# Looking at Everything in Context:

## *Community-Scale Data Integration for Real*


Zachary G. Ives University of Pennsylvania

with **Z. Yan, N. Zheng,** B. Litt, J. Wagenaar


CIDR 2015 / January 5, 2015


Funded in part by
NSF IIS-1217798,
NIH 5U24NS063930,
and a gift from Google

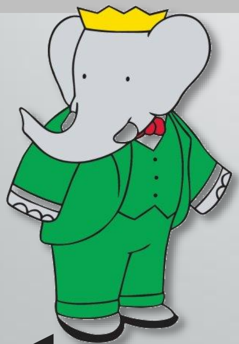
# The Spectrum of Data Management

**Database / Warehouse-ETL / EII**

Mandated standards

Requires **human**-developed ETL, curation

**Central authority, $$$**



***"Open" Data Integration***

Structured data with an uncertain scope / domain

Requires semi-automated solutions!


**Web Search / WebTables**

Heterogeneous, partly structured data, spam

Exploits **machine learning**, pattern matching

**Scale, workload, link struct.**

Closed-domain

Open, mid-scale, *dynamic* domain

Open, large-scale domain-agnostic

# Open Data Integration: Much Progress, or Little Progress?

Many fundamental advances the past decade to *semi-automate* certain layers of the open integration "stack"!

- Machine learning, better **matching/linking** algorithms (LSD, COMA, etc.; Tamer), better **extraction** algorithms (DeepDive; System T)
- **Human-machine**: Pay-as-you-go (dataspaces, etc.), crowdsourcing, p2p mediation, …
- Scalable compute platforms (cloud, cluster), more robust Internet infrastructure, …

Yet: few community-scale, end-to-end integration success stories

- *Why are they so rare?*

[Applications] *Lack of access to, and experience with, real data & problems!*
[Platforms] *Lack of platforms combining best-of-breed components!*
[Users] *Lack of ability to build user communities*

# Real Applications as Community Resources

# How Do We Create a Lens into Real Community Data Sharing?

Data is now easy to get – but we are missing the *context* of how it's used!

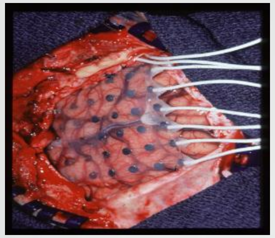
- How do we get access to enough users to learn where the bottlenecks are?
- Consider that Google, Facebook, etc. credit access to workloads, A/B testing as a huge enabler of improvement in their systems

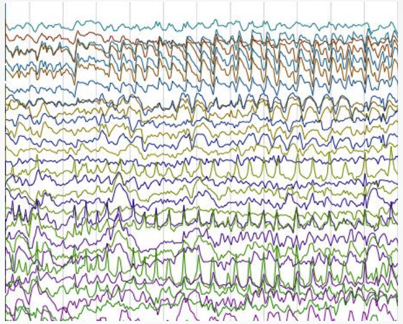
Can **a few of us** build "research instruments" that the **community** can leverage to evaluate new data integration algorithms?

- analogous to PlanetLab, EmuLab in networking

Key to applications: collaborators with vision, influence on diverse communities!

# Our Efforts in this Space:
# Neuroscience/Electrophysiology as a 1st Foothold

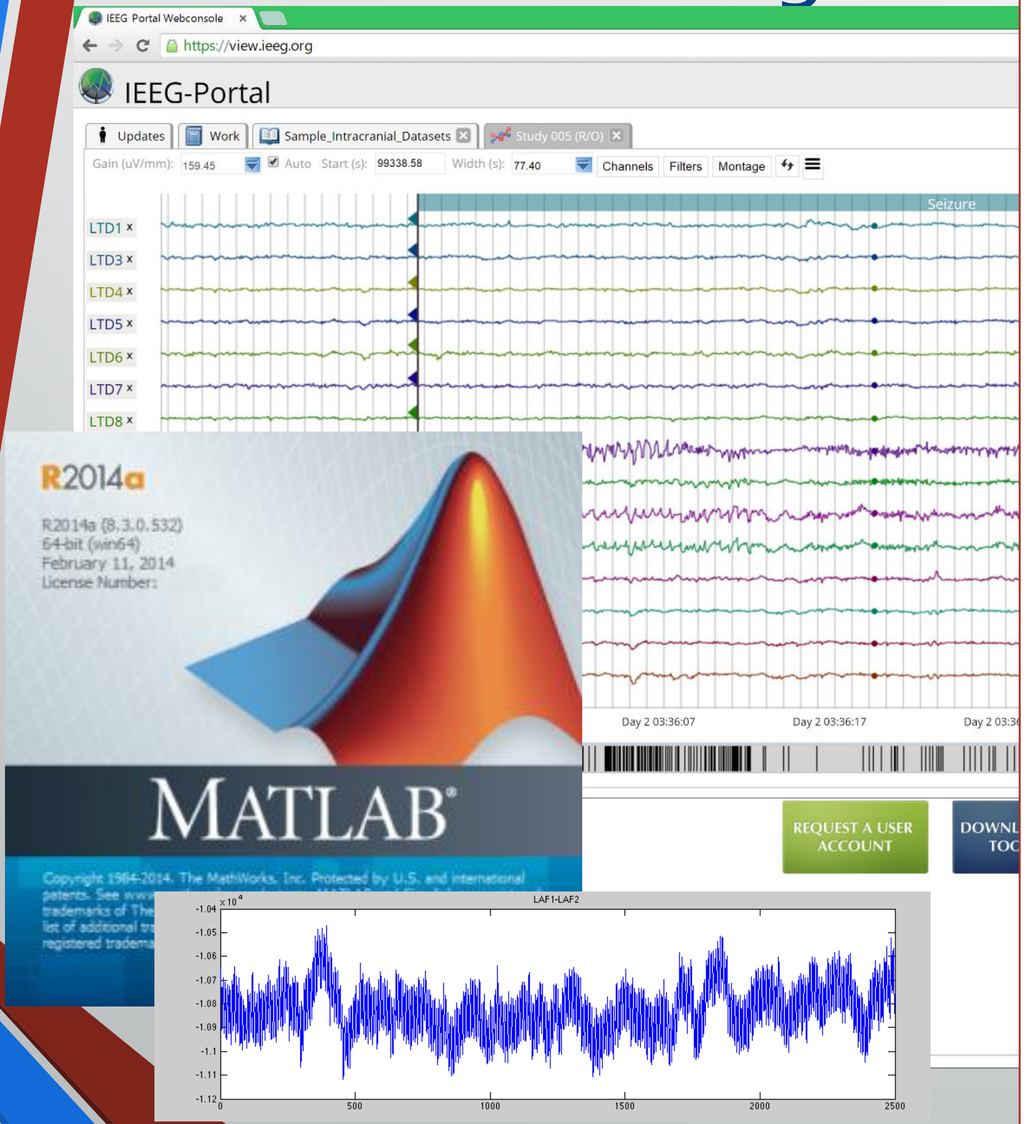


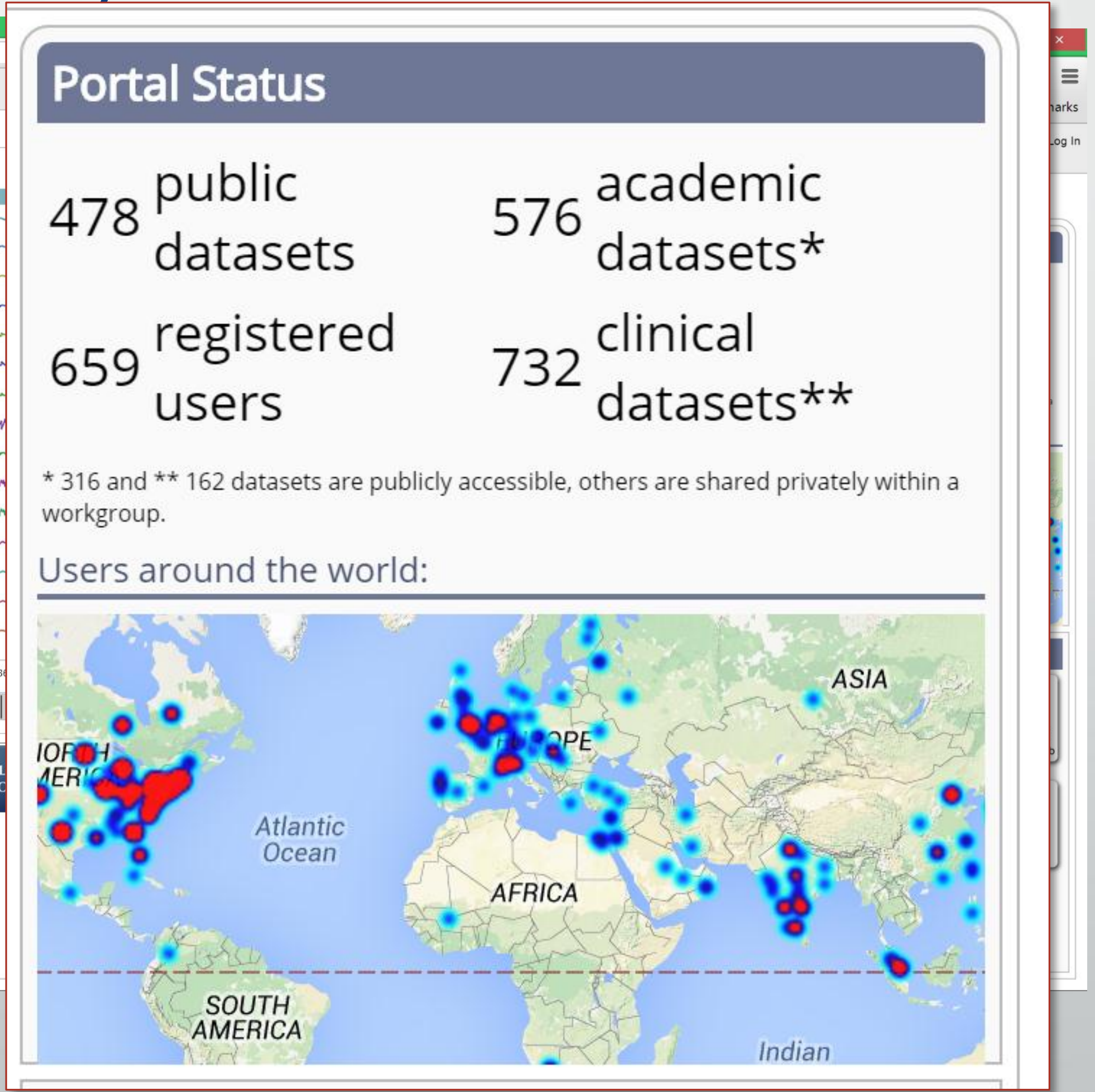



Electrophysiology – key to understanding many brain activities and developing treatments

- No practice of data sharing
- Limited infrastructure to displace, “hunger” for new solutions!

# IEEG.org: Neuroscience Data Sharing & Analysis on the Cloud



# Neuroscience as a Lens into Real Scientific Community Data Sharing

Many aspects of IEEG.org are standard cloud/Web/DBMS, but gives us:

- multi-modal **data and metadata**
  - (10+TB, 25+ academic, device partners)
- over **600 real users** in heterogeneous **communities**
  - (epilepsy, behavioral neuroscience, brain-computer interface, implantable/wearable devices)

Goal: testbed and user community to enable **user studies**

- Evaluate, improve **algorithms for automating integration** tasks
  - Each new lab, data modality → new integration task
- Evaluate **query answering and learning-from-feedback** techniques

More broadly: can we build a new **architecture** for facilitating such evaluations in context?

# A Proposed Platform for *in situ* Evaluation of Data Integration Techniques

# Supporting Experiments with Real Users: Proposed HABITAT Platform

1. "Pay as you go" integration (i.e., user-driven, iterative process)

2. Modular, pluggable architecture

3. *Evaluation management* to recruit users, do A/B testing

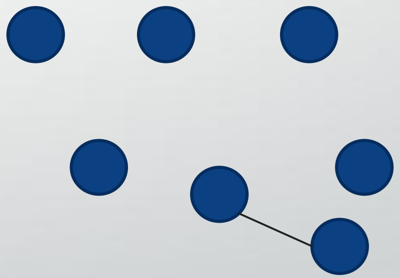
   Figure out what works based on real workloads, usage

# Pay-as-You-Go / Search-Driven Integration

**Ingest**: Offline "partial ETL" as data is discovered / loaded

- Data gets loaded (as feasible) into a weighted "search graph" (~ "data lake")
- Data and metadata as nodes, relationships as edges

**Periodic** workload-driven improvement of data, e.g., when new extractor is developed



# Pay-as-You-Go / Search-Driven Integration

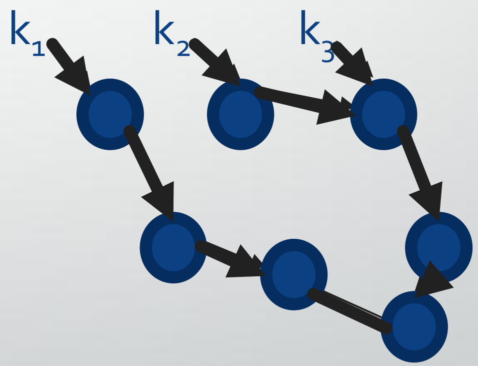
**Ingest**: Offline "partial ETL" as data is discovered / loaded

- Data gets loaded (as feasible) into a weighted "search graph"
- Data and metadata as nodes, relationships as edges

**Periodic** workload-driven improvement of data, e.g., when new extractor is developed

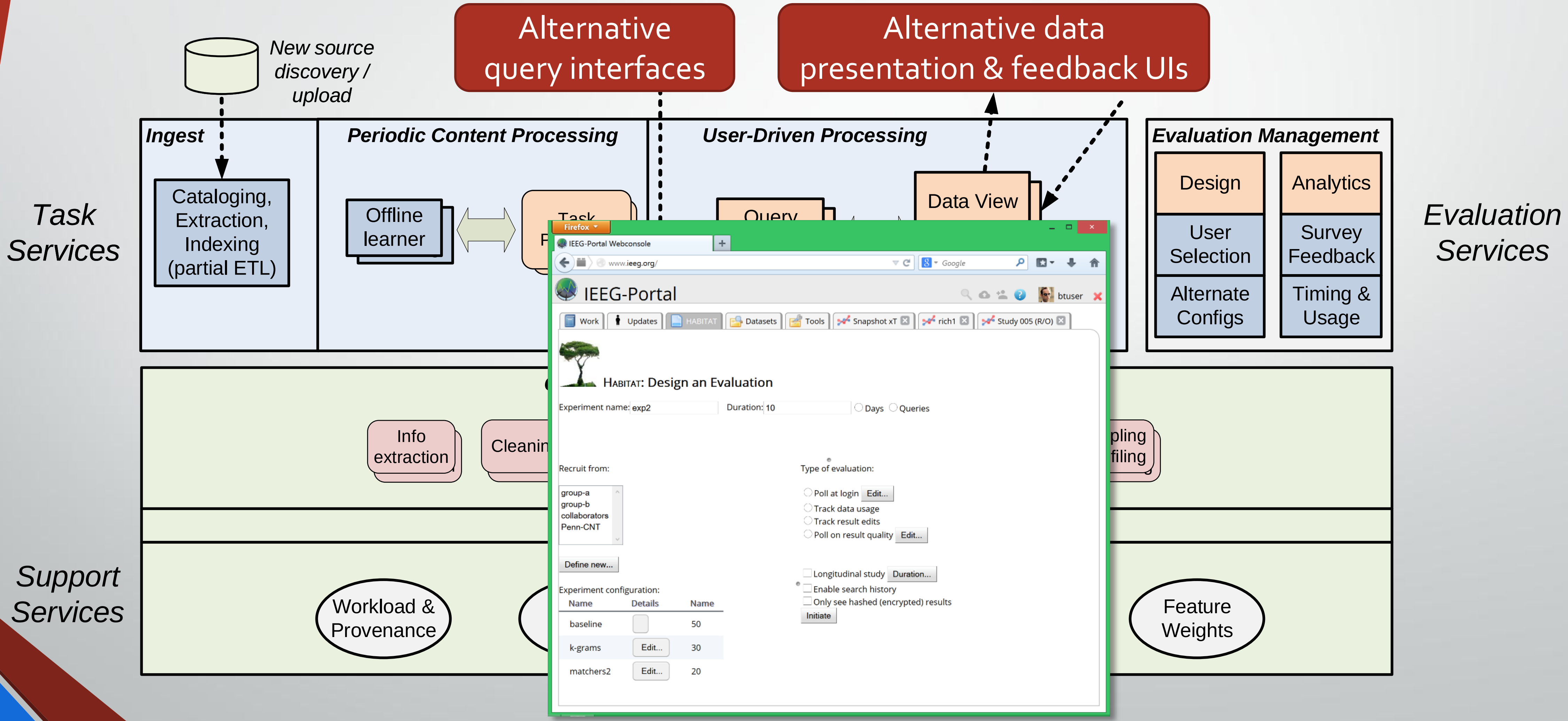
**User-driven** integration: users pose keyword searches over data and metadata [Talukdar+08,10][Yan+13,15]

- Keywords match nodes
- Record linking, schema matching algorithms link nodes
- Query result: a Steiner tree whose leaves are the keywords – presented in a domain-specific way




- User marks answers as good or bad, and the system learns to repair mistakes [Talukdar+10][Yan+13]!

# HABITAT Modular System Architecture



# Status

Current status of HABITAT:  integrating components within IEEG

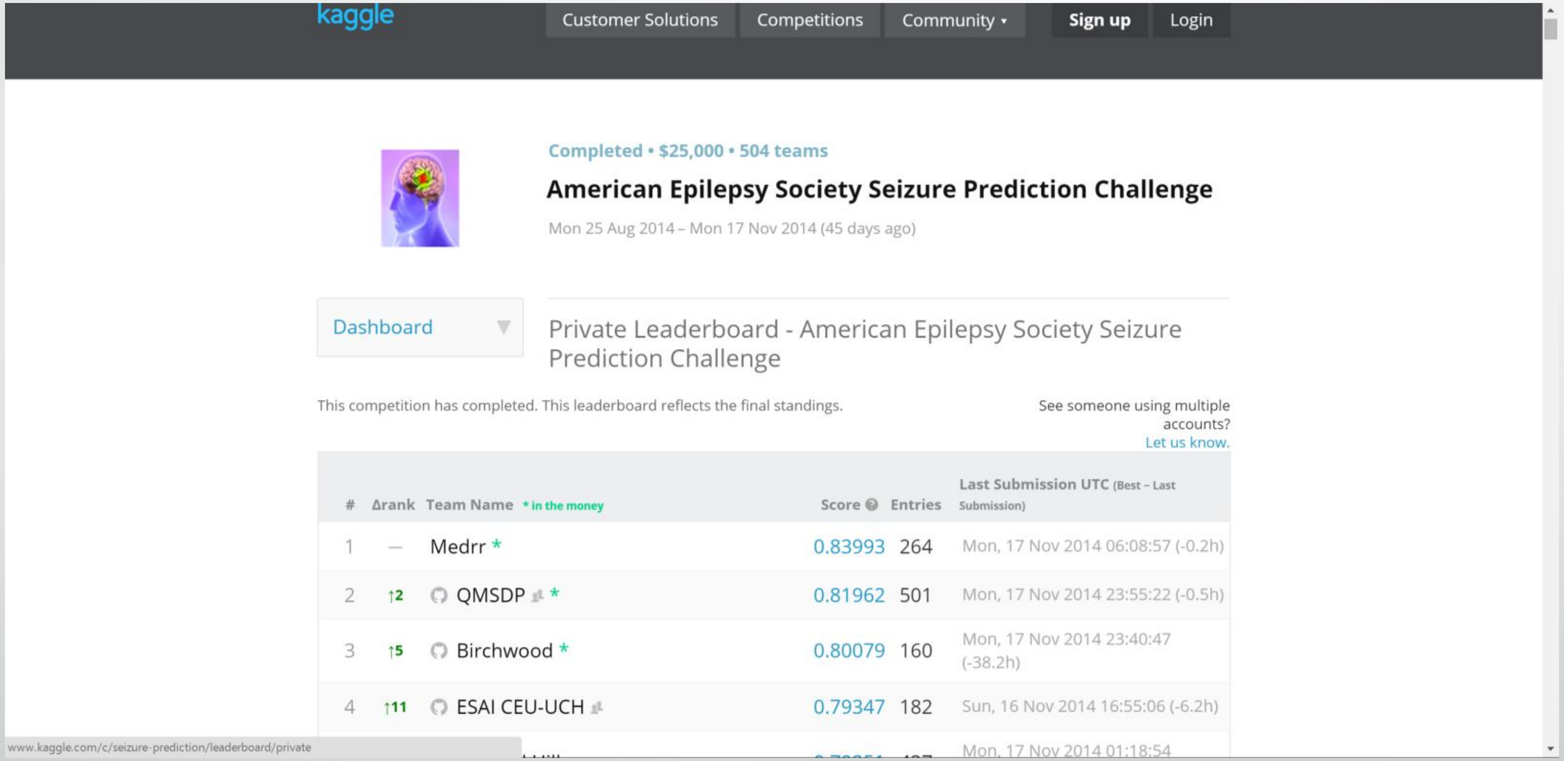
- **Modular components** for linking, query processing, query, and presentation
- Capabilities for recruiting users into groups, conducting A/B testing and surveys using different components

Meanwhile – many lessons learned on the way to this point!

# Highlights of Lessons Learned and Open Challenges
# *(See Paper for More)*

# Public Data Doesn't Lead to Users!

Simply offering data  is very different from engaging the community and changing the culture.  "If you build it, they won't necessarily come."




We need to sponsor challenges, show successes, and highlight benefits.

# “Passive Sharing” Is a Major Hurdle

**In the life sciences, many are required to make their data available.**

But in many sciences, data is **very costly to obtain**, thus there is perception of **risk** in sharing.

Tendency to make a **token effort to share**.  **Posting files** on an FTP site vs. ensuring the data is documented, includes provenance, and is usable by others!

**We need to offer rewards (and reduce the costs) to encourage sharing.**

# Open Research Challenge: Data Sharing Metrics & Incentives

How do we get past the practice of measuring impact by citation counts and h-indices?

Need a "Sharing-index" (S-index) for data, databases, and users:

- We can capture data usage in a **provenance graph** [Green+07]

  Adapt h-index, PageRank, ObjectRank?

- But data isn't atomic; how do we account for joins, aggregation, net impact?
  - Perhaps generalize from notions like **responsibility** (Meliou, Gatterbauer, Suciu)?

# Open Research Challenge: Privacy Preserving User Studies

There has been much progress in privacy-preserving computations, e.g., differential privacy

But how do we facilitate user studies in a way that:

- assures privacy (of user queries, workloads, data)
- yet *enables us to determine what techniques are most effective under what conditions?*

A key challenge: the algorithms we're testing may not be data-independent!

# Conclusions

Community-scale data integration will *only happen* if we have infrastructure that lets us evaluate, improve our techniques *in context* of real usage

- One “launching pad” in this effort, for neuroscience
- A *platform* for evaluating data integration techniques

Our journey has led to numerous lessons learned:

- Perceived risks and inertia
- Encouraging adoption
- Key research challenges:
  - data sharing metrics & incentives
  - privacy-preserving user experiments

More lessons in the paper – but hopefully more to come if we as a community can work together to get our techniques evaluated in the real world