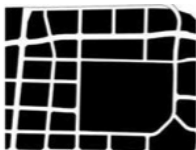## Outline

Reality is predictable

Los Angeles Times | ARTICLE COLLECTIONS

Stopping crime before it starts

Sophisticated analysis of data can sometimes tell police where criminals are headed. It's academic now, but the LAPD plans to get involved.

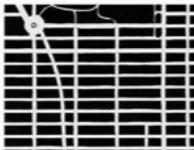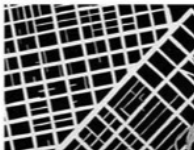# Reality is visualizable

MISSISSAUGA · BARCELONA · COPENHAGEN · LONDON · NEW YORK · PARIS · ROME · SAN FRANCISCO · TORONTO

Reality is multidimensional

Bad Religion (317 plays) David Bowie (220 plays) Isis (202 plays) Horace Andy (178 plays) Army of the Pharoahs (177 plays) Biosphere (168 plays)
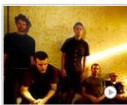
Bibio (162 plays) Antonio Vivaldi (143 plays) Neil Young (125 plays) King Crimson (116 plays) H.P. Lovecraft (116 plays) Virgin Prunes (115 plays)

Motorama (109 plays) Wax Tailor (104 plays) Lou Reed (103 plays) Kate Bush (102 plays) 7L & Esoteric (102 plays) Gonzales (100 plays)

Reality is relational

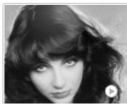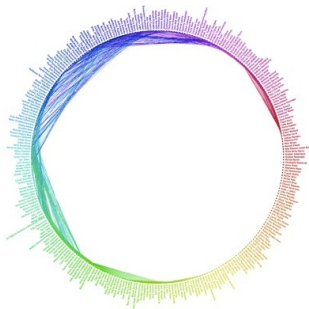Friendship ties on Facebook
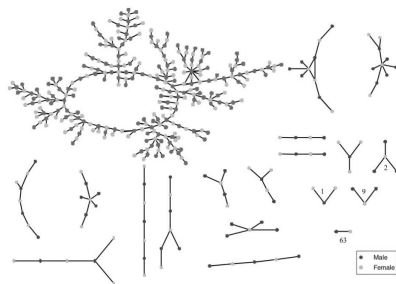
Fig. 2.—The direct relationship structure at Jefferson High

Sexual ties in high school

Data stand as professional assets

OECD Health Data 2010: Statistics and Indicators

AVAILABLE NOW - October 21st - Internet update for OECD Health Data 2010

OECD Health Data 2010*, released on 29 June 2010, offers the most comprehensive source of comparable statistics on health and health systems across OECD countries. It is an essential tool for health researchers and policy advisors in governments, the private sector and the academic community, to carry out comparative analyses and draw lessons from international comparisons of diverse health care systems.

- What is OECD Health Data 2010

Data stand as policy expertise

Graphique 1 – Vue d'ensemble de l'ONDAM
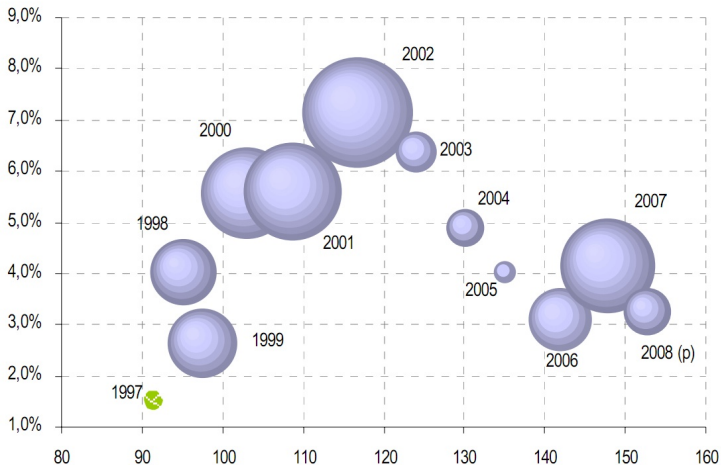
Interpretation is key to all analysis

Lonely **Neutral** Not Lonely **Friends, Spouses** Siblings ● Women

FIGURE 1

Jesus is the Christ
self-preservation
omnium et singulorum
} fundamental

a, b, c sets of beliefs of different
Christian sects
jura in omnia —given up
} superstruction
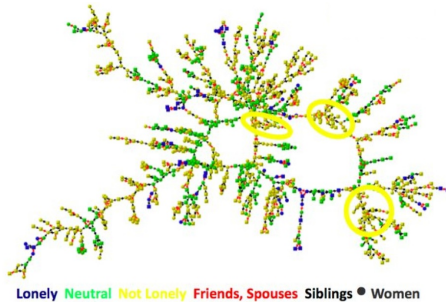
Loneliness in social networks
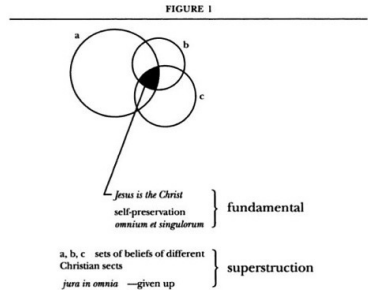
Sets of Christian beliefs

Interpretation is difficult

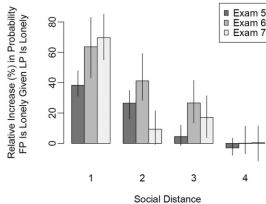*Figure 2.* Social distance and loneliness in the Framingham Social Network. This figure shows for each exam the percentage increase in the likelihood a given focal participant (FP) is lonely if a friend or family member at a certain social distance is lonely (where lonely is defined as feeling lonely more than once a week). The relationship is strongest between individuals who are directly connected, but it remains significantly greater than zero at social distances up to three degrees of separation, meaning that a person's loneliness is associated with the loneliness of people up to three degrees removed from them in the network. Values are derived by comparing the conditional probability of being lonely in the observed network with an identical network (with topology and incidence of loneliness preserved) in which the same number of lonely participants are randomly distributed. Linked participant (LP) social distance refers to closest social distance between the LP and FP (LP = Distance 1, LP's LP = Distance 2, etc.). Error bars show 95% confidence intervals.

With explanation
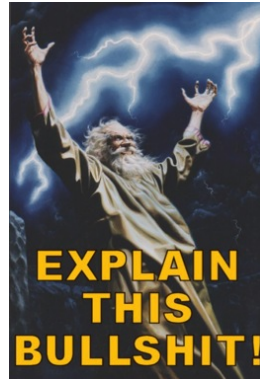
Without explanation

Interpretation is what this course is eventually about
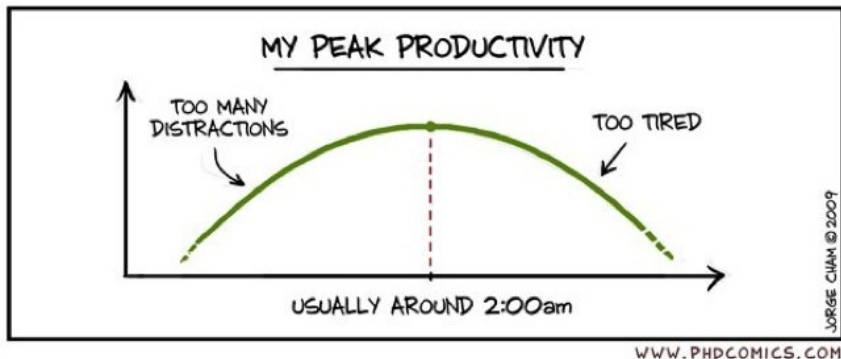
- What is the **measurement** of the axes?
- What is the **probability** of 2am being the "usual" cutoff point?
- What is the **shape** of the time/productivity relationship?

Internal biases

Statistical reasoning and quantitative methods are services to enhance your interpretation of reality with measurements and probability levels.

We will use observational data from social surveys, which have internal limitations:

- Survey design and **sampling strategy**:

    "Please fill in this 367-page long questionnaire"
    "Answer our questions and win an iPad! (perhaps)"
    "We surveyed knowledge of GLMM among toddlers."

- Question design and **measurement error**:

    "Are you a racial supremacist pig?"
    "What do you like most, inflation or Star Wars?"
    "Do you support Aliina Koospürati's new coalition?"

Further external biases

- Media coverage
- Political spin

Signs of a warming world
Sea level (mm)
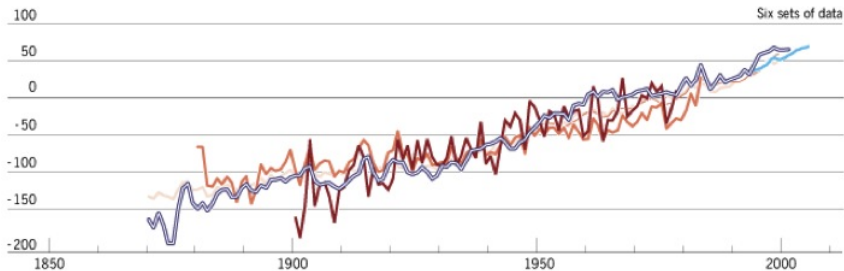
Further WEIRD biases

The weirdest people in the world?

"The findings suggest that members of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies, including young children, are among the least representative populations one could find for generalizing about humans."

Joseph Henrich
Department of Psychology and Department of Economics, University of British Columbia, Vancouver V6T 1Z4, Canada
joseph.henrich@gmail.com
http://www.psych.ubc.ca/~henrich/home.html

Steven J. Heine
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
heine@psych.ubc.ca

Ara Norenzayan
Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada
ara@psych.ubc.ca

Learning objectives

- **Frequentist statistics**
  - □ Normal distribution
  - □ Probability levels
- **Data management**
  - □ Variable transformations
  - □ Dataset manipulation
- **Analytical techniques**
  - □ Univariate statistics
  - □ Bivariate statistics
  - □ Regression modelling

The course offers an introduction to "Statistical Reasoning and Quantitative Methods" using Stata. It leaves out much of survey design and nonparametric statistics. It stops before advanced regression modelling and post-frequentist statistics.

Requirements

- **Homework**
  - □ Handbook readings
  - □ Session replication

  Homework deals with statistical theory and learning Stata programming for quantitative analysis.

- **Assignments**
  - □ Draft Paper No. 1: data preparation and variable descriptions
  - □ Draft Paper No. 2: association

  Assignments deal with the cumulative steps of research design that are reflected in your final paper.

- **Final paper**
  - □ Complete do-file
  - □ Multiple regression model
  - □ Interpretation

Attendance

Regular attendance is compulsory, but most importantly, homework needs to be done **on a weekly basis**, in order to write up Assignments No. 1 and 2.

The final paper comes out of this regular coursework and **cannot**, from experience, be rushed as overnight sessions and other methods to catch up with late work. Sorry.

There is plenty of course material to achieve all requirements and find help, but you will need to read and practice weekly.
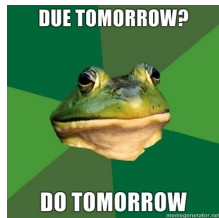
This won't work.

Logistics

- Teaching Pack
- Course Website
- Student Rep

No estimation without representation, one (wo)man, one vote, etc.

Any questions so far?

Do not worry about deadlines, grades and instruction sets: these will be discussed in class and sent by email.

Computers

- Music players
- Facebook terminals
- Porn stashes

Despite the appearances, computers are not just that.

They can also be used to do, like, serious stuff.

Computer skills

- **Section 2 of the Stata Guide** will walk you through the most essential notions that you must be confident with:
  - □ Files (size, names, extensions, compression)
  - □ System (memory, folders/directories, file/folder paths)
  - □ Code (programming, replication, debugging)
- **Using your personal computer** is preferrable due to limitations in administrator privileges with university computer lab machines.
- **Backup your files every week** on at least two different locations. Using a USB key or a Gmail account will do.

Personal pledge:

I hereby declare that, if no single case of data loss is reported throughout the semester, I will buy a drink to the whole class.

Example: Saving files

Do not save online files by single-clicking them. This is good for opening holiday pictures attached to an email on the fly, not for working with files that often need to be decompressed and archived.
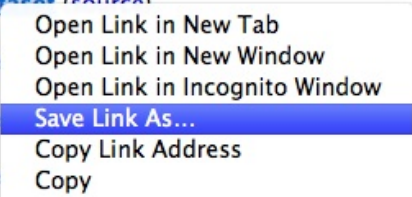
Instead, use the " Save As…" or "Download As…" right-click menu item of your browser to save the file at a precise location.

- **World Values Survey 2000:** dataset (source)

  Replication of Session 3: do-file

  Replication of Session 6: do-file

  Replication of Session 7: do-file

| Open Link in New Tab |
| Open Link in New Window |
| Open Link in Incognito Window |
| Save Link As... |
| Copy Link Address |
| Copy |

Computer skills

It would be best for all to start the course with computer issues
solved. Let's pause here and solve as many issues right now.

© Scott Adams, Inc./Dist. by UFS, Inc.

Computer skills

Everything you will learn in this course will serve with other projects.
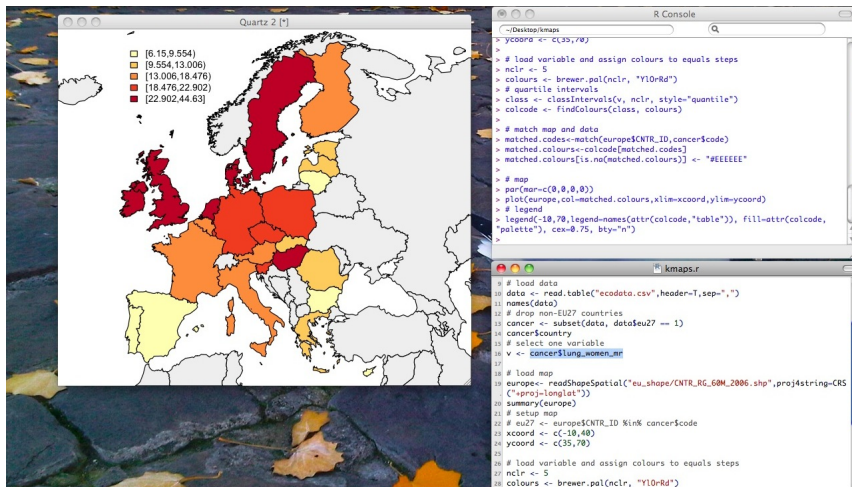
First Stata run

Open Stata and type the following commands, shown `in blue`, in the
□ **Command** window:

- `set mem 500m` will show changes in □ **Results** and □ **Review**.
  (ignore this command if you are running Stata 12+)
- `sysuse lifeexp` loads data and variables in □ **Variables**.
- `edit` opens the □ **Data Editor** with the `lifeexp.dta` dataset.
  Close the Data window.
- `scatter lexp safewater` creates your first **Stata graph**!
  Close the Graph window.
- `doedit` opens an empty, untitled do-file in the □ **Do-file Editor**.

First Stata run

- Write the following lines in your do-file:
  ```
  * First do-file.
  sysuse lifeexp, clear
  scatter lexp safewater
  clear
  ```
- Select the four lines of your do-file and use the **Tools ▷ Execute Selection (do)** menu to run them. To finish, save up your work:

  □ Save the Graph with **File ▷ Save**, using the `.jpg` format.
  □ Save the do-file, also with **File ▷ Save**, using the `.do` format.

  The do-file allows to replicate your results. Using that do-file, you can share your results on life expectancy and access to clean water with any other Stata user.

Stata application

- **Recent PC/Macs can run Stata 10/12.**
  Make sure that you have a reasonable amount
  of memory and disk space available.
- **"SE" stands for 'Special Edition.'**
  Other editions of Stata come with slightly
  different software characteristics.
- **Additional packages are downloadable.**
  The main online source is the Statistical
  Software Components (SSC) server.

Software details:
http://www.stata.com/

Operationalization

```
* Allocate 500MB memory (use in Stata 11-).
set mem 500m, perm
* Suppress screen breaks permanently.
set more off, perm
* Install the 'fre' package.
ssc install fre
```

- Lines starting with * are **comments**, others are **commands**.
  Syntax colouring (supported by Stata 11+) differentiates them.

- The perma option requires **system administrator privileges**.
  Without the options, the commands have to be run at each launch.

- Installing packages also requires administrator system privileges
  and **Internet access** to reach the SSC server.

Stata datasets

- A Stata dataset is a specific file format of quantitative data that ends with the `.dta` file extension.
- Stata opens datasets from the **File ▷ Open** menu or with the `use` command, often with the `clear` option.
- Other formats are supported through import commands, e.g. `insheet` for [comma-separated values] CSV files.

Examples:

- **Datasets** in the SRQM folder
- **Course Website**

Operationalization

Load the National Health Interview Survey (2009):

- Using the **File ▷ Open** menu:
    - □ Select "**Stata dataset (.dta)**" in the file extension menu.
    - □ Open **Teaching Pack ▷ Datasets ▷ nhis2009.dta**.

- Using the Stata command line:
  ```
  * Set the path to your Teaching Pack folder.
  cd "/Users/fr/Documents/SRQM/"
  * Load the dataset by calling its file name.
  use "Datasets/nhis2009.dta"
  ```

Note: the working directory setting called by the `cd` command is system-dependent and must reflect your own folder hierarchy.

Stata do-files

- A do-file is a plain text file holding a list of Stata commands, that can be edited from any **plain text editor**.
- Stata opens and edits do-files from the **File ▷ Open** menu or through the `doedit` command.
- The `do` and `run` commands will try to execute a full do-file in order to replicate its results.

Examples:

- **Teaching Pack ▷ Replication**
- **Course Website**

Operationalization

- Using the **File ▷ Open** menu:
  - □ Select "**Stata do-file (.do)**" in the file extension menu.
  - □ Open **Teaching Pack ▷ Replication ▷ week1.do**.
- Locate the following commands:
  ```
  d year sex weight raceb // describe a few variables
  keep if year==2009 // keep observations for year 2009
  ```
- Select both command lines and run (execute) them together:
  - □ You can select * commented lines when running a command, as they have no effect on how Stata executes the rest of the code.
  - □ Type **Ctrl-D** (PC) or **Cmmd-Shift-D** (Mac) to run the selected command(s) from the keyboard.
- Alternately, use the **"Execute (do)" icon** from the do-file toolbar:

Other components

- **Log files** are plain text files that hold commands and their results, used for replication purposes.
  - □ Stata has its own SMCL format but we will prefer plain text `.log` files.
  - □ Log files are optional but recommended to keep track of your work.

- **Graphs** are picture files that you produce through specific Stata commands, for visualization purposes.
  - □ Stata has its own GPH format but we will prefer saving as `.jpg`.
  - □ Use the `name()` option to store a graph and leave its window open, and then use the `graph export` command to save it on disk.
  - □ Alternately, save a graph manually with the **File ▷ Save** menu, or with the **"Save" icon** from the graph toolbar.
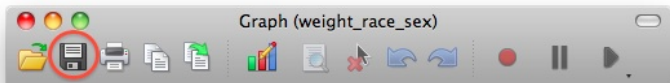
Operationalization

Return to Stata and to the **week1.do** file.

- To log your work, type:
  ```
  log using week1.log, replace
  ```
  - ☐ When opening a log, remember to use the `.log` extension.
  - ☐ The log file will be saved in your working directory. The `replace` option overwrites any previously log file at that location.
  - ☐ Stata automatically stops logging when you quit it. The `log close` and `exit` commands respectively perform the same operations.

- To plot a variable as a histogram, type:
  ```
  histogram weight, bin(20) name(weight, replace)
  ```
  - ☐ When saving the graph, remember to select the **JPG** format.
  - ☐ The `replace` option is required to overwrite any previous graph with that name. Use it systematically to avoid errors.
  - ☐ This graph uses the `bin` and `name` options, and might require further options to produce an appropriate visualization of the variable.

Getting help

Quantitative methods systematically require extensive documentation.
Locating help resources is a course requirement and a skill in itself.
When asked for help, we will frequently refer to these resources:

- **Course material**:
  - □ Stata Guide
  - □ Course slides and emails
  - □ Handbook chapters

  Start reading the **Stata Guide.**

- **Other material**:
  - □ Stata `help` pages
  - □ Stata tutorials

  Start exploring the **Course Website.**

Stata Guide

The Stata Guide is a course handbook which we co-write with Ivaylo. Its chapters cover virtually all course sessions.

The Guide is work in progress: please email comments, questions and corrections, and watch for updates during the course.

Statistics with Stata

Student guide

Version 0.9.6.7, by François Briatte

Contents

Course slides

The course slides contain very basic guidance and are absolutely insufficient to complete the course requirements.

More generally, slideware is a teaching aid, not a learning tool. Watching slides does not compensate documentation.

Copyright © 2000 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited

Course emails

The course emails will be our lifeline, with insanely detailed instructions about your homework, and more. Read them carefully.

We will also answer all emails about the course as quickly as possible. **Include both instructors on emails.**

Do not forget to email with a subject that starts with **"SRQM:"** to help us parse our mailboxes, as in "SRQM: Question on recoding."

Possible email subjects:

"SRQM: Issue with plotting sex, drugs and rock'n'roll"
"SRQM: Assignment No. 2, Briatte and Petev"
"SRQM: Where has the cherry blossom of our youth gone?"
"SRQM: Why are scatterplots with ordinal data so ugly?"

Handbook chapters

"The fact that they looked it up in a book just shows that they don't get the idea of truthiness at all... You don't look up truthiness in a book, you look it up in your gut."
(Stephen Colbert to the Associated Press)

- **Understanding the statistical theory** that underlies the rest of your practical tasks will inevitably show up in your analysis.
- **Truthiness will just not work with us**. It might with a sloppy newspaper or polling institute, but not with academics.
- **Aim at understanding the theory once**, which is enough at the introductory level. Once is enough, but once, no less.

Stata `help` pages

Type `help su` and skim-read the page:

- **Syntax indications** help correcting a great deal of coding errors.
- **Options lists** help improving your code, and especially graphs.
- **Examples** illustrate how the commands can be used effectively.

Remarks on Stata command syntax:

- `su` is the abbreviated version of the `summarize` command.
  Command shorthands are underlined in the Stata help pages.
  You can abbreviate tons of Stata commands and options: `h li`,
  for example, opens the `help` page for the `list` command.
- Variable names cannot be abbreviated, but you can refer to groups
  of variables: `su v1-v5` is equivalent to `su v1 v2 v3 v4 v5`, and
  `su var*` will treat all variables named `var1 var2 var3` etc.

Stata tutorials

Reading documentation goes hand in hand with lots of practice:

- **Replicate all course sessions** to get familiar with the commands introduced during class.
- **Read the Stata Guide** and work through the examples, while working on your own data and project.
- **Find online tutorials** on the specific aspects that you find most problematic, such as recoding.

Links to the most useful resources:
http://f.briatte.org/teaching/quanti/#stata

Examples of excellent online tutorials:
http://www.princeton.edu/wwac/academic-review/stata/
http://www.ats.ucla.edu/stat/stata/default.htm

Welcome, and thank you.