Modern Research Requires Collaboration:

Incorporating Statistics and Bioinformatics

USDA, ARS, Northeast Area

- Statisticians
  - Bryan Vinyard
  - Mary Camp
  - Matt Kramer
- Bioinformaticists
  - Jonathan Shao
  - Nadia Darwish

ARS-NEA-StatGroup@ars.usda.gov

Art is 'I', Science is 'We' *–Claude Bernard*

- Modern scientific work is not solitary
- Collaboration is essential to successful research
- Collaboration is multidisciplinary

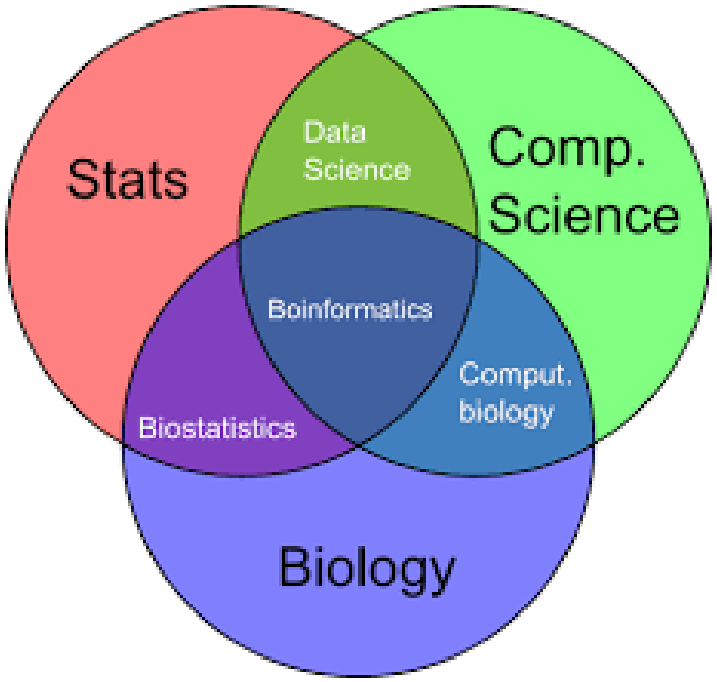
Data in the 21st Century

- Amounts of data being collected has expanded, greatly
- Information in Data has greater potential /ability to influence direction of research investigations, than ever before
- Essential need to organize and interpret the data
  - Becomes more difficult with each passing year; as amount of data increases

"Buzz Words" – each has Various Definitions

- "Big Data"
- "Data Mining"
- "Machine Learning"
- "Artificial Intelligence"
- "Data Science" ⇨ a broad discipline for making sense of numbers

Data Science – A Big Tent



- None of the disciplines wholly contained under *Data Science*

Data Science

- Why are these disciplines useful?
- What do you need to know about them?
- Understanding allows you to know when you need to seek assistance from an expert

Data Exploration Goals:

Typically about Finding and/or Verifying Patterns

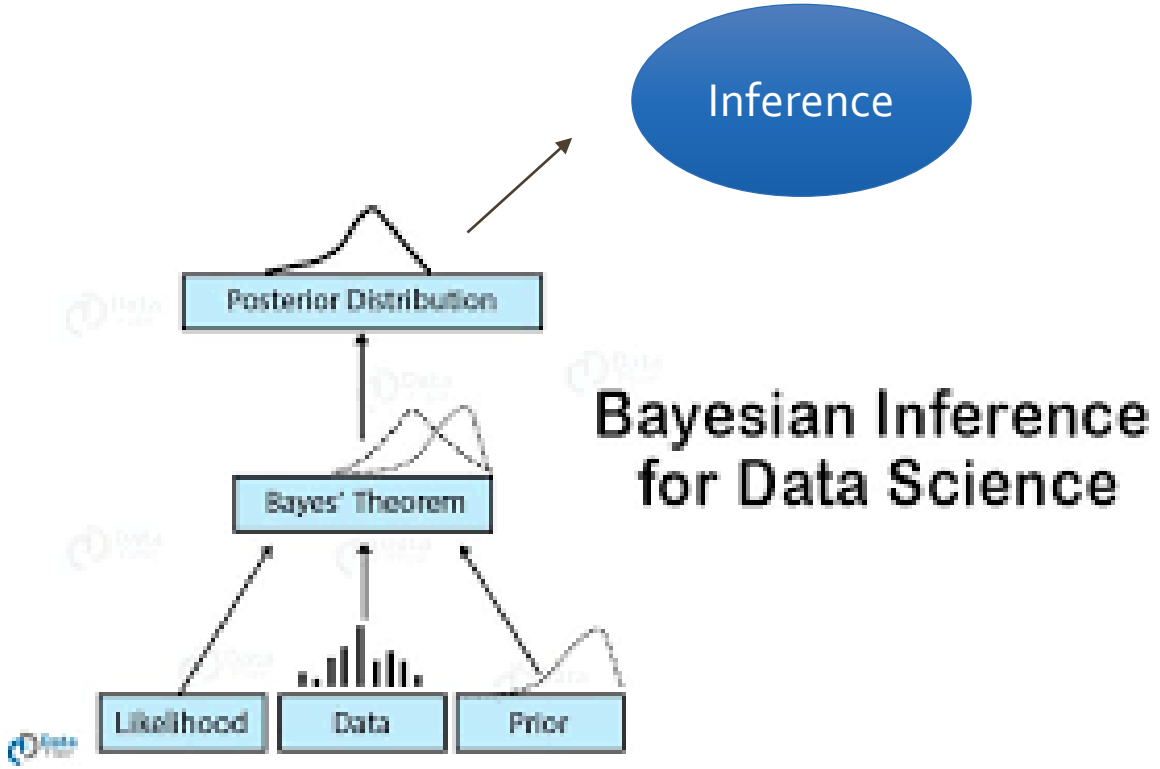
It is what we think we know already that prevents us from learning – *Claude Bernard*

- Graphics can often assist in finding the patterns in the data
- Algorithms are used to model observed patterns [Machine Learning]
- Predict future outcomes using model based on observed patterns
- Produce a graphical representation of the pattern to communicate and interpret its associations and causes

Statistics – A Crucial Part of Research

- Statistics is one of the ***oldest*** of the Data Science components
- Thomas Bayes, Karl Friedrich Gauss, Gregor Mendel were major contributors to the field pre-20th century
- Advent of Modern Statistics began in the 1920s under R.A. Fisher

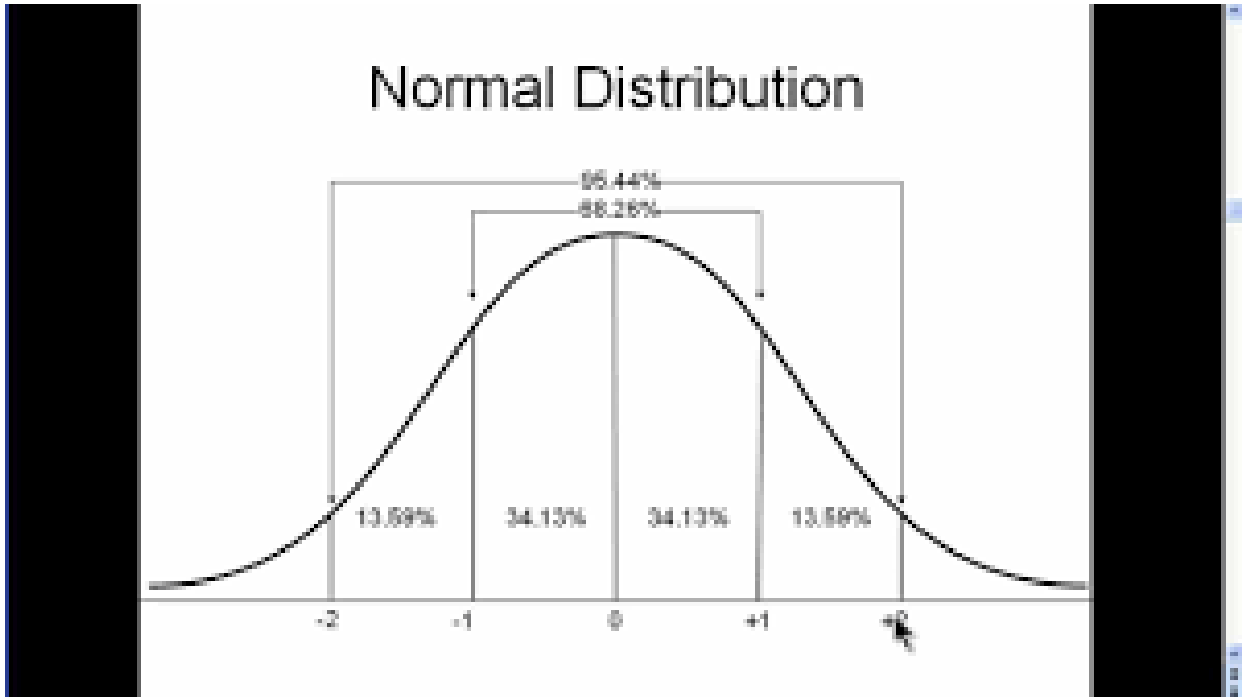
Bayesian Methods use "Prior" Assumptions



Early 18th Century

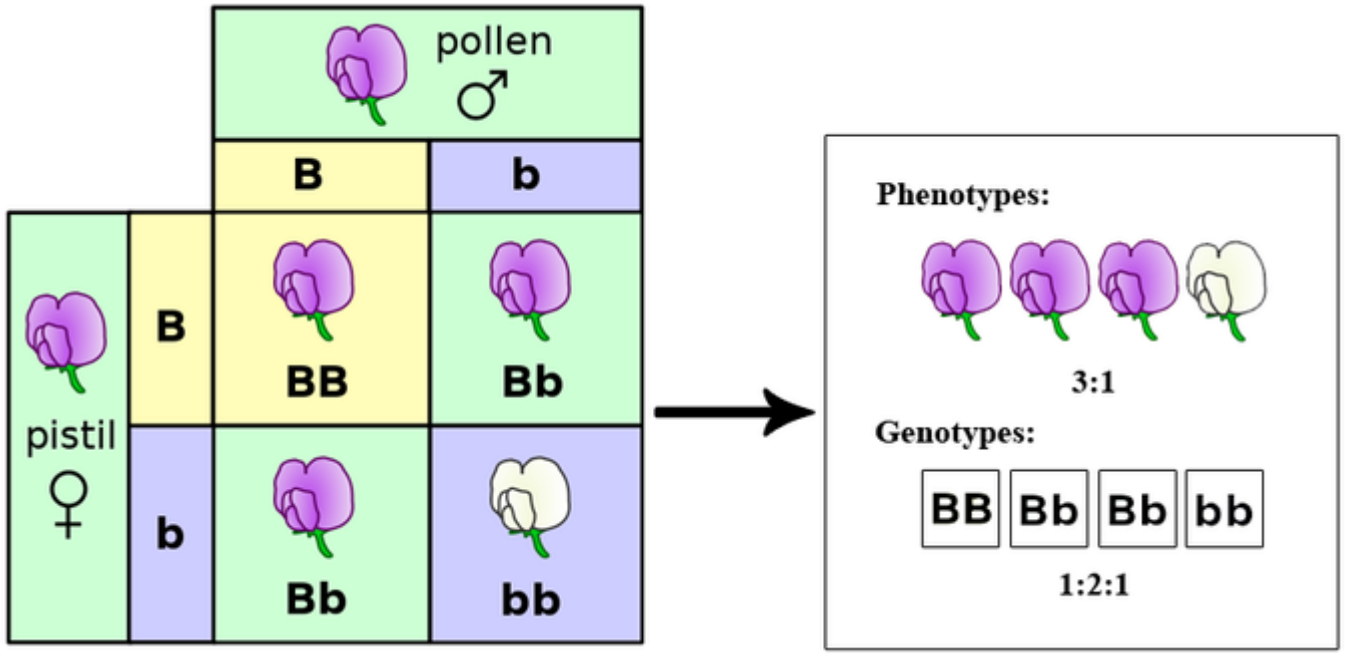
Gaussian "Normal" Distribution

Early 19th Century



Mendelian Genetic Probabilities

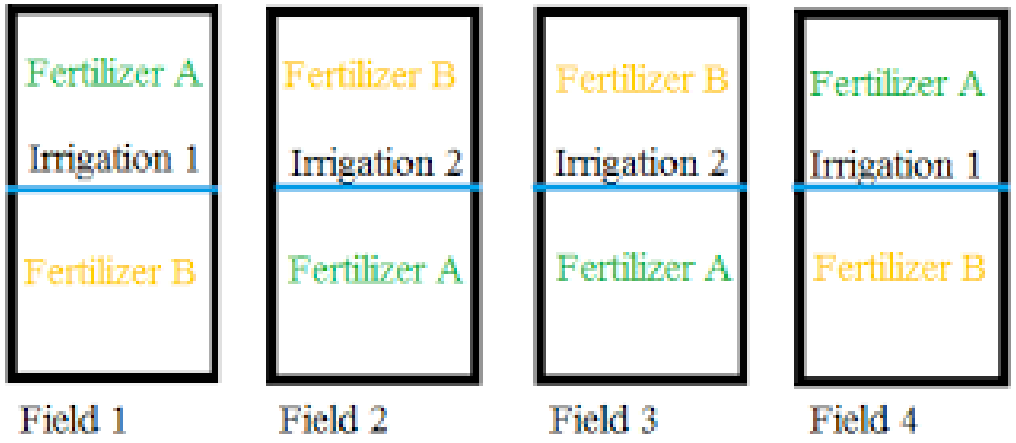
19th Century



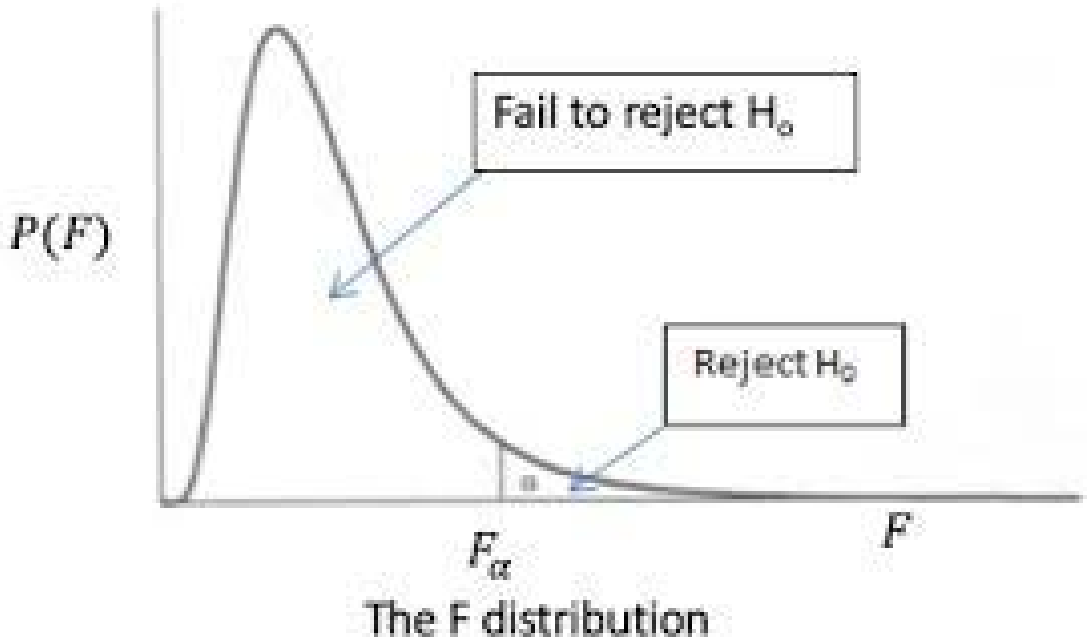
R.A. Fisher – Modern Statistical Methods

Early 20th Century

Experimental Design



ANOVA and Hypothesis Testing



The F distribution

Statistics Classes

They know enough who know how to learn – Henry Adams

- The statistics class(es) required for your major:
  - Will ***not*** provide you enough knowledge to work as a statistician
  - Will provide you enough knowledge to converse with a statistician

"Statistical Portion" of the Scientific Process

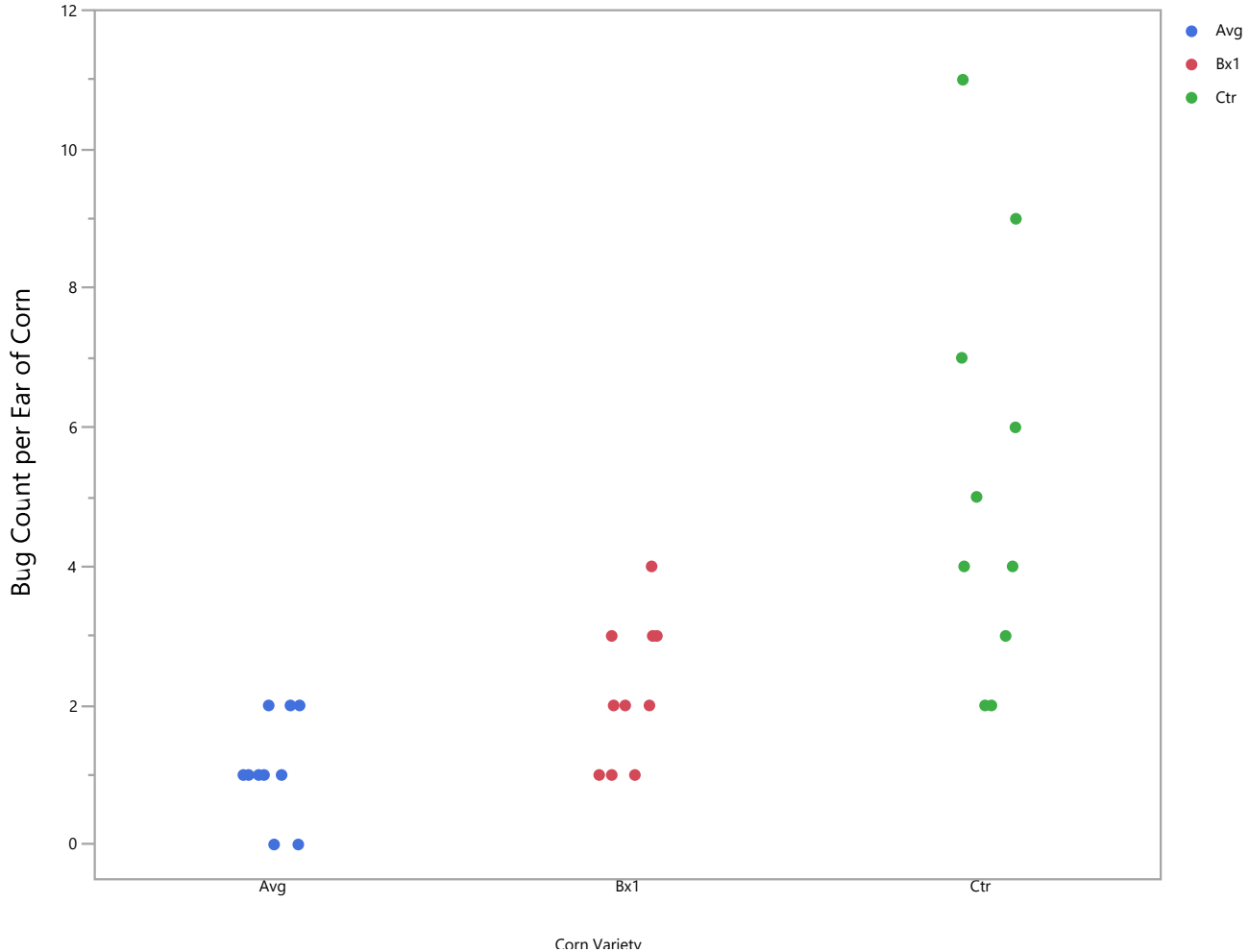
- Plan Experiments
- Assure Data Quality
- Structure the data for use in software
- Data analysis
  i. Graphical
  ii. Statistical Summaries
  iii. Statistical modeling and diagnostics
  iv. Formal statistical tests of comparisons
- Interpret analysis results
- Communicate results

Statistical Analysis: An Illustrative Example

Bacillus thuringiensis (BT) produces a toxin often used as a biological alternative to insecticide. One method of use is to genetically modify crops to include the BT gene.

- Three BT Corn Varieties are evaluated for resistance to tarnished plant bugs
- Corn varieties are randomly planted in a large field
- Ten plants per variety are randomly select at optimal size
- Resistance is measured by counting the number of tarnished plant bugs on the largest ear of corn on each selected plant

Tarnished Plant Bugs on BT Modified Corn

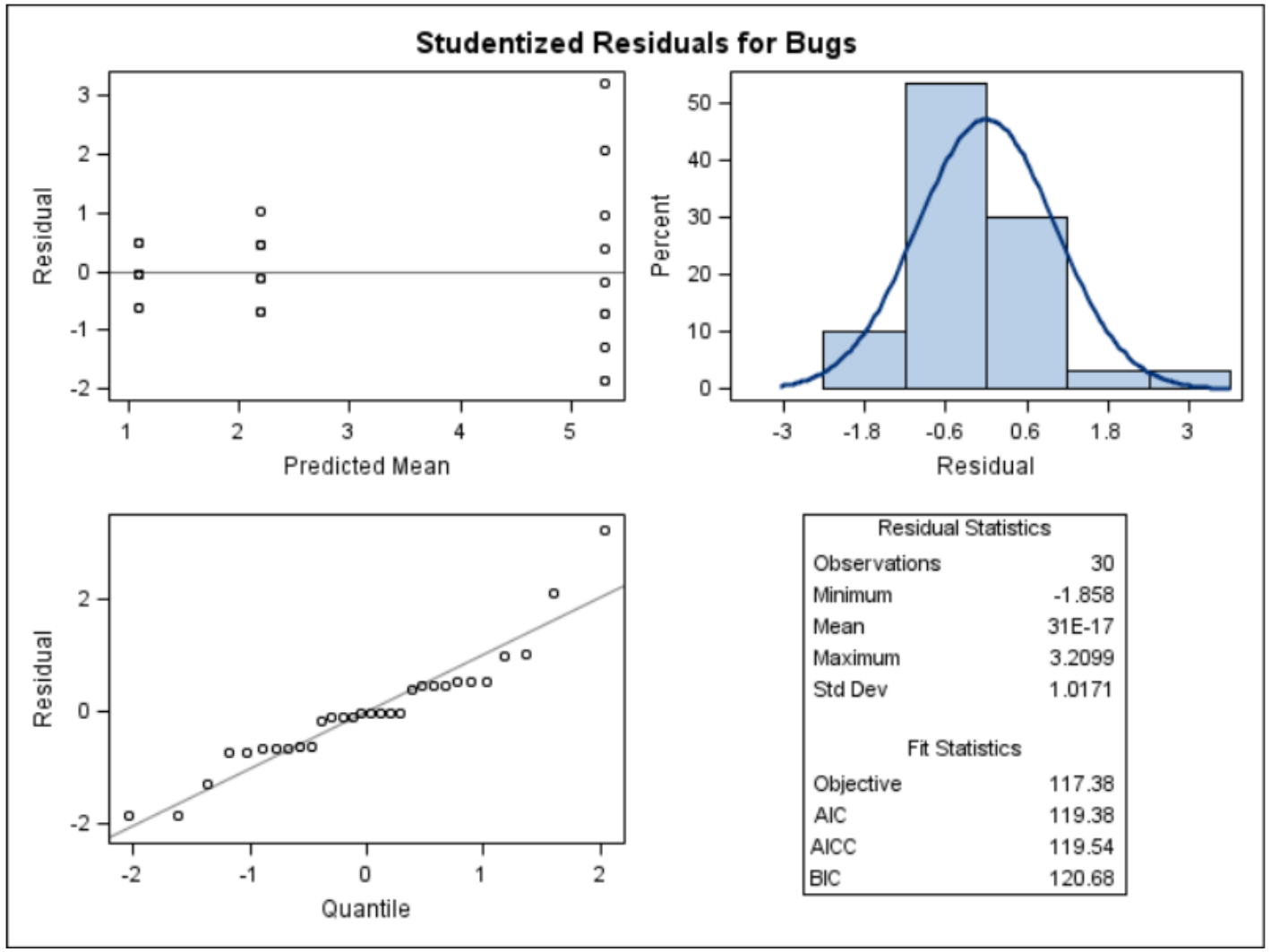


Step 1: Analyze the data using an One-Way "Corn Variety" ANOVA

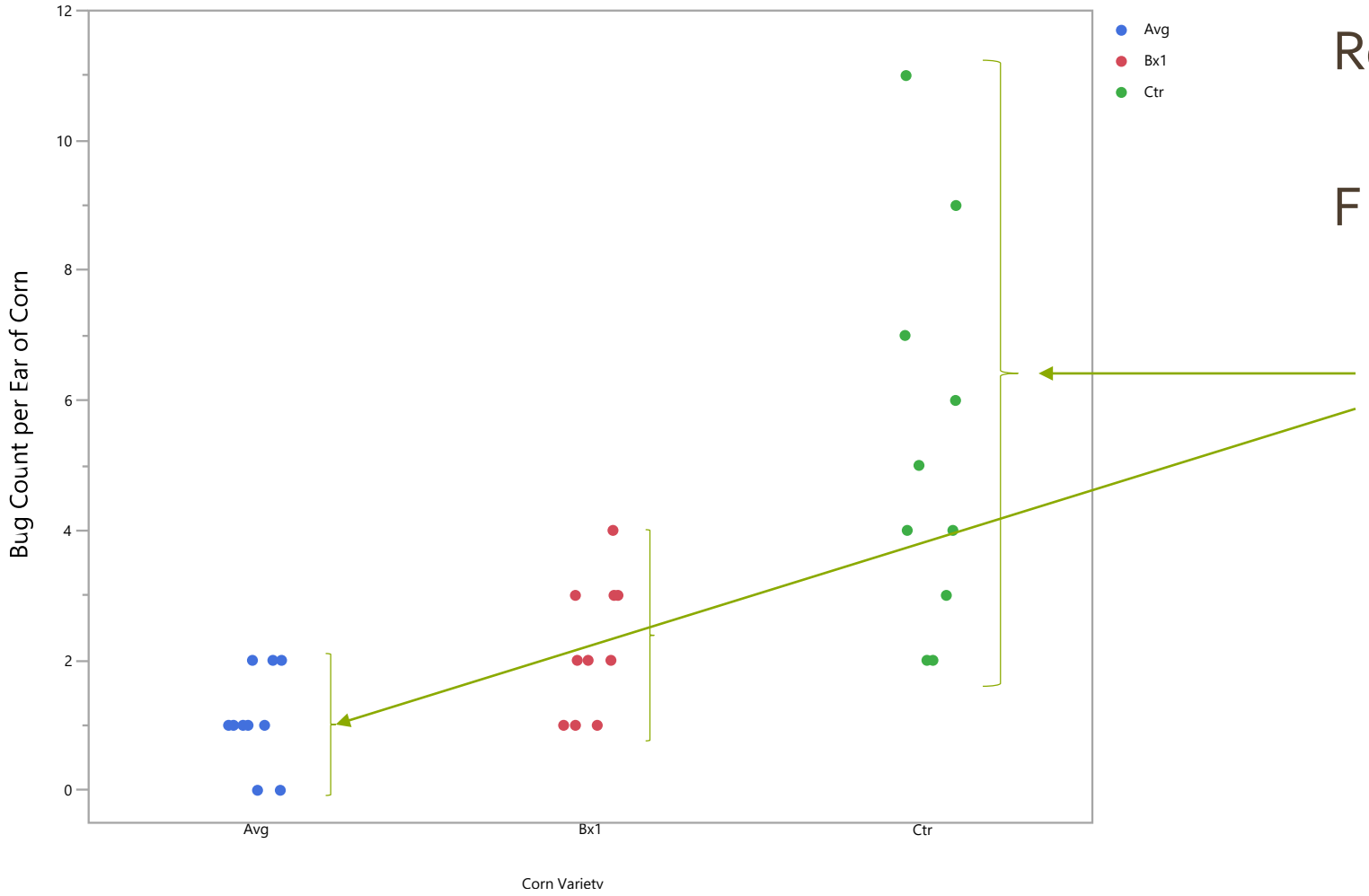
- assumes the 10 bug counts observed on a corn variety exhibit a "Normal" Bell-Shaped, distribution.

- Residual = [ModelPredicted - Observed]

Residuals from the Normal Distribution ANOVA



Observed Data: Tarnished Plant Bugs on BT Modified Corn



Recall:

$$F = \frac{MS[Among\ Varieties]}{MS[Residual]}$$

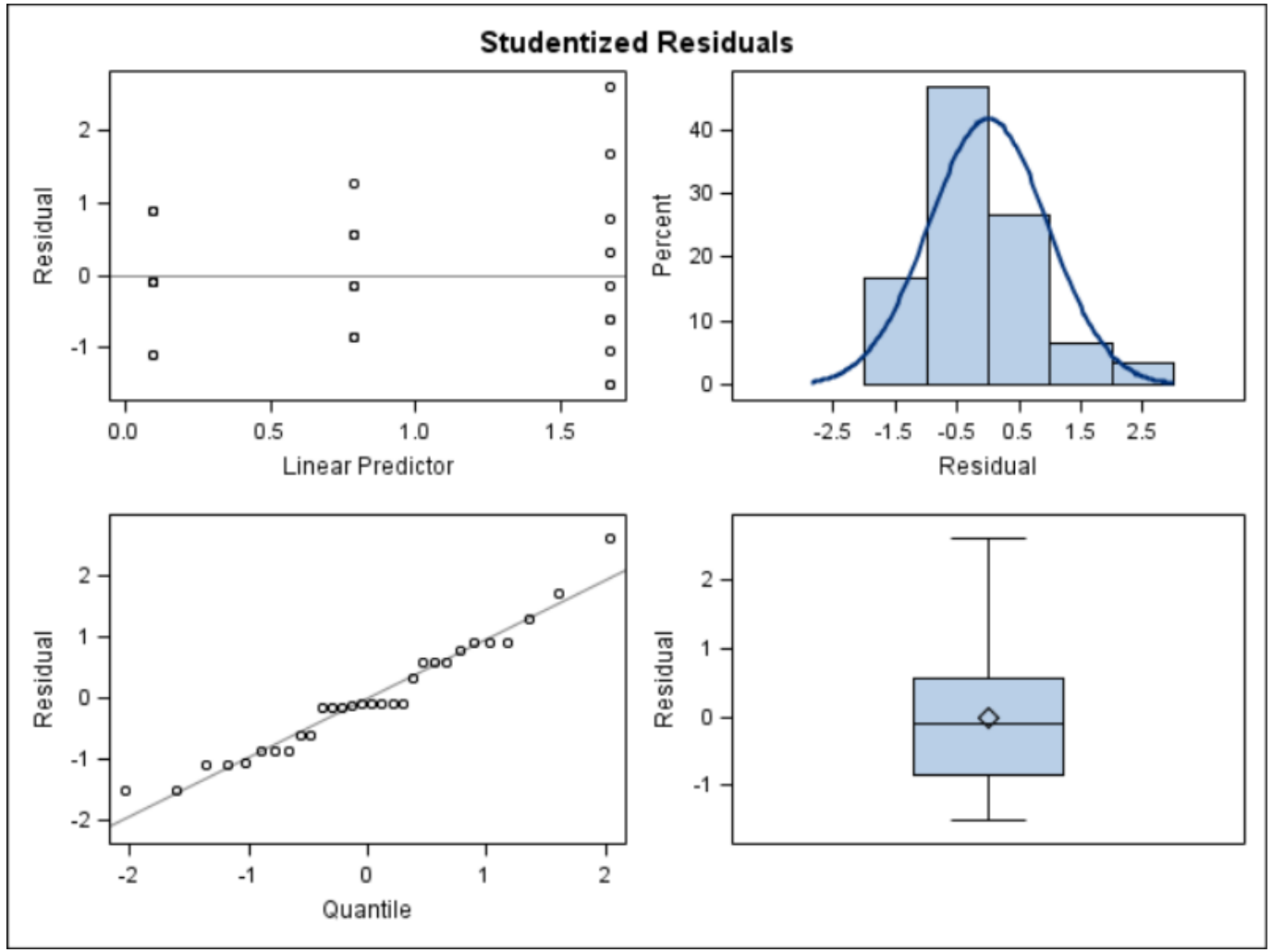
Variation within varieties are not of same magnitude.

Pooling across varieties will not give an accurate estimate of model residual variance.

Step 2: Fit the ANOVA Model using a Poisson Distribution

- Poisson distribution:
  - accurately describes "count data"
  - Variance increases as Mean increases

Residuals from Poisson ANOVA Model



Step 3: Conduct Valid Statistical Tests

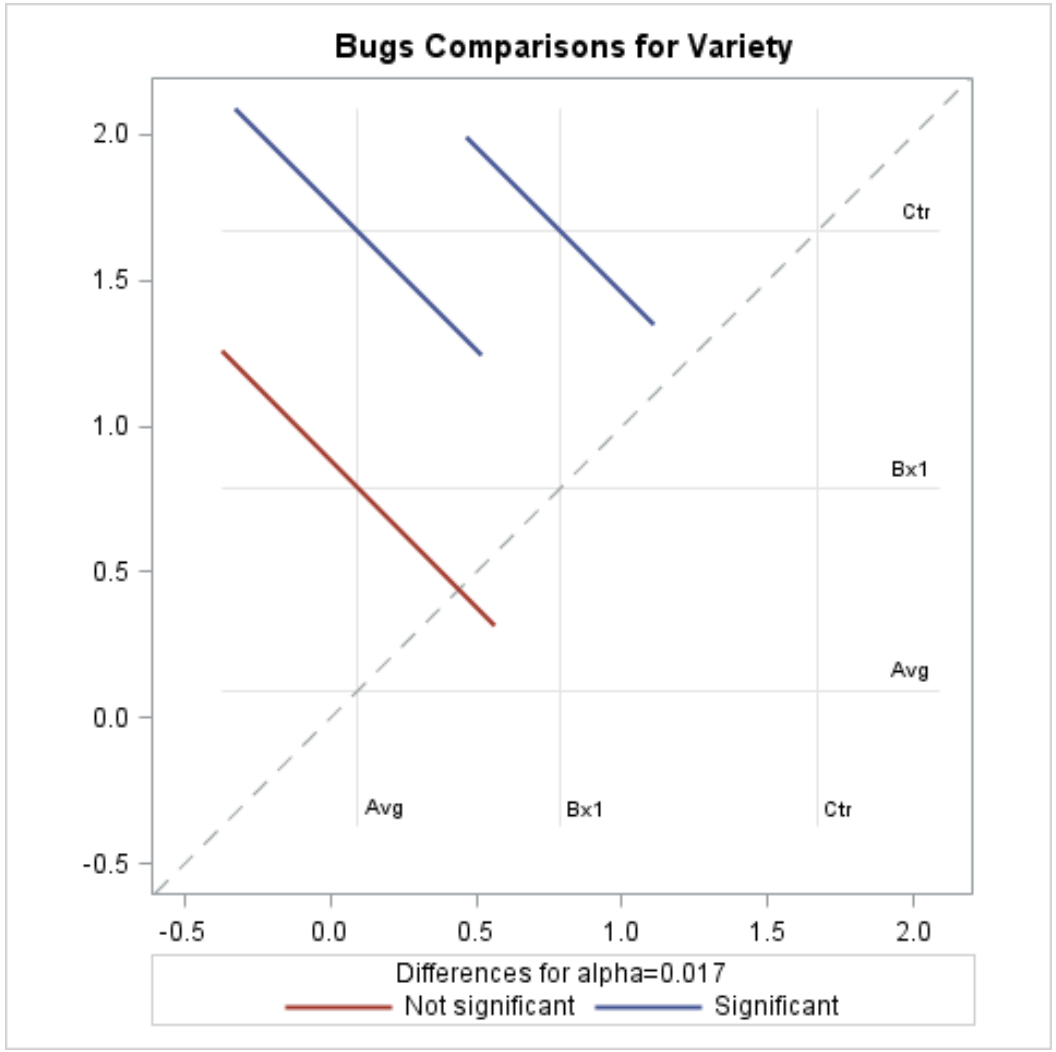
- The Poisson ANOVA model accurately describes the observed data (i.e., all model assumptions are realistic)

- (F-value = 14.3, P-value < 0.0001) indicates statistical difference in # tarnished plant bugs among the 3 varieties

- Poisson model produces *accurate estimates of standard errors*:

Variety	Mean	StdErr[Poisson]	StdErr [Pooled]
Afh	1.1	0.33	0.59
Bx1	2.2	0.47	0.59
Ctr	5.3	0.73	0.59

Mean Comparison Plot for the 3 Varieties – Poisson Model



Communication Skills – Vital to Collaboration

- Informal Communication:
  - Primarily Verbal Discussions / Collaborative Consultations
- Formal Communication: Based on Evidence (statistical tests)
  - Verbal Presentations
  - Written Materials: Research Proposals / Manuscripts / Posters
- Miscommunicating – talking at cross purposes

https://www.youtube.com/watch?v=Hz1fyhVOjr4

Modern Research Requires Collaboration:

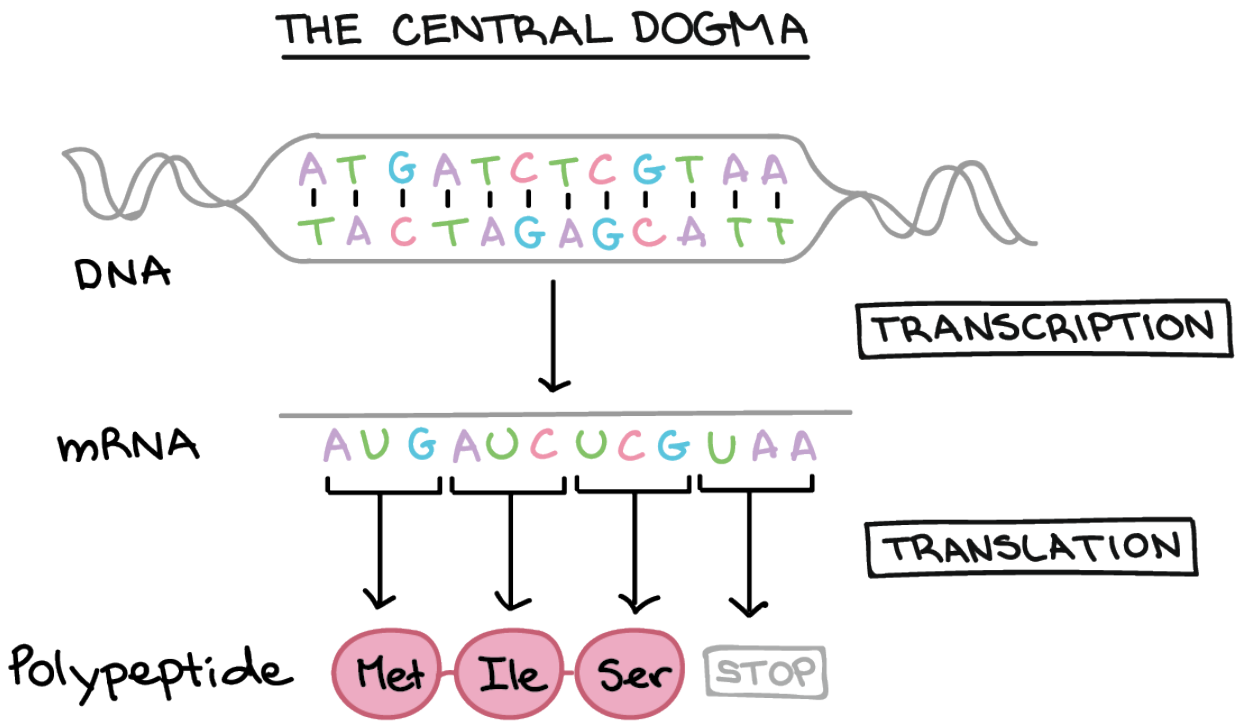
Incorporating
Statistics
and
Bioinformatics

Bioinformatics - What is it?

- "Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied maths, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale." – Luscombe et al. 2001

ATGCATGCATGAGAGAGATAAATAGGCCAAGGAAA

Bioinformatics - Central Dogma



https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

Bioinformatics - Puzzles

5′ ATGCATGCATGAGAGAGATAAATAGGCCAAGGAAA 3′

AAGGAAA

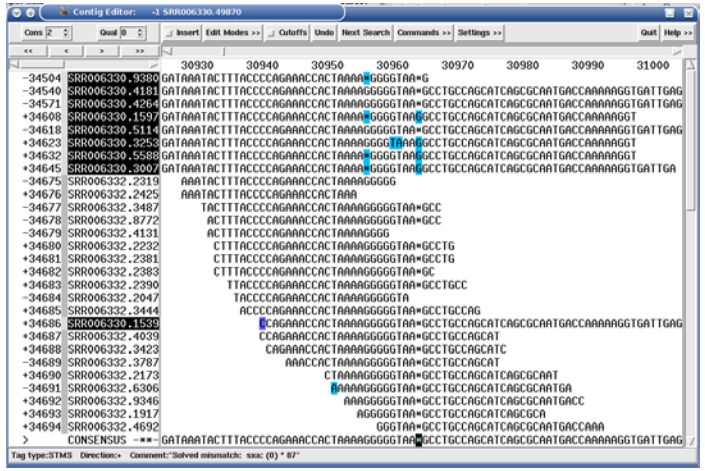
ATAGGCC

GCATGAGAGAGA

GAGAGATAA

ATGCATGCAT

AGGCCAA



http://gyra.ualg.pt/_static/mira/

https://www.biostars.org/p/16049/

In Silico Analysis –Scientific Method

- Define a question
- Gather information and resources (observe)
- Form an explanatory hypothesis
- Test the hypothesis by performing an experiment and collecting data in a reproducible manner
- Analyze the data (control and variables)
- Interpret the data and draw conclusions that serve as a starting point for new hypothesis
- Publish results
- Retest (frequently done by other scientists)

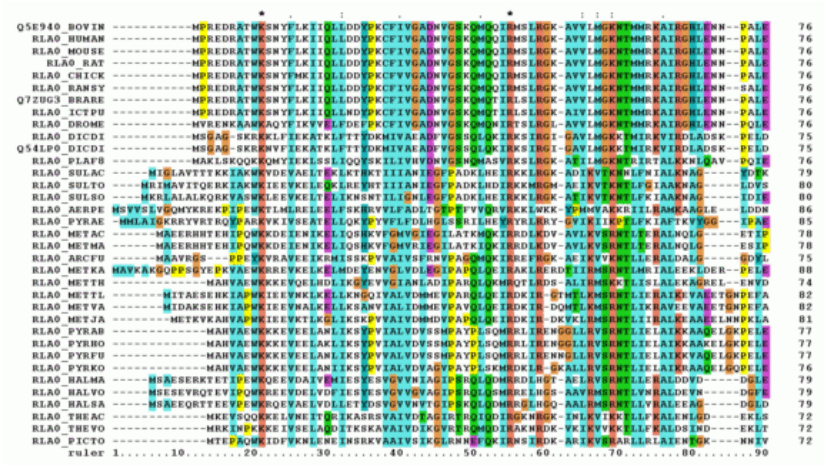
https://en.wikipedia.org/wiki/Scientific_method

Biology and Bioinformatics

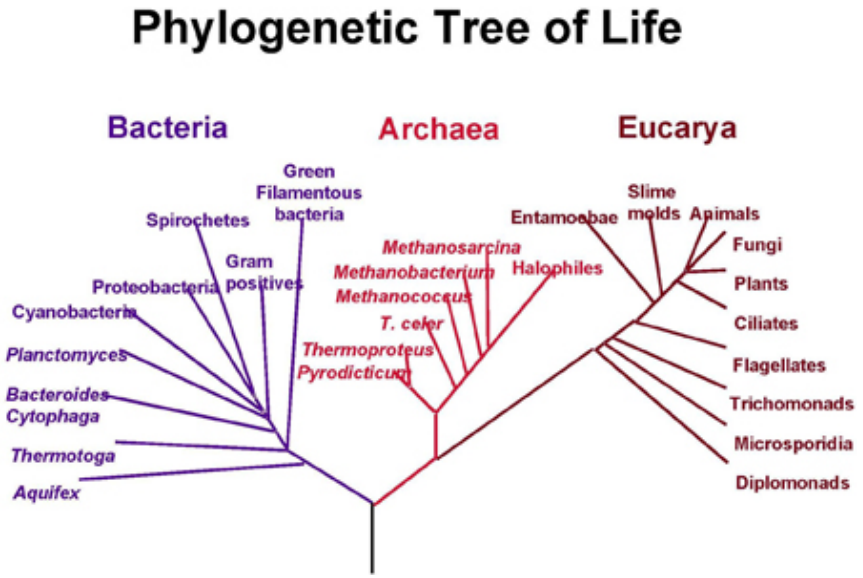
- Biochemistry
- Biology
- Chemistry
- Molecular Biology
- BioPhysics
- Genetics
- Protein and Structure analysis
- Machine Learning
- Gene Finding
- Statistics
- Expression analysis – Microarrays and RNAseq
- Phylogenetics
- Systems Biology
- Genome Annotation
- SNP Analysis
- Genome Assembly
- Numerical Methods
- Computational Biology and Algorithms

Calculus
Differential Equations

Bioinformatics – Alignments and Phylogenetics

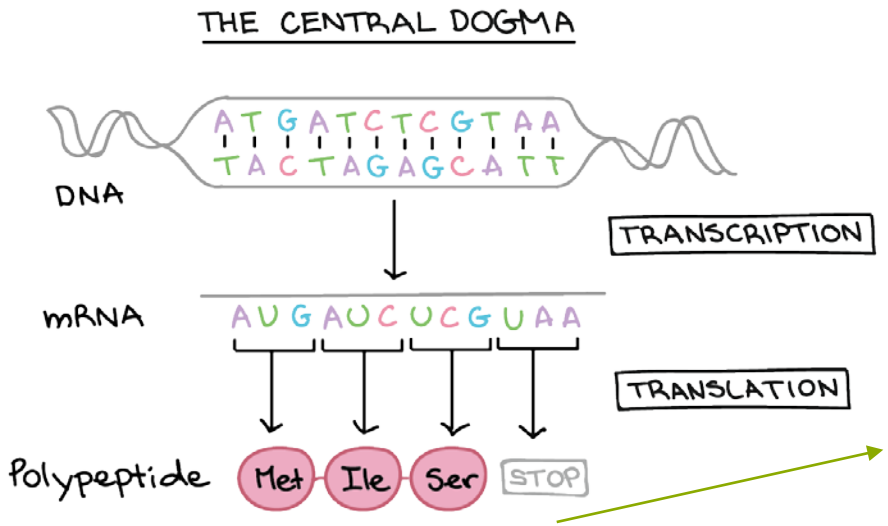


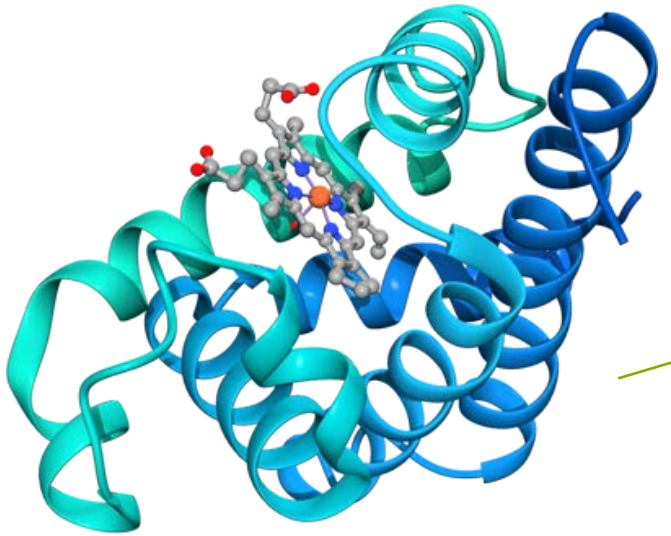
https://en.wikipedia.org/wiki/Multiple_sequence_alignment

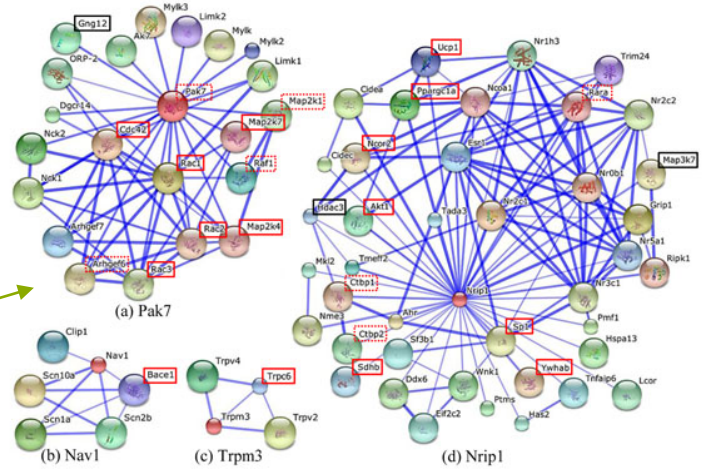


https://plus.maths.org/content/reconstructing-tree-life

Bioinformatics –Protein Structure and Networks





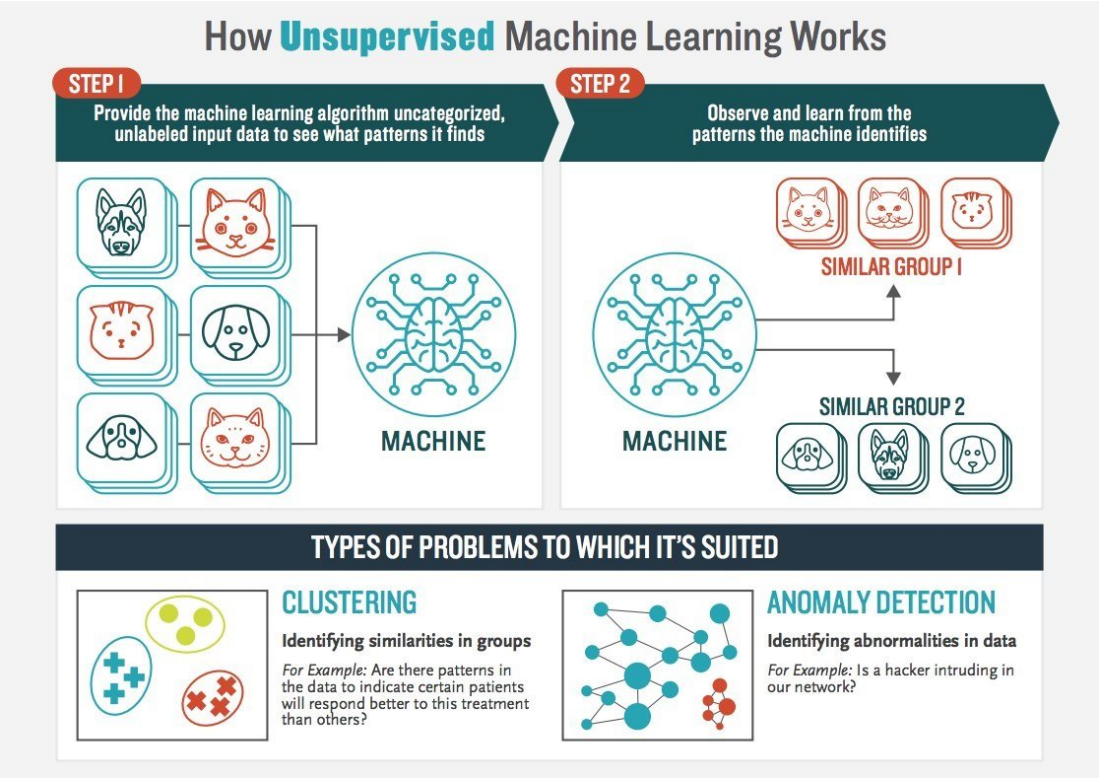


https://www.wwpdb.org/

https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

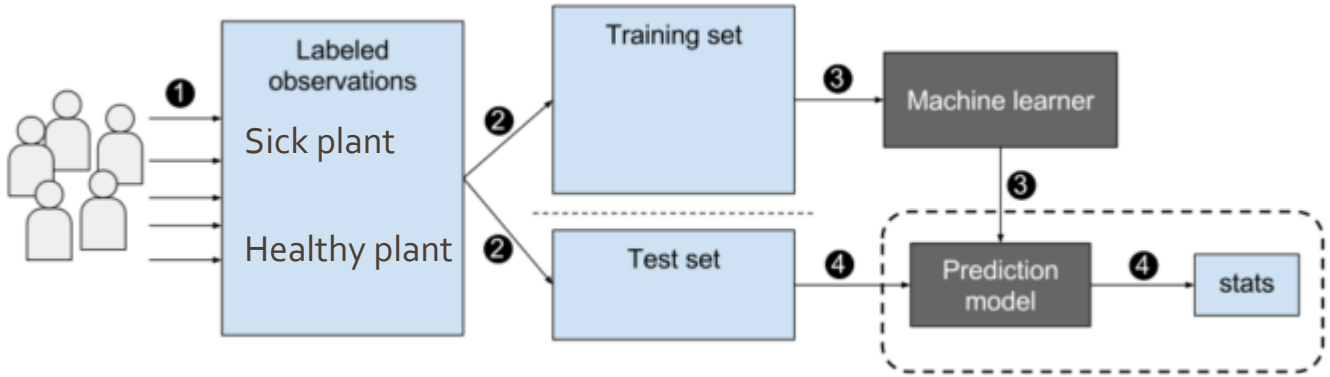
https://www.researchgate.net/figure/Gene-networks-constructed-from-interacting-protein Solid-lines-in-red-stand-for-genes_fig7_230610340

Machine Learning



https://www.newtechdojo.com/list-machine-learning-algorithms/

Supervised Machine Learning



https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Recommended Computational Tools

- R – programming and statistical modeling environment
- Python – efficient and powerful scripting language
- SAS – for purposes of "legacy" knowledge only
  - (proprietary) statistical software standard (1976 – early 2000's)

- Operating System
  - Windows or MAC [at least]
  - Linux – may be necessary for computationally intensive data analyses:
    - Genomics
    - Remote Sensing
    - Bayesian Statistical Methods

Which operating system should I use?

Linux

- an operating system (i.e., software) that manages all of the hardware resources associated with your desktop or laptop. Basically– the operating system manages the communication between your software and your hardware
- Perfect for beginners , learn as your first OS
- Very quick to learn , many users can learn Linux OS basics within a few days
- Free and portable to any hardware
- Easy to edit/modify scripts
- Most bioinformatic programs are written for the Linux platform , require some Linux "pipelining"

https://www.linux.com/what-is-linux

Important Programming Languages to Know

Python

- Object- oriented programming language, and also known as general language, which means it can be used to build just about anything.
- Code is easy to read
- It can be quickly written
- Portable – Works on most operating systems (Linux , Win, MacOS)
- It's commonly used in data science, machine learning fields & Web Development
- Opens a lot of career opportunities
- Data structures and Algorithms

Which programming language should I learn first?

R

- A free open source statistical programming language that is perfect for working on statistical tests, models, and analyses
- Cross platform which means it works on almost every operating system
- Fairly easy to learn and similar to other languages
- Has over a thousand bioinformatic packages
- Perfect for creating visualizations, will make almost any graph you can imagine

Online "Data Science" Training Resources

Data Science Specialization Certificate – by Johns-Hopkins University:

https://www.coursera.org/specializations/jhu-data-science

Free to view, certificates and grades have a cost.

Comprehensive Selection of ***R*** and ***Python*** Learning Modules on Data Science Topics

https://www.datacamp.com

Choice between payment per course or an (annually renewable) subscription that grants access to all courses.

Application Domain-Specific Teaching Materials Free to view

https://datacarpentry.org/lessons/

Bioinformatics course Fall 2019

Beginning Bioinformatics courses , Fall 2019, Free to view, certificates and grades have a cost.

https://www.edx.org/course/dna-sequences-alignments-and-analysis

https://www.edx.org/course/proteins-alignment-analysis-and-structure-1

https://www.edx.org/course/statistical-analysis-in-bioinformatics-1

https://www.edx.org/course/principles-statistical-and-computational-tools-for-reproducible-data-science

https://www.coursera.org/learn/bioinformatics

https://www.coursera.org/learn/genome-sequencing

https://www.coursera.org/learn/comparing-genomes

https://www.coursera.org/learn/molecular-evolution

https://www.coursera.org/learn/genomic-data

https://www.coursera.org/learn/bioinformatics-project

Machine Learning courses

https://www.edx.org/course/machine-learning-fundamentals-4

https://www.edx.org/course/machine-learning

Commonly used bioinformatic resources

- **Biostars** – Platform to ask any bioinformatic question https://www.biostars.org/
- **Stackoverflow** – Blog for answering and posting Computational Questions https://stackoverflow.com/
- **NCBI/Genbank** – a database of publicly available sequences and its part of the **NCBI** (National Center for Biotechnology Information) – series of databases that references biology and biomedical topics https://www.ncbi.nlm.nih.gov/
- **EMBOSS** (European Molecular Biology Open Software Suite) https://www.ebi.ac.uk/
- **KEGG** (Kyoto Encyclopedia of Genes and Genomes) https://www.genome.jp/kegg/kegg1.html

Take-Home Messages

- *Embrace every opportunity* to Become more "**Numerically Literate**"
  - Statistical Methods
  - Programming Languages and Operating Systems
  - Computational Biology Methods
  - Bioinformatics Applications

- *Understand YOUR discipline well*
  - Be able to communicate scientific concepts using straight-forward English
  - Avoid relying on technical jargon
  - Can you explain your research to someone unfamiliar with your discipline, so that it is clearly understood?

- *Develop Strong Interpersonal Communication Skills*

- *Statisticians and Bioinformaticists* – get to "play in everyone's backyard"

Thank You for Coming!

We hope you're enjoying your time at ARS