



Supporting Personalized Cancer Medicine

@

MD Anderson Cancer Center

Krishna Sankhavaram

Director, Research Information Systems & Technology
Development

The University of Texas MD Anderson Cancer Center
Houston, TX

THE UNIVERSITY OF TEXAS

**MD Anderson
Cancer Center**

Making Cancer History®

Brief Introduction to MD Anderson

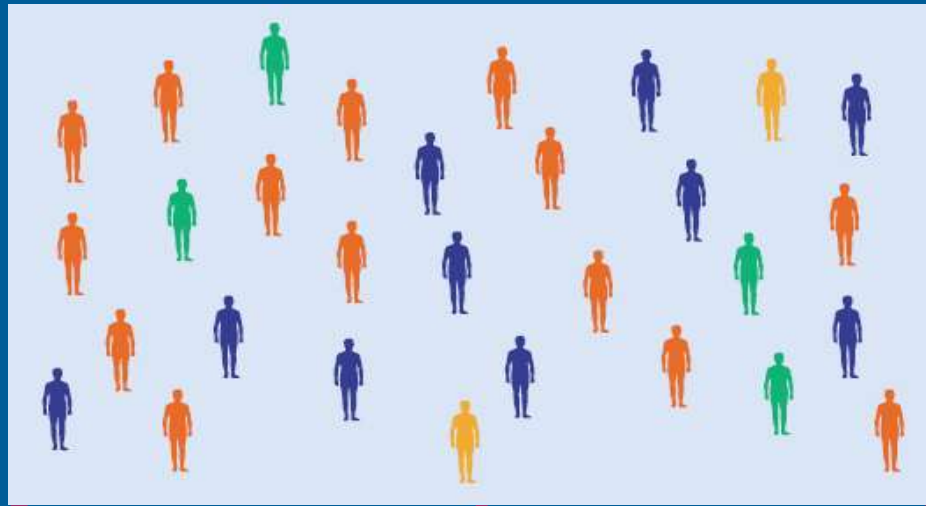
► Mission

The mission of The University of Texas M. D. Anderson Cancer Center is to eliminate cancer in Texas, the nation, and the world through outstanding programs that integrate patient care, research and prevention, and through education for undergraduate and graduate students, trainees, professionals, employees and the public.

**MD Anderson is ranked #1 Cancer Center
in the US**

Personalized Molecular Medicine

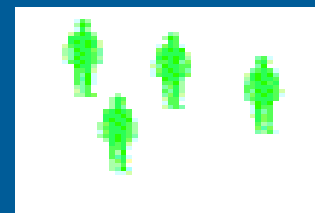
Patients with the same diagnosis . . . are not the same



Predicted good response to drug or combination of drugs



Predicted poor or no response to drug or combination of drugs
CHANGE DRUGS



Increased likelihood of toxicity of drug or combination of drugs
CHANGE DRUGS

Major Challenges for Healthcare Information Technology

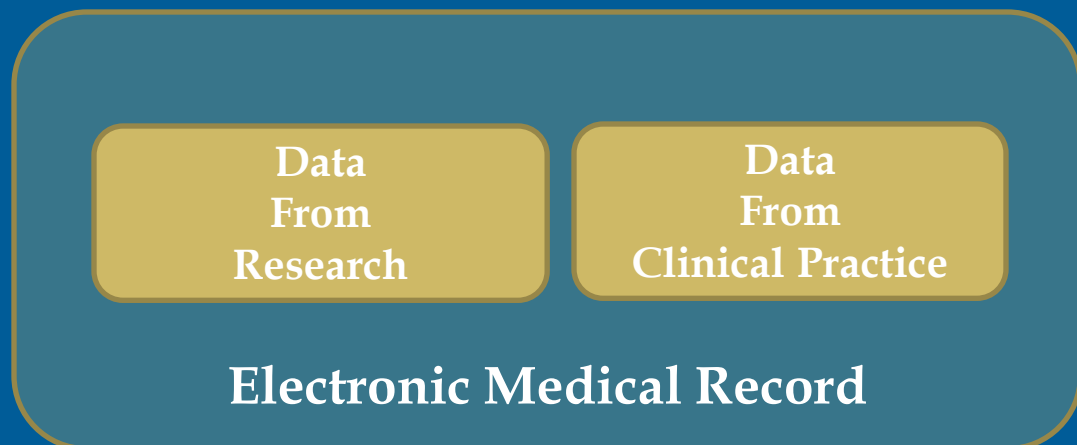
- ▣ Huge volume of data and availability of Computational Resources
 - Storage: 2.2 PB of storage, backed up & mirrored across town
 - Processing: 11 shared memory servers (512GB RAM, 32 CPUs each); 4 node Oracle 11g;
 - High Performance Computing Cluster with 8,064 cores, and 24 cores/server (AMD Magny Cours), 1GB memory/core;
 - Staff: 4 PhDs, 9 Computer Science Masters, 3 Engineering Masters, 3 with Biomedical Informatics Certificates, 3 pursuing Biomedical Informatics PhDs.
- ▣ Interoperability in order to access and utilize data from diverse clinical and research sources at the point of care
- ▣ Complexity and data validation

Initial Observations about IT

- ▣ Historically, our IT investments have supported silos of research and clinical data
 - Investigator-initiated research
 - Clinical care in hospitals versus physician offices
- ▣ Moving beyond these silos is both a technical and a cultural challenge
- ▣ In spite of these barriers, IT is now pervasive in almost every organization
- ▣ BUT, if clinical care and scientific discovery are increasingly intertwined, then IT is a very important mechanism for facilitating this process.

How we frame the questions:

Source of our Challenge



How we describe in IT terminology

Interoperability

Where the solution lies

Architecture

The fundamental question for Scientists, Clinicians, and Patients

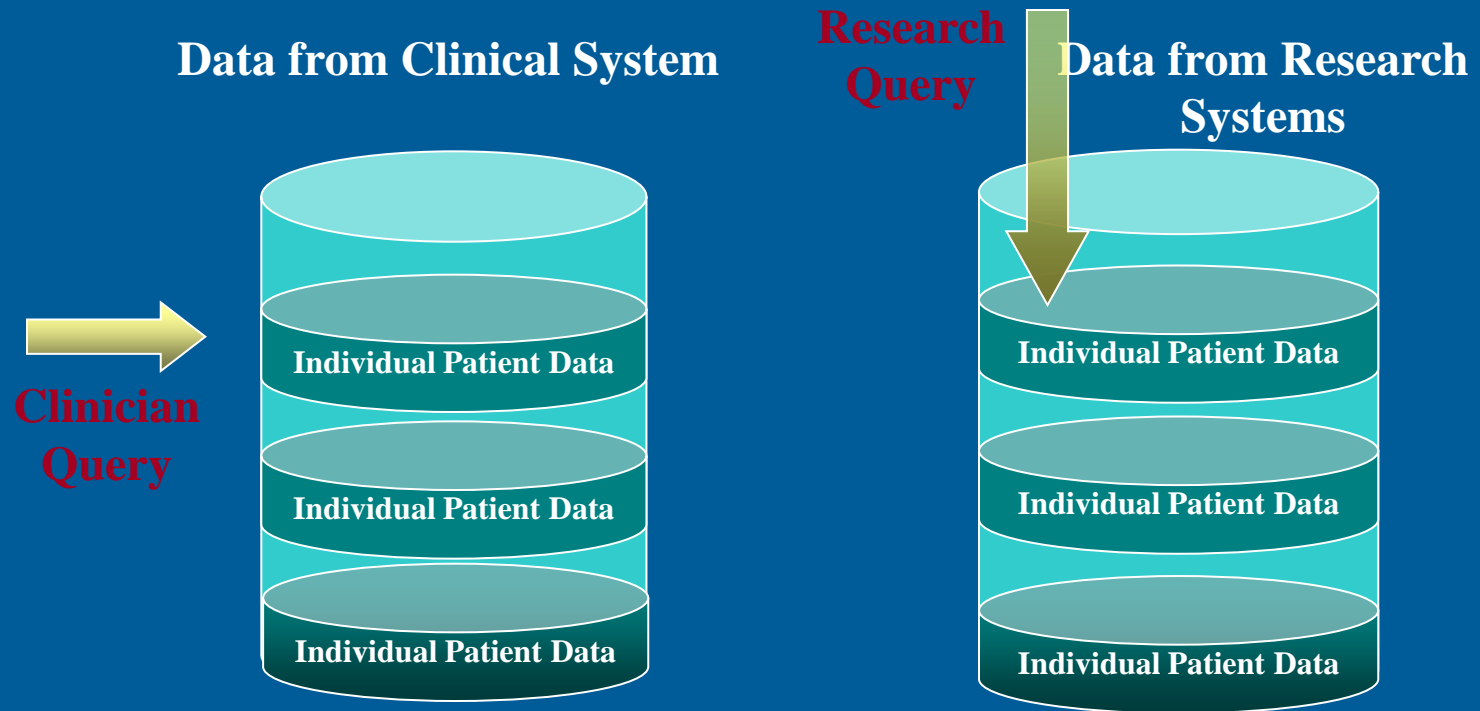
What happened to the last 100 patients
who had the same diagnosis,
who have similar characteristics, and
who were treated with the regimen that the
clinician is recommending?

**Outcomes
from
Evidence-
based
Medicine**

**Personalized
Medicine**

**Translational
Research**

Challenges of integrating Data from Clinical Practice and Research



**Focus on Single Patient,
Many Attributes**

**Focus on Many Patients,
Few Attributes, Algorithm-based**

The Stream of Molecular Data

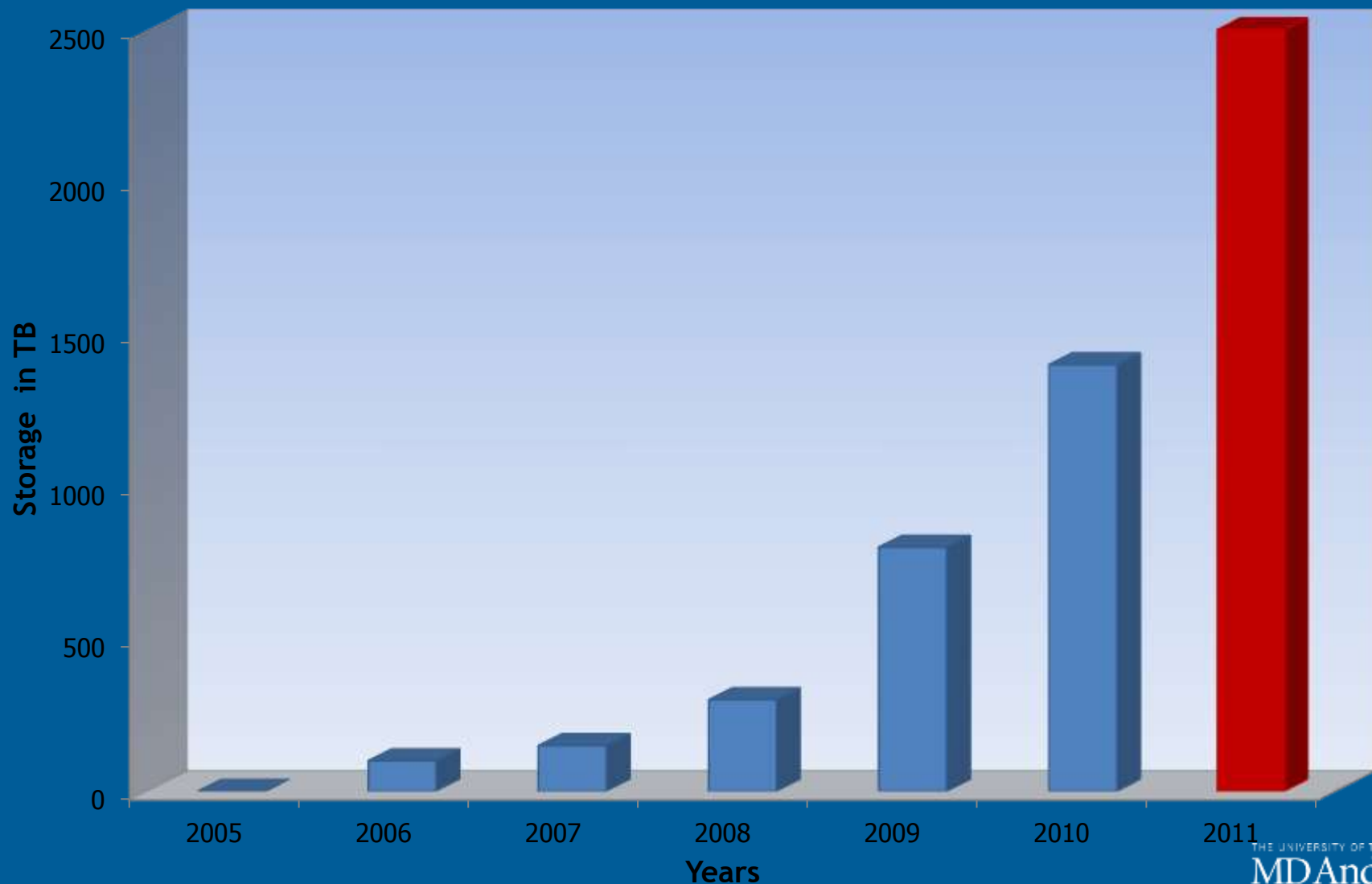


Thanks to Dr. John Weinstein

2008

2010

Challenge of Managing the Growth of Data



What to keep & what to throw away

- ▣ Life Sciences data is persistent
 - In a Cancer center, this is extremely critical !
- ▣ We are required to keep data for a long time
 - At the moment we keep all data on-line
 - In most of our cases, you will not get a chance to repeat the experiment
- ▣ In the next several years as we expect to manage data in the order of 10-15 Peta Bytes/year
 - How to handle this much data?

What do MD Anderson Researchers ask for?

- ▣ Secure access and storage of their research data
- ▣ Ability to grant access to their data themselves
- ▣ Access to as many analysis tools as possible
 - Ability to add their own
- ▣ Access to as much compute power as they need
 - *They don't know how much !! But ALWAYS want more!!*
- ▣ Ability to access clinical attributes of the samples per IRB approvals.
- ▣ Eventual Goal:
 - Improve Cancer Treatment, *personalize it.*

How we approached this problem

- ▣ Users need to have Services Provisioned as a package
 - Storage space
 - Computing resources
 - Analytical tools for analysis
- ▣ Users need to feel confident that
 - Data Security is per Corporate policy
 - Only the Data Owner has the right to give access to his data
- ▣ Users need to have a simple way
 - To access data, upload/retrieve data (numerous formats)
 - To run applications on appropriate servers
 - To share projects, data, results with collaborators
- ▣ Users need a place to store result datasets
 - Should support numerous output formats
 - Access to other clinical data sources

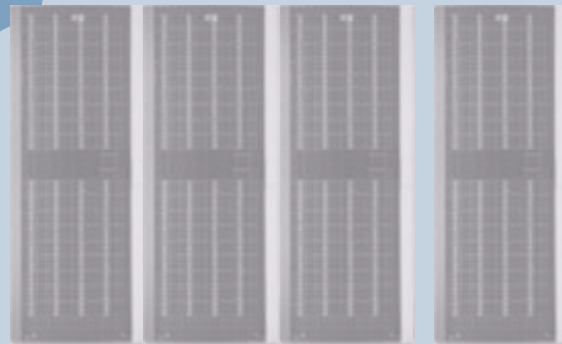
Our Solution



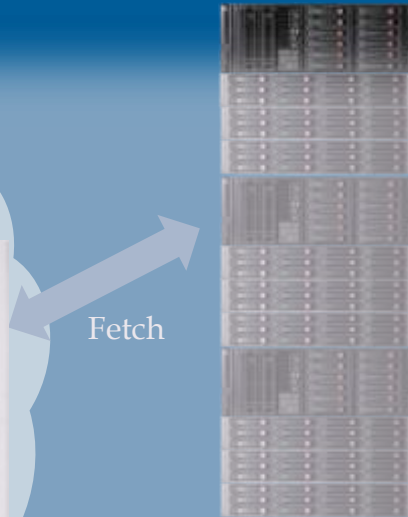
Sequencing Instruments



Archive



Research SAN



Large Memory Servers

Fetch

Fetch

Computational Cluster



Private Cloud

ResearchStation



Compute Capacity for Research

- 8064 cores blade cluster (BL465G7-based)
 - Infiniband interconnect (2:1 oversubscribed)
 - Used for I/O intensive and Message Passing codes: NGS analysis, Dosing calculations for Radiation therapy, Epidemiology, Outcomes Research
- 11 servers with 32 cores each with 512 GB memory each (DL585G7)
 - Large amount of NGS pre-processing, Molecular modeling, SAS, several standard packages
- Parallel file system (HP X9000)
 - *Allows for the file system to be accessible to the compute nodes and all the large memory servers*
 - Don't have to copy large datasets (TBs) from point a to point b for analysis

Storage Environment

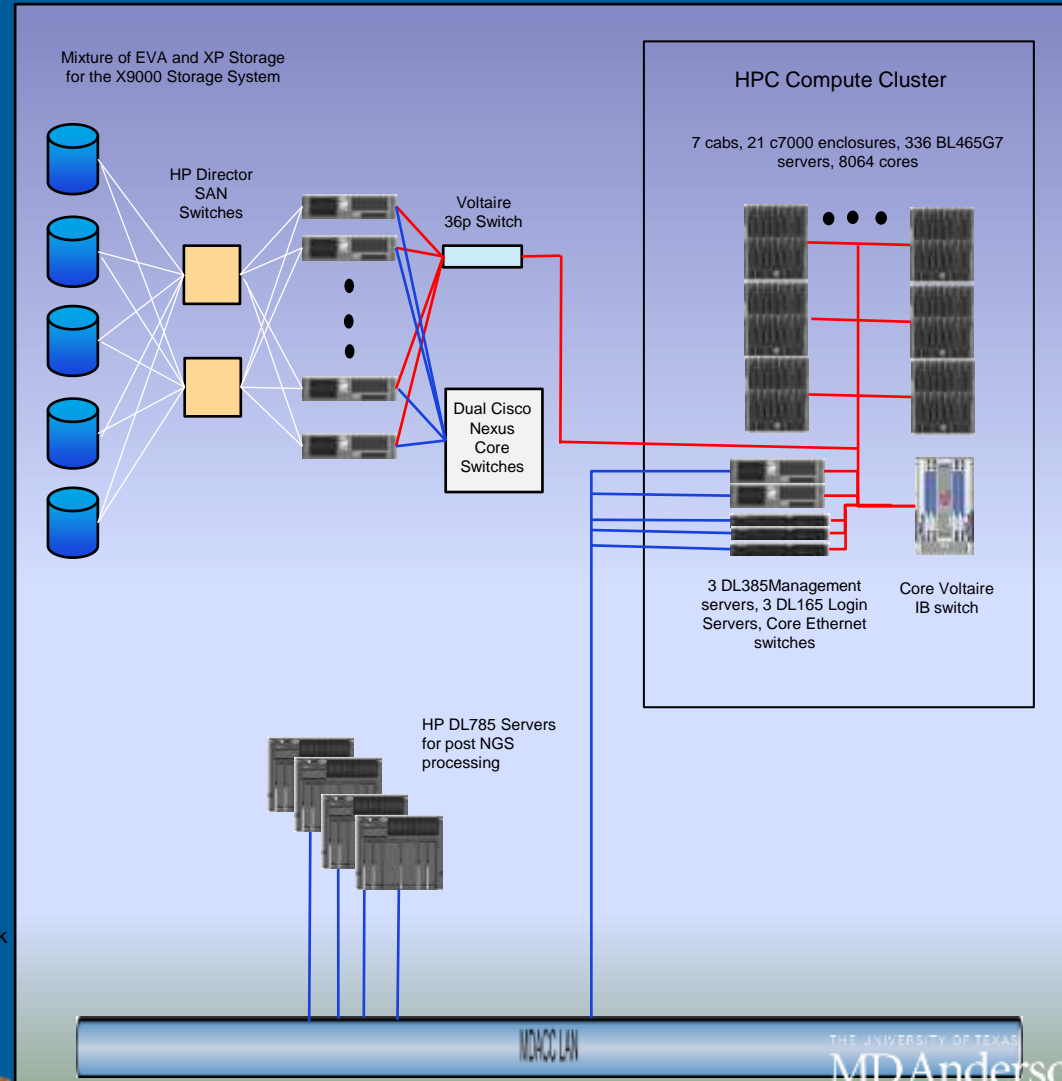
- ▣ Tier-1: Used for real-time data acquisition, fast access
 - HP XP24000
- ▣ Tier-2: Data repositories, most of our SAN
 - HP EVA 8000s, 8100s and 8400s
- ▣ Tier-3: Older data
 - Backup to disks on HP X9720
 - Backup mirrored across metro Houston into another X9720
- ▣ 16 X9300 gateway servers
 - The entry point to the entire storage for all servers
 - A specific file can be presented to all servers and HPC

MDA Research Computing Environment

MDACC Main Campus Datacenter



Guhn Road Colo Site (DCG)



10G WAN Link
Approx. 26 miles

Who uses high performance computing at MDACC?

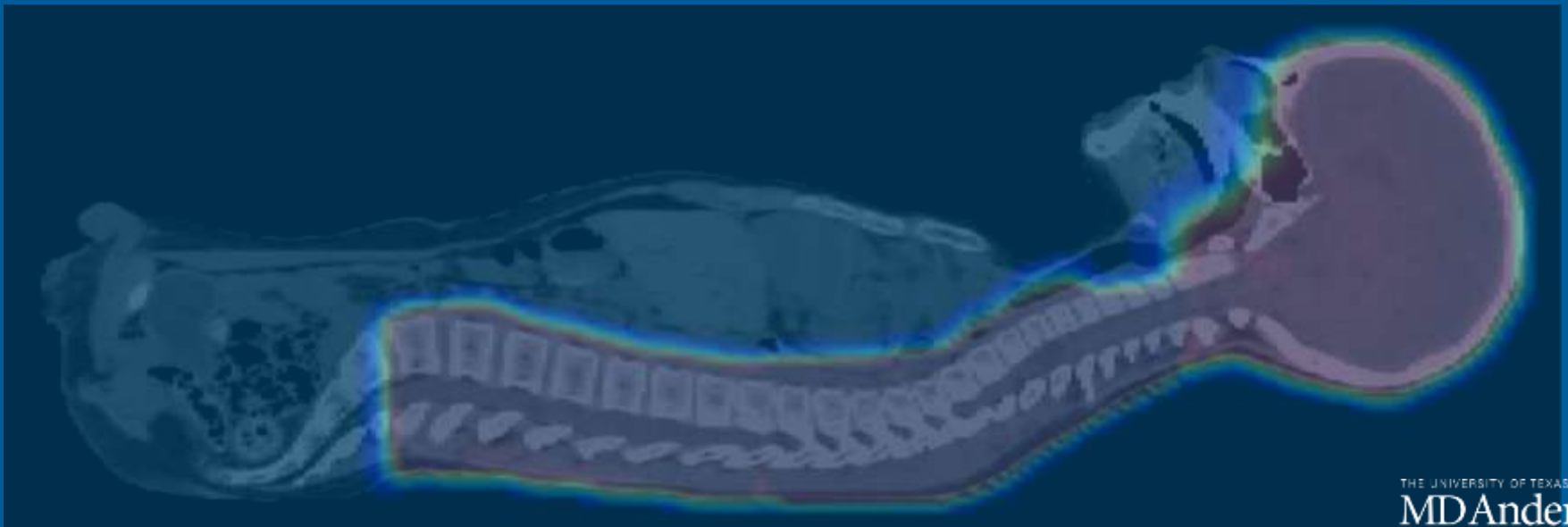
- ▣ Broadly diversified community: Epidemiology, Bioinformatics, Computational Biology, Systems Biology, Radiation Physics, Pathology, Biostatistics, and Others.
- ▣ In recent years, cancer researchers at MDACC have increasingly relied on computational methods to perform their NCI-funded investigations.

Types of Analyses Requiring High Performance Computation

- ❑ Crystallographic reconstruction of protein structures
- ❑ Identifying molecule interaction sites for small molecules – drug discovery
- ❑ Computational Genomics
- ❑ Epidemiological Population Based Simulation Studies
- ❑ Evaluating Characteristics of Tests for Arrays
- ❑ Whole Body Reconstruction of Dosing- Radiation Physics

How much does HPC help MDACC?

- ▣ An internal study revealed that there were 31 **active** NIH-funded projects that will continue to utilize high performance computing resources at MDACC in 2010
- ▣ Around 140 publications have resulted from work that used the cluster as the main computational resource



What is over the horizon?

- ▣ A much bigger machine due to increased demand
- ▣ Big enough for personalized cancer prevention, diagnosis, intervention, and surveillance. Not just for research anymore ...
- ▣ Big enough to keep MDAnderson #1 in computational cancer research
- ▣ Constrained only by capital and operational costs

Acknowledgements

- ▣ Dan Jackson, Sr. UNIX Systems Administrator, MDACC
- ▣ Dr. Bradley Broom, Professor, MDACC
- ▣ Dr. Lynn Vogel, VP/CIO & Professor, MDACC
- ▣ Dr. Randall Splinter, HPC Architect, HP
- ▣ Greg Mazzu, Storage Architect, HP
- ▣ Joe Sherar, Storage Architect, OnX
- ▣ Nancy Hemmen, Enterprise Account Manager, HP
- ▣ Dena Pharr, Enterprise Account Manager, OnX

Thank you !!