UAB Research Computing Day
September 15, 2011

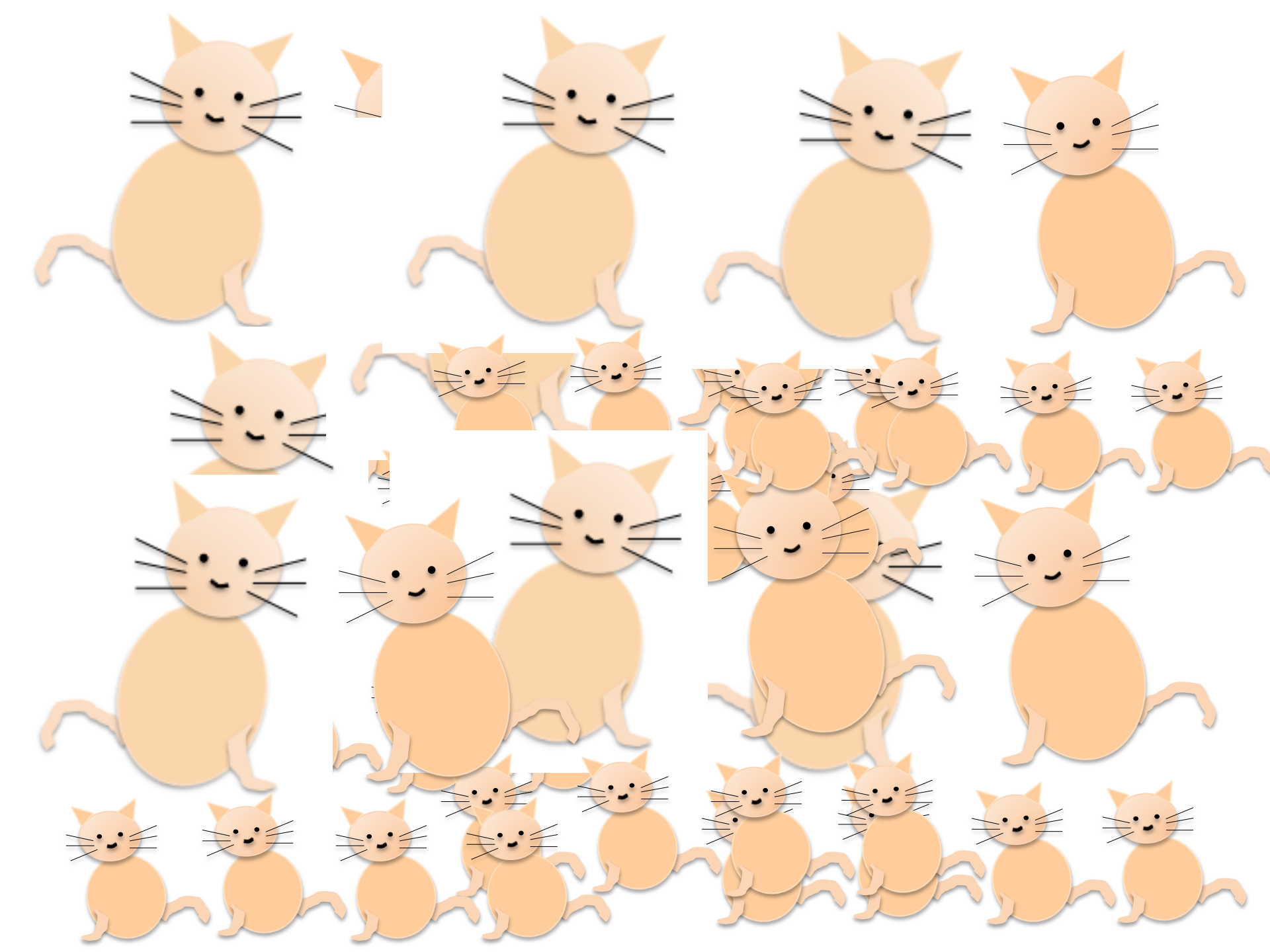
Challenges in data acquisition, storage and processing for NIH funded studies

Stephen Barnes, PhD

Department of Pharmacology & Toxicology and the Targeted Metabolomics and Proteomics Laboratory

Synopsis

- Proteomic and genomic “cats”
- Federal rules for maintaining data from funded grants
- Economics of storing and transferring data
  - Local Cloud vs Commercial Cloud
  - Media for Cloud storage



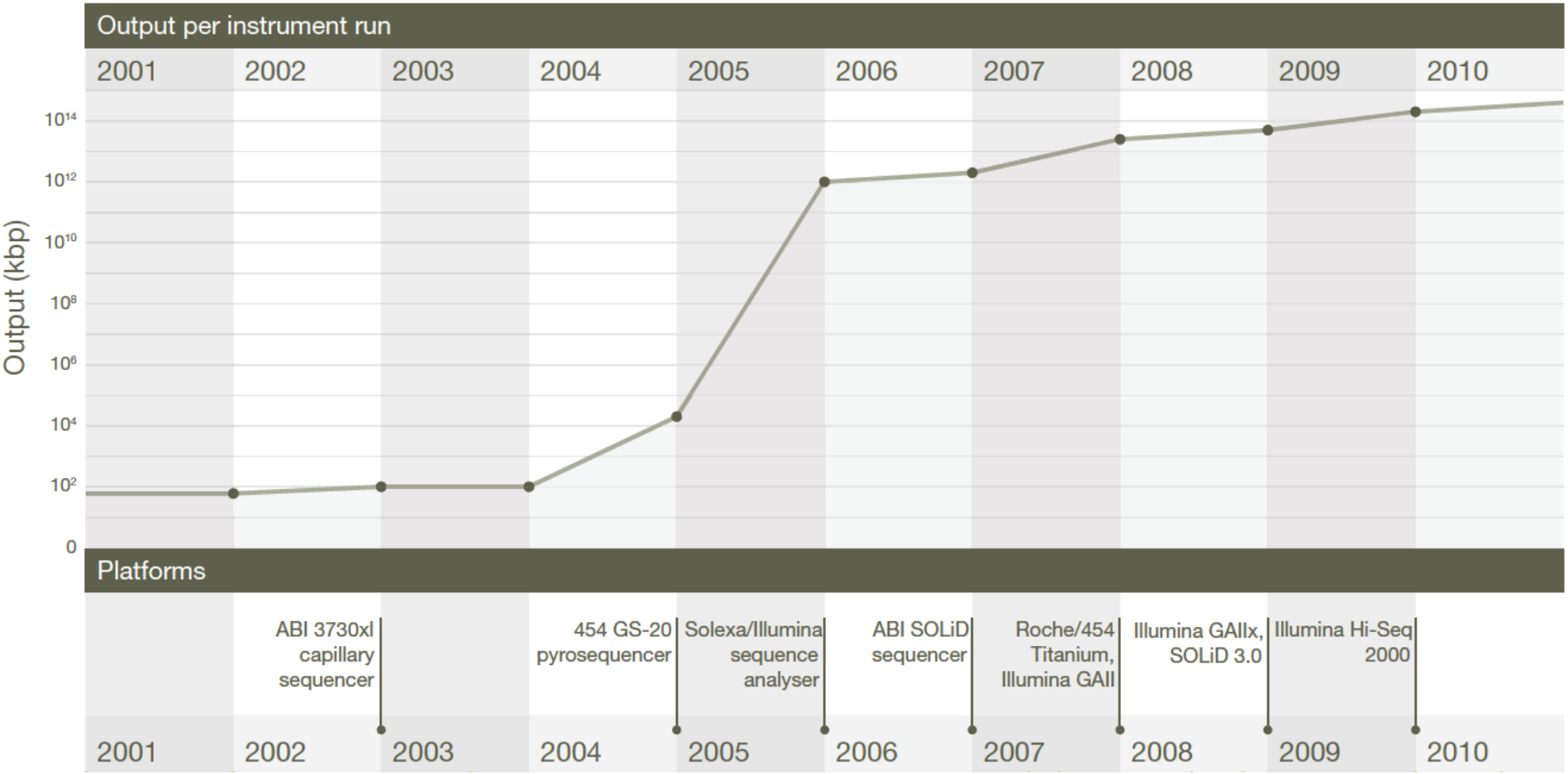
How many cats in 2011?

- **If we start with 2 cats (male and female) in 2001 and cats breed every 3 months producing a litter of 4 kittens, how many cats will we have today?**
  - **At 3 months, we'll have 4 kittens**
  - **At 6 months, we'll have 8 kittens**
  - **At 12 months we'll have 32 kittens (expanded by 16)**
  - **At 5 years, we'll have 16^5 kittens = (2^{4*5}) = one million**
  - **At 10 years, we'll have 16^{10} kittens = 2^{40}**
  - **Transfer to OMG units or as they're better known in IT departments as terabytes**

The problem

- **Analysis is exploding**
  - **Imaging**
  - **NextGen sequencing**
  - **Proteomics**
  - **Metabolomics**
- **Have created a world where there are "routine" TB datasets**

Recent increases in speed of DNA sequencing



Since 2005, DNA sequencing rates have increased 500-fold and are headed higher. The annualized increase is 4.5-fold, i.e., 22.5-fold every two years, an order of magnitude more than Moore's law in computing.

Mardis ER Nature 470:198, 2011

Expense of Deep Sequencing

- **Deep DNA sequencing is leading to Terabytes of data per month**
  - **If the genomes of the population of the USA were sequenced, this would amount to >1,000 Petabytes of data**
  - **1 Terabyte of storage is ~$100**
  - **If simple scaling is possible, then it would cost $100 million a year (expected lifetime of the drives) to store the data or approaching $7 billion for the average person's lifetime (assuming no further population increase)**
  - **And that's without backups, using the data in any way and assuming that Microsoft doesn't structure your saved files!**

NIH requirements for data collected from funded grants

- **NIH requires you to make available datasets created from federally supported studies**
- **How long should this be after termination of the grant?**
  - **Federal rules about "data" state that you must keep them for 3 years**
  - **http://grants.nih.gov/grants/policy/nihgps_2010/nihgps_ch8.htm#_Toc271264950**
  - **So, who pays for keeping TB datasets?**
  - **The investigator doesn't have financial authority once a grant is over**
    - **So, UAB?**

Does the “cloud” present a viable option?

Yes, if we can transfer the data to other computers with greater processing power and cheaper long-term storage, but……..

Whither the cloud?

- Depends on the size and capacity of the pipe from the acquisition computer to the computers in the cloud
- If we can get data to the cloud, can the software, particularly commercial software, be used for analysis in the cloud?
  - Opportunity for companies to go to a different business model where there is one, always up-to-date version of their software in the cloud and users pay a small fee for each time they use it

Principal issues posed at Bio-IT 2011

- **Can cloud infrastructure can support high-throughput data analysis pipelines designed for next-generation sequencing data?**
  - ***"The cloud is a valuable option for small research centers that lack the resources to purchase and maintain in-house infrastructure"***
  - ***"For large centers like the Broad Institute (or UAB), which require almost constant compute power to manage and move files ranging from 1 gigabyte to 1 terabyte in size, the cloud, at least for the present, does not seem to be a cost-effective option."***

Economics of the cloud

- Estimated cost of traditional storage is $3.75 per Gigabyte-month
  - Amazon cost – $0.15 Gigabyte-month
- CPU cost estimated at $2.63-$3.33 per CPU-hour
  - Microsoft Azure cost - $0.12 CPU-hour
- The cloud, if you can get there, offers substantial savings

From Virtualization and Cloud Computing – Digital Realty Trust, February 2011

Tb storage costs in the Cloud

- $0.15/GB/month → $150/TB/month
- Annual cost → $1,800/TB/year
- Commercially, “life-time storage” is $3,000
- For our group, ONE machine generates 2 TB each month
- Annualized cost $78,000
- Conclusion: Cloud HD storage is not viable

Other models to consider

- Do we really need to have high speed access to old data?
- Is a tape back up system viable?
  - Google still uses it as their long-term storage system
  - One tenth of the costs of HD storage
  - This reduces 1 TB storage to $15/month or $180/year

Costs of tape storage

http://www.bitbunker.com/pricing

Pricing

Our pricing is up to 10x less expensive than competing cloud storage products because we offer a fundamentally different approach to storage. See what features make this possible.

Monthly	Storage	Additional
$49.95	600 GB	< 1¢ / GB

If your usage surpasses what is included for free, you are billed the rates detailed below.

Provision (new files)	Persist (stored files)	Transfer (in & out)	PUTs	GETs
$99.95 / TB same as 9.761¢ / GB	$9.95 / TB same as 0.972¢ / GB*	10¢ / GB	Free	5¢ / Each

* Not a typo. Store your data for less than 1¢ / GB / month.

If an investigator uploads a 200 GB file, this costs $20. It also costs $20 each time it is downloaded. This will be a cost borne by the investigator.

The robotic tape storage system

Bunker Alpha

Model
StorageTek L700e
Cells
678
Drives
3
Media
Ultrium LTO-4
Capacity
1.08 Petabytes

Status
✔ Operational

Read Latency
< 3.0 mins
* Time to first byte over last 1 month.

Started	Finished	Duration	Seek / Data Ratio	Drive	Volume	Tasks	Data Read	Data Written
2011-08-24 11:57:05	3 hrs ago	3 mins	60% / 40%	1	A 0139	1	—	303 MB

Download time is consistent with the time to get a coffee or a Coke, or take a quick bathroom break.

Summary and the future

- **Biomedical research is generating *volumes* of data that strain the current system for data transport and storage**
  - **There are two options**
    - **Gaining access to very fast pipes from UAB NextGen and other UAB data generating centers to existing regional fast pipes and on to the commercial cloud**
    - **Creating very fast pipes from UAB NextGen and other UAB data generating centers to a UAB cloud**
- **Storage costs of large data sets has become an economic heavyweight**
  - **Is a tape system the solution for NIH data?**
- **Software may not be transferrable to the cloud**
- **Security issues need good solutions for all parties**

Acknowledgements

- David Shealy, PhD
- Chiquito Crasto, PhD
- Jonas Almeida, PhD
- John-Paul Robinson
- Scott Sweeney
- Landon Wilson
- Mikako Kawai
- Chandrahas Narne

- NCCAM R21 AT004661
- NCRR S10 RR027822