Arizona State University’s Next Generation Cyber Capability (NGCC)

Complex Data Integration in Biomedical Research

Ken Buetow
Director, Computational Science and Informatics
Human Genome Sequence Data is Growing Exponentially

Orcutt  Technology Review 2012, data NHGRI
The cure for cancer must be in there somewhere...
Genome Data is only one component of the Expansion

```
phenome
```

```
big data
```

```
exposome
```
Phenome Data

• Diverse types
  • Clinical Observation
  • Clinical Laboratory
  • Imaging
  • Registry
  • Biospecimens
  • Reference

• Distributed sources
  • Research Center
  • Care Delivery Setting
    • Hospital
    • Practice
    • Laboratory
  • Registry
  • Industry
  • Consumer
Google’s “smart” Contact Lens to measure glucose levels

http://www.engadget.com/2014/01/17/google-health-smart-contact-lenses-diabetes/
Real time consumer data: the next Big Data Challenge

*shipments of telehealth devices grow to about 7 million by 2018*

4th Paradigm Science

- A new method of pushing forward the frontiers of knowledge, enabled by new technologies for gathering, manipulating, analyzing and displaying data.
- Complementing data-generating science with data-driven science
- Ecumenical
  - Astronomy
  - Physics
  - Economics
  - Climate
  - Genomics
- Transdisciplinary
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil
Data Science is increasingly the driving force/rate-limiting step in biomedical problem solving

Genome Biology 12: 125 (2011)
Building a Data Science research platform
Traditional Research Platforms
one size does not fit all…
different users, different approaches…
The ASU NGCC **Data Science** platform

*an elemental whole composed of:*

- **Physical Capacity**
  - Ultra-high bandwidth Networks
  - Large-scale storage
  - Multiple “flavors” of computation
- **Logical Capabilities**
  - Software
  - Metadata
  - Semantics
- **Human Resources**
  - Transdisciplinary Teams
NGCC Physical Infrastructure

Data Reservoir

Transactional

High Speed Connectivity

HPC parallel

HPC SMA

Scratch Space

Big Data
M1000e Chassis
20 M420 Blade Servers
Intel® Xeon® Processor E5-2430
(15M Cache, 2.20 GHz, 7.20 GT/s Intel® QPI)
(40S, 240C, 640Thread)
32GB DDR3 Memory
2*200GB SSD RAID 0
2 16GB FC Controllers for SAN
Brocade 6505 FC Fabric 16G
2 Force 10 MXL 10/40Gbe
Dual Flex IO Modules (iSCSI)
64 Internal 10 Ports
Bright Cluster Manager
CMC/OM/iDRAC
Utility

TRANSACTIONAL CLUSTER COMPONENT IS BUILT FROM A 20 NODE CLUSTER OFFERING 240 CORES, 640GB OF DDR3 AND 40-200GB SOLID STATE DRIVES IN AN M1000E CHASSIS WITH 40Gbe UPLINK CONNECTS, CONVERGED NETWORK ADAPTERS FOR ISCSI OFFLOAD.

HA STORAGE

HIGHLY AVAILABLE COMPELLENT STORAGE WITH CLUSTERED NAS/NFS/CIFS CAPABILITIES, 30 ENCLOSURES WITH 12-3TB (360) 7.2K DRIVE, 1PB REPLICATED TO 2PB CAPACITY ARRAY VIA FCIP FABRIC EXTENSION BRIDGE.

1080TB

FCIP Replication via Fabric Extension

HIGHLY AVAILABLE COMPELLENT STORAGE CONSISTING OF 21 ENCLOSURES WITH 12-41TB (252) 7.2K DRIVES. REPLICATION TARGET FOR PRIMARY ETL INGESTION ARRAY.

1008TB

16GB FC SAN Connection(s)
100 NODE / 800 CORE CLUSTER WITH
32GB MEMORY PER NODE CONNECTED VIA INFINIBANC QDR INTERCONNECTS.
CLUSTER HAS 250Tb OF FAST
SCRATCH STORAGE SPACE IN THE
LUSTRE FILE SYSTEM. A HIGH SPEED
40Gbe INTERCONNECT TO THE SUPPORTING ENVIRONMENT IS PROVIDED.
Big Data cluster configuration consisting of 40 - 10Gbe interconnected nodes with dual 8 core servers with 128Gb DDR3 memory, dual 10Gbe interconnects, 24-2Tb drives and dual PSU’s. High speed 40Gbe connections to the supporting environment are provided.
100 node / 800 core cluster with 32 GB memory per node connected via InfiniBand DDR interconnects. Cluster has 250 TB of fast scratch storage space in the Lustre file system. A high-speed 40 GbE interconnect to the supporting environment is provided.

Transactional Cluster component is built from a 20 node cluster offering 240 cores, 840 GB of DDR3 and 40-200 GB solid state drives in an M1000E chassis with 40 GbE uplink connects, converged network adapters for (SCSI) offload.

Big data cluster configuration consisting of 40 - 100 GbE interconnected nodes with dual 8 core servers with 128 GB DDR3 memory. Dual 100 GbE interconnects, 24-20 TB drives and dual PSU's. High speed 40 GbE connections to the supporting environment are provided.

Highly available compellent storage with clustered NAS/IFS capabilities. 30 enclosures with 12-3TB (360) 7.2k drive, 1 petabyte replicated to 2 petabyte capacity array via FCIP Fabric Extension Bridge 1080TB.
Next Generation Cyber Infrastructure Elastic Capabilities

On-demand Capabilities

physical

Standing

Capabilities

staff

logical
NGCC Business Architecture

“Collaboratory”

• Participants
  • Academia
  • Government
  • Industry

• Contributions
  • Resources
  • Capabilities
“need to focus not on Big Data, but Big Answers”

Harper Reed – CTO Obama for America 2012
DIABESITY: THE NEW EPIDEMIC
REAL PEOPLE

Percentages and predictions can mask the enormity of the diabetes problem. Large numbers of people with diabetes are unaware they have the disease because they have not been diagnosed (shown as the shaded ridge in the country bubbles). The imperative for public-health professional is to diagnose and treat people as soon as possible.

Source: http://www.zerohedge.com (based on 2012 WHO statistics)
Obesity

- Type II Diabetes
- Cardio-vascular disease
- Liver Disease
- Cancer
Finding Solid Footing:

Inherited DNA Variation
**“Available” Obesity-related Data**

<table>
<thead>
<tr>
<th>Study Description</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMERGE Genome-Wide Association Studies of Obesity:</td>
<td>982</td>
</tr>
<tr>
<td>Population Architecture Using Genomics and Epidemiology (PAGE): Multiethnic Cohort:</td>
<td>486</td>
</tr>
<tr>
<td>The Thrifty Microbiome: The Role of the Gut Microbiota in Obesity in the Amish:</td>
<td>2802</td>
</tr>
<tr>
<td>Whole Genome Association Study of Visceral Adiposity in the Health, Aging and Body Composition (Health ABC) Study:</td>
<td>3563</td>
</tr>
<tr>
<td>Northwestern Nugene Project: Type 2 Diabetes:</td>
<td>1980</td>
</tr>
<tr>
<td>Starr County Health Studies’ Genetics of Diabetes Study:</td>
<td></td>
</tr>
<tr>
<td>A Whole Genome Association Search for Type 2 Diabetes Genes in African Americans:</td>
<td>2004</td>
</tr>
<tr>
<td>T2D-GENES Project 2: San Antonio Mexican American Family Studies:</td>
<td>2802</td>
</tr>
<tr>
<td>GENEVA Genes and Environmental Initiatives in Type 2 Diabetes (Nurses’ Health Study/ Health Professionals Follow-Up Study):</td>
<td>6033</td>
</tr>
<tr>
<td><strong>TOTAL:</strong></td>
<td><strong>21086</strong></td>
</tr>
</tbody>
</table>
BMI, age, status, glucose, diabetes, fasting, cholesterol, blood pressure, HDL, insulin, A1c, triglycerides, concentration, LDL, education, race, ethnicity, medical history, weight, hypertension, gender, family, alcohol, physical activity, smoking, diabetes type, fasting plasma glucose, hemoglobin A1c, waist-to-hip ratio, waist circumference, waist-to-height ratio, systolic blood pressure, diastolic blood pressure, cholesterol levels, HDL cholesterol, LDL cholesterol, triglycerides, blood glucose, hemoglobin, blood pressure, diabetes history, hypertension status, family history, ethnicity, gender, physical activity, smoking status, alcohol consumption.
Summary

• While much attention is focused on DNA sequence data a large, diverse collection of biomedical is also emerging

• These resources are driving the development of a new discipline – Data Science

• Data Science requires an extended definition of High Performance Computing

• ASU is constructing a first generation Data Science research platform – the NGCC
Seeking Collaborators and Partners
Come Join the Effort!
Kenneth.Buetow@ASU.edu