INTRODUCTION

Applied Analytics: Frameworks and Methods 1

No Phone
No Photographing
No Audio Recording Except by Instructor
No Video Recording Except by Instructor

Lecture Content and Materials (including Audio and Video) Should Not Be Posted or Shared Online or Offline Without Explicit Permission

Outline

- Introduction
- Forces Shaping the Growth of Analytics
- Domain of Analytics and its Impact
- Review the Course Structure
- Overview of R

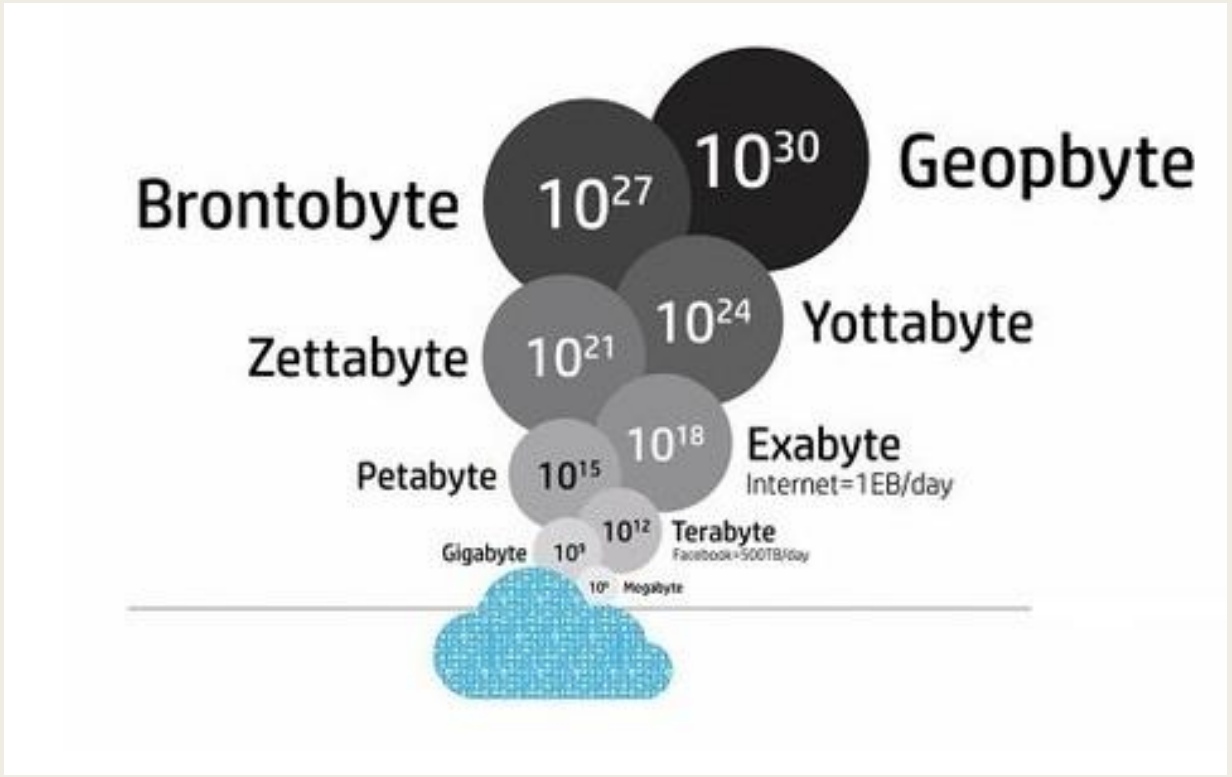
FORCES SHAPING THE GROWTH OF ANALYTICS

Fundamental Forces Changing Everything

- Data
- Analytical techniques
- Software
- Hardware

Fundamental Forces Changing ….. …. *Everything*

Data, Data, Data



Data

- Traditional data: E.g., transactional data and survey data
- People generating Data
  - *Email, Social Media, …*
  - *Personal devices such as Fitbit, smartphone, Nest, iRobot, Nike*
- Other Devices
  - *IoT: On board computers on machines from cars to airplanes*
- Democratization of data
  - *Open Data revolution is afoot led by governments. Data.gov shares over 315,000 datasets (as of 2021)*
  - *Websites share data for free directly from website or through an API. E.g., Google Trends, Yahoo Finance, Twitter*

Data

- Governments and companies have made it easier for people to share data with them.
  - *Software and hardware diagnostics*
  - *Reporting potholes, graffiti, crime*
  - *Service complaints*
- Harvesting of data
  - *Previously unused data is being captured. E.g., Electronic Health Records*
  - *Speech, pictures, video*
  - *Web scraping*

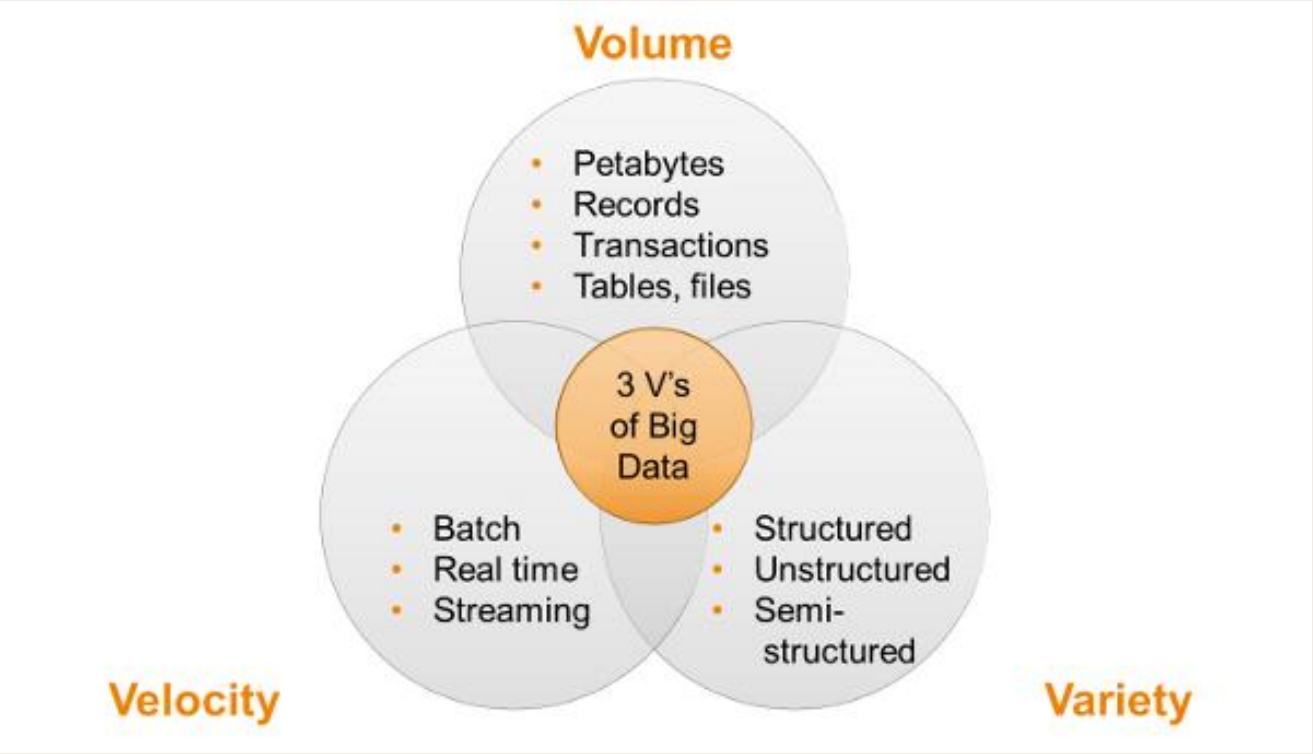
Data:
Consider SmartCity program in Boston

- Citizen-connect makes it easy to report problems such as graffiti
- Street Bump gathers pothole information from vibration data
- Rubbish bins equipped with solar panels and sensors signal when it is full
- A/B Tests to do things such as prioritizing buses at traffic lights and assess with data on congestion from Waze
- Coming soon: Small robots crawling through sewers gathering samples to reveal what people eat, how many have flu and excess salt intake
- All this data gathered is used for, CityScore a single number that indicates Boston's Overall Health

Source: https://www.economist.com/special-report/2016/03/23/how-cities-score

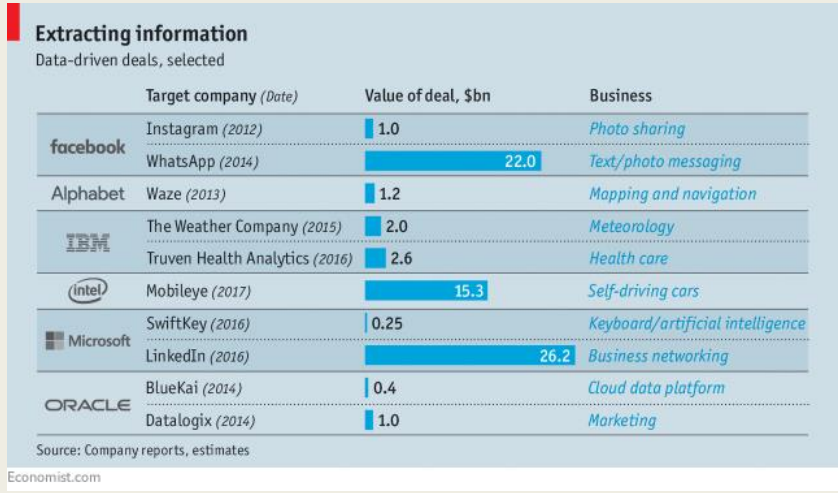

BIG
DATA
USE
ONE
SOFTWARE
USING
GARTNER
MASSIVELY
WORKING
BIOLOGICAL
HUNDREDS
PERFORMANCE
ALSO
DISK
SHARED
TIME
GENOMICS
SET
TOOLS
WITHIN
PROCESS
DEFINITION
SEARCH
INCLUDE
SYSTEMS
NETWORKS
INFORMATION
RECORDS
RELATIONAL
TENS
CAPACITY
PRACTITIONERS
ANALYTICS
BUSINESS
MOVING
SAN
PARALLEL
TOLERABLE
SIZE
NEEDED
PETABYTES
INTERNET
PROCESSING
LOGS
TECHNOLOGIES
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
EVERY
CASE
LARGER
STORAGE
COST
CONTINUES
TERABYTES
MPP
RESEARCH
CITATION
COMPLEX
DATABASES
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
ELAPSED
CURRENT
SENSOR
ARCHIVES

What Makes Big Data Unique



Data

- But also posing a challenge because,
  - *It is too big to be housed in a single computer*
  - *It is streaming*
  - *Much of it does not fit neatly into a spreadsheet*
  - *And, it is messy*
- Data is the new Oil, giving rise to a new economy
  - *Making big companies bigger. E.g., Facebook, Google*
  - *Fueling data driven companies such as Uber, Google Adwords*
  - *Stimulating acquisitions:*



Source: https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy

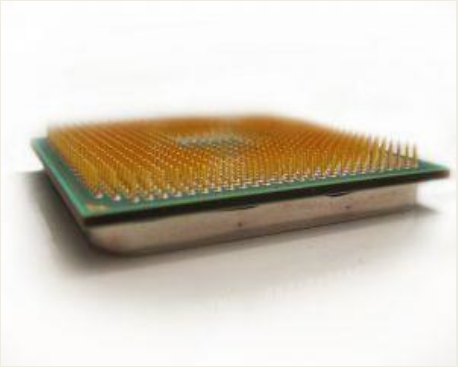
Analytical Techniques

- Analytical techniques is the common denominator across a wide array of disciplines
  - *Statistics, Mathematics, Computer Science, Information Systems, Physics, Computational biology, Business, etc.*
- The community benefits from input from many disciplines
  - *Regression from traditional Statistics*
  - *Time series analysis from Economics and Finance*
  - *Neural Networks (the basis for Deep Learning) from Biology*
  - *Computer science is helping scale up and optimize algorithms*

Software

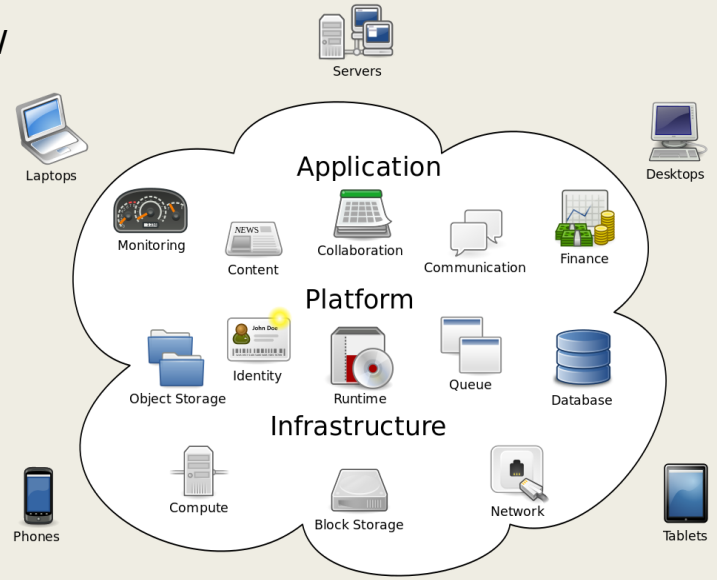
- Ecosystem of software to crunch data
- Open source
  - *An open source architecture has accelerated the advancement of open source software*
  - *Individual has idea, writes code for it and shares on GitHub, following feedback improves software, and applies for inclusion in software library*
  - *R and Python are the fastest growing data analysis software – both are open source.*

Hardware



Moore's Law

- Faster computing devices
  - *Moore's Law*
- Cloud storage and processing
  - *AWS, Google Big Table, Microsoft Azure*
- Distributed processing systems
  - *Hadoop, Spark, ..*



Cloud Computing

Spark + Hadoop

Value

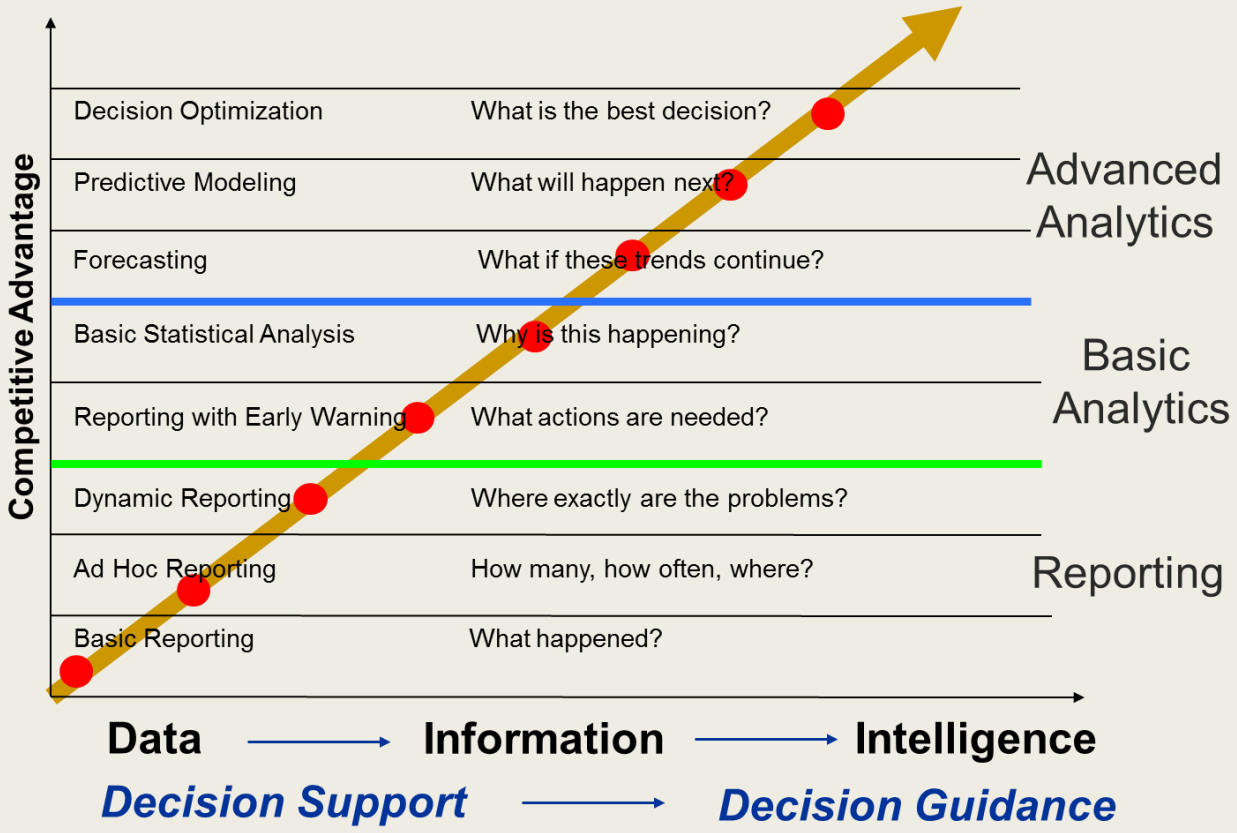
- Today's Innovative companies differentiate themselves through Analytics
  - *Netflix – Recommendations*
  - *Pandora and Spotify – Recommendations*
  - *Google – PageRank*
  - *Google Advertising*
  - *eHarmony – date matching*
  - *Uber - matching*
  - *LinkedIn - networking*
  - *Obama'08, '12, Trump'16*

DOMAIN OF ANALYTICS AND ITS IMPACT

Analytics

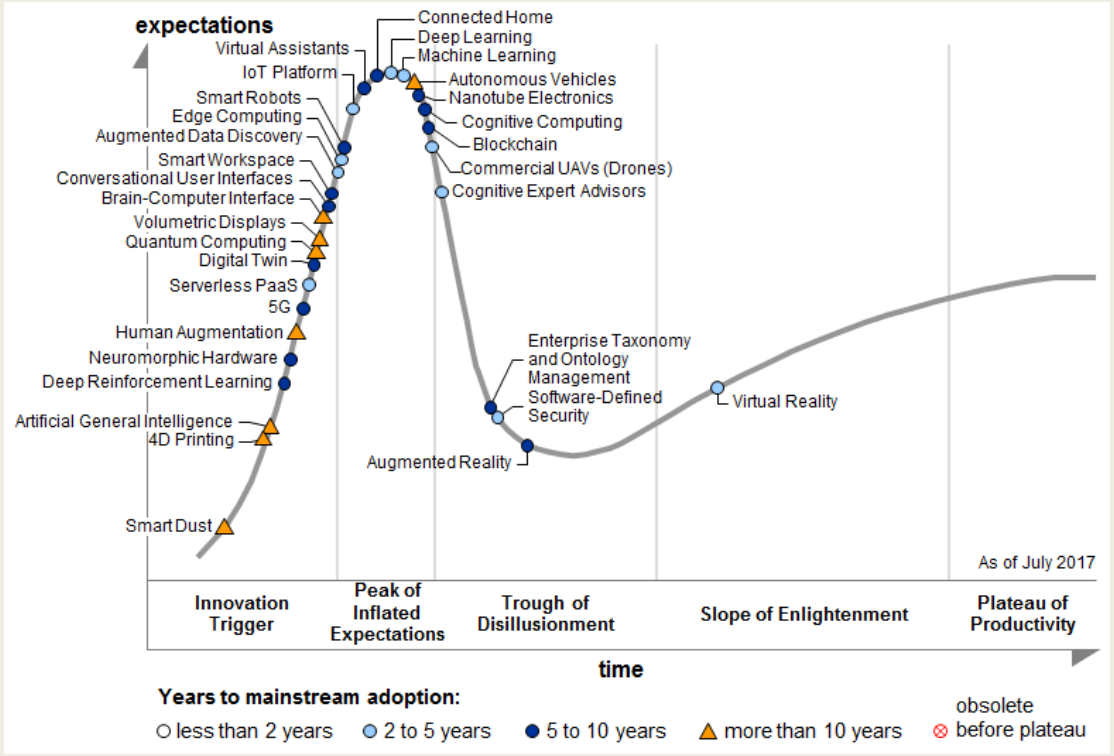
- "*We are drowning in data but starving for knowledge!*" (John Naisbitt, 1982)
- Analytics is
  - *about deriving insights from data*
  - *is the discovery and communication of meaningful patterns in data (Wikipedia, 2018)*

Levels of Analytics

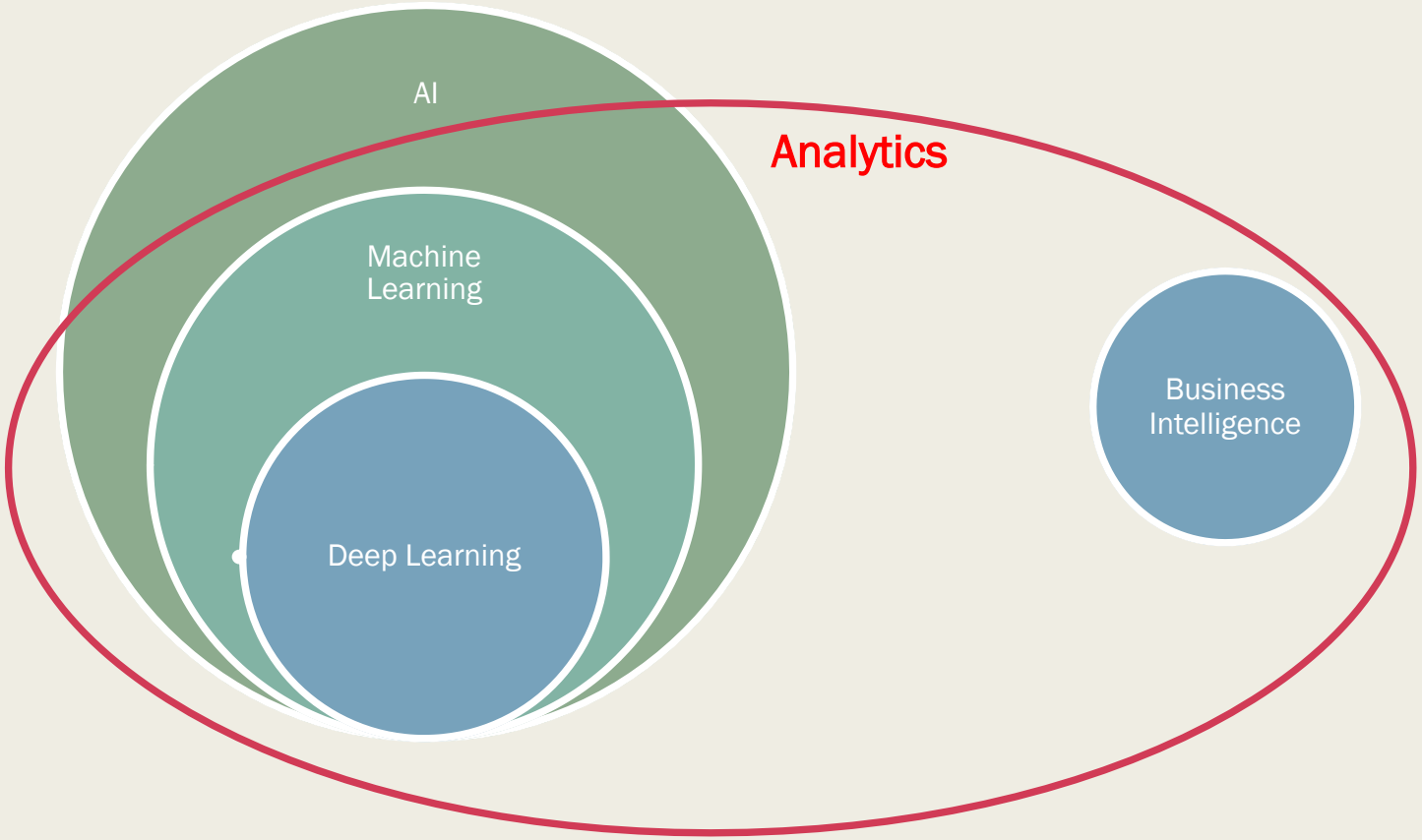


Source: Competing on Analytics by Thomas H. Davenport and Jeanne G. Harris

Many Buzz Words, Where Does Analytics Fit? Gartner Hype Cycle



Many Buzz Words, Where Does Analytics Fit? Gartner Hype Cycle



Analytics Domains in Business

- Marketing Analytics
- Workforce Analytics
- People Analytics
- HR Analytics
- Business Analytics
- Strategic Analytics
- Sales Analytics
- Web Analytics

Everyone is doing Analytics

Analytics is a part of our day-to-day lives

- Image tagging on Facebook
  - *Facial recognition based on supervised learning*
- Siri, Alexa, automatic customer service
  - *Speech recognition*
- Self-driving cars
  - *Image and object recognition*
- Credit card fraud alert, email spam filters
  - *Predictive Models*

Analytics is a part of our day-to-day lives

- Machine Learning
  - *Siri and even auto-correct learn from users past corrections.*
  - *Machine learns that an Indian name, Vishal ≠ Visual*
- Predictive Models
  - *Google is able to predict what language a webpage is in and offer to translate it.*
- Clustering
  - *Google News articles are categorized automatically.*
- Recommender Systems
  - *Amazon and NetFlix make recommendations based on past behavior; Spotify and Pandora explicate preferences from song feedback to decide on what to play.*

Analytics is Changing the World

- Analytical techniques applied to large amounts of data using powerful computers are changing the way we do things at a rapid pace.
  - *Large volumes of stocks are traded by algorithms*
  - *Facial recognition technology borne out of Deep Learning algorithms are being used to tag pictures in social media, track movement of citizens, and process MRI scans to identify tumors*
  - *Machine vision is driving self-driving cars*
  - *Predictive models are being used for purposes ranging from weather forecasts to predictive policing*

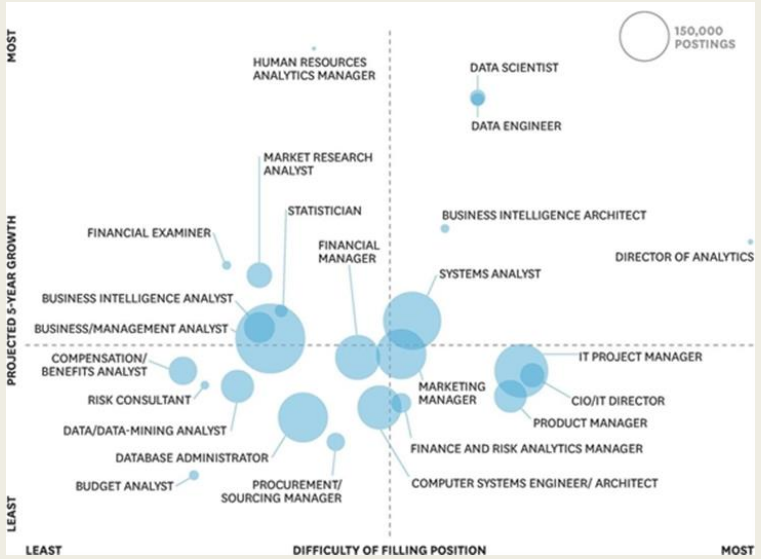
Analytics is Changing Government Too

- (Theme of the Hollywood movie) Minority Report is becoming a reality
  - *Predictive Policing*
- Boston mayor's office is testing just such an approach, using data from Yelp reviews. This has led to a 25% rise in the number of spot inspections that uncover violations.
- London borough is developing an algorithm to predict who might become homeless.
- In India Microsoft is helping schools predict which students are at risk of dropping out.
- Researchers behind an algorithm designed to help judges make bail decisions claim it can predict recidivism so effectively that the same number of people could be bailed as are at present by judges, but with 20% less crime.
- Of course, some of the predictions are unpalatable
  - *ProPublica, an investigative-journalism outfit, claims that a risk assessment in Broward County, Florida, wrongly labelled black people as future criminals nearly twice as often as it wrongly labelled whites.*

Source: https://www.economist.com/leaders/2016/08/18/the-power-of-learning

Analytics is Changing the Job Landscape

- Analytics driven innovations are threatening many occupations.
- The future will belong to those who can leverage the power of analytics



DATA SCIENCE IS DISRUPTING THE JOB MARKET.
Source: Harvard Business Review

Catalogue of fears

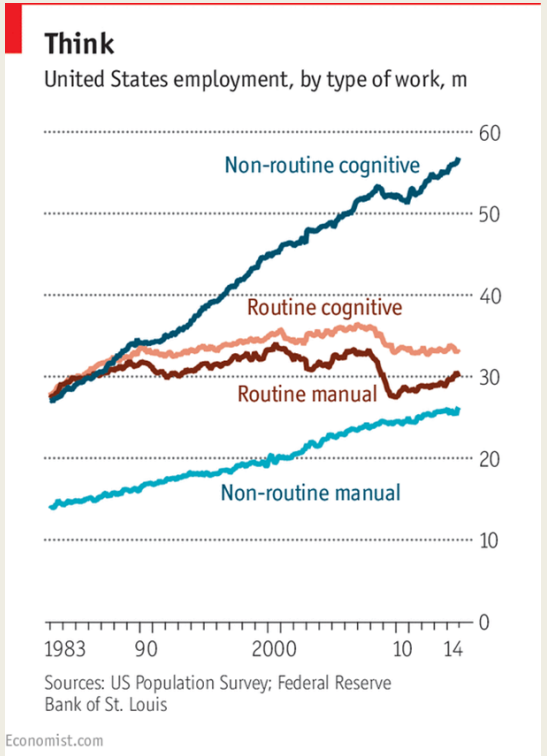
Probability of computerisation of different occupations, 2013
(1 = certain)

Job	Probability
Recreational therapists	0.003
Dentists	0.004
Athletic trainers	0.007
Clergy	0.008
Chemical engineers	0.02
Editors	0.06
Firefighters	0.17
Actors	0.37
Health technologists	0.40
Economists	0.43
Commercial pilots	0.55
Machinists	0.65
Word processors and typists	0.81
Real-estate sales agents	0.86
Technical writers	0.89
Retail salespeople	0.92
Accountants and auditors	0.94
Telemarketers	0.99

Source: "The Future of Employment: How Susceptible are Jobs to Computerisation?", by C. Frey and M. Osborne (2013)

Economist.com

Source: Economist



Source: Economist

There isn’t a better time to embrace Applied Analytics

INTRODUCTIONS

Student Introductions

- Please share your
  - *Name, and*
  - *Absolutely anything you know about Analytics*

ABOUT THE COURSE

About the Course

- Overview
- Learning Objectives
- Readings
- Resources
- Assessment
- Grading
- Policies
- Schedule

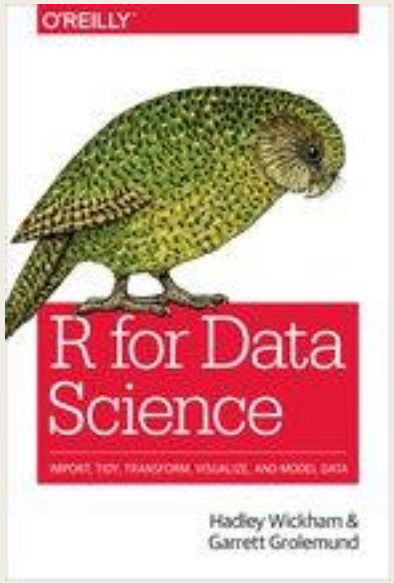
About the Course

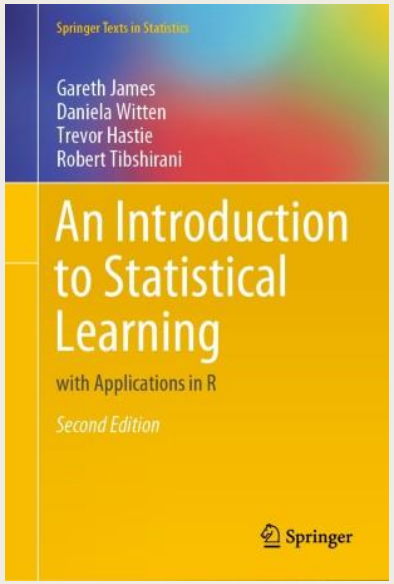
Overview

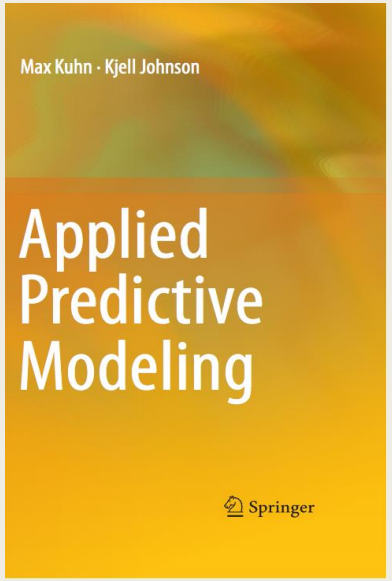
- Course will
  - *develop knowledge and skills in the use of predictive models*
  - *cover a wide array of supervised learning techniques*
  - *develop skills in data wrangling*
  - *use R*
- Approach of the course is to help you to "*learn by doing*"
  - *Discuss analytical technique at high level*
  - *Implement on a dataset in class*
  - *Hands-on assignment to tackle a different problem*
  - *Exam and predictive analysis competition to demonstrate learning*

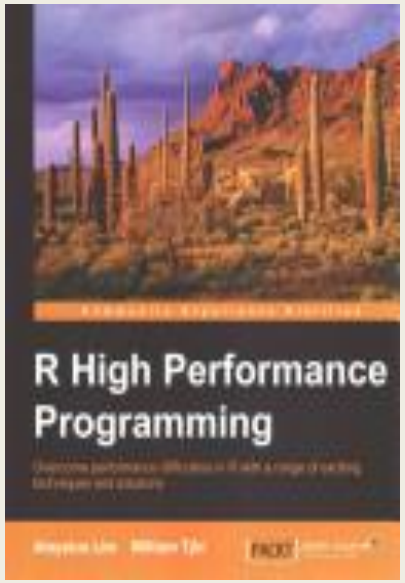
About the Course

Readings









Resources

- Columbia University Information Technology
  - *University Provided discounted software downloads*
- *Columbia University Library*
- *SPS Academic Resources*

Assessment

- Assignments
  - *Eight assignments, each worth 5%: 40%*
- Exam: 40%
- Kaggle Project: 15%
- Class Engagement: 5%

Grading Scale

Grade	Percentage
A+	98–100 %
A	93–97.9 %
A-	90–92.9 %
B+	87–89.9 %
B	83–86.9 %
B-	80–82.9 %
C+	77–79.9 %
C	73–76.9 %
C-	70–72.9 %
D	60–69.9 %
F	59.9% and below

Policies

- Course Policies
  - *Participation and Attendance*
  - *Late Work*
  - *Citation and Submission*
- School Policies
  - *Copyright policy*
  - *Academic Integrity*
  - *Accessibility*

Schedule and Availability

- On the Syllabus page on Canvas, you will find
  - *Detailed Course Schedule*
  - *Contact Information and Office Hours*

OVERVIEW OF R

Analysis Software

- R
- Python
- SQL
- SAS
- Julia
- MATLAB
- SPSS
- Excel
- …

- A large and growing list
- Vary based on
  - *Paid (e.g., SAS) vs Open Source (R, Python)*
  - *Menu driven (e.g., Excel) vs code-driven*
  - *Workflow (e.g., SAS Enterprise Miner, RapidMiner)*
  - *Statistical Domain focus (e.g., IBM SPSS, MATLAB)*
  - *Scalability*
  - *Reporting capability (e.g., Tableau)*

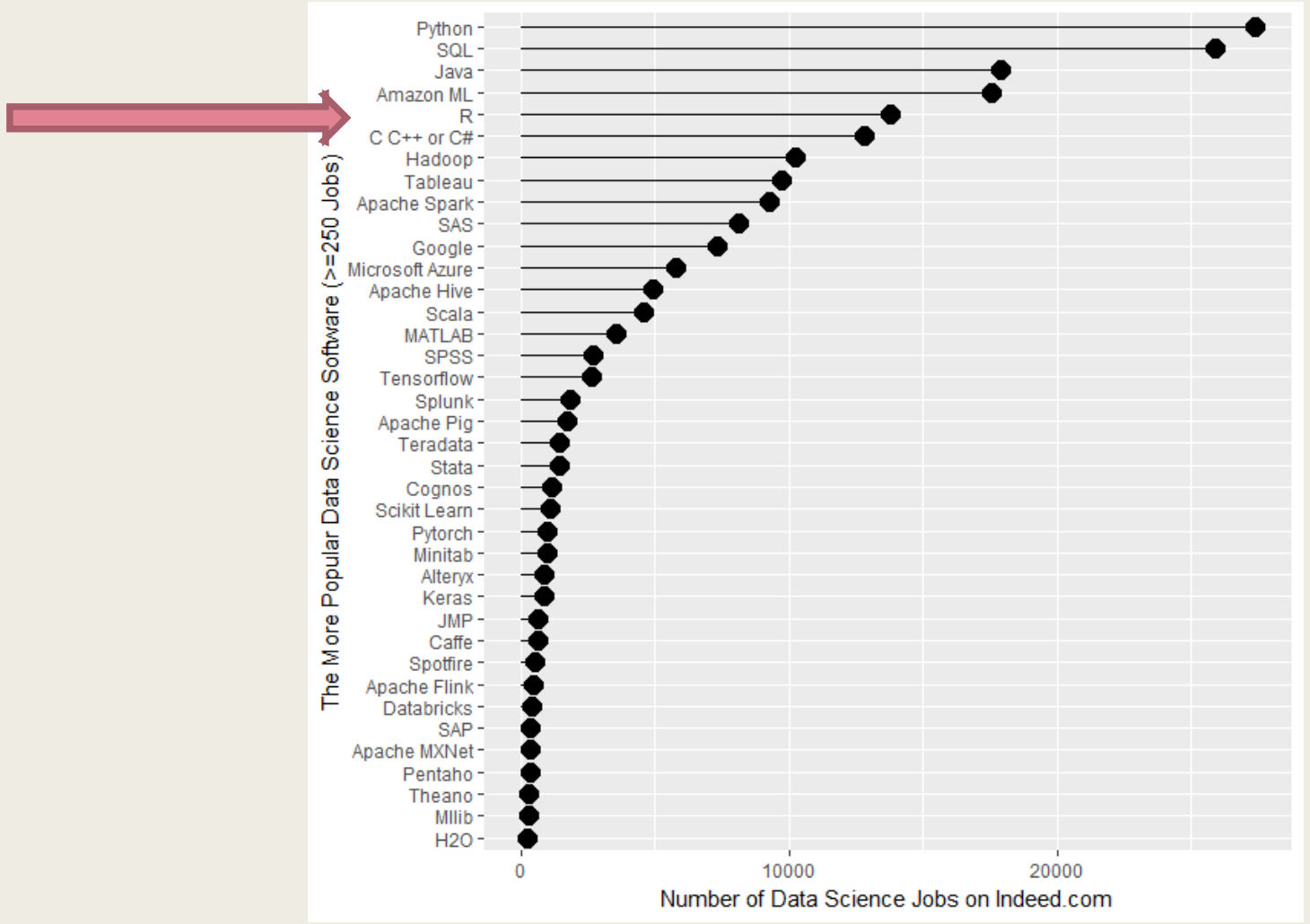
R

Software for conducting data preparation, analyses, and visualization

&

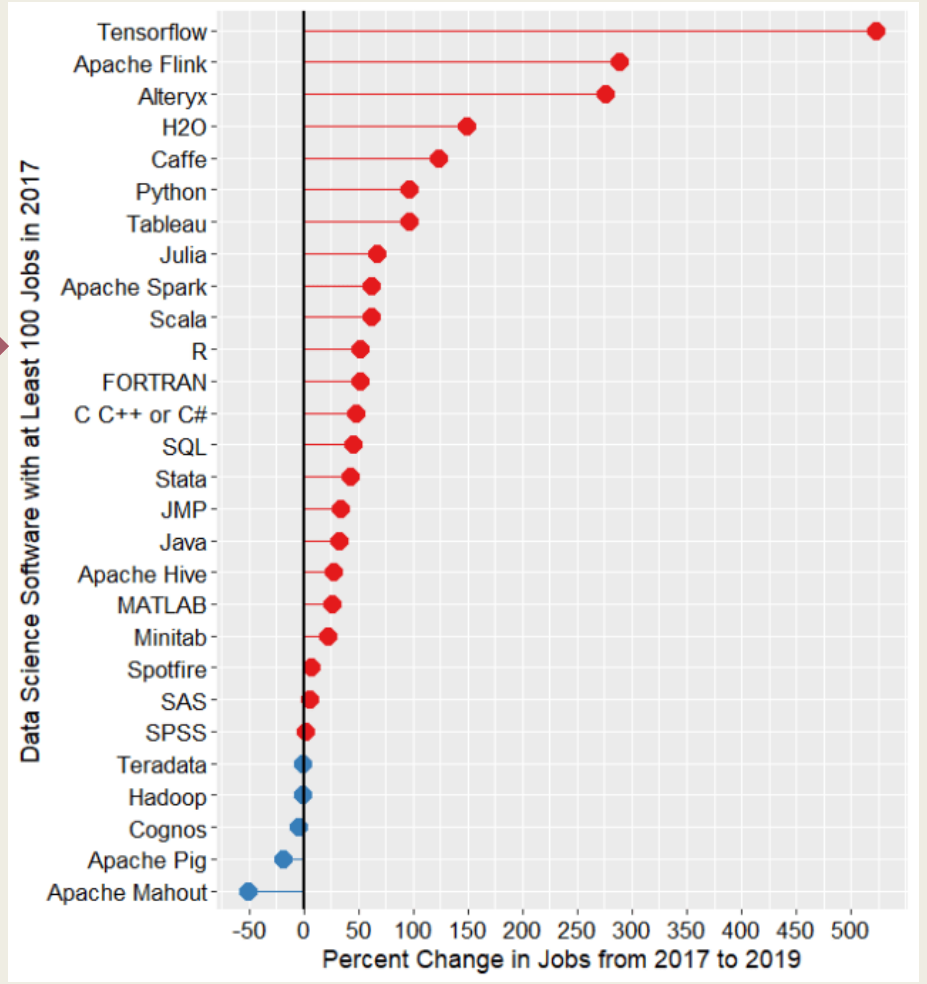
Programming language

Data Science Jobs by Software

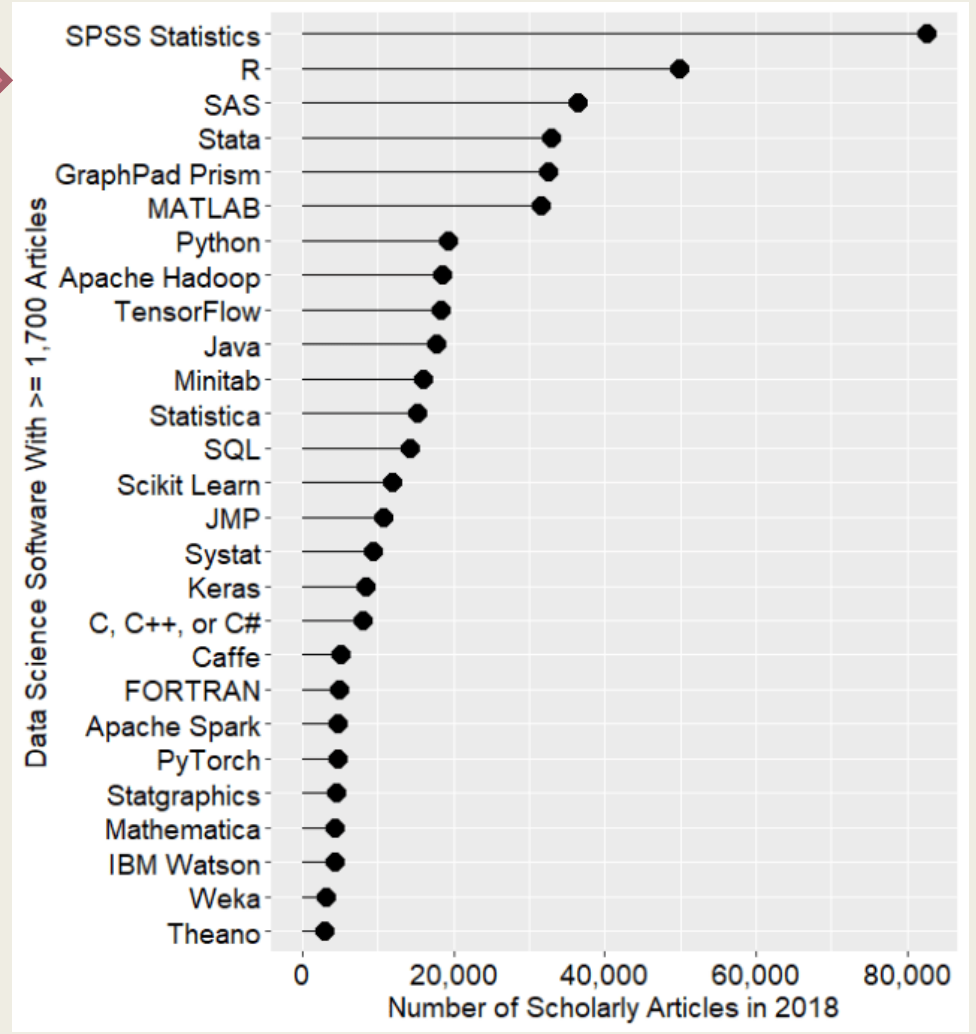


Source:r4stats

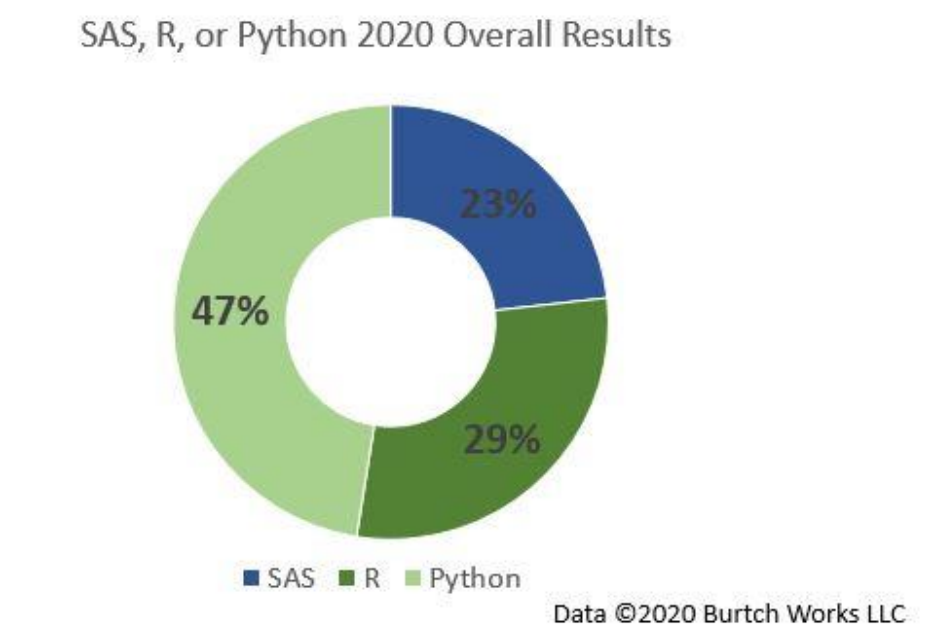
% Change in Data Science Jobs: 2017 vs. 2019

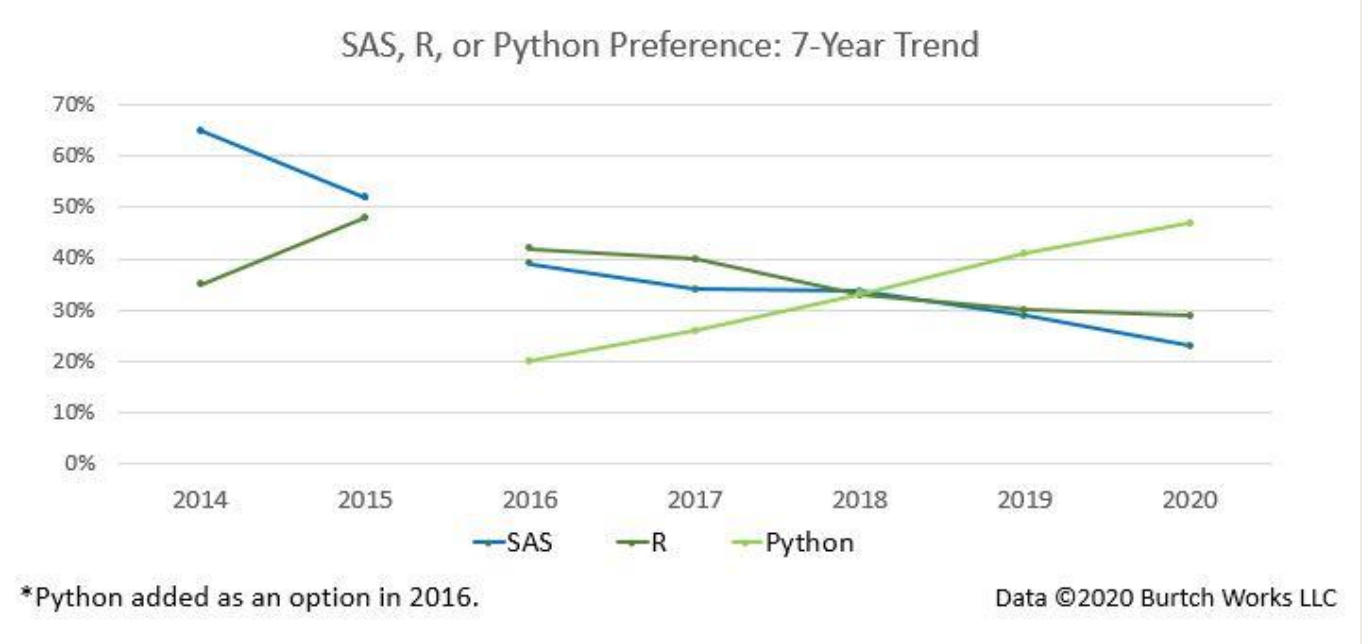


In Academic Research



R vs. Python vs. SAS
Survey of Analytics Professionals





Source: Burtch Works Survey (2020)

Why R?

- Open Source, so Free
- It is not only a statistical software, but a language.
- Comprehensive statistical platform. Can do just about any type of data analysis
- Interactive programming language
- Functional programming language
- Exceptional graphics capabilities
- Able to load many types of data and access from multiple sources
- Number of GUIs
- Works on a wide array of platforms (Runs on a wide array of platforms and interfaces (Windows, Mac, Unix, iPhone..)
- Growing capabilities (by way of Packages)
- Very active and vibrant user community

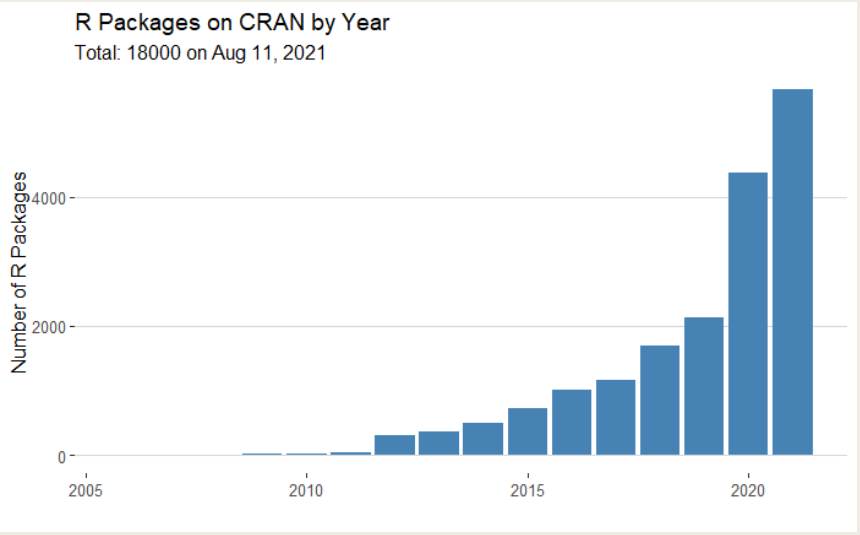
What’s not good about R

- Not point and click
- Size of data is limited by RAM of computer. But, there are ways to work around this
- Functionality is based on user demand and contributions
- No customer support and no one to sue

R System

Base R

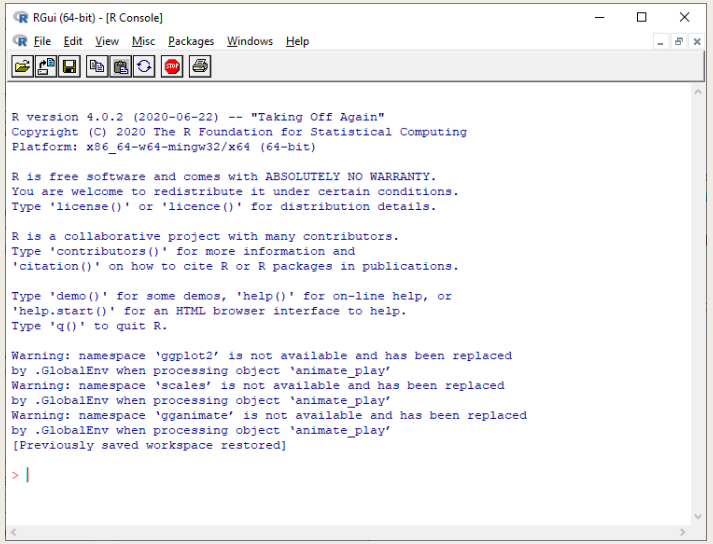
With packages



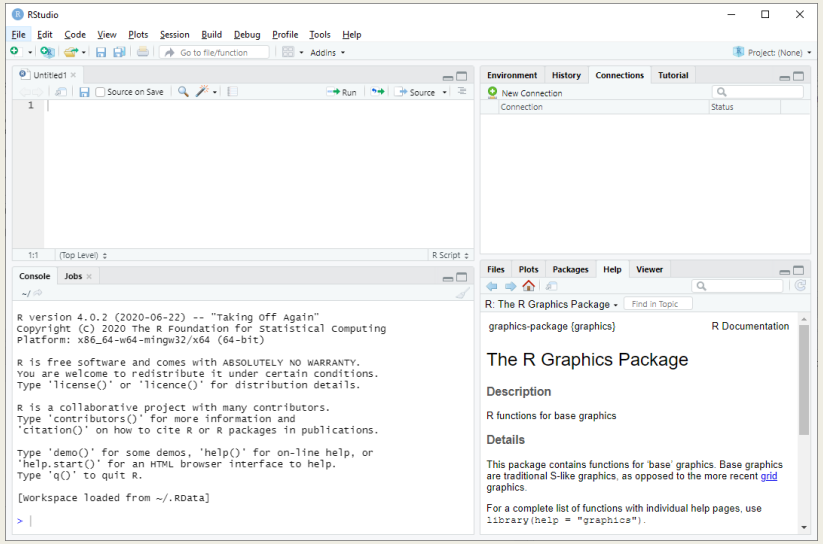
List of R packages

Using R

R Console



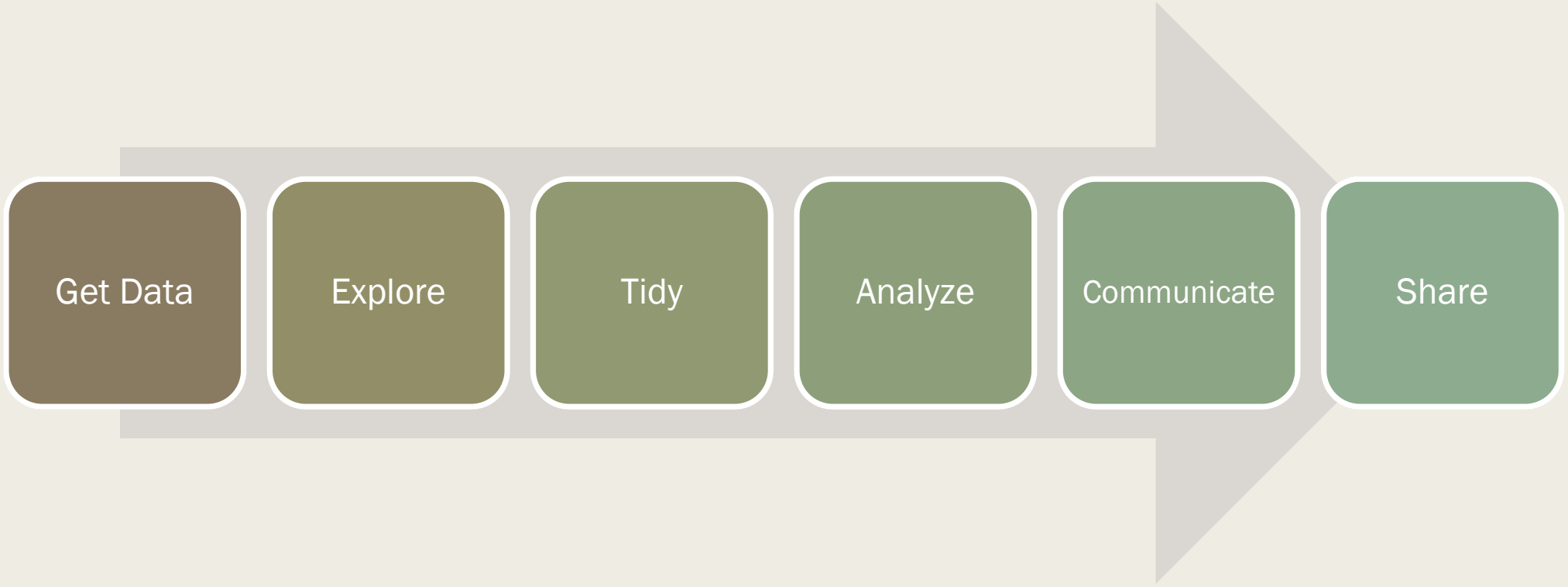
RStudio



Setup

- Ensure you have
  - *R (version 4.1.1 or later) and*
  - *RStudio (version 1.4.1717 or later)*
- If not
  - *Install R*
  - *Install RStudio Desktop*

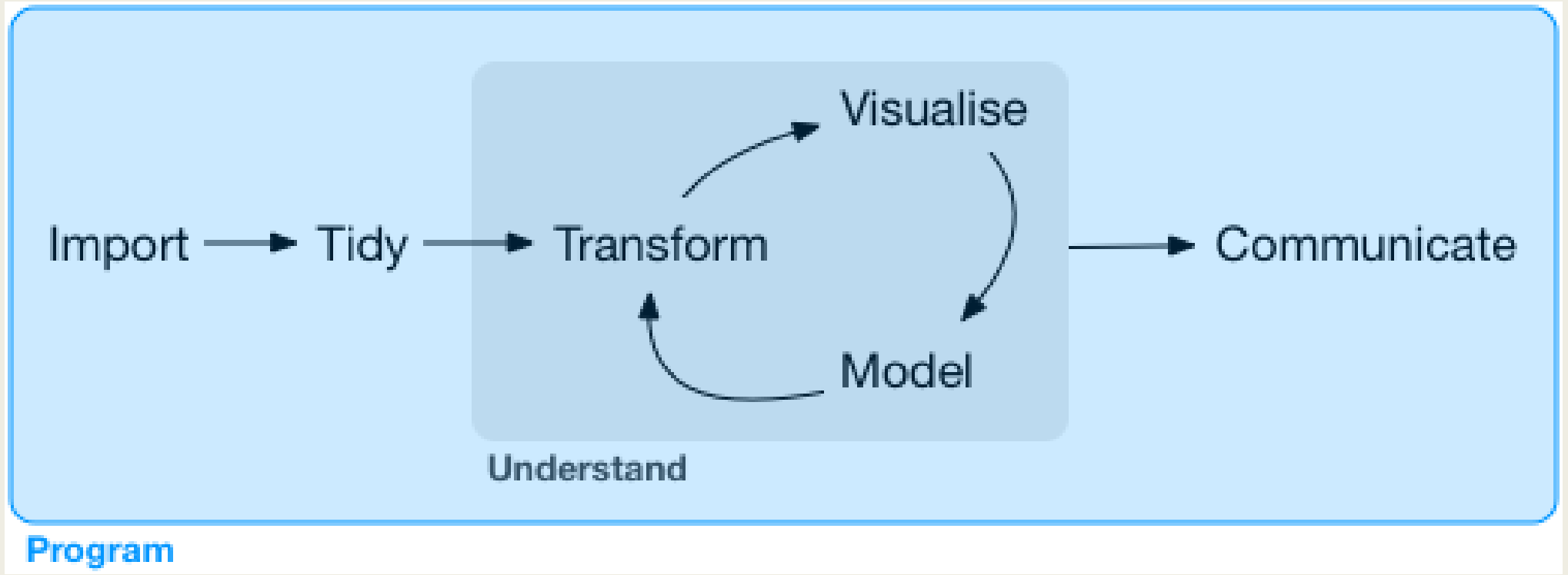
Data Analysis Pipeline



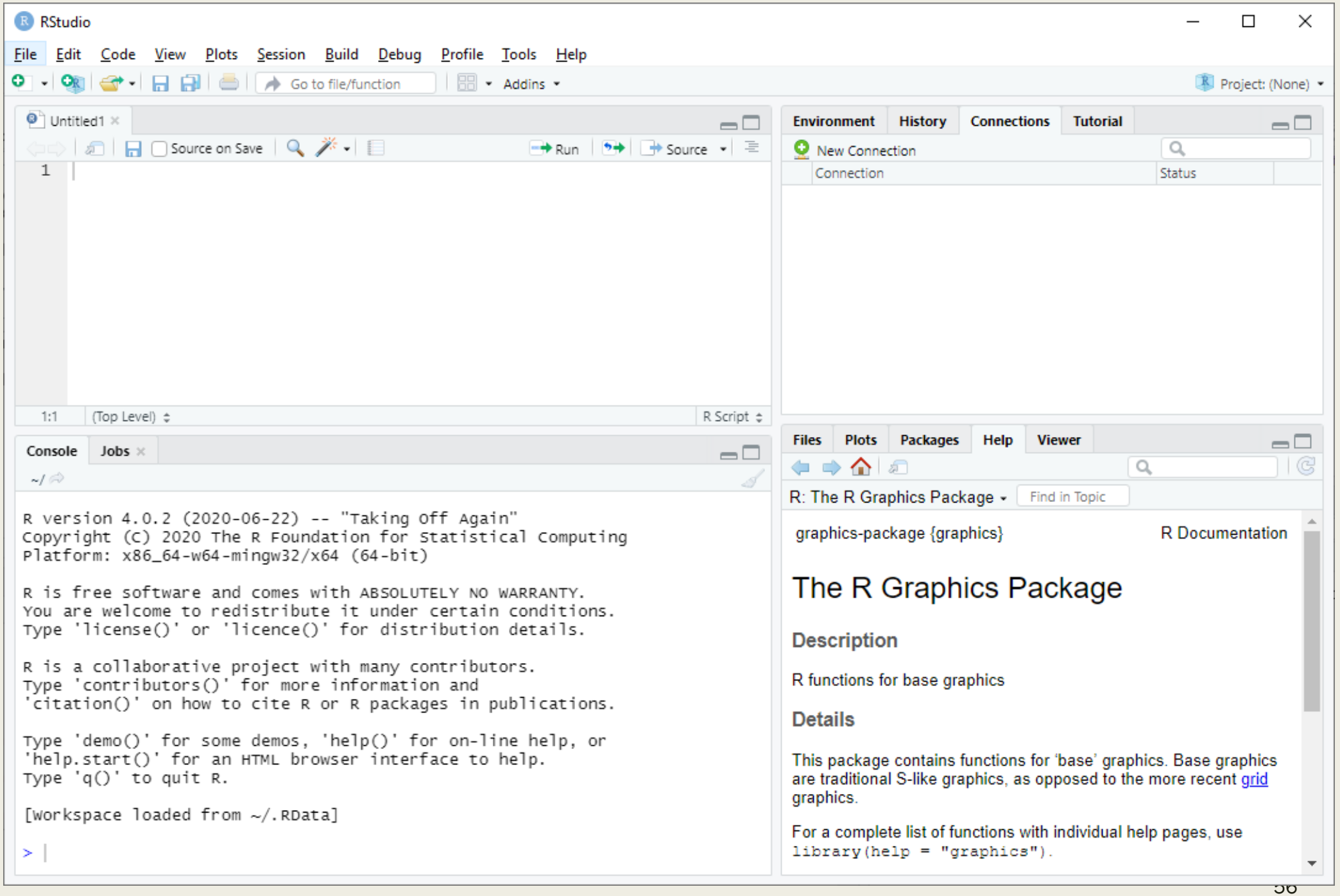
Data Analysis Pipeline

With representative R Packages

Data Analysis Process



Source: R for Data Science


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console
Jobs
~/
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment History Connections Tutorial
New Connection
Connection
Status
Files Plots Packages Help Viewer
R: The R Graphics Package
Find in Topic
graphics-package {graphics}
R Documentation
The R Graphics Package
Description
R functions for base graphics
Details
This package contains functions for 'base' graphics. Base graphics are traditional S-like graphics, as opposed to the more recent grid graphics.
For a complete list of functions with individual help pages, use library(help = "graphics").

Working with RStudio

- Write code in either R Script or R Markdown.
- At the end of your work, save script file as .R or .rmd.
  - *By default, R will save the file to your working directory. To see your working directory, run getwd(). You can change your working directory in R using setwd. E.g. setwd("c:/coolClass/awesomeProf/ILoveR")*
- RStudio does auto-save session but relying on this is not a good practice.
- Avoid using menu options of RStudio.

Getting Help

- Search for answers
  - *Access R Help by typing "?" before a function (e.g., ?mean). See Examples*
  - *Stack Overflow, R Bloggers , R Mailing List: r-help@r-project.org*
  - *But, by far the best is Google*
    - copy-paste error messages from R
    - type function followed by R
- Ask for help on
  - *Class Discussions Board, Stack Overflow, Github*
  - *But, be sure to share a reproducible example. library(reprex) generates shareable code and result.*

Foundations for Applied Analytics Frameworks and Methods

- You are expected to have a working knowledge of R
- The DataCamp courses you were asked to complete are meant to provide you a foundation for this Course
- **If you have not completed those Data Camp courses, it is absolutely imperative that you complete them before the next session.**

Summary

- In this session we
  - *discussed the forces shaping the growth of analytics*
  - *examined the domain of analytics and its impact*
  - *reviewed the course structure*
  - *took a look at R*