Programming and p-values

Michael Schatz

Sept 6, 2017 – Lecture 2
EN.601.452 Computational Biomedical Research
AS.020.415 Advanced Biomedical Research

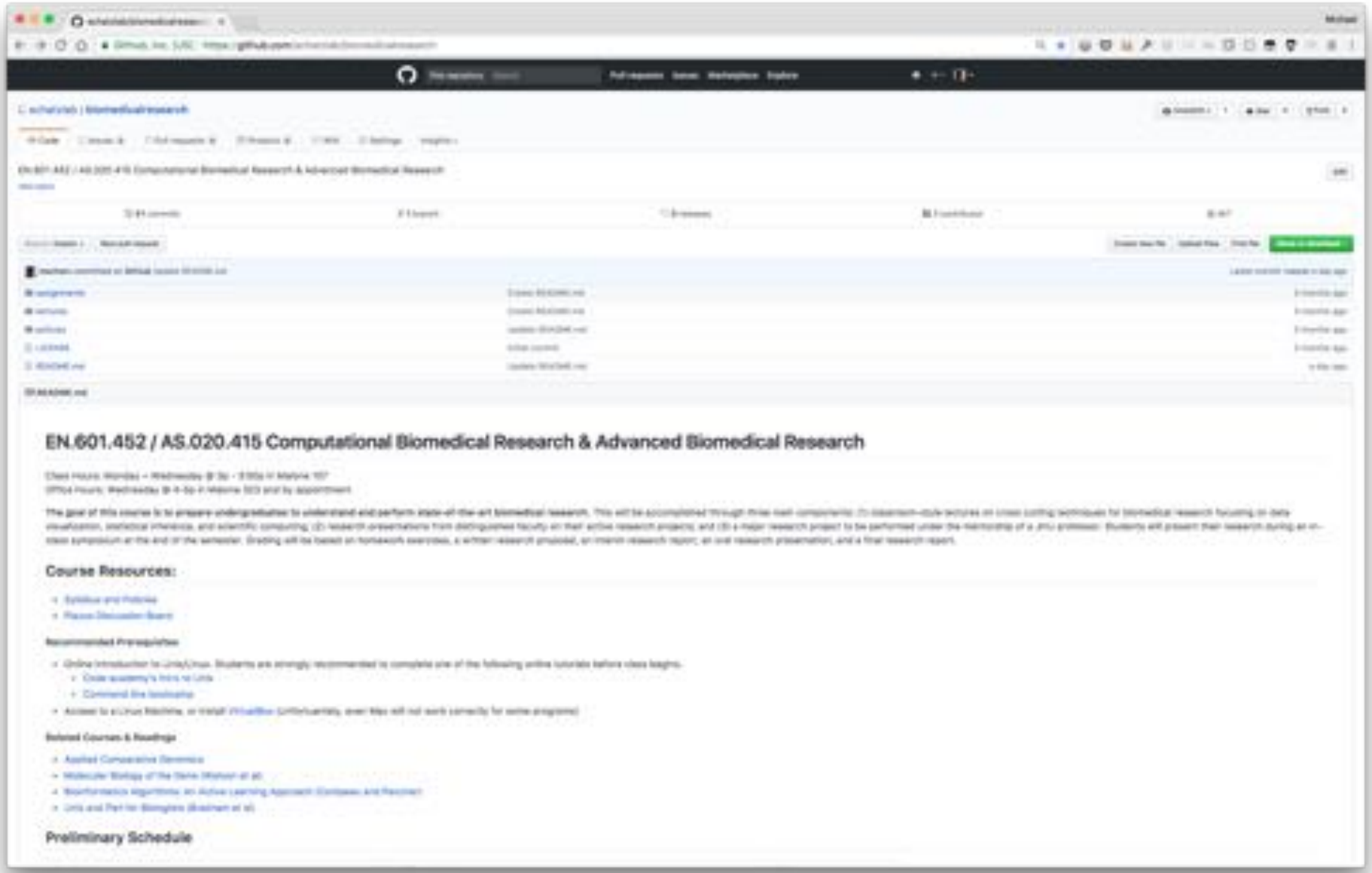
Welcome!

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.

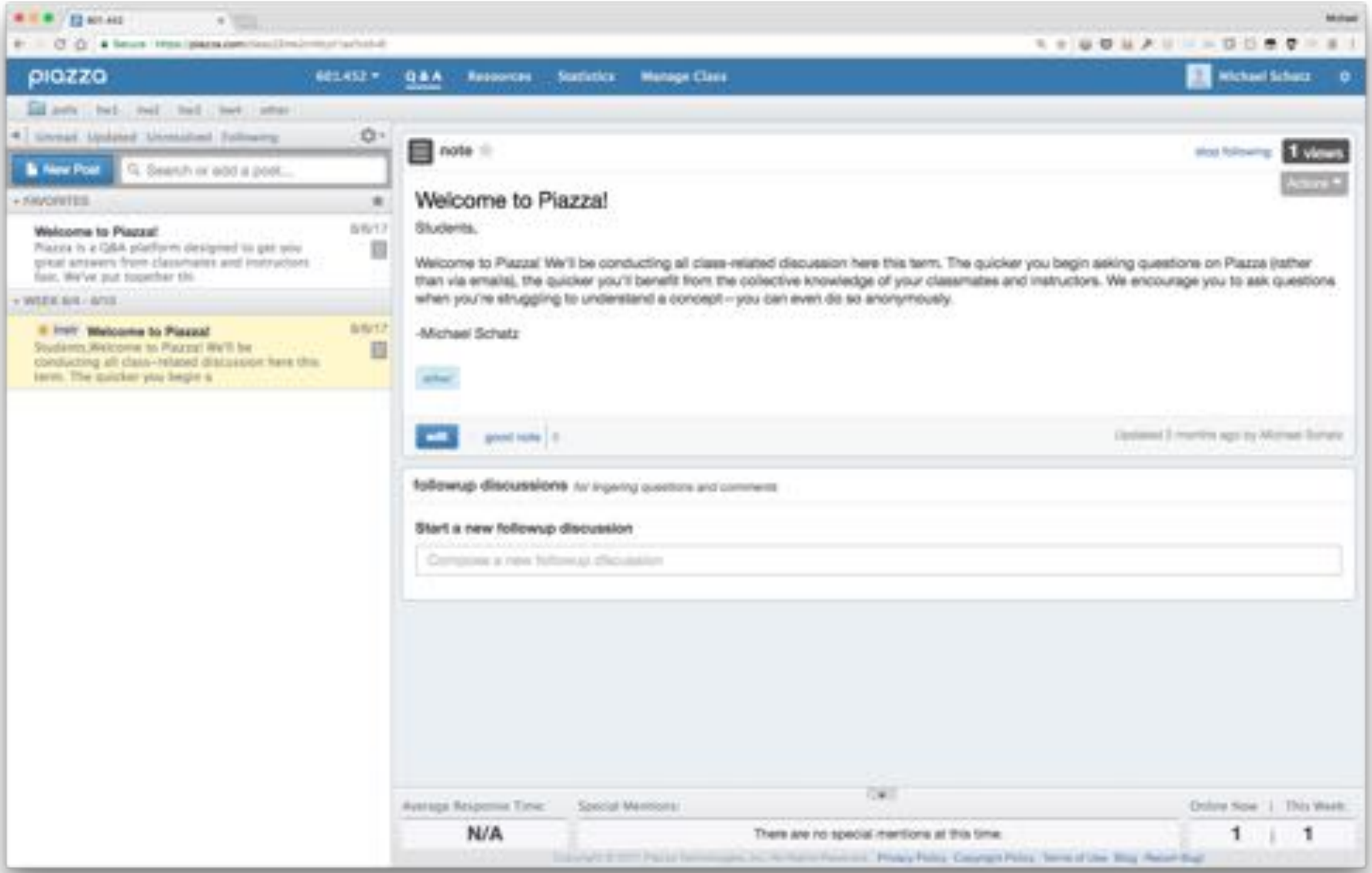
Course Webpage:	https://github.com/schatzlab/biomedicalresearch
Course Discussions:	http://piazza.com
Class Hours:	Mon + Wed @ 3p – 3:50p Malone 107
Office Hours:	Wed @ 4-5p and by appointment Please try Piazza first!

Course Webpage



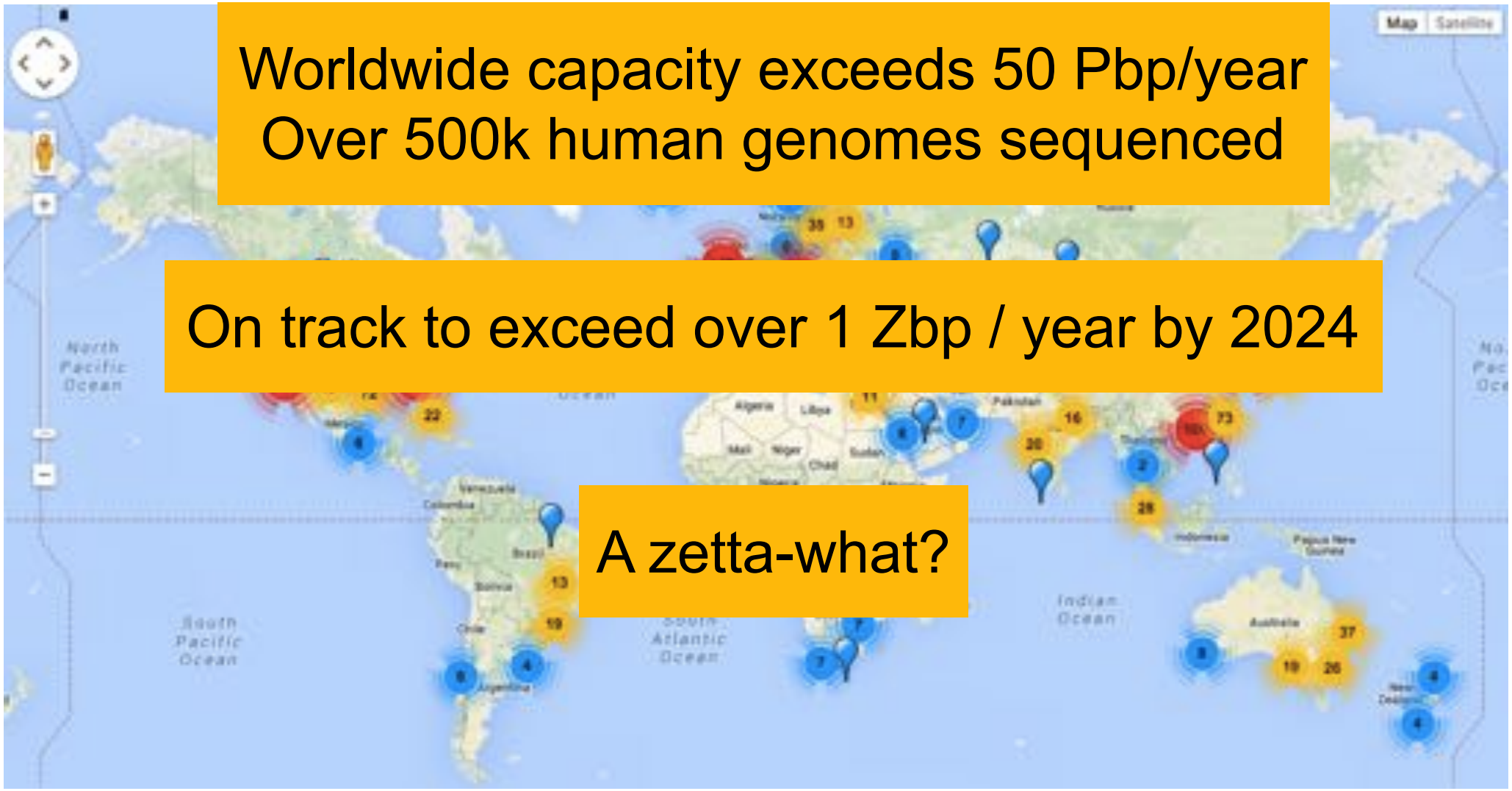
https://github.com/schatzlab/biomedicalresearch

Piazza



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

Sequencing Centers



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

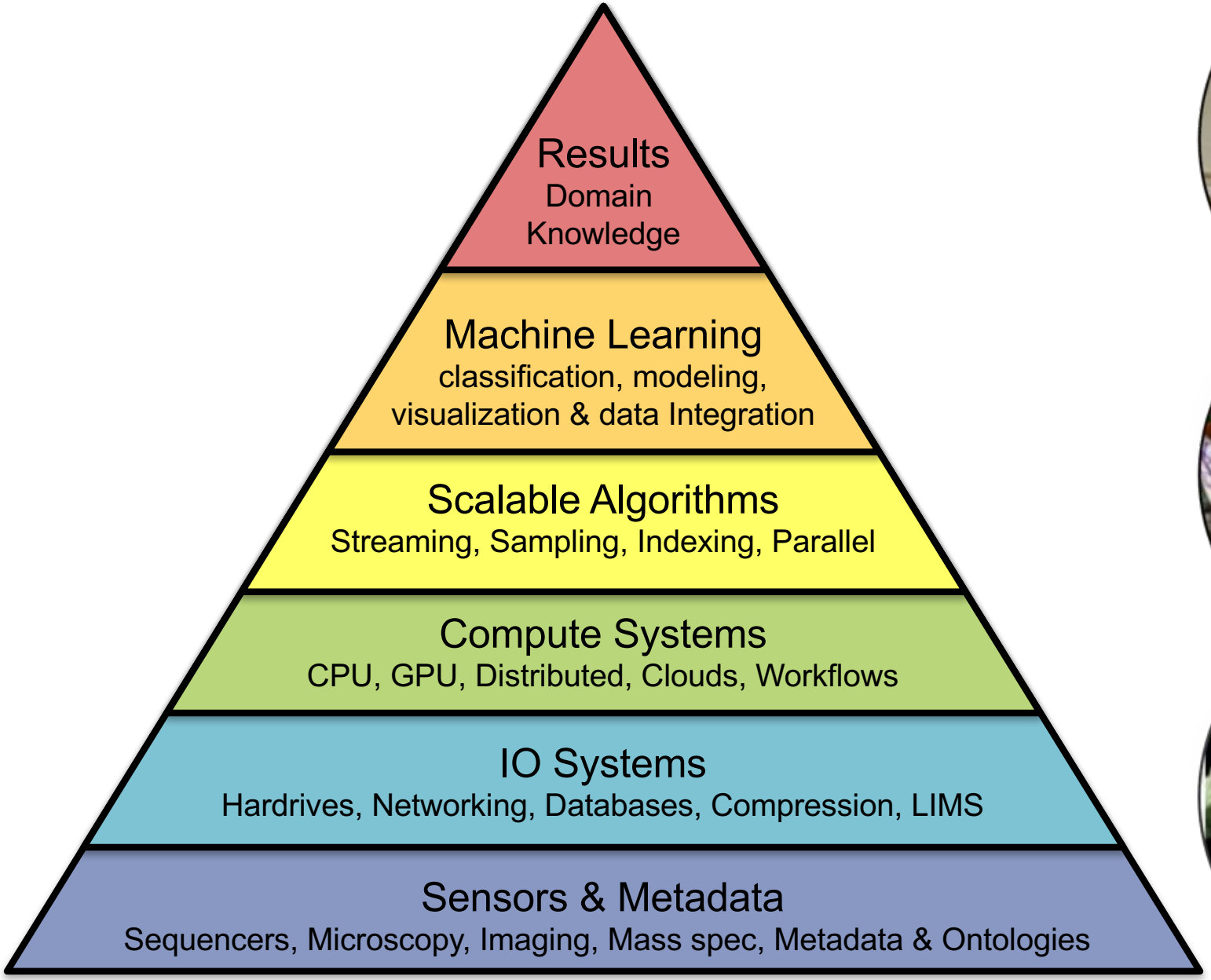
- What is your genome sequence?
-
-
-
-
-
-
-
-
-
-
-
- ***Plus thousands and thousands more***

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

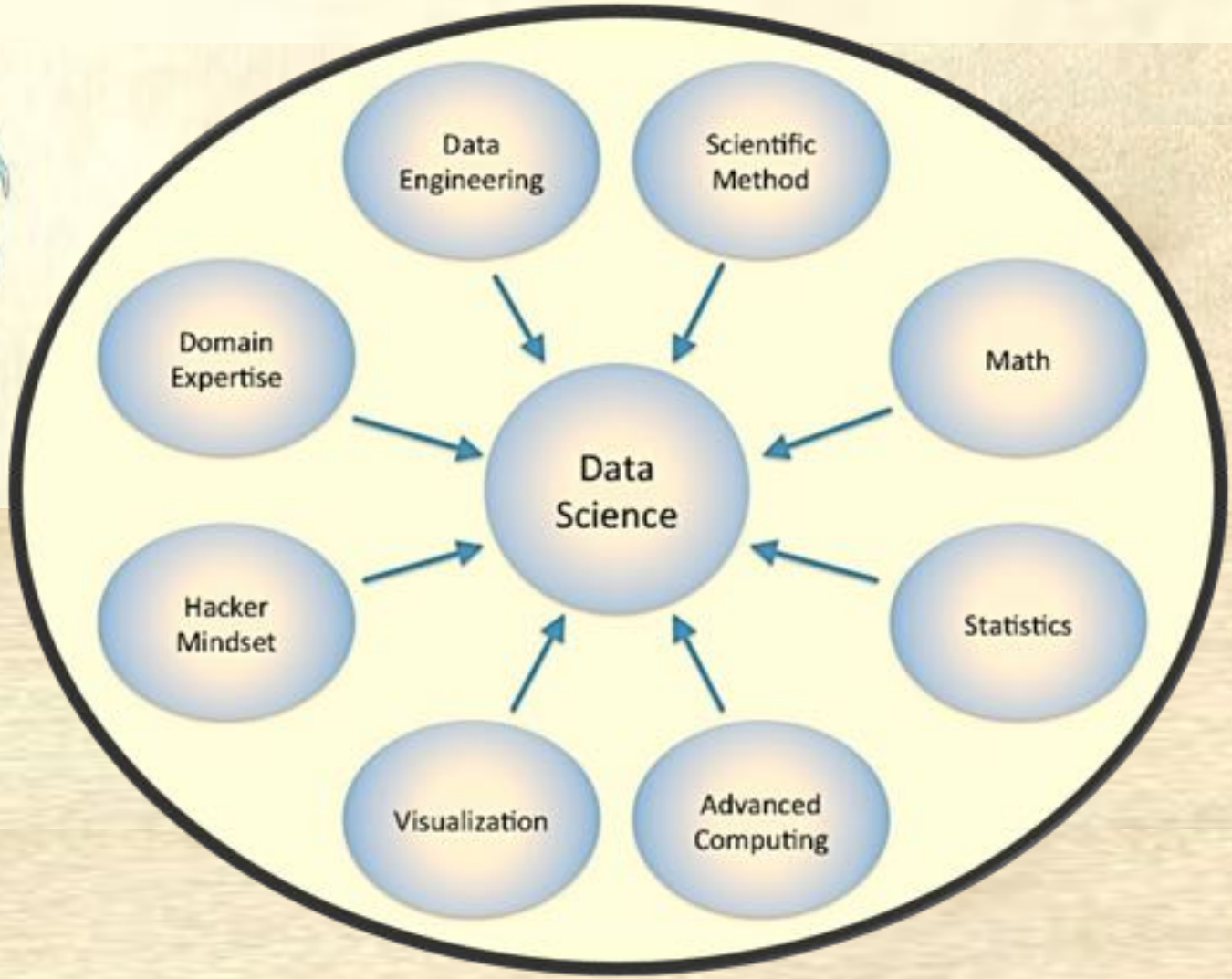
What software and systems will?

And who will create them?

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.

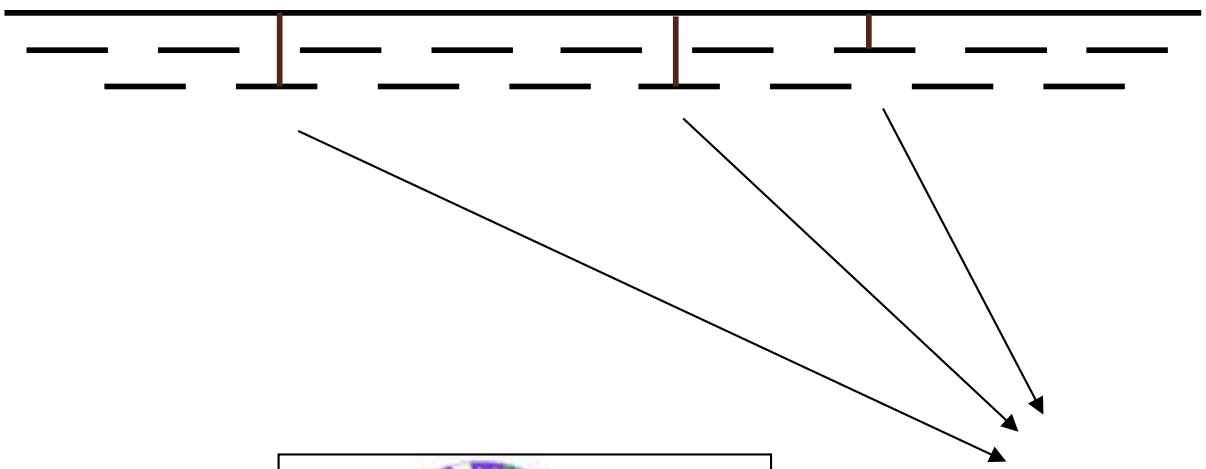
U.S. CDC identify around 1 in 68 American children as on the autism spectrum

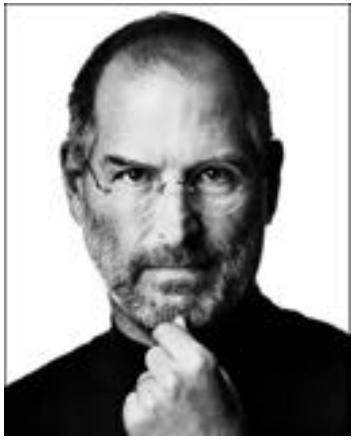
- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

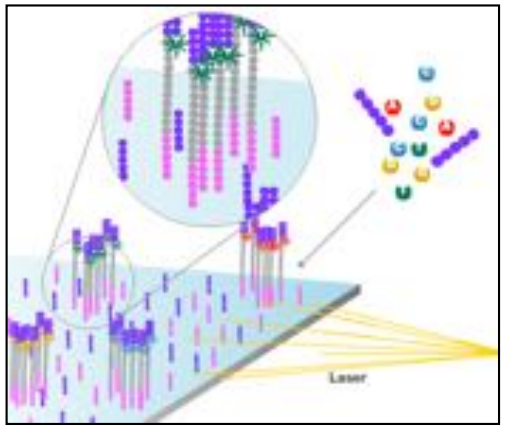
What is Autism?
http://www.autismspeaks.org/what-autism

Personal Genomics

How does your genome compare to the reference?







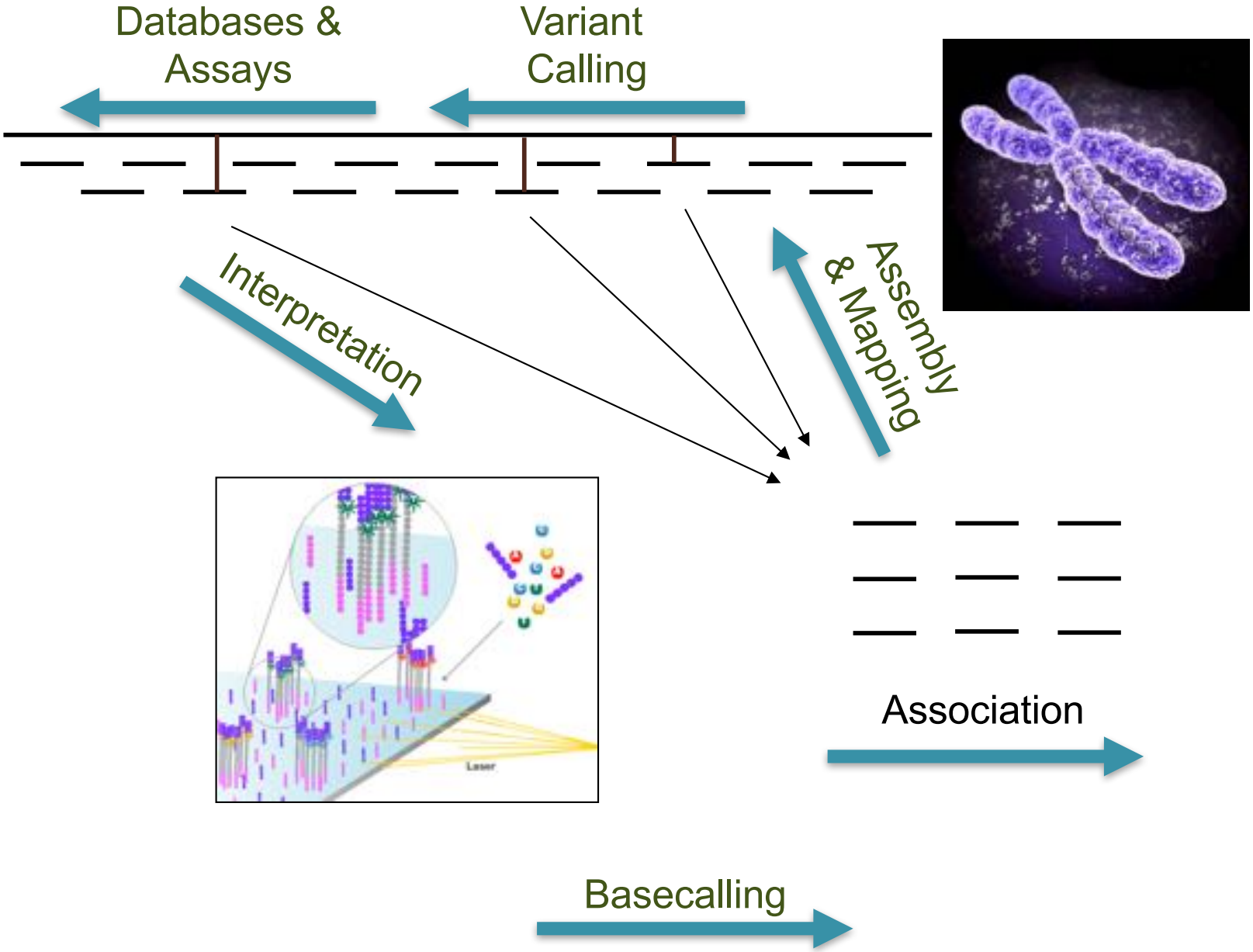
Heart Disease

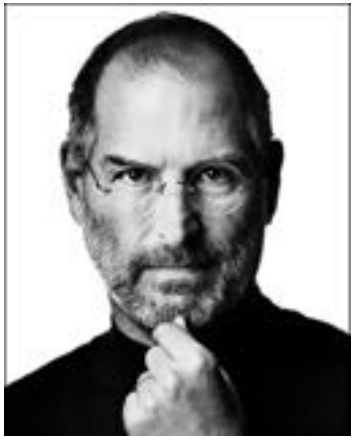
Cancer

Autism?

Personal Genomics

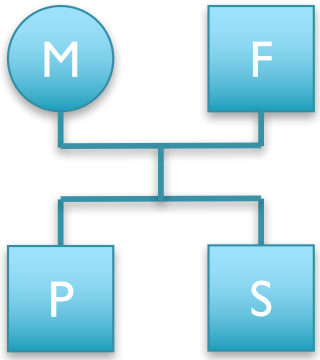
How does your genome compare to the reference?





De novo mutation discovery and validation

Concept: Identify mutations not present in parents.



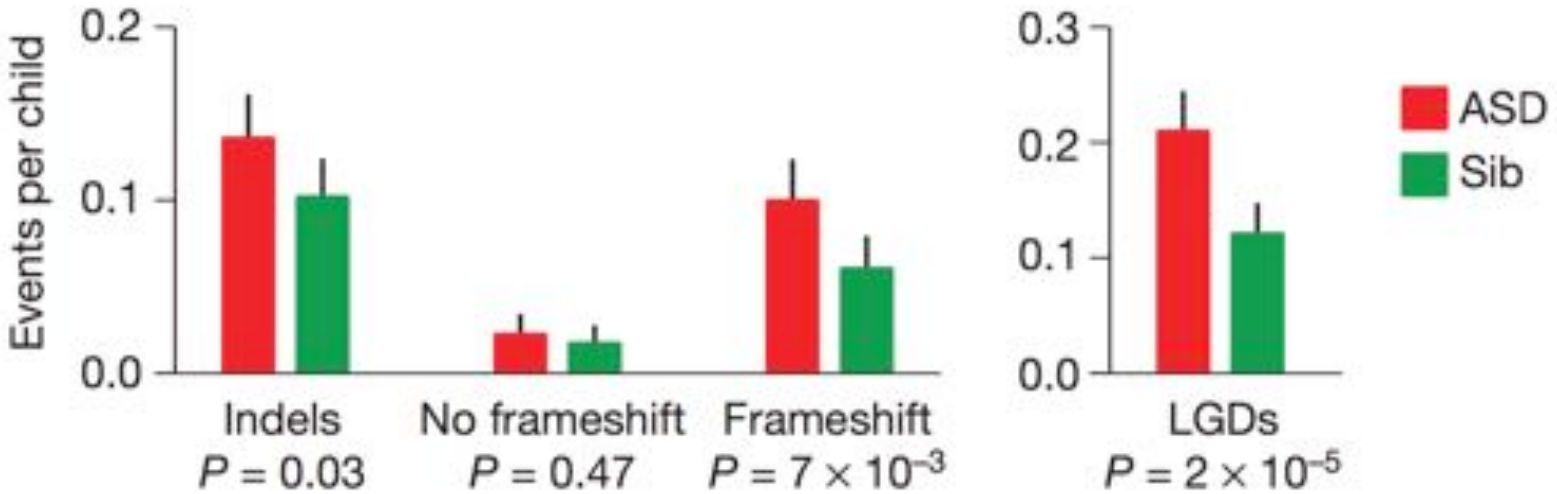
Challenge: Sequencing errors in the child or low coverage in parents lead to false positive de novos

```
Reference:   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling:     ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(1):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):  ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

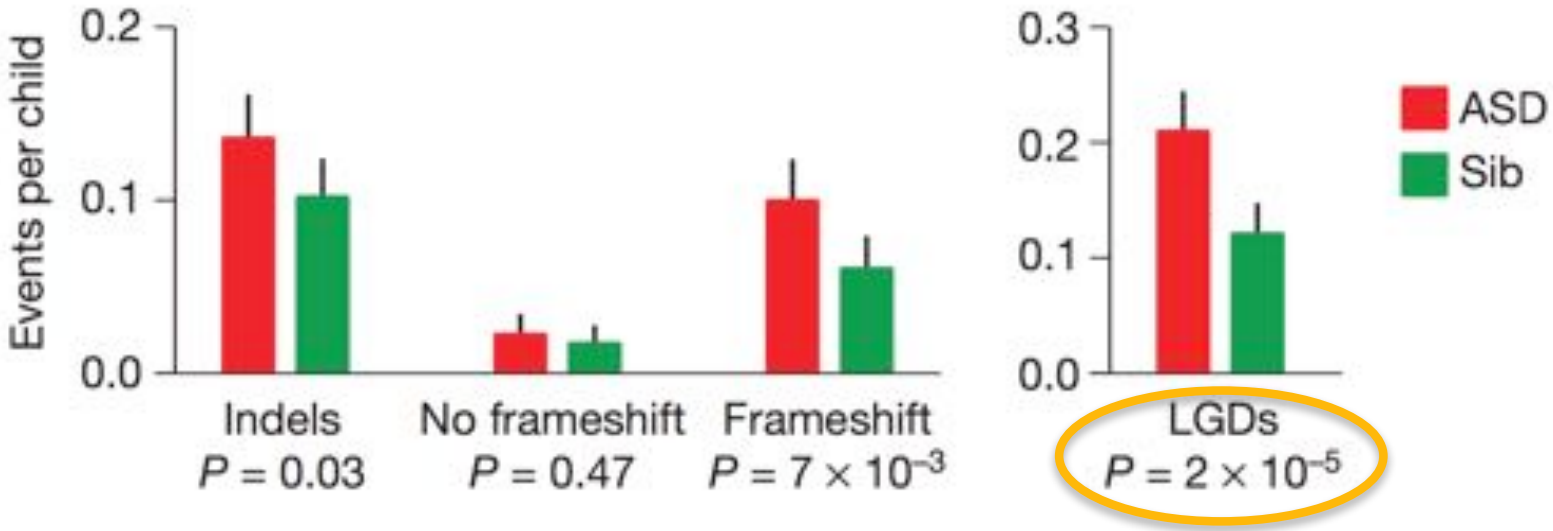
De novo Genetics of Autism



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

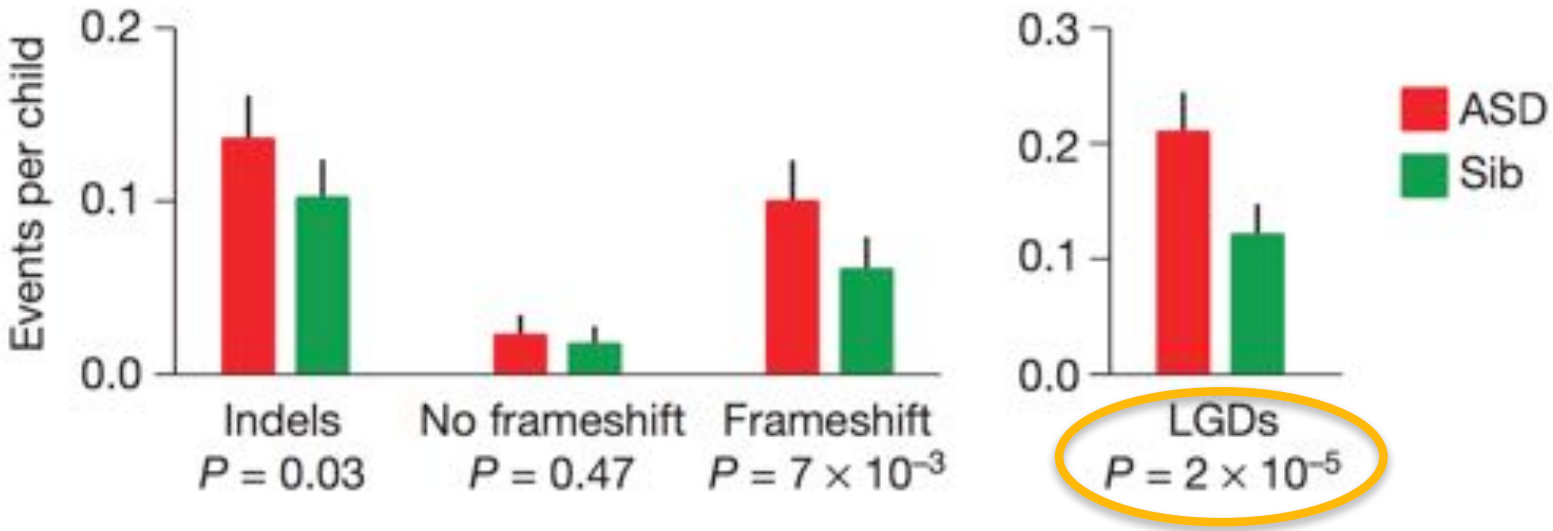
De novo Genetics of Autism



- 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.
Iossifov *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 Im not surprised if I get 49 heads and 51 tails.
  - On the other hand Im extremely surprised if I get 1 heads and 99 tails!
  - What about 25 heads? or 15 heads? Or 5?
  - Im more surprised when the probability is smaller and smaller (exponent is more and more negative)