Quantitative Biology Bootcamp

Michael Schatz, Justin Kinney, Mickey Atwal

Aug 27, 2014
WSBS

Unsolved Questions in Biology

- What is your genome sequence?
- How does your genome compare to my genome?

- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- How do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs and treatments should we give you?

- ***Plus thousands and thousands more***

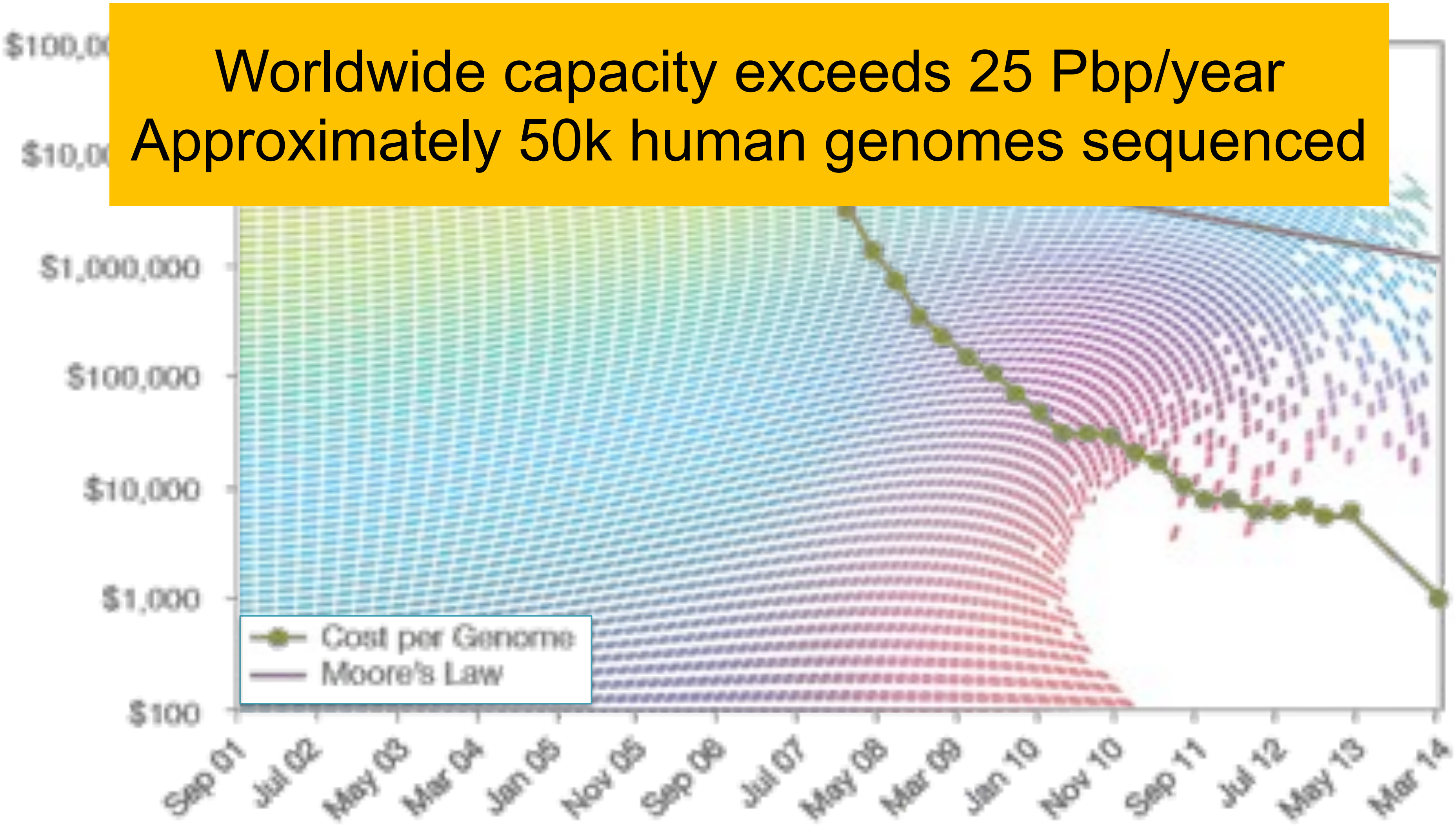
Data types across the NIH

Genomic
Other 'omic
Imaging
Phenotypic
Exposure
Clinical
Complexity
Courtesy of NHGRI

Phil Bourne, Associate Director of Data Science for NIH
http://www.slideshare.net/pebourne/wiki-mania080914

Cost per Genome

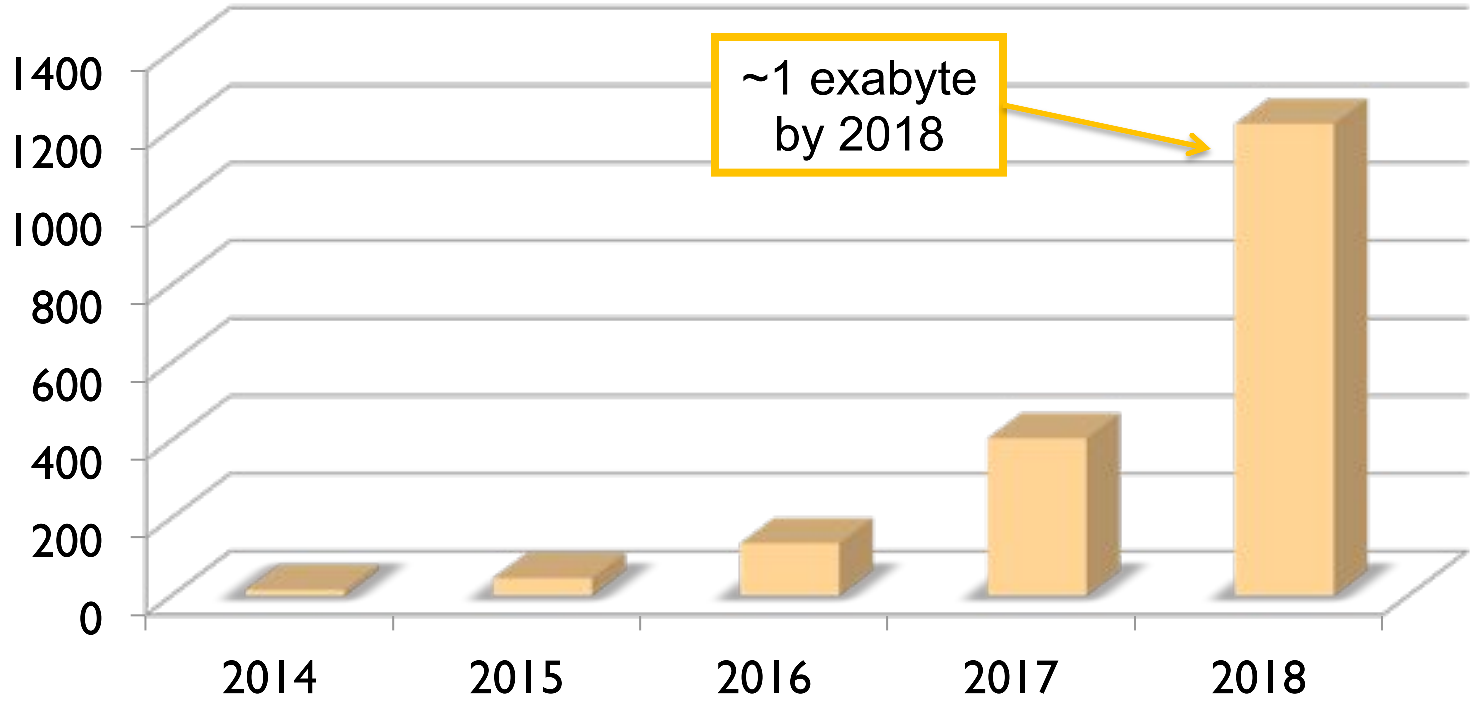
Worldwide capacity exceeds 25 Pbp/year
Approximately 50k human genomes sequenced



http://www.genome.gov/sequencingcosts/

DNA Data Tsunami

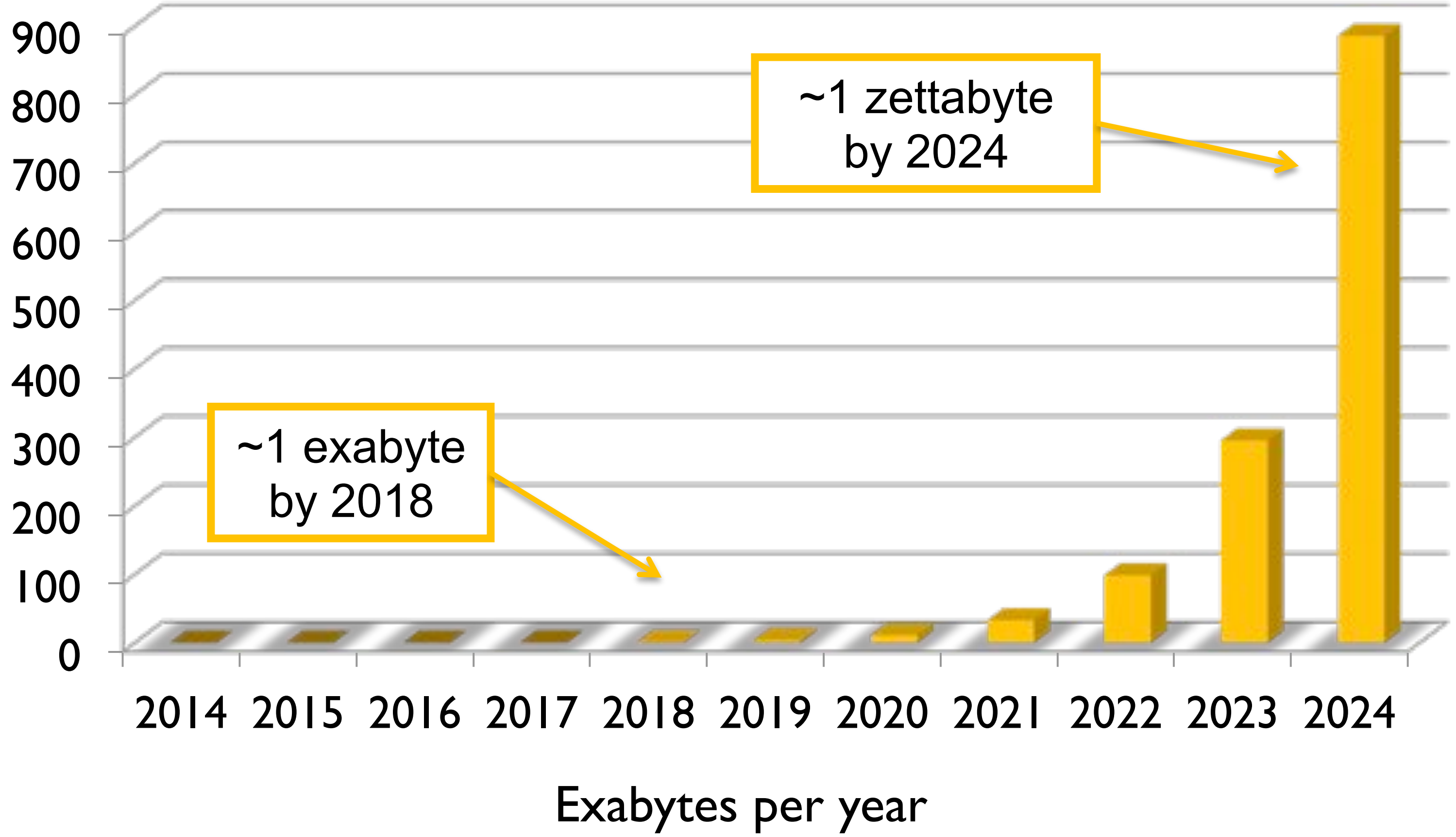
Current world-wide sequencing capacity is growing at ~3x per year!



Petabytes per year

DNA Data Tsunami

Current world-wide sequencing capacity is growing at ~3x per year!



~1 zettabyte
by 2024

~1 exabyte
by 2018

How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

Both currently ~100Pb
And growing exponentially

Unsolved Questions in Biology

- What is your genome sequence?

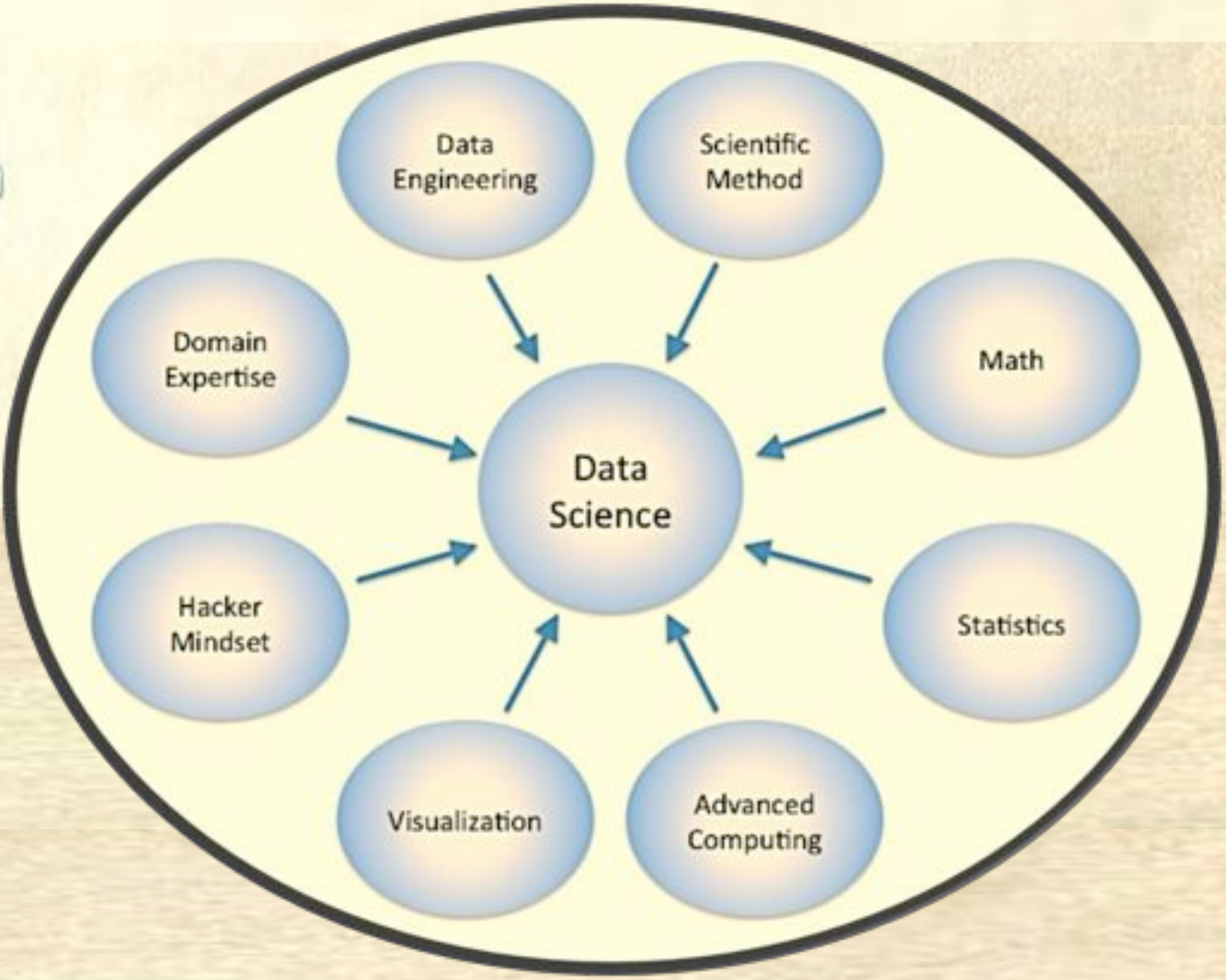
The instruments provide the data, but none of the answers to any of these questions.

What software and systems will?

And who will create them?

- ***Plus thousands and thousands more***

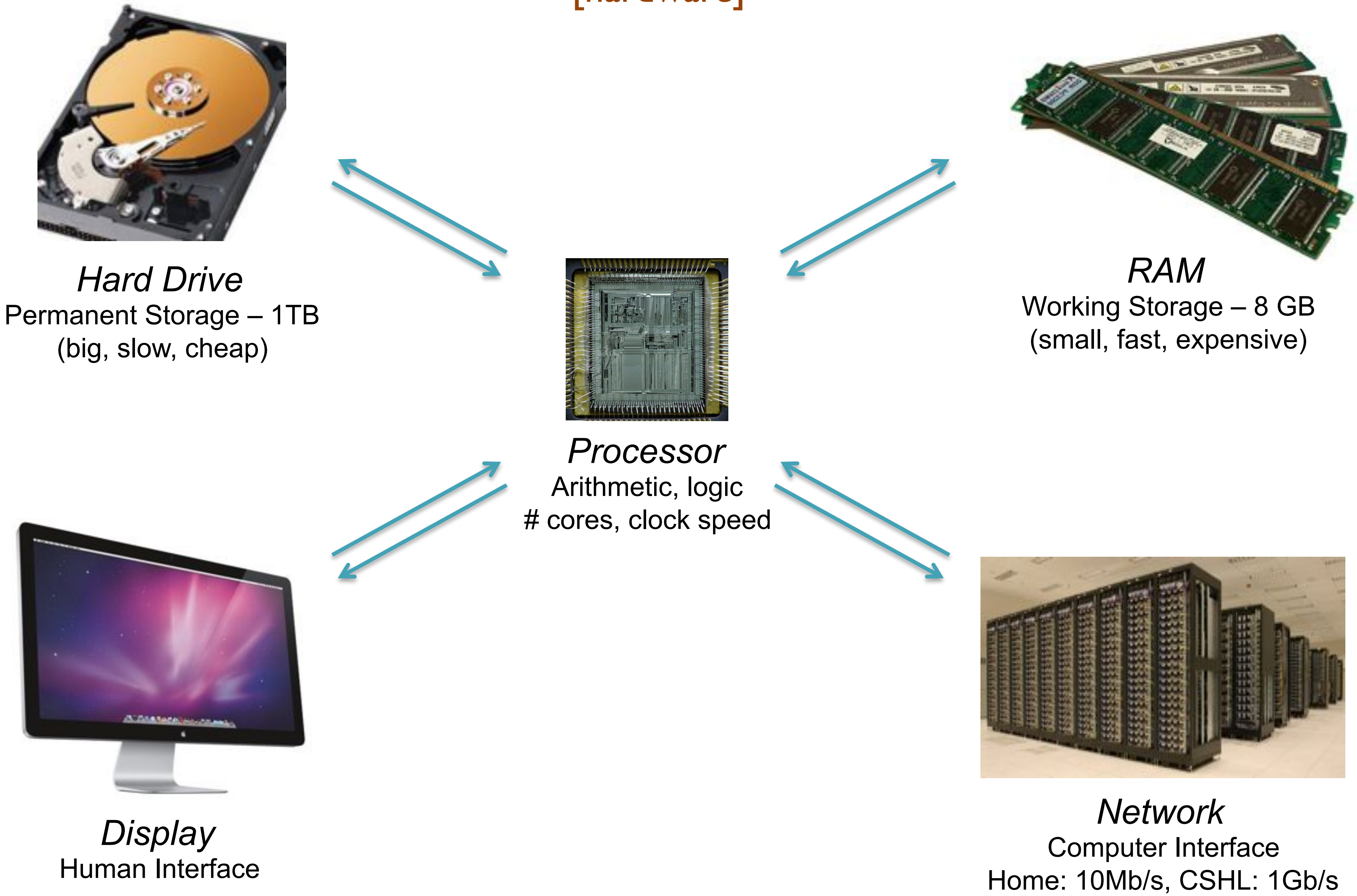
Who is a Data Scientist?



http://en.wikipedia.org/wiki/Data_science

What is a computer?

[hardware]



Hard Drive
Permanent Storage – 1TB
(big, slow, cheap)

RAM
Working Storage – 8 GB
(small, fast, expensive)

Processor
Arithmetic, logic
# cores, clock speed

Display
Human Interface

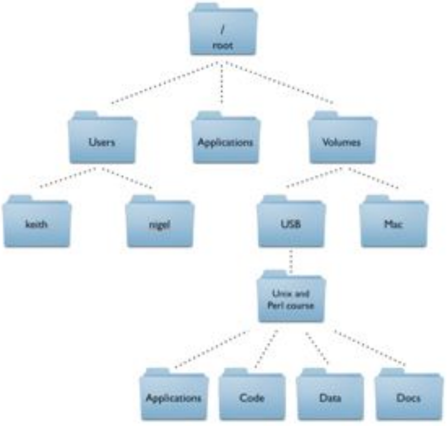
Network
Computer Interface
Home: 10Mb/s, CSHL: 1Gb/s

What is a computer?

[software]

Office Applications

Presentations, Documents
Simple statistics and plots

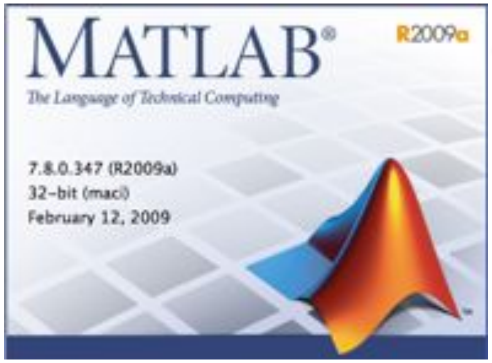


Files / Data

Papers, sequences, measurements

Operating System

Mission Control
Windows, Mac, Unix, iOS



MATLAB® R2009a

The Language of Technical Computing

7.8.0.347 (R2009a)
32-bit (maci)
February 12, 2009

Scientific Applications

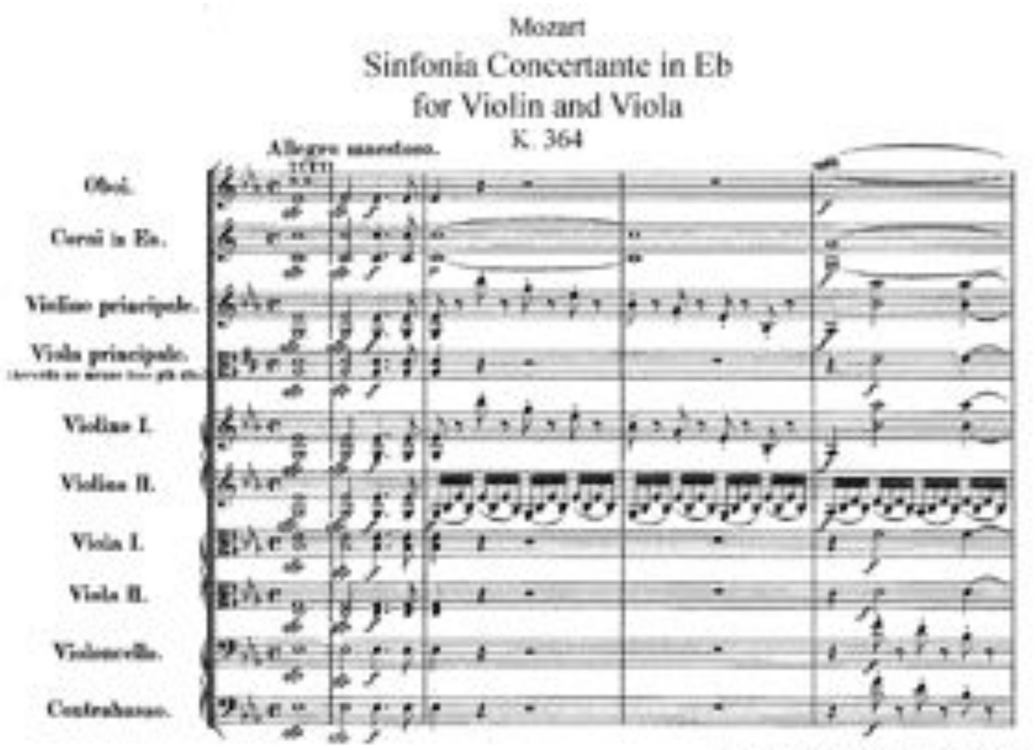
Specialized Analysis
Commercial

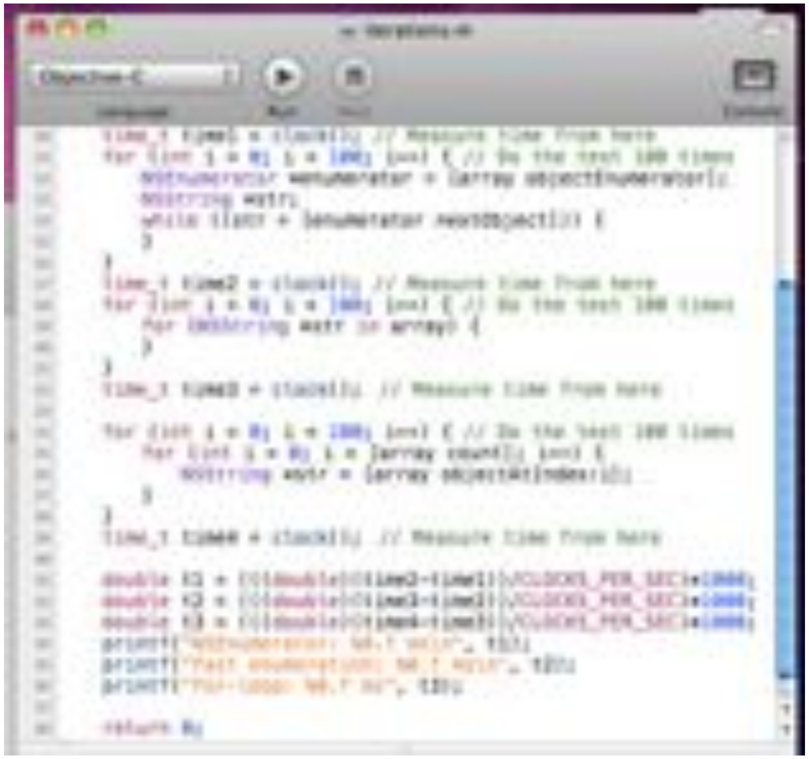
```
time_t time1 = clock(); // Measure time from here
for (int i = 0; i < 100; i++) { // Do the test 100 times
    NSEnumerator *enumerator = [array objectEnumerator];
    NSString *str;
    while ((str = [enumerator nextObject])) {
    }
}
time_t time2 = clock(); // Measure time from here
for (int i = 0; i < 100; i++) { // Do the test 100 times
    for (NSString *str in array) {
    }
}
time_t time3 = clock(); // Measure time from here

for (int i = 0; i < 100; i++) { // Do the test 100 times
    for (int j = 0; j < [array count]; j++) {
        NSString *str = [array objectAtIndex:j];
    }
}
time_t time4 = clock(); // Measure time from here

double t1 = (((double)(time2-time1))/CLOCKS_PER_SEC)*1000;
double t2 = (((double)(time3-time2))/CLOCKS_PER_SEC)*1000;
double t3 = (((double)(time4-time3))/CLOCKS_PER_SEC)*1000;
printf("NSEnumerator: %0.f ms\n", t1);
printf("Fast enumeration: %0.f ms\n", t2);
printf("For-loop: %0.f ms\n", t3);

return 0;
```

Code / Scripts

Research Applications
Academic

Programming 101

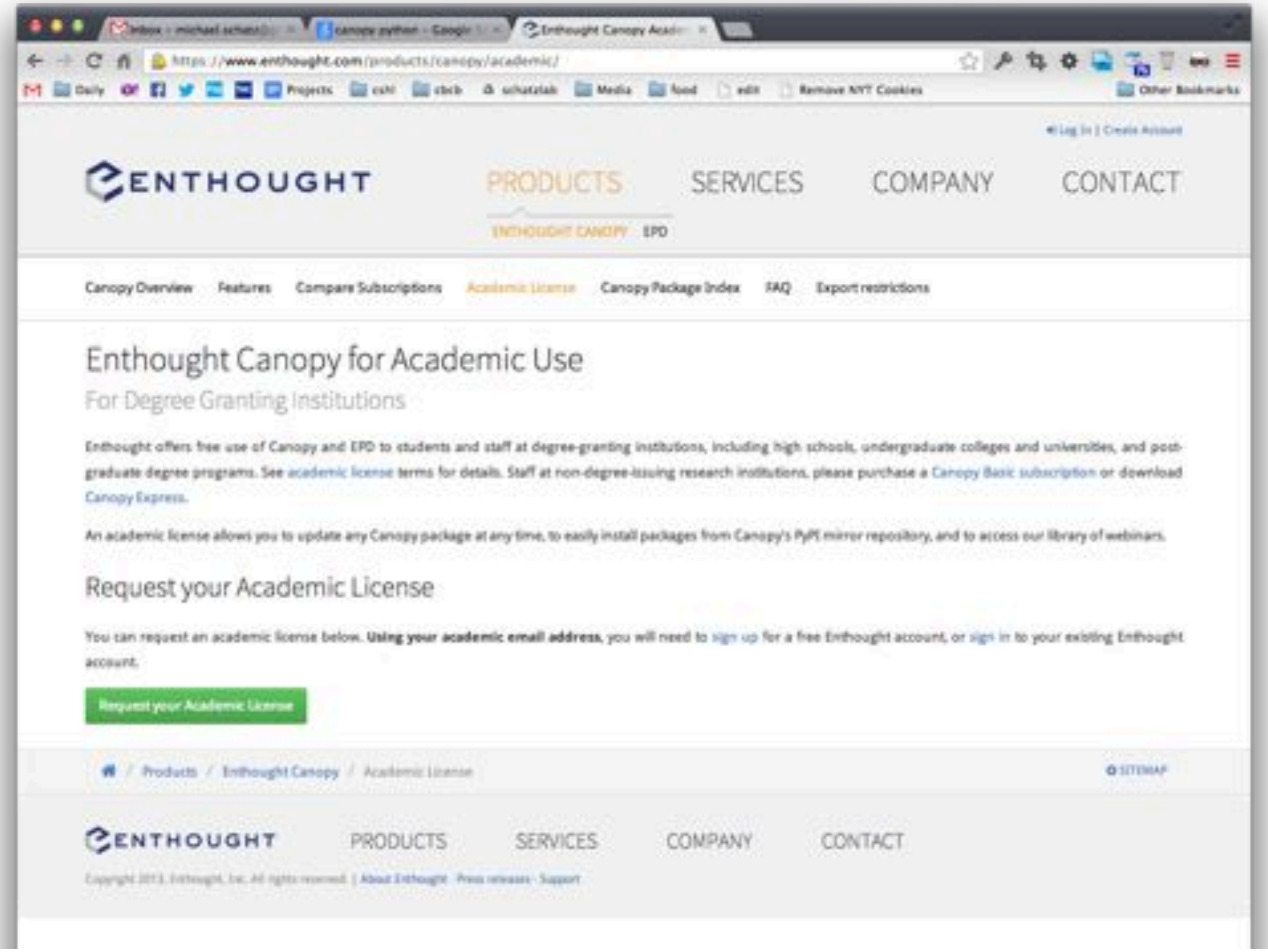




```
    time_t time1 = clock(); // Measure time from here
    for (int i = 0; i < 100; i++) { // Do the test 100 times
        NSEnumerator *enumerator = [array objectEnumerator];
        NSString *str;
        while (str = [enumerator nextObject]) {
        }
    }
    time_t time2 = clock(); // Measure time from here
    for (int i = 0; i < 100; i++) { // Do the test 100 times
        for (NSString *str in array) {
        }
    }
    time_t time3 = clock(); // Measure time from here

    for (int i = 0; i < 100; i++) { // Do the test 100 times
        for (int j = 0; j < [array count]; j++) {
            NSString *str = [array objectAtIndex:j];
        }
    }
    time_t time4 = clock(); // Measure time from here

    double t1 = ((double)(time2-time1))/CLOCKS_PER_SEC*1000;
    double t2 = ((double)(time3-time2))/CLOCKS_PER_SEC*1000;
    double t3 = ((double)(time4-time3))/CLOCKS_PER_SEC*1000;
    printf("NSEnumerator: %0.f ms\n", t1);
    printf("Fast enumeration: %0.f ms\n", t2);
    printf("For-loop: %0.f ms", t3);

    return 0;
```

A software program is like sheet music for the orchestra inside your computer
Static, written representations of an active process

Programming with Python



Enthought Canopy for Academic Use

For Degree Granting Institutions

Enthought offers free use of Canopy and EPD to students and staff at degree-granting institutions, including high schools, undergraduate colleges and universities, and post-graduate degree programs. See academic license terms for details. Staff at non-degree-issuing research institutions, please purchase a Canopy Basic subscription or download Canopy Express.

An academic license allows you to update any Canopy package at any time, to easily install packages from Canopy's PyPI mirror repository, and to access our library of webinars.

Request your Academic License

You can request an academic license below. **Using your academic email address**, you will need to sign up for a free Enthought account, or sign in to your existing Enthought account.

Request your Academic License
Products / Enthought Canopy / Academic License
SITEMAP
ENTHOUGHT
PRODUCTS
SERVICES
COMPANY
CONTACT

https://www.enthought.com/products/canopy/academic/
http://www.codecademy.com/tracks/python