

Programming and p-values

Michael Schatz

Sept 6, 2017 – Lecture 2

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research



Welcome!

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components:

1. **Lectures** on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing
2. **Research presentations** from distinguished faculty on their active research projects
3. **A major research project** to be performed under the mentorship of a JHU professor.

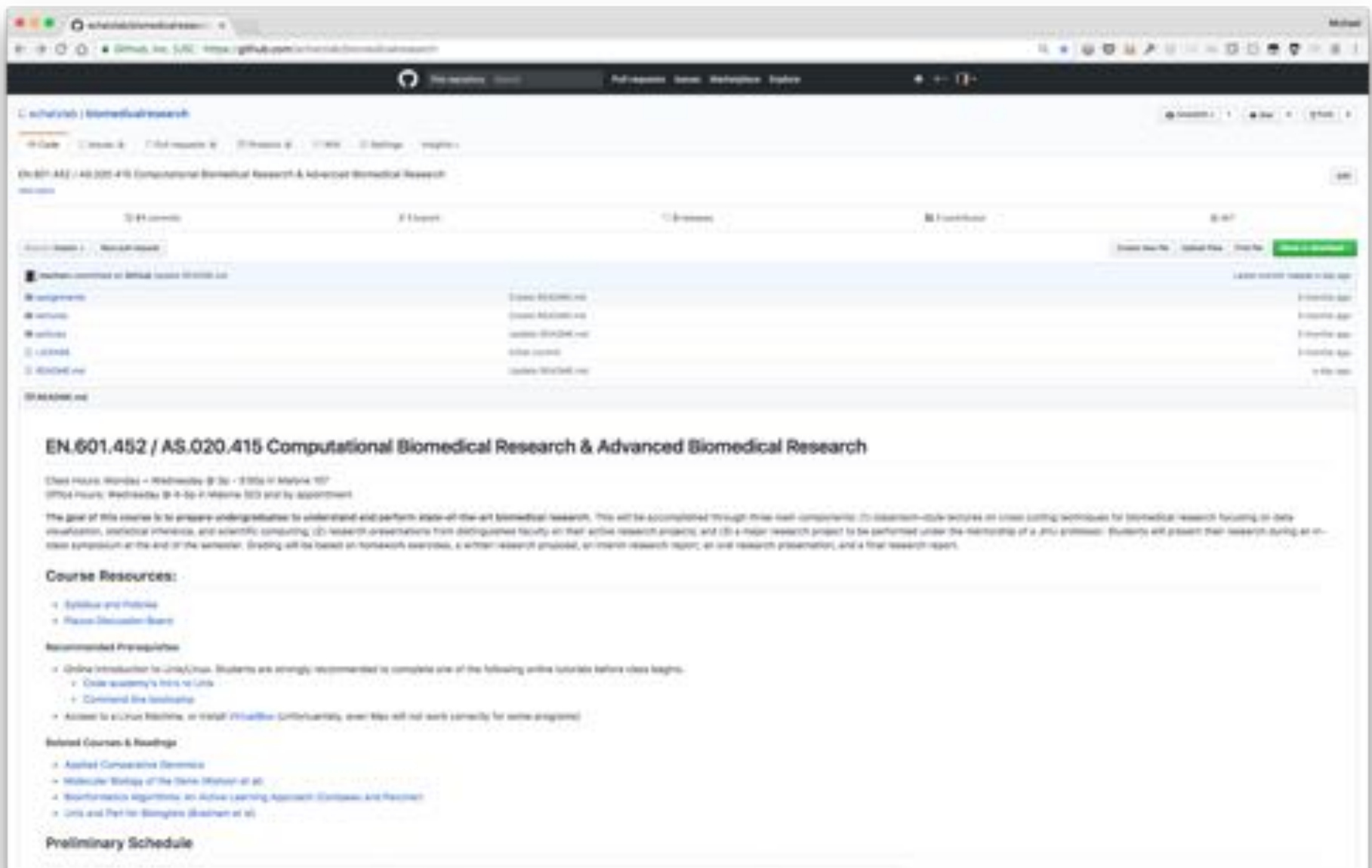
Course Webpage: <https://github.com/schatzlab/biomedicalresearch>

Course Discussions: <http://piazza.com>

Class Hours: Mon + Wed @ 3p – 3:50p Malone 107

Office Hours: Wed @ 4-5p and by appointment
Please try Piazza first!

Course Webpage



The screenshot shows the GitHub repository page for 'schatzlab/biomedicalresearch'. The repository is titled 'EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research'. It has 24 commits, 1 branch, 11 issues, and 1 pull request. The repository is a public repository for the course. The main content area displays the repository's README, which includes the course title, a brief description of the course, and a list of course resources and recommended prerequisites.

EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research

Class Hours: Monday - Wednesday 9:30 - 11:00 in Marina 107
Office Hours: Wednesday 9:30 - 11:00 in Marina 107 and by appointment

The goal of this course is to prepare undergraduate students to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components: (1) classroom-style lectures on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing; (2) research presentations from distinguished faculty on their active research projects; and (3) a major research project to be performed under the mentorship of a JHU professor. Students will present their research during an on-site symposium at the end of the semester. Grading will be based on homework exercises, a written research proposal, an interim research report, an oral research presentation, and a final research report.

Course Resources:

- Syllabus and Policies
- Piazza Discussion Board

Recommended Prerequisites:

- Online Introduction to Linux/Unix. Students are strongly recommended to complete one of the following online tutorials before class begins:
 - Codecademy's Intro to Linux
 - Command Line Hero
- Access to a Linux Machine, or install VirtualBox. Unfortunately, over time we will not work correctly for some programs.

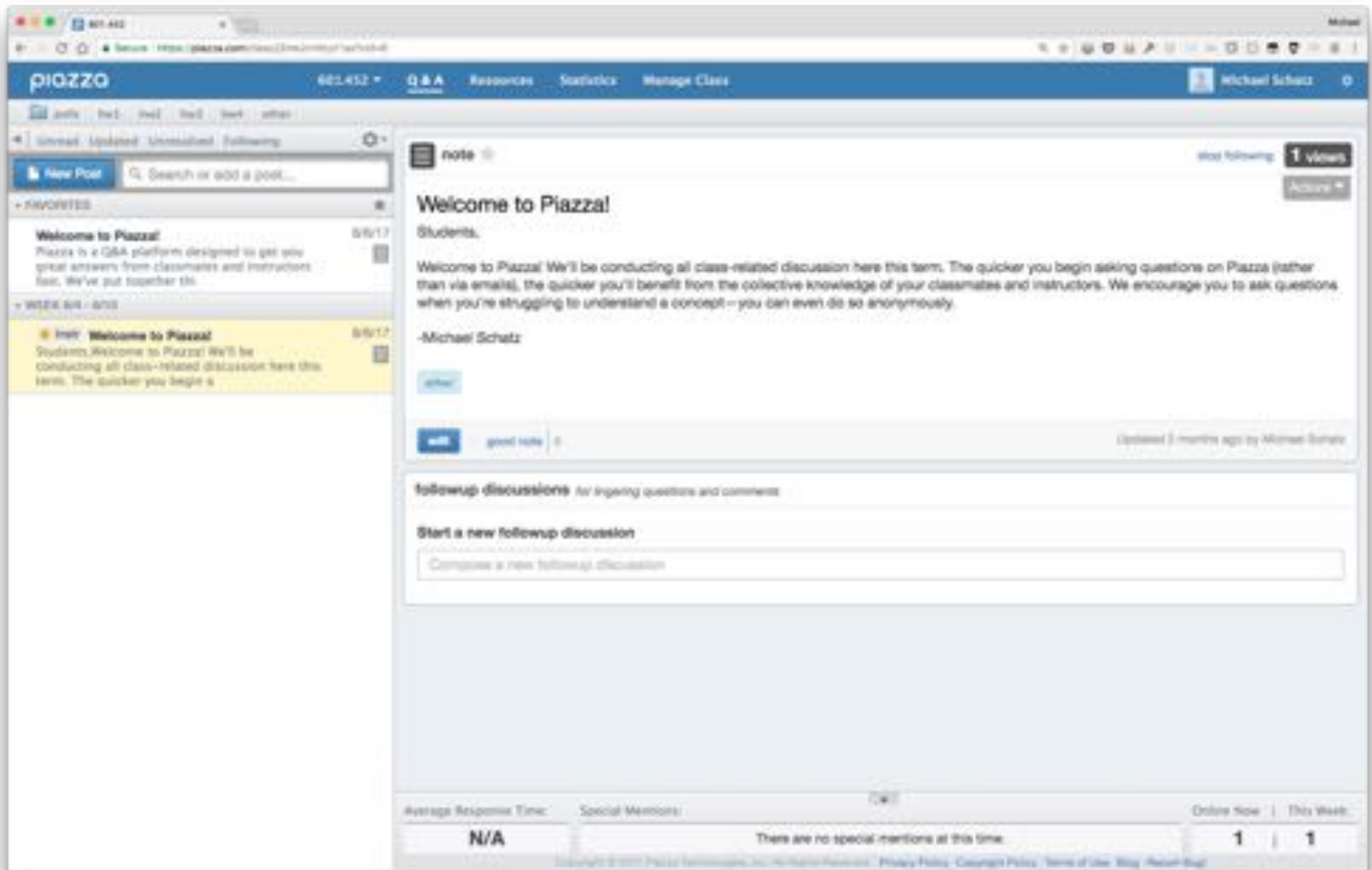
Related Courses & Readings:

- Applied Comparative Genomics
- Molecular Biology of the Gene (Watson et al)
- Bioinformatics Algorithms: An Active Learning Approach (Compeau and Pevzner)
- Unix and Perl for Biologists (Bioinformatics)

Preliminary Schedule

<https://github.com/schatzlab/biomedicalresearch>

Piazza



<http://piazza.com/jhu/fall2017/601452/home>

Sequencing Centers

A world map with blue pins indicating sequencing centers. Each pin has a number next to it, representing the number of sequencing centers in that region. The numbers are: North America (38, 13), Europe (11, 7, 8, 16, 20, 28, 73), Africa (22, 13, 19, 4, 8), Asia (16, 20, 28, 73), Australia (37, 19, 20, 4, 4), and South America (13, 19, 4, 8).

Worldwide capacity exceeds 50 Pbp/year
Over 500k human genomes sequenced

On track to exceed over 1 Zbp / year by 2024

A zetta-what?

Next Generation Genomics: World Map of High-throughput Sequencers

<http://omicsmaps.com>

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

Unsolved Questions in Biology

- What is your genome sequence?

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

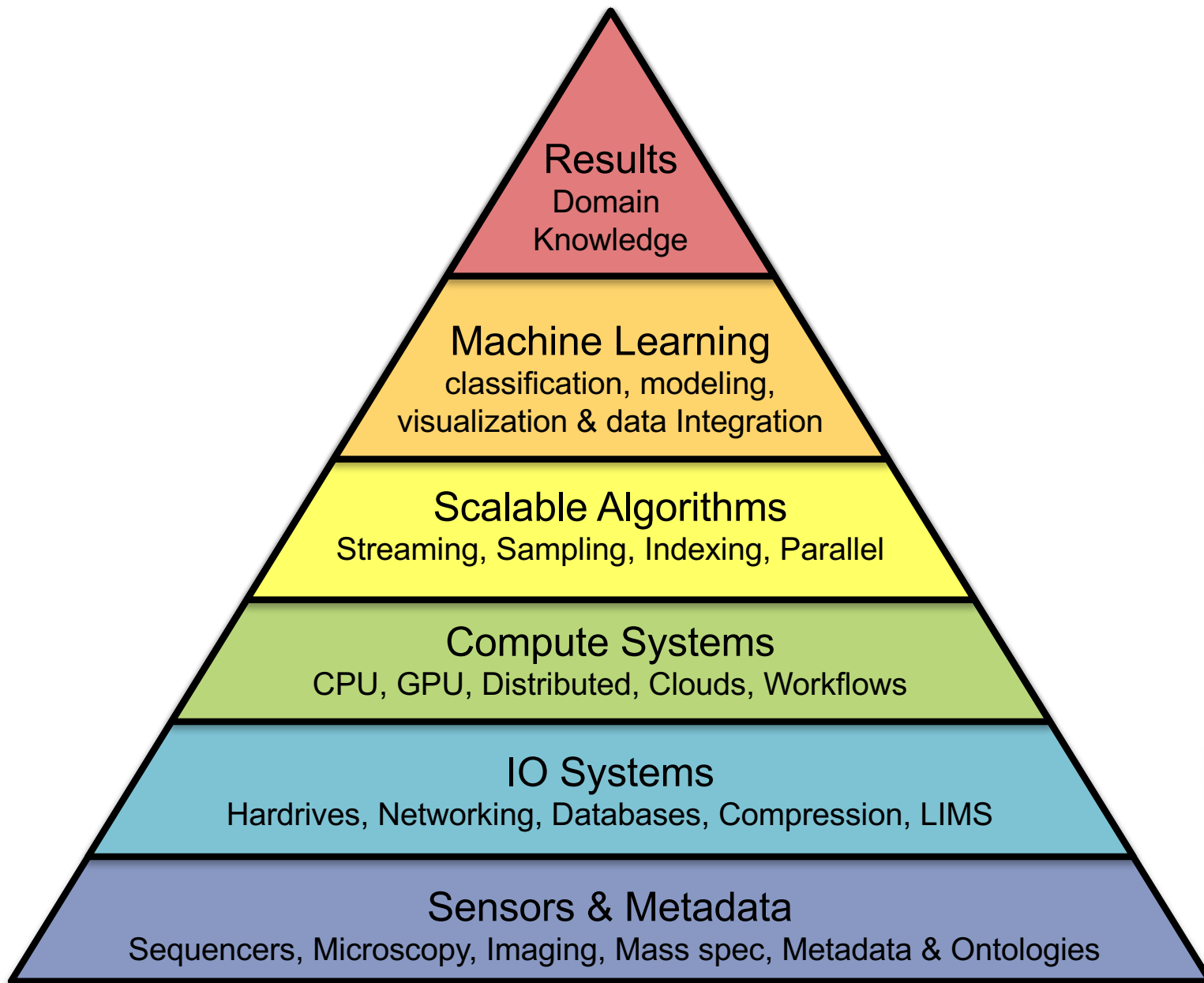
What software and systems will?

And who will create them?

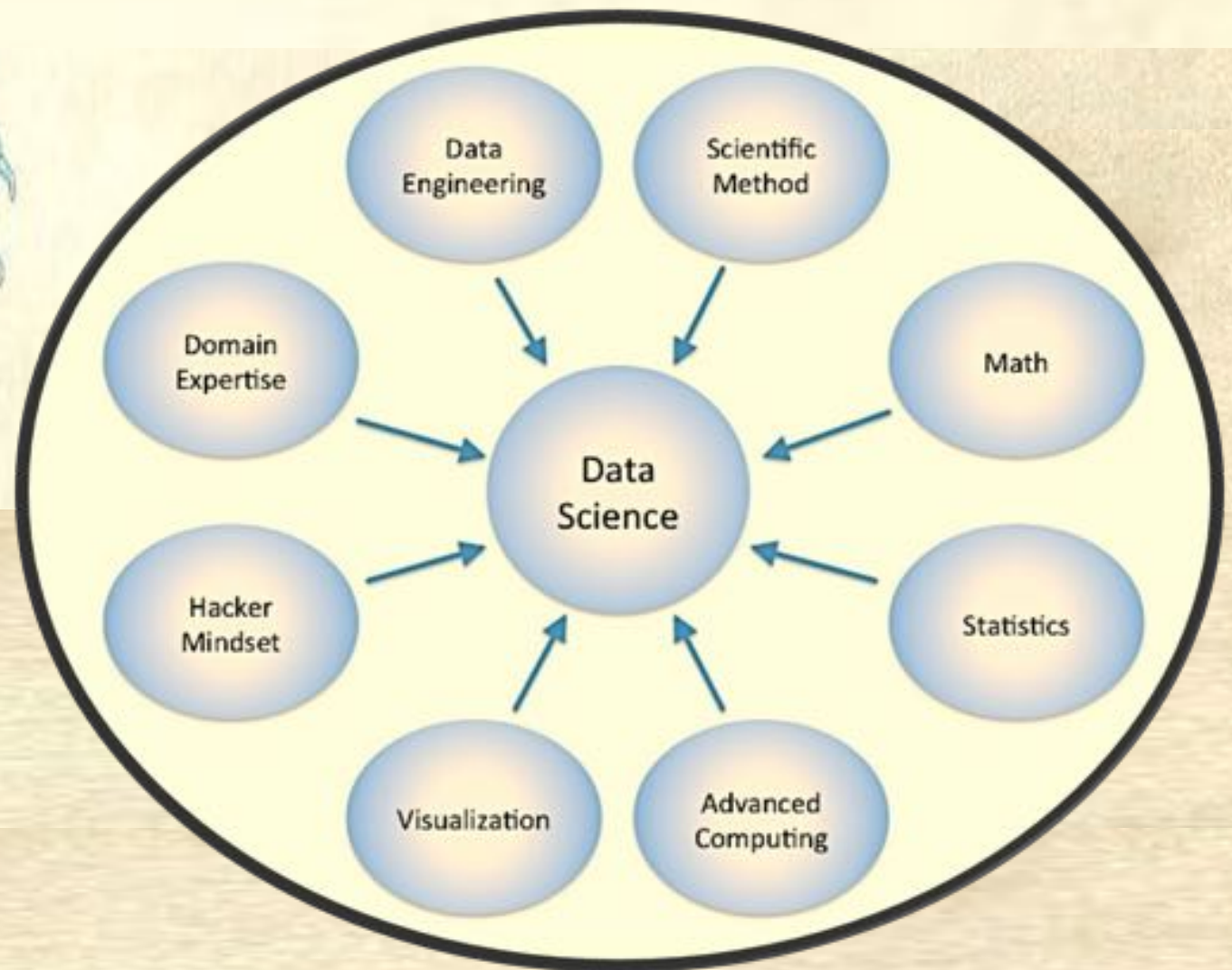
- ***Plus thousands and thousands more***



Biological Data Science Technologies



Who is a Data Scientist?



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



Genetic Basis of Autism Spectrum Disorders



Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

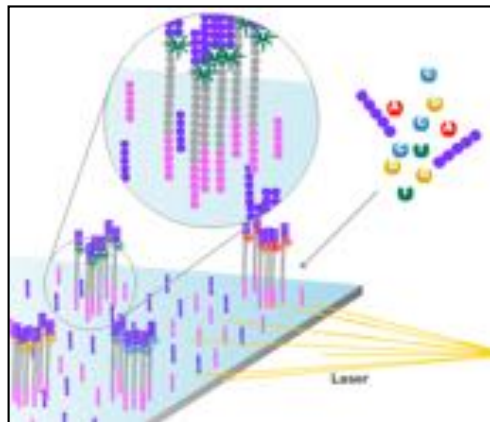
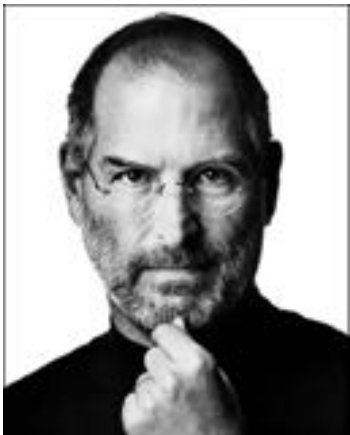
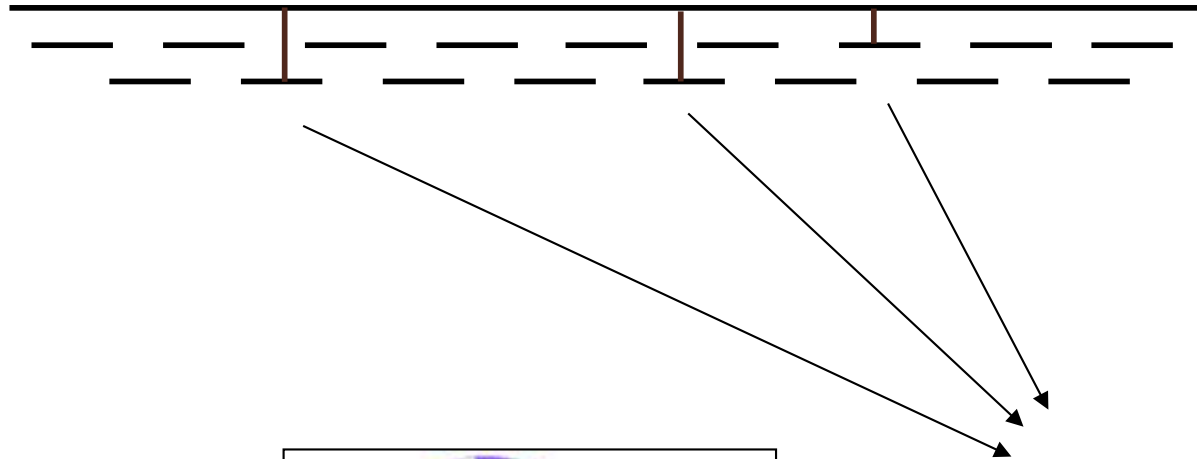
- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?

<http://www.autismspeaks.org/what-autism>

Personal Genomics

How does your genome compare to the reference?



Heart Disease

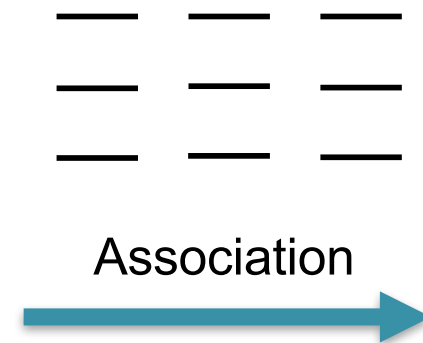
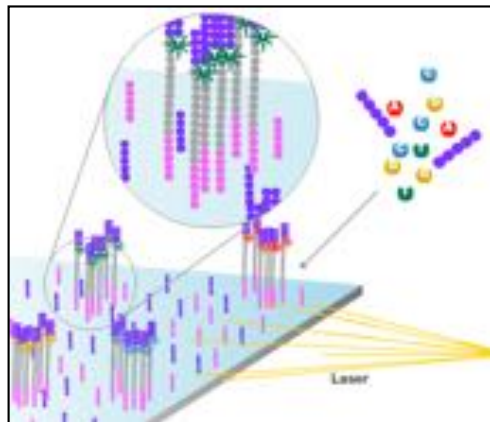
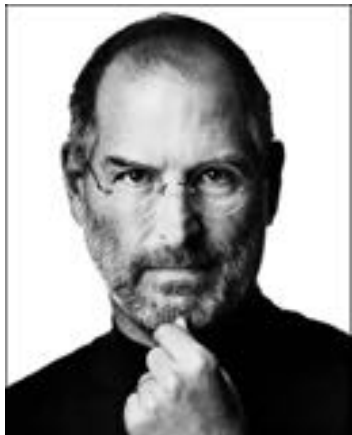
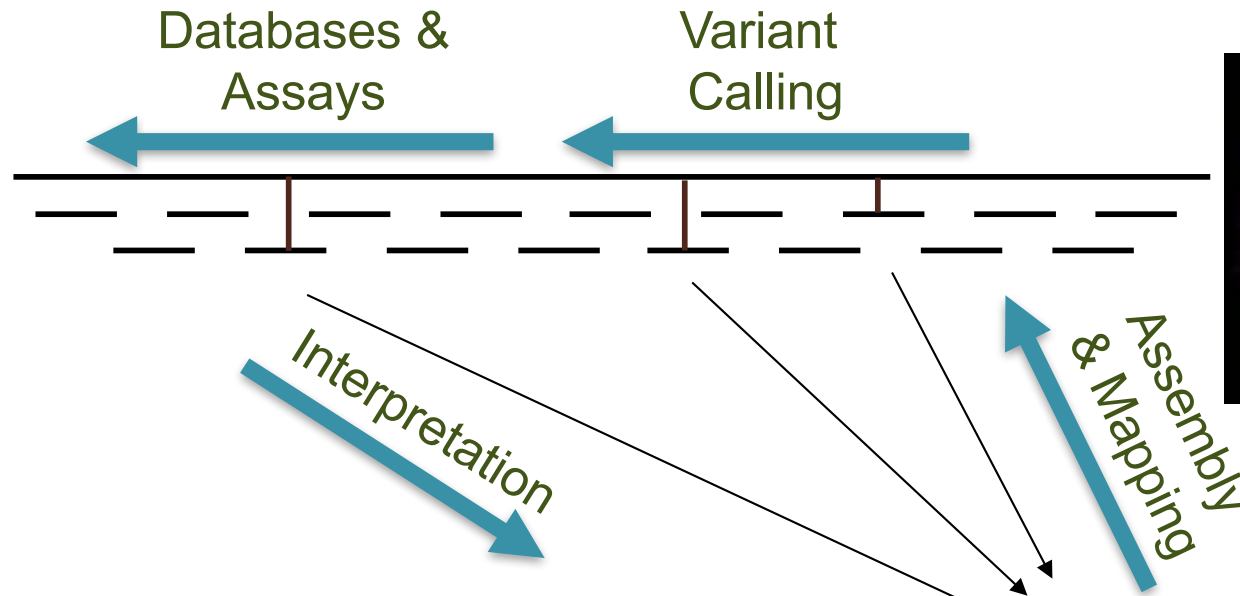
Cancer

Autism?

_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

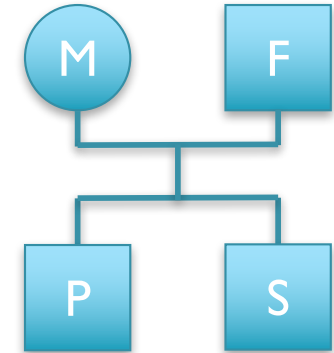
Personal Genomics

How does your genome compare to the reference?



De novo mutation discovery and validation

Concept: Identify mutations not present in parents.



Challenge: Sequencing errors in the child or low coverage in parents lead to false positive de novos

Reference: ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father: ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Mother: ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

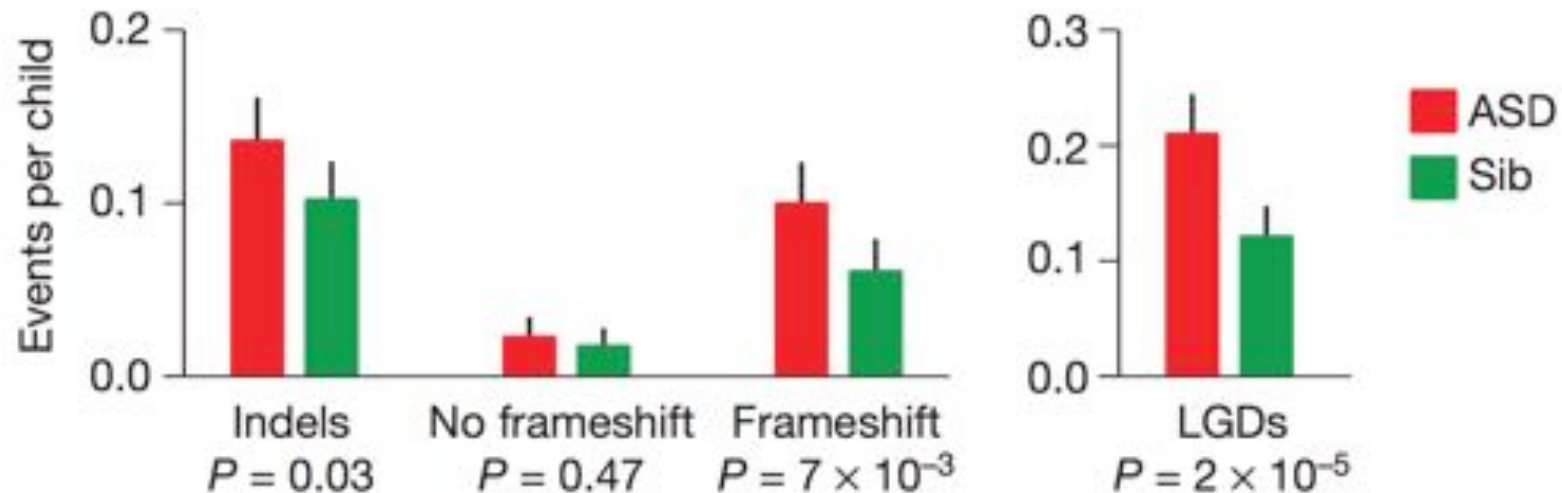
Sibling: ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Proband(1): ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Proband(2): ...TCAAATCCTTTTAAT****AAGAGCTGACA...

4bp heterozygous deletion at chr15:93524061 CHD2

De novo Genetics of Autism

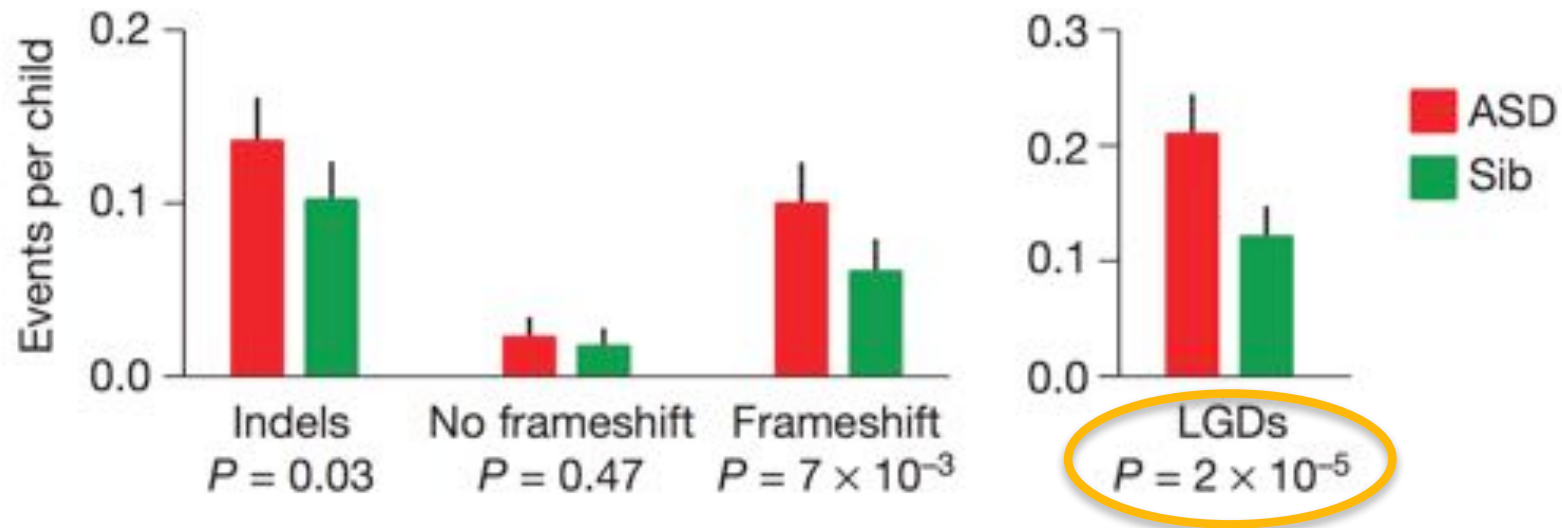


- In 2,500 family quads we see significant enrichment in de novo **likely gene disruptions (LGDs)** in the autistic kids
 - Overall rate basically 1:1
 - 2:1 enrichment in frameshift indels
 - Contributed dozens of new autism candidate genes, many associated with neuron development or chromatin formation

The burden of de novo coding mutations in autism spectrum disorders.

lossifov et al (2014) *Nature*. doi:10.1038/nature13908

De novo Genetics of Autism

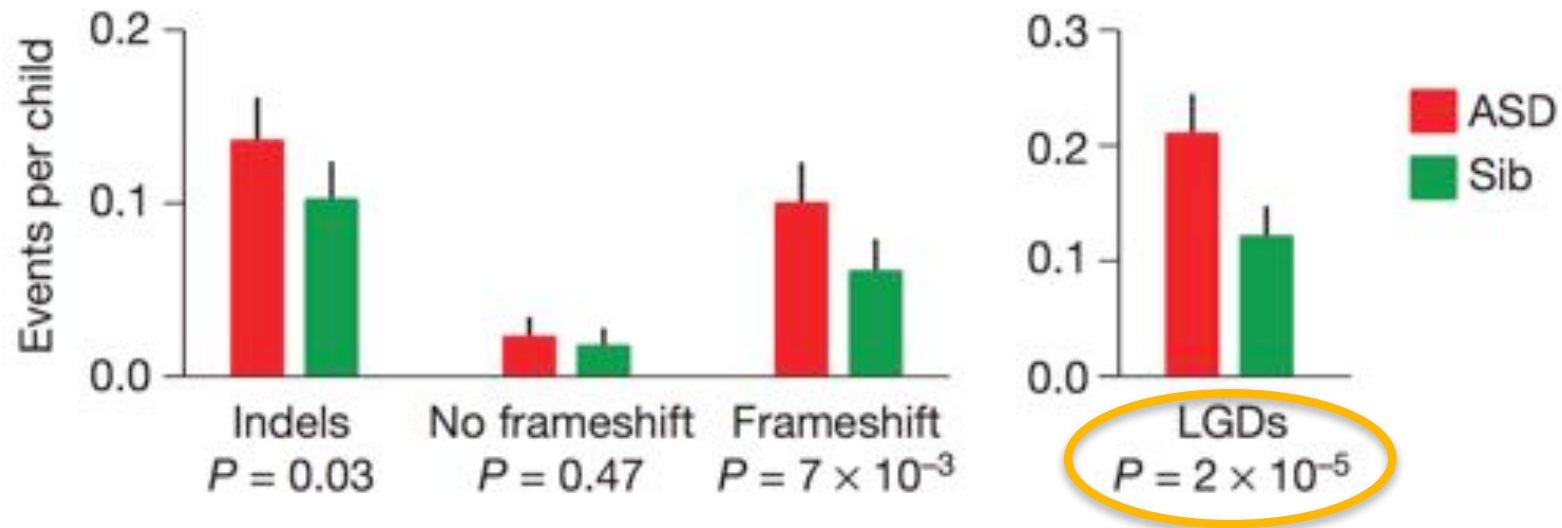


- In 2,500 family quads we see significant enrichment in de novo **likely gene disruptions (LGDs)** in the autistic kids
 - Overall rate basically 1:1
 - 2:1 enrichment in frameshift indels
 - Contributed dozens of new autism candidate genes, many associated with neuron development or chromatin formation

The burden of de novo coding mutations in autism spectrum disorders.

lossifov et al (2014) *Nature*. doi:10.1038/nature13908

P-value



- The “p-value” is the probability of observing a difference with the same or larger magnitude as observed but completely by chance (under the null hypothesis)
 - Maybe kids with ASD genuinely have a larger number of gene disrupting mutations, or maybe we just got a slightly skewed sample?
- If I flip a coin 100 times, I expect 50 heads and 50 tails, but I'm not surprised if I get 49 heads and 51 tails.
 - On the other hand I'm extremely surprised if I get 1 heads and 99 tails!
 - What about 25 heads? or 15 heads? Or 5?
 - I'm more surprised when the probability is smaller and smaller (exponent is more and more negative)