Predictive Learning from Data

LECTURE SET 1

INTRODUCTION and OVERVIEW

OUTLINE of Set 1

1.1 Overview: what is this course about

1.2 Prerequisites and Expected outcomes

1.3 Big Data and Scientific Discovery

1.4 Related Data Modeling Methodologies

1.5 Experimental Procedure (for predictive modeling)

1.6 Discussion: computer vs human intelligence

1.1 Overview

Predictive Learning from Data

- **Prediction** is difficult

Examples: science, simple vs complex systems

- **Learning**, knowledge ~ vaguely defined

Issues: what is knowledge/learning/intelligence?

Scientific (~ certain) vs. empirical knowledge

- **Data**: sense perceptions and digital data

Bottom Line – these are all difficult concepts, so we can only hope to understand their relationship under well-defined assumptions (~ scientific understanding)

Uncertainty and Learning

- Decision making under uncertainty
- Biological learning (adaptation)
- Epistemology: logical inference & classical science vs. plausible (uncertain) inference
- Uncertain knowledge is regarded as inferior in science (until recently) in philosophy and science
- Probability and statistics are fairly new disciplines
- Inductive inference in Statistics and Philosophy

  Ex. 1: Many old men are bald

  Ex. 2: Sun rises on the East every day

(cont'd) Many old men are bald

- *Psychological Induction:*
  - inductive statement based on experience
  - also has certain predictive aspect
  - no scientific explanation

- *Statistical View:*
  - the lack of hair = random variable
  - estimate its distribution (depending on age) from past observations (training sample)

- *Philosophy of Science Approach:*
  - find scientific theory to explain the lack of hair
  - explanation itself is not sufficient
  - true theory needs to make non-trivial predictions

Conceptual Issues

- An explanation (model) has two aspects:

  1. explanation of past/ known observations (data)
  2. prediction of future events (data)

- Achieving (1) is easy, but (2) is hard
- Important issues to be addressed:

  - good quality indices for explanation and prediction

  - if two models explain past data equally well, which one is better?

The main conceptual issues (addressed in this course):

(a) can statistical model that explains past data *also* provide good predictions for future data?

(b) Under what math conditions (a) is possible?

Philosophical connections

Men have lower life expectancy than women

- *Because* they choose to do so
- *Because* they make more money (on average) and experience higher stress managing it
- *Because* they engage in risky activities
- *Because* …..

Demarcation problem in philosophy

Induction

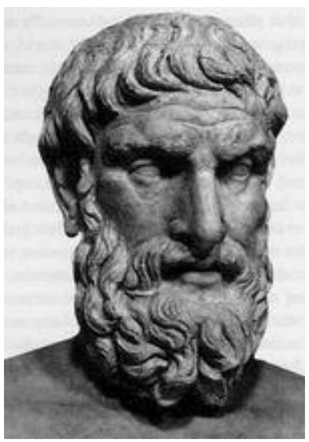
- *From Oxford English dictionary:*

  Induction is the process of inferring a general law or principle from the observations of particular instances.
- Clearly related to Predictive Learning.
- All science and (most of) human knowledge involves (some form of) induction
- How to form 'good' inductive theories?

  → inductive principles ~ general rules

Philosophical Inductive Principles

William of Ockham: entities should not be multiplied beyond necessity



Epicurus of Samos: If more than one theory is consistent with the observations, keep all theories

Note: all philosophical ideas from pre-digital era have useful interpretation in machine learning)

Expected Outcomes + Prerequisites

Scientific / Technical:

- Learning = generalization, concepts and issues
- Math theory: Statistical Learning Theory aka VC-theory
- Conceptual basis for various learning algorithms

Methodological:

- How to use available statistical/machine learning/ data mining s/w
- How to compare prediction accuracy of different learning algorithms
- Are you getting good modeling results because you are smart or just lucky?

Practical Applications:

- Financial engineering
- Biomedical + Life Sciences
- Security
- Image recognition etc., etc.

What is this course NOT about

Grading

HOMEWORK ~ 40% (4 HW assignments)

- Application of existing s/w to real-life and synthetic data sets
- minor programming
- emphasis on understanding underlying algorithms, experimental procedure and interpretation of results

COURSE PROJECT ~ 35%

- Variety of topics, ranging from research to SW development
- Individual Projects
- List of possible topics will be posted on the web page this week
- Student-initiated project topics are *allowed* subject to instructor's approval

MIDTERM EXAM ~ 25%

- Open book / Open notes

CLASS PARTICIPATION (extra credit) – up to 5%

- I will occasionally ask open-ended questions during lectures

1.2 Prerequisites and Hwk1

- Math: working knowledge of basic Probability + Linear Algebra
- Introductory course on ML, i.e. EE4389W or CSci5525 or equiv. or consent of instructor
- Statistical or machine learning software

  - MATLAB, also R-project, Mathematica etc.

  Note: you will be using s/w implementations of learning algorithms(not writing programs)
- Software available on course website:

  - Matlab-based for Windows

Homework 1 (background)

- Purpose: testing background on probability + computer skills + common sense.
- Modeling financial data from Yahoo! Finance
- Real Data: *X*=daily price change of an index

i.e. $X(t) = \frac{Z(t) - Z(t-1)}{Z(t-1)} * 100\%$ where *Z(t)* = closing price

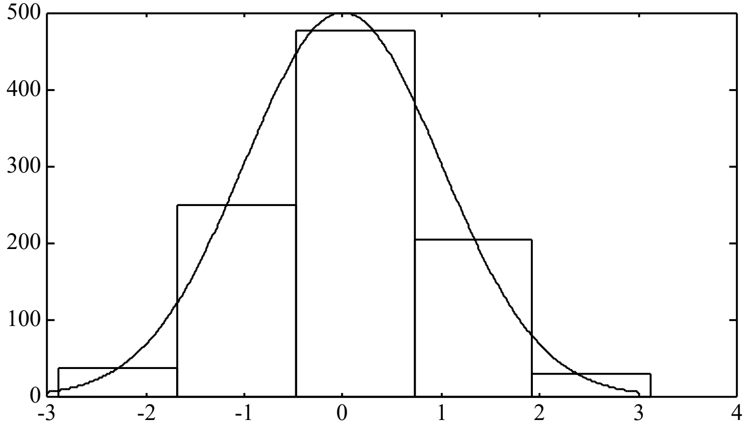
- Is the stock market *truly random?*
- Modeling assumption: price changes *X* are i.i.d.

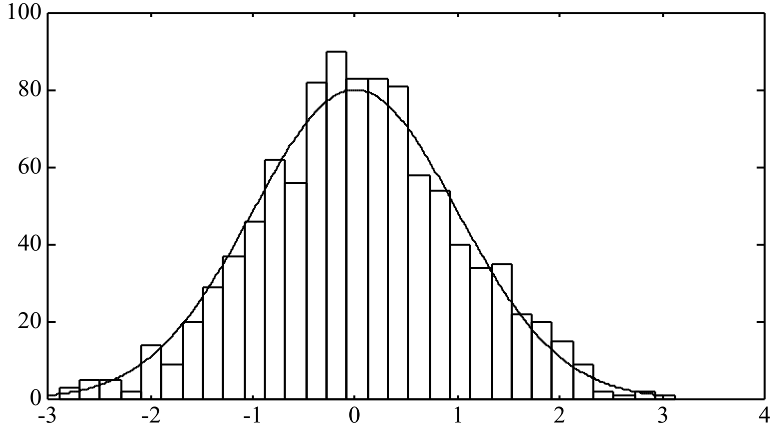
→ leads to certain analytic relationship that can be verified using empirical data.

Understanding Daily Price Changes

Histogram = estimated pdf (from data)

- Example: histograms of 5 and 30 bins to model *N(0,1)*
  also mean and standard deviation (estimated from data)

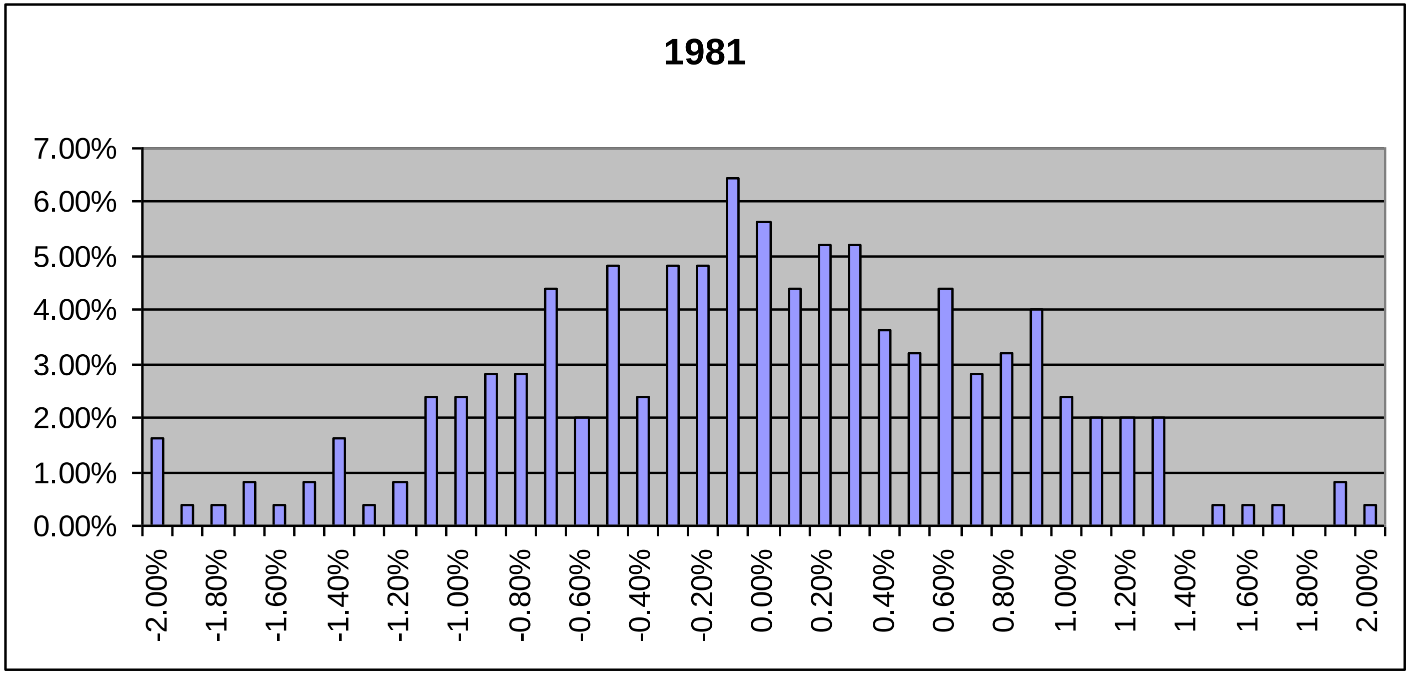




Histogram of daily price changes in 1981

NOTE: histogram ~ empirical pdf, i.e. scale of y-axis scale is in % (frequency).

Histogram of SP500 daily price changes in 1981:



OUTLINE of Set 1

1.1 Overview: what is this course about

1.2 Prerequisites and Expected outcomes

1.3 Big Data and Scientific Discovery

- scientific theories and fairy tales
- promise of Big Data
- characteristics of scientific knowledge
- dealing with uncertainty and risk

1.4 Related Data Modeling Methodologies

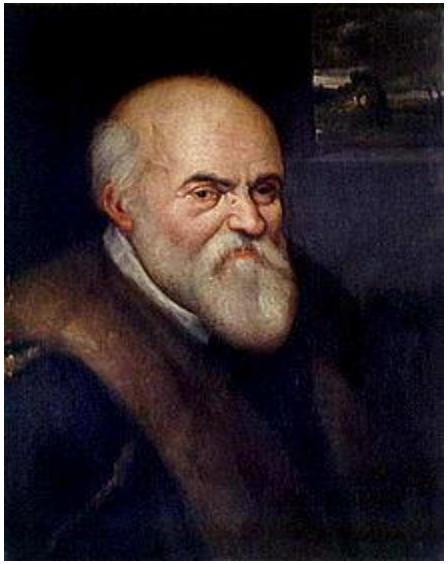
1.5 General Experimental Procedure

1.6 Discussion: Computer vs. Human intelligence

Historical Example

Ulisse Aldrovandi,16th century wrote

Natural History of Snakes



Promise of Big Data

- **Technical fairy tales in 21st century**

  ~ marketing + more marketing

- **Promise of Big Data:**

  s/w program + DATA → knowledge

  ~ More Data → more knowledge

- **Yes-we-Can !**

Examples from Life Sciences…

- Duke biologists discovered an unusual link btwn the popular singer and a new species of fern, i.e.

  - bisexual reproductive stage of the ferns;

  - the team found the sequence GAGA when analyzing the fern’s DNA base pairs

Scientific Discovery

- **Combines ideas/models and facts/data**
- **First-principle knowledge:**

  hypothesis → experiment → theory

  ~ deterministic, causal, intelligible models
- **Modern data-driven discovery:**

  s/w program + DATA → knowledge

  ~ statistical, complex systems
- **Many methodological differences**

Invariants of Scientific Knowledge

- **Intelligent questions**
- **Non-trivial predictions**
- **Clear limitations/ constraints**

- **All require human intelligence**

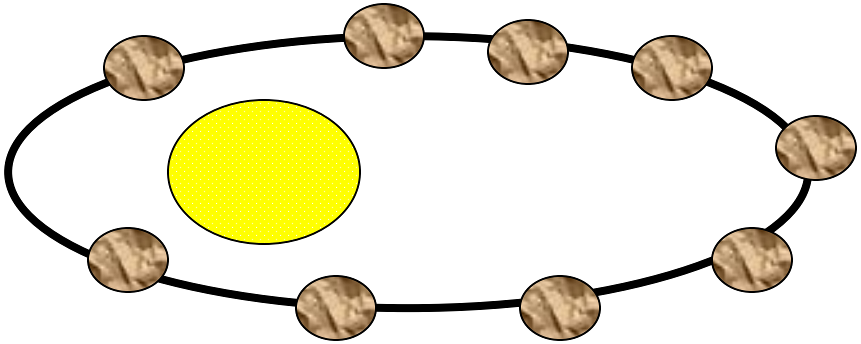
- missing/ lost in Big Data?

Historical Example: Planetary Motions

- How planets move among the stars?
  - Ptolemaic system (geocentric)
  - Copernican system (heliocentric)
- Tycho Brahe (16 century)
  - measure positions of the planets in the sky
  - use experimental data to support one's view
- Johannes Kepler:
  - used volumes of Tycho's data to discover three remarkably simple laws

First Kepler's Law

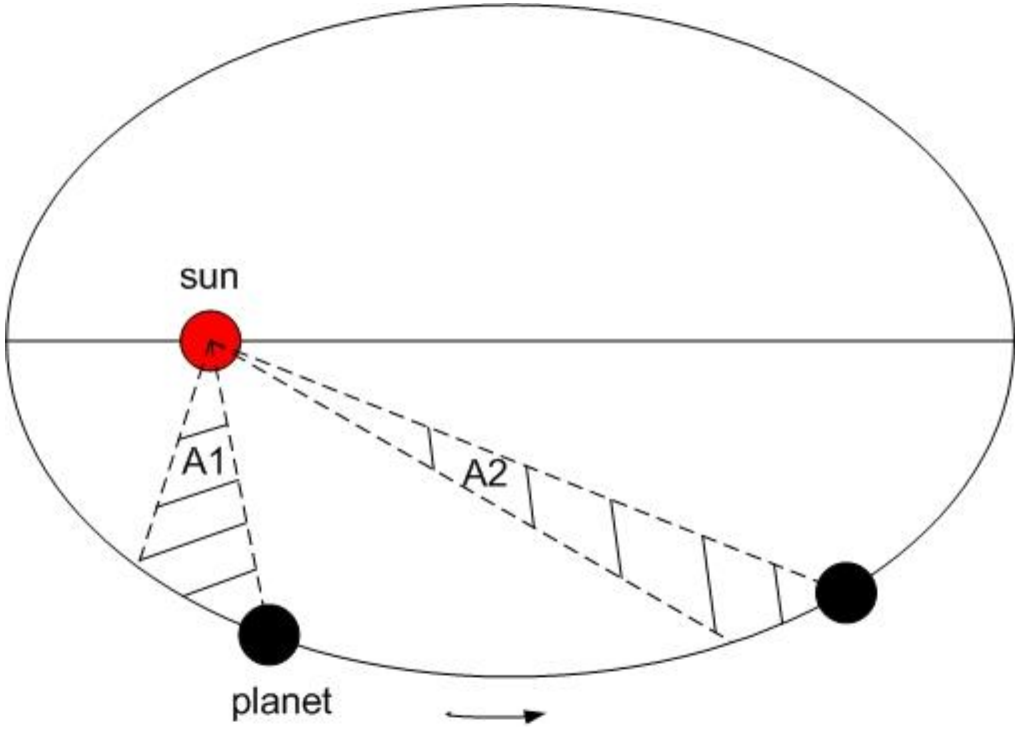
- Sun lies in the plane of orbit, so we can represent positions as (x,y) pairs
- An orbit is an ellipse, with the sun at a focus



$$c_1 x^2 + c_2 y^2 + c_3 xy + c_4 x + c_5 y + c_6 = 0$$

Second Kepler's Law

- The radius vector from the sun to the planet sweeps out equal areas in the same time intervals



Third Kepler's Law

	P	D	P^2	D^3
Mercury	0.24	0.39	0.058	0.059
Venus	0.62	0.72	0.38	0.39
Earth	1.00	1.00	1.00	1.00
Mars	1.88	1.53	3.53	3.58
Jupiter	11.90	5.31	142.0	141.00
Saturn	29.30	9.55	870.0	871.00

P = orbit period   D = orbit size (half-diameter)

For any two planets: $P^2 \sim D^3$

Empirical Scientific Theory

- Kepler's Laws can
  - explain experimental data
  - predict new data (i.e., other planets)
  - *BUT* do not explain *why planets move*.
- Popular explanation
  - planets move because there are invisible angels beating the wings behind them
- **First-principle scientific explanation**

  Galileo and Newton discovered laws of motion and gravity that explain Kepler's laws.

OUTLINE of Set 1

1.1 Overview: what is this course about

1.2 Prerequisites and Expected outcomes

1.3 Big Data and Scientific Discovery

1.4 **Related Data Modeling Methodologies**

- growth of empirical knowledge
- empirical vs first-principle knowledge
- handling uncertainty and risk
- related data modeling methodologies

1.5 General Experimental Procedure.

Scientific knowledge

- **Knowledge**

  ~ stable relationships between facts and ideas (mental constructs)

- **Classical first-principle knowledge**:

  - rich in ideas
  - relatively few facts (amount of data)
  - simple relationships

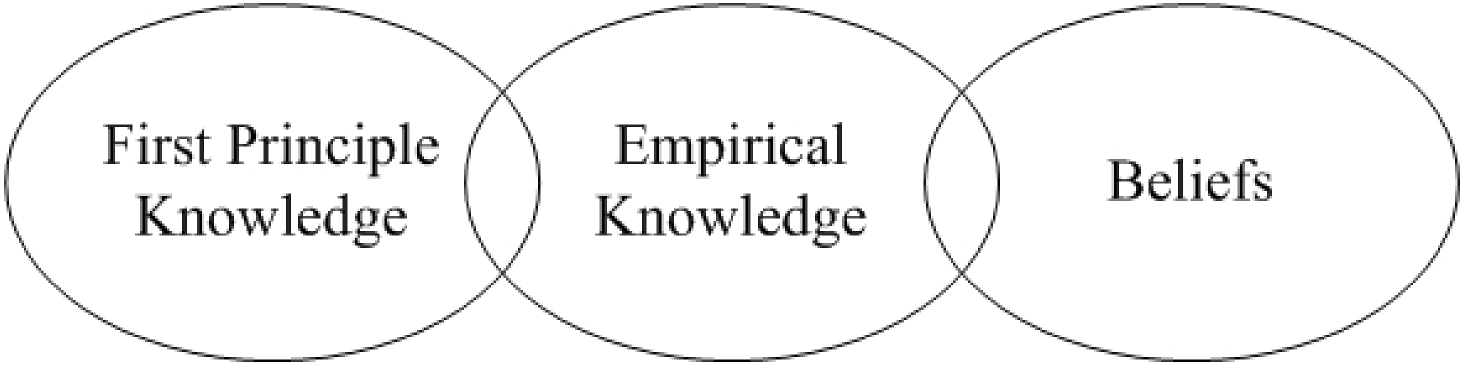
Growth of empirical knowledge

- Huge growth of the amount of data in 20th century (computers and sensors)
- Complex systems (engineering, life sciences and social)
- Classical first-principles science is inadequate for empirical knowledge
- Need for new Methodology:

  How to estimate good predictive models from noisy data?

Different types of knowledge

- Three types of knowledge
  - scientific (first-principles, deterministic)
  - empirical (uncertain, statistical)
  - metaphysical (beliefs)



- Boundaries are poorly understood

Handling Uncertainty and Risk(1)

- Ancient times
- Probability for quantifying uncertainty
  - degree-of-belief
  - frequentist (Cardano-1525, Pascale, Fermat)
- Newton and causal determinism
- Probability theory and statistics (20th century)
- **Modern classical science** (A. Einstein)

→Goal of science: estimating a **true model** or **system identification**

Handling Uncertainty and Risk(2)

- Making decisions under uncertainty
  ~ *risk management, adaptation, intelligence…*
- **Probabilistic approach**:
  - estimate probabilities (of future events)
  - assign costs and minimize expected risk
- **Risk minimization** approach:
  - apply decisions to known past events
  - select one minimizing expected risk
- **Biological learning + complex systems**

Summary

- First-principles knowledge:

  deterministic relationships between a few concepts (variables)

- *Importance of empirical knowledge*:

  - statistical in nature

  - (usually) many input variables

- Goal of modeling: to act/perform well, rather than system identification

Other Related Methodologies

- **Estimation of empirical dependencies** is commonly addressed many fields

  \- *statistics, data mining, machine learning, neural networks, signal processing* etc.

  \- each field has its own methodological bias and terminology → confusion

- Quotations from popular textbooks:

  The field of *Pattern Recognition* is concerned with the automatic discovery of regularities in data.

  Data Mining is the process of automatically discovering useful information in large data repositories.

  Statistical Learning is about learning from data.

- All these fields are concerned with estimating predictive models from data.

Other Methodologies (cont'd)

- **Generic Problem**

  Estimate (learn) useful models from available data

- **Methodologies differ** in terms of:

  - what is useful

  - (assumptions about) available data

  - goals of learning

- Often these important notions are not well-defined.

Common Goals of Modeling

- **Prediction (Generalization)**
- **Interpretation ~ descriptive model**
- **Human decision-making** using both above
- **Information retrieval,** i.e. predictive or descriptive modeling of unspecified subset of available data

Note:

- These goals usually ill-defined
- Formalization of these goals in the context of application requirements is THE MOST IMPORTANT aspect of ‘data mining’

Three Distinct Methodologies (section 1.5)

- **Statistical Estimation**

  - from classical statistics and fct approximation
- **Predictive Learning (~ machine learning)**

  - practitioners in machine learning /neural networks

  - Vapnik-Chervonenkis (VC) theory for estimating predictive models from empirical(finite) data samples
- **Data Mining**

  - exploratory data analysis, i.e. selecting a subset of available (large) data set with interesting properties

OUTLINE of Set 1

1.1 Overview: what is this course about

1.2 Prerequisites and Expected outcomes

1.3 Big Data and Scientific Discovery

1.4 Related Data Modeling Methodologies

1.5 General Experimental Procedure for Estimating Models from Data

1.6 Discussion: computer vs human intelligence

1.5 General Experimental Procedure

1. **Statement of the Problem**
2. **Hypothesis Formulation** (Problem Formalization) – *different from classical statistics*
3. **Data Generation/ Experiment Design**
4. **Data Collection and Preprocessing**
5. **Model Estimation** (learning)
6. **Model Interpretation, Model Assessment and Drawing Conclusions**

Note:

- each step is complex and requires several iterations
- estimated model depends on all previous steps
- **observational data** *(not experimental_design)*

Data Preprocessing and Scaling

- Preprocessing is required with observational data (***step 4*** in general experimental procedure)

Examples: …

- Basic preprocessing includes

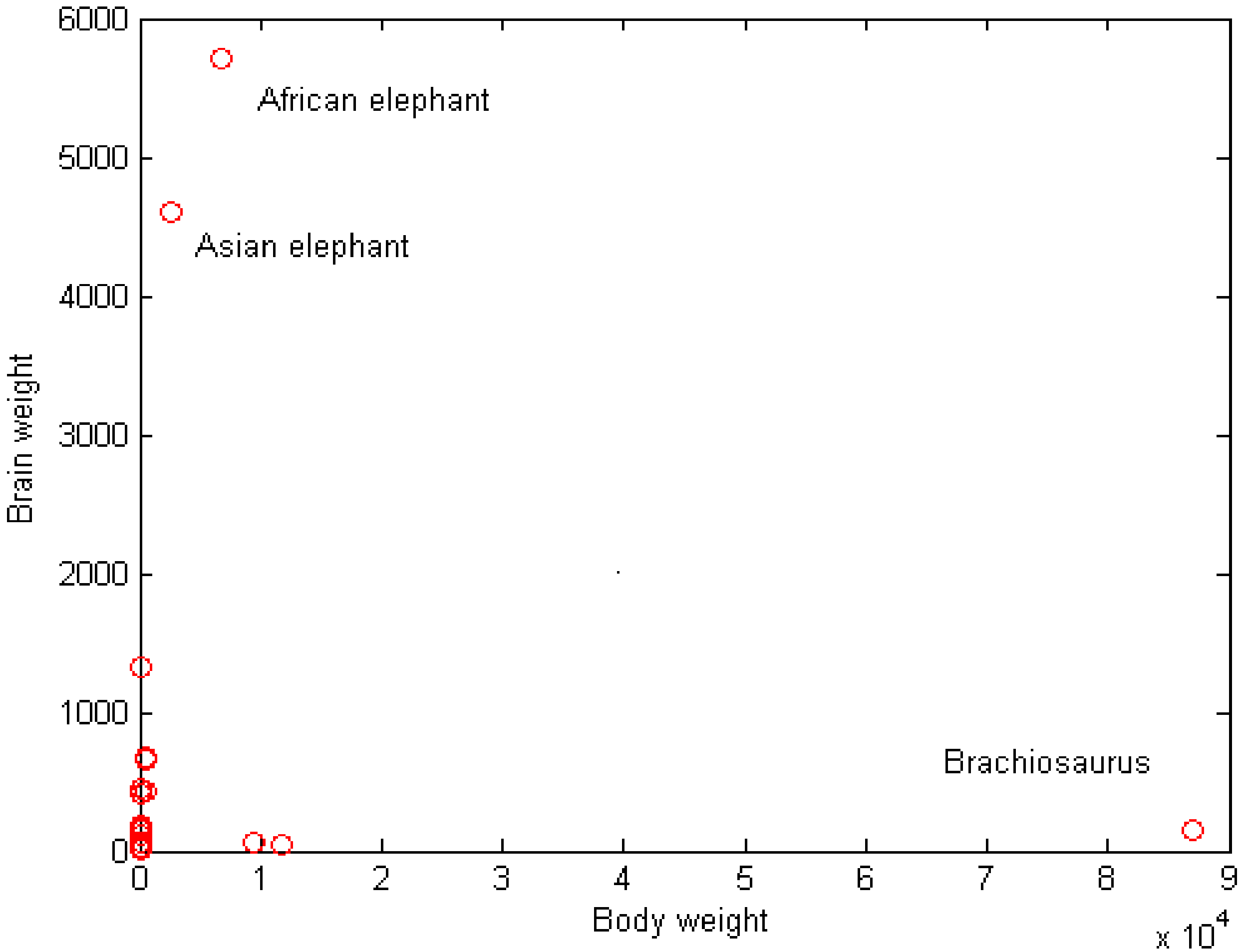
  - summary univariate statistics: *mean, st. deviation, min + max value, range, boxplot* performed independently for each input/output

  - *detection (removal) of outliers*

  - *scaling* of input/output variables (may be *necessary* for some learning algorithms)

- Visual inspection of data is tedious but useful

Original Unscaled Animal Data



Cultural + Ethical Aspects

- **Cultural and business aspects** usually affect:
  - problem formalization
  - data access/ sharing (i.e., in life sciences)
  - model interpretation
- *Examples:* …

- **Possible solution approach**
  - to adopt common methodology
  - critical for interdisciplinary projects

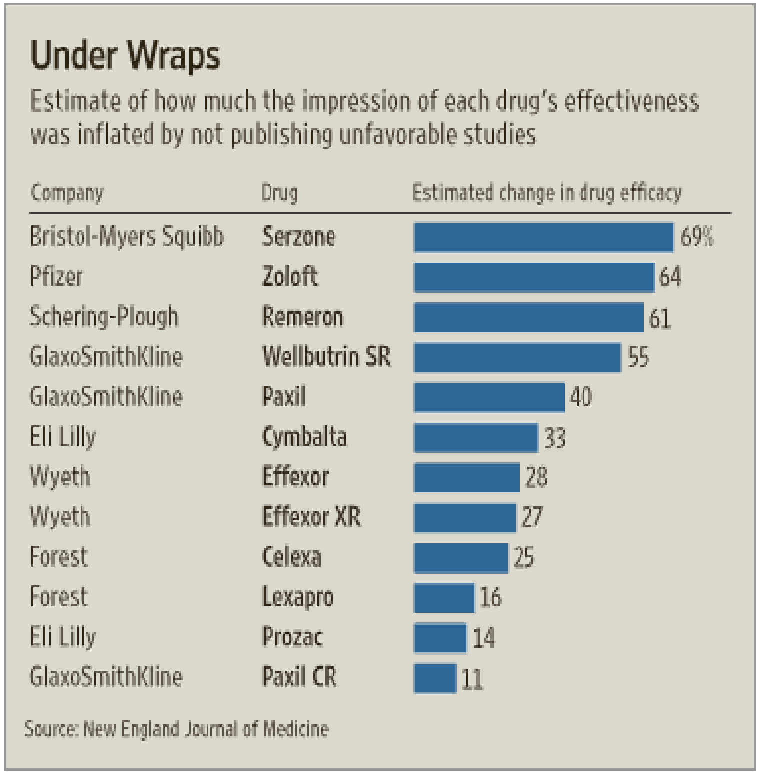
Honest Disclosure of Results

- Recall Tycho Brahe + Kepler (16th century)
- **Modern drug studies**

Review of studies submitted to FDA

- Of 74 studies reviewed, 38 were judged to be positive by the FDA. All but one were published.
- Most of the studies found to have negative or questionable results *were not published*.

Source: The New England Journal of Medicine, WSJ Jan 17, 2008)



Publication bias: common in modern research

Topic for Discussion

Read the paper by Ioannidis (2005) about the danger of *self-serving data analysis*. Explain how the general experimental procedure can help to safeguard against such biased data modeling. Then give a specific example of a recent misleading research finding based on incorrect interpretation of data. Try to come up with an example from your own application domain (i.e., the technical field you are interested in/ or working in).

Ioannidis (2005) paper is available on-line at

http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

OUTLINE of Set 1

1.1 Overview: what is this course about

1.2 Prerequisites and Expected outcomes

1.3 Big Data and Scientific Discovery

1.4 Related Data Modeling Methodologies

1.5 Experimental Procedure (for predictive modeling)

1.6 Discussion: computer vs human intelligence

Human vs Computer Intelligence

- **Human/ biological brain** good for:
  - visual recognition/ pattern recognition
  - natural language understanding
  - complex environment (~ driving a car)
- **Digital computers** good for:
  - storing/ manipulating large amounts of data
  - mathematically well-defined problems

→ machine intelligence ~ solving ill-posed problems

Caveat: computer intelligence usually means *imitating* human intelligence, not *understanding* it

Some Examples

- **Computers can be programmed** to do well:

  - play chess

  - fly commercial airplane

- **Example** of natural language understanding:

  Original: **Cheese eating surrender monkeys**

  Computer Translation to French: **Primates capitulars et toujours en quete de fromages**

  Back to English: **Primates who capitulate and who are constantly in search of cheese**

- **Common misunderstanding:** Turing test

Growth of Technology vs Knowledge

- **Two problems in history of human knowledge**

  (1) How to store+disseminate human knowledge

  (2) How to create new knowledge
- **Digital Technology** solves Problem (1) very well.

  But can technology help to generate knowledge?
- **Where does knowledge come from?**

  Central problem in *Philosophy of Science:*

  Idealism ~ knowledge comes from human mind

vs. Materialism ~ from observations of Nature

Modern variant of naïve materialism ~ Big Data

Many Discussions in Mass Media

- Some representative articles on AI+ML:

  by Y. Harari:
  https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/

  by H. Kissinger:
  https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/

- These articles usually discuss:
- doomsday future with AI dominating humans
- written by authors lacking any technical knowledge

→ Many strange assumptions, such as:

"AI establishes its own goals"