Statistical Language Modeling

- Statistical language models assign **probabilities** to **sequences of words**

  \[ P(\text{“the dog barked”}) = 4.203 \times 10^{-9} \]

- **Applications**
  - Speech Recognition
  - Machine Translation
  - Spelling Correction
  - Information Extraction
Information Extraction

- IE: Text $\rightarrow$ machine-understandable data

Paris, the capital of France, ...

$(Paris, France) \in \text{CapitalOf}, p = 0.85$

- Applied to Web: better search engines, semantic Web, step toward human-level AI
IE Automatically?

Intractable to get human labels for every concept expressed on the Web

Idea: extract from semantically tractable sentences

...Edison invented the light bulb...
(Edison, light bulb) ∈ Invented
  \[ x \ V y \implies (x, y) \in V \]

...Bloomberg, mayor of New York City...
⇒ (Bloomberg, New York City) ∈ Mayor
  \[ x, C \text{ of } y \implies (x, y) \in C \]
But...

Extraction patterns make errors:

“Erik Jonsson, CEO of Texas Instruments, mayor of Dallas from 1964–1971, and...”

- Empirical fact:
  - Extractions you see over and over tend to be correct
  - The problem is the “long tail”
Challenge: the “long tail”

e.g., (Bloomberg, New York City)  
Tend to be correct

A mixture of correct and incorrect

e.g., (Dave Shaver, Pickerington)  
(Ronald McDonald, McDonaldland)
Mayor McCheese
Assessing Sparse Extractions

Idea:
Use statistical language models to determine which sparse extractions are more likely to be correct
Project

- Work in teams of 2-4
  - E-mail me w/ team names and members
- Submit distributions over words for blanks in sentences (demo)
- Do whatever you want, but use probabilistic graphical models
  - We’ll discuss a few candidate ideas in class
- Record what works, what doesn’t
- Presentations Dec 2, 4 (last week of class)
  - 8 mins + 4 mins Q/A
- Final Report (~2 pages of text + figures/tables)
The Distributional Hypothesis

*Terms in the same class tend to appear in similar contexts.*

<table>
<thead>
<tr>
<th>Context</th>
<th>Hits with Chicago</th>
<th>Hits with Twisp</th>
</tr>
</thead>
<tbody>
<tr>
<td>“cities including __”</td>
<td>42,000</td>
<td>1</td>
</tr>
<tr>
<td>“__ and other cities”</td>
<td>37,900</td>
<td>0</td>
</tr>
<tr>
<td>“__ hotels”</td>
<td>2,000,000</td>
<td>1,670</td>
</tr>
<tr>
<td>“mayor of __”</td>
<td>657,000</td>
<td>82</td>
</tr>
</tbody>
</table>