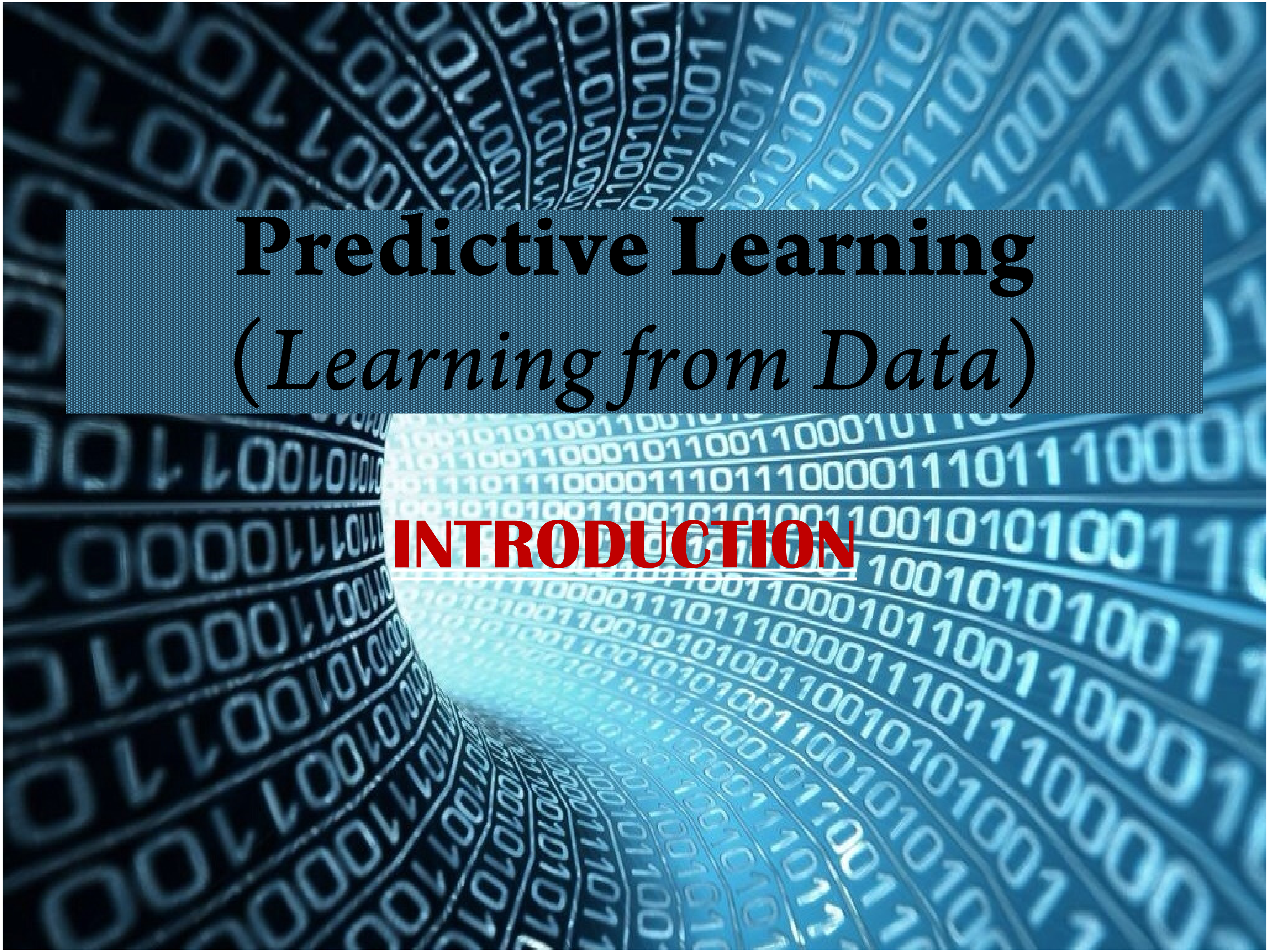
Machine Learning

CS6316



Predictive Learning

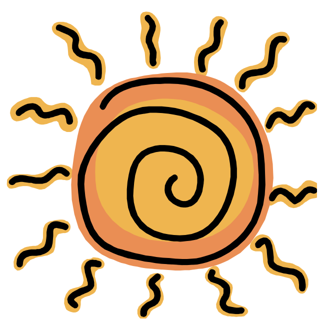
(*Learning from Data*)

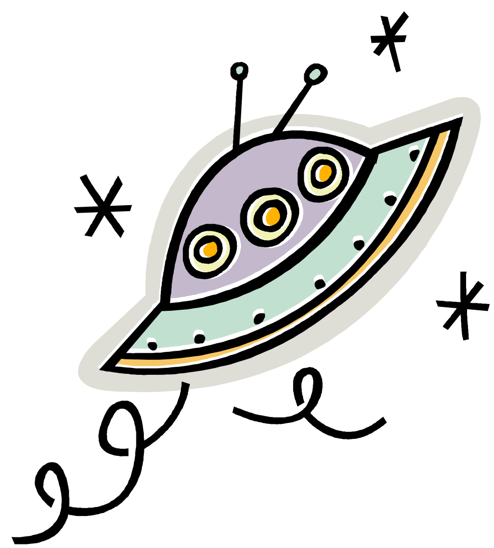
INTRODUCTION

Predictive Learning from Data

Uncertainty and Learning

- Decision making under uncertainty
- Biological learning (adaptation)
  - Hot stove
  - Cats vs Dogs
- Induction in Statistics and Philosophy
  - Ex. 1: Many elderly males are bald
  - Ex. 2: Sun rises on the East every day





Statement: "Many elderly 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

Explanation and Prediction

- Every theory (or model) has two aspects:
  1. EXPLANATION – of past data (*observations*)
  2. PREDICTION – of future (*unobserved/unknown*) data
- Achieving both goals perfectly is *not possible*
- Important issues to be addressed:
  - Quality of explanation and prediction
  - Is good prediction possible at all?
  - If two methods explain past data equally well, which one is better?
  - How to distinguish between true scientific and pseudo-scientific theories?

Beliefs vs True Theories

- "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
- The demarcation problem in the philosophy of science is *about how to distinguish between science and nonscience*, including between science, pseudoscience, and other products of human activity, like art and literature, and *beliefs*.

Philosophical Connections

- Oxford English dictionary:

  <u>INDUCTION</u> is the process of inferring a general law or principle from the observations of particular instances

- Clearly related to <u>PREDICTIVE LEARNING</u>
- All science and (most of) human knowledge involves induction
- How to form 'good' inductive theories?

Challenge of Predictive Learning

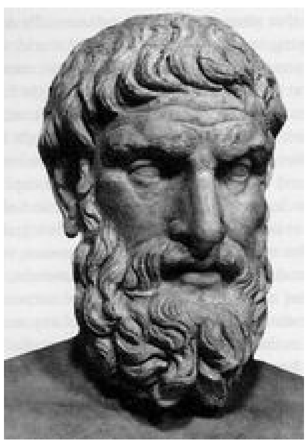
- Explain the past *and* predict the future

Does everybody understand this concept?

Explain the past *and* predict the future

Background: philosophy

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

Background: philosophy

Thomas Bayes: How to update / revise beliefs in light of new evidence

Karl Popper: Every true (inductive) theory prohibits certain events or occurrences, i.e. it should be **falsifiable**

Historical Perspective

Historical Perspective – Handling Uncertainty and Risk

- Since 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

Historical Perspective – Handling Uncertainty and Risk

- Making decisions under uncertainty involves
  - Risk management, and
  - Adaptation
- 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
- Common in all living things: learning, generalization

Human Generalization

- "All men by nature desire knowledge" – Aristotle
- Example 1: continue the given sequence

  6, 10, 14, 18, …
- Example 2:

Sceitnitss osbevred: it is nt inptrant how lteters are msspled isnide the word. It is ipmoratnt that the fisrt and lsat letetrs do not chngae, tehn the txet is itneprted corrcetly

Scientific Example: Planetary Motions

Historical Example: Planetary Motions

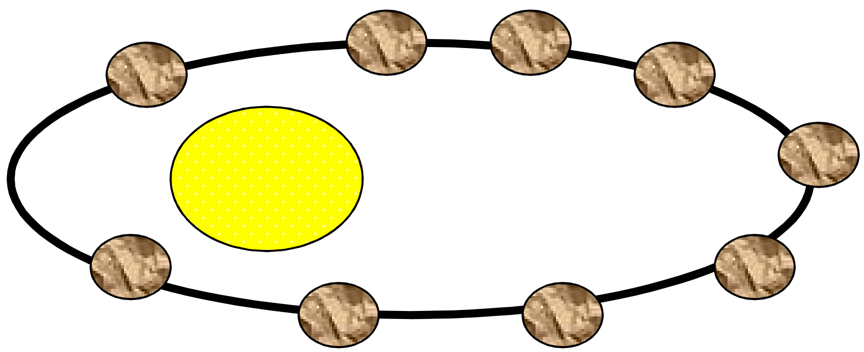
- How planets move among the stars?
  - Ptolemaic system (geocentric) – *earth-centered universe*
  - Copernican system (heliocentric) – *sun-centered solar system*

- 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

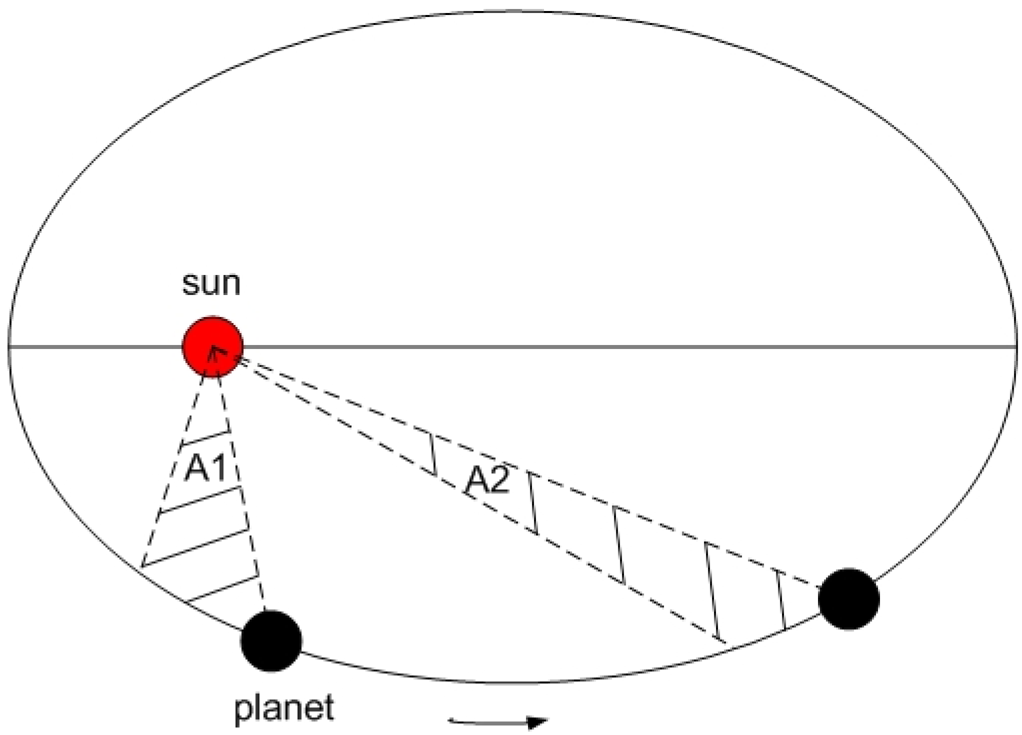
- Sun lies in the plane of orbit, so we can represent positions as (x,y) pairs
- An orbit is an ellipse, with the sun at a focus



$$c_1x^2 + c_2y^2 + c_3xy + c_4x + c_5y + c_6 = 0$$

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: $P^2 \approx D^3$

Empirical Scientific Theory

- Kepler's Laws can
  - Explain experimental data
  - Predict new data (i.e. other planets)
  - BUT does **not** explain why planets move
- Popular explanation (belief)
  - Planets move because there are *invisible angels* beating their wings behind them (!!!!)

- First Principle scientific explanation
  - Galileo and Newton discovered laws of motion and gravity that **explain** Kepler's laws

Motivation for Empirical Knowledge

Motivation for Empirical Knowledge

- Human (scientific) knowledge
- Growth of empirical knowledge
- The nature of human knowledge

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

First Principles

- A first principle is a basic, foundational, self-evident proposition or assumption that cannot be deduced from any other proposition or assumption. It represents the fundamental concepts or assumptions on which a theory, system, or method is based.

- Modern science and engineering are based on using first-principle models to describe physical, biological, and social systems. → Starts with a basic scientific model (e.g. Newton's laws of mechanics) and builds upon it.

First Principles

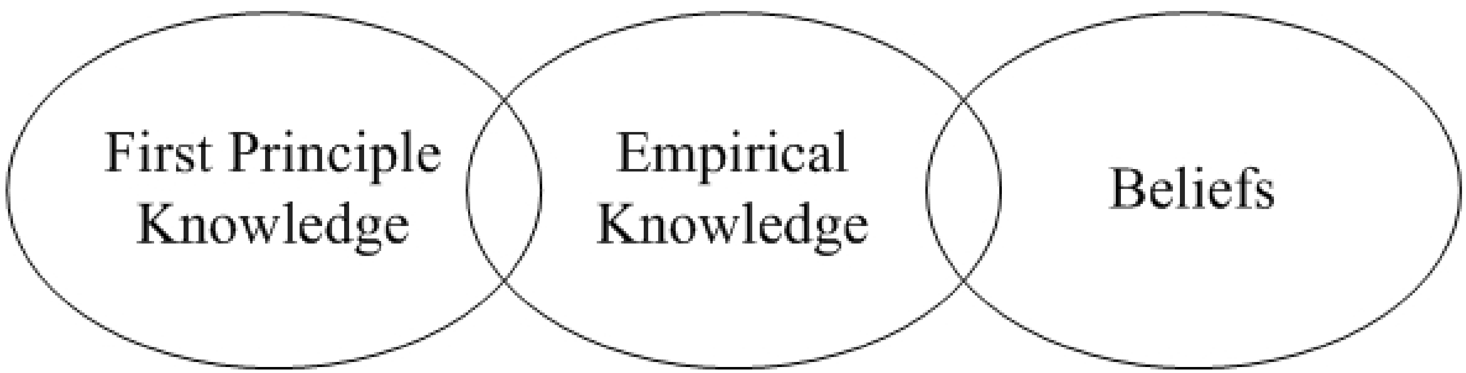
- However, in many applications the underlying first principles are unknown or the systems under study are too complex to be mathematically described.
- With the growing use of computers and low-cost sensors for data collection, there is a great amount of data being generated by such systems. ***In the absence of first-principle modes, such readily available data can be used to derive models by estimating useful relationships between system's inputs and outputs***
- → paradigm shift from the classical modeling based on first principles to **developing empirical data-driven models**.

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:
  - Data-Analytic Modeling:
    How to estimate good predictive models from noisy data?

Nature of Human Knowledge

- Three types of Knowledge:
  - Scientific (first-principles, deterministic)
  - Empirical
  - Metaphysical (beliefs)



  - Boundaries are poorly understood

Empirical Knowledge & Beliefs

- Empirical knowledge: a belief that is learned by observing it using our *empirical knowledge*; e.g. sight, hearing, touch etc.
- Empirical: Empirical or ***a posteriori*** knowledge is possible only subsequent, or posterior, to certain *sense* **experiences** (in addition to the use of reason) Often thought of as data driven

Empirical Knowledge & Beliefs

- Beliefs: Non-empirical or ***a priori*** knowledge is possible independently of, or prior to, any experience, and requires **only the use of reason**; examples include knowledge of logical truths such as the law of non-contradiction, as well as knowledge of abstract claims (such as ethical claims or claims about various conceptual matters)

Summary

- First-principles knowledge (taught at school):
  - 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 in many fields/areas
  - Statistics
  - Data mining
  - Machine learning
  - Neural networks
  - Signal processing
  - … etc.
  - Each field has its own methodological bias and terminology → confusion

Other Related Methodologies

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***

Common Goals of Modeling

Data-driven models have two main components:

1. Explanation / Interpretation (of past/known data) (*Descriptive*)
2. Prediction (of future data) (*Generalization*)

Also can involve:

- Human decision-making (using both above)
- Information retrieval
  - i.e. predictive or descriptive modeling of unspecified subset of available data

General Experimental Procedure for Estimating Models from Data

General Experiment Procedure

It is important to realize that the problem of learning/estimation of dependences from data is only <u>one part</u> of the general experimental procedure used by scientists, engineers, medical doctors, and others who apply statistical (*machine learning data mining, etc.*) methods to make inferences from the data.

General Experiment Procedure

1. Statement of the problem (*goals and requirements*)
2. Hypothesis Formulation (*learning problem statement*)
3. Data Generation/ Experiment Design
4. Data cleaning, encoding, and preprocessing
5. Model Estimation (*learning*)
6. Model Interpretation and Drawing Conclusions

Note:

- each step is complex and usually involves several iterations
- estimated model depends on all previous steps

Cultural and Ethical Issues

- Concerns relate to intellectual integrity of researchers who perform data modeling
- Ethical problems are most evident in life sciences and medical research (where financial implications of data-analytic models are very high)
- [Ioannidis (2005)] *"most published research findings (in clinical research) are false"* **&** *"over-eagerness to find anything that seems significant"*
- Not outright fraud but due to self-serving data analysis
- Over-eagerness → inherent bias in interpreting statistically insignificant differences and reporting them as significant findings!

Honest Disclosure of Results

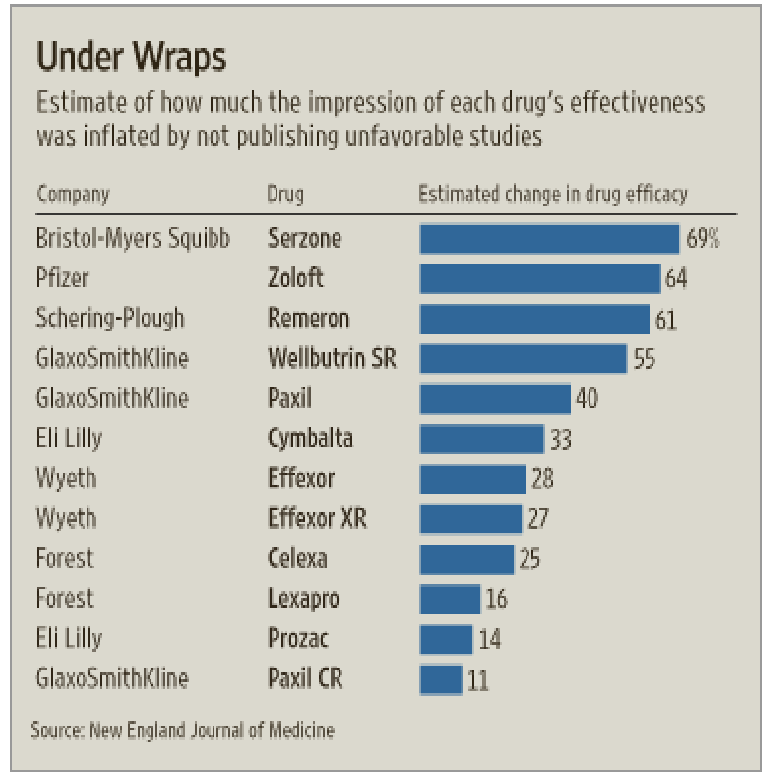
- **Modern drug studies**

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.*

Publication bias:
common in modern research



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