ββMC
Biomedical and Behavioral Methodology CORE

OSCTR
Oklahoma Shared Clinical
& Translational Resources

A Clinical Data Warehouse Ecosystem to Support Research & QAQI

DAVID E. BARD, PHD[1]
WILLIAM H. BEASLEY, PHD[1]
ZSOLT NAGYKALDI, PHD[2]

[1]DEPARTMENT OF PEDIATRICS
[2]DEPARTMENT OF FAMILY AND PREVENTIVE MEDICINE
UNIVERSITY OF OKLAHOMA HEALTH SCIENCES CENTER

THE UNIVERSITY OF OKLAHOMA College of Medicine | Department of Pediatrics

THE UNIVERSITY OF OKLAHOMA College of Medicine | Department of Family and Preventive Medicine

Objectives

Define and describe clinical data warehouses

Provide examples of data warehouse uses in medicine and public health

Describe the connection between data warehousing and Big Data

Challenges of building a clinical data warehouse (legal, architectural, procedural)

Limitations of warehouse projects

Birth and development of an OUHSC CDW

Brief demonstration of current CDW design and process

Future directions for the OUHSC CDW

What is a Clinical Data Warehouse?

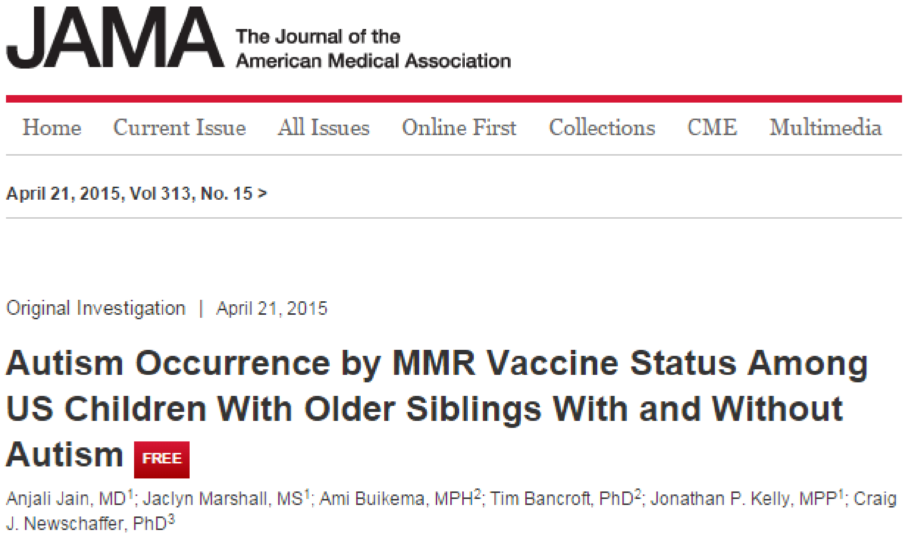
A data warehouse is a repository of historical data organized for reporting and analysis. It facilitates data access by having data from many sources in one place, linked together, and easily searchable.

Common CDW Characteristics

- A database containing data from multiple sources
- Data extracted from the databases of clinical software packages
- A database containing data related to each other with some unique identifier
- A database that is only as good as the data entered

Slide courtesy of BUMC Rozen & Saitz
http://www.bumc.bu.edu/crro/files/2010/01/Rosen-4-11-07.pdf

Examples of CDW Uses

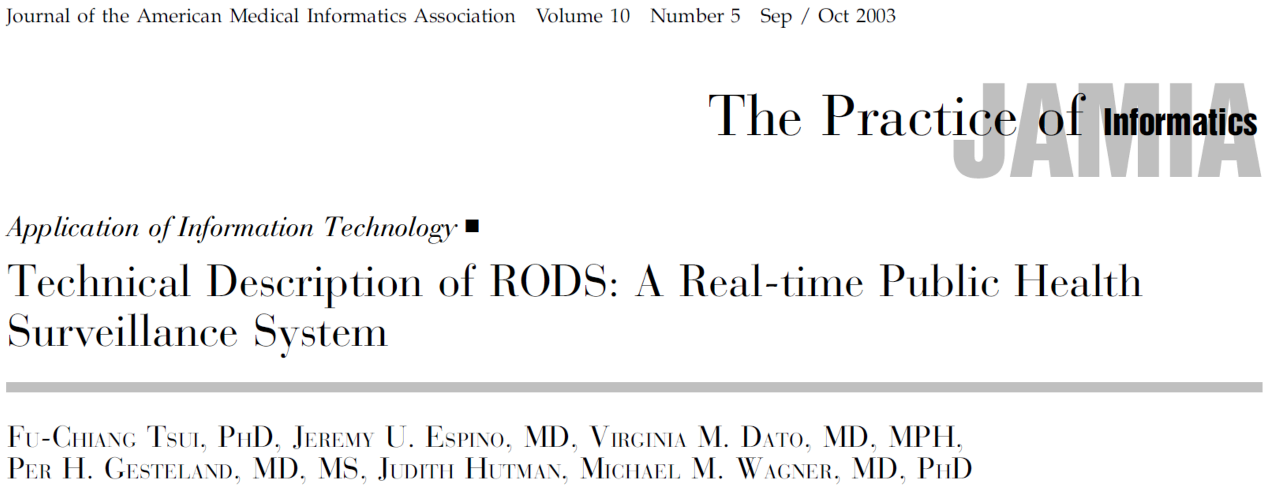


April 21, 2015, Vol 313, No. 15 >

Original Investigation | April 21, 2015

Autism Occurrence by MMR Vaccine Status Among US Children With Older Siblings With and Without Autism

Anjali Jain, MD[1]; Jaclyn Marshall, MS[1]; Ami Buikema, MPH[2]; Tim Bancroft, PhD[2]; Jonathan P. Kelly, MPP[1]; Craig J. Newschaffer, PhD[3]



The Practice of Informatics
JAMIA

Journal of the American Medical Informatics Association Volume 10 Number 5 Sep / Oct 2003

Application of Information Technology ■

Technical Description of RODS: A Real-time Public Health Surveillance System

Fu-Chiang Tsui, PhD, Jeremy U. Espino, MD, Virginia M. Dato, MD, MPH, Per H. Gesteland, MD, MS, Judith Hutman, Michael M. Wagner, MD, PhD

Recent publications using RODS

Ye, Y., Tsui, F. R., Wagner, M., Espino, J. U. and Li, Q. (2014), Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers, J Am Med Inform Assoc.

Wagner, M. M., Levander, J. D., Brown, S., Hogan, W. R., Millett, N. and Hanna, J. (2013), Apollo: giving application developers a single point of access to public health models using structured vocabularies and Web services, AMIA Annu Symp Proc, 2013: 1415--1424.

Liu, T. Y., Sanders, J. L., Tsui, F. C., Espino, J. U., Dato, V. M. and Suyama, J. (2013), Association of over-the-counter pharmaceutical sales with influenza-like-illnesses to patient volume in an urgent care setting, PLoS ONE, 8, 3: e59273.

Lee, B. Y., Tai, J. H., McGlone, S. M., Bailey, R. R., Wateska, A. R., Zimmer, S. M., Zimmerman, R. K. and Wagner, M. M. (2012), The potential economic value of a 'universal' (multi-year) influenza vaccine, Influenza Other Respi Viruses, 6, 3: 167--175.

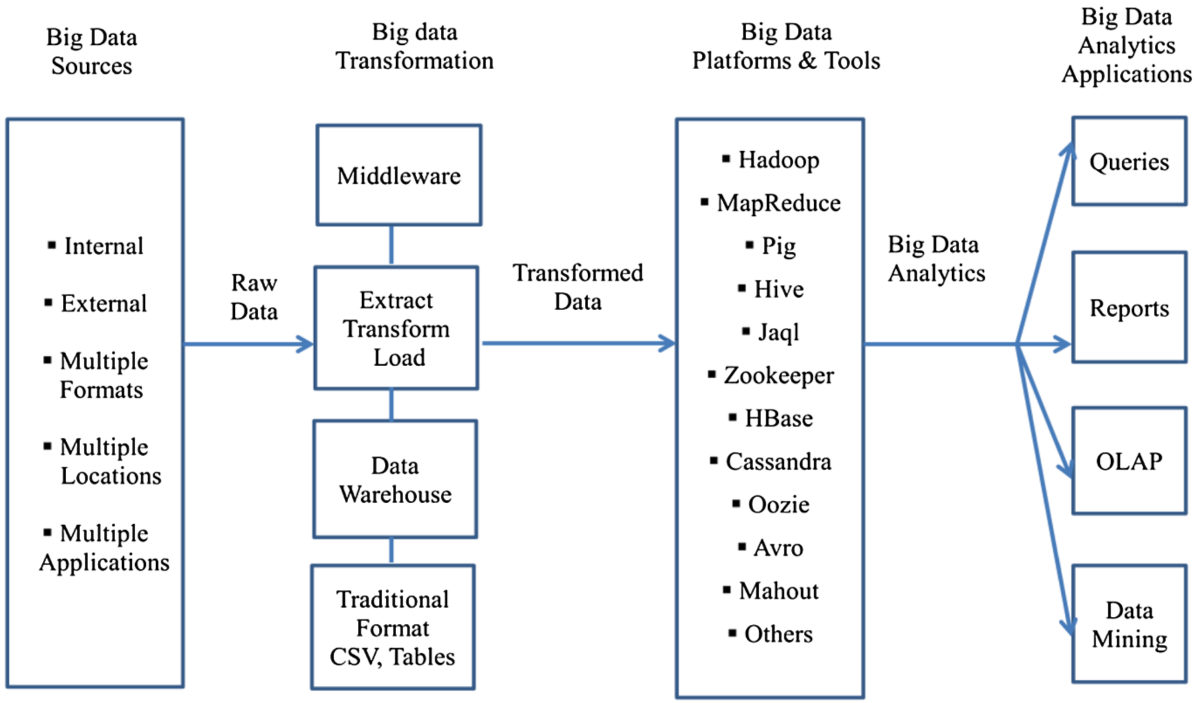
Wagner, M., Cooper, G., Tsui, Fuchiang, Espino, J.U., Harkema, H., Levander, J., Villamarin, R., Millett, N., Brown, S.T. and Gallaggher, A. (2012), A Decision-Theoretic Model of Disease Surveillance and Control and a Prototype Implementation for the Disease Influenza, Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on: 49-54.

Lee, B. Y., Stalter, R. M., Bacon, K. M., Tai, J. H., Bailey, R. R., Zimmer, S. M. and Wagner, M. M. (2011), Cost-effectiveness of adjuvanted versus nonadjuvanted influenza vaccine in adult hemodialysis patients, Am. J. Kidney Dis., 57, 5: 724--732.

Marrying CDW and Big Data

One step in the process of whipping “Big Data” into shape for analysis

- Aggregates data from various sources
- Data is usually not real-time (but not too far behind)
- Standardization, cleansing, and relational restructuring are common tasks that immediately precede CDW storage



Raghupathi and Raghupathi *Health Information Science and Systems* 2014 **2**:3

Big Data Analytics

What is Big Data?

- Extremely large data sets (usually containing raw data) that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Why is it important?

- Improve care, save lives, and lower costs (some estimates reach $300 Billion/year in U.S.)
- The promise of predictive analytics
  - Predict cost-effective best practices, disease surveillance, etc.
- The promise of prescriptive analytics
  - Adaptive interventions that redirect patient health trajectories that drift off-course
  - Enables personalized medicine via advanced clinical decision support

What is the future of Big Data?

- As data capture technologies (e.g,. fitness monitoring devices, the eventual Internet of Things) continue to evolve, we be accumulating more and more Big Data
- Healthcare isn’t really at the Big Data stage, yet, but with the advent of medical devices and cloud data capture technologies, we are certainly on the doorstep

Examples of Big Data Analytics Use in Healthcare

Columbia University Medical Center’s “complex correlations” project

- Using physiological data from patients with brain injuries, provides information to aggressively treat complications
- Reputedly diagnoses serious complications 48 hours earlier in patients suffering a bleeding stroke from a ruptured brain aneurysm

California-based Kaiser Permanente discovery of adverse drug effects and subsequent withdrawal of Vioxx

PCOR Context

The potential of fully functional EHRs that are aggregated in a CDW- all members of the clinical team have ready access to the ~*latest* information allowing for more **coordinated, patient-centered care**.

- The information gathered by the primary care provider tells the emergency department clinician about the patient’s life threatening allergy, so that care can be adjusted appropriately, even if the patient is unconscious.
- A patient can log on to his own record and see the trend of the lab results over the last year, which can help motivate him to take his medications and keep up with the lifestyle changes that have improved the numbers.
- The lab results run last week are already in the record to tell the specialist what she needs to know without running duplicate tests.
- The clinician’s notes from the patient’s hospital stay can help inform the discharge instructions and follow-up care and enable the patient to move from one care setting to another more smoothly.

Challenges of building a CDW

Legal

- Data protection (consenting policies)
- Data security (HIPAA, FERPA, FISMA regs)
- Ethical dilemmas (revealing previously unknown Dx)

Architectural

- Interoperability of existing systems
- Record linkage (Master patient index)
- Record anonymization, pseudonymization
- Standardization
- Data security
- Updating, version control, adjudication decisions

Challenges of building a CDW

Procedural

- Governance
- Maintenance of institutional support
- Team of custodians
- Training and education
- Regulatory compliant uses of the data

Limitations of CDW Research

All the usual quasi-experimental (lack of randomization) threats to validity (ala Cook & Campbell, 1979)

- Selection bias (understand the sample and population from which data were drawn; differs across site, clinic, database)
- Missing data bias (patient “drop-out” issues)

Data entry errors (garbage in, garbage out)

Vagaries of standardization decisions and measurement bias

- Non-commensurate measures (How was the source data collected and coded)
- Crudeness of measurement

Typical limitations of archival analysis

- Availability of information
- Changes in measurement procedures and instruments over time, places, & persons

Statistical anomalies with large data

- Statistical vs Clinical significance
- Data fishing influences on Type I error

Strengths of CDW Research

Share many of the advantages of Integrative Data Analysis (if you're careful; see Curran & Hussong, *Psychological Methods*, *14*, 2009; Hofer & Piccinin, *Psychological Methods*, 14, 2009)

- Larger sample sizes and increased statistical power
- Increased sample heterogeneity
- Increased frequency of low bas-rate outcomes
- Breadth of measurement
- Longitudinal hypotheses

CDW team establishes procedures for data sharing and data use so you can concentrate on the hypotheses and analysis

Cost effective research (even for projects collecting new data but requiring EMR information for recruitment)

The Birth of an OUHSC CDW

Dr. Zsolt Nagykaldi has been advocating for years

- NagyKaldi & Shay initial HSC CDW sketch

Conversation changed when EHR came asking for assistance

Important for OUHSC's ability to

- compete for grants
- care for patients

Present needs assessed: Warehouses for Research and QAQI

- identification and recruitment of potential research participants
- epidemiologic evaluation of clinical information
- surveillance of disease
- quality of care assessments
- cost-effectiveness of prevention and treatment strategies

Progress to Date

OSCTR funding secured for initial development and deployment

Initial design complete

Computing infrastructure up and running

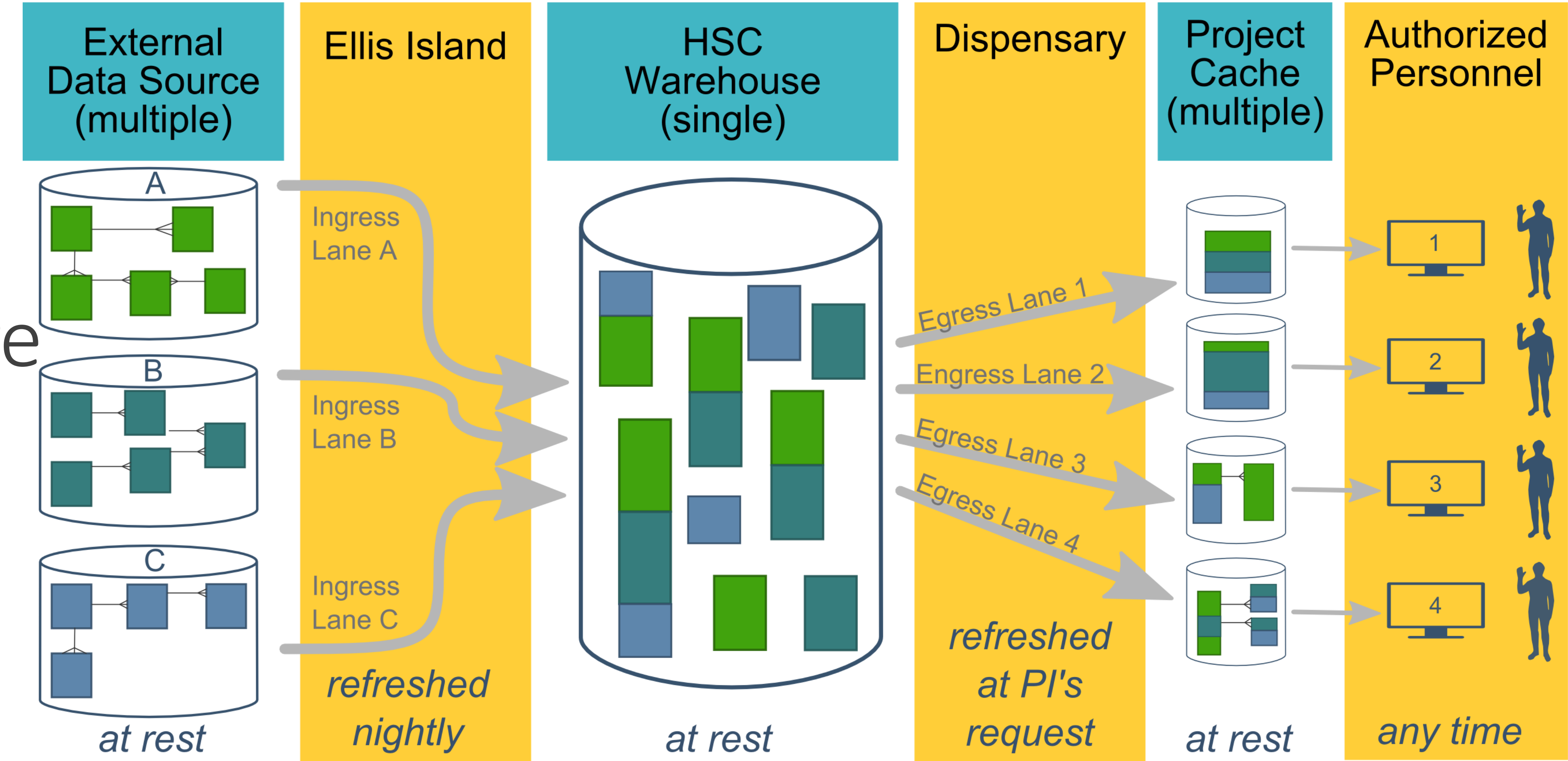
Product security review is complete

Product support plan is complete

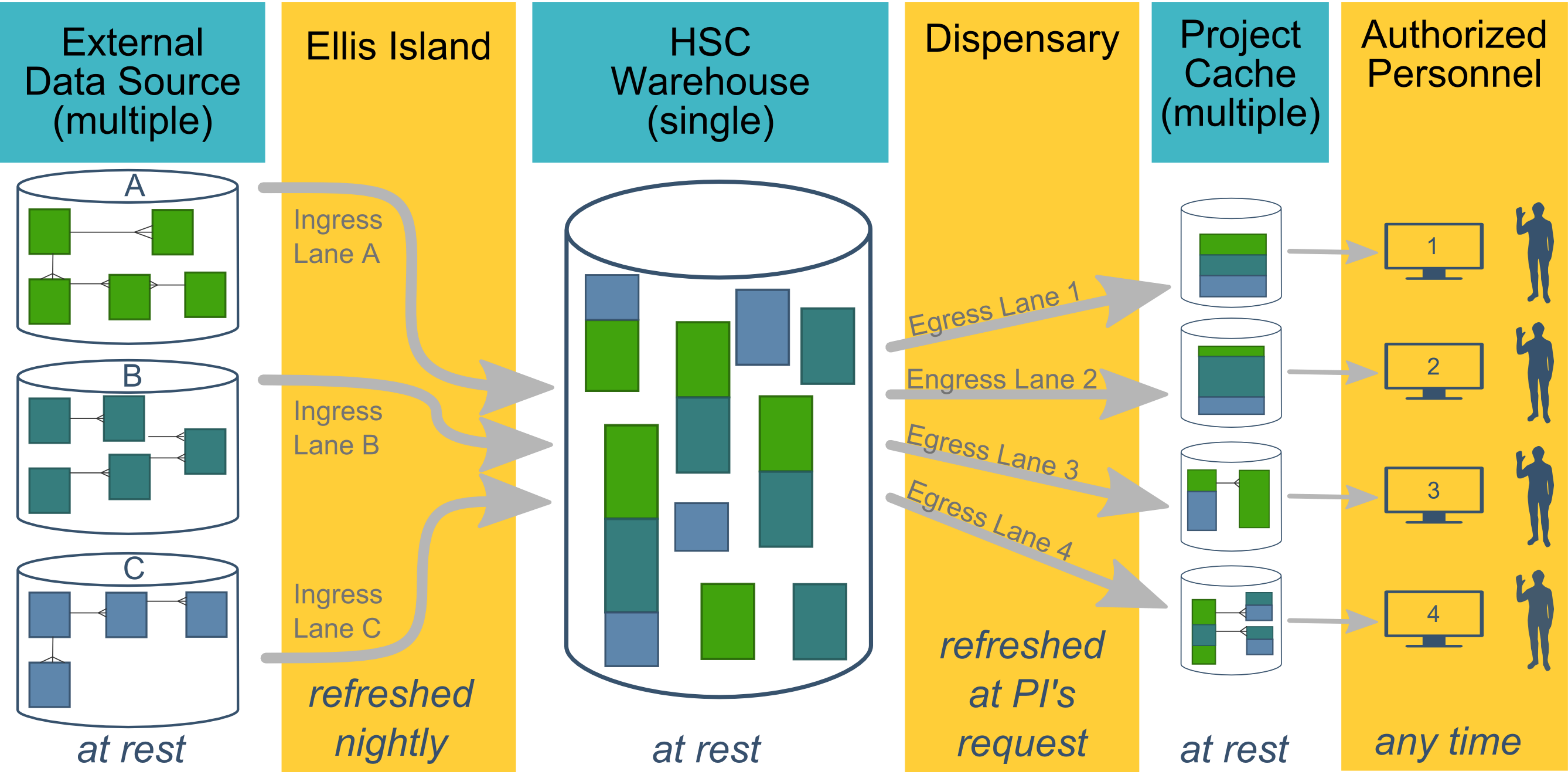
CDW development team is assembled

- 3 data science faculty members
- Campus IT support personnel identified
- New dedicated DBA hired
- New dedicated system navigator hired

Ecosystem Architecture



- **Data Source** (column 1): contains unique info
- **Warehouse** (column 3): contains copy after manipulation
- **Project Cache** (column 5): contains copy of copy after a lot of manipulation



External Data Source (multiple)

A

B

C

at rest

Ellis Island

Ingress Lane A

Ingress Lane B

Ingress Lane C

refreshed nightly

HSC Warehouse (single)

at rest

Dispensary

Egress Lane 1

Engress Lane 2

Egress Lane 3

Egress Lane 4

refreshed at PI's request

Project Cache (multiple)

at rest

Authorized Personnel

1

2

3

4

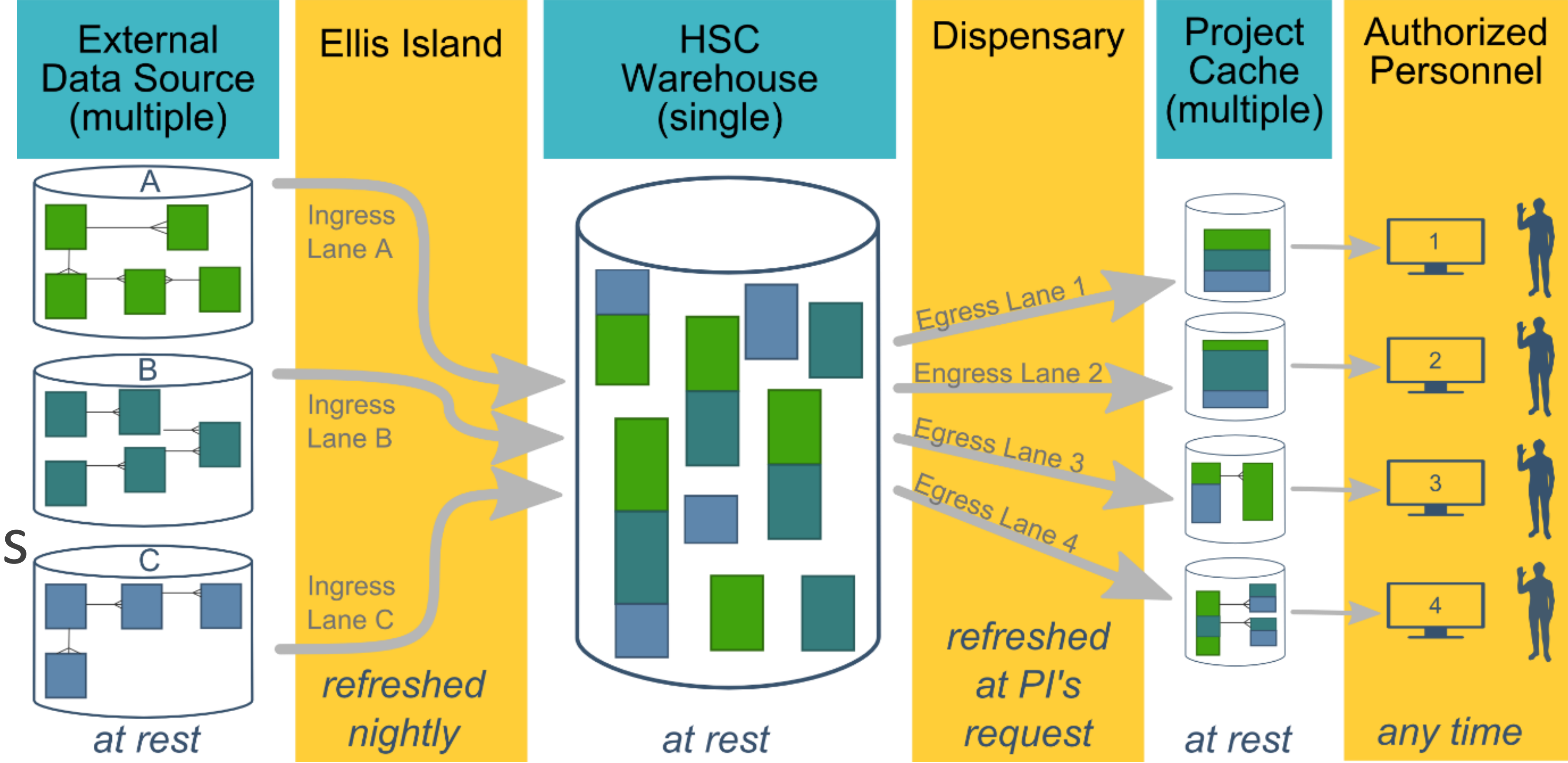
any time

Benefits over ad-hoc EMR Extracts

In contrast to the current approach of using EMR data in research & QA, this new system will provide:

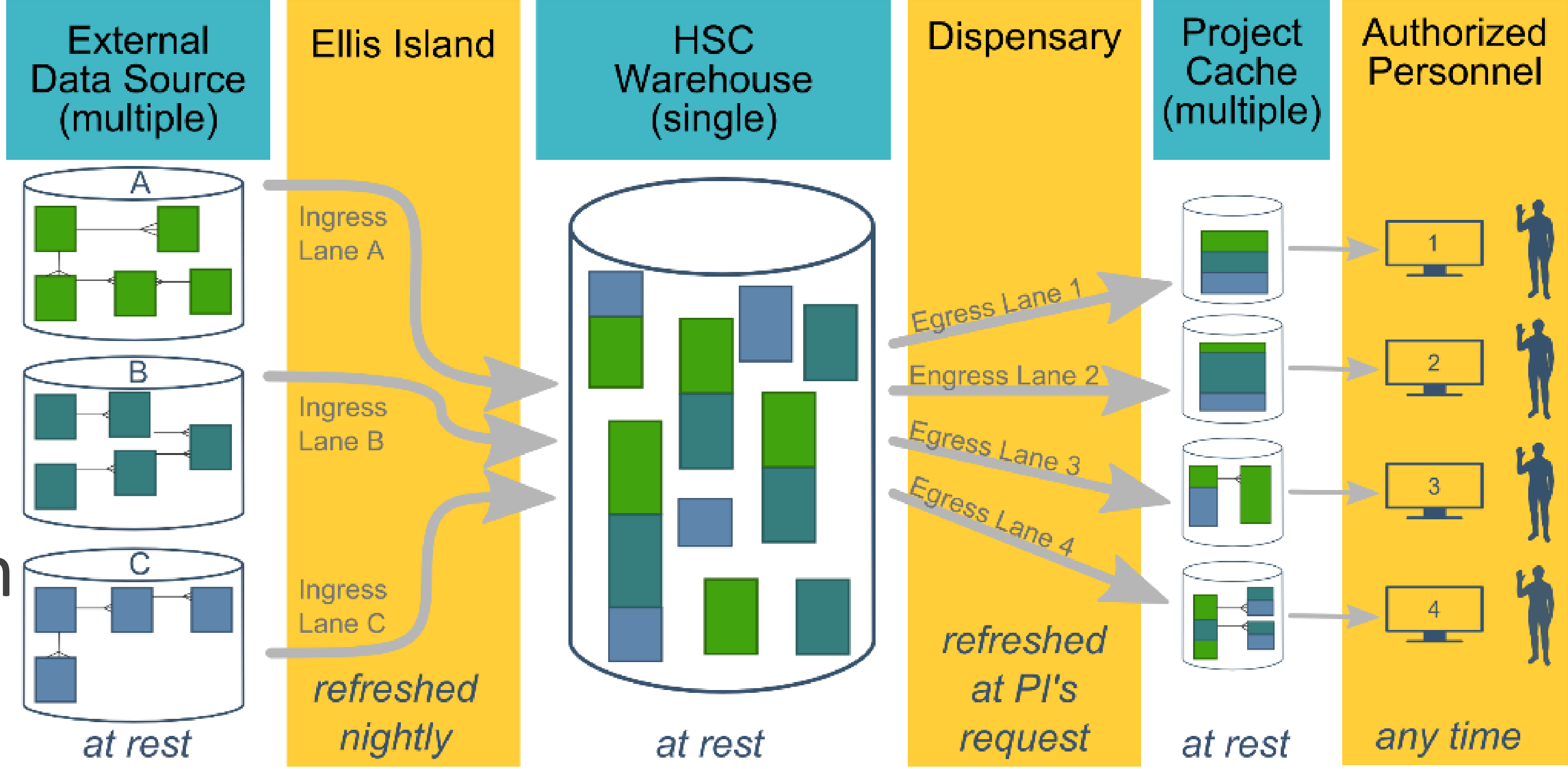
- Coordinated datasets across sources (combining patients and variables)
- Automated updates to datasets (important for on-going QA)
- More streamlined IRB approval (hopefully)
- More secure delivery through REDCap (instead of loose Excel files)
- More consistent experience for multiple research/QA investigations

Overall Goals



- Existing investigations can proceed quicker and more frequently.
- The process is more user-friendly and consistent across different investigations & EMRs.
- More complicated investigation are possible (by combining patients & measurements across data sources).

Instantiation Activities



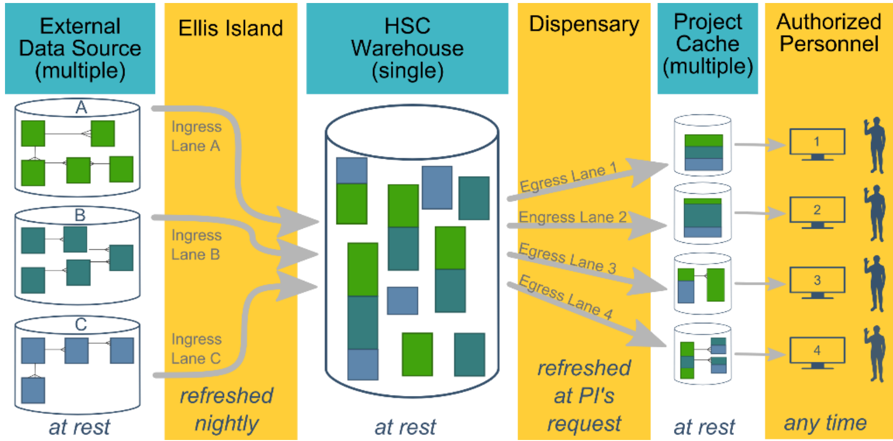
Creating and maintaining the warehouse ecosystem involves three different levels:

- **Ecosystem** (eg, configuring the database VM)
- **Data source** (eg, connecting an EMR's feed to the warehouse)
- **Project** (eg, processing data from specific patients for a single IRB-approved investigation)

Requirements Per Ecosystem

These steps are required for the overall ecosystem. This work is required once for our **campus**, and doesn't need to be replicated for each additional data sources or investigation.

1. Planning meetings with Campus IT and shared services
2. Campus IT security review (and annual re-reviews)
3. Creating multiple VMs (virtual machines)
4. Installing and patching software on each machine

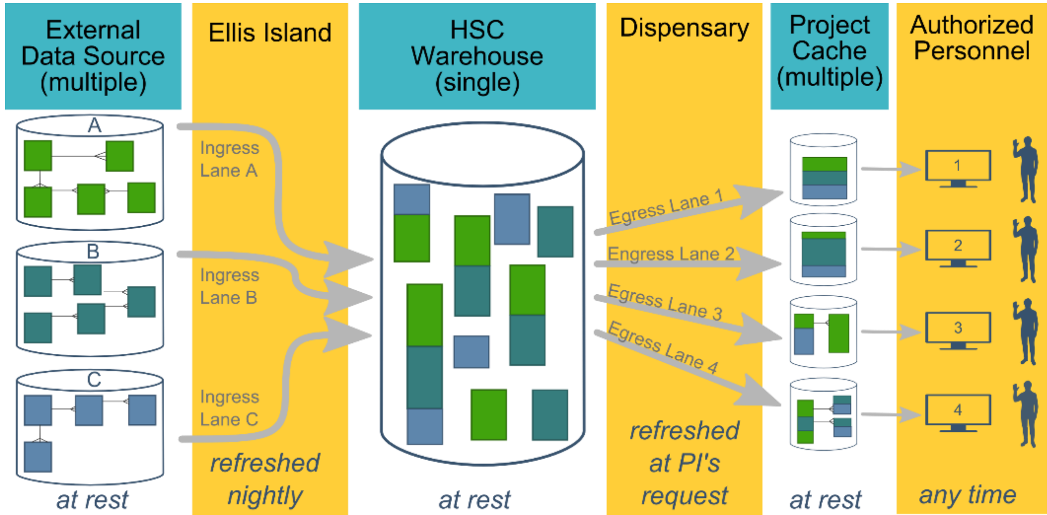


5. Establish and maintaining secure connections between the 2nd-5th cols.
6. Assembling team (DBA, EMR reporter, REDCap admin, statistician)

Requirements Per Data Source

Necessary for each data source (eg, once for Centricity, once for Meditech…)

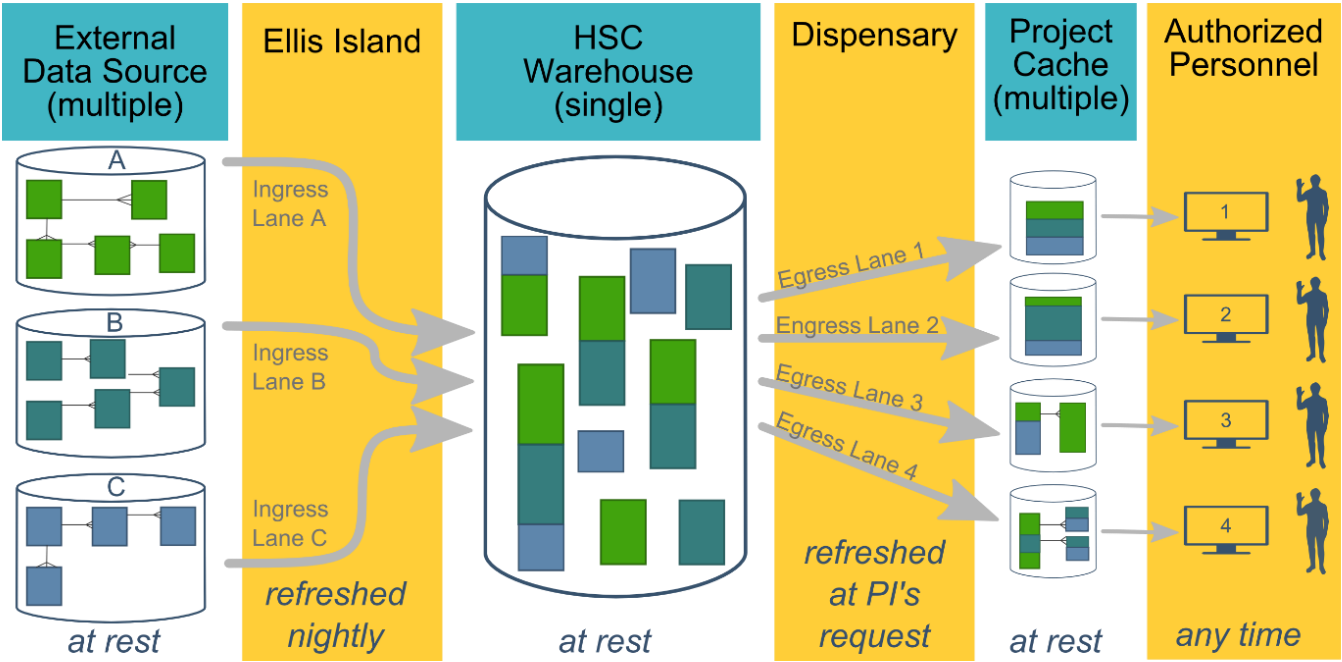
1. Agreement with data source owner.
2. Establish a periodic automated data feed.
3. Secure licenses for GUI and specialized software (eg, Crystal Reports & Centricity front-end).
4. Learn data source.
   - Database schema
   - Sampling plan (and source of missingness)
5. Write code to:
   - Transform EAV schema into warehouse-friendly schema.
   - Combine measures from different data sources
   - Link clients from different data sources



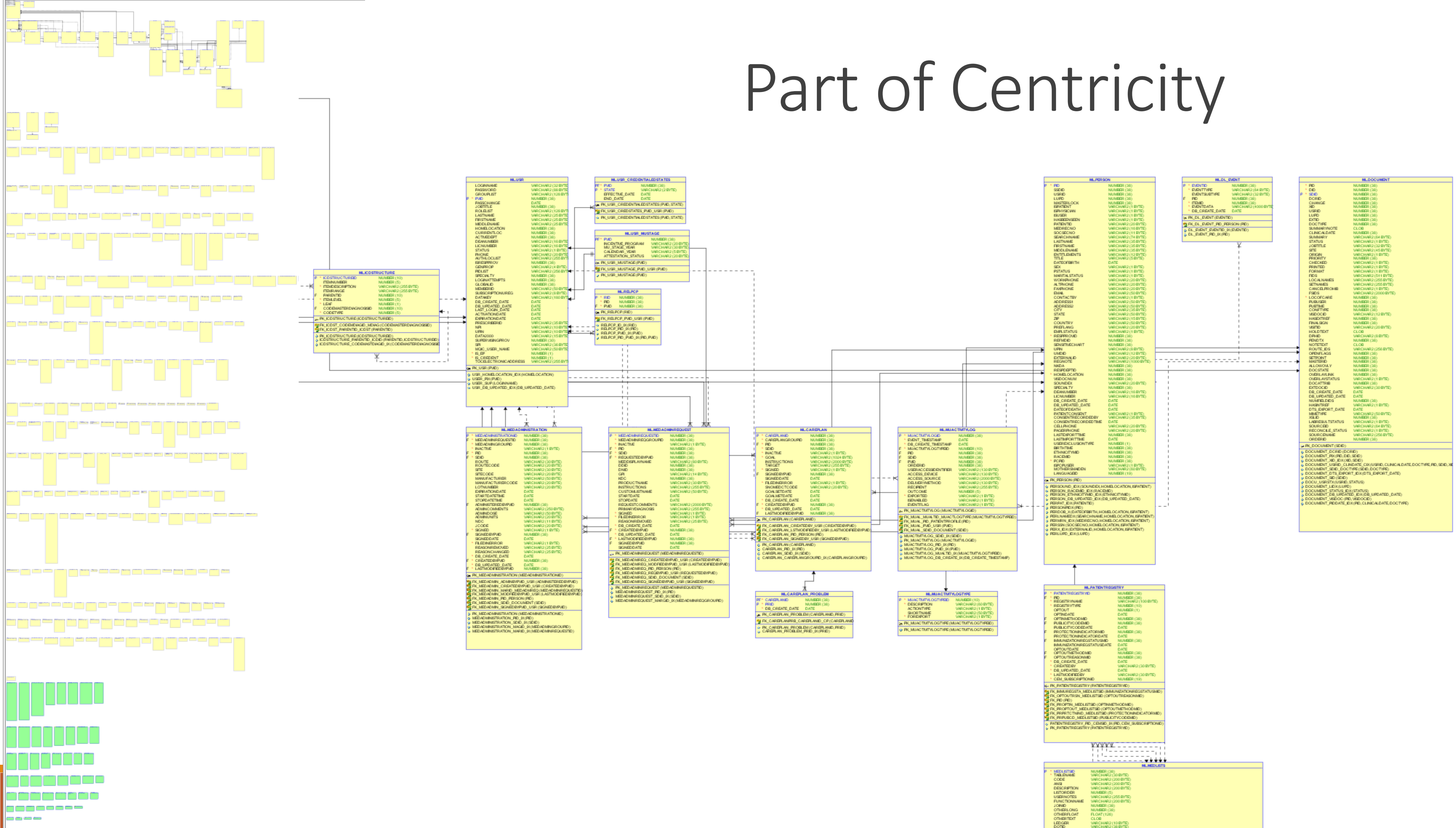
Requirements Per Project/Investigation

Necessary for each investigation (eg, once for Gillaspy/Weedn, once for Miller…)

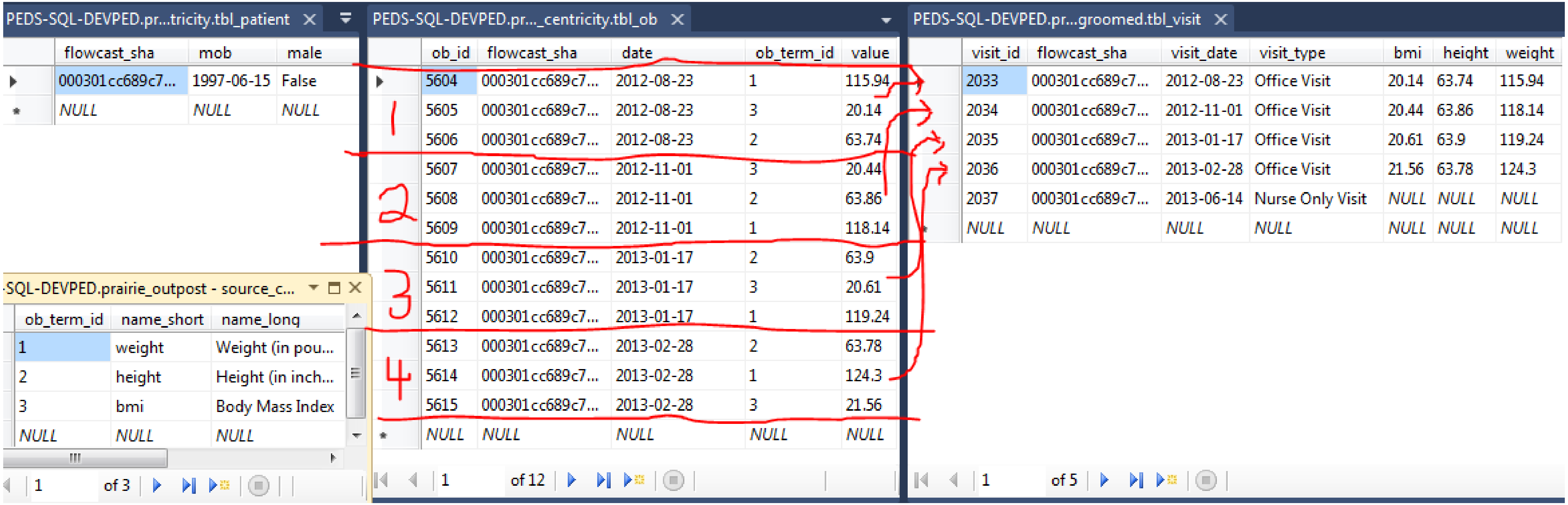
1. IRB approval for investigation
2. IRB approval for warehouse collaborators
3. Teach us research goals & specifics
4. Write code to transform warehouse into research-friendly schema.
5. Establish REDCap cache
6. Write code to transfer data from warehouse to cache
7. Educate researchers how REDCap & cache work



Part of Centricity



Transforming (one patient's collection of) records for statistical software



PEDS-SQL-DEVPED.pr...tricity.tbl_patient

flowcast_sha	mob	male
000301cc689c7...	1997-06-15	False
NULL	*NULL*	*NULL*

-SQL-DEVPED.prairie_outpost - source_c...

ob_term_id	name_short	name_long
1	weight	Weight (in pou...
2	height	Height (in inch...
3	bmi	Body Mass Index
NULL	*NULL*	*NULL*

1 of 3

PEDS-SQL-DEVPED.pr..._centricity.tbl_ob

ob_id	flowcast_sha	date	ob_term_id	value
5604	000301cc689c7...	2012-08-23	1	115.94
5605	000301cc689c7...	2012-08-23	3	20.14
5606	000301cc689c7...	2012-08-23	2	63.74
5607	000301cc689c7...	2012-11-01	3	20.44
5608	000301cc689c7...	2012-11-01	2	63.86
5609	000301cc689c7...	2012-11-01	1	118.14
5610	000301cc689c7...	2013-01-17	2	63.9
5611	000301cc689c7...	2013-01-17	3	20.61
5612	000301cc689c7...	2013-01-17	1	119.24
5613	000301cc689c7...	2013-02-28	2	63.78
5614	000301cc689c7...	2013-02-28	1	124.3
5615	000301cc689c7...	2013-02-28	3	21.56
NULL	*NULL*	*NULL*	*NULL*	*NULL*

1 of 12

PEDS-SQL-DEVPED.pr...groomed.tbl_visit

visit_id	flowcast_sha	visit_date	visit_type	bmi	height	weight
2033	000301cc689c7...	2012-08-23	Office Visit	20.14	63.74	115.94
2034	000301cc689c7...	2012-11-01	Office Visit	20.44	63.86	118.14
2035	000301cc689c7...	2013-01-17	Office Visit	20.61	63.9	119.24
2036	000301cc689c7...	2013-02-28	Office Visit	21.56	63.78	124.3
2037	000301cc689c7...	2013-06-14	Nurse Only Visit	*NULL*	*NULL*	*NULL*
NULL	*NULL*	*NULL*	*NULL*	*NULL*	*NULL*	*NULL*

1 of 5

From Centricity
(one row per patient/measurement)

To REDCap Cache
(one row per visit)

Live demo of REDCap Project Cache

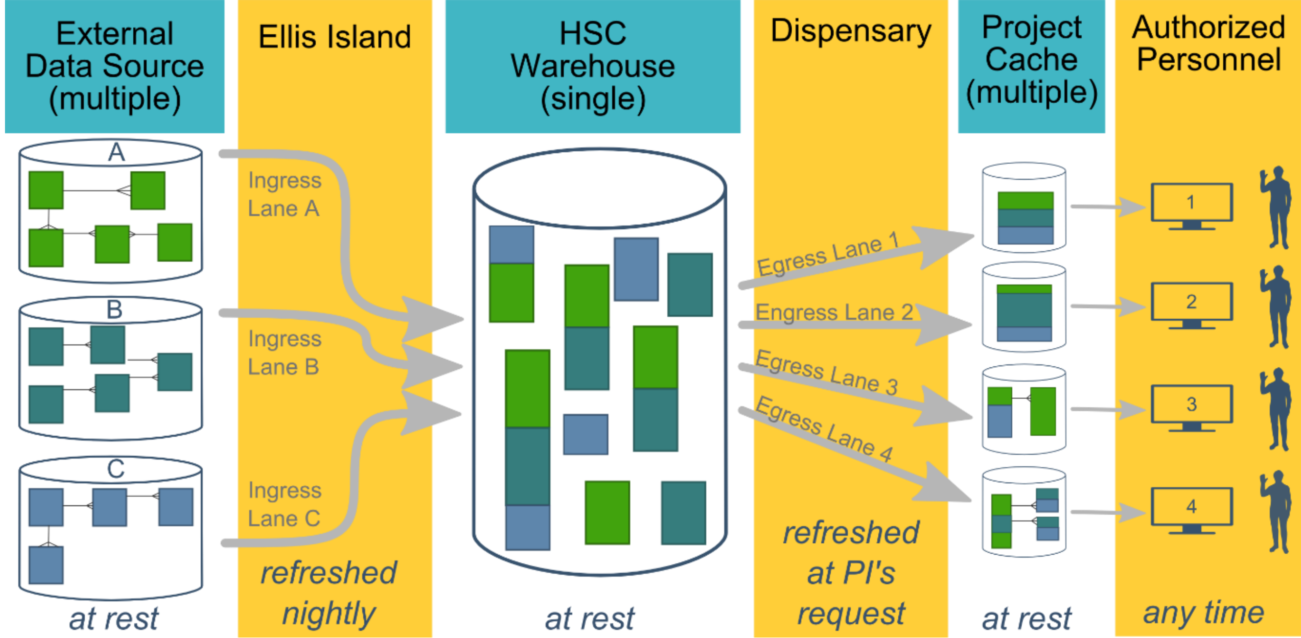
- REDCap GUI
- Access data with R/SAS

SAS Example

```
filename in  "E:\encrypted_drive\aslan\apicall.csv";
filename out "E:\encrypted_drive\aslan\redcap.csv";

%let token = xxxxx;
"%NRStr(content=record&type=flat&format=csv&fields=last_first_mi&token=)&token%NRStr(&_x=1)";

%let url    = "https://bbmc.ouhsc.edu/api/";

data _null_ ;
file in ;
put "%NRStr(content=record&type=flat&format=csv&token=)&token";
run;

proc http
   in=in
   out=out
   url=&url
   method="post";
run;
```

Adapted from Linda Carlin’s and Kevan Essmyer’s posts at https://groups.google.com/forum/#!topic/project-redcap/MS_4_l1q-WU

Open to any Campus Collaborators



With the appropriate IRB approval, this system is open to anyone on campus, including:

- Statisticians analyzing the project caches (eg, BERD & BBMC)
- Applied researchers or QA investigators (last column)
- Departments adding their own data source to the ecosystem (first column)
- Departments who want a separate warehouse & data sources, but want to reuse the distribution mechanism.

Possible Service Tiers

Tier 0:	Education	(free & composed of fake data)
Tier 1:	Dept Cookie-cutter	(free)
Tier 2:	3D box	(hourly fee)
Tier 3:	Multilevel	(retainer + hourly fee)

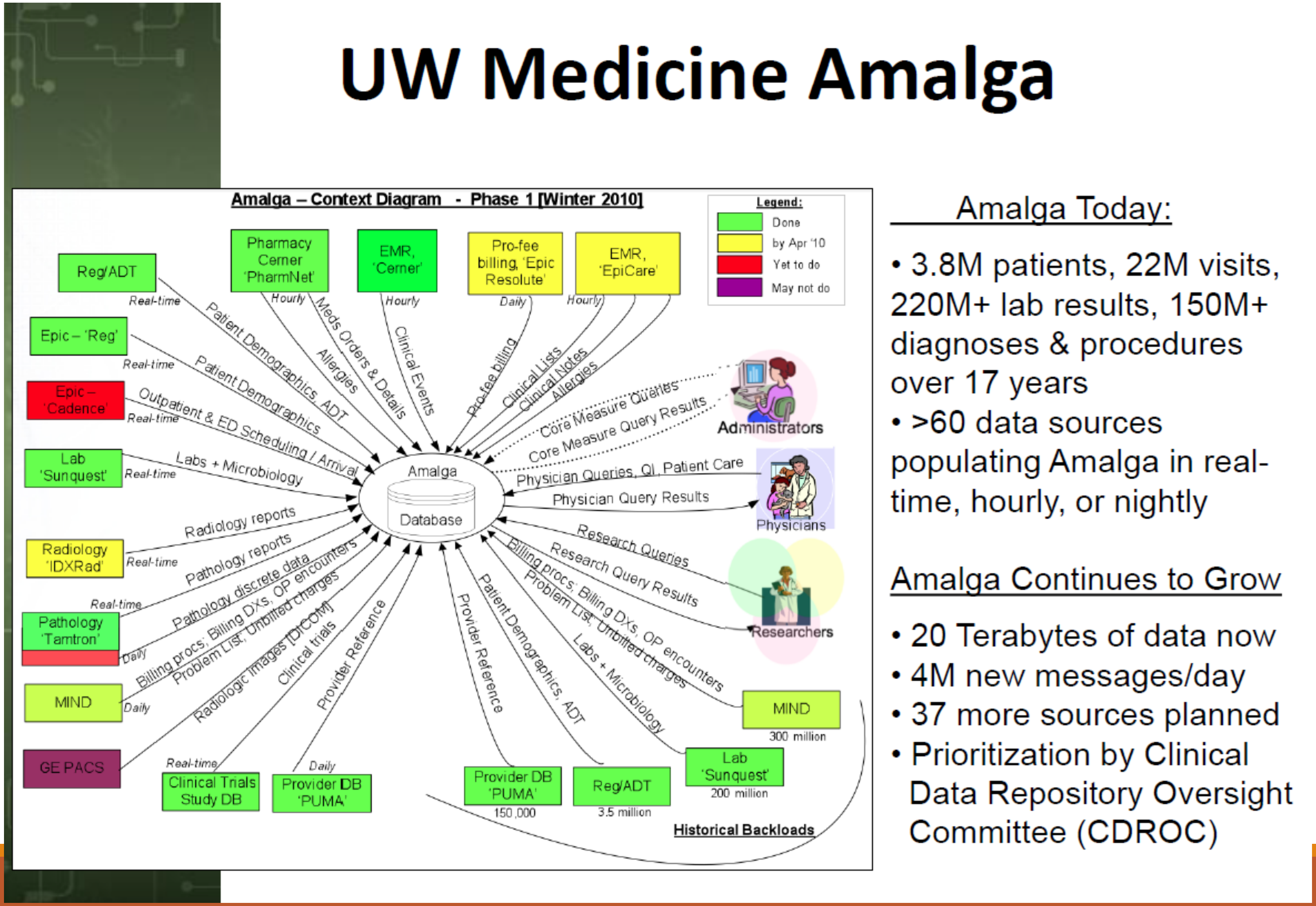
Misc

- Design benefits
  - For the subspecialties, a wider collection/net starts to escape limitation of case-control-like designs
- Advantages of starting small, compared to MyHealthNet
  - Pitfalls of relying on standardization that exists in those biggies
  - NOT missing context of a standardized tag
- Not a black box. We'll share the code that took the data from datasource to project cache

Future

- Scalability
  - More projects
  - More data per project
  - More frequent updates per project
- Federal reporting benefits of combining data sources
- i2b2 possible (https://www.i2b2.org/)
  - Learn from/with other institutions
  - Cheaper code development & maintenance.

Long-term Vision



UW Medicine Amalga

Amalga – Context Diagram - Phase 1 [Winter 2010]

Amalga Today:

- 3.8M patients, 22M visits, 220M+ lab results, 150M+ diagnoses & procedures over 17 years
- >60 data sources populating Amalga in real-time, hourly, or nightly

Amalga Continues to Grow

- 20 Terabytes of data now
- 4M new messages/day
- 37 more sources planned
- Prioritization by Clinical Data Repository Oversight Committee (CDROC)

Immediate Next Steps

Finalize roster and convene the CDW governance body

- Develop standard operating procedures for CDW access and use
- Determine regulatory review processes for CDW projects

Obtain data sharing agreements with EHR systems

Build system components to accommodate a few more select pilot projects

Continually collect and monitor feedback from end-users and adapt system accordingly

Invite and select new CDW projects

OUHSC CDW Data Possibilities

Currently available

- Centricity data from OKC-HSC clinics
  - Patient demographics
  - Visit information
  - Biometrics
  - Diagnoses
  - Lab results
  - Prescription information
  - Limited billing information

Planned additions in near future (negotiations pending)

- Meditech data
- CribNotes data (neonatal/perinatal specific EMR)
- Select OSIIS data from OK State Dept of Health
- HSC Pharmaceutical data
- IDX billing information

Not the Only CDW in Town/State

Coming Soon: OSDH Health-e-Oklahoma system

- Orion Health vendor

OPSR Data Systems & Coordination workgroup

- http://smartstartok.org/opsr

OU Centricity maintains a business operations warehouse

MediTech participates in pooled warehouse

Health Information Exchanges (HIEs)

- MyHealth Access Network
- Coordinated Care Oklahoma
- OSU Center for Health Systems Innovation pays for Cerner HIE

ββMC

THE UNIVERSITY OF OKLAHOMA College of Medicine | Department of Family and Preventive Medicine

OSCTR Oklahoma Shared Clinical & Translational Resources

Current Collaborators

Design and Management Team

- Will Beasley (BBMC)
- David Bard (BBMC)
- Zsolt Nagykaldi (HSC FPM)
- Sreeharsha Mandem (BBMC & OU CS)
- Sabrina Antry (BBMC)

Centricity Support

- Kevin Elledge
- Cynthia Proctor
- Adam McGann
- June Pearson

Institutional Support & Guidance

- Judith James (OSCTR)
- Tim VanWagoner (OSCTR)
- Joel Guthridge (OSCTR)
- Robert Roswell (COM)
- Darrin Akins (COM)

OU IT

- Mark Ferguson
- Scott DeWitt
- Cliff Mack
- Tony Miller
- Randy Moore
- April Lee