Course Introduction: GSND 5355Q

Machine Learning for Biomedical Data (MLBD)

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Associate Director, Center for Biomedical Informatics and Health AI
Rutgers University – New Jersey Medical School
w.evan.johnson@rutgers.edu

2025-10-14

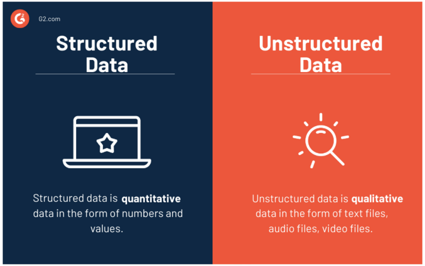
Things you should know about MLBD

- Click here for the Zoom link
- GitHub vs Canvas:
  - https://github.com/wevanjohnson/2025_Fall_MLBD
- Link to Syllabus
- Background experience
  - Introductory statistics and molecular biology
- Prerequisites (Machine Learning for Biomedical Data)
  - Basic Unix scripting
  - Amarel access and experience (ondemand, submissions)
  - Basic R programming: tidyverse, ggplot2, R Markdown
  - Working knowledge of git and GitHub

Introduction to Machine Learning and Data Science

Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data



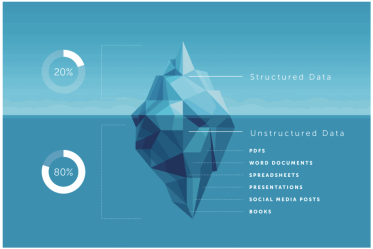
G2.com

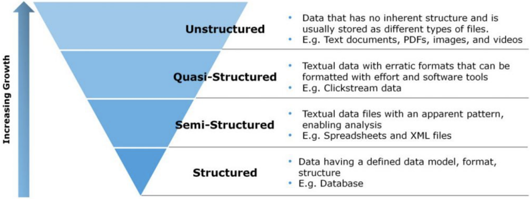
Structured Data

Structured data is **quantitative** data in the form of numbers and values.

Unstructured Data

Unstructured data is **qualitative** data in the form of text files, audio files, video files.





Unstructured

- Data that has no inherent structure and is usually stored as different types of files.
- E.g. Text documents, PDFs, images, and videos

Quasi-Structured

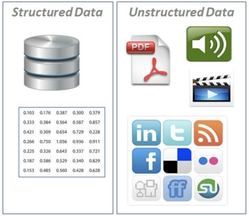
- Textual data with erratic formats that can be formatted with effort and software tools
- E.g. Clickstream data

Semi-Structured

- Textual data files with an apparent pattern, enabling analysis
- E.g. Spreadsheets and XML files

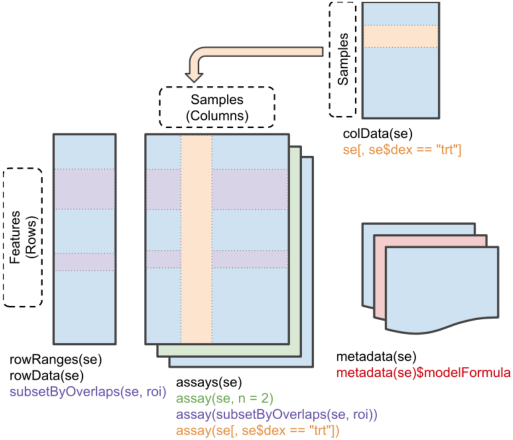
Structured

- Data having a defined data model, format, structure
- E.g. Database



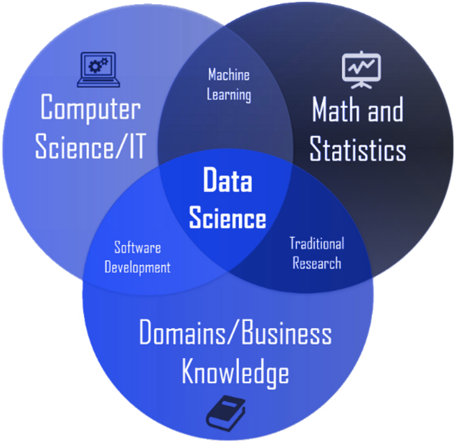
0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Structured vs. Unstructured data

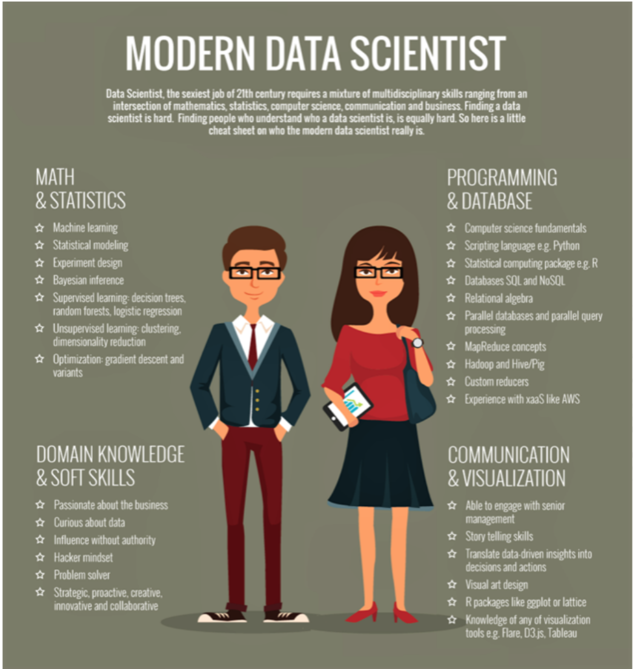


Summarized Experiment

Data Science Revolution



- Few have all the skills
- Flexibility in area (business, strategy, health care) and conditions
- Data science makes companies and data better!



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Science Process

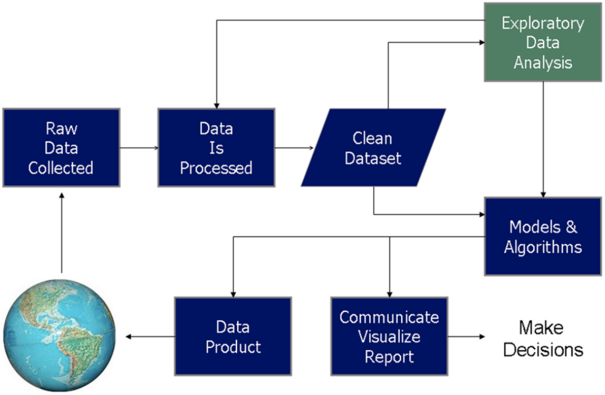
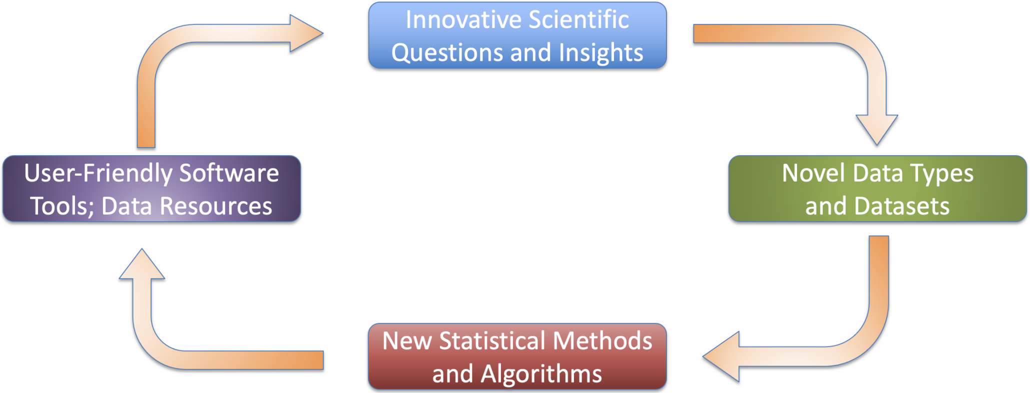


Image: https://en.wikipedia.org/wiki/Data_science

WEJ LAB

Scientific Cycle for Data Science

Johnson Lab Approach to Science:



Innovative Scientific Questions and Insights

Novel Data Types and Datasets

New Statistical Methods and Algorithms

User-Friendly Software Tools; Data Resources

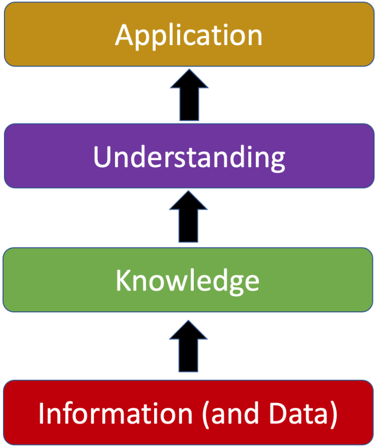
Center for Data Science

Keeping the “Science” in Data Science

Domain Knowledge

Domain knowledge is knowledge of a specific, specialized discipline or field, in contrast to general (or domain-independent) knowledge. For example, in describing a software engineer may have general knowledge of computer programming as well as domain knowledge about developing programs for a particular industry. People with domain knowledge are often regarded as specialists or experts in their field. (Wikipedia!)

Analytics Hierarchy



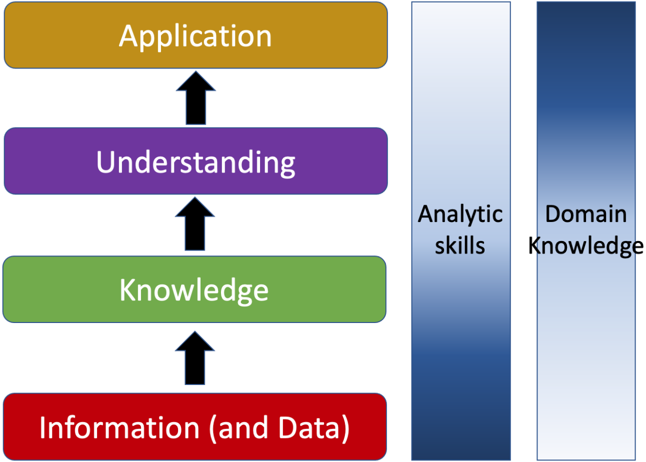
Application

Understanding

Knowledge

Information (and Data)

Analytics Hierarchy



Formal definitions

Machine learning is a computer's way of learning from examples, and it is one of the most useful tools we have for the construction of *artificially intelligent* systems

Artificial intelligence is a term used when machines (or computers) can mimic functions that humans can do.

- Example: Learning and problem solving

Formal definitions

An **Algorithm** is a sequence of actions that are “self-contained”

- Effective method for calculation
  - Finite steps/instructions
  - When applied, it produces a correct answer
  - The instructions need to be followed
  - In principle, it can be done by a “human”

Formal definitions

A **Computer Algorithm** is a sequence of actions that are “self-contained”

- Effective method for calculation
  - Finite steps/instructions
  - When applied, it produces a correct answer
  - The instructions need to be followed
  - In principle, it can be done by a “computer”

Formal definitions

- Computer Code:
  - Human readable text
  - Fully executable description of a software system
- What's in the data?
  - Datum: "(thing) given"
  - Data is useful only when it has been analyzed

Supervised vs unsupervised learning

Machine learning algorithms are generally classified into two categories:

1. **Supervised:** Outcomes used to create the predictor, e.g., classification.
2. **Unsupervised:** Don't know outcomes, rather interestedin discovering groups, e.g., clustering.

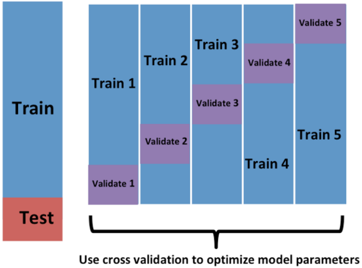
Formal definitions

Training and Test sets: We usually split our dataset into two parts, one for model creation (train) and one for validation (test)

Formal definitions

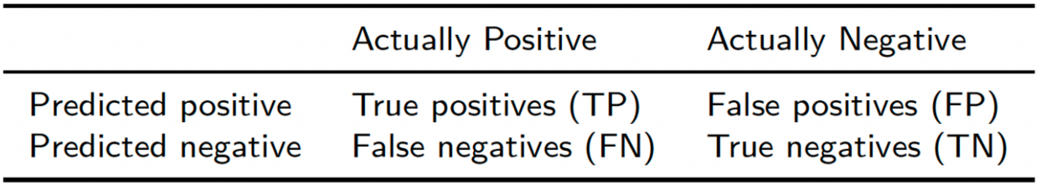
Training and Test sets: We usually split our dataset into two parts, one for model creation (train) and one for validation (test)

Cross-validation: Iterative retraining of the predictor on multiple partitions of training set



Formal definitions

Confusion matrix: tabulates each combination of predicted and actual values:



	Actually Positive	Actually Negative
Predicted positive	True positives (TP)	False positives (FP)
Predicted negative	False negatives (FN)	True negatives (TN)

Overfitting: Dangerously over-optimistic assessments—this is a *big problem* in machine learning

The caret package in R

The `caret` package in R has several useful functions for building and assessing machine learning methods.

- *Examples:*

Regularization in Machine Learning

In regression analysis, the features are estimated using coefficients while modeling. In small sample sizes or noisy data coefficient estimates could be anecdotally incorrect (e.g., overfitting) or innacurate.

If the estimates can be restricted, penalized, or shrunk towards zero, then the impact of insignificant features might be reduced and would prevent models from high variance with a stable fit.[1]

[1] Adapted from: https://www.analyticssteps.com/blogs/l2-and-l1-regularization-machine-learning

Regularization in Machine Learning

Regularization is the most used technique to penalize complex models in machine learning, it is deployed for reducing overfitting (or, contracting generalization errors) by putting small network weights into the model (adding a small amount of biad). Also, it enhances the performance of models for new inputs.[2]

[2] Source: https://www.analyticssteps.com/blogs/l2-and-l1-regularization-machine-learning

Regularization in Machine Learning

Examples of regularization in machine learning, include:

- Regression: Penalizing coefficients to create parsimonious models (variable selection)
- K-means: Restricting the segments for avoiding redundant groups.
- Neural networks: Confining the complexity (weights) of a model.
- Random forests: Reducing the depths of tree and branches (new features)

Kernel-based machine learning

Kernel machines are a class of methods for pattern analysis that use linear classifiers to solve nonlinear problems.

Kernel methods only require a user-specified kernel, i.e., a similarity function over all pairs of data points computed using inner products.

In contrast, many other algorithms require the explicit transformation of the raw data.

Kernel based machine learning

Algorithms that operate with kernels include:

- Kernel perceptron
- Support-vector machines (SVM)
- Gaussian processes
- Principal components analysis (PCA)
- Canonical correlation analysis
- Ridge regression
- Spectral clustering
- Linear adaptive filters

Session info

```r
sessionInfo()
```

```
## R version 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.6.1
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.1
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: America/Denver
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## loaded via a namespace (and not attached):
##  [1] compiler_4.5.1    fastmap_1.2.0     cli_3.6.5         tools_4.5.1      
##  [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29   
##  [9] knitr_1.50        xfun_0.52         digest_0.6.37     rlang_1.1.6      
## [13] evaluate_1.0.4
```