Introduction to R and RStudio

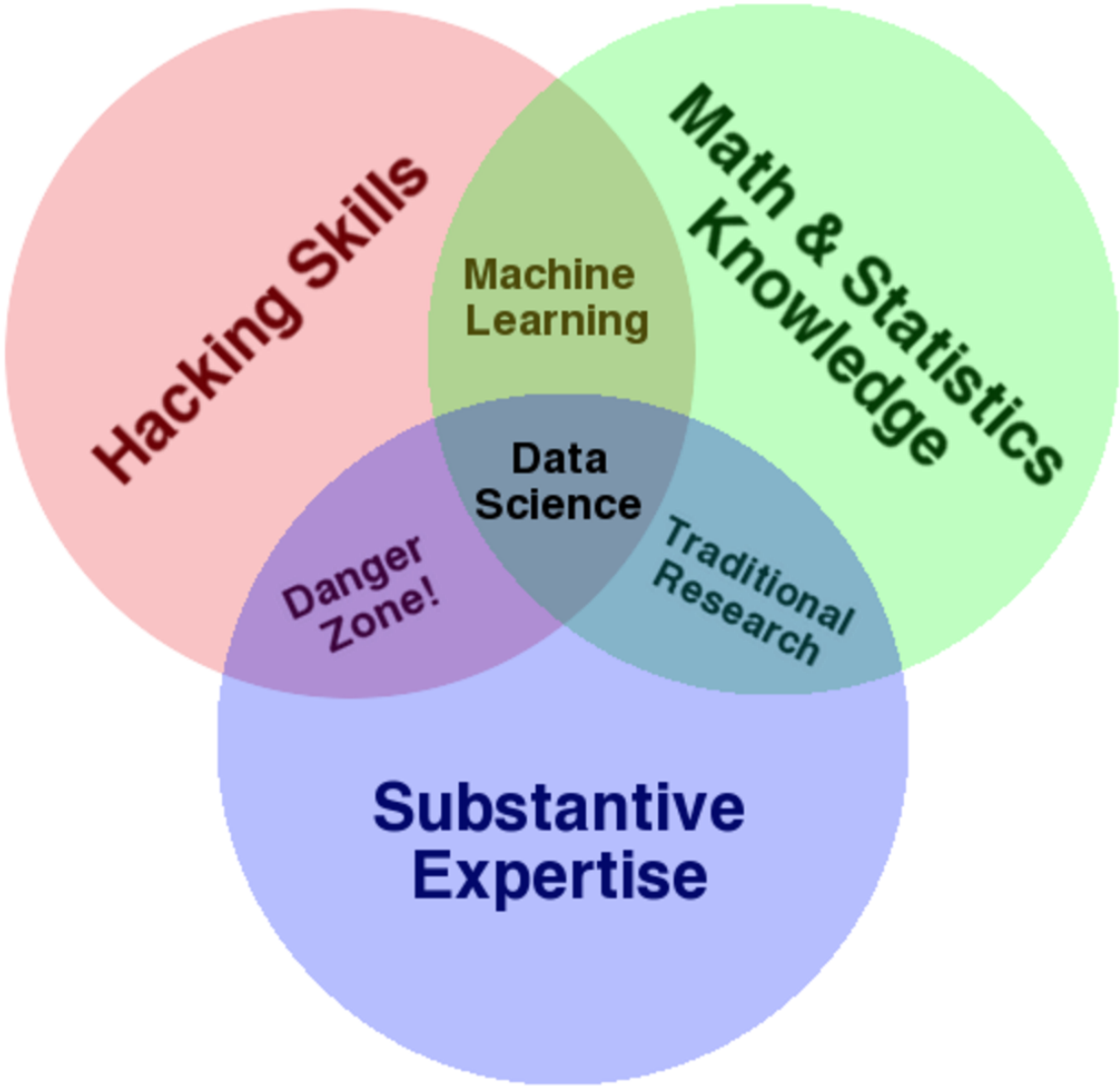
Chien-Liang Liu

August 22, 2016

Data Scientist: The Sexiest Job of the 21st Century

Source: `https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/`

Data Science Venn Diagram

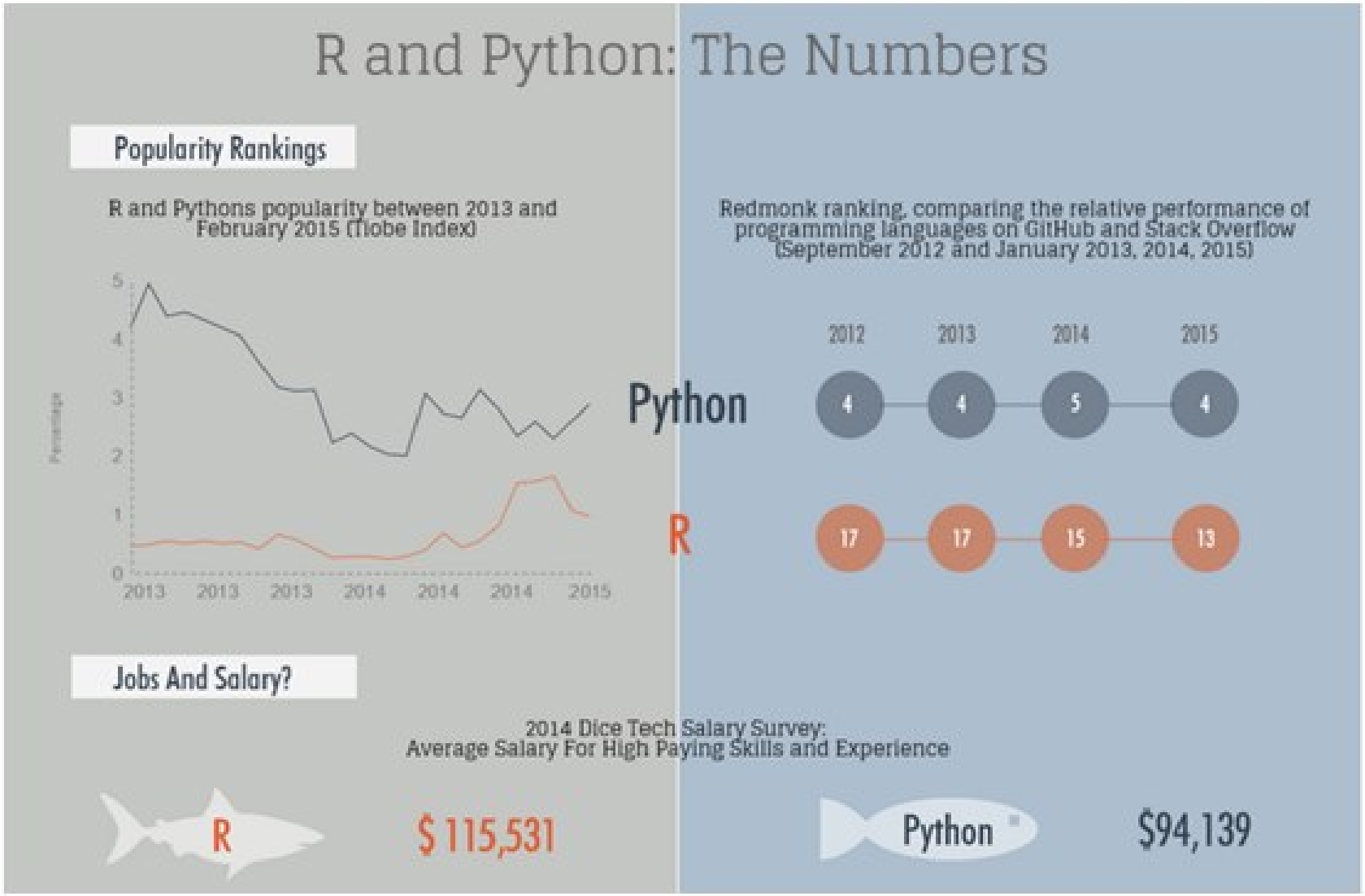


Source: `http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram`

Objective of this course

- This course focuses on hacking skill
- We use R as the programming language
- Learning by doing is the best approach
- R is simpler than C/C++, Java, and Python
- Besides R, we will also cover basic data analysis

R vs. Python



Source: `http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html`

When and how to use R?

- R is mainly used when the data analysis task requires standalone computing or analysis on individual servers
- It's great for exploratory work, and it's handy for almost any type of data analysis because of the huge number of packages and readily usable tests that often provide you with the necessary tools to get up and running quickly
- R can even be part of a big data solution.

IDE and packages for R

- When getting started with R, a good first step is to install the amazing RStudio IDE
- dplyr, plyr and data.table to easily manipulate packages
- stringr to manipulate strings
- zoo to work with regular and irregular time series,
- ggvis, lattice, and ggplot2 to visualize data, and
- caret for machine learning

When and how to use Python?

- You can use Python when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database
- Being a fully fledged programming language, it's a great tool to implement algorithms for production use

IDE and Packages for Python

- Python has no clear “winning” IDE. We recommend you to have a look at Spyder, IPython Notebook and Rodeo to see which one best fits your needs
- NumPy /SciPy (scientific computing)
- pandas (data manipulation) to make Python usable for data analysis
- matplotlib to make graphics
- scikit-learn for machine learning

Machine Learning

- Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence
- Machine learning explores the study and construction of algorithms that can learn from and make predictions on data
- Machine learning is closely related to and often overlaps with computational statistics; a discipline that also specializes in prediction-making. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field

Source: `https://en.wikipedia.org/wiki/Machine_learning`

Where is Data From?

Everywhere

Transaction Data

Government Data

- Traffic
- Real Estate
- ...

Scientific Data

- Genome
- Climate
- ...

UGC

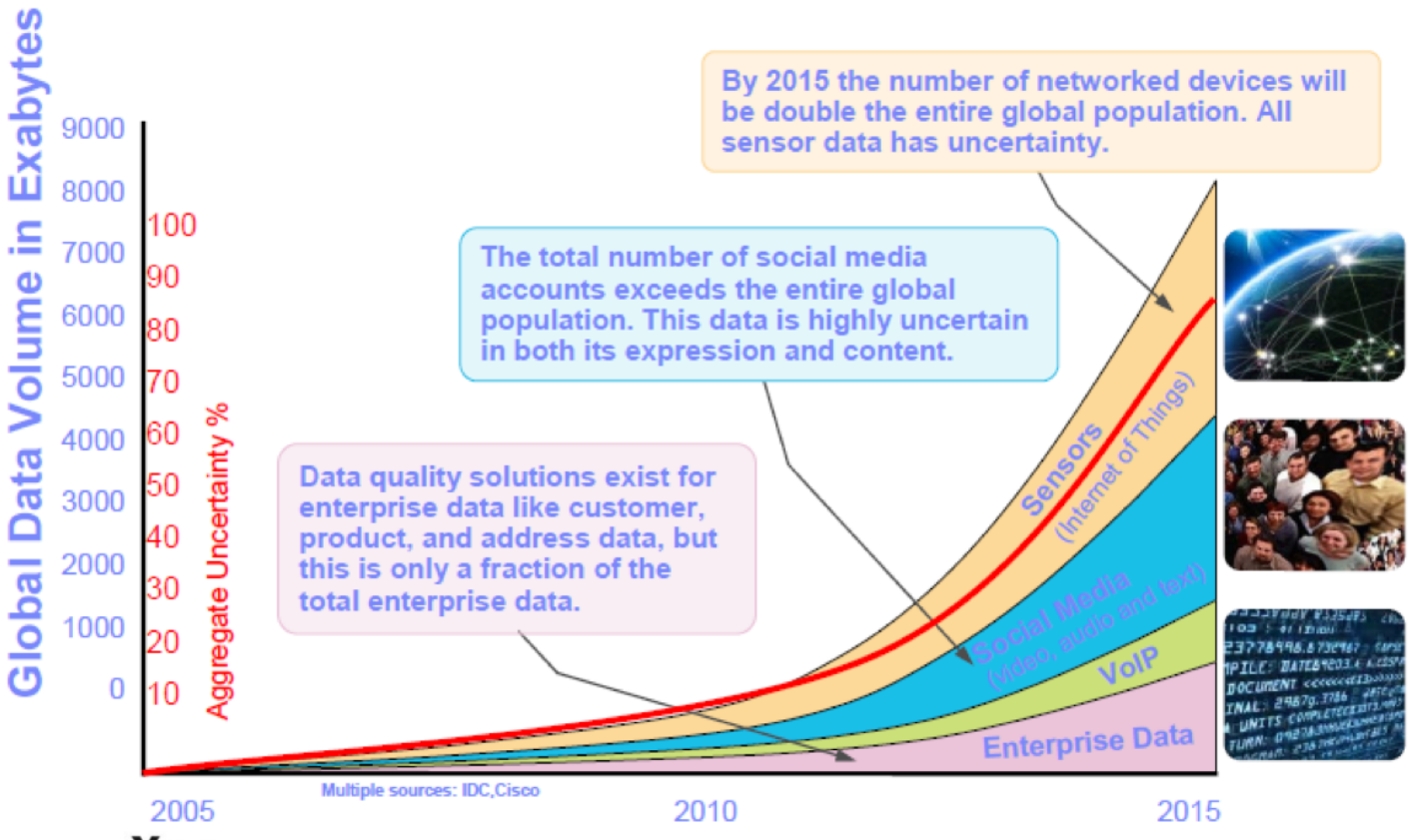
- Text
- Multimedia
  - Video
  - Image
  - Audio

Sensor Data

- GPS
- Gyroscope
- Accelerometer
- Thermometer
- ...

Log Data

Data is growing exponentially



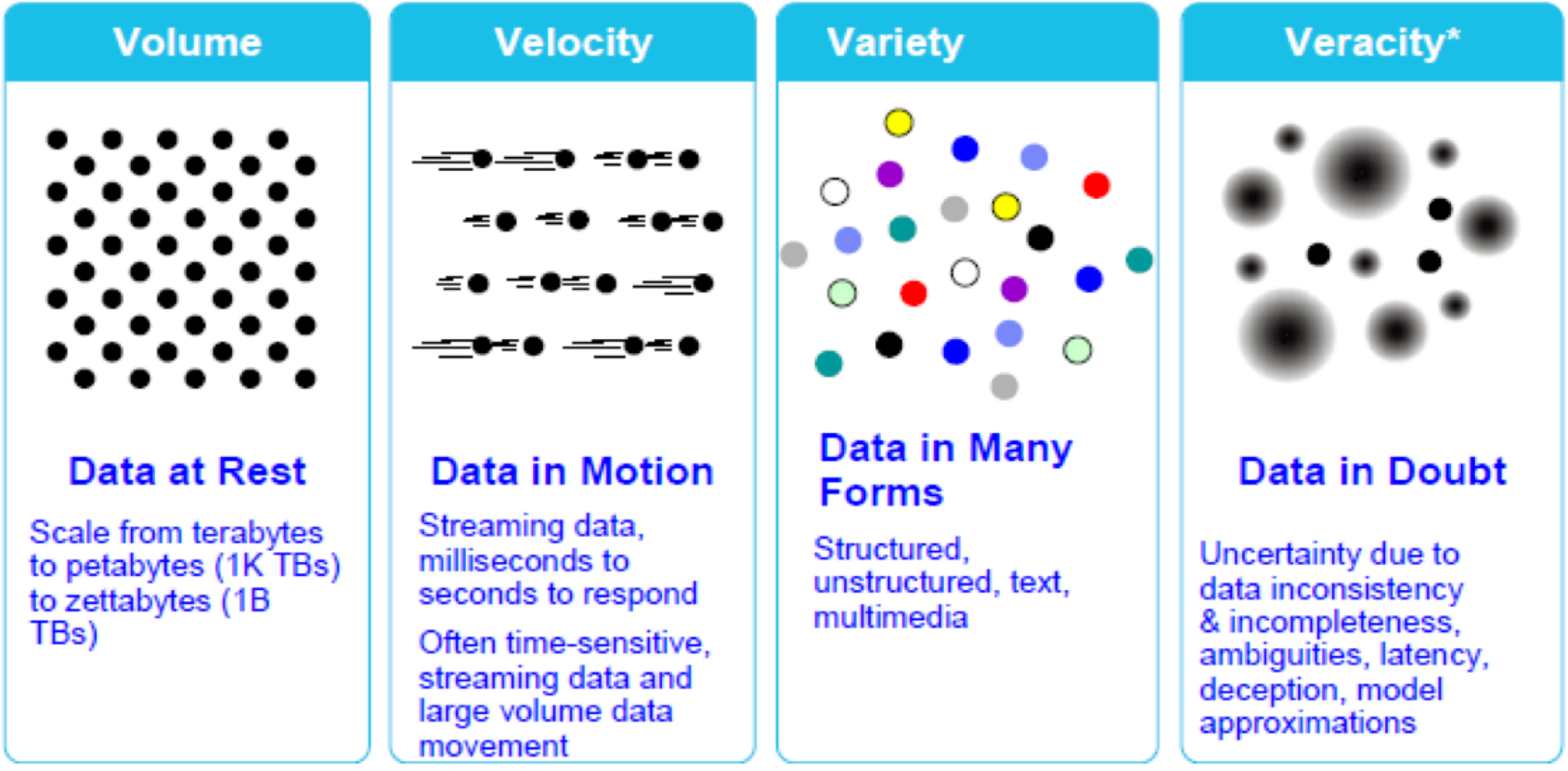
YouTube Video length 2006: <1hr/min
60hrs/min 2010
2005.02

tweets
2006.03 400k/Q 2007 | 100M/Q 2008 | 50M/Day 2010 | 140M/Day 2011

users
2004.02 100M 2008 | 300M 2009 | 500M 2010 | 800M 2011 | 1B 2012

What's Big Data

"Big Data" refers to data that grows so large that it is difficult to capture, store, manage, share, analyze and visualize with the typical hardware environments and database software tools



Source : Solutions Big Data IBM, 2012

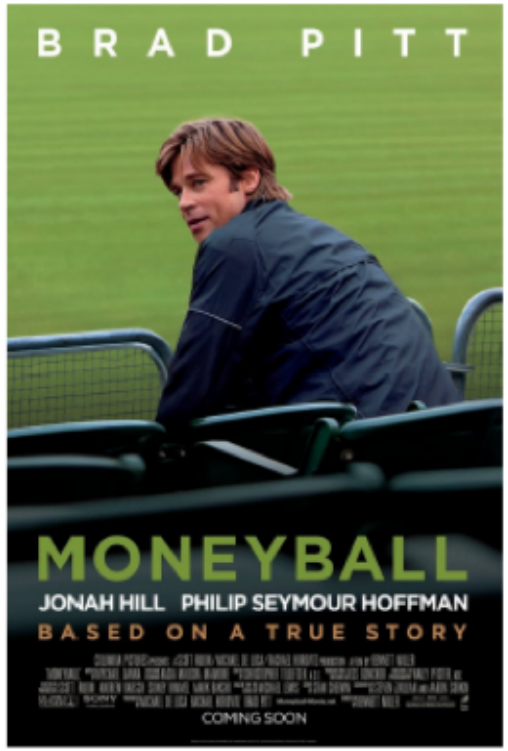
Data-Driven Decision Making

Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself `http://www.cra.org/ccc/docs/init/bigdatawhitepaper.pdf`

Corporations (baseball teams) need more data and advanced analytics to remain competitive.

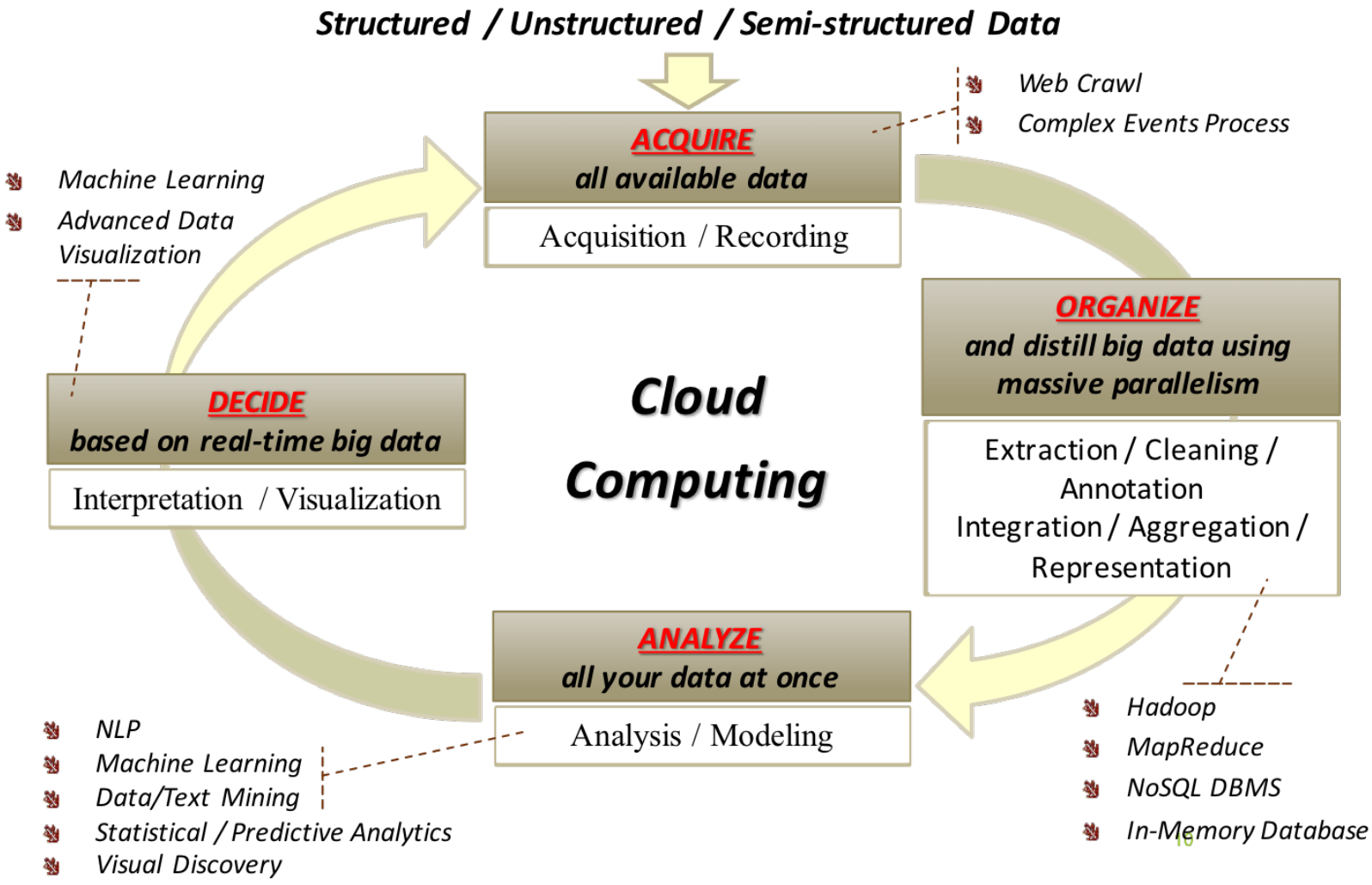
E.R.A evolved into ... Skill-Interactive Earned Run Average

$$SIERA = 6.145 - 16.986*(SO/PA) + 11.434*(BB/PA) - 1.858*((GB-FB-PU)/PA) + 7.653*((SO/PA)^2) +/- 6.664*(((GB-FB-PU)/PA)^2) + 10.130*(SO/PA)*((GB-FB-PU)/PA) - 5.195*(BB/PA)*((GB-FB-PU)/PA)$$



Source : New York Times

Data Processing Flow

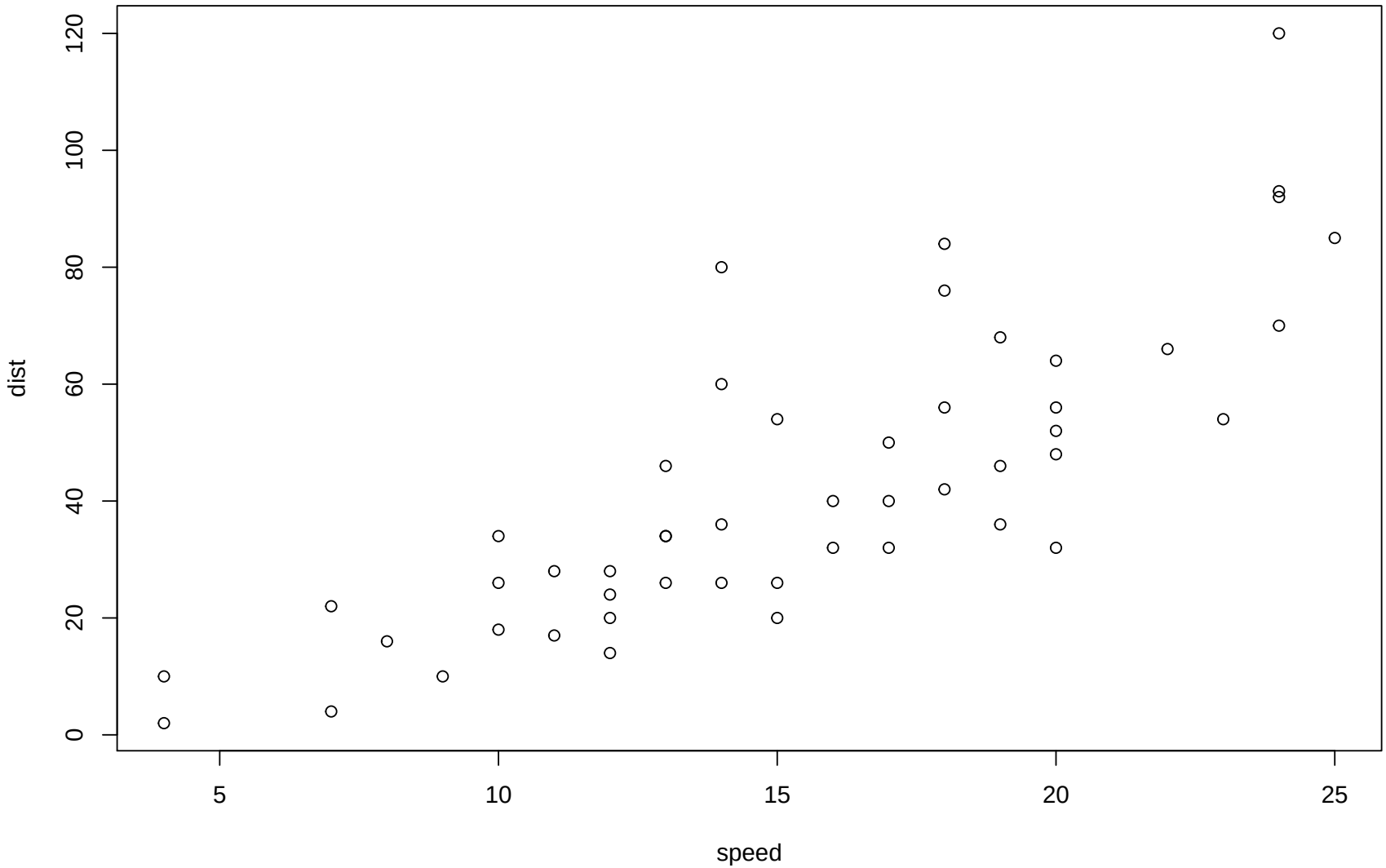


R & RStudio

What's R

Why R?

Visualization



R Installation

RStudio

Howework 0

- Install R & RStudio
- Customize your own RStudio environment and try to be familiar with it
- Start coding