

Introduction to the Course

Welcome to EPA 1315

Data Analysis and Visualization

- Set forth the objectives and grading procedures for the course
- Brief story about the new world of data
- Break
- Walk through of data science distributions
- Specifics about the upcoming module
- Check-in

Course Team

- [Trivik Verma](#), Associate Professor
- [Ruchik Patel](#) and
- [Fabio Hernan Tejedor](#), Teaching Assistants

Course Description

- This is a course about modern techniques of machine learning
- Most lectures involve live-coding, programming exercises, or small group work with computers, so having a laptop and bringing it to lecture is recommended.

Tasks and Responsibilities

- Module Manager
- Instructor
- Teaching Assistants
- You / Course Evaluation

Learning Objectives

Students are able to

- **apply** the R language to deliver essential data analysis
- **produce** and **critique** informative visualisations
- **design** and **compare** statistical models
- **analyse** mathematical problems involving probabilistic processes.
- **reason** and **communicate** about uncertainty
- **apply** and **diagnose** at least probabilistic programming frameworks
- **manage** the data analysis lifecycle using CRISP-DM

The Parts of the Course

The course consists of three modules:

- Module 1. Data science in R
- Module 2. Data visualization
- Module 3. Data analysis

Learning Objectives Module 1

Students are able to

- **apply** the R language to deliver essential data analysis

Learning Objectives Module 2

Students are able to

- **produce** and **critique** informative visualisations

Learning Objectives Module 3

Students are able to

- **design** and **compare** statistical models
- **analyse** mathematical problems involving probabilistic processes.
- **reason** and **communicate** about uncertainty
- **apply** and **diagnose** at least probabilistic programming frameworks
- **manage** the data analysis lifecycle using CRISP-DM

Assessment Principles

- Small course assignments occur throughout the course
- Each module is continuously assessed
- There are opportunities to be graded individually and in groups

Grading Procedures

<u>No</u>	<u>Module</u>	<u>Grade %</u>	<u>Type</u>	<u>Specifics</u>	<u>Group/Individual</u>
1	Module 1	10%	Homework	Install Anaconda	Individual
2		10%		Dataframes Homework	Individual
3		11%	Coding Challenge	Agent-Based Tournament	Group
4	Module 2	11%		Neogeography Lab	Individual
5		6%		Visual Analysis	Group
6	Module 3	12%	Project	Visualization Elements	Group
7		21%		Analysis Elements	Group
8		19%	Exam	Probability Processes	Individual
		100%			

Course Schedule

- The course meets two or three days a week, depending on the week.
- There is a detailed agenda for each lecture
- We'll post updates for each module and each week
- Although attendance for all lectures is not mandatory, substantial self-study will be required to keep up with the course.

Ways of Working

- Attendance is not mandatory, but we do want you to offer value and for you to be here
- Generally we want the face-to-face meetings to involve live coding, examples, small group work and demonstrations

Funnel

- Live coding
- Sample scripts
- Exercises
- Demonstrations
- Group work
- Project work

Previous Knowledge

- For a student who has already had a bachelor's course on statistical hypothesis testing, this will expand upon basic knowledge of point estimation.
- It also expands the multilinear regression to a range of general models.
- Prior knowledge of probability, and an introductory course on linear regression is expected.
- Previous courses on programming (in R or other languages) or scientific computing are helpful but not required.

Why Me

- STEM Student
- International Student
- Ph.D. in Technology Policy
- Worked in Industry
- Worked in an entrepreneurial environment
- Plugged in to Dutch cities and data needs

Jump Start

- The course explores model experimentation techniques to fit a range of credible models to data.
- Specifically, the course covers Bayesian statistics in R.
- This course will offer a jump-start for further professional or academic learning in the field of data science.

Traditional Statistics

- Traditionally data analysts have worked in a controlled, experimental setting
- Historically the statistician was subsidiary to the scientist
- The scientist proposed the research designs, and entertained the hypotheses
- The statistician simply ran the numbers, and confirmed or denied the results

Social History of Statistical Computing

- 1750. Early modern state
- 1910. Vienna programme of modern science
- 1950. Proliferation of database systems
- 1970. New machine learning methodologies
- 1990. Administrative frameworks
- 2000. Open source programming
- 2010. Deep learning



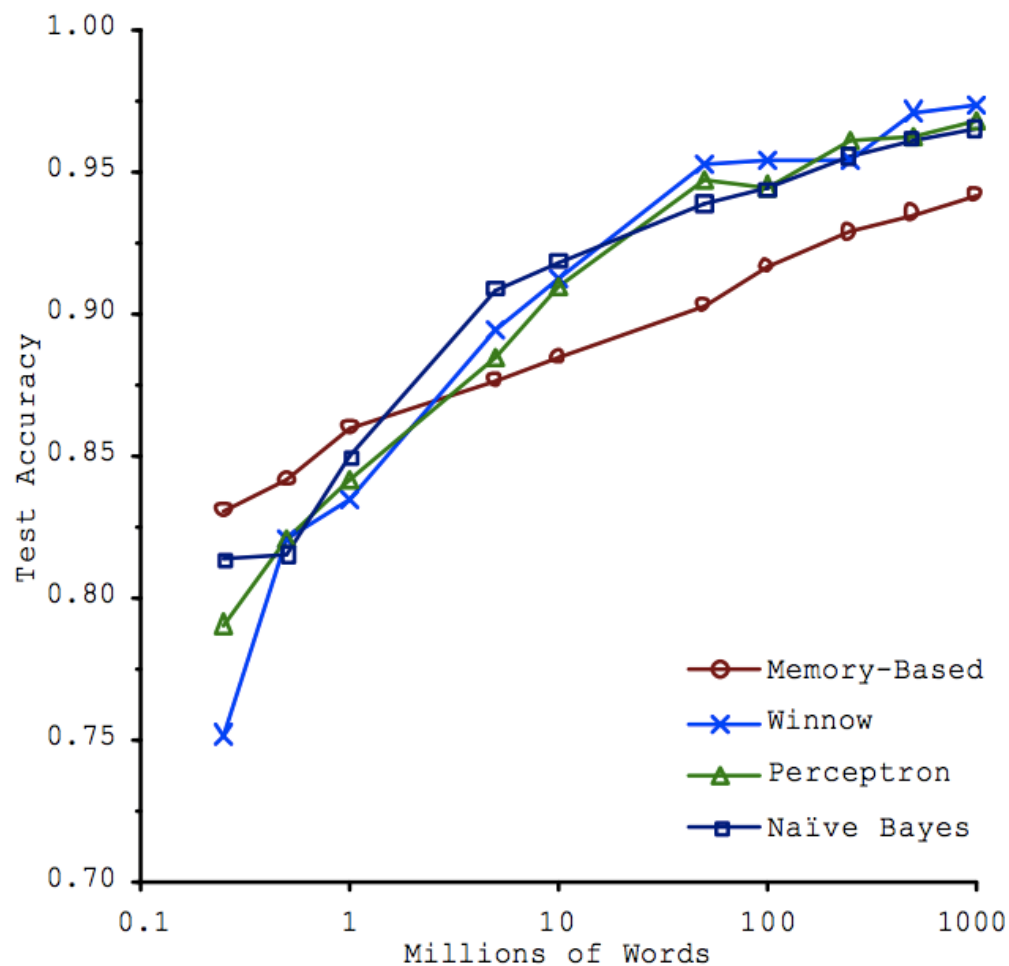
Johan Pieter Süssmilch, cc Alchetron 2019





Early IBM Computers, Wikimedia cc 2019

The Unreasonable Effectiveness of Data



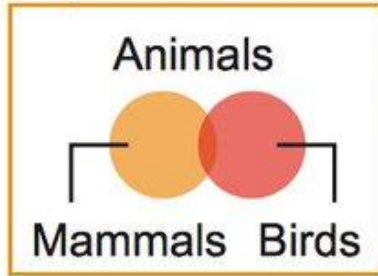
(Banko and Brill, 2011)

The World of Data Today

- Today the world of data is both open and incomplete
- The processes that produce data are varied and uncertain
- Data scientists play a critical role in proposing experiments
- Data scientists assist in formulating new scientific hypotheses

What are the five tribes?

Symbolists



Use symbols, rules, and logic to represent knowledge and draw logical inference

Favored algorithm

Rules and decision trees

Bayesians

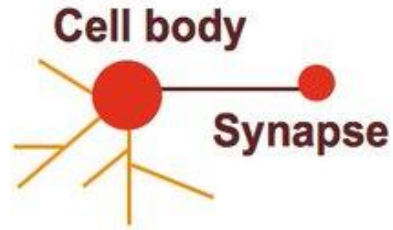


Assess the likelihood of occurrence for probabilistic inference

Favored algorithm

Naive Bayes or Markov

Connectionists

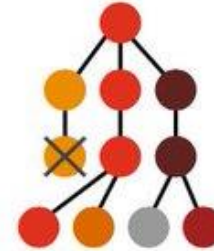


Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons

Favored algorithm

Neural networks

Evolutionaries

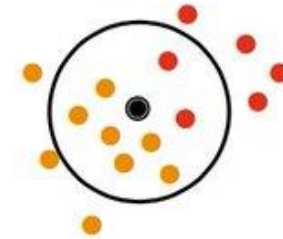


Generate variations and then assess the fitness of each for a given purpose

Favored algorithm

Genetic programs

Analogizers



Optimize a function in light of constraints (“going as high as you can while staying on the road”)

Favored algorithm

Support vectors

(Domingos, 2015)

Machine Learning Themes of the Course

- Computer experimentation
- Role of likelihood and belief
- Communicating uncertainty
- Drawing causal inferences

Second Half

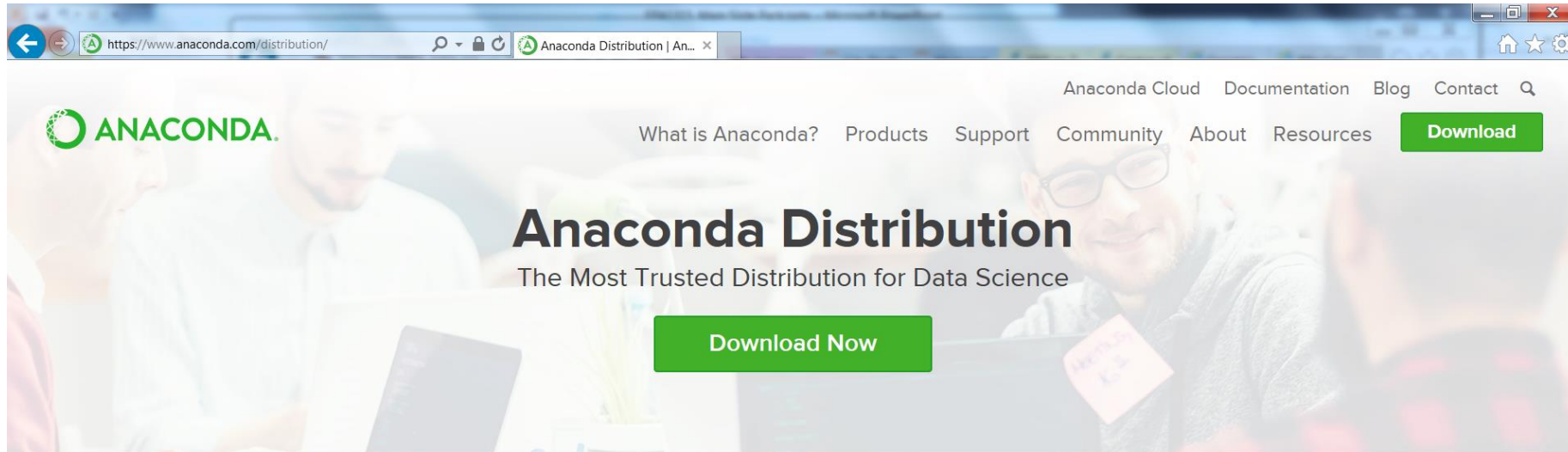
- Break
- Set-up in R
- Beginning module 1

Break

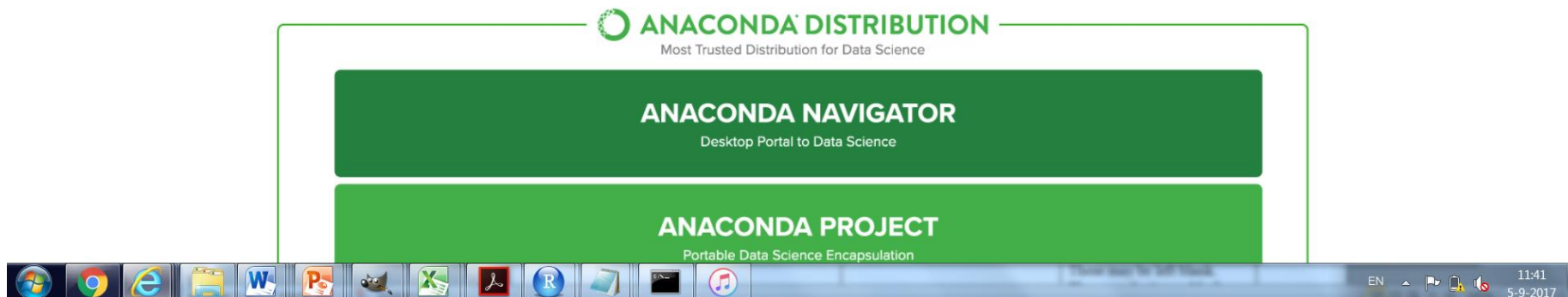
The Anaconda Distribution



Download R



Easily install 1,000+ data science packages and manage your packages, dependencies and environments—all with the single click of a button



Working Distribution Needed

- You need to have a running Anaconda distribution on your own laptop to pass this course
- Pass/Fail
- Drop by and I can confirm or help you

Other Degree Programs

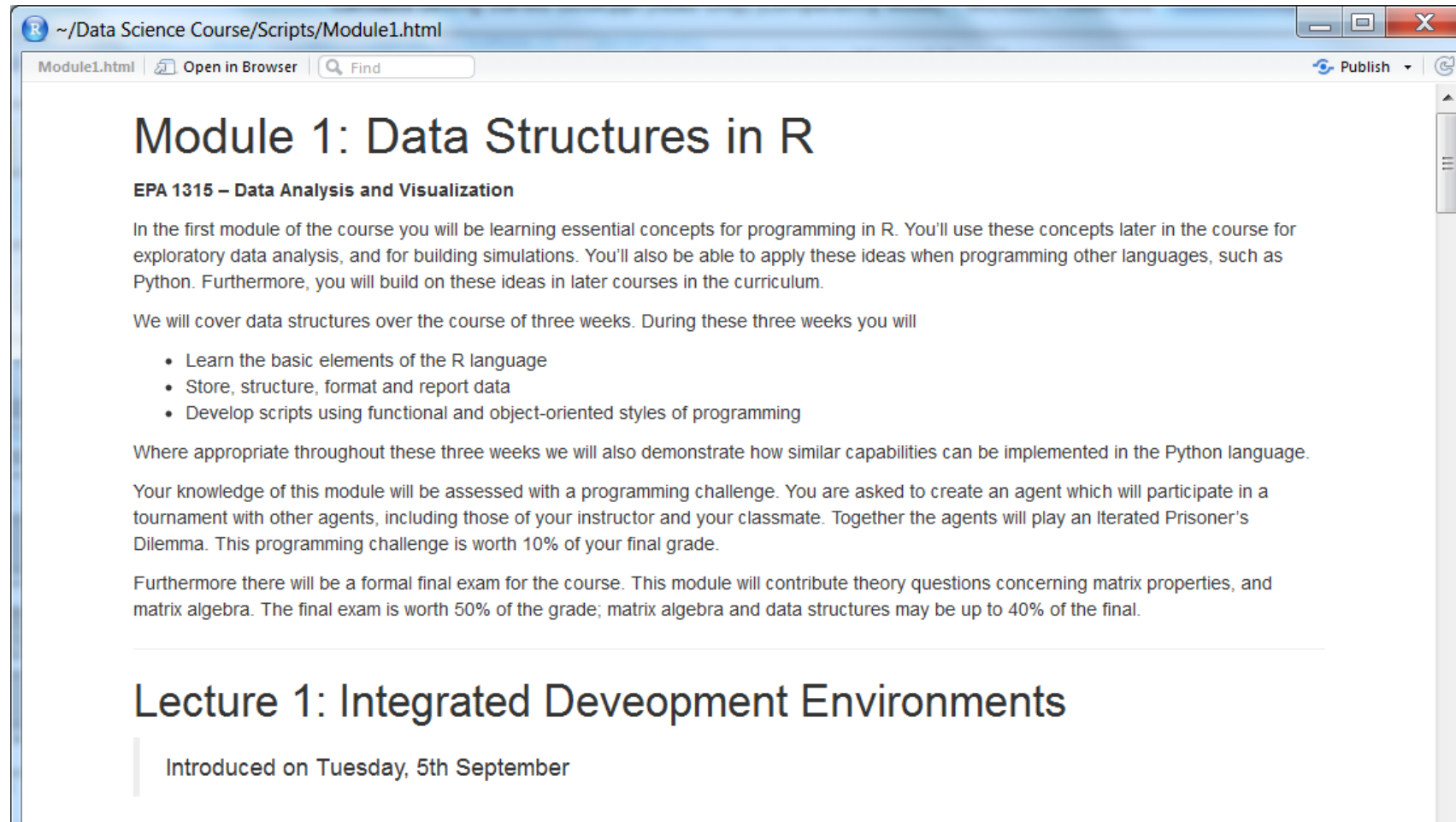
- This course is required for the [Engineering and Policy Analysis](#) master's degree
- Although I think data analysis and visual communication are critical skills for policy analysts I don't think there is anything in this course which limits its accessibility to other degree programs

Quick Tour of R Studio

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for a lecture introduction and a data setup. The code includes comments about the course structure and a series of assignment statements for variables like `salary`, `position`, `team`, `name.last`, and `name.first`.
- Environment Pane:** Shows the current environment with objects like `sampling` (19 obs. of 3 variables) and `students` (10000 obs. of 6 variables). It also lists the values of several objects, including `edu_margin`, `ggp`, `header`, `model`, `mytable`, `name.first`, `name.last`, `obj`, `position`, `pov_margin`, and `salary`.
- Console:** Displays the output of the R commands, showing the workspace loaded from `~/RData` and the execution of the assignment statements.

Quick Tour of R Markup



The screenshot shows a web browser window displaying an R Markdown document. The title bar indicates the file path is ~/Data Science Course/Scripts/Module1.html. The browser's address bar shows 'Module1.html' and includes buttons for 'Open in Browser', 'Find', and 'Publish'. The document content is as follows:

Module 1: Data Structures in R

EPA 1315 – Data Analysis and Visualization

In the first module of the course you will be learning essential concepts for programming in R. You'll use these concepts later in the course for exploratory data analysis, and for building simulations. You'll also be able to apply these ideas when programming other languages, such as Python. Furthermore, you will build on these ideas in later courses in the curriculum.

We will cover data structures over the course of three weeks. During these three weeks you will

- Learn the basic elements of the R language
- Store, structure, format and report data
- Develop scripts using functional and object-oriented styles of programming

Where appropriate throughout these three weeks we will also demonstrate how similar capabilities can be implemented in the Python language.

Your knowledge of this module will be assessed with a programming challenge. You are asked to create an agent which will participate in a tournament with other agents, including those of your instructor and your classmate. Together the agents will play an Iterated Prisoner's Dilemma. This programming challenge is worth 10% of your final grade.

Furthermore there will be a formal final exam for the course. This module will contribute theory questions concerning matrix properties, and matrix algebra. The final exam is worth 50% of the grade; matrix algebra and data structures may be up to 40% of the final.

Lecture 1: Integrated Deveopment Environments

Introduced on Tuesday, 5th September

Key Dates

<u>No</u>	<u>Specifics</u>	<u>Assigned</u>	<u>Due</u>
1	Install Anaconda	Tue, 3rd Sep	Tue, 10th Sep
2	Dataframes Homework	Fri, 13th Sep	Fri, 20th Sep
3	Agent-Based Tournament	Tue, 24th Sep	Tue, 8th Oct
4	Neogeography Lab	Tue, 24th Sep	Tue, 1st Oct
5	Visual Analysis	Tue, 8th Oct	Tue, 15th Oct
6	Visualization Elements	Tue, 22nd Oct	Tue, 5th Nov
7	Analysis Elements	Tue, 22nd Oct	Tue, 5th Nov
8	Probability Processes	Wed, 9th Oct	Tue, 5th Nov

Module 1. Recommended Text

- Adler, J. (2012), R in a Nutshell, O'Reilly: Sebastopol, CA.

Related Courses

- Also running [EPA 1333 Computer Engineering for Scientific Computing](#)
- Python course
- Many of the same topics
- Extended treatment weeks 4+
- Intended for our own bachelor's students who may not have had prior experience with programming

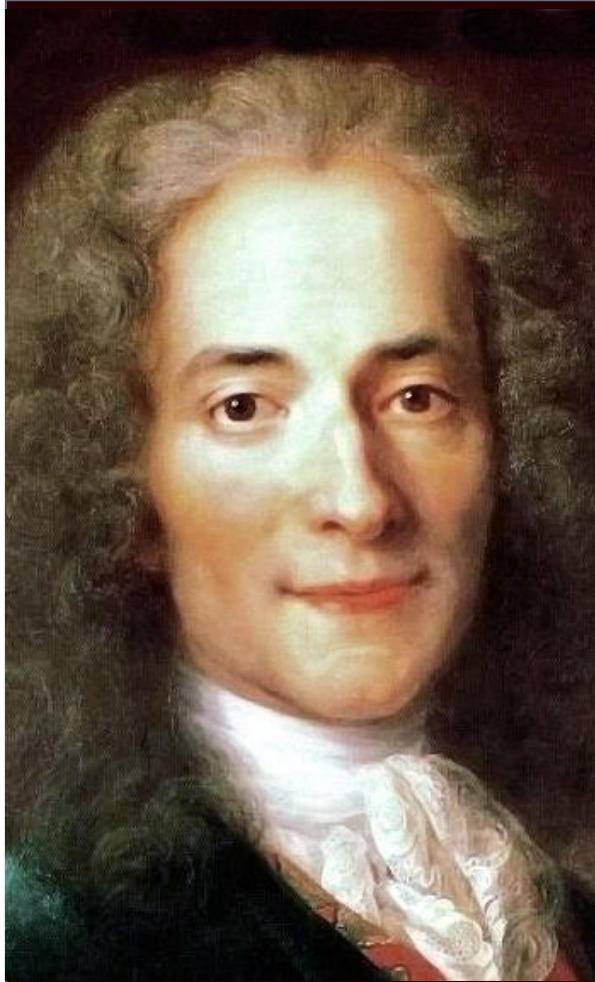


Coding Challenge: Agent-Based Tournament

Next Time -- Friday

- Module 1, Algorithms and Data Structures in R
- Downloaded R Studio
- Walkthrough of the Software
- Work with R Markup Language
- Hands-on examples data entry and output in R

Extra



Berkeley



Locke



Hume

The Empiricists



The Two Arches of Knowledge

A low-angle photograph of classical columns against a blue sky with wispy clouds. The columns are white with fluted shafts and papyrus capitals. The perspective is looking up, making the columns appear to converge towards the top of the frame. A dark blue horizontal band is superimposed over the middle of the image, containing the title text.

A Priori and A Posteriori

Empiricism

Drivers of the World of Data

There are five drivers for the new world of data

- Driver 1. Open source software
- Driver 2. Bayesian approaches to data
- Driver 3. New graphics software
- Driver 4. New computing hardware
- Driver 5. Connectionism

The First Driver

- The first and major driver of the data revolution is cheap and readily available open source software
- Without suitable computer languages to harness the power of our powerful computers, where would we be?

The R Programming Language

- For this reason a significant portion of the course will be devoted to using the R programming language
- R is now the major language in the world for data science applications
- The language is also renowned for beautiful visualization
- There will be homework assignments to test your growing knowledge of R

The Bayesian Revolution

- Bayesian inference is giving rise to fundamental new advances which are changing the economy
- One of the examples of this revolution is the development of the self-driving car
- Other examples include routine voice recognition and image recognition technologies

Another Driver of the Revolution

- Another driver of the revolution are ever faster, ever cheaper computers
- This enables high-tech, high-touch graphics
- And also increasingly more complex models of data
- Fast computers enable us to find and fit credible models in a suitable amount of time

An Objective of the Course

- Graphics cards, originally developed for computer games, are offering improved new graphics capabilities
- Graphics are opening up the world of data creating new fields like data journalism, crowd-sourced data, neogeography
- In this course, we begin with you. Exploratory data analysis, procedures for exploring hypotheses.

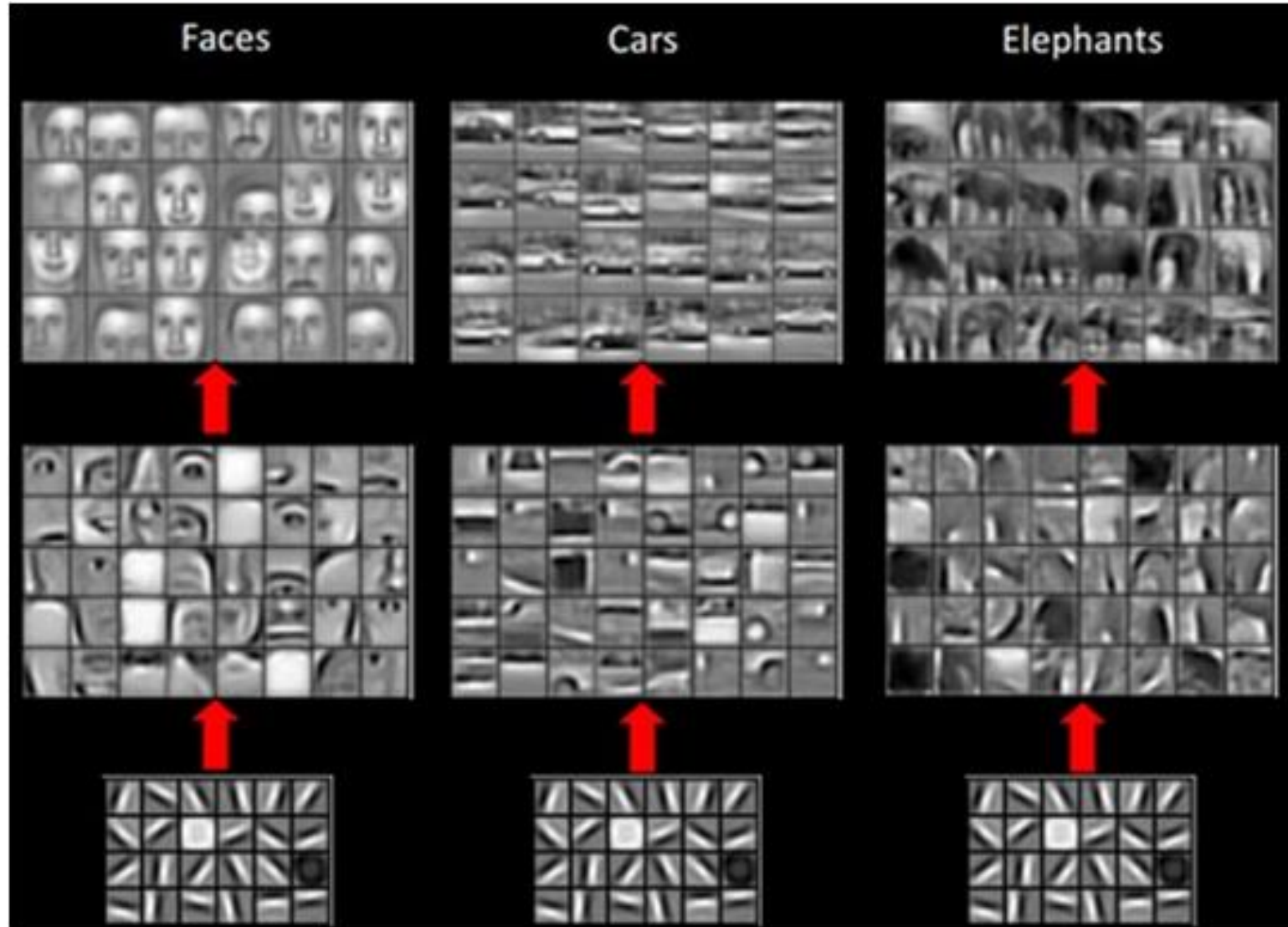
The Fourth Objective of the Course

- The next objective of the course is to deploy computer experimentation techniques to search through millions of possible parameters in pursuit of the best possible models of data
- We will be using a technique called Gibbs sampling to speed and structure our search
- This objective will be met by in-class working groups with your laptops, where we search through models

Connectionism

- Networked and modular architectures
- As a result of the computing and software revolutions it is no longer useful to consider statistical as individual, unrelated, silo-ed models
- Instead we will be building an architecture of models using the general linear framework
- Our goal is to grow or adapt our models to meet the complexity of real world challenges

Convolutional Neural Networks



An Objective of the Course

- A final objective of the course involves applying your knowledge to complex, real-world problems
- For that reason you'll have an opportunity to apply your work to a problem of your choice based on policy databases
- You will work with a team to develop a small, policy report using a statistical model of your choice

Guiding You in Your Project

- To better guide you in your report writing process I will introduce the leading data mining process
- The Cross-Industry Standard Process for Data Mining (CRISP-DM) involves a structured process for performing data mining
- This involves particular phases including problem discovery, data discovery, model selection, and deployment

So Let's Summarize

- Statistics has been transformed from a subject where analysis was subservient to scientific team leaders
- Practice has shifted from closed-form and analysis to a full practice of computer experimentation
- The resultant proliferation of powerful hardware and cheap software has sped a revolution
- The revolution is accompanied by powerful, general purpose learning algorithms