View From The Trenches

2010 AIRI Petabyte Challenge

BIOTEAM
Enabling Science

Housekeeping Notes

- Some Acknowledgements
- Fair warning #1
  - I speak fast and travel with a large slide deck
- Fair warning #2
  - Unrepentant PowerPoint fiddler
  - Latest slides (as delivered) will be posted at http://blog.bioteam.net

BIOTEAM
Enabling Science

Topics

Science-driven Storage
IT exists to enable the researcher

Field Observations
Trends & war stories

Tips & Tricks
An attempt at some practical advice

BIOTEAM
Enabling Science

BioTeam Inc.

- **Independent Consulting Shop:**
  Vendor/technology agnostic

  Staffed by:
  - Scientists forced to learn High Performance IT to conduct research

- **Our specialty:**
  Bridging the gap between Science & IT

BIOTEAM
Enabling Science

A brief note on our client base

- Very few of our customers are in this room
  - *With a few cool exceptions, of course*
- The fact that you are here today speaks volumes
  - Chances are:
    - You have forward-looking research IT roadmaps
    - You have dedicated research IT staff
    - You have dedicated storage gurus
    - You have research datacenter(s)
- With a few notable exceptions, many of our customers do not have the level of expertise, experience and resources that an AIRI affiliated lab would have
- This tends to bias what I say and think

BIOTEAM
Enabling Science

Science Driven Storage

BIOTEAM
Enabling Science

Photo Tour - Lab Local / Single Instrument

Self-contained lab-local cluster & storage for Illumina

BIOTEAM
Enabling Science

Photo Tour: Single Lab Solution

100 Terabyte storage system and 40 core Linux Cluster supporting multiple instruments in a single lab

BIOTEAM
Enabling Science

Photo Tour: Large Genome Center

BIOTEAM
Enabling Science

Setting the stage

- Data Awareness
- Data Movement
- Data Management

BIOTEAM
Enabling Science

The Stakes …

180+ TB stored on lab bench

The life science “data tsunami” is no joke.

BIOTEAM
Enabling Science

Flops, Failures & Freakouts

How we've seen storage go bad …

BIOTEAM
Enabling Science

#1 - Unchecked Enterprise Architects

- Scientist: "*My work is priceless, I must be able to access it at all times*"
- Storage Guru: "*Hmmm…you want H/A, huh?"*

- *System delivered:*
  - 40TB Enterprise FC SAN
  - Asynchronous replication to remote DR site
  - Can't scale, can't do NFS easily
  - $500K/year in maintenance costs

BIOTEAM
Enabling Science

#1 - Unchecked Enterprise Architects

- Lessons learned
- Corporate storage architects may not fully understand the needs of HPC and research informatics users
- End-users may not be precise with terms:
  - “Extremely reliable” means “no data loss”, not 99.999% uptime at a cost of millions
- When true costs are explained:
  - Many research users will trade a small amount of uptime or availability for more capacity or capabilities

BIOTEAM
Enabling Science

#2 - Unchecked User Requirements

- Scientist: "*I do bioinformatics, I am rate limited by the speed of file IO operations. Faster disk means faster science.* "
- System delivered:
  - Budget blown on top tier 'Cadillac' system
  - Fast *everything*
- Outcome:
  - System fills to capacity in 9 months

BIOTEAM
Enabling Science

#2 - Unchecked User Requirements

- Lessons learned
  - End-users demand the world
  - Necessary to really talk to them and understand their work, needs and priorities
- You will often find
  - The people demanding the “fastest” storage don’t have actual metrics to present
  - Many groups will happily trade some level of performance in exchange for a huge win in capacity or capability

BIOTEAM
Enabling Science

#3 - D.I.Y Cluster/Parallel File systems

- Common source of storage unhappiness

- Root cause:
  - Not enough pre-sales time spent on design and engineering
- System as built:
  - Not enough metadata controllers
  - Poor configuration of key components
- End result:
  - Poor performance or availability

BIOTEAM
Enabling Science

#3 - D.I.Y Cluster/Parallel File systems

- Lessons learned:
- Software-based parallel or clustered file systems are non-trivial to *correctly* implement
- Essential to involve experts in the initial design phase
  - *Even if using 'open source' version …*
- Commercial support is essential
  - *And I say this as an open source zealot …*

BIOTEAM
Enabling Science

Science Driven Storage

Back on track …

BIOTEAM
Enabling Science

Data Awareness

- First principals:
  - Understand science changes faster than IT
  - Understand the data you will *produce*
  - Understand the data you will *keep*
  - Understand how the data will *move*
- Second principals:
  - One research type or many?
  - One instrument type or many?
  - One lab/core or many?

BIOTEAM
Enabling Science

Data You Produce

- Important to understand data sizes and types throughout the organization
  - 24x7 core facility with known protocols?
  - Wide open “discovery research” efforts?
  - Mixture of both?

- Where it matters:
  - Big files or small files?
  - File types & access patterns?
  - Hundreds, thousands or millions of files?
  - Does it compress well?
  - Does it deduplicate well?
  - Where does the data have to move?

BIOTEAM
Enabling Science

Data You Will Keep

- Instruments producing terabytes/run are the norm, not the exception
- Data triage is real and here to stay
  - Triage is the norm, not the exception these days
  - I think the days of “unlimited storage” are likely over
- *What bizarre things are downstream researchers doing with the data ?*
- Must decide what data types are kept
  - And for how long …

BIOTEAM
Enabling Science

Data You Will Keep

- Raw data ⇒ Result data
  - Can involve 100x reduction in some cases
- Result data ⇒ Downstream derived data
  - Often overlooked and trend-wise the fastest growing area
  - Researchers have individual preferences for files, formats and meta-data
  - Collaborators have their own differences & requirements
  - The same data can be sliced and diced in many ways when used by different groups

BIOTEAM
Enabling Science

Data Movement

- Facts
  - Data captured does not stay with the instrument
  - Often moving to multiple locations (and offsite)
  - Terabyte volumes of data could be involved
  - Multi-terabyte data transit across networks is rarely trivial no matter how advanced the IT organization
  - Campus network upgrade efforts may or may not extend all the way to the benchtop …

BIOTEAM
Enabling Science

Data Movement - Personal Story

- One of my favorite '09 consulting projects …
  - Move 20TB scientific data out of Amazon S3 storage cloud
- What we experienced:
  - Significant human effort to swap/transport disks
  - Wrote custom DB and scripts to verify all files each time they moved
    - Avg. 22MB/sec download from internet
    - Avg. 60MB/sec server to portable SATA array
    - Avg. 11MB/sec portable SATA to portable NAS array
  - At 11MB/sec, moving 20TB is a matter of *weeks*
  - *Forgot to account for MD5 checksum calculation times*
- Result:
  - Lesson Learned: data movement & handling took 5x longer than data acquisition

BIOTEAM
Enabling Science

Things To Think About

An attempt at some practical advice …

BIOTEAM
Enabling Science

Storage Landscape

- Storage is a commodity
- Cheap storage is easy
- Big storage getting easier every day
- Big, cheap & *SAFE* is much harder …
- Traditional backup methods may no longer apply
  - Or even be possible …

BIOTEAM
Enabling Science

Storage Landscape

- Still see extreme price ranges
  - Raw cost of 1,000 Terabytes (1PB):
    - $125,000 to $5,000,000 USD
- Poor product choices exist in all price ranges

BIOTEAM
Enabling Science

Poor Choice Examples

- On the low end:
  - Use of RAID5 (unacceptable in since 2008)
  - Too many hardware shortcuts result in unacceptable reliability trade-offs

BIOTEAM
Enabling Science

Poor Choice Examples

- And with high end products:
  - Feature bias towards corporate computing, not research computing - pay for many things you won't be using
  - Unacceptable hidden limitations (size or speed)
  - Personal example:
    - $800,000 70TB (raw) Enterprise NAS Product
    - *… can't create a NFS volume larger than 10TB*
    - *… can't dedupe volumes larger than 3-4 TB*

BIOTEAM
Enabling Science

One slide on RAID 5

- I was a RAID 5 bigot for many years
  - Perfect for life science due to our heavy read bias
  - Small write penalty for parity operation no big deal
- RAID 5 is no longer acceptable
  - Mostly due to drive sizes (1TB+), array sizes and rebuild time
  - *In the time it takes to rebuild an array after a disk failure there is a non-trivial chance that a 2nd failure will occur, resulting in total data loss*
- Today:
  - Only consider products that offer RAID 6 or other “double parity” protection methods
  - Even RAID 6 is a stopgap measure …

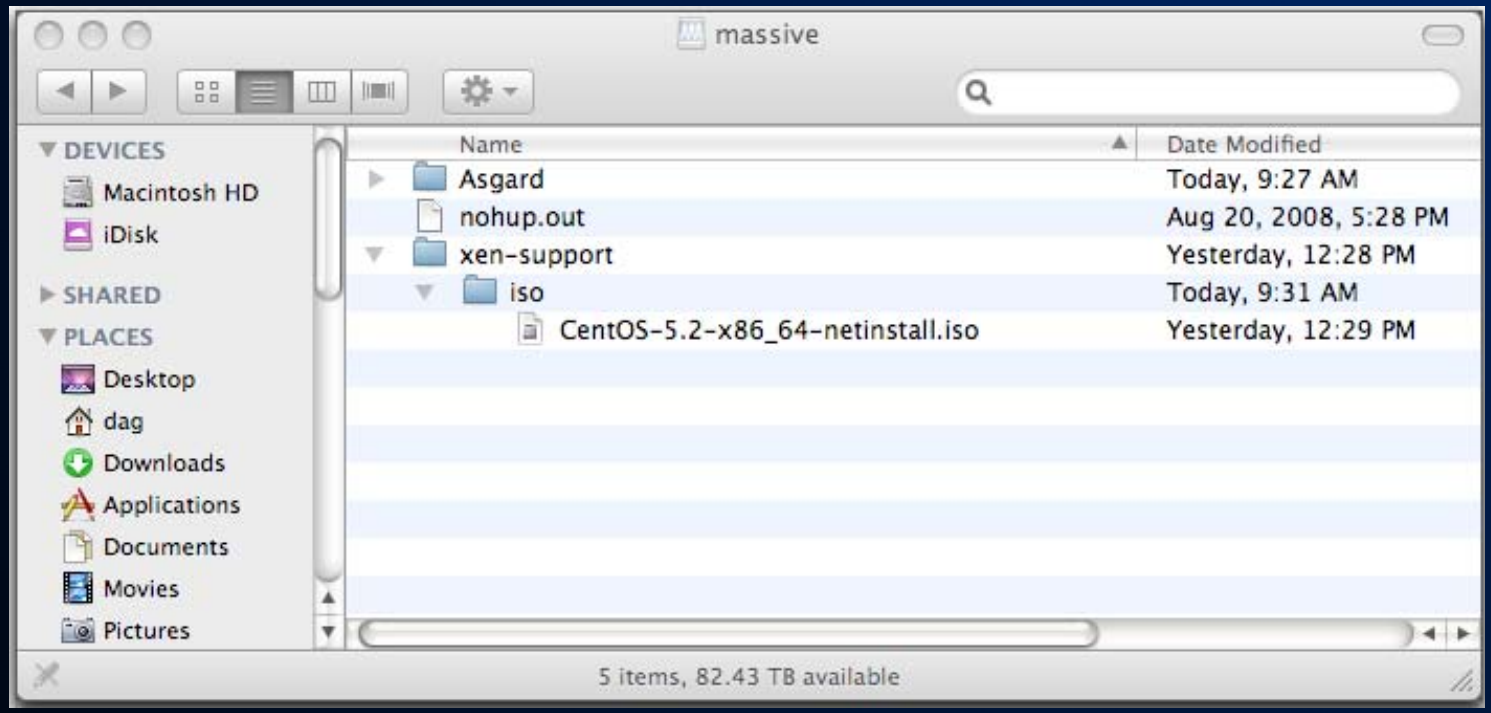
BIOTEAM
Enabling Science

Observations & Trends

BIOTEAM
Enabling Science

Trends: Single Namespace

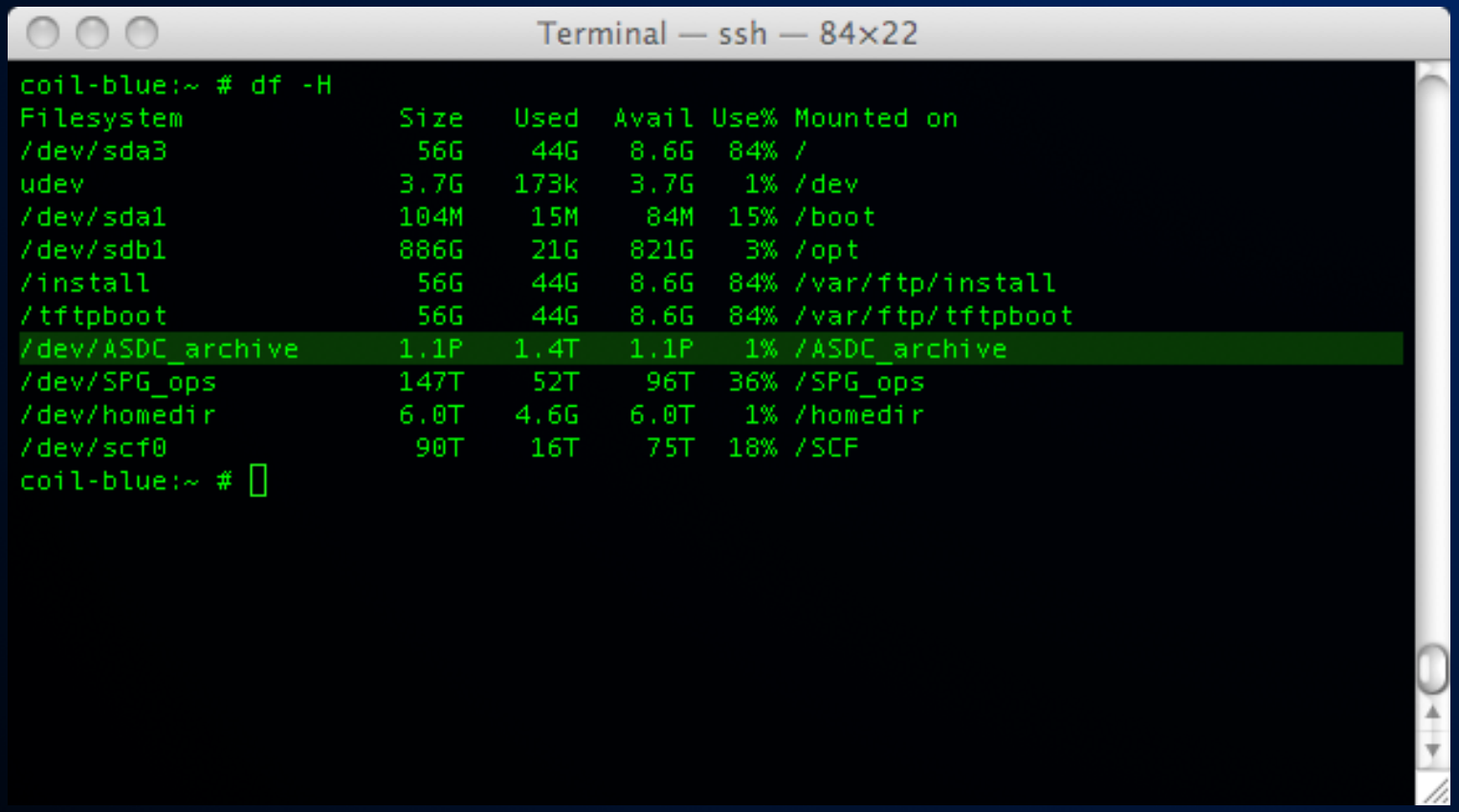
- 82TB - Very Satisfying



BIOTEAM
Enabling Science

Trends: Single Namespace

- 1PB - *More Satisfying*



BIOTEAM
Enabling Science

Single Namespace Matters

- Non-scalable storage islands add complexity
- Also add “data drift”

- Example:
  - Volume “Caspian” hosted on server “Odin”
  - “Odin” replaced by “Thor”
  - “Caspian” migrated to “Asgard”
  - Relocated to “/massive/”

- Resulted in file paths that look like this:

```
/massive/Asgard/Caspian/blastdb
/massive/Asgard/old_stuff/Caspian/blastdb
/massive/Asgard/can-be-deleted/do-not-delete…
```

BIOTEAM
Enabling Science

User Expectation Management

- End users still have no clue about the true costs of keeping data accessible & available

- "I can get a terabyte from Costco for $220!" (Aug 08)
- "I can get a terabyte from Costco for $160!" (Oct 08)
- "I can get a terabyte from Costco for $124!" (April 09)
- "I can get a terabyte from NewEgg for $84!" (Feb 10)

- IT needs to be involved in setting expectations and educating on true cost of keeping data online & accessible

BIOTEAM
Enabling Science

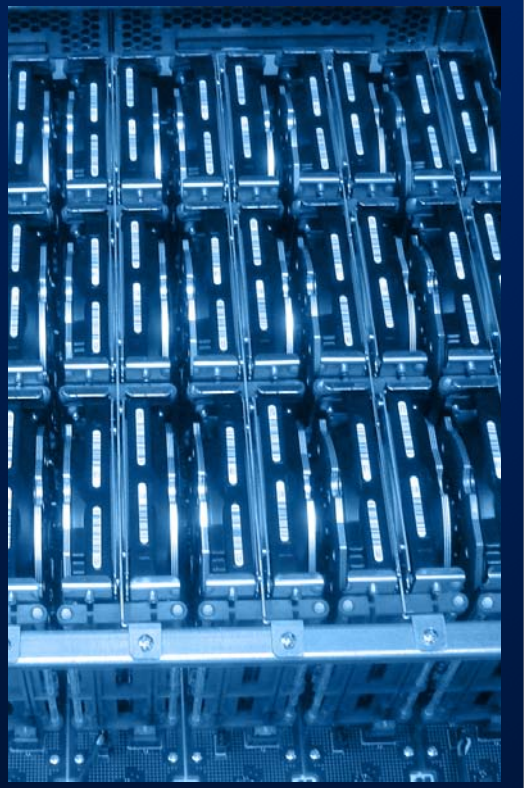
Storage Trends

- In 2008 …
  - First 100TB single-namespace project
  - First Petabyte+ storage project
  - 4x increase in “technical storage audit” work
  - First time witnessing 10+TB catastrophic data loss
  - First time witnessing job dismissals due to data loss
  - Data Triage discussions are spreading well beyond cost-sensitive industry organizations

BIOTEAM
Enabling Science

Storage Trends

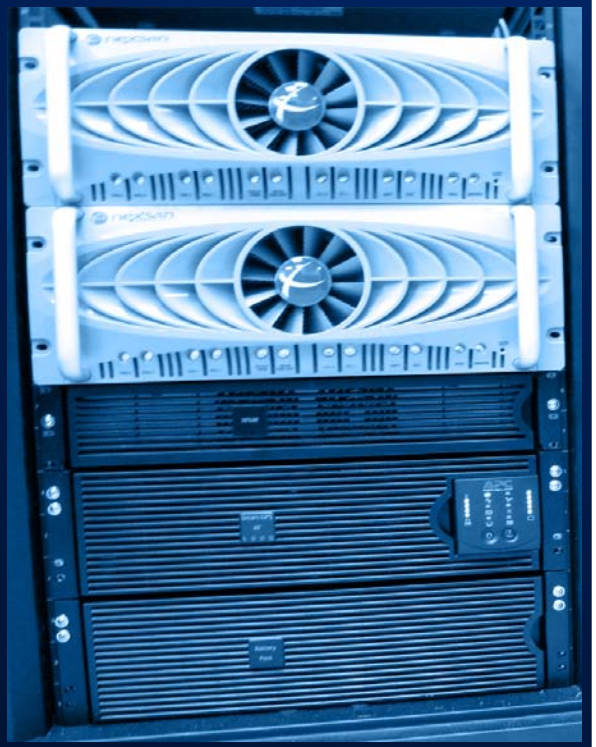
- In 2009 …
  - More of the same
  - 100TB not a big deal any more
  - Even smaller organizations are talking (or deploying) petascale storage
- Witnessed spectacular failures of Tier 1 storage vendors:
  - $6M 1.1PB system currently imploding under a faulty design.
  - $800K NAS product that can't supply a volume larger than 10TB
    - Even less with dedupe enabled



BIOTEAM
Enabling Science

Going into 2010 …

- Peta-scale is no longer scary
- A few years ago 1PB+ was somewhat risky and involved significant engineering, experimentation and crossed fingers
  - Especially single-namespace
- Today 1PB is not a big deal
  - Many vendors, proven architectures
  - Now it's a capital expenditure, not a risky technology leap



BIOTEAM
Enabling Science

Going into 2010 …

- Biggest Trend
  - Significant rise in storage requirements for post-instrument downstream experiments and mashups
  - The decrease in instrument generated data flows may be entirely offset by increased consumption from users working downstream on many different efforts & workflows
    - *… this type of usage is harder to model & predict*

BIOTEAM
Enabling Science

Cloud Storage

I’m a believer (maybe)

BIOTEAM
Enabling Science

Why I drank the kool-aid

- I am known to be rude and cynical when talking about over hyped “trends” and lame cooption attempts by marketing folk
  - Wide-area Grid computing is an example from dot com days
  - “Private Clouds” - another example of marketing fluff masking nothing of actual useful value in 2010
- I am also a vocal cheerleader for things that help me solve real customer-facing problems
  - *Cloud storage might actually do this …*

BIOTEAM
Enabling Science

Cloud Storage: Use Case 1

- Amazon AWS “downloader pays” model is extremely compelling
- Potentially a solution for organizations required to make large datasets available to collaborators or the public
  - Costs of local hosting, management & public bandwidth can be significant resource drain
  - Cloud-resident data sets where the downloader offsets or shares in the distribution cost feels like a good match

BIOTEAM
Enabling Science

Cloud Storage: Use Case 2

- Archive, deep or cold storage pool
- Imagine this scenario:
  - Your 1PB storage resource can't be backed up via traditional methods
  - Replication is the answer
  - However just to be safe you decide you need:
    - Production system local to campus
    - Backup copy at Metro-distance colo
    - Last resort copy at WAN-distance colo
  - Now you have 3PB to manage across three different facilities
    - *Non trivial human, facility, financial and operational burden costs …*

BIOTEAM
Enabling Science

Cloud Storage: Use Case 2

- James Hamilton has blogged some interesting figures
  - Site: http://perspectives.mvdirona.com
  - Cold storage geographically replicated 4x can be achieved *at scale* for $.80 GB/year (and falling quickly)
  - With an honest accounting of all your facility, operational and human costs can you *really* approach this figure?

BIOTEAM
Enabling Science

Cloud Storage: Use Case 2

- Google, Amazon, Microsoft, etc. all operate at efficiency scales that few can match
  - Cutting-edge containerized data-centers with incredible PUE values
  - Fast private national and trans-national optical networks
  - Rumors of "1 human per XX,000 servers" automation efficiency, etc.
  - Dozens or hundreds of datacenters and exabytes of spinning platters
  - My hypothesis:
    - Not a single person in this room can come anywhere close to the IT operating efficiencies that these internet-scale companies operate at every day
    - Someone is going to eventually make a compelling service/product offering that leverages this …

BIOTEAM
Enabling Science

Cloud Storage: Use Case 2

- Cheap storage is easy, we all can do this
- Geographically replicated, efficiently managed cheap storage is not very easy (or not cheap)

- When the price is right …
- I see cloud storage as being a useful archive or deep storage tier
  - Probably a 1-way transit
  - Data only comes “back” if a disaster occurs
  - Data mining & re-analysis done in-situ with local ‘cloud’ server resources if needed

BIOTEAM
Enabling Science

Final Thoughts

- Yes the “data deluge” problem is real
- Many of us have peta-scale storage issues today
- “Data Deluge” & “Tsunami” are apt terms
- But:
  - The problem does not feel as scary as it once did
  - Many groups have successfully deployed diverse types of peta-scale storage systems - Best practice info is becoming available
  - Chemistry, reagent cost, date movement & human factors are natural bottlenecks
  - Data Triage is an accepted practice, no longer heresy
  - Data-reduction starting to happen within instruments
  - Customers starting to trust instrument vendor software more
  - We see large & small labs dealing successfully with these issues
  - Many ways to tackle IT requirements

BIOTEAM
Enabling Science

End;

- Thanks!
- Comments/feedback:
  - chris@bioteam.net

BIOTEAM
Enabling Science