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***

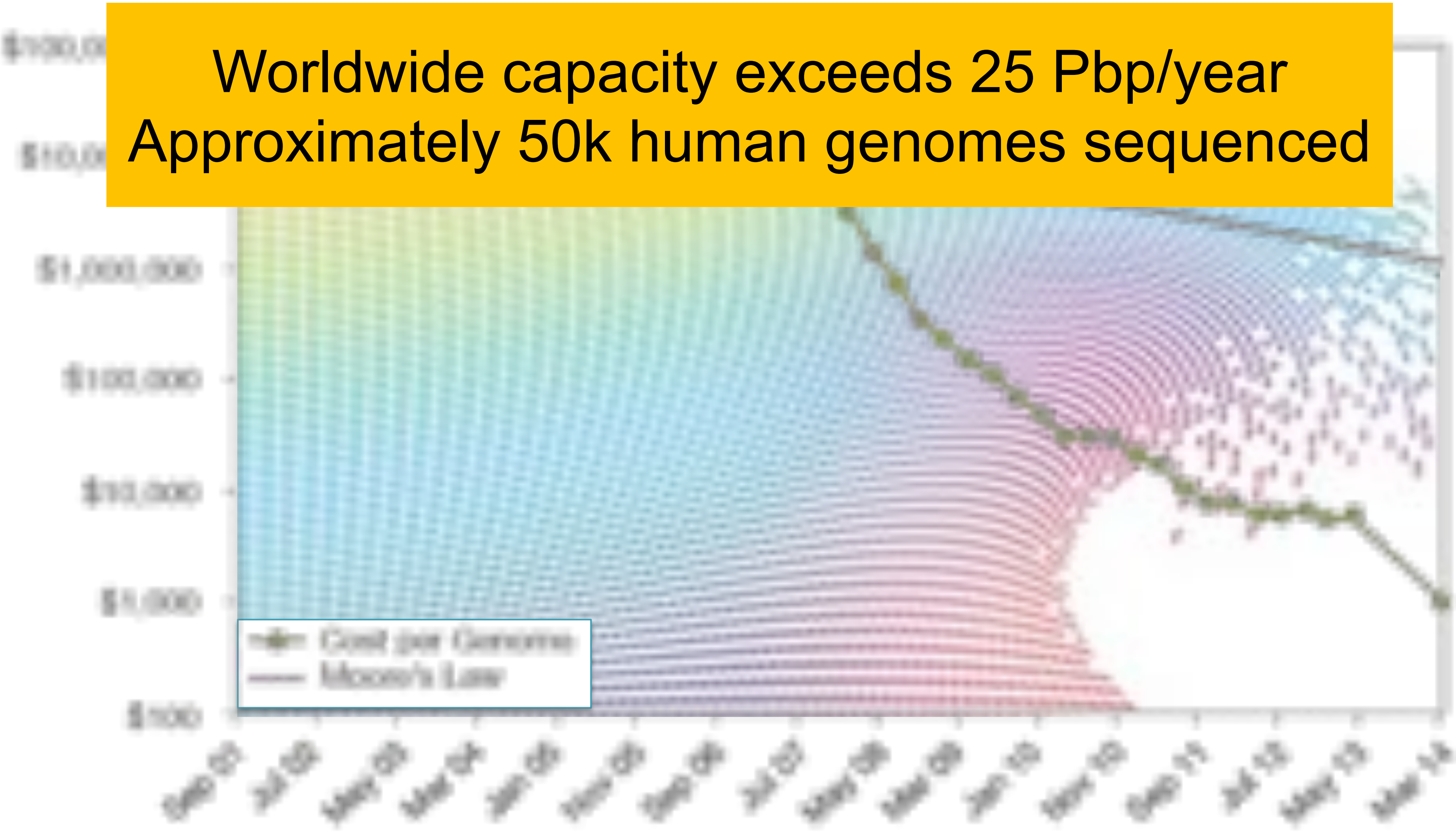
Data types across the NIH

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

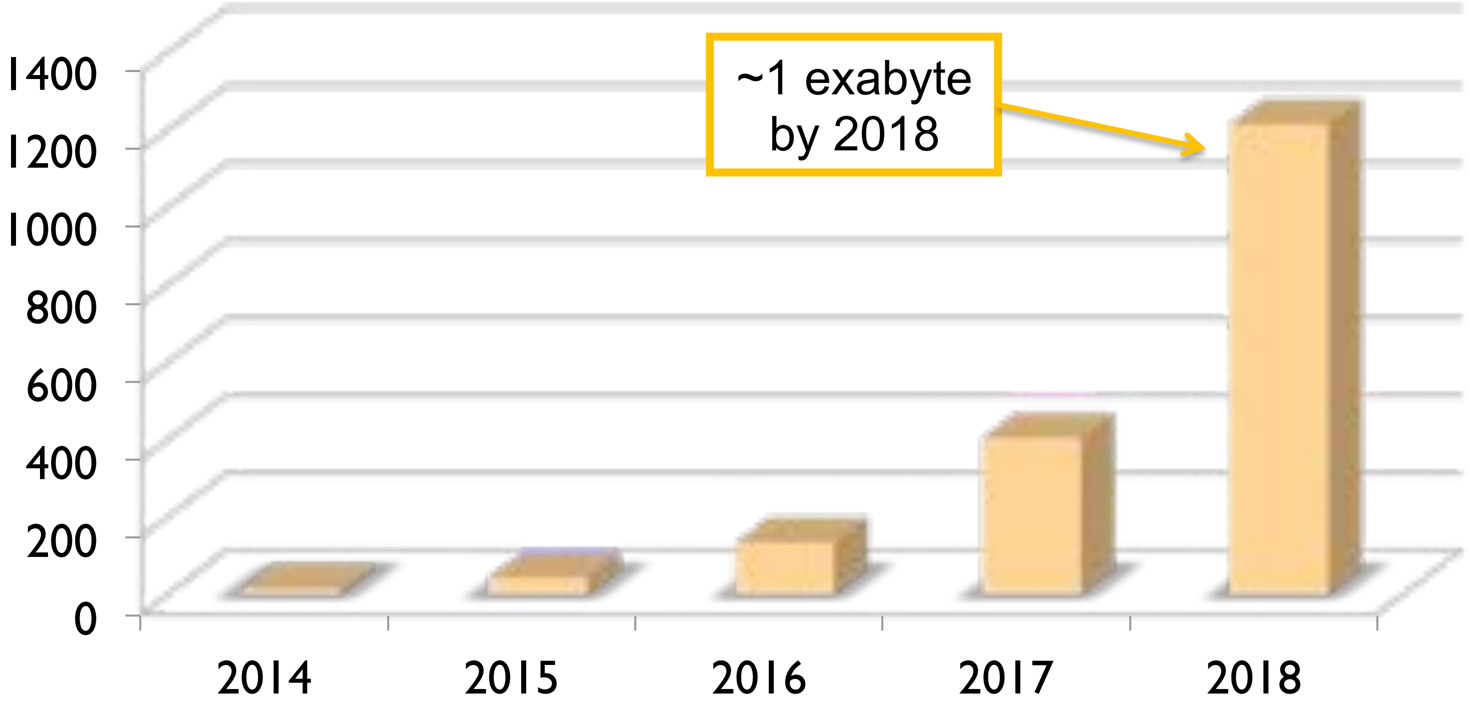
Worldwide capacity exceeds 25 Pbp/year
Approximately 50k human genomes sequenced



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

DNA Data Tsunami

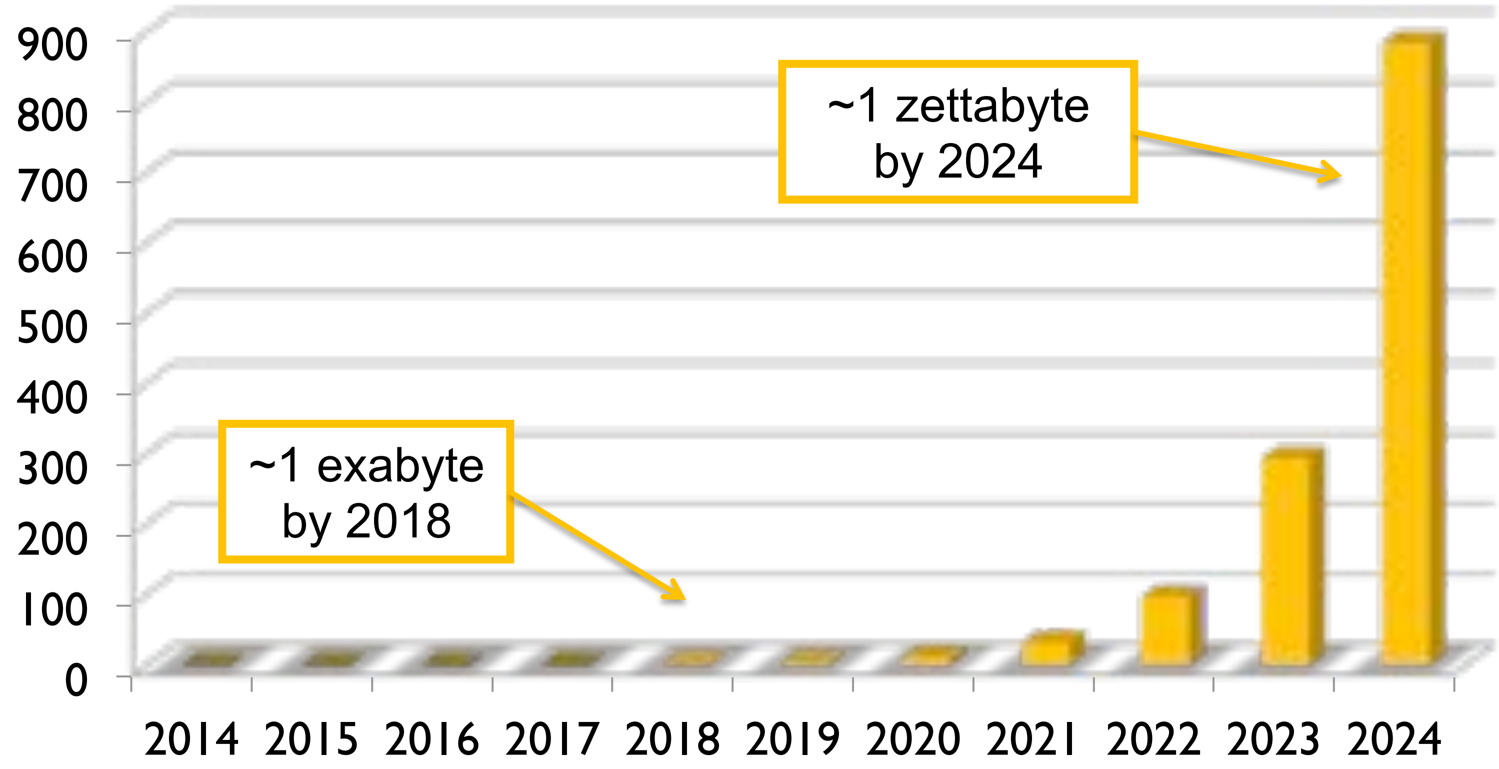
Current world-wide sequencing capacity is growing at ~3x per year!



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

Exabytes per year

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?

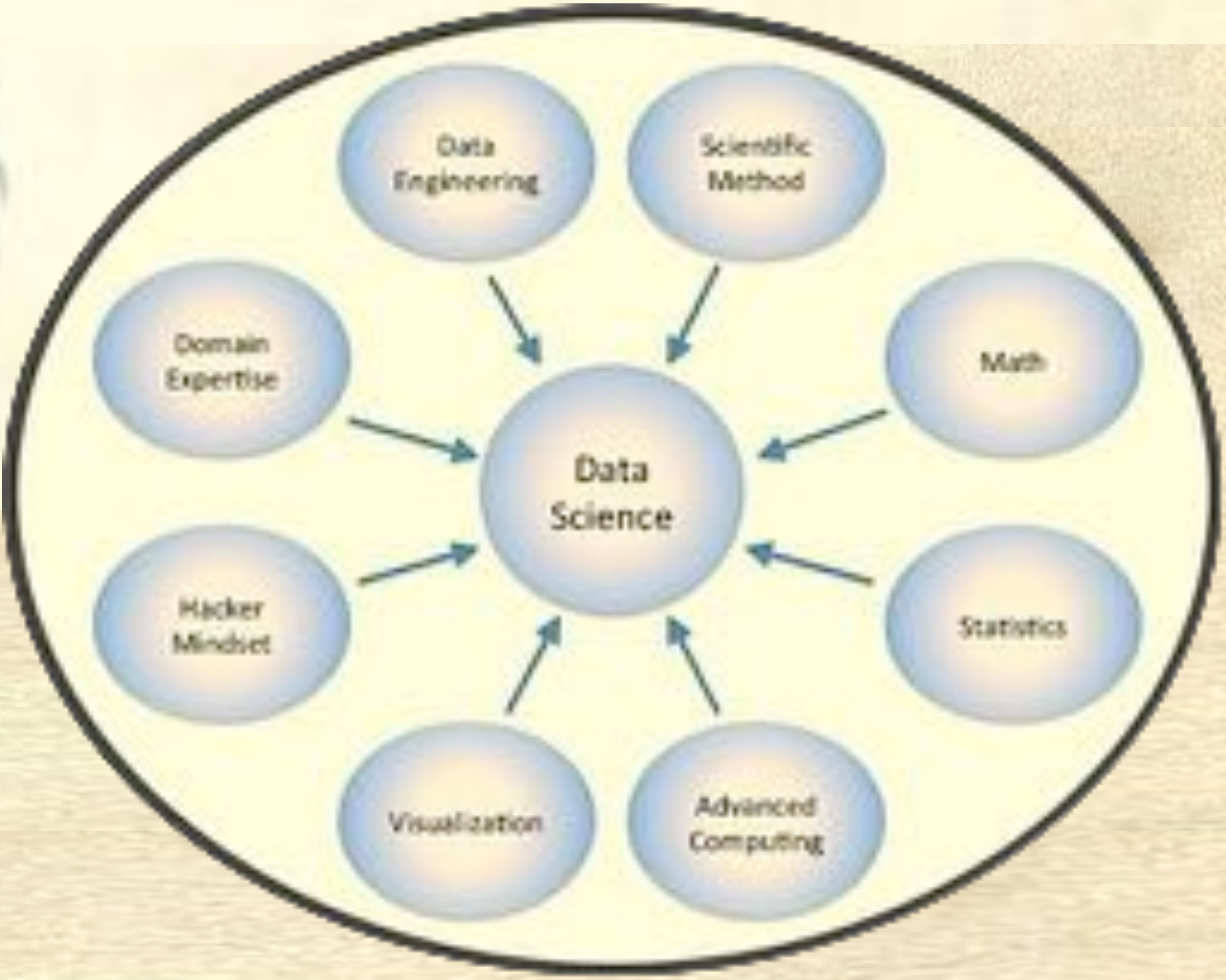
The instruments provide the data, but none of the answers to any of these questions.

What software and systems will?

And who will create them?

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

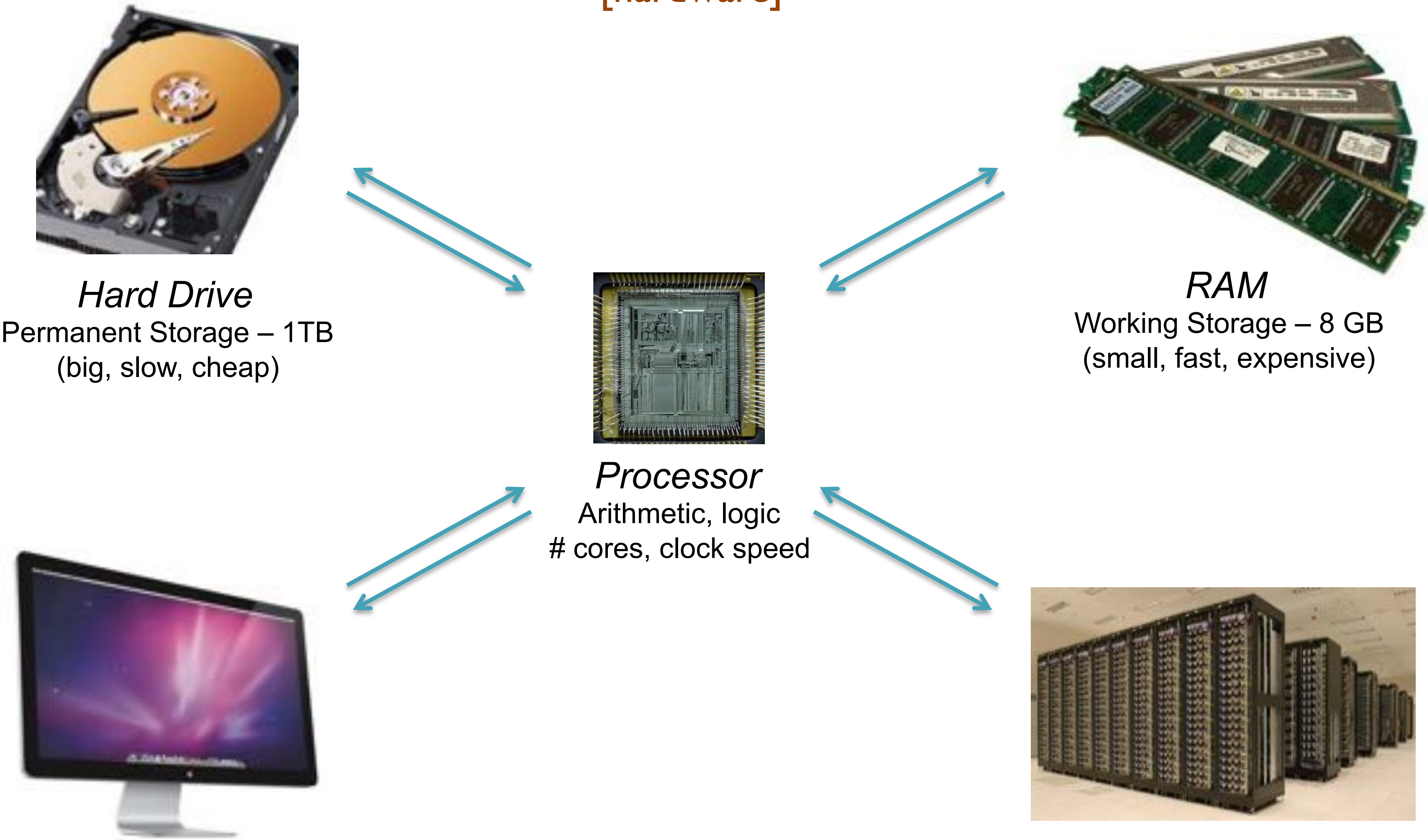
Who is a Data Scientist?



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

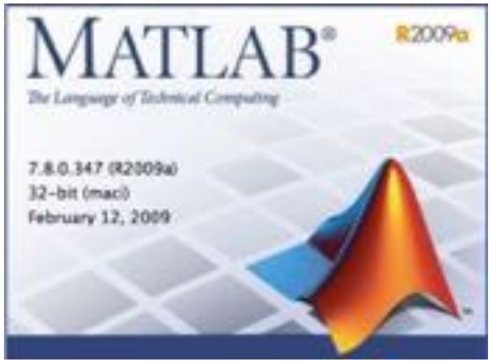
What is a computer?

[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

Code / Scripts
Research Applications
Academic

Programming 101

Mozart
Sinfonia Concertante in Eb
for Violin and Viola
K. 364
Allegro maestoso
Oboi
Corni in Eb
Violino principale
Viola principale
Violino I
Violino II
Viola I
Viola II
Violoncello
Contrabasso

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

Request your Academic License

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