1980-1995: A sea change in AI technologies

Example:
Natural Language Processing

*The Great Wave off Kanagawa* by Hokusai, ~1830
**SHRDLU: 1969 – NLP solved??**

Person: *PICK UP A BIG RED BLOCK.*

Computer: OK. (does it)

Person: *GRASP THE PYRAMID.*

Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

Person: *FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.*

Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

Computer: OK. (does it)

Person: *WHAT DOES THE BOX CONTAIN?*

Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.

Person: *WHAT IS THE PYRAMID SUPPORTED BY?*

Computer: THE BOX.
SHRDLU: A demonstration proof

1990: Parsing Research in Crisis…

Informal IBM study (Jelinek) in 1990:

- No existing parsers worked for short sentences from news text.
- Methodology:
  - Compared a range of best “broad coverage” parsers in U.S.
  - All but one: < 40% correct (hand checked) for 10 word sentences
  - One claimed 60% - (I don’t believe it…)

How could this be true?

- Most NLP work in interactive systems, with major user adaptation

Proposed Solution (IBM, me, a few others):

- A hand-built corpus of examples (10^6 words of WSJ)
- Some unknown statistical model trained on that set of examples.
- IT WORKED!
An Early Robust Statistical NLP Application

• A Statistical Model For Etymology (Church ’85)

• Determining etymology is crucial for text-to-speech

<table>
<thead>
<tr>
<th>Italian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>AldriGHetti</td>
<td>lauGH, siGH</td>
</tr>
<tr>
<td>IannuCCi</td>
<td>aCCept</td>
</tr>
<tr>
<td>ItaliAno</td>
<td>hAte</td>
</tr>
</tbody>
</table>
An Early Robust Statistical NLP Application

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angeletti</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Iannucci</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Italiano</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Lombardino</td>
<td>58%</td>
<td>Italian</td>
</tr>
<tr>
<td>Asahara</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Fujimaki</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Umeda</td>
<td>96%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Anagnostopoulos</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Demetriadis</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Dukakis</td>
<td>99%</td>
<td>Russian</td>
</tr>
<tr>
<td>Annette</td>
<td>75%</td>
<td>French</td>
</tr>
<tr>
<td>Deneuve</td>
<td>54%</td>
<td>French</td>
</tr>
<tr>
<td>Baguenard</td>
<td>54%</td>
<td>Middle French</td>
</tr>
</tbody>
</table>

- A very simple statistical model (an upcoming homework) solved the problem, despite a wild statistical assumption
The Initial Driver: Success of Statistical Models in Automatic Speech Recognition

http://recognize-speech.com/acoustic-model/knn/benchmarks-comparison-of-different-architectures
Recent Significant Advances In NLP

- Web-scale information extraction & question answering
  - IBM’s Watson

- Interactive Dialogue Systems
  - Apple’s Siri
  - Microsoft Cortana
  - Amazon Echo

- Google Translate
  - Automatic Machine Translation
The State of NLP (and all of AI)

NLP Past before 1990:
- Rich Representations

NLP Present:
- *Powerful Statistical Disambiguation*
Fast Review of Basic Probability
Basic Review of Probability

From Jean Gallier’s
*Discrete Mathematics Ch 6*
on Discrete Probability

*The Great Wave off Kanagawa* by Hokusai, ~1830
Semi-Rigorous Definition of Probability

• **Standard account**: assumes some (potentially abstract) *experiment*

• **Sample space** $\Omega$: The set of *all possible outcomes* of an experiment.
  - must be known
  - could be uncountably or countably infinite, or finite

• **Example experiment**: Rolling two dice
  - Sample space $\Omega = D \times D$, where $D = \{1,2,3,4,5,6\}$
  - $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$
  - Each member $\omega \in \Omega$ is called an *elementary event*

• **Example (non-elementary) event**: Rolling doubles
  - $B = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$

(This and several following from Jean Gallier’s *Discrete Mathematics* Ch 6 on Discrete Probability)
Our problems: **Finite** Probability Space

- **Finite Probability Space:** \((\Omega, P)\)
- 1. A nonempty, finite set \(\Omega\) of elementary events.
- 2. \(P: \Omega \rightarrow R\), a function (probability measure) satisfying ‘Kolomogorov’ axioms:
  - \(0 \leq P(\omega) \leq 1\) \(\forall \omega \in \Omega\)
  - \(\sum_{\omega \in \Omega} P(\omega) = 1\)
  - For any event \(A \subseteq \Omega\), \(P(A) = \sum_{\omega \in A} P(\omega)\)
- **Results**
  - \(P(\emptyset) = 0, P(\Omega) = 1\)
  - If \(A \subseteq B\) then \(P(A) \leq P(B)\)
  - Since \((A \cup B) = (A \cap B) \cup (A - B) \cup (B - A)\):
    - \(P(A \cup B) = P(A) + P(B) - P(A \cap B)\)
    - \(P(\bar{A}) = 1 - P(A)\)
- **To be more careful with countably infinite \(\Omega\), need more structure, but we’ll stay with finite probability spaces**
Independence

• Given a discrete probability space \((\Omega, P)\), two events \(A\) and \(B\) are *independent* if: \(P(A \cap B) = P(A)P(B)\).

• Two events are *dependent* if they are not independent.
Independence Example

- **Example:** \( \Omega = \{HH, HT, TH, TT\} \)
- **Assume a uniform probability measure:** \( \forall \omega \in \Omega, P(\omega) = \frac{1}{|\Omega|} \)
- **Event A: First flip is H:** \( A = \{HH, HT\} \)
- **Event B: Second flip is H:** \( B = \{HH, TH\} \)
  - \( P(A \cap B) = P(HH) = \frac{1}{4} \)
  - \( P(A)P(B) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4} \)
- *Independent* events

- **Event A: First flip is H:** \( A = \{HH, HT\} \)
- **Event B: Contains a T:** \( B = \{HT, TH, TT\} \)
  - \( P(A \cap B) = P(HT) = \frac{1}{4} \)
  - \( P(A)P(B) = \frac{1}{2} * \frac{3}{4} = \frac{3}{8} \)
- *Non-Independent* events
Random Variables (Standard version)

Given some Probability Space \((\Omega, P)\),

- A random variable \(X: \Omega \rightarrow R\) is a function defined from the sample space to the Reals
- RVs are functions! Never forget! When Max Mintz asks you what a RV is... you say “a function from the sample space to the reals!”
- RVs take on values in an experiment (e.g. temperature, growth in inches, weight, ...)
- For some \(a \in R\) want to know \(P(X = a)\). What does this mean? We defined probability only over \(\Omega\)
  - “\(P(X = a)\)” = \(P(X^{-1}(a)) = P(\{\omega \in \Omega \mid X(\omega) = a\})\)
  - “\(P(X \leq a)\)” = \(P(X^{-1}((-\infty, a])) = P(\{\omega \in \Omega \mid X(\omega) \leq a\})\)
Example of Random Variables

- Example: $\Omega = \text{Sample space of 5 coin flips}$. $\{HH, HT, TH, TT\}$ is the sample space of 2 coin flips. So $|\Omega| = 32$. Assume uniform probability measure.
- $X(\omega)$ is the number of heads in a given flip, ie $X(\text{HHTHT}) = 3$
- $Y(\omega)$ is 1 if H appears in $\omega$, otherwise is 0

Random Variables have distributions
- $P(X = 0) = \frac{1}{32}, P(X = 1) = \frac{5}{32}, P(X = 2) = \frac{10}{32} \ldots$
- $P(X = a) = \binom{5}{a}$
- $P(Y = 0) = \frac{1}{32}$
- $P(Y = 1) = \frac{31}{32}$
PMF, CDF

- **Probability Mass Function (PMF) for a *discrete* distribution is:**
  - \( f: R \rightarrow [0,1] \)
  - \( \forall a \in R, \ f_X(a) = P(X = a) = P(\{\omega \in \Omega: X(\omega) = a\}) \)

- **Cumulative Distribution Function is:**
  - \( F: R \rightarrow [0,1] \)
  - \( \forall a \in R, \ F_X(a) = P(X \leq a) \)
  - **\( F \) is monotonic nondecreasing:** \( \forall x \leq y, F(x) \leq F(y) \)
Examples: Probability Mass Function and Cumulative Distribution Function

- Experiment: roll 2 dice
- $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$
- RV $X$ is the sum of the numbers on the throw
- $X(\omega) = X((d_1, d_2)) = d_1 + d_2$

![Histogram](image1.png)

![Cumulative Distribution Function](image2.png)

*Fig. 6.5* The cumulative distribution function for the sum of the numbers on two dice
Conditional Probability

Given a probability space \((\Omega, P)\),

- For any two events \(A\) and \(B\), if \(P(B) \neq 0\), we define the conditional probability that \(A\) occurs given that \(B\) occurs \([P(A \mid B)]\) as

\[
P(A \mid B) = \frac{P(A \land B)}{P(B)}
\]

- From this follows Chain Rule:
  - \(P(A \land B) = P(A \mid B)P(B)\)
  - \(P(A \land B \land C) = P(A \mid B \land C)P(B \land C) = P(A \mid B \land C)P(B \mid C)P(C)\)
  - Etc…
Example with Conditional Probability

- A family has two children. What is the probability that both are girls, given at least one is a girl?
- Try it!
Cond. Prob Example Cont’d

• $\Omega = \{GG, GB, BG, BB\}$
• $\forall \omega \in \Omega, P(\omega) = \frac{1}{|\Omega|}$
• **Event** $B = \{GB, BG, BB\}$ of having at least one girl
• **Event** $A = \{BB\}$ of having two girls
• **Event** $A \cap B = \{BB\}$

• $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$

• **Check out Monty Hall Problem in Gallier**
Probability for AI: An approach to uncertainty

Artificial Intelligence
AIMA, Chapter 13.1

Some slides adapted from CMSC 421 (U. Maryland) by Bonnie Dorr
Outline

• A major problem: *uncertainty*
• Probability: the *right* way to handle uncertainty
• Frequentists vs. Bayesians
• Review: Fundamentals of probability theory
• Joint Probability Distributions
• Conditional Probability
• Probabilistic Inference
Uncertainty and Truth Values: a mismatch

- Let action $A_t = \text{leave for airport t minutes before flight}$.  
  Will $A_{15}$ get me there on time? true/false  
  Will $A_{20}$ get me there on time? true/false  
  Will $A_{30}$ get me there on time? true/false  
  Will $A_{200}$ get me there on time? true/false

- Problems: The world is not
  - Fully observable (road state, other drivers’ plans, etc.)
  - Deterministic (flat tire, etc.)
  - Single agent (immense complexity modeling and predicting traffic)
  - And often incoming information is wrong: noisy sensors (traffic reports, etc.)
A purely logical approach:

- Risks falsehood: “$A_{25}$ will get me there on time”
- Leads to conclusions that are too weak for decision making:
  - “$A_{1440}$ will get me there on time”  
    — almost certainly true but I’d have to stay overnight at the airport!
  - “$A_{25}$ will get me there on time if there is no accident on the bridge and it doesn’t rain and my tires remain intact, etc.”  
    — True but useless
AI Methods for handling uncertainty

Default or nonmonotonic logic:

- Assume my car does not have a flat tire
- Assume $A_{25}$ works unless contradicted by evidence

- Issues: What assumptions are reasonable? How to handle contradiction?

Rules with ad-hoc fudge factors:

- $A_{25} \rightarrow_{0.3}$ get there on time
- $Sprinkler \rightarrow_{0.99} WetGrass$
- $WetGrass \rightarrow_{0.7} Rain$

- Issues: Problems with combination, e.g., Sprinkler causes Rain??

→ Probability

- Model agent's degree of belief
  — “Given the available evidence, $A_{25}$ will get me there on time with probability 0.04”

- Probabilities have a clear calculus of combination
Our Alternative: Use Probability

- Given the available evidence, $A_{25}$ will get me there on time with probability 0.04

- Probabilistic assertions summarize the ignorance in perception and in models
  - Theoretical ignorance: often, we simply have no complete theory of the domain, e.g. medicine
  - Uncertainty (partial observability): Even if we know all the rules, we might be uncertain about a particular patient
  - Laziness: Too much work to list the complete set of antecedents or consequents to ensure no exceptions
Probabilities as *Degrees of Belief*

- **Subjectivist (Bayesian) (Us)**
  - Probability is a model of *agent’s own* degree of belief

- **Frequentist (Not us, many statisticians)**
  - Probability is *inherent* in the process
  - Probability is *estimated* from measurements
Frequentists: Probability as expected frequency

- **Frequentist** *(Not us, many statisticians)*
  - Probability is *inherent* in the process
  - Probability is *estimated* from measurements

- $P(A) = 1$: $A$ will always occur.
- $P(A) = 0$: $A$ will never occur.
- $0.5 < P(A) < 1$: $A$ will occur more often than not.
Bayesians: Probabilities as Degrees of Belief

- **Subjectivist (Bayesian) (Us)**
  - Probability is a model of *agent’s own* degree of belief

- $P(A) = 1$: Agent completely *believes* $A$ is true.
- $P(A) = 0$: Agent completely *believes* $A$ is false.
- $0.5 < P(A) < 1$: Agent *believes* $A$ is more likely to be true than false.

- *Increasing evidence*
  - strengthens belief
  - therefore changes probability estimate
Frequentists vs. Bayesians I

• Actual dialogue one month before 2008 election
  • Mitch (Bayesian): I think Obama now has an 80% chance of winning.
  • Sue (Frequentist): What does that mean? It’s either 100% or 0%, and we’ll find out on election day.

Why be Bayesian?

• Often need to make inferences about singular events
  • (See above)

• De Finetti: It’s good business
  • If you don’t use probabilities (and get them right) you’ll lose bets – see discussion in AIMA
Making decisions under uncertainty

Suppose I believe the following:

- \( P(A_{25} \text{ gets me there on time} \mid \ldots) = 0.04 \)
- \( P(A_{90} \text{ gets me there on time} \mid \ldots) = 0.70 \)
- \( P(A_{120} \text{ gets me there on time} \mid \ldots) = 0.95 \)
- \( P(A_{1440} \text{ gets me there on time} \mid \ldots) = 0.9999 \)

- Which action to choose?
  - It still depends on my preferences for missing flight vs. time spent waiting, etc.
Decision Theory

- **Decision Theory** develops methods for making optimal decisions in the presence of uncertainty.
  - Decision Theory = utility theory + probability theory

- **Utility theory** is used to represent and infer preferences: Every state has a degree of *usefulness*

- An agent is rational if and only if it chooses an action $A$ that yields the highest *expected* utility (expected usefulness).

- Let $O$ be the possible outcomes of $A$, $U(o)$ be the utility of outcome $o$, and $P_A(o)$ be the probability of $o$ as an outcome for action $A$, then the Expected Utility of $A$ is $EU(A) \equiv \sum_{o \in O} P_A(o)U(o)$