The future of arXiv and knowledge discovery in open science

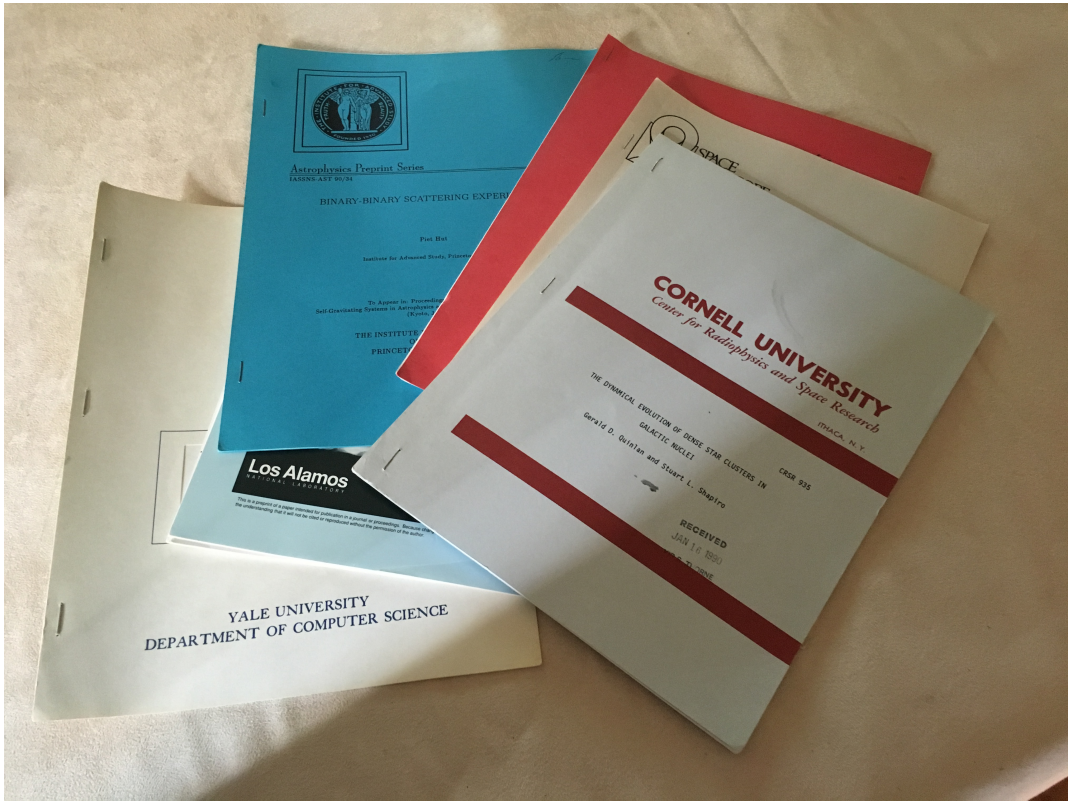
First Workshop on Scholarly Document Processing (SDP 2020)

November 19, 2020

Steinn Sigurðsson, Scientific Director

arXiv.org

Preprints



arXiv.org

Journals

Phys Rev - David Mermin noted that the shelf space will soon be expanding faster than the speed of light, but…
will not violate Relativity as no information will be transmitted!

Cornell University

arXiv.org

arχiv

arXiv.org

arXiv.org e-Print archive
arxiv.org

Login

Search... All fields Search

Help | Advanced Search

arXiv is a free distribution service and an open-access archive for 1,796,336 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse:

Physics | Search | Form Interface | Catchup

News

arXiv now processes new submissions and replacements with TeX Live 2020. Learn more.

Read about recent news and updates on arXiv's blog. (View the former "what's new" pages here). Read robots beware before attempting any automated download.

COVID-19 Quick Links

See COVID-19 SARS-CoV-2 preprints from

- arXiv
- medRxiv and bioRxiv

Important: e-prints posted on arXiv are not peer-reviewed by arXiv; they should not be relied upon without context to guide clinical practice or health-related behavior and should not be reported in news media as established information without consulting multiple experts in the field.

Physics

- Astrophysics (**astro-ph** new, recent, search)
  includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics
- Condensed Matter (**cond-mat** new, recent, search)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, search)
- High Energy Physics - Experiment (**hep-ex** new, recent, search)
- High Energy Physics - Lattice (**hep-lat** new, recent, search)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, search)
- High Energy Physics - Theory (**hep-th** new, recent, search)
- Mathematical Physics (**math-ph** new, recent, search)

arXiv.org

Cornell Tech

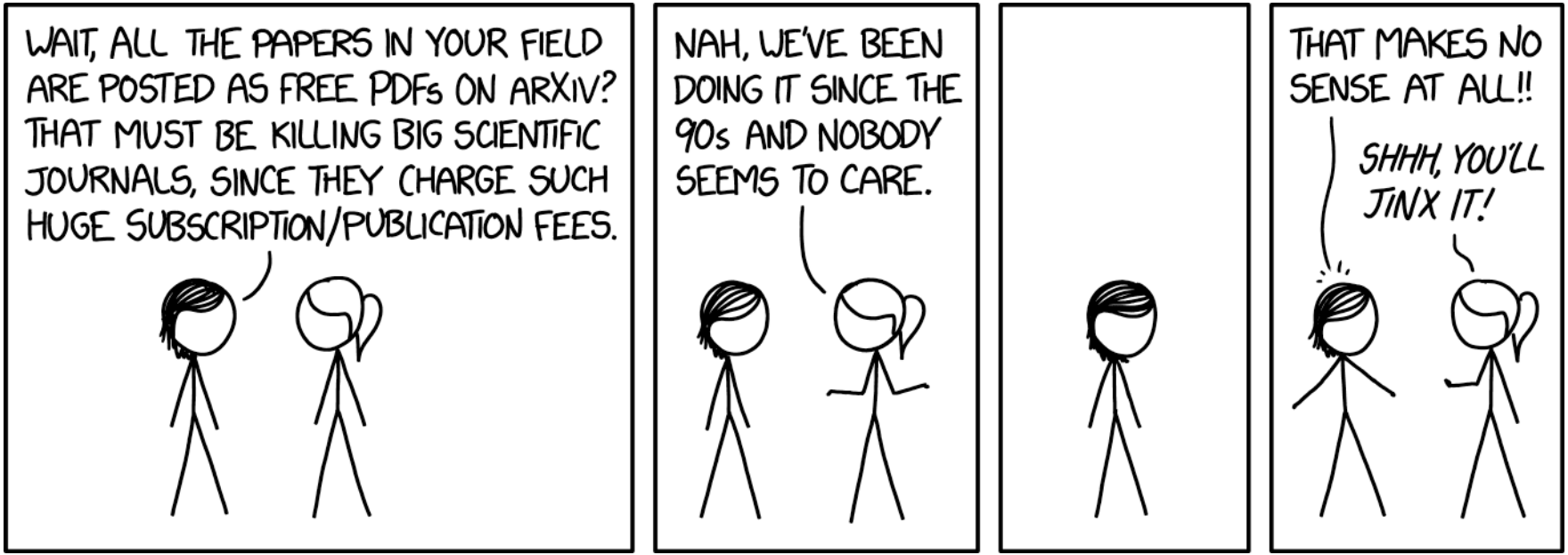
arXiv.org

arXiv on the move

- **Moved to Cornell Tech**
  - spring #2020…
- **Vice Provost and Dean Prof. Greg Morrisett**
- **new Executive Director - Dr Eleonora Presani**
  - 12 staff - 4 part time…
  - ~ 200 volunteer moderators
  - Subject Advisory Committees
  - Science Advisory Board, Member Advisory Board
- Funding:
  - Members == Institutional and University Libraries
  - Simons Foundation
  - Cornell
  - Assorted donations and foundations

arXiv.org

Basics



https://xkcd.com/2085/

arXiv.org

speed of research

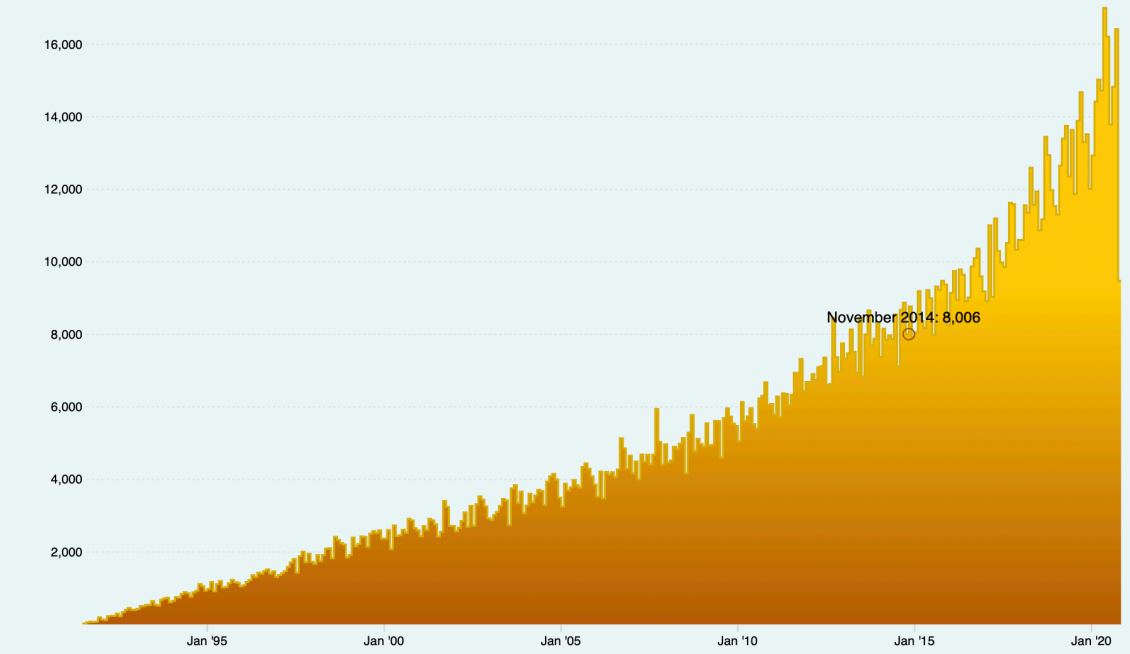
- receive research e-prints email and/or check web
- every morning at coffee ~~or arriving in office~~
- clean simple interface
- - “technical” and $T_E\chi$ — authors vs readers
- source and/or printable
- stable arXiv identifier
- papers on arXiv cited approx. twice as much

arXiv.org

Submissions

Total number of submissions as of November 18, 2020 = **1,796,491**.

Download CSV

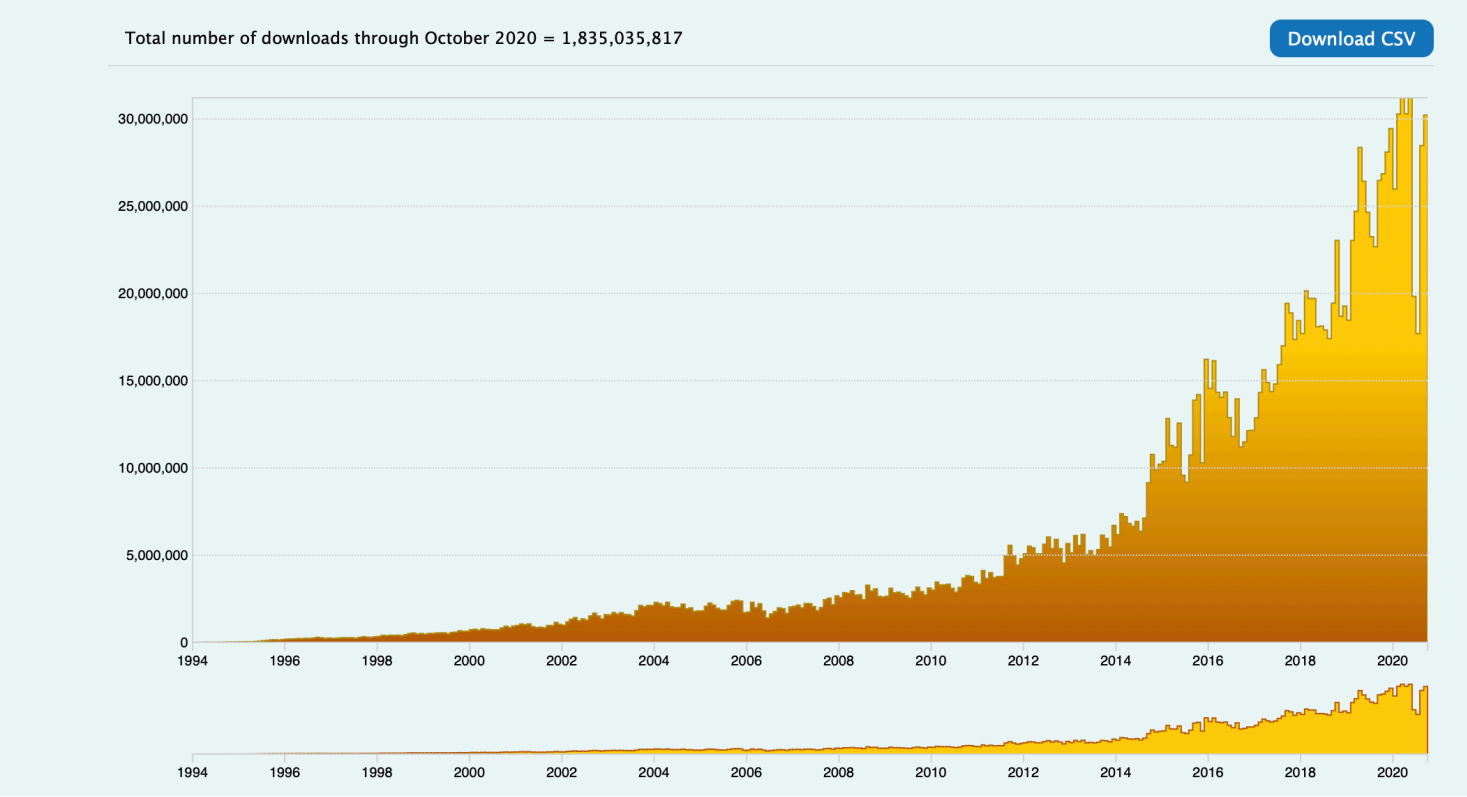


arXiv.org by the numbers

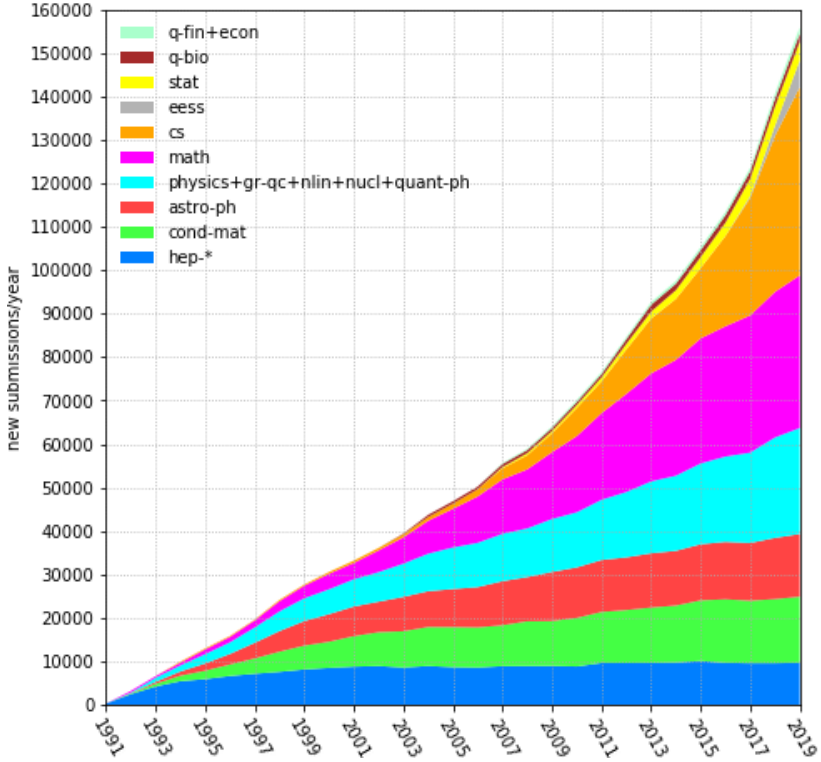
- **1,796,336 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics**
  - ~ 750+ tagged EMNLP 2020
- **over 1,800,000,000 downloads**
- **4-,8000,000 per week**
  - == ~10 downloads per second
- **over 150,000 new submissions per year**
  - ~ 1/4 in CS and growing

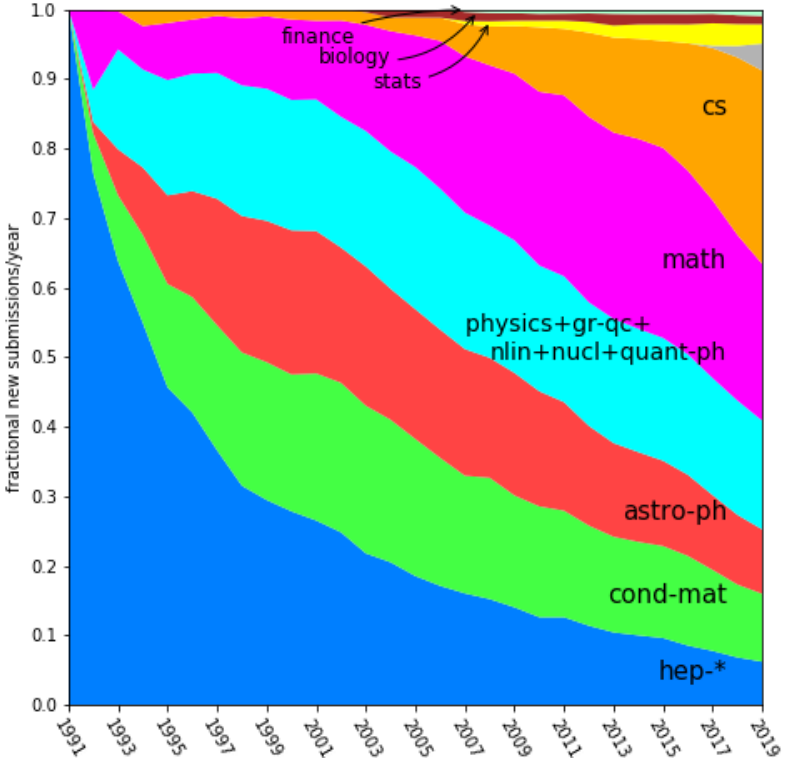
arXiv.org

Downloads



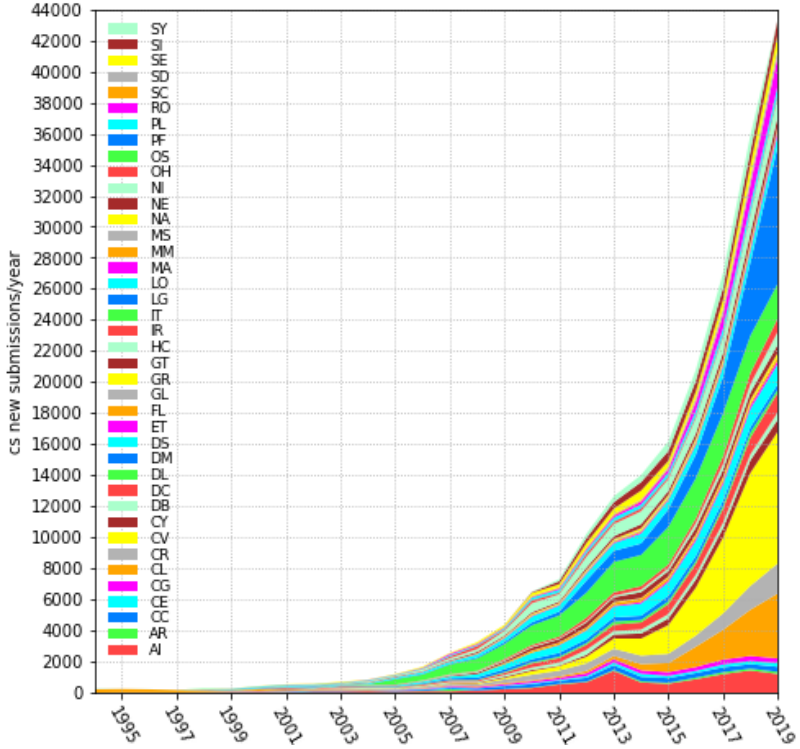
arXiv.org cs.* - big picture

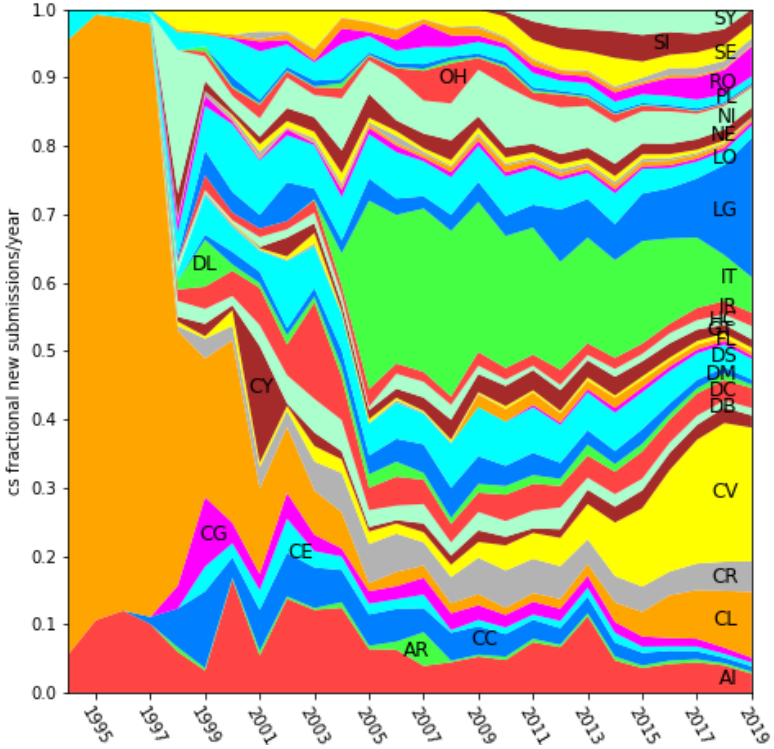




arXiv.org

cs Categories





arXiv.org

The arXiv is not the Internet

- **arXiv moderation**
  - light touch, registered users
- **curated collection!**
- **provides identifier, indexing, archive**
  - version retention
- **heterogenous standards by subject/category…**
  - *interesting or correct?*

arXiv.org Moderation

- **Hard to scale well.**
  - over-reliant on friends-of-friends
- **automation and ML/AI**
  - classifier for categorization
  - currently 3 operational; PwC classifier primary closes loop to authors
- **Need to automate most flow if expanding**
  - *edge cases always there and not important*
  - well, except to the authors…
- Holds… Need to head off logjams
- Normative issues - eg Covid-19 papers
- **Beware Gatekeeping**

arXiv.org

Classification

- **critical for arXiv to be interesting**
  - edge cases not uncommon
  - Sturgeon's Law - keep S/N up
- **judicious cross-listing invaluable**
  - auto cross-lists
  - vulnerable to blocking interdisciplinary knowledge transfer
  - or overwhelming information transfer
- **lateral knowledge transfer**
- **choice of categories may trigger Arrow's Theorem…** 😉
  - or equivalent…

arXiv.org

What's Next

- **refactor code base in situ**
  - dockerized python, portable and robust
  - lift to cloud - google cloud
- **streamline submission interface**
- **improved moderation tools**
- provide updated API for metadata and e-prints
- open source modules
  - ~ 1/3 code base now open source
- https://github.com/arXiv

arXiv.org

But, wait, there's more!

- **Working with:**
  - ADS, INSPIRE-HEP
  - Semantic Scholar & Google Scholar
  - Papers with Code
  - CORE
  - Kaggle
  - and others
- **expanded options outside core arXiv function**
- **Soliciting partners**
  - through arXiv Labs
- **third party and arxiv source**
  - eg. hepth.io arxiv-sanity.com, arxiv-vanity.com

arXiv.org

arXiv Labs

arXiv Labs

arXiv is surrounded by a community of researchers and developers working at the cutting edge of information science and technology.

While the arXiv team is focused on our core mission—providing rapid dissemination of research findings at no cost to readers and submitters—we are excited to be experimenting with a small number of collaborators on projects that add value for our stakeholders and advance research.

Here are some of the projects that our collaborators are working on right now.

arXiv Links to Code

Collaborators:

Robert Stojnic
Papers with Code / Facebook AI Research

Viktor Kerkez
Papers with Code / Facebook AI Research

Ludovic Viaud
Papers with Code / Facebook AI Research

arXiv Links to Code aims to provide an easy and convenient way to find relevant code for a paper. It is using data from Papers with Code - a free resource that links papers, code and results in Machine Learning. Papers with Code is the biggest such resource and is licensed under an open license.

Code: https://github.com/arXiv/arxiv-browse/tree/develop/browse/static/js/paperswithcode.js

CORE Recommender

Explore relevant open access papers from across a global network of research repositories while browsing arXiv. Research

https://labs.arxiv.org

arXiv.org experimentation

- **Looking to provide additional functionality outside core services**
  - Metadata and full text search - Kaggle - **for NLP**
  - https://www.kaggle.com/Cornell-University/arxiv
  - software - Papers with Code, preliminary Code Ocean
  - Knowledge Discovery - CORE, ADS ++
  - Author links and bibliography services ++: Semantic, Google Scholars
- **Planned**
  - DOIs
  - improved metadata
  - Supplemental data links
  - dark archive
  - ++

arXiv.org

Future

- **Ambitions**
  - content type
  - customized content service and interfaces
  - personalized discovery
  - smart discovery - “the unknown unknowns…”
  - subject area expansion
  - overlay journals
  - third party comment and discussion options

Caveat: $$$

We need people and resources

The Future of arXiv

arXiv.org

Free to readers, free to authors!

arXiv.org

this slide intentionlly left blank

- **Did I forget anything?!**
  -

arXiv.org

cs.*