A Blind Date with Big Data

Goal

- Measure correlations
- Rank relatedness

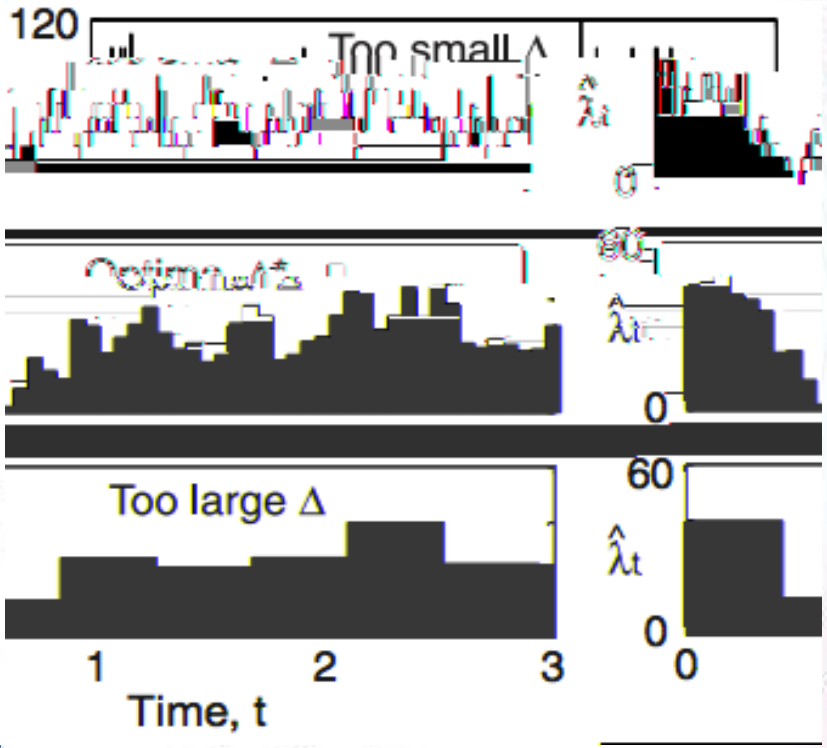
Challenges

- Conditional dependencies
- Missing data

Where We Are

Correlation Calculations

- Mutual Information
- Optimal Histograms
- Efficient Algorithm
  - Ranks NHANES in hours



$$\mathrm{MI}_{XY} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \frac{n_{(i,j)}}{n_{(:,:)}} \log \frac{n_{(:,:)} * n_{(i,j)}}{n_{(i,:)} * n_{(:,j)}}$$

Shimazaki H. and Shinomoto S. A method for selecting the !in size of a time histogram. Neural Computation. "#$$%& 'ol. ()"*& (+$,-(+#%

Where We're Going

Imputation

- 8e1eloping ne0 method !ased on optimal !ins

Correlations

- E6plore 7ernel 8ensit2 Estimation

Conditionals

- .ro!a!ilistic /raphical Models

Personal Goals

- .ractice programming 0ith 1ariet2 of languages
- Stud2 statistics !ioinformatics information theor2
- 3earn a!out all pro4ects of the Sa!eti 3a!
- Ha1e fun5