Planning for informatics in your grant applications

Samuel Volchenboum

Julie Johnson

November 14, 2017

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

CENTER FOR RESEARCH INFORMATICS

Center for Research Informatics

Applications - Systems - Bioinformatics - Data warehousing - Clinical trials

At the end of this talk, you will…

- Know what parts of a grant need informatics consideration
- Understand how important it is to seek help early
- Feel comfortable reaching out to CRI and asking for help

CRI vs. ??

- CBIS
- Research Computing Center (RCC)
- Computation Institute (CI)
- ITS
- CDIS
- Biostatistics core
- Center for Health Delivery Sciences and Innovation

The idealized process…

- Have an idea
- Get preliminary data
- Write a proposal
- Get funding
- Do work
- Repeat

What often happens…

- Have an idea or an extension of current work
- Apply for grant using old preliminary data
- Get award for new work
- Figure out how to actually do (and pay for) the work

What often happens…

- Have an idea or an extension of current work
- Apply for grant using old preliminary data
- Get award for new work
- Figure out how to actually do (and pay for) the work

Scenario #1 - The sequencer

- Researcher gets a pilot grant to study colon cancer patients using ChIP-Seq
- Pilot grant only for cost of sequencing
- No provisions made for **analysis** and **interpretation**

Scenario #2 - The multi-center trial

- Researcher gets U grant for testing a new survey tool at 30 cooperative sites
- Grant has no provisions for any research informatics **support**

Scenario #3 - The Big Data™ user

- Researcher gets funding to sequence 1000 whole genomes
- Gets funding for sequencing but then needs 20TB of storage space
- No grant provisions for **storage** or **backup**

Scenario #4 - The simulator

- Researcher gets funding to design, perform, and test molecular simulations on millions of drug-target combinations
- Requires millions of hours of **HPC** usage
- No funding for HPC

Scenario #5 - The analyzer

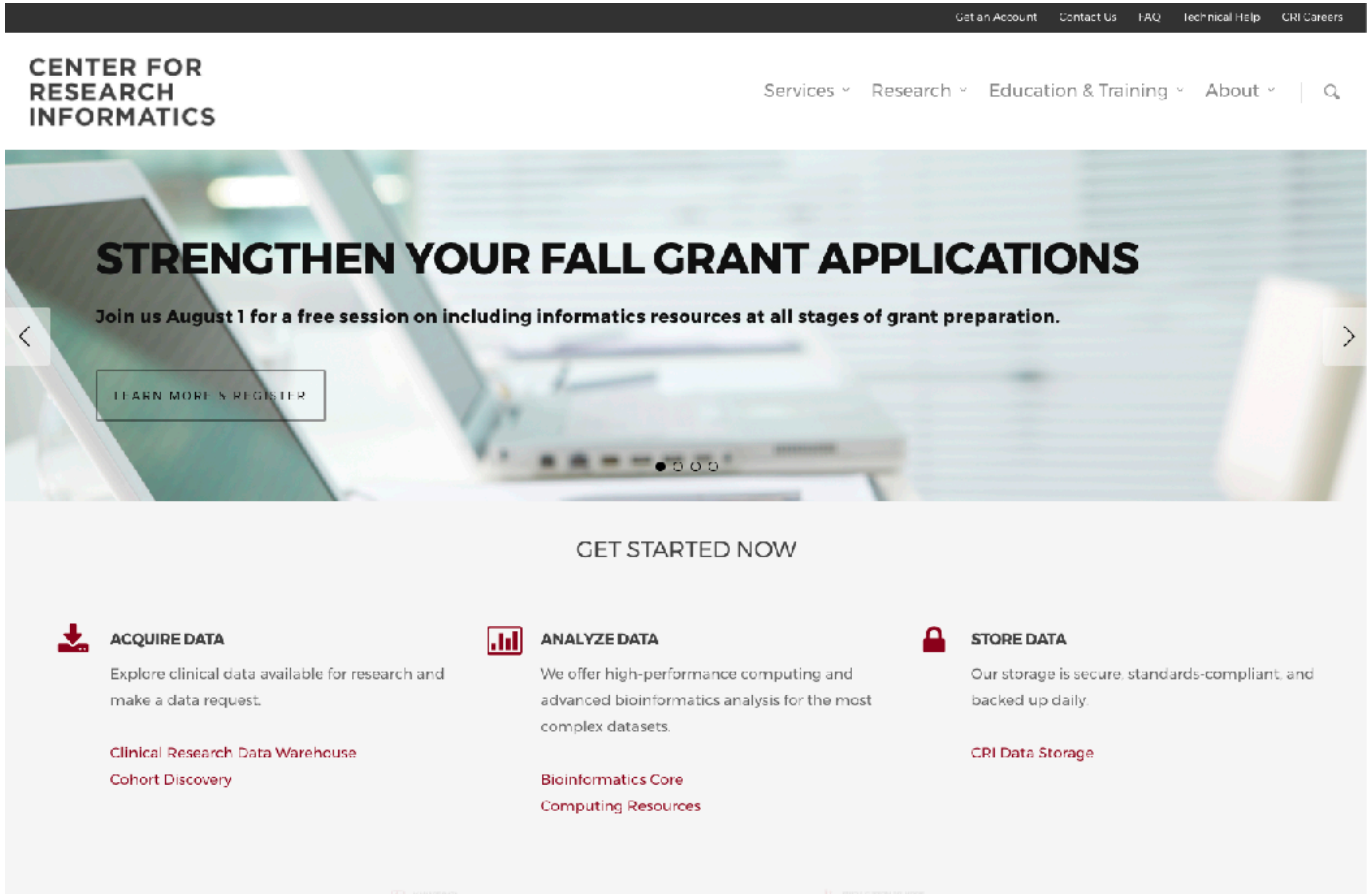
- Funding secured for pulling a large comprehensive data set from the data warehouse to perform disease modeling
- Data is pulled and given to research team but there is no one to **analyze** the data

There are many opportunities to consider informatics resources

The best time is when you're just thinking about a project or writing about it.

Getting informatics help

http://cri.uchicago.edu



STRENGTHEN YOUR FALL GRANT APPLICATIONS

Join us August 1 for a free session on including informatics resources at all stages of grant preparation.

GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources

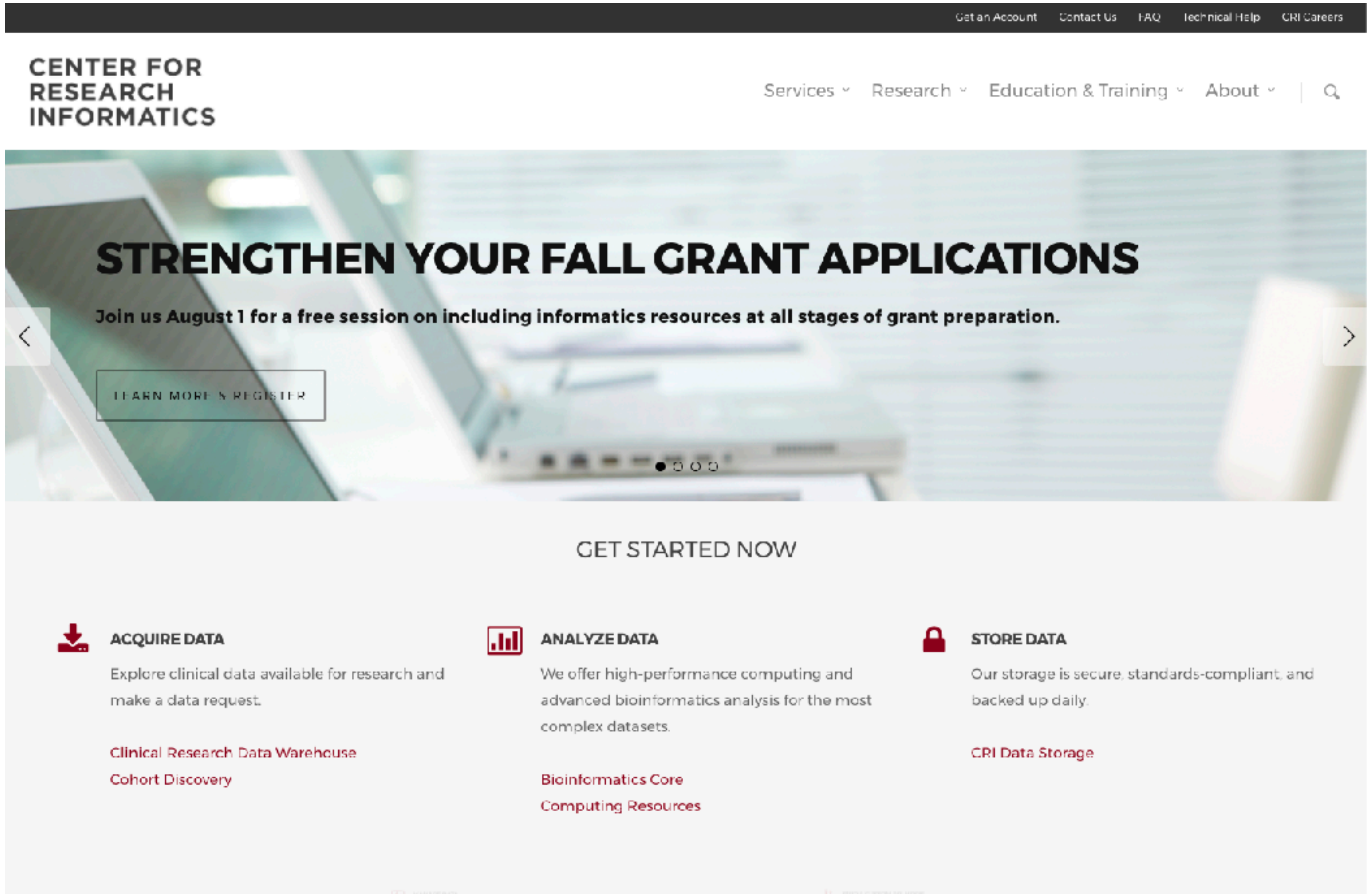
STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

CRI Data Storage

Getting informatics help

http://cri.uchicago.edu



Get an Account Contact Us FAQ Technical Help CRI Careers

CENTER FOR RESEARCH INFORMATICS

Services Research Education & Training About

STRENGTHEN YOUR FALL GRANT APPLICATIONS

Join us August 1 for a free session on including informatics resources at all stages of grant preparation.

LEARN MORE & REGISTER

GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources

STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

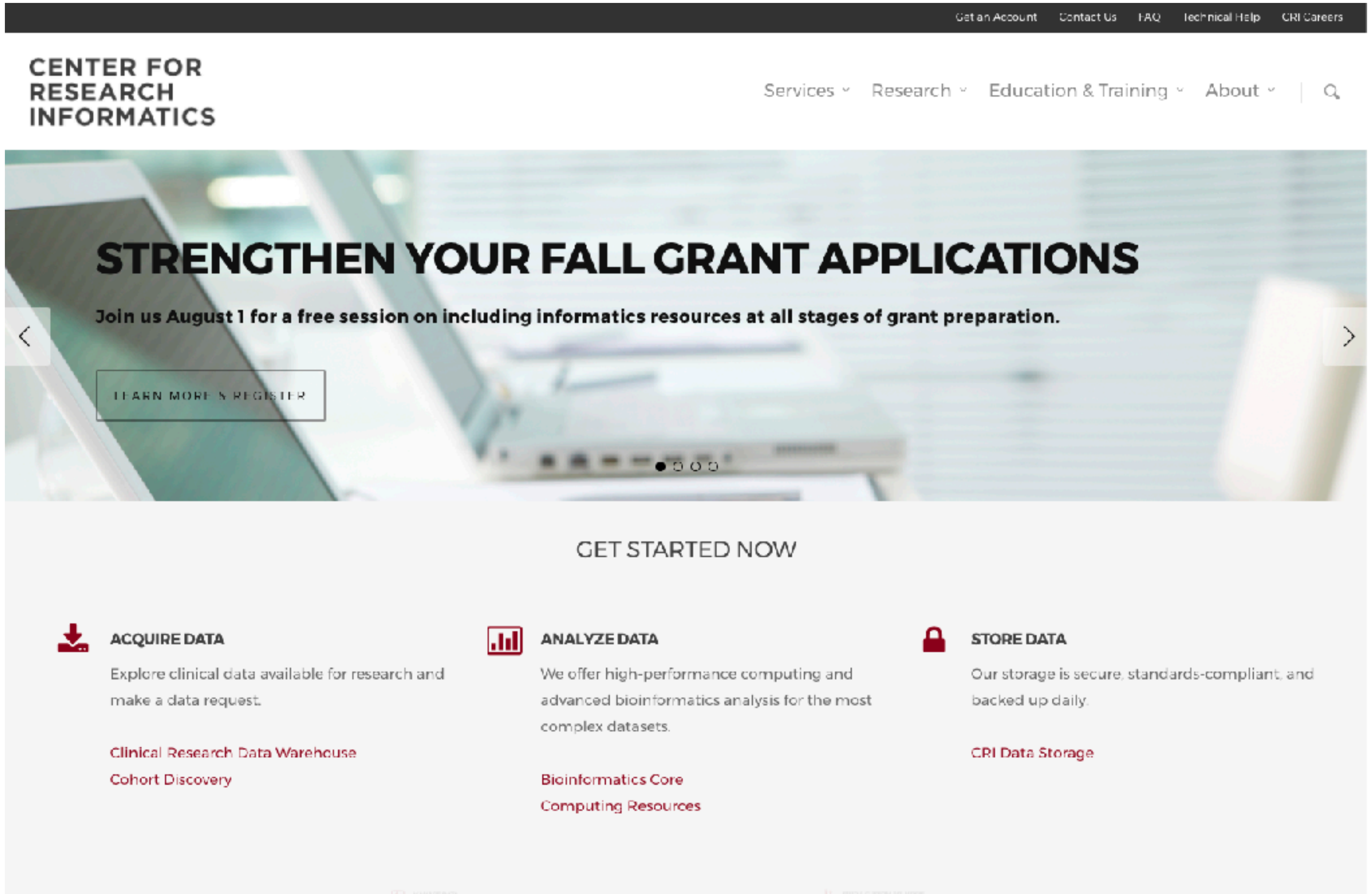
CRI Data Storage

slv@uchicago.edu

support@rt.cri.uchicago.edu

Getting informatics help

http://cri.uchicago.edu



STRENGTHEN YOUR FALL GRANT APPLICATIONS

Join us August 1 for a free session on including informatics resources at all stages of grant preparation.

GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources

STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

CRI Data Storage

slv@uchicago.edu

support@rt.cri.uchicago.edu

Common to all proposals

- IRB writing / consideration
- Contracts, data use agreements
- Data storage, movement, backup
- Data security
- Letters of support
- Facilities and resources documentation
- Data governance and stewardship
- Data sharing / software dissemination

IRB

- Do you need and IRB? An exemption?
- CRI has extensive experience in writing IRB protocols and shepherding them through the process
- Many of the issues have already been encountered for other proposals
- Engage the CRI **early** in the process

Contracts and data use agreements

- Sharing data outside the BSD requires an agreement
- Contracts may be needed for IP, data use, etc.
- The Data Use and Innovation Group meets monthly to address these issues **proactively** (CRI, OCR, IRB, legal, security, privacy)

Data storage, movement, backup

- CRI has extensive storage and backup capabilities
- Every investigator gets 2TB storage and backup for "free" as a lab share
- More extensive data usage needs to have a budget

A word about storage

WD - easystore® 8TB External USB 3.0 Hard Drive - Black

Model: WDBCKA0080HBK-NESN SKU: 5792401 4.8 (211)

24 Questions, 63 Answers

PRICE MATCH GUARANTEE

$199.99

SAVE $100 Was $299.99

Included Free: 1 item

FREE 2-DAY SHIPPING
on orders $35 and up

WE FIX IT OR REPLACE IT
Normal Wear & Tear • Power Surges

○ 2-Year Geek Squad Product Replacement - $19.99

ive Capacity: 8TB

ing: FREE Two Day
by Wed, Aug 2 for 60601 Change
der within 5h 43m

tore Pickup: JOHN HANCOCK CTR IL
lable Today (Ready in 1 hour) - FREE
rby stores

These aren't good places to store your data.

Why?

A word about storage

WD - easystore® 8TB External USB 3.0 Hard Drive - Black

Model: WDBCKA0080HBK-NESN SKU: 5792401 ★★★★★ 4.8 (211)
24 Questions, 63 Answers

ONLY @ BEST BUY

Click or tap to enlarge

Add to Cart

PRICE MATCH GUARANTEE

$199.99

SAVE $100 Was $299.99

Included Free: 1 item

FREE 2-DAY SHIPPING
on orders $35 and up

WE FIX IT OR REPLACE IT
Normal Wear & Tear • Power Surges
Plan Details

○ 2-Year Geek Squad Product Replacement - $19.99

Save for Later | Add to Registry

ive Capacity:

8TB

ing: FREE Two Day
by Wed, Aug 2 for 60601 Change
der within 5h 43m

tore Pickup: JOHN HANCOCK CTR IL
able Today (Ready in 1 hour) - FREE
rby stores

These aren't good places to store your data.

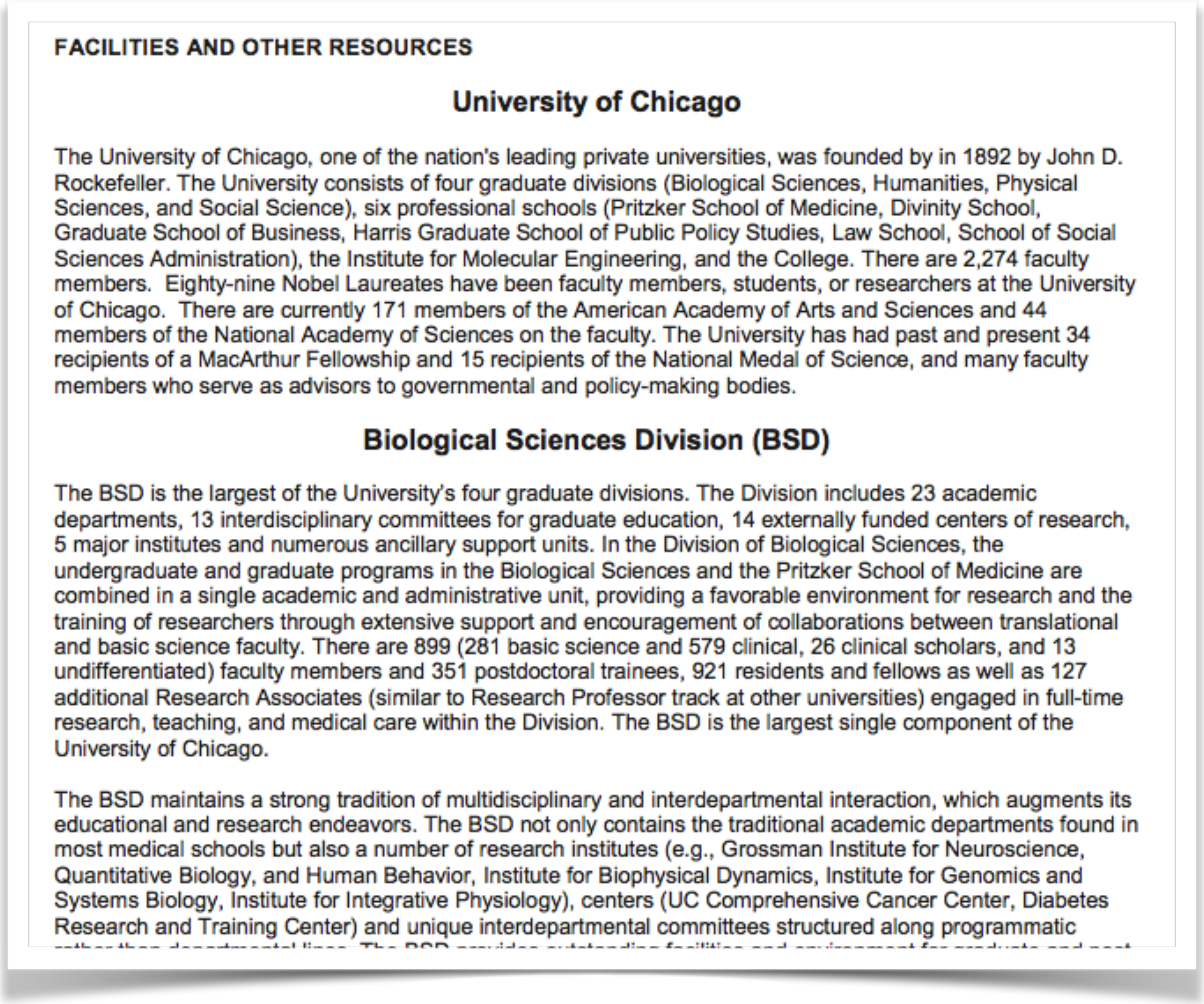
Why?

- Not HIPAA compliant
- Insecure
- No redundant backup
- Little chance of recovery if loss

Letters of support

- General letter from CRI
- Specific support for project from CRI leadership
- Contact the CRI director service line director for any LoS issues
- Do this early. A draft is always appreciated.

Facilities and resources pages



FACILITIES AND OTHER RESOURCES

University of Chicago

The University of Chicago, one of the nation's leading private universities, was founded by in 1892 by John D. Rockefeller. The University consists of four graduate divisions (Biological Sciences, Humanities, Physical Sciences, and Social Science), six professional schools (Pritzker School of Medicine, Divinity School, Graduate School of Business, Harris Graduate School of Public Policy Studies, Law School, School of Social Sciences Administration), the Institute for Molecular Engineering, and the College. There are 2,274 faculty members. Eighty-nine Nobel Laureates have been faculty members, students, or researchers at the University of Chicago. There are currently 171 members of the American Academy of Arts and Sciences and 44 members of the National Academy of Sciences on the faculty. The University has had past and present 34 recipients of a MacArthur Fellowship and 15 recipients of the National Medal of Science, and many faculty members who serve as advisors to governmental and policy-making bodies.

Biological Sciences Division (BSD)

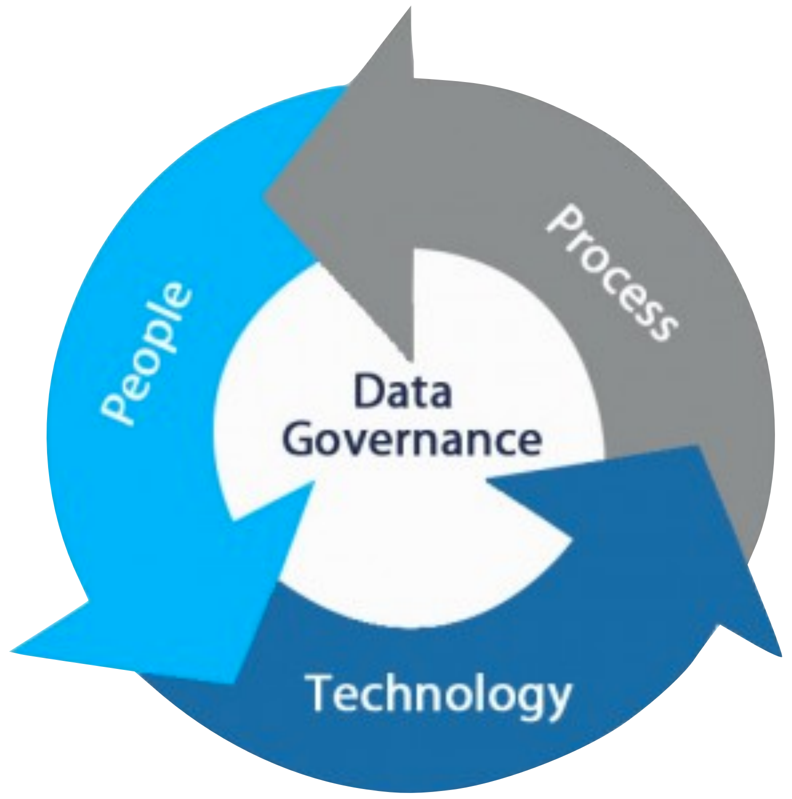
The BSD is the largest of the University's four graduate divisions. The Division includes 23 academic departments, 13 interdisciplinary committees for graduate education, 14 externally funded centers of research, 5 major institutes and numerous ancillary support units. In the Division of Biological Sciences, the undergraduate and graduate programs in the Biological Sciences and the Pritzker School of Medicine are combined in a single academic and administrative unit, providing a favorable environment for research and the training of researchers through extensive support and encouragement of collaborations between translational and basic science faculty. There are 899 (281 basic science and 579 clinical, 26 clinical scholars, and 13 undifferentiated) faculty members and 351 postdoctoral trainees, 921 residents and fellows as well as 127 additional Research Associates (similar to Research Professor track at other universities) engaged in full-time research, teaching, and medical care within the Division. The BSD is the largest single component of the University of Chicago.

The BSD maintains a strong tradition of multidisciplinary and interdepartmental interaction, which augments its educational and research endeavors. The BSD not only contains the traditional academic departments found in most medical schools but also a number of research institutes (e.g., Grossman Institute for Neuroscience, Quantitative Biology, and Human Behavior, Institute for Biophysical Dynamics, Institute for Genomics and Systems Biology, Institute for Integrative Physiology), centers (UC Comprehensive Cancer Center, Diabetes Research and Training Center) and unique interdepartmental committees structured along programmatic rather than departmental lines. The BSD provides outstanding facilities and environment for graduate and post

CRI has boilerplate language for grants

Data governance and stewardship

- Grant readers are looking for documentation of data governance procedures
- CRI can help document these procedures for your proposal



Examples of data governance considerations

- Why controls access to data?
- How is security documented?
- Will people have encrypted laptops?
- Is the storage HIPAA compliant?
- Are data being backed up regularly?
- How is data being moved securely between researchers?

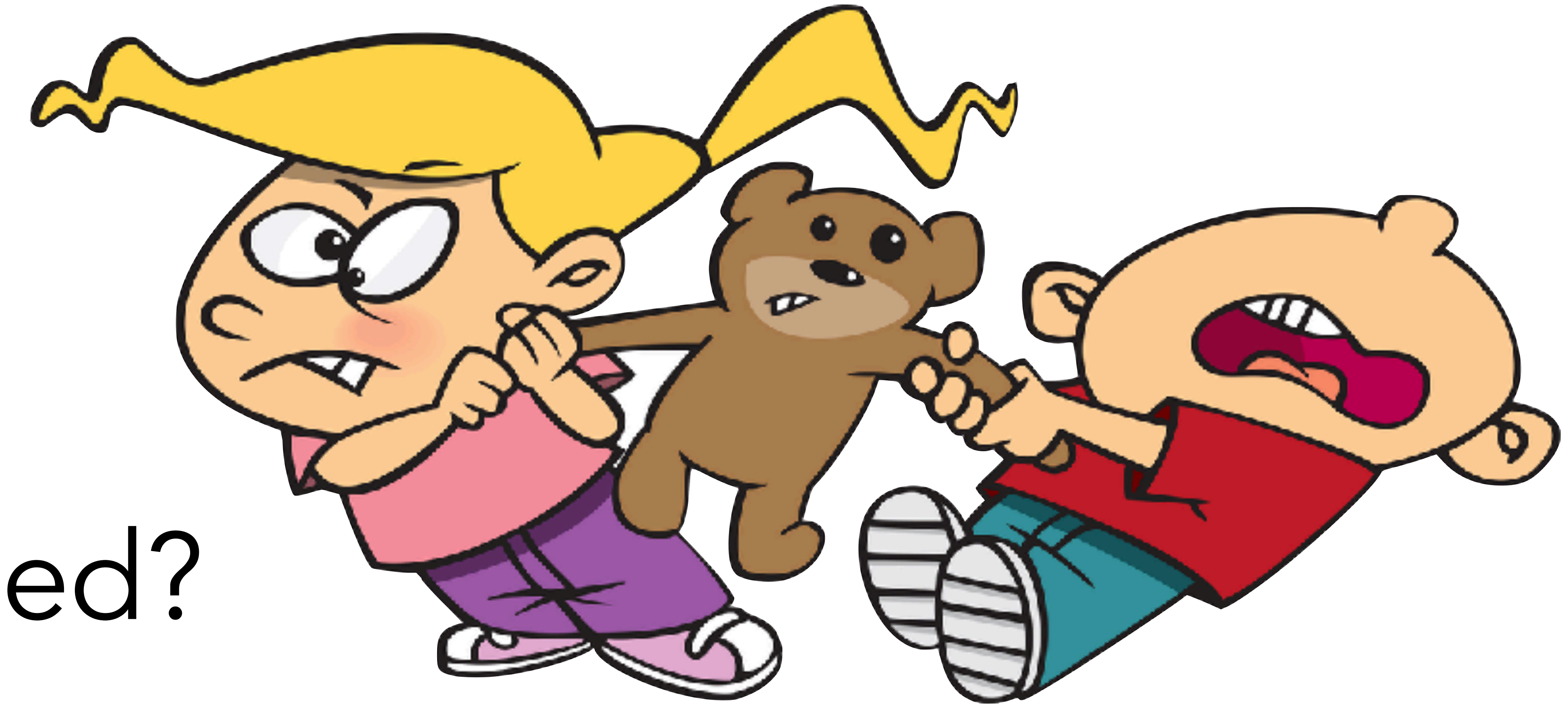
Examples of data governance considerations

- Why controls access to data?
- How is security documented?
- Will people have encrypted laptops?
- Is the storage HIPAA compliant?
- Are data being backed up regularly?
- How is data being moved securely between researchers?

Failure to address these questions can doom a proposal.

Data sharing plan

- Data sharing
  - Discussion of how data will be deposited in common repositories and shared
- Software dissemination
  - How will software be shared?
  - What kind of license will be used?
- CRI will help with this



Security considerations

- The Information Security Office in the BSD can help you with securing your data
- http://security.bsd.uchicago.edu/
- All proposals must adhere to policies and procedures (University and hospital)

Bioinformatics considerations

Bioinformatics - Methods and study design

- What kind of analysis?
  RNA-Seq? ChIP-seq? WGS? WES?
- What depth of coverage?
- Power calculations: How many samples?
  Technical replicates? Biological replicates?

Bioinformatics - Budget planning for data generation

- How many chips? What cost to run?
- How about sample collection and preparation?
- CRI can help broker this process

Bioinformatics - Grant writing

- CRI can help with all phases of grant writing
  - Background
  - Preliminary data
  - Methods
  - Research plan

Bioinformatics - Data storage, movement, backup

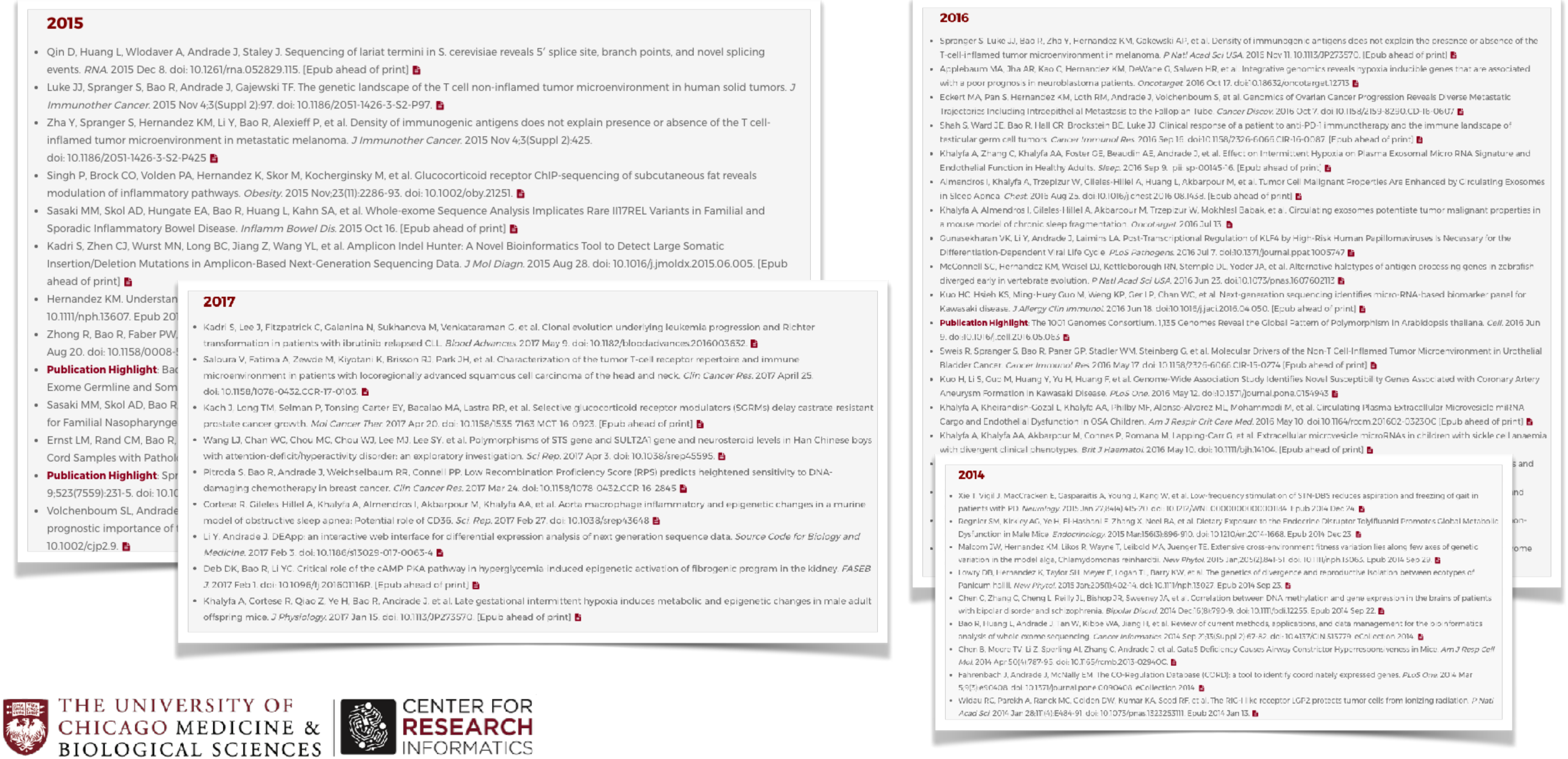
- How much storage is needed?
- How will data be transferred between investigators?
- Are data being redundantly backed up?
- CRI can help ensure that all phases are secure

Bioinformatics - Analysis and interpretation

Bioinformatics - Analysis and interpretation

- Best to involve a bioinformatician from the start
- Partnership is key for a successful collaboration
- Project time is charged on an hourly basis or through dedicated time on grants
- Co-authorship is expected, where appropriate

Publications



2015

- Qin D, Huang L, Wlodaver A, Andrade J, Staley J. Sequencing of lariat termini in S. cerevisiae reveals 5′ splice site, branch points, and novel splicing events. *RNA*. 2015 Dec 8. doi: 10.1261/rna.052829.115. [Epub ahead of print]
- Luke JJ, Spranger S, Bao R, Andrade J, Gajewski TF. The genetic landscape of the T cell non-inflamed tumor microenvironment in human solid tumors. *J Immunother Cancer*. 2015 Nov 4;3(Suppl 2):97. doi: 10.1186/2051-1426-3-S2-P97.
- Zha Y, Spranger S, Hernandez KM, Li Y, Bao R, Alexieff P, et al. Density of immunogenic antigens does not explain presence or absence of the T cell-inflamed tumor microenvironment in metastatic melanoma. *J Immunother Cancer*. 2015 Nov 4;3(Suppl 2):425. doi: 10.1186/2051-1426-3-S2-P425
- Singh P, Brock CO, Volden PA, Hernandez K, Skor M, Kocherginsky M, et al. Glucocorticoid receptor ChIP-sequencing of subcutaneous fat reveals modulation of inflammatory pathways. *Obesity*. 2015 Nov;23(11):2286-93. doi: 10.1002/oby.21251.
- Sasaki MM, Skol AD, Hungate EA, Bao R, Huang L, Kahn SA, et al. Whole-exome Sequence Analysis Implicates Rare Il17REL Variants in Familial and Sporadic Inflammatory Bowel Disease. *Inflamm Bowel Dis*. 2015 Oct 16. [Epub ahead of print]
- Kadri S, Zhen CJ, Wurst MN, Long BC, Jiang Z, Wang YL, et al. Amplicon Indel Hunter: A Novel Bioinformatics Tool to Detect Large Somatic Insertion/Deletion Mutations in Amplicon-Based Next-Generation Sequencing Data. *J Mol Diagn*. 2015 Aug 28. doi: 10.1016/j.jmoldx.2015.06.005. [Epub ahead of print]
- Hernandez KM. Understan
10.1111/nph.13607. Epub 201
- Zhong R, Bao R, Faber PW,
Aug 20. doi: 10.1158/0008-
- **Publication Highlight**: Bac
Exome Germline and Som
- Sasaki MM, Skol AD, Bao R,
for Familial Nasopharynge
- Ernst LM, Rand CM, Bao R,
Cord Samples with Pathol
- **Publication Highlight**: Spr
9;523(7559):231-5. doi: 10.10
- Volchenboum SL, Andrade
prognostic importance of t
10.1002/cjp2.9.

2017

- Kadri S, Lee J, Fitzpatrick C, Galanina N, Sukhanova M, Venkataraman G, et al. Clonal evolution underlying leukemia progression and Richter transformation in patients with ibrutinib-relapsed CLL. *Blood Advances*. 2017 May 9. doi: 10.1182/bloodadvances2016003632.
- Saloura V, Fatima A, Zewde M, Kiyotani K, Brisson RJ, Park JH, et al. Characterization of the tumor T-cell receptor repertoire and immune microenvironment in patients with locoregionally advanced squamous cell carcinoma of the head and neck. *Clin Cancer Res*. 2017 April 25. doi: 10.1158/1078-0432.CCR-17-0103.
- Kach J, Long TM, Selman P, Tonsing-Carter EY, Bacalao MA, Lastra RR, et al. Selective glucocorticoid receptor modulators (SGRMs) delay castrate-resistant prostate cancer growth. *Mol Cancer Ther*. 2017 Apr 20. doi: 10.1158/1535-7163.MCT-16-0923. [Epub ahead of print]
- Wang LJ, Chan WC, Chou MC, Chou WJ, Lee MJ, Lee SY, et al. Polymorphisms of STS gene and SULT2A1 gene and neurosteroid levels in Han Chinese boys with attention-deficit/hyperactivity disorder: an exploratory investigation. *Sci Rep*. 2017 Apr 3. doi: 10.1038/srep45595.
- Pitroda S, Bao R, Andrade J, Weichselbaum RR, Connell PP. Low Recombination Proficiency Score (RPS) predicts heightened sensitivity to DNA-damaging chemotherapy in breast cancer. *Clin Cancer Res*. 2017 Mar 24. doi: 10.1158/1078-0432.CCR-16-2845
- Cortese R, Gileles-Hillel A, Khalyfa A, Almendros I, Akbarpour M, Khalyfa AA, et al. Aorta macrophage inflammatory and epigenetic changes in a murine model of obstructive sleep apnea: Potential role of CD36. *Sci. Rep*. 2017 Feb 27. doi: 10.1038/srep43648
- Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code for Biology and Medicine*. 2017 Feb 3. doi: 10.1186/s13029-017-0063-4
- Deb DK, Bao R, Li YC. Critical role of the cAMP-PKA pathway in hyperglycemia-induced epigenetic activation of fibrogenic program in the kidney. *FASEB J*. 2017 Feb 1. doi: 10.1096/fj.201601116R. [Epub ahead of print]
- Khalyfa A, Cortese R, Qiao Z, Ye H, Bao R, Andrade J, et al. Late gestational intermittent hypoxia induces metabolic and epigenetic changes in male adult offspring mice. *J Physiology*. 2017 Jan 15. doi: 10.1113/JP273570. [Epub ahead of print]

2016

- Spranger S, Luke JJ, Bao R, Zha Y, Hernandez KM, Gakowski AP, et al. Density of immunogenic antigens does not explain the presence or absence of the T-cell-inflamed tumor microenvironment in melanoma. *P Natl Acad Sci USA*. 2016 Nov 11. 10.1113/JP273570. [Epub ahead of print]
- Applebaum MA, Jha AR, Kao C, Hernandez KM, DeWane G, Salwen HR, et al. Integrative genomics reveals hypoxia inducible genes that are associated with a poor prognosis in neuroblastoma patients. *Oncotarget*. 2016 Oct 17. doi:10.18632/oncotarget.12713
- Eckert MA, Pan S, Hernandez KM, Loth RM, Andrade J, Volchenboum S, et al. Genomics of Ovarian Cancer Progression Reveals Diverse Metastatic Trajectories Including Intraepithelial Metastasis to the Fallopian Tube. *Cancer Discov*. 2016 Oct 7. doi:10.1158/2159-8290.CD-16-0607
- Shah S, Ward JE, Bao R, Hall CR, Brockstein BE, Luke JJ. Clinical response of a patient to anti-PD-1 immunotherapy and the immune landscape of testicular germ cell tumors. *Cancer Immunol Res*. 2016 Sep 15. doi:10.1158/2326-6066.CIR-16-0087. [Epub ahead of print]
- Khalyfa A, Zhang C, Khalyfa AA, Foster GE, Beaudin AE, Andrade J, et al. Effect on Intermittent Hypoxia on Plasma Exosomal Micro RNA Signature and Endothelial Function in Healthy Adults. *Sleep*. 2016 Sep 9. pii: sp-00145-16. [Epub ahead of print]
- Almendros I, Khalyfa A, Trzepizur W, Gileles-Hillel A, Huang L, Akbarpour M, et al. Tumor Cell Malignant Properties Are Enhanced by Circulating Exosomes in Sleep Apnea. *Chest*. 2016 Aug 25. doi:10.1016/j.chest.2016.08.1438. [Epub ahead of print]
- Khalyfa A, Almendros I, Gileles-Hillel A, Akbarpour M, Trzepizur W, Mokhlesi Babak, et al. Circulating exosomes potentiate tumor malignant properties in a mouse model of chronic sleep fragmentation. *Oncotarget*. 2016 Jul 13.
- Gunasekharan VK, Li Y, Andrade J, Laimins LA. Post-Transcriptional Regulation of KLF4 by High-Risk Human Papillomaviruses Is Necessary for the Differentiation-Dependent Viral Life Cycle. *PLoS Pathogens*. 2016 Jul 7. doi:10.1371/journal.ppat.1005747
- McConnell SC, Hernandez KM, Weisel DJ, Kettleborough RN, Stemple DL, Yoder JA, et al. Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution. *P Natl Acad Sci USA*. 2016 Jun 23. doi:10.1073/pnas.1607602113
- Kuo HC, Hsieh KS, Ming-Huey Guo M, Weng KP, Ger LP, Chan WC, et al. Next-generation sequencing identifies micro-RNA-based biomarker panel for Kawasaki disease. *J Allergy Clin Immunol*. 2016 Jun 18. doi:10.1016/j.jaci.2016.04.050. [Epub ahead of print]
- **Publication Highlight**: The 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*. 2016 Jun 9. doi:10.1016/j.cell.2016.05.063
- Sweis R, Spranger S, Bao R, Paner GP, Stadler WM, Steinberg G, et al. Molecular Drivers of the Non-T Cell-Inflamed Tumor Microenvironment in Urothelial Bladder Cancer. *Cancer Immunol Res*. 2016 May 17. doi: 10.1158/2326-6066.CIR-15-0274 [Epub ahead of print]
- Kuo H, Li S, Guo M, Huang Y, Yu H, Huang F, et al. Genome-Wide Association Study Identifies Novel Susceptibility Genes Associated with Coronary Artery Aneurysm Formation in Kawasaki Disease. *PLoS One*. 2016 May 12. doi:10.1371/journal.pone.0154943
- Khalyfa A, Kheirandish-Gozal L, Khalyfa AA, Philby MF, Alonso-Alvarez ML, Mohammadi M, et al. Circulating Plasma Extracellular Microvesicle miRNA Cargo and Endothelial Dysfunction in OSA Children. *Am J Respir Crit Care Med*. 2016 May 10. doi:10.1164/rccm.201602-0323OC [Epub ahead of print]
- Khalyfa A, Khalyfa AA, Akbarpour M, Connes P, Romana M, Lapping-Carr G, et al. Extracellular microvesicle microRNAs in children with sickle cell anaemia with divergent clinical phenotypes. *Brit J Haematol*. 2016 May 10. doi: 10.1111/bjh.14104. [Epub ahead of print]
- s and
- nd
- on-
- ome

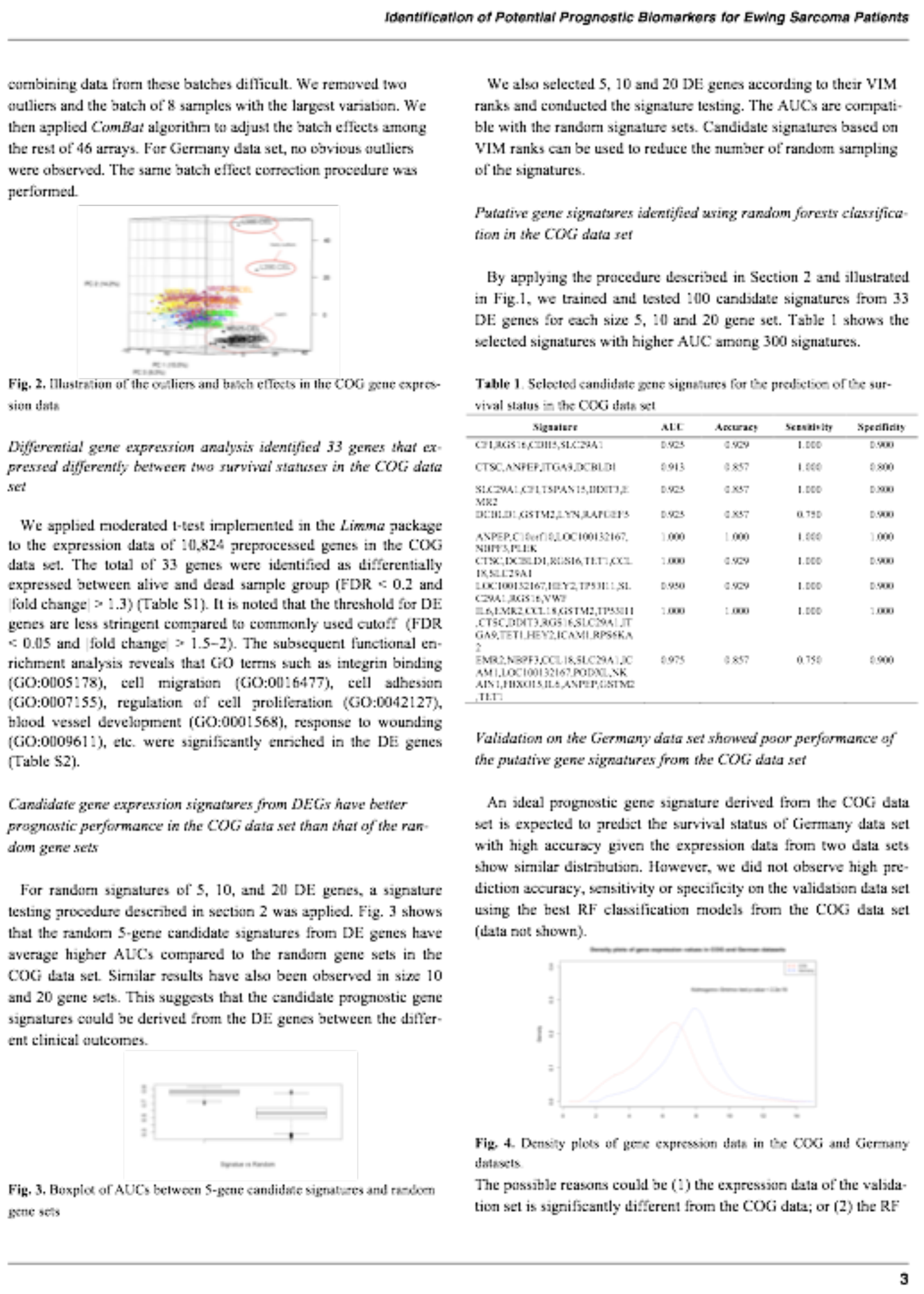
2014

- Xie T, Vigil J, MacCracken E, Gasparaitis A, Young J, Kang W, et al. Low-frequency stimulation of STN-DBS reduces aspiration and freezing of gait in patients with PD. *Neurology*. 2015 Jan 27;84(4):415-20. doi: 10.1212/WNL.0000000000001184. Epub 2014 Dec 24.
- Regnier SM, Kirkley AG, Ye H, El-Hashani E, Zhang X, Neel BA, et al. Dietary Exposure to the Endocrine Disruptor Tolylfluanid Promotes Global Metabolic Dysfunction in Male Mice. *Endocrinology*. 2015 Mar;156(3):896-910. doi: 10.1210/en.2014-1668. Epub 2014 Dec 23.
- Malcom JW, Hernandez KM, Likos R, Wayne T, Leibold MA, Juenger TE. Extensive cross-environment fitness variation lies along few axes of genetic variation in the model alga, Chlamydomonas reinhardtii. *New Phytol*. 2015 Jan;205(2):841-51. doi: 10.1111/nph.13063. Epub 2014 Sep 29.
- Lowry DB, Hernandez K, Taylor SH, Meyer E, Logan TL, Barry KW, et al. The genetics of divergence and reproductive isolation between ecotypes of Panicum hallii. *New Phytol*. 2015 Jan;205(1):402-14. doi: 10.1111/nph.13027. Epub 2014 Sep 23.
- Chen C, Zhang C, Cheng L, Reilly JL, Bishop JR, Sweeney JA, et al. Correlation between DNA methylation and gene expression in the brains of patients with bipolar disorder and schizophrenia. *Bipolar Disord*. 2014 Dec;16(8):790-9. doi: 10.1111/bdi.12255. Epub 2014 Sep 22.
- Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Informatics*. 2014 Sep 21;13(Suppl 2):67-82. doi: 10.4137/CIN.S13779. eCollection 2014.
- Chen B, Moore TV, Li Z, Sperling AI, Zhang C, Andrade J, et al. Gata5 Deficiency Causes Airway Constrictor Hyperresponsiveness in Mice. *Am J Resp Cell Mol*. 2014 Apr;50(4):787-95. doi: 10.1165/rcmb.2013-0294OC.
- Fahrenbach J, Andrade J, McNally EM. The CO-Regulation Database (CORD): a tool to identify coordinately expressed genes. *PLoS One*. 2014 Mar 5;9(3):e90408. doi: 10.1371/journal.pone.0090408. eCollection 2014.
- Widau RC, Parekh A, Ranck MC, Golden DW, Kumar KA, Sood RF, et al. The RIG-I-like receptor LGP2 protects tumor cells from ionizing radiation. *P Natl Acad Sci*. 2014 Jan 28;111(4):E484-91. doi: 10.1073/pnas.1323253111. Epub 2014 Jan 13.

Bioinformatics - Data integration

- Consider both phenotype and genotype data
- How will the clinical data be collected?
- Who is integrating these data into the analysis?
- CRI can get the clinical data **and** integrate it with the genomics information - this may require engaging the CRDW

Bioinformatics - Manuscript preparation and submission



Identification of Potential Prognostic Biomarkers for Ewing Sarcoma Patients

combining data from these batches difficult. We removed two outliers and the batch of 8 samples with the largest variation. We then applied *ComBat* algorithm to adjust the batch effects among the rest of 46 arrays. For Germany data set, no obvious outliers were observed. The same batch effect correction procedure was performed.

Fig. 2. Illustration of the outliers and batch effects in the COG gene expression data

Differential gene expression analysis identified 33 genes that expressed differently between two survival statuses in the COG data set

We applied moderated t-test implemented in the *Limma* package to the expression data of 10,824 preprocessed genes in the COG data set. The total of 33 genes were identified as differentially expressed between alive and dead sample group (FDR < 0.2 and |fold change| > 1.3) (Table S1). It is noted that the threshold for DE genes are less stringent compared to commonly used cutoff (FDR < 0.05 and |fold change| > 1.5–2). The subsequent functional enrichment analysis reveals that GO terms such as integrin binding (GO:0005178), cell migration (GO:0016477), cell adhesion (GO:0007155), regulation of cell proliferation (GO:0042127), blood vessel development (GO:0001568), response to wounding (GO:0009611), etc. were significantly enriched in the DE genes (Table S2).

Candidate gene expression signatures from DEGs have better prognostic performance in the COG data set than that of the random gene sets

For random signatures of 5, 10, and 20 DE genes, a signature testing procedure described in section 2 was applied. Fig. 3 shows that the random 5-gene candidate signatures from DE genes have average higher AUCs compared to the random gene sets in the COG data set. Similar results have also been observed in size 10 and 20 gene sets. This suggests that the candidate prognostic gene signatures could be derived from the DE genes between the different clinical outcomes.

Fig. 3. Boxplot of AUCs between 5-gene candidate signatures and random gene sets

We also selected 5, 10 and 20 DE genes according to their VIM ranks and conducted the signature testing. The AUCs are compatible with the random signature sets. Candidate signatures based on VIM ranks can be used to reduce the number of random sampling of the signatures.

Putative gene signatures identified using random forests classification in the COG data set

By applying the procedure described in Section 2 and illustrated in Fig.1, we trained and tested 100 candidate signatures from 33 DE genes for each size 5, 10 and 20 gene set. Table 1 shows the selected signatures with higher AUC among 300 signatures.

Table 1. Selected candidate gene signatures for the prediction of the survival status in the COG data set

Signature	AUC	Accuracy	Sensitivity	Specificity
CFI,RGS16,CDH5,SLC29A1	0.925	0.929	1.000	0.900
CTSC,ANPEP,ITGA9,DCBLD1	0.913	0.857	1.000	0.800
SLC29A1,CFI,TSPAN15,DDIT3,EMR2	0.925	0.857	1.000	0.800
DCBLD1,GSTM2,LYN,RAPGEF5	0.925	0.857	0.750	0.900
ANPEP,C10orf10,LOC100132167,NBPF3,PLEK	1.000	1.000	1.000	1.000
CTSC,DCBLD1,RGS16,TET1,CCL18,SLC29A1	1.000	0.929	1.000	0.900
LOC100132167,HEY2,TP53I11,SLC29A1,RGS16,VWF	0.950	0.929	1.000	0.900
IL6,EMR2,CCL18,GSTM2,TP53I11,CTSC,DDIT3,RGS16,SLC29A1,ITGA9,TET1,HEY2,ICAM1,RPS6KA2	1.000	1.000	1.000	1.000
EMR2,NBPF3,CCL18,SLC29A1,ICAM1,LOC100132167,PODXL,NKAIN1,FBXO15,IL6,ANPEP,GSTM2,TET1	0.975	0.857	0.750	0.900

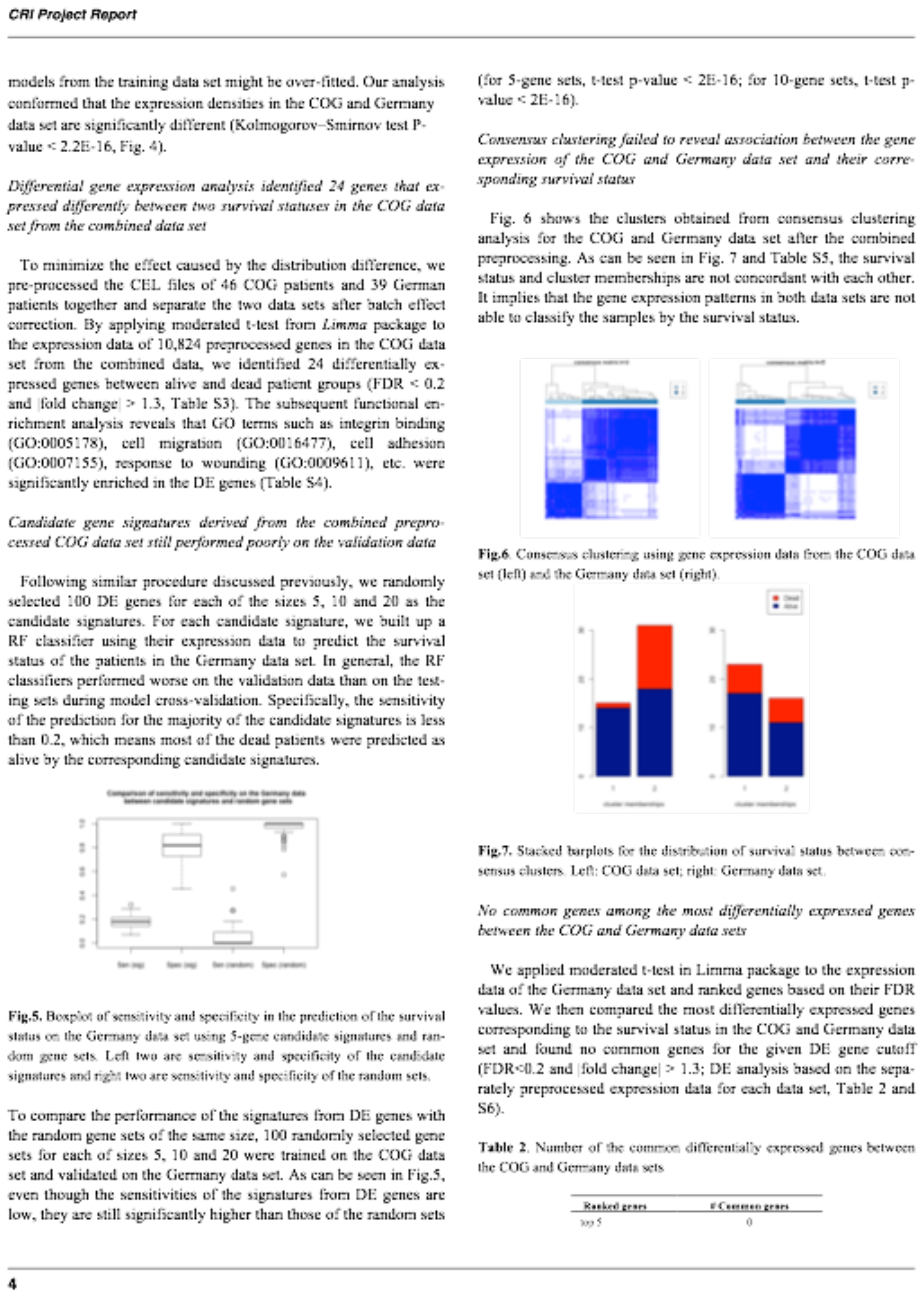
Validation on the Germany data set showed poor performance of the putative gene signatures from the COG data set

An ideal prognostic gene signature derived from the COG data set is expected to predict the survival status of Germany data set with high accuracy given the expression data from two data sets show similar distribution. However, we did not observe high prediction accuracy, sensitivity or specificity on the validation data set using the best RF classification models from the COG data set (data not shown).

Fig. 4. Density plots of gene expression data in the COG and Germany datasets.

The possible reasons could be (1) the expression data of the validation set is significantly different from the COG data; or (2) the RF

3



CRI Project Report

models from the training data set might be over-fitted. Our analysis conformed that the expression densities in the COG and Germany data set are significantly different (Kolmogorov–Smirnov test P-value < 2.2E-16, Fig. 4).

Differential gene expression analysis identified 24 genes that expressed differently between two survival statuses in the COG data set from the combined data set

To minimize the effect caused by the distribution difference, we pre-processed the CEL files of 46 COG patients and 39 German patients together and separate the two data sets after batch effect correction. By applying moderated t-test from *Limma* package to the expression data of 10,824 preprocessed genes in the COG data set from the combined data, we identified 24 differentially expressed genes between alive and dead patient groups (FDR < 0.2 and |fold change| > 1.3, Table S3). The subsequent functional enrichment analysis reveals that GO terms such as integrin binding (GO:0005178), cell migration (GO:0016477), cell adhesion (GO:0007155), response to wounding (GO:0009611), etc. were significantly enriched in the DE genes (Table S4).

Candidate gene signatures derived from the combined preprocessed COG data set still performed poorly on the validation data

Following similar procedure discussed previously, we randomly selected 100 DE genes for each of the sizes 5, 10 and 20 as the candidate signatures. For each candidate signature, we built up a RF classifier using their expression data to predict the survival status of the patients in the Germany data set. In general, the RF classifiers performed worse on the validation data than on the testing sets during model cross-validation. Specifically, the sensitivity of the prediction for the majority of the candidate signatures is less than 0.2, which means most of the dead patients were predicted as alive by the corresponding candidate signatures.

Fig.5. Boxplot of sensitivity and specificity in the prediction of the survival status on the Germany data set using 5-gene candidate signatures and random gene sets. Left two are sensitivity and specificity of the candidate signatures and right two are sensitivity and specificity of the random sets.

To compare the performance of the signatures from DE genes with the random gene sets of the same size, 100 randomly selected gene sets for each of sizes 5, 10 and 20 were trained on the COG data set and validated on the Germany data set. As can be seen in Fig.5, even though the sensitivities of the signatures from DE genes are low, they are still significantly higher than those of the random sets (for 5-gene sets, t-test p-value < 2E-16; for 10-gene sets, t-test p-value < 2E-16).

Consensus clustering failed to reveal association between the gene expression of the COG and Germany data set and their corresponding survival status

Fig. 6 shows the clusters obtained from consensus clustering analysis for the COG and Germany data set after the combined preprocessing. As can be seen in Fig. 7 and Table S5, the survival status and cluster memberships are not concordant with each other. It implies that the gene expression patterns in both data sets are not able to classify the samples by the survival status.

Fig.6. Consensus clustering using gene expression data from the COG data set (left) and the Germany data set (right).

Fig.7. Stacked barplots for the distribution of survival status between consensus clusters. Left: COG data set; right: Germany data set.

No common genes among the most differentially expressed genes between the COG and Germany data sets

We applied moderated t-test in Limma package to the expression data of the Germany data set and ranked genes based on their FDR values. We then compared the most differentially expressed genes corresponding to the survival status in the COG and Germany data set and found no common genes for the given DE gene cutoff (FDR<0.2 and |fold change| > 1.3; DE analysis based on the separately preprocessed expression data for each data set, Table 2 and S6).

Table 2. Number of the common differentially expressed genes between the COG and Germany data sets

Ranked genes	# Common genes
top 5	0

4

The Uchicago Clinical Research Data Warehouse

Epic

sunquest

GE Healthcare

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

The Uchicago Clinical Research Data Warehouse

850,000 patients

Epic

sunquest

GE Healthcare

The Uchicago Clinical Research Data Warehouse

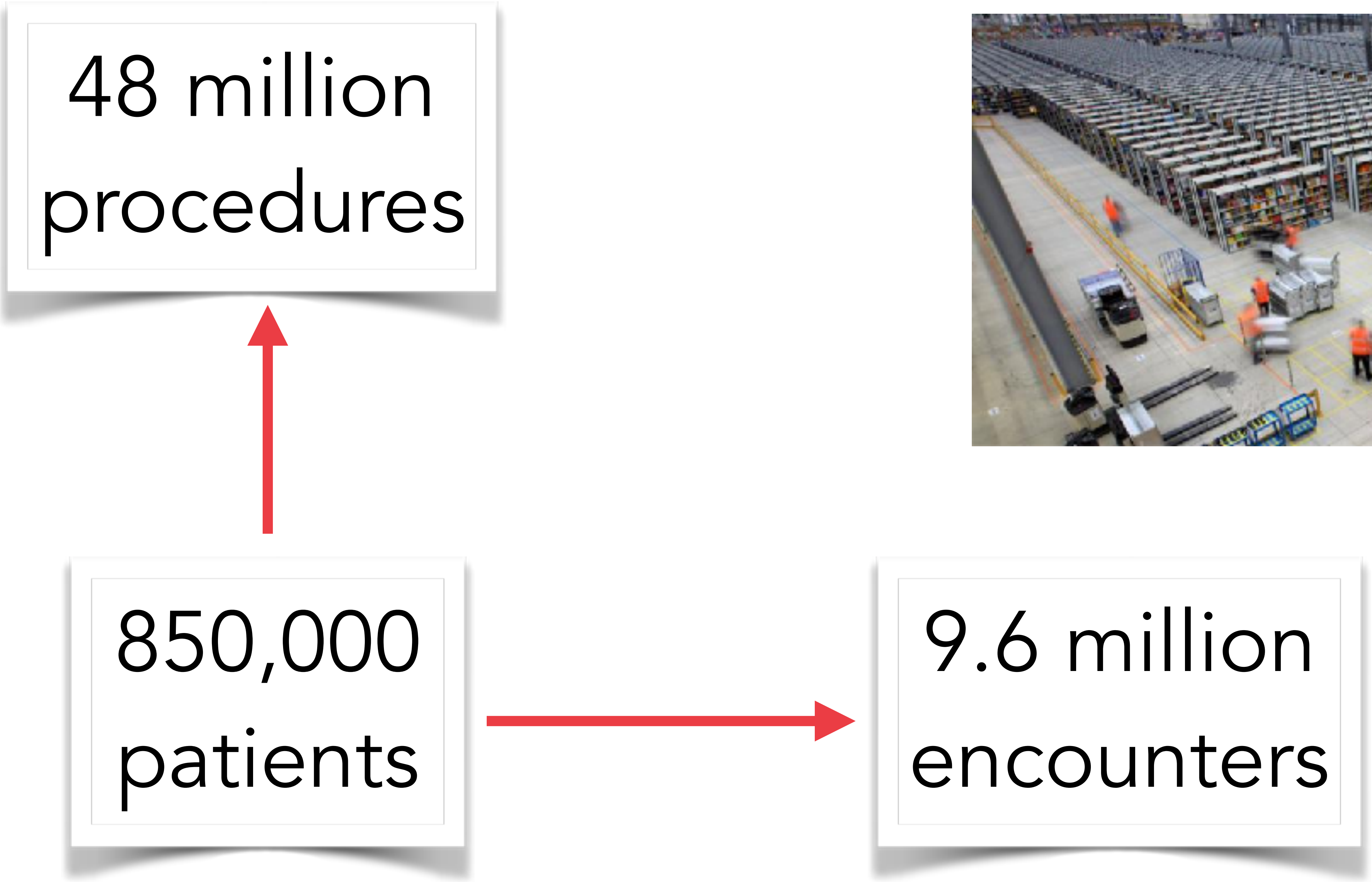
850,000 patients → 9.6 million encounters

Epic

sunquest

GE Healthcare

The Uchicago Clinical Research Data Warehouse


48 million procedures

850,000 patients

9.6 million encounters

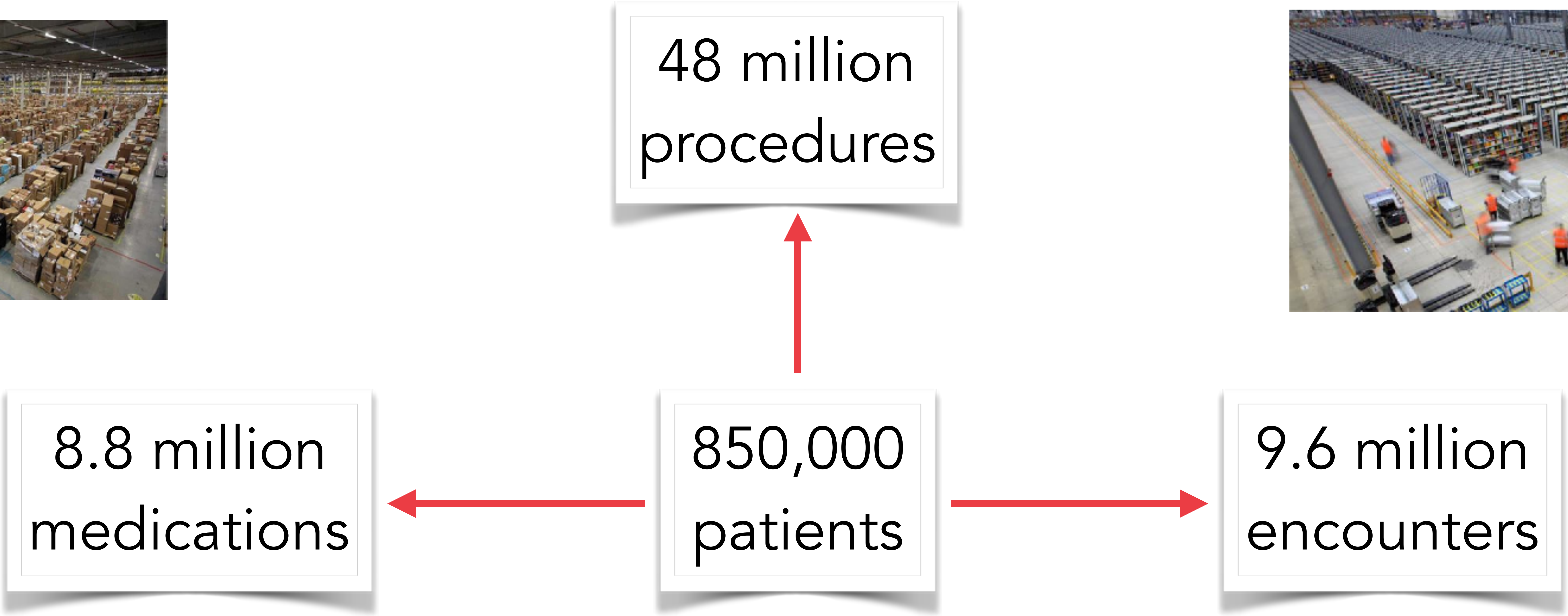
Epic

sunquest

GE Healthcare

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

The Uchicago Clinical Research Data Warehouse



48 million procedures

8.8 million medications

850,000 patients

9.6 million encounters

Epic

sunquest

GE Healthcare

The Uchicago Clinical Research Data Warehouse

48 million procedures

8.8 million medications

850,000 patients

9.6 million encounters

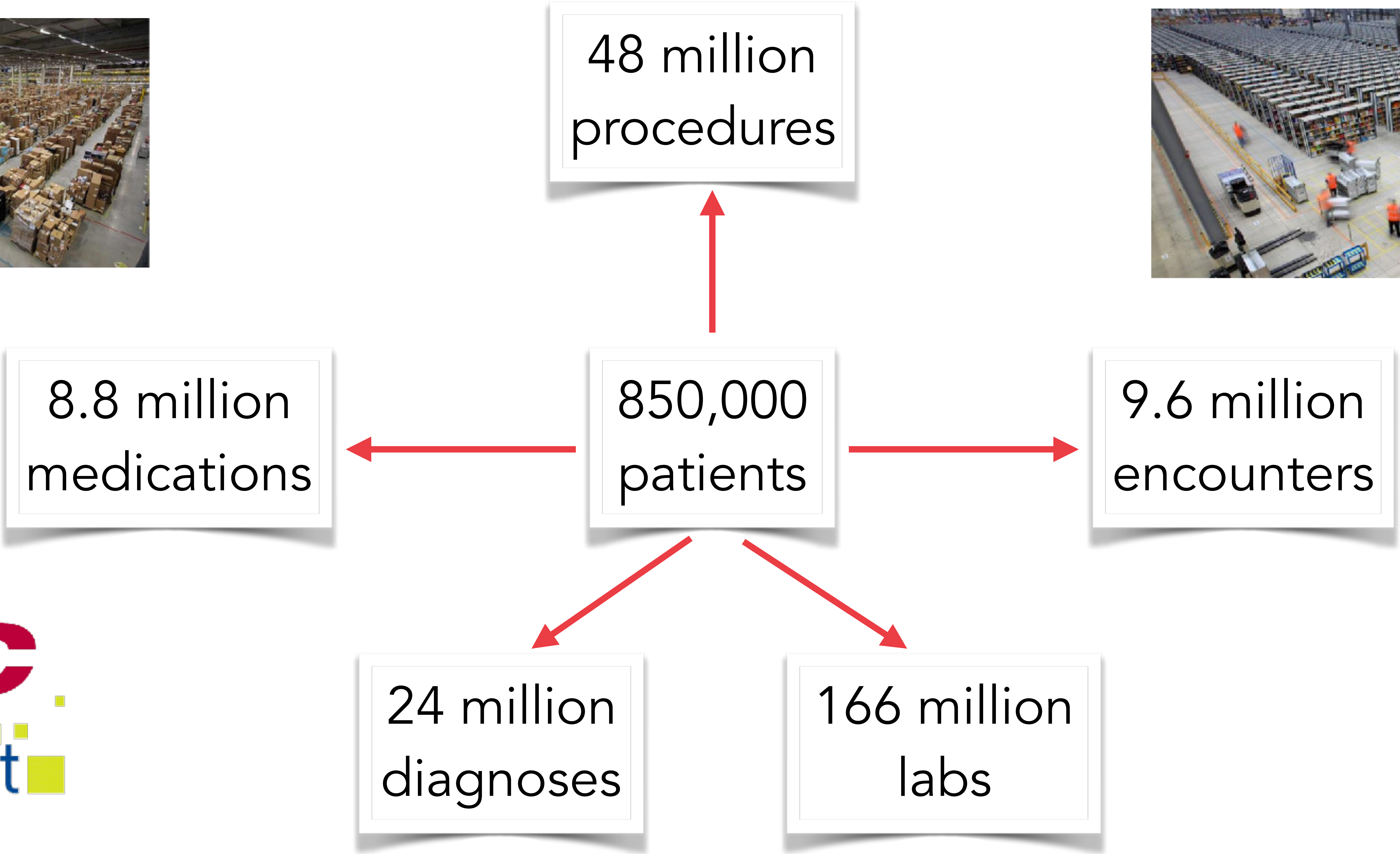
24 million diagnoses

Epic

sunquest

GE Healthcare

The Uchicago Clinical Research Data Warehouse



850,000 patients

48 million procedures

8.8 million medications

9.6 million encounters

24 million diagnoses

166 million labs

Epic

sunquest

GE Healthcare

Clinical Research Data Warehouse

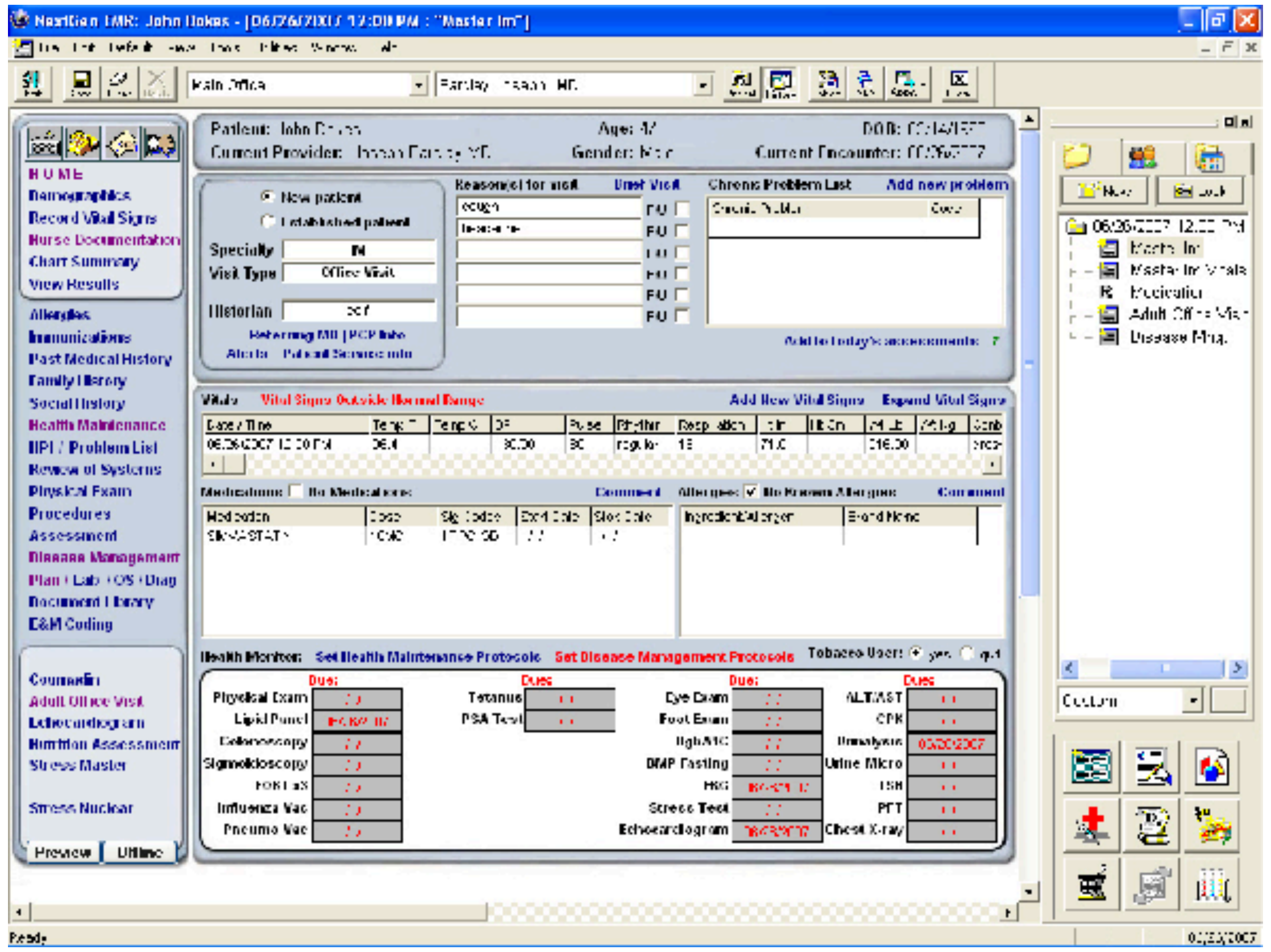
CRDW - Preliminary data / Cohort identification

- It can be hard to identify cohorts
- CRI has specialists to help

- Reviewers like to see preliminary identification of cohorts - “Can they really get the data”
- Sometimes new data has to be sourced

CRDW - Data element identification

- Identifying elements to pull from the CRDW is an iterative process
- Requires input from the CRDW and the investigator
- Delineation of data elements in the grant is essential



Patient: John [illegible] Age: 47 DOB: [illegible]
Current Provider: [illegible] Gender: Male Current Encounter: [illegible]

● New patient
○ Established patient
Specialty: IM
Visit Type: Office Visit

Vitals Vital Signs Outside Normal Range

Date / Time	Temp F	Temp C	BP	Pulse	Rhythm	Respiration	Ht in	Ht cm	Wt lb	Wt kg
06/26/2007 12:00 PM	98.4		90/60	90	regular	18	71.0		216.00	

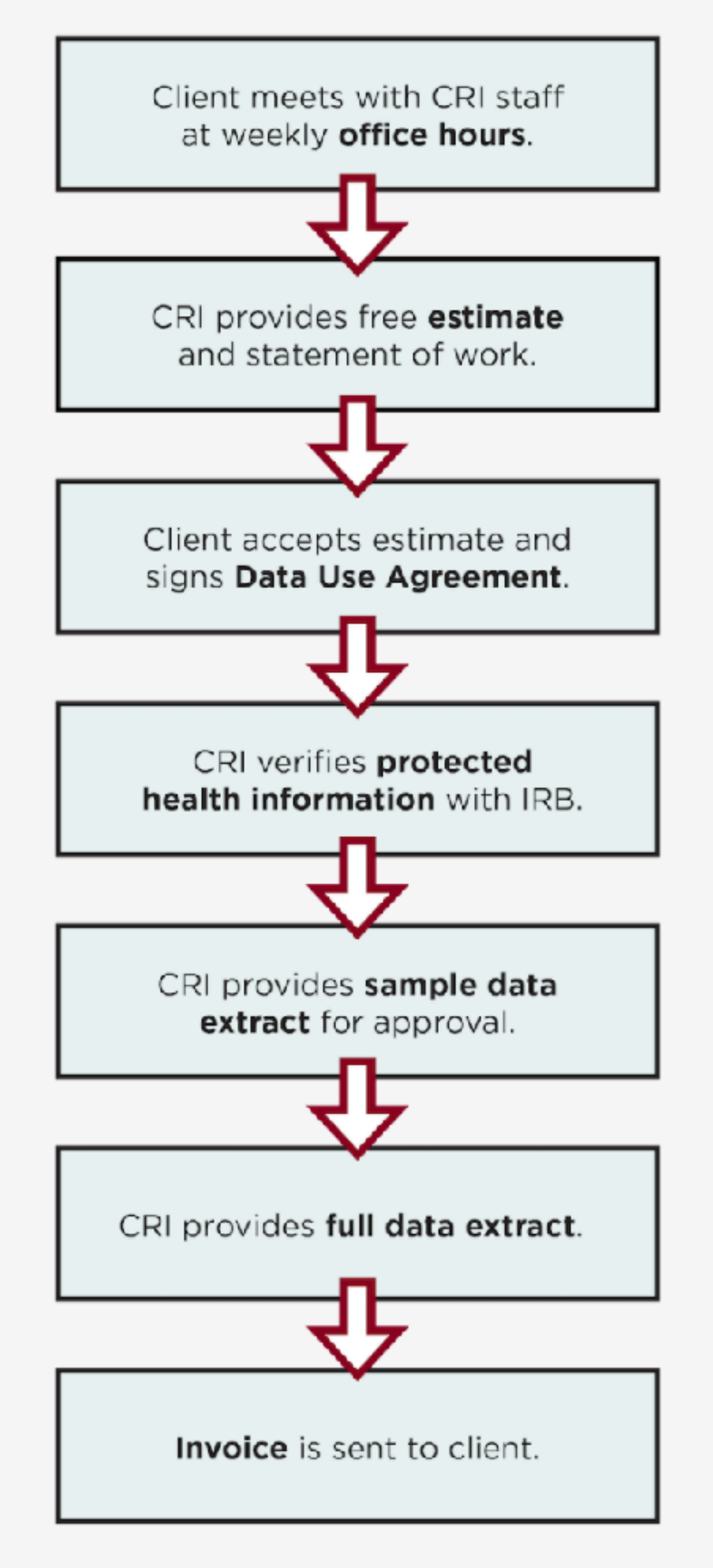
Medications □ No Medications; Allergies ☑ No Known Allergies

Medication	Dose	Sig Codes
SIMVASTATIN	[illegible]	[illegible]

Health Monitors: Tobacco User: ● yes ○ quit

Due:		Due:		Due:		Due:	
Physical Exam		Tetanus		Eye Exam		ALT/AST	
Lipid Panel	[illegible]	PSA Test		Foot Exam		CPK	
Colonoscopy				HgbA1C		Urinalysis	03/20/2007
Sigmoidoscopy				BMP Fasting		Urine Micro	
FOBT x3				EKG	[illegible]	TSH	
Influenza Vac				Stress Test		PFT	
Pneumo Vac				Echocardiogram	[illegible]	Chest X-ray	

Data request process



Client meets with CRI staff at weekly **office hours**.

CRI provides free **estimate** and statement of work.

Client accepts estimate and signs **Data Use Agreement**.

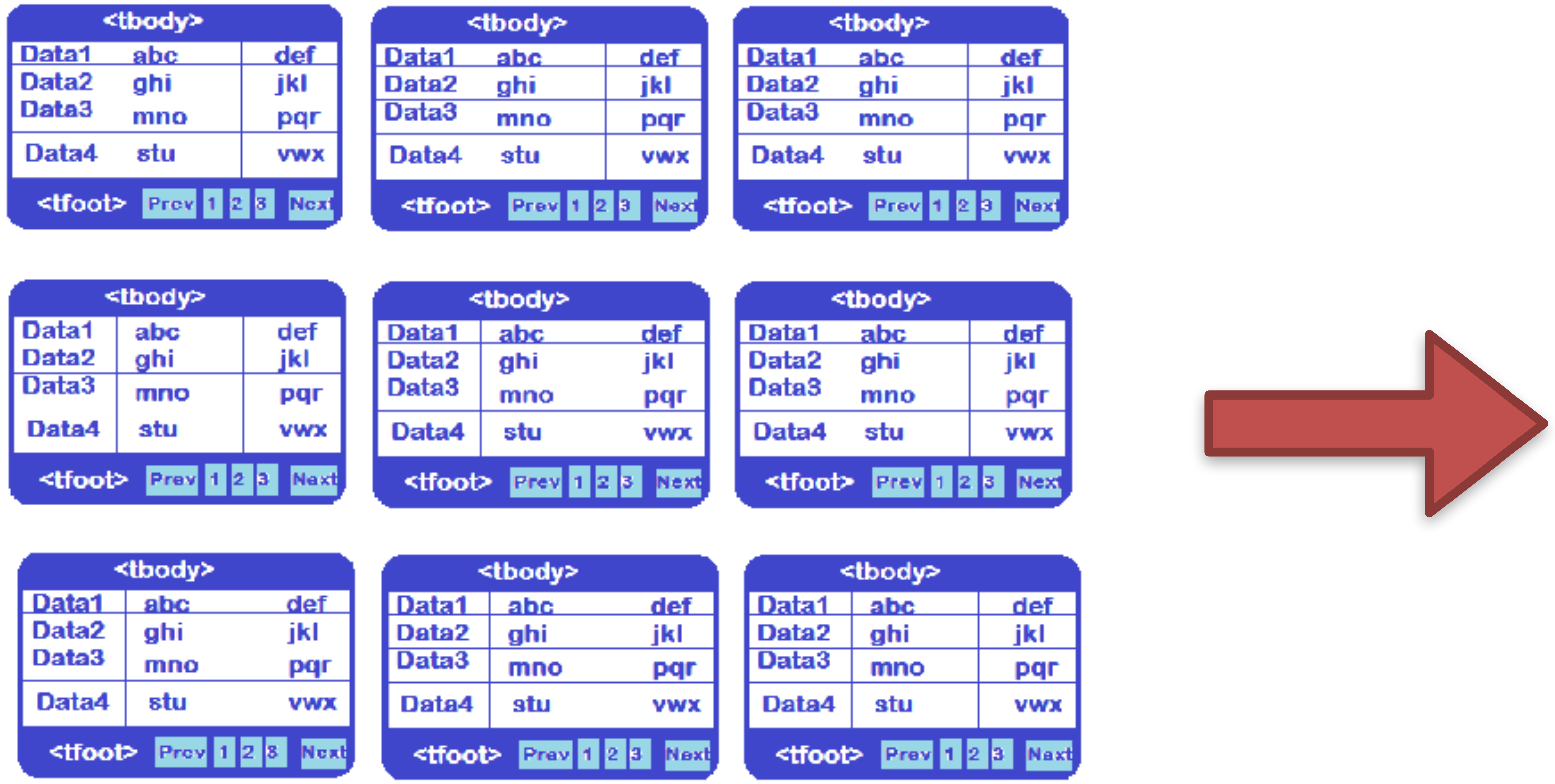
CRI verifies **protected health information** with IRB.

CRI provides **sample data extract** for approval.

CRI provides **full data extract**.

Invoice is sent to client.

CRDW - Data aggregation and normalization



<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

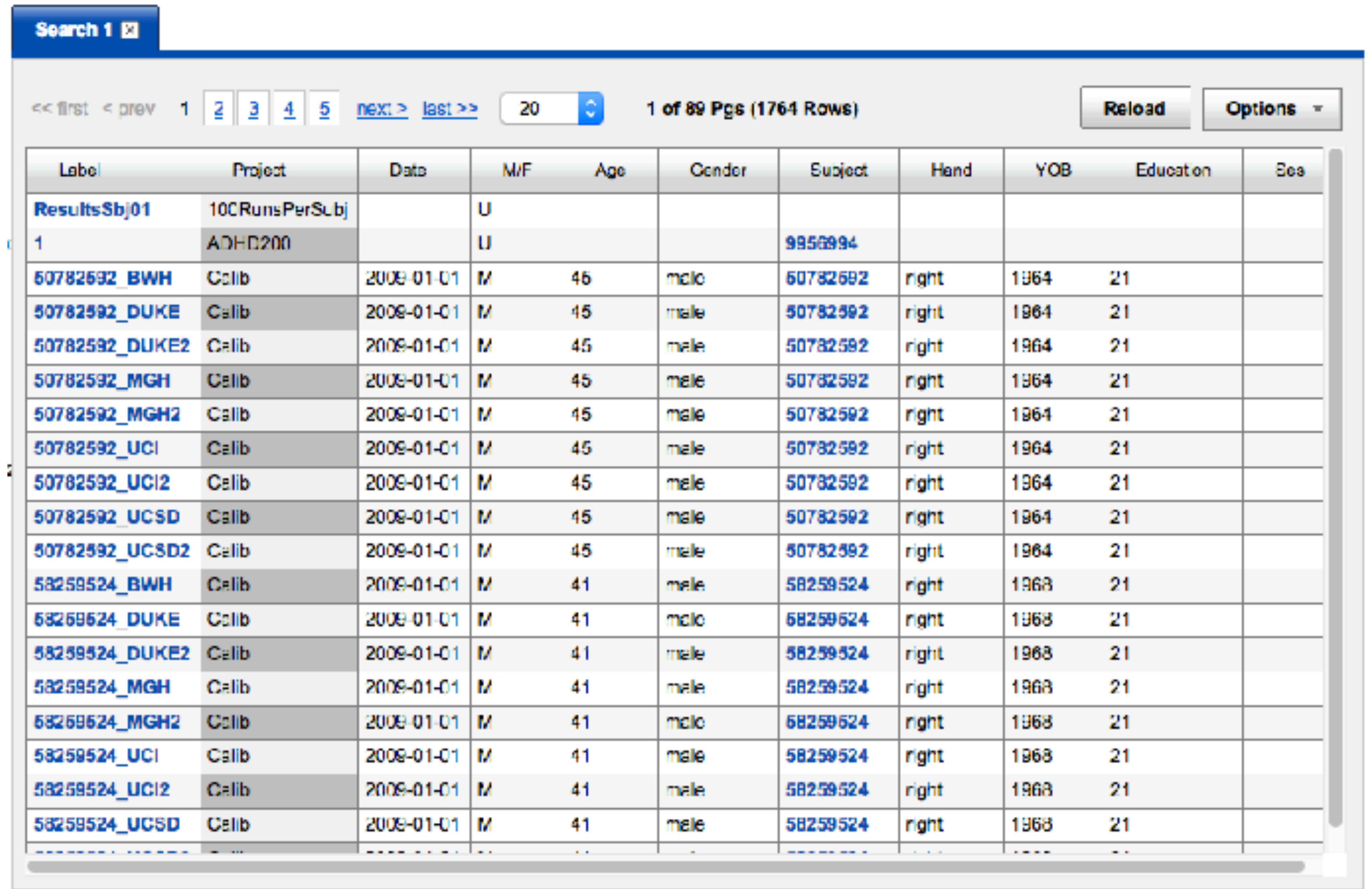
<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

<tbody>		
Data1	abc	def
Data2	ghi	jkl
Data3	mno	pqr
Data4	stu	vwx
<tfoot>	Prev 1 2 3	Next

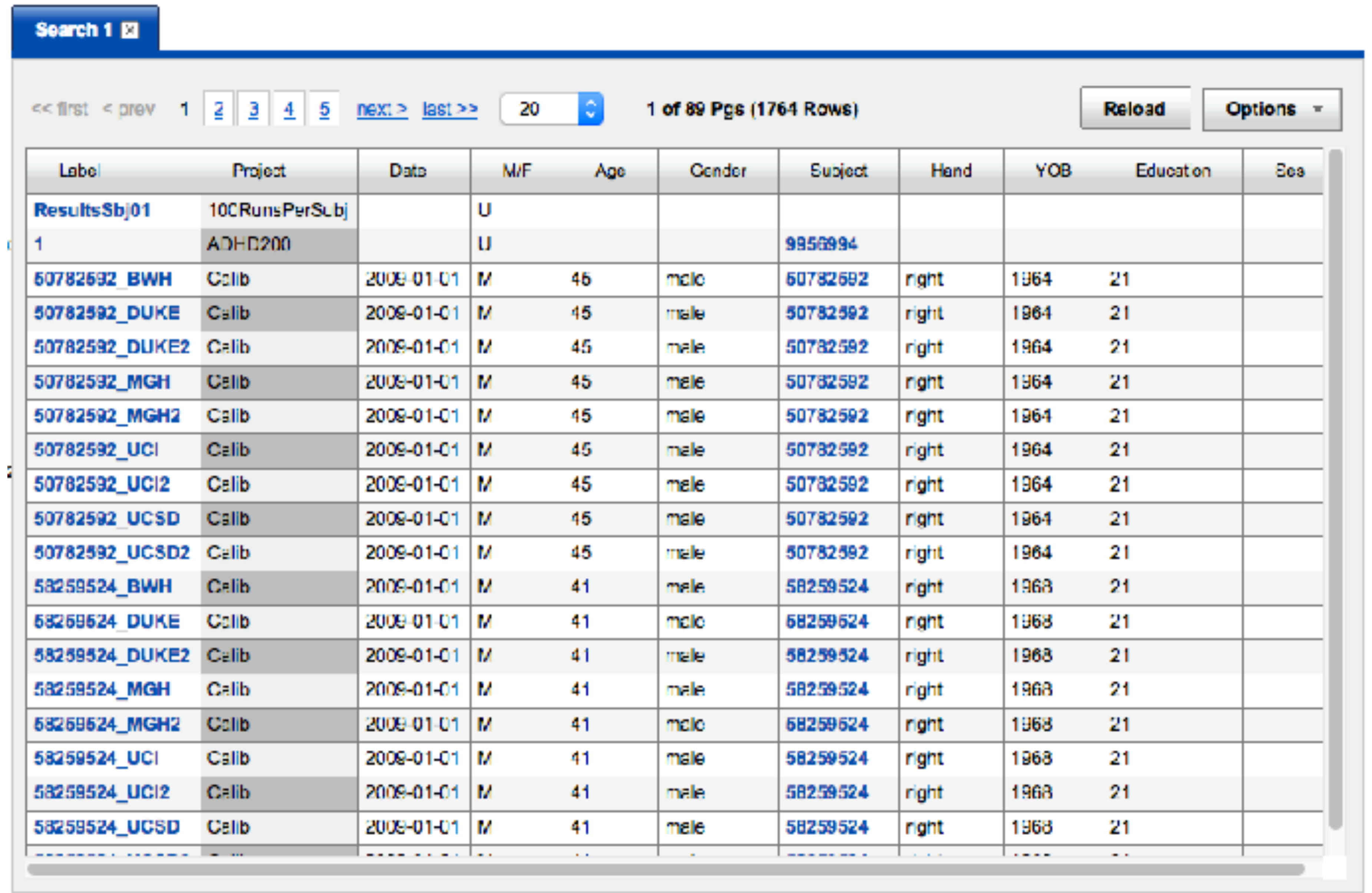


1 of 89 Pgs (1764 Rows)

Label	Project	Date	M/F	Age	Gender	Subject	Hand	YOB	Education	Ses
ResultsSbj01	10CRunsPerSubj		U							
1	ADHD200		U			9956994				
50782592_BWH	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_DUKE	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_DUKE2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_MGH	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_MGH2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCI	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCI2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCSD	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCSD2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
58259524_BWH	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_DUKE	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_DUKE2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_MGH	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_MGH2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCI	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCI2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCSD	Calib	2009-01-01	M	41	male	58259524	right	1968	21	

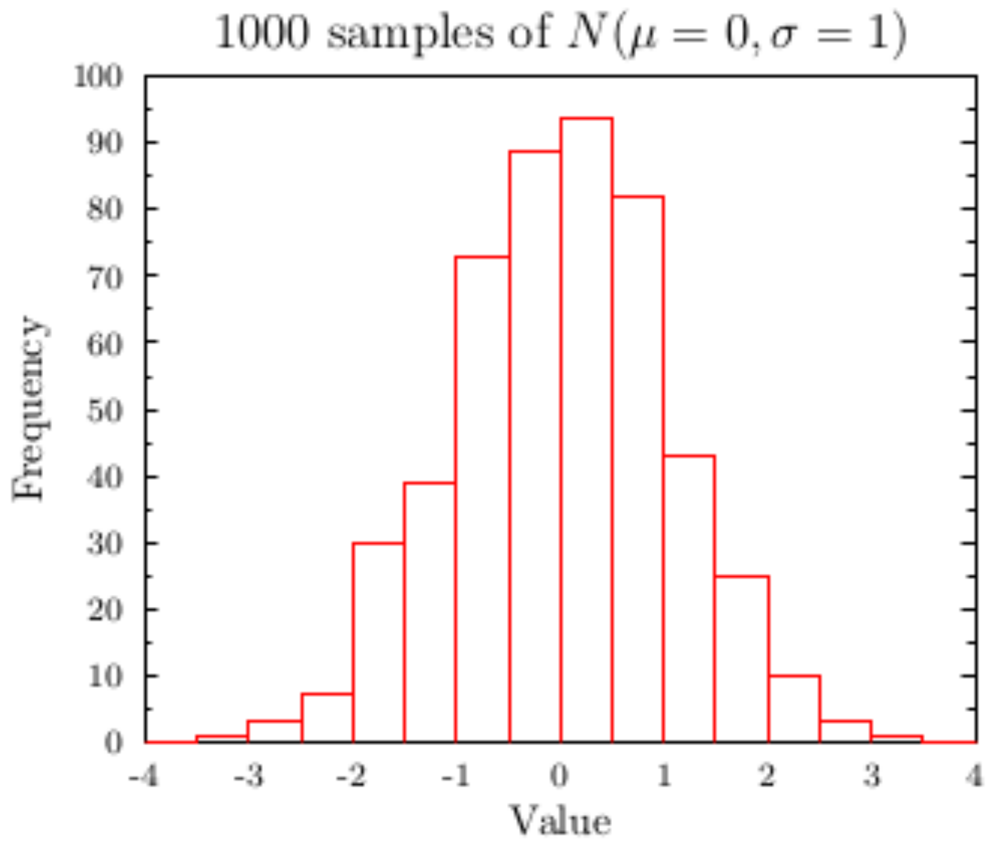
Taking the complex, multidimensional data from the CRDW and creating a usable data set for subsequent analysis requires special skills and should be included in the budget for data acquisition

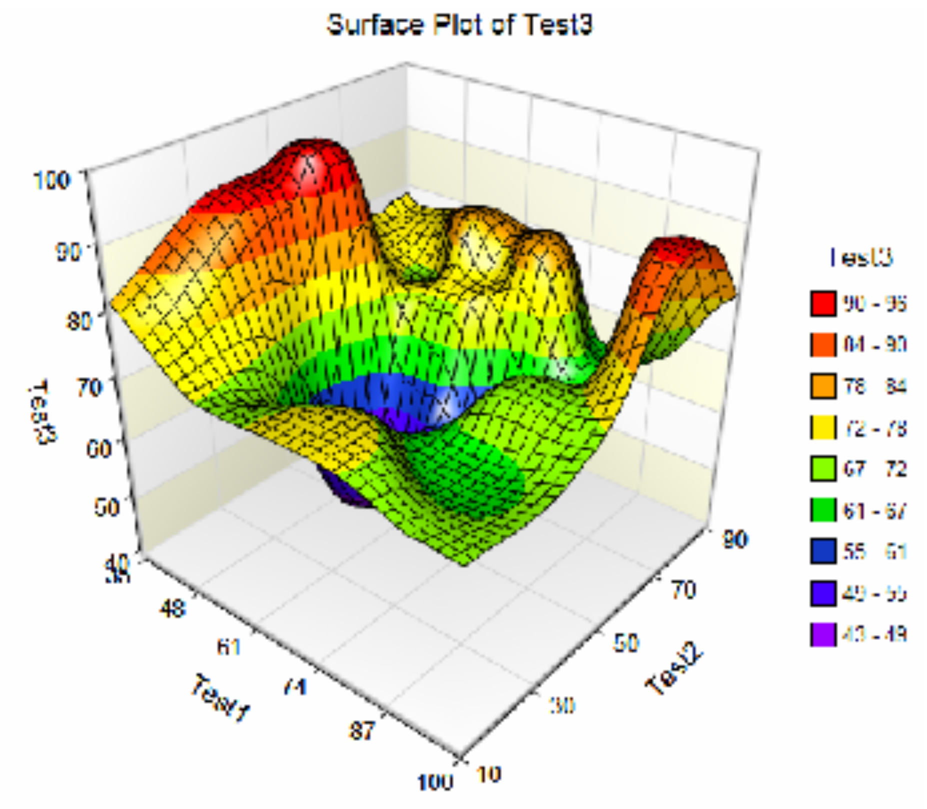
CRDW - Data analysis and interpretation

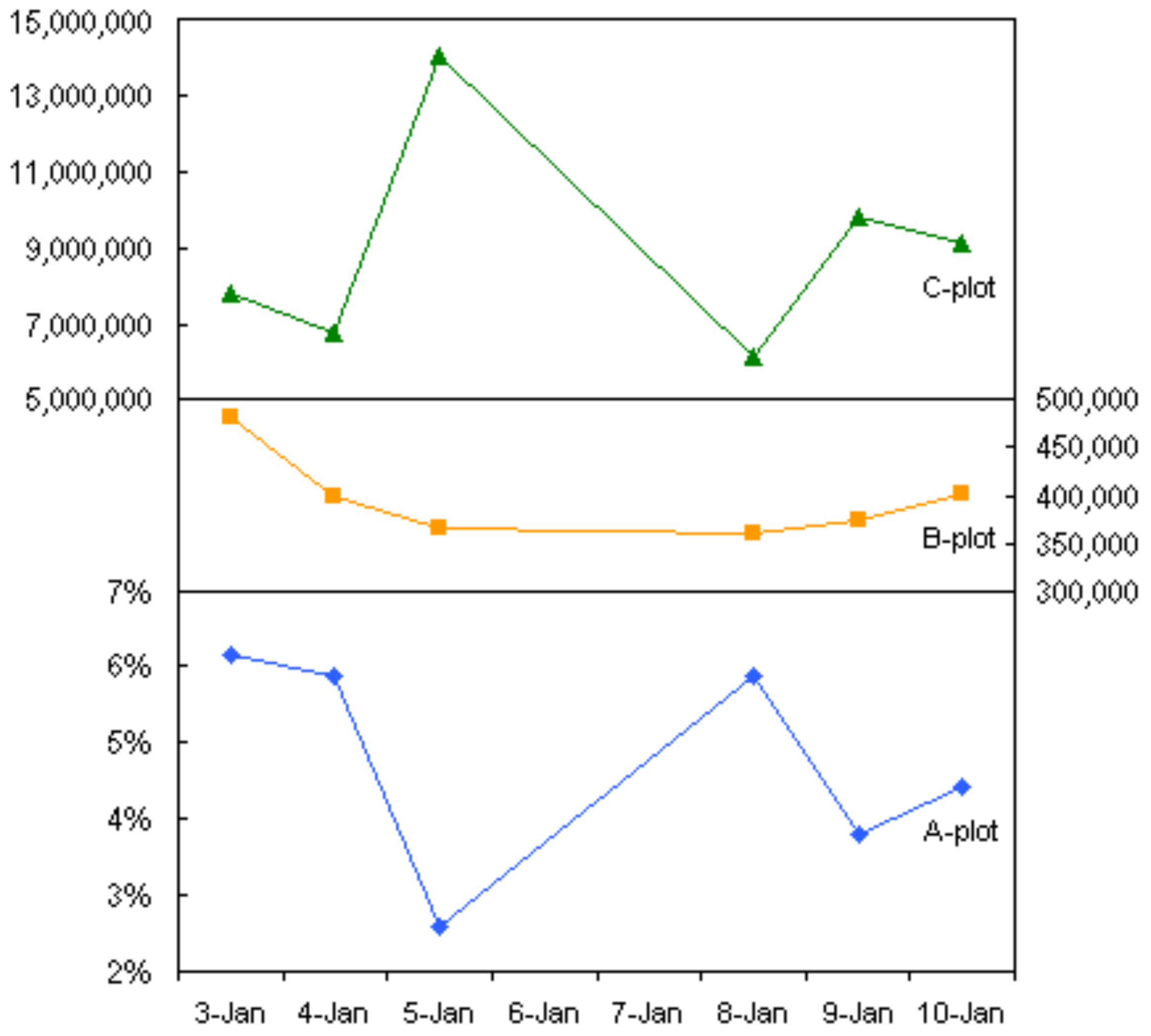


1 of 89 Pgs (1764 Rows)

Label	Project	Date	M/F	Age	Gender	Subject	Hand	YOB	Education	Ses
ResultsSbj01	10CRunsPerSubj		U							
1	ADHD200		U			9956994				
50782592_BWH	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_DUKE	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_DUKE2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_MGH	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_MGH2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCI	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCI2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCSD	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
50782592_UCSD2	Calib	2009-01-01	M	45	male	50782592	right	1964	21	
58259524_BWH	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_DUKE	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_DUKE2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_MGH	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_MGH2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCI	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCI2	Calib	2009-01-01	M	41	male	58259524	right	1968	21	
58259524_UCSD	Calib	2009-01-01	M	41	male	58259524	right	1968	21	







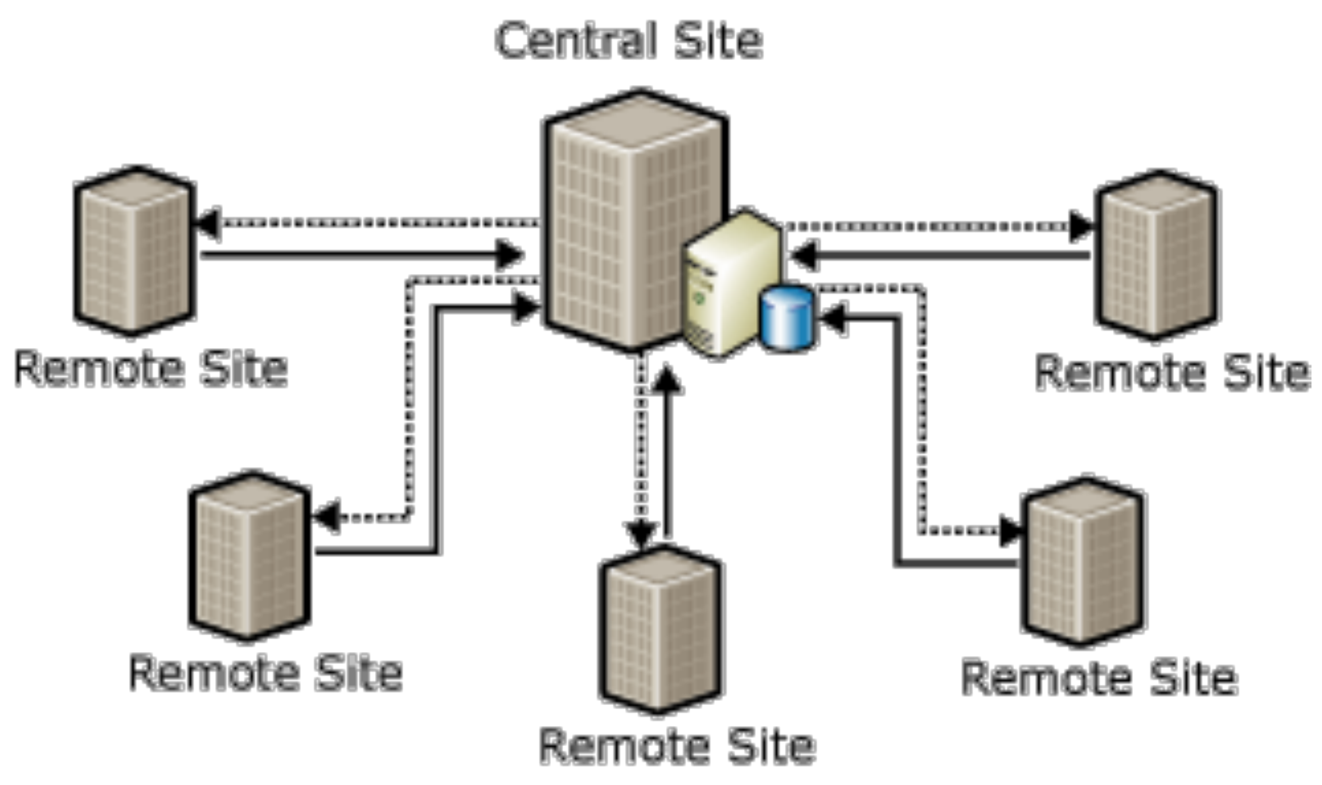
Data analysis can be costly and time-consuming.

Consider adding an analyst to your budget vs. chargeback

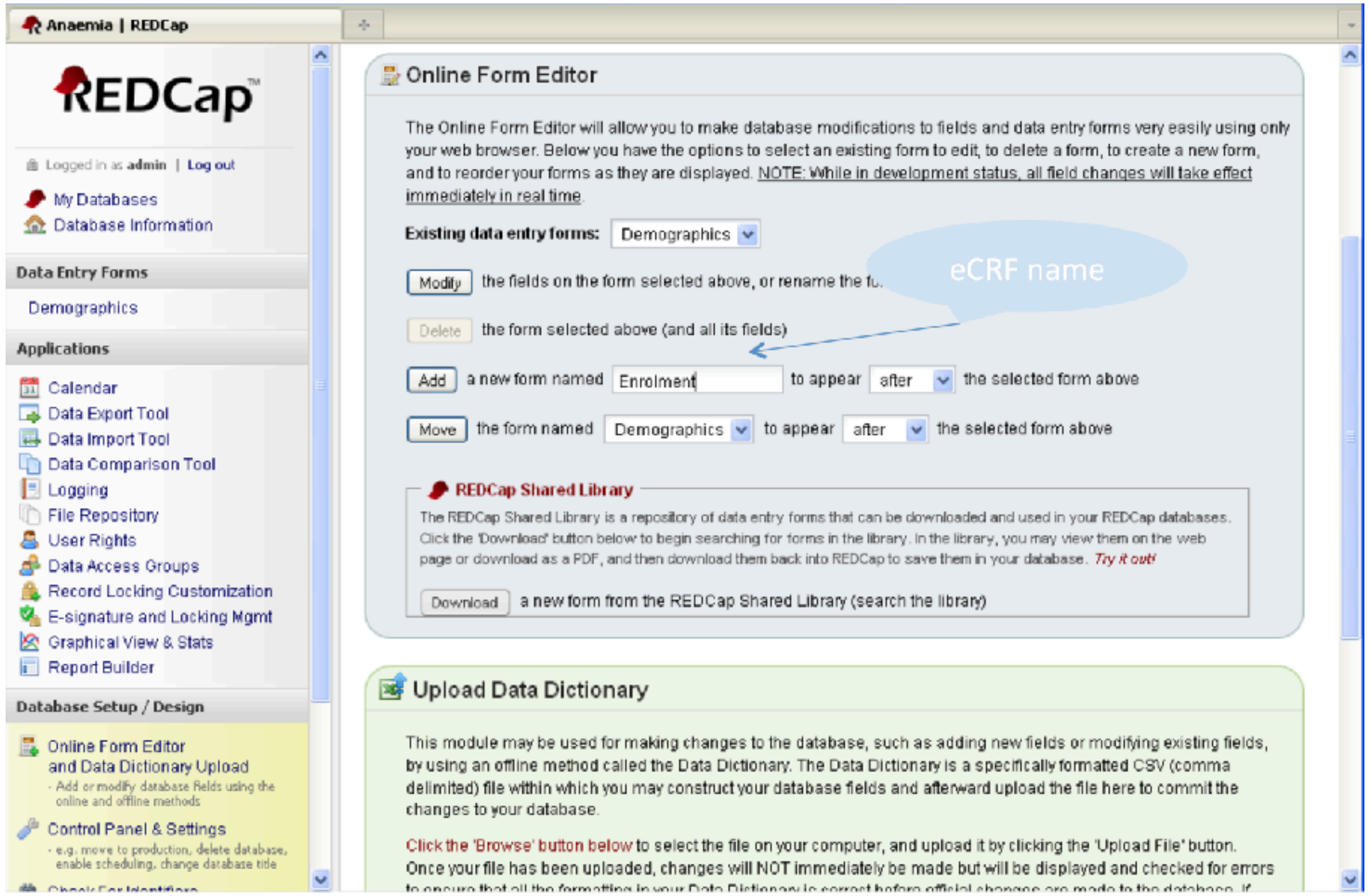
Application Development

Multi-site data collection, transfer, and storage

- If multiple sites enrolling, there are special considerations for IRB, contracts, data use agreements, and application development
- These must be tackled long before your proposal is submitted



REDCap



Online Form Editor

The Online Form Editor will allow you to make database modifications to fields and data entry forms very easily using only your web browser. Below you have the options to select an existing form to edit, to delete a form, to create a new form, and to reorder your forms as they are displayed. NOTE: While in development status, all field changes will take effect immediately in real time.

Existing data entry forms: Demographics

Modify the fields on the form selected above, or rename the fo

Delete the form selected above (and all its fields)

Add a new form named Enrolment to appear after the selected form above

Move the form named Demographics to appear after the selected form above

REDCap Shared Library

The REDCap Shared Library is a repository of data entry forms that can be downloaded and used in your REDCap databases. Click the 'Download' button below to begin searching for forms in the library. In the library, you may view them on the web page or download as a PDF, and then download them back into REDCap to save them in your database. *Try it out!*

Download a new form from the REDCap Shared Library (search the library)

Upload Data Dictionary

This module may be used for making changes to the database, such as adding new fields or modifying existing fields, by using an offline method called the Data Dictionary. The Data Dictionary is a specifically formatted CSV (comma delimited) file within which you may construct your database fields and afterward upload the file here to commit the changes to your database.

Click the 'Browse' button below to select the file on your computer, and upload it by clicking the 'Upload File' button. Once your file has been uploaded, changes will NOT immediately be made but will be displayed and checked for errors

HIPAA-compliant data collection and storage

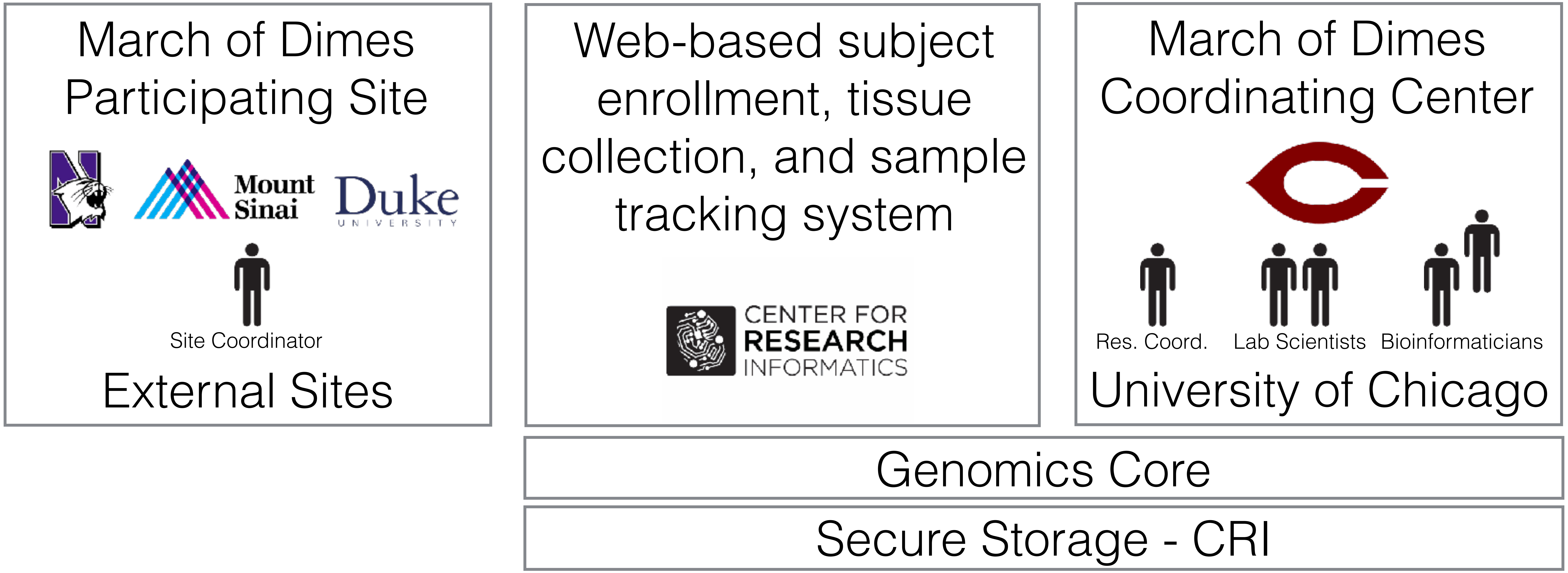
REDCap

- It's easy to get a REDCap account - and it's free
- Non-BSD collaborators will need BSD accounts - and this can take time (start early)
- Most form generation can be performed by the investigators
- CRI helps with complex forms and other needs
- We can help with boilerplate grant language for REDCap

Custom application development / programming

- Do you need a website?
- How about a customized platform for data collection?
- Online tools?
- The CRI can build anything you need, but there must be budget for programmer costs
- We can help estimate the budget and write up the relevant parts of the proposal

March of Dimes Prematurity Research Center



March of Dimes Participating Site

Site Coordinator

External Sites

Web-based subject enrollment, tissue collection, and sample tracking system

March of Dimes Coordinating Center

Res. Coord.

Lab Scientists

Bioinformaticians

University of Chicago

Genomics Core

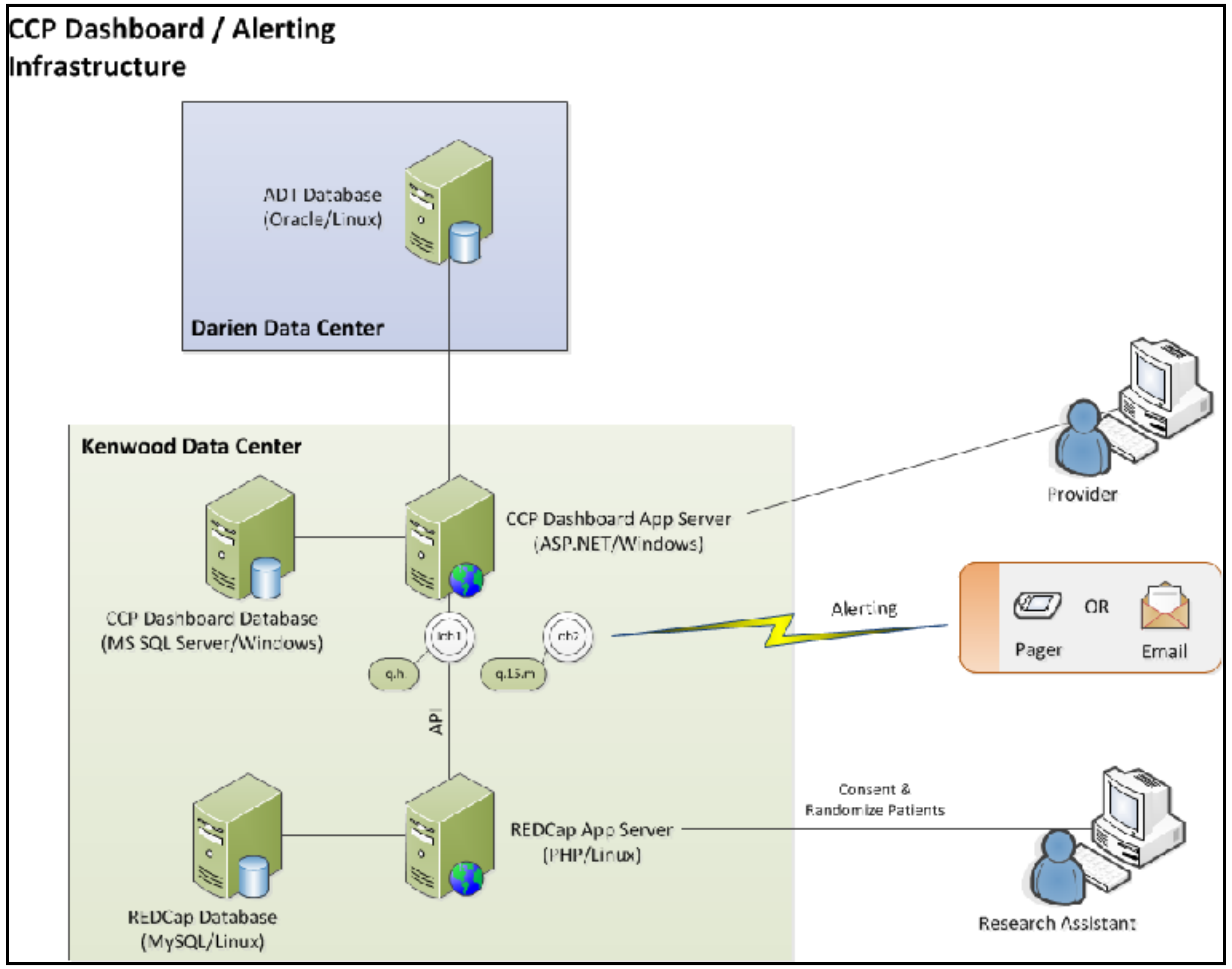
Secure Storage - CRI

1200 Patients GPS

- Web application developed for UChicago pharmacogenomics study
- Interfaces with PubMed and CRI survey engines
- Produces a patient-specific eligibility list on a per clinic level
- Patient enrollment directly in application

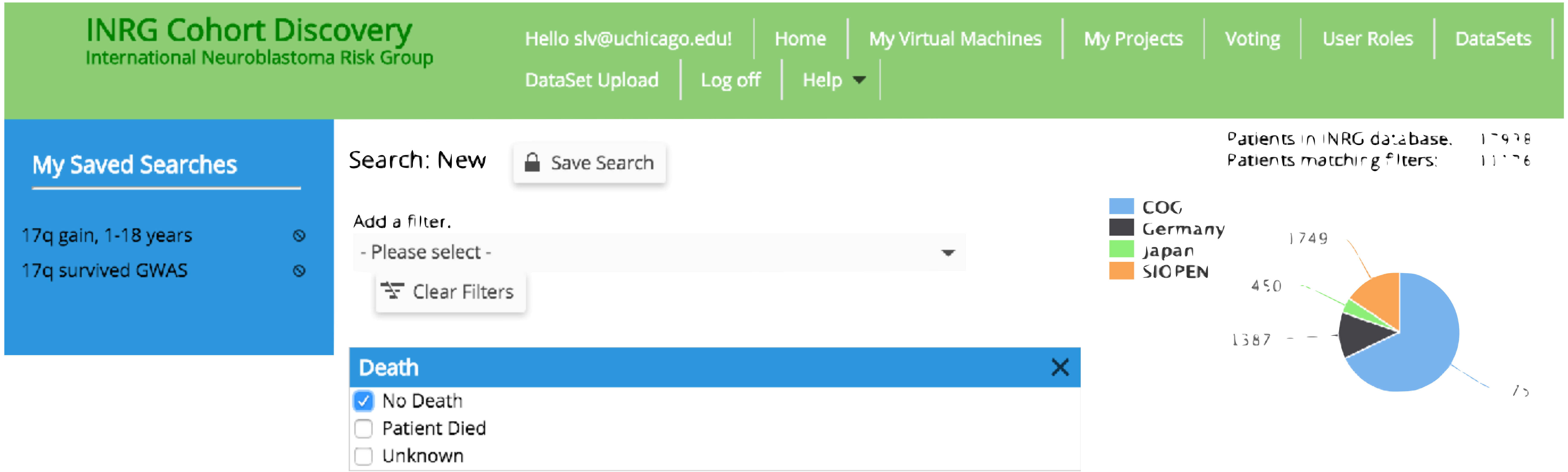
Comprehensive Care Program

- CRI developed applications to support enrollment and management of patients to this multi-arm CMMS-sponsored UChicago study
- Web application-enabled aggregation of data from multiple sources (REDCap, Epic Clarity, Centricity Billing, ADT)
- Alerting infrastructure built to notify providers of patient events


CCP Dashboard / Alerting Infrastructure
ADT Database (Oracle/Linux)
Darien Data Center
Kenwood Data Center
CCP Dashboard Database (MS SQL Server/Windows)
CCP Dashboard App Server (ASP.NET/Windows)
Job1
Job2
q.h
q.15.m
API
Alerting
Provider
Pager
OR
Email
REDCap Database (MySQL/Linux)
REDCap App Server (PHP/Linux)
Consent & Randomize Patients
Research Assistant

International Neuroblastoma Risk Group Database

- Web portal for cohort discovery
- International data governance
- Strict auditing of data revisions



INRG Cohort Discovery
International Neuroblastoma Risk Group

Hello slv@uchicago.edu! | Home | My Virtual Machines | My Projects | Voting | User Roles | DataSets | DataSet Upload | Log off | Help

My Saved Searches

17q gain, 1-18 years
17q survived GWAS

Search: New
Save Search

Add a filter.
- Please select -
Clear Filters

Death
☑ No Death
☐ Patient Died
☐ Unknown

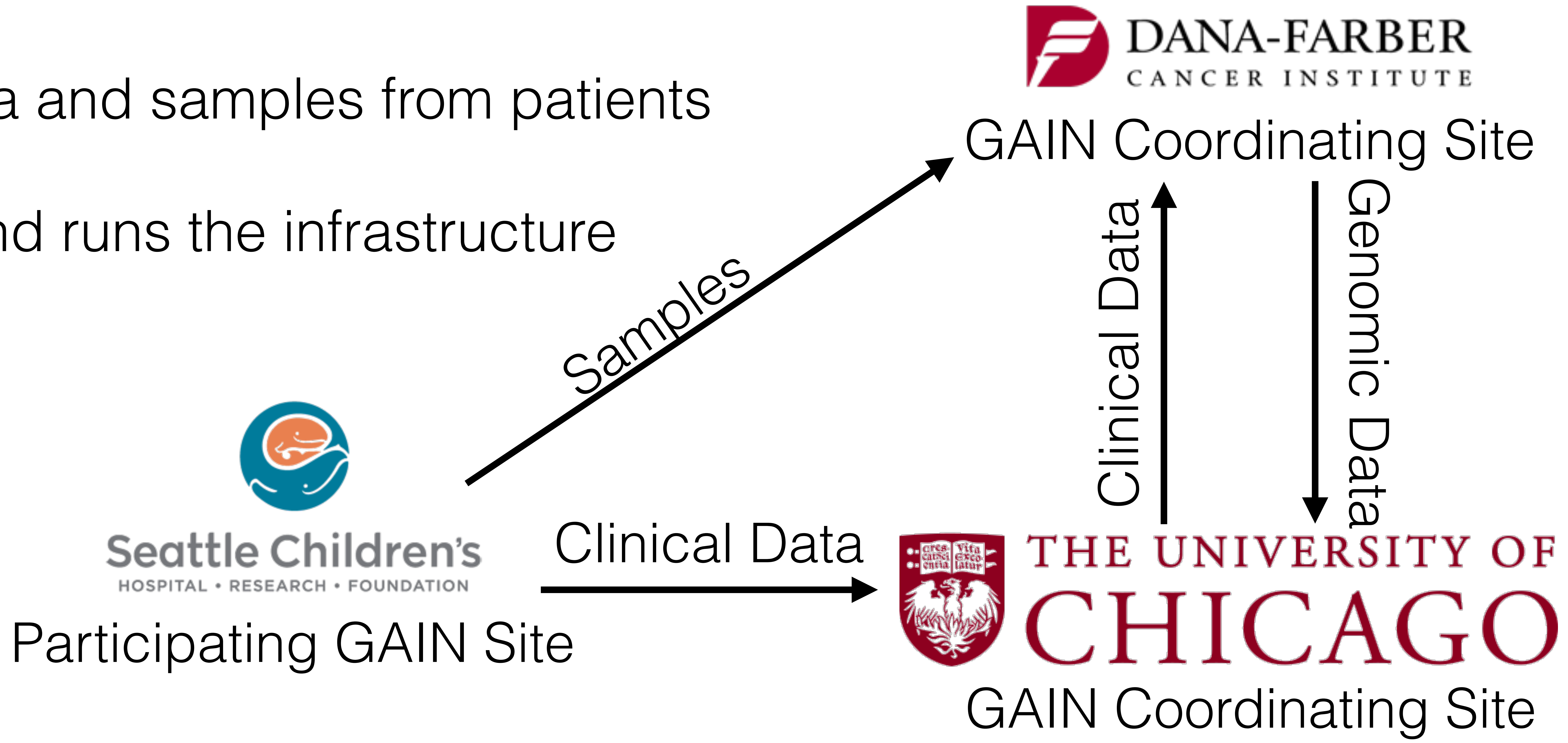
Patients in INRG database. 17928
Patients matching filters: 11176

Pediatric GAIN Project

- Multi-center (20+) pediatric cancer trial
- Collect data and samples from patients
- CRI built and runs the infrastructure

Pediatric GAIN Project

- Multi-center (20+) pediatric cancer trial
- Collect data and samples from patients
- CRI built and runs the infrastructure



DANA-FARBER CANCER INSTITUTE

GAIN Coordinating Site

Samples

Clinical Data

Genomic Data

Seattle Children's HOSPITAL • RESEARCH • FOUNDATION

Participating GAIN Site

Clinical Data

THE UNIVERSITY OF CHICAGO

GAIN Coordinating Site

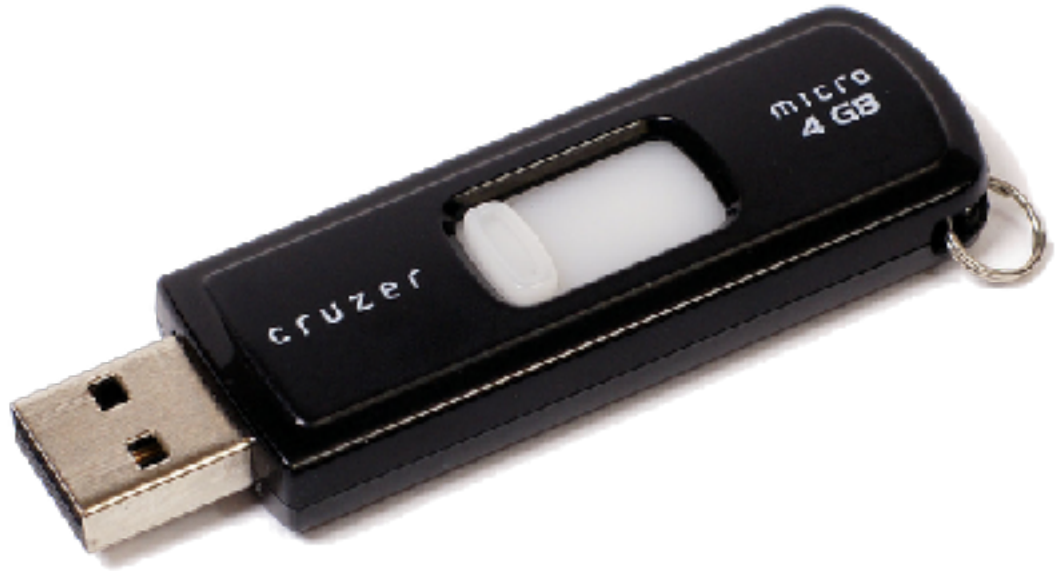
Systems and infrastructure

Systems - Offsite access

- Do researchers outside of UChicago need access to your data?
- Collaborator accounts take time to obtain - and the CRI can help

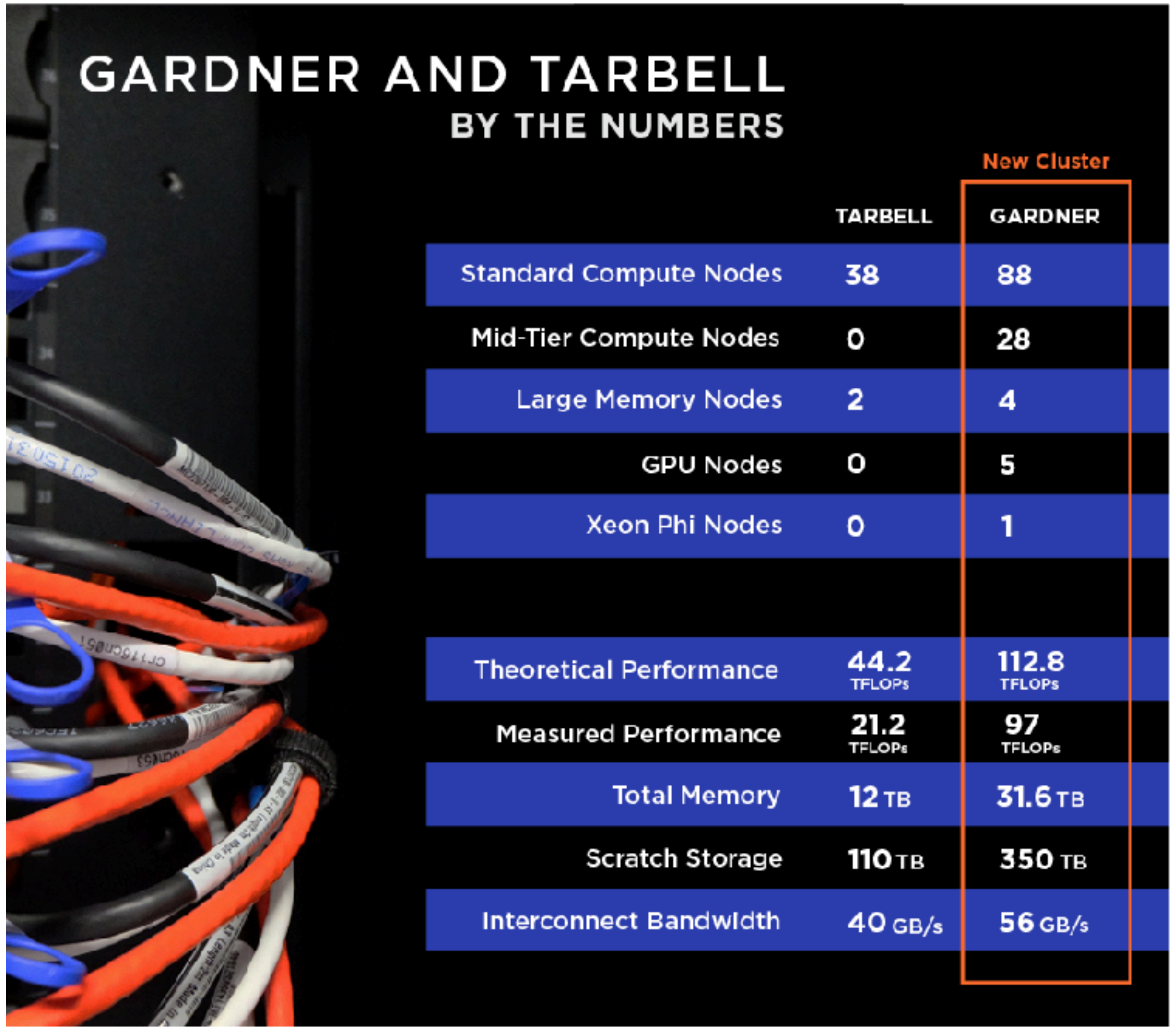
Systems - Growing / flexible storage needs

- Some projects do not require much storage in the beginning, but needs grow
- Consider the entire project, not just the first year when crafting the budget for systems



Systems - HPC

- There are many options for HPC. CRI has one of the biggest and fastest clusters on campus
- CRI also has dedicated support for helping your prepare your grant and complete your research



GARDNER AND TARBELL
BY THE NUMBERS

New Cluster

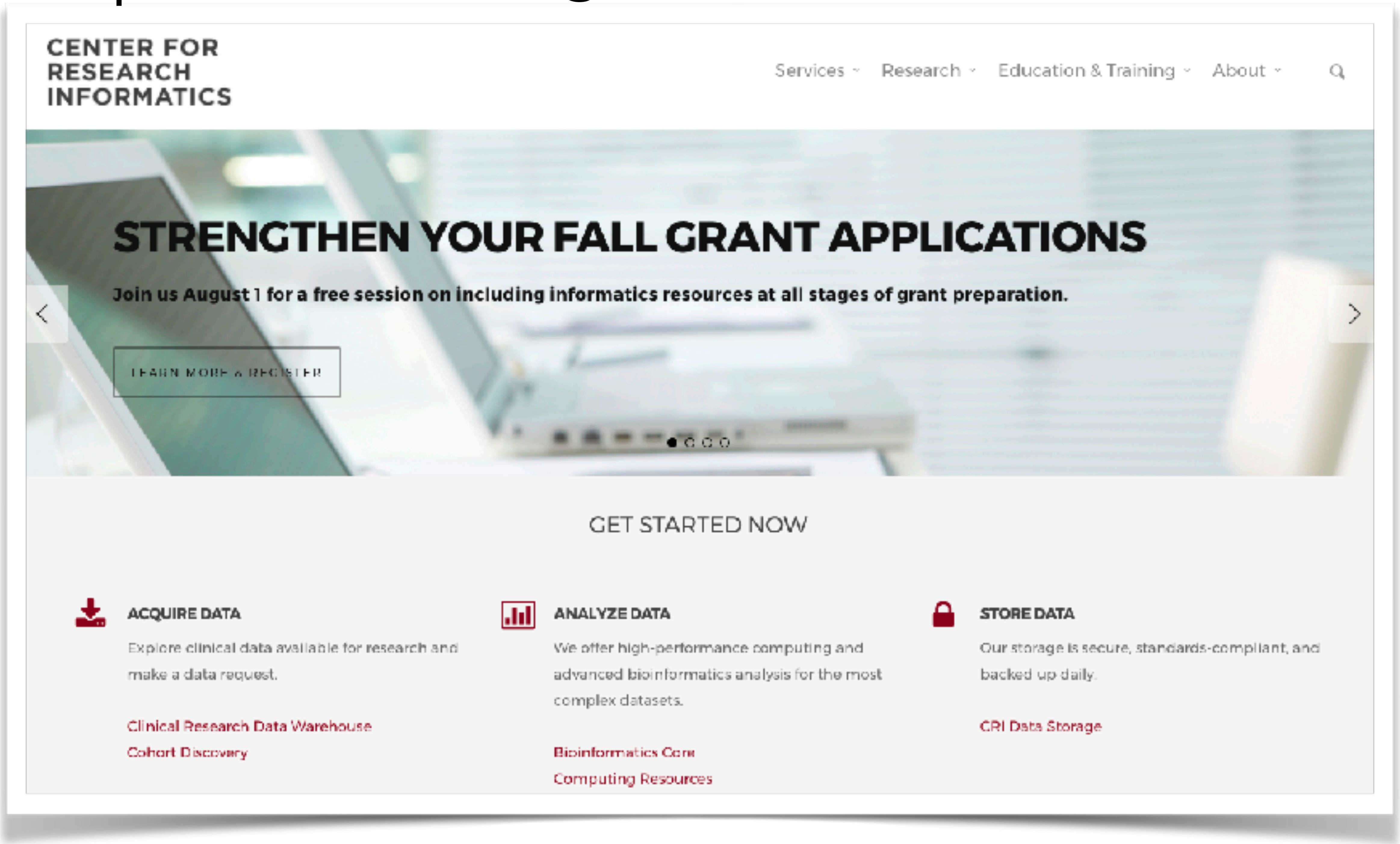
	TARBELL	GARDNER
Standard Compute Nodes	38	88
Mid-Tier Compute Nodes	0	28
Large Memory Nodes	2	4
GPU Nodes	0	5
Xeon Phi Nodes	0	1
Theoretical Performance	44.2 TFLOPs	112.8 TFLOPs
Measured Performance	21.2 TFLOPs	97 TFLOPs
Total Memory	12 TB	31.6 TB
Scratch Storage	110 TB	350 TB
Interconnect Bandwidth	40 GB/s	56 GB/s

Systems - Virtual machines / servers

- Setting up and maintaining VMs is expensive
- CRI will help you develop your budget
- This is commonly left out of grant applications/budgets

Ways to get help

http://cri.uchicago.edu



CENTER FOR RESEARCH INFORMATICS

STRENGTHEN YOUR FALL GRANT APPLICATIONS

Join us August 1 for a free session on including informatics resources at all stages of grant preparation.

LEARN MORE & REGISTER

GET STARTED NOW

ACQUIRE DATA

Explore clinical data available for research and make a data request.

Clinical Research Data Warehouse
Cohort Discovery

ANALYZE DATA

We offer high-performance computing and advanced bioinformatics analysis for the most complex datasets.

Bioinformatics Core
Computing Resources

STORE DATA

Our storage is secure, standards-compliant, and backed up daily.

CRI Data Storage

Sam Volchenboum
Director

Michael Baltasi
Deputy Director

Brian Furner
Applications

Thorbjorn Axelsson
Systems

Jorge Andrade
Bioinformatics

Tim Holper
Data Warehouse

Julissa Acevedo
REDCap

Julie Johnson
Data Warehouse

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES | CENTER FOR RESEARCH INFORMATICS

University of Chicago

Center for Research Informatics

Applications - Systems - Bioinformatics - Data Warehousing

Questions?