Course Introduction: GSND 5340Q

High Throughput Biomedical Data Analysis (BMDA)

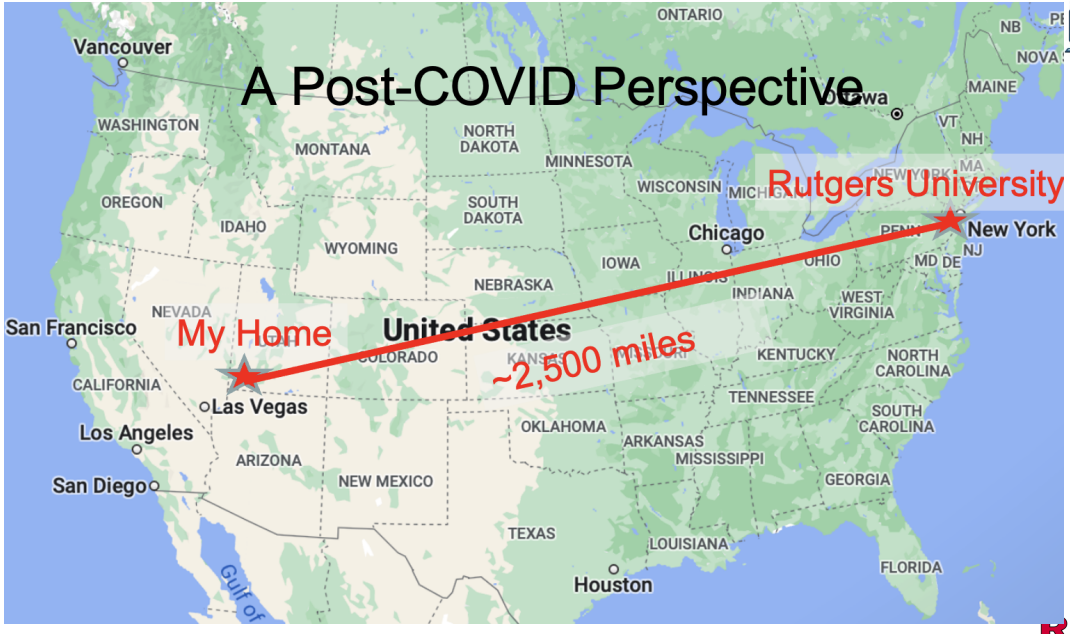
W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School
w.evan.johnson@rutgers.edu

2025-04-22

A Post-COVID Perspective

WEJ LAB

Center for Data Science



WEJ LAB

Center for Data Science

WEJ LAB

Center for Data Science

WEJ LAB

Center for Data Science

Johnson Lab Research

Here is a link to the Johnson Lab Research Page

Center for Data Science Updates: Courses

1. GSND 5345Q: Fundamentals of Data Science (Jan 2025)

- Command-line coding, literate programming, software development, version control, data wrangling and management, and visualization.

2. GSND 5340Q: High Throughput Biomedical Data Analysis (April 2025)

- Sequence alignment/QC, GWAS, gene expression and proteomics, epigenetics, metagenomics, and imaging data analysis.

3. Machine Learning for Biomedical Data (October 2025)

- Model training and validation, regression and regularization, unsupervised learning and clustering, dimension reduction and smoothing, supervised learning and classification, neural networks, and Bayesian learning

Things you should know about BMDA

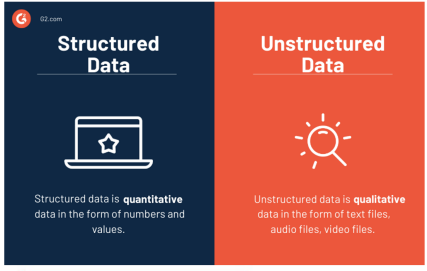
- Click here for the Zoom link
- GitHub vs Canvas:
  - https://github.com/wevanjohnson/2025_Spring_BMDA
- Link to Syllabus
- Background experience
  - Introductory statistics and molecular biology
- Prerequisites (Fundamentals of Data Science)
  - Basic Unix scripting
  - Amarel access and experience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

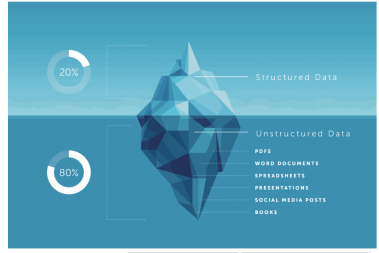
Introduction to Data Science

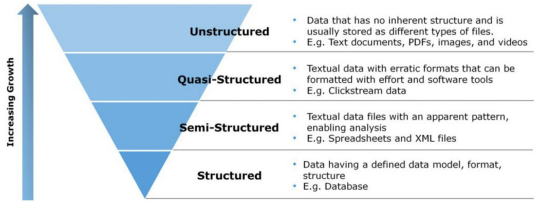
BIG DATA

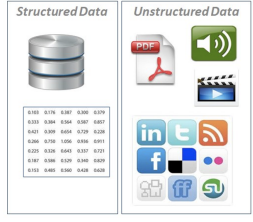
Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data

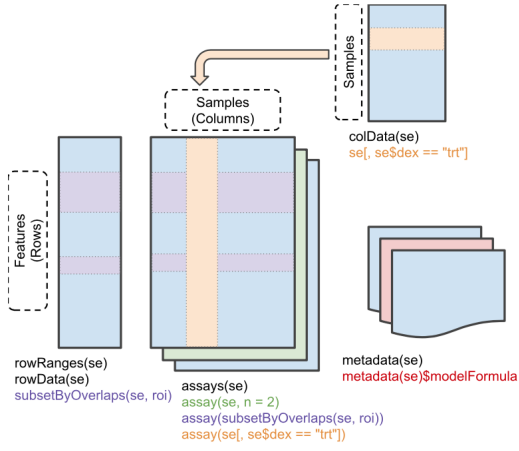




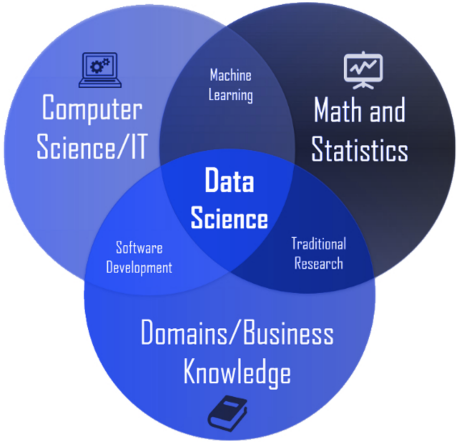




Structured vs. Unstructured data



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

WEJ LAB

Center for Data Science

Data Science Process

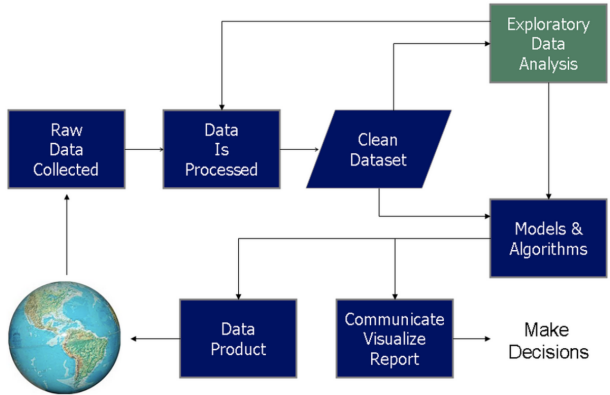
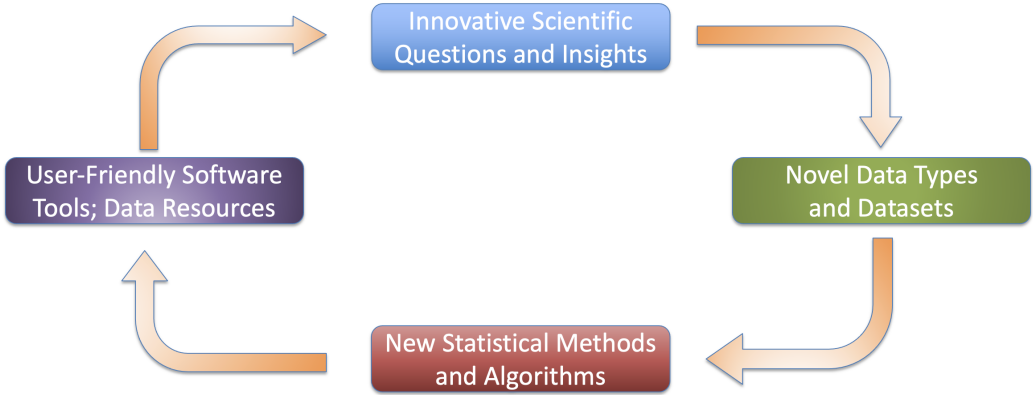


Image: https://en.wikipedia.org/wiki/Data_science

Scientific Cycle for Data Science

Johnson Lab Approach to Science:

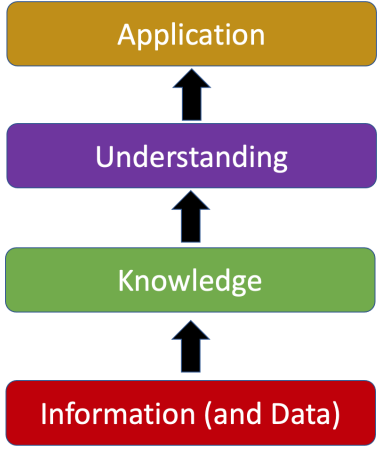


Keeping the “Science” in Data Science

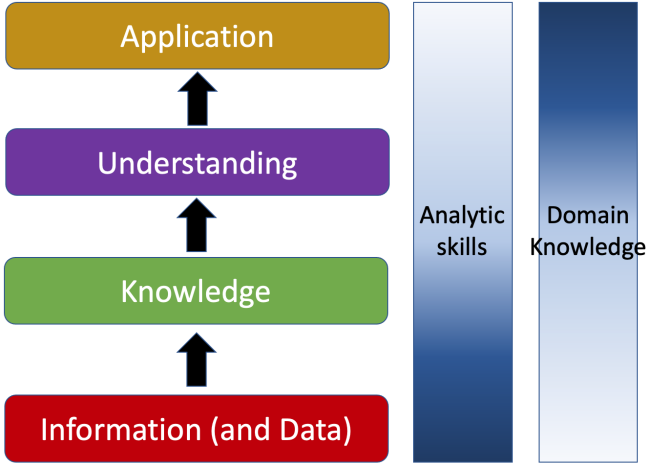
Domain Knowledge

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



Analytics Hierarchy



Session info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.3.2
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## 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/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.4.2    fastmap_1.2.0     cli_3.6.4         tools_4.4.2
##  [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
##  [9] knitr_1.50        xfun_0.51         digest_0.6.37     rlang_1.1.5
## [13] evaluate_1.0.3
```