

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

2024-04-22

Section 1

Course Details

Introductions!



A Post-COVID Perspective





The poster features a yellow header with the text "2024 STUDENT DATA SCIENCE CONFERENCE". Below this is a large red starburst containing the date "21 - 24 MAY" and the location "SOUTHERN UTAH UNIVERSITY". To the left of the starburst is a photograph of a desert landscape with red rock formations. On the right side of the poster, there is a "CO-SPONSORED BY" section featuring the Rutgers Health logo and text about the Rutgers Center for Data Science, Southern Utah University, and Red Rock Data Science. At the bottom, there is a "REGISTRATION:" section with a "DETAILS COMING SOON" button. On the far left, there are three circular icons with corresponding text: "POSTER SESSION", "DATA SCIENCE WORKSHOP", and "NETWORKING", each accompanied by a small photograph.

Goals of the Red Rock Data Science Student Conference

- 1. Student-focused (graduate and undergraduate) conference in Data Science:** a venue for students to present their own research and learn about the scholarly work of others.
- 2. Network with Data Science students at other universities:** The conference will provide workshops, poster sessions, oral presentations, and outdoor excursions for this purpose.
- 3. Connect students with companies and recruiters who are actively hiring graduates with Data Science related degrees.**

Go to: <https://www.redrockdatascience.com>



Go to <https://www.redrockdatascience.com>

Things you should know about this course

- Please fill out the following survey:
<https://forms.gle/XSPjJjMsUa9EWXo2A>
- Zoom:
 - Click here for the Zoom link
- GitHub:
 - https://github.com/wevanjohnson/2024_Spring_BDMA
- Link to Syllabus
- Lots of diverse material
 - Not a spectator sport!

Section 2

Installation Details

Important installations

You will need to install the following:

Mac Users

- R and R Studio
- Know how to access a terminal (Rstudio or Terminal)
- git (type `git --version` in the terminal)

Windows Users:

- R and R Studio
- A terminal app (Git Bash, MobaXterm, Putty)
- Git for Windows

R and Rstudio

See instructions at:

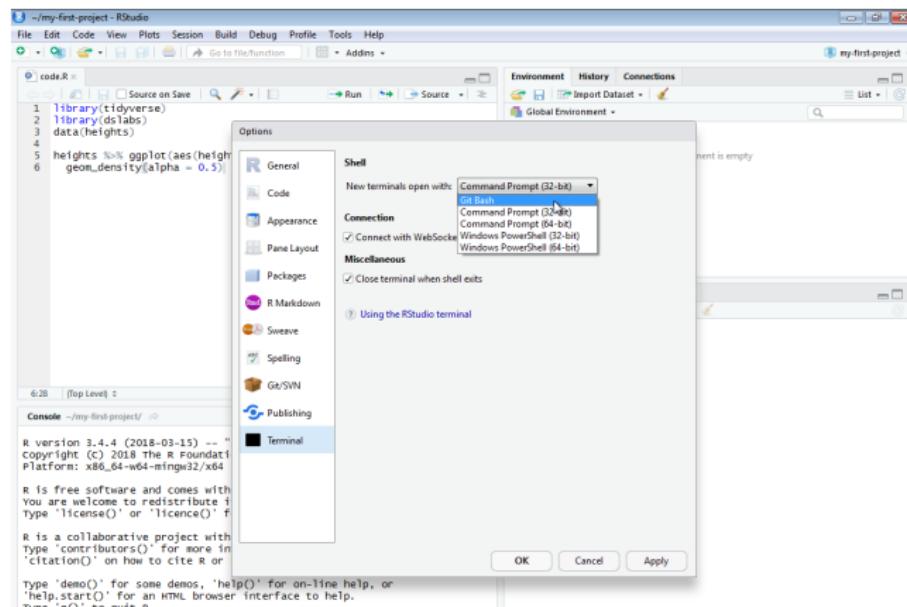
<https://rafalab.github.io/dsbook/installing-r-rstudio.html>

Accessing the terminal and installing Git

See instructions at: <https://rafalab.github.io/dsbook/accessing-the-terminal-and-installing-git.html>

For Windows: link Git Bash and RStudio

We can access the terminal either through RStudio or by opening Git Bash directly. For RStudio, set Git Bash as the default Unix shell: go to preferences (under the File pull down menu), then select Terminal, then select Git Bash:



Section 3

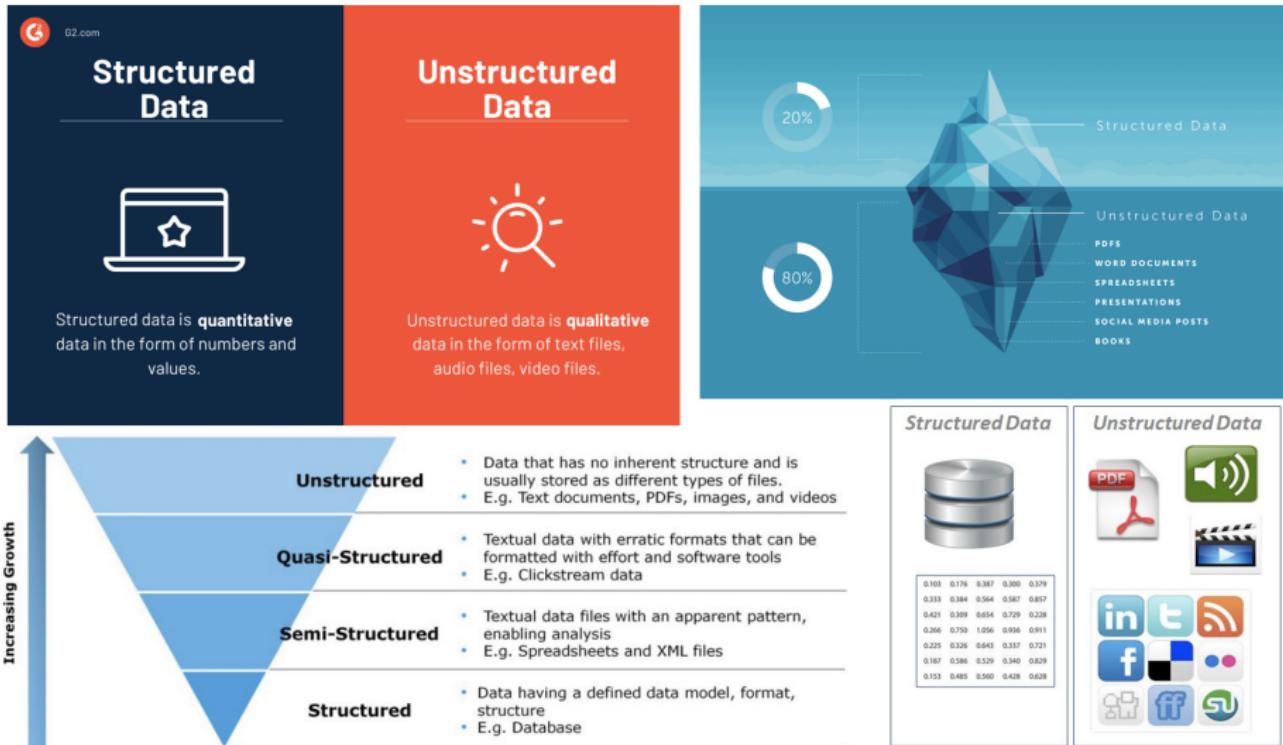
Introduction to Data Science

BIG DATA

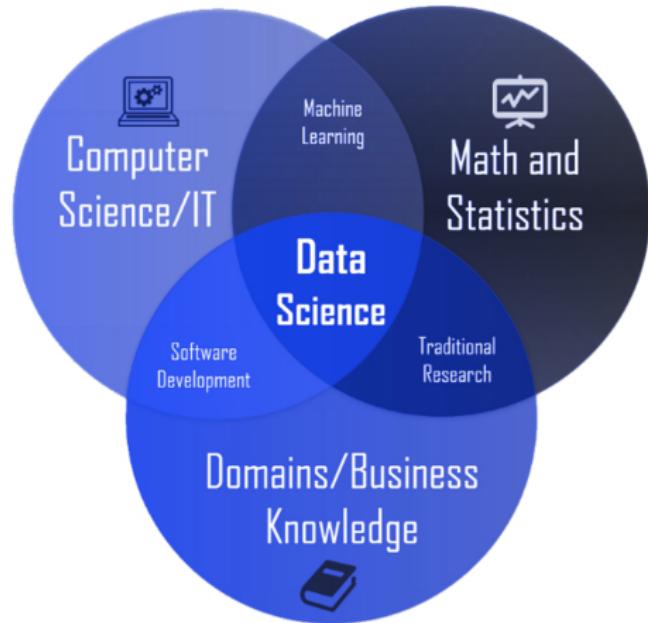


Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data



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 21st 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

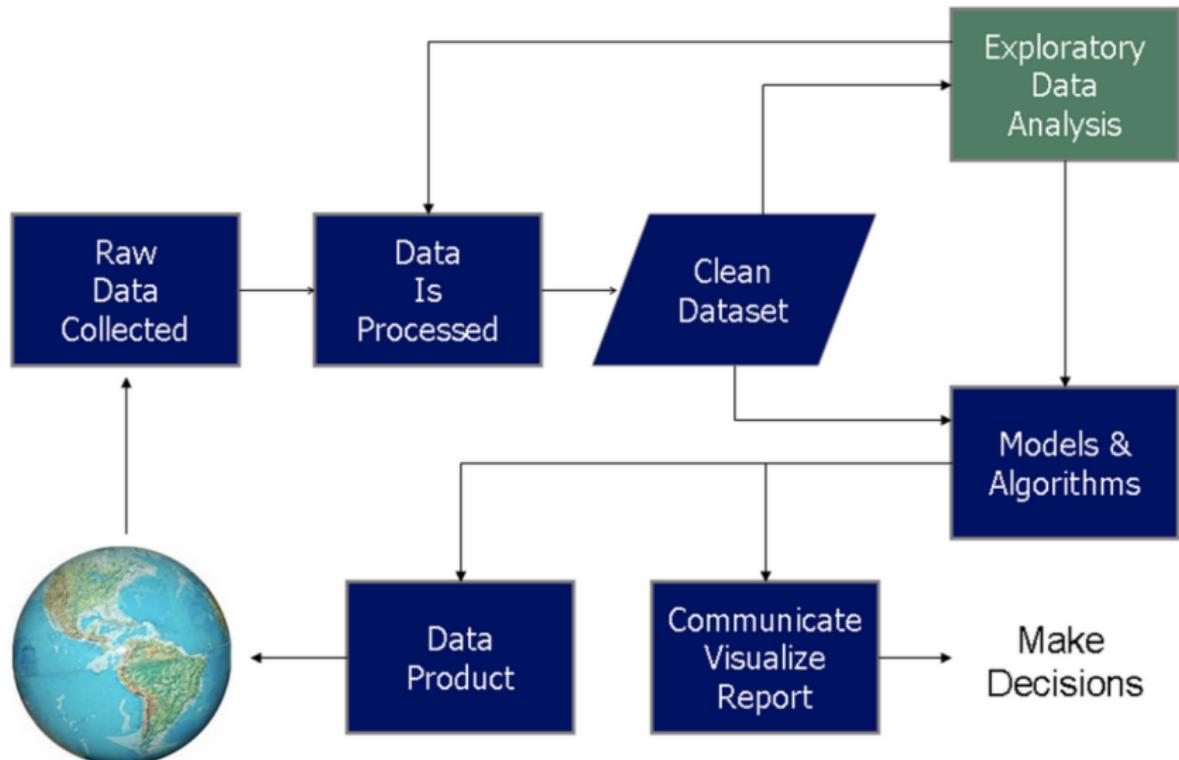
- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative



COMMUNICATION & VISUALIZATION

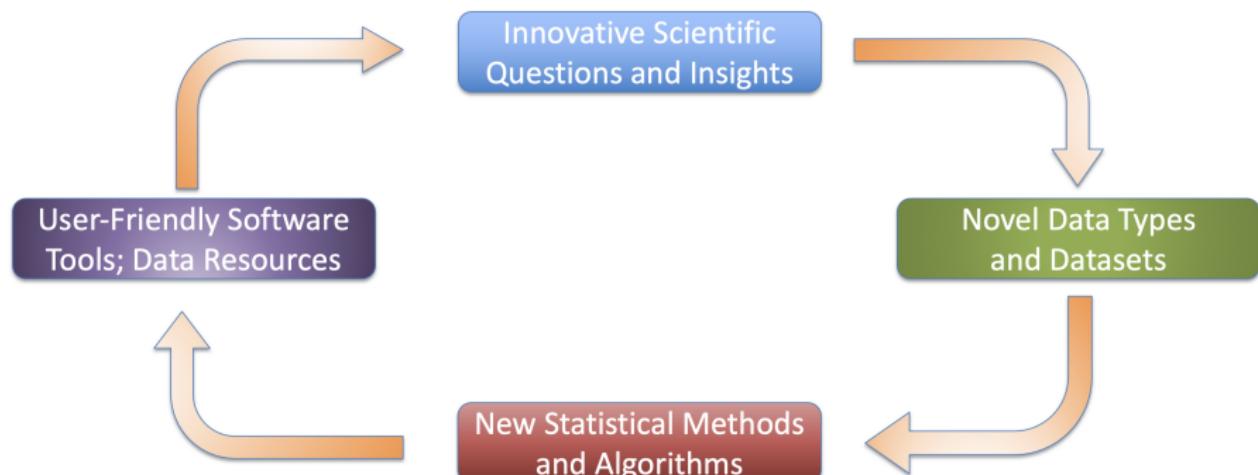
- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Science Process



Scientific Cycle for Data Science

Johnson Lab Approach to Science:



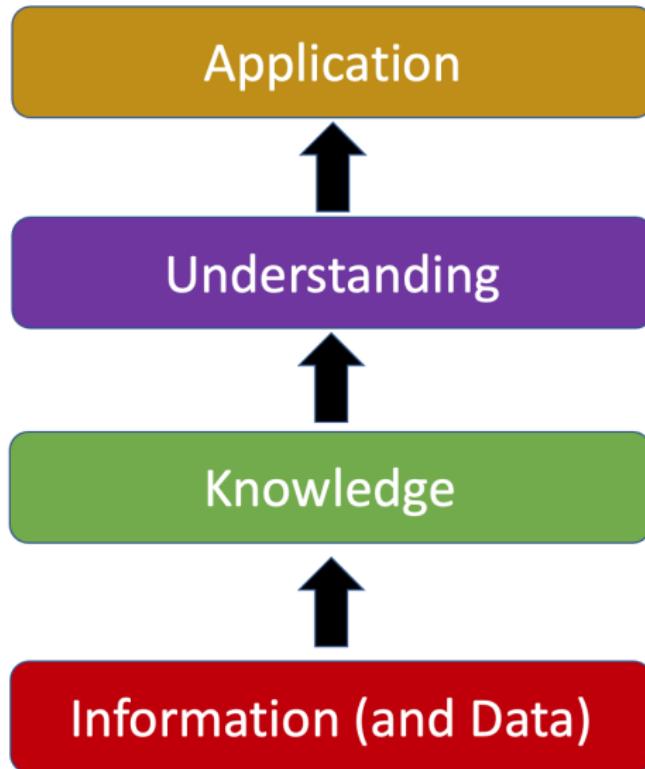
Section 4

Keeping the “Science” in Data Science

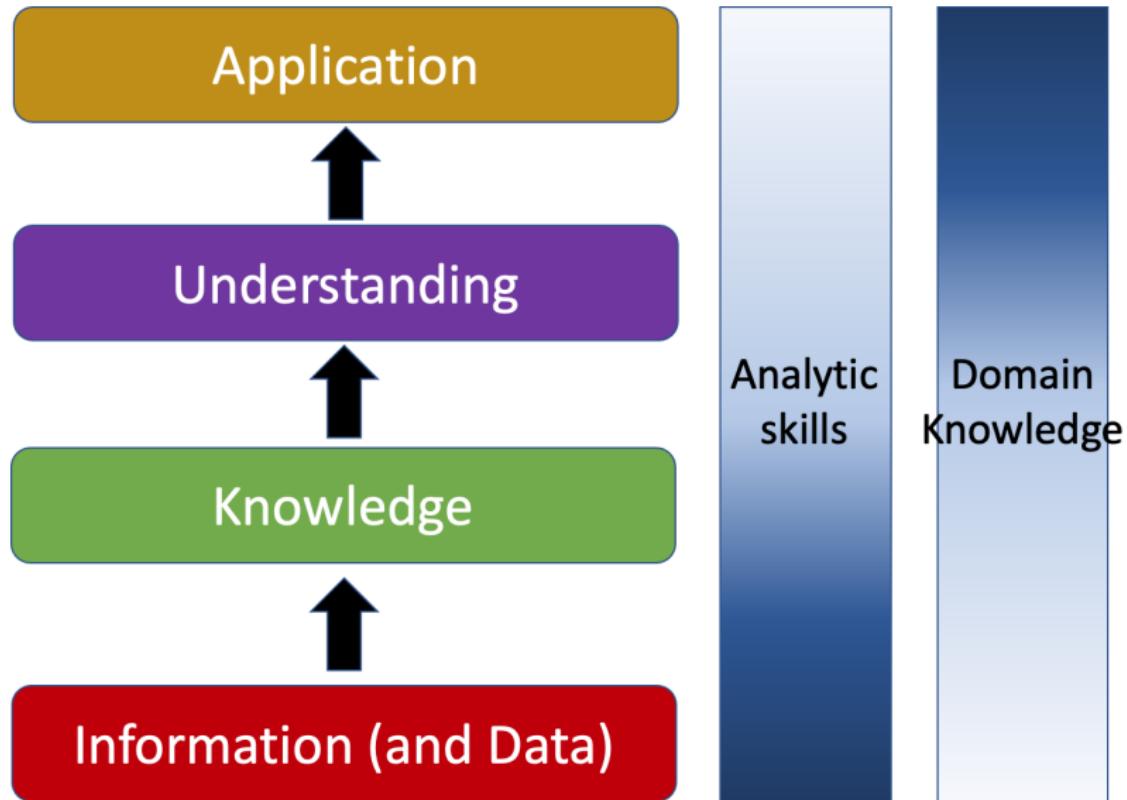
Domain Knowledge

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 4.3.2 (2023-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3
##
## 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_4.3.2    fastmap_1.1.1    cli_3.6.2       tools_4.3.2
## [5] htmltools_0.5.8   rstudioapi_0.16.0 yaml_2.3.8     rmarkdown_2.26
## [9] knitr_1.45        xfun_0.43       digest_0.6.35   rlang_1.1.3
## [13] evaluate_0.23
```