Challenges in Privacy-Preserving Learning for Collaborative Research Consortia

Anand D. Sarwate
Rutgers, The State University of New Jersey

Collaborative research on human health

There are many data sharing challenges in human health research:

- Secondary use of clinical data for research: can we use existing hospital records for tasks such as comparative effectiveness research?

- Designing multi-site studies: multi-site clinical trials, meta-analyses on original data, etc.

- *Collaborative research/data sharing initiatives to get population statistics from research subjects.*

Research consortia for human health

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

- focused on specific conditions: Alzheimer's, autism, breast cancer, etc.

- strong mandate to share data (e.g. from the NIH)

- significant concerns about privacy and ethics

Privacy technologies can help research consortia

Offering privacy protections can incentivize researchers to join research consortia:

- Allow research groups to hold and maintain “control” over their data.
- Need to design software systems to allow consortium members to run analyses
- What is “privacy” in this context?

State of the art: ENIGMA

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

http://enigma.ini.usc.edu

- Improve reproducibility, sample sizes by allowing easier meta-analyses.
- Example : genetic variation associated with intercranial and hippocampal volumes.
- 30+ working groups on a wide range of conditions and topics.

ENIGMA Workflow

- Study proposal is approved by ENIGMA managers.
- Analyses performed on local sites and emailed to ENIGMA manager as Excel spreadsheets.

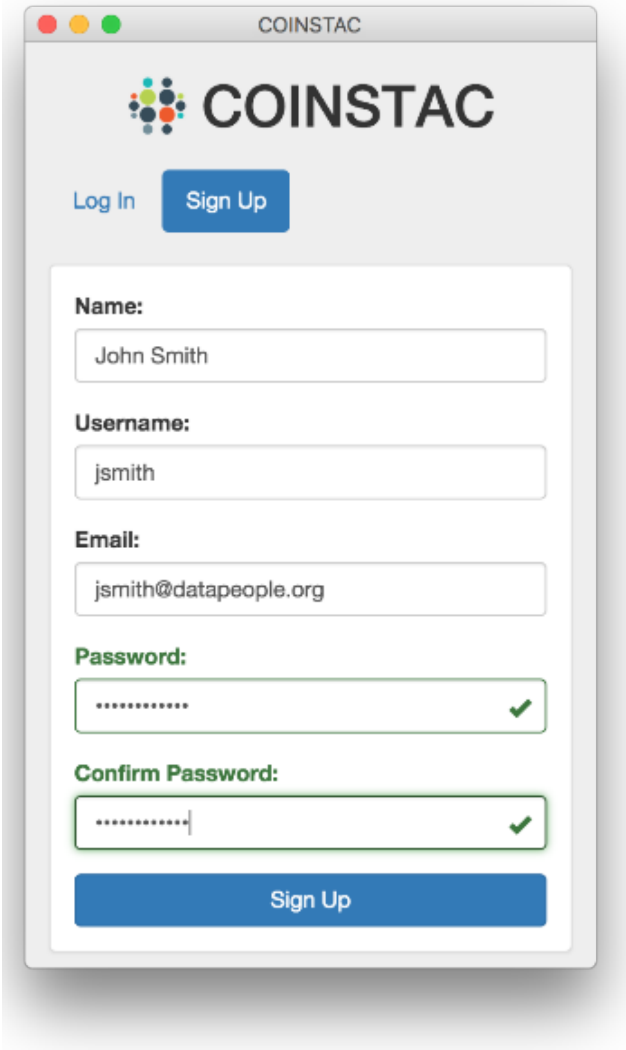
- Manager has to perform ``manual'' meta-analysis.

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

COINSTAC

A. Account creation and login

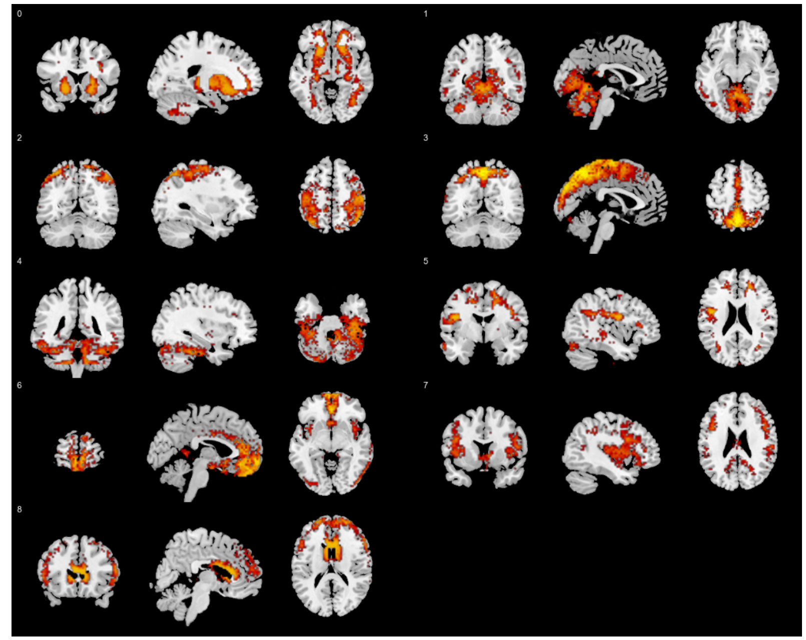


Extend COINS to to allow automated analyses:

- register data sets in COINSTAC
- perform automated analyses using message passing
- data held locally, analyses run automatically

Typical applications

Focus on popular neuroimaging tools:



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

What about privacy?

What sort of privacy can we guarantee in a system like COINSTAC?

- hand-waving: “data is held locally”

- formal: develop DP algorithms for neuroimaging tasks

Building DP into COINSTAC

- Designing decentralized/distributed versions of some of these algorithms is sometimes open.
- Given a distributed algorithm, we can apply “the DP toolkit” to make a DP version.
- “Utility first” approach to setting ε.

Challenges

Types of challenges in using DP:

- 📈 Statistical
- ⚒️ Definitional
- ⚙️ Algorithmic
- ⚖️ Policy

Small n, large p

The goal is to leverage multiple data sets to get larger sample size to learn about the population:

- Number of samples is still small. MRIs are big.
- Constants matter, log factors matter.
- Algorithm performance is very data dependent.
- How can we understand non-asymptotic performance?

Types of algorithms

Much of the work in differentially private learning has been driven by trends in “big data:”

- Other domains often have preferred tools/methods.
- Visualization is very important.
- How should we expand the “basic toolkit” to allow easier development of these tools?

ε- versus (ε,δ)-DP

Practitioners want stronger privacy guarantees: $\delta = 0$.

- Risk averse: nonzero δ is seen as unacceptable.
- Practically: choosing $\delta \approx 1/n$ destroys utility.
- Strong composition rules are nice, but may not help as much: can we get better $(\varepsilon,0)$ algorithms or help make smaller δ practical?

Multi-stage algorithms

Computational analyses in neuroimaging involve processing *pipelines*:

- Many (or all) stages need to guarantee DP.
- How should we think about allocating privacy risk across stages? Is there something better than empirical?
- Pipelines are used more than once: can we reuse parameter tuning to ease overall privacy loss?

Prior domain knowledge

Domain experts either “know” or assume “w.l.o.g.” many things about their data:

- Priors are a good way to incorporate this information, but knowledge may not be explicitly encoded Bayes-style.
- Restricting the data domain (or database schema) seems like a good start, but many prior assumptions are about the “population.”
- What kind of property testing methods should we use/develop? Is local sensitivity enough?

Trust models in consortia

Research consortia have different trust models and assumptions.

- Extreme view 1: everyone is trusted here, this is just between “friends” etc.

- Extreme view 2: I’m not going to let those #$%^# look at my hard-earned data.

- In reality, we operate somewhere in between…

Less pessimistic models

Much of the utility loss comes from conservative (strong) threat modeling in DP:

- Real workflows will require significant interaction with the data

- Privacy budgets may need to be renewed, privacy restrictions may expire.

- Are there some relaxations or different threat models (or modified privacy definitions) that are appropriate for these systems?

Lessons learned

- Good application domains have (i) mandate or desire for data sharing that is (ii) hampered by privacy concerns
- Not all algorithms/problems may be appropriate for differential privacy (at least for now).
- Accept large ε, at least initially.

Thank you!

Vince Calhoun
(MRN)

Sergey Plis
(MRN)

Jessica Turner
(Georgia State)