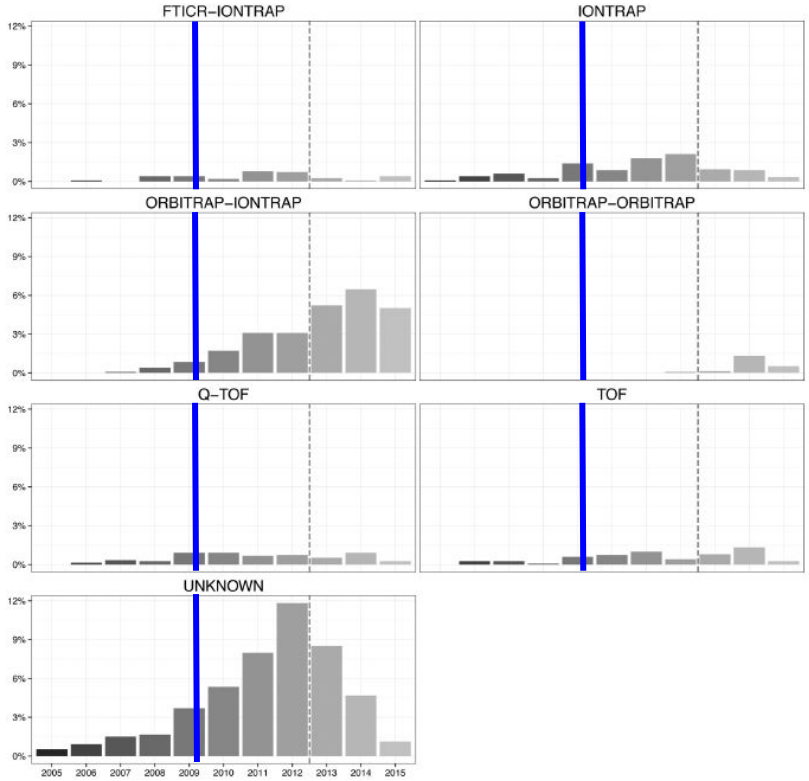


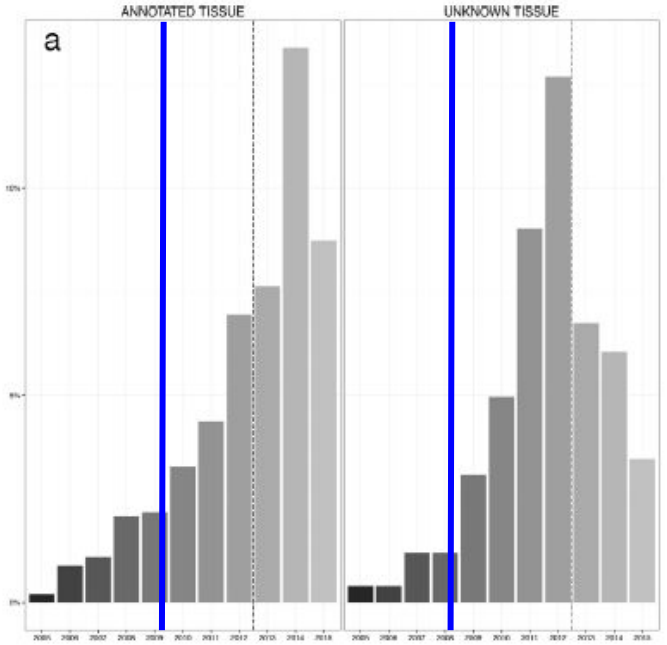
Metadata is often the key issue, as it requires manual work



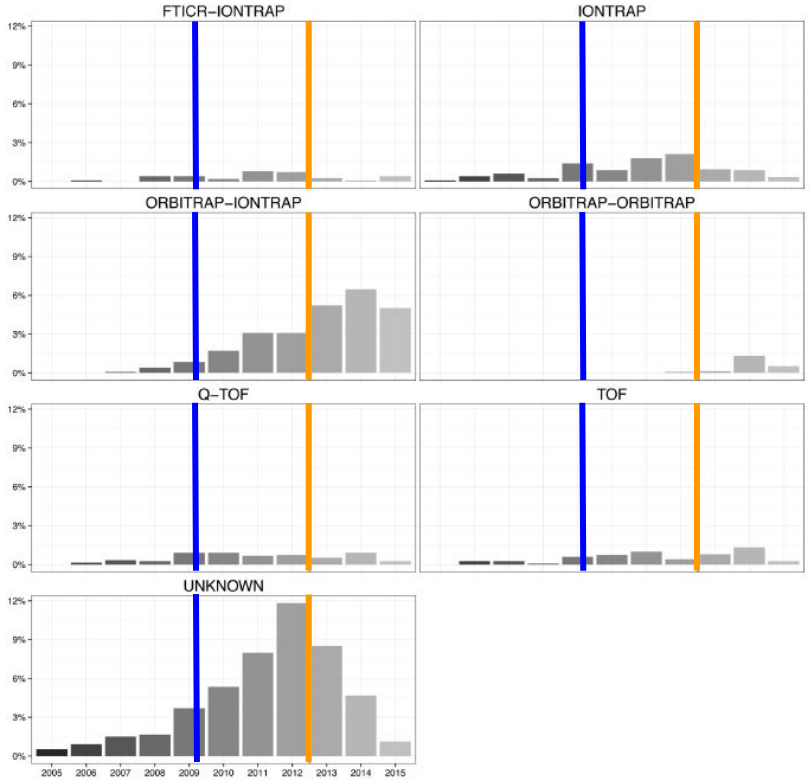


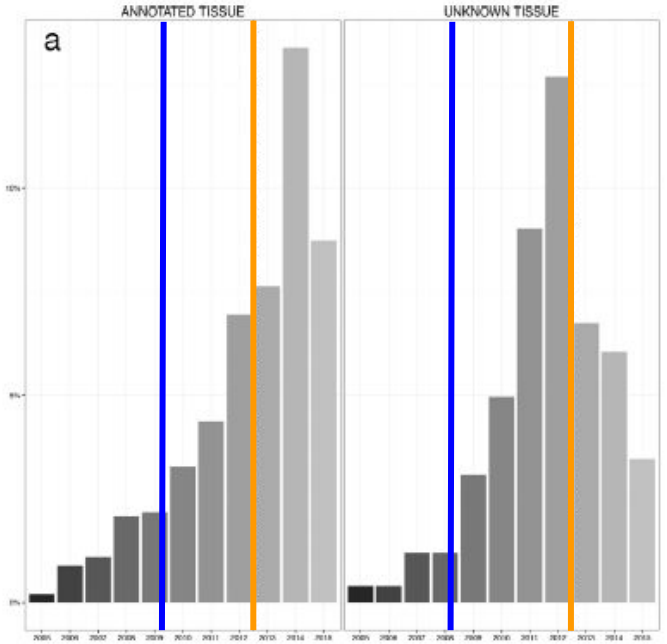
Even user-friendly submission tools cannot correct for a lack of elementary motivation





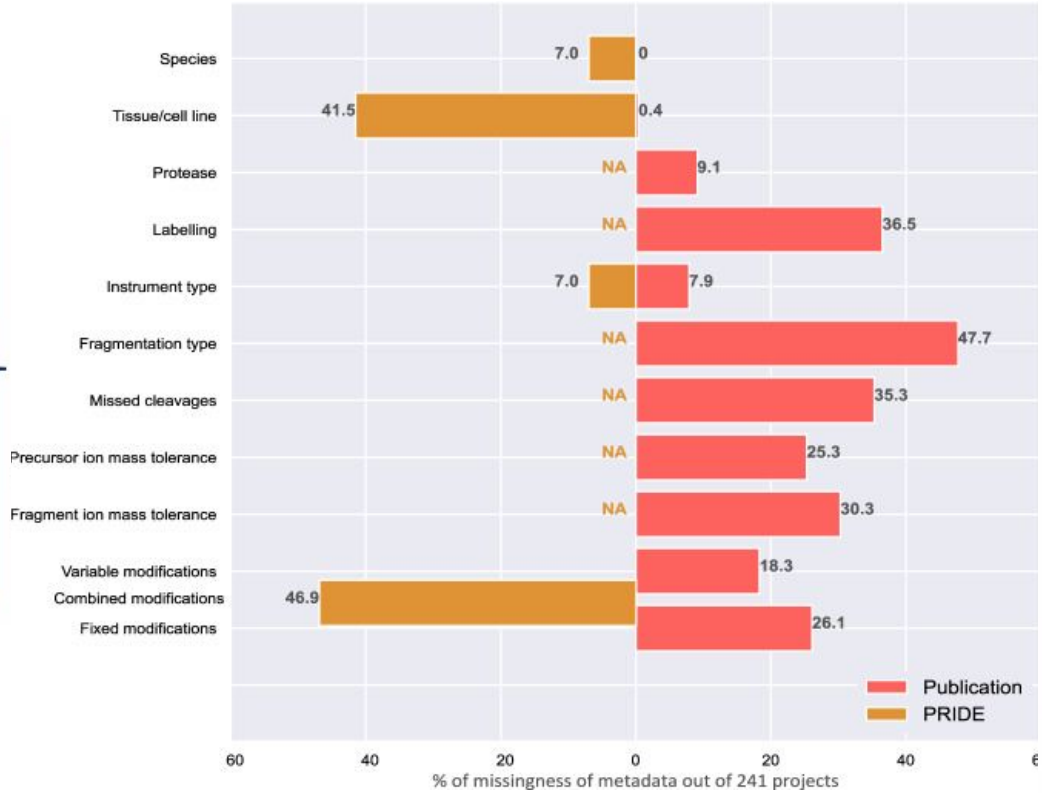
Manual curation of submission, equivalent to restrictive policing, does help





Metadata annotation in both PRIDE and articles(!) remains a major problem in proteomics

PRIDE
PRoteomics IDEntifications Database



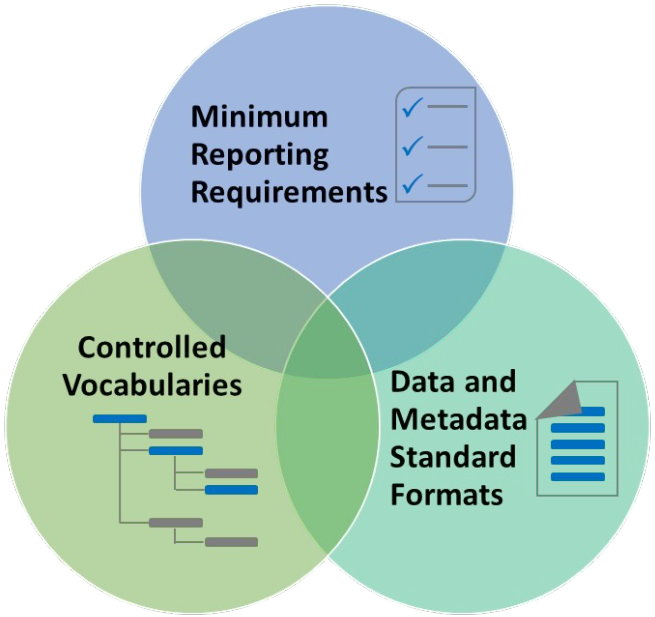
Ontologies

SDRF guidelines
Local metadata

lesSDRF
freely available
open source
no install required

10.1038/s41467-023-42543-5

As AI becomes more powerful, we can use it to annotate missing metadata and help complete reports



EuBIC NCEMS

AI

HAMLET: re-annotating PRIDE

Manuscript
Abstract + M&M → *Biological metadata*

Agentic pipeline:
Multiple LLM agents,
hallucination checks,
ontology normalization

Raw files → *Technical metadata*

.mzML extraction
De novo + Peptonizer
organism inference

We may want to take a choice of how we frame open data

Show me your data!

I don't trust you!

I'll find your mistakes!

This will not end well!

Could I look at your data?

Ok, this is pretty cool!

Look what I found in here!

Your data is so useful!

And let us not forget that your data will most likely live a much longer and more useful life than your publication!

Why should we be re-using data ?

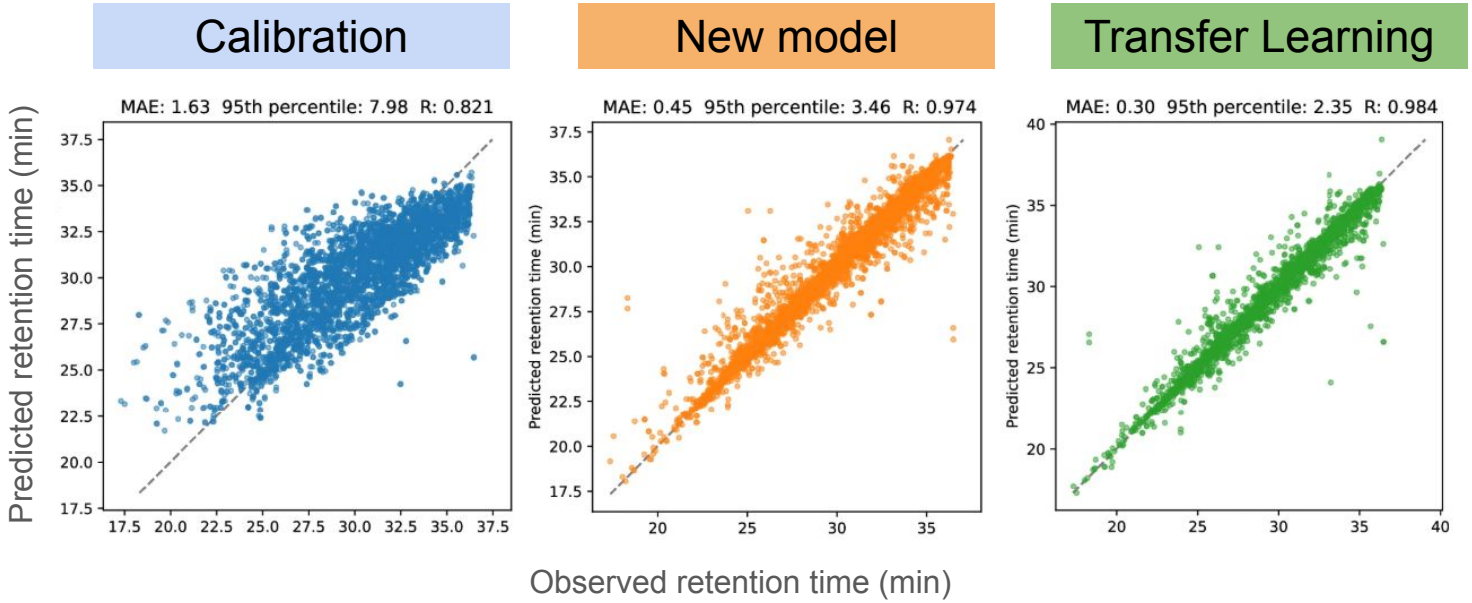
Four types of data re-use

Re-using data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

Advances in modeling enable detection of previously undetectable peptides



DeepLC

https://doi.org/10.1038/s41592-021-01301-5

By merging multiple new tools we achieve enhanced peptide identification performance

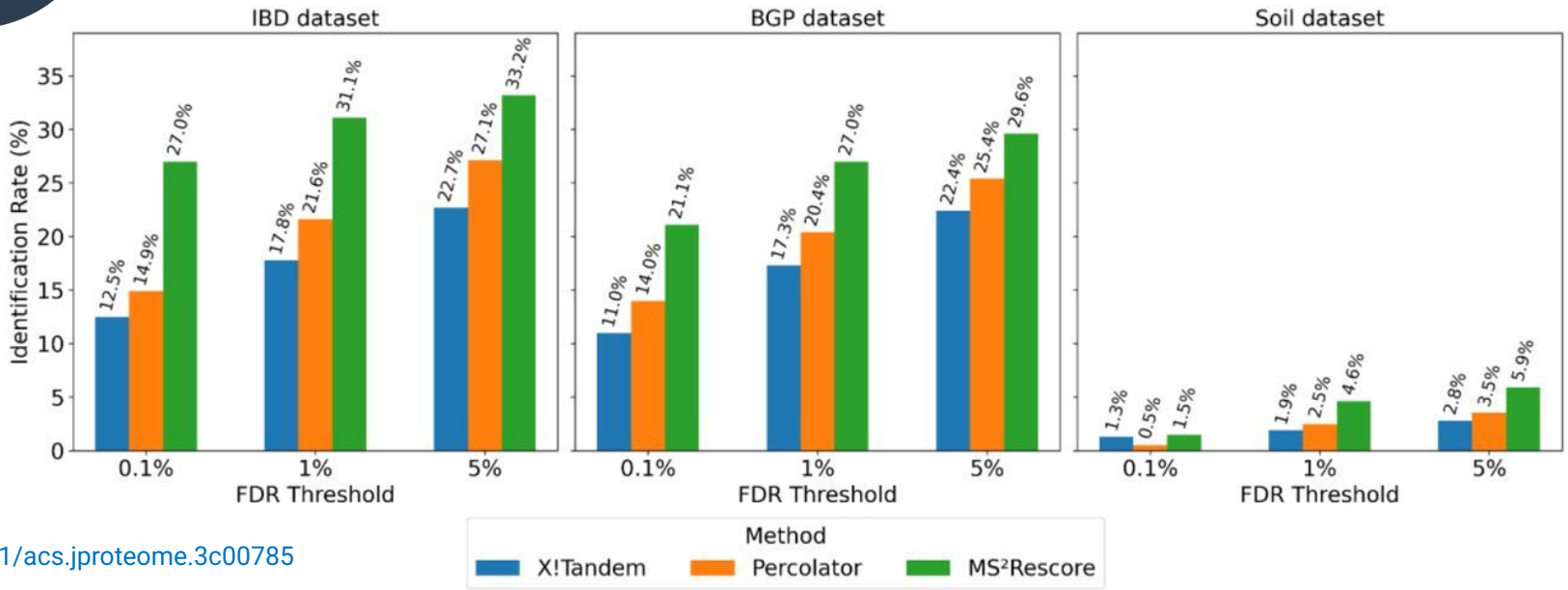
MS²Rescore

MS²PIP

DeepLC

IM2Deep

Percolator



10.1021/acs.jproteome.3c00785

Why should we be re-using data ?

Four types of data re-use

Re-using data to build machine learning models

Reprocessing data with new models for new insights

Repurposing large-scale data for new knowledge

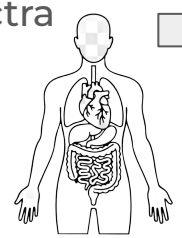
Public human and mouse data reprocessing using our tools has revealed a hidden world of protein modifications

277 mouse data sets (15%)
appr. 600 million spectra
14.5 k raw files

SAGE → MS²Rescore

2.9 million modified sites
16 808 proteins
(99% proteome coverage)

539 human datasets (5%)
appr. 924 million spectra
25 k raw files



SAGE → MS²Rescore

5.0 million modified sites
20 205 proteins
(99% proteome coverage)

And, in case you're wondering about this, here's how my PI felt like when he saw these results!

Why should we be re-using data ?

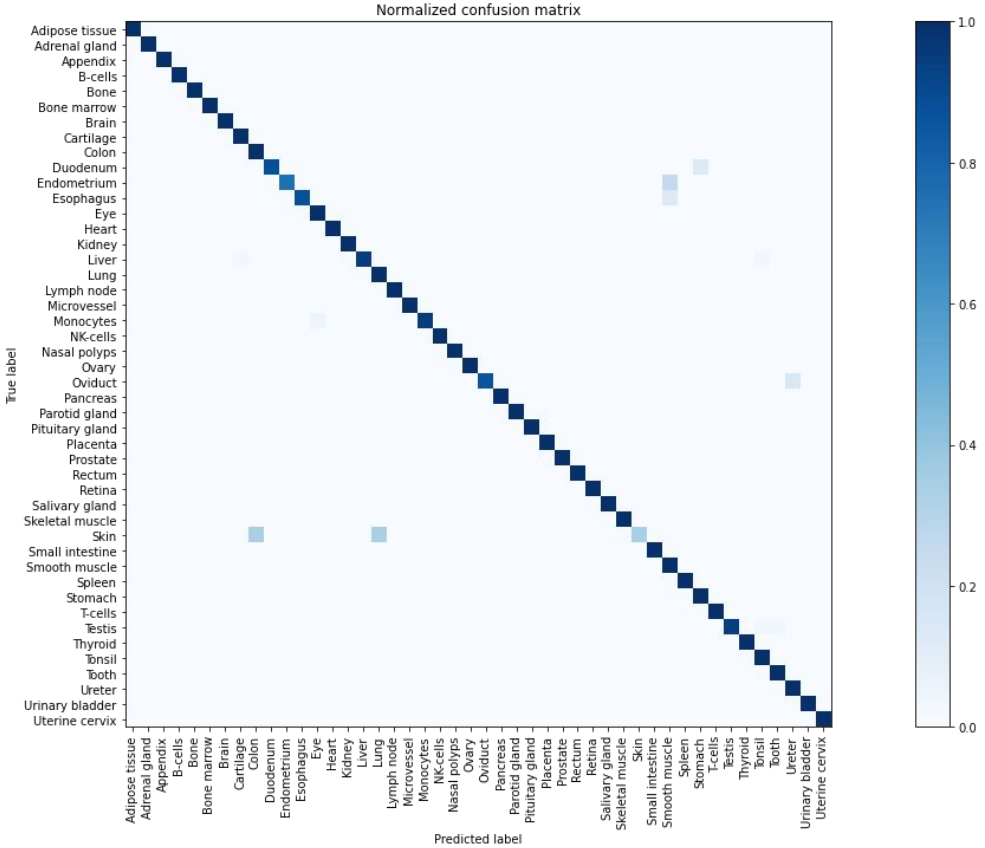
Four types of data re-use

Re-using data to build machine learning models

Reprocessing data with new models for new insights

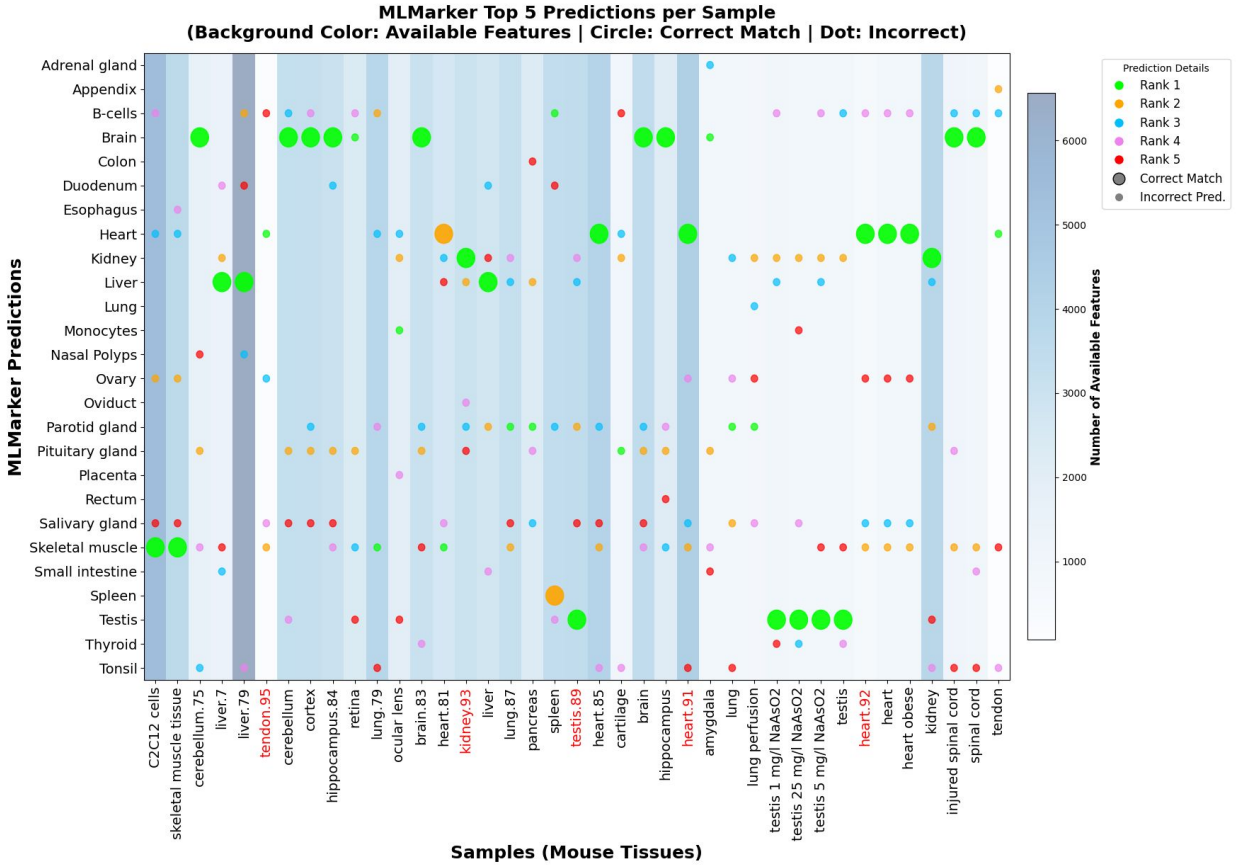
Repurposing large-scale data for new knowledge

We built an AI model that predicts tissue or cell type purely based on the proteome, at 98% accuracy

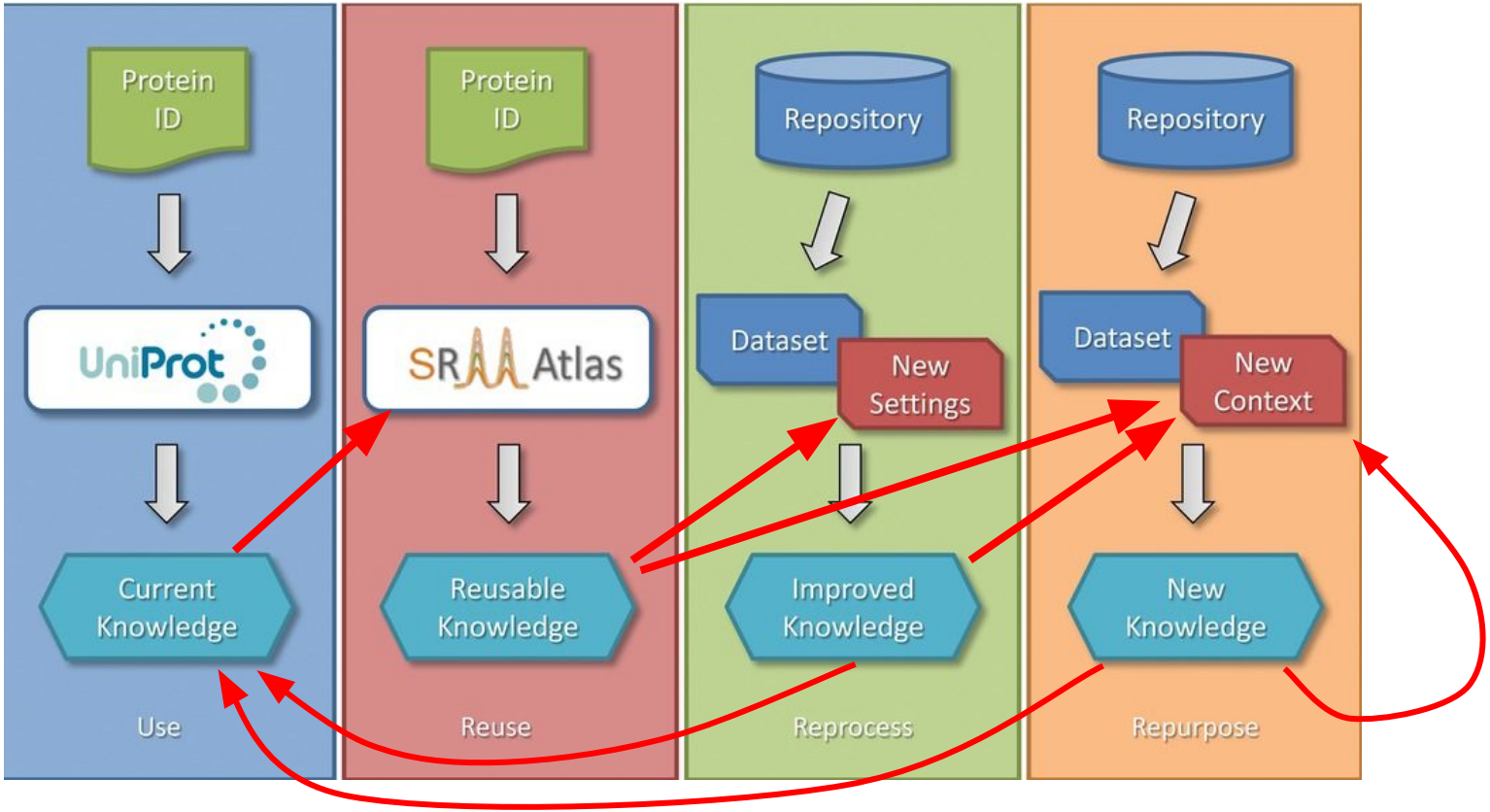


https://doi.org/10.1101/2025.06.10.658642

The human model applied on mouse proteome seems to works also pretty good



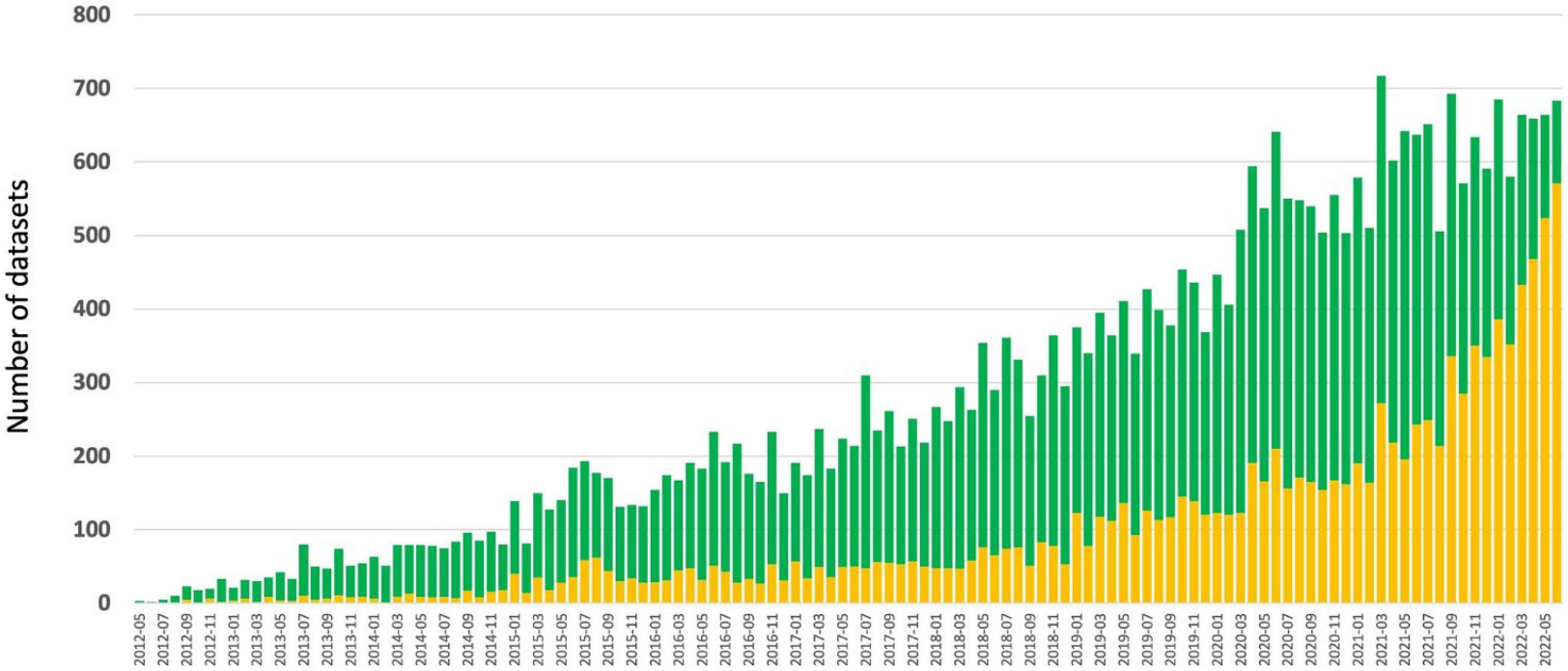
Open data enables endless reuse and maximize scientific ROI



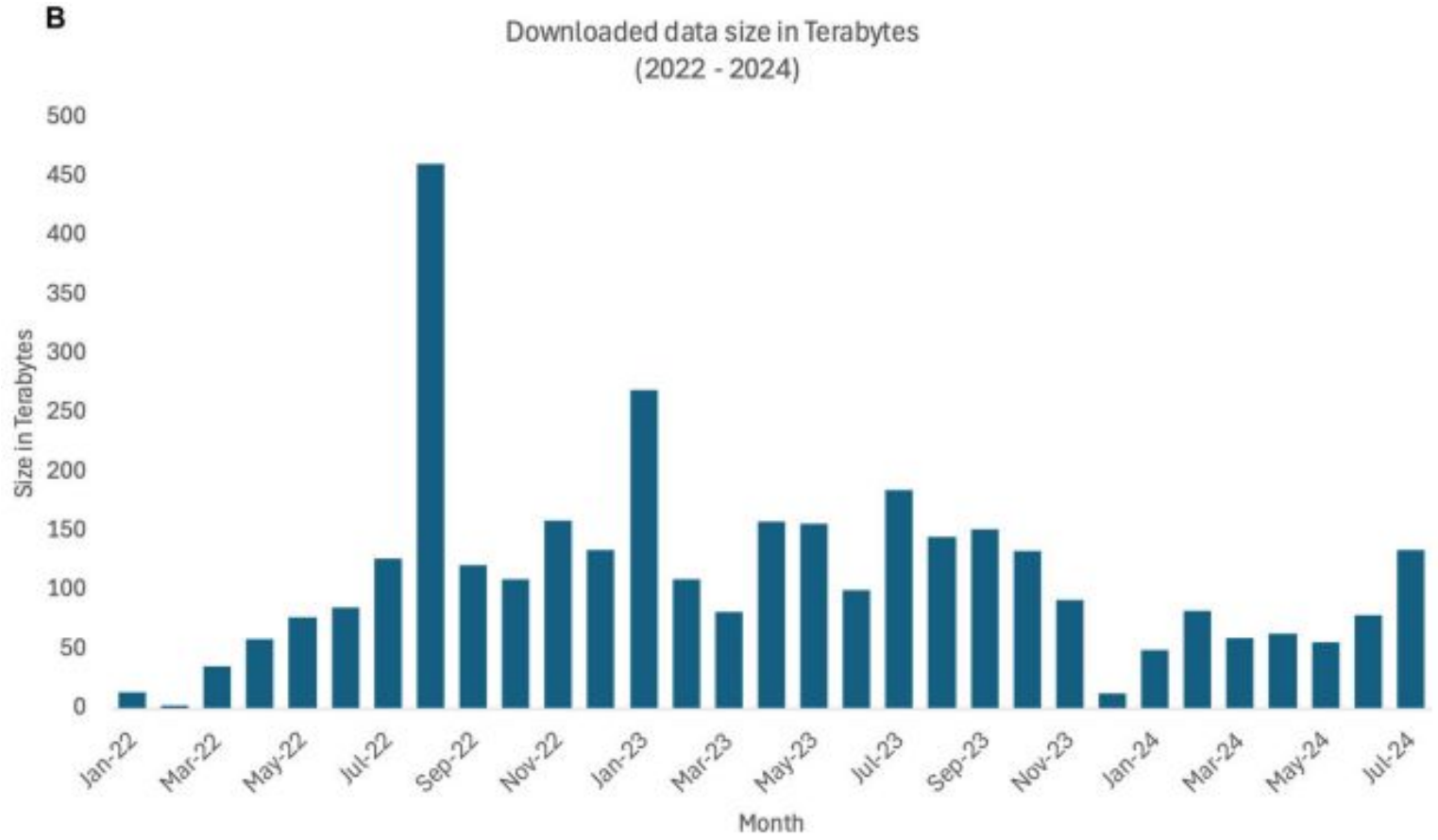
THANK YOU!

LSMBO

Number of submitted datasets per month to PX resources

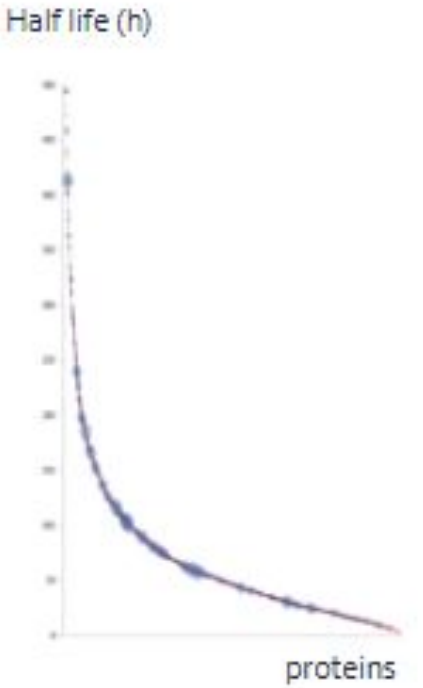


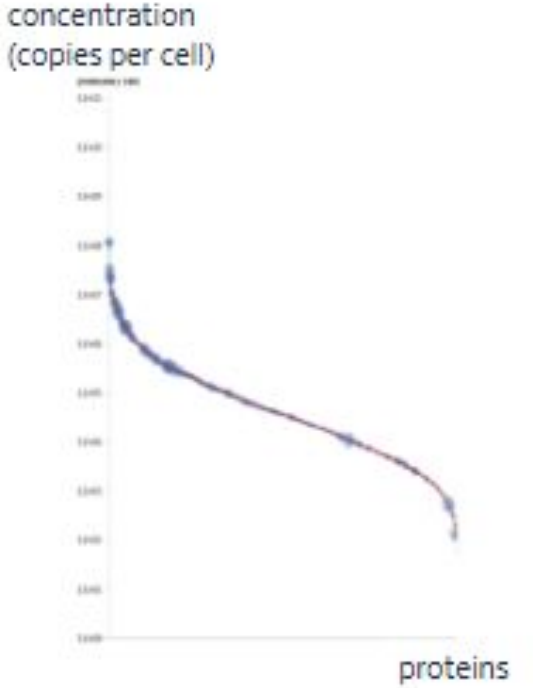
https://doi.org/10.1093/nar/gkac1040

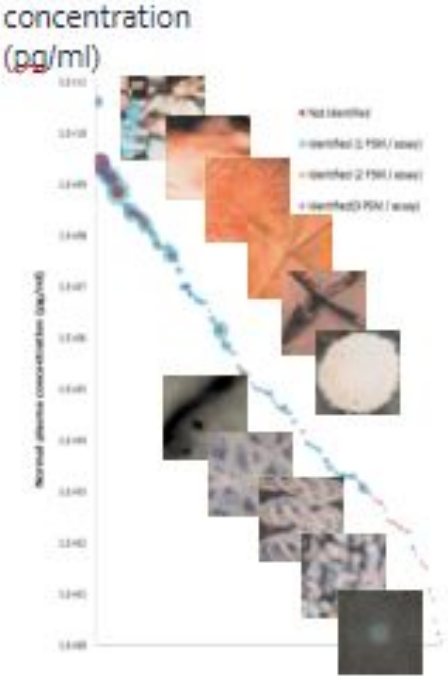


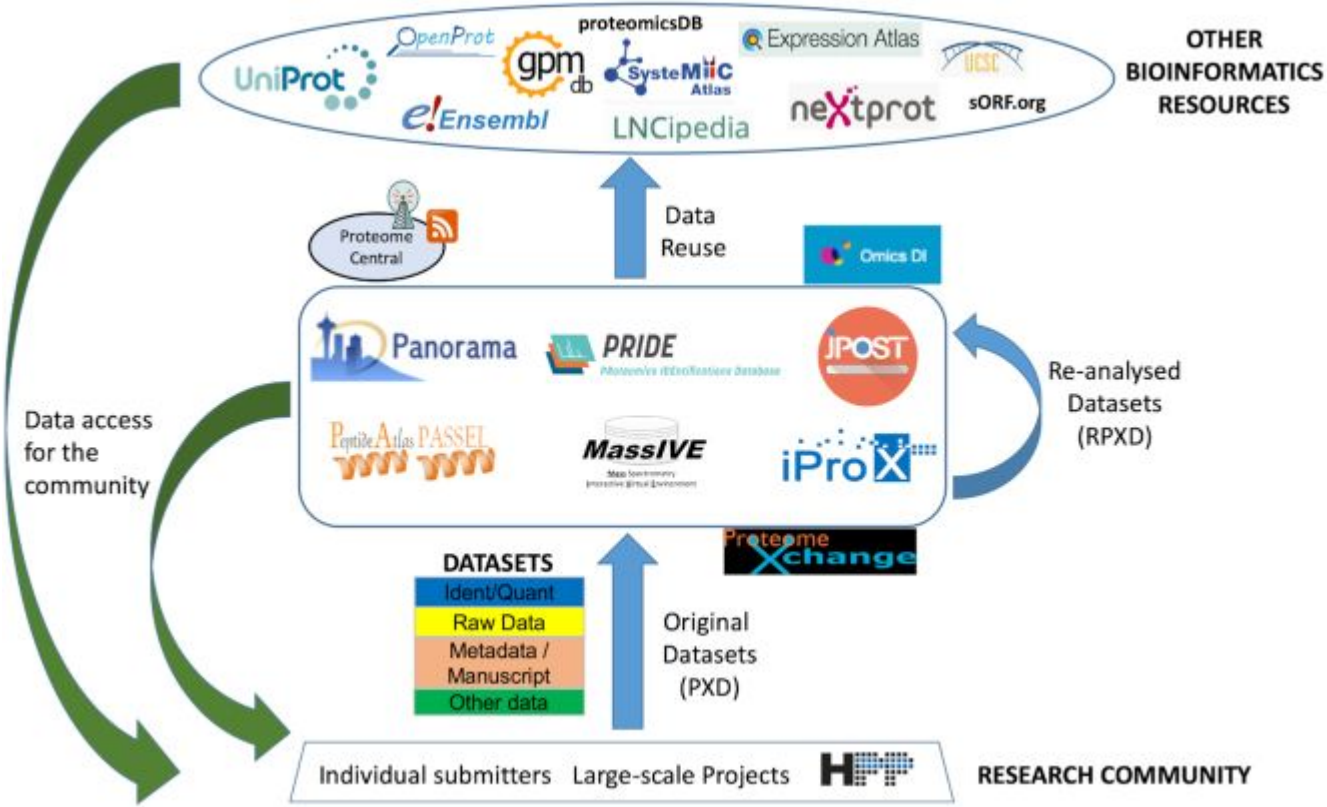
https://doi.org/10.1093/nar/gkae1011

Large-scale data reprocessing can harness heterogeneity to dig very deep into the proteome

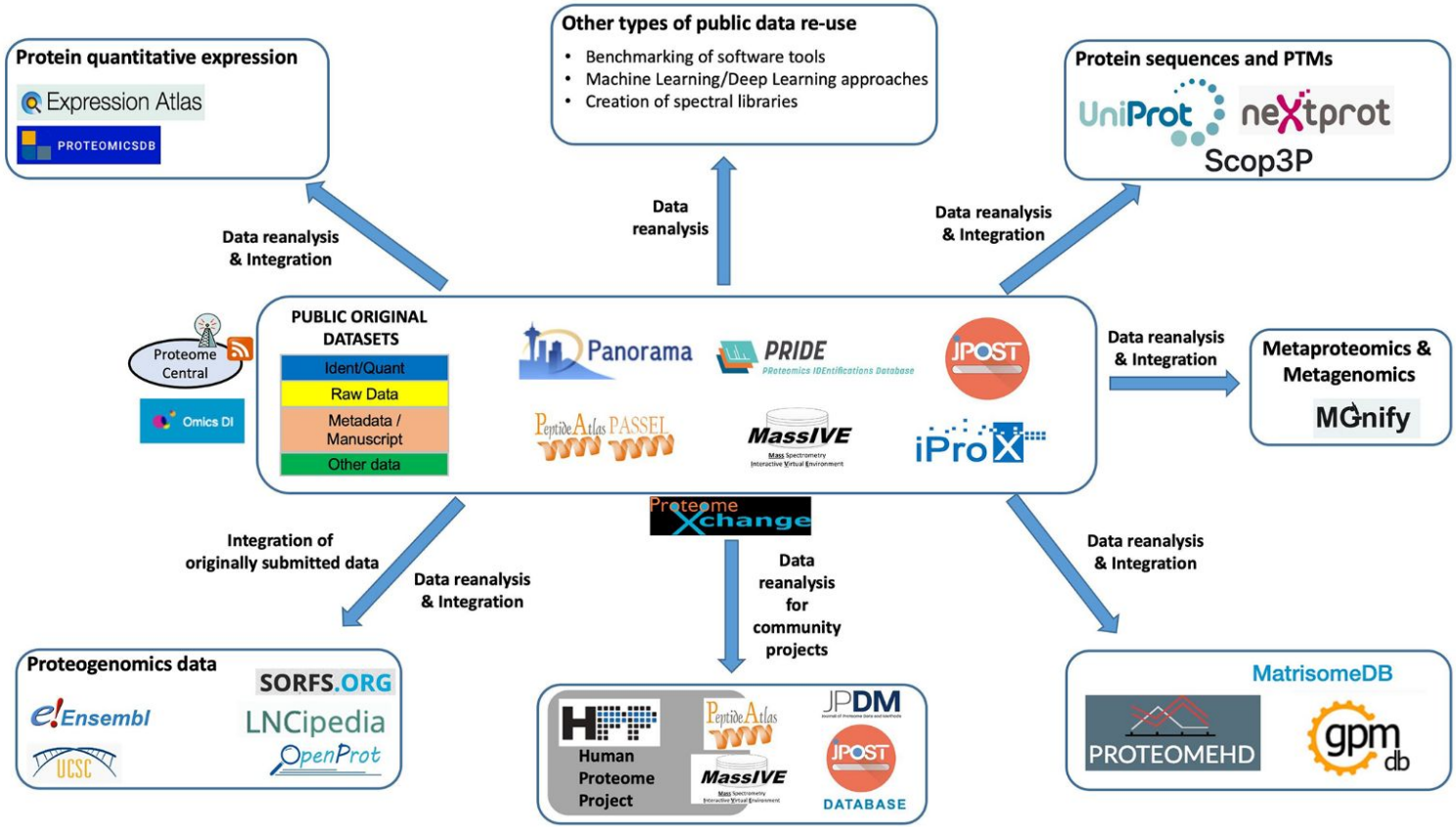








10.1093/nar/gkz984



10.1093/nar/gkac1040