Math Models for Data Science

Kristin P. Bennett
Mathematical Sciences Department
Rensselaer Polytechnic Institute
Spring 2006
The VIEW from MATHLAND
Traditional Hypothesis Driven Science Paradigm

- Hypothesis
- Design
- Experiment
- Data
- Data analysis
- Result
“If you experiment needs statistics, you ought to have done a better experiment”

Ernest Rutherford
Data Driven Science/Engineering

Process/Experiment

Data

No Prior Hypothesis
Data Science Research

- Extracting knowledge and information from complex data
- Posing and answering questions regarding data
- Providing performance in the face of massive amounts of data
Enabling Technology

- Mathematics
  - Mathematical Modeling
  - Statistics
  - Optimization

- Computer Science
  - Machine Learning
  - Data Mining
  - Data Bases
Data Science Problems

- Drug Discovery
- Microarray Gene Analysis
- SPAM Email Filtering
- Molecular Epidemiology of Tuberculosis
- Sudoku
Sudoku
Mathematical Models

- Optimization models
  - Linear programming
  - Nonlinear programming

- Probability models
  - Principal component analysis
  - Naïve-Bayes
  - Mixture Models

- Integer Programming
**Required Background**

- **Multivariate Calculus**
  - partial derivatives

- **Linear Algebra**
  - Matrix and vector arithmetic, eigenvalues, inverse of a matrix

- **Basic Probability**
  - Derivation and use of continuous and discrete distribution functions such as normal and binomial distributions. Expectation and covariance.

Will use simple Matlab for assignments – review tutorials if you don’t know it already.

Computational investigation required for final project but you can choose language and platform.
Class information

- See course web page
  www.rpi.edu/~bennek/class/mds/index.html

Name
Where you’re from
Grad/undergrad major year
Math and machine learning experience
Interesting fact about you
Course Requirements

- Commentaries
- Assignments
- Participation – do readings and contribute to discussion in each class.
- Final Project and Presentation (may be group for undergrads)
- (Grad only) Class presentation and lead discussion
Data Science Revolution Discussion

Next Generation for Life Sciences:
“Computers Replace Petri Dishes”

Global IT expenditures in the life sciences
Large pharmaceutical companies make up the biggest, and fastest-growing, customer segment.

Five-year compound annual growth rates
- Lg pharmaceutical firms: 24.3 percent
- Med pharmaceutical: 18.3
- Sm pharmaceutical: 12.6
- Biotech firms: 21.1
- Govt/Academic: 15.1
- Agribusiness: 17.1

Source: IDC
Assignment for Monday

- Readings
- Mandatory Commentary
- Why is drug discovery so hard and what can we do about it?
- Lecturer: Kristin Bennett