Course Introduction: GSND 5345Q

Fundamentals of Data Science (FDS)

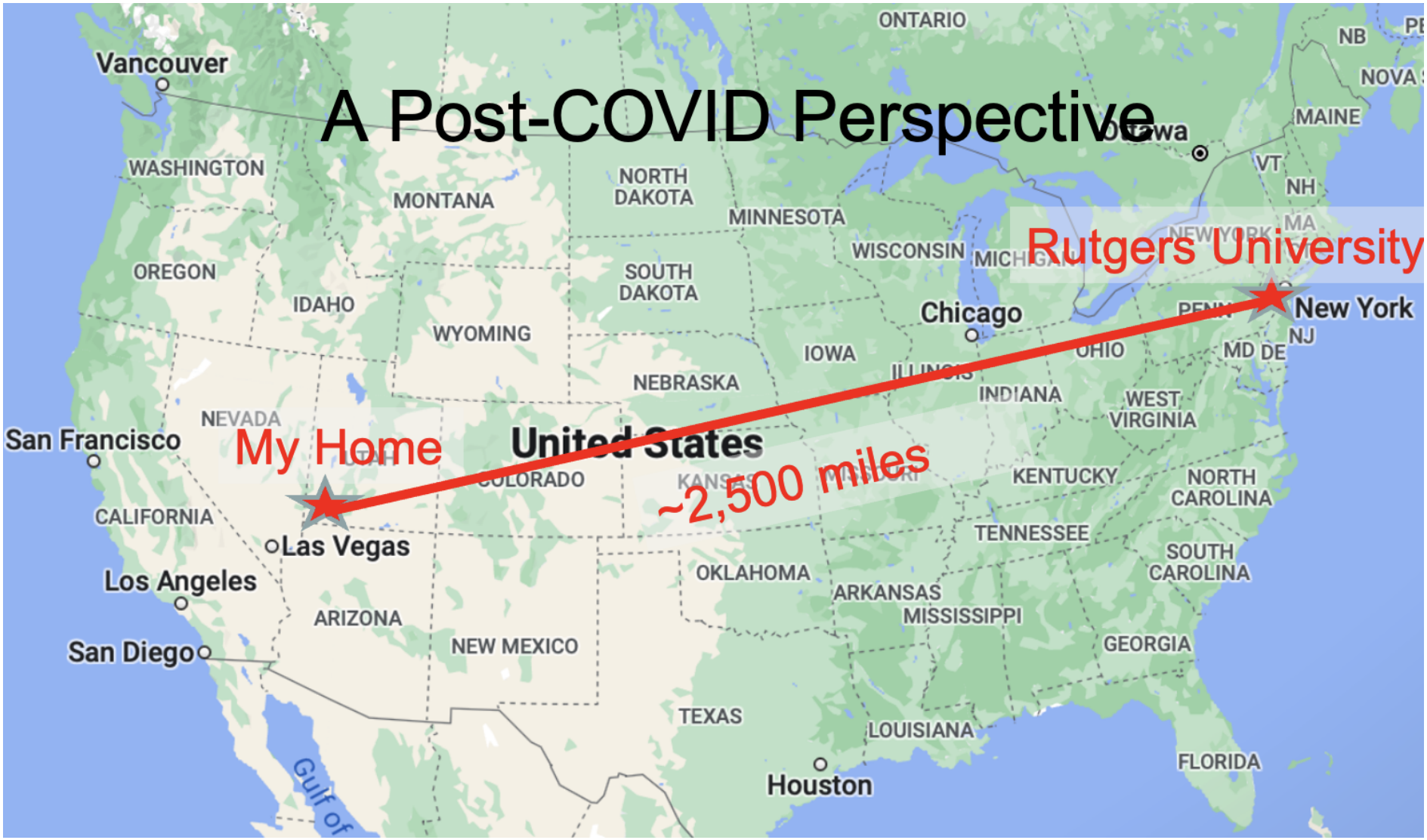
W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School
w.evan.johnson@rutgers.edu

2025-01-06

Section 1

Introductions


A Post-COVID Perspective


A Post-COVID Perspective
Vancouver
ONTARIO
NB
NOVA
MAINE
WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
VT
NH
MA
Rutgers University
New York
NJ
OREGON
IDAHO
SOUTH DAKOTA
WISCONSIN
Chicago
WYOMING
IOWA
OHIO
MD
DE
NEBRASKA
INDIANA
WEST VIRGINIA
NEVADA
San Francisco
My Home
United States
~2,500 miles
KENTUCKY
NORTH CAROLINA
CALIFORNIA
Las Vegas
TENNESSEE
SOUTH CAROLINA
Los Angeles
OKLAHOMA
ARKANSAS
MISSISSIPPI
ARIZONA
San Diego
NEW MEXICO
GEORGIA
TEXAS
LOUISIANA
FLORIDA
Houston
Gulf of

Johnson Lab Research

Here is a link to the Johnson Lab Research Page

Section 2

Course Introduction

Things you should know about this course

- Lots of diverse material
  - Not a spectator sport!
- Zoom Meeting ID for all sessions is 95398689633, passcode: 065918:
  - Click here for the direct Zoom link
  - Lectures will be recorded and posted in the "Announcements" (Canvas)
- Canvas:
  - The Canvas page will be limited confidential items: course announcements, communication, homework submissions, grades, etc.
- GitHub: https://github.com/wevanjohnson/2025_Spring_FDS
  - Course information, schedule, lecture notes, homework, etc.
  - Link to Syllabus
- You need to have a basic understanding of statistics
- Learning to program in R is a requirement of this course.

Resources for learning statistics

Here are some resources to learn basic statistics (and in some cases R simultaneously):

- Data Analysis with R Specialization (Coursera/Duke University)
- Introduction to statistics (Coursera/Stanford)

Resources for learning R

For learning R:

- RStudio Education
- R Programming (Coursera/Johns Hopkins)
- Data Science R Basics (edx/Harvard University)
- R Training Course (LinkedIn)
- R Programming A - Z: R for Data Science (Udemy)
- Programming with R (Pluralsight)

Dr. Johnson's R Tutorials (GitHub and YouTube)

Dr. Johnson will provide an online R tutorial on GitHub:
https://github.com/wevanjohnson/2025_01_R_tutorial

R Tutorials (GitHub and YouTube)

Lectures 1-6 will give you the R basics you need for the course. Lectures 7-12 will have some overlap with this course (Weeks 5-7).

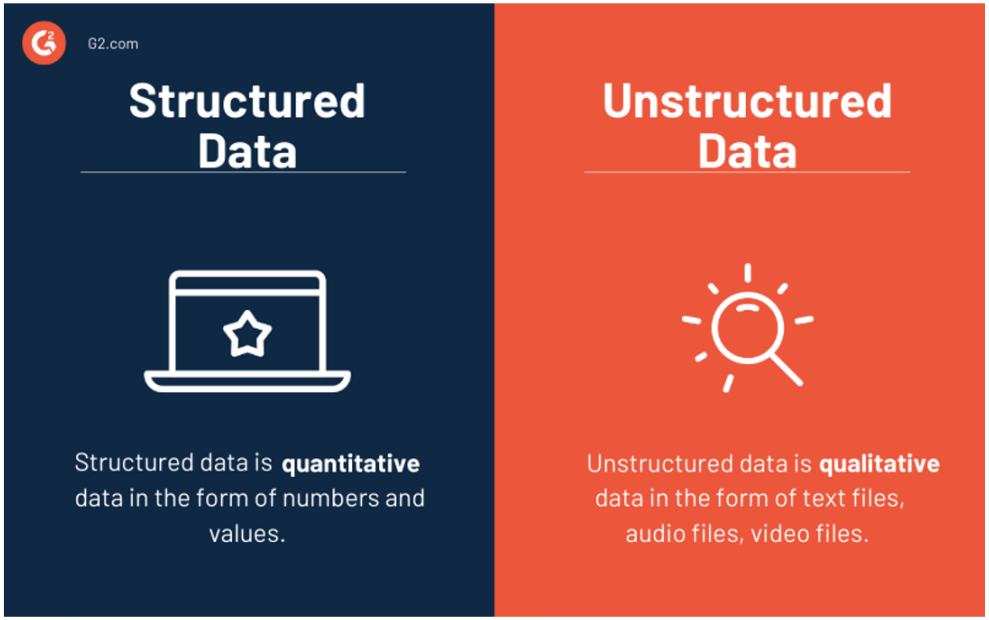
Lecture	Topics
Lecture 1	Installing R, RStudio, and R packages
Lecture 2	Objects, workspace, functions, scripts
Lecture 3	Vectors, factors, lists, data frames
Lecture 4	Sorting, arithmetic, basic plots
Lecture 5	Logic and Conditionals
Lecture 6	Loops, vectorization, functions
Lecture 7	Literate programming with R Markdown
Lecture 8	Input/output data, Data structures
Lecture 9	The tidyverse
Lecture 10	Visualization with ggplot2
Lecture 11	Creating R Packages
Lecture 12	Shiny Programming

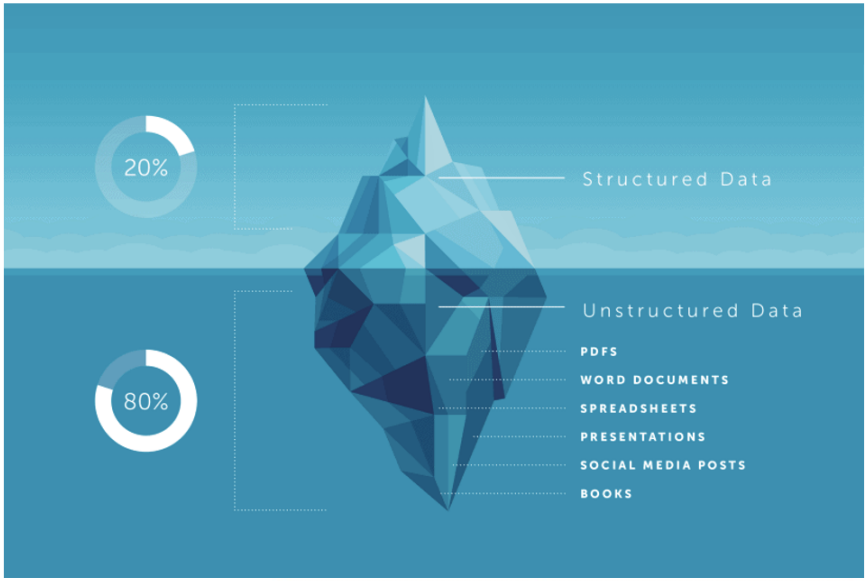
Section 3

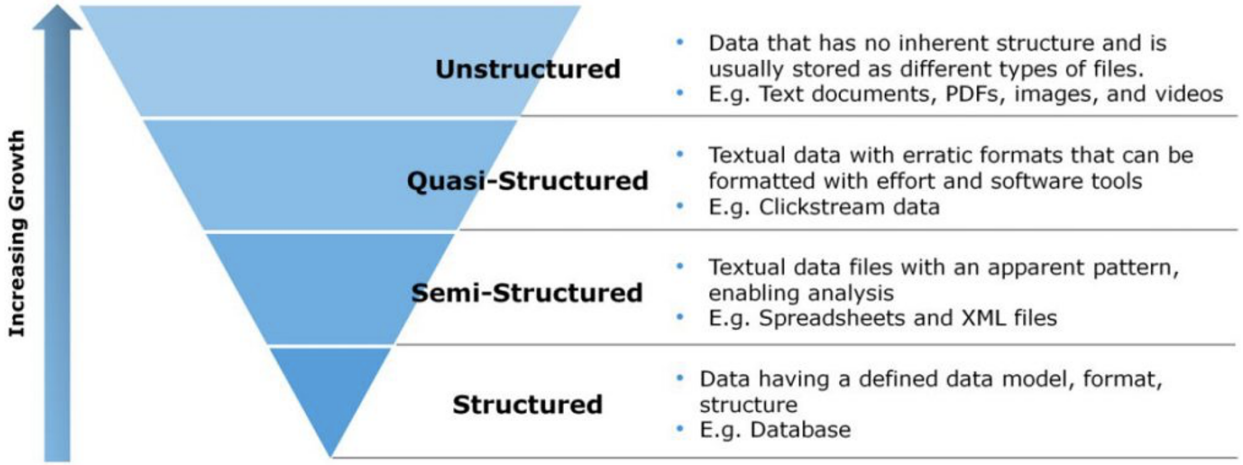
Introduction to Data Science

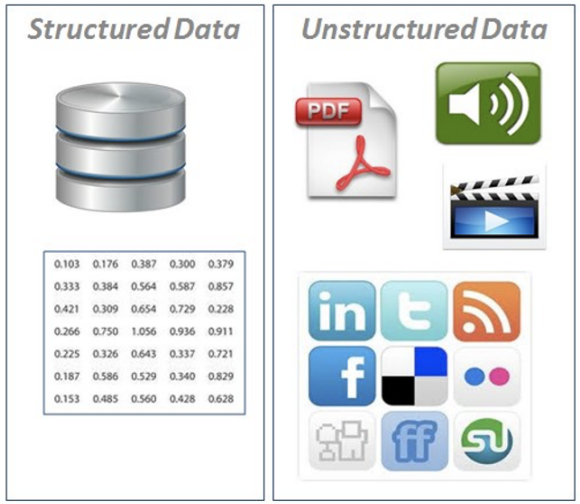
Big Data has fundamentally changed how we look at science and business. Along with advances in analytic methods, they are providing unparalleled insights into our physical world and society

Structured vs. Unstructured data

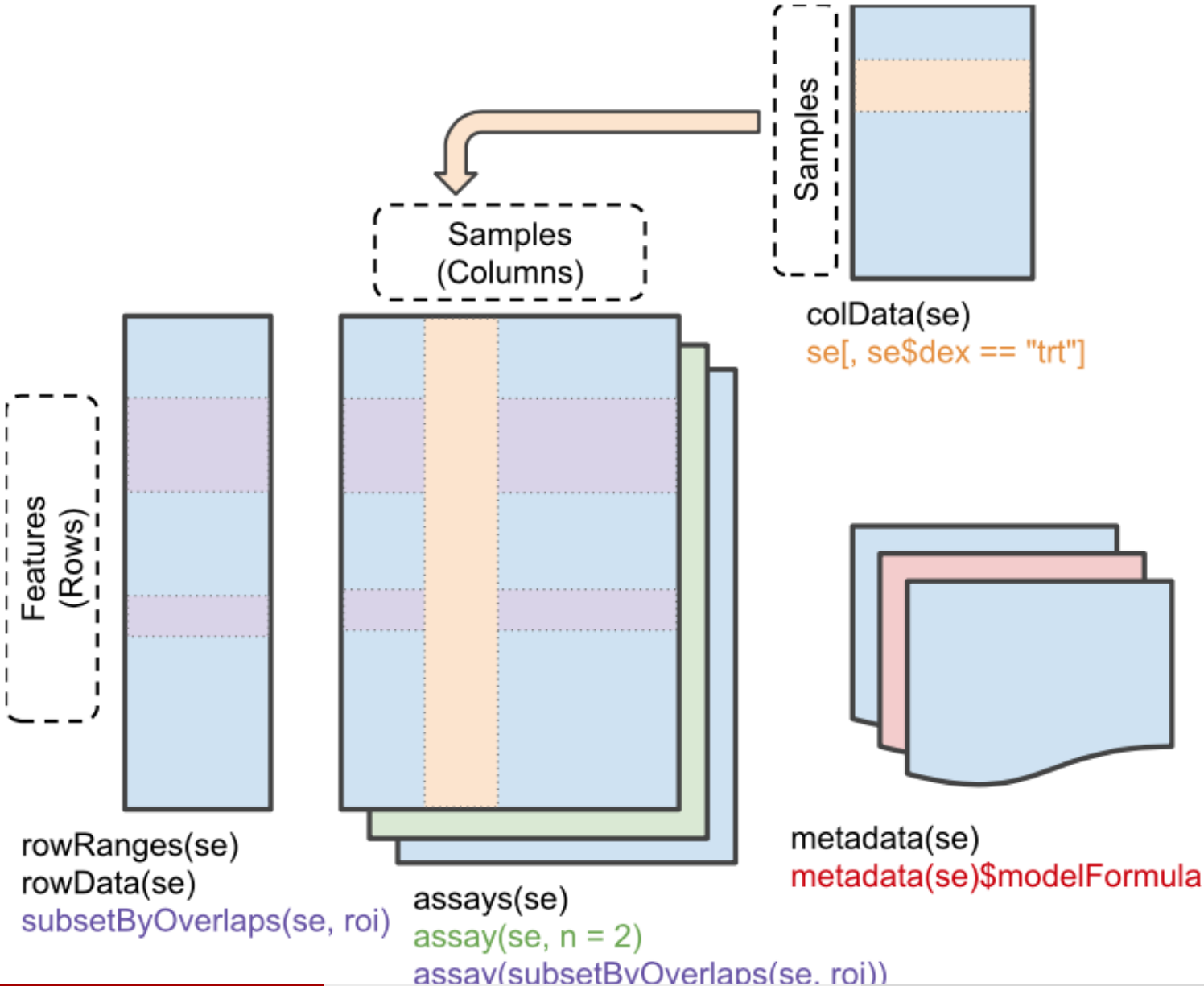




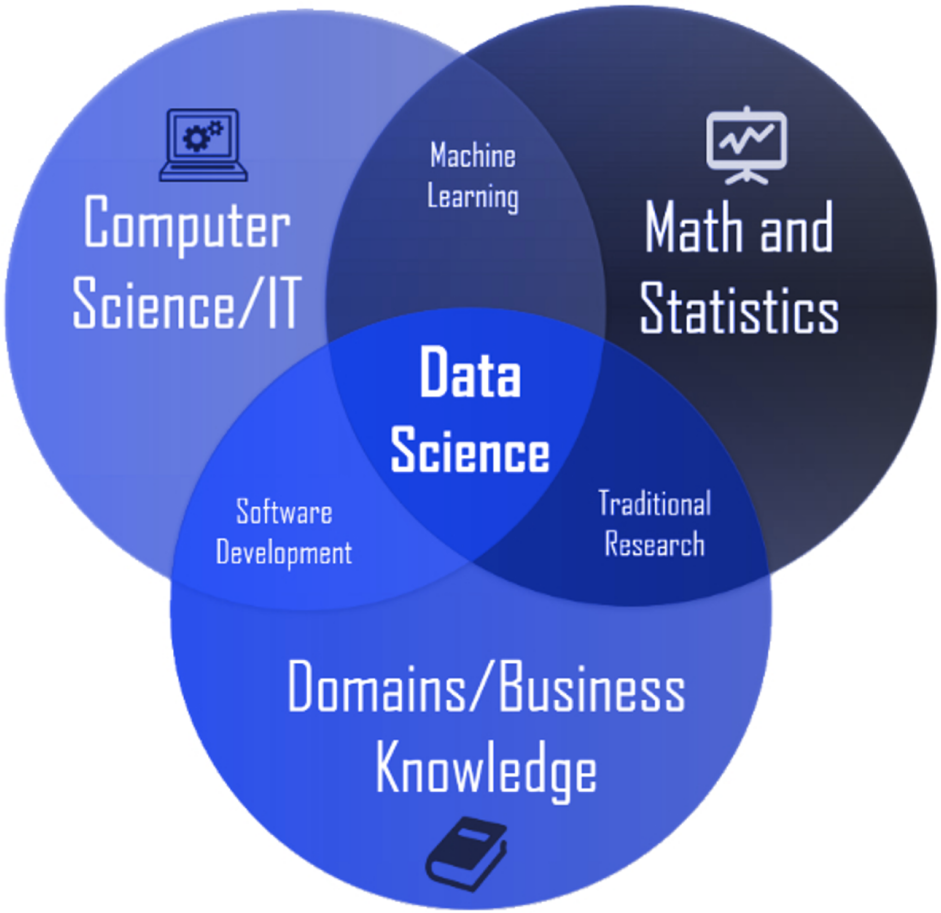




Structured vs. Unstructured data



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

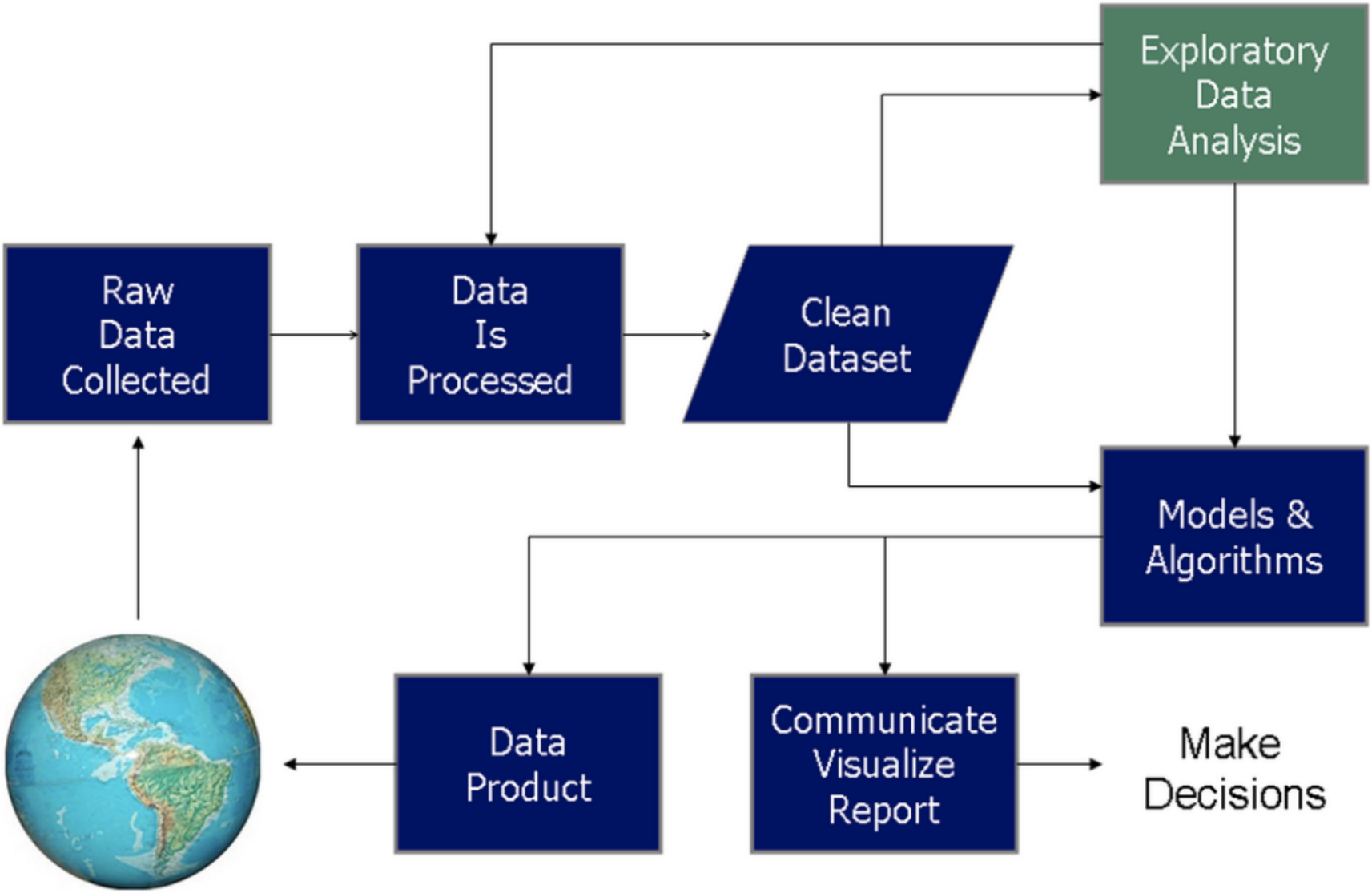
DOMAIN KNOWLEDGE & SOFT SKILLS

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

COMMUNICATION & VISUALIZATION

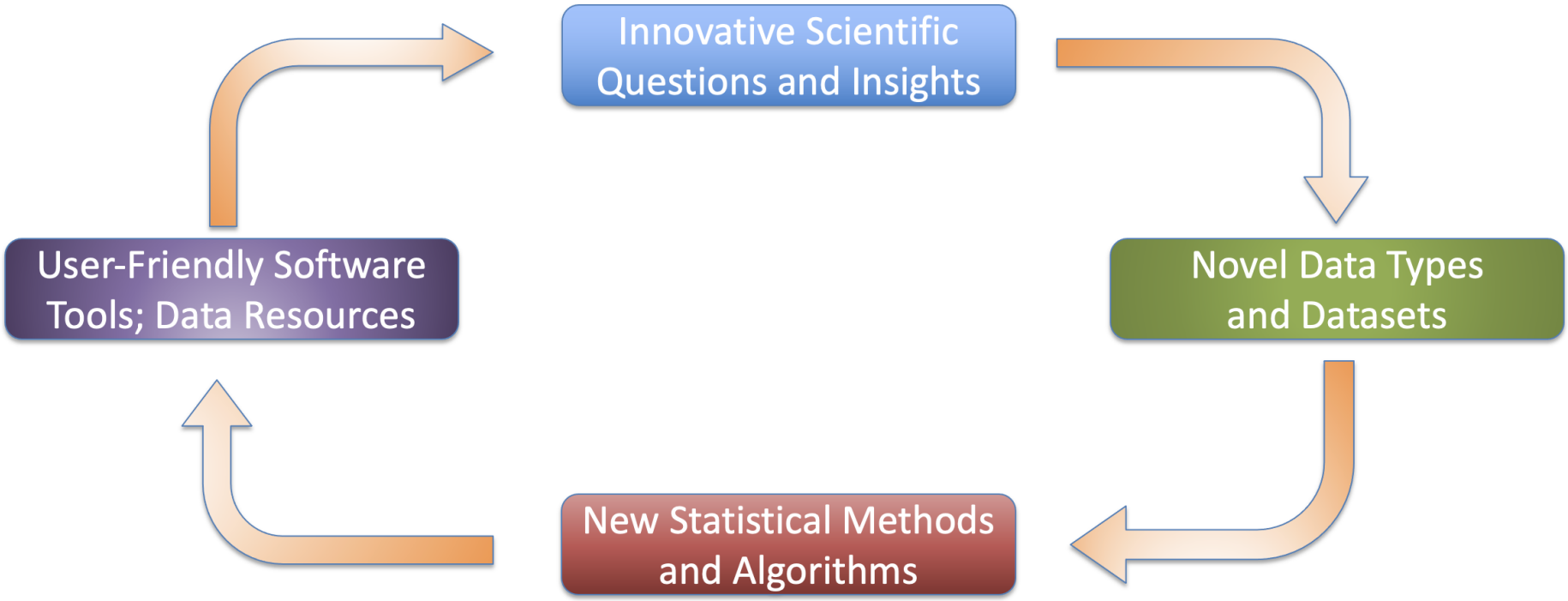
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Data Science Process



Scientific Cycle for Data Science

Johnson Lab Approach to Science:



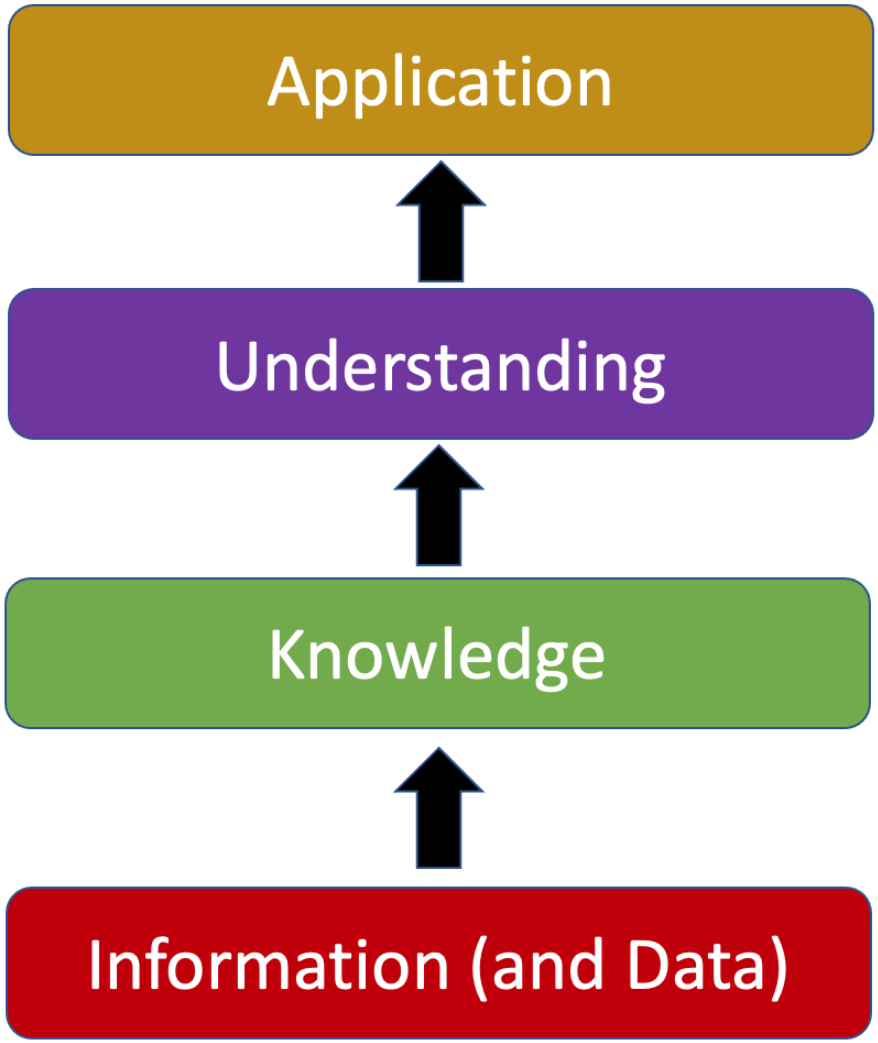
Section 4

Keeping the “Science” in Data Science

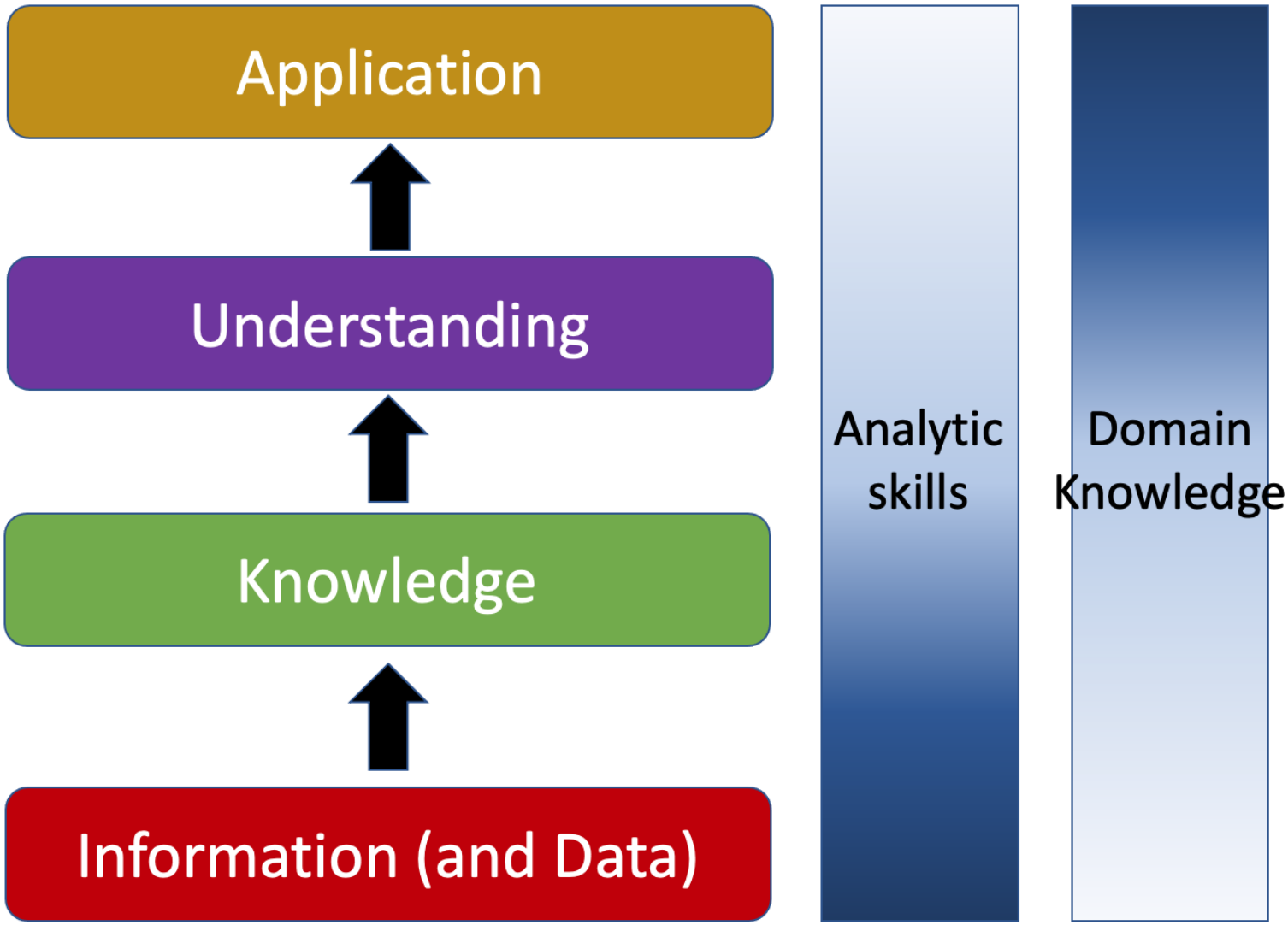
Domain Knowledge

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 4.4.2 (2024-10-31)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.2.1
##
## 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
##
## 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_4.4.2    fastmap_1.2.0     cli_3.6.3         tools_4.4.2
##  [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.10       rmarkdown_2.28
##  [9] knitr_1.48        xfun_0.47         digest_0.6.37     rlang_1.1.4
## [13] evaluate_1.0.0
```