Data Science

Antecedents and Consequents

Bharat Bhushan Verma

Data Science is a Fast Evolving Field

- Robert Geary (1947). Normality is a myth; there never was, and never will be, a normal distribution.

Data Drives Decision Making

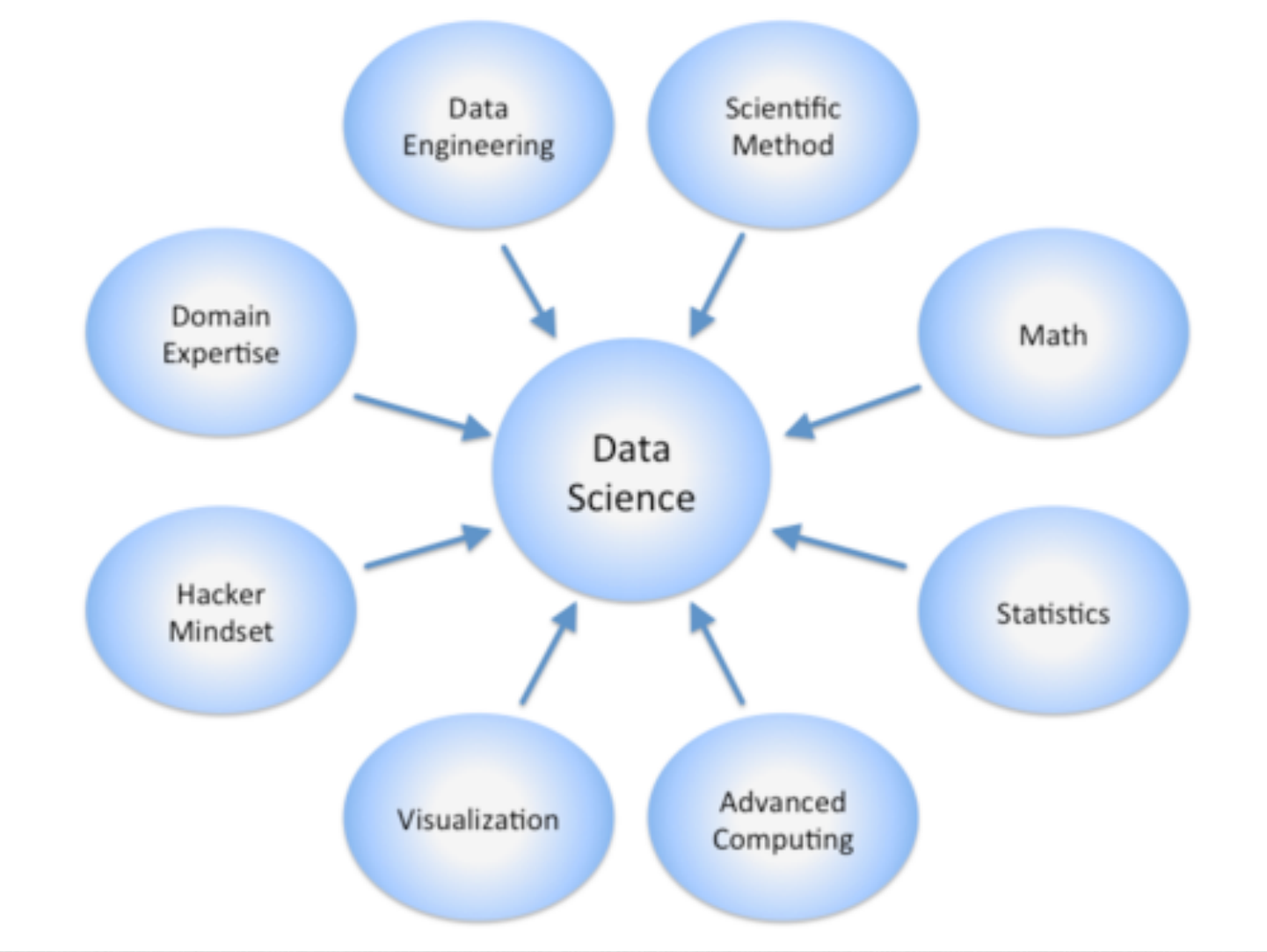
Age of Data Revolution

- Michael Dell (2014). Data is arguably the most important natural resource in this century.
- Sean Murphy (2013). Scientists have long known that data could create new knowledge but now the rest of the world has realized that data can create value.
- William Cleveland (2001). Data Science is an extended field of statistics to incorporate advances in computing.
- Edwards Deming (1942). Data are not taken for museum purposes; they are taken as a basis for doing something.

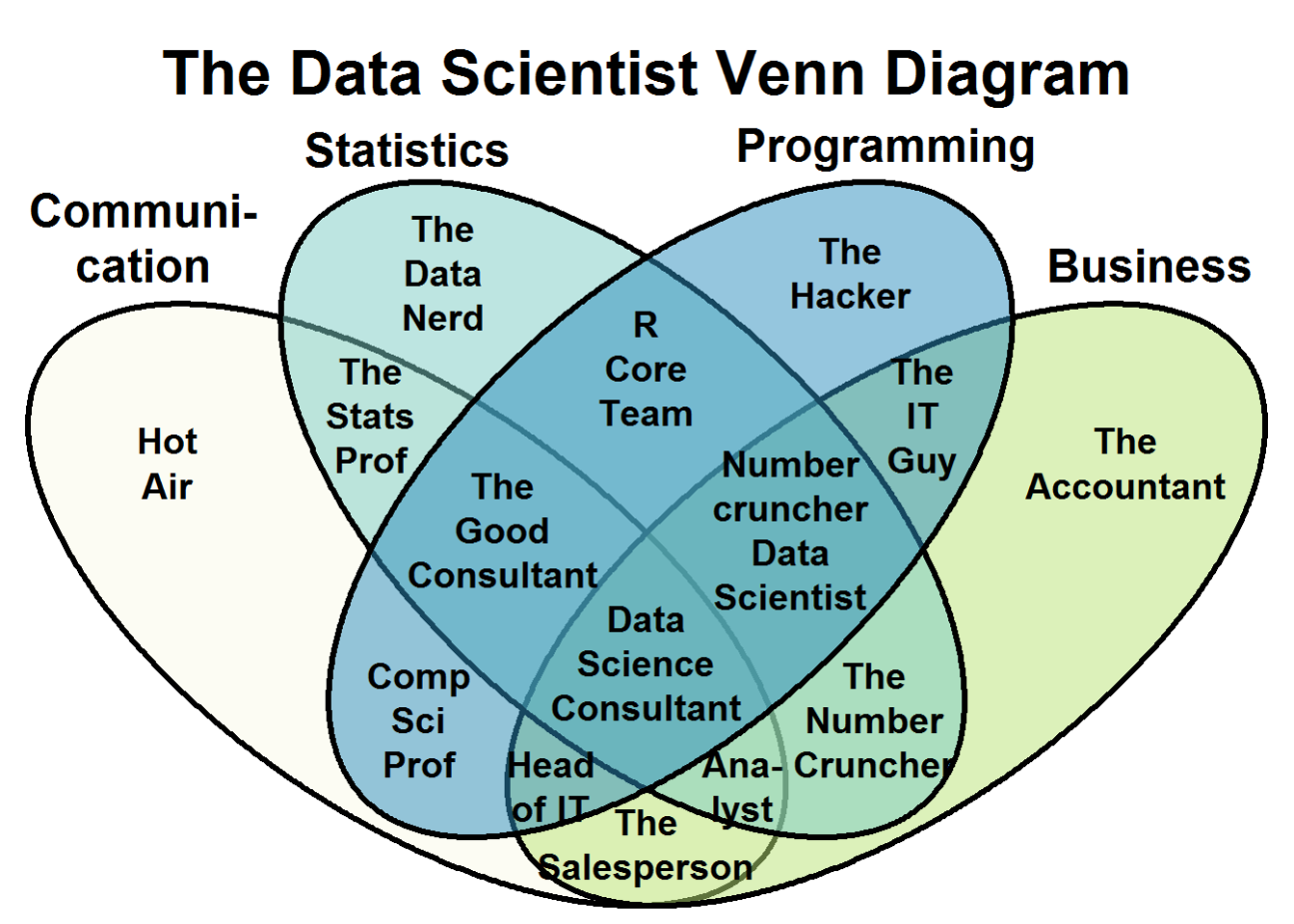
Data Science Vs. Statistics

- Statistics is concerned with analyzing experimental data that have been collected to explain and check the validity of specific existing ideas (hypotheses).
  - Top Down (Explanatory or Confirmatory) Analysis
  - Hypothesis testing
- Data Science is concerned with analyzing secondary data that has been collect for other reasons to create new ideas (hypotheses).
  - Bottom up (Exploratory and Predictive) analysis
  - Idea generation and knowledge discovery

Data Science



Data Science is about Data Mining



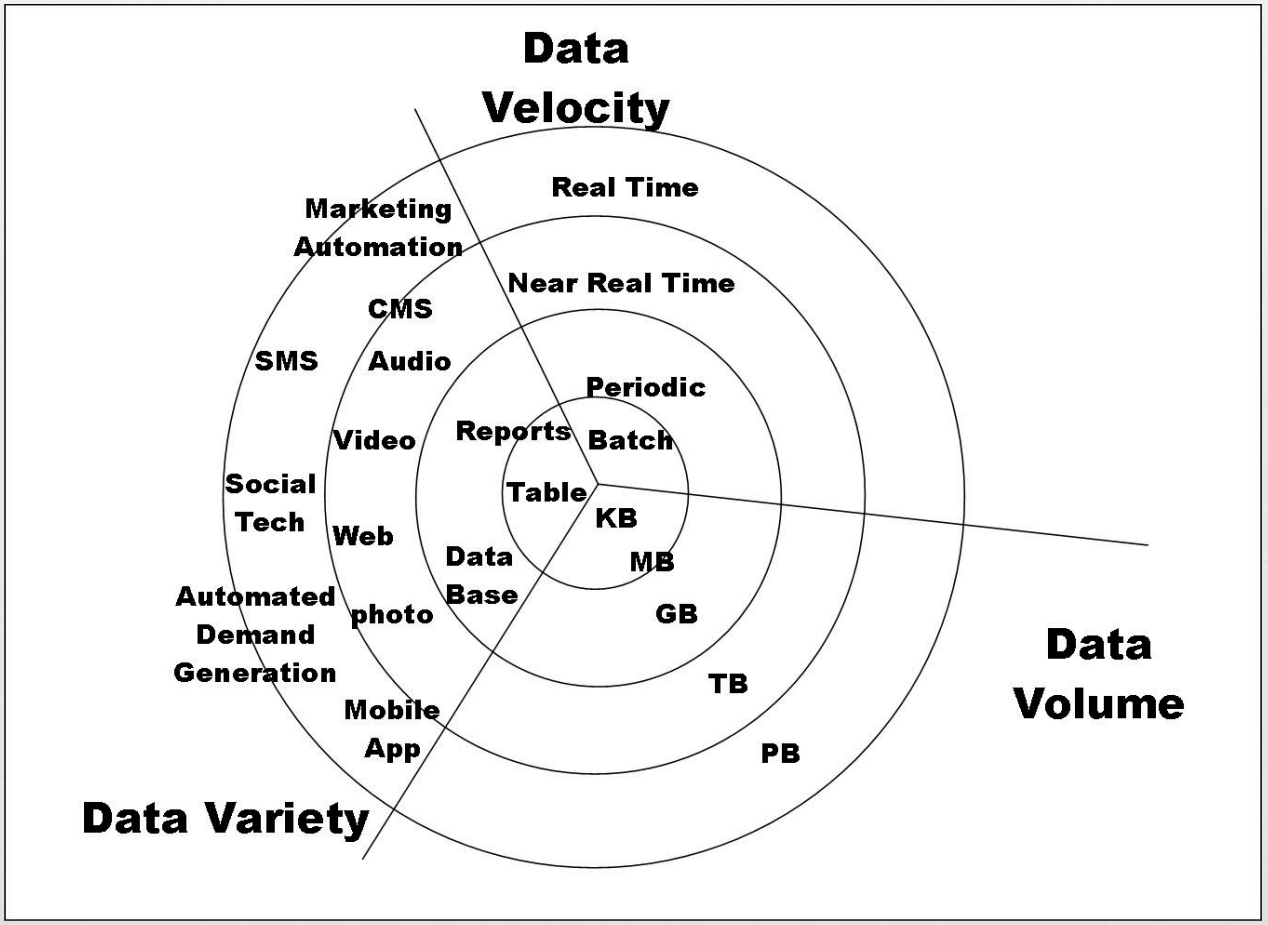
Data Science is

- Broad – broader than any one existing discipline
- Interdisciplinary: Computer Science, Statistics, Databases, Mathematics
- Applied focus on extracting knowledge from data to inform decision making.
- Focuses on the skills needed to collect, manage, store, distribute, analyze, visualize, and reuse data.

- the scientific study of creation, validation and transformation of data to create meaning.

- Though most definitions of data science underplay substantive theory, meta data, privacy and ethics.

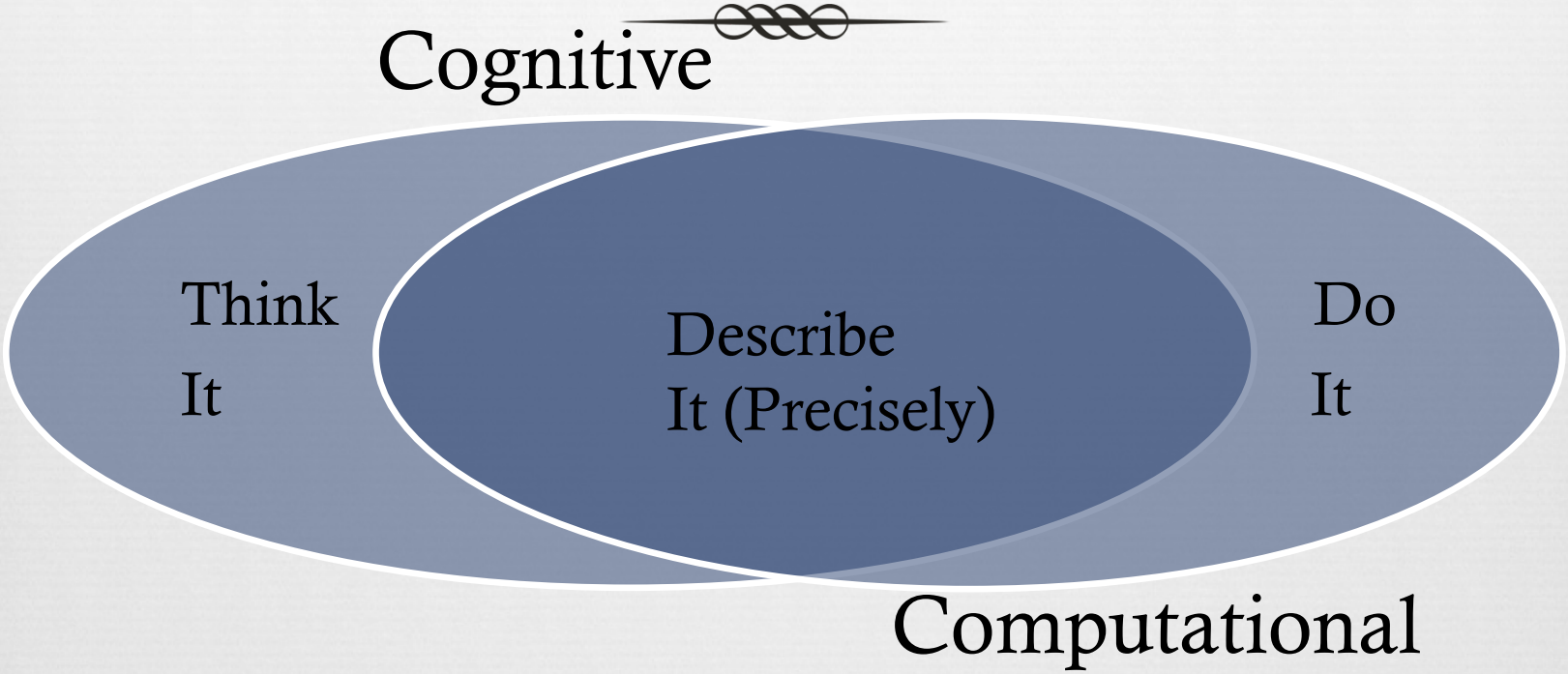
Data Explosion – Big Data



New Behaviors require New Methods

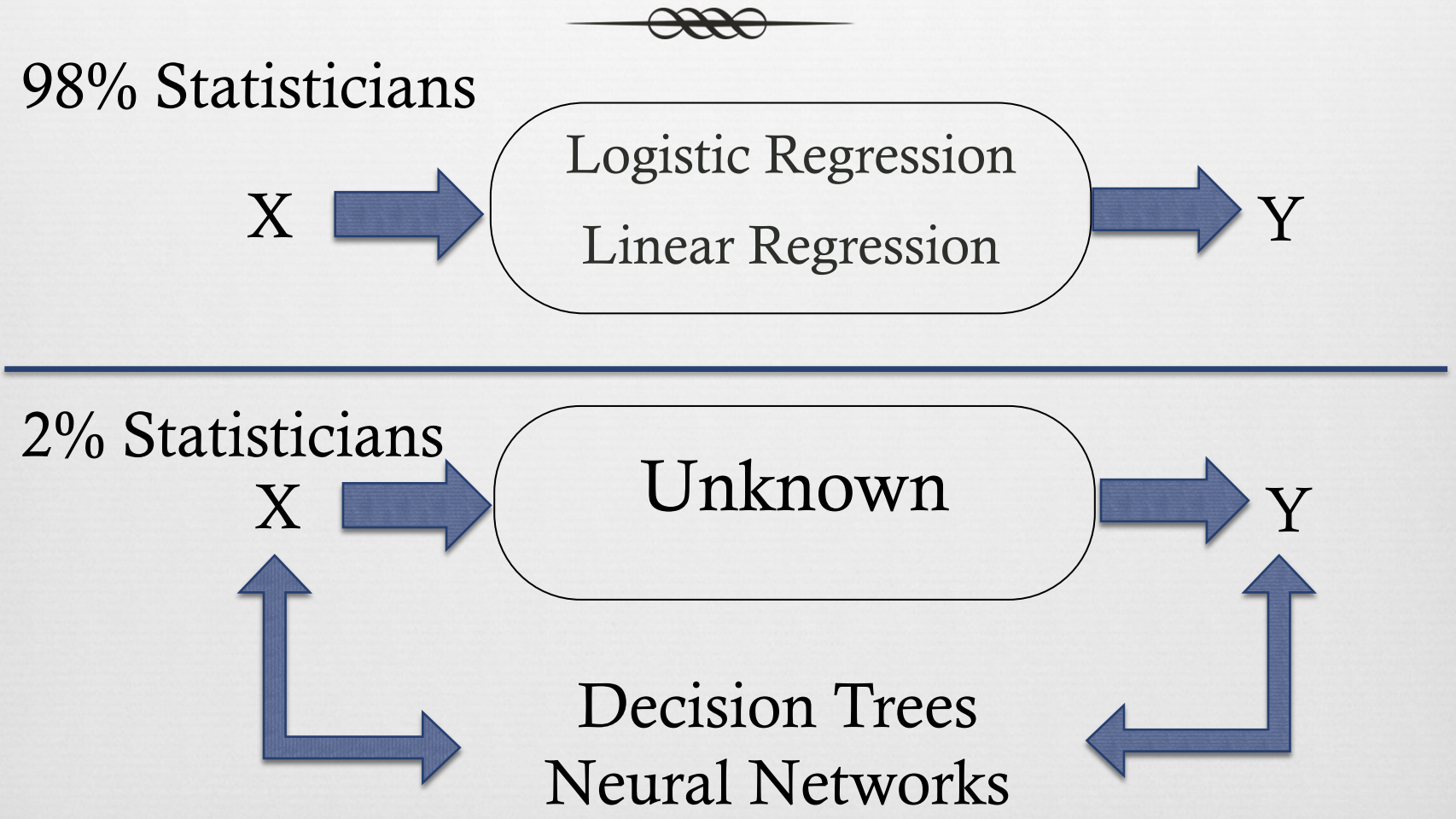
- New Behaviors generate new form of data
  - Online shopping
  - Cell phone usage
  - Crowd sourced recommendation systems
  - Facebook, Google searching, etc.
  - Online mobilization of social protests
- Patterns without understanding are at best uninformative and at worst deeply misleading.
- Real discovery is NOT about modeling patterns in observable data. It is about understanding the processes that produced that data.

The New Paradigm



Small Data: Cognition Time >> Computational Time
Big Data: Cognitive Time << Computational Time

Data Vs. Algorithmic Culture



Data Models

- Concludes on Model Mechanism and not about natures mechanism. Thus conclusions may be wrong.

- In the Current era of big data
  - Little power of Goodness of Fit tests
    - Bickel, Ritov, Stoker (2001). Tailor-made tests for goodness of fit for semi parametric hypotheses.
  - Residual Analysis fails to uncover lack of fit beyond 4 – 5 dimensions.
    - Cleveland, Grousse (1991). Computational methods for local regression.
  - False Positive significance levels (p-values)
    - Galit Shmueli (2012). Too big to fail: Large samples and p-value problem

- Other References
  - Mosteller, F. and Tukey, J. (1977). Data Analysis and Regression.
  - Freedman, D. (1991). Some issues in the foundations of statistics.
  - Freedman, D. (1991). Statistical models and shoe leather.
  - Freedman, D. (1994). From association to causation via regression.

Multiplicity of Data Models

- McCullah and Nelder (1989) - Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this.
- Mountain and Hsiao (1989) – It is difficult to formulate a comprehensive model capable of encompassing all rival models.
- Nobody really believes that multivariate data is multivariate normal
- A priori assumption that nature generates data through a parametric model can result in questionable conclusions.
- If all a man has is a hammer, then every problem looks like a nail.
- Statistics is falling short of dealing with increasing ability of computers to store and manipulate data which is giving rise to non-statisticians finding solutions more often.

Algorithmic Models

- Led by Physicists, Biologists & Computer Scientists
  - Physicists: Wavelet Theory, Spectroscopy etc.
  - Biologists: Epidemic Spread Models, Swarm Intelligence etc.
  - Computer Scientists: Computer Vision, Information Theory etc.
- Focused on predictive accuracy of algorithms
- Problems: Speech, handwriting and Image Recognition, Financial Markets
- New Age Tools
  - Bayesian Methods with Markov Chains
  - Neural Nets & Decision Trees.
  - Vladmir Vapnik's work on SRM and SVM.
  - Breiman's work on Tree Ensembles

Curse/Blessings of Dimensionality

- Support Vector Machines by V Vapnik
- Shape Recognition Forest by Amit and Geman
  - Algorithm superseded the human error in handwriting recognition.
- Challenge of Interpretability Vs. Accuracy
- Algorithmic models can give better predictive accuracy than data models, and provide better information about the underlying mechanism.

Role of Other Fields

- Behavioral Psychology – Intrinsic Biases
- Behavioral Economics – Prospect Theory
- Inherent Characteristics of objects in the world.

Simple Data Visualization

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating which cities are most at risk for spread of the Ebola virus.

Data Makes Everything Clearer?

Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]

1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

*** E-mail: Corresponding spechler@princeton.edu**

Abstract

The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

Data Makes Everything Clearer

Searches for
"MySpace"

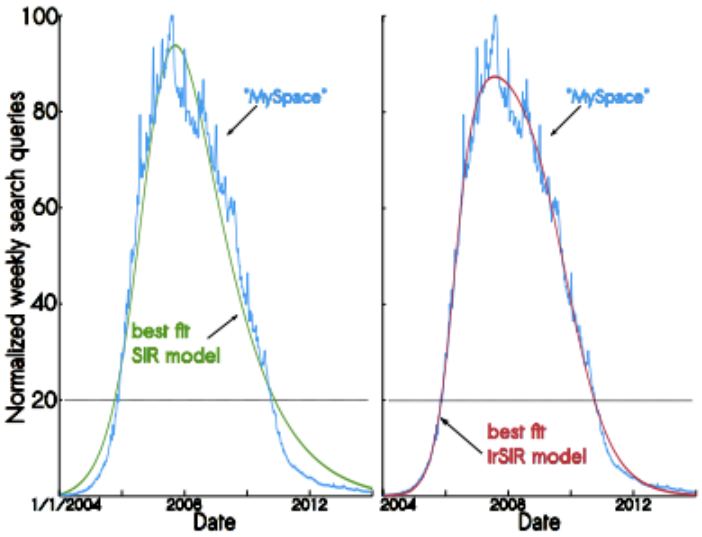
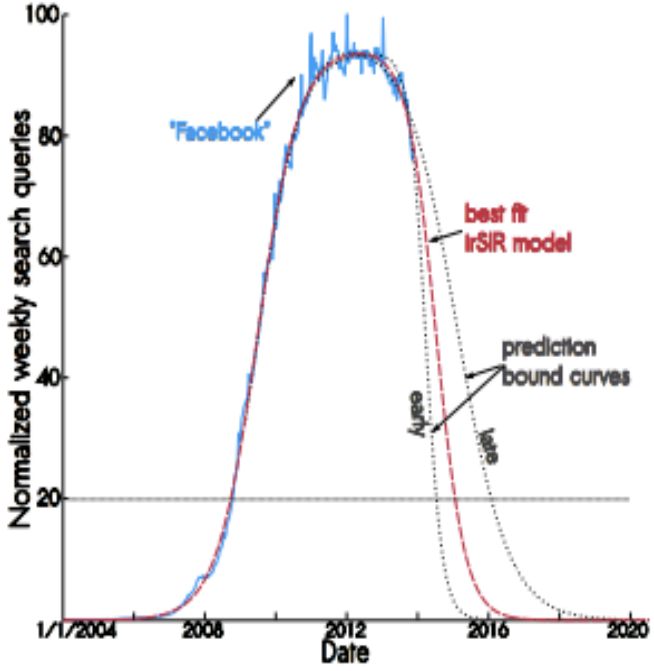


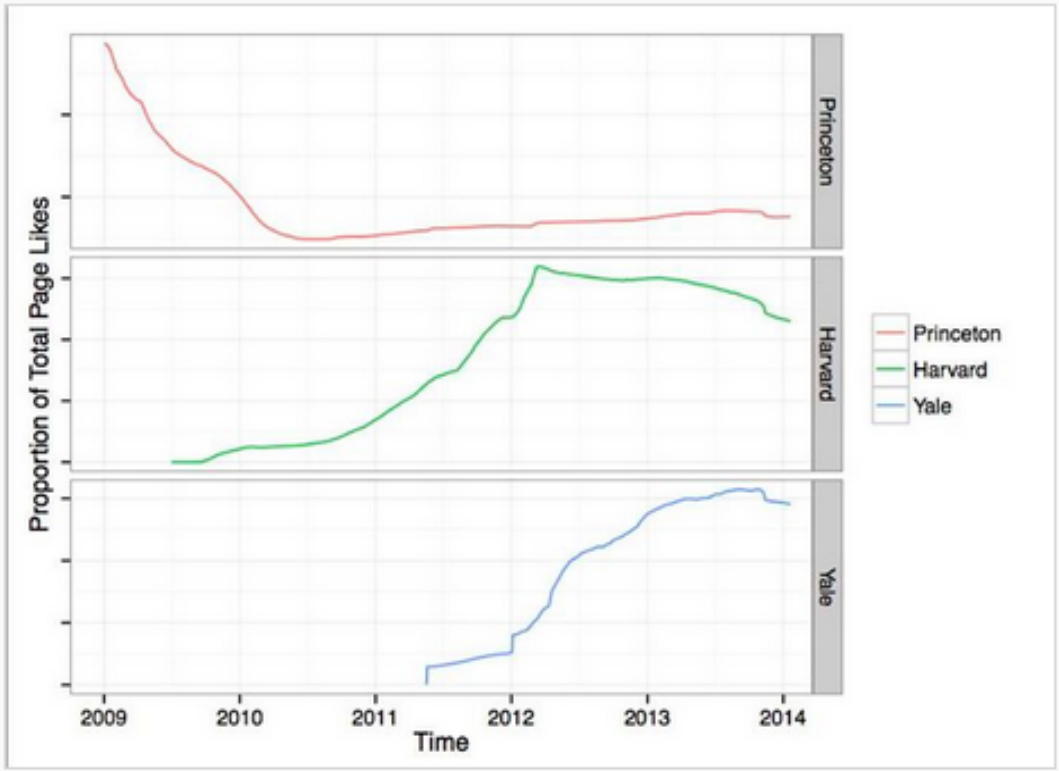
Figure 3: **Data for search query "Myspace" with best fit (a) SIR and (b) irSIR models overlaid.** The search query data are normalized such that the maximum data point corresponds to a value of 100.

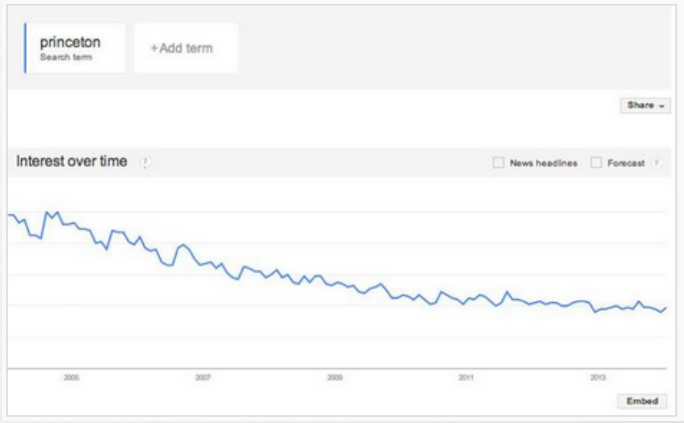


Searches for
"Facebook"

Data Makes Everything Clearer

In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:





and based on Princeton search trends:

"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…

http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/

Few Cases of Behavioral Insights

- Insurance Premium (Premium Amounts)
- Telecom Reload Behavior (Reload Amount)
- Peer Group Influence (Product choices)
- Same response to a given externality (Oreo)

Challenges and Caution

- Tim Hartford (2014). Statisticians have spent the past 200 years figuring out what traps lie in wait when we try to understand the world through data. The data are bigger, faster and cheaper these days – but we must not pretend that the traps have all been made safe. They have not.
- Spiegelhalter (2014). There are a lot of small data problems that occur in big data. They do not disappear because you have got lots of stuff. They get worse.
- Nate Silver (2012). The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning. Before we demand more of our data, we need to demand more of ourselves.
- Tukey (1980). Neither exploratory nor confirmatory is sufficient alone. To try to replace either by other is madness. We need them both.

Challenges and Caution

- Ethics of using and linking big data
- Visualization of the data
- Spurious Associations
- Identification of Confounding factors
- Multiple Hypothesis testing
- Validity of Generalization