Words of Wisdom on Data and Analysis

Senior Seminar
February 4, 2016

Frank Nitsche and Jacqueline Klopp

General Advice

- **Look at the methods used in related work**
  (other papers, especially those by your mentor and their colleagues)

- **Talk to your mentor**
  (or their lab managers, graduate students, postdocs, etc.)

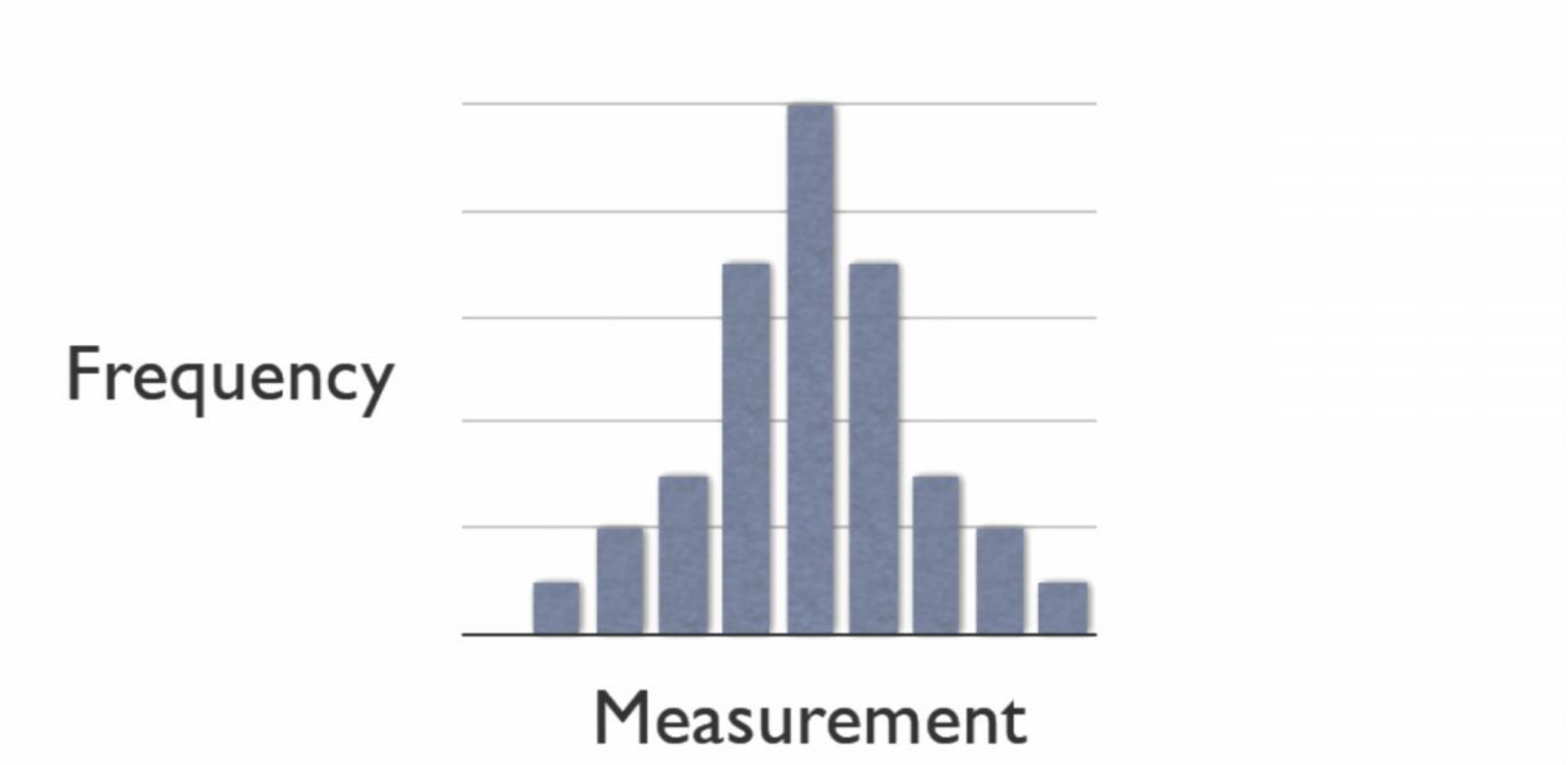
- **Use the statistical consulting service:**
  consult@stat.columbia.edu

- **Read a (portion of a) book!**

Author's Responsibility

- **Know what your statistical methods do!**
- **Be aware of the assumptions and limitations of your statistical tests**
- **Report all the proper results**
- **Understand what your results mean (Plot and look at your data, do the data make sense?)**

Variation



This will usually be a Normal Distribution, but not always

Reporting Variation

- **Every measure which summarizes a distribution (e.g., a mean) should include some measure of spread (e.g. a standard deviation)**

- **A graph without error bars is incomplete and potentially misleading!**

- **(Need to show whether differences between data are significant)**

Hypothesis Testing

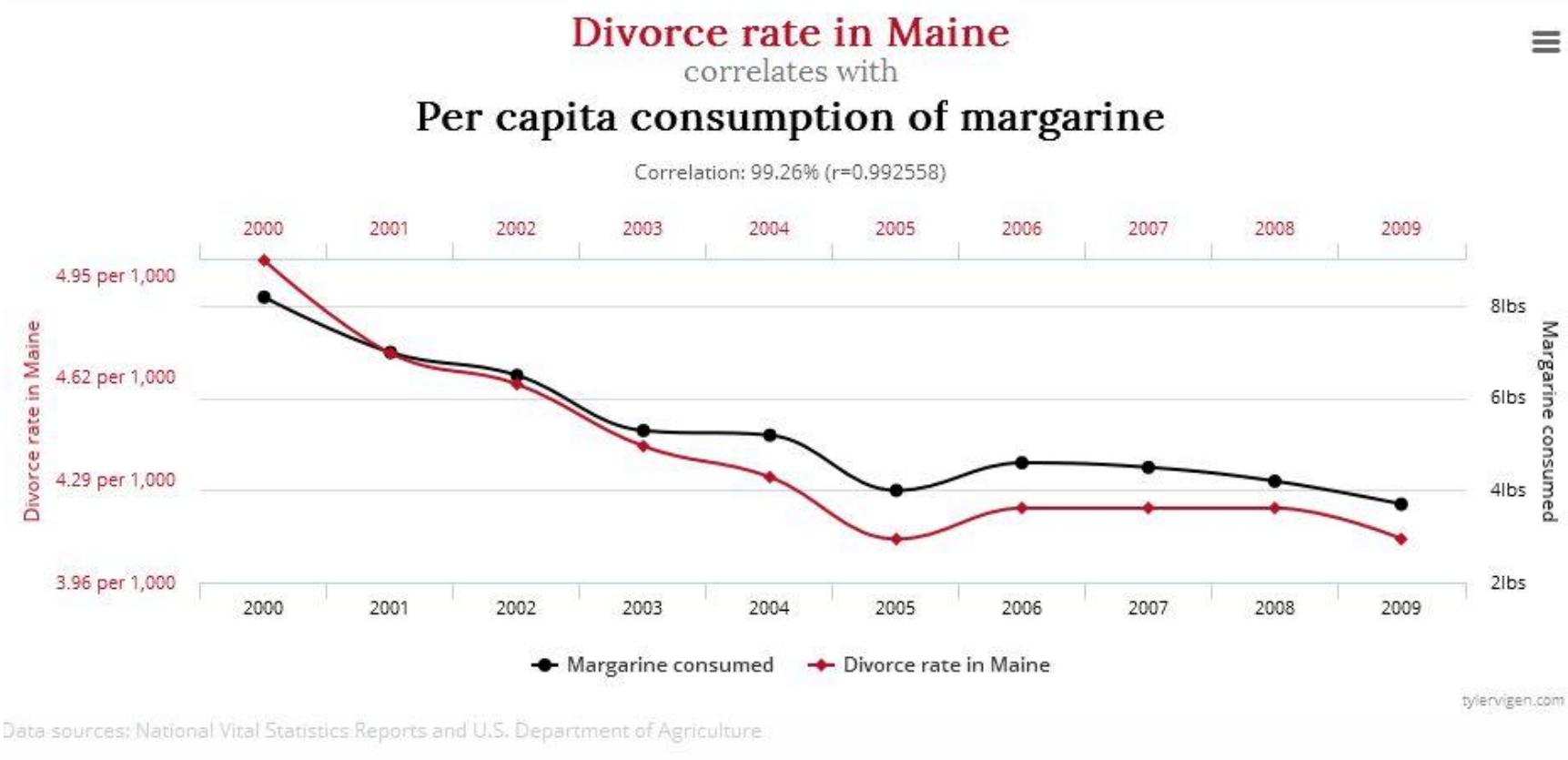
- **Comparing two or more hypotheses in light of the data**
- **Scientists generally make a null hypothesis of no effect - any variation in the data is just random**
- **We reject the null when the data deviate strongly from random. This lends support to the hypothesis that some phenomenon is responsible for part of the variation**

Testing and Probability

$$P = \frac{\text{number of outcomes}}{\text{number of trials}}$$

So P=0.05, means you expect an outcome one time in 20 trials, by chance.

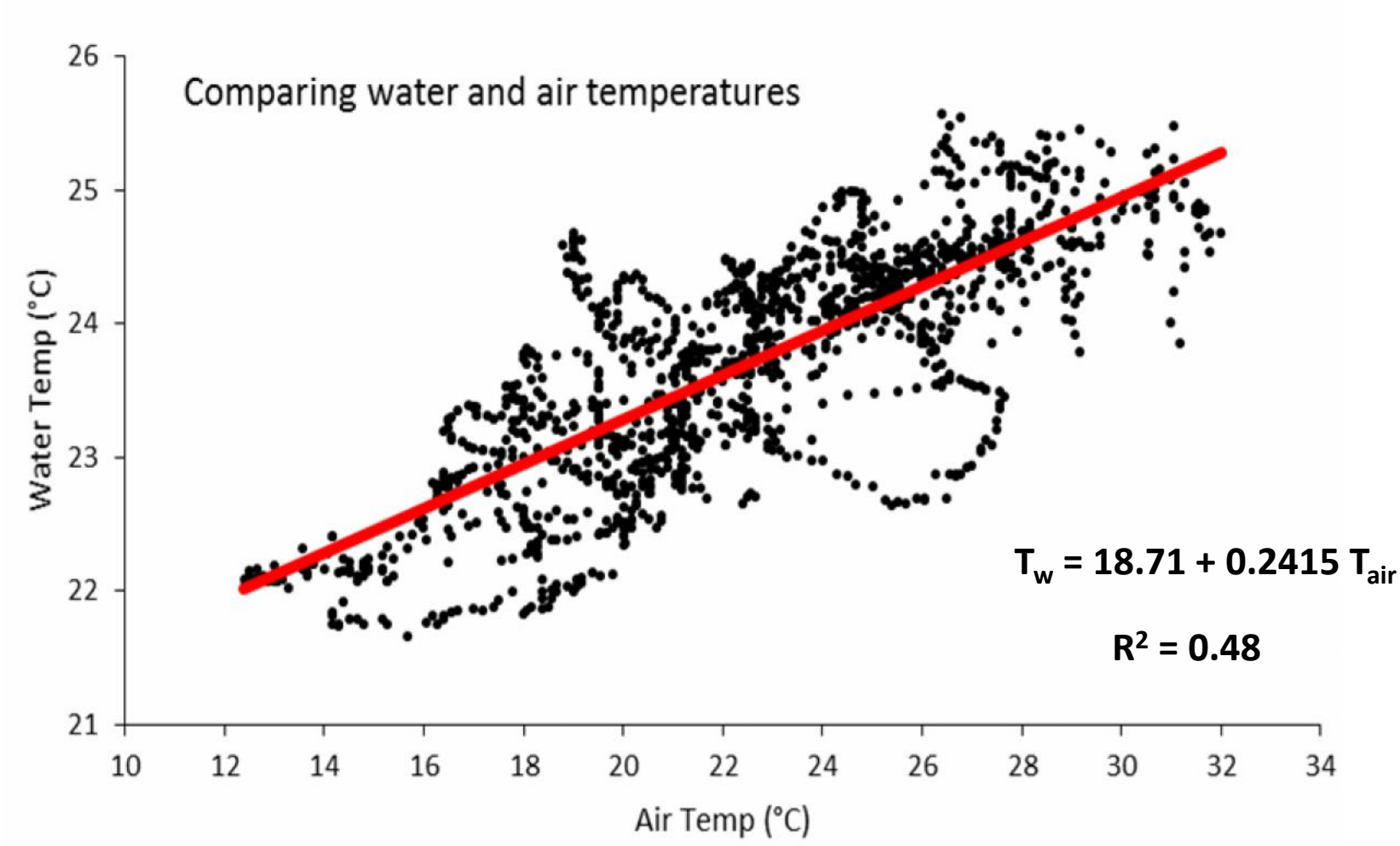
Correlation



Be careful:

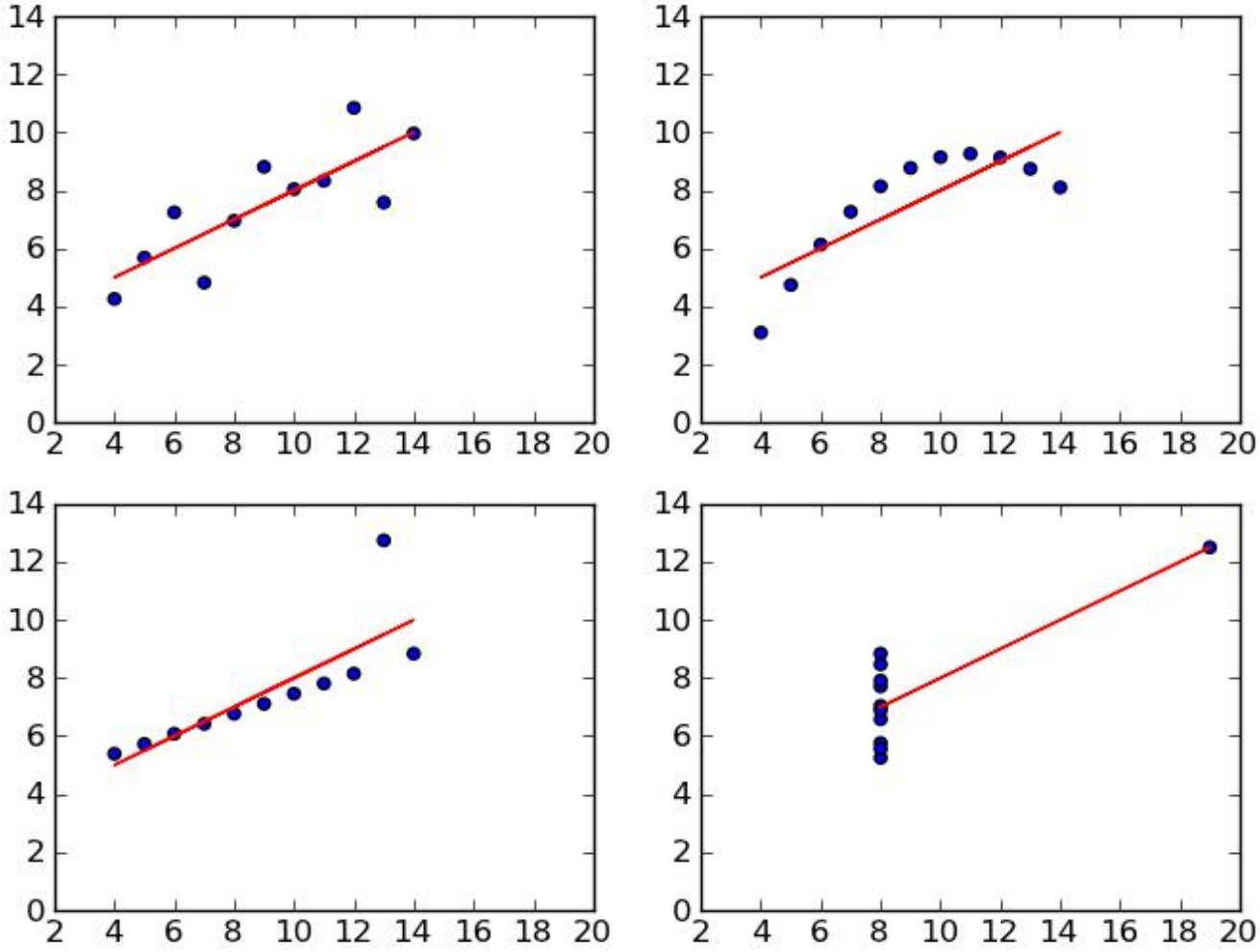
- **Correlation ≠ Causation**
- **Small number of data points could lead to false apparent correlation**

Regression



Always Present Regression with R^2 value

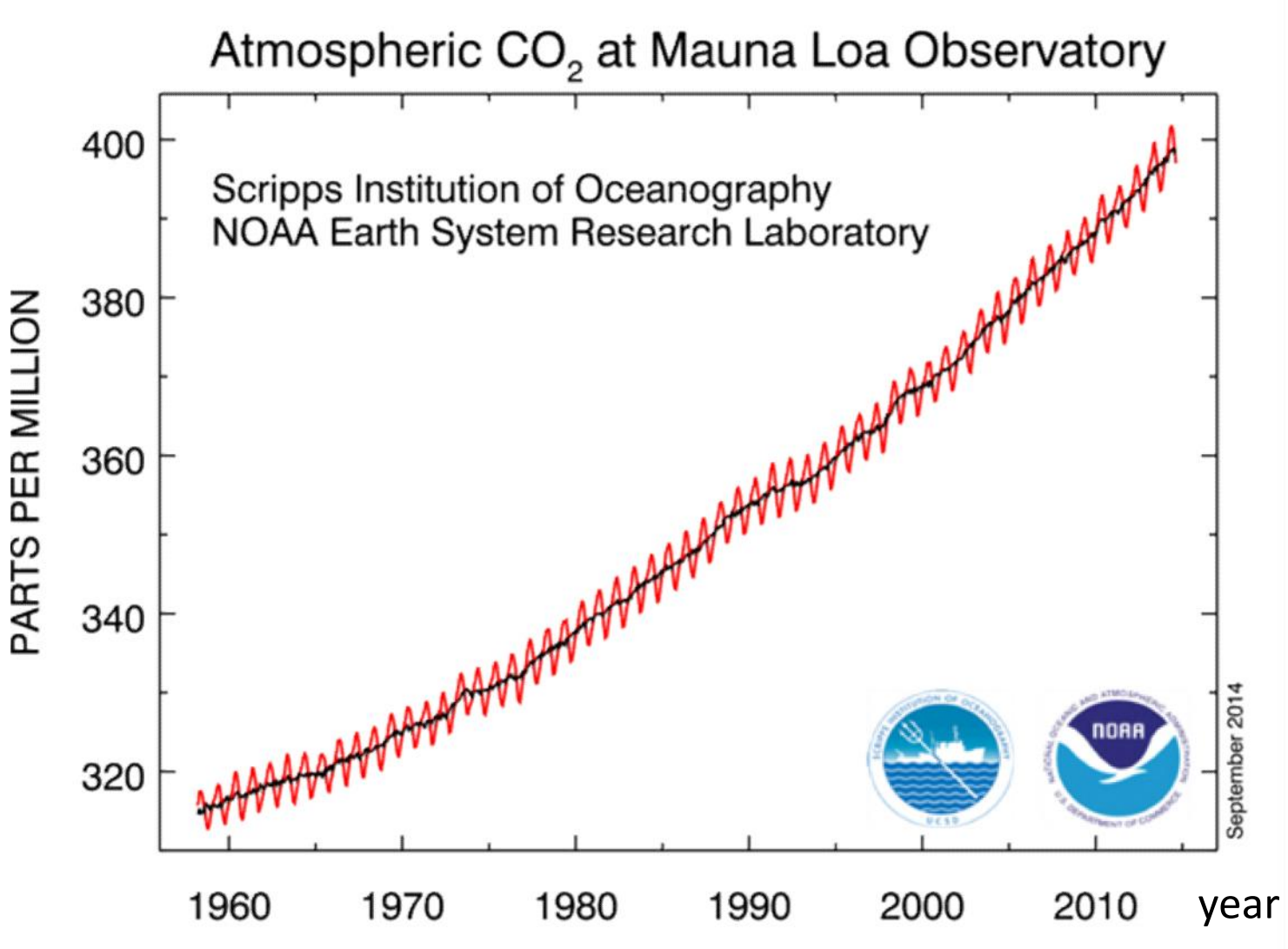
Regression



All of these datasets have the same regression and correlation

=> Plot your data to check, if regression and correlation make sense

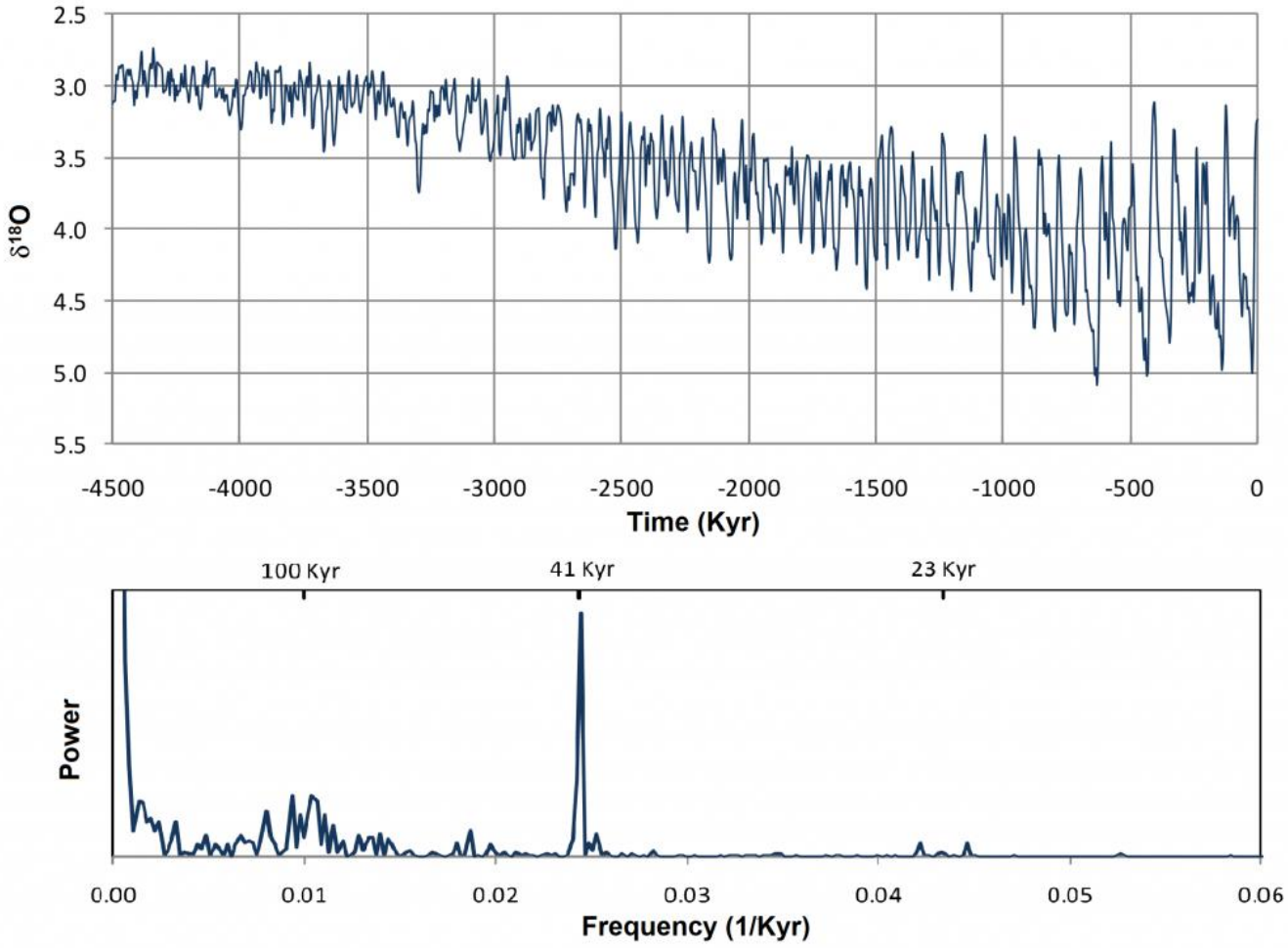
Time Series



Cycles (e.g. seasons), trends, smoothing

Time Series

Time series and power spectrum of the Earth's climate record for the past 4.5 Myr.



Be aware of limits: total time span, sampling rates, aliasing

Software Tools - Statistics

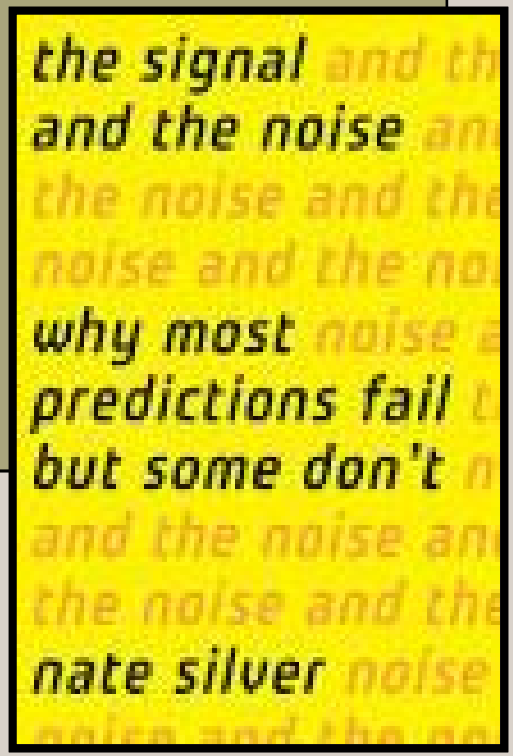
- **Excel (data analysis ToolPak add-on)**
- **LibreOffice / Open Office**

- **Stata**
- **SPSS**
- **Origin**

- **R**
- **Matlab**
- **Python**

"The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning."

-Nate Silver

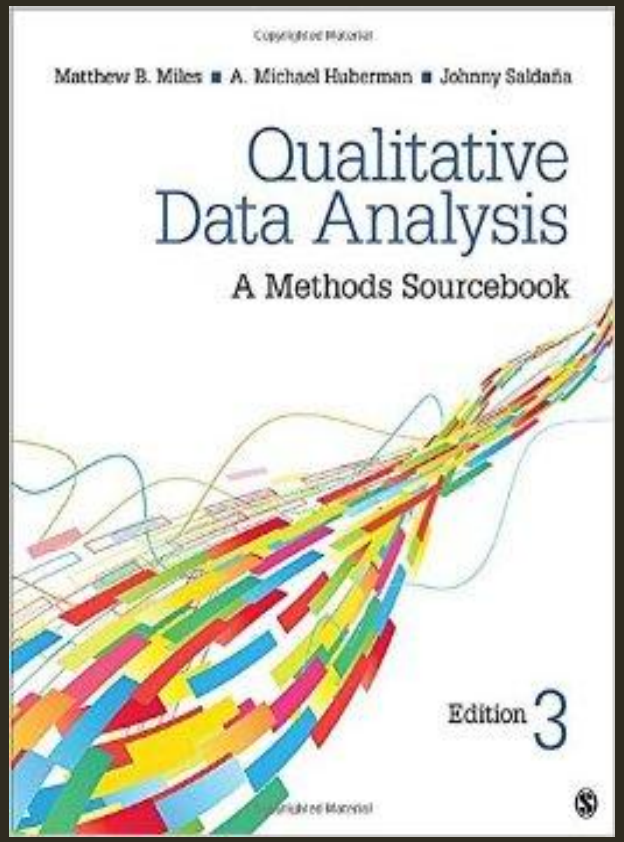


Power of Qualitative Analysis

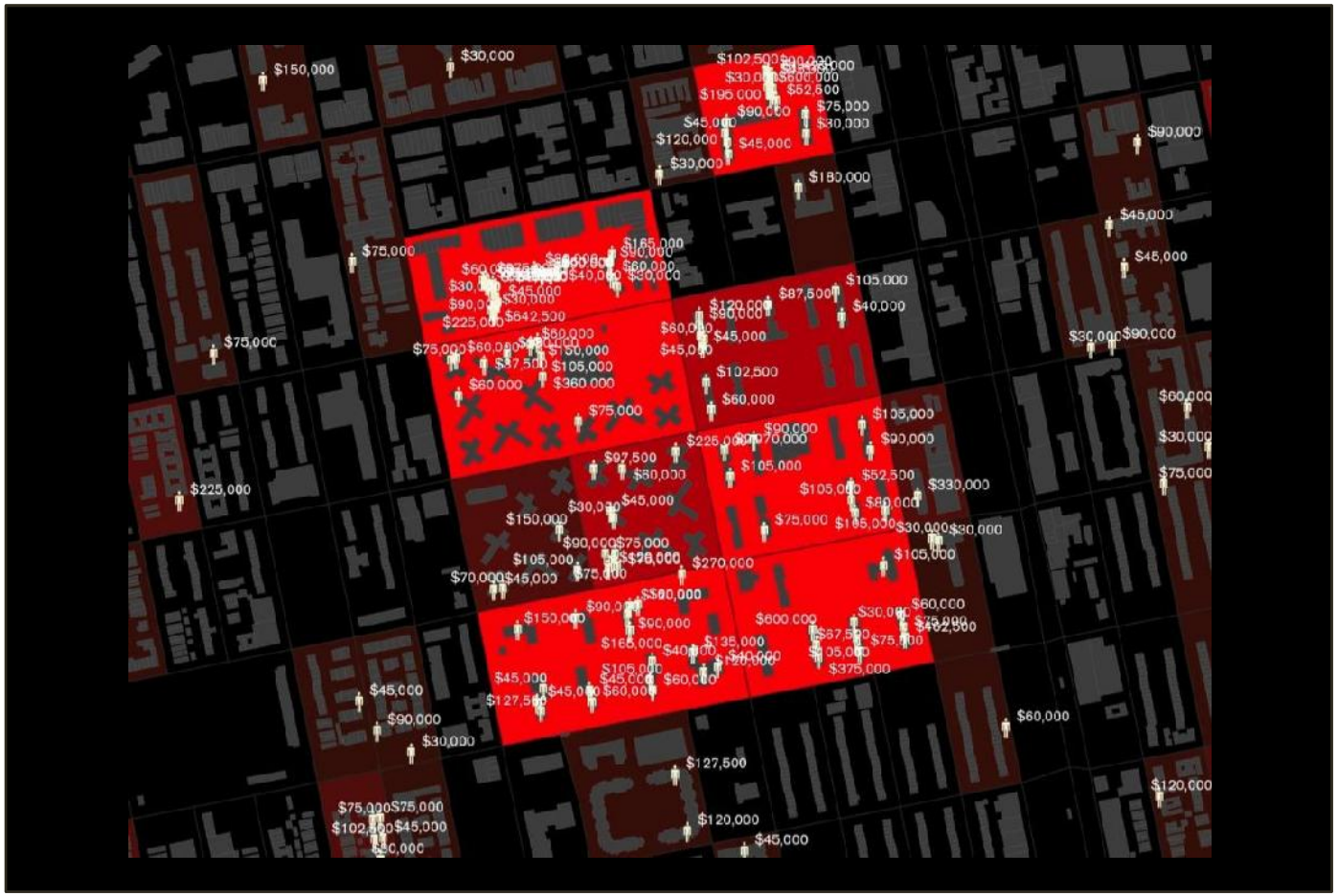
- Corroboration via triangulation
- Develop analysis, detail
- Surprises, paradoxes, new insights

"Not everything that can be counted counts, and not everything that counts can be counted"

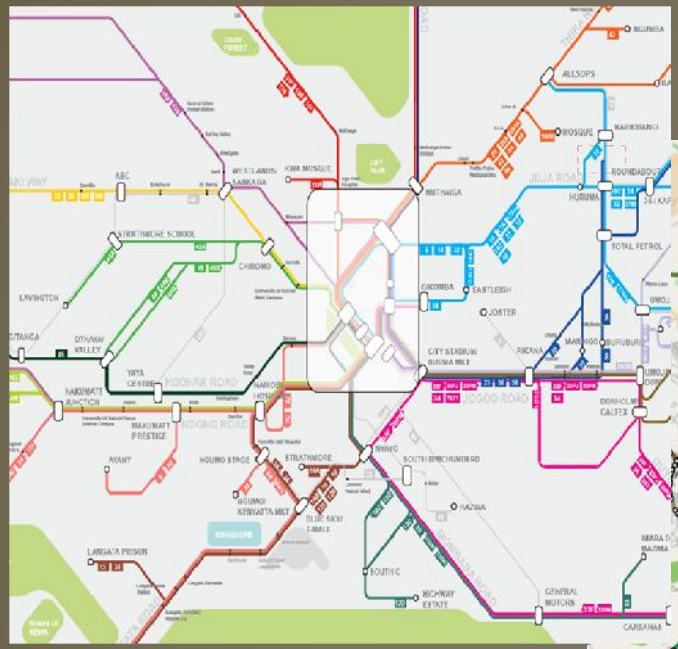
– William Bruce Cameron

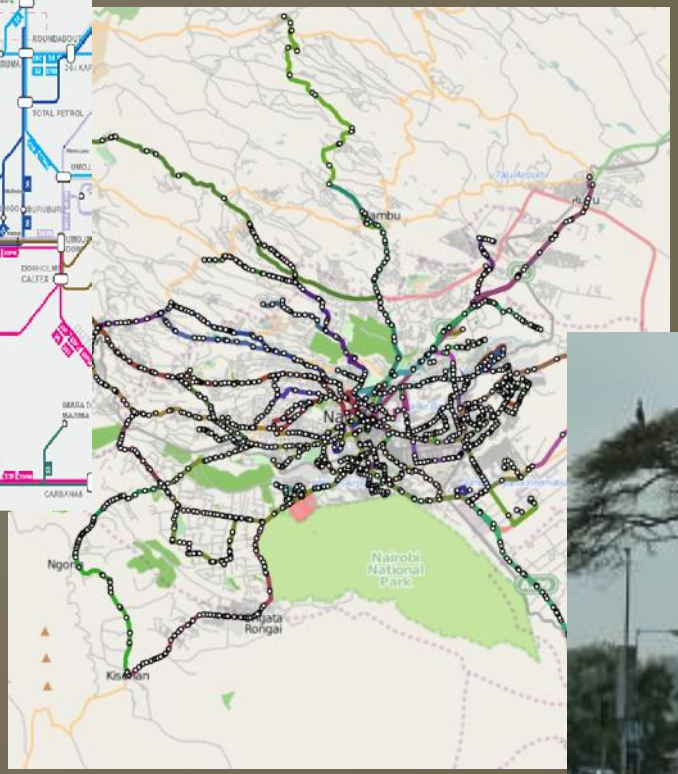


POWER OF SPATIAL ANALYSIS



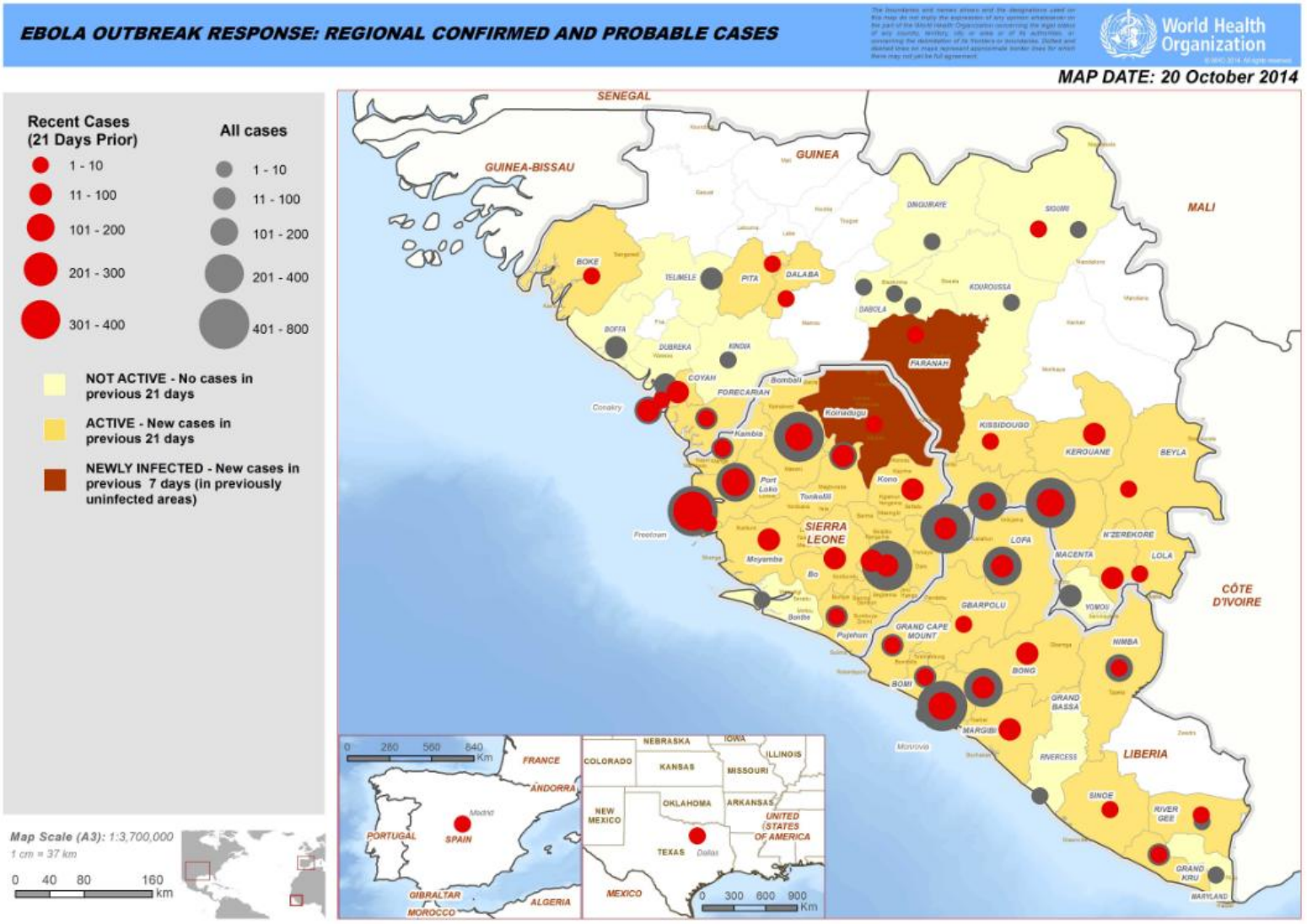
The Power of Visualization



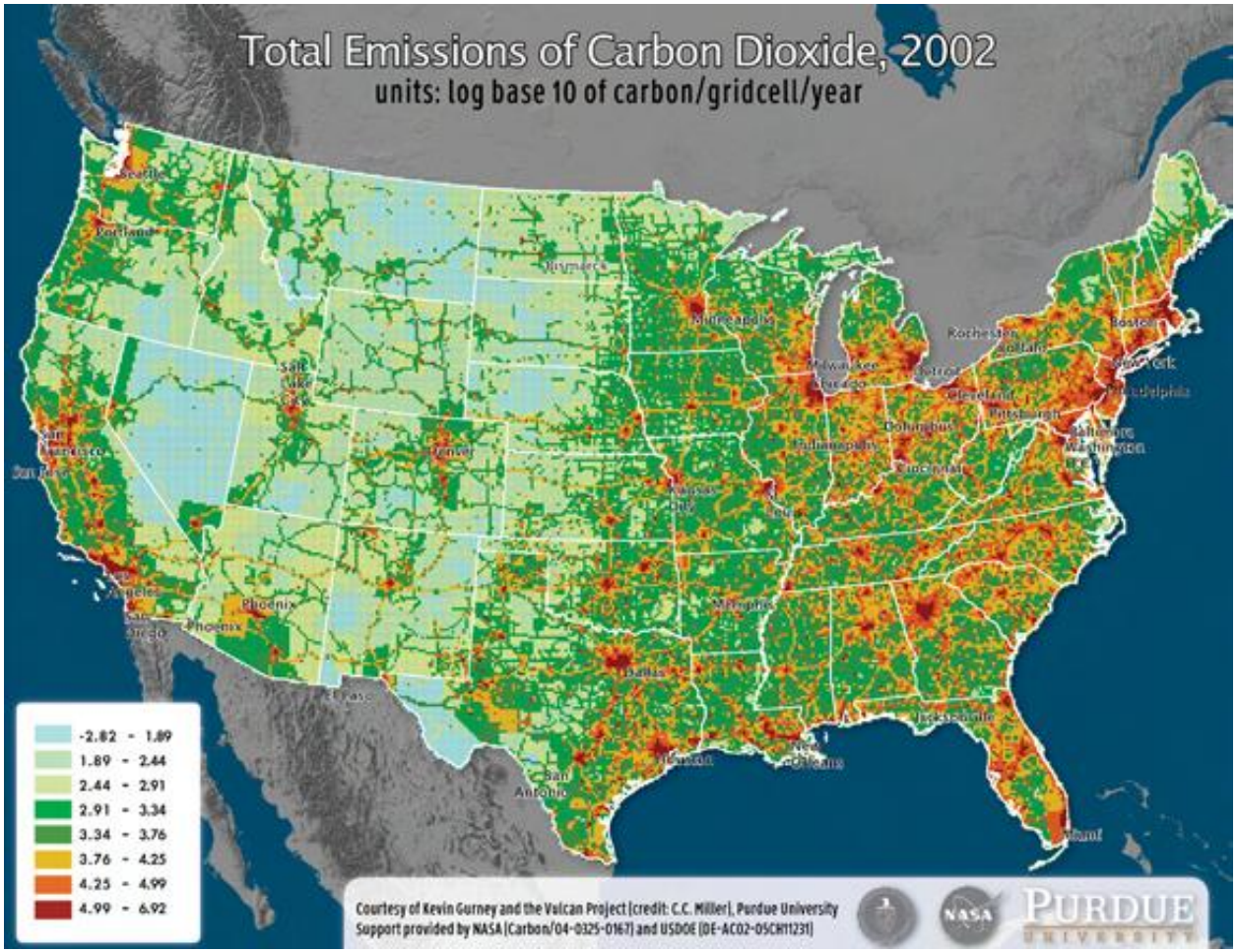


GIS / Spatial Analysis

Visualizing Data



GIS / Spatial Analysis



- **Maps are good tools to visualize spatial data**
- **GIS allows complex spatial analysis**

Software Tools – GIS

- ArcGIS
- QGIS (Open Source alternative)

- ENVI (Remote Sensing)

Resources:

- **Digital Social Science Center**
  http://library.columbia.edu/locations/dssc.html
- **CIESIN GIS resources**
  http://www.ciesin.org/gisservicecenter/resources.html

E.g. DSSC ArcGIS intro workshop Feb 10
http://library.columbia.edu/research/workshops.html

Software Tools – Advice

- **Check with your mentor**
  (What is she/he/the group using?)

- **Find somebody who can guide you**

- **Online Resources**
  - “How-To” instructions online
  - online tutorials

- **Introductory books**

Other Resources on Campus

(1) Barnard College Empirical Reasoning Lab:

(located in the Barnard Library)
http://erl.barnard.edu/

(2) CU - Digital Social Science Center

(Lehman Social Science Library)
http://library.columbia.edu/locations/dssc.html

(3) CU Dept. of Statistics

(they offer statistics consulting)
http://stat.columbia.edu/consulting-information/

(4) Applied Statistic Center

http://applied.stat.columbia.edu/
(see their consulting tab)

Playroom time

On Tuesdays from 2:15pm-5:15pm, people from the Applied Statistics Center are available in the Playroom (IAB 707)..

For software and basic data problems try the two library options (1) and (2) first.
They also offer GIS advice.