A Framework for Sharing High-Risk Data

Open Science in Action

Isabella Chu, MPH
Associate Director, Data Core, Stanford Center for Population Health Sciences
November 18, 2019

Open Science, to accelerate discovery and deliver cures

neuro

#OpenScienceinAction | www.theneuro.ca

The challenge: Large, high-risk datasets

- Commercial Claims
  - 58 million, 8 years
  - 149 million, 7 years
- Medicare
  - 11 million, 8 years
- EMRs
  - 80 million

	Traditional Model	PHS Data Center
Discoverability	Personal relationship or literature review	Searchable using standard ontologies such as Medical Subject Headings (MeSH)
Access		vailable across sciplines
Administration		ed processes
Data Quality		mmon model for interoperability
Hosting		d ontologies such adings (MeSH)
Computing Capacity	Limited bandwidth and ad hoc security for large-scale analysis	Secure, cloud computing for performance and scalable use

Goal: Research innovation—*not disparate data access*—defines the competitive advantage.

	Traditional Model	Optimal Model
Discoverability	Personal relationship or literature review	Searchable using standard ontologies such as Medical Subject Headings (MeSH)
Access	Limited to single teams and/or discipline	Data and agreements available across teams and/or disciplines
Administration	Dependent on research team knowledge and management	Centralized, automated processes
Data Quality	Fragmented, un-harmonized data sets requiring individual cleaning	Transformed data into common model for ease of replication and interoperability
Hosting	Individual computers or lab-specific servers	Cloud storage for ease of access and customization
Computing Capacity	Limited bandwidth and ad hoc security for large-scale analysis	Secure, cloud computing for performance and scalable use

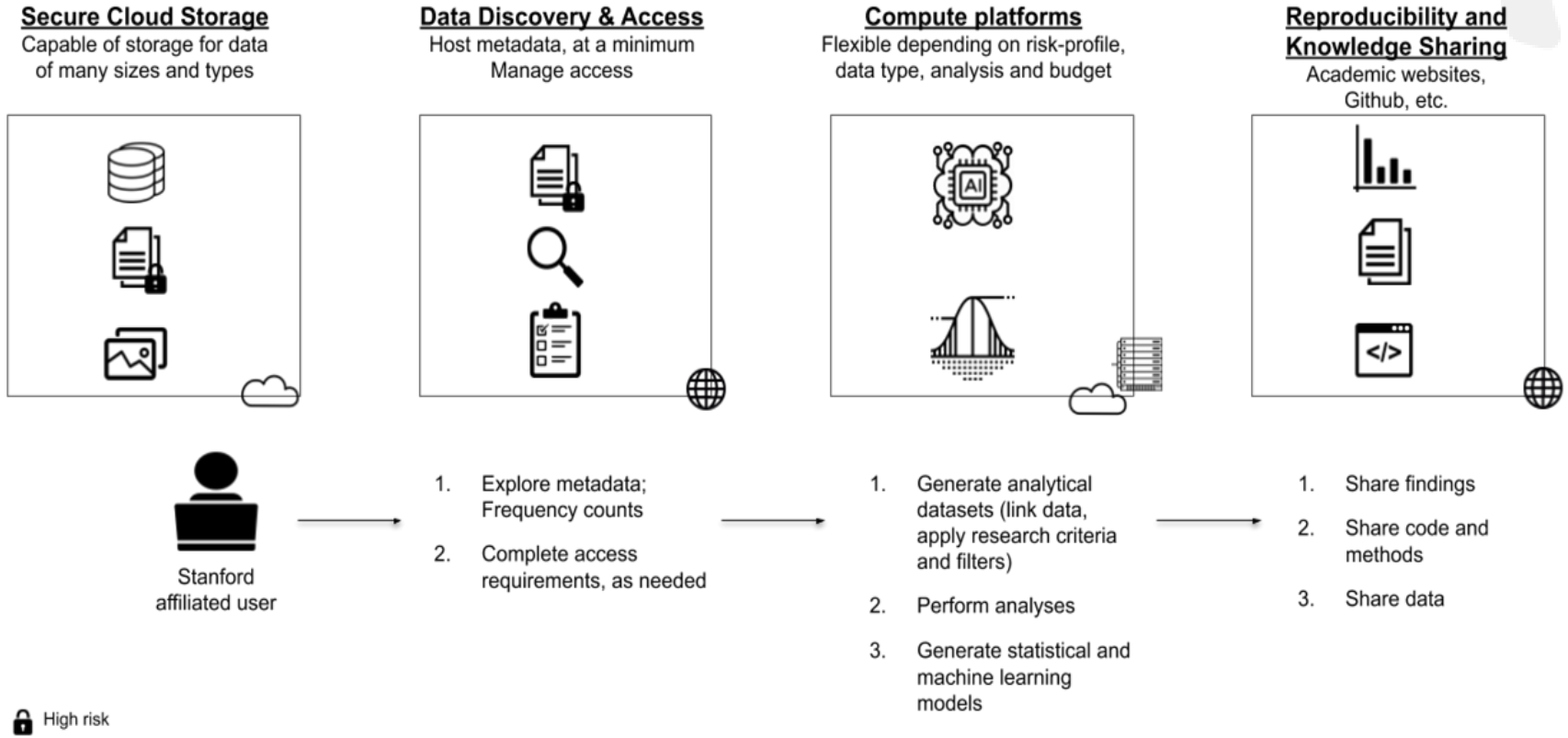
A Data Ecosystem for High-Risk Data

Stanford | Center for Population Health Sciences
Community Engagement | Data | Education | Research

Model Data Ecosystem

Goal : Combine disparate datasets together to answer population health questions

Strategic Priorities: ✓ *Data Governance* ✓ *Data Management* ✓ *Infrastructure* ✓ *Incentives* ✓ *Capacity Building*



Stanford affiliated user

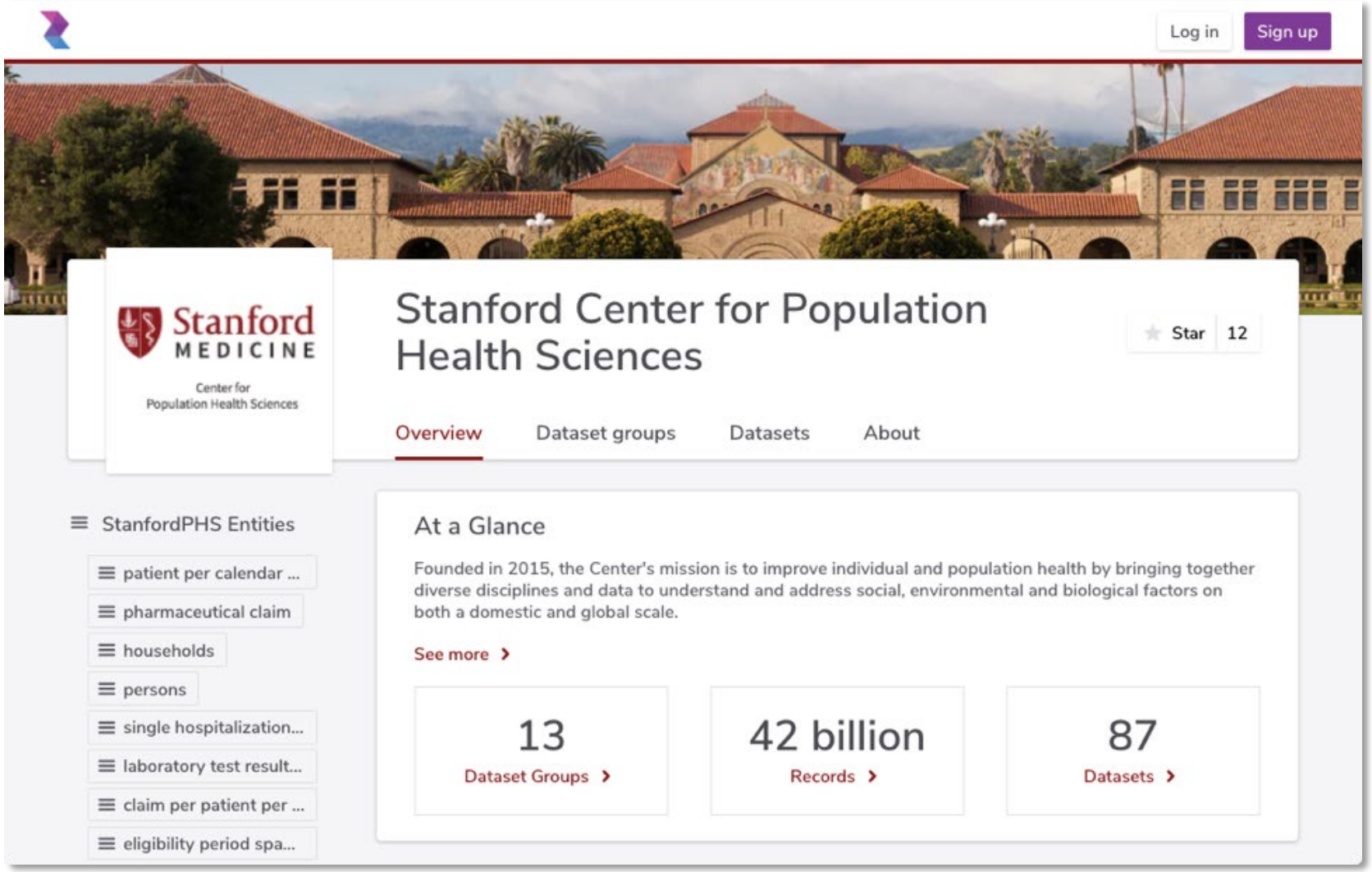
1. Explore metadata; Frequency counts
2. Complete access requirements, as needed

1. Generate analytical datasets (link data, apply research criteria and filters)
2. Perform analyses
3. Generate statistical and machine learning models

1. Share findings
2. Share code and methods
3. Share data

High risk

PHS Data Portal



phsdata.stanford.edu

Looking forward

- Research on incentives
- New tools
- Grants to share what we've learned.
- New ontologies
- Data management and annotation
- Data landscaping

Join the conversation online!

@TheNeuro_MNI
#OpenScienceinAction