1. Machine Learning
2. Local generalization
3. The curse of highly-variable functions
4. Inspiration from the brain:
   1. Distributed representations
   2. Deep architectures
Statistical Learning

• New multi-disciplinary field

• Numerous applications
What is Learning?
Learn underlying and previously unknown structure, from examples

= CAPTURE THE VARIATIONS
Supervised Learning
Supervised Training Example

Entrée X \rightarrow six

Sortie f(X)
deux!

Cible Y
Learn by heart

- Easy for machines
- Hard for humans (!? !?)

or

Generalize

- Mathematically: fundamentally difficult

Easier for humans!!!!
Supervised Learning and Local Generalization

\[ * = \text{example (x,y)} \]

\[ \text{vraie fonction: inconnue!} \]

\[ \text{fonction apprise: prediction} = f(x) \]
Locally capture the variations

\[ \text{prediction } f(x) \]

\[ \text{point test } x \]

\[ *= \text{example d'apprentissage} \]

\[ \text{vraie } = \text{inconnue} \]

\[ \text{apprise } = \text{interpolaee} \]
Easy when there are only a few variations

* = example (x,y)

vraie fonction: inconnue!

fonction apprise: prediction = f(x)
Curse of dimensionality

To generalize locally, need examples representative of each possible variation.
Learning Brains

- $10^{11}$ neurons, $10^{14}$ synapses
- Complex neural network
- Learning: modify synapses
Visual System

Sequence of transformations / abstraction levels
Deep Architecture in the Brain

Area V4 → Higher level visual abstractions
Area V2 → Primitive shape detectors
Area V1 → Edge detectors
Retina → pixels

Primitive shape detectors

Deep Architecture in the Brain
Distributed Representations

Many neurons active simultaneously in the brain: around 1%

The input is represented by the activation of a set of features that are not mutually exclusive.
Can be exponentially more efficient than local representations
Neurally Inspired Language Models

• Classical statistical models of word sequences: local representations
• Input = sequence of symbols, each element of sequence = 1 of N possible words
• Distributed representations: learn to embed the words in a continuous-valued low-dimensional semantic space
Neural Probabilistic Language Models

Successes of this architecture and its descendents: beats localist state-of-the-art in NLP in most tasks (language model, chunking, semantic role labeling, POS)
Embedding Symbols

- MAY, WOULD, COULD, SHOULD, MIGHT, MUST, CAN, CANNOT, Couldn’t, Won’t, WILL

- ONE, TWO, THREE, FOUR, FIVE, SIX, SEVEN, EIGHT, NINE, TEN, ELEVEN, TWELVE, THIRTEEN, FOURTEEN, FIFTEEN, SIXTEEN, SEVENTEEN, EIGHTEEN

- MILLION BILLION

- ZERO

- MONDAY TUESDAY WEDNESDAY THURSDAY FRIDAY SATURDAY SUNDAY

- JANUARY FEBRUARY MARCH APRIL JUNE JULY AUGUST SEPTEMBER OCTOBER NOVEMBER DECEMBER
### Nearby Words in Semantic Space

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>Jesus</td>
<td>XBOX</td>
<td>Reddish</td>
<td>Scratched</td>
</tr>
<tr>
<td>Spain</td>
<td>Christ</td>
<td>Playstation</td>
<td>Yellowish</td>
<td>Smashed</td>
</tr>
<tr>
<td>Italy</td>
<td>God</td>
<td>Dreamcast</td>
<td>Greenish</td>
<td>Ripped</td>
</tr>
<tr>
<td>Russia</td>
<td>Resurrection</td>
<td>PS###</td>
<td>Brownish</td>
<td>Brushed</td>
</tr>
<tr>
<td>Poland</td>
<td>Prayer</td>
<td>SNES</td>
<td>Bluish</td>
<td>Hurled</td>
</tr>
<tr>
<td>England</td>
<td>Yahweh</td>
<td>WH</td>
<td>Creamy</td>
<td>Grabbed</td>
</tr>
<tr>
<td>Denmark</td>
<td>Josephus</td>
<td>NES</td>
<td>Whitish</td>
<td>Tossed</td>
</tr>
<tr>
<td>Germany</td>
<td>Moses</td>
<td>Nintendo</td>
<td>Blackish</td>
<td>Squeezed</td>
</tr>
<tr>
<td>Portugal</td>
<td>Sin</td>
<td>Gamecube</td>
<td>Silvery</td>
<td>Blasted</td>
</tr>
<tr>
<td>Sweden</td>
<td>Heaven</td>
<td>PSP</td>
<td>Greyish</td>
<td>Tangled</td>
</tr>
<tr>
<td>Austria</td>
<td>Salvation</td>
<td>Amiga</td>
<td>Paler</td>
<td>Slashed</td>
</tr>
</tbody>
</table>
Insufficient Depth

Insufficient depth =
May require exponential-size architecture

Sufficient depth =
Compact representation

O(n)
Breakthrough!

Before 2006
Failure of deep architectures

After 2006
Train one level after the other, \textit{unsupervised}, extracting abstractions of gradually higher level
Success of deep distributed neural networks

Since 2006

• Record broken on MNIST handwritten character recognition benchmark
• A component of winning NetFlix entry
• State-of-the-art beaten in language modeling
• NSF et DARPA are interested…
Challenges and Ongoing Research

**Challenge:**
- Non-convex optimization
- Deep: huge number of local minima
- Recurrent nets / context learning

**Humans as inspiration:**
- Curriculum: first show easy examples, like teachers with students
- Collective computation: genetic-like parallel search through many communicating individuals
- Learning context: multiple time scales
Musical Curriculum Project

• Digitize scores in a musical teaching method
• Train a model, one lesson at a time, each time rehearsing only the past lessons
• Hypothesis: can get better model then training with everything in one go
Conclusions

1. Intelligence → knowledge → learning

2. Curse of dimensionality / highly varying functions: limitations of local generalization

3. Solution: distributed + deep architectures

4. Challenge: local minima

5. Cognitively inspired solutions: guide training to learn intermediate concepts