Planning for informatics in your grant applications

August 3, 2018

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

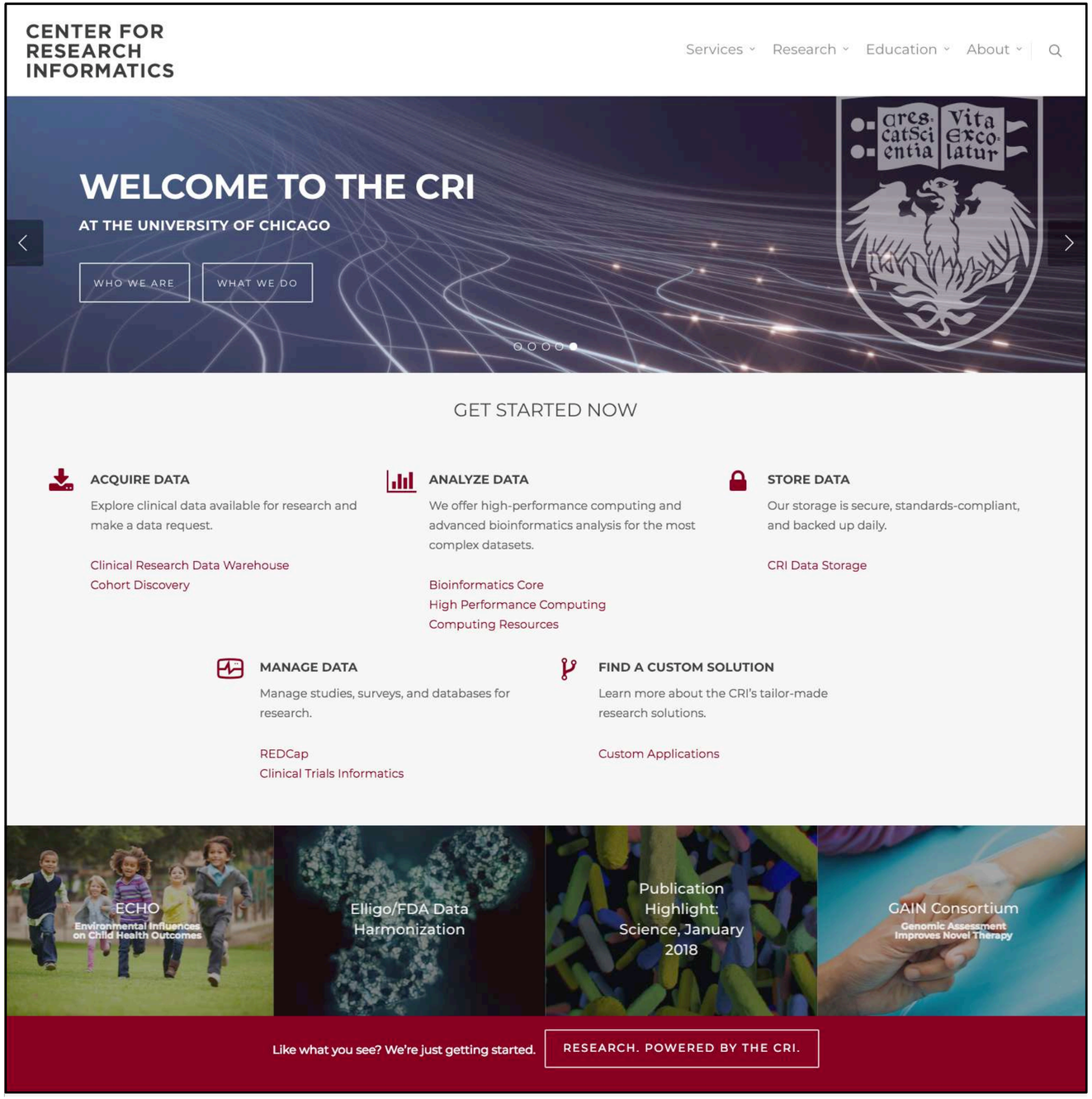
Center for Research Informatics

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

http://cri.uchicago.edu



CRI vs. ??

- CBIS
- Research Computing Center (RCC)
- Computation Institute (CI)
- ITS
- CDIS
- Biostatistics core
- CHDSI

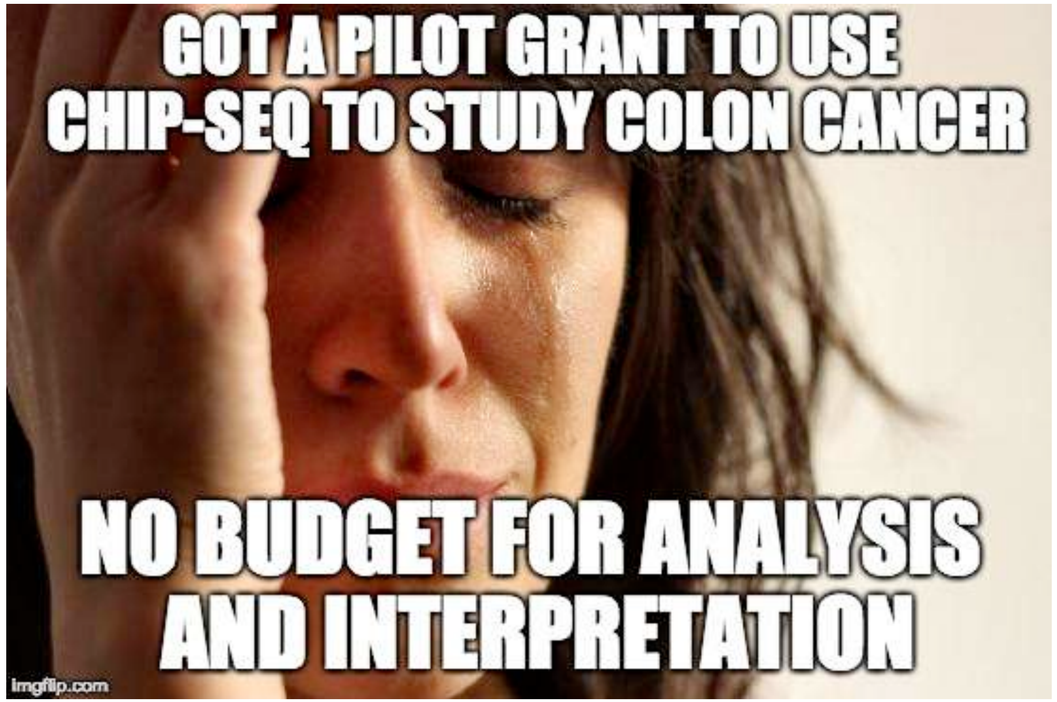
The idealized process…

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

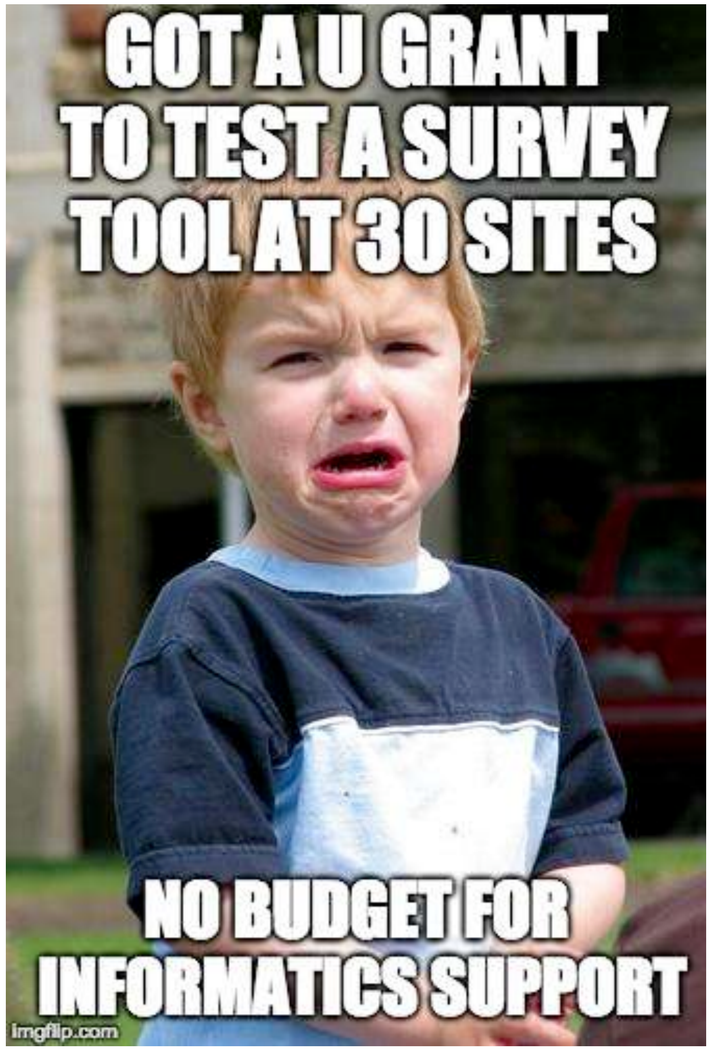
A more realistic process

- 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

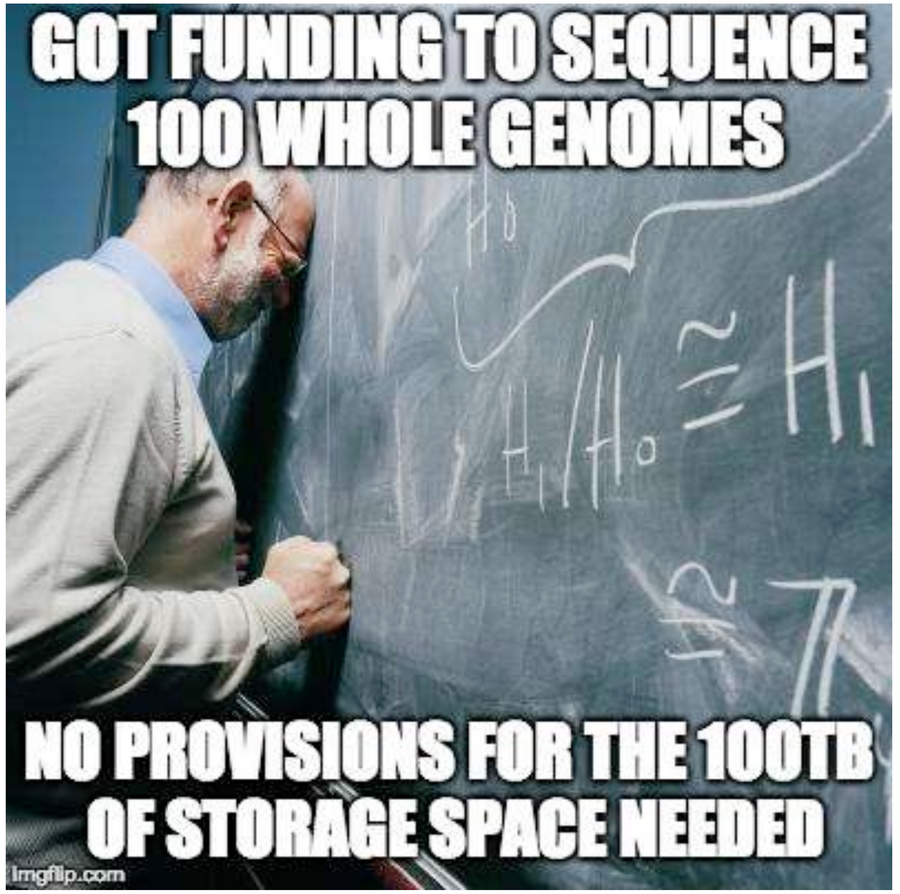
Scenario #1 - The sequencer



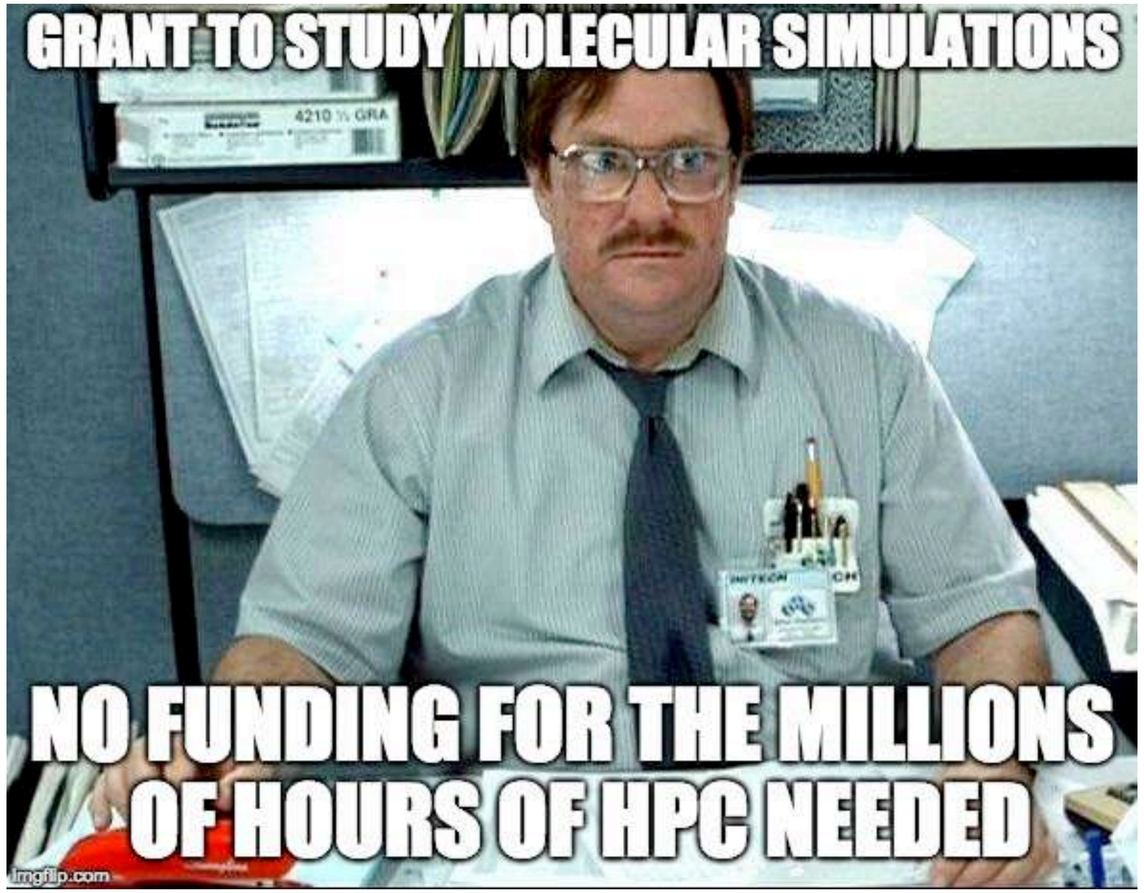
Scenario #2 - The multi-center trial



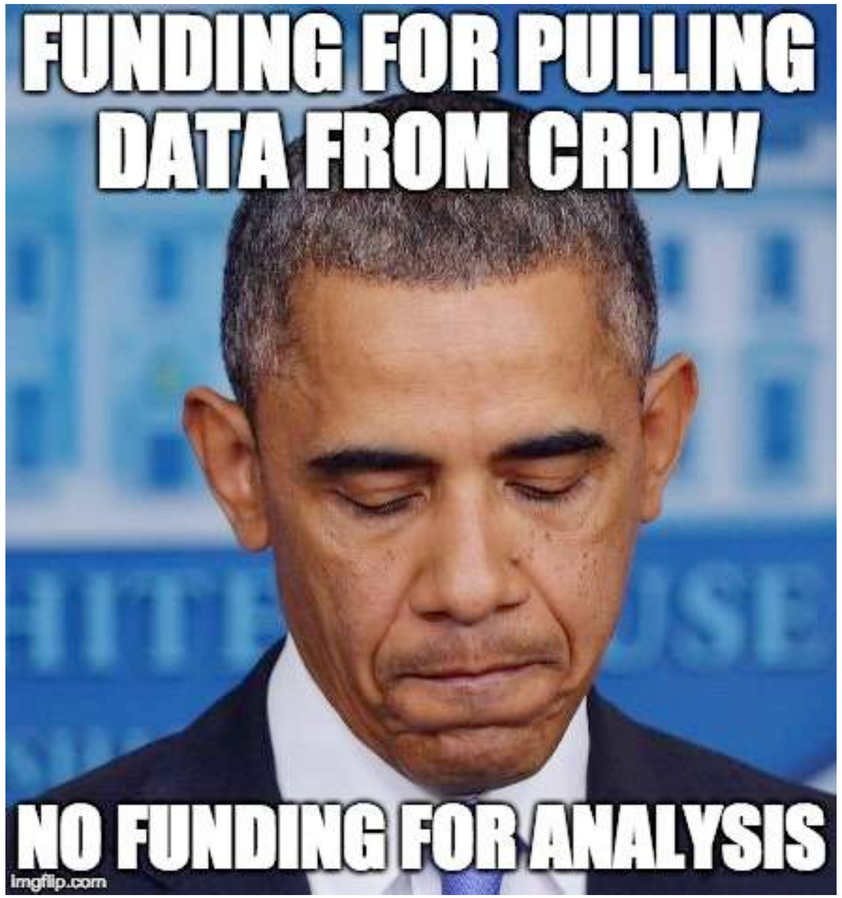
Scenario #3 - The Big Data™ user



Scenario #4 - The simulator



Scenario #5 - The analyzer

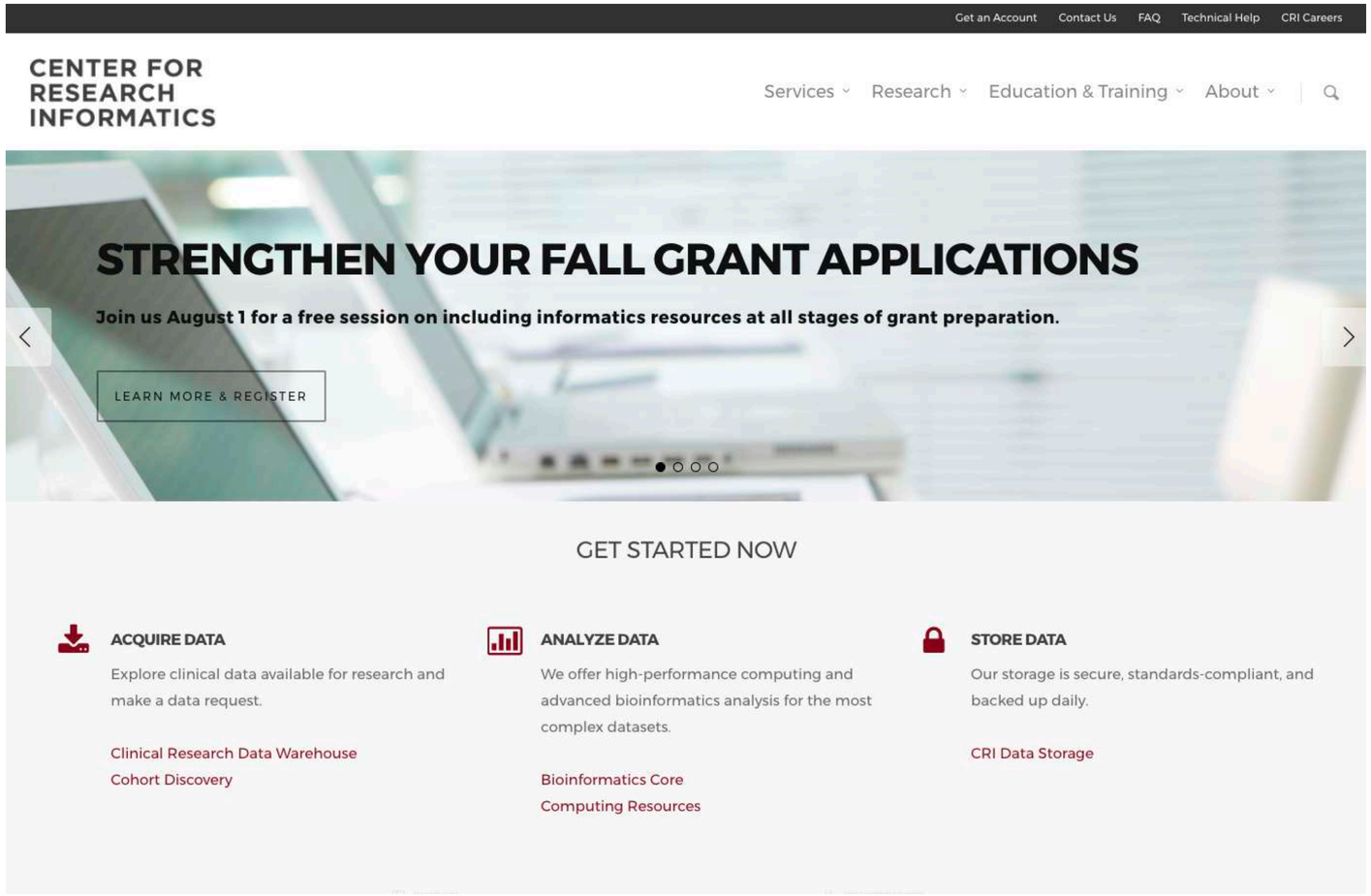


There are many opportunities to consider informatics resources

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

Getting informatics help

http://cri.uchicago.edu



slv@uchicago.edu

support@rt.cri.uchicago.edu

Common to all proposals

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

IRB writing / positioning

- 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 on 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.
- Monthly meeting with CRI, OCR, IRB, legal, and security to discuss and address these issues **proactively**

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

These aren't good places to store your data.

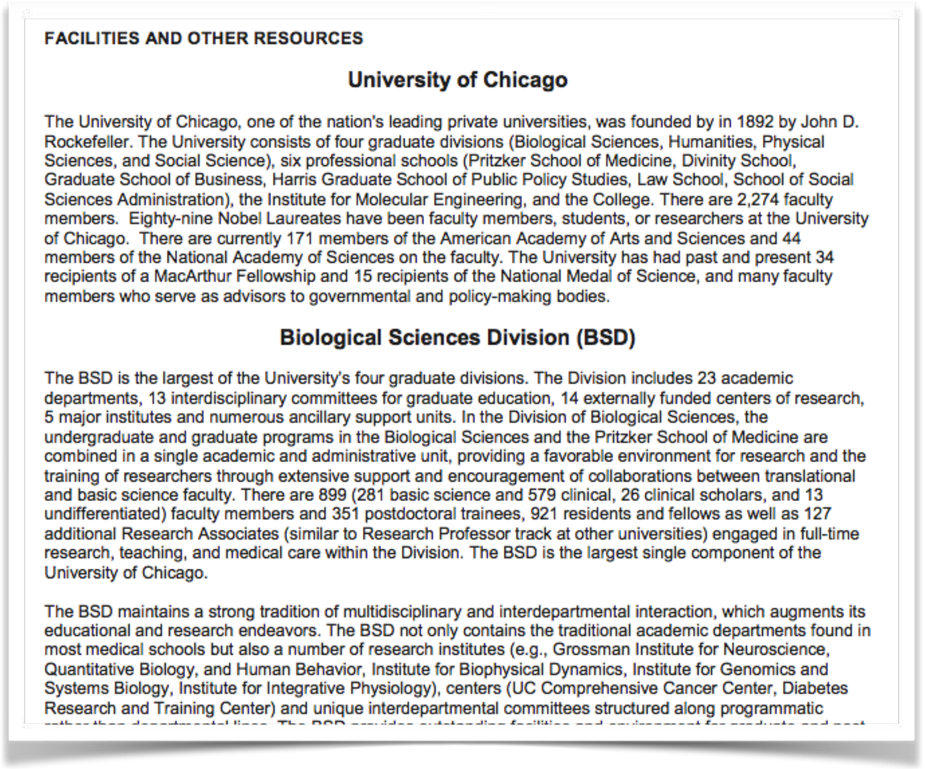
Why?

- 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)

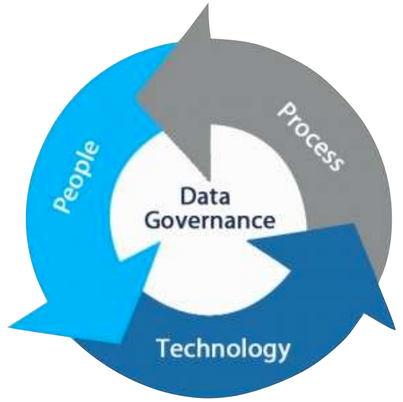
The BSD is the largest of the University's four graduate divisions. The Division includes 23 academic departments, 13 interdisciplinary committees for graduate education, 14 externally funded centers of research, 5 major institutes and numerous ancillary support units. In the Division of Biological Sciences, 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

CRI has boilerplate language for grants

Data governance and stewardship

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



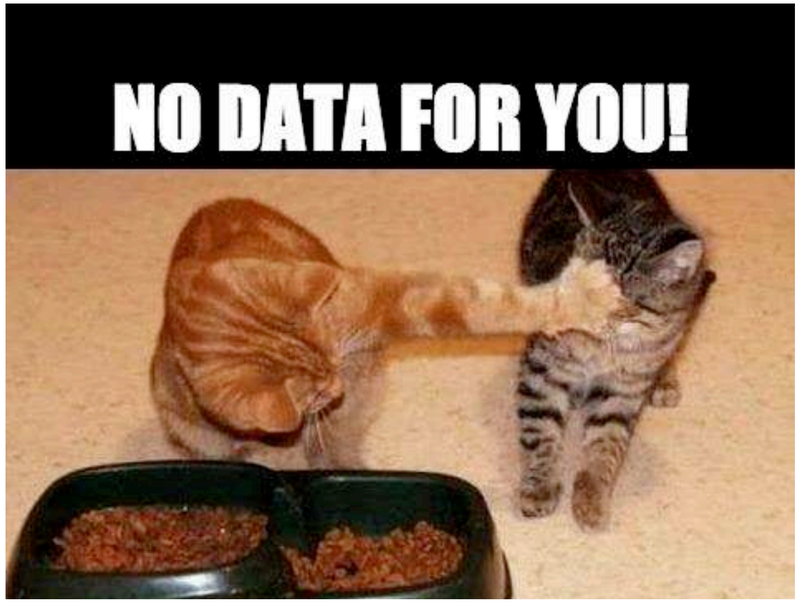
Examples of data governance considerations

- Who 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 are data being moved securely between researchers?

Failure to address these questions adequately 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



Bioinformatics considerations

- Methods and study design
- Budget planning for data generation
- Grant writing - preliminary data, methods, research plan
- Data storage, movement, backup
- Analysis and interpretation
- Integration of multiple data sources
- Manuscript preparation and submission

Bioinformatics - Methods and study design

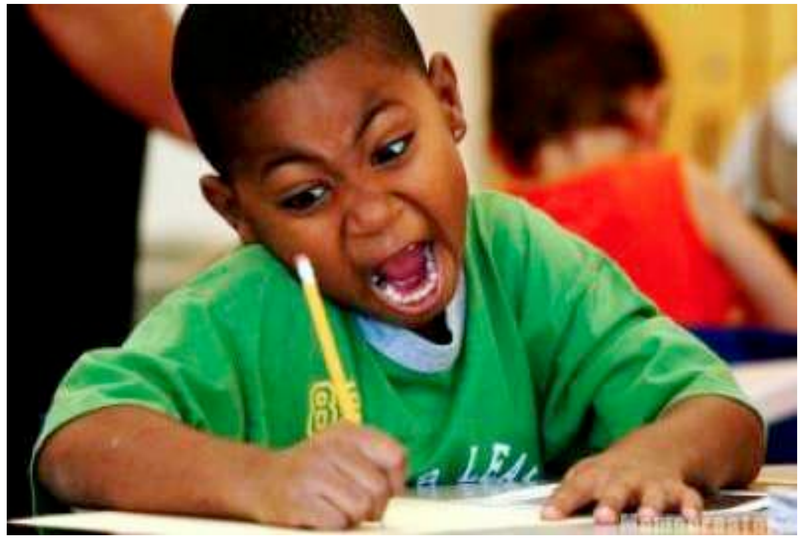
- What kind of analysis?
  RNA-Seq? ChIP-Seq? WGS? WES?
- What depth of coverage?
- Power calculations: How many samples?
  Technical replicates? Biological replicates?

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, Herna… modulation of inflammatory pathwa…
- Sasaki MM, Skol AD, Hungate EA, Ba… Sporadic Inflammatory Bowel Diseas…
- Kadri S, Zhen CJ, Wurst MN, Long BC… Insertion/Deletion Mutations in Amp… ahead of print]
- Hernandez KM. Understanding the g… 10.1111/nph.13607. Epub 2015 Aug 26.
- Zhong R, Bao R, Faber PW, Bindokas… Aug 20. doi: 10.1158/0008-5472.CAN-…
- **Publication Highlight**: Bao R, Herna… Exome Germline and Somatic Mutat…
- Sasaki MM, Skol AD, Bao R, Rhodes L… for Familial Nasopharyngeal Carcino…
- Ernst LM, Rand CM, Bao R, Andrade J… Cord Samples with Pathologic and C…
- **Publication Highlight**: Spranger S, B… 9;523(7559):231-5. doi: 10.1038/nature…
- Volchenboum SL, Andrade J, Huang … prognostic importance of tumour-st… 10.1002/cjp2.9.

2016

- Spranger S, Luke JJ, Bao R, Zha Y, Hernandez KM, Gakewski 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 16. doi:10.1158/2326-6066.CIR-16-0087. [Epub ahead of print]
- …halyfa A, Zhang C, Khalyfa AA, Foster GE, Beaudin AE, Andrade J, et al. Effect on Intermittent Hypoxia on Plasma Exosomal Micro RNA Signature and …ndothelial Function in Healthy Adults. *Sleep*. 2016 Sep 9. pii: sp-00145-16. [Epub ahead of print]
- …lmendros I, Khalyfa A, Trzepizur W, Gileles-Hillel A, Huang L, Akbarpour M, et al. Tumor Cell Malignant Properties Are Enhanced by Circulating Exosomes … Sleep Apnea. *Chest*. 2016 Aug 25. doi:10.1016/j.chest.2016.08.1438. [Epub ahead of print]
- …halyfa A, Almendros I, Gileles-Hillel A, Akbarpour M, Trzepizur W, Mokhlesi Babak, et al. Circulating exosomes potentiate tumor malignant properties in … mouse model of chronic sleep fragmentation. *Oncotarget*. 2016 Jul 13.

2017

- Shah S, Wood K, Labadie B, Won B, Brisson R, Karrison T, et al. Clinical and molecular features of innate and acquired resistance to anti-PD-1/PD-L1 therapy in lung cancer. *Oncotarget*. 2017 Dec 15. doi:10.18632/oncotarget.23315
- Vougiouklakis T, Bao R, Nakamura Y, Saloura V. Protein methyltransferases and demethylases dictate CD8+ T-cell exclusion in squamous cell carcinoma of the head and neck. *Oncotarget*. 2017 Nov 22. doi: 10.18632/oncotarget.22627
- Gray JA, MacKillop J, Weafer J, Hernandez KM, Gao J, Palmer AA, et al. Genetic analysis of i… candidates and genome-wide variation. *Psych Res*. 2017 Oct. doi: 10.1016/j.psychres.2017.10.0…
- Samant SA, Kanwal A, Pillai VB, Bao R, Gupta MP. The histone deacetylase SIRT6 blocks my… *Rep*. 2017 Sep 19. doi:10.1038/s41598-017-10838-5
- Girard R, Zeineddine HA, Fam MD, Mayampurath A, Cao Y, Shi C, et al. Plasma Biomarkers … of Cerebral Cavernous Malformations. *Transl Stroke Res*. 2017 Aug 17. doi: 10.1007/s12975-017…
- Kadri S, Lee J, Fitzpatrick C, Galanina N, Sukhanova M, Venkataraman G, et al. Clonal evolut… transformation in patients with ibrutinib-relapsed CLL. *Blood Advances*. 2017 May 9. doi: 10…
- Gozal D, Gileles-Hillel A, Cortese R, Li Y, Almendros I, Qiao Z, et al. Visceral white adipose tis… mice. *Am J Respir Cell Mol Biol*. 2017 Apr. doi: 10.1165/rcmb.2016-0243OC.
- Saloura V, Fatima A, Zewde M, Kiyotani K, Brisson RJ, Park JH, et al. Characterization of the … microenvironment in patients with locoregionally advanced squamous cell carcinoma of t… doi: 10.1158/1078-0432.CCR-17-0103.
- Kach J, Long TM, Selman P, Tonsing-Carter EY, Bacalao MA, Lastra RR, et al. Selective gluco… resistant prostate cancer growth. *Mol Cancer Ther*. 2017 Apr 20. doi: 10.1158/1535-7163.MCT-1…
- Wang LJ, Chan WC, Chou MC, Chou WJ, Lee MJ, Lee SY, et al. Polymorphisms of STS gene … boys with attention-deficit/hyperactivity disorder: an exploratory investigation. *Sci Rep*. 201…
- Pitroda S, Bao R, Andrade J, Weichselbaum RR, Connell PP. Low Recombination Proficienc… damaging chemotherapy in breast cancer. *Clin Cancer Res*. 2017 Mar 24. doi: 10.1158/1078-0…
- 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]

2018

- Bhanvadia RR, VanOpstall C, Brechka H, Barashi NS, Gillard M, McAuley EM, et al. MEIS1 and MEIS2 Expression and Prostate Cancer Progression: A Role For HOXB13 Binding Partners in Metastatic Disease. *Clin Cancer Res*. 2018 May 1. doi: 10.1158/1078-0432.CCR-17-3673. [Epub ahead of print]
- Friedrich M, Bao R, Dia S, Issa HA, Alhussein D. Comparative evidence of an exceptional impact of gene duplication on the developmental evolution of Drosophila and the higher Diptera. *Front Ecol Evol*. 2018 May. doi: 10.3389/fevo.2018.00063
- **Publication Highlight**: Pitroda SP, Khodarev NN, Huang L, Uppal A, Wightman SC, Ganai S, et al. Integrated molecular subtyping defines a curable oligometastatic state in colorectal liver metastasis. *Nature Communications*. 2018 May 4. doi: 10.1038/s41467-018-04278-6
- **Publication Highlight:** Kang W, Kadri S, Puranik R, Wurst MN, Patil SA, Mujacic I, et al. System for Informatics in the Molecular Pathology Laboratory (SIMPL): An Open-Source End-to-End Solution for Next-Generation Sequencing Clinical Data Management. *J Mol Diagn*. 2018 Apr 24. doi: 10.1016/j.jmoldx.2018.03.008
- Xie T, Bloom L, Padmanaban M, Bertacchi B, Kang W, MacCracken E, et al. Long-term effect of low frequency stimulation of STN on dysphagia, freezing of gait and other motor symptoms in PD. *J Neurol Neurosurg Psychiatry*. 2018 Apr 13. doi: 10.1136/jnnp-2018-318060. [Epub ahead of print]
- **Publication Highlight**: Han J, Perez JT, Chen C, Li Y, Benitez A, Kandasamy M, et al. Genome-wide CRISPR/Cas9 Screen Identifies Host Factors Essential for Influenza Virus Replication. *Cell Rep*. 2018 Apr 10. doi: 10.1016/j.celrep.2018.03.045
- **Publication Highlight**: Matson V, Fessler J, Bao R, Chongsuwat T, Zha Y, Alegre M, et al. The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science*. 2018 Jan 5. doi: 10.1126/science.aao3290

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.

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

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

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.

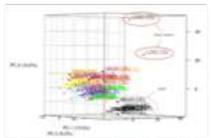


Fig. 2. Illustration of the outliers and batch effects in the COG gene expression data

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
CFLIR,GS1-6,CDH5,SLC29A1	0.925	0.929	1.000	0.900
CTSC,ANPEP,ITGA9,DCBLD1	0.913	0.857	1.000	0.800
SLC29A1,CFL1,SPAN15,DDIT3,EMR2	0.925	0.857	1.000	0.800
DCBLD1,GSTM2,LYN,RAPGEF5	0.925	0.857	0.750	0.900
ANPEP,C1orf162,LOC100132167,NDP52,PLEK	1.000	1.000	1.000	1.000
CTSC,DCBLD1,RGS16,TET1,CCL19,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,CCL3,GSTM2,TP53I11,CTSC,DDIT3,RGS16,SLC29A1,ITGA9,TET1,HEY2,ICAM1,RPS6KA2	1.000	1.000	1.000	1.000
EMR2,NR0F3,CCL18,SLC29A1,QCAM1,LOC100132167,PODXL,NKAIN1,HBX015,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).

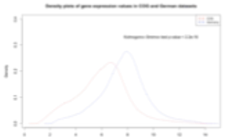


Fig. 4. Density plots of gene expression data in the COG and Germany datasets

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

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.

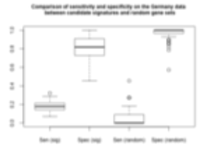


Fig.5. Boxplot of sensitivity and specificity in the prediction of the survival status on the Germany data set using 5-gene candidate signatures and random gene sets. Left two are sensitivity and specificity of the candidate signatures and right two are sensitivity and specificity of the random sets.

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.

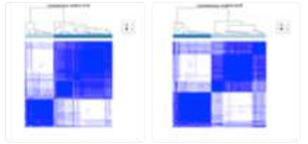


Fig.6. Consensus clustering using gene expression data from the COG data set (left) and the Germany data set (right).

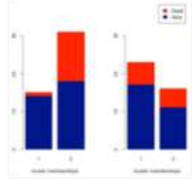


Fig.7. Stacked barplots for the distribution of survival status between consensus clusters. Left: COG data set; right: Germany data set.

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

Data warehouse and business intelligence

CRDW - Special considerations

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

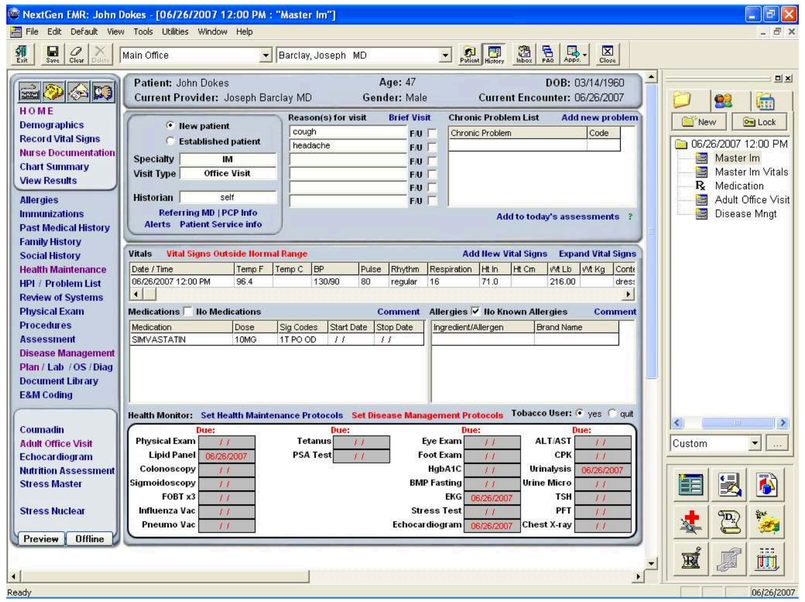
CRDW - Preliminary data / Cohort identification

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

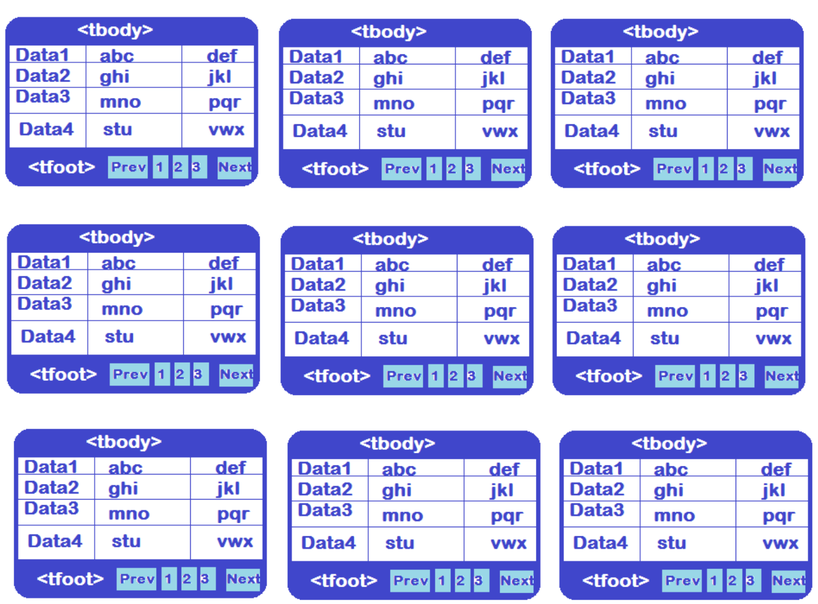
- Reviewers like to see preliminary identification of cohorts - “Can they really get the data”
- Sometimes new data have to be sourced

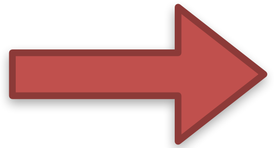
CRDW - Data element identification

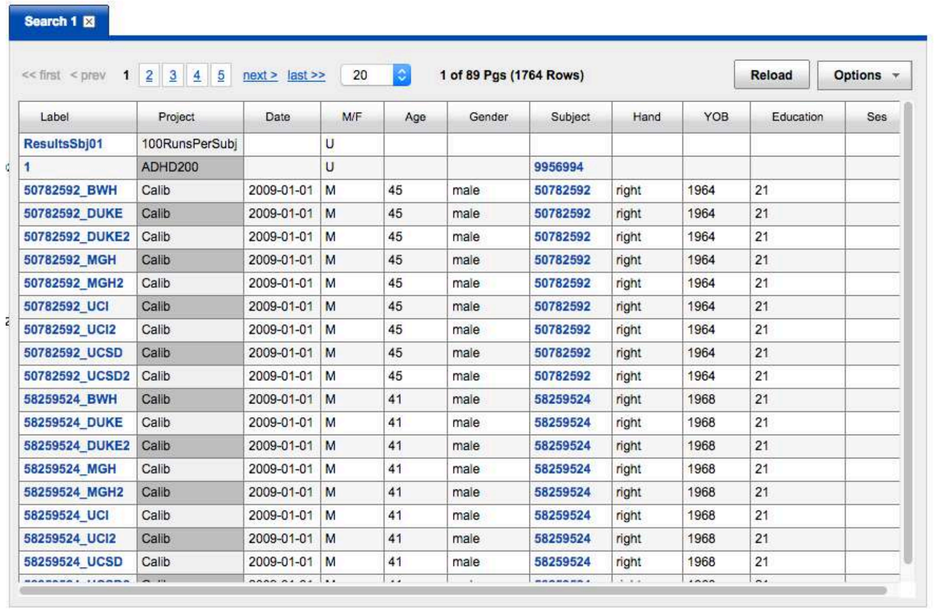
- Identifying elements to pull from the CRDW is an iterative process
- Requires input from the CRDW and the investigator
- Delineation of data elements in the grant is essential



CRDW - Data aggregation and normalization







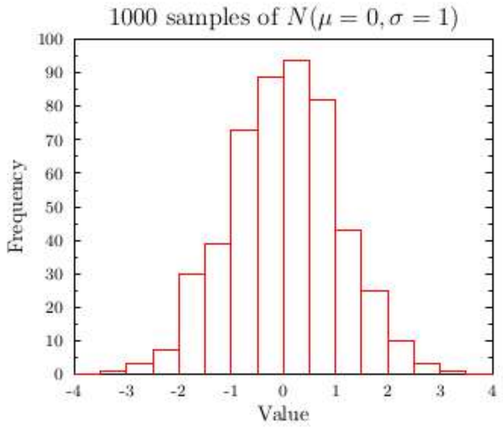
Taking the complex, multidimensional data from the CRDW and creating a usable data set for subsequent analysis requires special skills and should be included in the budget for data acquisition

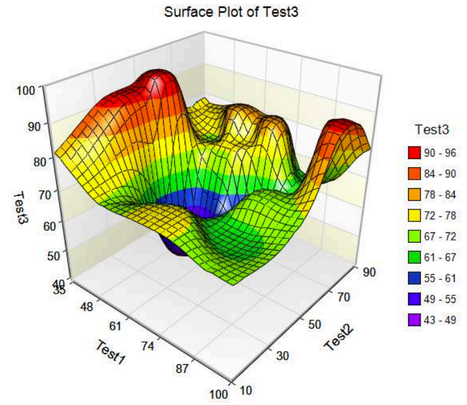
CRDW - Data analysis and interpretation

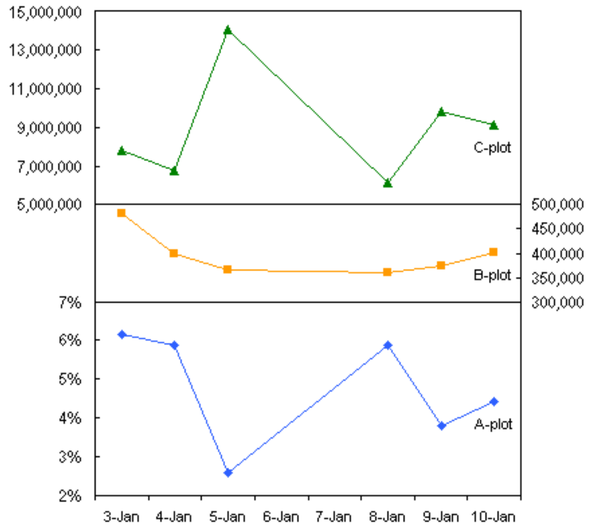
Search 1

<< 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	







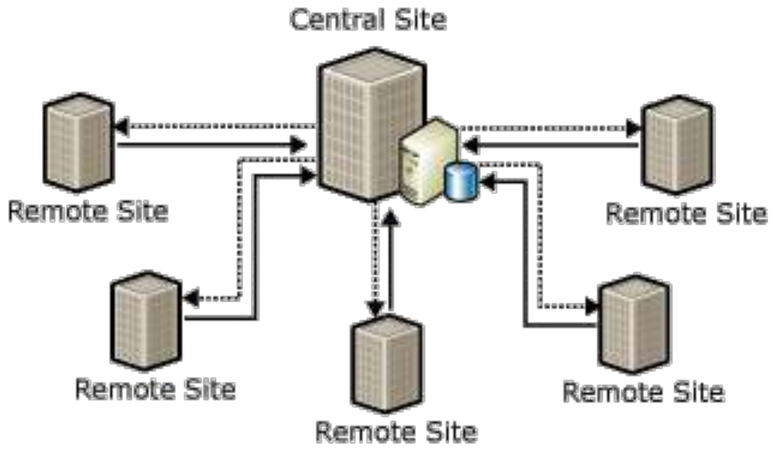
Data analysis can be costly and time-consuming.

Consider adding an analyst to your budget vs. chargeback

Applications / Special projects

Applications - Special considerations

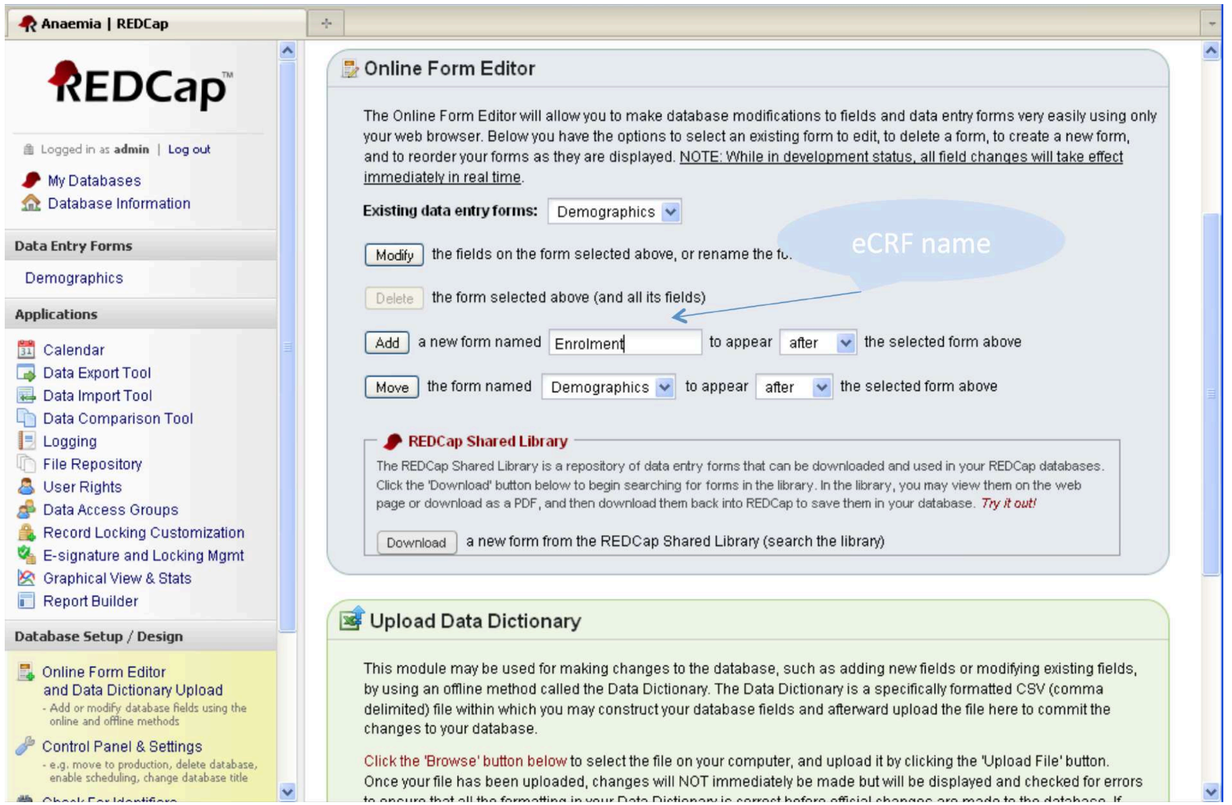
- Multi-site data collection, transfer, and storage
- REDCap usage
- Application development and programming



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

Examples projects

- Data commons for pediatric cancer
- ECHO
- March of Dimes
- GAIN
- Thirty Million Words
- Genomic Prescribing System

Systems and infrastructure - special considerations

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

Systems - Offsite access

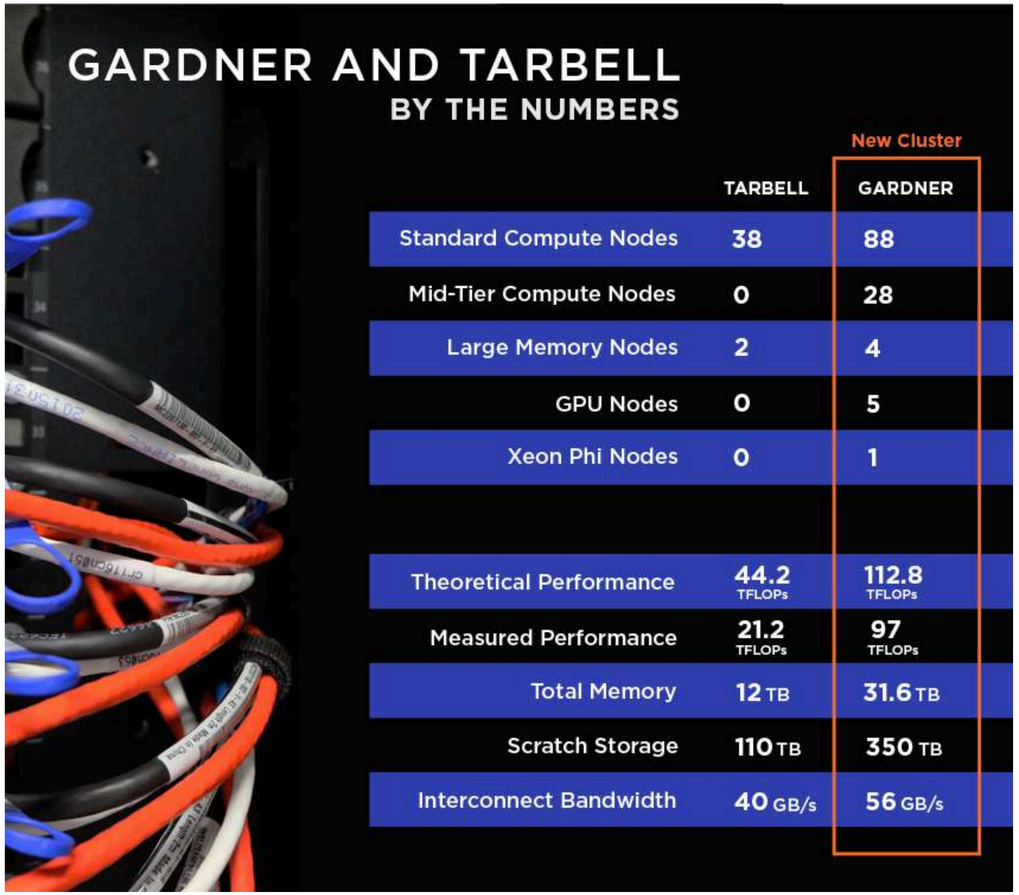
- Do researchers outside of UChicago need access to your data?
- Collaborator accounts take time to obtain - and the CRI can help

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

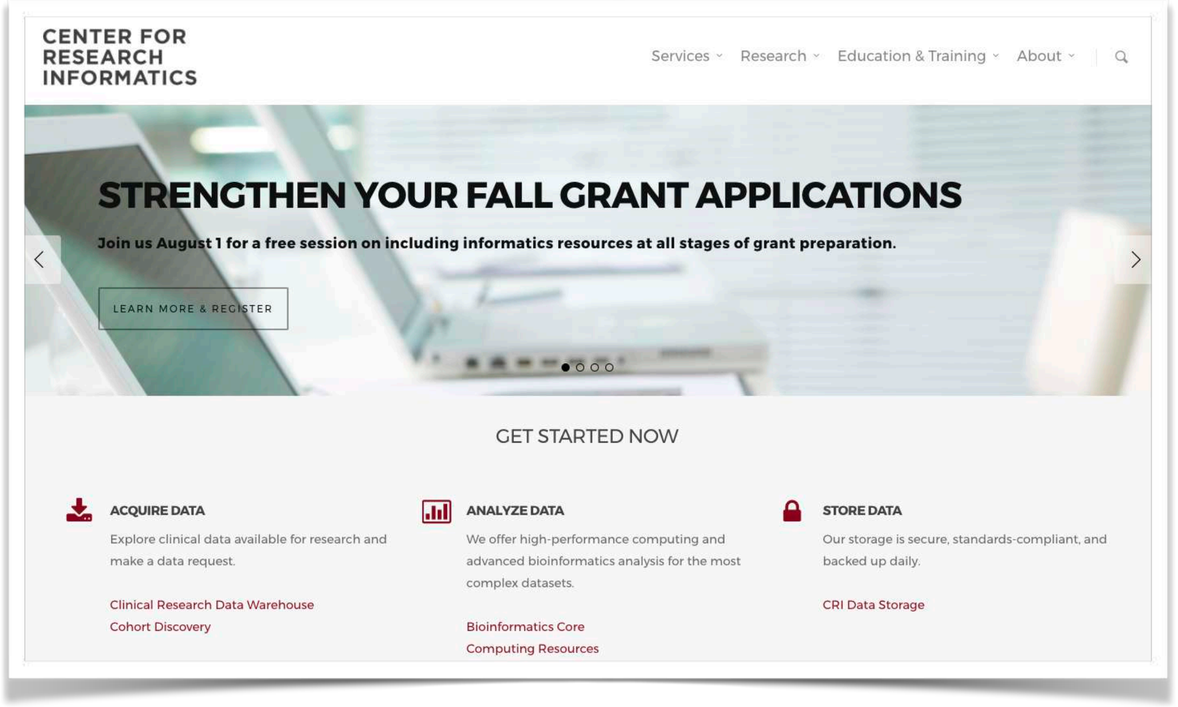


Systems - Virtual machines / servers

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

Center for Research Informatics

Applications - Systems - Bioinformatics - Data warehousing - Clinical trials

Questions?