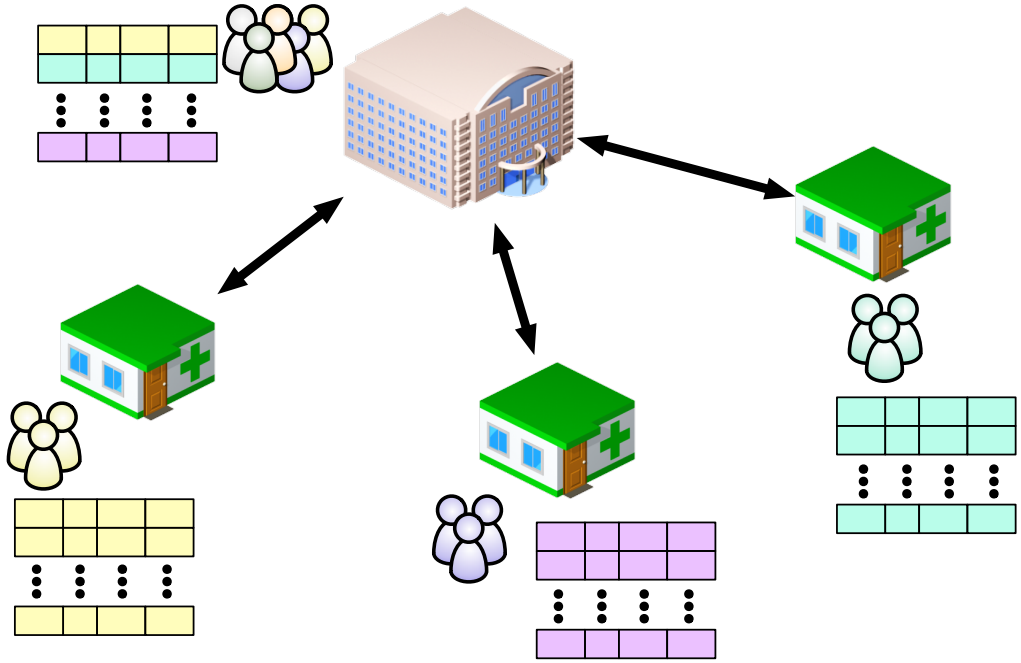
Privacy Protections as an Incentive for Collaborative Research on Human Health

Anand D. Sarwate

Department of Electrical and Computer Engineering
Rutgers, the State University of New Jersey

April 24, 2017

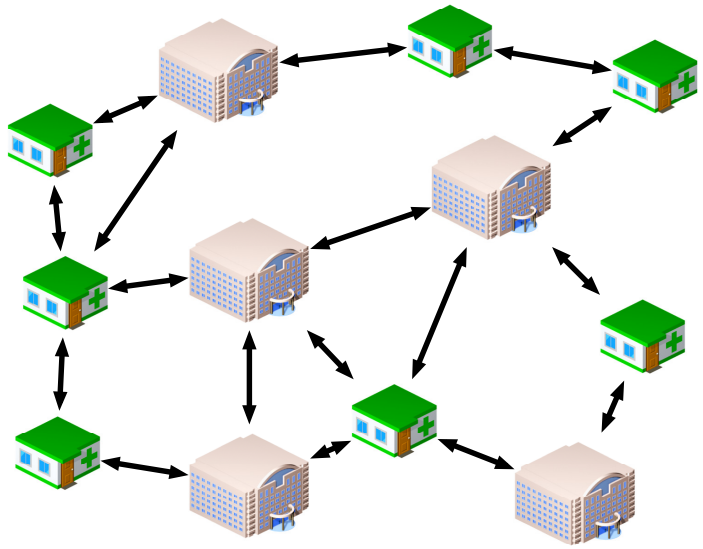
Human health research



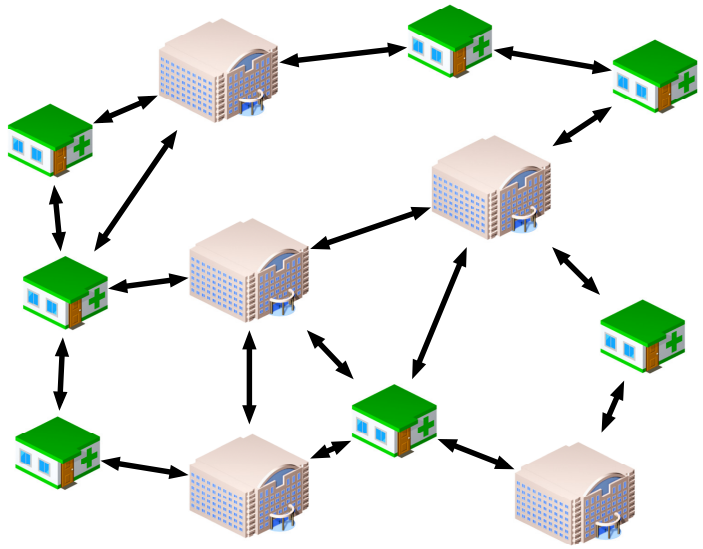
There are many data sharing challenges in human health research

- Secondary use of clinical data for research
- Multi-site studies on QA or comparative effectiveness
- *Joint (secondary) analyses on aggregated research data*

Institutions often want to share data

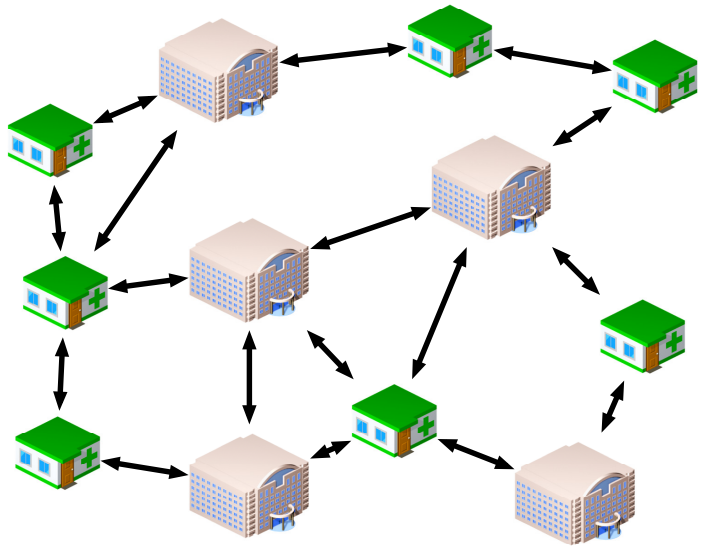


Institutions often want to share data



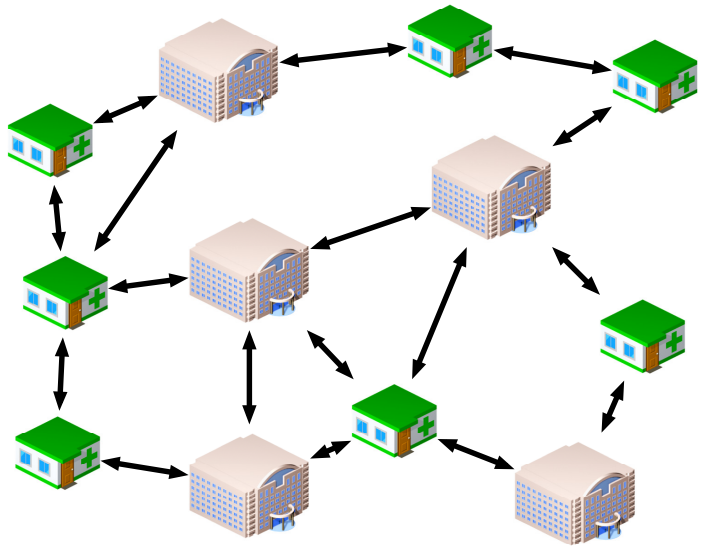
- Different research groups using the same type of measurements want to do a joint analysis.

Institutions often want to share data



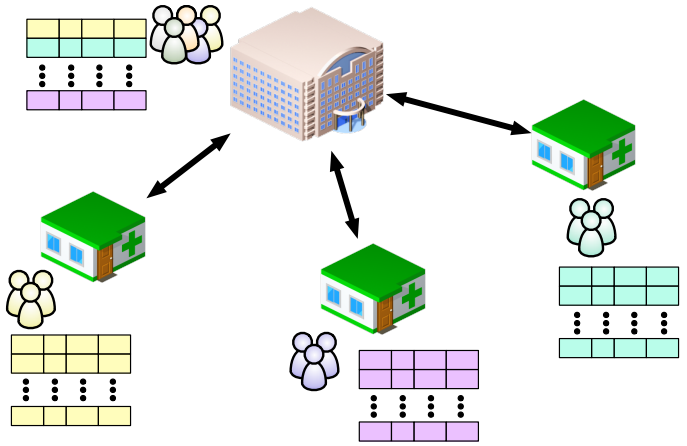
- Different research groups using the same type of measurements want to do a joint analysis.
- Sharing requires lawyers at each institution to generate Data Use Agreements.

Institutions often want to share data



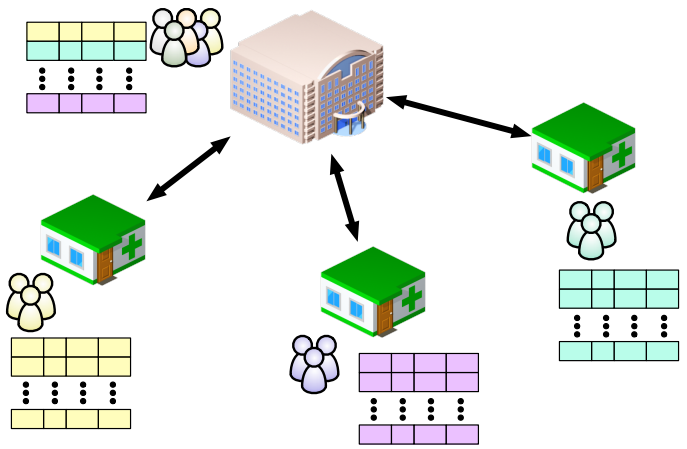
- Different research groups using the same type of measurements want to do a joint analysis.
- Sharing requires lawyers at each institution to generate Data Use Agreements.
- Resulting months of negotiation makes even small-scale collaboration too complicated.

Collaborative research systems



Research consortia are common in many research areas involving human health:

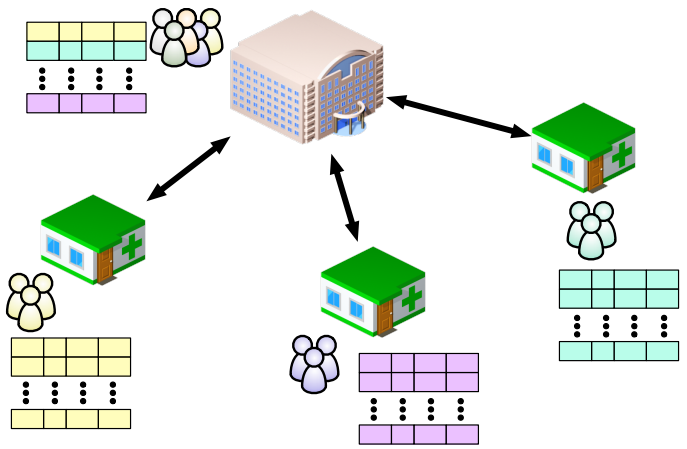
Collaborative research systems



Research consortia are common in many research areas involving human health:

- Foster collaborative research about a particular condition (Alzheimer's, autism, breast cancer, etc.)

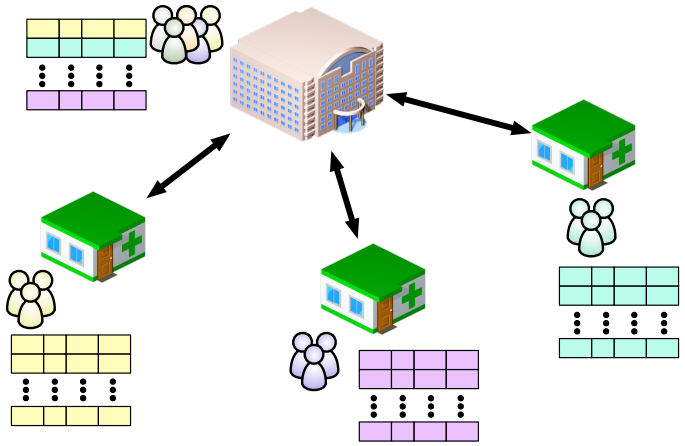
Collaborative research systems



Research consortia are common in many research areas involving human health:

- Foster collaborative research about a particular condition (Alzheimer's, autism, breast cancer, etc.)
- Automated sharing is challenging, but this is changing.

Collaborative research systems



Research consortia are common in many research areas involving human health:

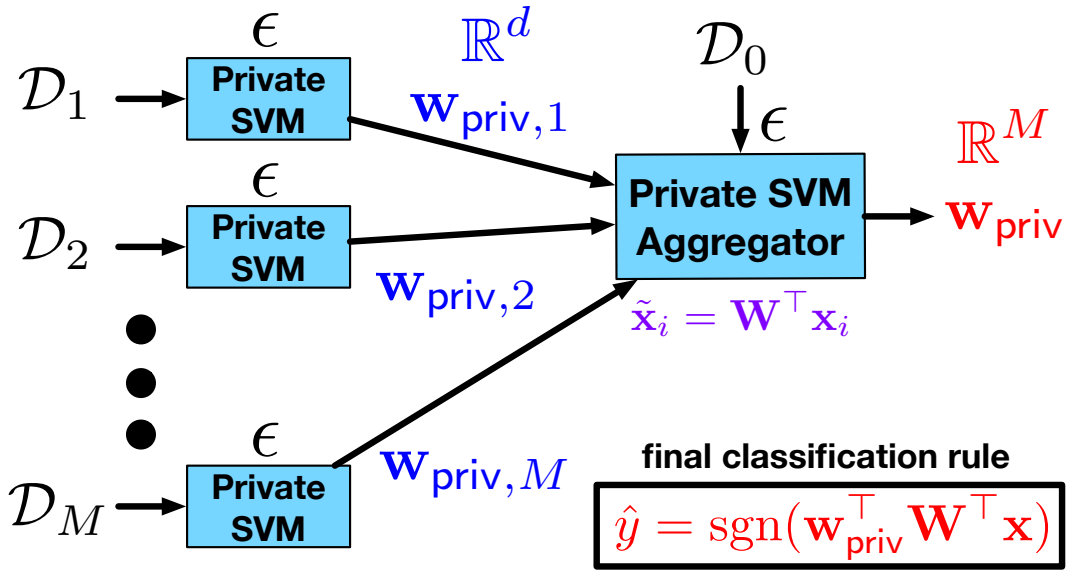
- Foster collaborative research about a particular condition (Alzheimer's, autism, breast cancer, etc.)
- Automated sharing is challenging, but this is changing.

Goal: use privacy protections to encourage consortium growth.

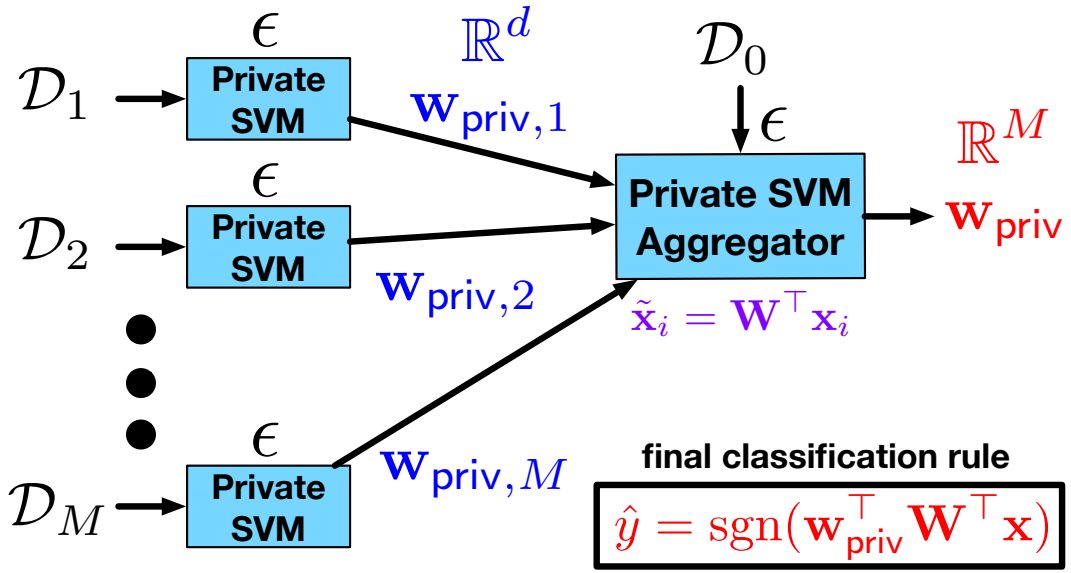
COllaborative Informatics Neuroimaging Suite

- End-to-end system for managing data for studies on the brain
- Current usage: 37,903 participants in 42,961 scan sessions from 612 studies for a total of 486,955 clinical assessments.
- Data from 34 states, 38 countries
- Partners with research consortia such as the Autism Brain Imaging Data Exchange (ABIDE)

Example: schizophrenia research

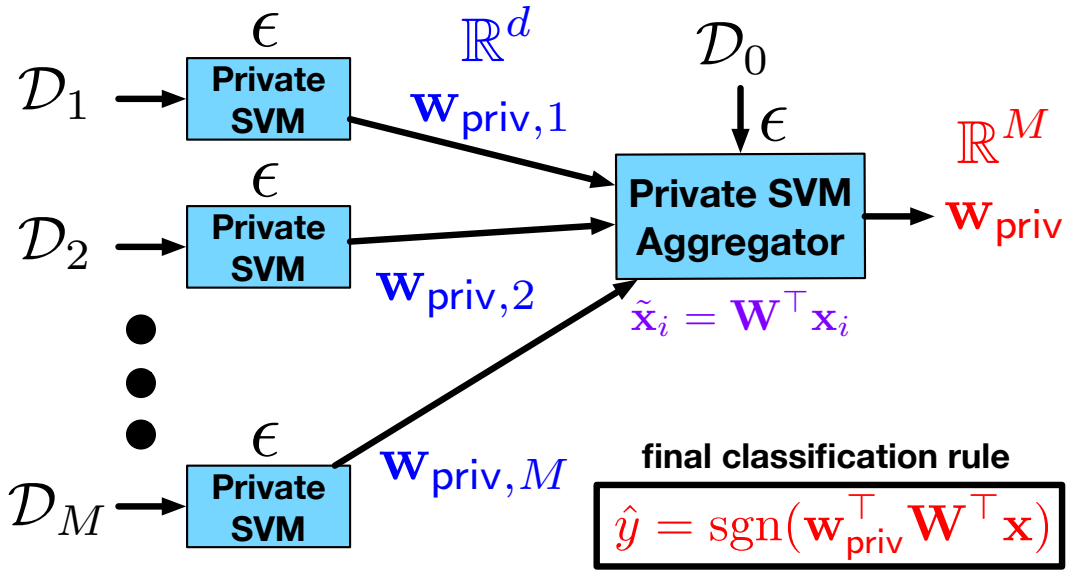


Example: schizophrenia research



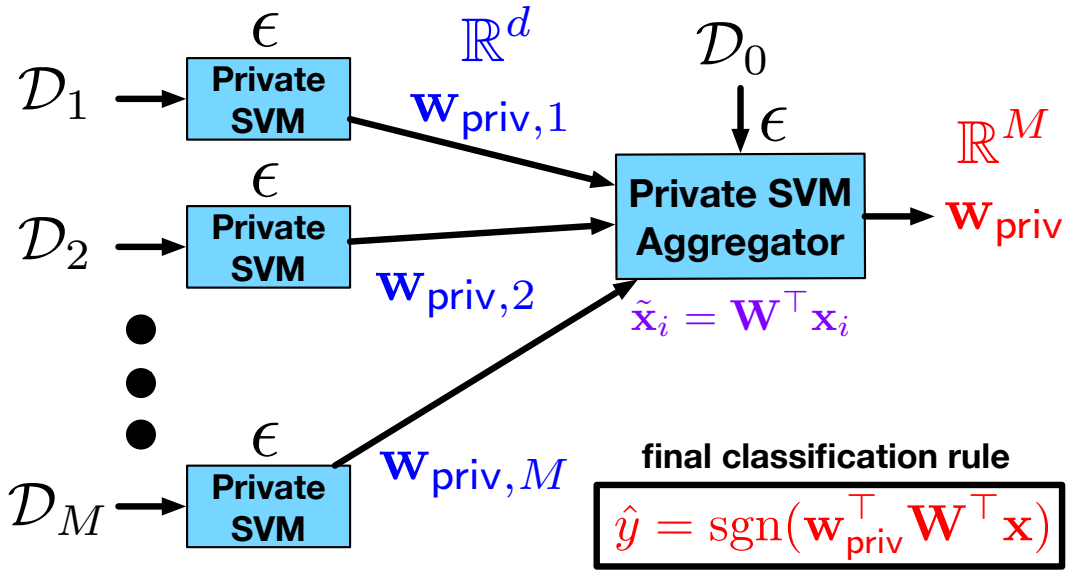
- **Goal:** build a system that can identify schizophrenia.

Example: schizophrenia research



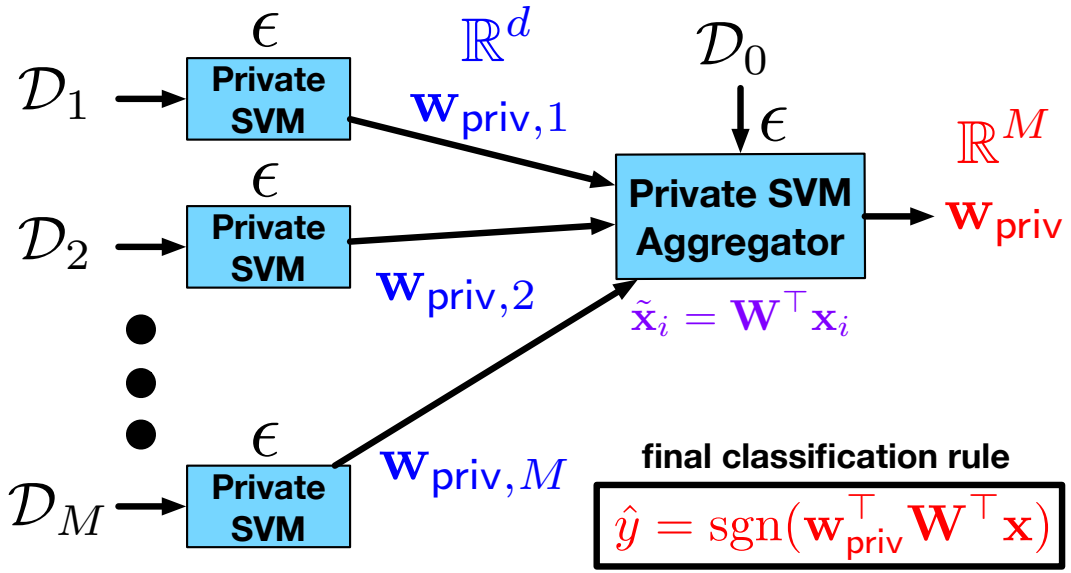
- **Goal:** build a system that can identify schizophrenia.
- **Data:** MRIs from multiple studies (healthy controls and schizophrenics).

Example: schizophrenia research



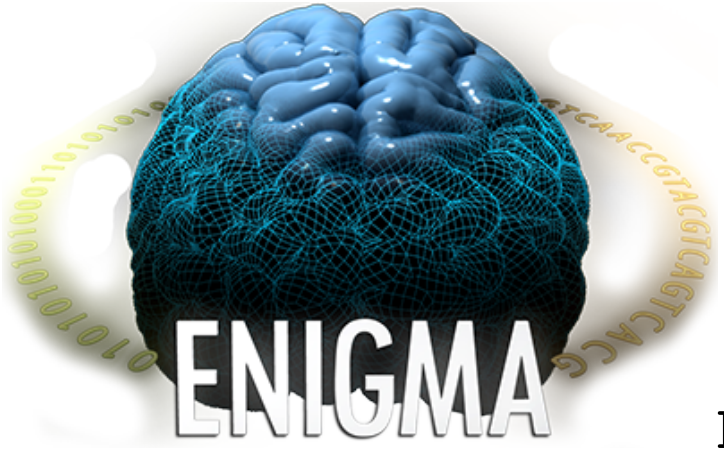
- **Goal:** build a system that can identify schizophrenia.
- **Data:** MRIs from multiple studies (healthy controls and schizophrenics).
- **Algorithm:** classification using machine learning (e.g. support vector machine).

Example: schizophrenia research



- **Goal:** build a system that can identify schizophrenia.
- **Data:** MRIs from multiple studies (healthy controls and schizophrenics).
- **Algorithm:** classification using machine learning (e.g. support vector machine).
- **Privacy risk:** each study has to allow access to sensitive subject data.

State of the art: ENIGMA

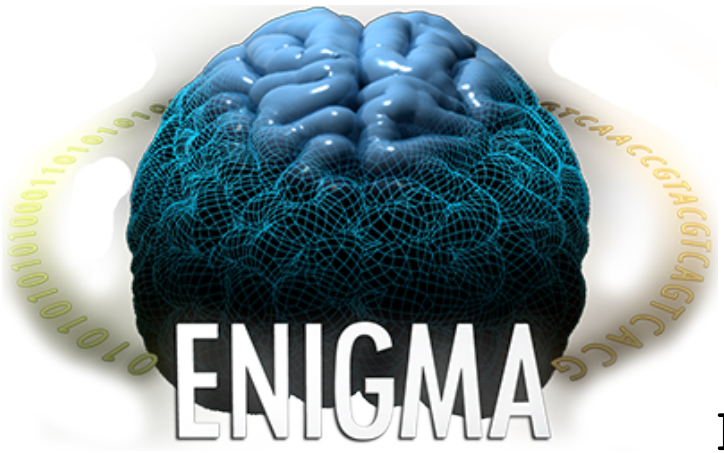


http://enigma.ini.usc.edu

"The ENIGMA Network brings together researchers in imaging genomics to understand brain structure, function, and disease, based on brain imaging and genetic data."

- MA = meta analysis : focused on
- Goals: improve reproducibility, sample sizes
- Validation: found genetic variations associated with neurophysiological characteristics (e.g. hippocampal/intercranial volumes)

Workflows in ENIGMA

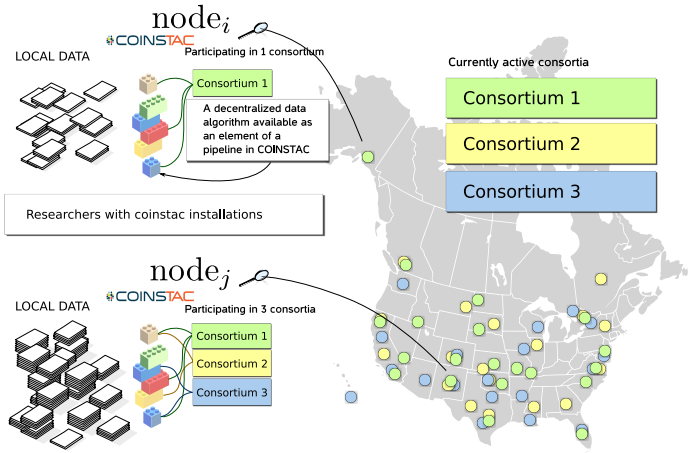


http://enigma.ini.usc.edu

ENIGMA has $30+$ working groups on diseases, genomics, population variation, and methods. To do a study:

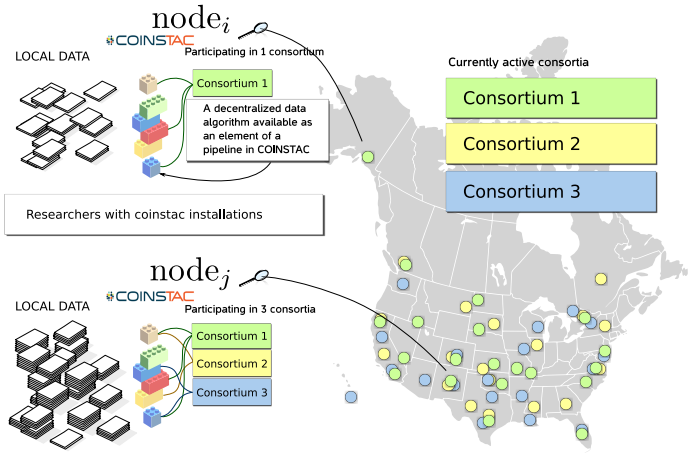
- Study proposal is approved by ENIGMA managers.
- Analyses performed on local sites and emailed to ENIGMA manager as Excel spreadsheets.
- Manager has to perform "manual" meta-analysis.

Low-hanging fruit: automate this



COINSTAC works in a different way: data is registered in the system and analyses are performed/aggregated automatically through message passing.

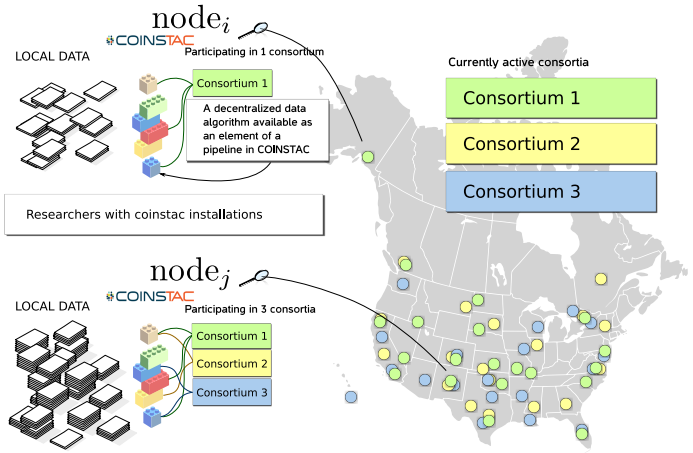
Low-hanging fruit: automate this



COINSTAC works in a different way: data is registered in the system and analyses are performed/aggregated automatically through message passing.

- Study is proposed specifying data needed.

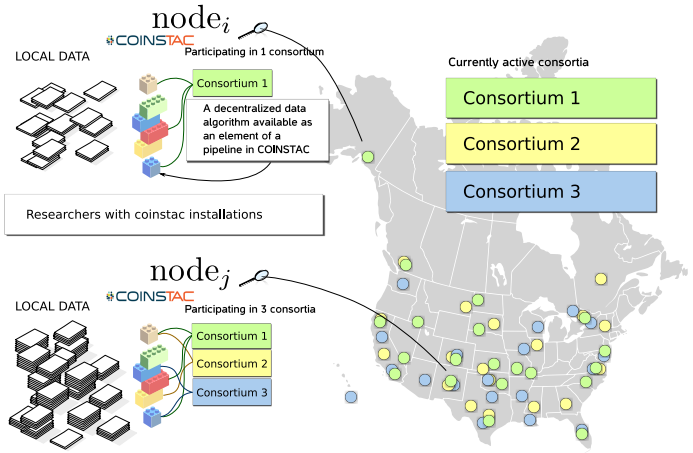
Low-hanging fruit: automate this



COINSTAC works in a different way: data is registered in the system and analyses are performed/aggregated automatically through message passing.

- Study is proposed specifying data needed.
- Local sites approve access to data.

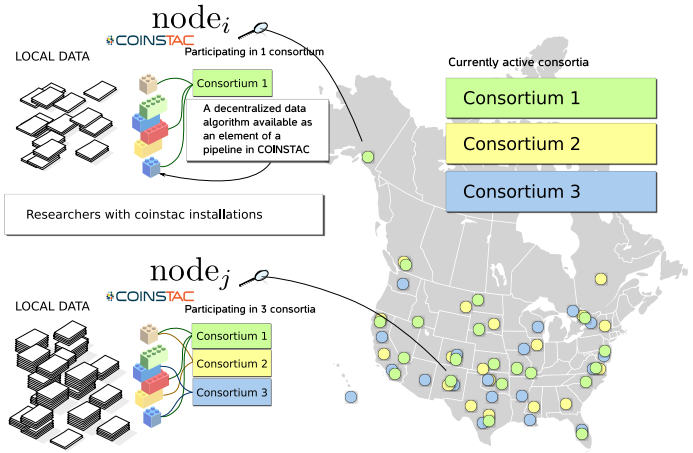
Low-hanging fruit: automate this



COINSTAC works in a different way: data is registered in the system and analyses are performed/aggregated automatically through message passing.

- Study is proposed specifying data needed.
- Local sites approve access to data.
- Analyses are run and aggregated automatically.

Low-hanging fruit: automate this



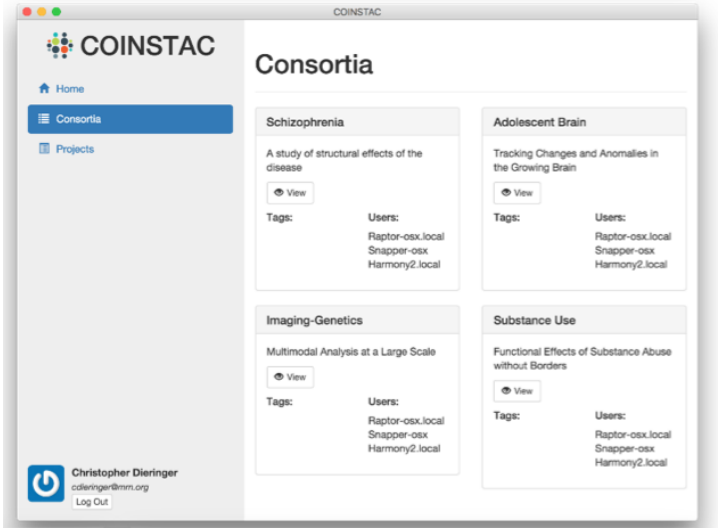
COINSTAC works in a different way: data is registered in the system and analyses are performed/aggregated automatically through message passing.

- Study is proposed specifying data needed.
- Local sites approve access to data.
- Analyses are run and aggregated automatically.

This can be significantly faster than the ENIGMA approach.

The COINSTAC workflow

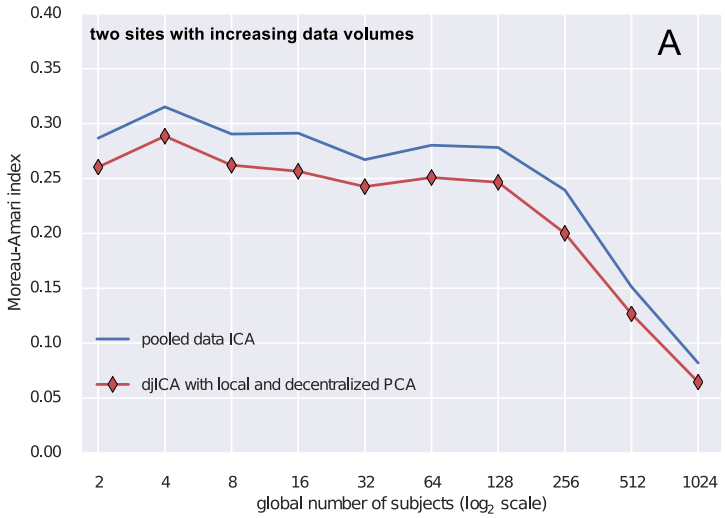
B. Consortia list view



In COINSTAC, research groups install the software and register their data in the system:

- Form ongoing and ad-hoc "consortia" (slow, requires approval)
- Once established, consortium members can initiate a joint analysis
- Computation is performed locally and messages passed between sites

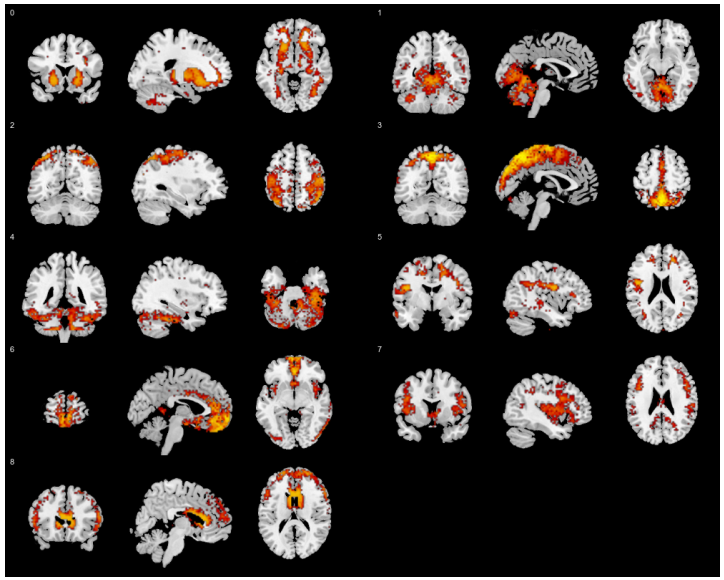
What's in the medium term



COINSTAC prototype is currently "demo-able" but not up and running.

- Compute more than summary statistics, ridge regression, etc.
- Improve user interface and usability for practitioners, including visualization tools.
- Initial subject focus for new results: addiction studies.
- Incorporate/test differentially private methods for machine learning.

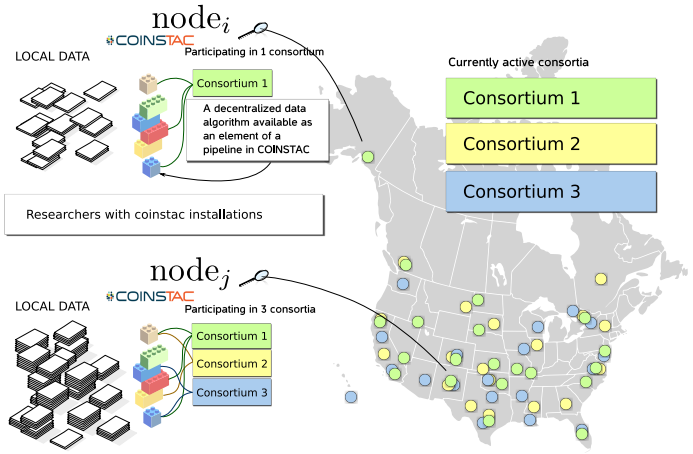
Focusing on "old" algorithms



Because the focus is on usability, we are working on methods popular in neuroimaging:

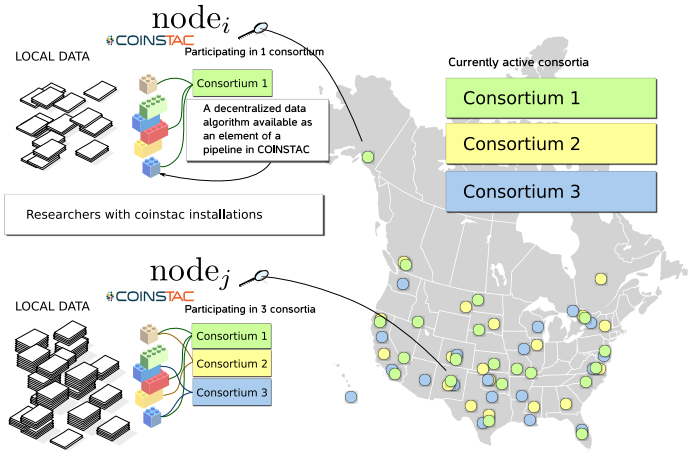
- Feature discovery: ICA, IVA, NMF, deep learning, etc.
- Regression and classification: ridge regression, LASSO, SVM, etc.
- Visualization: t-SNE, network visualization, etc.

COINSTAC vs. other health data systems



COINSTAC is a solution that works for typical neuroimaging research initiatives.

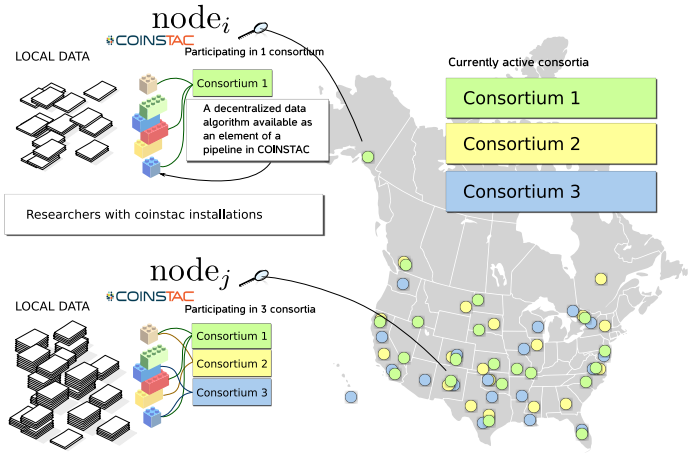
COINSTAC vs. other health data systems



COINSTAC is a solution that works for typical neuroimaging research initiatives.

1. Data is "big" from the *perspective of the domain area*.

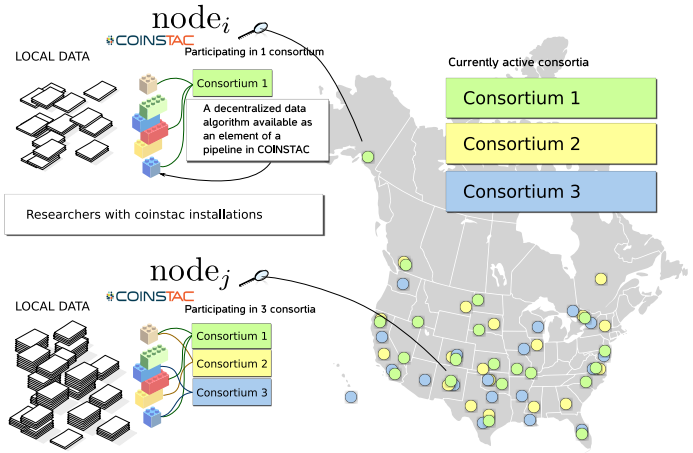
COINSTAC vs. other health data systems



COINSTAC is a solution that works for typical neuroimaging research initiatives.

1. Data is "big" from the *perspective of the domain area*.
2. Methods with asymptotic guarantees may not be ideal.

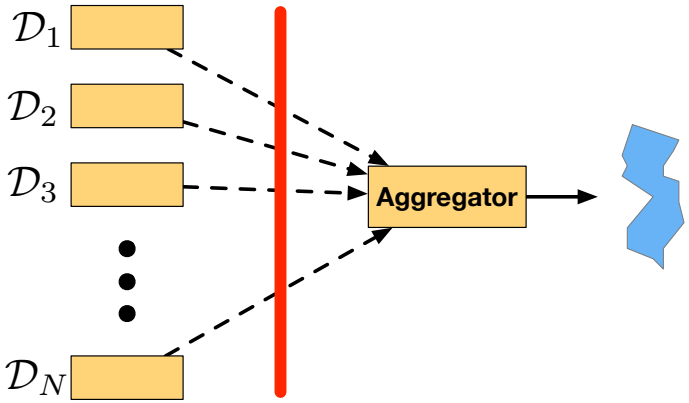
COINSTAC vs. other health data systems



COINSTAC is a solution that works for typical neuroimaging research initiatives.

1. Data is "big" from the *perspective of the domain area*.
2. Methods with asymptotic guarantees may not be ideal.
3. Strong formal privacy may be trumped by utility requirements.

Sad news: no privacy is enough?



From the perspective of IRBs and other regulatory bodies, decentralized/distributed algorithms may be "good enough."

- Getting them to work on the computing infrastructure is itself challenging.
- Threat models and surface are different than "typical" data sharing scenarios.
- Provides a useful test case for "newer" privacy technologies: differential privacy, multiparty computation.

Making formal privacy guarantees

Currently working on making *differentially private* versions of existing algorithms. Differential privacy involves introducing randomization (e.g. noise) in computations.

- Small number of subjects $\rightarrow$ larger noise $\rightarrow$ more error.
- Neuroimaging data is high-dimensional: need some dimension reduction.
- Preference for stronger $(\varepsilon, 0)$ guarantees, but improved analyses give (ε, δ).

Dealing with federated infrastructure

Uploading all the data to EC2 or Azure is not an acceptable (yet?)

Dealing with federated infrastructure

Uploading all the data to EC2 or Azure is not an acceptable (yet?)

- Local storage overhead can be challenging.

Dealing with federated infrastructure

Uploading all the data to EC2 or Azure is not an acceptable (yet?)

- Local storage overhead can be challenging.
- Local processing costs are heterogeneous.

Dealing with federated infrastructure

Uploading all the data to EC2 or Azure is not an acceptable (yet?)

- Local storage overhead can be challenging.
- Local processing costs are heterogeneous.
- Communication can act as a real bottleneck.

Compromises, compromises

At the moment we are making many compromises:

- Utility first: practical values of ε for differential privacy are large.
- Low communication: focus on one-shot aggregation over iterative methods.
- Simple tasks: stick with developing distributed methods for known algorithms.

Policy and privacy and systems, oh my!

Data sharing in health research may be different than open sharing or industry/academia sharing.

- Different regulations around human subjects for experimental data or for PHI in clinical data.
- Informed consent model allows *subject-level* and *study-level* privacy preferences.
- Data sharing is contingent and possibly transient: revert to access only.

Recap

Shared-access models with "privacy protections" (formal or not) can encourage researchers to join consortia.

- Benefits/risks align with the desires of data holders/researchers.
- Data holders retain control over access and allowed computations.
- Data users can use automated computations for hypothesis generation.

Some lessons learned



Some lessons learned



- start small: variability of problem types is large

Some lessons learned



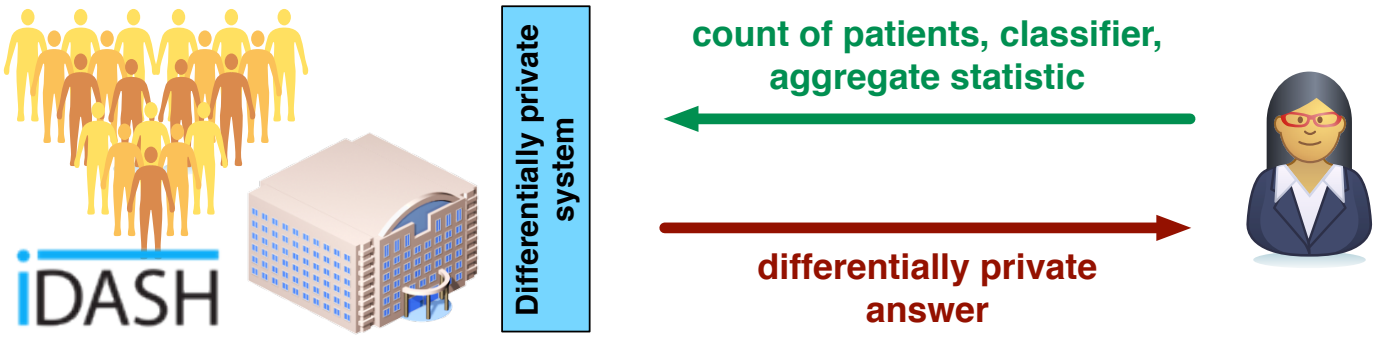
- start small: variability of problem types is large
- challenging to bridge gaps between algorithmist and developer

Some lessons learned



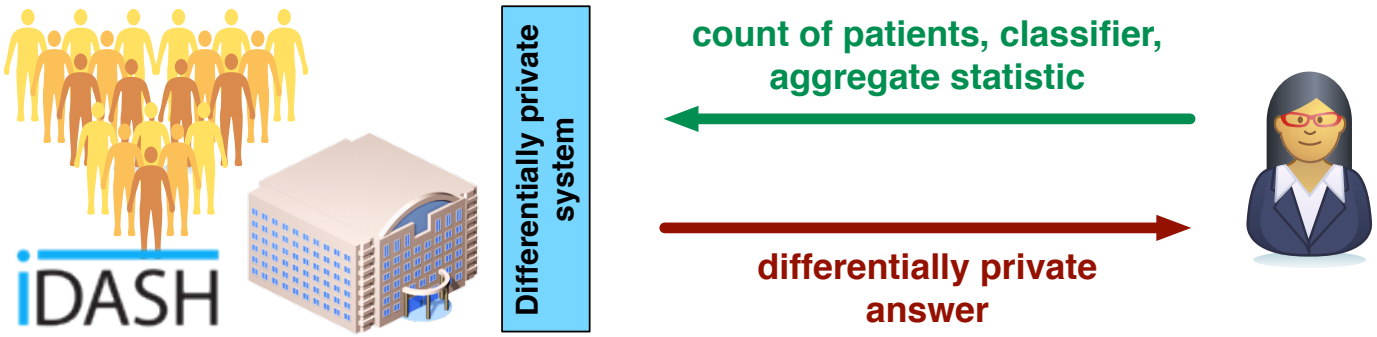
- start small: variability of problem types is large
- challenging to bridge gaps between algorithmist and developer
- communication requirements are nontrivial

Another application: secondary use of clinical data



The iDASH center at UCSD is working on larger-scale human health research involving clinical records.

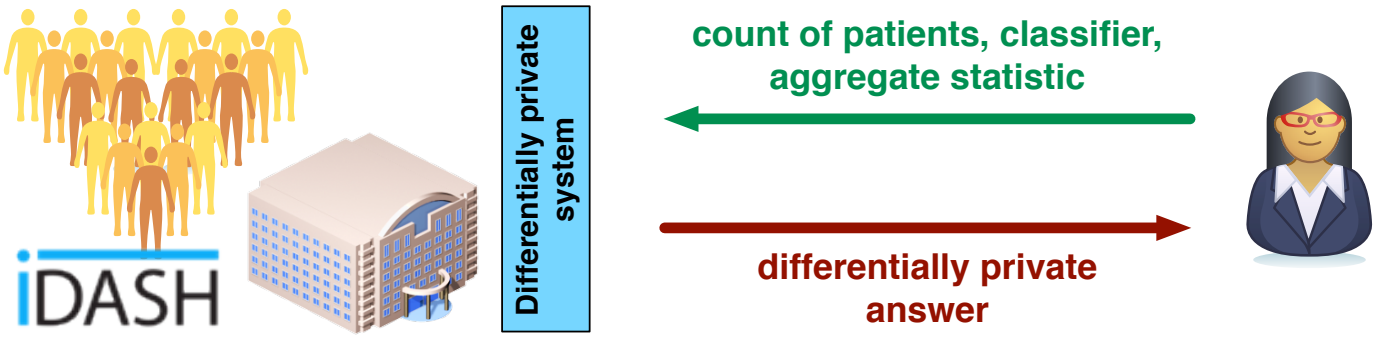
Another application: secondary use of clinical data



The iDASH center at UCSD is working on larger-scale human health research involving clinical records.

- **Goal:** to make clinical data warehouse more useful to researchers

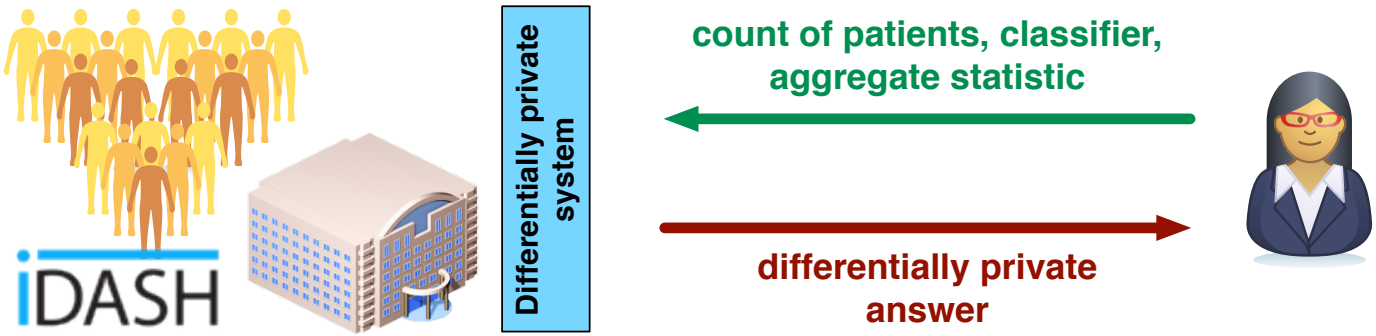
Another application: secondary use of clinical data



The iDASH center at UCSD is working on larger-scale human health research involving clinical records.

- **Goal:** to make clinical data warehouse more useful to researchers
- Diverse range of problems in compression, genomics, NLP, etc. with privacy

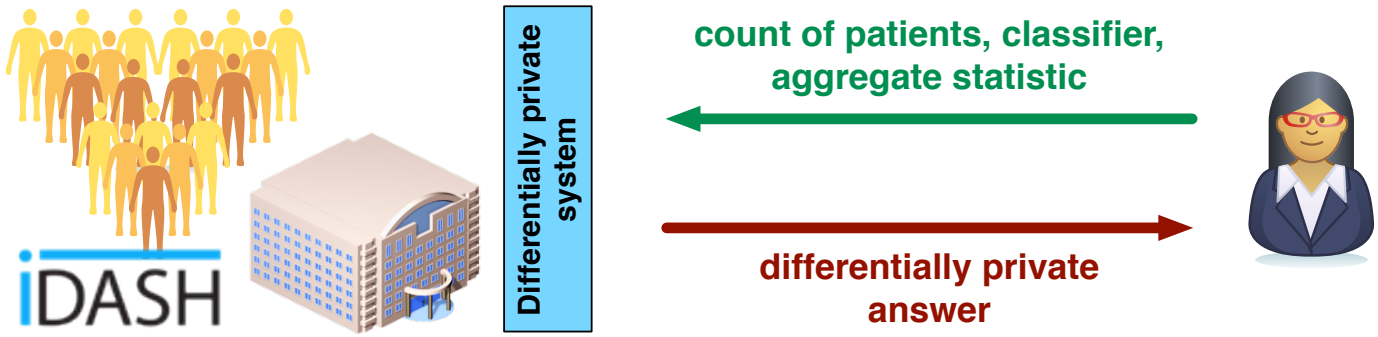
Another application: secondary use of clinical data



The iDASH center at UCSD is working on larger-scale human health research involving clinical records.

- **Goal:** to make clinical data warehouse more useful to researchers
- Diverse range of problems in compression, genomics, NLP, etc. with privacy
- Spurring research transition through data challenges, internships, etc.

Another application: secondary use of clinical data

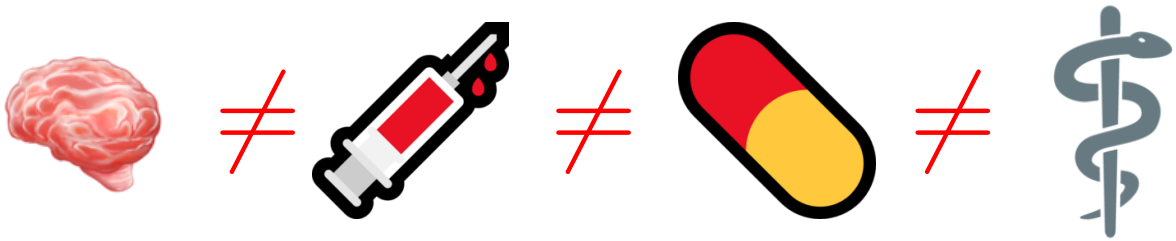


The iDASH center at UCSD is working on larger-scale human health research involving clinical records.

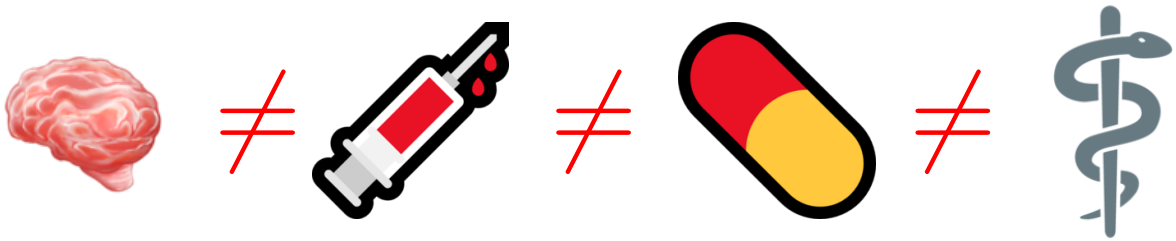
- **Goal:** to make clinical data warehouse more useful to researchers
- Diverse range of problems in compression, genomics, NLP, etc. with privacy
- Spurring research transition through data challenges, internships, etc.

The features of these problems are very different!

Moving the hub forward w.r.t. health

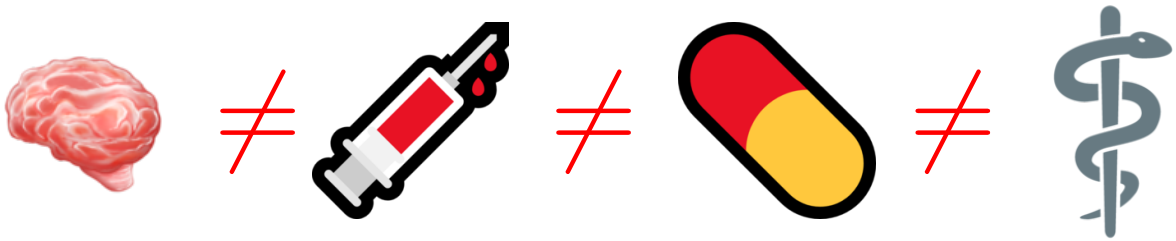


Moving the hub forward w.r.t. health



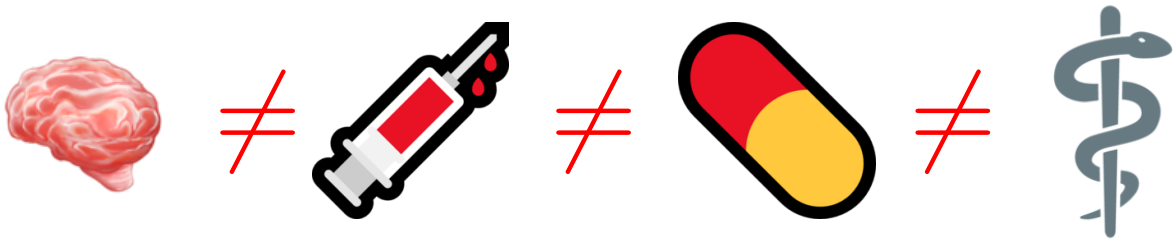
- Recognize that "medical data" is at best a placeholder and at worst semantically void.

Moving the hub forward w.r.t. health



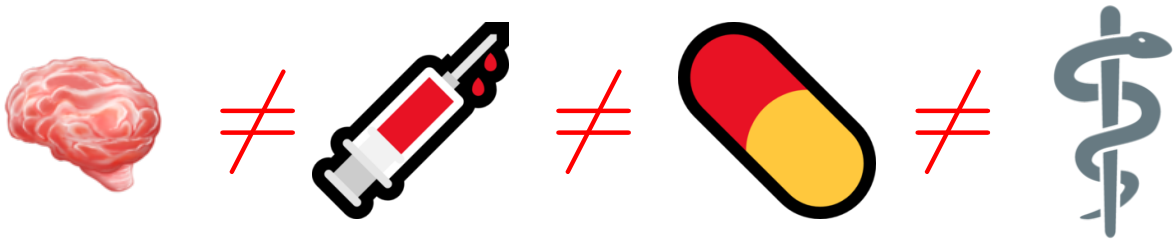
- Recognize that "medical data" is at best a placeholder and at worst semantically void.
- Spend some time delineating the problem space and domain-specific challenges.

Moving the hub forward w.r.t. health



- Recognize that "medical data" is at best a placeholder and at worst semantically void.
- Spend some time delineating the problem space and domain-specific challenges.
- For theorists: can we get out of asymptopia?

Moving the hub forward w.r.t. health



- Recognize that "medical data" is at best a placeholder and at worst semantically void.
- Spend some time delineating the problem space and domain-specific challenges.
- For theorists: can we get out of asymptopia?
- For practitioners: what do you want to *do* versus *how* do you want to do it?