Northeastern University

INFO 6105
Data Sci Eng Methods & Tools
Lecture 1

6 January 2020

Do *now*: Standalone R

- **Install the R environment:**
  - **https://cran.rstudio.com/**

Do *now*: RStudio IDE

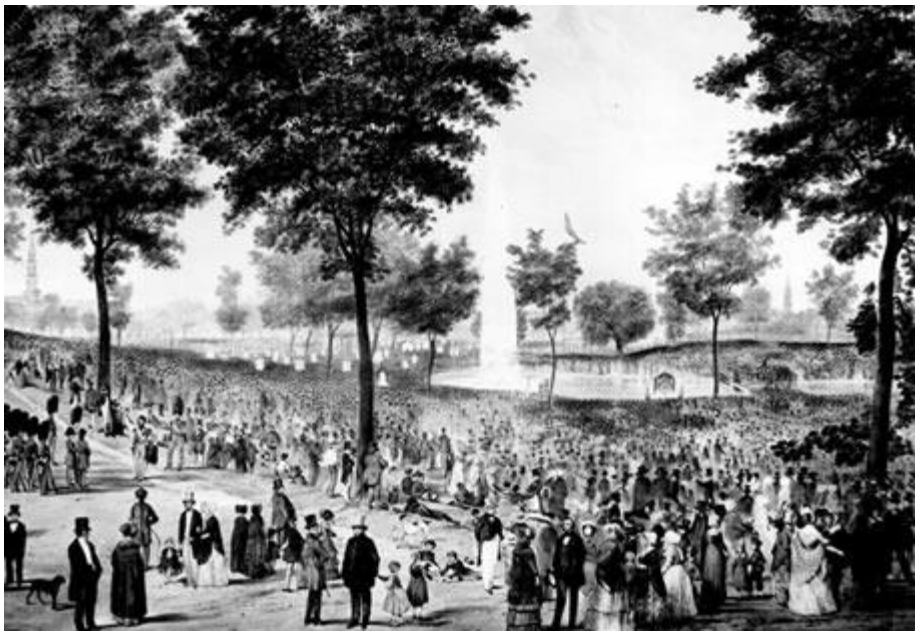
- **Install *RStudio* (IDE for using R)**
  - **Install for your appropriate system from the list at: http://www.rstudio.com/products/rstudio/download/**

Peloponnesian War

- In the 5th century BCE. The historian *Thucydides* in his History of the Peloponnesian War describes how the Athenians calculated the height of the wall of Platea by counting the number of bricks in an unplastered section of the wall sufficiently near them to be able to count them
- The count was repeated several times by a number of soldiers. The most frequent value so determined was taken to be the most likely value of the number of bricks
- Multiplying this value by the height of the bricks used in the wall allowed the Athenians to determine the height of the ladders necessary to scale the walls
- The Mean

Statistics

- By the 18th century, the term *statistics* designated the systematic collection of demographic and economic data by states
- The birth of statistics is often dated to beginning 18th century when statistical and census methods provided a framework for modern demography that involved giving probabilities of survival to each age

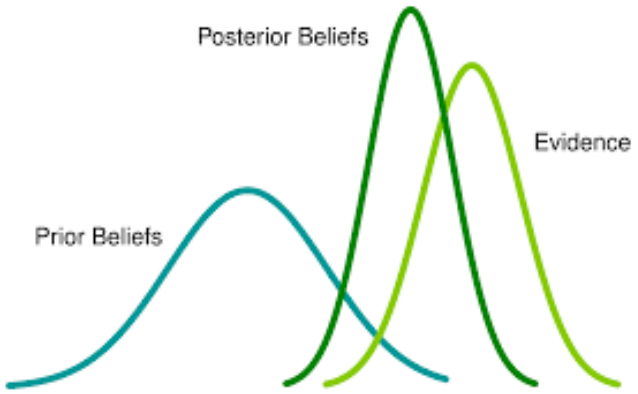


Modern Statistics

- The modern field of statistics only emerged in the early-20th century in three stages
- The first wave, at the turn of the century, was led by the work of *Karl Pearson*, who transformed statistics into a rigorous mathematical discipline used for analysis, not just in science, but in industry and politics as well
- The second wave reached its culmination in the insights of *Ronald Fisher*. This involved the development of better design of experiments models and techniques for use with *small data samples*
- The final wave, which mainly saw the refinement and expansion of earlier developments, emerged from work by the son of Karl Pearson, with modern concepts of the *confidence interval*, *statistical hypothesis testing*, and the *null hypothesis*

Bayesian Statistics

- The term Bayesian refers to Thomas Bayes (1702–1761), who proved that *probabilistic limits* could be placed on an unknown event
- Later on, Pierre-Simon Laplace (1749–1827) introduced what is now called *Bayes' theorem* and applied it to celestial mechanics, medical statistics, reliability, and law
- In the 1980s, there was a dramatic growth in research and applications of Bayesian methods, mostly attributed to the discovery of *Markov chain Monte Carlo (MCMC)* methods, which removed many of the computational problems
- Despite growth of Bayesian research, most undergraduate teaching is still based on frequentist statistics. Nonetheless, Bayesian methods are widely accepted and used, such as in the field of *machine learning (ML)*



Statistics today

- Today, statistical methods are applied in all fields that involve decision making in the face of uncertainty
- For example, should I ask this person out on a date because i'm not sure if they like me?
- And in Machine Learning (ML)..

Foundations

- Machine Learning is built on top of statistics & Linear Algebra
- You can study ML without a foundation in statistics and Linear Algebra, but that would be a little bit like learning how to fly a plane in a flight simulator
- You never develop the quality of airworthiness, and over time, this will show and hinder your professional development
- This class is dedicated to teaching you this airworthiness by giving you a very solid foundation in statistics and linear algebra, and then showing you how it is applied in the field of Machine Learning

Class

- This class has no *prerequisites*, and it's a great class to learn python, but you will be competing with other students who may already know python, so having some experience in a managed programming language like java, C#, or Python is almost *necessary*
- There will be homework every week, a midterm, a final, and a final project
- You will have experienced TAs to help you, and an experienced instructor to motivate you

LIKE A DIVA.

Part 1

DATA SCIENCE IS SEXY

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

MENU Data Scientist: The Sexiest Job of the 21st Century

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 12 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see possibilities. He began forming theories, testing hunches, and finding patterns that allowed him to predict whose networks a given profile would

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

The sexiest job of the 21st Century

- *George Roumeliotis, the head of a data science team at Intuit in Silicon Valley, holds a doctorate in astrophysics*
- *Roumeliotis begins his search for data scientists by asking candidates if they can develop prototypes in a mainstream programming language. Roumeliotis seeks both a skill set—a solid foundation in, statistics, probability, linear algebra, and computer science—and certain habits of mind. He wants people with a feel for business issues and empathy for customers.*
- *Pay will of course be a factor. A good data scientist will have many doors open to him or her, and salaries will be bid upward*
- *Survey of the priorities of data scientists revealed something more fundamentally important. They want to be "on the bridge." The reference is to the 1960s television show Star Trek, in which the starship captain James Kirk relies heavily on data supplied by Mr. Spock. Data scientists want to be in the thick of a developing situation*

On the bridge

The sexiest job of the 21st Century

- *Data scientists say they want to build things, not just give advice to a decision maker. One described being a consultant as "the dead zone — all you get to do is tell someone else what the analyses say they should do". By creating solutions that work, they can have more impact and leave their marks as pioneers of their profession*
- *Hal Varian, the chief economist at Google, is known to have said, "The sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"*
- *The advance of big data shows no signs of slowing. The more data we have, the more we need data scientists to analyze it*

The tools..

- *Foundations in **probability & statistics**, **linear algebra**, and **computer science***
  - ***Programming***

- ***Not** apps you can download*
  - ***Using***

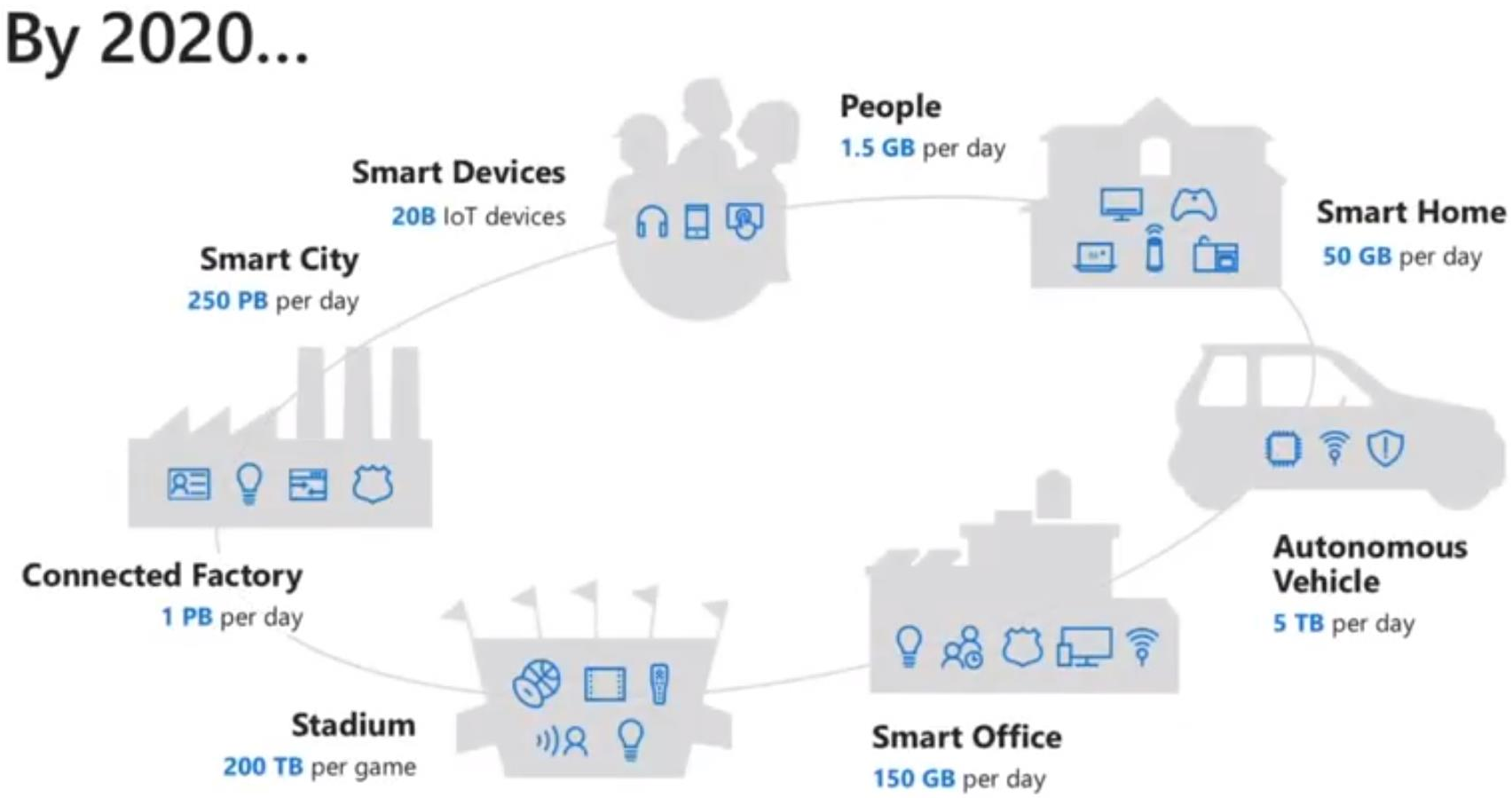
After you take this class..

- ..you will date the partner of your dreams ☺

But first..

- 1. Linear Algebra- Vector Spaces and Norms, Basis, Tensors, Vector and Tesnor Operations, Singular Value Decomposition (SVD), Eigendecomposition of a matrix, LU Decomposition, QR Decomposition/Factorization, Symmetric Matrices, Orthogonalization/Orthonormalization, Matrix Operations, Projections, Eigenvalues & Eigenvectors
- 2. Probability & Statistics - Probability Rules & Axioms, Bayes' Theorem, Random Variables, Variance and Expectation, Conditional and Joint Distributions, Standard Distributions (Bernoulli, Binomial, Multinomial, Uniform and Gaussian), Moment Generating Functions, Maximum Likelihood Estimation (MLE), Prior and Posterior, Maximum a Posteriori Estimation (MAP) and Sampling Methods
- 3. Multivariate Calculus- Differential and Integral Calculus, Partial Derivatives, Vector-Values Functions, Directional Gradient, Hessian, Jacobian, Laplacian and Lagrangian Distribution
- 4. Algorithms and Complex Optimizations- Complexity theory, data structures (Binary Trees, Hashing, Heap, Stack etc.), Dynamic Programming, Randomized & Sublinear Algorithm, Graphs, Gradient/Stochastic Descents

Data Science Because..



Machine Learning is..

- ***Statistics***
  - ***Finding patterns in data***
  - Covariates become ***features***, non-linear regression becomes ***neural networks***, transformations of random variables becomes ***normalizing flow***, placing a gaussian prior on parameters becomes ***L2 regularization,*** *etc.*
- *For the opposing viewpoint, read here:*
  - *https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3*

Data Science with Machine Learning..

- Before *Tensorflow*, *Keras*, and *Torch*..
- You will become familiar with *NumPy*, *Pandas*, *SciPy*, and *SciKit-learn*
- Along with knowledge of *probability & statistics*, *and linear algebra*
  - Not simple, but *necessary*
- *And you will learn black-belt python along the way*

Why Python

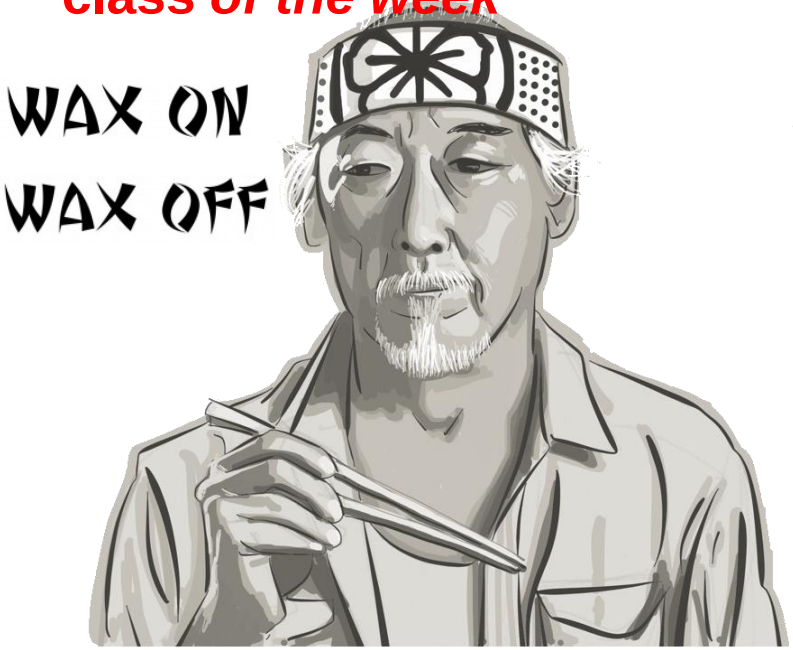
- Part of Python's success is ease of integration with C, C++, and Fortran code, the other part is that it *hides* its OO heritage
- As a result, Python has inherited Fortran legacy
  - Fortran used to solve ODEs and PDEs
  - Terse, functional and not very Object Oriented
  - All legacy Fortran code now in Python
- Python has become the de-facto language of scientists

Why R?

- **Easier than Python**
- **Lots of libraries written for R**
- **An overview of class today**
  - Today's class we learn how to program with *rows* and *columns*
- ***Start working with vectors and matrices***
- **R** first appeared in 1995 and served as an implementation of the **S** statistical programming language
- Roger Peng, an 18-year R programming veteran who teaches **R** both at the university and on Coursera notes: *"**R** is the most popular language used in the field of statistics"*
  - *"I like **R** because it's very easy to program"*
  - **R**'s advantages include its package ecosystem. *"The vastness of package ecosystem is definitely one of R's strongest qualities -- if a statistical technique exists, odds are there's already an **R** package out there for it"*
  - *"There's a lot of functionality built in that's built for statisticians"*

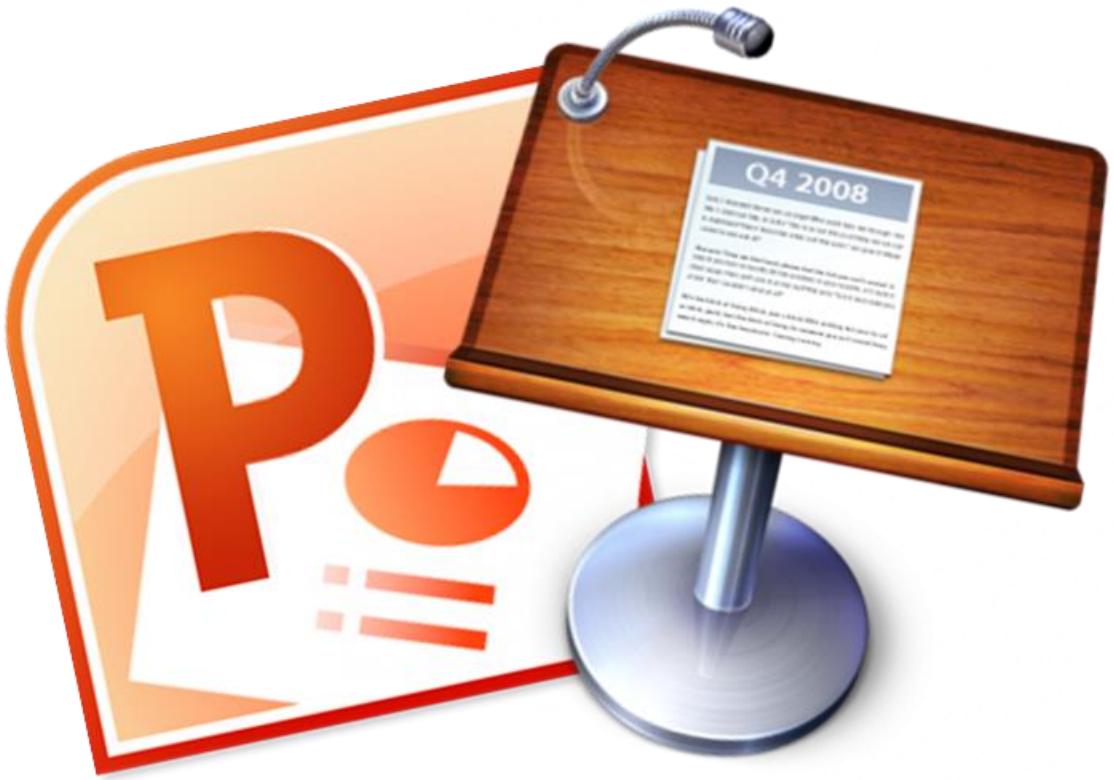
Class Policy

- **Grade: 30% homework, 30% mid-term, 30% final project, 10% final exam**
  - **Why do universities have grades? To separate the good from the bad?**
    - **No, to improve your *skills***
- **Do your homework, ask your TA how to deliver, due before *first* class *of the week***



Questions, Slides

- Ask questions: email TAs or dino.k@northeastern.edu
- Lecture slide-decks on Blackboard, about an hour before lecture



Part 2

ABOUT ME

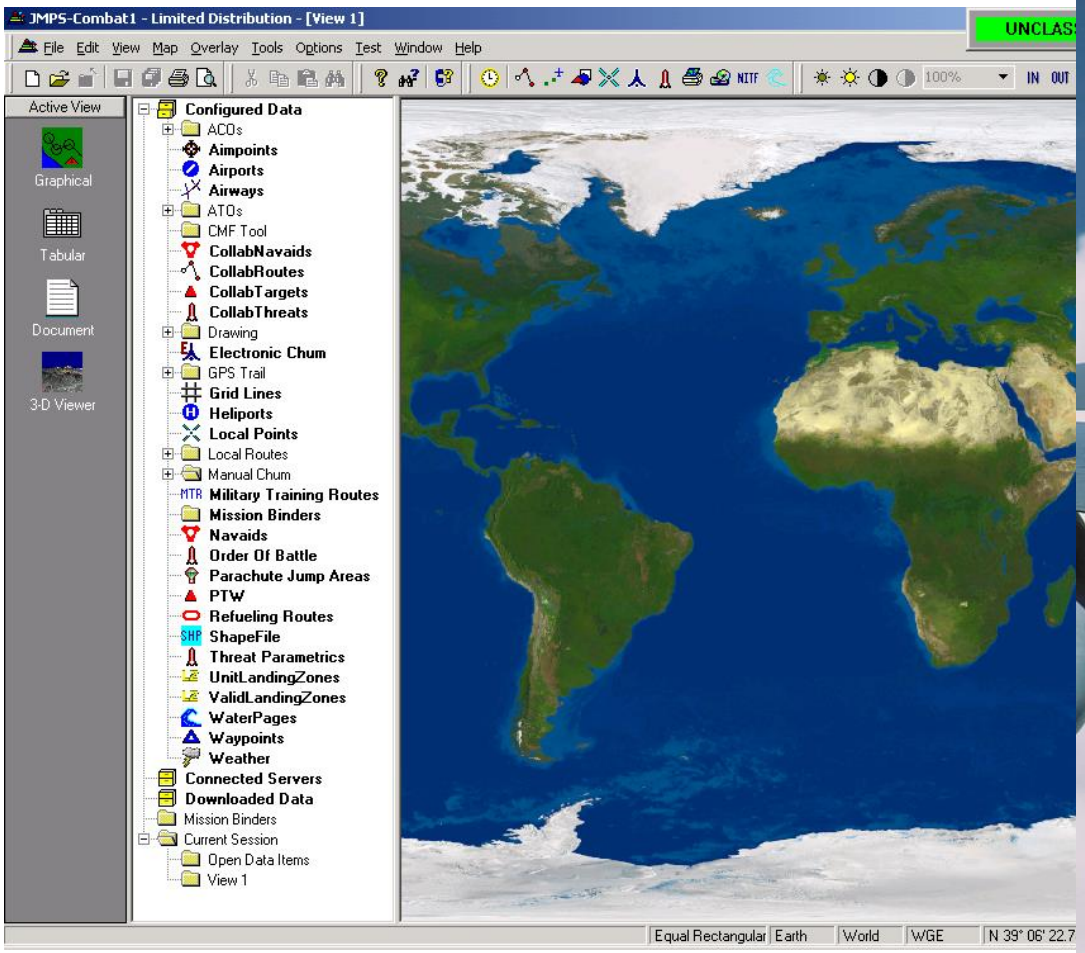
Me

Industry

- Xerox, Operating System research (headless)

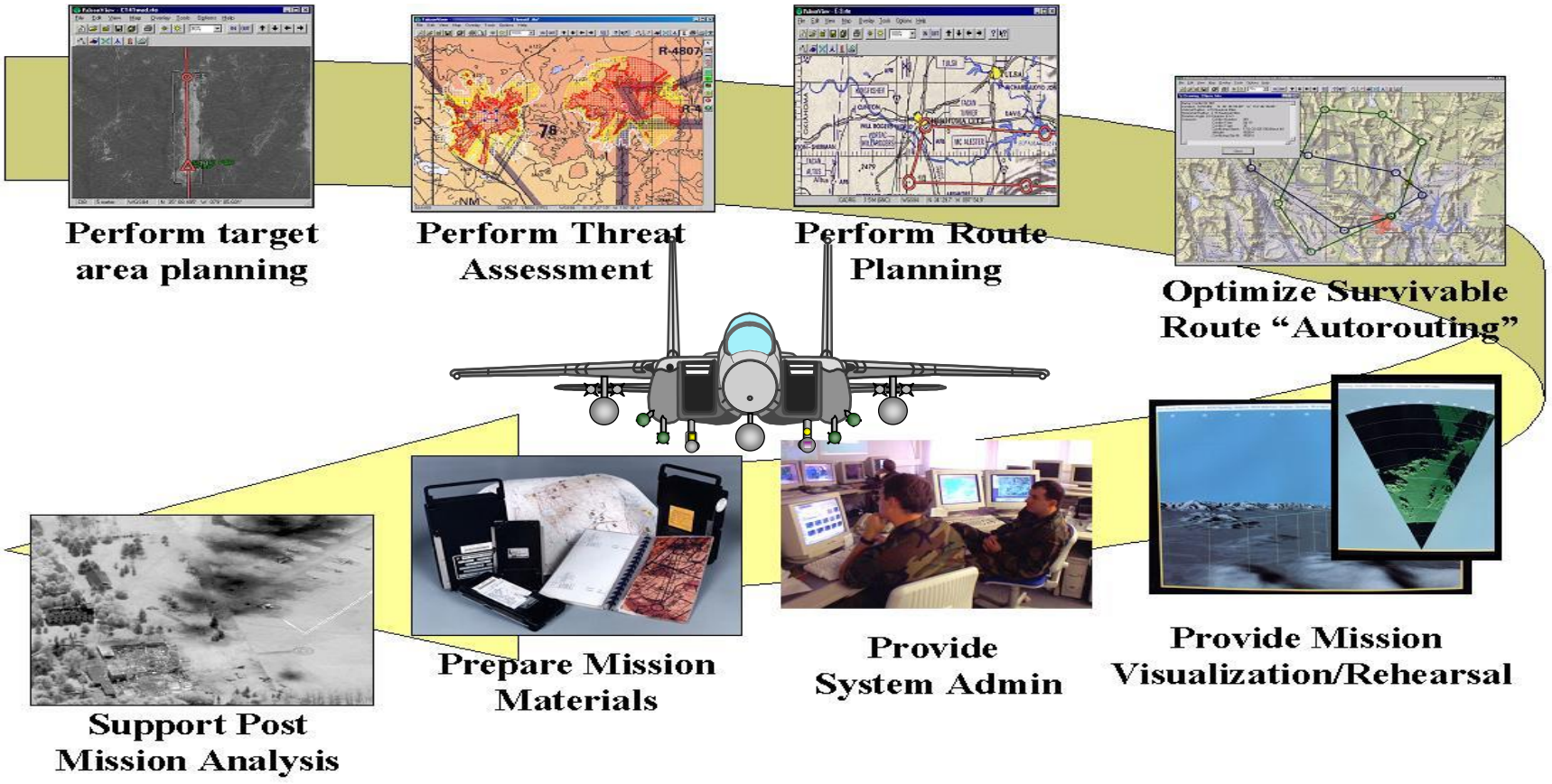
xerox ®

Government


JMPS-Combat1 - Limited Distribution - [View 1]
UNCLAS
File Edit View Map Overlay Tools Options Test Window Help
Active View
Graphical
Tabular
Document
3-D Viewer
Configured Data
ACOs
Aimpoints
Airports
Airways
ATOs
CMF Tool
CollabNavaids
CollabRoutes
CollabTargets
CollabThreats
Drawing
Electronic Chum
GPS Trail
Grid Lines
Heliports
Local Points
Local Routes
Manual Chum
Military Training Routes
Mission Binders
Navaids
Order Of Battle
Parachute Jump Areas
PTW
Refueling Routes
ShapeFile
Threat Parametrics
UnitLandingZones
ValidLandingZones
WaterPages
Waypoints
Weather
Connected Servers
Downloaded Data
Mission Binders
Current Session
Open Data Items
View 1
Equal Rectangular
Earth
World
WGE
N 39° 06' 22.7


© Ted Carlson

Mission Planning

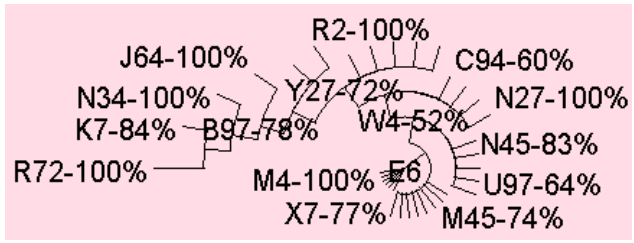


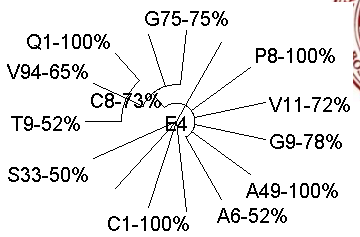
Army

- Use of microwave sensors for imaging continues to grow for applications such as target detection, airport screening, soil moisture estimation, etc.
- Hence, the need for indoor/small theater microwave imaging facilities
- A multistatic system is one where multiple transmit sites or multiple receive sites, or both, are used to construct the radar picture

IRS


INTERNAL
REVENUE
SERVICE

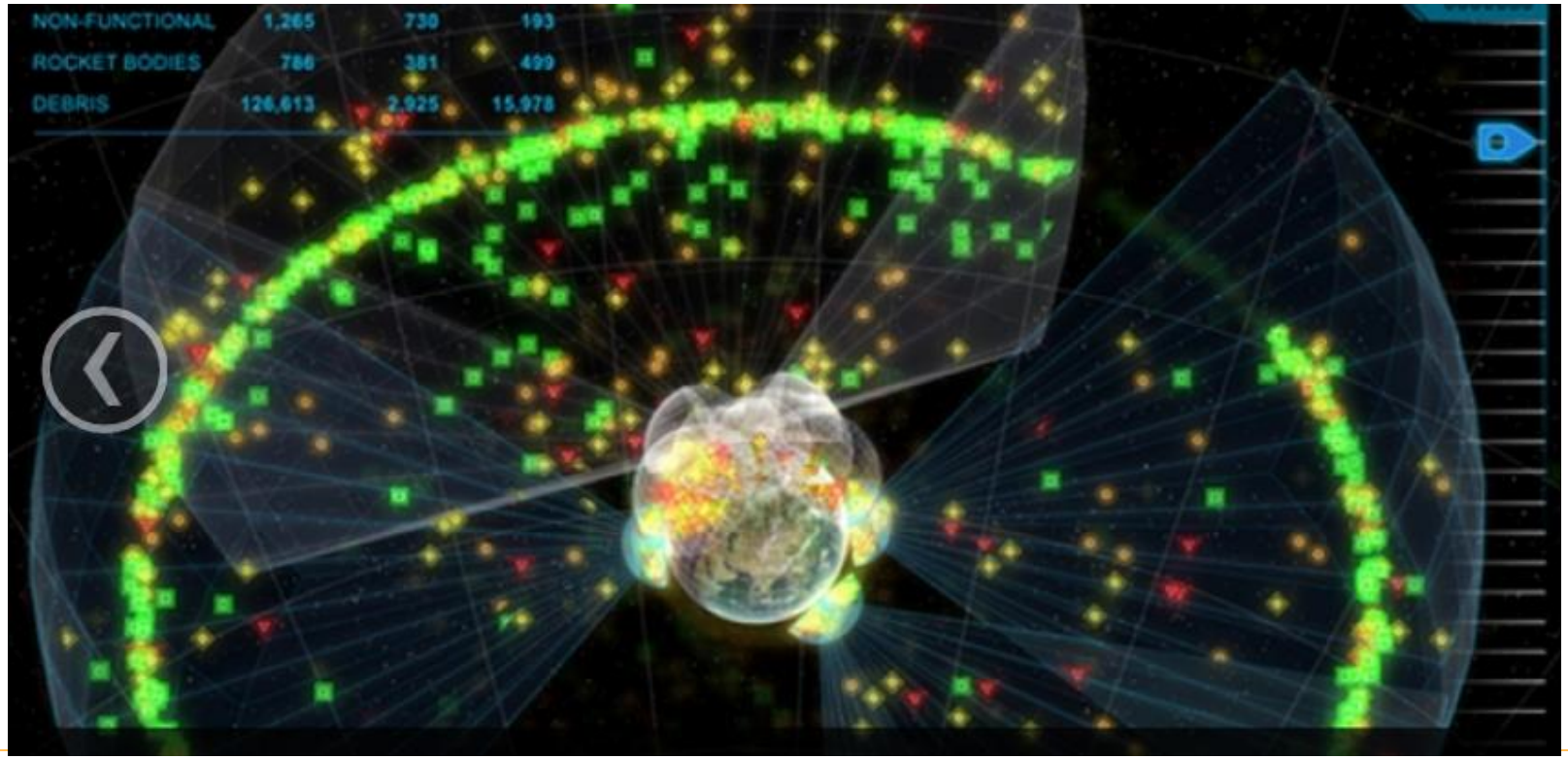

R2-100%
J64-100%
C94-60%
N34-100%
Y27-72%
N27-100%
K7-84%
B97-78%
W4-52%
N45-83%
R72-100%
M4-100%
F6
U97-64%
X7-77%
M45-74%


G75-75%
Q1-100%
P8-100%
V94-65%
C8-73%
V11-72%
T9-52%
F4
G9-78%
S33-50%
A49-100%
C1-100%
A6-52%

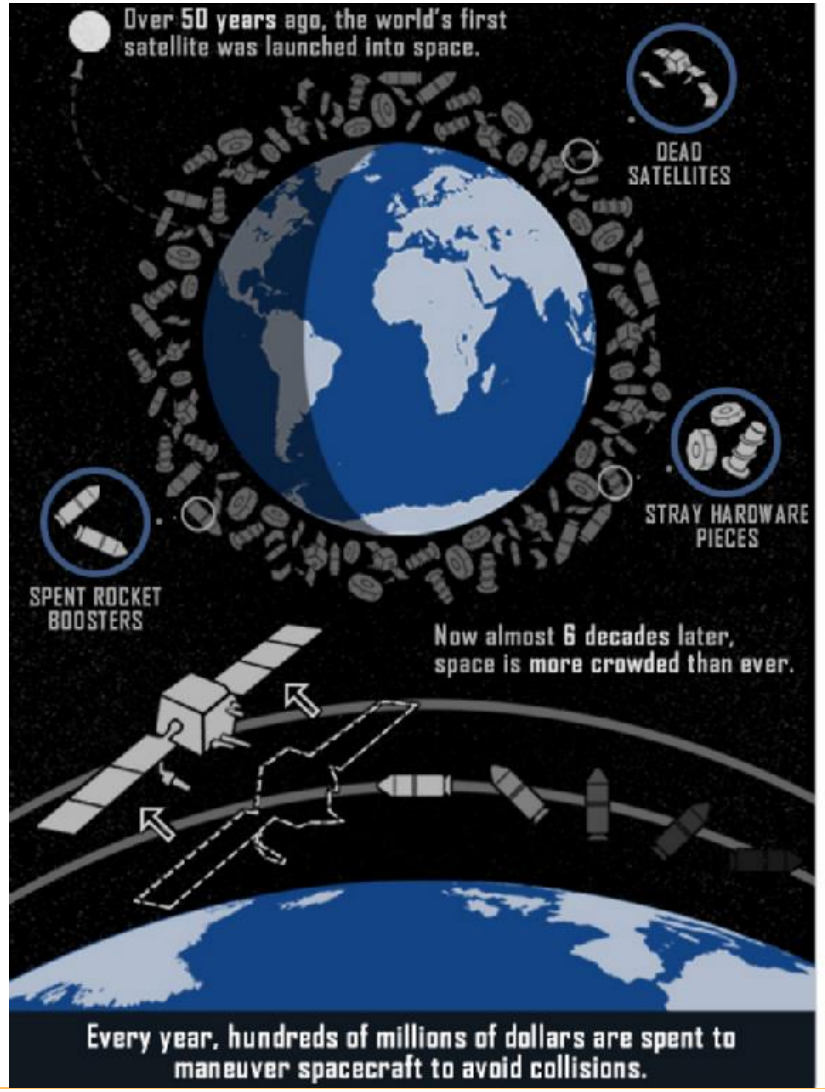
X55-16%.1
J38-17%.1
N96-36%.2
X97-72%
D4-48%.2
B5-100%
F9-18%.4
T1-100%.14
S4-1%.15
F9-5%.11
Z91-89%.15
C51-97%.15
T8-8%.90
Y58-79%.6
J4-55%.31
F4
N58-30%
L41-95%.3
T1-100%.27
S4-44%
Y47-33%.27
Y58-21%.39
N6-24%
H3-17%.16
Y58-79%.3
J4-55%.12
W6-15%.43
T8-8%.44
Z81-14%.28
T9-100%.27
V2-100%.27
C38-20%.7
Y58-21%.40
G55-44%.26
N58-45%
M4-16%
W6-15%.86
F6-32%.12
T1-100%.26
U4-96%
C23-86%.25
G55-44%.25
A53-5%.26
C8-100%
J38-17%.2
D8-49%.37
V9-98%.25
T1-100%.25
X55-16%.2
T9-100%.26
L2-39%.11
M54-93%
V5-100%.2
S55-76%.3
Y4-41%.2
C23-86%.28
S5-33%.25
H3-17%.25
D4-48%.3
Z81-14%.25
W6-15%.82
J4-55%.25
V43-59%.118
H3-40%.25
K8-100%.82
P61-88%.2
D8-49%.36
Y58-21%.34
V43-59%.133
T1-100%.24
R5-82%.2
N96-64%.2
T8-92%.133
W6-15%.80
C51-97%.24
S5-33%.23
A5-54%.2
R5-7%.3
H3-36%.81
K8-100%.76
K8-100%.73
L41-95%.5
W9-73%.6
A5-46%.6
L2-59%.2
Y2-100%-cycle.4
L2-2%.4
D8-49%.38
V43-59%.134
T8-92%.127
N58-25%.24
Y58-79%.5
F9-5%.14
F9-5%.13
H3-40%.24
K8-100%.72
J9-92%.6
C38-20%.6
Y2-100%.4
Y47-33%.19
T8-92%.134
T8-92%.129
C23-86%.26
S5-33%.22
F6-32%.10
C38-80%.4
Y58-79%.4
Y47-33%
D8-51%.40
W6-15%.71
F9-18%.6
Z91-89%.19
F6-32%.9
T8-92%.131
T8-92%.126
G55-44%.23
S4-55%.10
V2-100%.19
G55-44%.18
V43-59%.130
T8-92%.121
Y47-33%.22
H3-17%.26
K8-100%.70
J4-18%.9
C51-97%.19
W6-15%.52

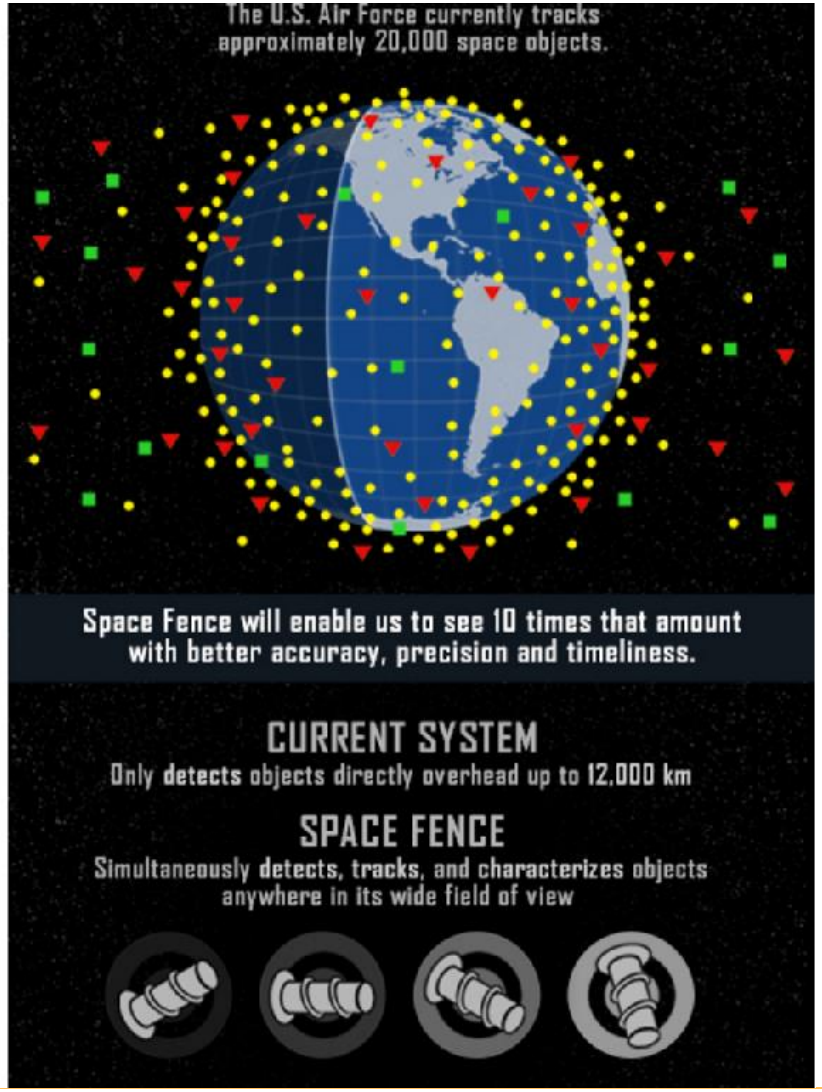
USAF Space Fence

- http://www.lockheedmartin.com/us/products/space-fence.html
- **Kwajalein Atoll in the Marshall Islands: Space Fence, the most powerful radar ever built, to identify and tracks objects in space. Initial operational capability 2018**

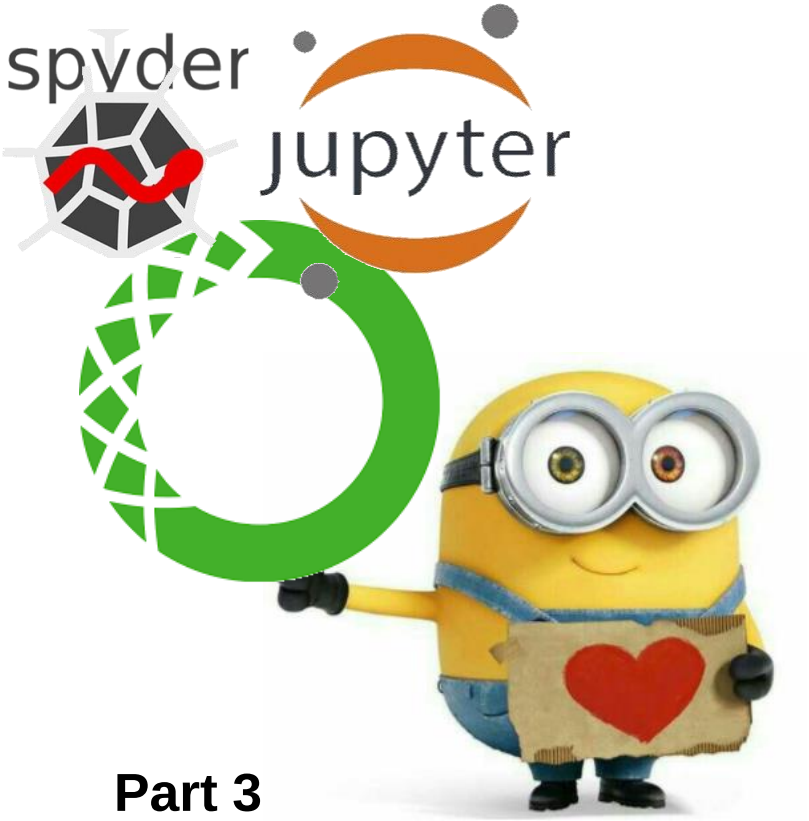


USAF Space Fence





Also



Part 3

INSTALLING ANACONDA PYTHON (*NOT* NOW, IT'S *YOUR HOMEWORK!)*

Anaconda

- **Enterprise-ready Python distribution with packages for Big Data processing, predictive analytics, and scientific computing**
- **https://www.continuum.io/anaconda**

ANACONDA
Powered by Continuum Analytics®

Install Anaconda

- Install anaconda, from:
  - https://www.continuum.io/downloads
  - https://repo.continuum.io/archive/index.html
- Option: You may install `miniconda` instead, if you don't have enough disk space
  - Pick Python 3.x distro
- Test drive anaconda by running all steps in:
  - http://conda.pydata.org/docs/test-drive.html
  - Learn to play with different python environments and manage packages

To create new conda environments

- **You are going to create <u>new</u> environments in this class:**
  - `conda create --name theano python=3.4`
    - **Because theano does not support python 3.5!**
- **Activate it:**
  - **Windows:** `activate theano`
  - **OSX/Linux:** `source activate theano`
- **List all & default (*) environment:**
  - `conda info –envs`
- **Python version?**
  - `conda version`

Python IDEs

- Anaconda *spyder*
- Visual Studio with *Python Tools for Visual Studio* (PVTS)
- Visual Studio Code
- PyCharm
- PyDev (on top of Eclipse)
- Wing IDE
- Ninja IDE
- https://docs.continuum.io/anaconda/ide_integration

Python notebook

- **Install notebook (for miniconda):**
  - `conda install notebook`
  - **That will install a number of dependencies (incl. ipython and jupyter)**
  - **It will allow us to do python in class like we did R in RStudio..**
- **Full conda should already have notebook installed**

Anaconda IDE

- `conda install spyder`
- **At a command prompt:** `spyder`
- **(like** `code`**)** ☺
- **Full anaconda should already have spyder installed, but probably old version**
  - **Update with:**
    - `conda update spyder`
  - **Windows: Administrator privileges console**
  - **OSX/Linux:** `sudo`

Anaconda add-ons

- Commercial packages from *Continuum Analytics* and other vendors into Anaconda:
  - `IOPro`: fast, memory-efficient Python interface for databases, data files, Amazon S3 and MongoDB
  - `Accelerate`: includes `NumbaPro`, a compiler that targets multi-core CPUs and GPUs directly from simple Python syntax
  - `MKL Optimizations`: accelerates NumPy, SciPy, scikit-learn and NumExpr using Intel's Math Kernel Library
- All commercial packages from *Continuum Analytics* are available for a free 30 day trial
- The Anaconda Add-Ons are free for individual Academic use
  - Visit the Anaconda Academic page to request an Academic license

On the Cloud: Wakari

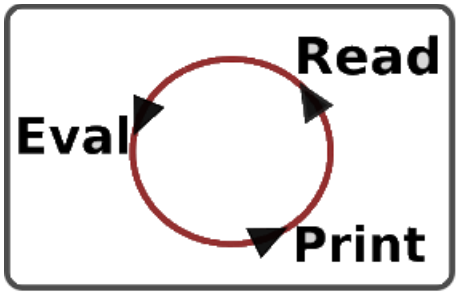
- **Web-based Python environment for collaborative data analysis, exploration and visualization**
  - **You can upload, create, and publish IPython Notebooks easily from your browser, and Wakari has Anaconda already installed**
  - **Create a free account for a full Python environment in the cloud**
- **http://wakari.io**

Google Colaboratory

- **Free Jupyter notebook environment that requires no setup and runs entirely on the Cloud**
- **https://colab.research.google.com/notebooks/welcome.ipynb**
- **Free GPU support**
  - **Able to cut down training time from few hours to well below 5 mins**
  - **Direct integration with Google drive, where you can download a lot of data or just upload your own**
  - **Share the link for the notebook and anyone can import it and run it for themselves**

How to run ipython notebook in a browser

- In a command shell: `ipython notebook`
- That will pop up a browser
- You cut and paste code in the `In[]` cells on the iPython page, and then click on the 'Play' button to run each cell
- The great part about the seamless integration of text and code in IPython Notebook is that it's entirely conducive to the process:
  1. Form hypothesis
  2. Evaluate data
  3. Form conclusion
- It's a REPL loop, and we'll use that to do ML

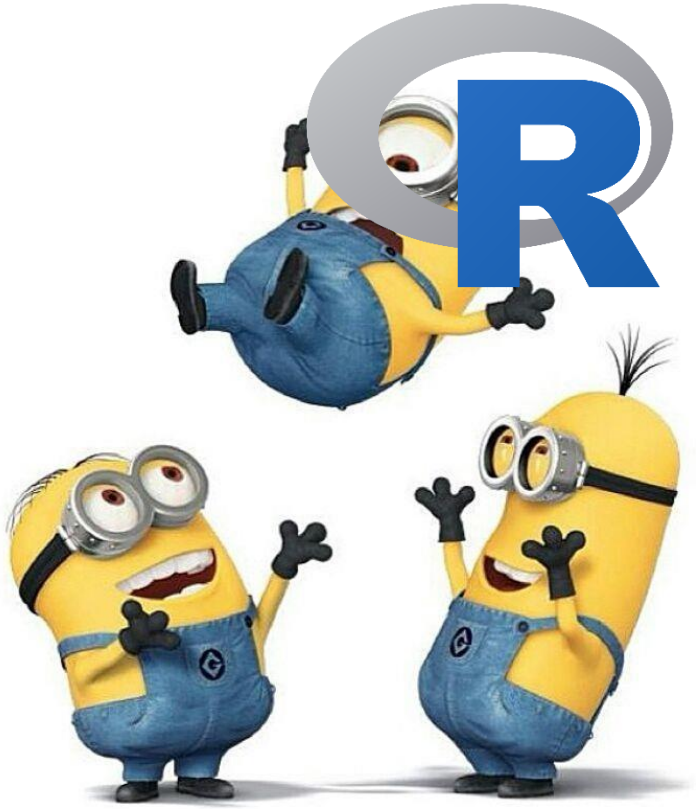


To load a notebook..

- .. That you downloaded from the Web:
  - Open up a command shell
  - Go to the folder where you downloaded your notebook
  - `ipython notebook mynotebook.ipynb`
- You can only run one notebook server at a time..
  - u..Unless you change the port for ipython, because the default port for conda is 8888

An important URL

- **Unofficial Windows Binaries for Python Extension Packages**
  - **https://www.lfd.uci.edu/~gohlke/pythonlibs**
  - **E.g. https://www.lfd.uci.edu/~gohlke/pythonlibs/#python-ldap**



Part 4

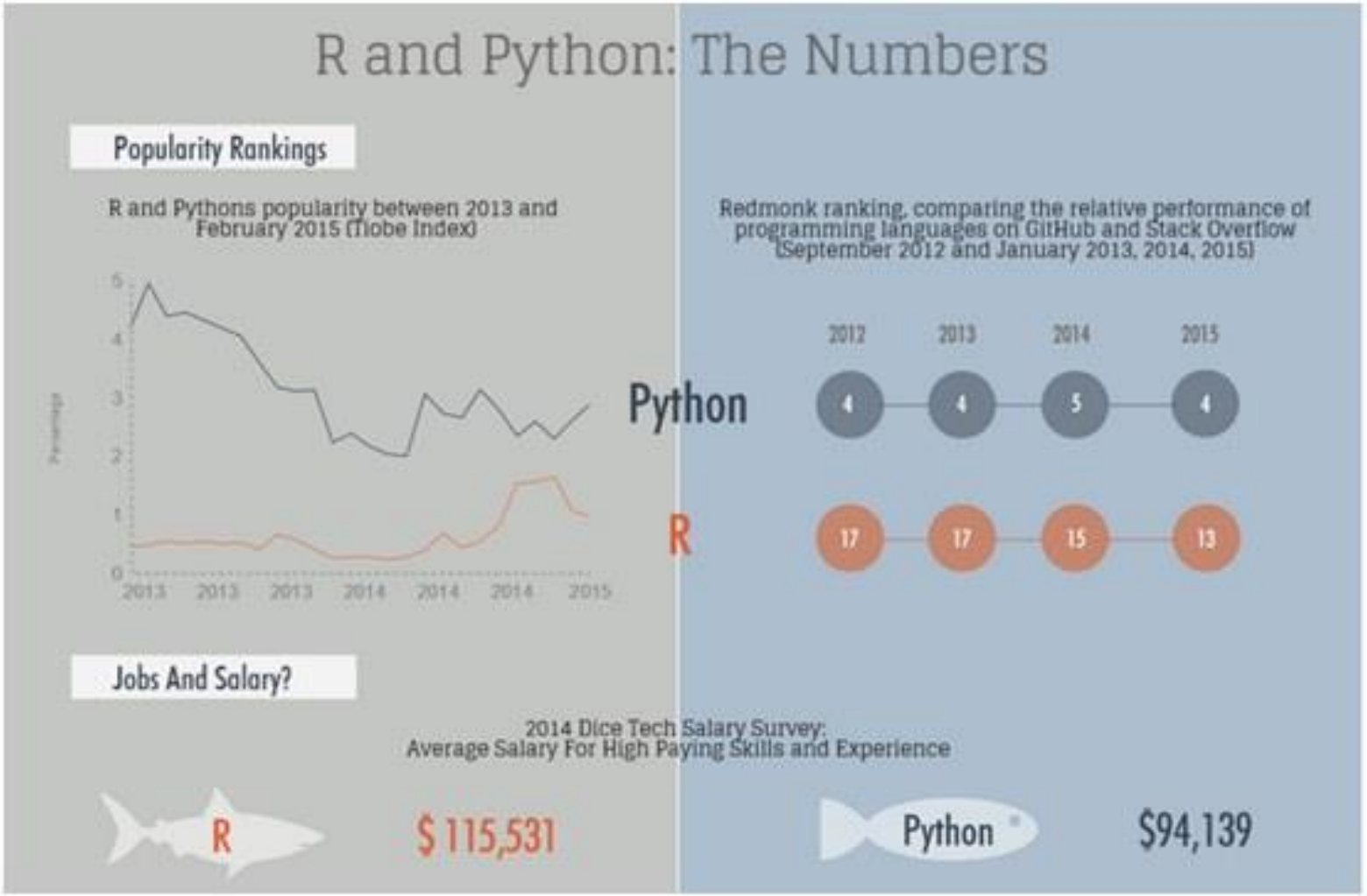
MACHINE LEARNING WITH R

Introduction to Statistical Modeling with R

R Introduction

- Ross Ihaka and Robert Gentleman created the open-source language R in 1995 as an implementation of the *S programming language*
  - Purpose was to develop a language that focused on delivering a better and more user-friendly way to do data analysis, statistics and graphical models
  - CRAN is a huge repository of curated R packages to which users can easily contribute
    - Comprehensive R archival network
    - https://cran.r-project.org/

R vs Python



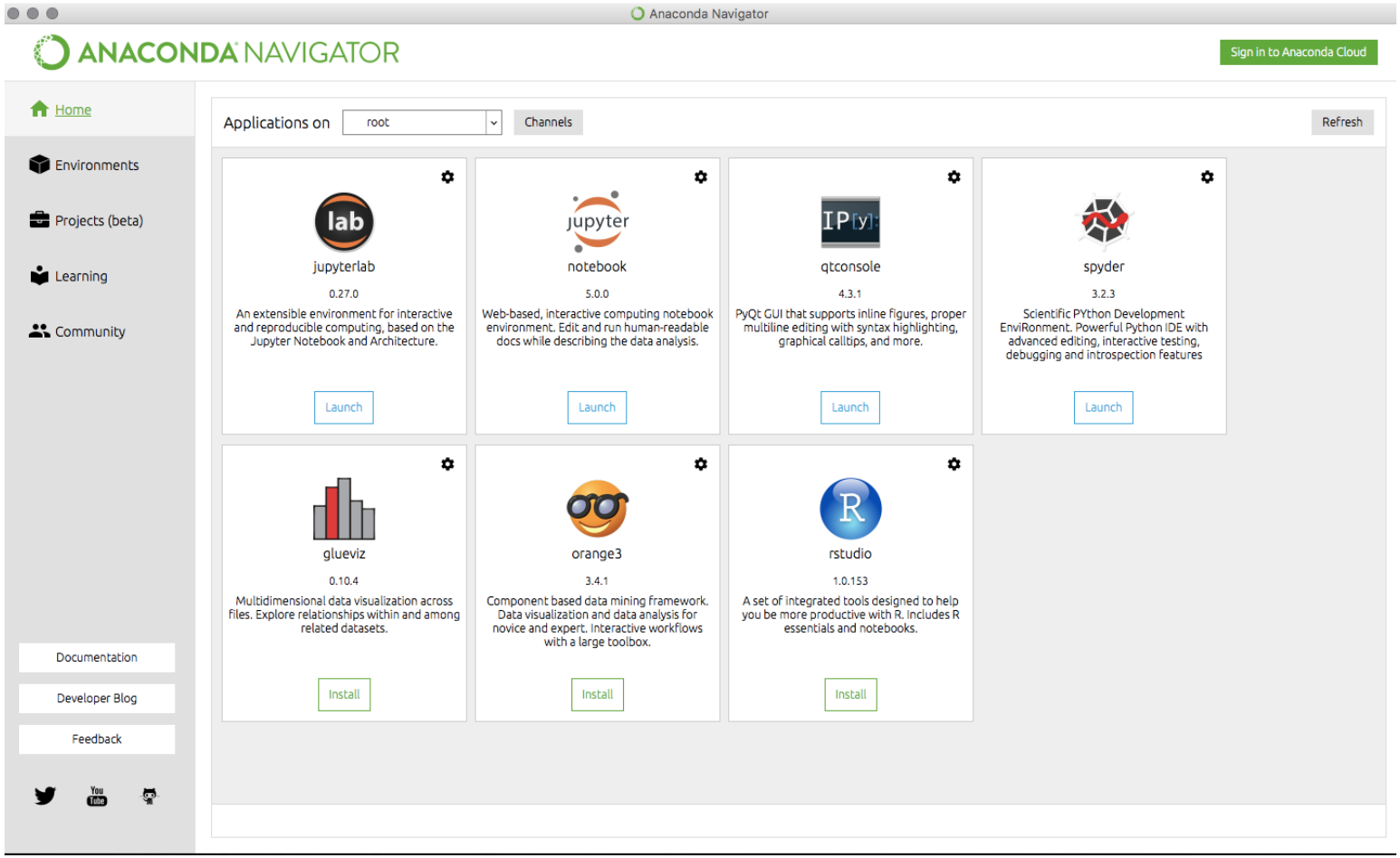
Source www.kdnuggets.com

What is R

- **R is a free software environment for statistical computing, data mining, and graphics**
  - **Hopefully replace `Matlab`, the crack-cocaine of scientists-engineers-turned-programmers**
- R is one of the major languages for data science
  - It provides excellent visualization features, which is essential to explore the data before submitting it to any automated learning, as well as assessing the results of the learning algorithm
  - Many R packages for machine learning are available off the shelf and many modern methods in statistical learning are implemented in R as part of their development
- **We use statistical analysis for:**
  - **inference - making conclusions based on data**
  - **prediction - what will happen when I observe new data?**
  - **and we create models to do both of those things**

To install R and RStudio, Prerequisites

- **Install it with Anaconda!**



Other option: Standalone R & RStudio

- **Installing the R environment:**
  - https://cran.rstudio.com/

RStudio IDE

- **Install *RStudio* (IDE for using R)**
  - **Install for your appropriate system from the list at: http://www.rstudio.com/products/rstudio/download/**

R = or ← ?

In R, both statements **`x = 3`** and **`x <- 3`** have the effect of assigning the value 3 to the variable x. So if they have the same effect, does it matter which you use?
When R (and S before it) was first created, `<-` **was the only choice for the assignment operator.** Old AT&T keyboards had arrow as a key!
But R uses = for yet another purpose: associating function arguments with values (as in `pnorm(1, sd=2)`, **to set sd to 2)**
To make things easier for new users, R added the capability in 2001 to *also* allow = be used as an assignment operator, on the basis that the intent (assignment or association) is usually clear by context. So,

```
x = 3
```

clearly means "assign 3 to `x`", **whereas**

```
f(x = 3)
```

clearly means "call function `f`, setting the argument `x` to 3".
There is one case where ambiguity might occur: if you wanted to assign a variable *during* a function call. The only way to do this in modern versions of R is:

```
f(x <- 3)
```

which means "assign 3 to `x`, **and call f with the first argument set to the value 3**

APL Character Set

IBM 2741 keyboard layout

Part 5

INTRODUCTION TO STATISTICAL LEARNING

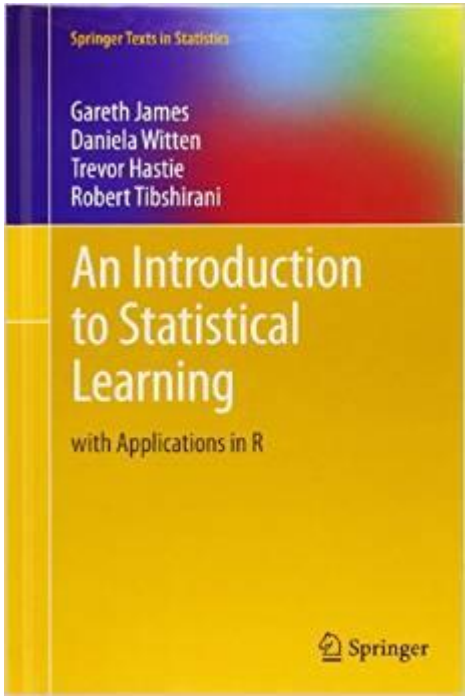
A.I.

- Today's "A.I." may be *artificial*, but it's not *intelligent*
- It is mostly just *statistics, and invented centuries ago*
- The new part is mostly to handle lots of data with new methods based on using algebra to create geometry

- So, what is AI?
  - To me, AI implies a certain amount of *independence* of the machine, to the programmer
    - The machine decides, to a certain extent, what to learn
  - It also implies a certain amount of deductive power
    - In other words, the machine can *reason*
    - *Not just statistical learning, even though it often can be reduced to statistical learning through genetic algorithms..*
  - *And a certain amount of free will*
    - *We feel something is intelligent when it does something unexpected*

Intro to Statistical Learning

- **An Introduction to Statistical Learning with Applications in R, by James et al: http://www-bcf.usc.edu/~gareth/ISL/**

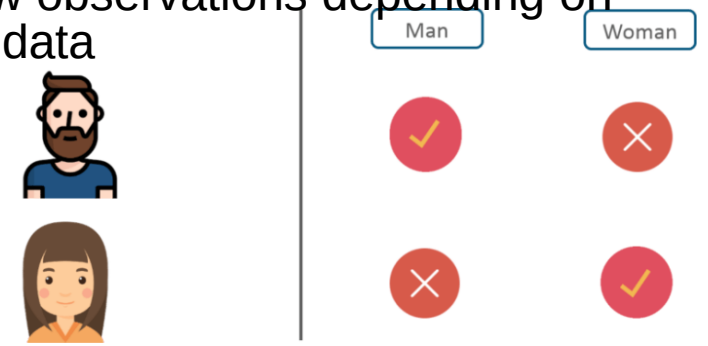


Overview of Machine Learning (ML)

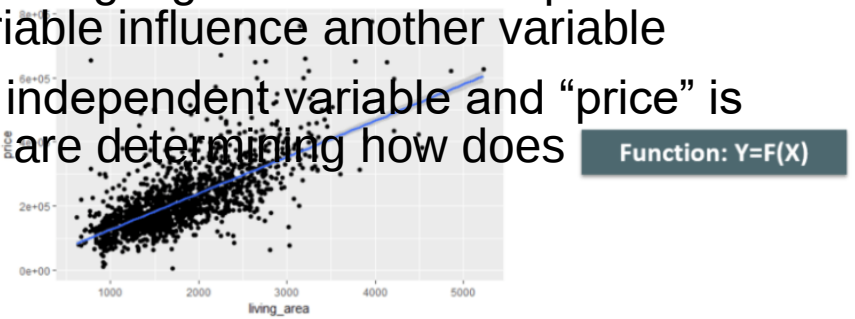
- In **supervised learning** (SL), the learning algorithm is presented with labelled example inputs, where the labels indicate the desired output
  - SML itself is composed of **classification**, where the output is categorical, and **regression**, where the output is numerical
- In **unsupervised learning** (UL), no labels are provided, and the learning algorithm focuses solely on detecting structure in unlabeled input data
- Note that there are also **semi-supervised learning** approaches that use labelled data to inform unsupervised learning on the unlabeled data to identify and annotate new classes in the dataset (also called *novelty detection*)
- **Reinforcement learning** (RL), the learning algorithm performs a task using feedback from operating in a real or synthetic environment, with rewards over actions

Supervised Learning

- Supervised Learning algorithm learns from a known data-set(Training Data) which has labels to make predictions
- **Classification:**
  - Classification determines to which set of categories does a new observation belongs i.e. a classification algorithm learns all the features and labels of the training data and when new data is given to it, it has to assign labels to the new observations depending on what it has learned from the training data

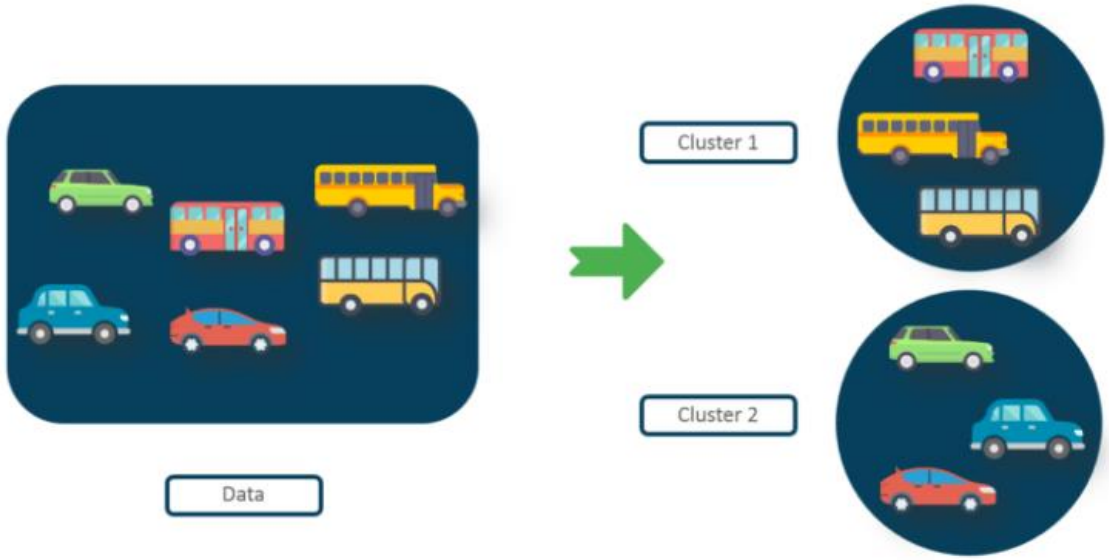


- **Regression:**
  - Regression is a supervised learning algorithm which helps in determining how does one variable influence another variable
  - Over here, “living_area” is the independent variable and “price” is the dependent variable i.e. we are determining how does “price” vary wrt “living_area”



Unsupervised Learning

- Unsupervised learning algorithm draws inferences from data which does not have labels



- In this example, the set of observations is divided into two clusters. Clustering is done on the basis of similarity between the observations. There is a high intra-cluster similarity and low inter-cluster similarity i.e. there is a very high similarity between all the buses but low similarity between the buses and cars.

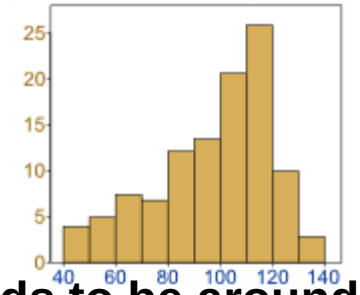
Reinforcement Learning

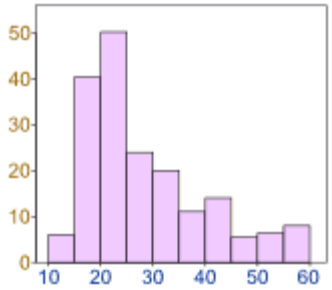
- Reinforcement Learning is a type of machine learning algorithm where the *machine/agent* in an *environment* learns ideal behavior in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior, this is known as the *reinforcement signal*

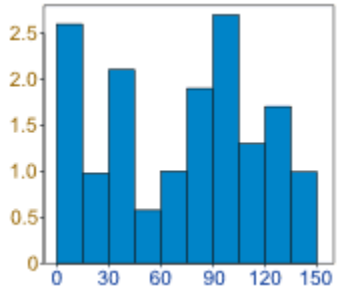
- *pacman* for example. As long as pacman keeps eating food, it earns points but when it crashes against a monster it loses it's life. Thus pacman learns that it needs to eat more food and avoid monsters so as to improve it's performance

How is data distributed?

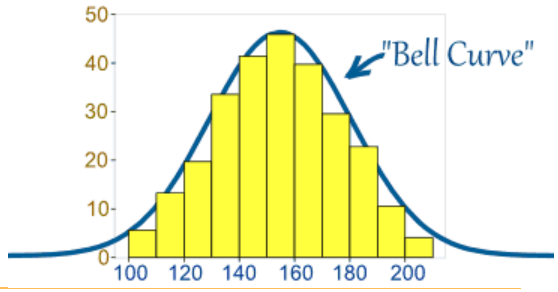
- Data can be "distributed" (spread out) in different ways
  - It can be spread out more on the left
  - Or more on the right
  - Or it can be all jumbled up
  - In many cases, the data tends to be around some bias left or right
  - It is often called a "Bell Curve" because it looks like a bell







- *How data is distributed* is the foundation of statistics

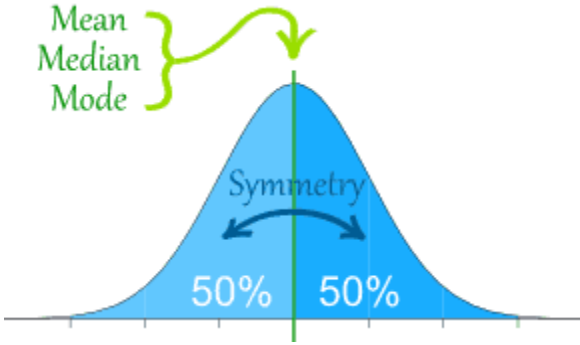


Normal distribution (`rnorm`)

- **We say the data is *normally distributed* when:**
  - **The probability density of the normal distribution is:**

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

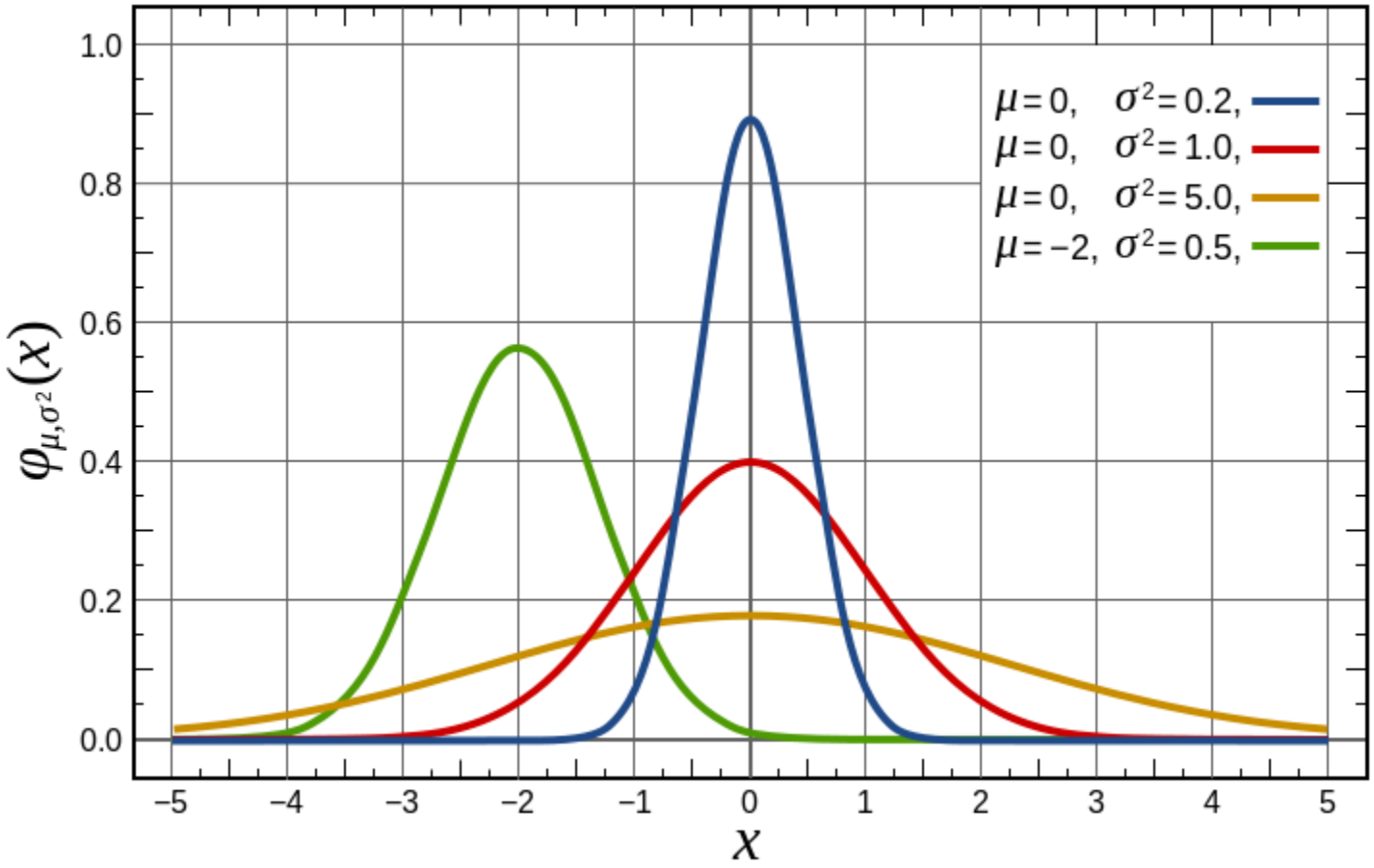
  - **Here, μ is the mean or *expectation* of the distribution (and also its median and mode). The parameter σ is its standard deviation; its variance is then σ^2**
  - **If $\mu = 0$ and $\sigma = 1$ the distribution is called the *standard normal distribution* or the *unit normal distribution***



symmetry about the center

50% of values less than the mean and 50% greater than the mean

Normal Distributions



Central Limit Theorem

- **In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately *normally distributed*, *regardless of the underlying distribution***
- **https://www.mathsisfun.com/data/quincunx.html**

Correlation

- **In statistics, *dependence* is any statistical relationship between two random variables or two sets of data**
  - **Correlation refers to any of a broad class of statistical relationships involving dependence**
  - **Familiar examples of dependent phenomena include the correlation between the demand for a product and its price**
  - **Correlations are useful because they can indicate a *predictive* relationship that can be exploited**
- **Statistical learning, or learning-by-experience, is all about correlation**
  - ***When I see my mum, I get good food***
  - ***If I climb high on a tree, I can fall and hurt myself***
  - ***When the girl winks at me, it means she likes me***

Linear relationships

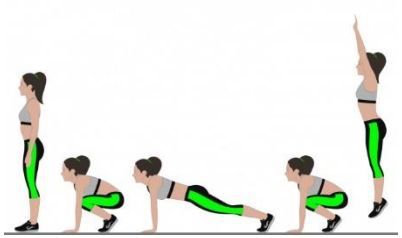
- **The Pearson correlation coefficient indicates the strength of a *linear relationship* between two variables**
  - **The *Pearson correlation coefficient*, which is sensitive *only* to a *linear relationship* between two variables**
  - **If we have a series of *n* measurements of *X* and *Y* written as x_i and y_i where *i* = 1, 2, ..., *n***

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where x and y are the sample means of *X* and *Y*, and s_x and s_y are the sample standard deviations of *X* and *Y*

Causation

- ***Correlation* does not imply *causation*!**
- **A correlation between `age` and `height` in children is fairly causally transparent (`age` → `height`), but a correlation between `mood` and `health` in people is less so**
  - **Does improved `mood` lead to improved `health` (`mood` → `health`), or does good `health` lead to good `mood` (`health` → `mood`), or both (`mood` ←→ `health`)? Or does some other factor underlie both (`factor X` ←→ `mood + health`)?**
  - **E.g.:**
    - **Exercise**



    - **Culture**
      - **Is local Culture in Crete, which includes a great diet and lots of music, responsible for elevated measures of both health and mood?**

Statistical Analysis

- **We use statistical analysis for:**
  - ***Inference* - making conclusions based on data**
  - ***Prediction* - what will happen when I observe new data?**
  - **And we create *models* to do both of those things**
- ***"All models are wrong - some are useful"* - George E. P. Box**

Ghost in the shell

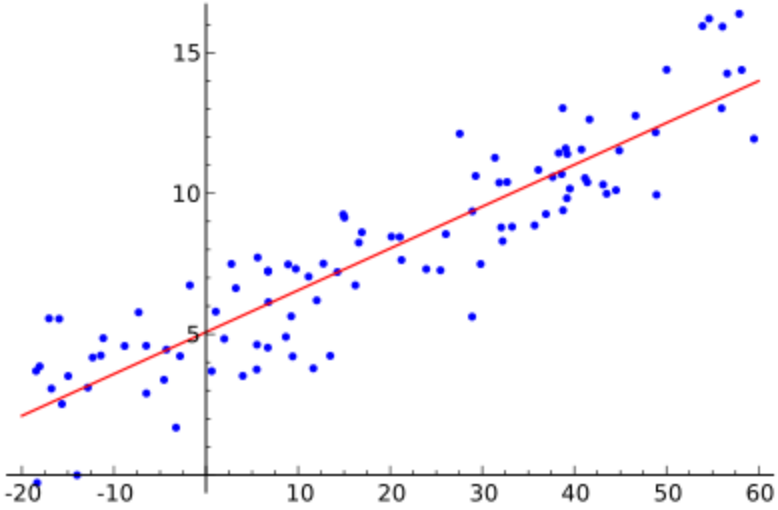
- Our brain simply consists of a bunch of *predictors*, based on *models* we build for ourselves in our lifetimes..
- In the last 20 years, we got very excited trying to find the ghost in the shell, learning how to use a machine to learn, only to uncover that ML is nothing more than plotting high-dimensional graphs using algebraic methods
- Then we realized that we too, we are just machines, with graphs in our brains, and there is no real ghost in the shell

Variables

- **The decision as to which variable in a data set is modeled as the *dependent variable* and which are modeled as the *independent variables* may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables**
- **Independent variables are also called *regressors*, *exogenous variables*, *explanatory variables*, *covariates*, *input variables*, and *predictor variables***

Linear Regression

- **Regression is an approach for modeling the relationship between a scalar *dependent* variable y and one or more explanatory variable (*independent* variable) x**
  - **In linear regression, data are modeled using linear predictor functions**
  - **Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways..**



$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

Logistic Regression

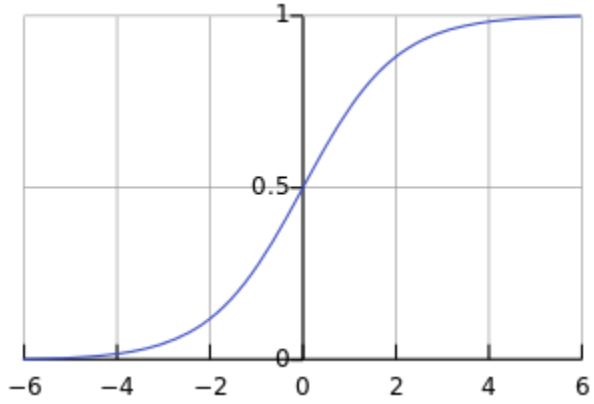
- Logistic regression describes a kind of classification model in which predictor variables are combined with linear weights and then passed through a soft-limit function that limits the output to the range [0, 1]
- Logistic regression is closely related to other models such as:
  - *Perceptron* (where the soft limit is replaced by a hard limit)
  - *Neural networks* (where multiple layers of linear combination and soft limiting are used)
  - *Naive Bayes* (where linear weights are determined strictly by feature frequencies assuming independence)
- Logistic regression can't separate all possible classes, but in very high dimensional problems or where you can introduce new variables by combining other predictors, this is less of a problem
- Mathematical simplicity of logistic regression allows very efficient and effective learning algorithms to be derived

Logistic Regression

- **The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function**
  - **A logistic function is a common "S" shape (sigmoid curve), with equation:**

$$f(x) = \frac{L}{1 + \mathrm{e}^{-k(x - x_0)}}$$

  - **Population growth: The initial stage of growth is approximately exponential, then, as saturation begins, the growth slows, and at maturity, growth stops**



The logistic function is useful because it can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one and hence is interpretable as a probability

Regression Example

- **A good grade in this class (the *dependent variable*) is directly influenced by how many hours per week (the *predictor* or *independent variable*) you review the slides and do the homework**
  - **For example:**
    1or less hours per week → F
    2 – 3 hours per week → C
    4 – 5 hours per week → B
    8 – 9 hours per week → A

Understanding Machine Learning

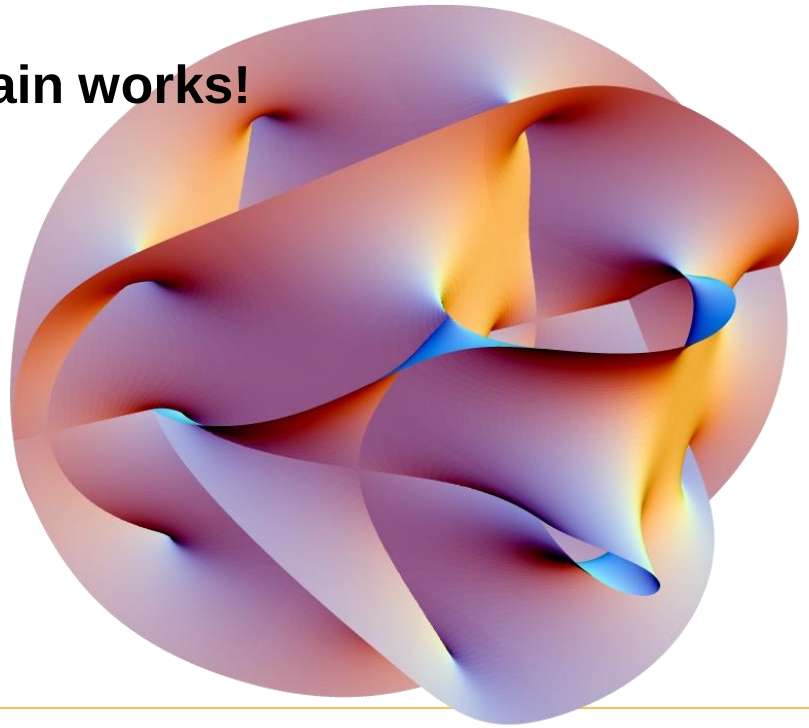
- **How do you know all of these are *cats*?**
- As a kid, you might have come across a picture of a cat and you would have been told by your kindergarten teachers or parents that this is a cat and it has some specific features associated with it like it has fluffy ears, a pair of cute eyes, a tail and so on
- Now, whenever your brain comes across an image with those set of features, it automatically registers it as a cat because your brain has *learned* that it is a cat

Machine Learning

- If the same image is fed to a machine, how will the machine identify it to be a cat?
- This is where M*achine Learning* comes in. We'll keep on feeding images of a cat to a computer with the tag "cat" until the *machine learns all the features associated* with a *cat*
  - *We'll also feed it images of things that are not a cat, like a dog for example, with the tag "not a cat"*
- Once the machine learns all the features associated with a cat, we will feed it new data to determine how much has it learned
- In other words, *Raw Data/Training Data* is given to the machine, so that it *learns* all the features associated with the *Training Data.* Once, the learning is done, it is given *New Data/Test Data* to determine how well the machine has learned

The graph in the Machine

- The machine decides how many features it will examine, say *n*, and then it will draw a cat graph in *n*-dimensional space, and a non-cat graph
- Given a new image, it will look at the *n* features, and plot them in *n*-dimensional space and look at the intersect with the cat graph, and the non-cat graph, and decide which is the best intersect
- That, is also how your brain works!



Part 6

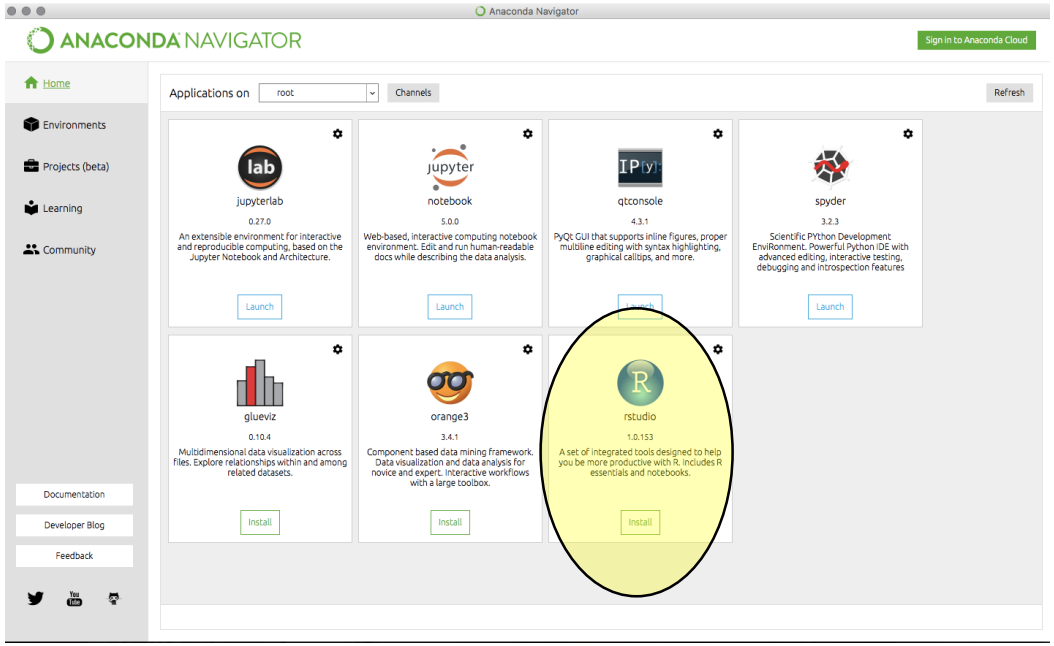
LEARNING R WITH ML

Why?

- **To learn to operate on vectors (lists) and matrices (spreadsheets) *all at the same time***
- ***To introduce ML and have fun***

Learning R

- Run `Anaconda`
- Click on `RStudio`



Preliminaries

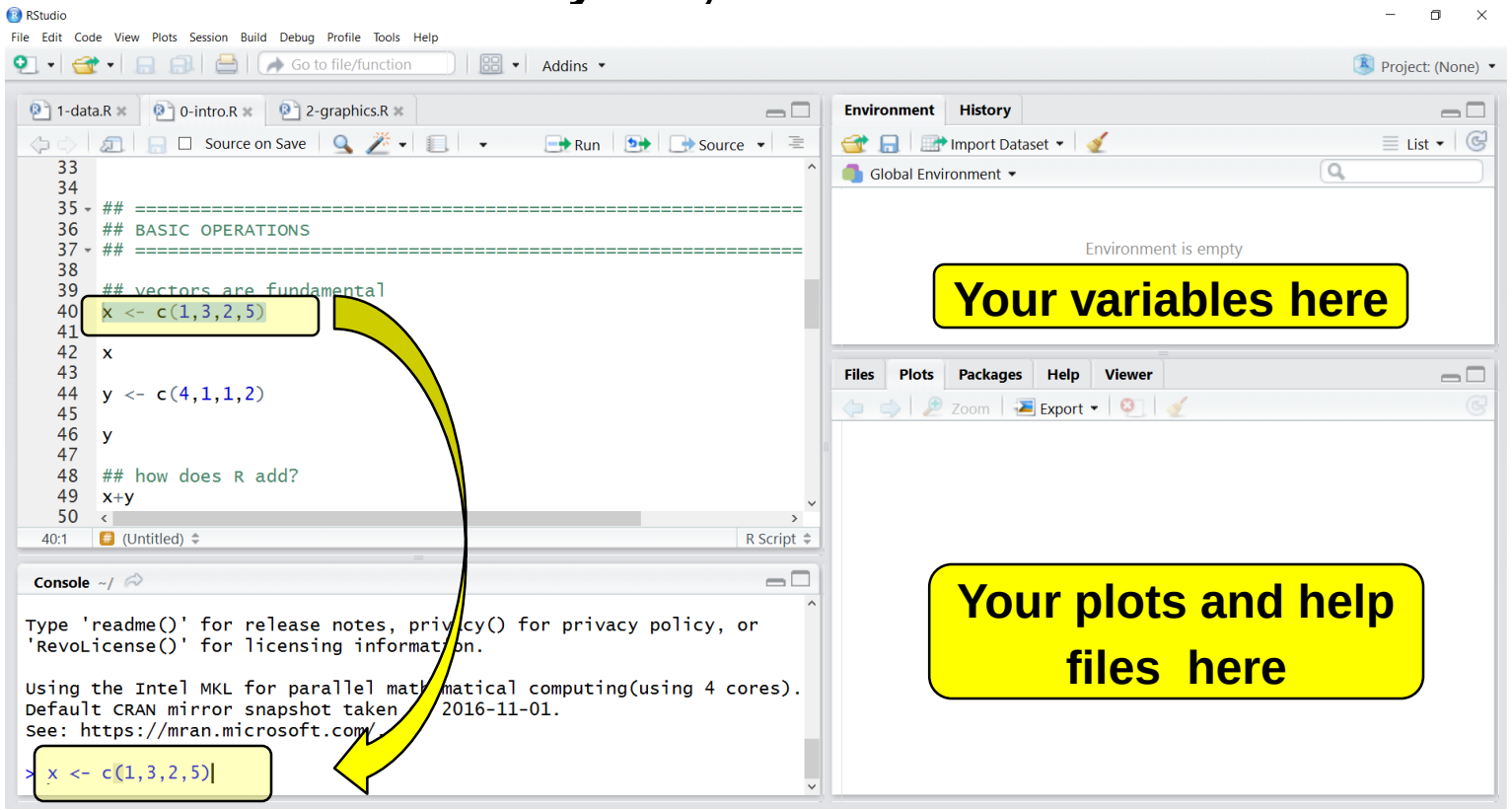
- Unzip the contents of file `Rlab6105.zip` to a new folder on your hard drive, let's say "Rlab"
- Set the `RStudio` working directory to that folder
  - `Session` → `Set Working Directory` → `Choose Directory...`
  - Navigate to Rlab, and Open the /programs directory
- *Focus console after executing from source* option: Moves the focus to the console after executing a line or selection of code within the source editor (you *will* thank me)
  - `Tools` → `Global Options` → `Code Editing...`
  - Check "`Focus console after executing from Source`"

Installing R packages:

- Install *dplyr* and *ggplot2* packages
  - `install.packages("dplyr")`
    `install.packages("ggplot2")`
- `Load the libraries`
  - `library(dplyr); library(ggplot2)`
- a) MAC OS X and Linux:
  - Open Rstudio, in Menu, go to `Packages and Data` → `Package Installer`
  - Search for and install the above two packages (may need to choose a "mirror" - click on something in the USA):
    - Type in the name of one package, click "`get list`", check "`Install Dependencies`" and then "`Install Selected`"
  - Do the same for the other package
- b) Windows:
  - Open Rstudio, in Menu, go to "`Packages`" → "`Install package(s)...`" and select each of the packages at top to install

Open R Lab Files

- In `RStudio`, pick `Open File` menu and open file `/programs/0-Intro.R`
- Read each line, copy *command line(s)* and paste them into the RStudio console line by line, then hit return to execute



Homework (due next week)

- **Finish homework embedded in all R data files we covered**
  - **The `group_by` homework may be a bit complicated if you never studied databases..**
- **Download and install `Anaconda with RStudio`, `Visual Studio Code`, install `ggplot2` and `dplyr` R packages**
- **Repeat `ml.R` using a different dataset from UCI or another repository to learn a model, then use that model to predict whether an observation belongs to the model or not**
- **Bonus credit Do the wordcloud lab `4-textmining-cloud.R` on the `Iliad` and the `Odyssey` classic**
  - **Then repeat with a text corpus in *your* native language**
  - **What do you have to modify to make this work?**


CLASS
DISMISSED
iStockphoto