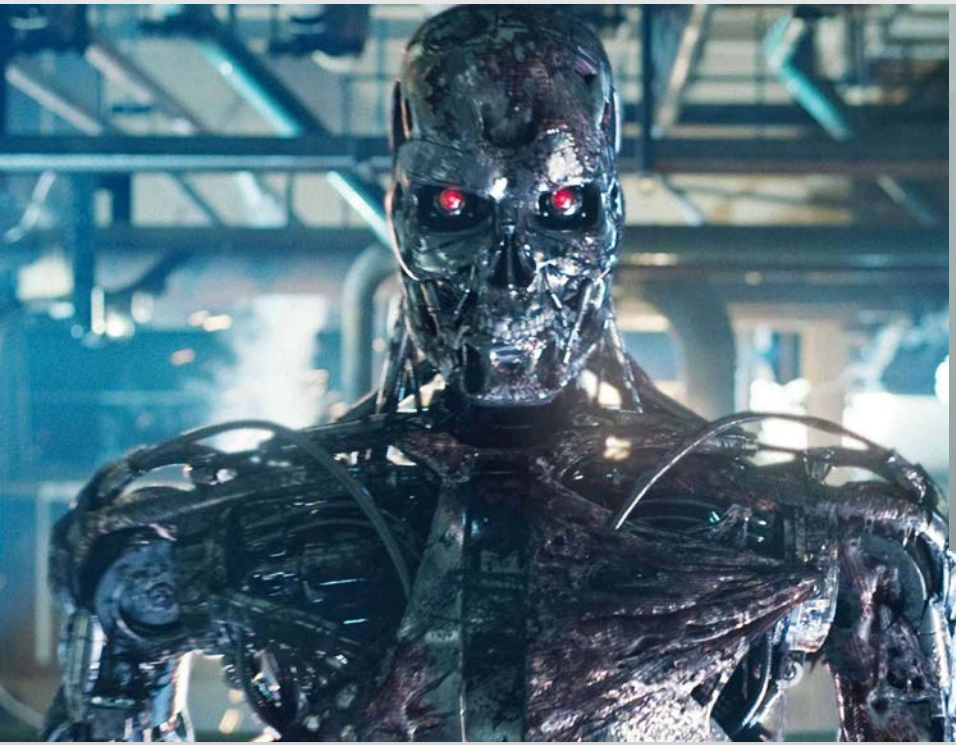
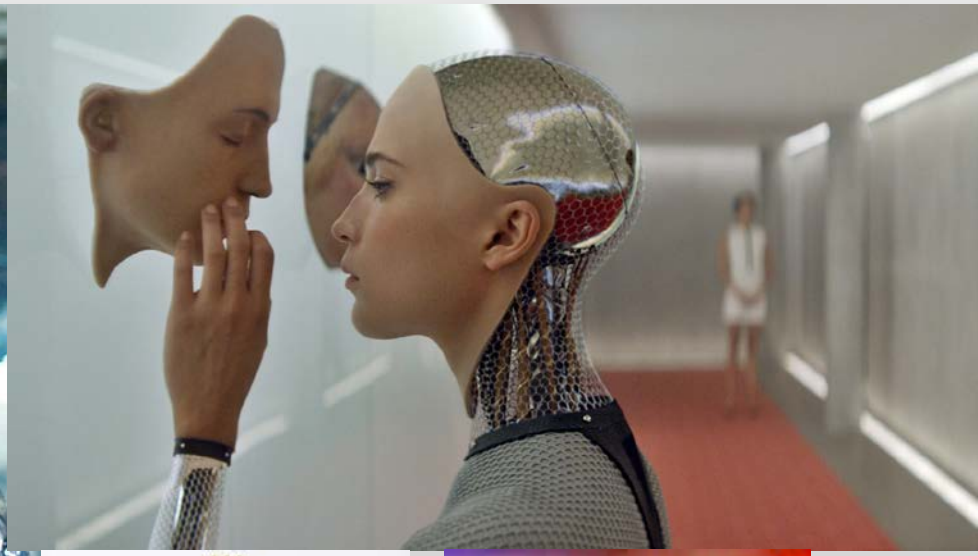
BIGDATA@CHALMERS
Hosted by INFORMATION AND COMMUNICATION TECHNOLOGY A CHALMERS AREA OF ADVANCE

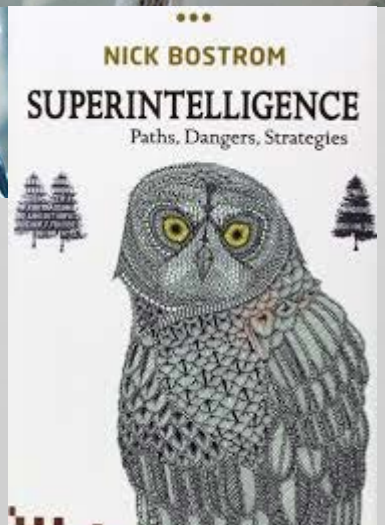
NLP Technologies for Cognitive Computing
Geilo Winter School 2017

Devdatt Dubhashi
LAB
(Machine Learning. Algorithms, Computational Biology)
Computer Science and Engineering
Chalmers

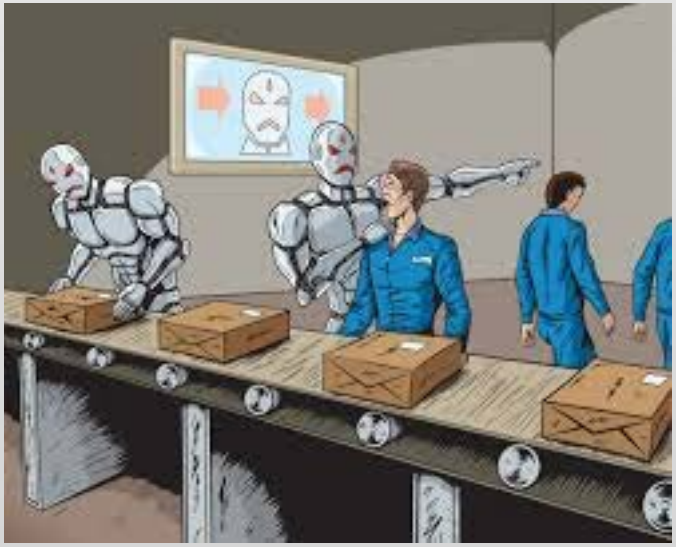
Horizon (100 years): Superintelligence







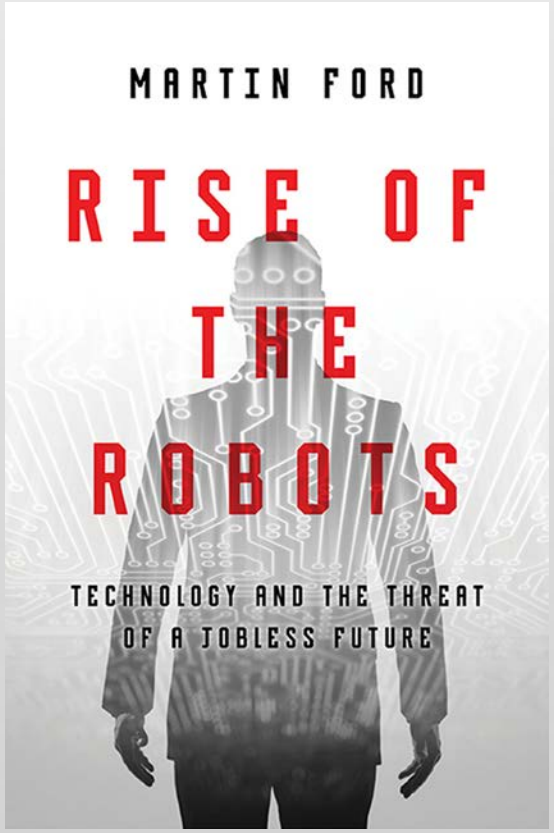
Horizon (20 years): Automation

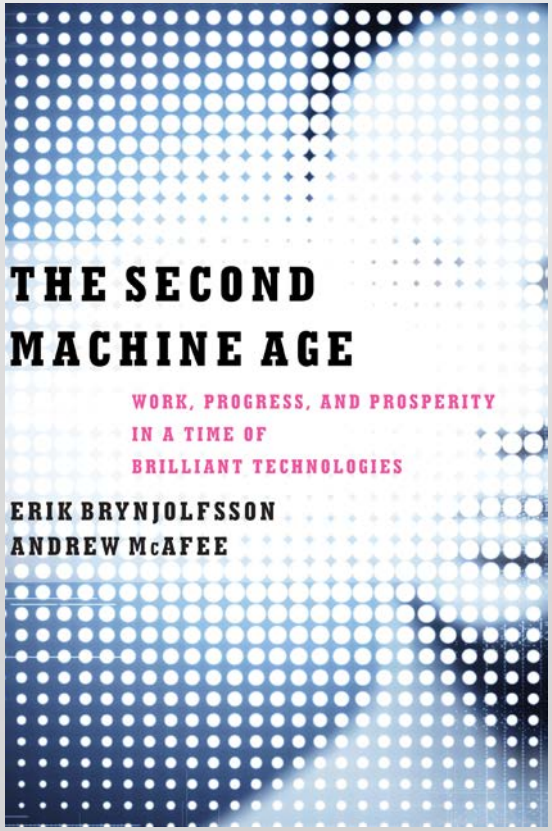


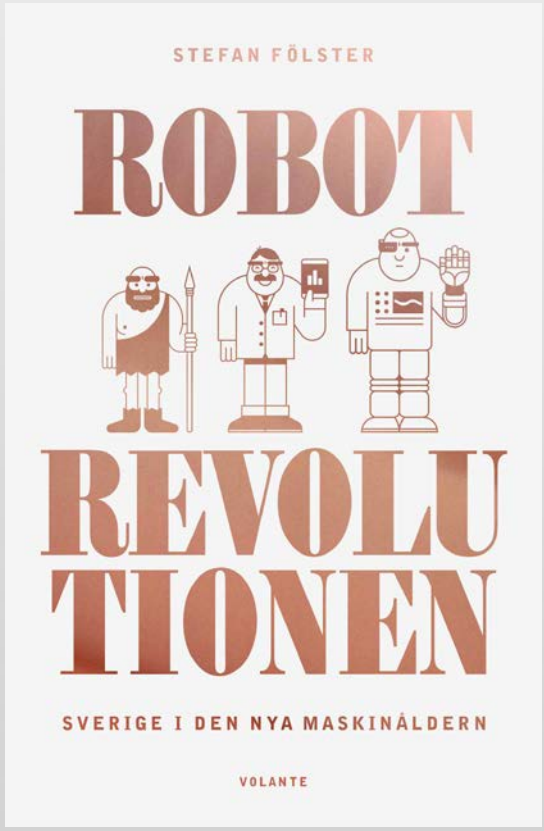
- " … we really have to think through the economic implications. Because most people aren't spending a lot of time right now worrying about singularity—they are worrying about "*Well, is my job going to be replaced by a machine*?" WIRED Nov. 2016

D. Dubhashi and S. Lappin,
"AI Dangers: Real and Imagined"
Comm. ACM (to appear)

A Spectre is Haunting the World



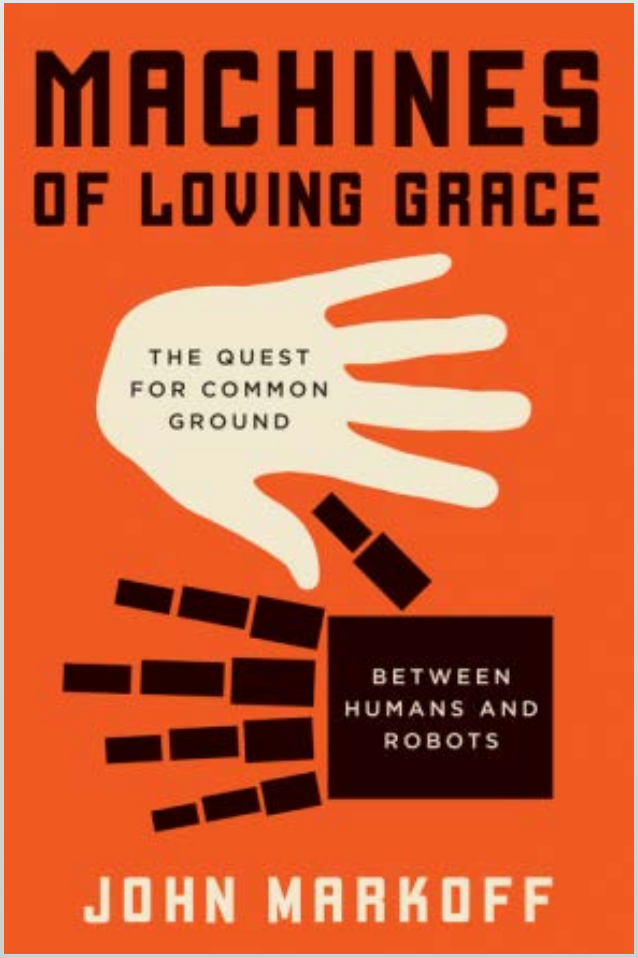




"Greatest problem of 21st century Economics is what to do with surplus humans."
Yuval Noah Harari, *Homo Deus: History of the Future* (2016)

A Tale of Two Stanford Labs

- Artificial Intelligence (AI John McCarthy)
- Intelligence Augmentation (IA Douglas Engelbart)



Why do we need Cognitive Assistants?

"The reason I was interested in interactive computing, even before we knew what that might mean, arose from this conviction that we would be able to **solve really difficult problems** only through **using computers to extend the capability of people** to collect information, create knowledge, manipulate and share it, and then to put that knowledge to work…**Computers** most radically and usefully extend our capabilities when they **extend our ability to collaborate** to solve problems beyond the compass of any single human mind.[1]"

[1] Improving Our Ability to Improve: A Call for Investment in a New Future. Douglas C. Engelbart, September 2003.

What is a Cognitive Assistant?

A software agent (cog) that

- "augments human intelligence" (Engelbart's definition[1] in 1962)
- Performs tasks and offer services (assists human in decision making and taking actions)
- Complements human by offering capabilities that is beyond the ordinary power and reach of human (intelligence amplification)

[1]Augmenting Human Intellect: A Conceptual Framework, by Douglas C. Engelbart, October 1962

From Cognitive Assistance at Work: Cognitive Assistant for Employees and Citizens, by Hamid R. Motahari-Nezhad, AAAI 2015 Fall Symposium.


What can I help you with?

Today

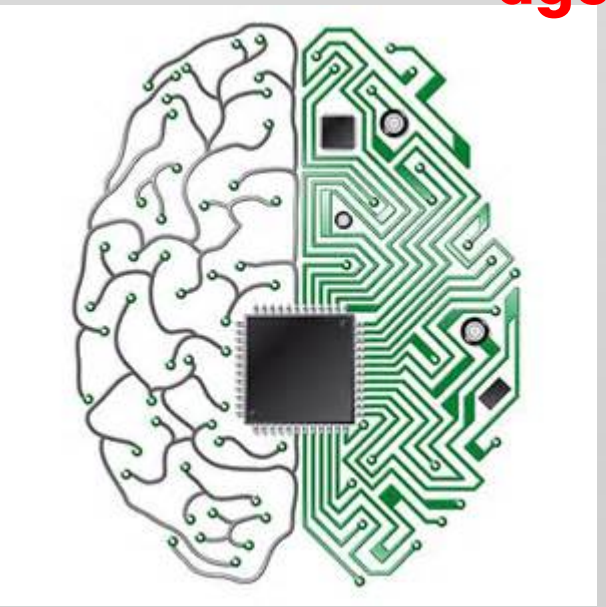

Hi, I'm Cortana.


Google Now UI

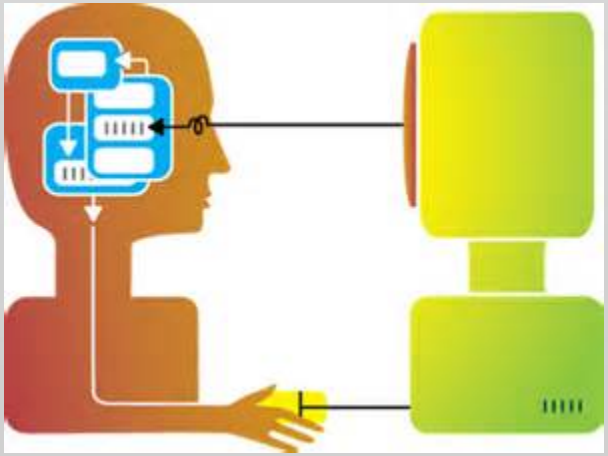

amazon echo
Always ready, connected, and fast. Just ask.
amazon

The Vision...

All pervasive cognitive computing agents .



IBM Watson



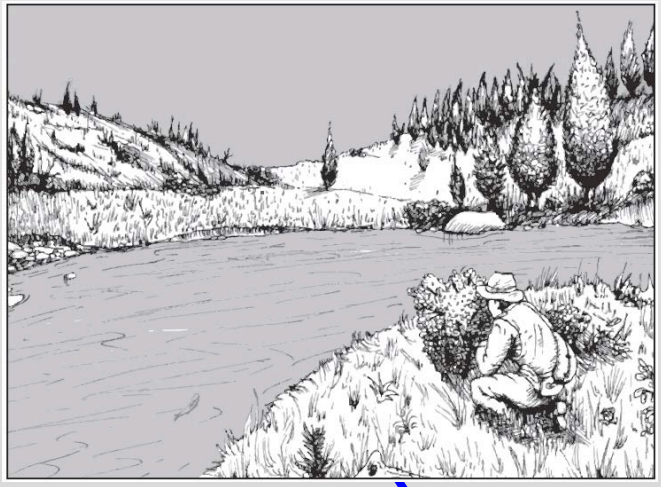
AI: Roadmaps to the Future

- B. Lake, J. Tennenbaum *et al*: "Building machines that learn and think like people" In press at *Behavioral and Brain Sciences*. 2016
- T. Mikolov, A. Joulin and M. Baroni. “A Roadmap towards Artificial Intelligence”, 2015 arxiv.
- J. Schmidthuber, “On Learning to think”, 2015 arxiv

How to Dance with the Robots

- Natural Language Processing (NLP) and Understanding
- Interaction, Feedback, Communication, Learning from the environment
- Causal reasoning
- Intuitive Physics
- Behavioural psychology

Why Language is difficult ..



polysemous

synonymous

Concept Layer

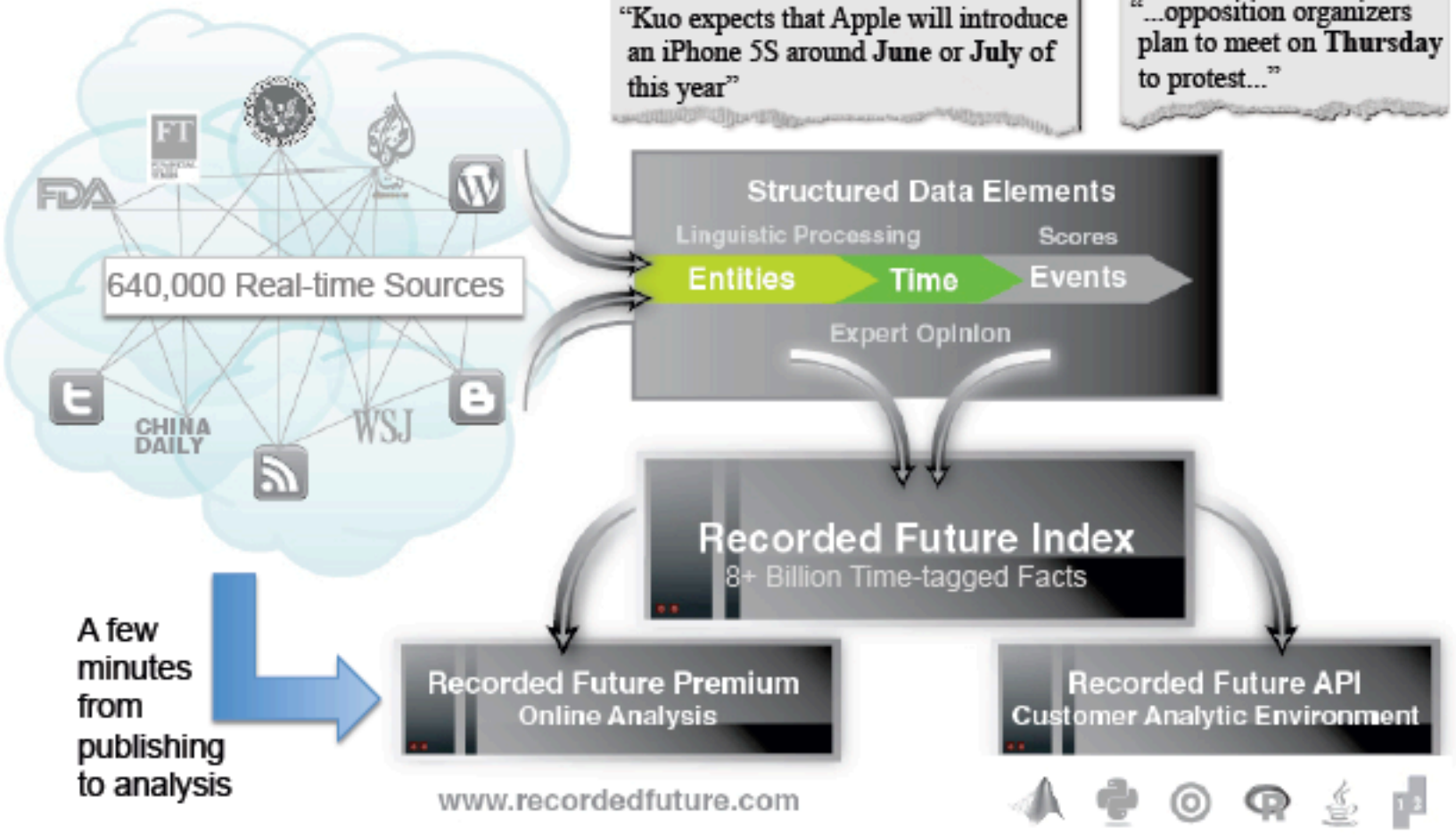
Lexical Layer

He sat on the river bank and counted his dough.

She went to the bank and took out some money.

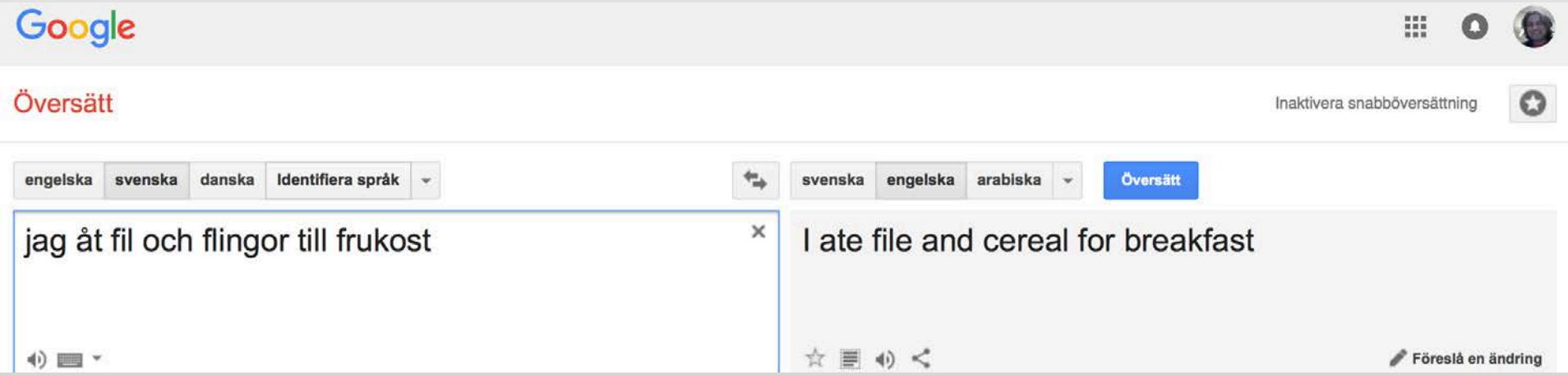
Inside the Web Intelligence Machine

Recorded Future

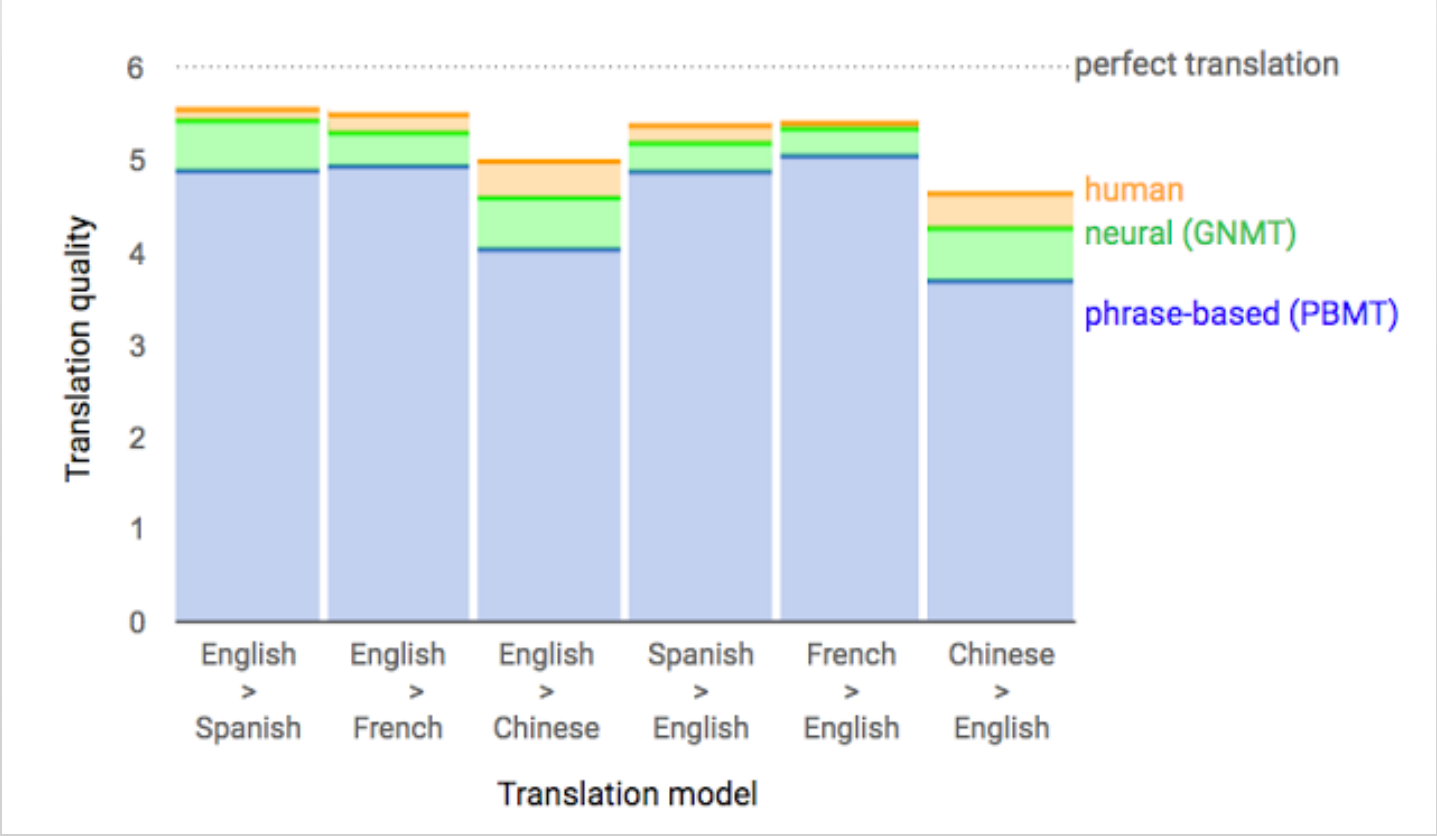



Geology
Name the three types of rock.
1. Classic
2. Punk
3. Hard

Word senses and Machine Translation



Google Neural Machine Translation



Google Translate

Input sentence:	*Translation (PBMT):*	*Translation (GNMT):*	*Translation (human):*
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

reduce translation errors across its Google Translate service by between 55 percent and 85 percent

Goals and Contents of Lectures

Core Machine Learning

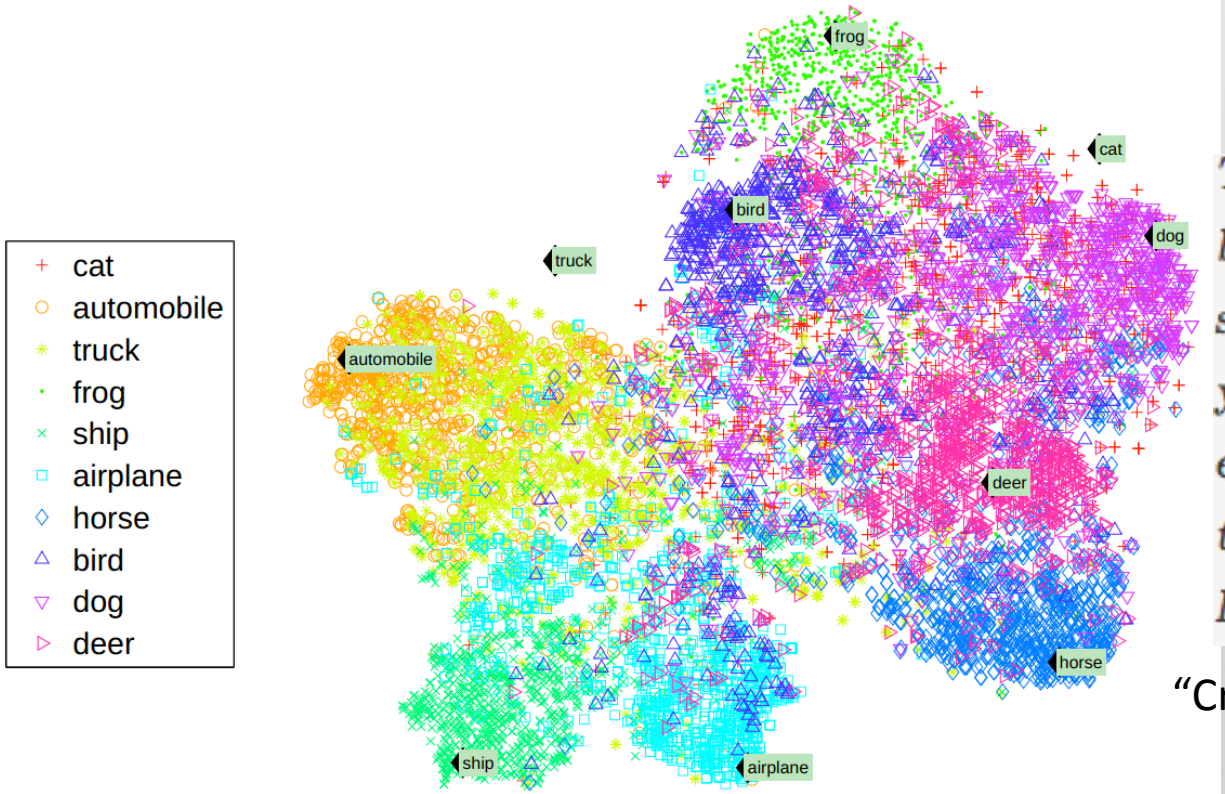
- Supervised learning: large scale logistic regression, neural networks
- Unsupervised learning: clustering
- Optimization: first order methods, submodular functions

NLP Applications

- Distributional semantics
- Summarization
- Word sense induction and disambiguation

WORD EMBEDDINGS

Word Embeddings

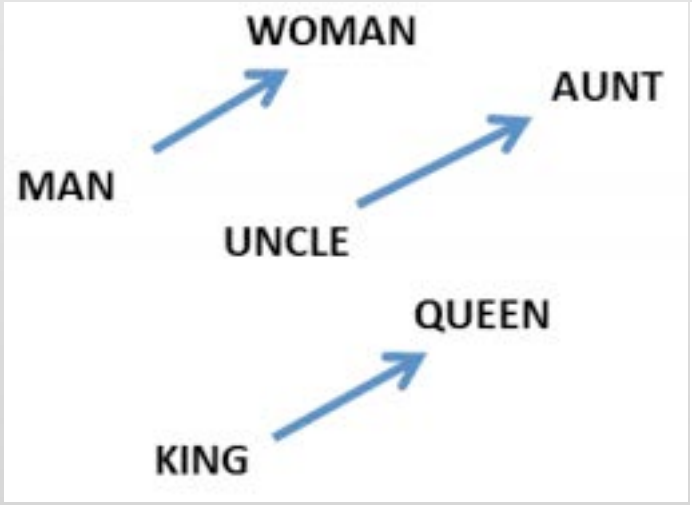


$$W : \text{words} \rightarrow \mathbb{R}^n$$

The use of word representation
become a key "secret sauce"
success of many NLP systems i
years, across tasks including
entity recognition, part-o
tagging, parsing, and seman
labeling. (Luong et al. (2013))

"Crown jewel of NLP", J. Howard (KD

Word Embeddings capture meaning

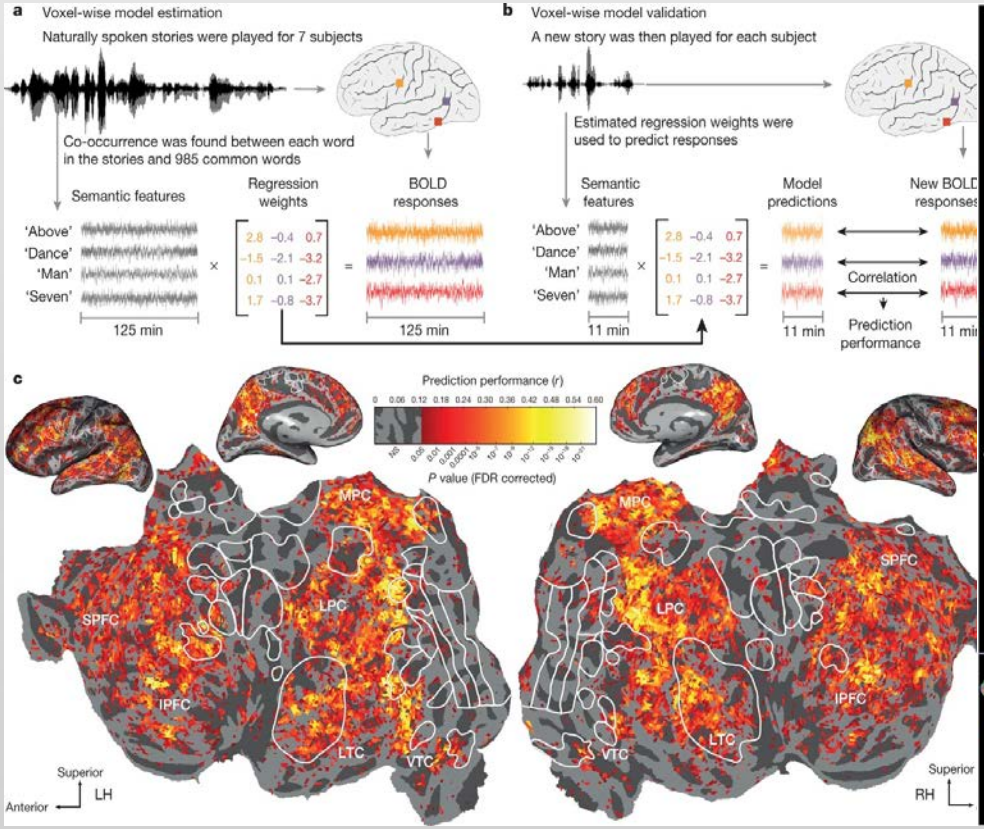


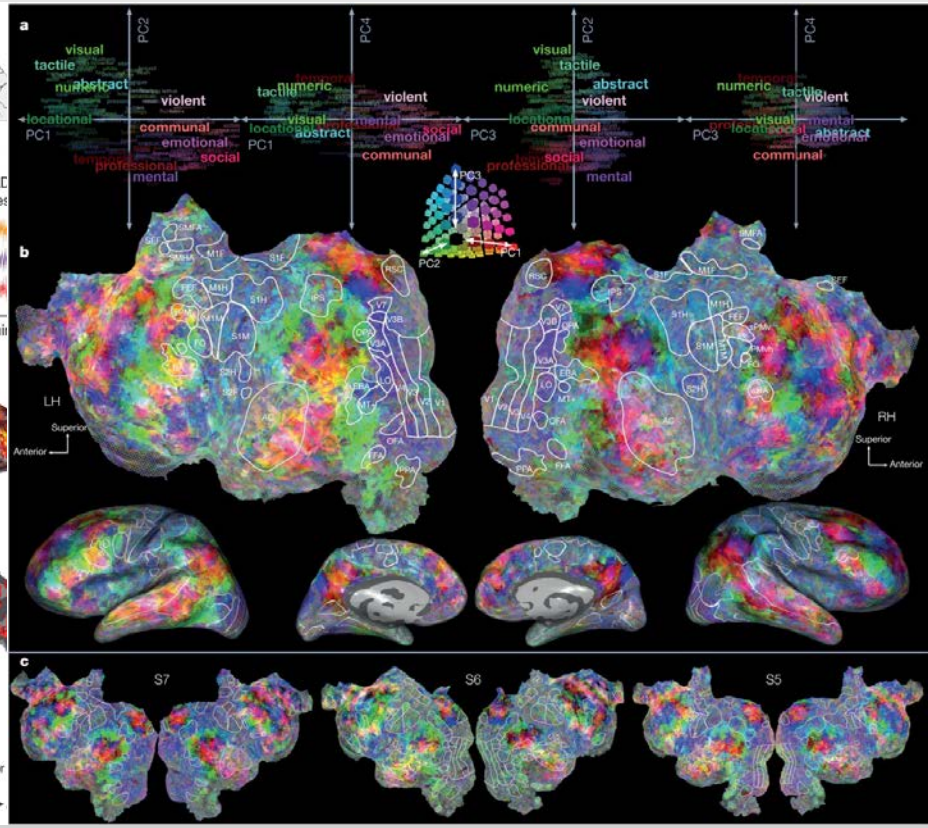
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

Voxel-wise modelling





A G Huth *et al. Nature* **532,** 453–458 (2016) doi:10.1038/nature17637

nature

Distributional Hypothesis

- “Know a man by the company he keeps”. (Euripedes)
- Distributional Hypothesis (Harris 54, Firth 57): if two words are similar in meaning, they will have similar distributions in texts, that is, they will tend to occur in similar linguistic contexts.

Distributional Models: LSA

Terms		d1	d2	d3		q
a		1	1	1		0
arrived		0	1	1		0
damaged		1	0	0		0
delivery		0	1	0		0
fire		1	0	0		0
gold	A =	1	0	1	q =	1
in		1	1	1		0
of		1	1	1		0
shipment		1	0	1		0
silver		0	2	0		1
truck		0	1	1		1

U

-0.4201	-0.075	-0.046
-0.2995	0.2001	0.4078
-0.1206	-0.275	-0.4538
-0.1576	0.3046	-0.2006
-0.1206	-0.275	-0.4538
-0.2626	-0.379	0.1547
-0.4201	-0.075	-0.046
-0.4201	-0.075	-0.046
-0.2626	-0.379	0.1547
-0.3151	0.6093	-0.4013
-0.2995	0.2001	0.4078

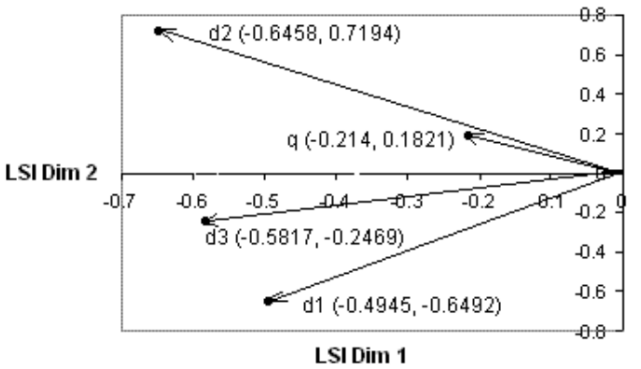
S

4.0989	0	0
0	2.3616	0
0	0	1.2737

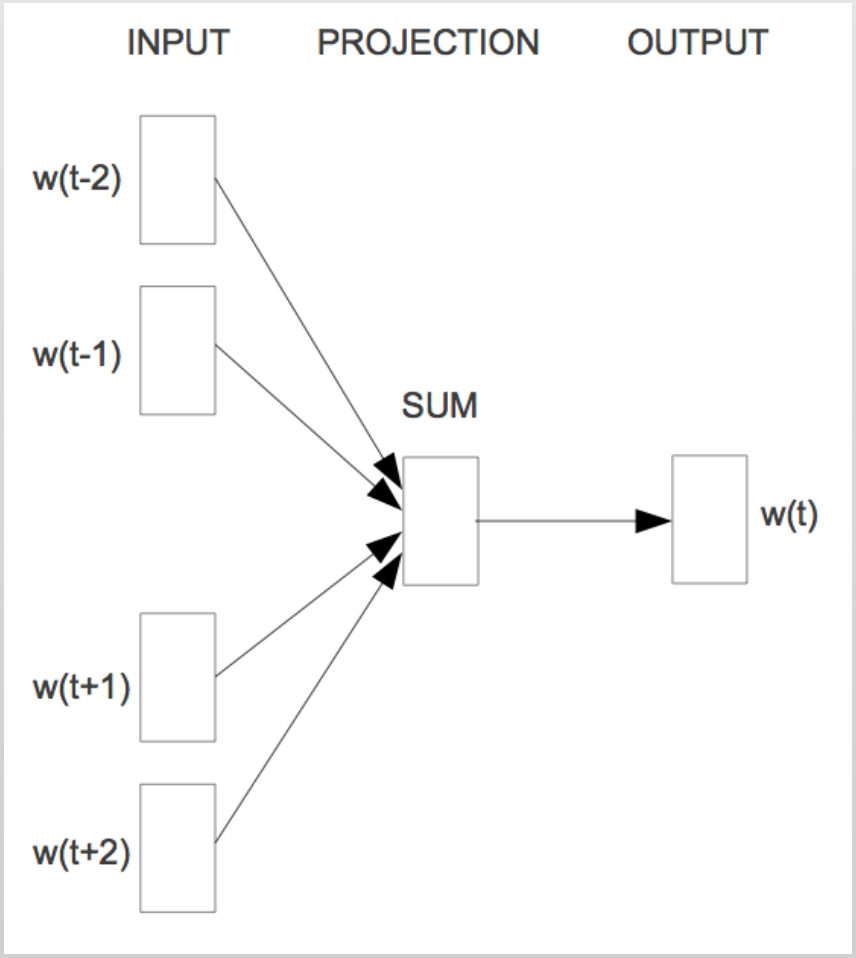
V^T

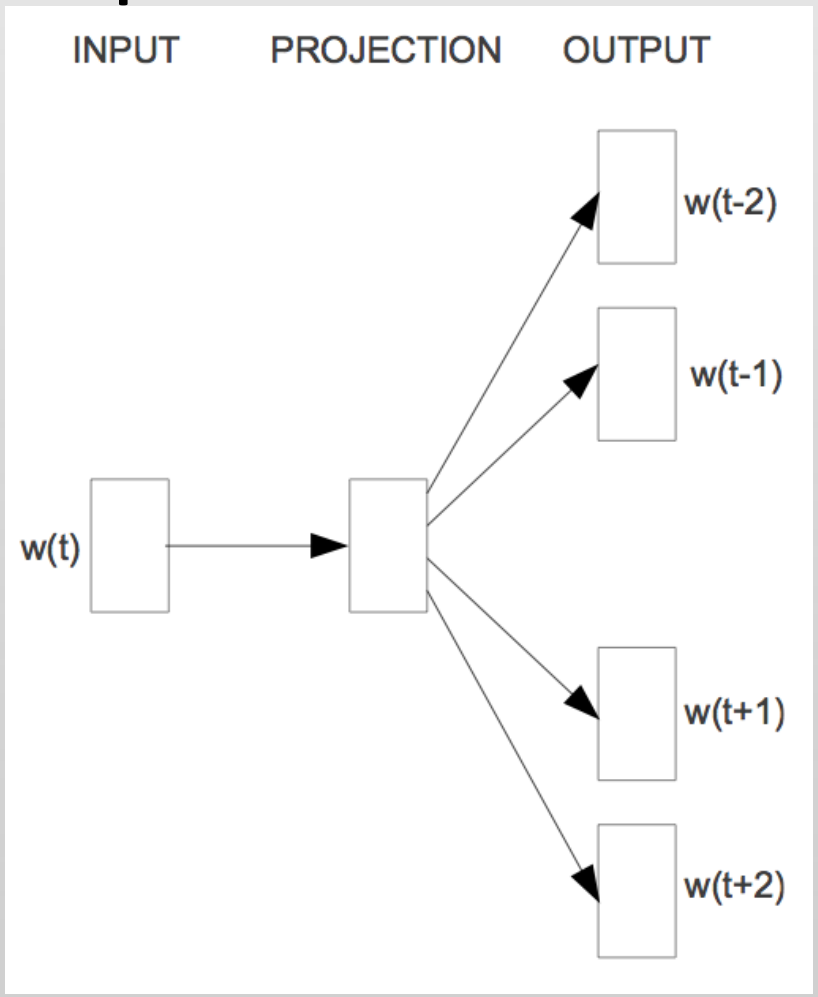
-0.4945	-0.6458	-0.5817
-0.6492	0.7194	-0.2469
-0.578	-0.2556	0.775

q = [-0.2140 0.1821]



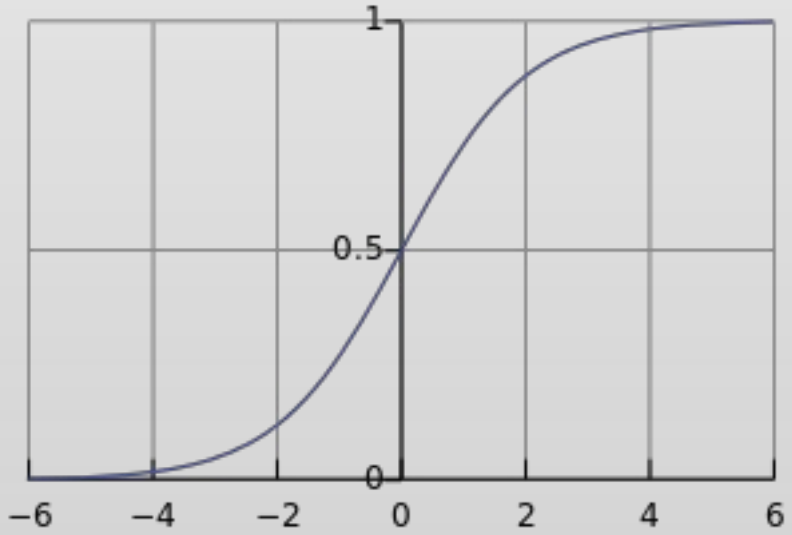
Predictive Distributional Models: CBOW vs SkipGram





Logistic Regression: Recap

$$P[Y = 1 \mid \mathbf{x}] = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$



- Optimize w to maximize log likelihood of training data.

Skipgram Model

- Dataset: `the quick brown fox jumped over the lazy dog`
- Context window:

```
([the, brown], quick), ([quick, fox], brown), ([brown, jumped], fox), ...
```

- Positive examples:

```
(quick, the), (quick, brown), (brown, quick), (brown, fox), ...
```

- Negative examples: (sheep, quick), generated at random

Context and Target Vectors

- Assign to each word *w,* a target vector $\boldsymbol{u}_w$ and a context vector $\boldsymbol{v}_w$ in $\boldsymbol{R}^d$

$$P[(w, w') \in D] = \sigma(\mathbf{u}_w^T \mathbf{v}_{w'})$$

$$P[(w, w') \notin D] = \sigma(-\mathbf{u}_w^T \mathbf{v}_{w'})$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid function

Log-likelihood Function

$$J(\{\mathbf{u}_w, \mathbf{v}_{w'}\}) = \sum_{(w',w)\in D} \log \sigma(-\mathbf{u}_w \cdot \mathbf{v}_{w'}) + \sum_{(w',w)\notin D} \log \sigma(\mathbf{u}_w \cdot \mathbf{v}_{w'})$$

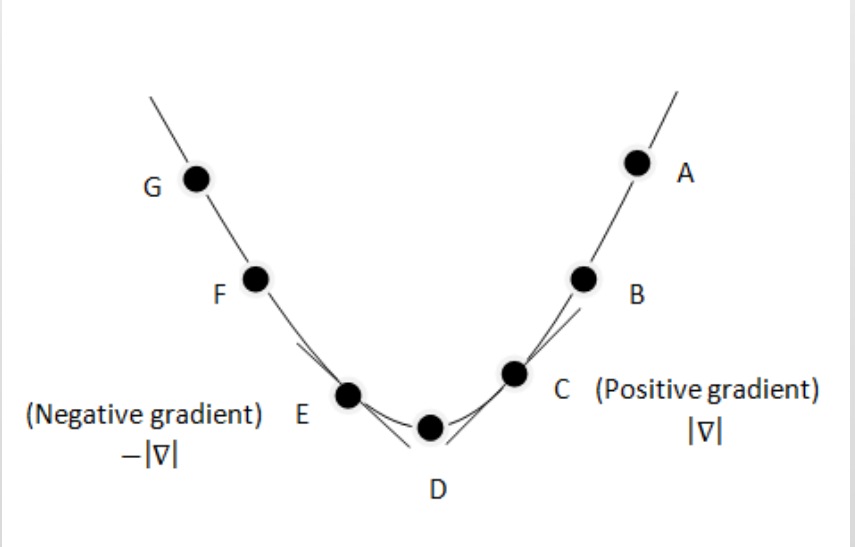
Negative Sampling: Use randomly generated pairs (w', w) in place of D'

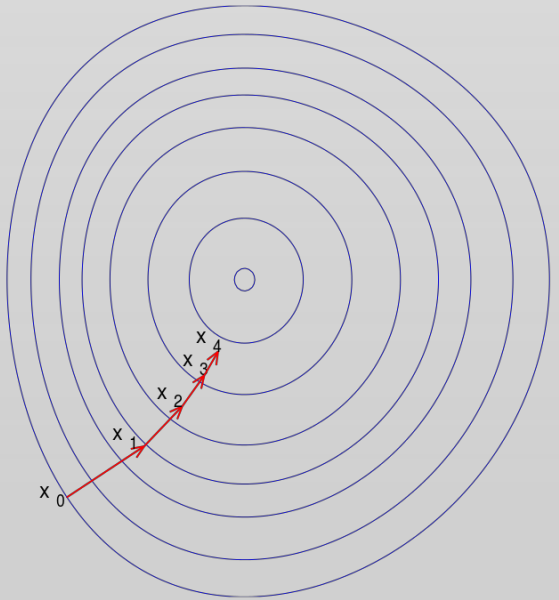
Quiz

- How do we train parameters for this likelihood function?

Gradient Descent

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \alpha_t \nabla f(\mathbf{u}_t)$$





(Stochastic) Gradient Descent

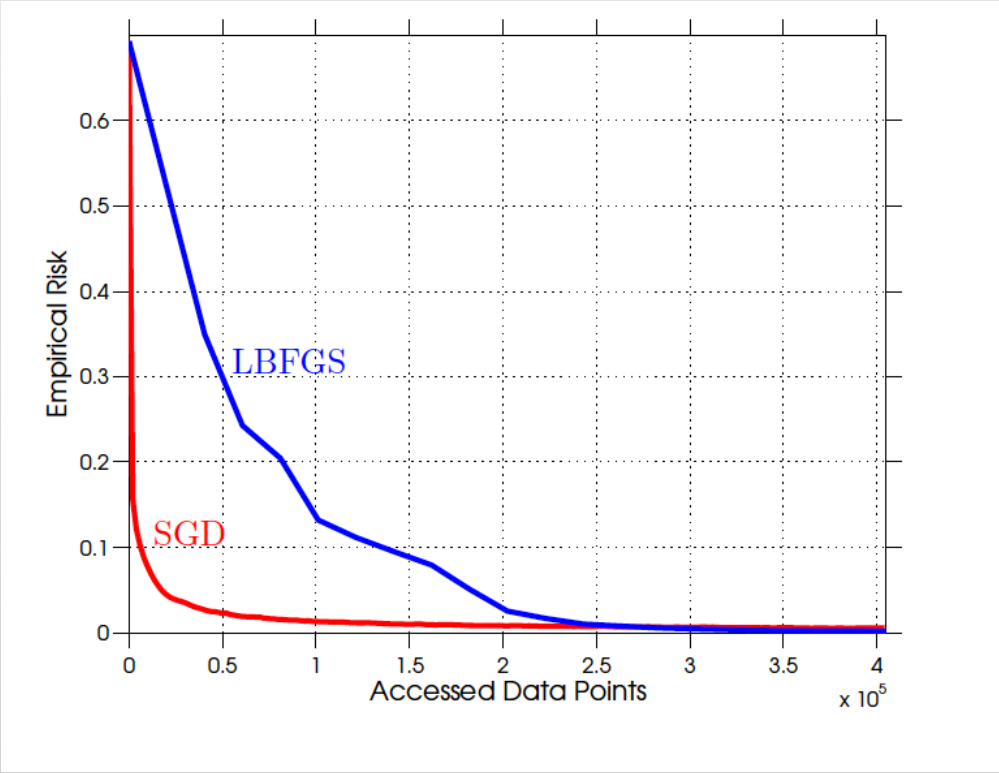
$$\mathbf{u}_{t+1} = \mathbf{u}_t - \alpha_t \nabla f(\mathbf{u}_t) \quad \mathbf{u}_{t+1} = \mathbf{u}_t - \alpha_t \nabla f_i(\mathbf{u}_t)$$

- Each iteration expensive as it needs to run through all data points
- Steady linear convergence
- Number of iterations $O(\log \frac{1}{\epsilon})$
- Total cost $O(n \log \frac{1}{\epsilon})$

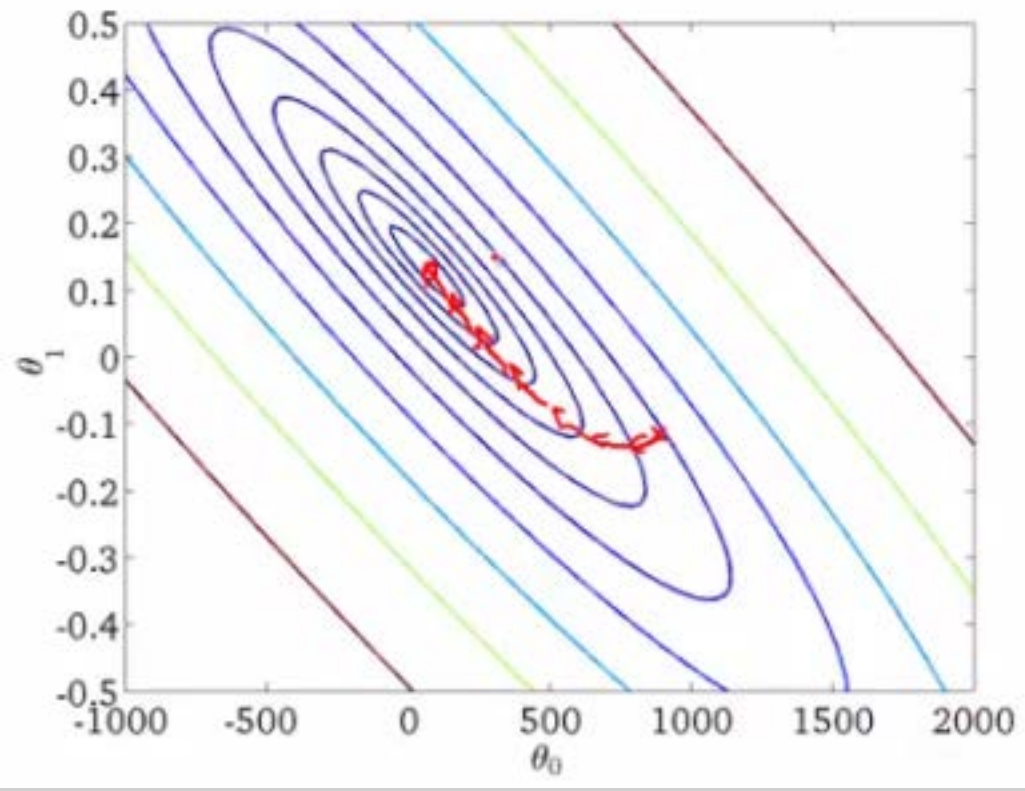
- Cheap iteration as it looks at only one data point
- Initial fast descent but slow at the end
- Number of iterations $O(\frac{1}{\epsilon})$
- Escape saddle points!
- Better suited for BigData

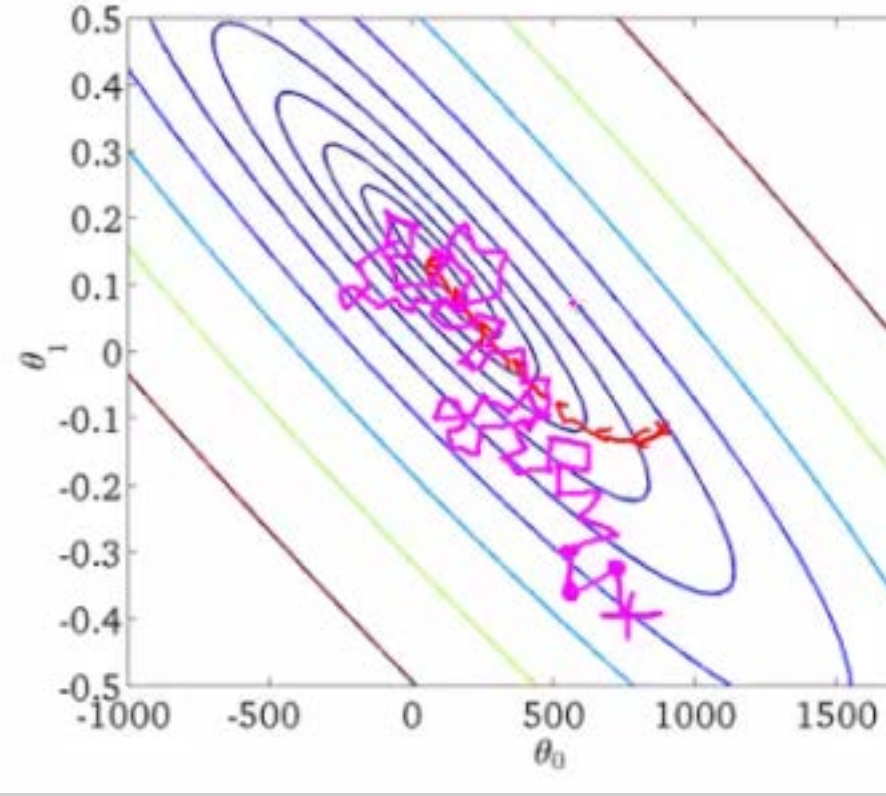
Error of SGD

- Initial fast decrease in error
- Slows down closer to optimum
- Sufficient to be close to opt or …
- … switch to deterministic variant



(Stochastic) Gradient Descent





Gradient Descent and Relatives

- Momentum
- Nesterov acceleration
- Mirror descent
- Conjugate gradient descent
- Proximal gradient descent …
- L. Bottou et al, “Optimization Methods for Large Scale Machine Learning”, 2016.

Convex vs Non-Convex

- unique global optimum
- Local opt = global opt
- Well understood: gradient descent methods guaranteed to converge to optimum, with known rates of convergence

- Complex landscape of optima
- Local opt ≠ global opt
- Gradient descent methods converge only to local opt.
- However, in practice gradient descent type methods converge to good optima

Quiz

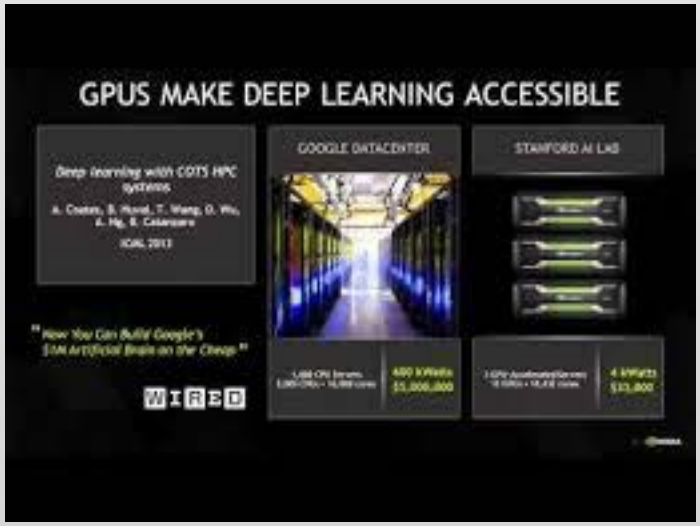
- How are neural networks trained?
- What about our objective? Is it convex?

$$J(\{\mathbf{u}_w, \mathbf{v}_{w'}\}) = \sum_{(w',w)\in D} \log \sigma(-\mathbf{u}_w \cdot \mathbf{v}_{w'}) + \sum_{(w',w)\notin D} \log \sigma(\mathbf{u}_w \cdot \mathbf{v}_{w'})$$

Gradient Descent for Non-convex

- Recent rigorous results showing that noisy/stochastic gradient descent can escape saddle points for certain classes of non-convex functions.
- R. Ge et al “Matrix Completion has no spurious local minimum”, NIPS 2016 (Best theoretical paper)
- NIPS 2016 workshop on Non-convex opt: https://sites.google.com/site/nonconvexnips2016

Why does it work well in practice?



Word2vec tutorial on TensorFlow:
https://www.tensorflow.org/tutorials/word2vec/

Why does word2vec work?

- Why are “similar” words assigned similar vectors?
- Why is

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“aunt”}) - W(\text{“uncle”})$$

$$W(\text{“woman”}) - W(\text{“man”}) \simeq W(\text{“queen”}) - W(\text{“king”})$$

word2vec as Matrix Factorization

- Levy and Goldberg (2014): word2vec can be viewed as implicit factorization of the pointwise mutual information matrix

$$PMI(w, w') = \log \frac{\#(w,w')\ D}{\#(w)\#(w')}$$

Relations = Lines

- Arora et al (2016): Posit a generative model such that for every relation R, there is a direction μ_R such that if $(a, b) \in R$ then v_a - v_b = $\alpha_{a,b}$ μ_R+ η, where η is a noise vector.

References

- Y. Goldberg and O. Levy, “word2vec Explained”, Arxiv 2014
- O. Levy, Y. Goldberg, “Neural Word Embedding as Implicit Matrix Factorization”, NIPS 2014.
- S. Ruder, “An Overview of Gradient Descent Optimization Algorithms”, Arxiv. 2016
- L. Bottou, F. Curtis and J.Nocedal, “Optimization Methods for Large Scale Machine Learning”
- S. Arora et al, “A Latent Variable Model Approach to PMI Based Word Embeddings”, TACL 2016.