Introduction to Data Science

Data Science Lecture Series

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

2023-03-13

Section 1

Course Details

Introductions!

Things you should know about this course

- Lots of diverse material
  - Not a spectator sport!
- Zoom:
  - https://rutgers.zoom.us/j/97374683455?pwd=cHJWaE92eTIHVGRUYTNaUkVqNkJhZz09
- Johnson Lab Slack:
  - #data-science-learning channel
  - Contact Brie Odom-Mabey to get access: aodom@bu.edu
- GitHub:
  - https://github.com/wevanjohnson/DataScienceLecturesSpring2023

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

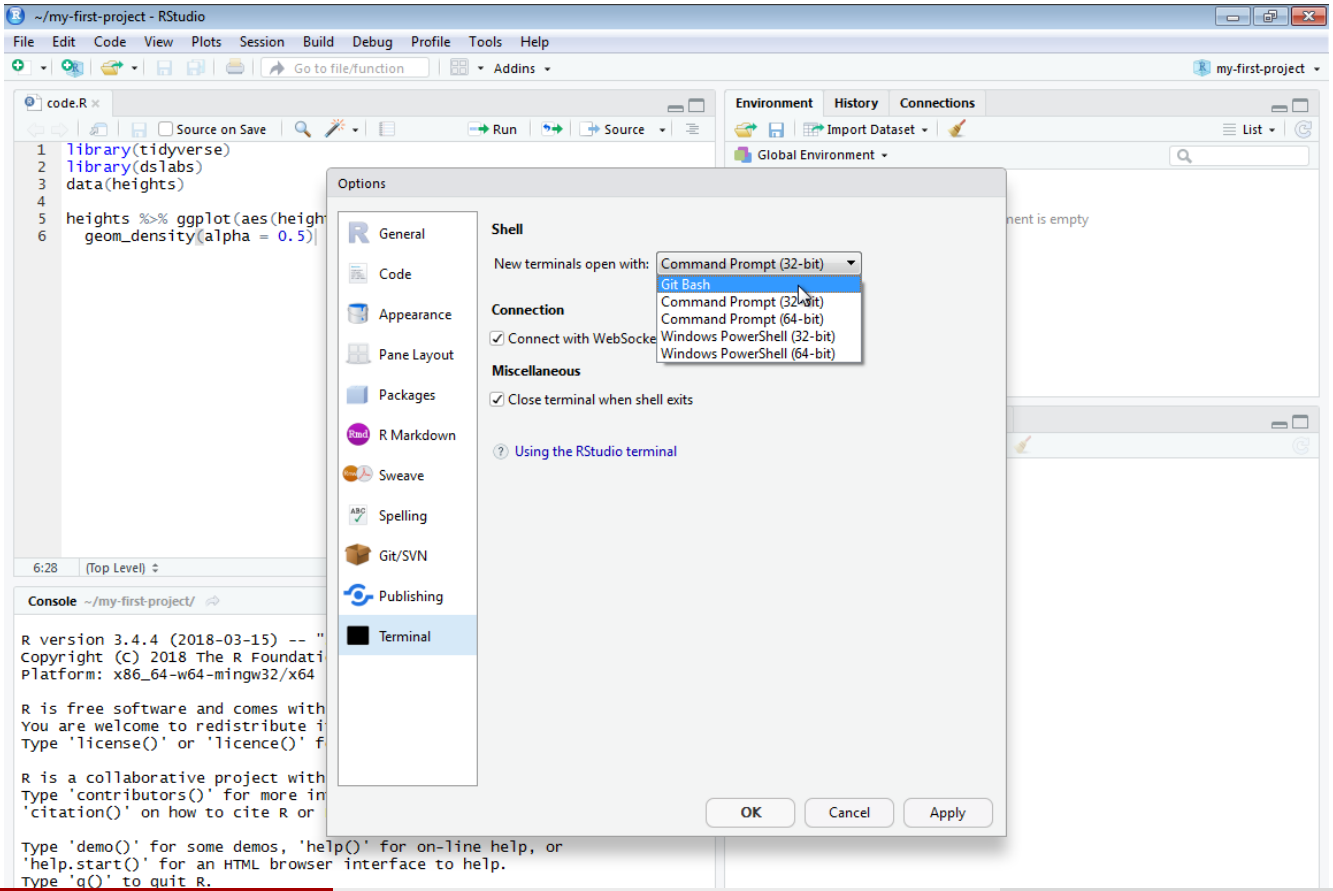
See instructions at:
https://rafalab.github.io/dsbook/installing-r-rstudio.html

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:

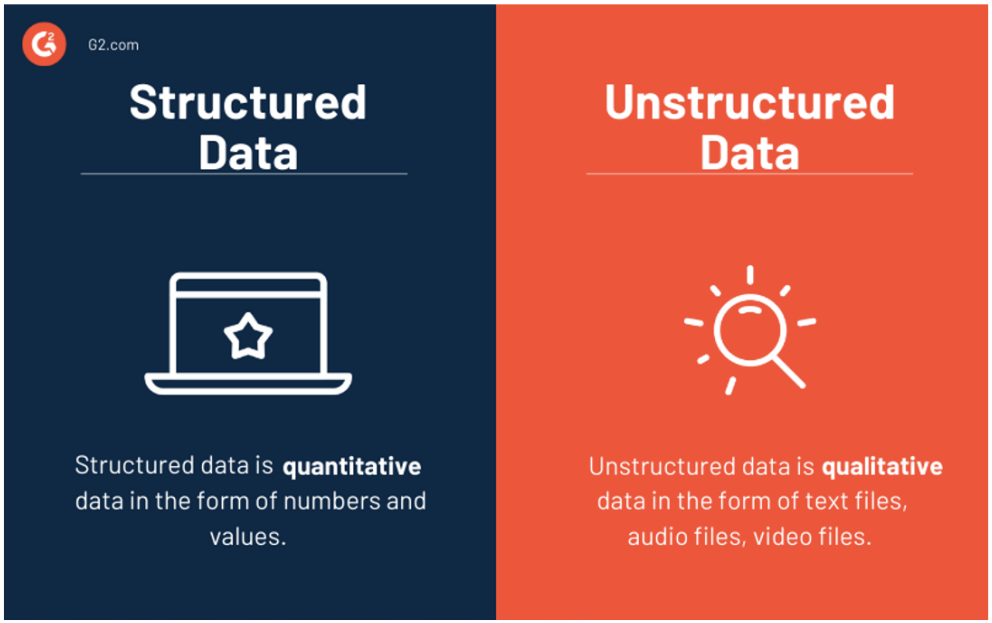


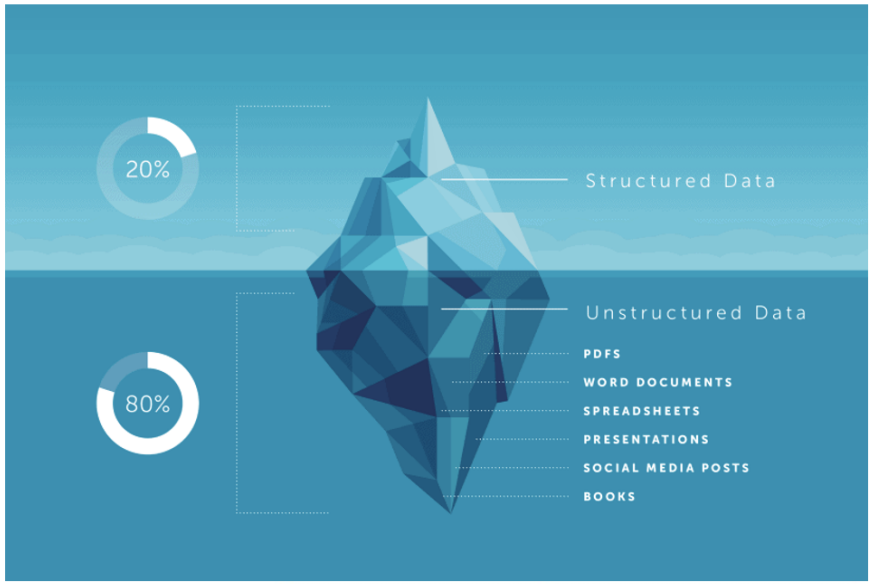
Section 3

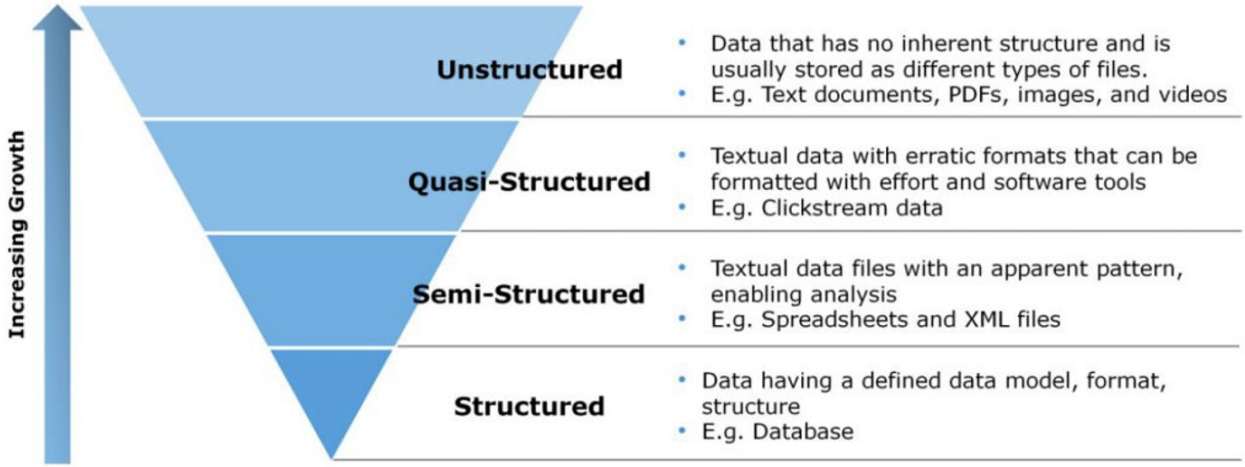
Introduction to Data Science

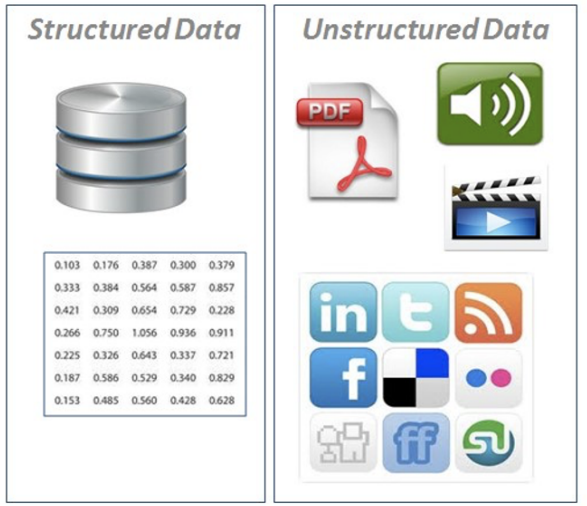
Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data

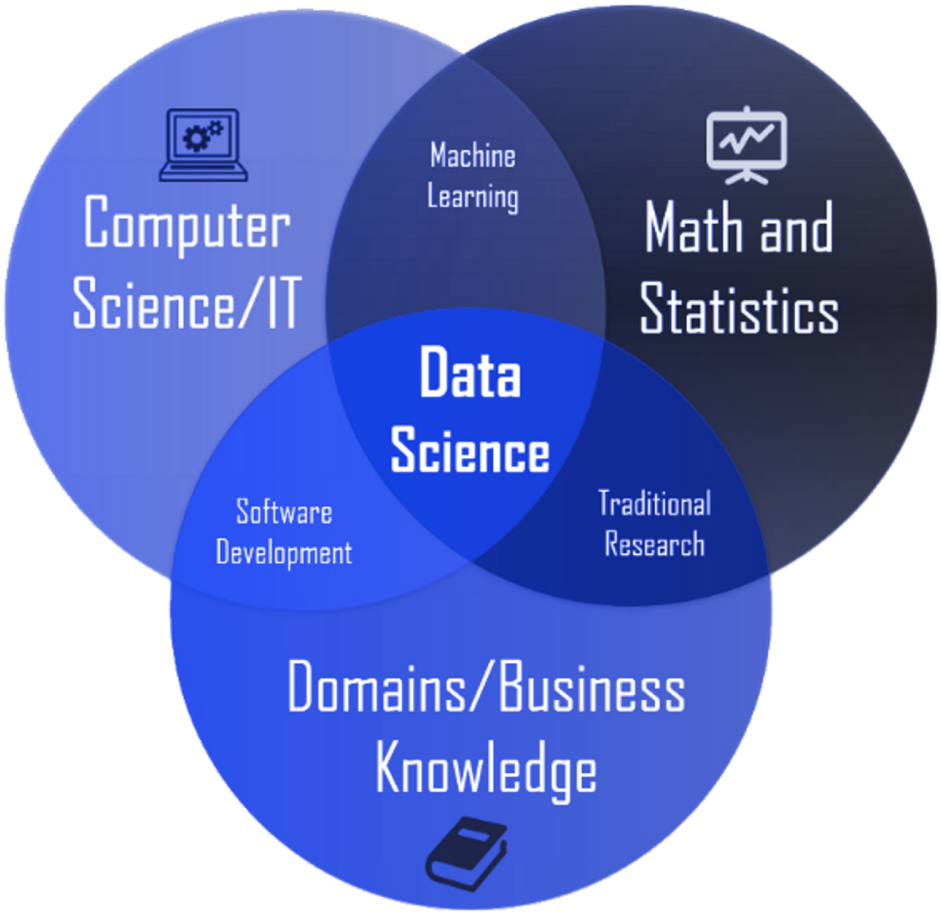








Data Science Revolution



- 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 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

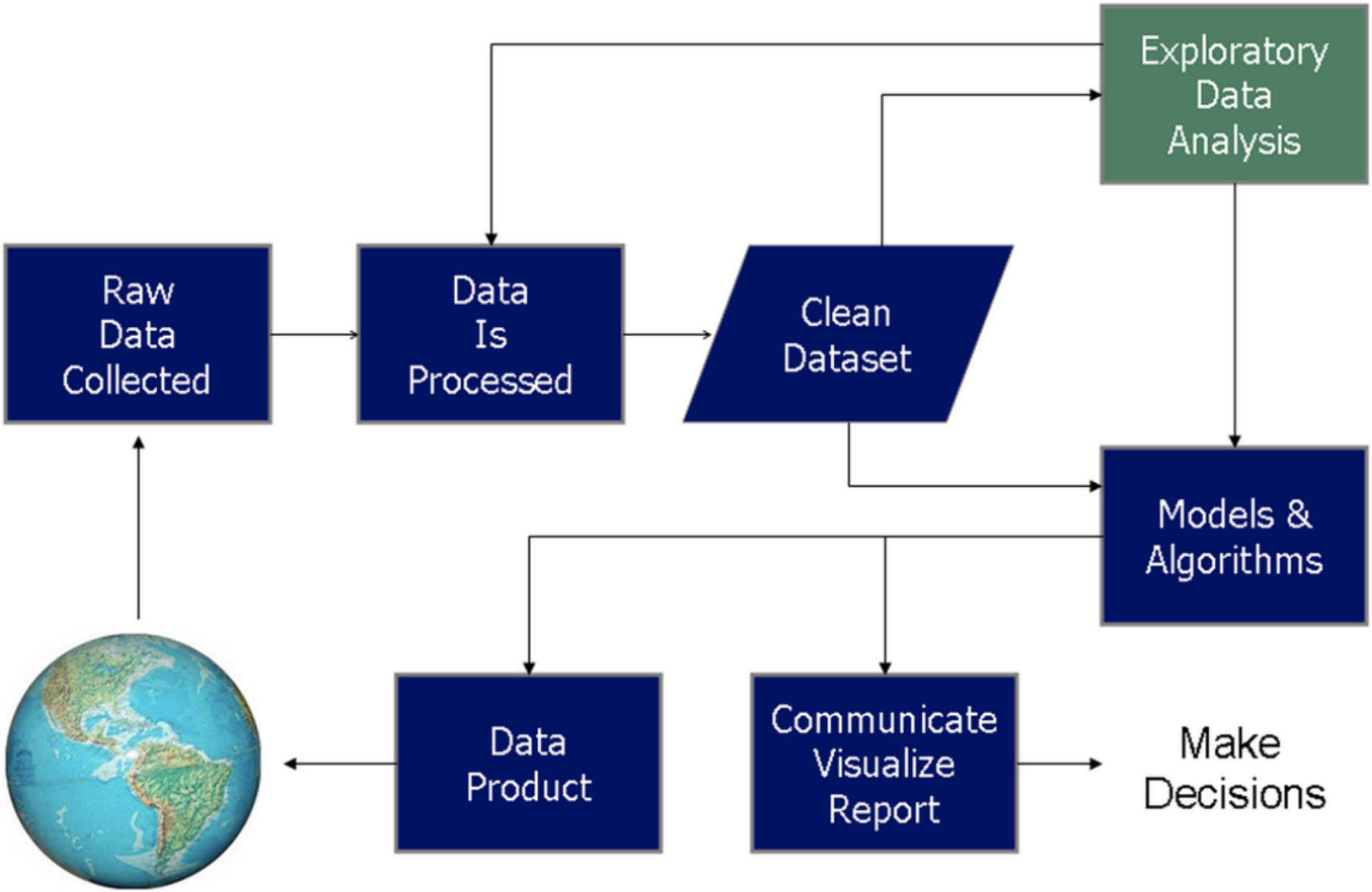
DOMAIN KNOWLEDGE & SOFT SKILLS

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

COMMUNICATION & VISUALIZATION

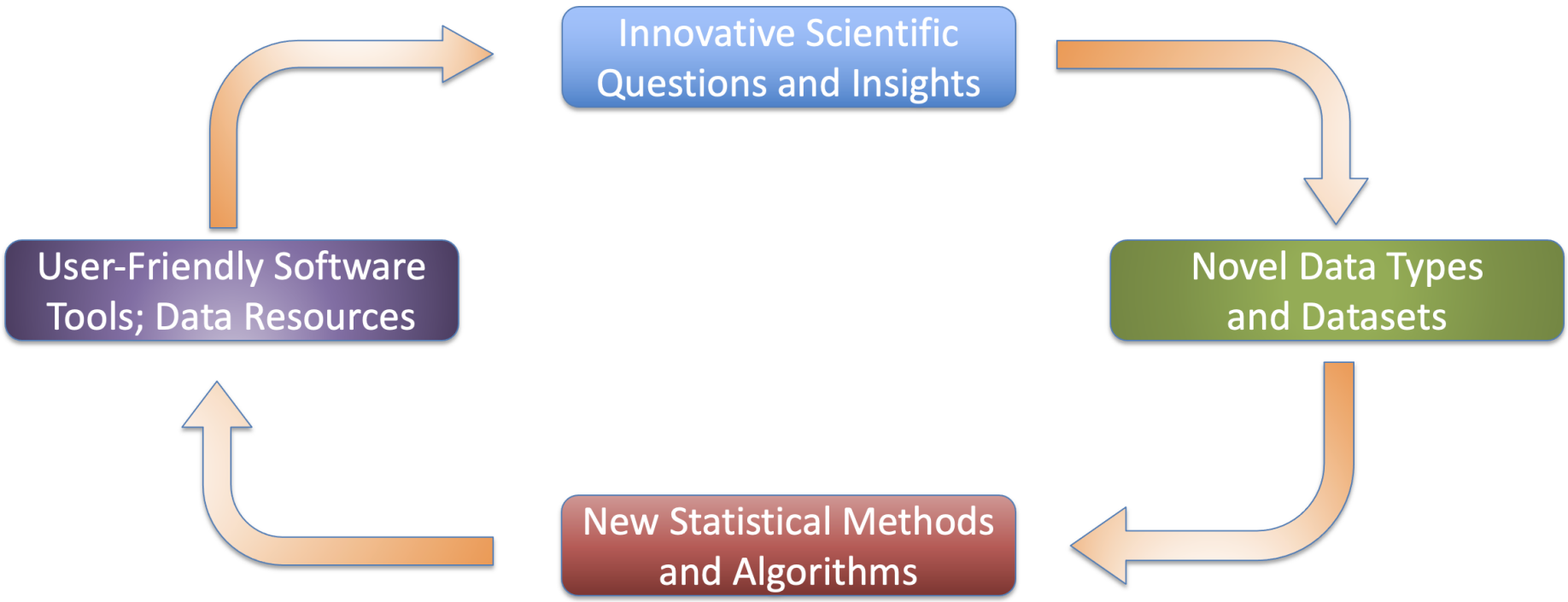
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization

Data Science Process



Scientific Cycle for Data Science

Johnson Lab Approach to Science:



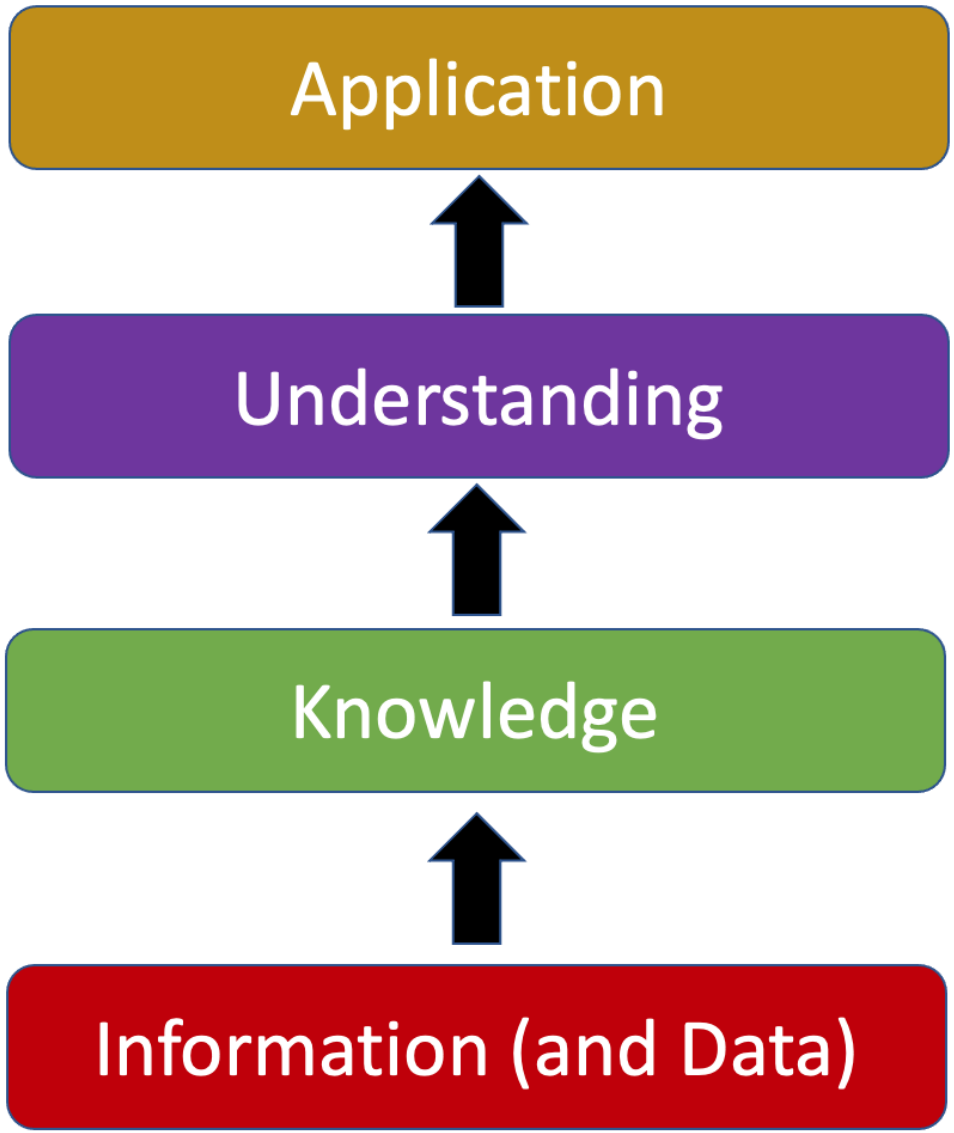
Section 4

Keeping the "Science" in Data Science

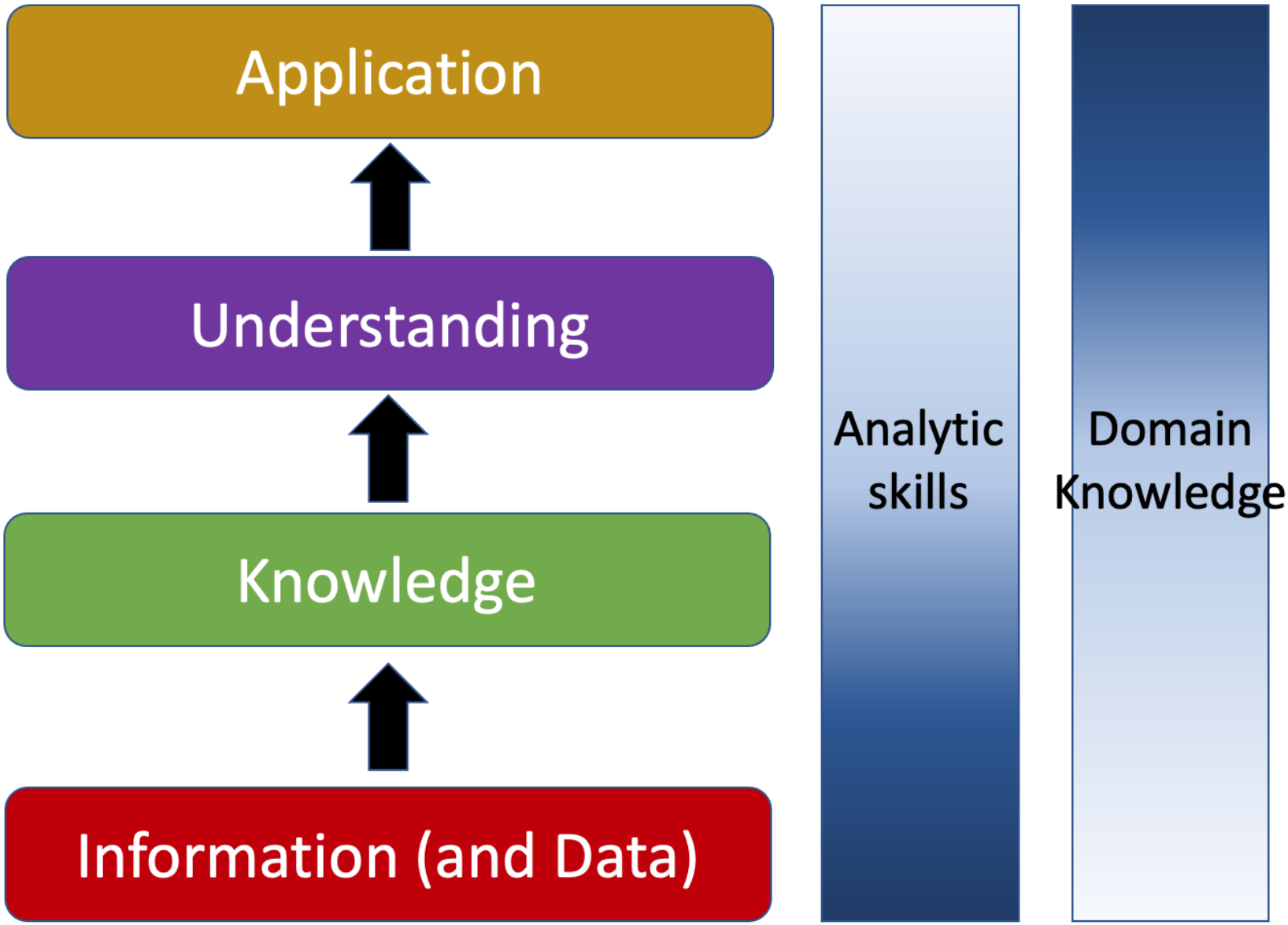
Domain Knowledge

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 4.2.2 (2022-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.31   lifecycle_1.0.3 magrittr_2.0.3  evaluate_0.19
##  [5] rlang_1.0.6     stringi_1.7.8   cli_3.5.0       rstudioapi_0.14
##  [9] vctrs_0.5.2     rmarkdown_2.19  tools_4.2.2     stringr_1.5.0
## [13] glue_1.6.2      xfun_0.36       yaml_2.3.6      fastmap_1.1.0
## [17] compiler_4.2.2  htmltools_0.5.4 knitr_1.41
```