MIS 4470/5470 – Practical Computing for Data Analytics

Welcome

Outline

- Introductions
- Overview of data science and business analytics
- Overview of our class
- The **pcda** computing appliance
  - How it works
  - Hands on practice to get familiar with **pcda**
- Steps to prepare for next class

Data science and data scientists

- To gain insights from data through computation, statistics and visualization
- “A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.” - Josh Blumenstock
- “Data Scientist = statistician + programmer + coach + storyteller + artist” - Shlomo Aragnon

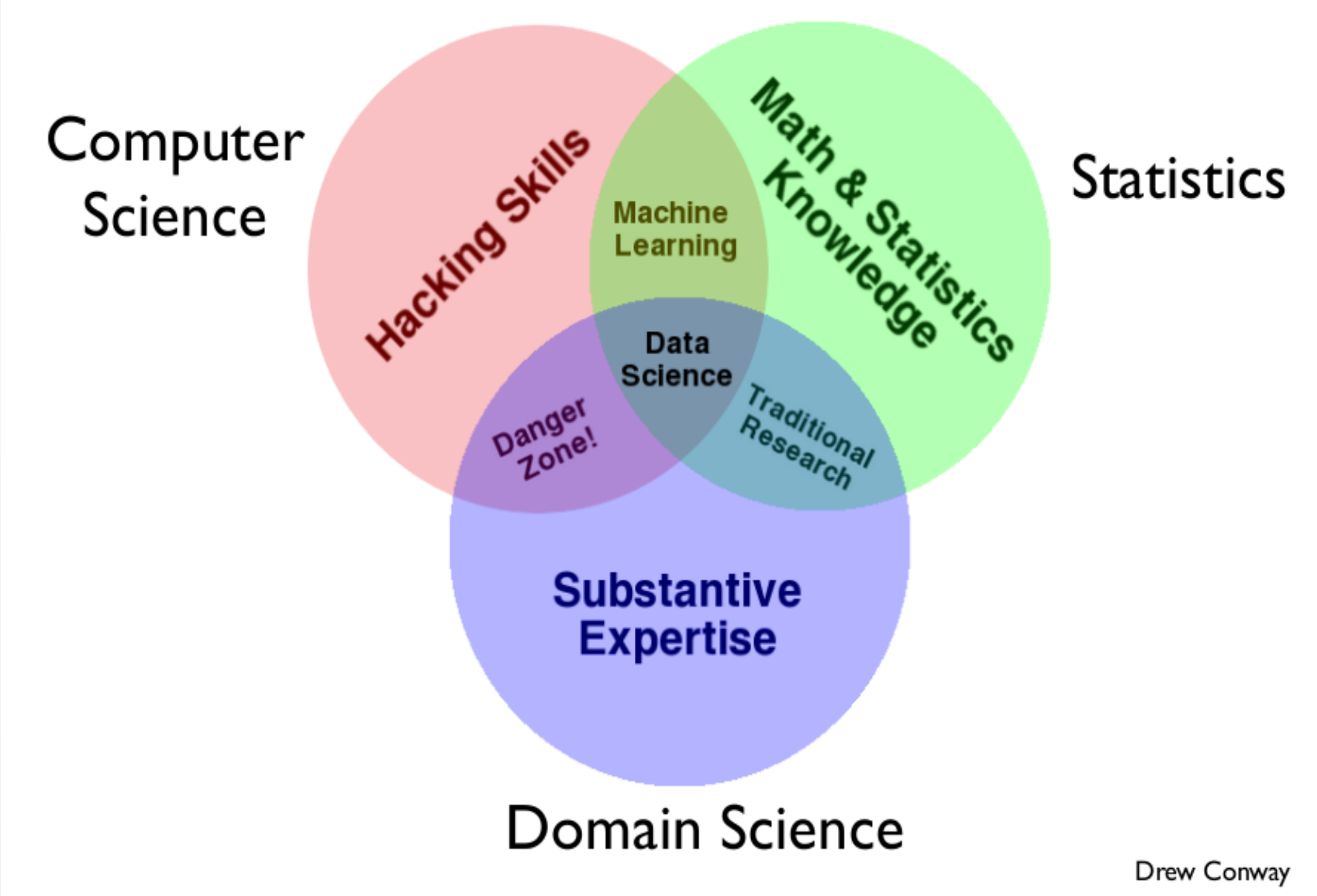
Big Data and Data Science Hype

- Ambiguity of terms like “big data” and “data science”
  - Meaning? Who? Where? How big is big?
- Lack of respect for the foundation
  - Decades of research and practice in statistics, computer science, engineering, science
  - The media would have you believe “data science” is a new invention
- Crazy hype regarding data science and what it can do
- Statisticians already work on the “science of data”
- “Anything that has to call itself a science, isn't”.
  - Like modeling, data science has a large element of “craft”

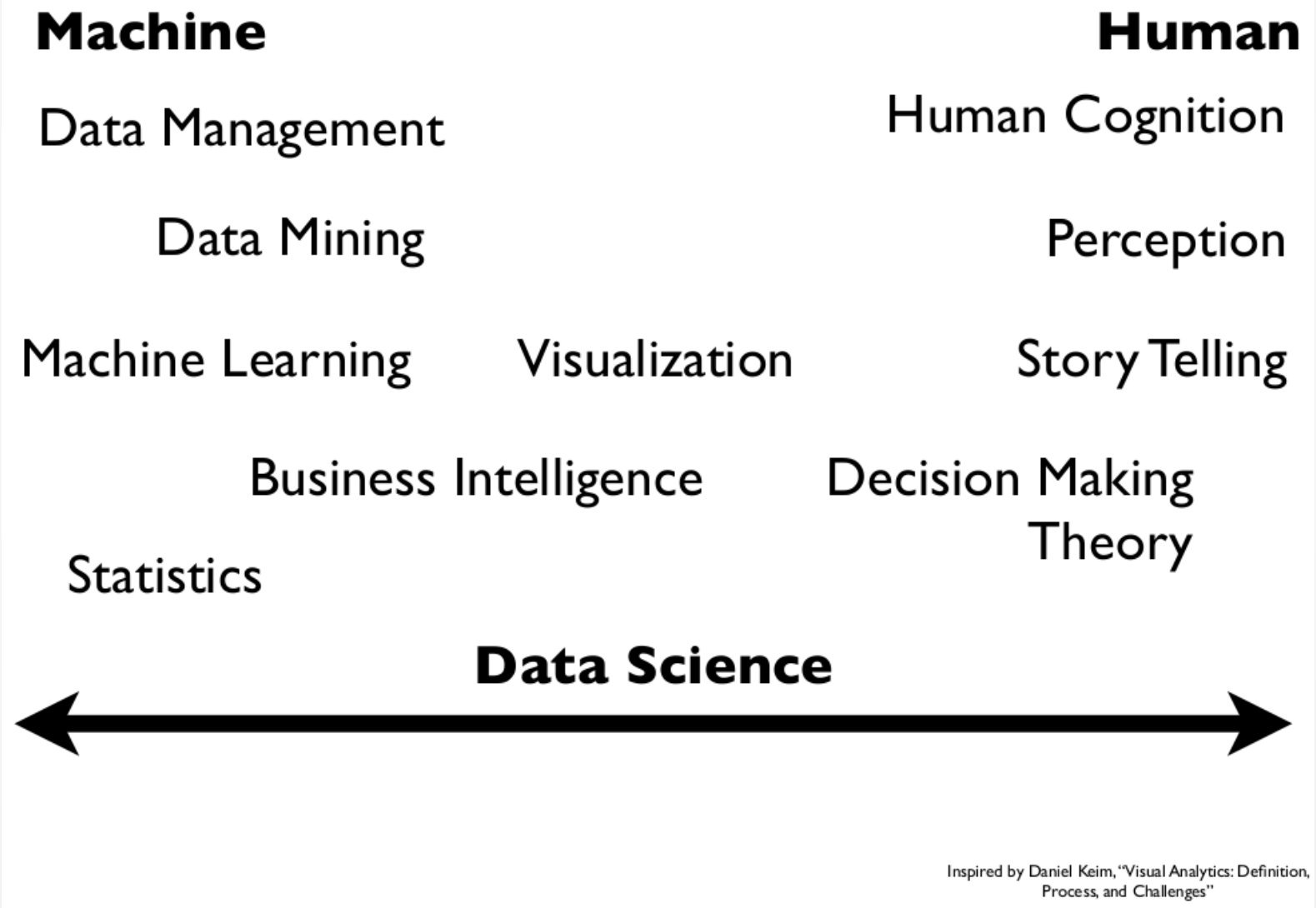
Getting Past the Hype

- There IS a gap between what has traditionally been taught in academia and the actual craft of analysis of large and complex data
  - Sanitized, nicely formatted and structured datasets of reasonable size is NOT reality
  - Exploratory data analysis and visualization given short shrift in stats classes
  - Academia tends to have disciplinary chimneys while the practice of data science and business analytics is inherently cross-disciplinary
- http://columbiadatascience.com/blog/
  - Cathy O'Neil's (one of the Doing Data Science authors) blog that complements her book and that continues to try to get us all past the hype
  - https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815

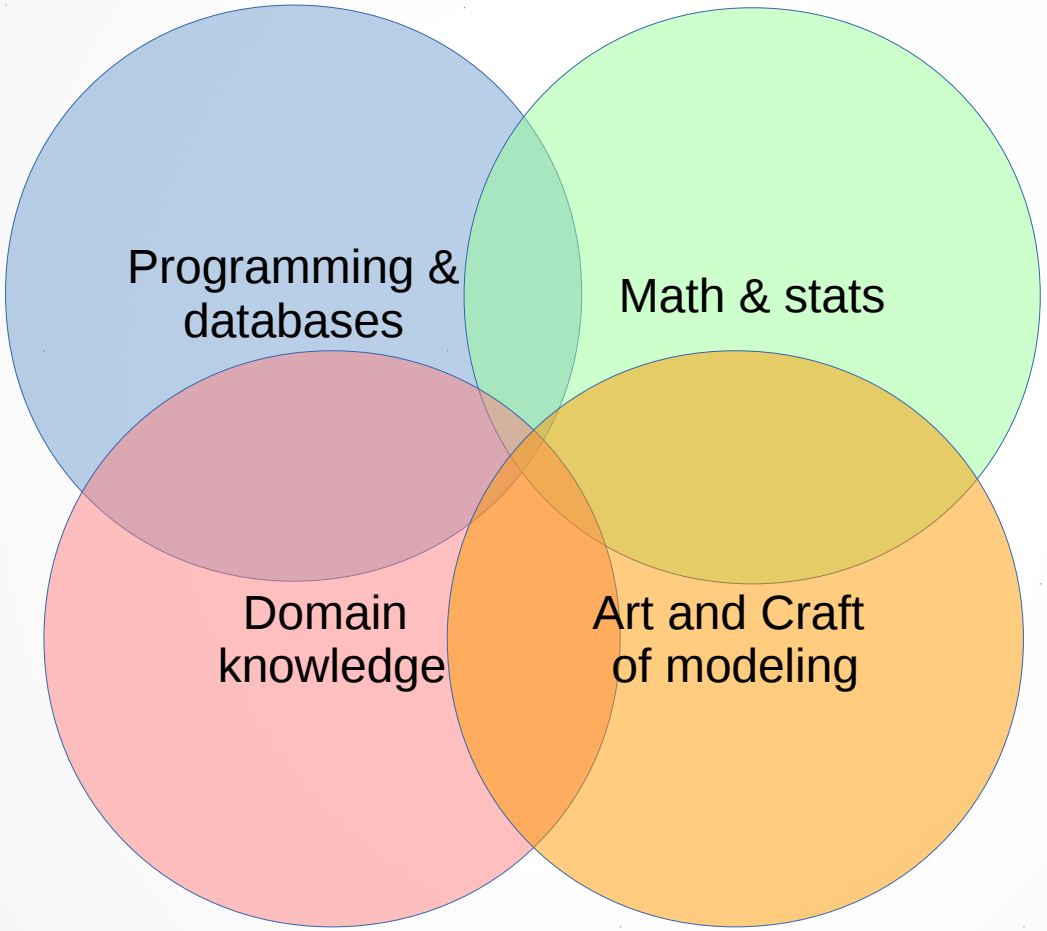
Conway's Data Science Venn Diagram



The machine - human continuum



Business analytics

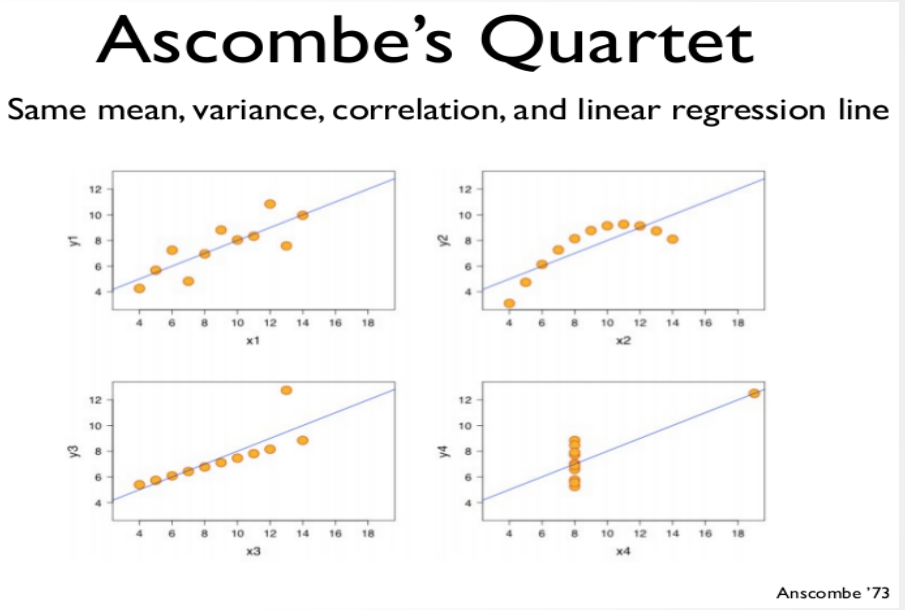


Historical perspective on the business analytics movement

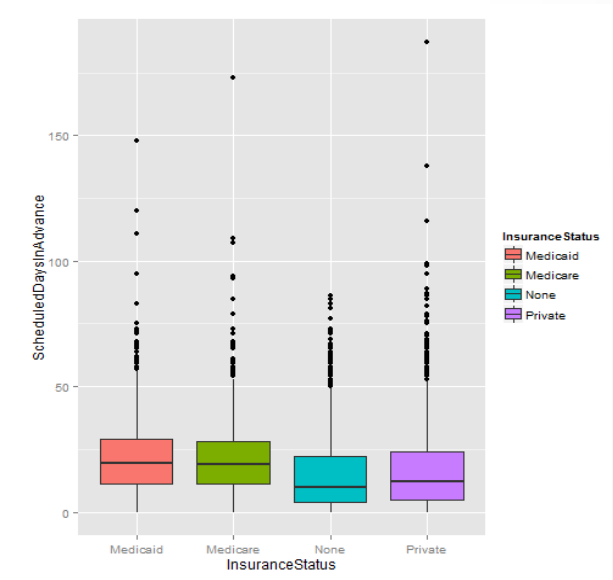
- Statistical analysis has been around for a long time
- Operations research out of WWII
- Decision support systems out of MIS
- Data warehousing and business intelligence
- Internet explodes
- Computing power gets cheaper and cheaper
- Machine learning & data mining out of CS
- “Competing on Analytics” by Davenport
- Connected devices and data sharing, “datafication”
- Commodity computing, cloud computing
- Nate Silver, the 538 Blog and 2012 election
- Data can be building block for *data products*
- Golden age for business analytics and data science

Descriptive Analytics

- Exploratory data analysis
- Statistics
- Data visualization

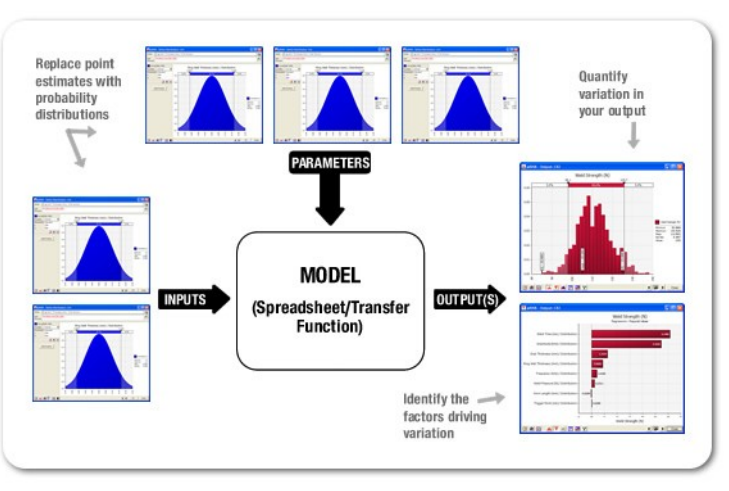


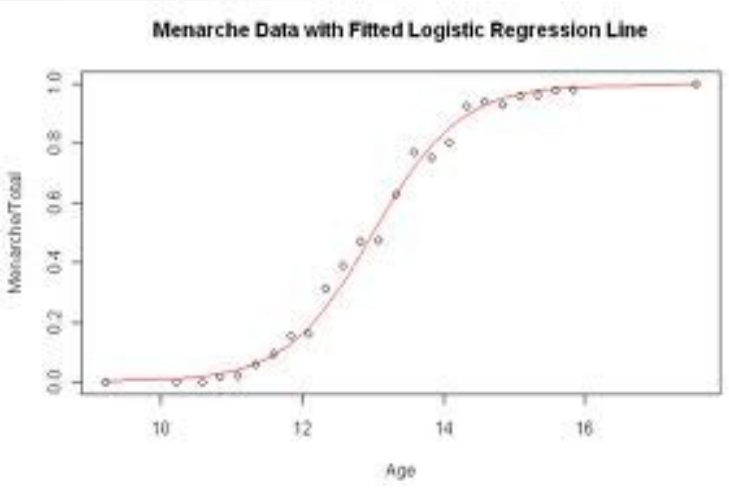
	Urgency	InsuranceStatus	mean_leadtime
1	Emergency	Medicaid	0.500
2	Emergency	Medicare	2.167
3	Emergency	None	1.880
4	Emergency	Private	2.318
5	FastTrack	Medicaid	148.000
6	FastTrack	None	4.500
7	FastTrack	Private	15.400
8	Routine	Medicaid	22.536
9	Routine	Medicare	21.786
10	Routine	None	15.946
11	Routine	Private	17.048
12	Urgent	Medicaid	2.467
13	Urgent	Medicare	2.240
14	Urgent	None	2.741
15	Urgent	Private	2.441

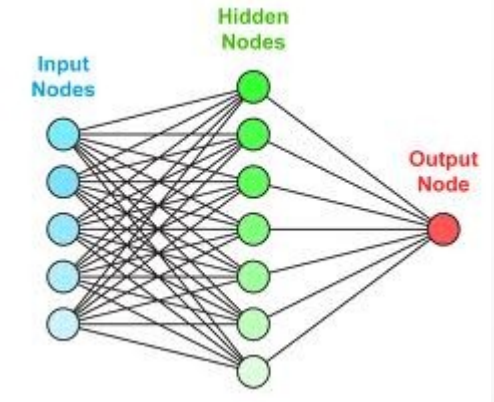


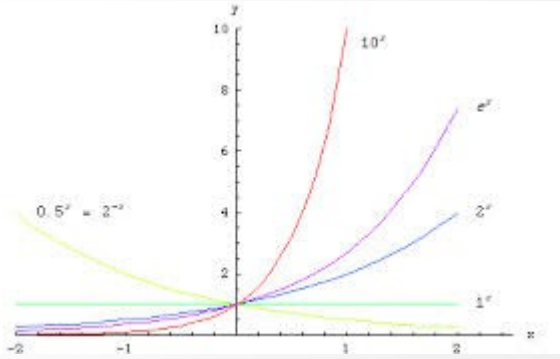
Predictive Analytics

- Statistical models
- Machine learning
- Simulation
- Other mathematical models





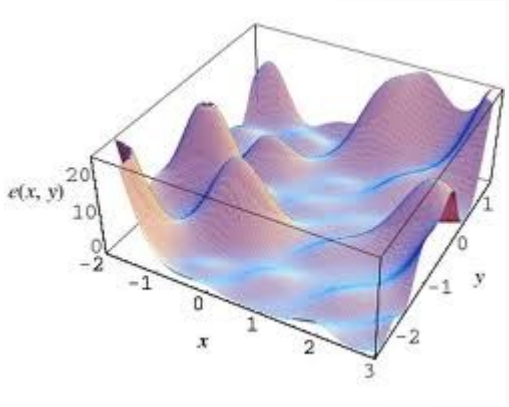


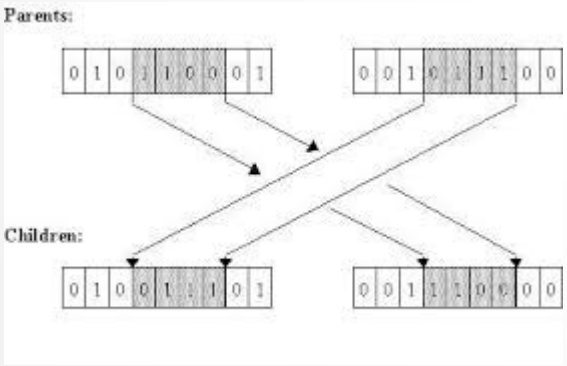


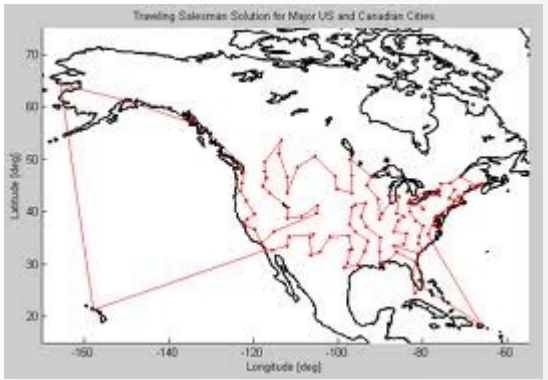
Prescriptive Analytics

- Optimization models and solution algorithms
- Optimizing heuristics

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{n} c_j x_j \\ \text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad i = 1, 2, \ldots, m \\ & x_j \in \mathbb{Z}^+, \quad j = 1, 2, \ldots, n \end{aligned}$$





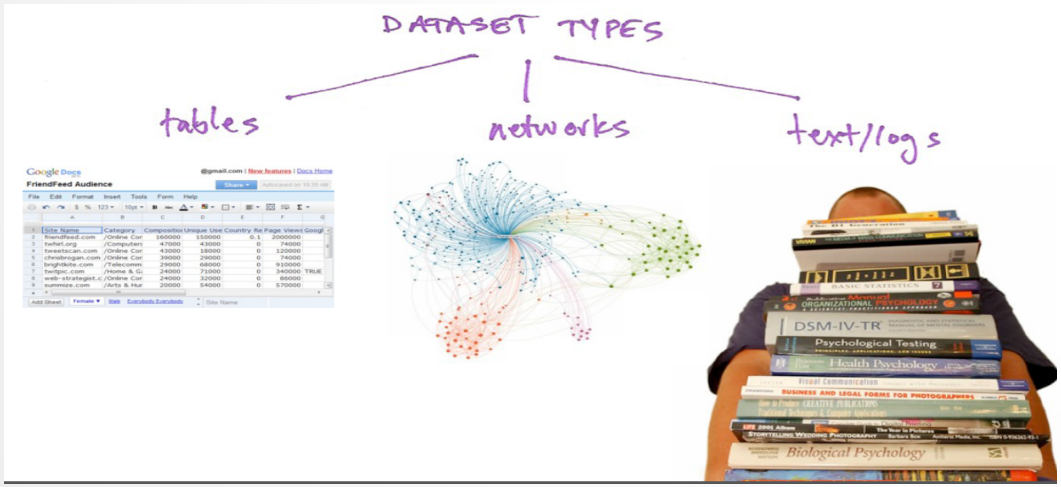


Beyond nice tables of data

Ben Shneiderman, 1996

- 1D (sequences)
- Temporal
- 2D (maps)
- 3D (shaped)
- nD (relational)
- Trees (hierarchical)
- Networks (graphs)
- Others?

The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization [Shneiderman, 96]



Data Access Schemes

- Bulk downloads
  - Wikipedia, IMDB, Million Song Database, etc.
  - See list of data web sites on the Resources page
- API access
  - NY Times, Twitter, Facebook, Foursquare, Google, ...
- Web scraping

JSON

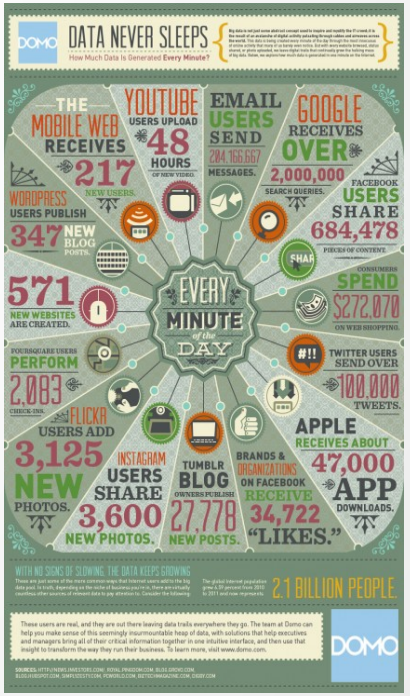
- Looks like Python dictionaries and arrays

```
{
    "kind": "grape",
    "color": "red",
    "quantity": 12,
    "tasty": true
}
```

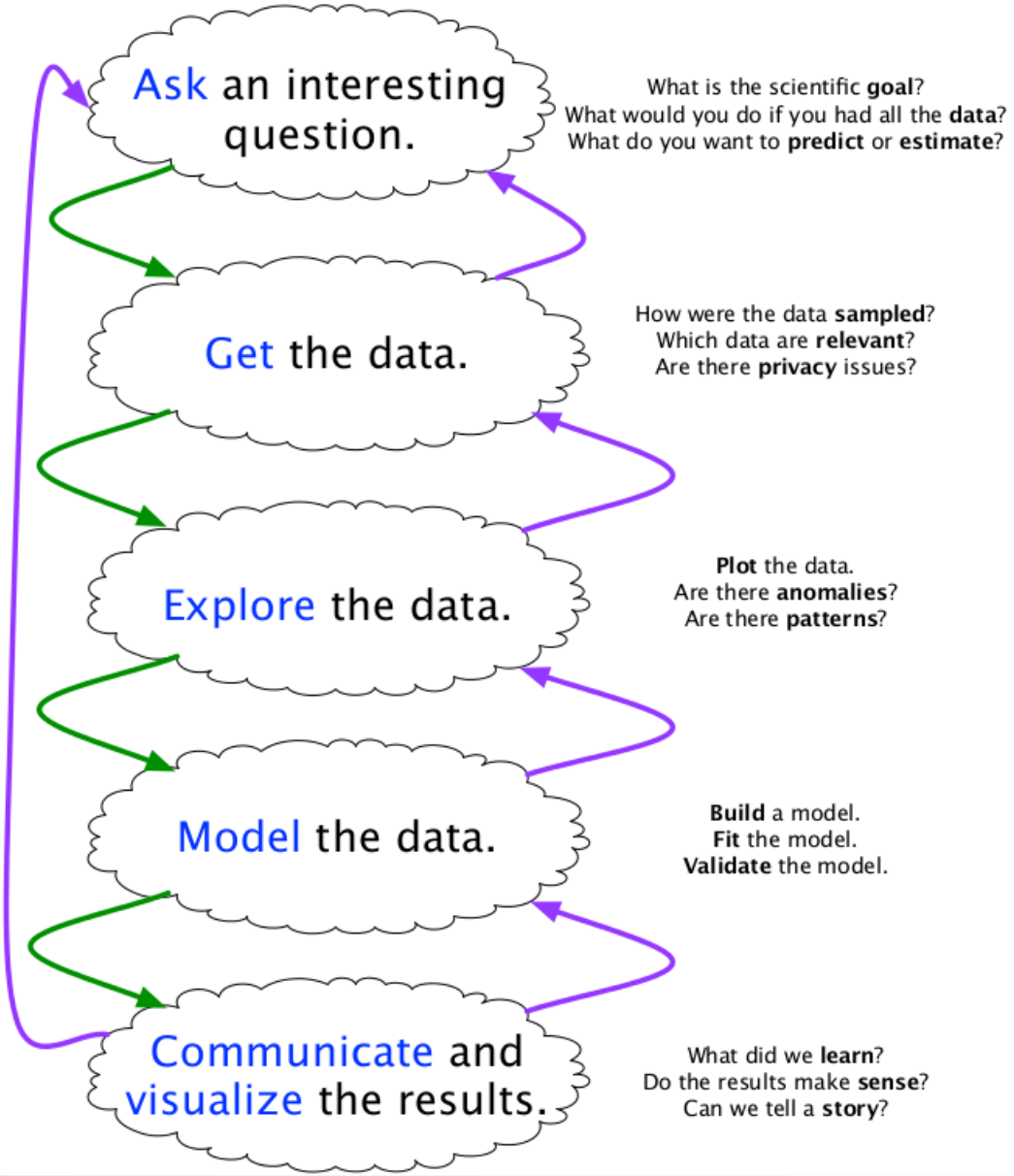
- Easy to parse in JS and Python

The Big Data problem

- Data, data, more data than a single machine can handle, data coming at you faster
  - **4V's**: Volume, velocity, variety, veracity
  - Twitter generates 1.39MB/s (~42Tb/yr)
  - Facebook sitting on PB's of data
  - Walmart: 1.5M trans/hr = 1.43GB/hr of data
  - **http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/**
  - **moving this data around and extracting info from this data is a challenge**



The Big Picture for Data Science
Abstractions AND ...



CS109 Key Facets

- *data munging/scraping/sampling/cleaning* in order to get an informative, manageable data set;
- *data storage and management* in order to be able to access data - especially big data - quickly and reliably during subsequent analysis;
- *exploratory data analysis* to generate hypotheses and intuition about the data;
- *prediction* based on statistical tools such as regression, classification, and clustering; and
- *communication* of results through visualization, stories, and interpretable summaries.

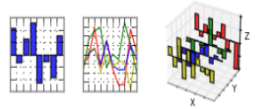
The Big Picture for Data Science

… Tools

IP[y]: IPython
Interactive Computing

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



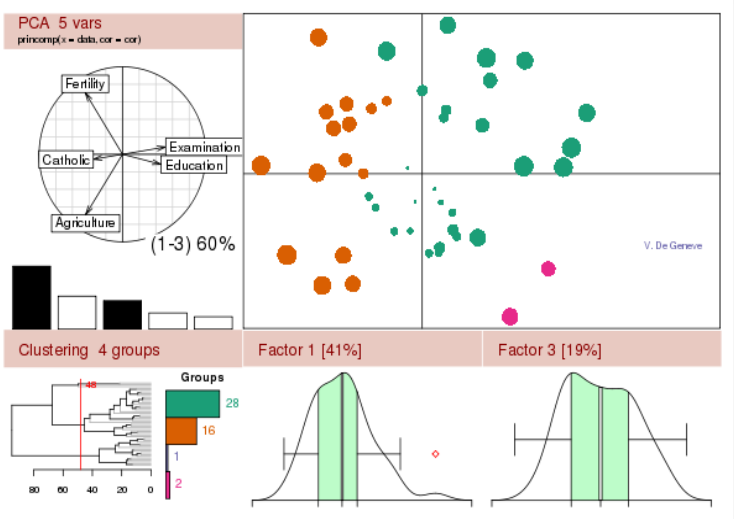
scikits learn
machine learning in Python

NumPy

SciPy.org ENTHOUGHT

matplotlib

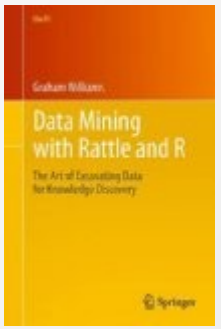
mrjob



R Studio

ggplot2

plyr
The split-apply-combine strategy for R

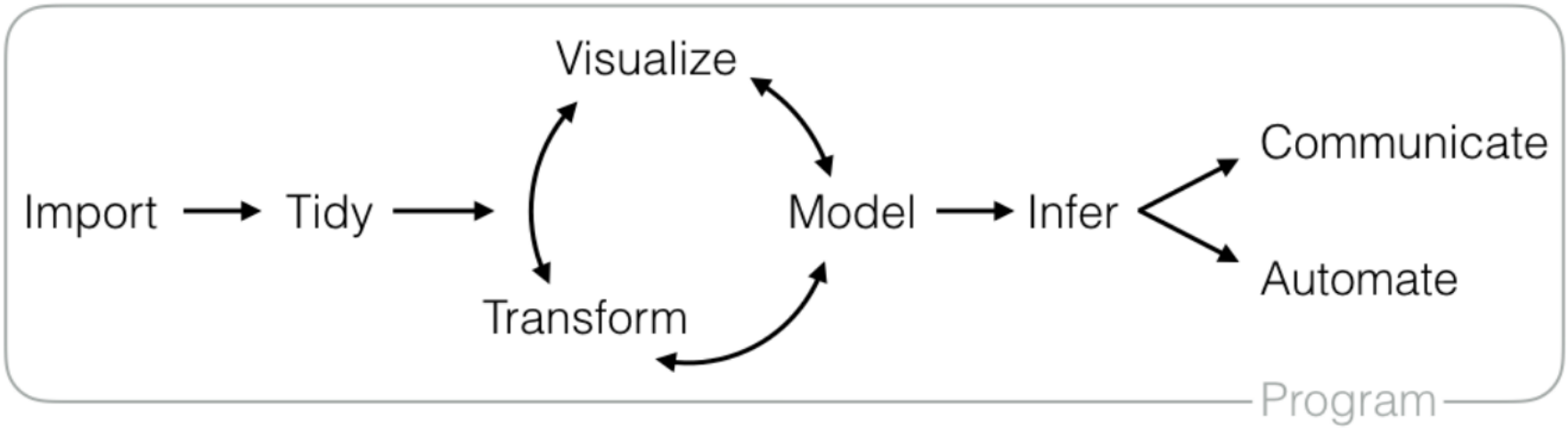


The Data Science workflow

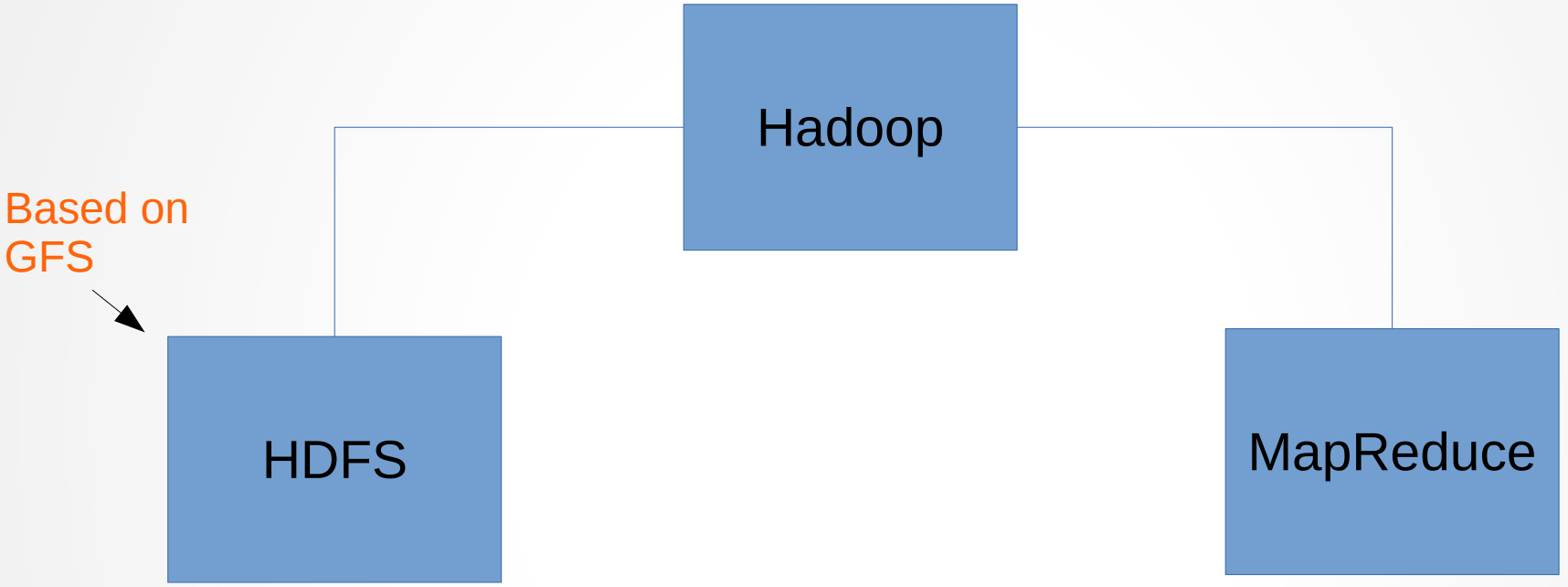
https://github.com/rstudio/RStartHere

Data Science Workflow

Each data science project is different, but each follows the same general steps. You:



So, about this “big data” thing...
Hadoop = HDFS + MapReduce



Hadoop Distributed File System

A fault tolerant, scalable, distributed file system that recovers gracefully from hardware and/or software failures and provides consistent results and performance

A programming model that lends itself to parallelization over a large cluster of commodity computers.

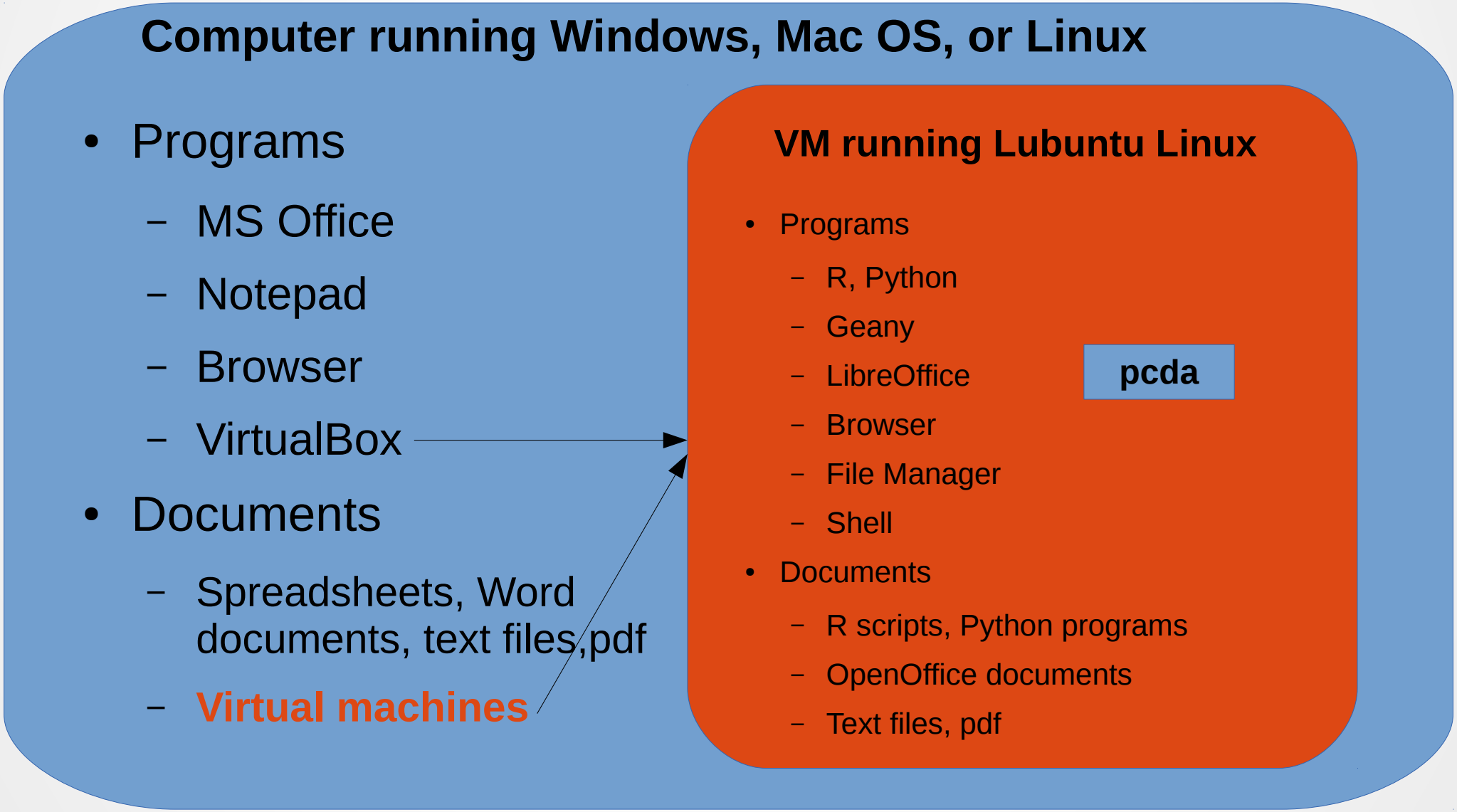
Syllabus and Moodle Site Review

- Overview
- Motivation
- Books and website
- Structure
- Requirements

Our computing “appliance” - **pcda**

- I created a virtual machine that is used via VirtualBox
- We will all start out with the same software
- You get a chance to learn some Linux
- Tools like R and Python are multi-platform
- You can download appliance and run on your own computer but will also be able to use in the lab

pcda – The Big Picture



Lubuntu Linux

- Lubuntu is a fast and lightweight operating system developed by a community of Free and Open Source enthusiasts.
- The core of the system is based on Linux and Ubuntu.
- Lubuntu uses the minimal desktop LXDE, and a selection of light applications.
- We focus on speed and energy-efficiency. Because of this, Lubuntu has very low hardware requirements.

For those new to Lubuntu Linux

- Ubuntu is widely used and the most Windows like Linux “distro”
  - I've been using it for years on multiple machines
  - Lubuntu is just a super lightweight version of Ubuntu
- It's got a GUI that is very Windows like
- It's got a Software Center that makes it easy to find, download and install new programs
- It's totally FREE, both as in freedom and beer
- There is a large Ubuntu user community

Why learn to use Linux for analytics?

- Linux widely used in the data science and analytics world
- Linux shell FAR superior to Windows command line application
  - Powerful shell scripting language
  - Tab completion
  - Command line is often way more efficient than GUI
- Linux is free and open source
- Sets you apart from other business analysts who only know Windows and Microsoft applications

Why R and Python?

- Both R and Python are widely used in the data science and business analytics worlds
- A quote from Enterprise Data Analysis and Visualization: An Interview Study on the growing need for technically adept analysts:

  > When discussing recruitment, one Chief Scientist said “analysts that can’t program are disenfranchised here”

- Both support a combination of interactive use via tools like R Studio and Ipython/Jupyter along with programmatic use via text scripting
- Huge communities and ecosystems supporting R and Python for analytics work
- Both facilitate reproducible analysis
- Some things that are simply hideously difficult to do in tools like Excel or a database, are simple in R and/or Python
  - Group By or Pivoting type analysis for operations such as percentiles
  - Small multiples and other complex graphing/charting/plotting
  - Documenting and reproducing complex series of data cleaning and transformations

First peek at R and Python

- Both can be used in multiple modes
  - “scripts” or “programs”
  - Interactive
- This is really useful for data science work
- In Downloads you’ll find a folder called `peek_r_python`
  - Let’s explore a bit...

What I hope to do in this class

- Introduce you to a new approach to doing data analysis and business analytics work
- Motivate you to continue to learn to use tools like R and Python and to become a competent analytics programmer
- Truly challenge you
- Give you a chance to apply these tools and approaches to a data set of your own choosing

Stuff to do now

- HW0_IntroToPCDA
  - In-class exercise for becoming familiar with **pcda** and some GUI oriented Linux things
  - Intro to data science
  - prep for next class
- Explore our Moodle site
- This class will move at a brisk pace.
  - Important you do the prep work for each class
  - The more comfortable you are with these new tools, the more valuable class time will be