Admin:
- We'll go over midterm on Tuesday
- Project 1 released (discussed shortly) Due 3/14

HDSC:
- Another podcast (via Marine C.): http://www.thetalkingmachines.com/episodes

Project 1
- Go over project description
- Questions and comments

Entity Resolution and Record Linkage
Exploratory Data Analysis (via visualization)

Record Linkage

Given: Entity sets $E_1$, $E_2$

Goal: Match linked entities $(e_1, e_2)$

1) Define a similarity function $s(e_1, e_2)$
\[ s(e_1, e_2) = \sum_{j \in A} s_j(e_{1j}, e_{2j}) \]

- **Categorical variables:**
  \[ s_j(e_{1j}, e_{2j}) = \begin{cases} 1 & \text{if } e_{1j} = e_{2j} \\ 0 & \text{otherwise} \end{cases} \]

- **Numerical variables:**
  \[ d_j(e_{1j}, e_{2j}) = (e_{1j} - e_{2j})^2 \]
\[ d_j \left( e_{ij}, e_{ij} \right) = e^{-d_j} \]

2) Compute \( S_j (e_1, e_2) \) \( \forall e_1, e_2 \in E_1 \cup E_2 \)

3) Match \( e_1, e_2 \) based on similarity \( S_j (e_1, e_2) \)

**Assumption**
1) Each \( e_1 \) matches to a single \( e_2 \), \( e_2 \) can match to multiple \( e_1 \).
Match $e_1$ to
$$\arg\max_{e_2} S(e_1, e_2)$$

Assumption 2) One-to-one mapping

Matching problem (451)

$G$ (optimal) Don't match $e_1$ if
$$\max_{e_2} S(e_1, e_2) < \text{too small}$$

Exploratory Data Analysis
Understanding the distribution of attribute values in an entity set

- Central tendency
- Spread
- Skew
- Outliers