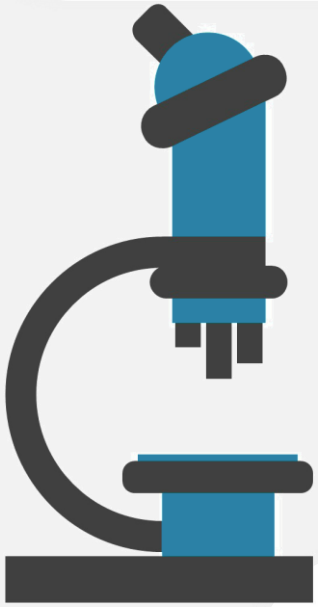
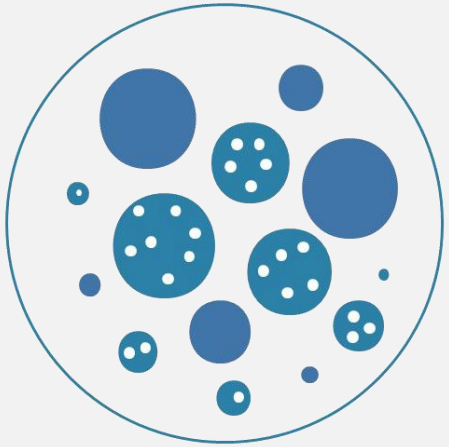
Data Collection
A view from 30,000 ft to <3 Å





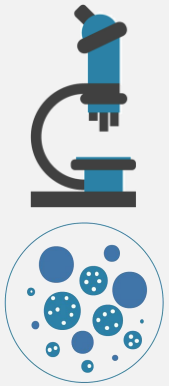
Alex J. Noble

Smart Data Collection Workshop!

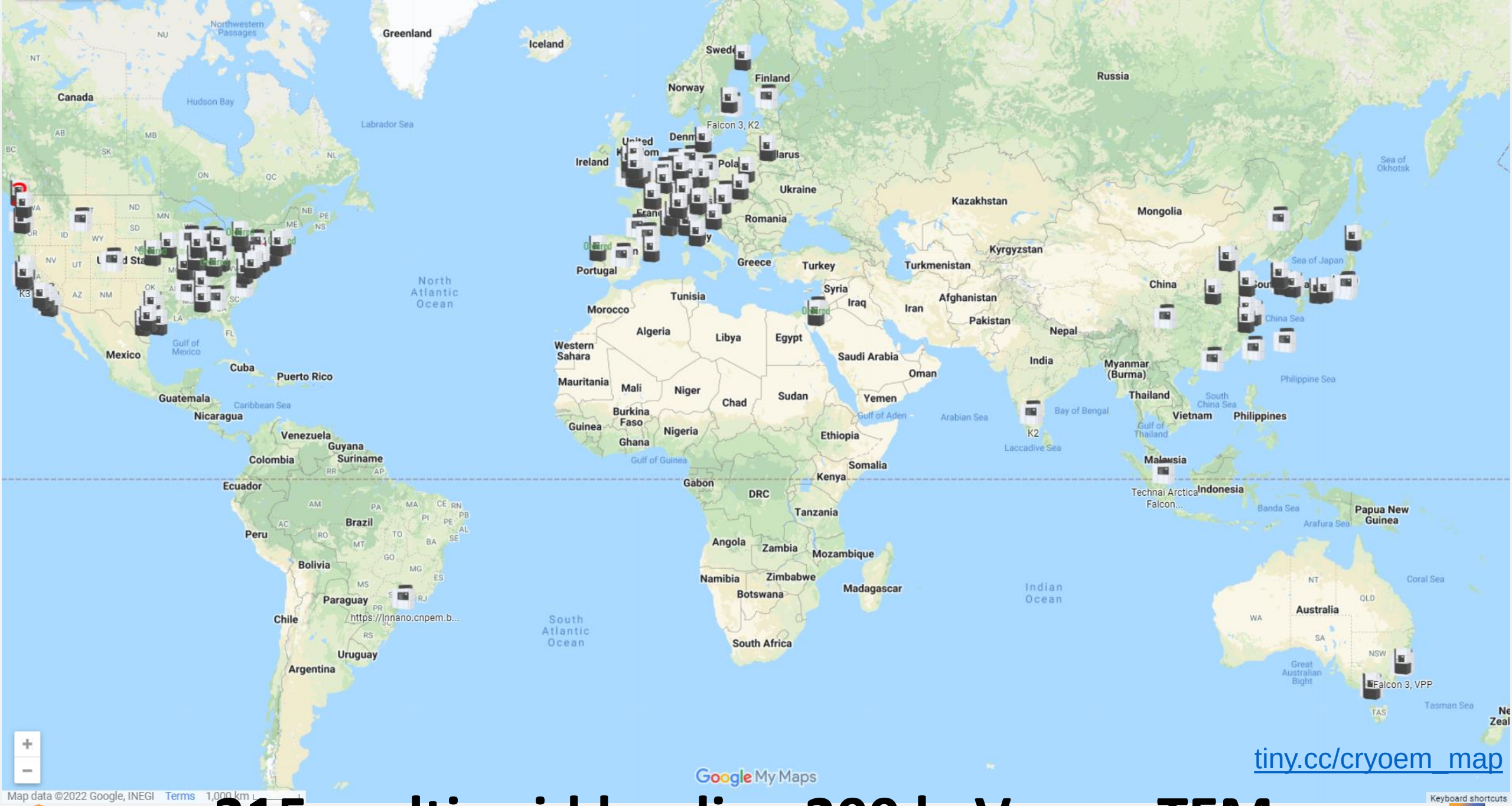
4-6-22

I want this to be a **brainstorming session** on **collection throughput**

- 30 minutes total
- 15-20 minutes of slides with ideas
- I'll put several ideas on a slide, then we'll brainstorm!

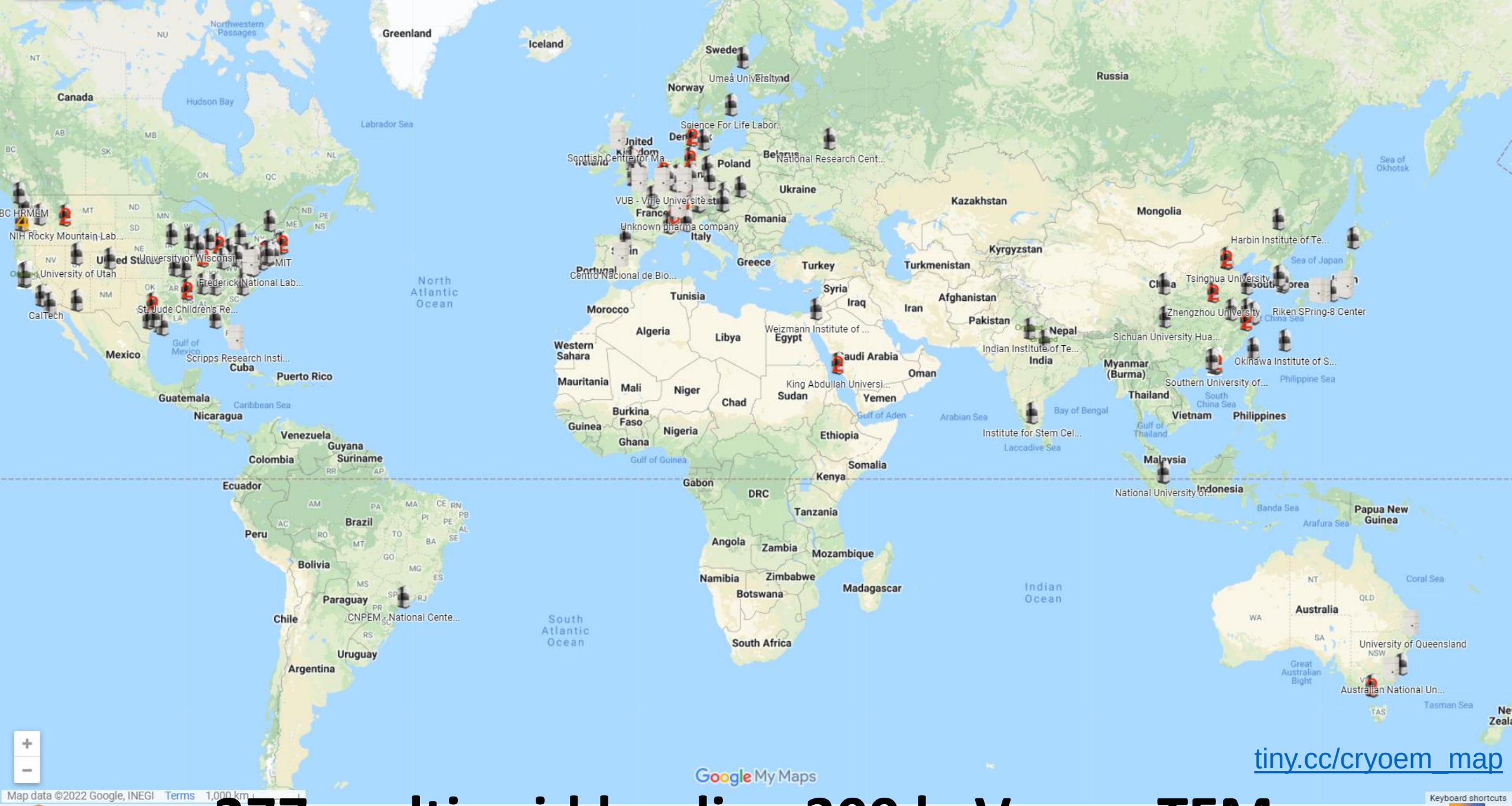


Who are our target audience?



tiny.cc/cryoem_map

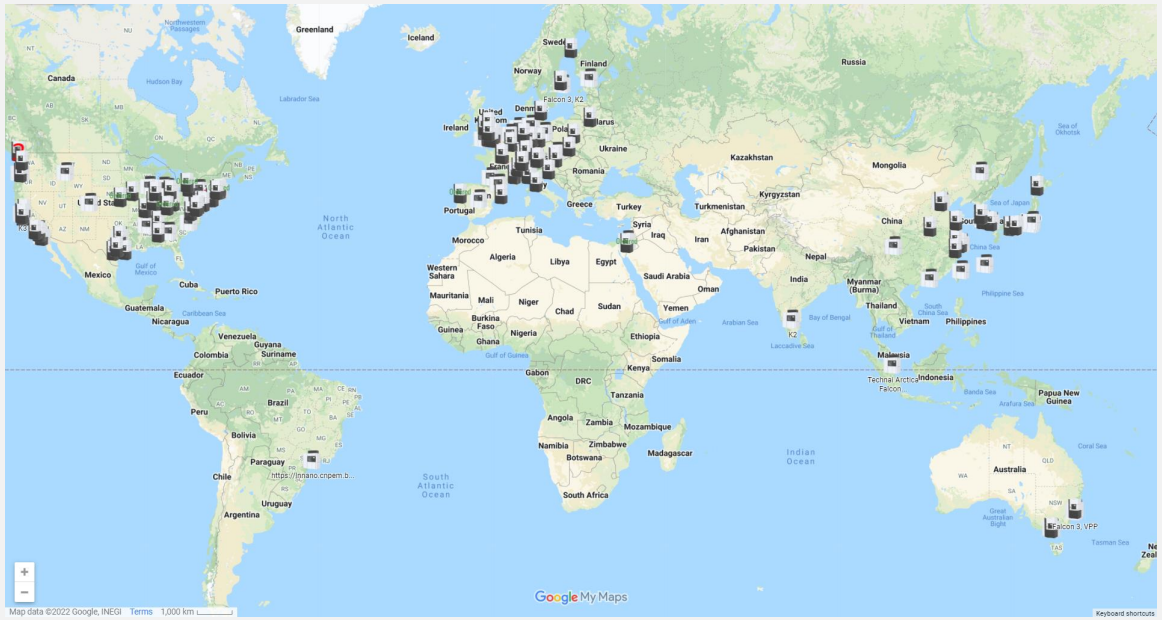
215 multi-grid loading 200 keV cryo-TEMs

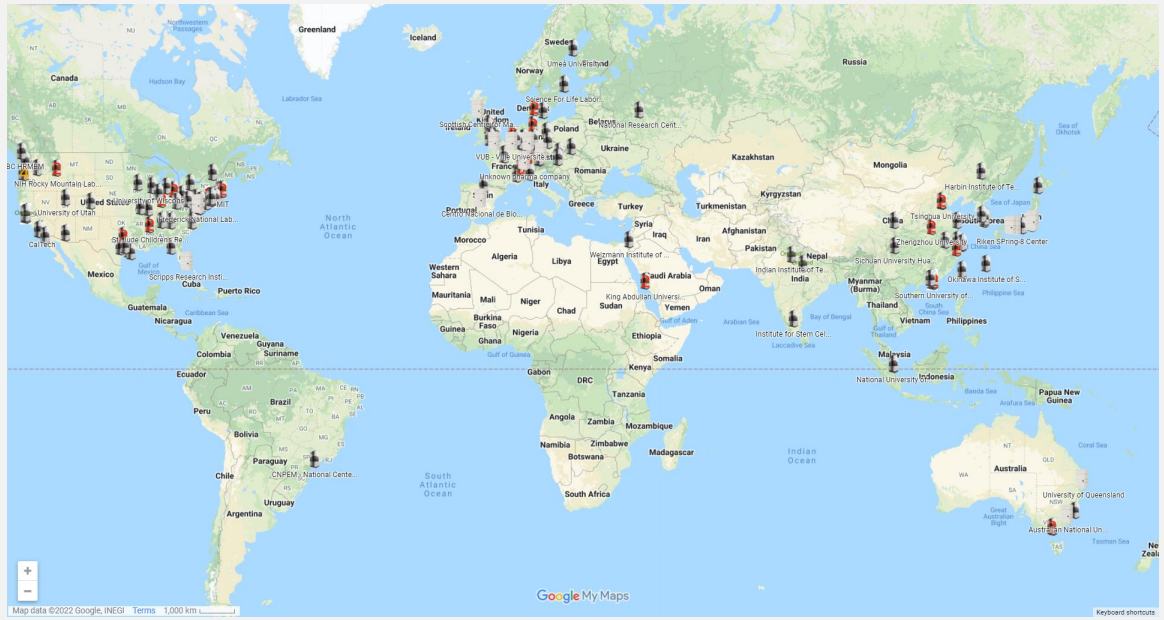


tiny.cc/cryoem_map

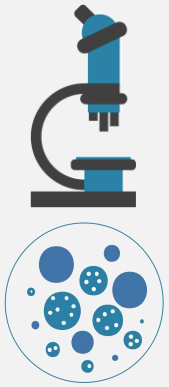
277 multi-grid loading 300 keV cryo-TEMs

Who are our target audience?





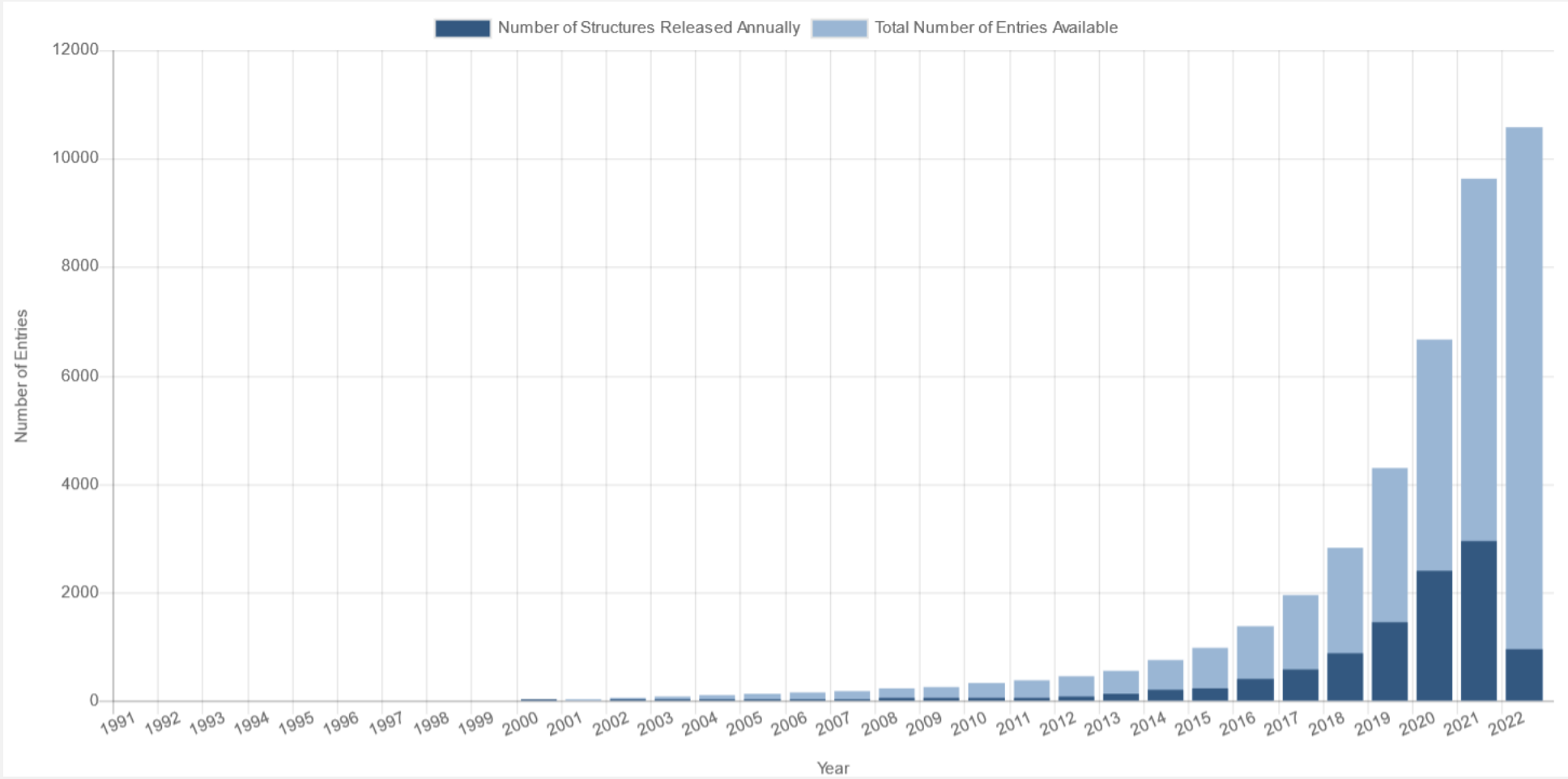
- Most places **with a 300keV** multi-grid scope **also have** a **200keV** multi-grid scope
  - Are multi-grid scopes our **target audience?**
- There are **hundreds** of **side-entry scopes** too
  - *Not* our main target?



Should we follow the synchrotron model?

Should we follow the synchrotron model?

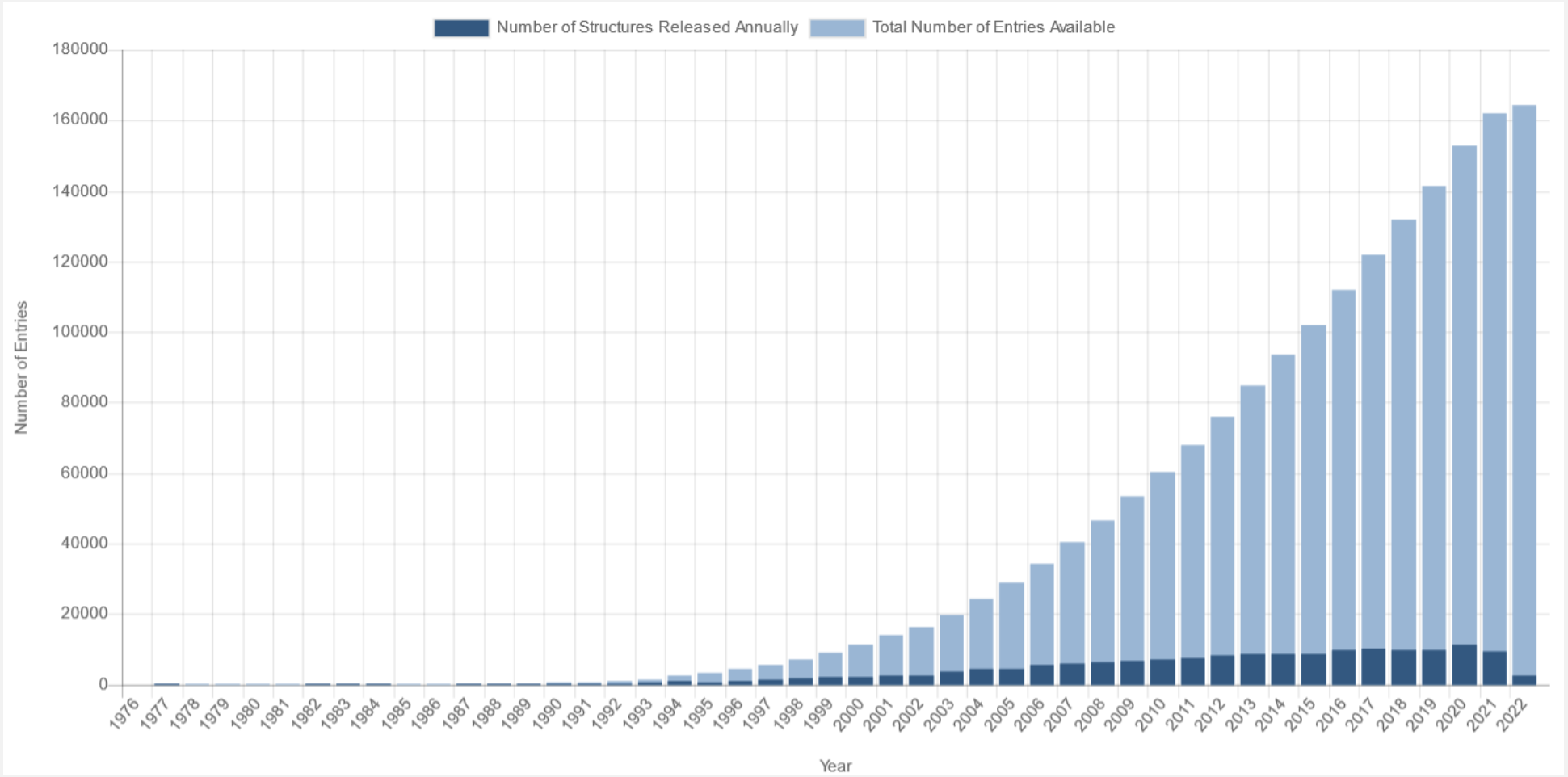
CryoEM annual structures vs. total

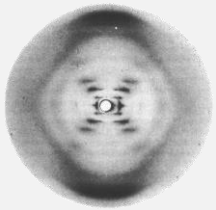


pdb.org

Should we follow the synchrotron model?

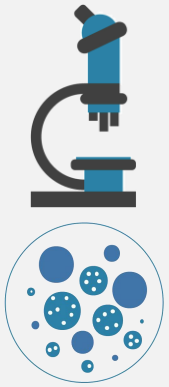
Xray annual structures vs. total

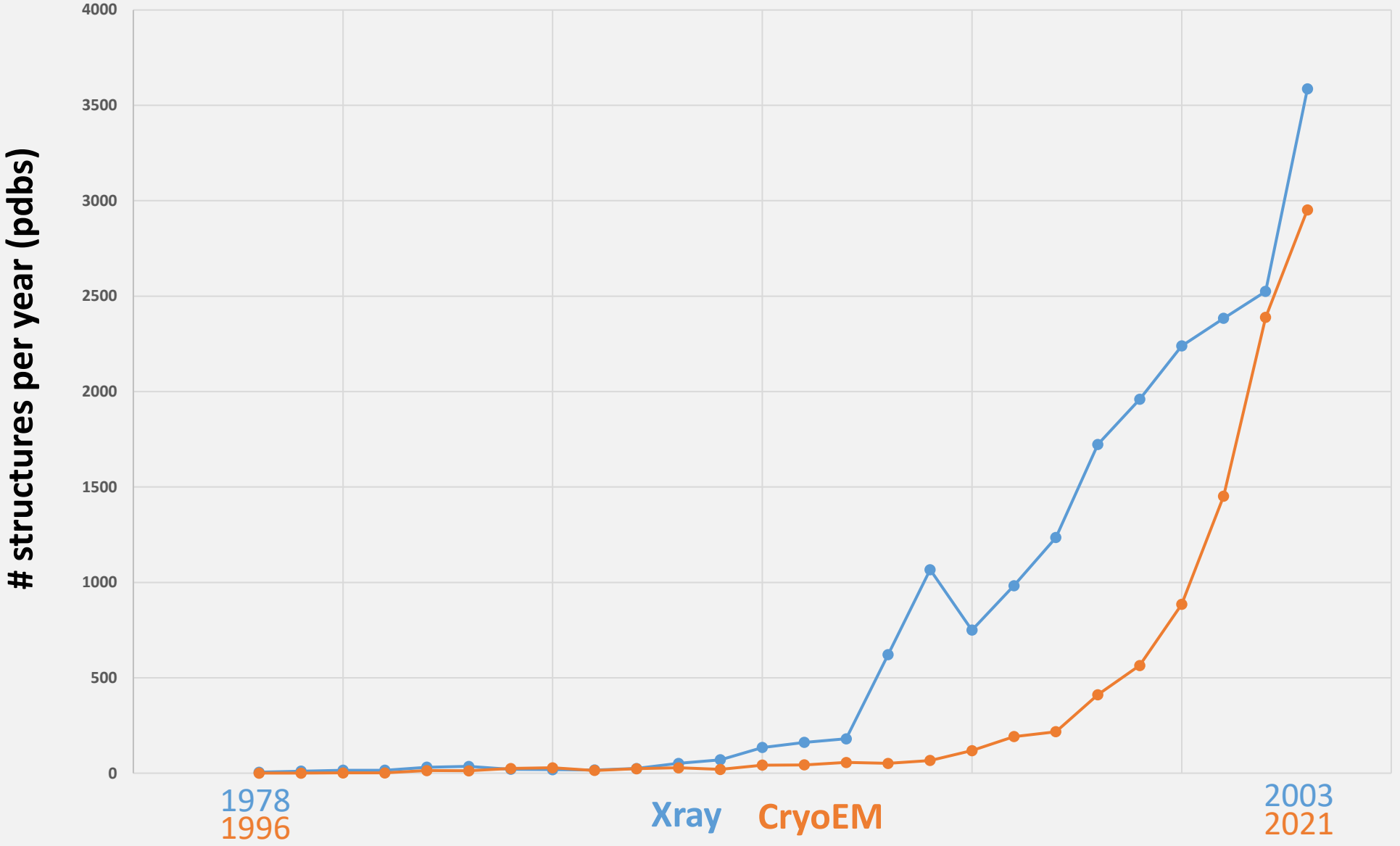




pdb.org

Should we follow the synchrotron model?



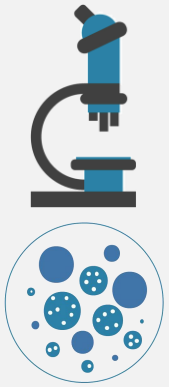


Note: Industry is not accounted for

pdb.org

Should we follow the synchrotron model?

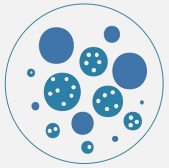
or just borrow some ideas?



- There are **~70 synchrotrons** & **hundreds** of smaller **screening** sources
- There are **277 300keV** multi-grid scopes, **215 200keV** multi-grid scopes, and **hundreds** more **screening** scopes
- Xray researchers get **8-hour blocks** = **hundreds** of crystals may be **screened each session**
  - Do we need to **increase our screening throughput?**

Should we follow the synchrotron model?

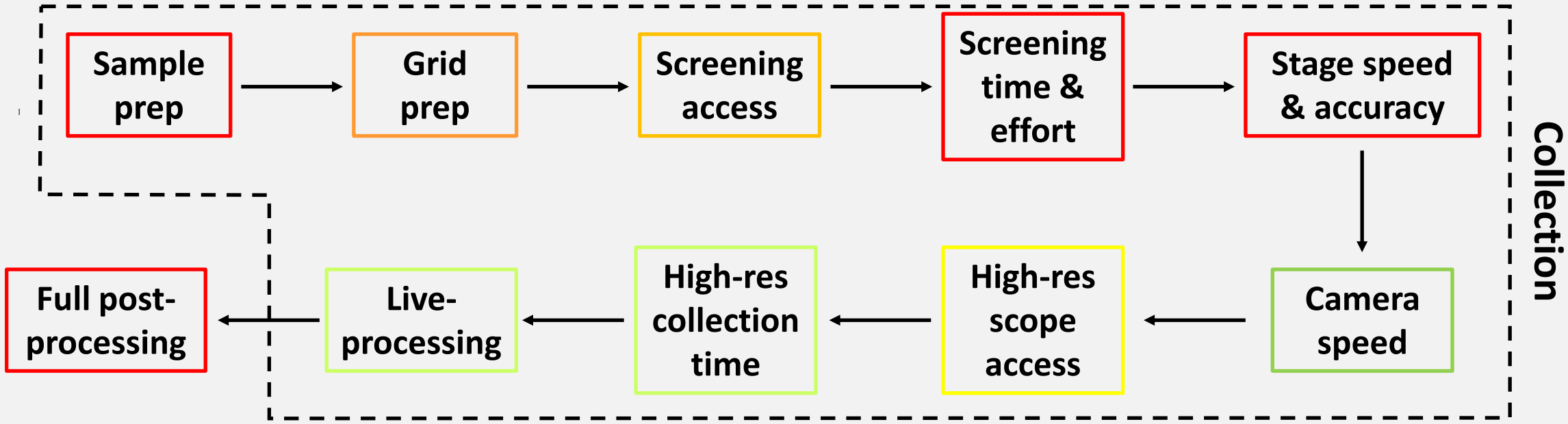
or just borrow some ideas?



- Like cryoEM, xray screening takes much longer than data collection
  - *Xray screening*: **hours+**, *high-quality data*: **minutes**
  - *CryoEM screening*: **hours+**, *high-quality data*: **hours**
  - **A synchrotron** can theoretically collect **hundreds** of datasets **overnight**, and **10k+** datasets per **year**
- Based on published structures, cryoEM **may be growing faster** now than xray has ever grown.

Where are our bottlenecks in cryoEM?

Where are our bottlenecks in cryoEM?



Miscellaneous

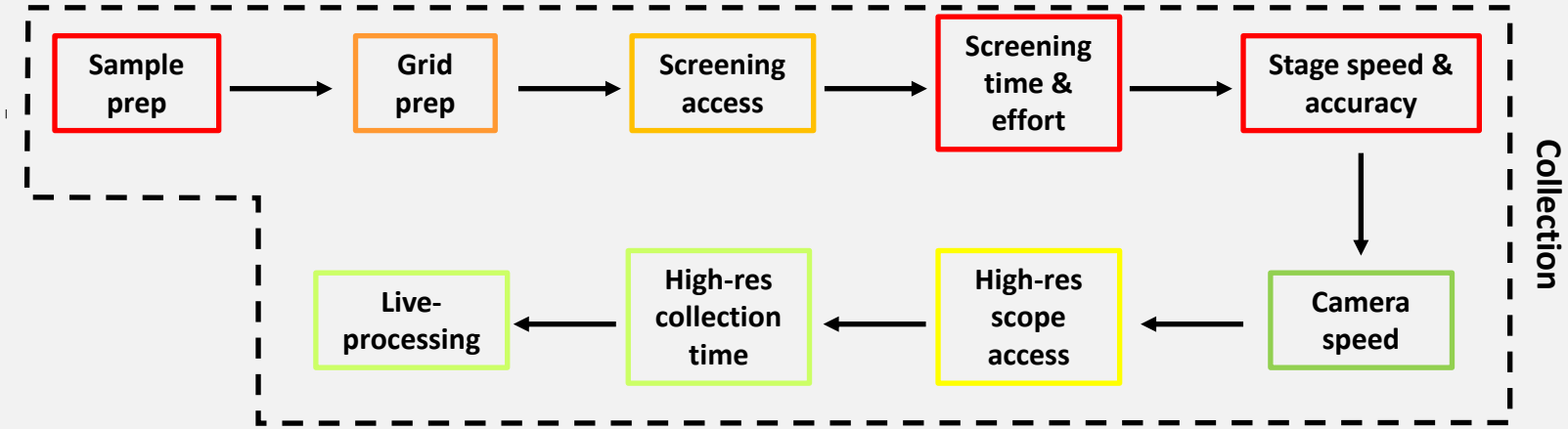
Cross-training facilities for efficiency

Cross-training users for proficiency

Finding un-tapped users (biologists, doctors/hospitals)

Not current bottleneck — Current bottleneck

Where are our bottlenecks in cryoEM?



Goals:

1) **Increase** screening **speed & feedback** to user
   - *lots of grids fast!* - to **optimize** sample/grid **conditions**
2) Collect high-quality data as **efficiently** as possible

- Use **live-processing feedback**
- Always do **better than human** efficiency & quality
- Other broad goals?

Where are our bottlenecks in cryoEM?

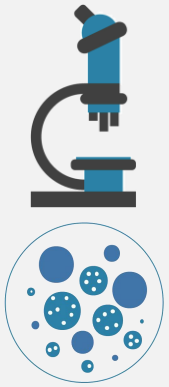
How can companies help?

- Bigger **autoloaders**, faster & more accurate **stages**
  - 96-grid autoloader = unattended weekend screening
- Bridget's idea: Put **holes closer** on grids!
  - Is there any reason this can't be done?
- Richard's idea: **Cheaper screening** cryo-TEMs
  - What are the minimum scope requirements?
- **Better** live- and post-processing **software**
  - Preferably open source and with clear file handling so researchers can extend and validate it.
- **Other ideas?**

Where are our bottlenecks in cryoEM?

How can researchers help?

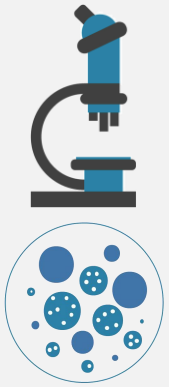
- **Better** live- and post-processing **software**
  - Preferably open source and with clear file handling so researchers can extend and validate it.
- Devise **orthogonal & complementary** screening **methods**
  - E.g. MS or photometry methods
- Label and release **curated training data**
  - Each labeled image will reap a **hundred-fold+ increase** in future throughput
  - Machine learning is ripe in cryoEM!
- **Other ideas?**

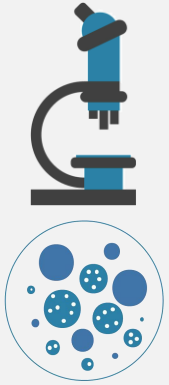


Quick machine learning primer

Quick machine learning primer

- Machine learning **models**:
  Train from labeled data, **apply** to **new data**
- Model **generalizability**:
  How well a model works on data **with different characteristics** from training data
- Model **re-training**:
  Continue training an **existing model** with **additional** training **data**
- Active learning:
  Real-time label and output **updating**



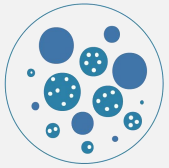


Screening as a machine learning problem

Screening as a machine learning problem

Break screening down into *de novo* and *prior knowledge* projects

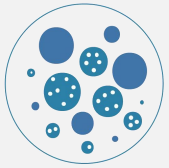
- *De novo*: You **know nothing** about the grid+sample prior to cryo



- *Prior knowledge*: **You do!** E.g. The particle likes thin ice = 'big' squares

Note: **High-res** collection usually comes after screening and thus **includes prior knowledge**

Can screening be **broken down differently?**



SEMC/SMLC/NYSBC's solution:

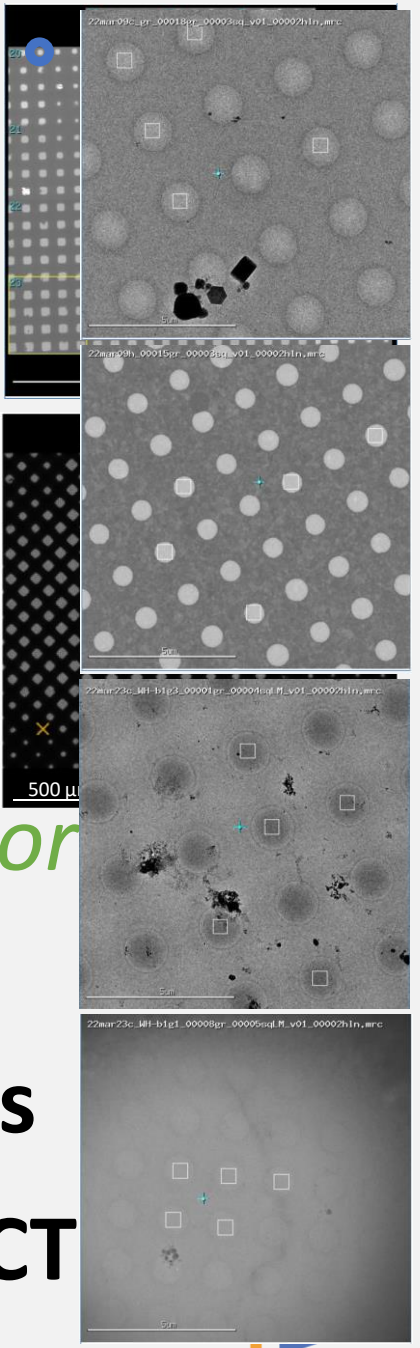
Smart Leginon

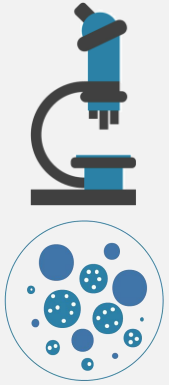
Smart Leginon screening

Current progress highlights

For **11 grids** in a Glacios:

- Hands-on operator time: **from 6 hrs to 10 min**
- Microscope time: **from 6 hrs to 5.4 hrs**
- Works for *de novo* projects & projects with *prior knowledge*
- Works for **different grid** types & **hole contrasts**
- Chooses squares w/ **comparable** good **holes**, **CT** and **ice thickness** to operators





Smart Leginon

Anchi and Paul will show the nuts and bolts=)

Acknowledgements

Paul Kim
Software Engineer

Huihui Kuang
Scientist

Joshua Mendez
Scientist

Kashyap Maruthi
Scientist

Hui Wei
Scientist

Eugene Chua
Scientist

Mahira Aragon
Research Associate

Slavic Serbynovskyi
Technician

Kasahun Neselu
Scientist

Edward Eng
Scientist
NCCAT Manager

Anchi Cheng
Senior Scientist

Alex Noble
Scientist / Group Leader SEMC

Tristan Bepler
SMLC Group Leader

Bridget Carragher
Director

Clint Potter
Director

NRAMM
National Resource for Automated Molecular Microscopy

SRAMM

National Center for CryoEM Access and Training

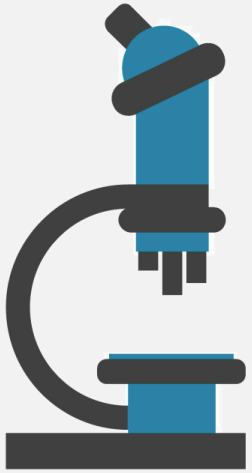
SIMONS ELECTRON MICROSCOPY CENTER

NEW YORK STRUCTURAL BIOLOGY CENTER

SIMONS FOUNDATION SF349347

NIH National Institutes of Health
Turning Discovery Into Health

NIH IGMS GM103310



Thank you

Time for more discussion?