

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.

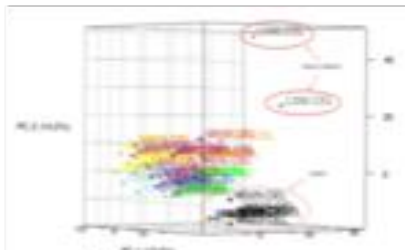


Fig. 2. Illustration of the outliers and batch effects in the COG gene expression data

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.

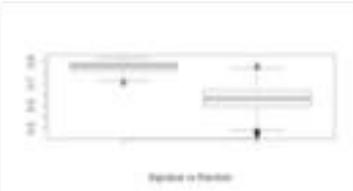


Fig. 3. Boxplot of AUCs between 5-gene candidate signatures and random gene sets

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
CTSL,IGF1R,CDH15,SLC29A1	0.925	0.929	1.000	0.900
CTSC,ANPEP,TGFA,DCBLD1	0.913	0.857	1.000	0.800
SLC29A1,CTLT,SPAN15,DDIT3,MR2	0.925	0.857	1.000	0.800
DCBLD1,GSTM2,LYN,RAF1,IFIT5	0.925	0.857	0.750	0.900
ANPEP,C1orf162,LOC100132167,NR1H3,PLEK	1.000	1.000	1.000	1.000
CTSC,DCBLD1,IGF1R,TEF,CCL18,SLC29A1	1.000	0.929	1.000	0.900
LOC100132167,HEY2,TP53,IL13L	0.950	0.929	1.000	0.900
C29A1,IGF1R,VWF,IL6,MR2,CCL18,GSTM2,TP53,CTSC,DDIT3,IGF1R,SLC29A1,TGFA,TEF,HEY2,IGF1R,RAF1,IFIT5	1.000	1.000	1.000	1.000
EMR2,NR1H3,CCL18,SLC29A1,CAM,LOC100132167,PODXL,NKAIN1,HSX1S,IL6,ANPEP,GSTM2,TEF	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).

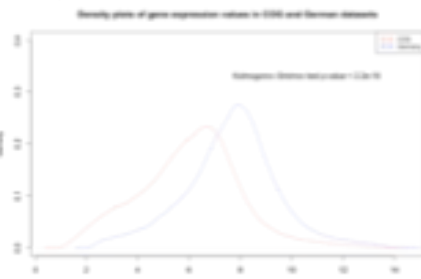


Fig. 4. Density plots of gene expression data in the COG and Germany datasets.

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

models from the training data set might be over-fitted. Our analysis confirmed 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.

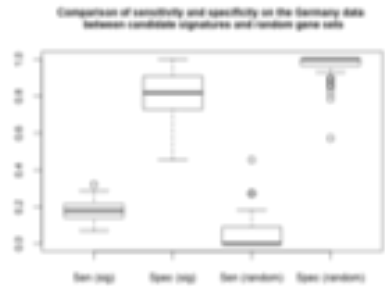


Fig.5. Boxplot of sensitivity and specificity in the prediction of the survival status on the Germany data set using 5-gene candidate signatures and random gene sets. Left two are sensitivity and specificity of the candidate signatures and right two are sensitivity and specificity of the random sets.

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.

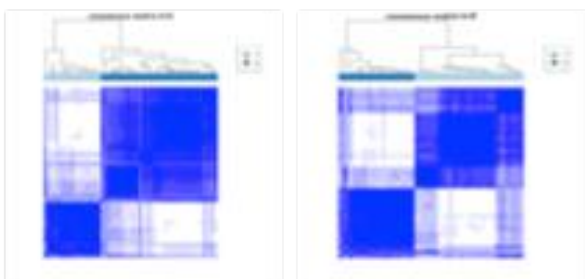


Fig.6. Consensus clustering using gene expression data from the COG data set (left) and the Germany data set (right).

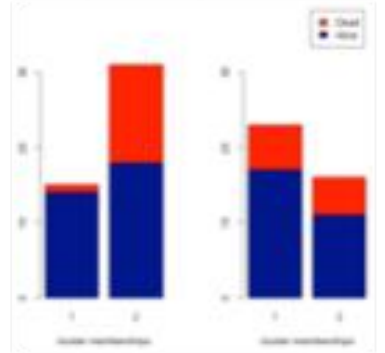


Fig.7. Stacked barplots for the distribution of survival status between consensus clusters. Left: COG data set; right: Germany data set.

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

CRDW - Special considerations

- Preliminary data
- Cohort definition
- Data element identification
- Aggregation / normalization
- Analysis and interpretation

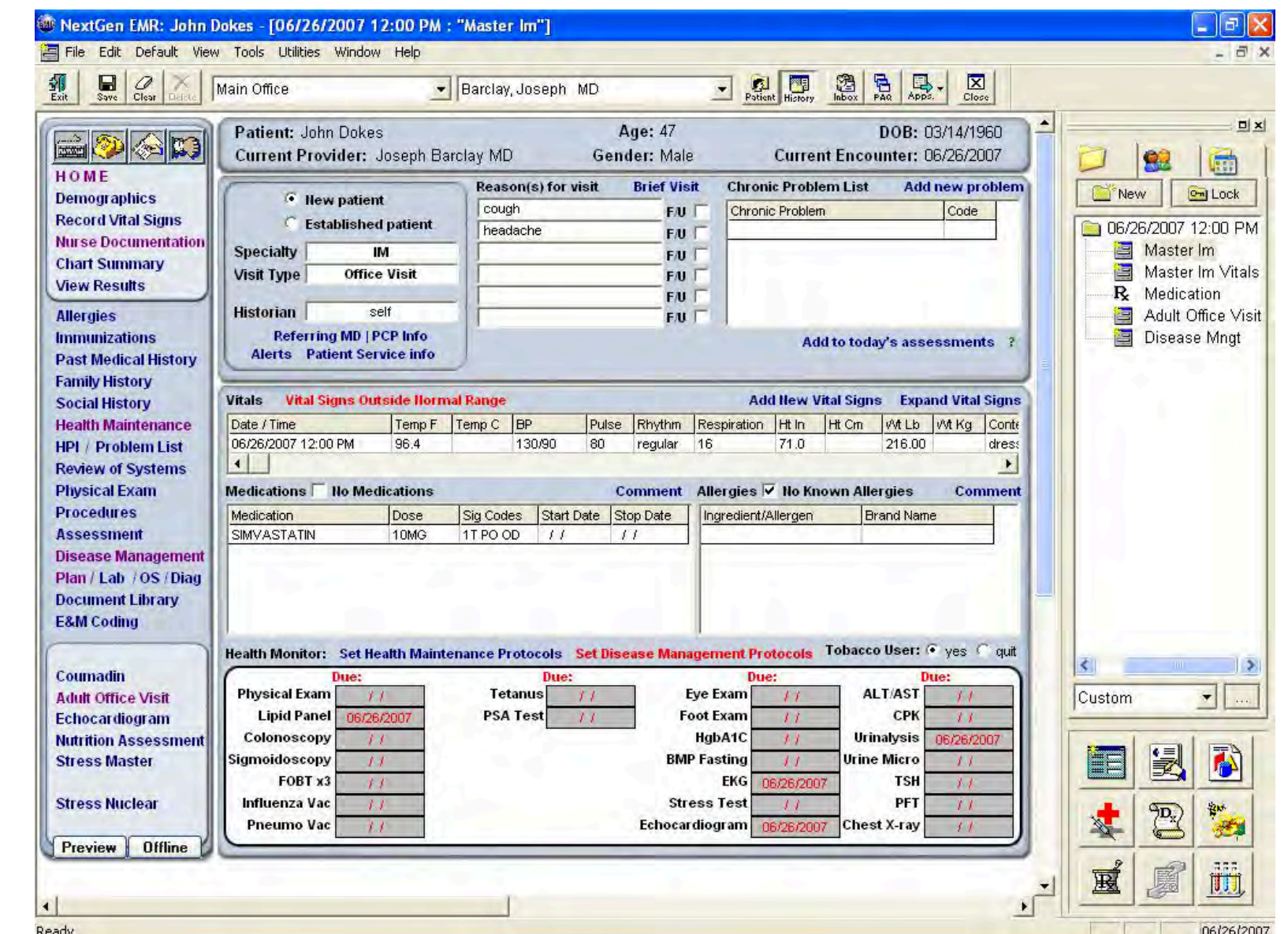


CRDW - Preliminary data / Cohort identification

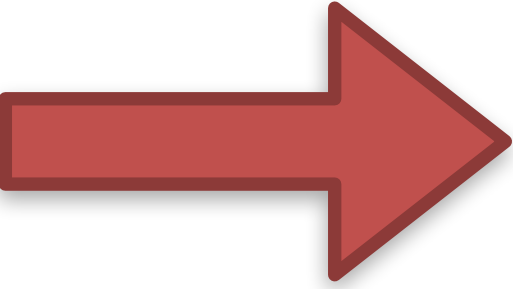
- It can be hard to identify cohorts
- CRI has specialists to help
- Reviewers like to see preliminary identification of cohorts - “Can they really get the data”
- Sometimes new data has to be sourced

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



CRDW - Data aggregation and normalization



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

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

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

Search 1										
<< first < prev 1 2 3 4 5 next > last >> 20 1 of 89 Pgs (1764 Rows) Reload Options										
Label	Project	Date	M/F	Age	Gender	Subject	Hand	YOB	Education	Ses
ResultsSbj01	100RunsPerSubj		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

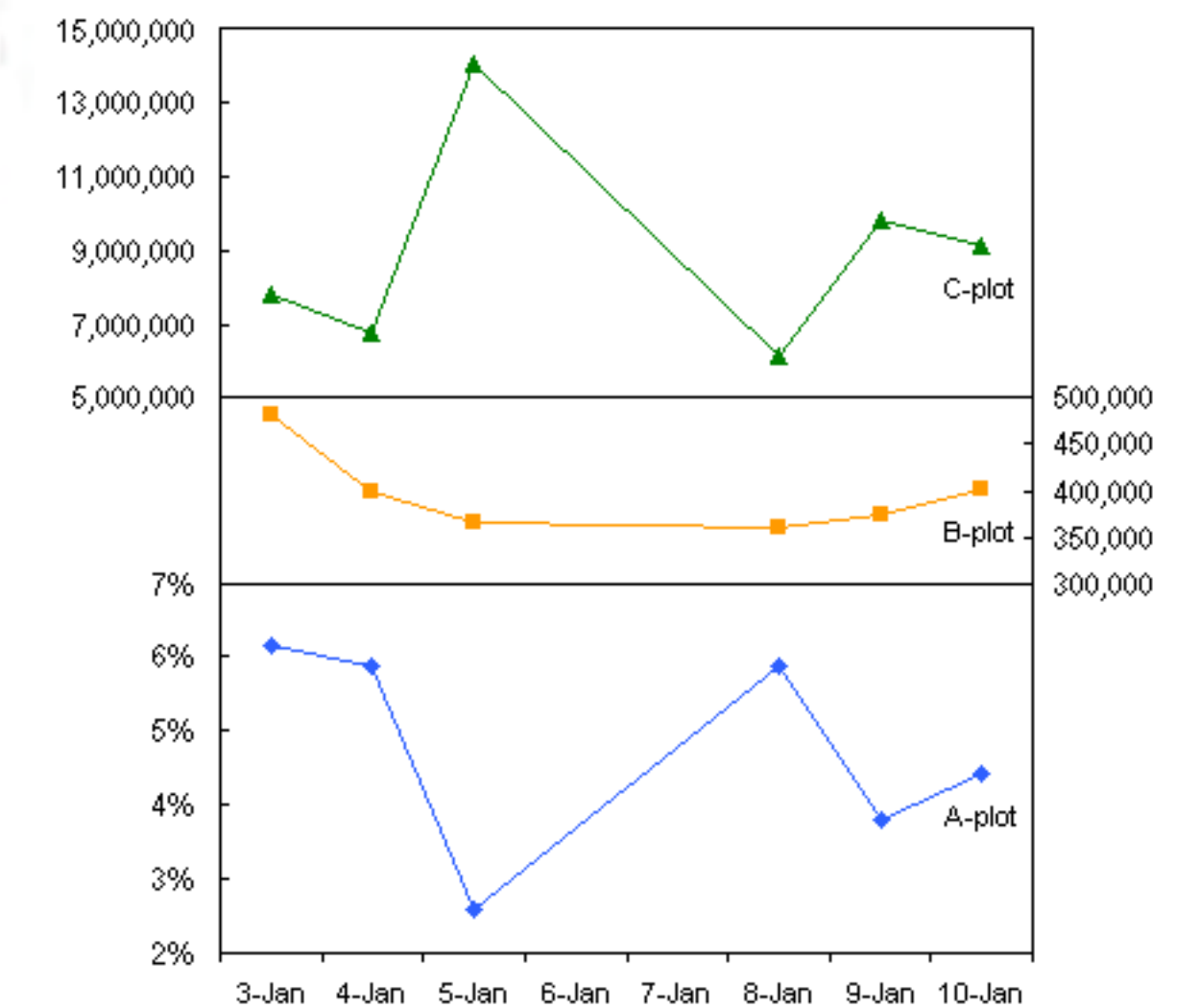
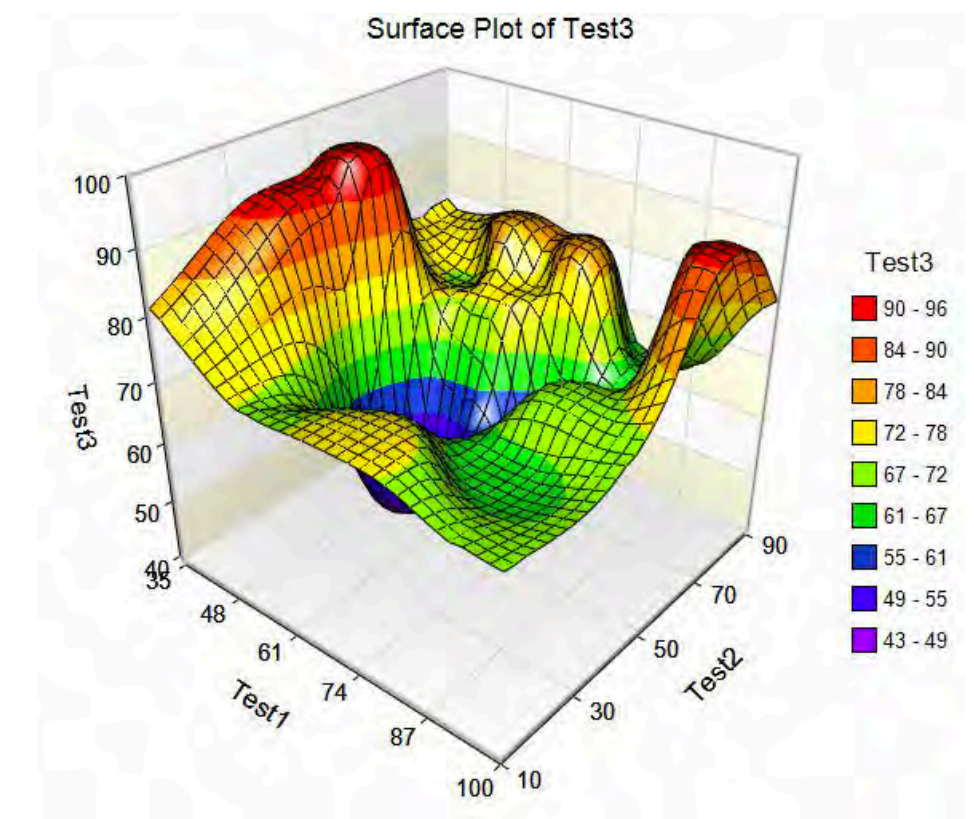
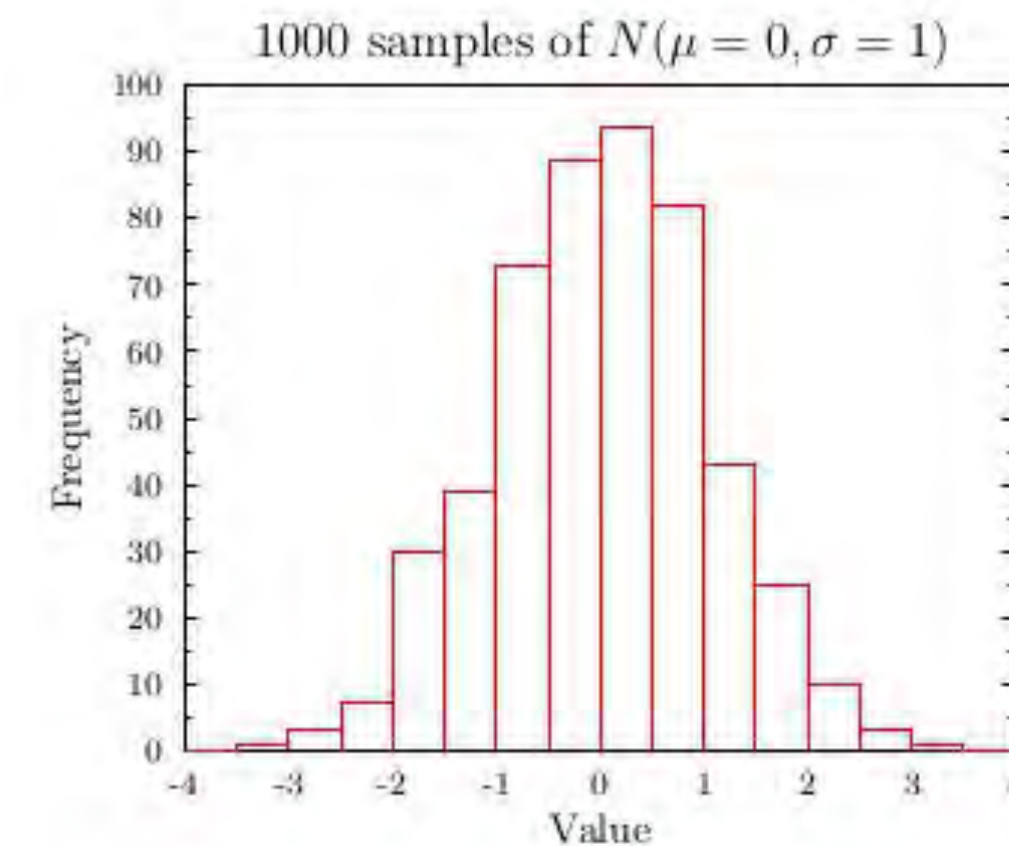
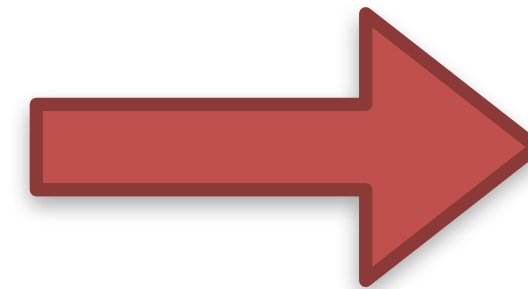


CRDW - Data analysis and interpretation

Search 1 X

<< first < prev 1 2 3 4 5 next > last >> 20 1 of 89 Pgs (1764 Rows) Reload Options

Label	Project	Date	M/F	Age	Gender	Subject	Hand	YOB	Education	Ses
ResultsSbj01	100RunsPerSubj		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	

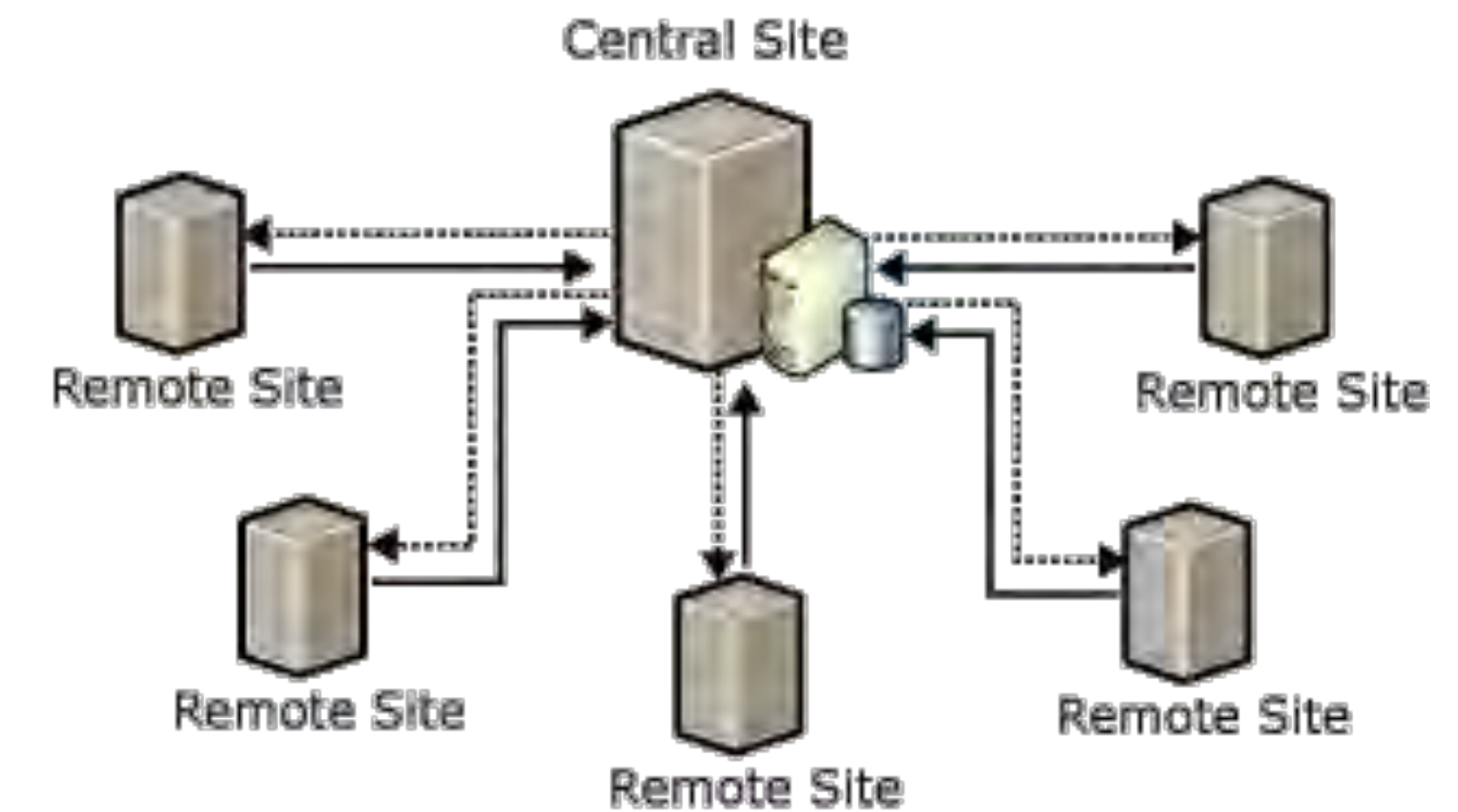


Data analysis can be costly and time-consuming.
Consider adding an analyst to your budget vs. chargeback



Applications - Special considerations

- Multi-site data collection, transfer, and storage
- REDCap usage
- Application development and programming

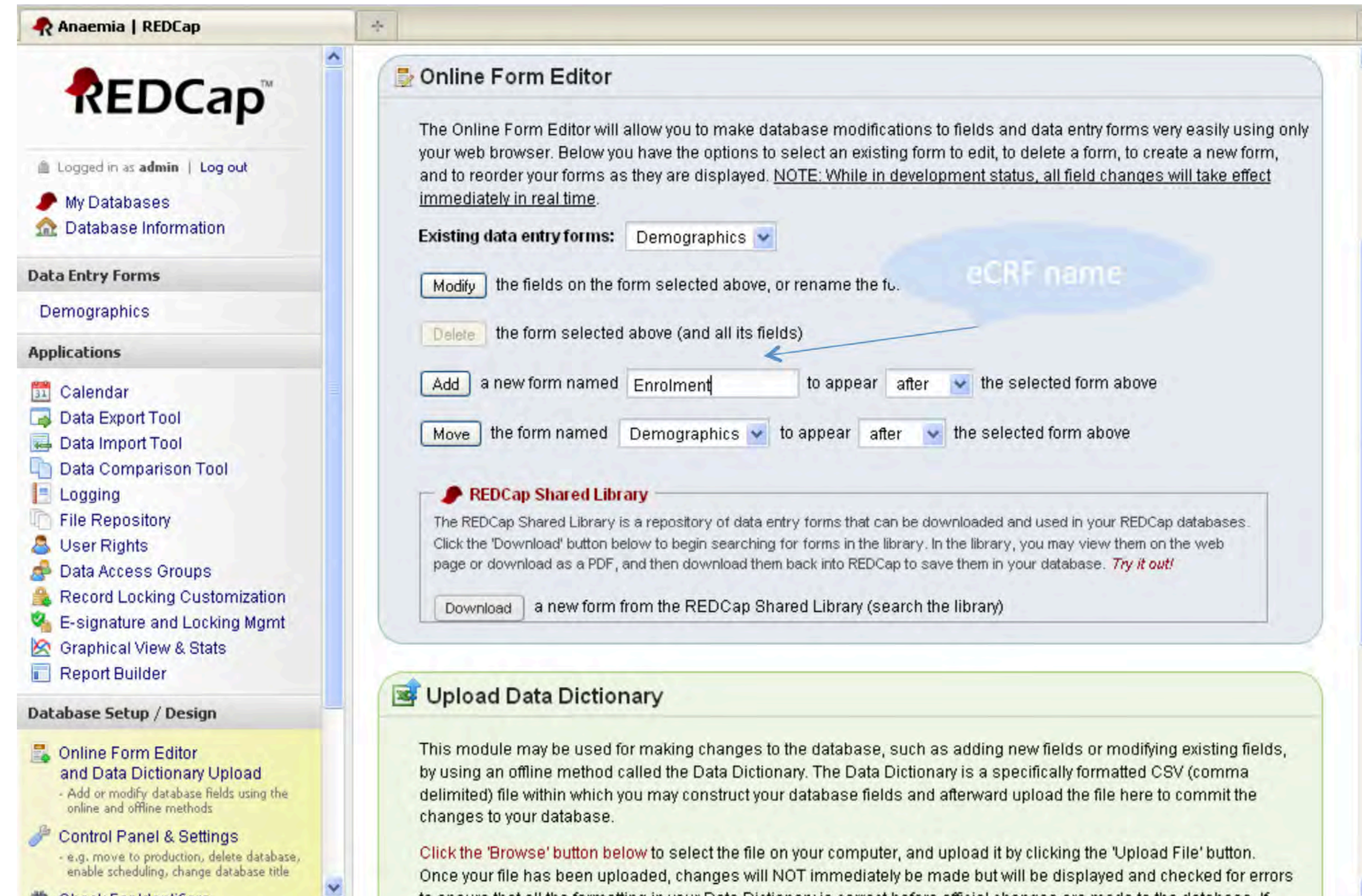


Applications - Multi-site data collection, transfer, and storage

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



REDCap



HIPAA-compliant data collection and storage



Applications - 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



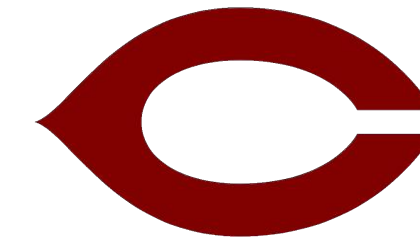
Site Coordinator

External Sites

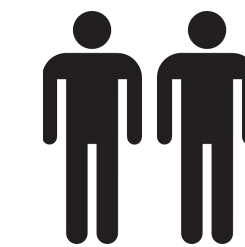
Web-based subject
enrollment, tissue
collection, and sample
tracking system



March of Dimes
Coordinating Center



Res. Coord.



Lab Scientists



Bioinformaticians

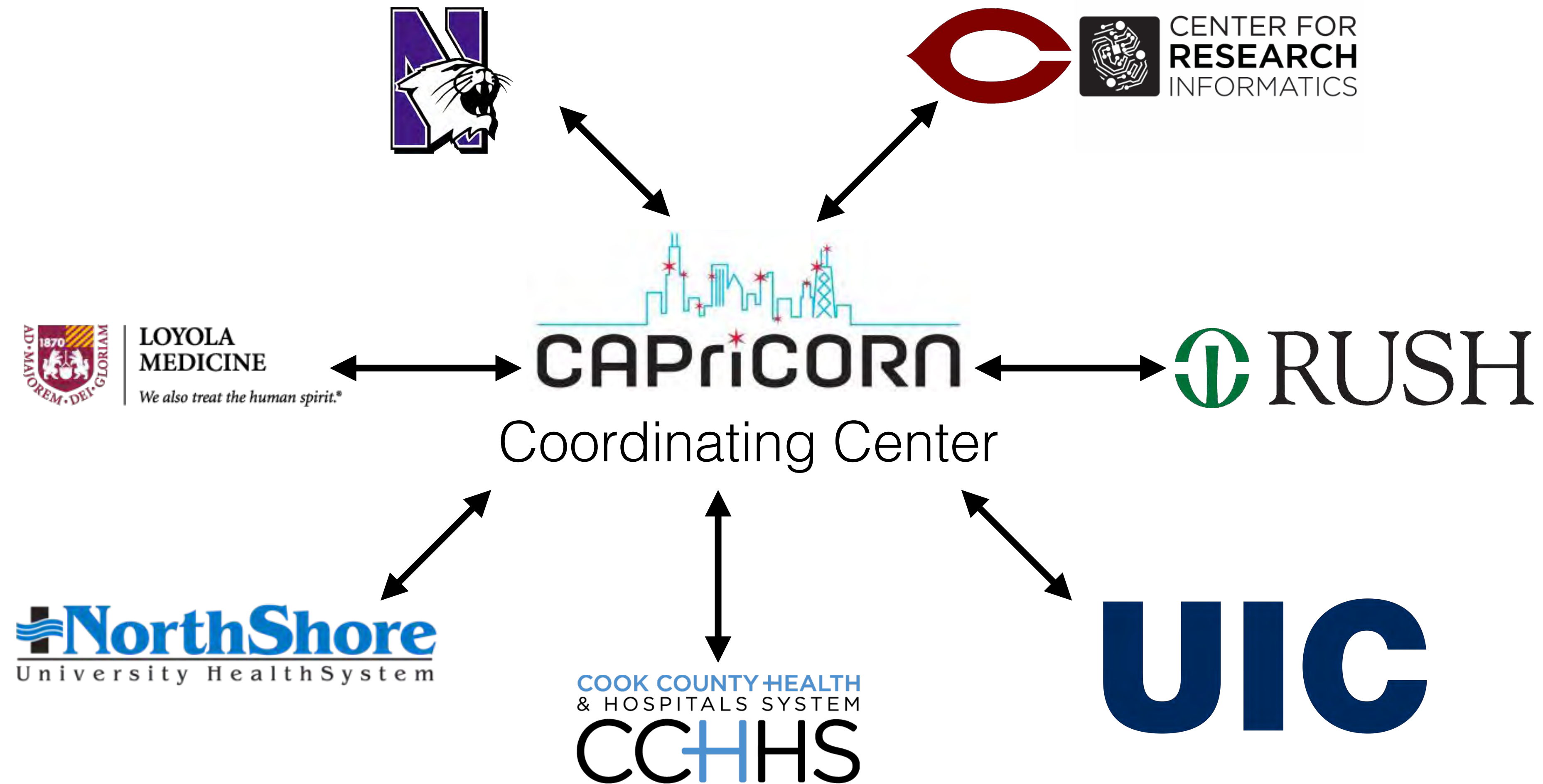
University of Chicago

Genomics Core

Secure Storage - CRI



CAPriCORN



Thirty Million Words



- Took over development from vendor
- Providing HIPAA compliant storage and analytics space
- Hosting the production of the analytic data sets



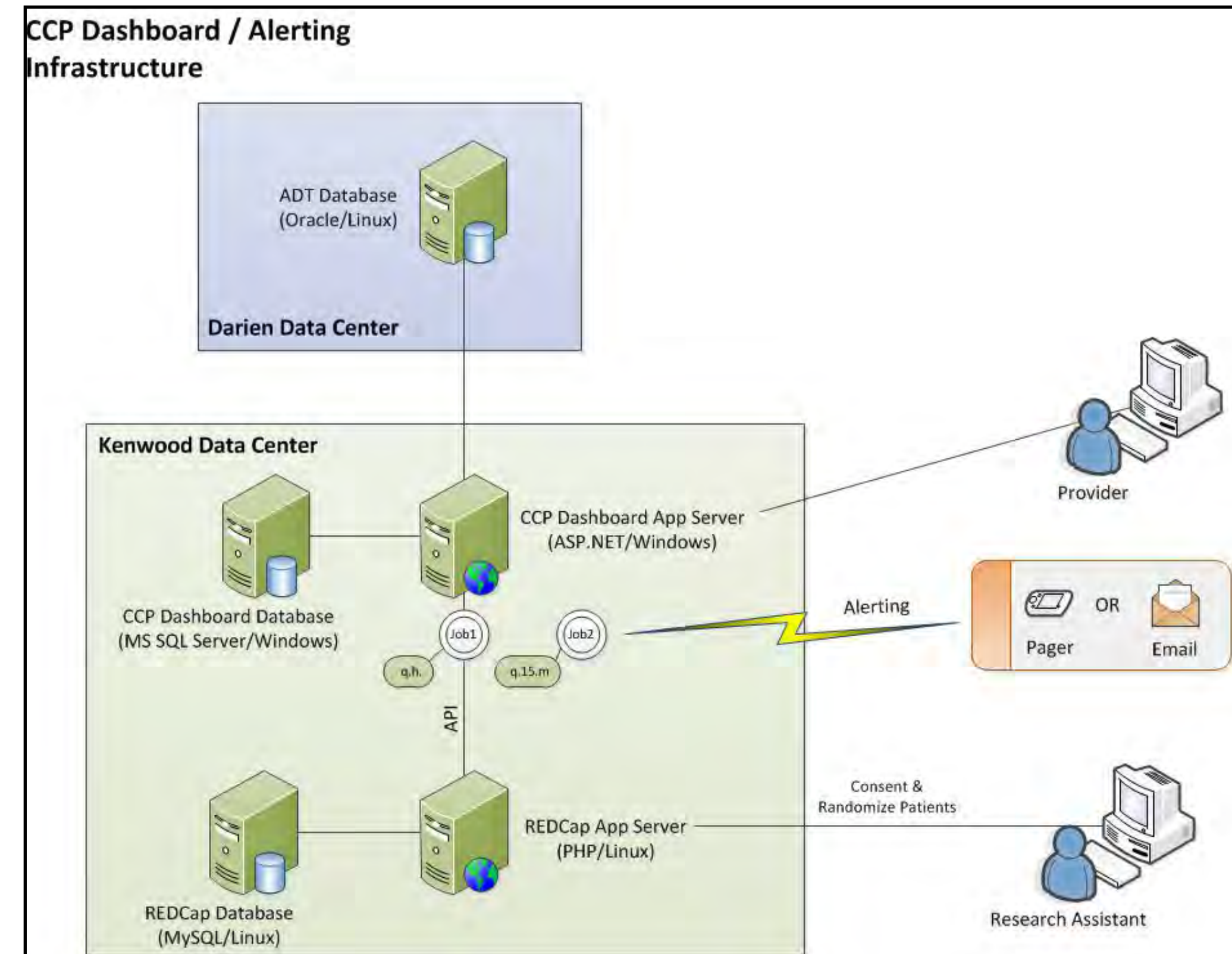
1200 Patients GPS

- 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



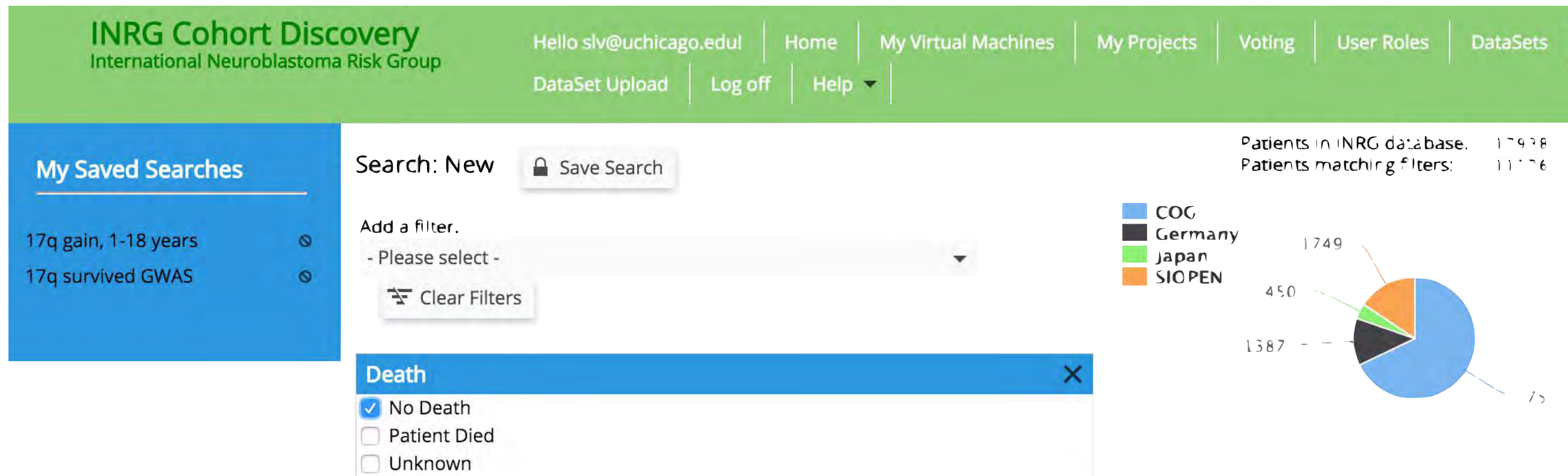
Comprehensive Care Program

- 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



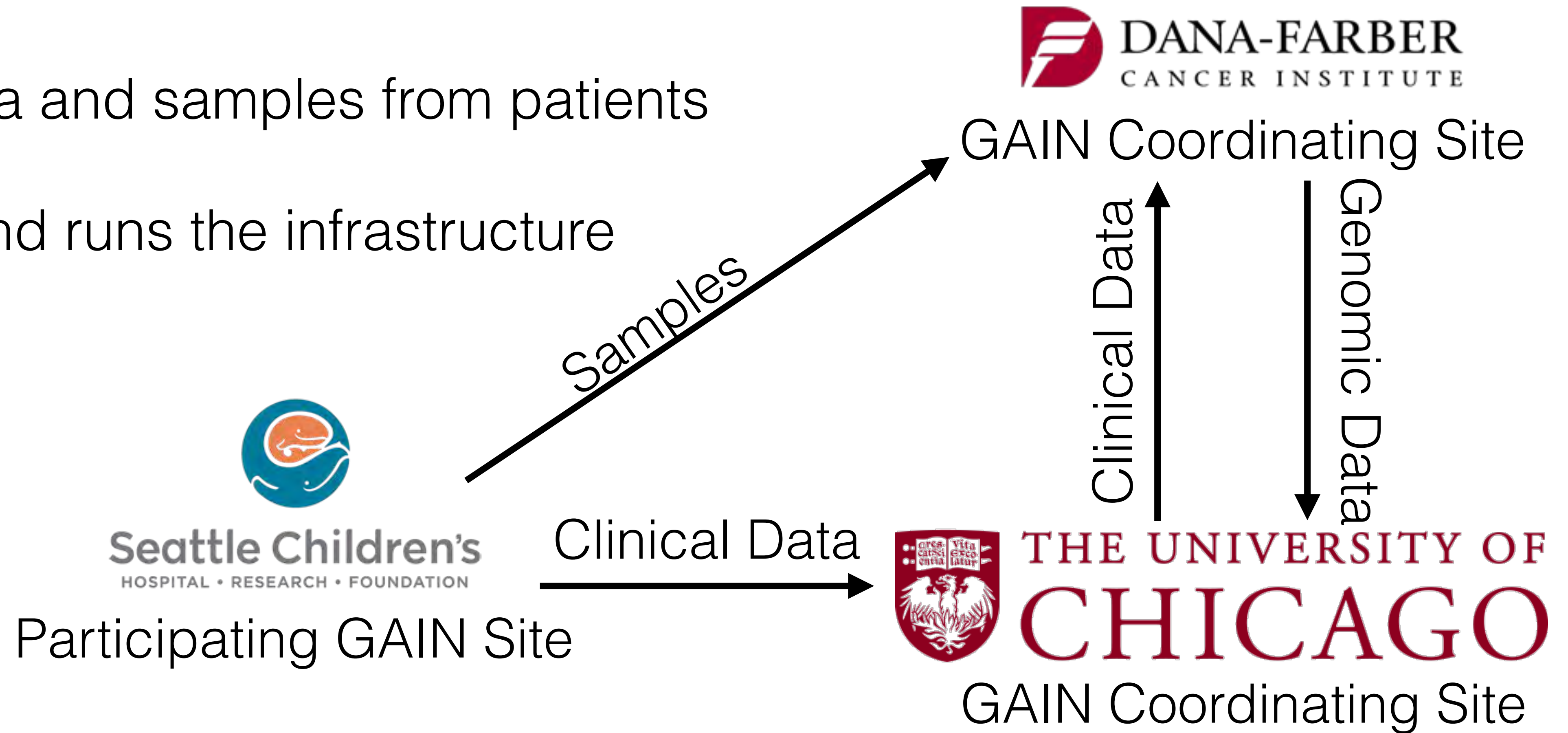
International Neuroblastoma Risk Group Database

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



Pediatric GAIN Project

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



Systems and infrastructure - special considerations

- Off-site access
- Flexible / growing storage needs
- HPC access
- HPC consulting
- Virtual machines / servers



Systems - Offsite access

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

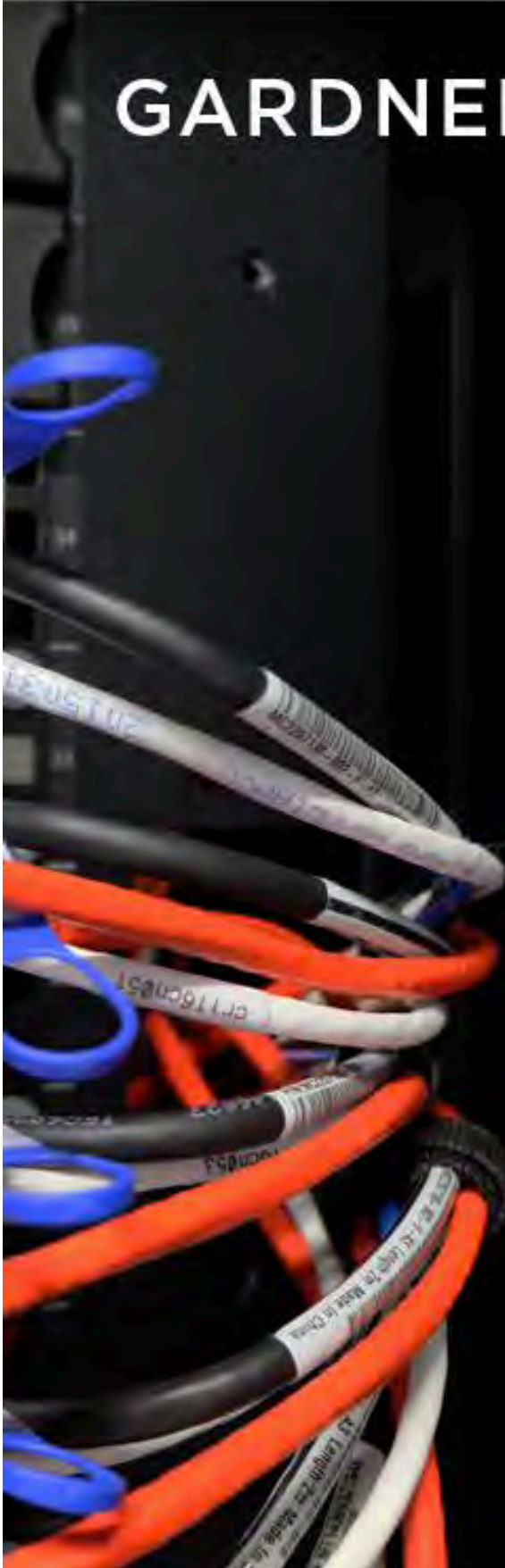
Systems - Growing / flexible storage needs

- 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



	TARBELL	New Cluster 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



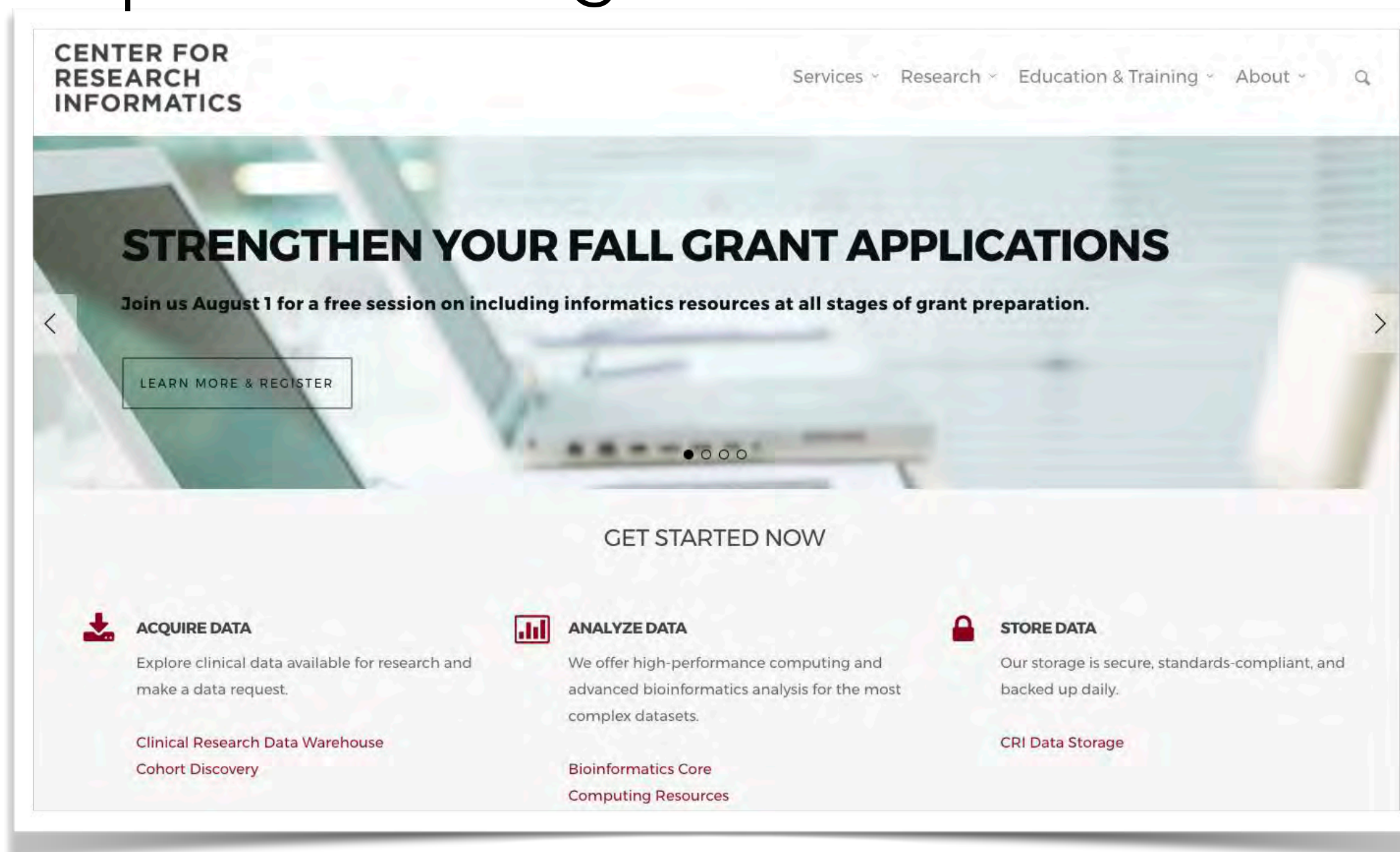
Systems - Virtual machines / servers

- 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>



The screenshot shows the homepage of the Center for Research Informatics (CRI). The header includes the CRI logo and navigation links for Services, Research, Education & Training, and About. The main banner features the text "STRENGTHEN YOUR FALL GRANT APPLICATIONS" and "Join us August 1 for a free session on including informatics resources at all stages of grant preparation." Below the banner is a "GET STARTED NOW" button. The page is divided into three columns: "ACQUIRE DATA" (Clinical Research Data Warehouse, Cohort Discovery), "ANALYZE DATA" (Bioinformatics Core, Computing Resources), and "STORE DATA" (CRI Data Storage).



Questions?



THE UNIVERSITY OF
CHICAGO MEDICINE &
BIOLOGICAL SCIENCES



CENTER FOR
RESEARCH
INFORMATICS