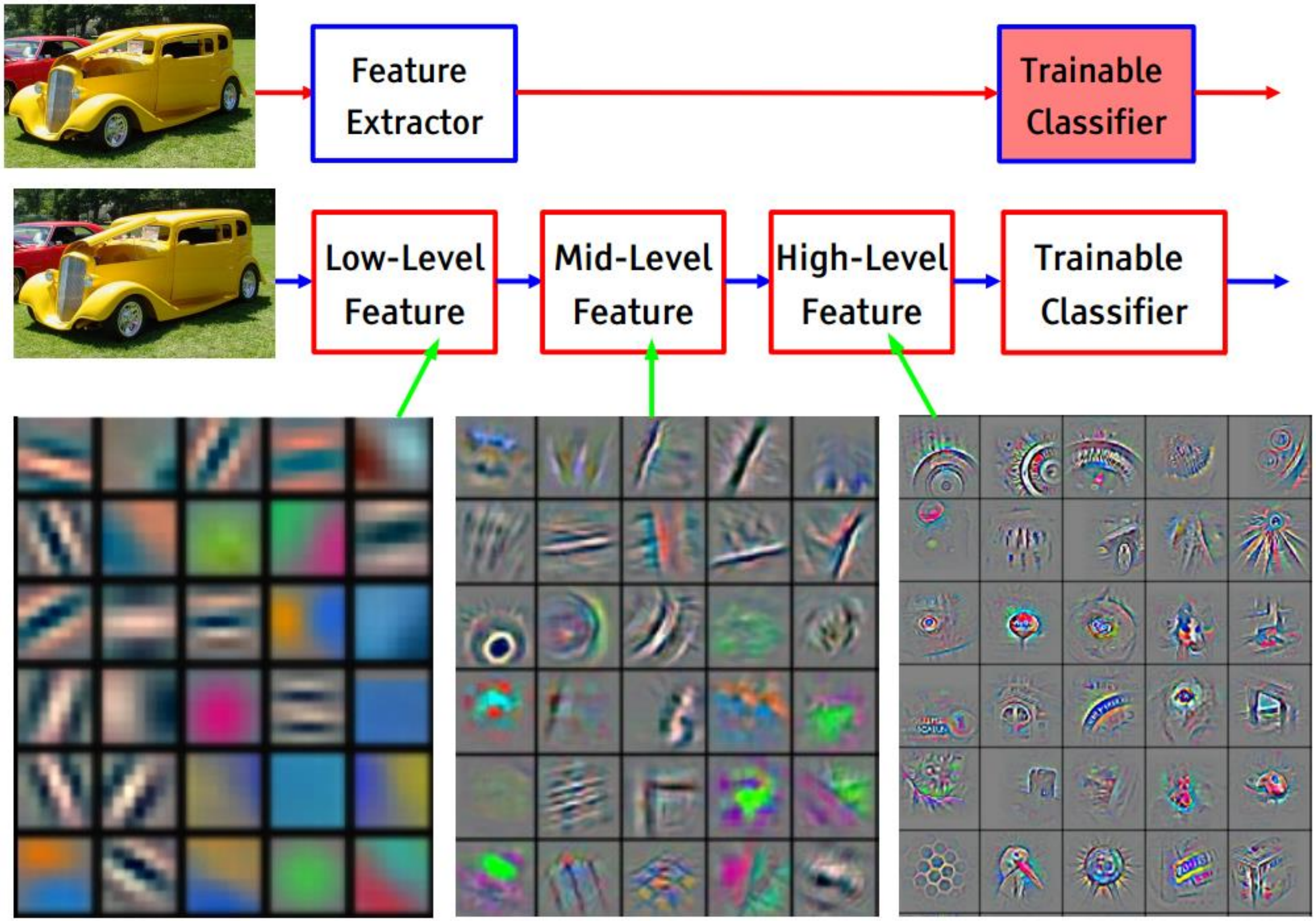
CONCLUSIONS

We reviewed machine learning methods



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

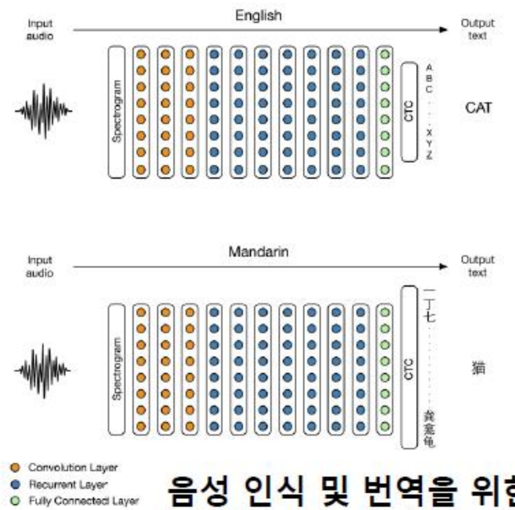
<table>
<tr><th colspan="5" rowspan="4"></th><th colspan="6">Tasks</th></tr>
<tr><th colspan="5">ADAS</th><th></th></tr>
<tr><th colspan="3">Self Driving</th><th></th><th></th><th></th></tr>
<tr><th>Localization</th><th>Perception</th><th>Planning/ Control</th><th>Driver state</th><th>Vehicle Diagnosis</th><th>Smart factory</th></tr>
<tr><td rowspan="7">Methods</td><td>Traditional</td><td colspan="3">Non-machine Learning</td><td>GPS, SLAM</td><td></td><td>Optimal control</td><td></td><td></td><td></td></tr>
<tr><td></td><td rowspan="6">Machine-Learning based method</td><td rowspan="4">Supervised</td><td>SVM
MLP</td><td></td><td>Pedestrian detection (HOG+SVM)</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="5">Deep-Learning based</td><td>CNN</td><td></td><td>Detection/ Segmentation/Classification</td><td>End-to-end Learning</td><td></td><td></td><td></td></tr>
<tr><td>RNN (LSTM)</td><td></td><td>Dry/wet road classification</td><td>End-to-end Learning</td><td>Behavior Prediction/ Driver identification</td><td></td><td>*</td></tr>
<tr><td>DNN</td><td></td><td></td><td></td><td></td><td>*</td><td>*</td></tr>
<tr><td colspan="2">Reinforcement</td><td></td><td></td><td>*</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">Unsupervised</td><td></td><td></td><td></td><td></td><td></td><td>*</td></tr>
</table>

Deep Learning?

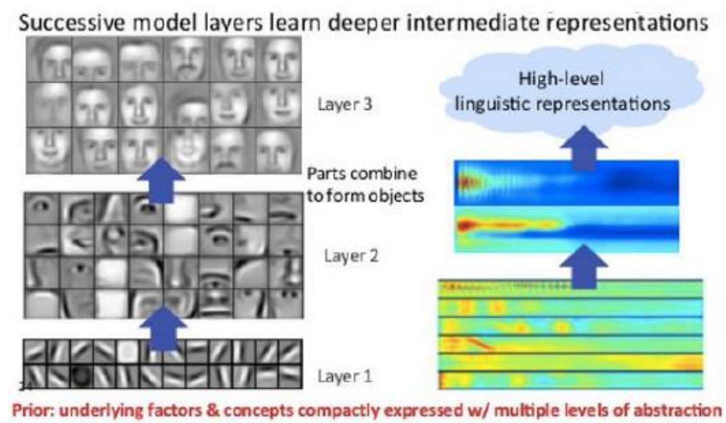
- Ranzato's definition
  - a method which makes predictions by using a sequence of non-linear processing stages. The resulting intermediate representations can be interpreted as feature hierarchies and the whole system is jointly learned from data. Some deep learning methods are probabilistic, others are loss-based, some are supervised, other unsupervised...

딥러닝 모델의 특징

- 다층구조 (multi layer)
  - 신경망의 구조 모사
  - 상위 층으로 갈 수록 추상화된 정보가 학습과정에서 자동으로 생성
- 문제 해결과정자동화
  - End-to-end learning
  - 사람의 개입을 배제하고 오직 raw input과 output 사이에 모든 과정을 데이터에서 학습하는 방향 추구
- 분산 표현
  - Distributed representation
  - 여러 뉴런이 협력하여 정보 저장/처리



음성 인식 및 번역을 위한 딥러닝 모델 사례(Baidu)



BACKUPS

WORD EMBEDDING

http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

Word representation

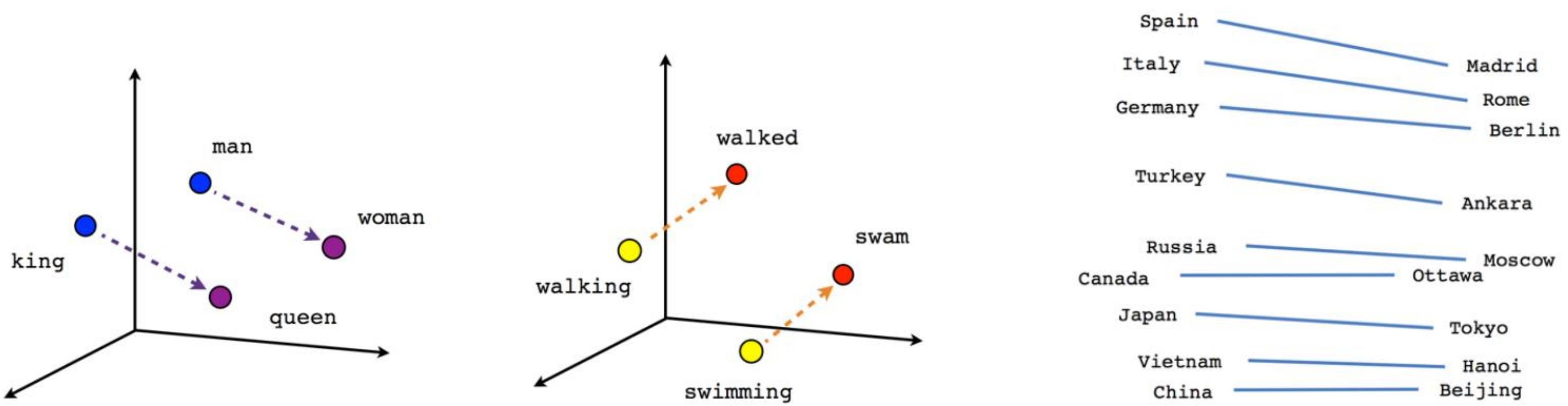
Traditional method

- Uses one hot encoding
- Each word in the vocabulary is represented by one bit position in a huge vector.
- Context information is not utilized

Word embeddings

- Stores each word in as a point in space, where it is represented by a vector of fixed number of dimensions (generally 300)
- Unsupervised, built just by reading huge corpus

Examples



Male-Female

Verb tense

Country-Capital

Demo (http://w.elnn.kr/search/)

Word Embedding

- A word embedding W:words$\rightarrow\Re^n$ is a paramaterized function mapping words in some language to high-dimensional vectors (perhaps 200 to 500 dimensions). For example, we might find:

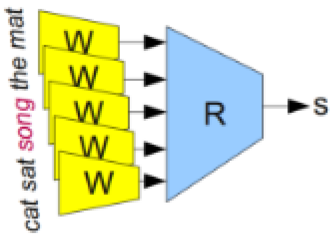
$$W(\text{“cat”}) = (0.2,\ \text{-}0.4,\ 0.7,\ \ldots)$$

$$W(\text{“mat”}) = (0.0,\ 0.6,\ \text{-}0.1,\ \ldots)$$

- Typically, the function is a lookup table, parameterized by a matrix, θ, with a row for each word: $W_\theta(w_n) = \theta_n$

Word Embedding Learning

- W is initialized to have random vectors for each word. It learns to have meaningful vectors in order to perform some task.
- Train a network for is predicting whether a 5-gram (sequence of five words) is 'valid.'
  - "cat sat on the mat" vs "cat sat **song** the mat"
- 5-gram -> (W, R) -> 'valid' vs 'broken'



$$R(W(\text{“cat”}),\ W(\text{“sat”}),\ W(\text{“on”}),\ W(\text{“the”}),\ W(\text{“mat”})) = 1$$

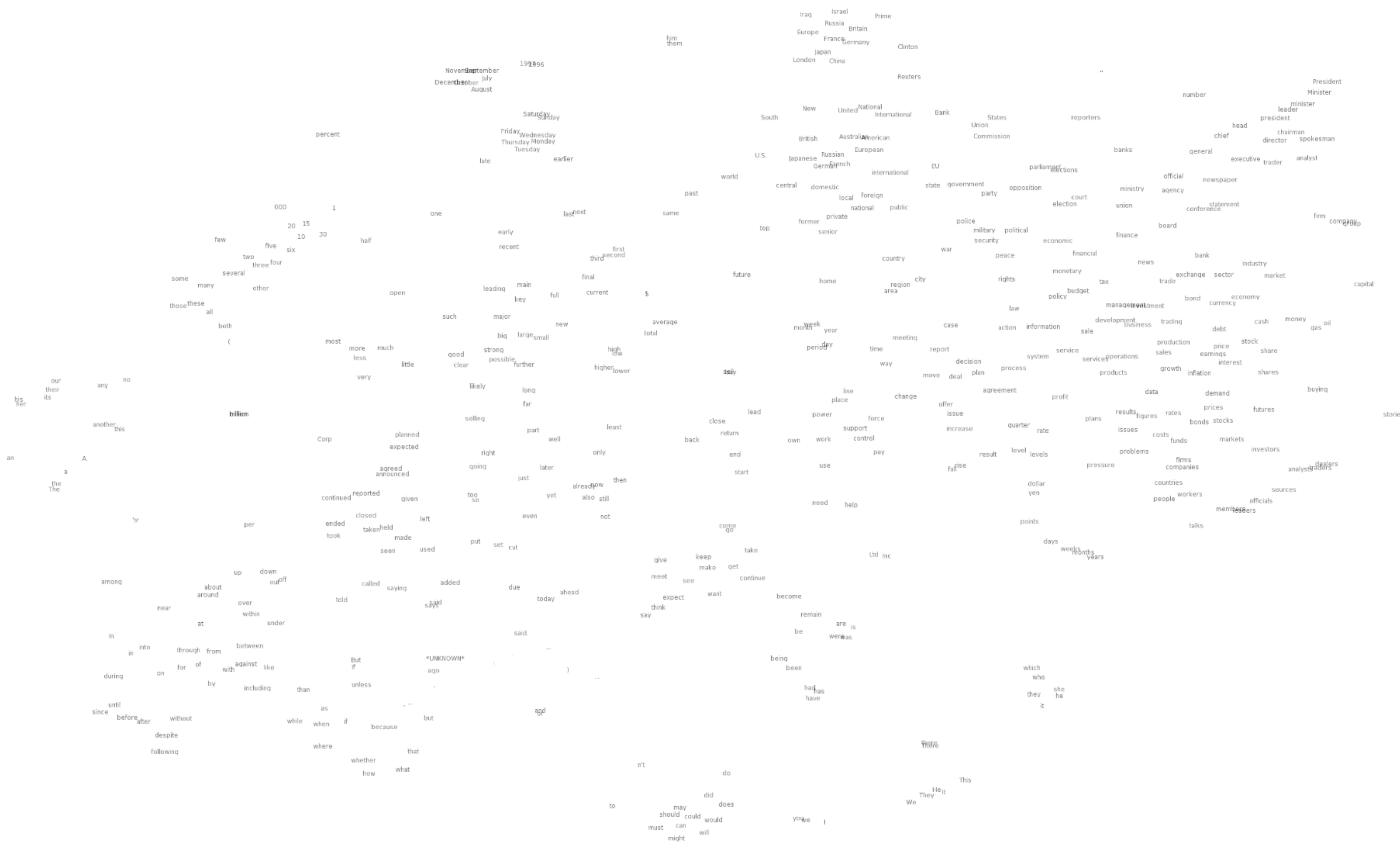
$$R(W(\text{“cat”}),\ W(\text{“sat”}),\ W(\text{“song”}),\ W(\text{“the”}),\ W(\text{“mat”})) = 0$$

Modular Network to determine if a 5-gram is 'valid' (From Bottou (2011))

Word Embedding Learning

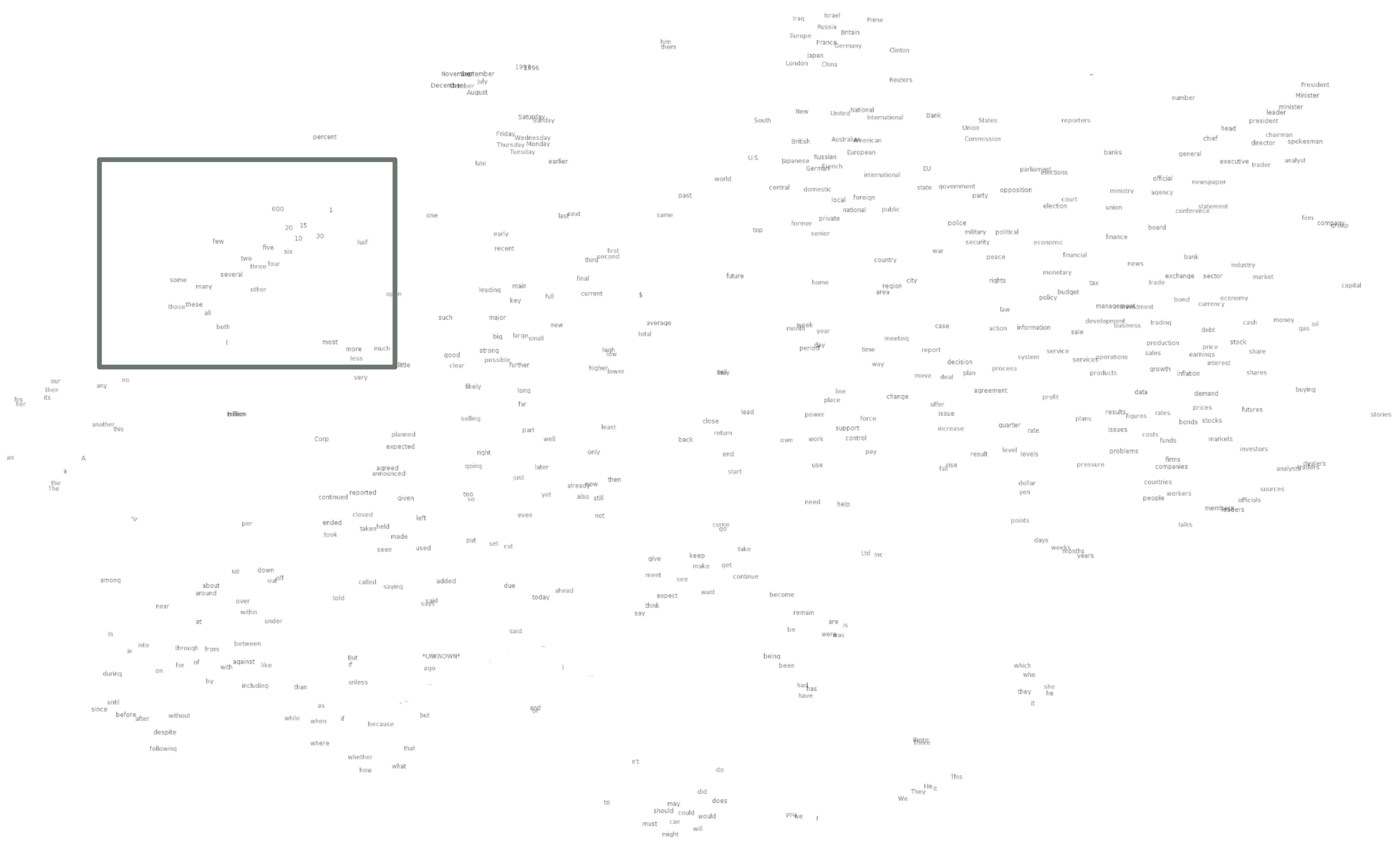
- In order to predict these values accurately, the network needs to learn good parameters for both W and R.

t-SNE Visualization

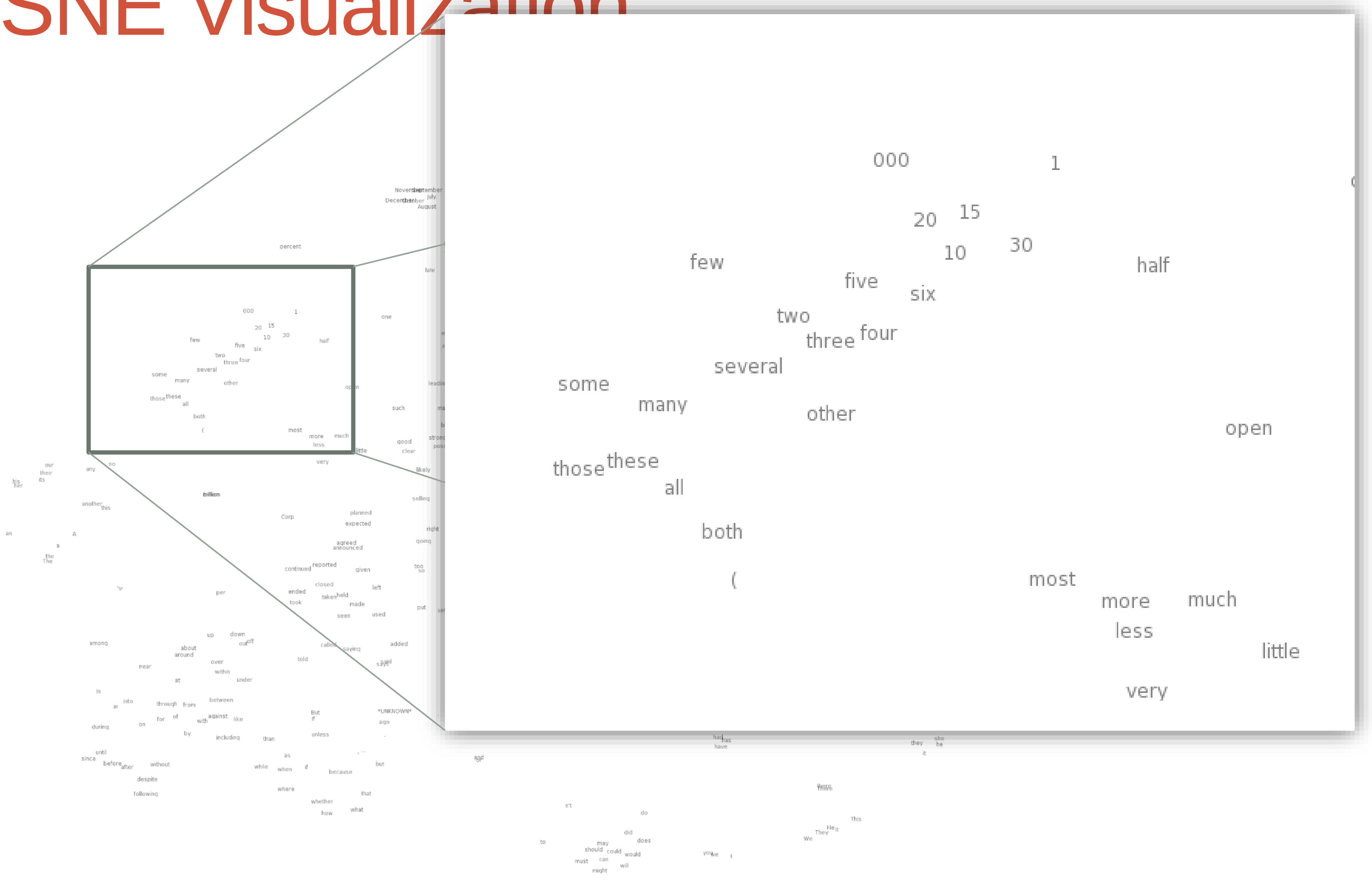


http://metaoptimize.s3.amazonaws.com/cw-embeddings-ACL2010/embeddings-mostcommon.EMBEDDING_SIZE=50.png

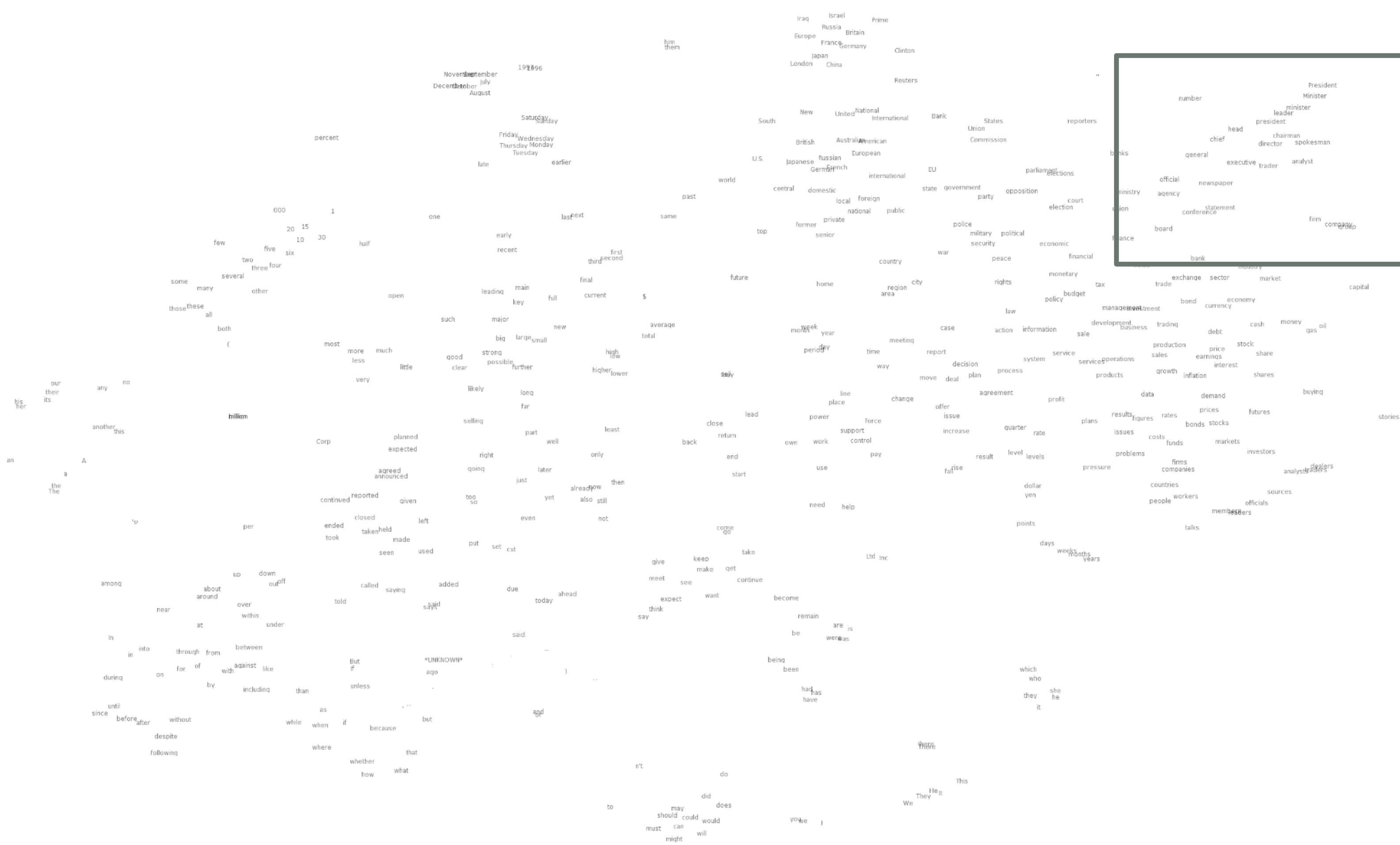
t-SNE Visualization



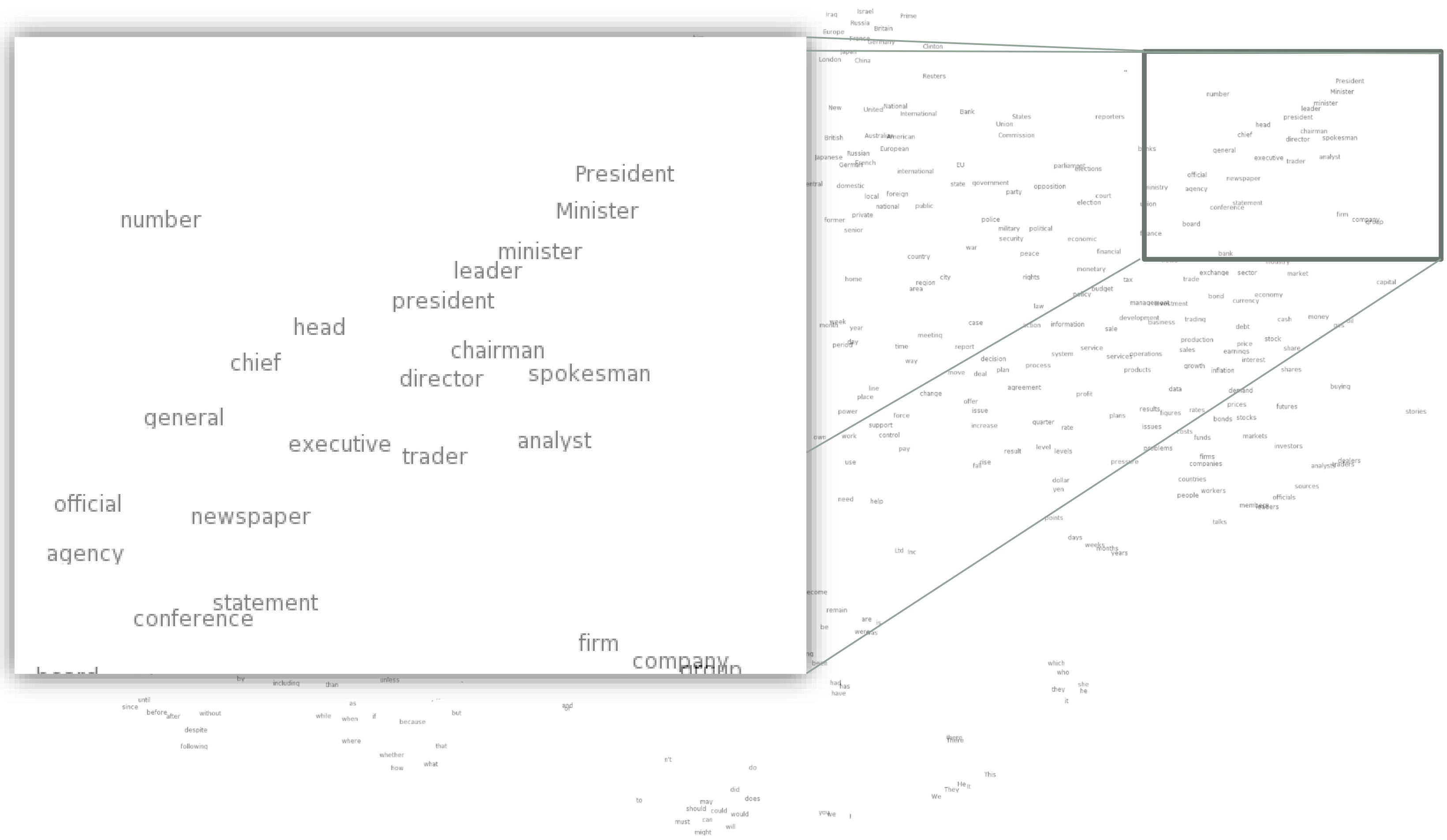
t-SNE Visualization



t-SNE Visualization



t-SNE Visualization



What words have embeddings closest to a given word?

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

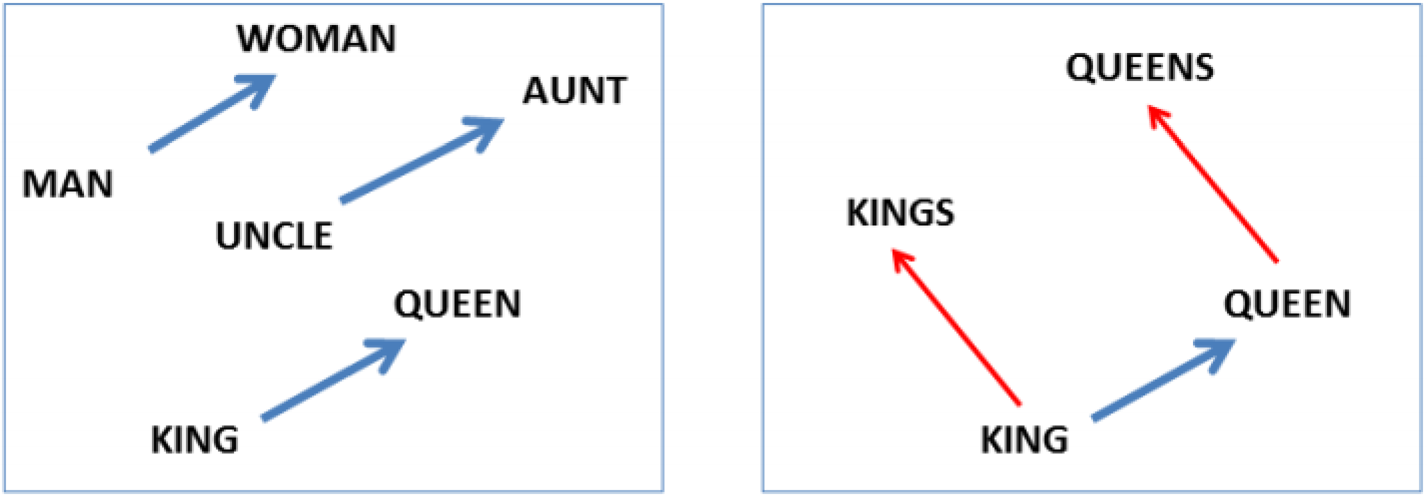
Gender dimension?

- Word embeddings exhibit an even more remarkable property: analogies between words seem to be encoded in the difference vectors between words. For example, there seems to be a constant male-female difference vector:

$$W(\text{``woman''}) - W(\text{``man''}) \simeq W(\text{``aunt''}) - W(\text{``uncle''})$$
$$W(\text{``woman''}) - W(\text{``man''}) \simeq W(\text{``queen''}) - W(\text{``king''})$$

- We say with hindsight, “the word embedding will learn to encode gender in a consistent way. In fact, there’s probably a gender dimension. Same thing for singular vs plural.

Examples



(Mikolov et al., NAACL HLT, 2013)

Much more sophisticated relationships

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

Summary

- All of these properties of W are side effects.
  - We didn't try to have similar words be close together.
  - We didn't try to have analogies encoded with difference vectors.
  - All we tried to do was perform a simple task, like predicting whether a sentence was valid. These properties more or less popped out of the optimization process.

- This seems to be a great strength of neural networks
  - They learn better ways to represent data, automatically.
  - Representing data well, in turn, seems to be essential to success at many machine learning problems.
  - Word embeddings are just a particularly striking example of learning a representation.