What is Data Science

2021.03.09

BiS800: Methods in functional genomics and
computational molecular biology

Survey

1. Did you listen to Jack Horner’s TED talk?
2. Do you think T-rex is a chicken or a reptile?
3. What best describes your impression towards Data Science?
4. How much did your follow this week’s reading material on “R for Data Science”?
5. What is your strong suit?

Last time

For T-rex the chicken:

No alternative hypothesis for this peptide identification that cannot be simply a coincidence. Out of all the species why was it chicken?

Much of biology is based on observations, and this is a simple report on the biological observations made by the others.

Against T-rex the chicken:

Peptide sequencing is more appropriate as the T-rex peptide information is not in the database.

Peptide identification was collagen-only. Expanding the database to any peptide may reveal peptides from other species.

Finding is **not statistically significant**. Need independent/complementary evidence.

The Origin of Data Science

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

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

Machine learning is the study of computer algorithms that improve through experience (in our case that’ll be biological data).

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

What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

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

Quotes from 🐦

‘Data Scientist’ is a Data Analyst who lives in California.

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

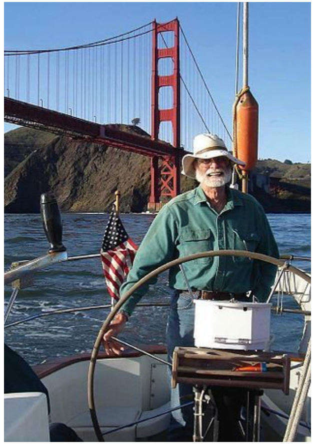
A data scientist is a business analyst who lives in New York.

A data scientist is a statistician who lives in San Francisco.

Data Science is statistics on a Mac.

https://datascopeanalytics.com/blog/what-is-a-data-scientist/

Fourth paradigm of science

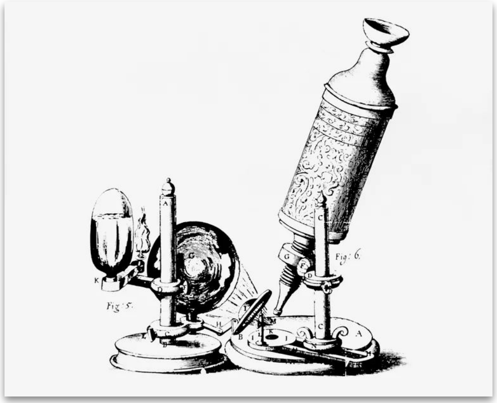


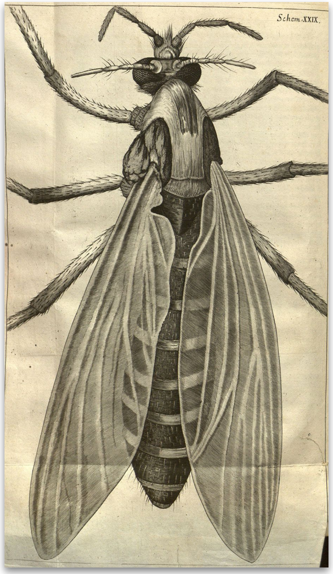
Turing award winner Jim Gray (1944 - 2007)

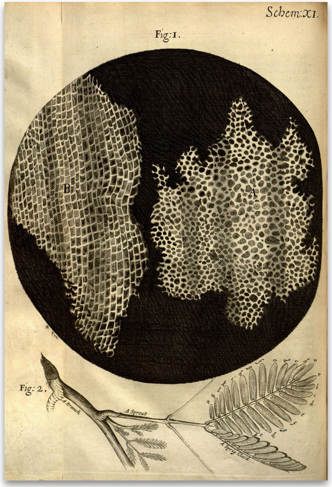
1. Empirical
2. Theoretical
3. Computational
4. Data-driven intensive research

In 1995, he advocated building clusters of “storage bricks,” consisting of inexpensive, balanced systems of central processing units, memory, and storage for data-intensive research.

First paradigm of science: Empirical (1665)





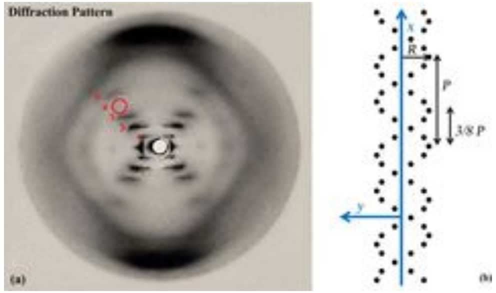


Robert Hooke's drawings of the cellular structure of cork and a sprig of sensitive plant from *Micrographia*.

While this and similar works of microscopists seem to lack a definite objective, the development of microscopes sparked the beginning of observation and experimentation during a time of hypothetical and philosophical speculations.

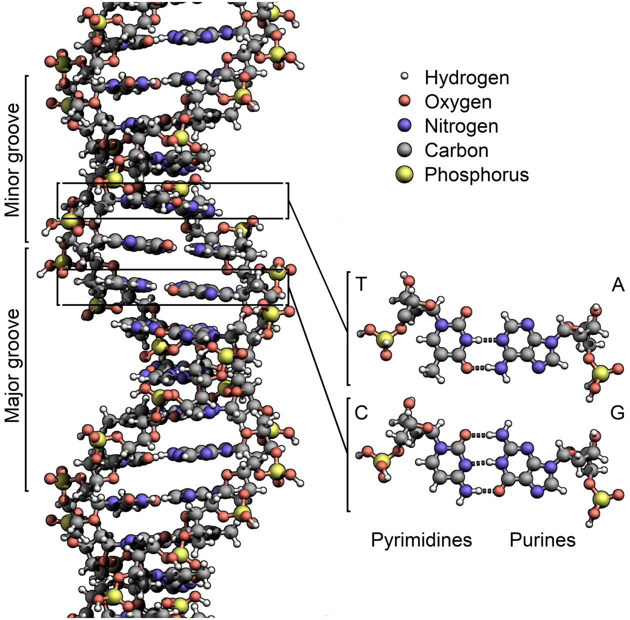
https://www.britannica.com/science/biology/The-discovery-of-cells

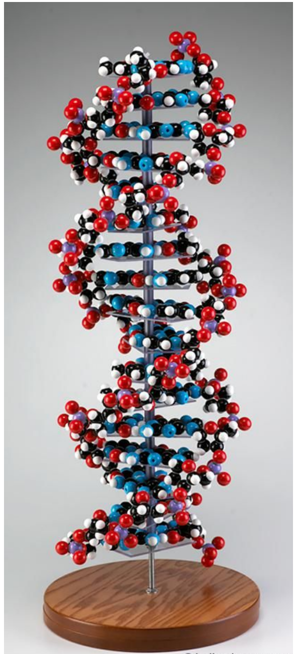
First paradigm of science: Empirical (1953)



The diffraction pattern from DNA in its so-called B configuration.

Rosalind Franklin used X-ray diffraction to determine the structure of DNA molecules



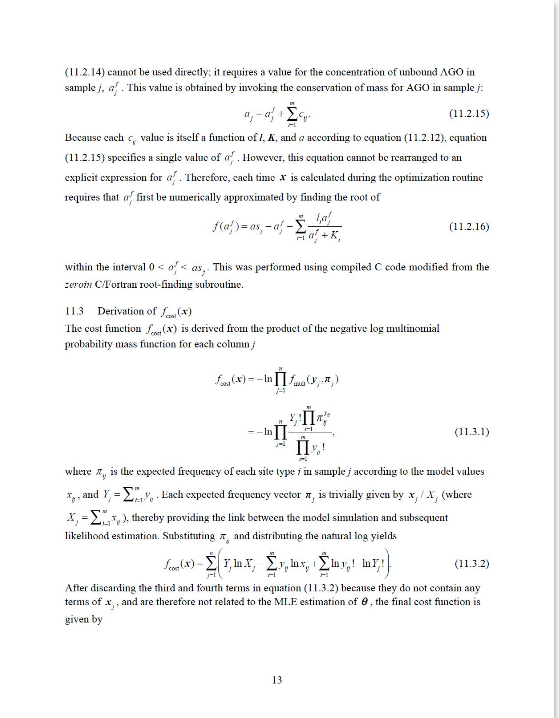


© Indigo Instruments

Thompson et al. (2018, Teaching material)

https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

Second paradigm of science: Theoretical



(11.2.14) cannot be used directly; it requires a value for the concentration of unbound AGO in sample *j*, a_j^f . This value is obtained by invoking the conservation of mass for AGO in sample *j*:

$$a_j = a_j^f + \sum_{i=1}^{m} c_{ij}. \qquad (11.2.15)$$

Because each c_{ij} value is itself a function of ***l***, ***K***, and *a* according to equation (11.2.12), equation (11.2.15) specifies a single value of a_j^f . However, this equation cannot be rearranged to an explicit expression for a_j^f . Therefore, each time $\boldsymbol{x}$ is calculated during the optimization routine requires that a_j^f first be numerically approximated by finding the root of

$$f(a_j^f) = as_j - a_j^f - \sum_{i=1}^{m} \frac{l_i a_j^f}{a_j^f + K_i} \qquad (11.2.16)$$

within the interval $0 < a_j^f < as_j$. This was performed using compiled C code modified from the *zeroin* C/Fortran root-finding subroutine.

11.3 Derivation of $f_{\text{cost}}(\boldsymbol{x})$

The cost function $f_{\text{cost}}(\boldsymbol{x})$ is derived from the product of the negative log multinomial probability mass function for each column *j*

$$f_{\text{cost}}(\boldsymbol{x}) = -\ln \prod_{j=1}^{n} f_{\text{mult}}(\boldsymbol{y}_j, \boldsymbol{\pi}_j)$$

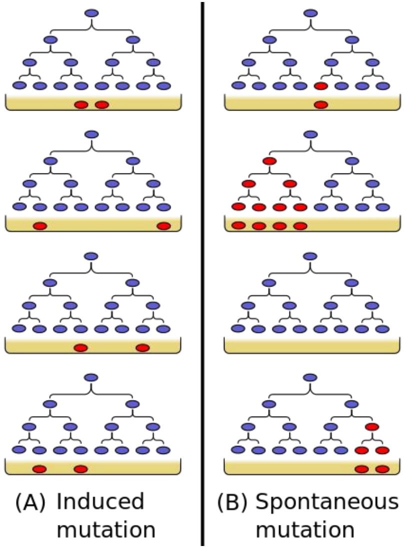
$$= -\ln \prod_{j=1}^{n} \frac{Y_j! \prod_{i=1}^{m} \pi_{ij}^{y_{ij}}}{\prod_{i=1}^{m} y_{ij}!}, \qquad (11.3.1)$$

where π_{ij} is the expected frequency of each site type *i* in sample *j* according to the model values x_{ij}, and $Y_j = \sum_{i=1}^{m} y_{ij}$. Each expected frequency vector $\boldsymbol{\pi}_j$ is trivially given by $\boldsymbol{x}_j / X_j$ (where $X_j = \sum_{i=1}^{m} x_{ij}$), thereby providing the link between the model simulation and subsequent likelihood estimation. Substituting π_{ij} and distributing the natural log yields

$$f_{\text{cost}}(\boldsymbol{x}) = \sum_{j=1}^{n} \left(Y_j \ln X_j - \sum_{i=1}^{m} y_{ij} \ln x_{ij} + \sum_{i=1}^{m} \ln y_{ij}! - \ln Y_j! \right). \qquad (11.3.2)$$

After discarding the third and fourth terms in equation (11.3.2) because they do not contain any terms of $\boldsymbol{x}_j$, and are therefore not related to the MLE estimation of $\boldsymbol{\theta}$, the final cost function is given by

13



Luria–Delbrück experiment (1943)
Available on KLMS

In 1940s, the ideas of inheritance and mutation were generally accepted, but the responsible biomolecule (i.e. DNA) was unknown.

At the time, it was thought that bacteria could develop heritable genetic mutations depending on the environment.

McGeary and Lin et al. 2020 from Bartel Lab

https://en.wikipedia.org/wiki/Luria–Delbrück experiment

Third paradigm of science: Computational

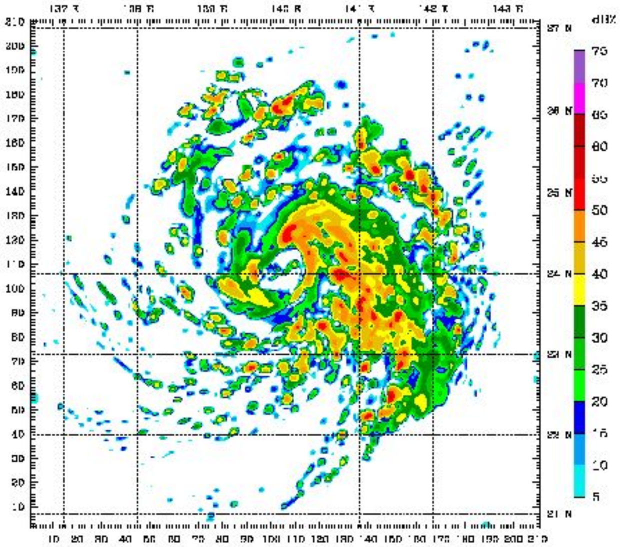
Complex systems

Weather forecast

Population dynamics

*Human cognition and performance

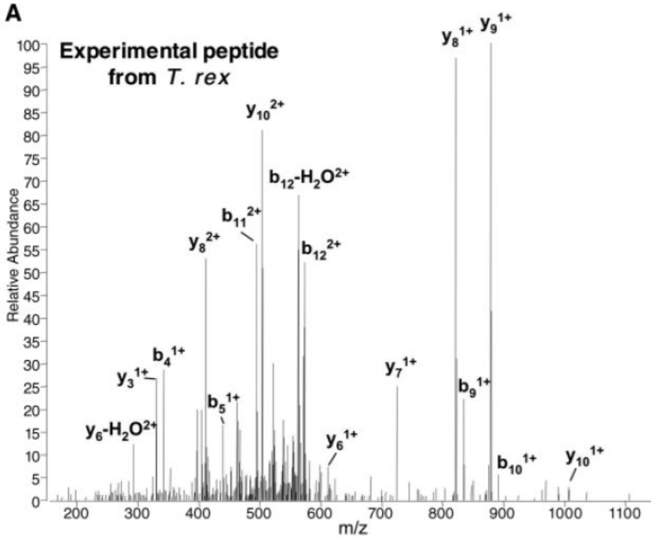
*Molecular modeling and more



Fourth paradigm of science: Data Science

Living complex systems

Abundance



Molecular charge and weight

1. Peptide sequencing (no database)
2. **Peptide identification (yes database)**
3. False discovery rate (decoy database)

Example from last lecture

1. First paradigm: Mass Spectrometry
2. Second paradigm: Analytical chemistry
3. Third paradigm: Peptide sequencing

   Similar to DNA/RNA/Protein sequence alignment, an scoring system must be implemented. Scoring system is based on analytical chemistry.

4. Fourth paradigm: Peptide identification

   Limit search space to peptides present in the proteome database.

Example from last lecture

1. First paradigm: Mass Spectrometry
2. Second paradigm: Analytical chemistry
3. Third paradigm: Peptide sequencing

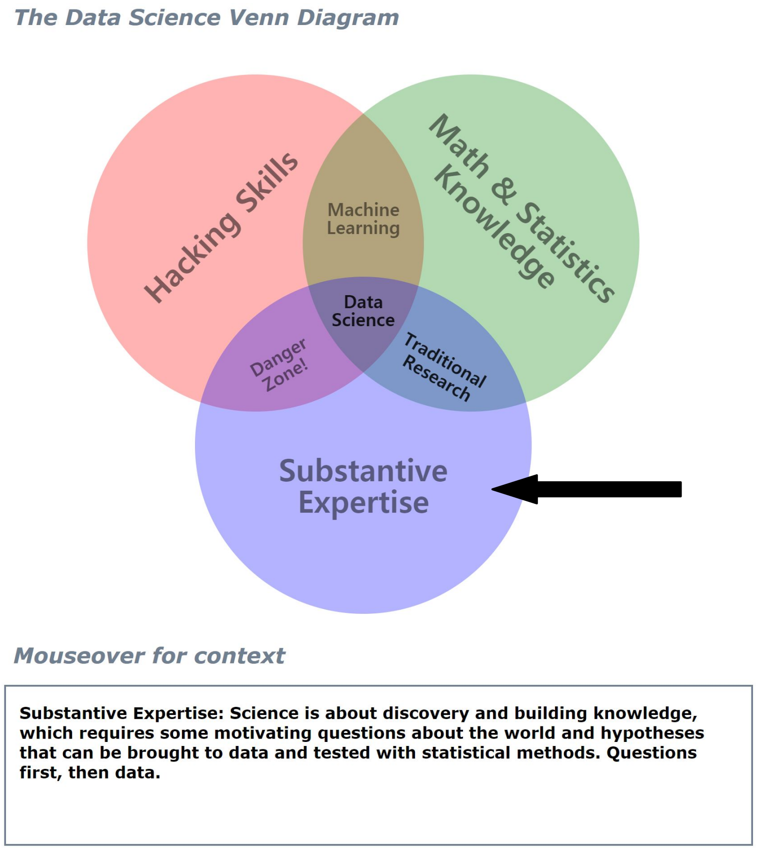
   Similar to DNA g system must be
   implemented. stry.

The advancement of techniques in biomedical engineering has enabled the democratization and subsequent explosion of bio-data generation.

4. Fourth paradigm: Peptide identification

   Limit search space to peptides present in the proteome database.

Corollary: Effective execution of data science depends on your understanding of the empirical, theoretical and computational aspect of your specific question.

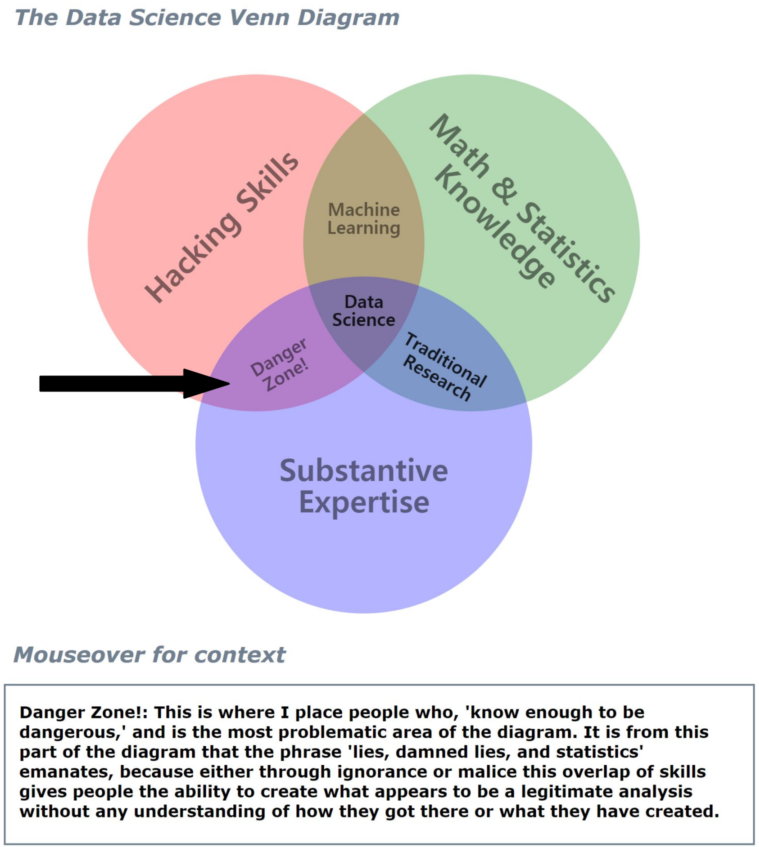


The Data Science Venn Diagram

Mouseover for context

Substantive Expertise: Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods. Questions first, then data.

https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html



The Data Science Venn Diagram

Mouseover for context

Danger Zone!: This is where I place people who, 'know enough to be dangerous,' and is the most problematic area of the diagram. It is from this part of the diagram that the phrase 'lies, damned lies, and statistics' emanates, because either through ignorance or malice this overlap of skills gives people the ability to create what appears to be a legitimate analysis without any understanding of how they got there or what they have created.

https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html

5 minute break

Fundamental concepts in Data Science

Import

Tidy

Transform

Data wrangling

Visualization

Models

Communicate

Tidy: organize data in consistent form

Tidying your data means storing it in a **consistent** form that matches the semantics of the dataset with the way it is stored.

For example for RNA-Seq data, each row is a gene and each column is the patient or sample ID.

Tidy data is important because the consistent structure lets you **focus** your struggle on questions about the data, not fighting to get the data into the right form for different functions.

Transform: organize data in consistent form

Transformation includes:

1. narrowing in on observations of interest
    a. all people in one city
2. creating new variables that are functions of existing variables
    a. computing speed from distance and time
3. calculating a set of summary statistics
    a. like counts or means

Together, tidying and transforming are called **wrangling** because getting your data in a form that's natural to work with often feels like a fight! But why go through this trouble will be clear at the end of this lecture.

Visualization (or exploration): Show me your data (week 4)

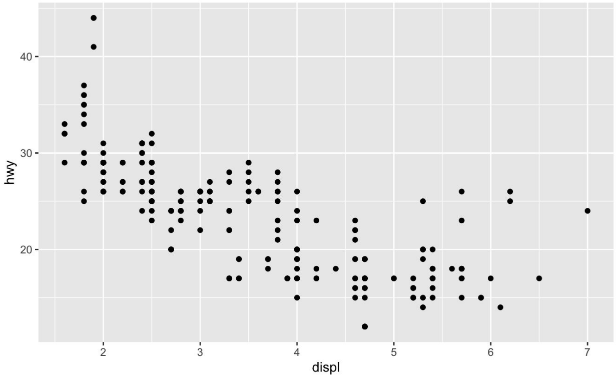
Visualisation is a fundamentally **human** activity. If other words, this is where your **expertise** in the subject of interest (i.e. biology) plays the biggest role.

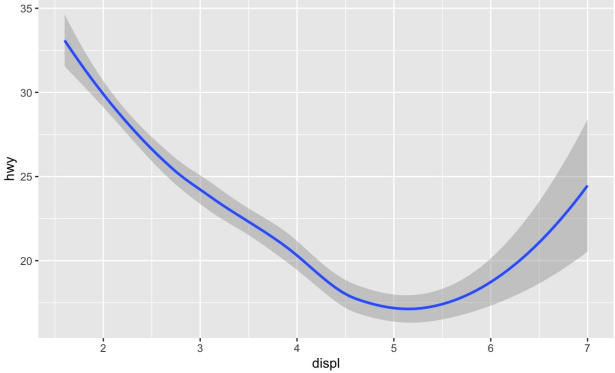
A good visualisation might also hint that you're asking the wrong question, or you need to collect different data.

For example, outliers! Are those outliers just technical artifacts or is it a fundamentally rare but essential feature related to your question?

Remember the first paradigm of science?

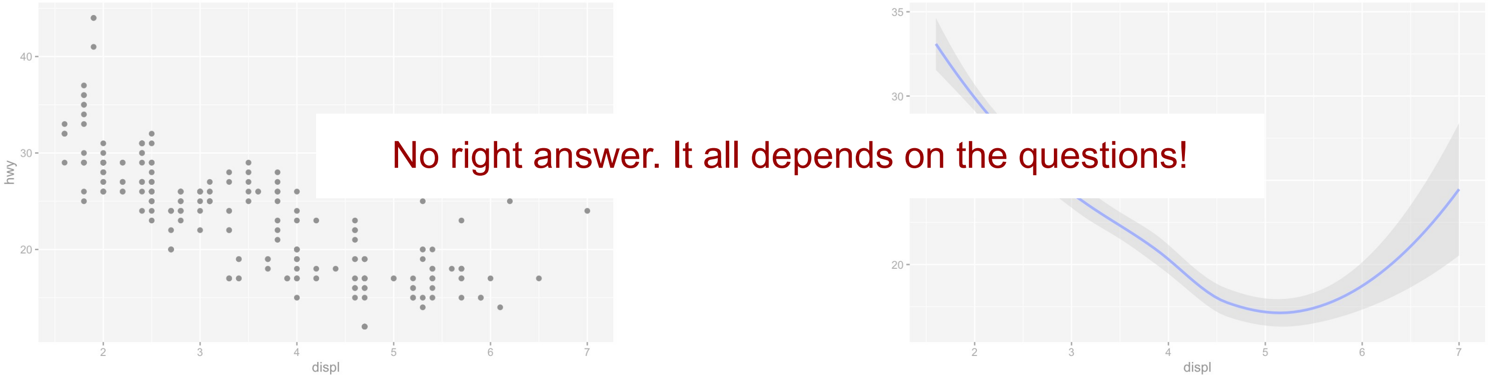
Visualization choice example and thoughts?





displ, a car’s engine size, in litres. And hwy, a car’s fuel efficiency on the highway, in miles per gallon (mpg). A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance.

Visualization choice example and thoughts?



No right answer. It all depends on the questions!

displ, a car’s engine size, in litres. And hwy, a car’s fuel efficiency on the highway, in miles per gallon (mpg). A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance.

Models: Show me your method (week 5)

Models are complementary tools to visualisation.

Once you have made your questions sufficiently precise, you can use a model to answer them.

Models are a fundamentally **mathematical** or **computational** tool. Remember the second and third paradigm?

Every model makes assumptions! If you compute the mean, fold change, etc, you are nonetheless (un)consciously making assumptions. There are also assumptions in nonparametric statistics.

“All models are wrong, but some are useful.”

George Box (1976)

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

“Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. ...”

George Box (1976)

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

“... Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.”

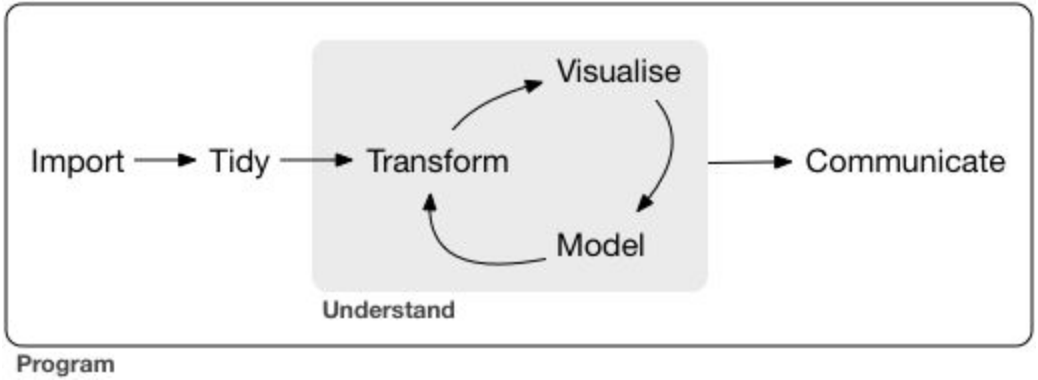
George Box (1976)

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

Typical workflow in Data Science

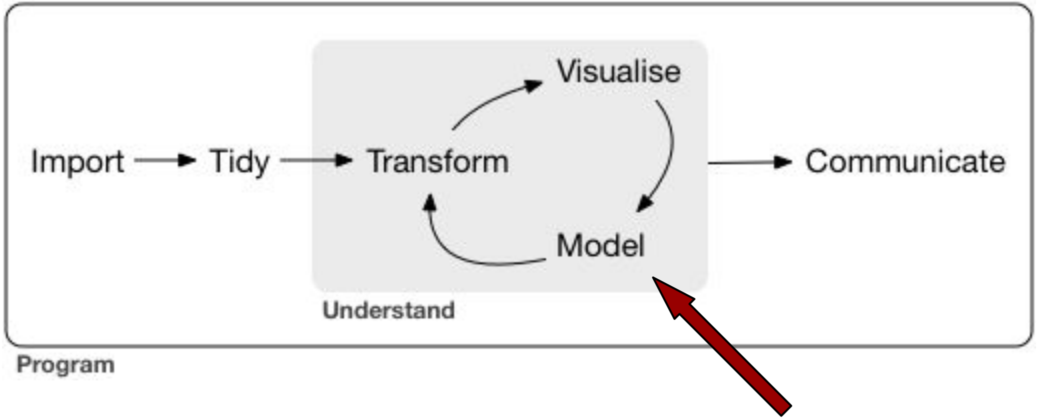
Tidying provides platform for iterative optimization in data analysis.

New computational methods enables this iterative process.



Another shameless advertisement ...

I’m interested in developing new algorithms to enable data-specific modeling and maximize information extraction.



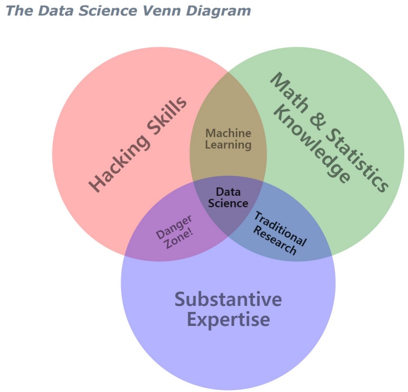
Summary

Data wrangling is the most time consuming step in data analysis, but is the foundation of your experience with the data.

Visualization provides an interactive medium between your substantive expertise and your data.

Modeling enables you to extract insight from your data.

Don’t be in the Danger Zone!



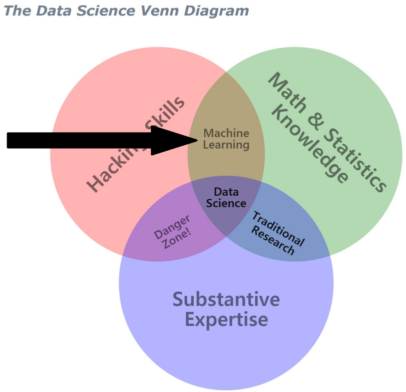
My story

Computer science and mathematics as an undergraduate, but little biology. The most advanced biology course I took was Genetics.

Statistics during my time as graduate student.

Molecular biology in 2016 when I started as a postdoc.

I'm more interested in the question itself than the computational technique.


The Data Science Venn Diagram

Possible directions for biologists

Import

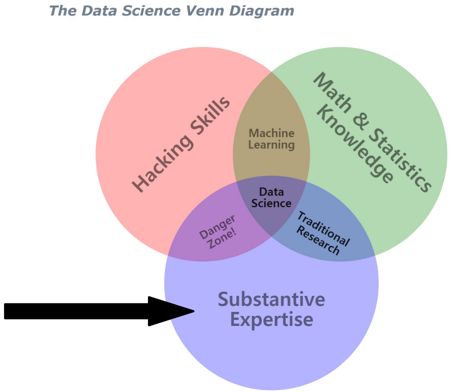
Tidy

Transform

Visualization*

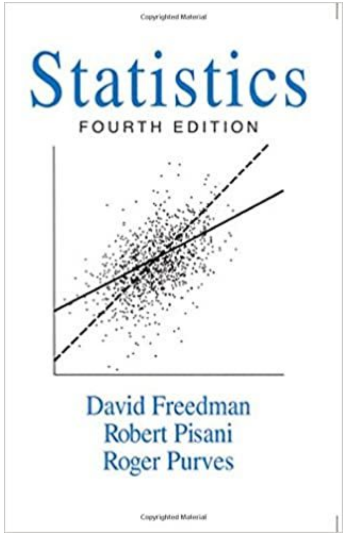
Models

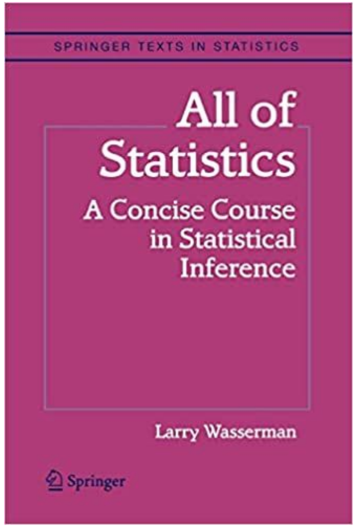
Communicate

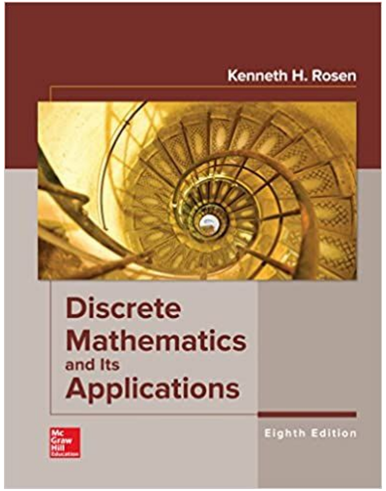


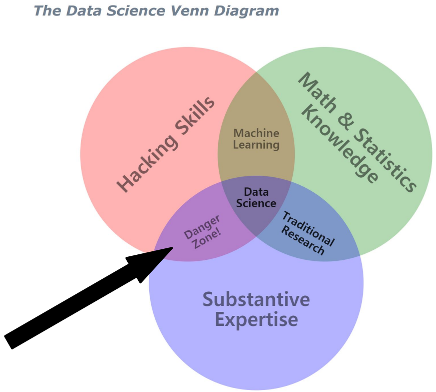
If you think you might be heading to the Danger Zone

It's not too late to learn math and statistics! (Also next week on statistics)




SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer





What does this have to do with the method section?

The computational method section must state:

1. How the data was transformed (or preprocessed) and why
2. What was the questions (i.e. why the particular visualization was used)
3. The rationale behind the assumptions used to model the data

Oral presentation for week 4/5

Slides are due via KLMS by Tuesday 3/23, 3/30 before class starts

Tuesday 3/23 and Thursday 4/1 speakers (n=8):

김병수, 문채영, 이동은, 김규희, 안혜진, Hamza A. Dar, 이한슬, 황희선

Thursday 3/25 and Tuesday 3/30 speakers (n=9):

장현수, 주재건, 양진욱, 임진아, 진주애, 이건용, 이기헌, 이경림, 김서현

Shuffle live? If schedule conflict, please contact TA

Next lecture

Exploratory data analysis

Data wrangling

Scribe volunteer?