The BioTeam

The BioTeam

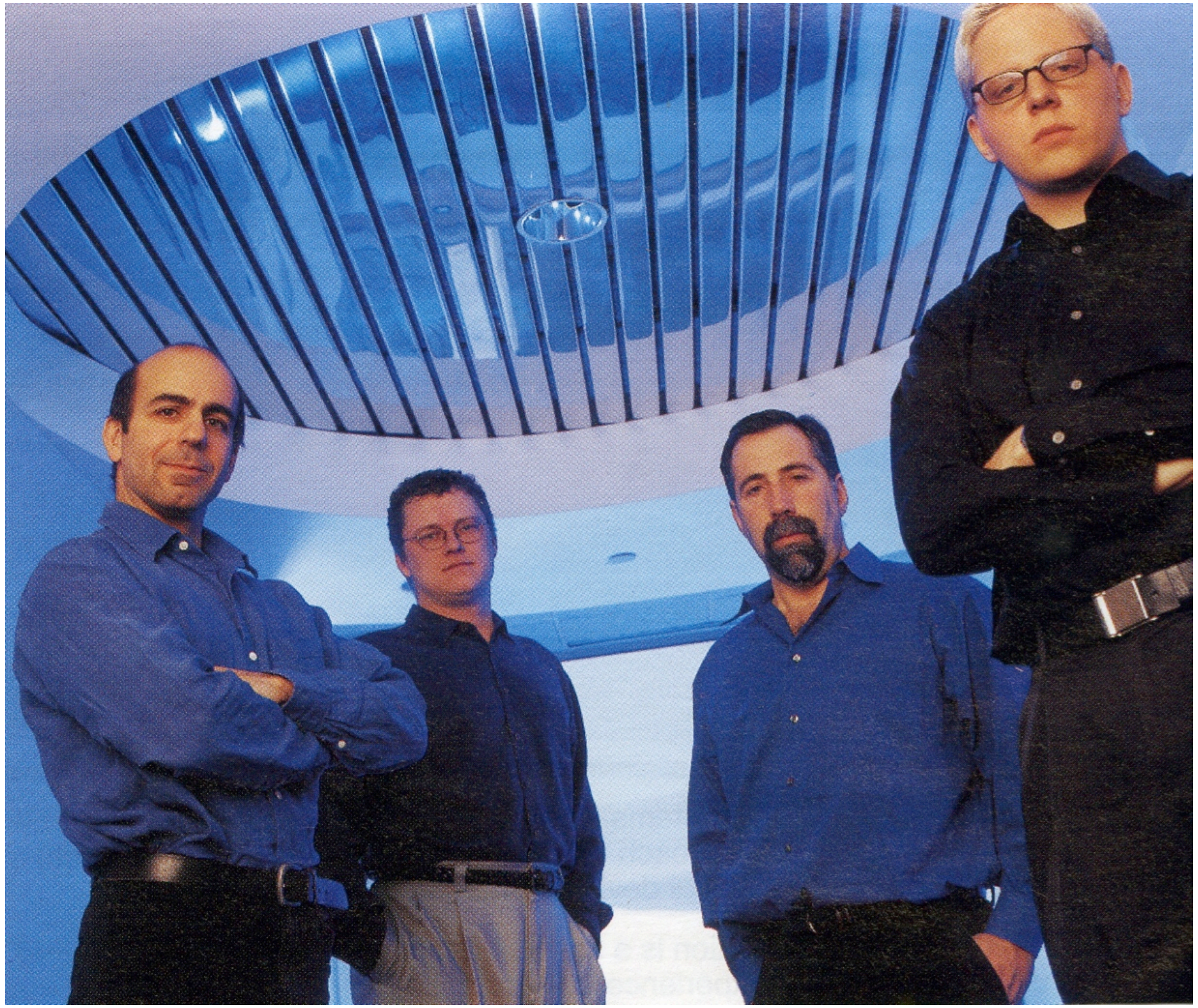
A SysAdmin’s Intro to Bioinformatics

William Van Etten, PhD
The BioTeam, Inc.
November 18, 2004

The BioTeam, Inc.

Who are they?

- Scientists, Developers, IT Professionals
- Objective, vendor agnostic professional services
- Principals
  - Michael Athanas
  - Chris Dagdigian
  - Stan Gloss
  - William Van Etten



http://bioteam.net/

Session Objectives

25 min	Genetics to Genomics
10 min	Informatics Algorithms Compared
10 min	Clustering for Informatics
15 min	iNquiry Demo "Live"
30 min	Q & A

Genetics to Genomics

Genetics to Genomics

- 1600's- Europe Emerges from the Dark Ages
- 1866- Genetic Theory Published (Mendel)
- 1869- DNA Discovered (Miescher)
- 1952- DNA is Genetic Material (Hershey)
- 1953- DNA Helical Structure Determined (W&C, Franklin)
- 1959- Protein Structure Determined (Perutz, Kendrew)
- 1966- Genetic Code (Nirenberg, Khorana)
- 1977- DNA Sequenced (Sanger)
- 1988- Human Genome Project Started
- 2001- Human Genome Draft Finished

Genetics Trivia

- “Everything you are, is either protein or the result of protein action.”
- Proteins: folded strings of Amino Acids (20)
- Genes: [regions of DNA that define a protein]
- 3 billion DNA letters (4) in the human genome
- ~5% of human DNA contain genes (~35K)
- 999/1000 DNA identity between any two people
- Human genes are 98% similar to those of a Chimpanzee
- Human genes are 50% similar to those of a Banana

Origin of Life

Francesco Redi: 1626-1697

- Prevailing Theory "Spontaneous Generation"
- Life arises spontaneously from non-living matter
- Meat makes maggots, Straw makes mice…
  - Meat in two jars, one open one sealed.
  - Observe flies -> eggs -> maggots -> flies
  - nothing happens to the closed jar meat
- Inference: Flies make flies

Father of Genetics

Gregor Mendel (1822-1884)

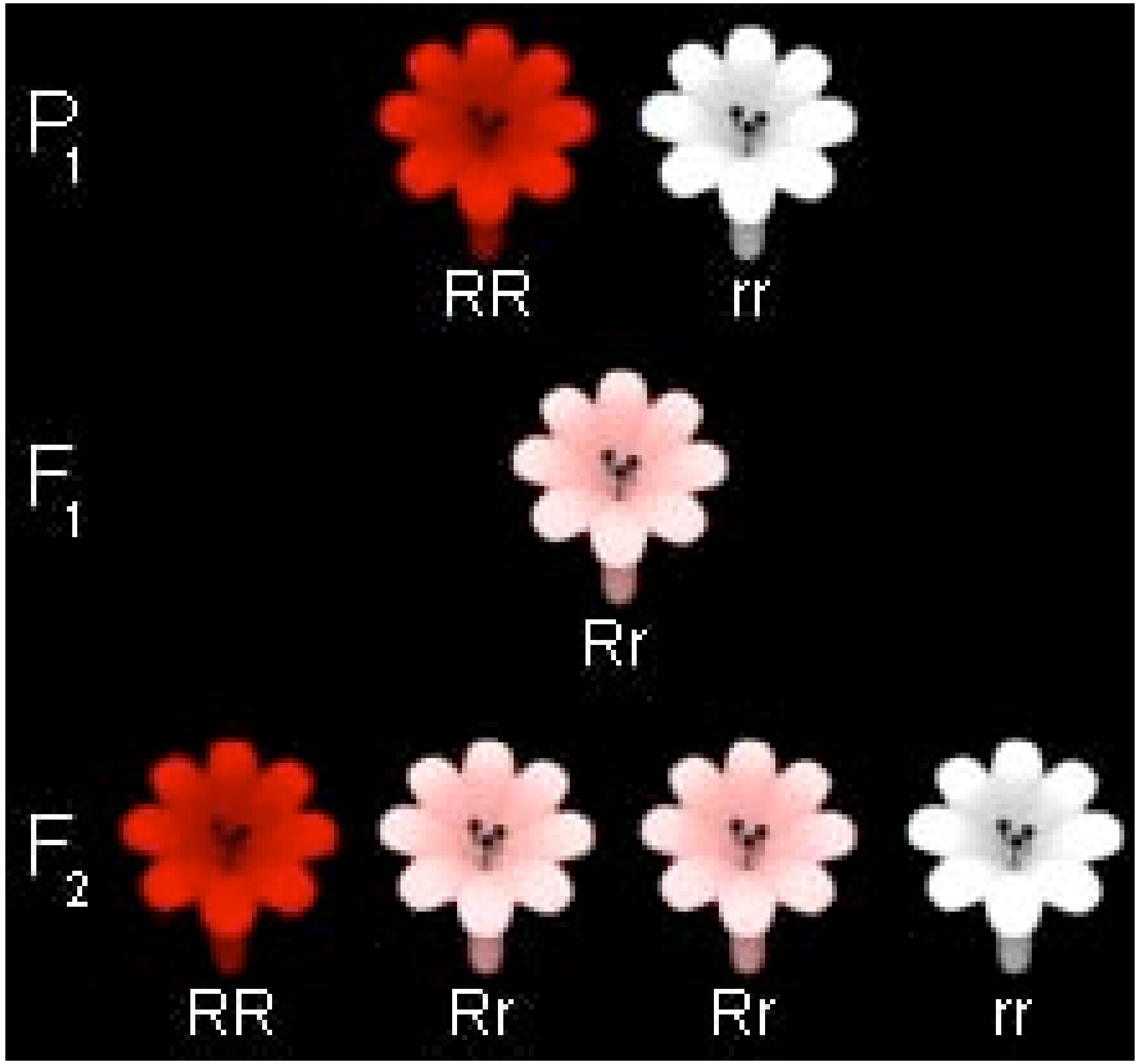
- Monk, Interested in math & gardening
- Selectively bred pea plants
  - 28,000 plants over 7 years
  - 7 distinct traits
- Studied one characteristic at a time:
  - Pea shape, color, seed-coat, flower color...
- Kept pedigrees and made several generations of crosses
- Kept track of the number and type of progeny from each cross

Mendelian Genetics

- Genetic "factors" (**genes**) determine phenotypic traits.

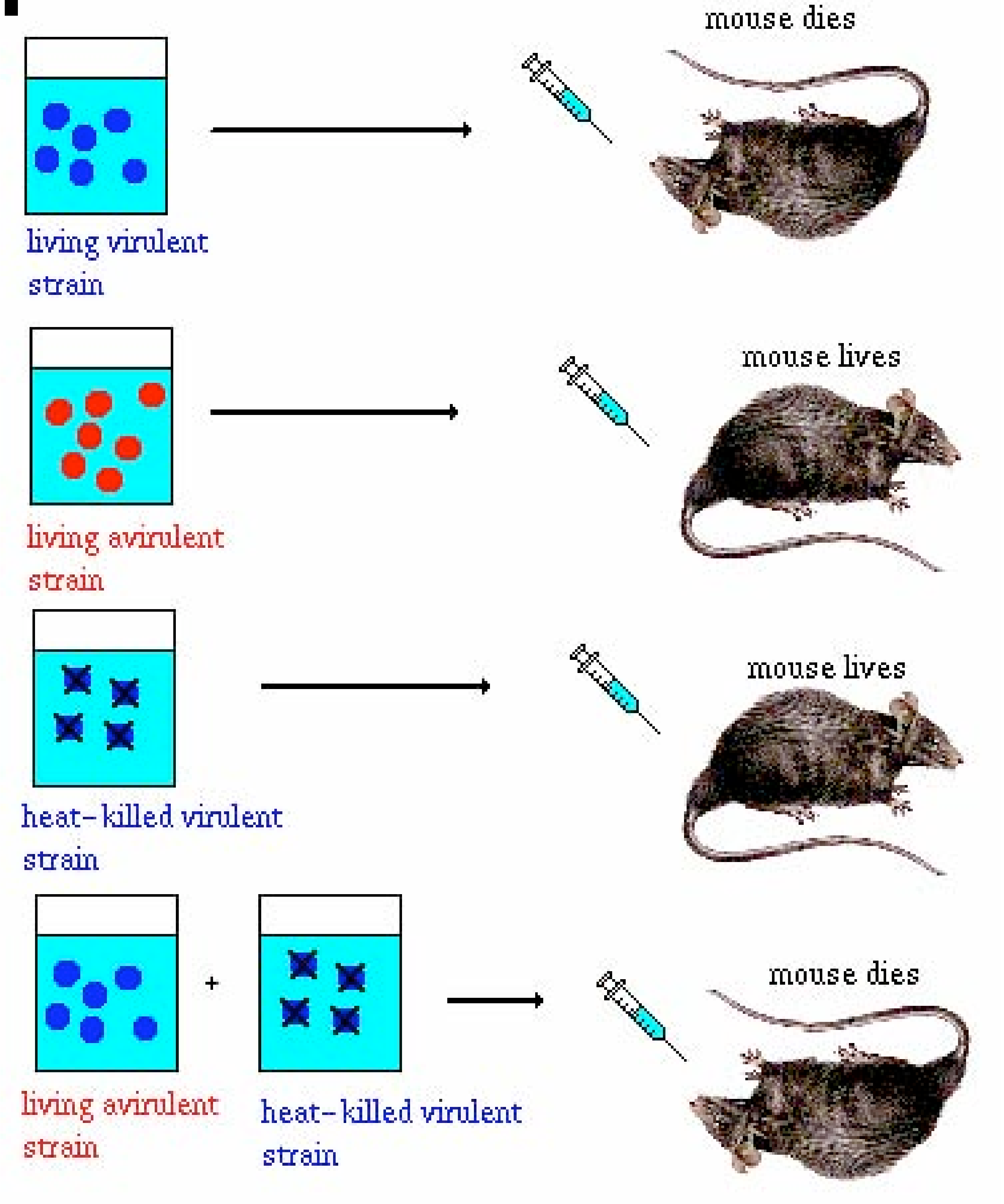
- Each organism has two instances (**alleles**) of each gene.

- Independent assortment: One copy from from each parent is (**selected at random**) is passed on to each progeny.



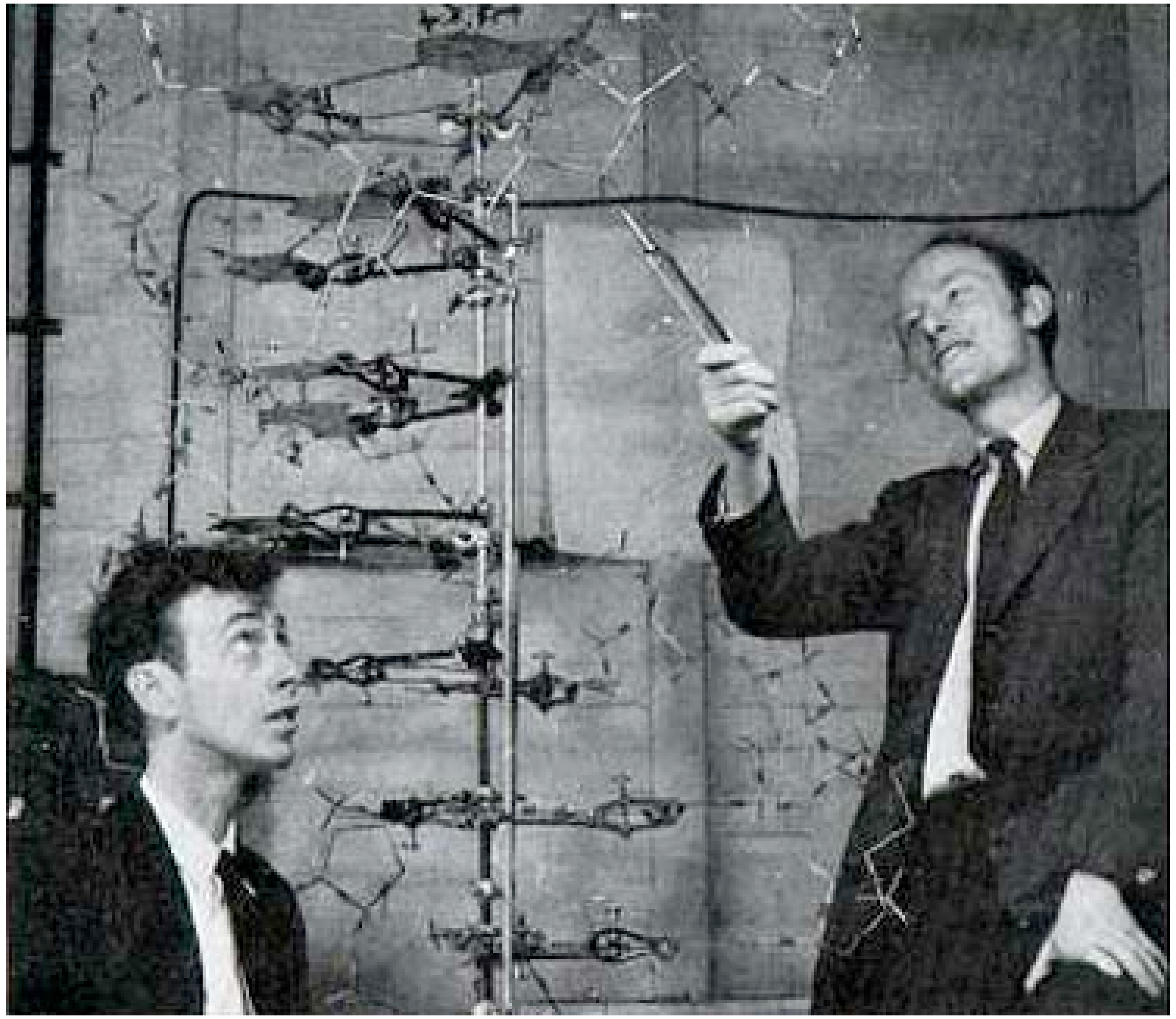
DNA is the Genetic Material

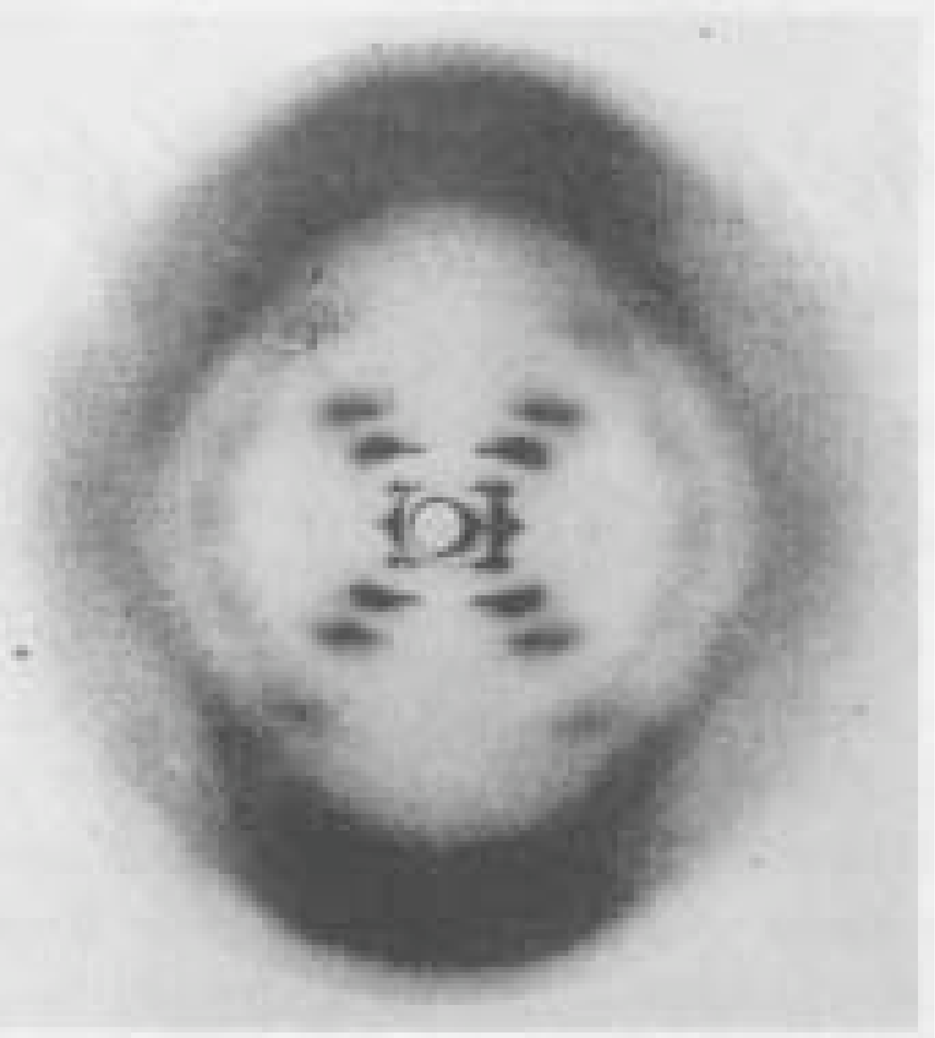
- 1943: Oswald Avery et. al. sacrifice mice to demonstrate that DNA could be the material for genes. (to one part in 6x108)
- 1952: Alfred Hershey and Martha Chase use viruses to prove it.
- "Perhaps we will be able to grind genes in a mortar and cook them in a beaker after all."
  -Hermann Muller



DNA Structure is a Double Helix

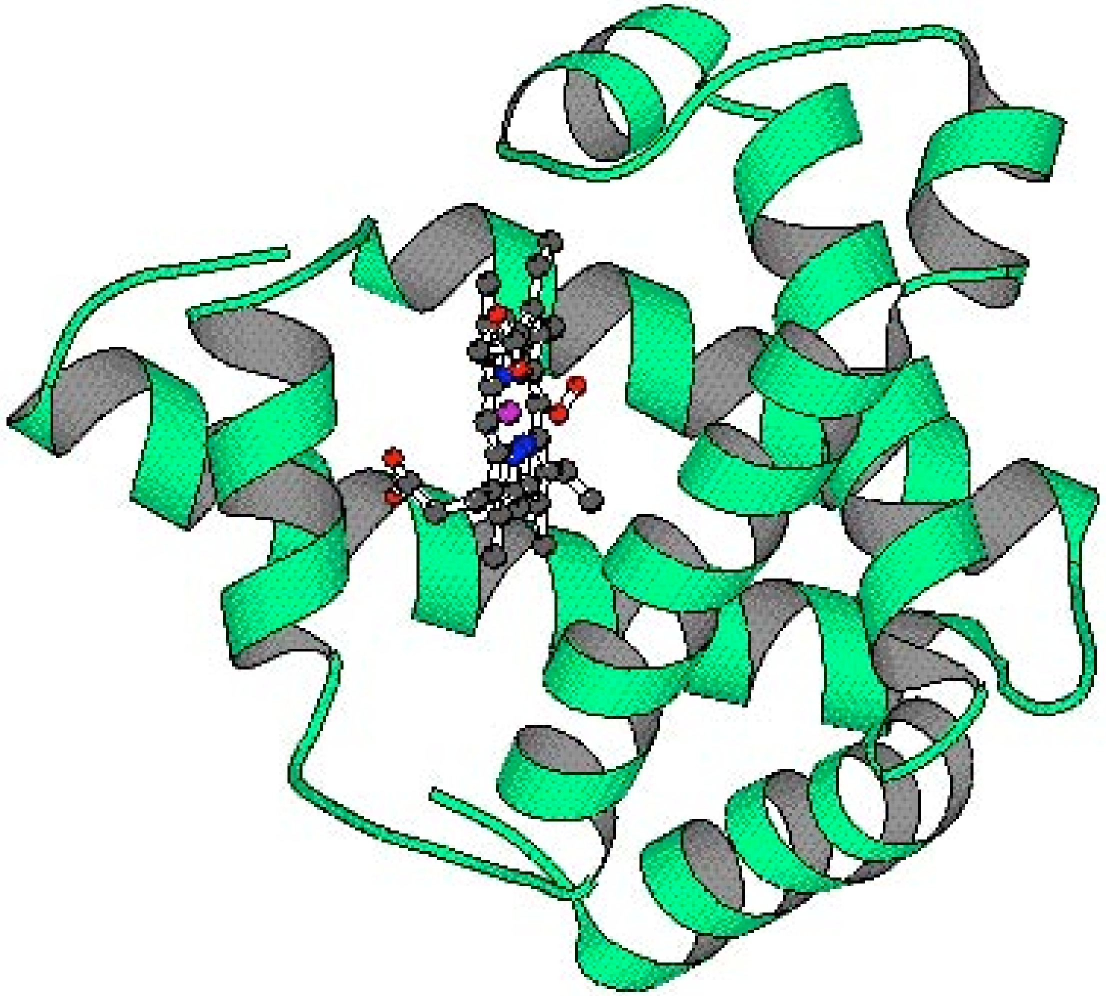
- 1953 - 3D Structure of DNA
  - Watson & Crick - model
  - Wilkins & Franklin -x-ray structure
  - Nobel in 1962

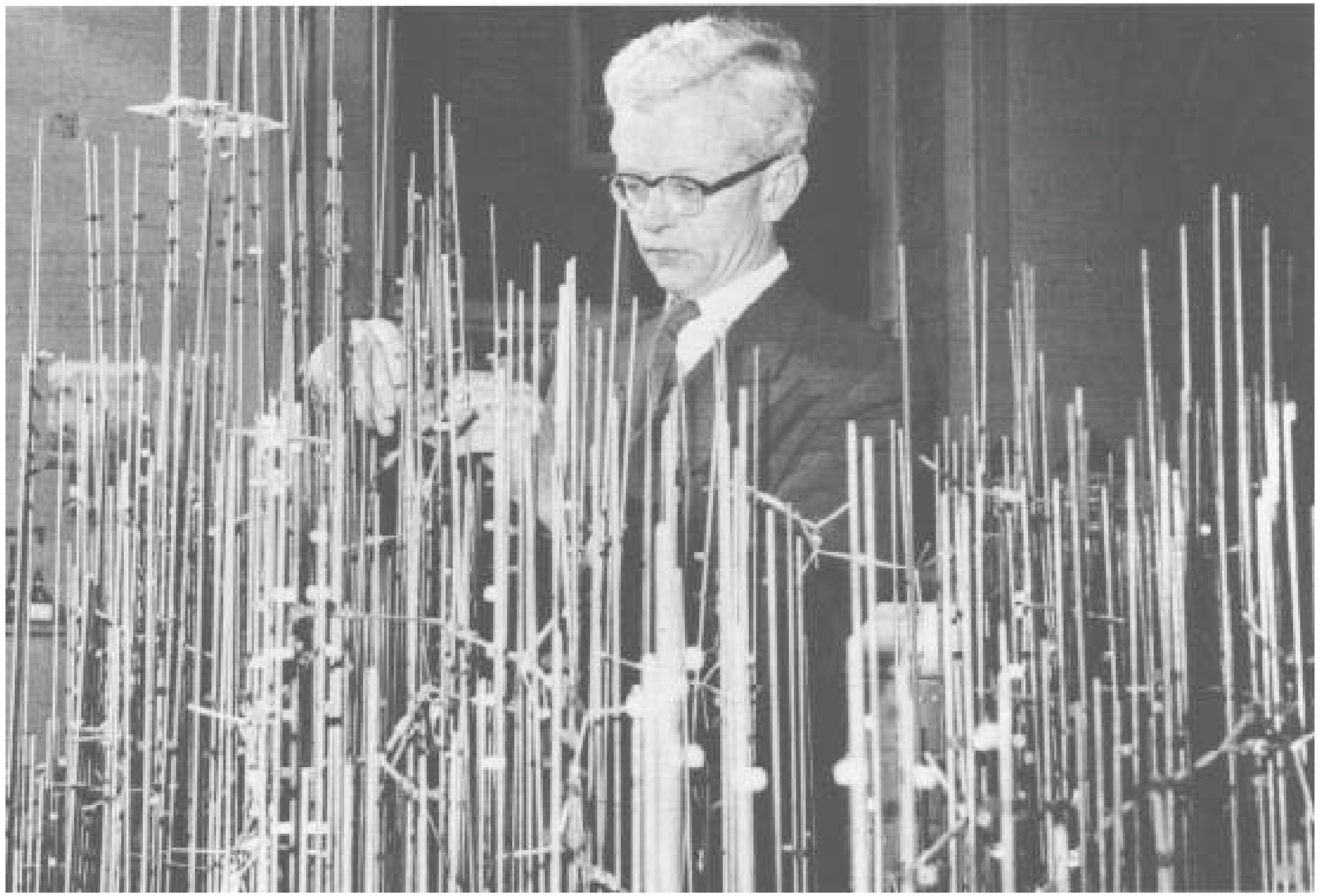




3D Protein Structure Determined

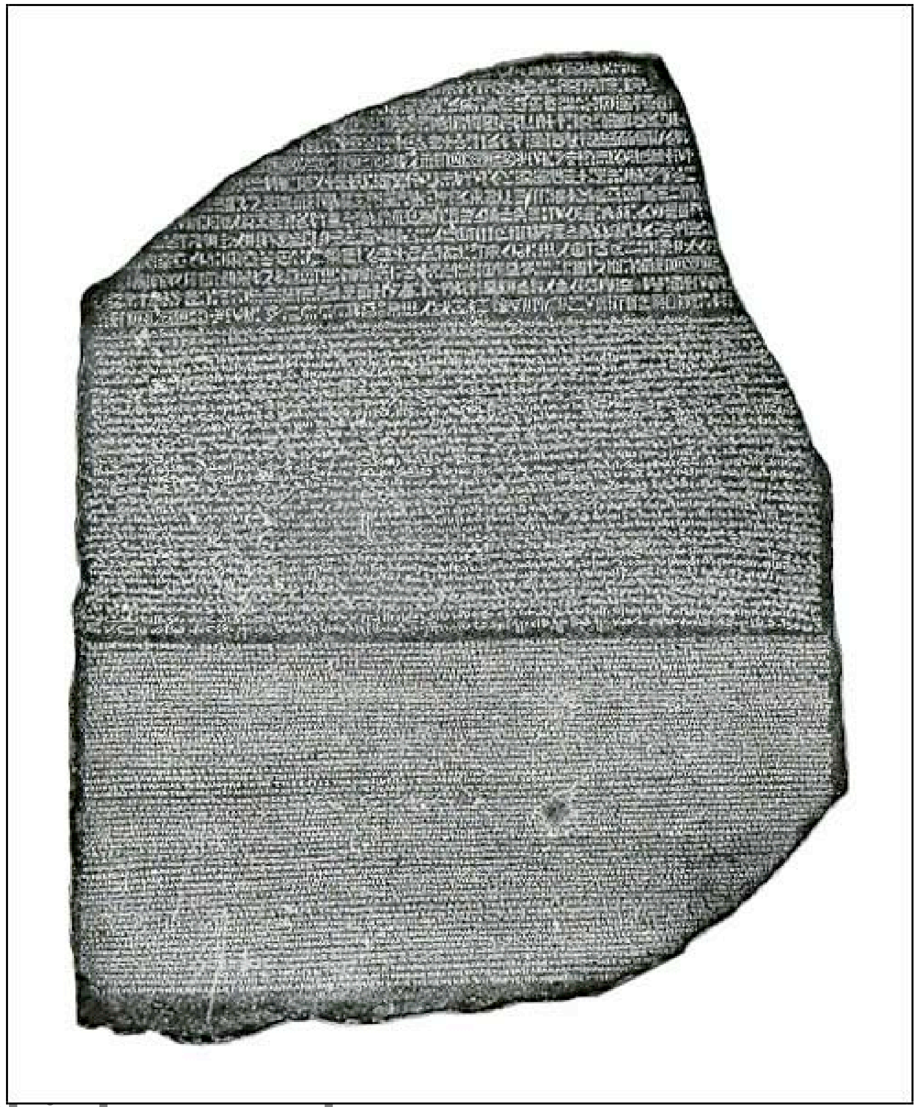
- 1959 – 3D Structure of a Protein
  - Perutz & Kendrew
  - Structure of myoglobin & hemoglobin
  - Nobel in 1962

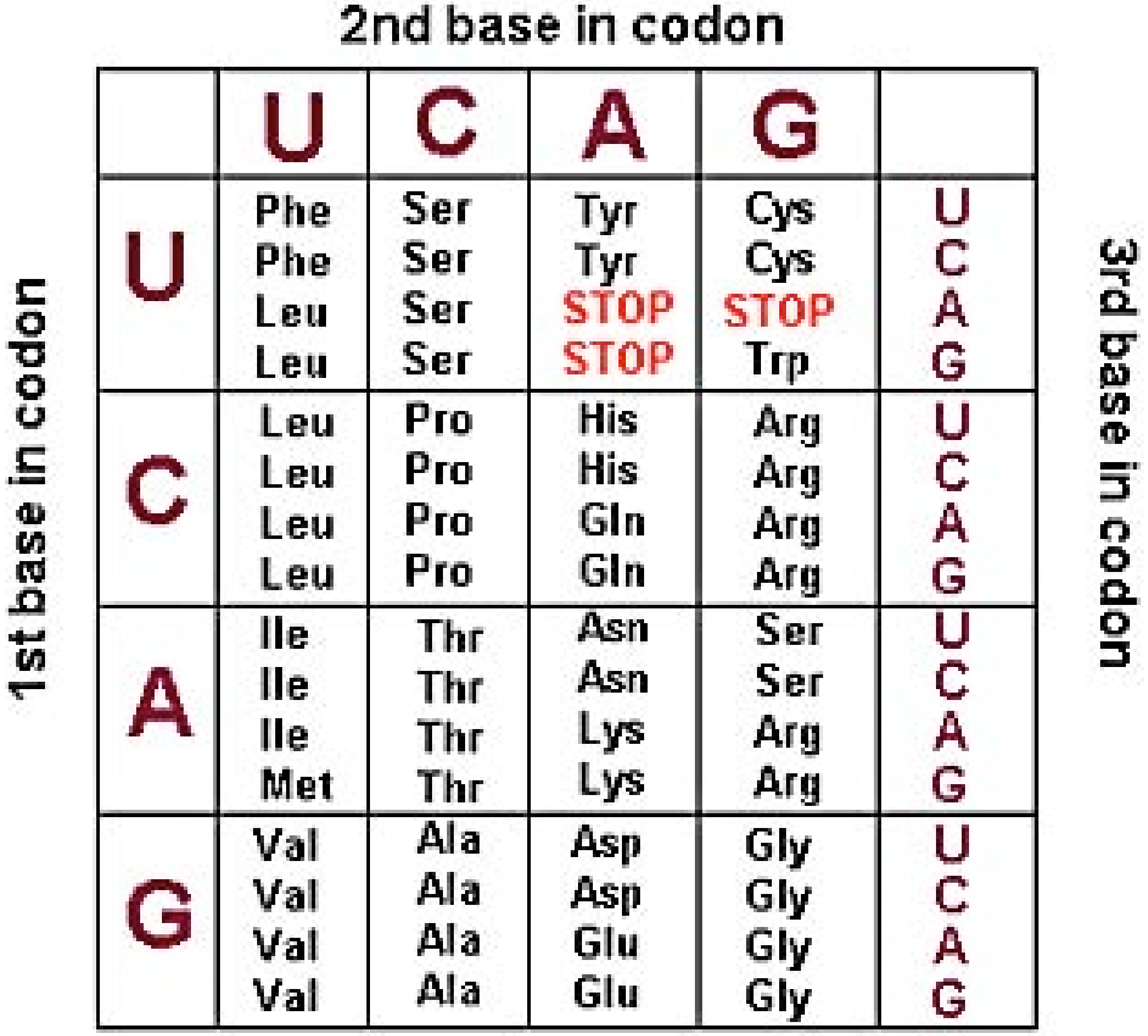




Genetic Code Broken

- 1960's – Genetic Code
  - Holley, Khorana and Nirenberg
  - Rosetta Stone of Life
  - Nobel in 1968





2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Sanger DNA Sequencing

- DNA of all possible lengths from a known starting point
- Each strand ends with a radioactive “didioxy” nucleotide which terminates the chain
- The strands are “weighed” using gel electrophoresis

...AGTCCTG

...AGTCCT

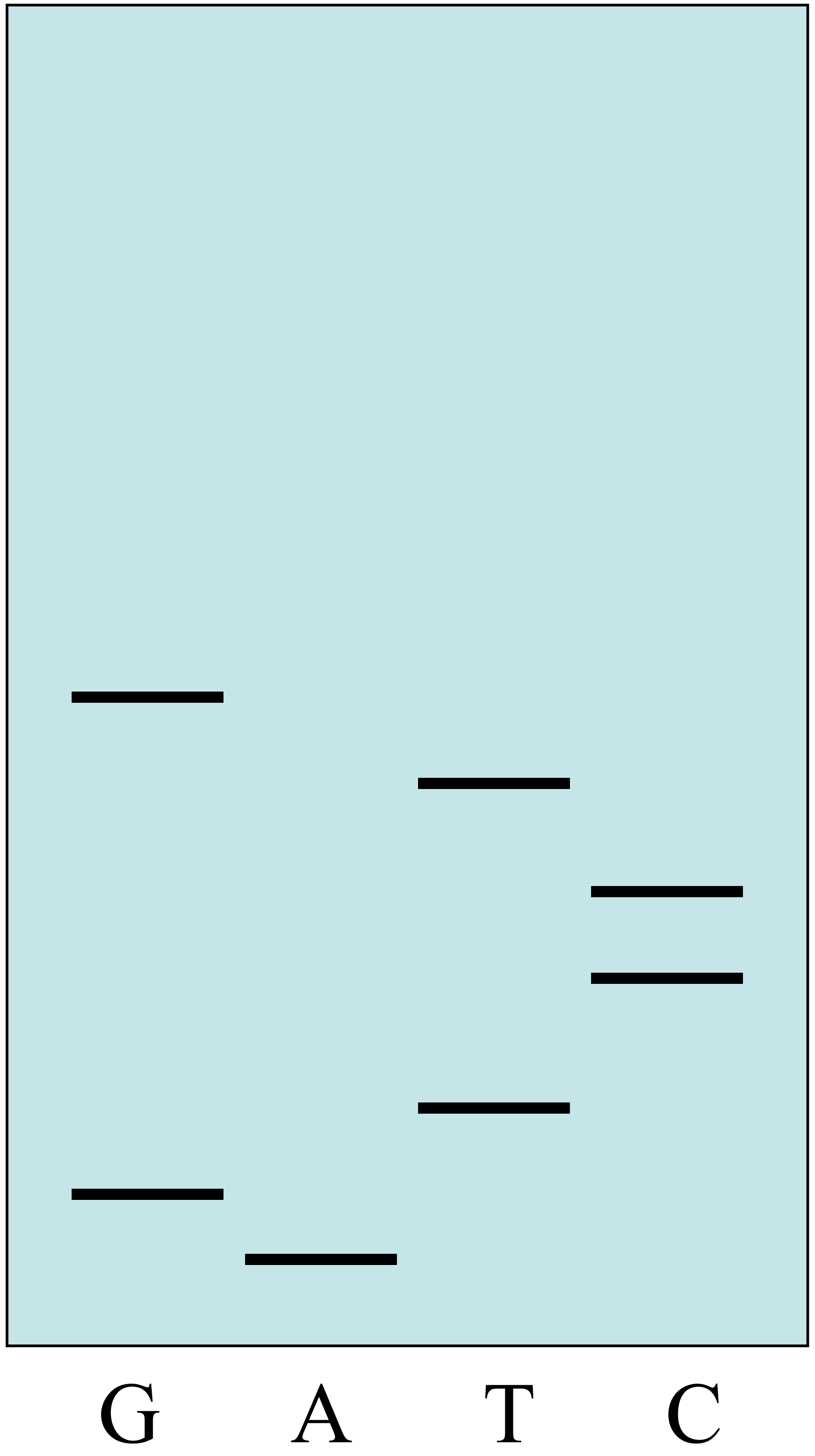
...AGTCC

...AGTC

...AGT

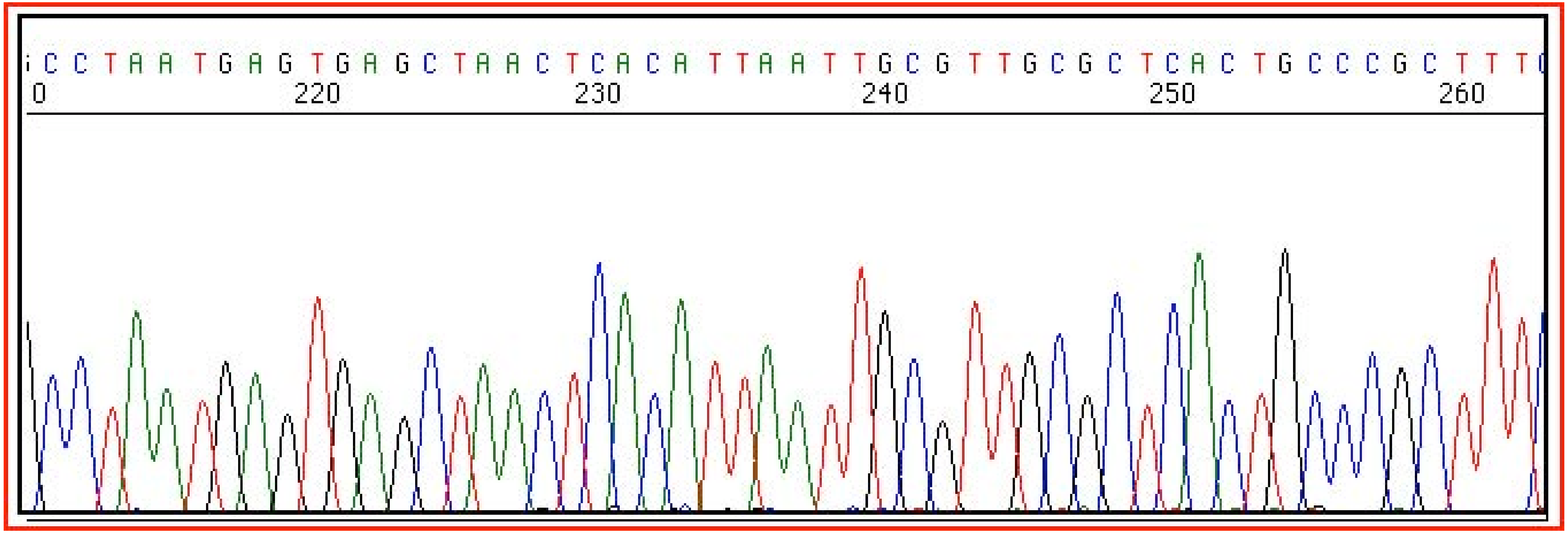
...AG

...A



Modern Sequencing

- Accomplished in a single capillary tube
- Results read via a laser spectrometer
- Accurate to ~700bp
- Completely automated (~$0.04 / bp)



High-Throughput Sequencing

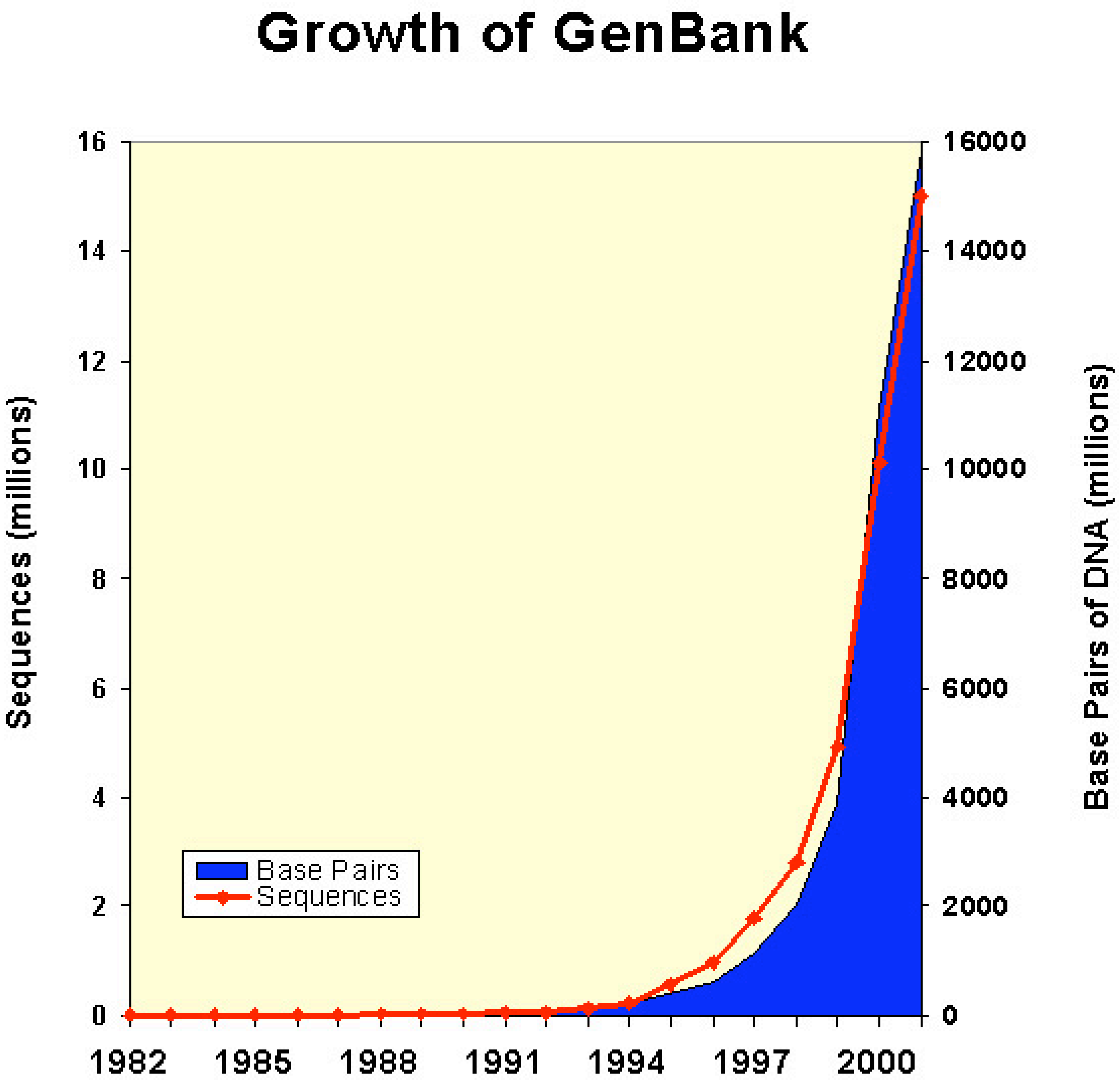
The Sanger Centre

Phil Butcher
Head of I.T.

The Sequencing Facility

http://www.sanger.ac.uk/Info/IT/

Growing Public Sequence DBs



*omics by Cartoon

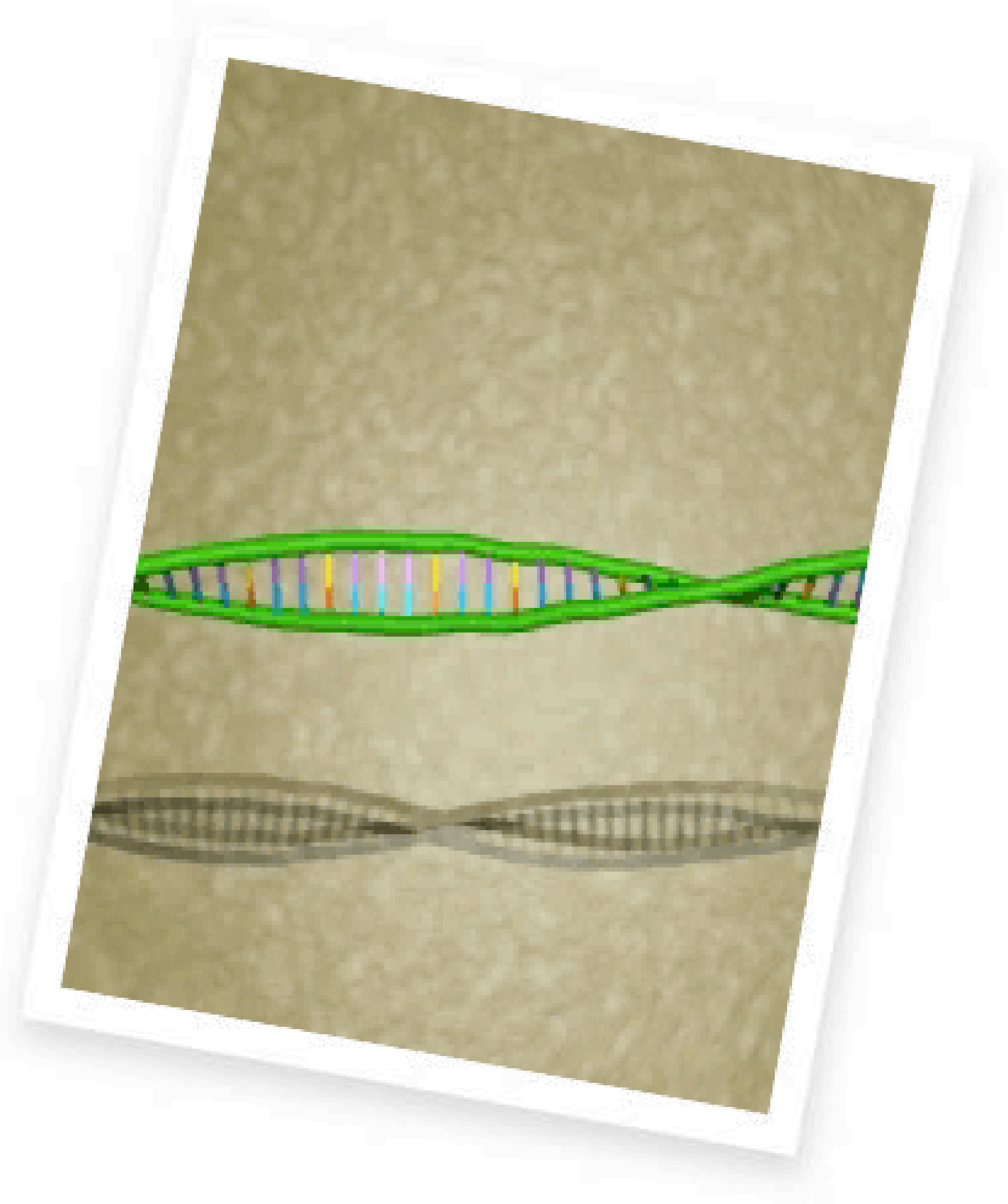
*omics Cartoon

Informatics Problems ...

- Genetic Mapping

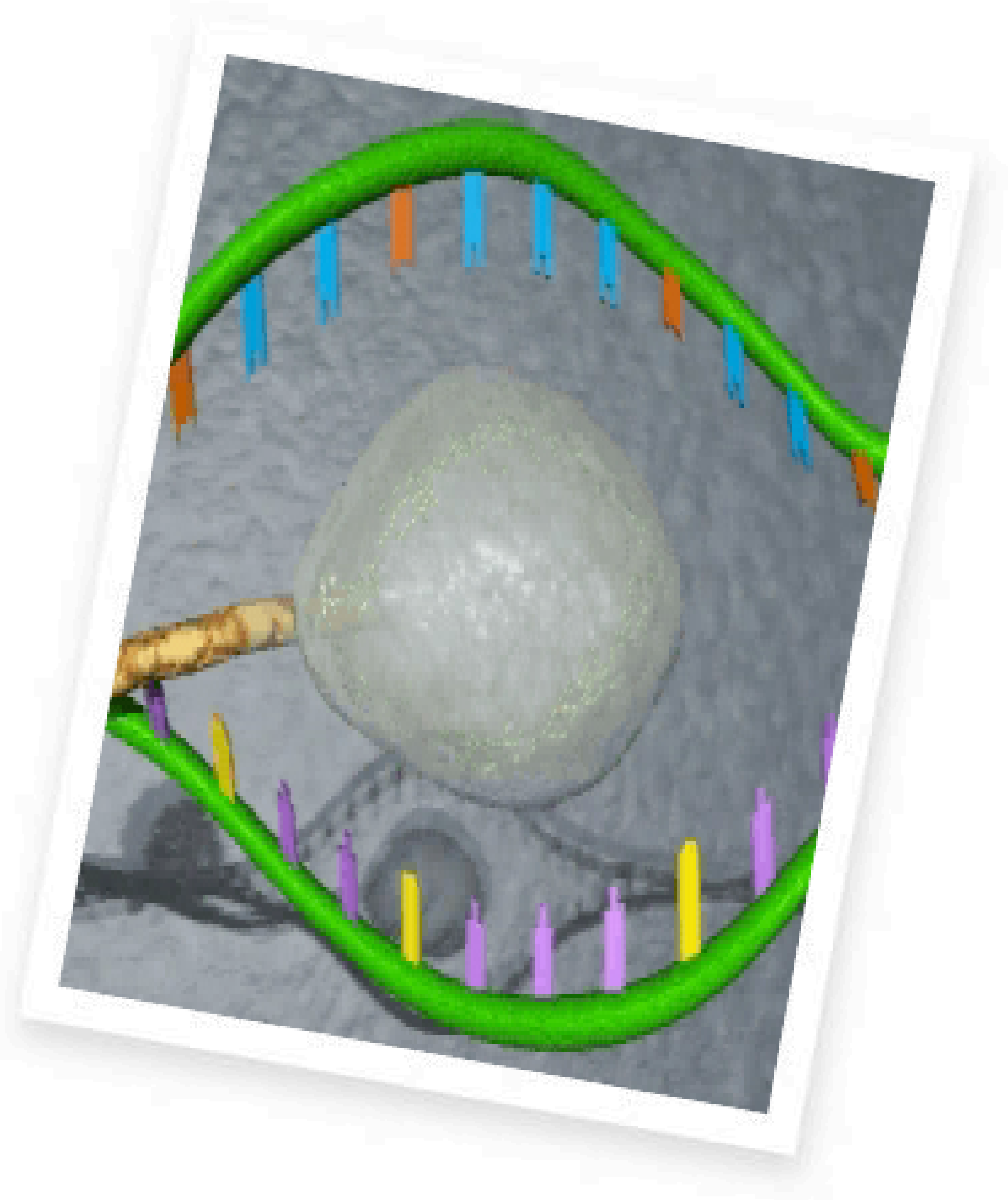
Informatics Problems ...

- Genetic Mapping
- Sequence Analysis
- Genome Annotation
- Functional Genomics
- Comparative Genomics



Informatics Problems ...

- Genetic Mapping
- Sequence Analysis
- Genome Annotation
- Functional Genomics
- Comparative Genomics
- Expression Analysis

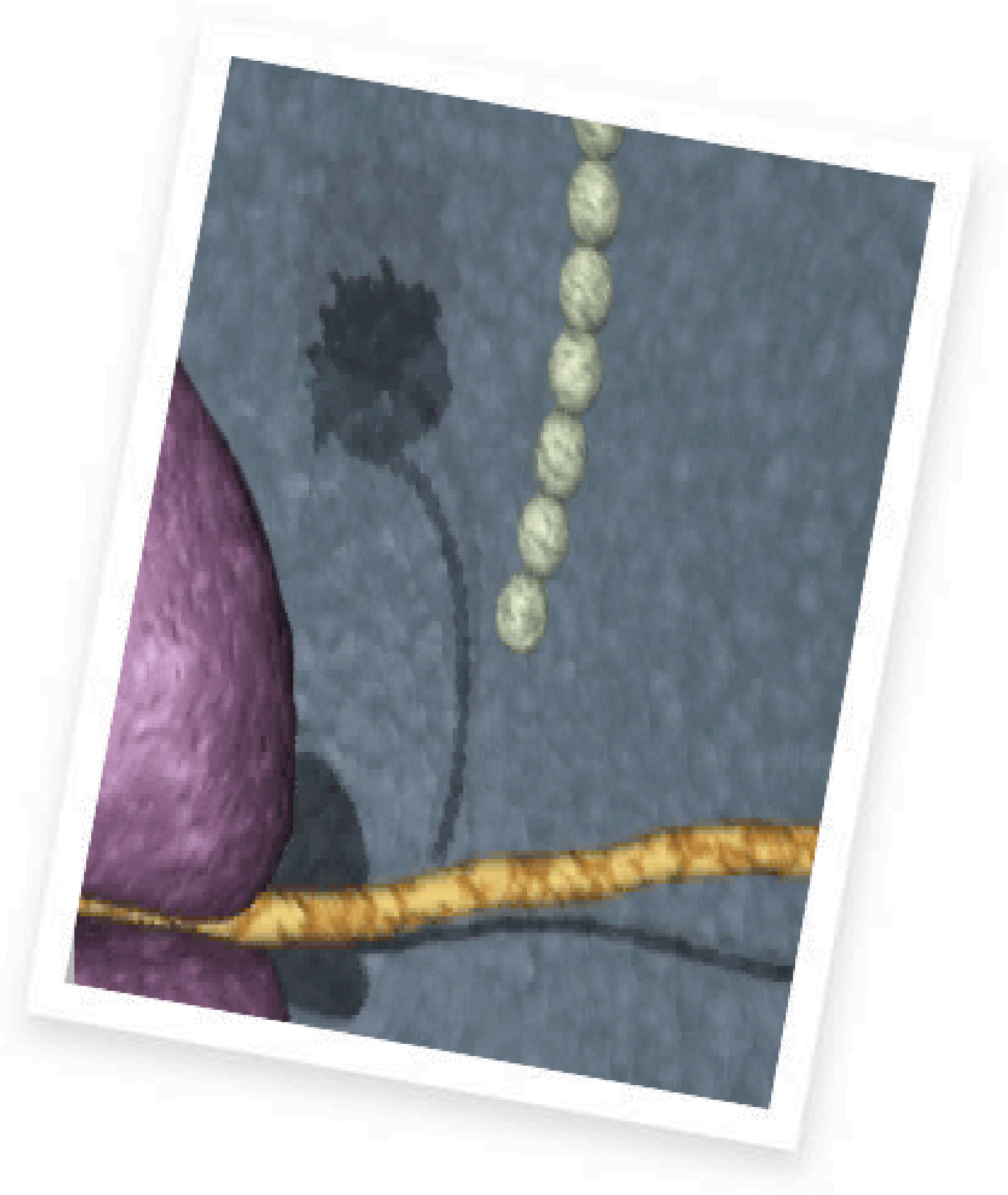


Informatics Problems ...

- Proteomics

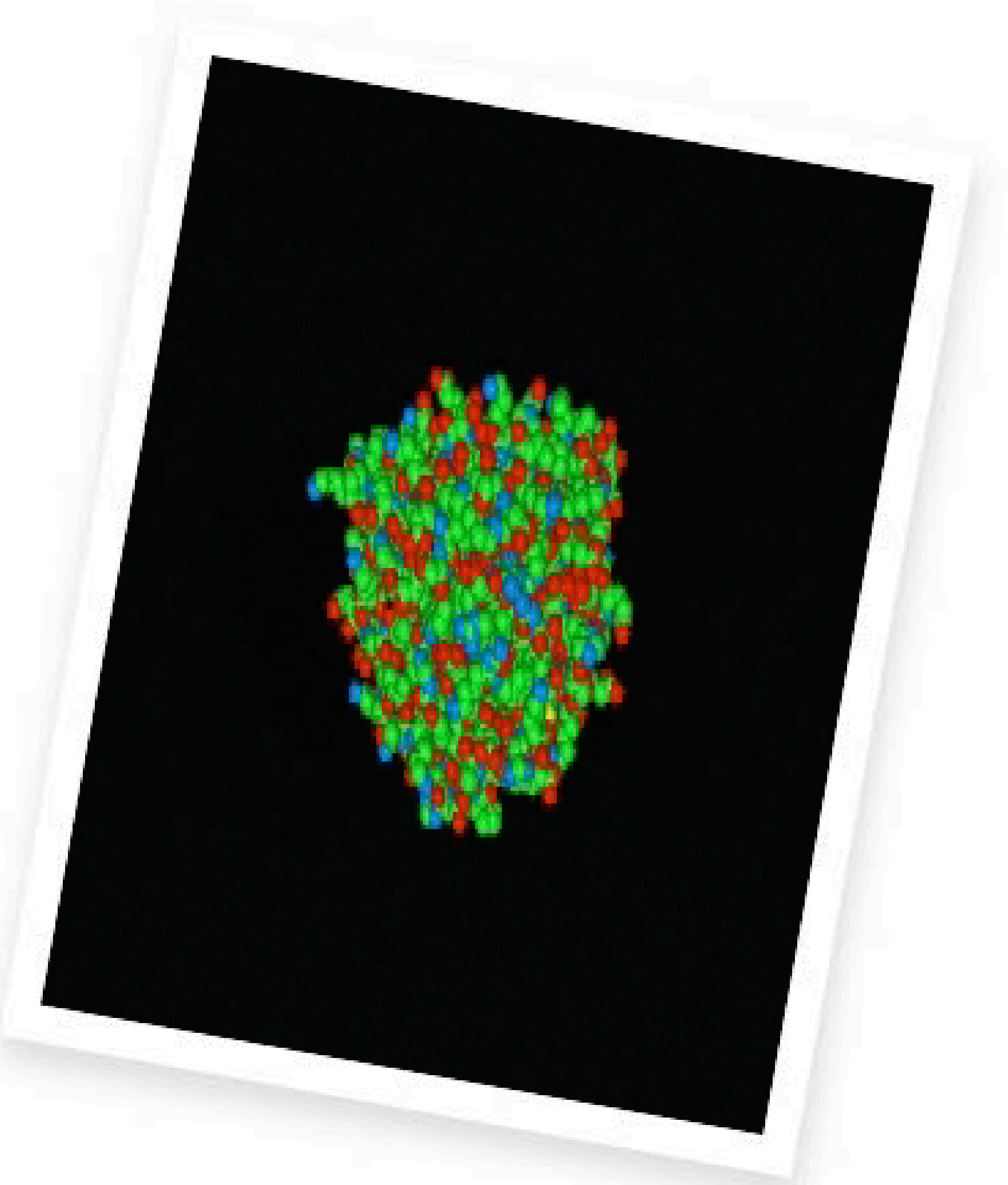
Informatics Problems ...

- Proteomics
- Protein Folding



Informatics Problems ...

- Proteomics
- Protein Folding
- Crystallography
- Chemical Modelling
- Docking
- Systems Biology



Informatics

Informatics

“Computer Aided Scientific Data Analysis”

Working Together

One would expect wet-lab scientists to have a healthy skepticism of any results, knowing how often experiments fail, and how much bad data has made it out into the literature, but many seem to have an almost mystical faith in anything produced by computation.

On the other hand, computational people seem to have an almost mystical faith in wet-lab verification---expecting experiments to be neat, quick deterministic tests like "if" statements in code.

- Gordon D. Pusch

The Data

DNA Code

```
A --> adenosine
C --> cytidine
G --> guanine
T --> thymidine
```

DNA/RNA Code

A --> adenosine

C --> cytidine

G --> guanine

T --> thymidine

U --> uridine

R --> G A (purine)

Y --> T C (pyrimidine)

K --> G T (keto)

M --> A C (amino)

S --> G C (strong)

W --> A T (weak)

B --> G T C

D --> G A T

H --> A C T

V --> G C A

N --> A G C T (any)

- gap of ? length

Amino Acid Code

A	alanine	P	proline
B	D or N	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	E or Q
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of ? length

Genetic Code

TTT	**F**	TCT	S	TAT	Y	TGT	C
TTC		TCC		TAC		TGC	
TTA	L	TCA		TAA	STOP	TGA	STOP
TTG		TCG		TAG	STOP	TGG	W
CTT	L	CCT	P	CAT	H	CGT	R
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Q	CGA	
CTG		CCG		CAG		CGG	
ATT	I	ACT	T	AAT	N	AGT	S
ATC		ACC		AAC		AGC	
ATA		ACA		AAA	K	AGA	R
ATG	M	ACG		AAG		AGG	
GTT	V	GCT	A	GAT	D	GGT	G
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	E	GGA	
GTG		GCG		GAG		GGG	

Sequence Searching Algorithms

Compared

BLAT

- Simple, “grep-like”, character matching algorithm
- Identifies matches > 90% identity

- chop database sequences into overlapping N-mers
- create an indexed HASH table of each database sequence

- chop query sequence into overlapping N-mers
- create a HASH table of query sequence words

- positive hit when >90% of query HASH entries are in DB HASH entry

BLAST

- Simple character matching algorithm
- Everything that BLAT does
- Plus...
-
- Identifies similarities (with gaps) of relative statistical significance
-
- After identifying significant query/database HASH overlap
-
- Extend matches, allowing gaps

HMMer ...

- Not like BLAT or BLAST

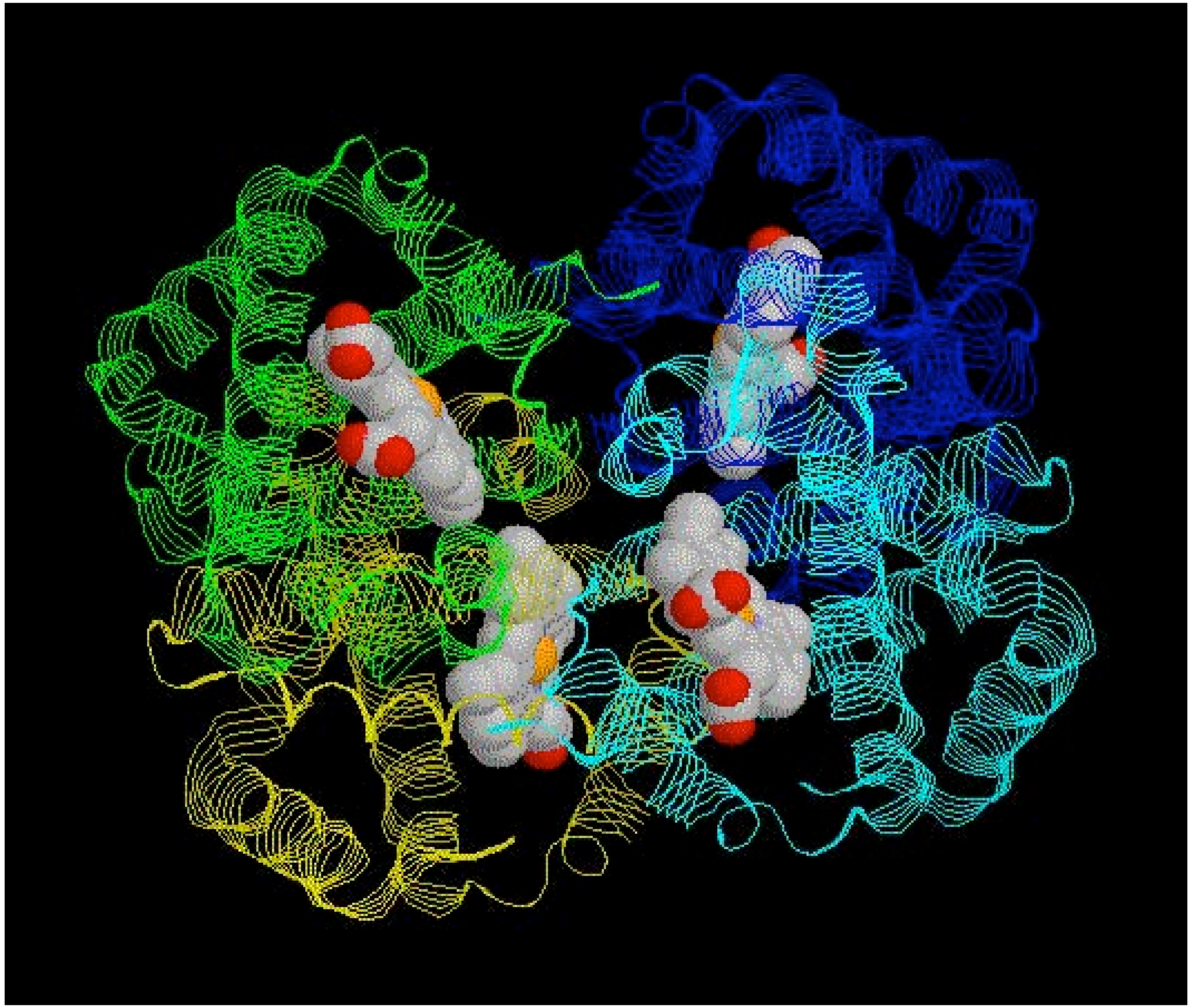
- More sophisticated, pattern matching algorithm
- “Voice recognition-like”

- Build statistical model of a multiple sequence alignment
- Search sequence databases with models
- Search model databases with sequences

Hmmer ...

- Build an HMM model from 50 globins
  - `% hmmbuild globin.hmm globins50.msf`
- Calibrate the model
  - `% hmmcalibrate globin.hmm`
- Search shrimp sequence DB with model
  - `% hmmsearch globin.hmm Artemia.fa`
- Search model database with shrimp sequence
  - `% hmmpfam globin.hmm Artemia.fa`

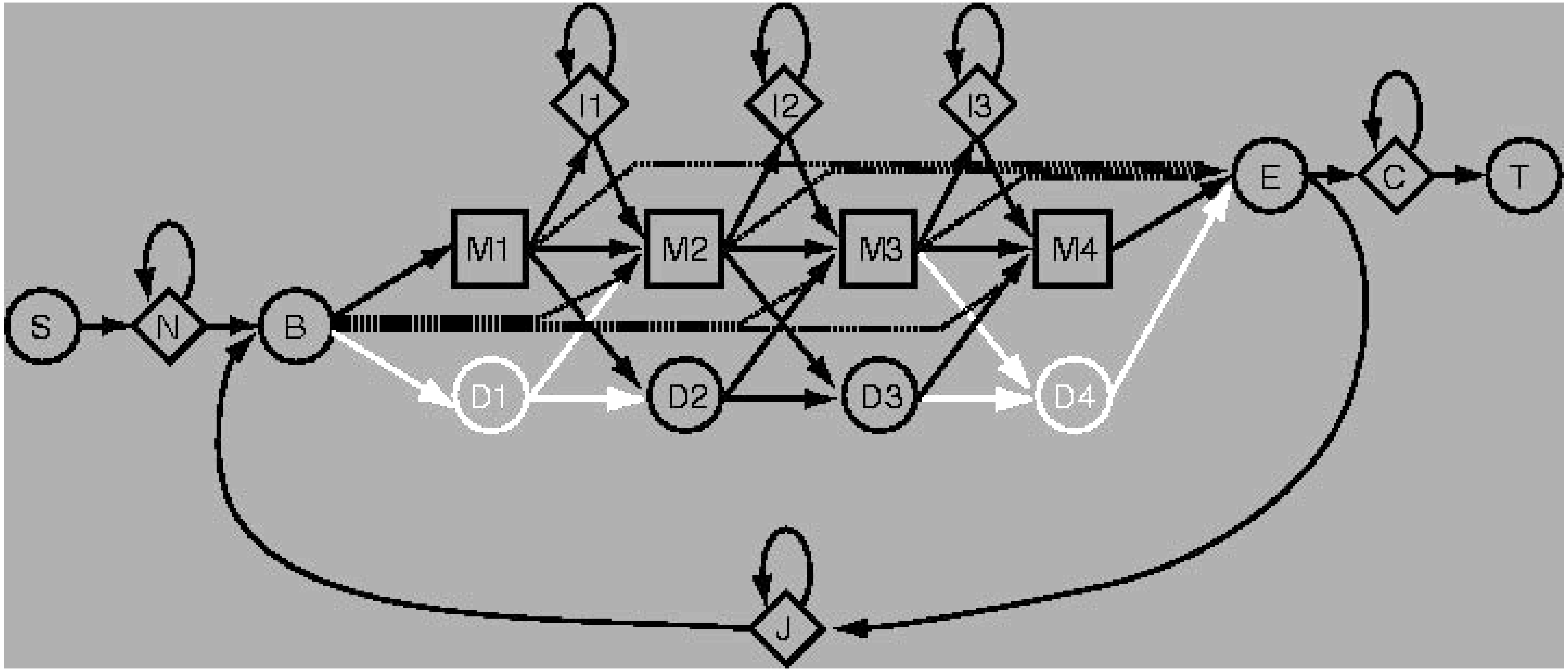
globin



MSF Format

```
DNA_MULTIPLE_ALIGNMENT 1.0
Three anthropoidea
MSF: 50  Type: N  Check: 2666 ..


Name: Homo_sapiens     Len: 50   Check: 8318   Weight: 1.00
Name: Pan_paniscus     Len: 50   Check: 7854   Weight: 1.00
Name: Gorilla_gorilla  Len: 50   Check: 7778   Weight: 1.00


//


Homo_sapiens        AGUCGAGUC...GCAGAAAC
Pan_paniscus        AGUCGCGUCG..GCAGAAAC
Gorilla_gorilla     AGUCGCGUCG..GCAGAUAC


Homo_sapiens        GCAUGAC.GACCACAUUUU.
Pan_paniscus        GCAUGACGGACCACAUCAU.
Gorilla_gorilla     GCAUCACGGAC.ACAUCAUC


Homo_sapiens        CCUUGCAAAG
Pan_paniscus        CCUUGCAAAG
Gorilla_gorilla     CCUCGCAGAG
```

hmm State Diagram



FASTA Format

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
>gi|532320|some other protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
…
```

hmm Format

```
HMMER2.0  [2.2g]
NAME  globins50
LENG  148
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   ../binaries/hmmbuild globin.hmm globins50.msf
COM   ../binaries/hmmcalibrate globin.hmm
NSEQ  50
DATE  Thu Jul 25 10:51:38 2002
CKSUM 9858
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313
   45    531    201    384  -1998   -644
EVD   -41.853970   0.212647
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q
   R      S      T      V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
         -661      *  -1444
     1     77   -228  -1302  -1020   -730  -1034   -756    578   -803   -375     82   -791  -1461   -720
   -959    364    -94   2204  -1315   -857     9
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45
   96    359    117   -369   -294   -249
     -    -39  -5807  -6849   -894  -1115   -701  -1378   -661      *
```

Sequence Analysis Algorithm Summary

- BLAT rapidly identifies nearly identical sequences (chars)
- BLAST less rapidly identifies similar sequences (chars)
- HMMer identifies "likeness" of protein families (pattern matching)

Clustering for Informatics

What’s Different?

- Many User Models
- Many Applications
- Mostly Open Source
- Cooperative and Collaborative
- Compute and Data Intensive
- Rapid Rate of Growth

Types Of Problems

Tightly Coupled	Embarrassingly Parallel
$1 + 1 = X$	$1 + 1 = X$
$X + 2 = Y$	$2 + 2 = Y$
$Y + 3 = Z$	$3 + 3 = Z$
$Z + N = W$	$N + N = W$

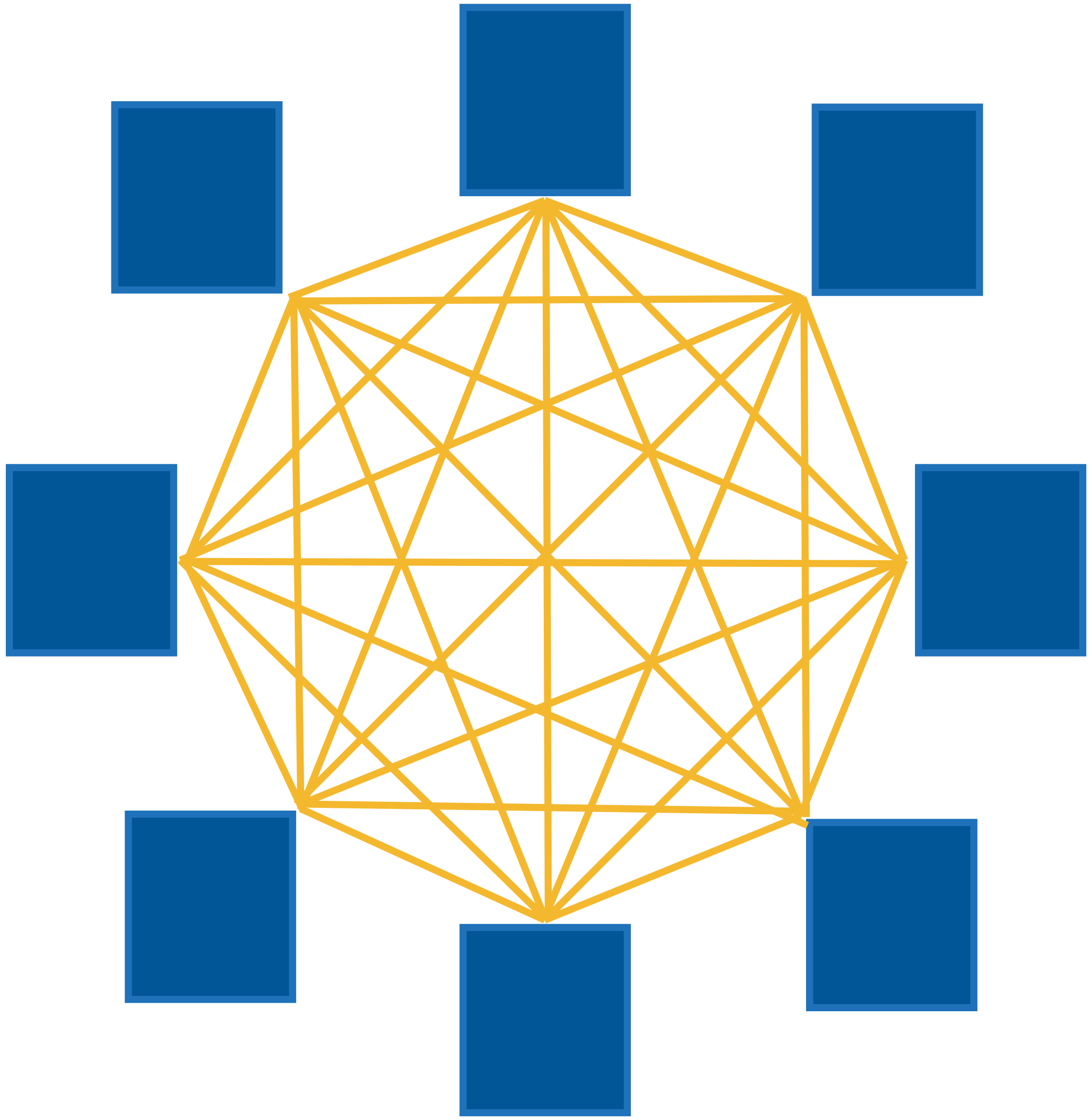
Solution Strategies

High Performance

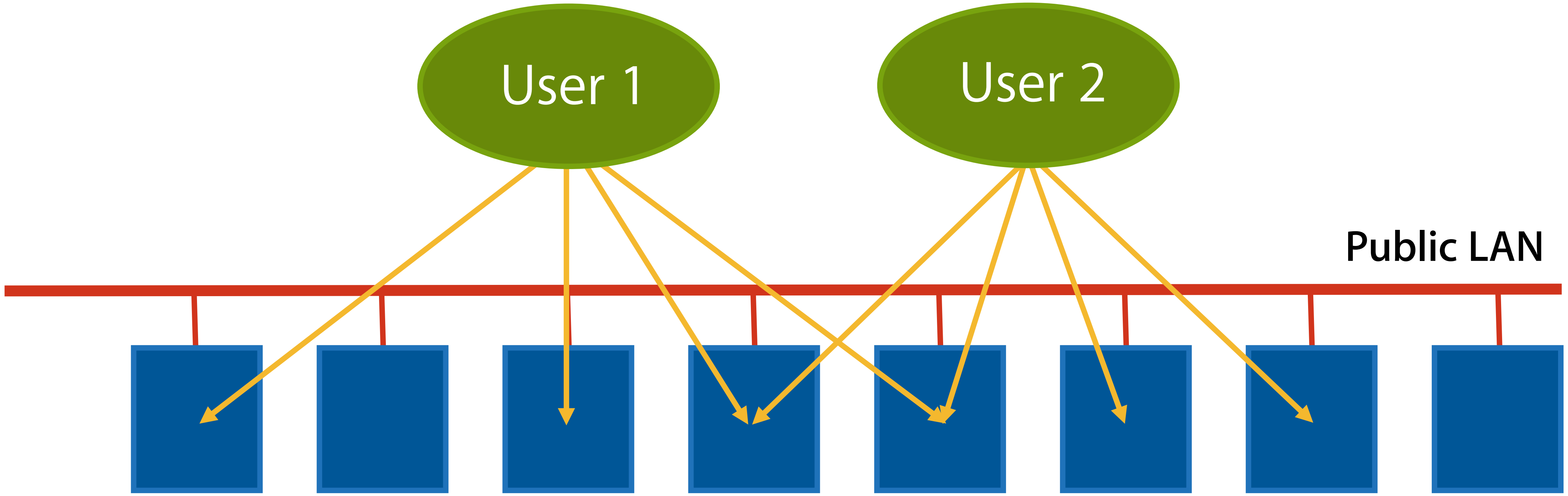
High Throughput

Types of Clusters

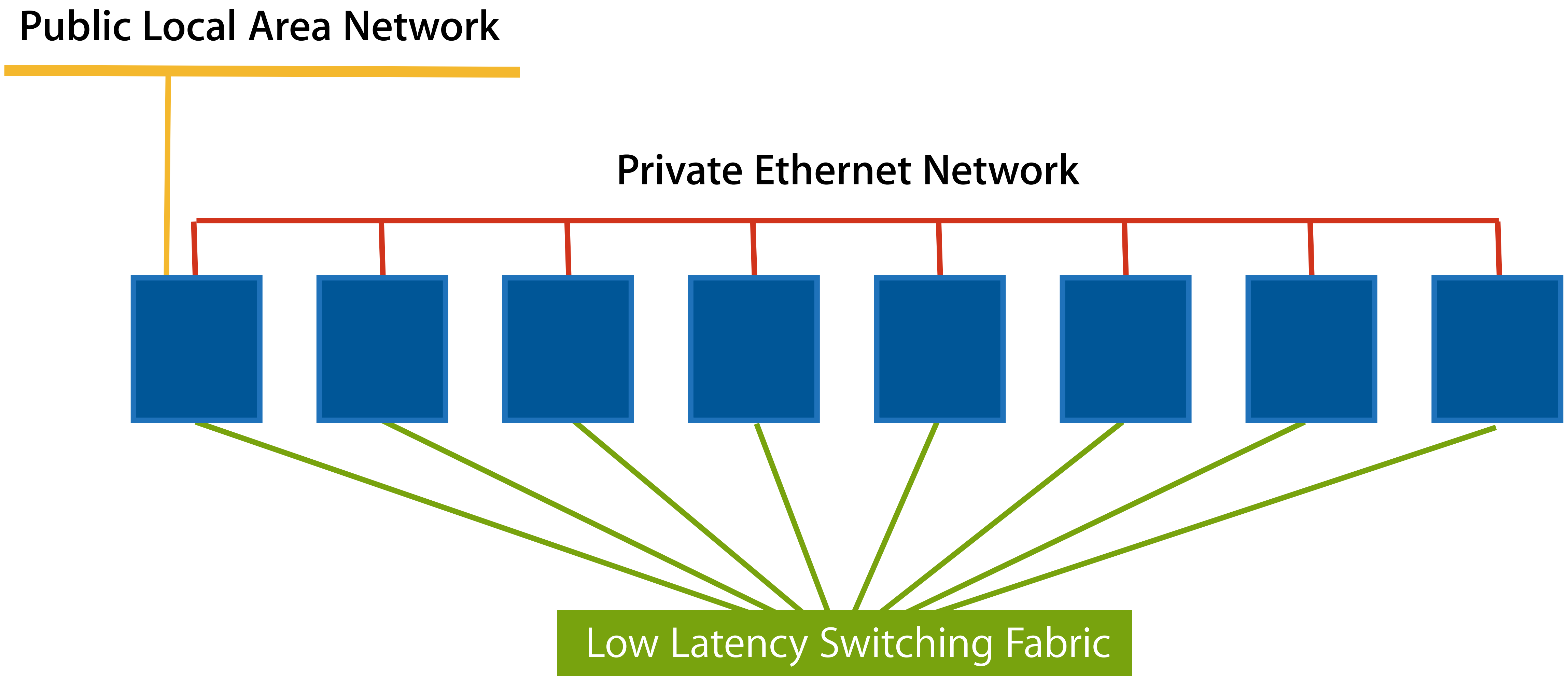
Beowulf



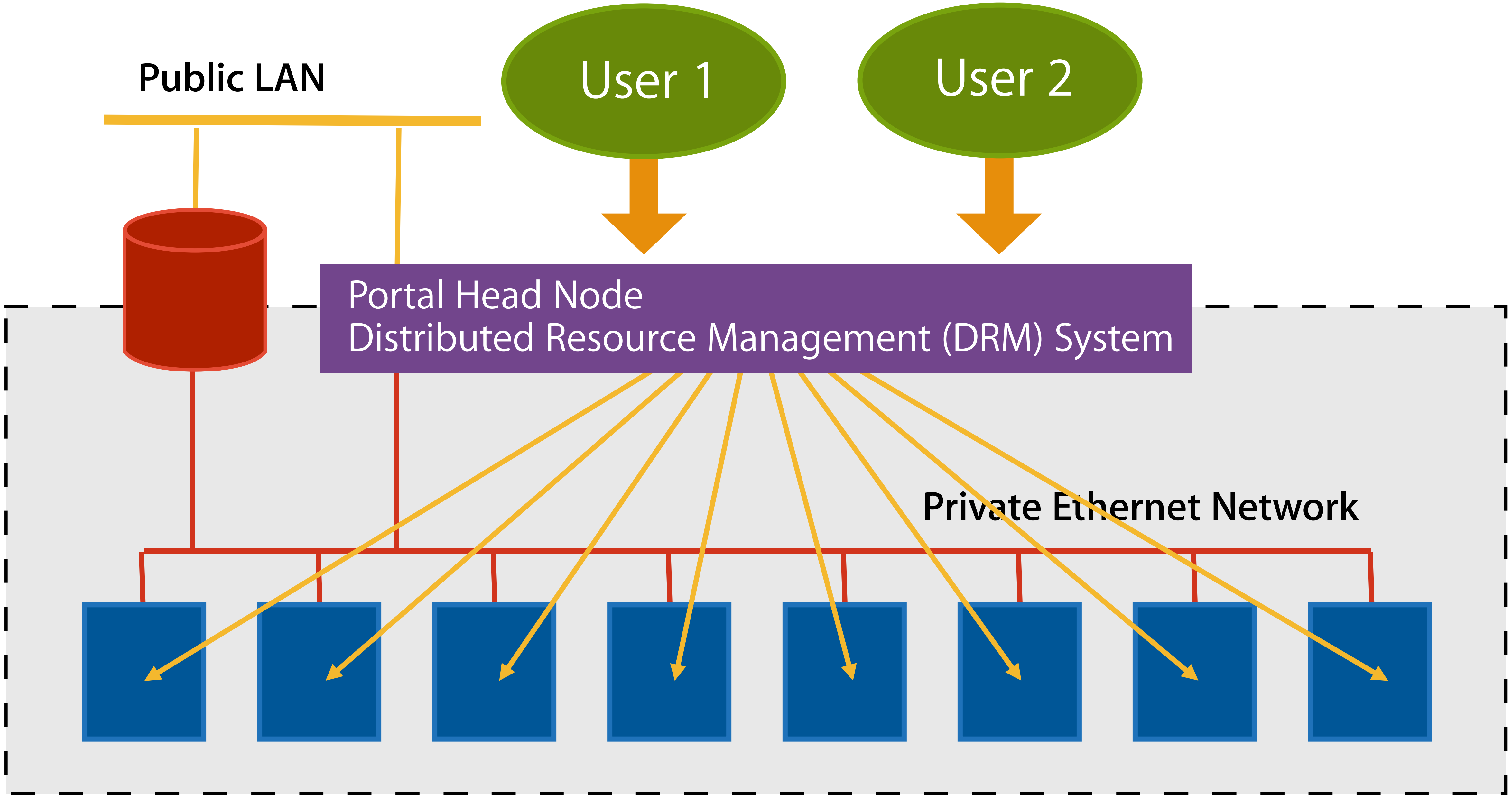
LAN Computing



Tightly Coupled



Portal Architecture



Infx Cluster Requirements

Infx Cluster Requirements

User Model

Applications

Compilers

Physical Environment

Know Bottlenecks

Network & Topology

Storage

Maintenance

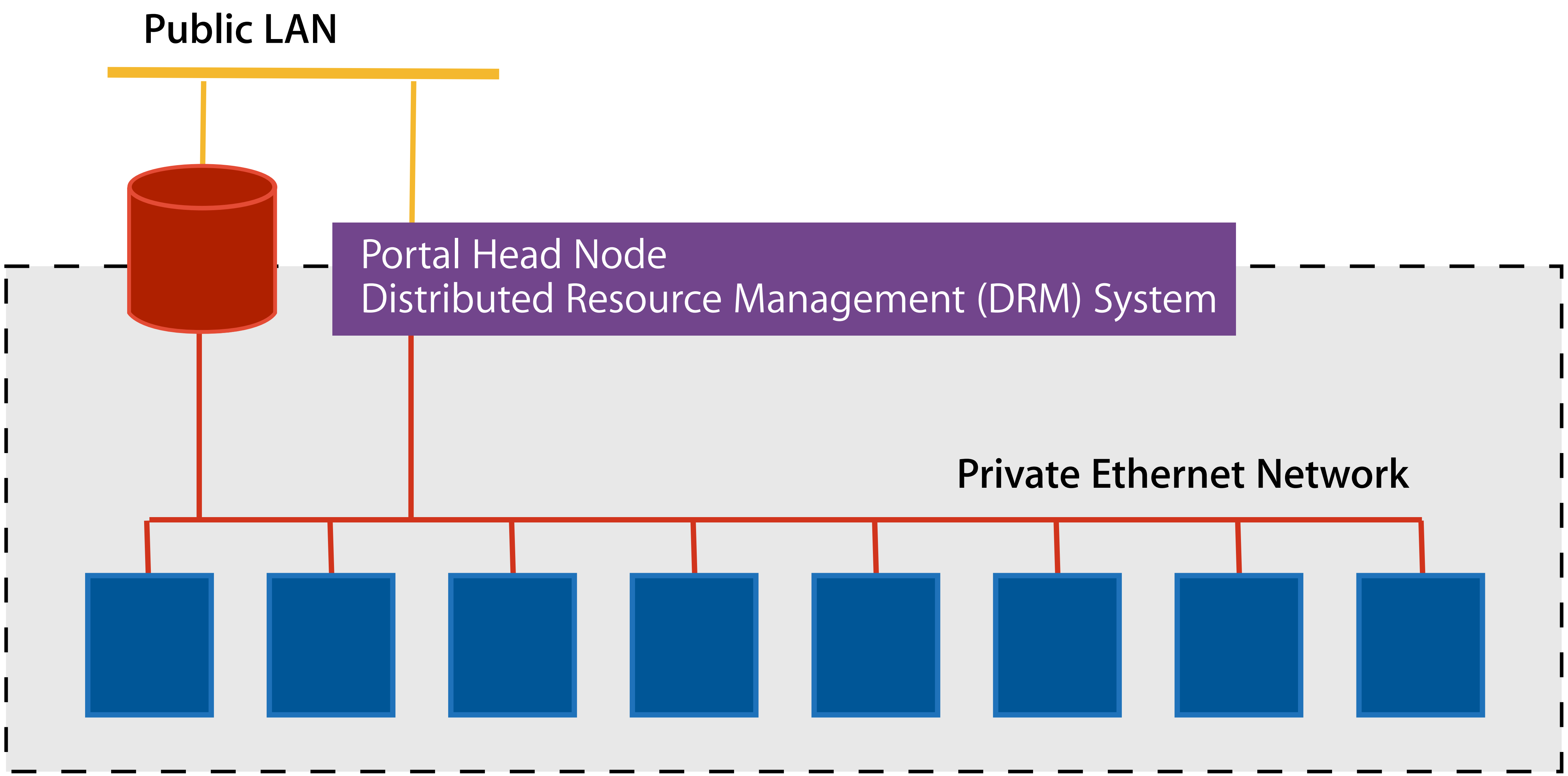
Administration & Monitoring

DRM

Common File System

Common User Environment

Portal Architecture



Infx Cluster Requirements

Network Services
-DHCP, DNS

Common User Environment
-NetInfo, NIS, LDAP, /etc/*

Common File System
-AFP, NFS, SMB, AFS

Network Translation
-NATd

Mail Services
-POP, IMAP, SMTP

Web Services
-HTTP, HTTPs

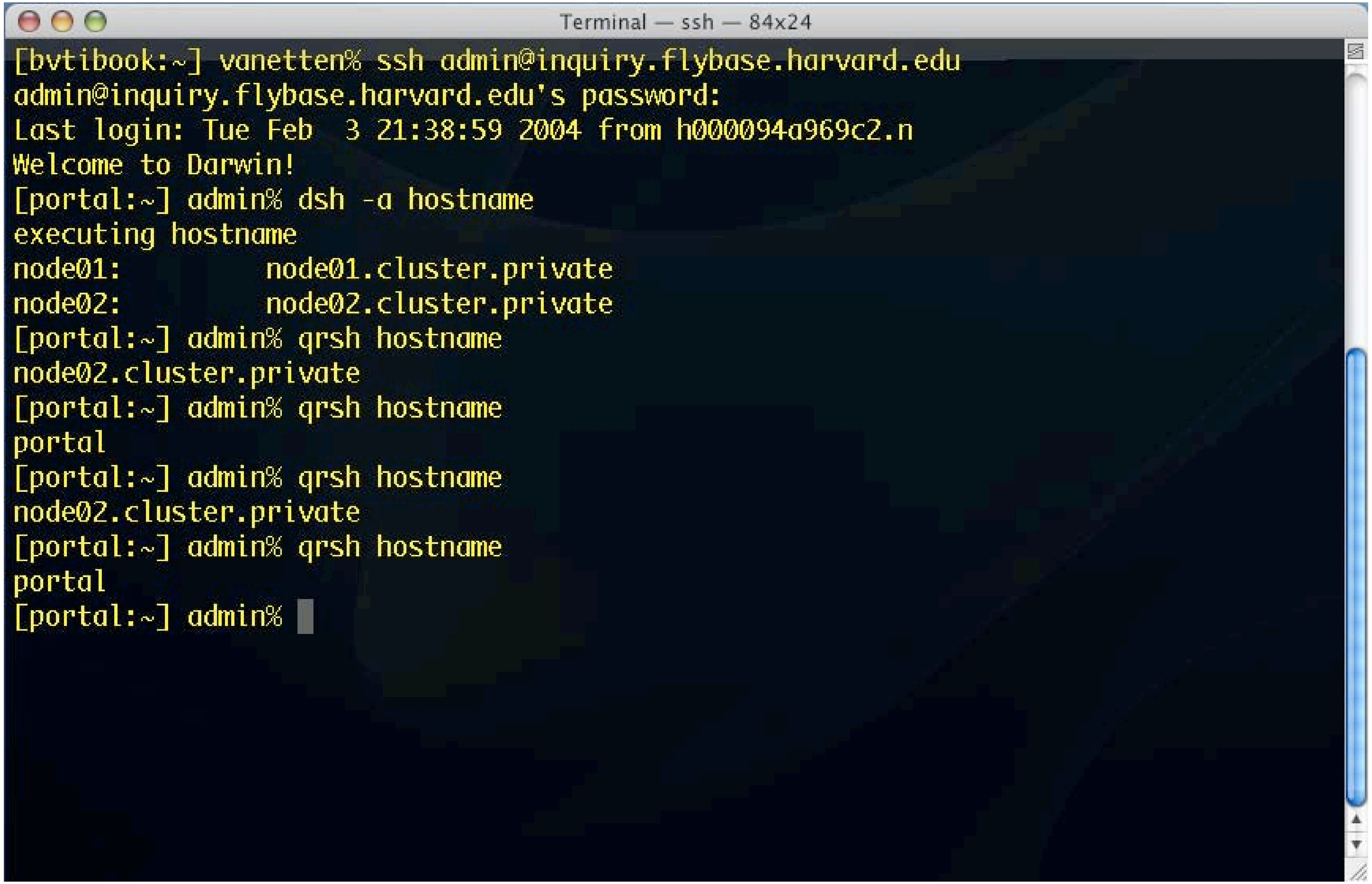
Time Synchronization
-NTP

DRM
-LSF, SGE, PBS

Image Server
-NetInstall, NetRestore, SI

Cluster Monitor
-Server Admin/Monitor, BB

Using a Cluster



```
[bvtibook:~] vanetten% ssh admin@inquiry.flybase.harvard.edu
admin@inquiry.flybase.harvard.edu's password:
Last login: Tue Feb  3 21:38:59 2004 from h000094a969c2.n
Welcome to Darwin!
[portal:~] admin% dsh -a hostname
executing hostname
node01:         node01.cluster.private
node02:         node02.cluster.private
[portal:~] admin% qrsh hostname
node02.cluster.private
[portal:~] admin% qrsh hostname
portal
[portal:~] admin% qrsh hostname
node02.cluster.private
[portal:~] admin% qrsh hostname
portal
[portal:~] admin% 
```

Clustering: PROS

- Affordable
- Scalable
- Flexible
- Reliable
- Consumable
- Sharable

Clustering: CONS

Challenging to ...

- Build
- Manage
- Use
- Map computing to computers
- Achieve high-throughput
- Achieve high-performance

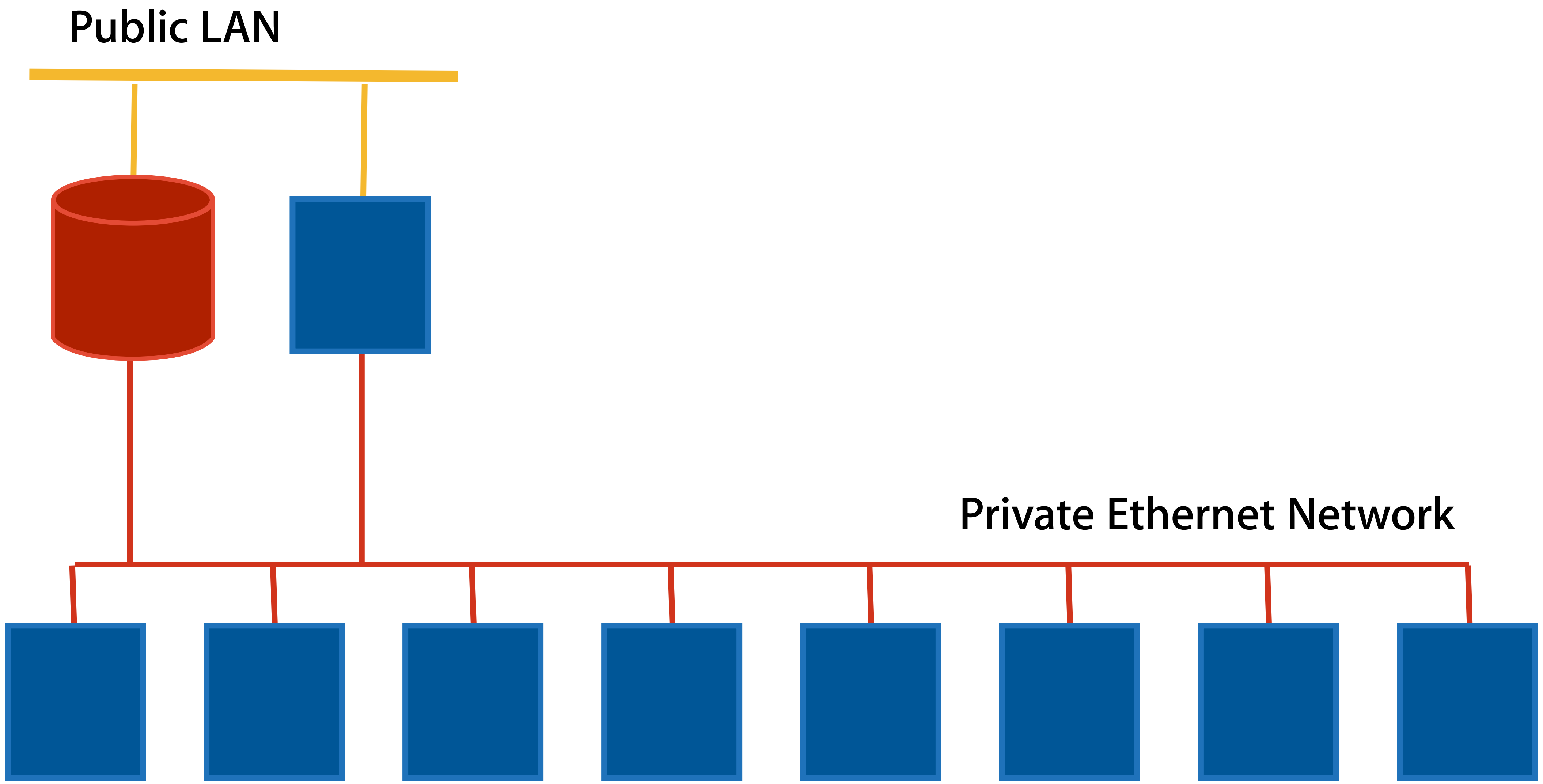
BioTeam iNquiry

BioTeam iNquiry

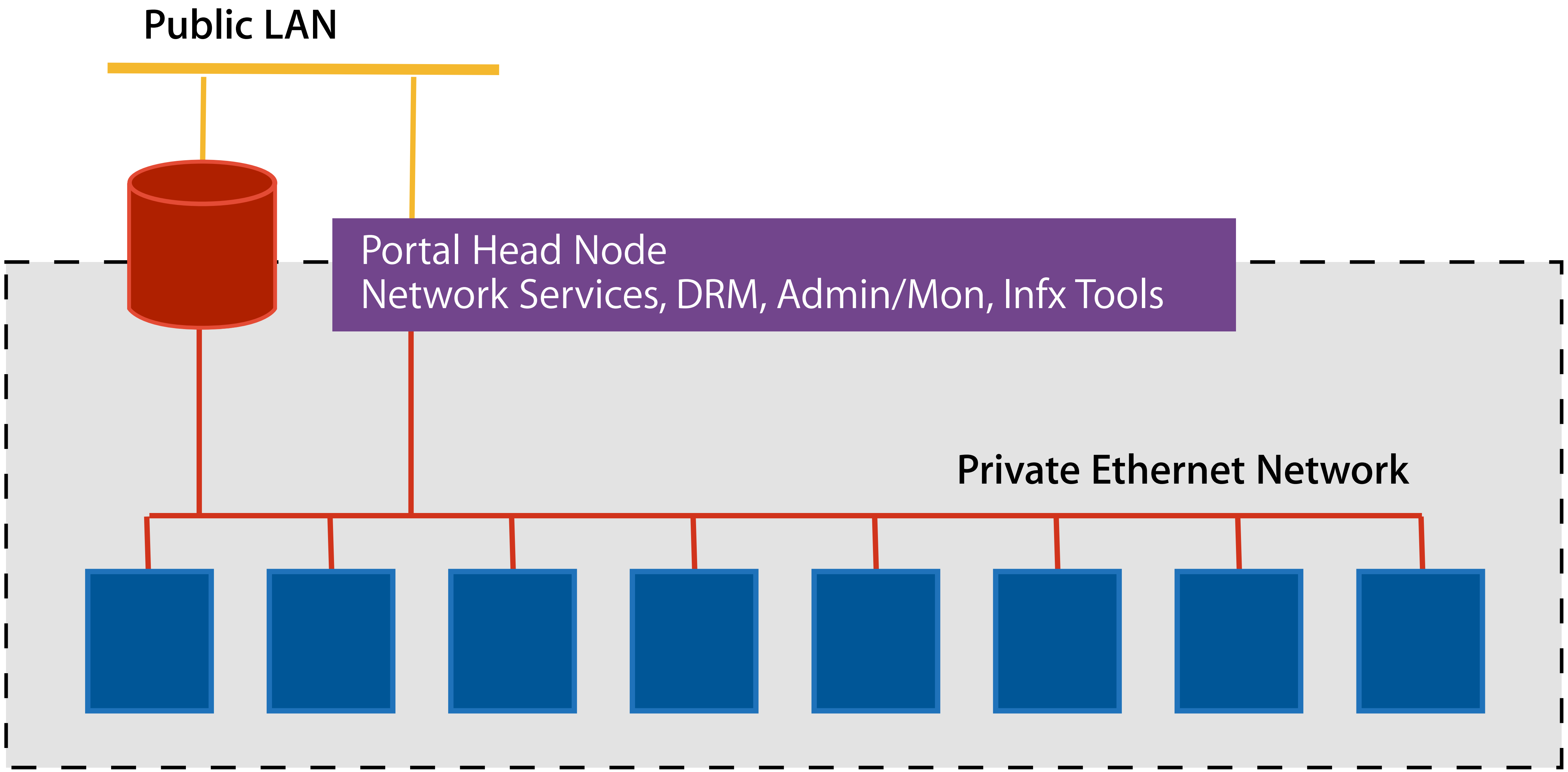
What's inside?

- Network Services
- DRM (Sun GridEngine, LSF compatible)
- Admin & Monitoring Tools
- 200+ Informatics Applications
  - Cluster enabled
  - UI for the scientist
  - Consistent & Extensible

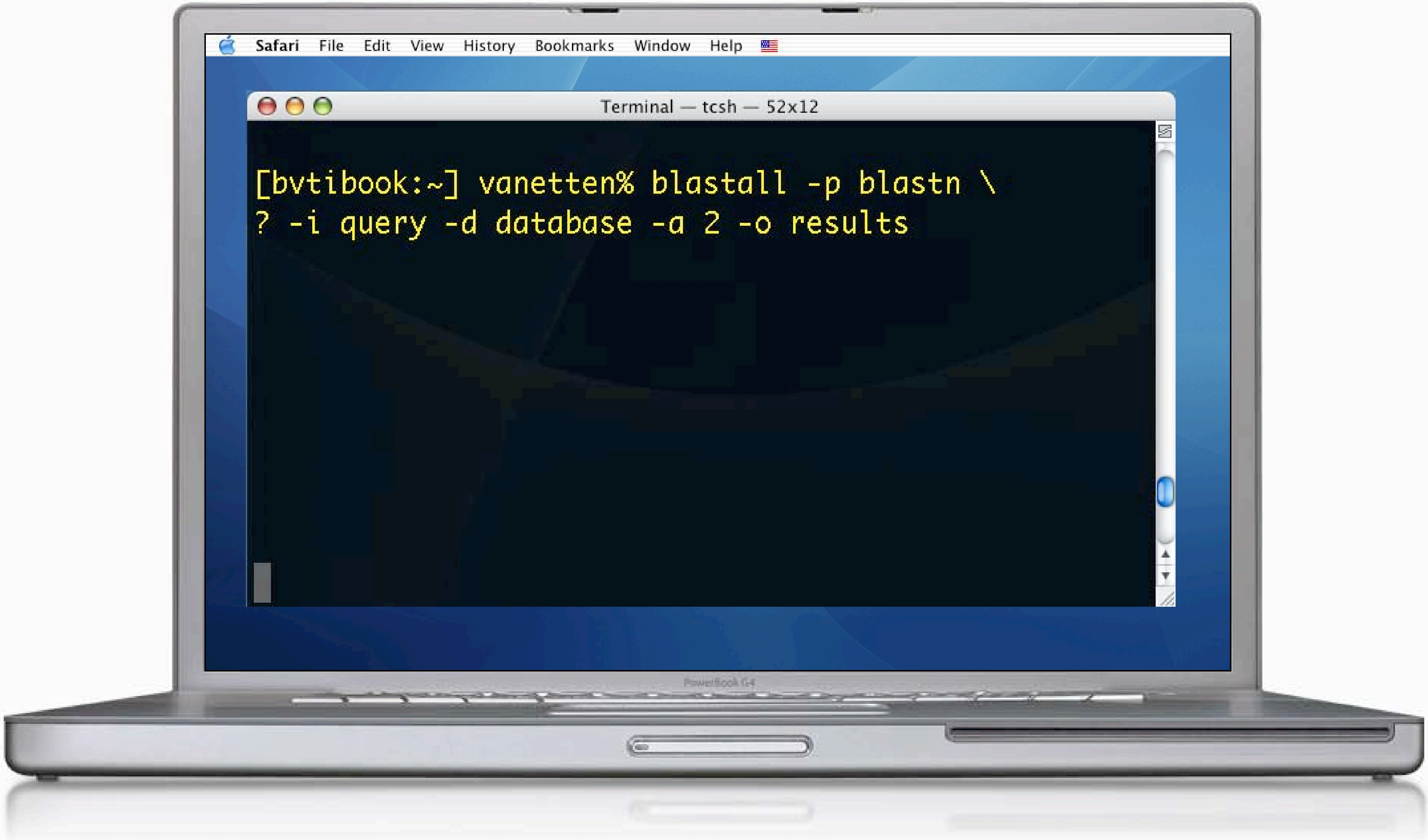
From This ...



To This

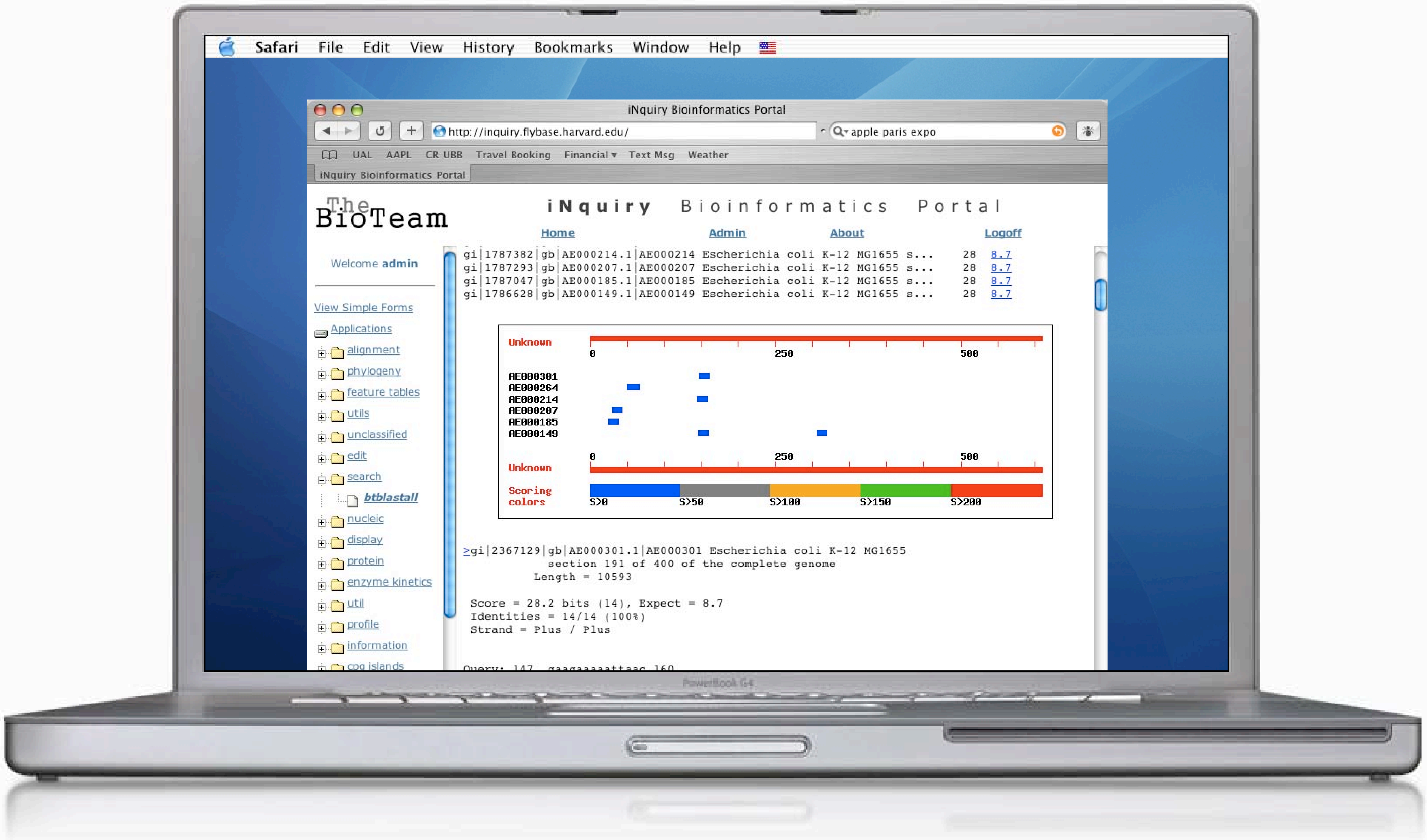

Public LAN
Portal Head Node
Network Services, DRM, Admin/Mon, Infx Tools
Private Ethernet Network

From This ...



```
[bvtibook:~] vanetten% blastall -p blastn \
? -i query -d database -a 2 -o results
```

To This



```
gi|1787382|gb|AE000214.1|AE000214 Escherichia coli K-12 MG1655 s...    28  8.7
gi|1787293|gb|AE000207.1|AE000207 Escherichia coli K-12 MG1655 s...    28  8.7
gi|1787047|gb|AE000185.1|AE000185 Escherichia coli K-12 MG1655 s...    28  8.7
gi|1786628|gb|AE000149.1|AE000149 Escherichia coli K-12 MG1655 s...    28  8.7

>gi|2367129|gb|AE000301.1|AE000301 Escherichia coli K-12 MG1655
           section 191 of 400 of the complete genome
         Length = 10593

 Score = 28.2 bits (14), Expect = 8.7
 Identities = 14/14 (100%)
 Strand = Plus / Plus


Query: 147  gaagaaaaattaac 160
```

iNquiry Demo

Screenshots in place of “Live” Demo

Web-based Tool Access

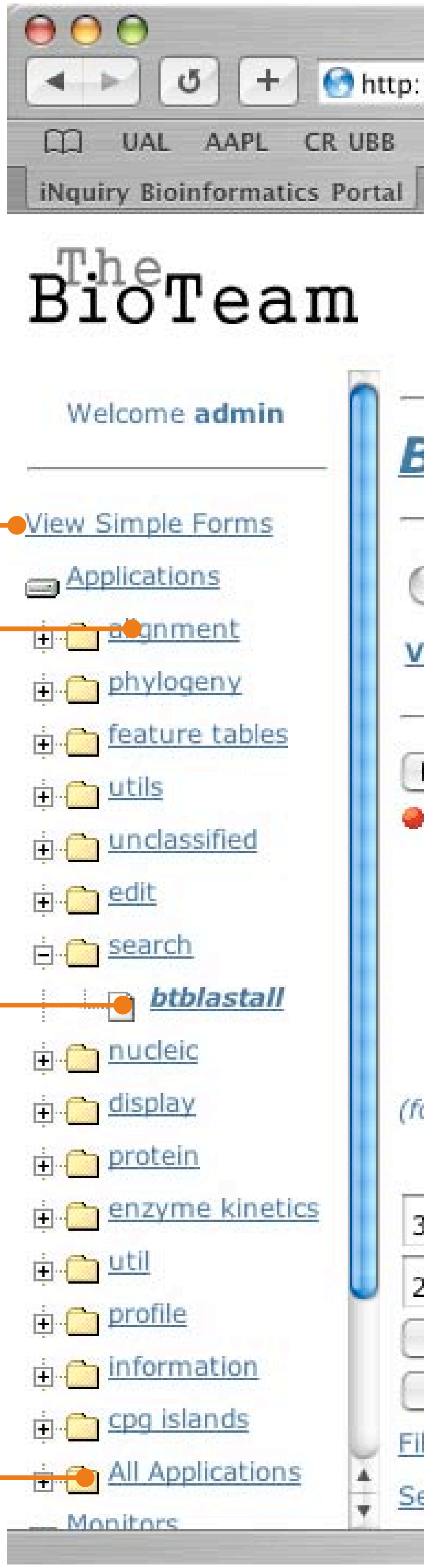
Extensible and modifiable

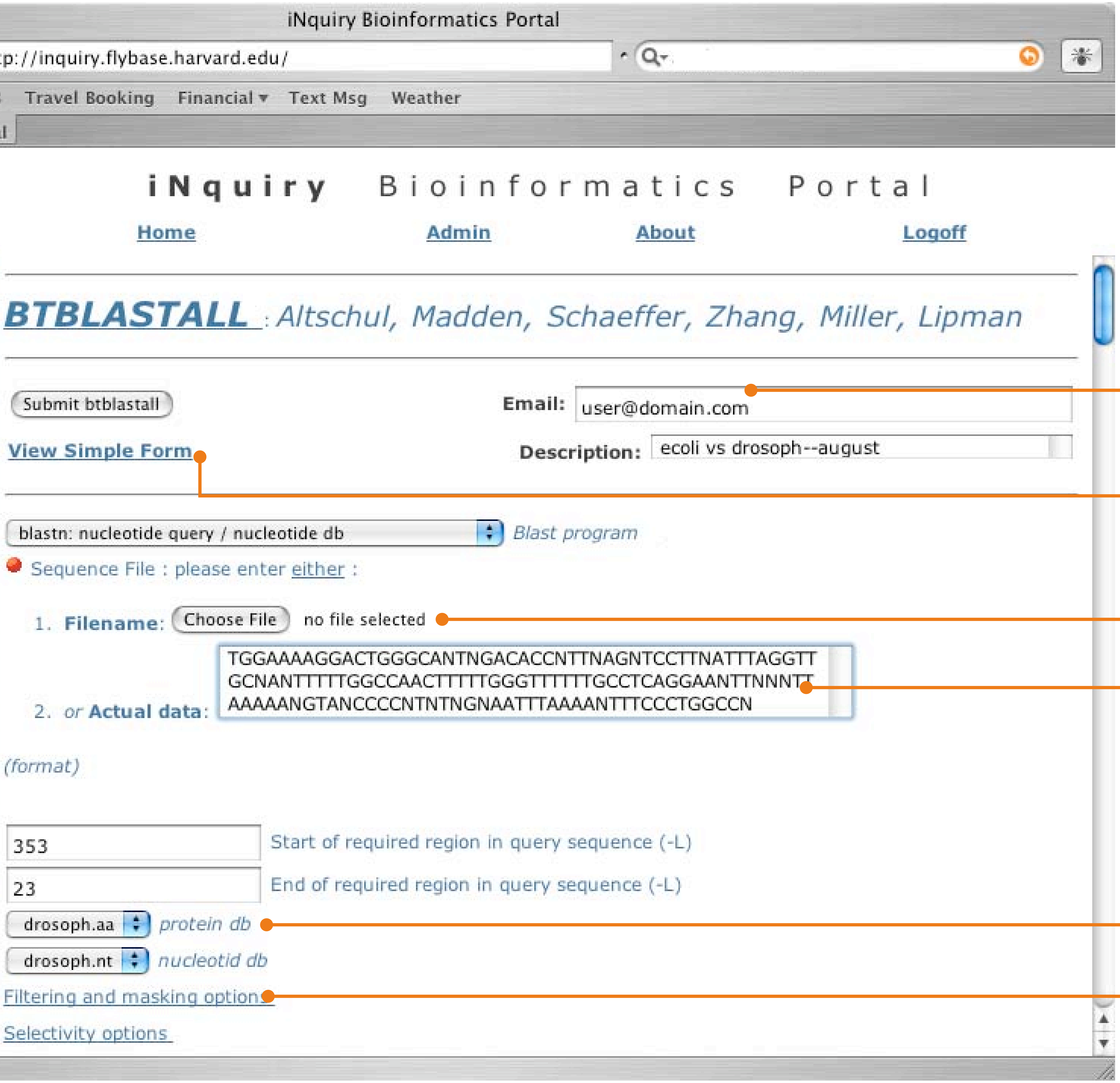
All tools have an "advanced" and "simple" form

Tools are organized into folders by function

The active tool is emboldened

Applications can be viewed in a flat list




iNquiry Bioinformatics Portal
p://inquiry.flybase.harvard.edu/
Travel Booking
Financial ▾
Text Msg
Weather

iNquiry Bioinformatics Portal

Home Admin About Logoff

BTBLASTALL : *Altschul, Madden, Schaeffer, Zhang, Miller, Lipman*

Email: user@domain.com

Description: ecoli vs drosoph--august

blastn: nucleotide query / nucleotide db — *Blast program*

● Sequence File : please enter either :

1. Filename: Choose File no file selected
2. *or* Actual data:

```
TGGAAAAGGACTGGGCANTNGACACCNTTNAGNTCCTTNATTTAGGTT
GCNANTTTTTGGCCAACTTTTTGGGTTTTTTGCCTCAGGAANTTNNNTT
AAAAANGTANCCCCNTNTNGNAATTTAAAANTTTCCCTGGCCN
```

(format)

353 — Start of required region in query sequence (-L)

23 — End of required region in query sequence (-L)

drosoph.aa — *protein db*

drosoph.nt — *nucleotid db*

Results are e-mailed to user
(and displayed in the web browser)

Quick access to simple forms for standard searches

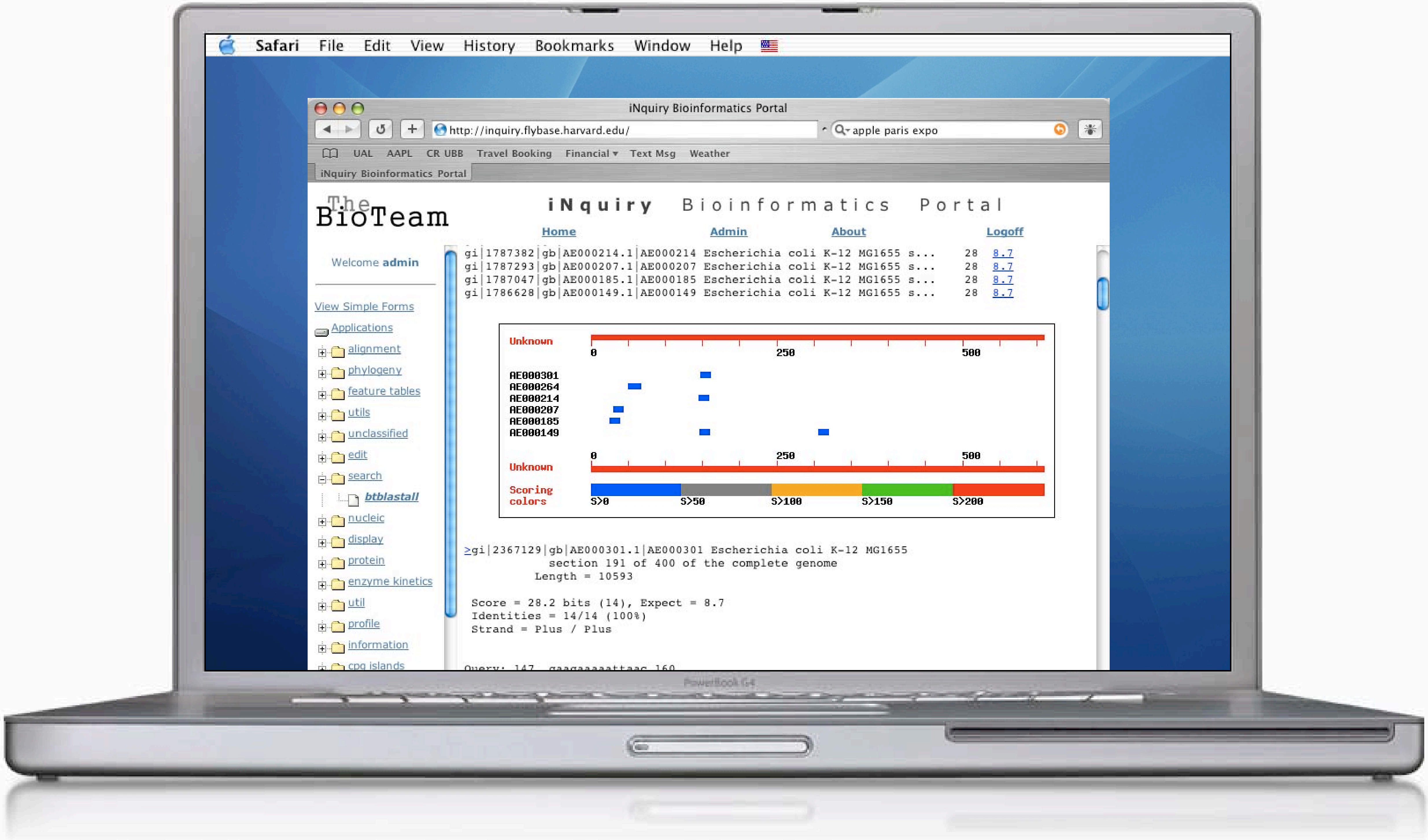
Data entry via file upload

Direct data entry

Quick access to installed databases

All command-line options are available in the web interface

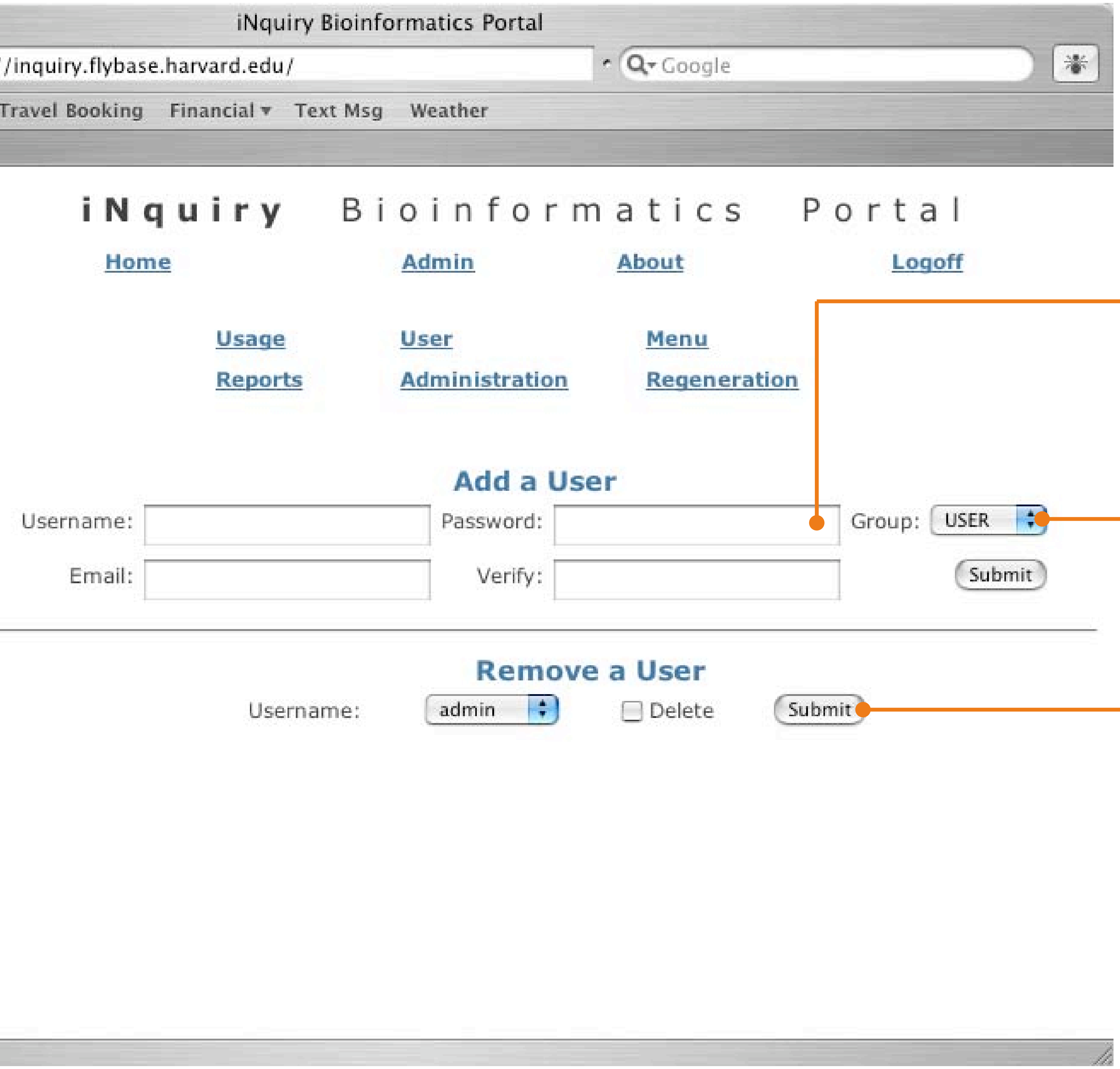
Web-based Results



```
gi|1787382|gb|AE000214.1|AE000214 Escherichia coli K-12 MG1655 s...    28   8.7
gi|1787293|gb|AE000207.1|AE000207 Escherichia coli K-12 MG1655 s...    28   8.7
gi|1787047|gb|AE000185.1|AE000185 Escherichia coli K-12 MG1655 s...    28   8.7
gi|1786628|gb|AE000149.1|AE000149 Escherichia coli K-12 MG1655 s...    28   8.7
```

```
>gi|2367129|gb|AE000301.1|AE000301 Escherichia coli K-12 MG1655
            section 191 of 400 of the complete genome
          Length = 10593

 Score = 28.2 bits (14), Expect = 8.7
 Identities = 14/14 (100%)
 Strand = Plus / Plus


Query: 147  gaagaaaaattaac 160
```

Simple User Management



iNquiry Bioinformatics Portal

/inquiry.flybase.harvard.edu/

iNquiry Bioinformatics Portal

Add a User

Username: ____ Password: ____ Group: USER

Email: ____ Verify: ____ Submit

Remove a User

Username: admin □ Delete Submit

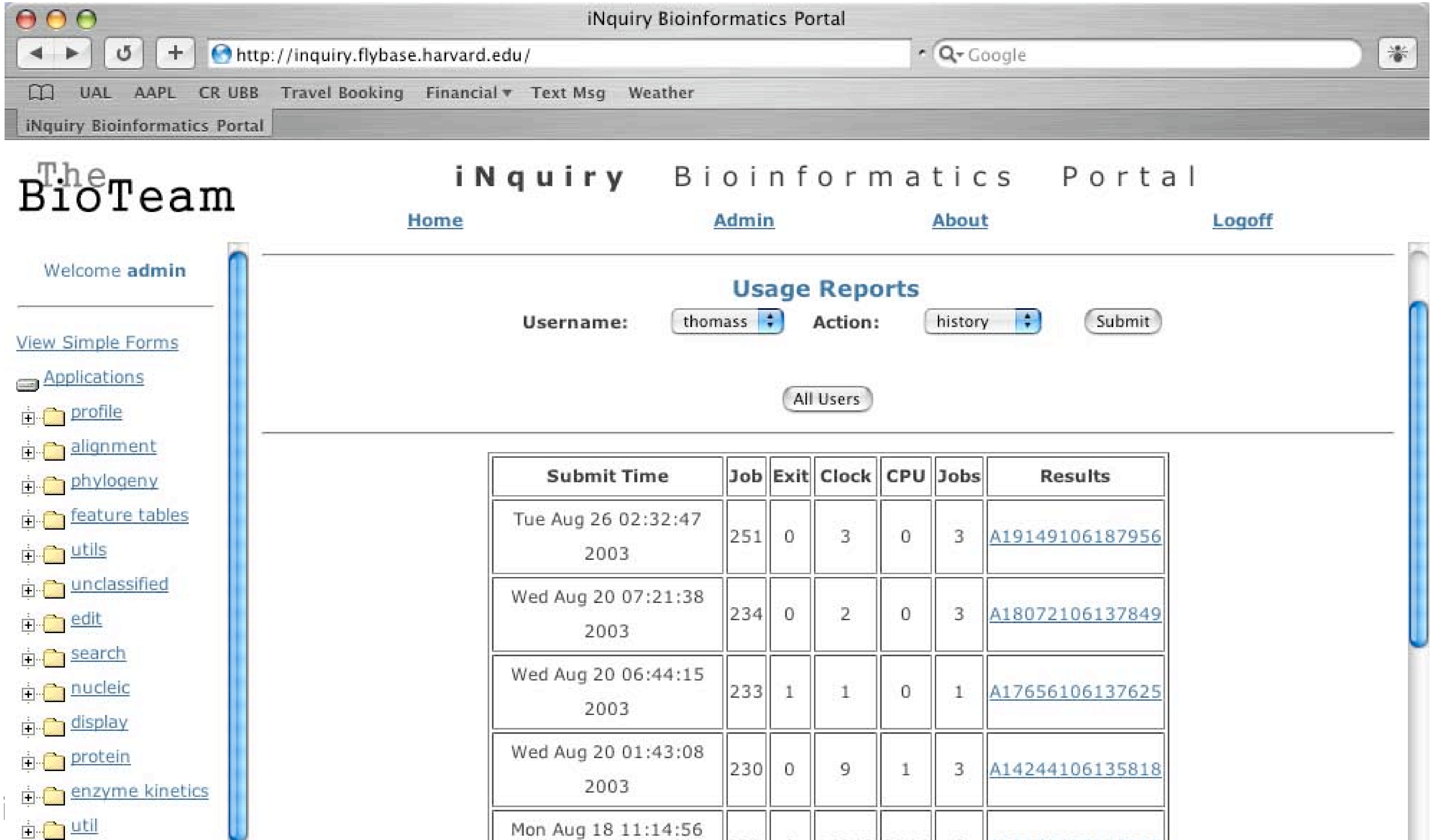
Minimal effort to create a new user

Choose between regular user or administrator

Single-click user disabling or deletion

Usage Reporting

Quickly see results and user statistics



Usage Reports

Username: thomass **Action:** history

Submit Time	Job	Exit	Clock	CPU	Jobs	Results
Tue Aug 26 02:32:47 2003	251	0	3	0	3	A19149106187956
Wed Aug 20 07:21:38 2003	234	0	2	0	3	A18072106137849
Wed Aug 20 06:44:15 2003	233	1	1	0	1	A17656106137625
Wed Aug 20 01:43:08 2003	230	0	9	1	3	A14244106135818
Mon Aug 18 11:14:56						

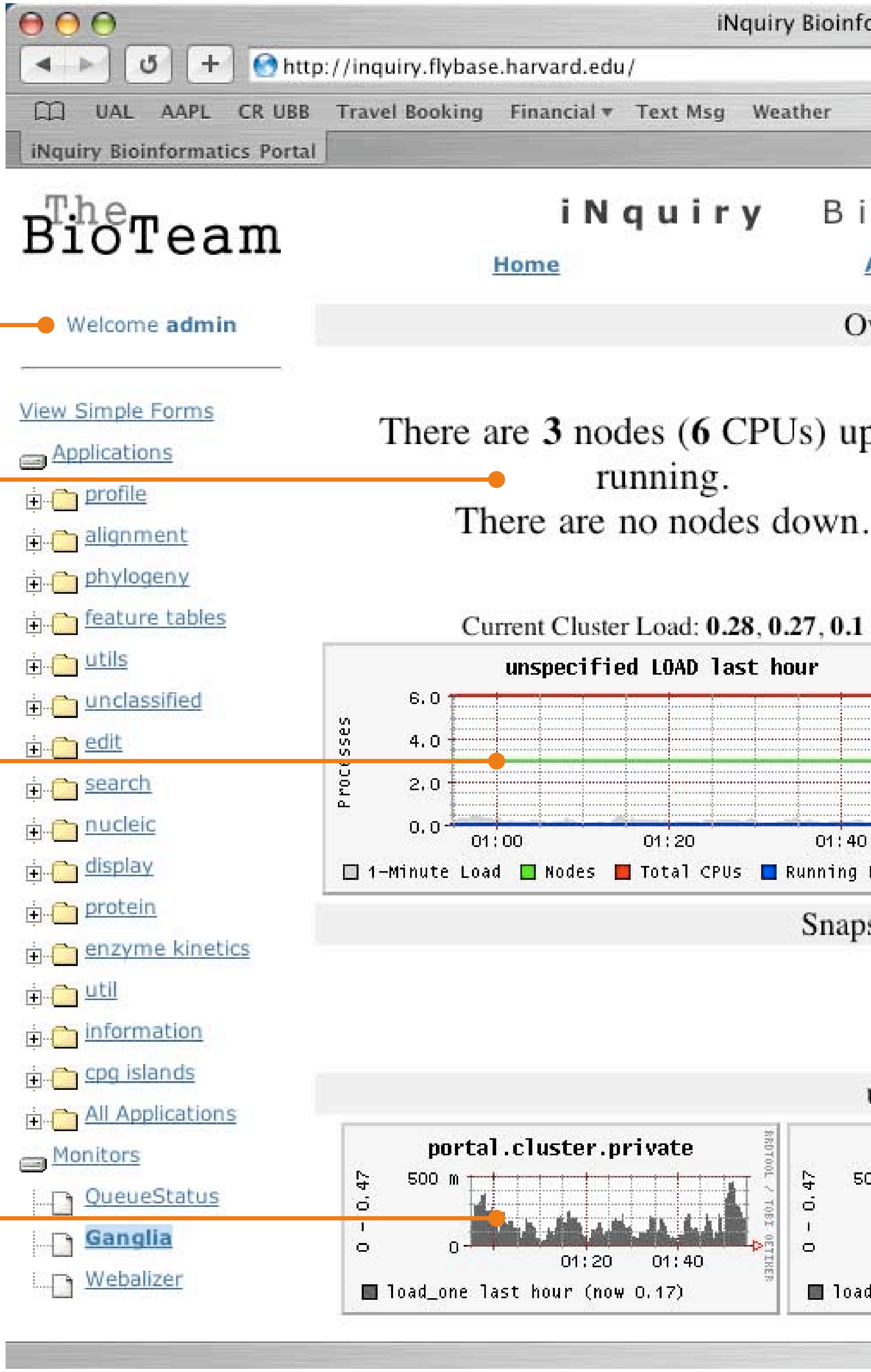
Cluster Monitoring

Only admin users have access to management tools

Instantly see if any cluster nodes are offline

See load statistics for the entire cluster

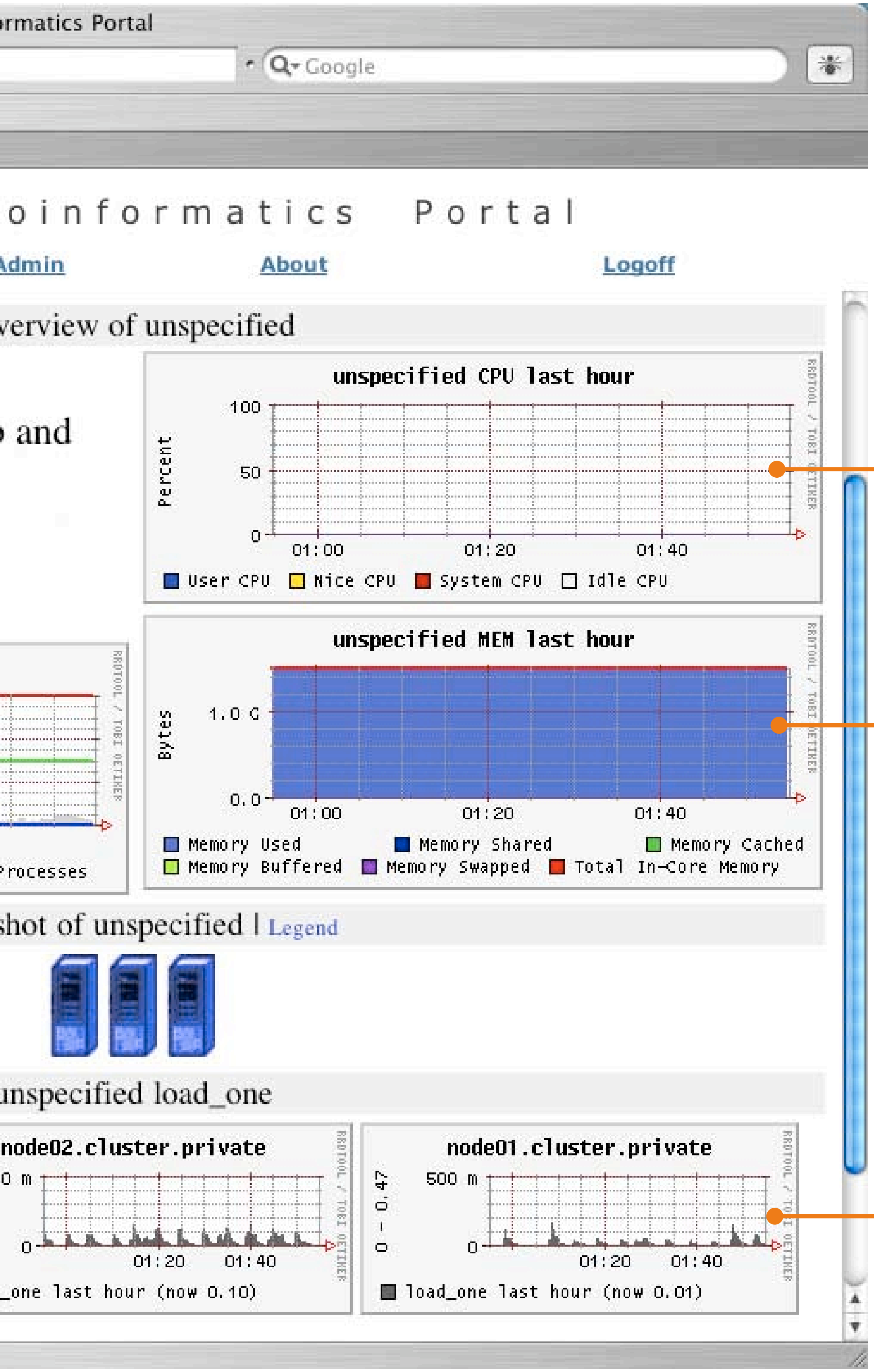
See load statistics for individual cluster nodes, including the portal (head) node



Welcome **admin**

There are **3** nodes (**6** CPUs) up running.
There are no nodes down.

Current Cluster Load: **0.28**, **0.27**, **0.1**



Histograms of CPU utilization by percentage

Histograms of memory usage by gigabyte

Load histograms for individual cluster nodes

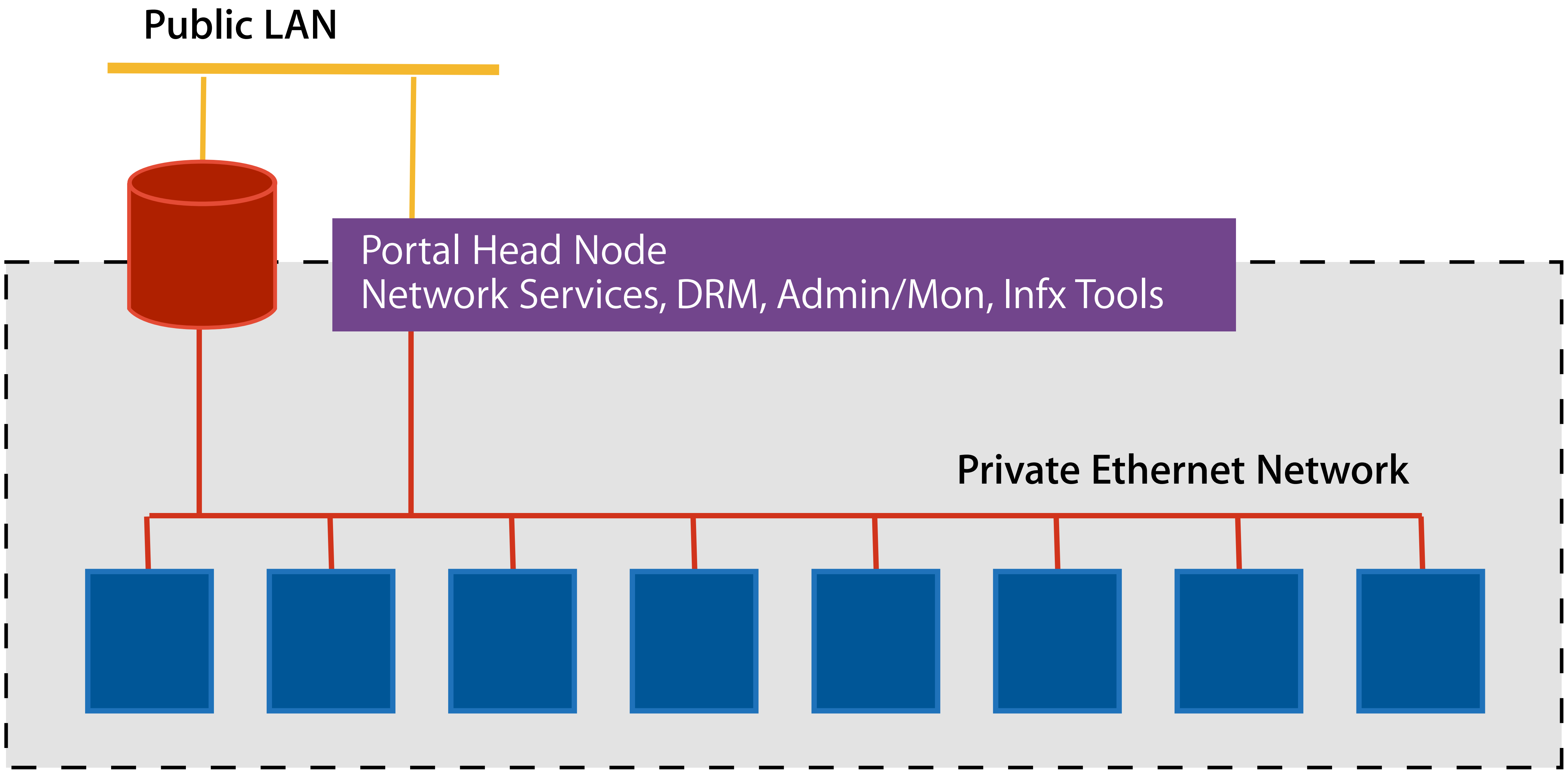
BioTeam's iNquiry - Solution Summary

Fully provisioned bioinformatics portal

INQUIRY

- Turnkey bioinformatics solution
- Web-based front end to bioinformatics cluster
- Useful for biologists that are non-bioinformaticists
- Sophisticated Unix level access for power users
- Many applications are Velocity Engine performance optimized
- Excellent solution for G4 Xserves
- Fast and simple deployment of a compute cluster

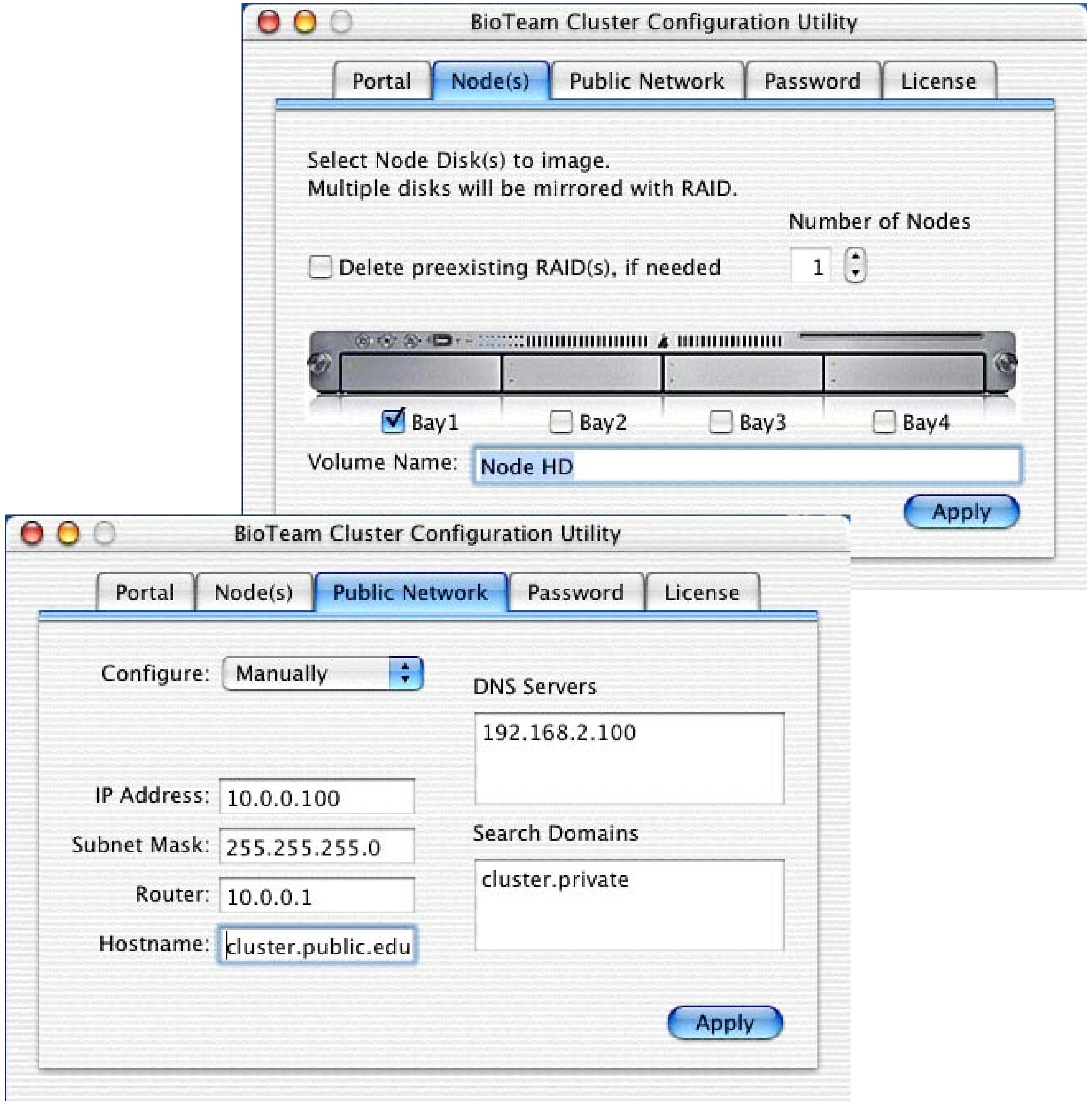
BioTeam iNquiry



iNquiry Installation ...

Step One

- Cluster Configuration



Select Node Disk(s) to image.
Multiple disks will be mirrored with RAID.

□ Delete preexisting RAID(s), if needed

Number of Nodes: 1

☑ Bay1 □ Bay2 □ Bay3 □ Bay4

Volume Name: Node HD

Configure: Manually

IP Address: 10.0.0.100

Subnet Mask: 255.255.255.0

Router: 10.0.0.1

Hostname: cluster.public.edu

DNS Servers: 192.168.2.100

Search Domains: cluster.private

iNquiry Installation ...

Step Two

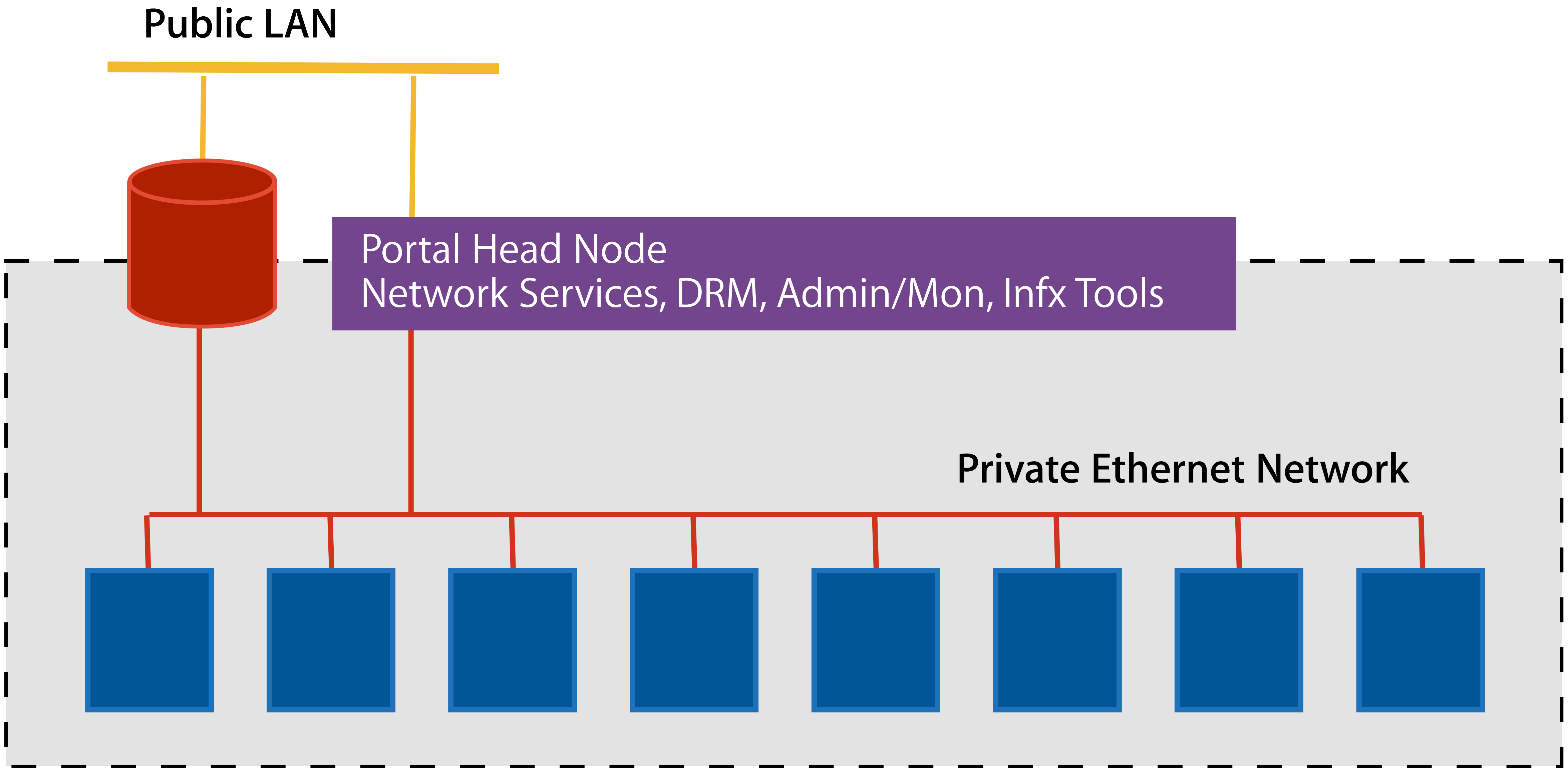
- Boot and Re-image head node from external FireWire device

INQUIRY

iNquiry Installation ...

Step Three

- Boot and re-image cluster from head node



Q&A

The
BioTeam