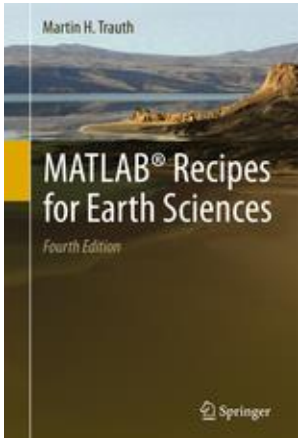
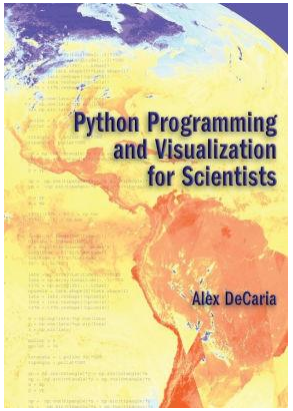
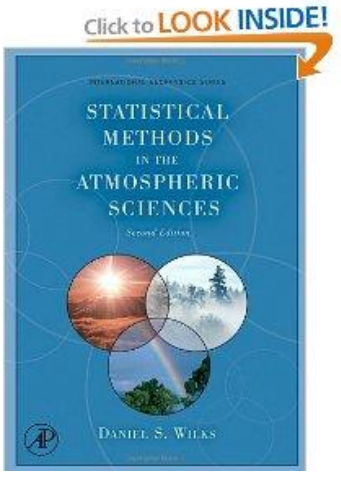
ATMOSPHERIC SCIENCES 5040/6040- Environmental Statistics

- Instructors: John Horel, Court Strong
- Grader: Chris Foster
- Required Text: *Matlab Recipes for Earth Sciences.*
- Available as eBook through library. The eBook can be read online or checked out for 24 hours
- Recommended if interested in python: *Python Programming*
- Recommended text for 6040: Statistical *Methods in the Atmospheric Sciences*







What you should be doing

- Assignment 1- Survey- Due Tonight
- Read Chapter 1 Notes
- Read Chapter 1 in text
- No class Wednesday
  - Spend time reading above
  - Complete Assignment 2 based on Chapter 1 notes and text
  - Due Sunday night midnight

Course Learning Objectives

- State and use basic statistical metrics to analyze environmental information
- Develop proficiency to program and use computer software to analyze environmental data sets
- State and demonstrate the characteristics of effective research; organize, quality control, and find relationship(s) among data

Official Syllabus

- In Canvas
- Direct link to course: https://utah.instructure.com/courses/541131

When you have questions…

- email: john.horel@utah.edu
- Use the class slack channel
  - (will be discussed next week)
- Office hours: by appointment
  - Send email, a message in slack or phone

5040/6040 Course Outline

- Week 1. Jan. 7, 9. Basic concepts
- Week 2. Jan. 14, 16. Exploratory Univariate Data Analysis
- Week 3. **Jan. 21. Holiday. No class.** Jan. 23. Transforming Data
- Week 4. Jan. 28, 30. Directional Data & Probability
- Week 5. Feb 4, 6. Parametric Distributions
- Week 6. Feb. 11, 13. Exploratory Multivariate Data Analysis. Correlations
- Week 7 **Feb. 18. Holiday**. **No class.** Feb 20. Exploratory Multivariate Data Analysis. Compositing
- Week 8. Feb. 25. First Half Wrap Up. Take Home Final-5040/Midterm-6040. Due March 6

6040 Course Outline

- Weeks 9-10. Topics in regression (multiple and curvilinear regression)
- Weeks 11-12. Harmonic analysis (spectral analysis and filtering)
- Weeks 13-14. Data reduction (empirical orthogonal function analysis)
-

Format

- Meet in 820 WBB on most** Mondays, MLIB 1110 on Wednesdays
- You must read the assigned text and class notes prior to the corresponding lecture
- During the first half semester, there will be an assignment due nearly every class day
- Assignments completed late will receive at most partial credit
- Take Home Exam due- March 6: (5040: final); (6040: midterm)
- 6040: Instead of a final exam, you will pursue a research question using statistical methods to analyze data

Class Policies and Grading

- 5040:
  - (1) final (33%)
  - (2) class assignments (66%)
- 6040:
  - (1) assignments (60%)
  - (2) mid-term exam (20%)
  - (3) research project (20%)

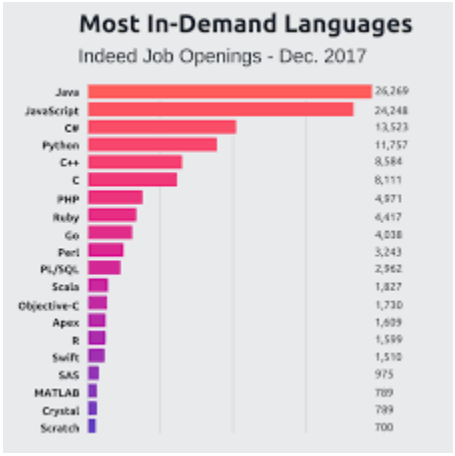
Programming Languages

- Programs written in a specific language are just variations on ways to pass instructions to a computer
- Each language has its own syntax (form) and semantics (meaning)
  - **Syntax**: Sequences of text including words, numbers, and punctuation using rules like written languages (grammar)
  - **Semantics**: The meaning given to the syntax - a sequence of words that makes sense to a computer

Popularity of programming languages

https://www.tiobe.com/tiobe-index/

Dec 2018	Dec 2017	Change	Programming Language	Ratings	Change
1	1		Java	15.932%	+2.66%
2	2		C	14.282%	+4.12%
3	4	↑	Python	8.376%	+4.60%
4	3	↓	C++	7.562%	+2.84%
5	7	↑	Visual Basic .NET	7.127%	+4.66%
6	5	↓	C#	3.455%	+0.63%
7	6	↓	JavaScript	3.063%	+0.59%
8	9	↑	PHP	2.442%	+0.85%
9	-	⇈	SQL	2.184%	+2.18%
10	12	↑	Objective-C	1.477%	-0.02%
11	16	⇈	Delphi/Object Pascal	1.396%	+0.00%
12	13	↑	Assembly language	1.371%	-0.10%
13	10	↓	MATLAB	1.283%	-0.29%



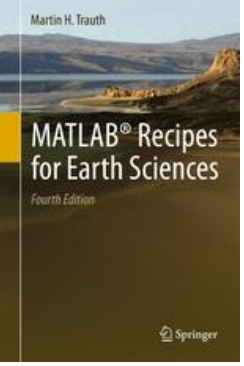
vs.

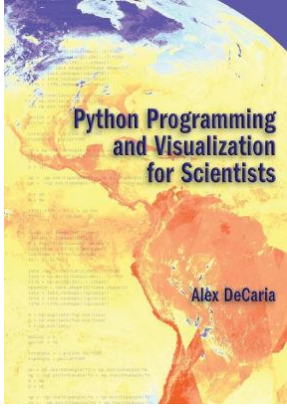
Matlab

- Proprietary
- Inexpensive for educational uses, expensive otherwise
- Flexible graphical user interface
- Many toolboxes
- Matrix oriented

Python

- Open source
- Free, although enterprise releases for commercial applications
- Notebooks are a convenient way to learn
- Many modules
- Object oriented





In this course:

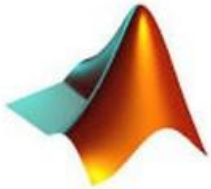
Matlab

- All codes available in Matlab
- What nearly everyone should use
- Matlab text required

Python

- Nearly all codes available in python
- Some may want to do some exercises using python
- Intro text recommended

- Everything in one github repository (notes, codes, data): https://github.com/johnhorel/atmos_5040_2019
- Programming language reviews from ATMOS 5020 in repository
- Notes independent of language
- It may be possible to have additional lab time on Wednesdays after the official class period ends



vs.

Matlab

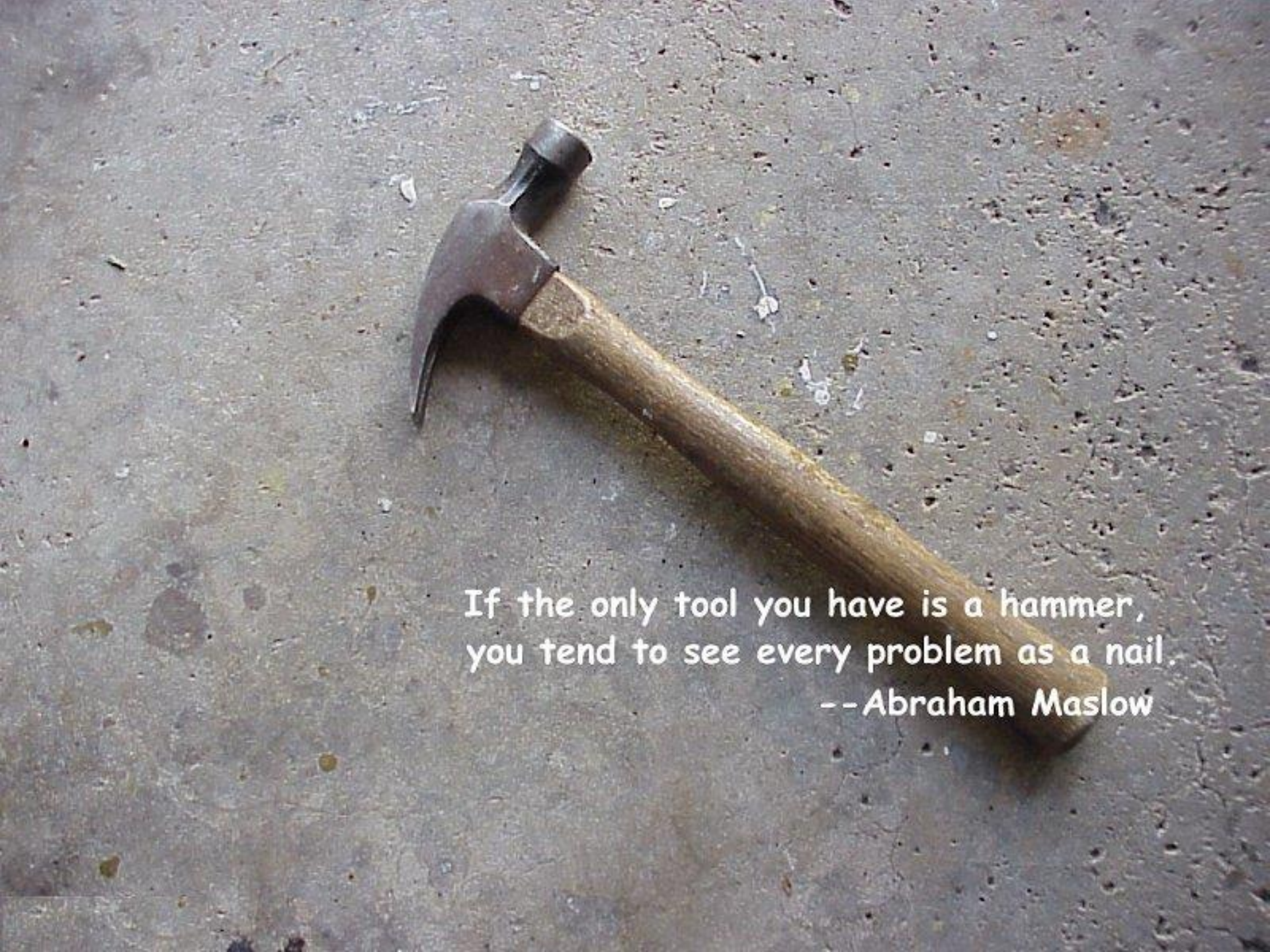
- We will be using Marriott Library classroom
- you will need to do assignments using the library or other locations that have Matlab with statistics tool box installed
- Not required to own a copy of Matlab but may be helpful. Costs $30
- Review Matlab material in Canvas and github repository

Python

- Python installed on Marriott Library computers
- We will use python notebooks at times to show how python handles statistical calculations
- Instructions provided on how to install python locally, but not required
- Not required to look at, but lots of python material in github repository

Problems with statistics?

- Oriented towards confirming preconceived ideas?
- Start with a technique and look for a data set to apply it to?
- There’s always two sides to every issue?
- What do you think? Other examples of poor statistics?


If the only tool you have is a hammer,
you tend to see every problem as a nail.
--Abraham Maslow

Questioning traditional statistical approaches

- **Nassim Nicholas Taleb**
- http://www.fooledbyrandomness.com/
- Black Swan: extreme impact of rare and unpredictable events and human tendency to search for simplistic explanations for these events retrospectively
- Antifragile: Some things benefit from shocks and thrive when exposed to volatility, randomness, disorder, stressors, and uncertainty

Oddball statistics?

https://www.statnews.com/2016/05/09/john-oliver-bad-science/

Warning: some inappropriate language

The Health Risks of Being Left-Handed

Lefties Face Chance Of ADHD, Other Disorders; Brain Wiring Holds Clues

- http://www.wsj.com/articles/SB10001424052970204083204577080562692452538
- Modern lefty lore says left-handers are smarter, more creative and have an advantage over righties.
- About 10% of people are left-handed, Six of the last 12 U.S. presidents, including Barack Obama and George H. W. Bush, have been lefties.
- Babies born to older mothers or at a lower birth weight are more likely to be lefties
- On average there is no difference in intelligence between right-and left-handed people.
- Left-handed people earn on average 10% lower salaries than righties
- lefties aren't more accident prone than right-handed people yet don't tend to die at a younger age.
- Left-handedness appears to be associated with a greater risk for a number of psychiatric and developmental disorders.
  - About 20% of people with schizophrenia are lefties even tho 10% of people are lefties
  - 1% of general population has schizophrenia

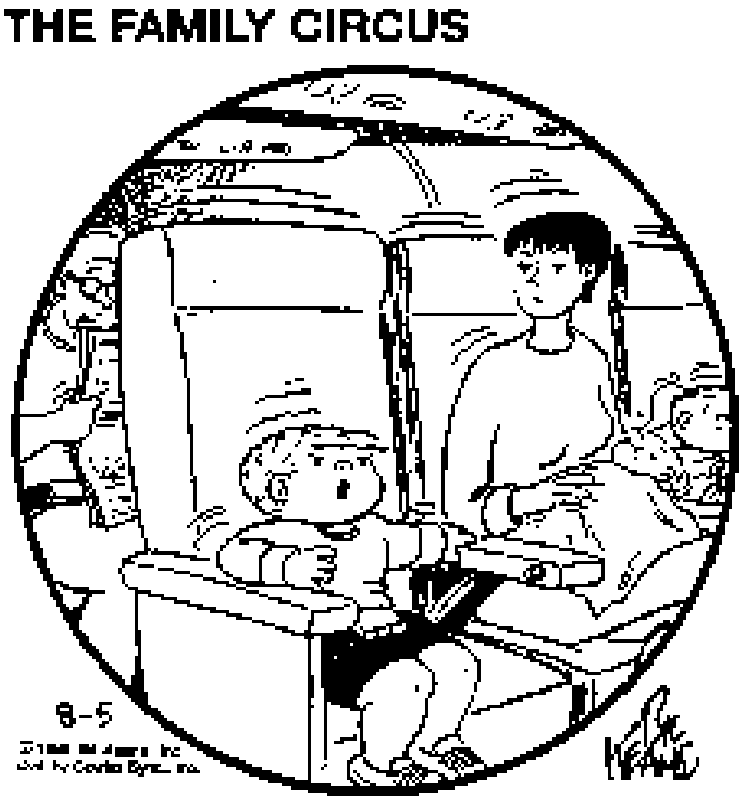
Scientific Explanations are:

- based on empirical observations or experiments
- tentative
- historical
- probabilistic
- assume cause-effect relationships
- limited
- made public
- influenced by individuals and culture

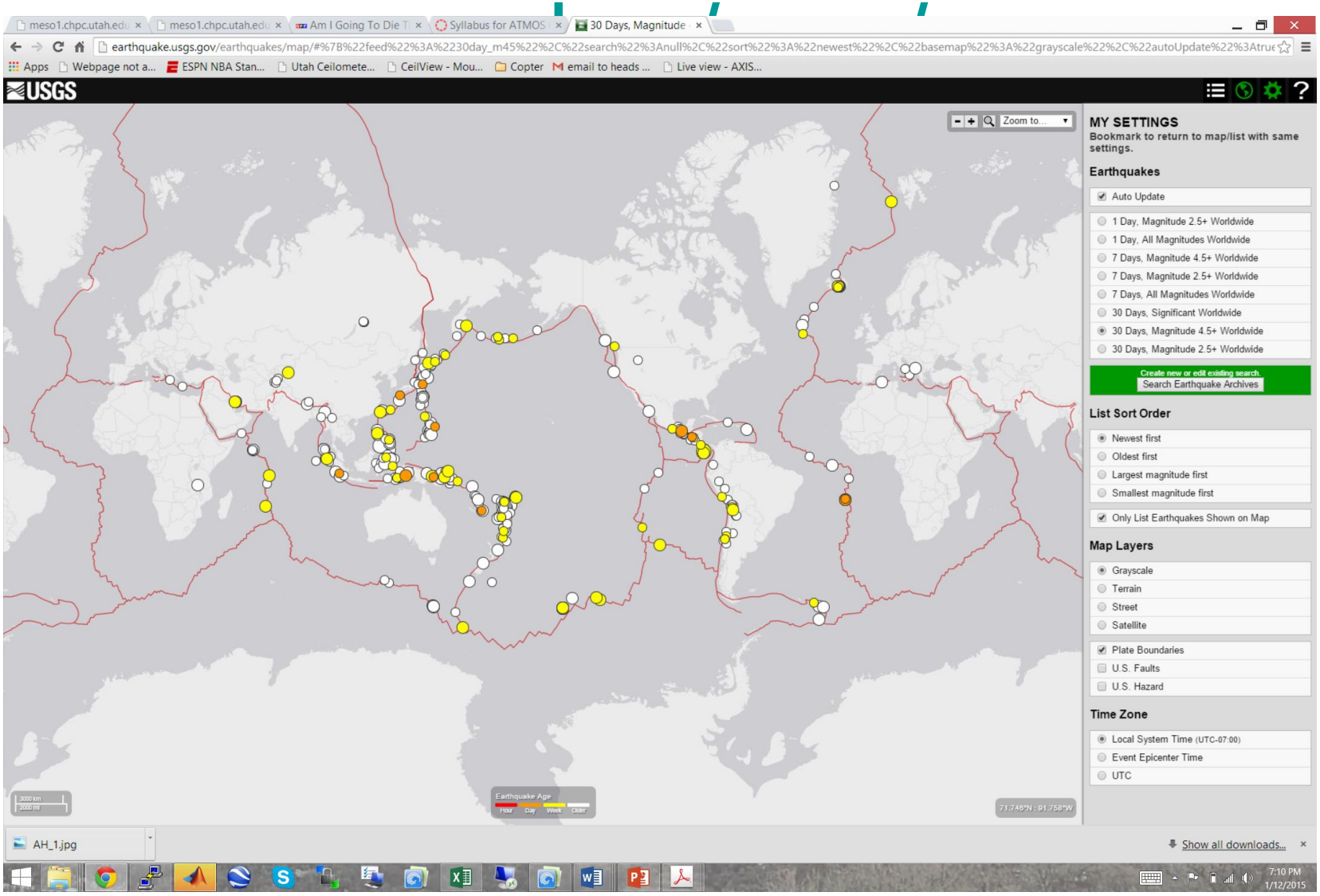
What's the Goal??

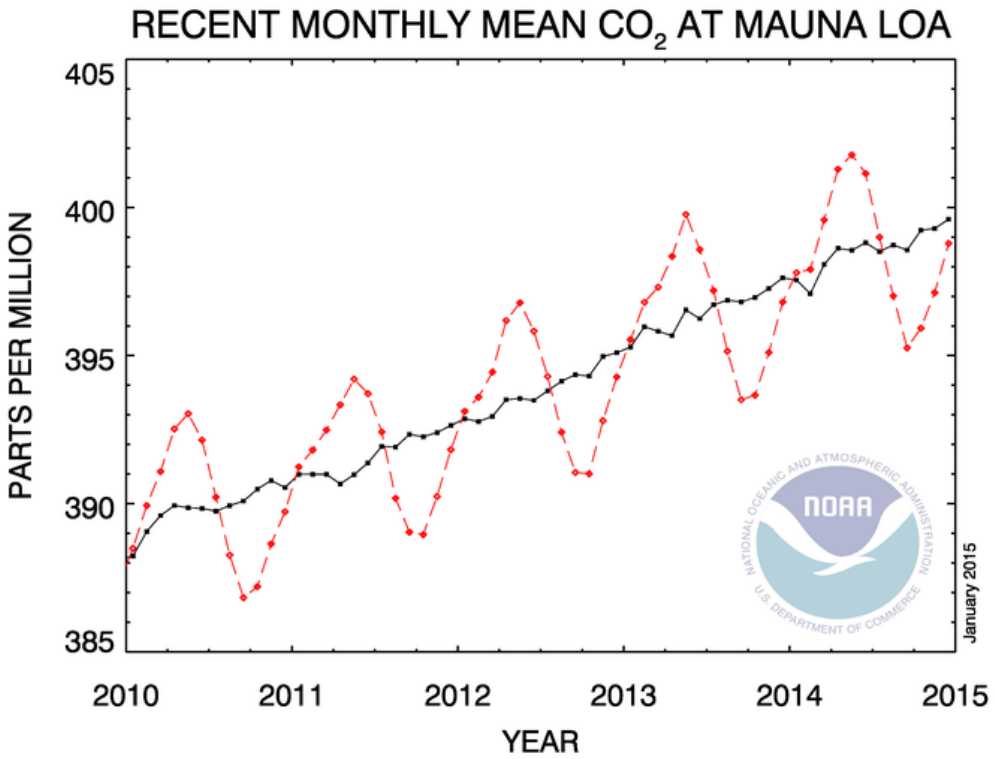
- Exploratory or descriptive statistics:
  - Organize and interpret volumes of data

- Inferential statistics:
  - Assess the underlying physical processes that generate environmental data



http://earthquake.usgs.gov/earth

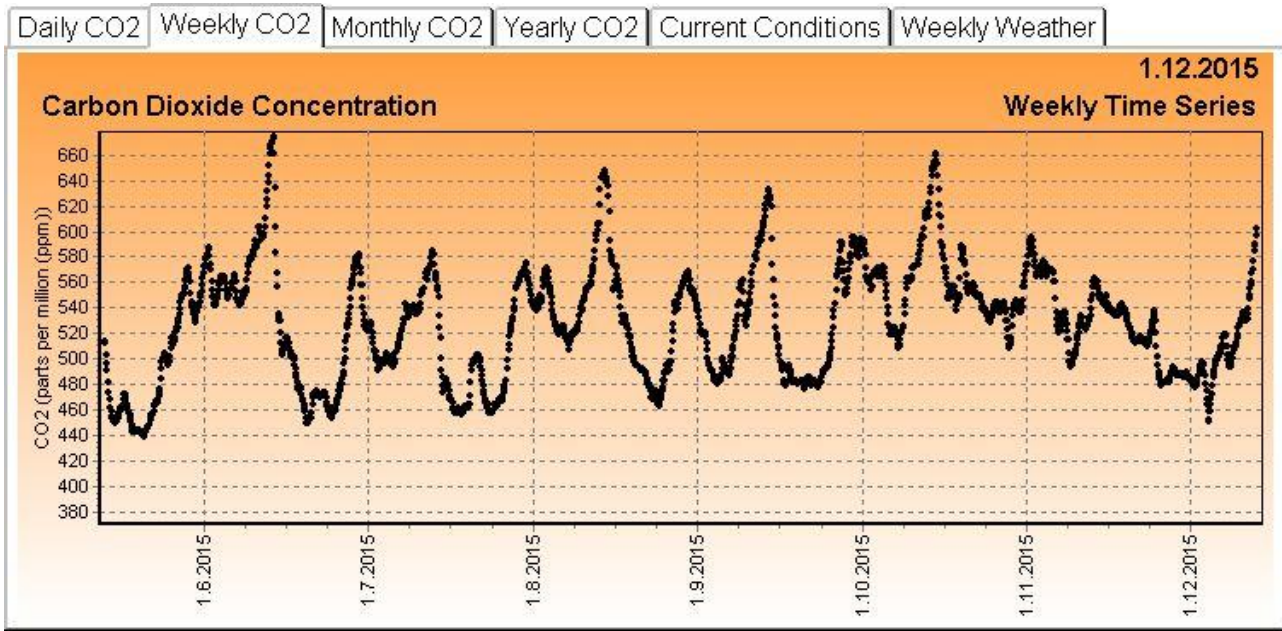




CO_2

Campus

co2.utah.edu



Observations and Truth

- True value- value of a quantity sought through measurement, but unknown usually in the field
- Truth depends on application
- Assumption: average of many unbiased observations should be same as expected value of truth
- However, accurate observations may be biased or *unrepresentative* due to siting or other factors

Rudy Giuliani says 'truth isn't truth'

By Caroline Kenny, CNN

Updated 4:50 PM ET, Sun August 19, 2018

MO

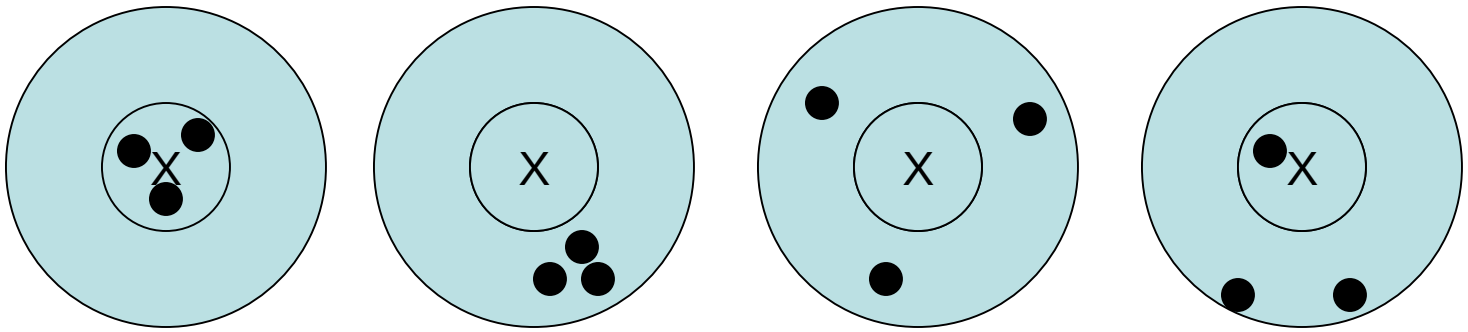
Causes of Uncertainty

- 1. we can never measure the environment with complete accuracy and precision
- 2. the environment is a chaotic system, which is a maddening combination of randomness and order arising from the characteristics of a complex nonlinear system,
- 3. our understanding of the environmental system is imperfect, so physical (and certainly statistical) models do not capture the complete behavior of the system.

Gauging Uncertainty

- Accuracy- difference in response between a standard and instrument in varying environmental conditions a measure of how close a measurement is to the “true” value

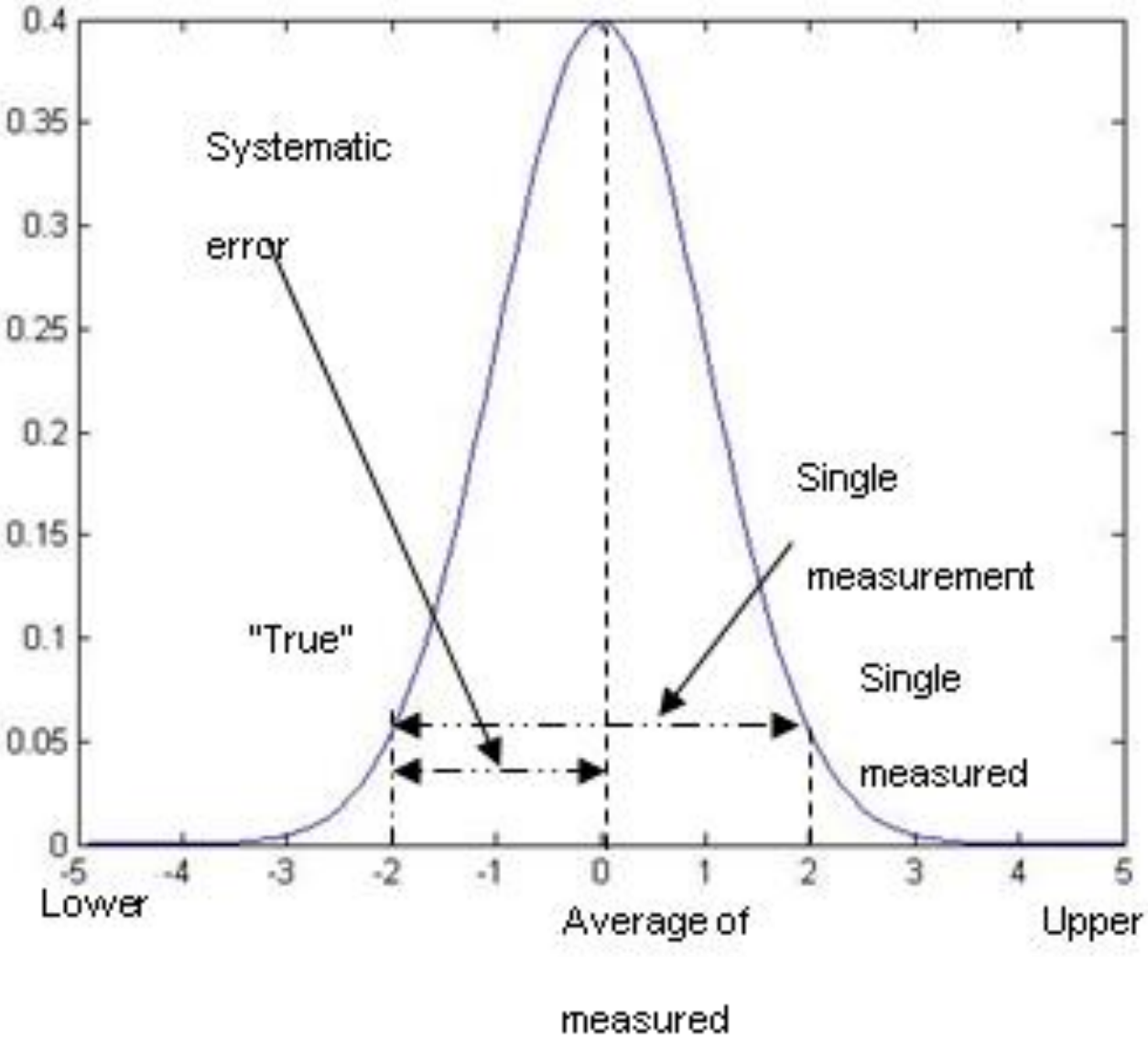
- Precision- how well repeated measurements of some quantity agree with each other. A precise instrument can be inaccurate



Systematic vs. Random Errors

- Random- that which is not precisely predictable or determinable
- Systematic- errors arising from a consistent response of a measuring device to environmental conditions or faulty characteristics of instrumentation that occurs frequently

Truth vs. single and large sample of observations



Population vs. Sample

- we never know the entire population of true values as the environmental conditions change in time or space.
- We hope that we choose a sample of observations for analysis such that each element in the population has an equal chance to be selected.
- Sampling issues
  - Trends
  - serial dependence of environmental data
  - model sample tend to be less variable than observed samples

Selecting a Sample

- Selecting the sample for analysis is a critical aspect of organizing the data and depends on the question to be addressed by the study
- rule of thumb: sample should be large enough to capture the phenomenon of interest many times
- “Degrees of freedom”: number of independent elements in the sample;
  - usually much smaller than the total number of members in the sample in environmental data sets
- Keeping your powder dry- saving data for an independent sample to evaluate and confirm your results.
- Tendency to assume sample is drawn randomly from the population, when sample grossly underestimates the variability inherent in the population

Selfies

"It's not surprising that men who post a lot of selfies and spend more time editing them are more narcissistic, but this is the first time it has actually been confirmed in a study,"

Study Links Selfies To Narcissism And Psychopathy

The Huffington Post | By Carolyn Gregoire

Posted: 01/12/2015 8:15 am EST | Updated: 01/12/2015 8:59 am EST

The sample included 800 men from age 18 to 40 who completed an online survey asking about their photo posting behavior on social media. The participants also completed standard questionnaires for anti-social behaviors and for self-objectification. (This study doesn't include women because the dataset, which Fox received from a magazine, did not have comparable data for women.):

Facebook updates may stave off loneliness, even if no one 'likes' you, study finds

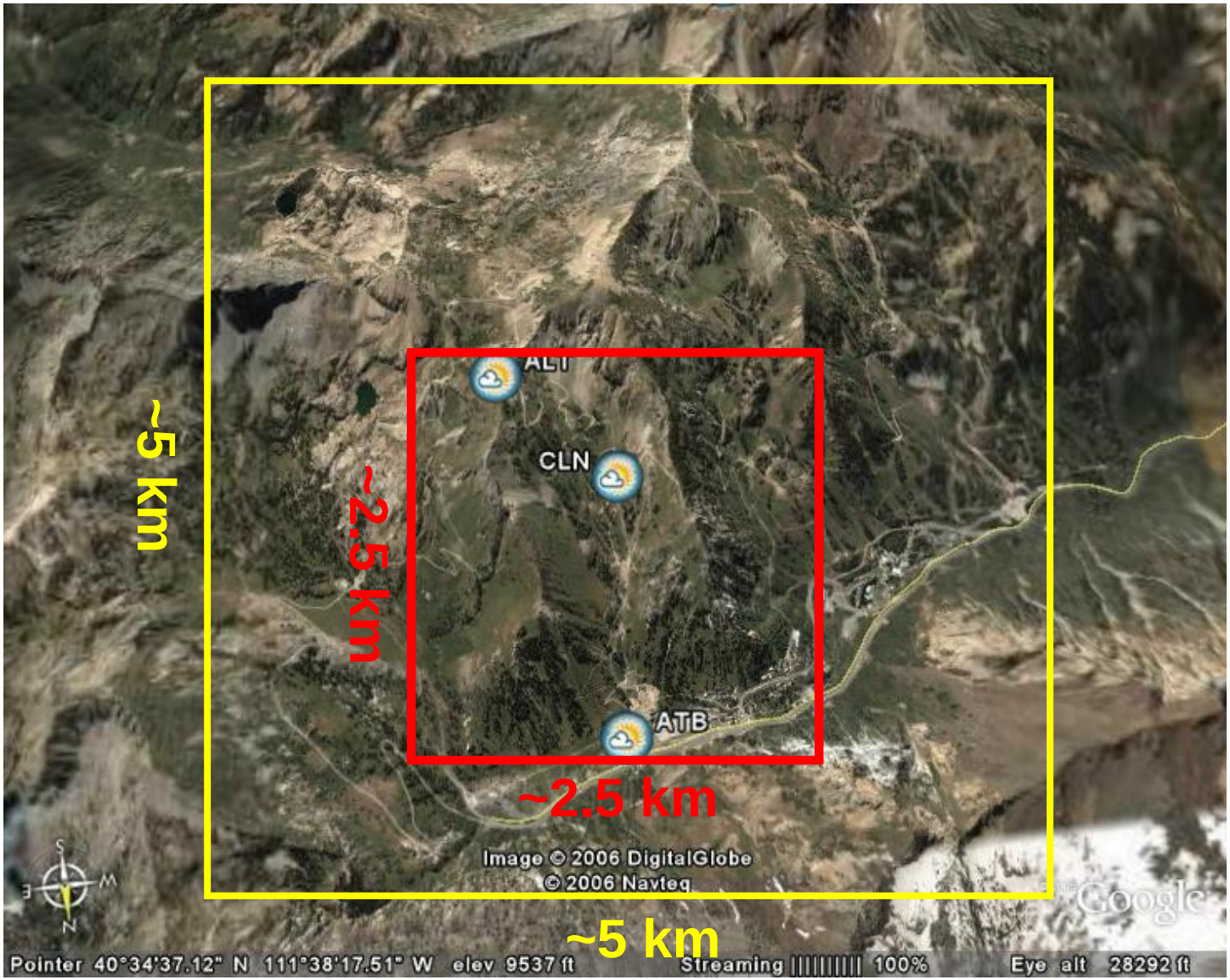
- Compared with other students, those who had been urged to go on a status-writing blitz felt less lonely
- happiness and depression levels unchanged, "suggesting that the effect is specific to experienced loneliness," Simply thinking about their friends can have a "social snacking" effect.
- "Similar to snack temporarily reducing hunger until next meal, social snacking may help tolerate lack of 'real' social interaction for a certain amount of time"

- Scientists have found clues to what compels people to constantly update their Facebook status. College students who posted more status updates than they normally did felt less lonely over course of a week, even if no one "liked" or commented on their posts
- researchers at Free University Berlin recruited 100 undergraduates at University of Arizona; paper published last month in Social Psychological and Personality Science.
- Participants filled out surveys to measure their levels of loneliness, happiness and depression, and gave researchers access to their Facebook
- students were sent an analysis of their average weekly status updates; some were told to post more updates than usual over next seven days. During that week, all completed a short online questionnaire at the end of each day about their mood and level of social connection.

Observations

- Observations are not perfect…
  - Gross errors
  - Local siting errors
  - Instrument errors
  - Representativeness errors

Representative errors to be expected in mountains
Alta Ski Area

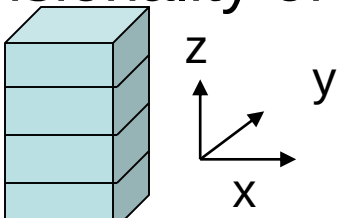


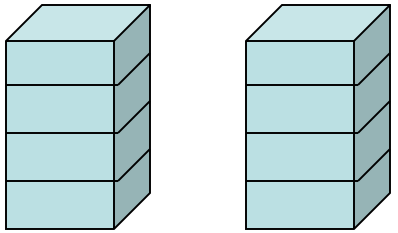
Common Goal is to Synthesize and Reduce Dimensionality

- Statistical analysis of environmental data typically involves reducing the dimensionality of the data to a manageable size.
- Which variable(s) do we need to consider?
- Can we consider one variable (univariate analysis) or must we consider multiple variables (multivariate analysis)?
- What time scales are we interested in? Hours, days, months, years? And, what region (local, regional, globally) or level in the vertical (surface, subsurface, upper air)?
- Are the data available on a spatial grid or at specific points?

Large Dimensionality of Geophysical Data Sets

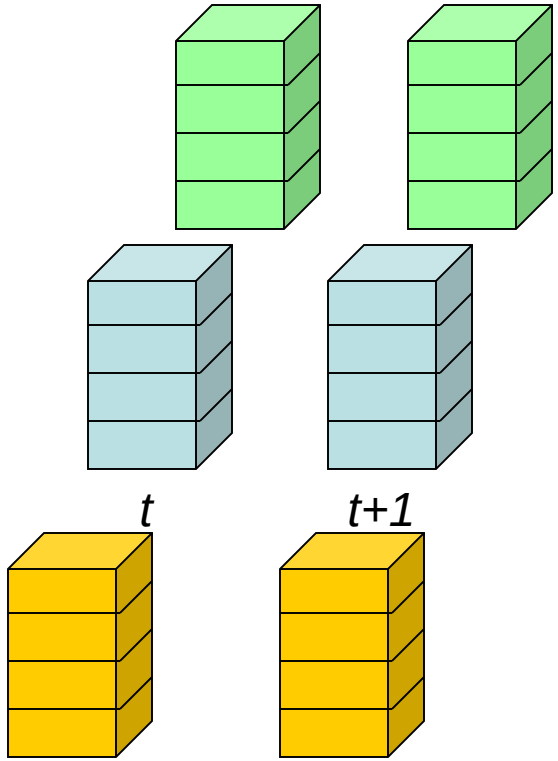
•Space: x, y, z





•Time: time (t) and forecast time (t_f)

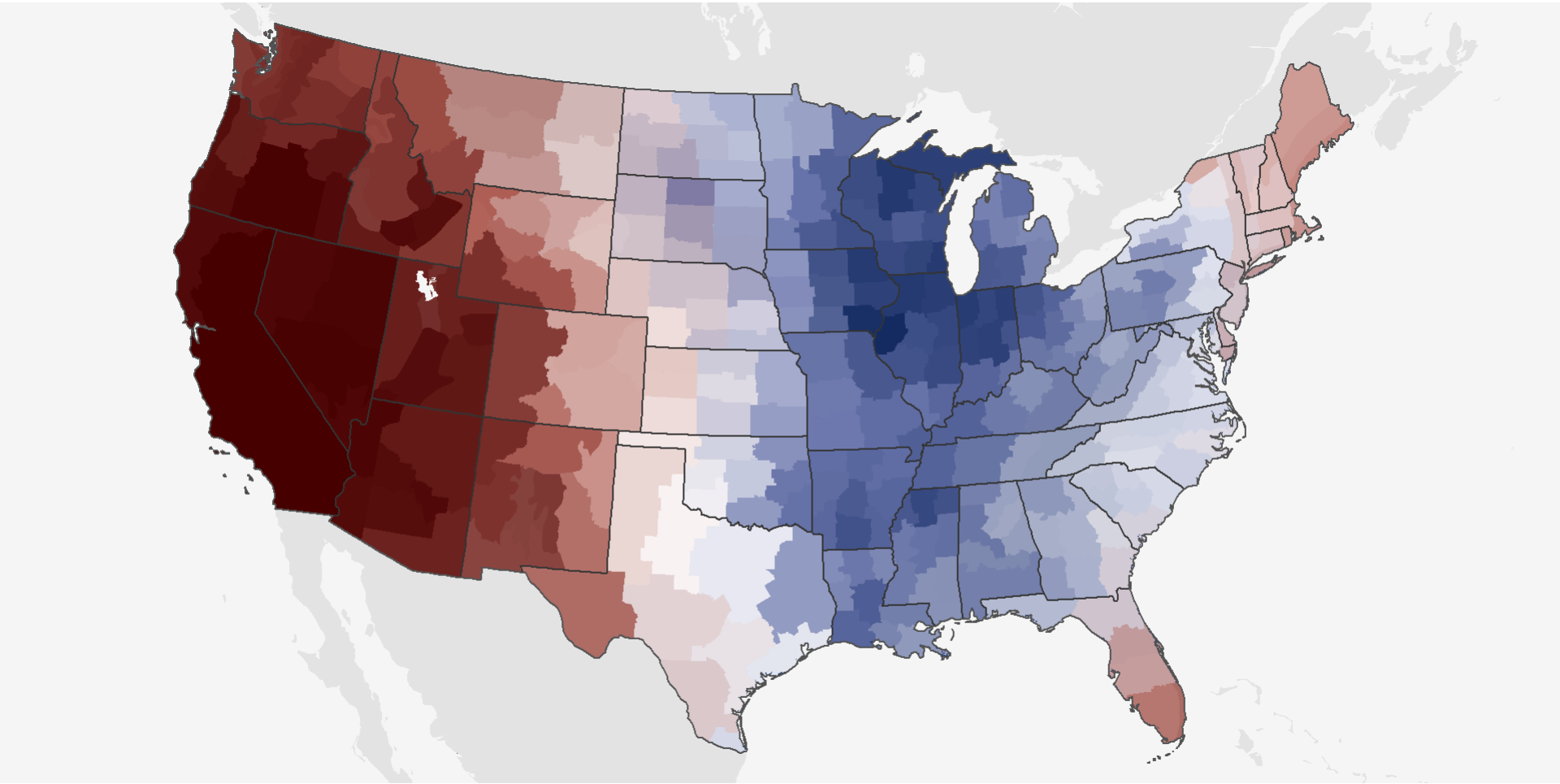
•Parameter & Source (temperature, winds, different models, measuring systems, perturbations)



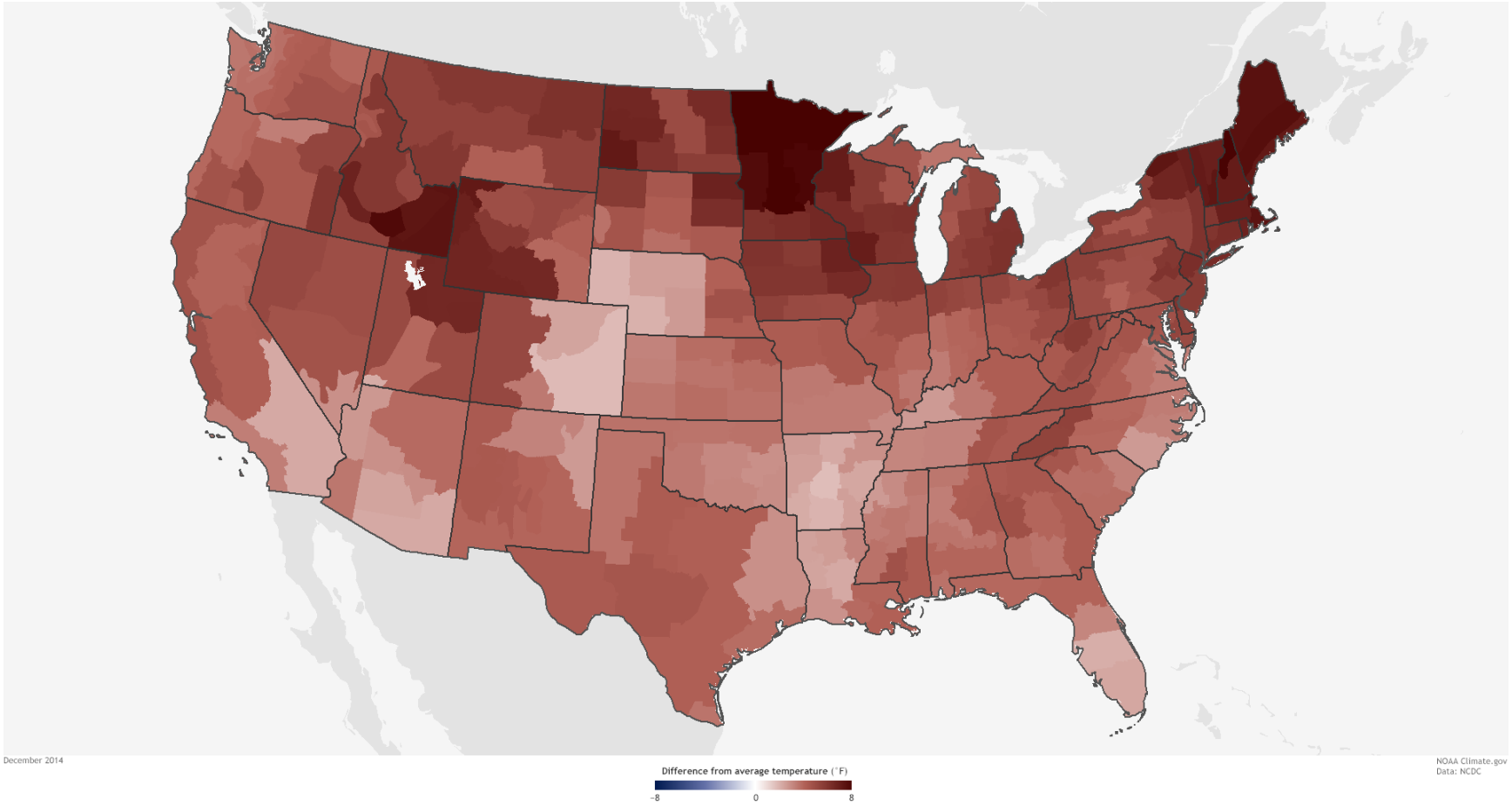
Climate

- Statistics often misinterpreted as bookkeeping.
  - What is the warmest temperature on record at Salt Lake City?
  - What is the biggest snow storm at Alta?
- weather and climate:
  - weather- state of the environment
  - climate- aggregate summary of the environment
  - Climate normal: arbitrarily defined reference state: 1981-2010

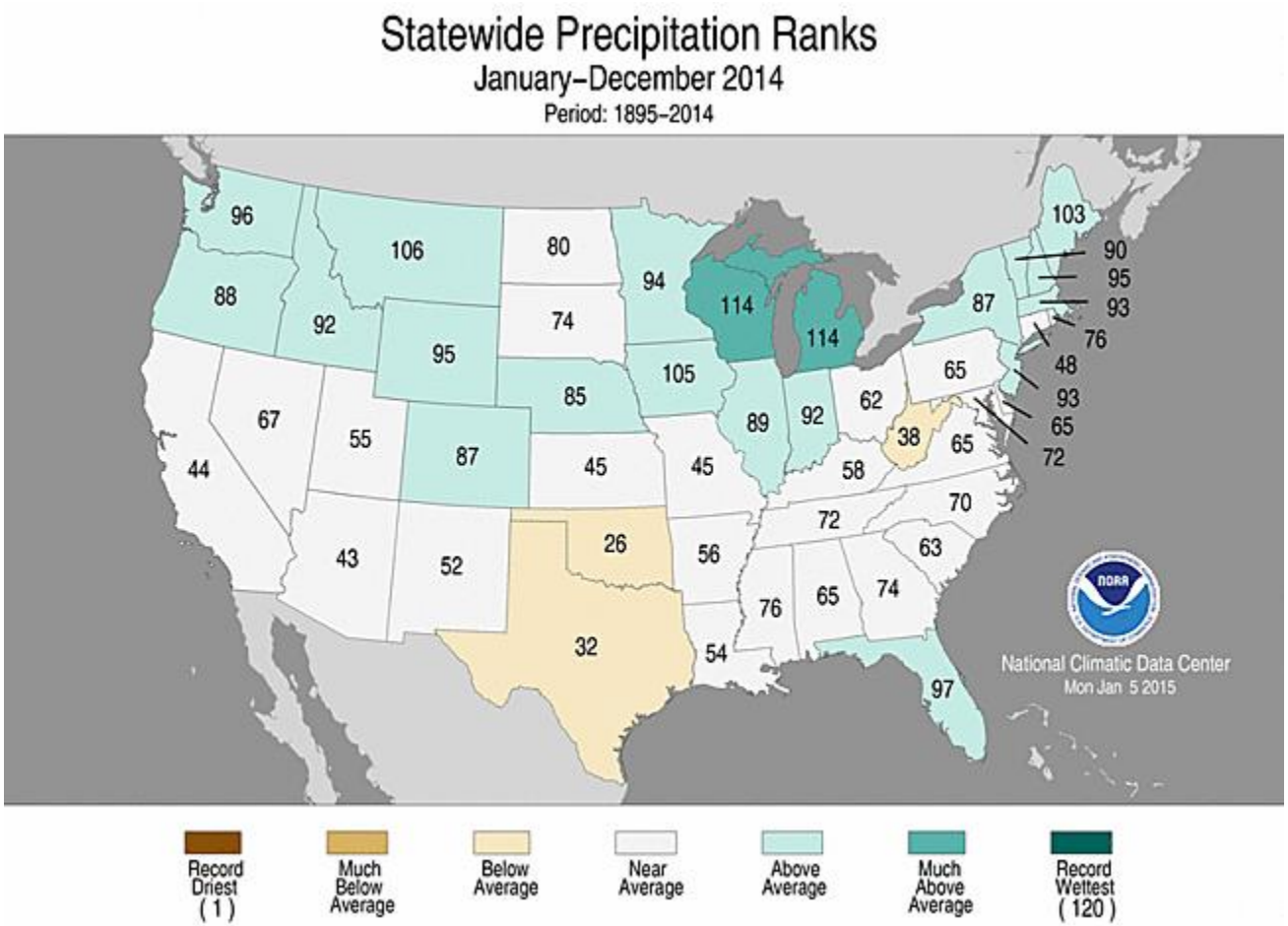
Climate.gov: 2014 temperature relative to climate normal



December 2014 temperature anomaly



Comparing 2014 to 1895-2013

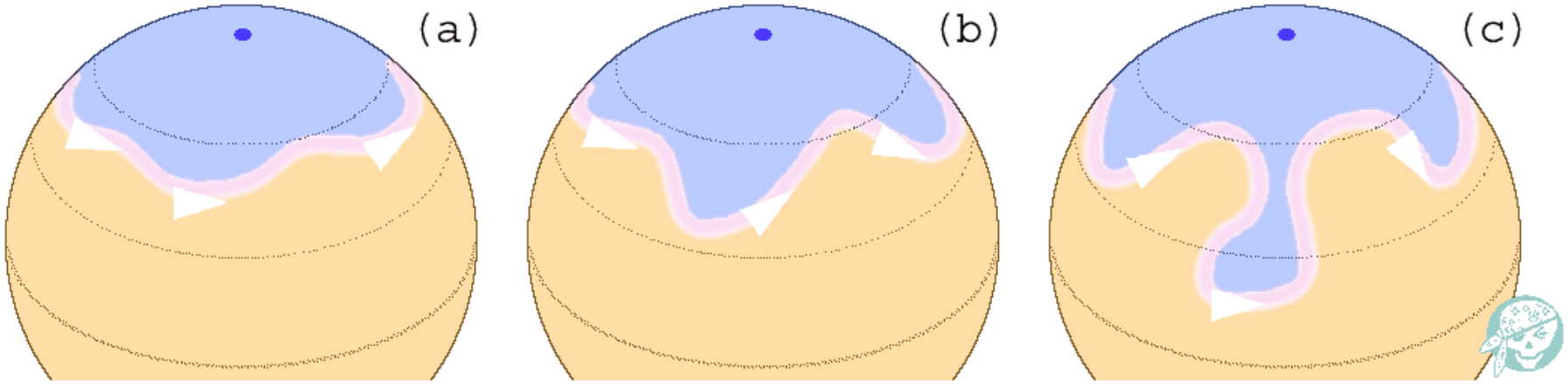


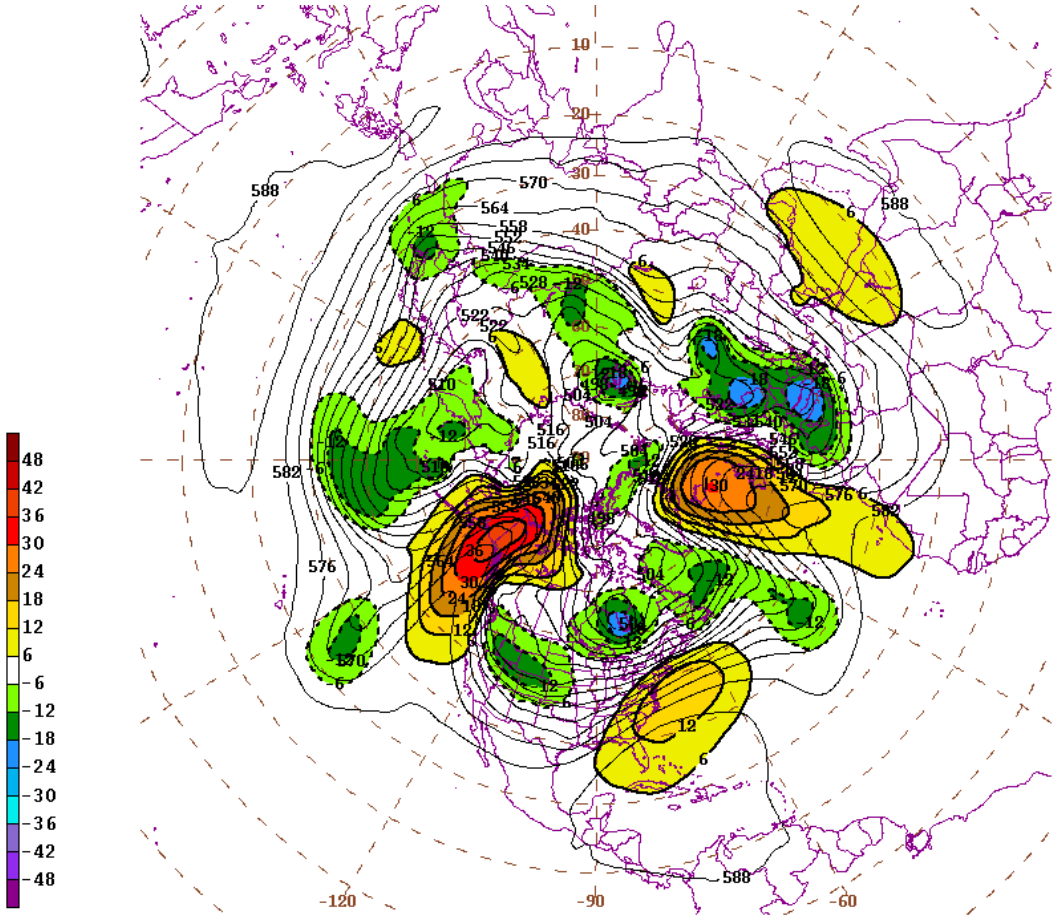
Why use statistics to describe the environment?

- Environment controlled by innumerable factors, which we hope to segregate into a few critical factors from the rest that, for the most part, simply contribute to background noise
- the characteristics of the system include linearly unstable processes such as baroclinic waves that cause growth of small features into larger ones
- the characteristics of the system (dynamics, thermodynamics) are nonlinear and include discrete step functions (i.e., rain/no rain) that can lead to the amplification of small errors into large ones
- the system is dissipative, which guarantees “stationarity”, i.e., the climate system will remain stable and not run away from the current state

Baroclininc Wave Growth

- Physical processes in atmosphere can be unstable at times but environmental system overall is “stationary”





141229/0000F000 500 hPa Height and Height Anomaly (dam), from GFS

Steps for Effective Research

- distill a general interest in a subject into a specific question/hypothesis that can be evaluated
- organize the data
- find relationship(s) among the data
- examine the significance of your results
- review thoroughly what you have done and document your analysis and results
- submit your results and study for independent evaluation

What you should be doing

- Assignment 1- Survey- Due Tonight
- Read Chapter 1 Notes
- Read Chapter 1 in text
- No class Wednesday
  - Spend time reading above
  - Complete Assignment 2 based on Chapter 1 notes and text
  - Due Sunday night midnight