Introduction to Text Mining

Text and Web Mining (IS6751)

School of Communication and Information

What is Text Mining?

- Is finding **interesting regularities** in large **textual** datasets.
  - Where **interesting** means non-trivial, hidden, previously unknown and potentially useful.
    有趣的含义是非平凡的、隐藏的、以前未知的和潜在有用的。
  - E.g., extract **relations** between drugs and diseases.
- Is finding semantic and abstract information from the surface form of textual data.
  - E.g., predict **sentiments** towards products 预测对产品的情绪
- The International Data Corporation estimated that approximately **80%** of the data in an organization is **text based**.
- Text Mining is also called **Text Analytics**.

Which areas are related to Text Mining?

- Data Mining
  - Structured Data Analysis
- Machine Learning
  - Data Analysis algorithms
- Natural Language Processing
  - Computational Linguistics 计算语言学
- Knowledge Management
  - Knowledge Representation and Reasoning (e.g., *born-in(Albert Einstein, Ulm Germany)*)
  - Used in Question & Answering systems
- Information Retrieval
  - Full-text indexing
  - Search in natural language

What is Natural Language Processing (NLP)?

- **Natural language processing** is a field at the intersection of
  - computer and information science
  - artificial intelligence
  - and linguistics.
- **Goal:** for computers to process or “understand” natural language in order to perform tasks that are useful, e.g.,
  - Performing Tasks, like making appointments, buying things
  - Language translation
  - Question Answering
    - Siri, Google Assistant, ChatGPT, etc.
- Fully **understanding and representing** the **meaning** of language (or even defining it) is a difficult goal.

NLP/Text Mining Applications

应用范围从简单到复杂:
- 查找同义词
- 从网站中提取信息，例如
- 产品价格、日期、位置、人员或公司名称
- 分类：学校课本的阅读水平、营销情绪分析
- 机器翻译
- 语音识别
- 口语对话系统：自动化客户支持
- 复杂问答

Applications range from simple to complex:

- Finding synonyms
- Extracting information from websites such as
  - product price, dates, location, people or company names
- Classifying: reading level of school texts, Sentiment analysis for marketing
- Machine translation
- Speech recognition
- Spoken dialog systems: automating customer support
- Complex question answering

Natural Language Processing Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG		LOC
Einstein	met	with	UN	officials in	Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco! 👍

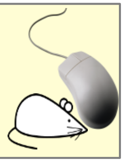
The waiter ignored us for 20 minutes. 👎

Coreference resolution 共引用分辨率

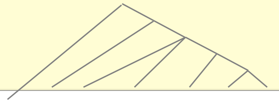
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD) 词义消歧义

I need new batteries for my ***mouse***.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

➡

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still hard but making good progress

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

He needed to make a quick decision in that situation.
The scenario required him to make a split-second judgment.

Summarization

The Dow Jones is up
The S&P500 jumped
Housing prices rose

➡ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Why is NLP hard?

- Complexity in representing, learning and using linguistic/situational/contextual/world/visual knowledge

语言/情境/语境/世界/视觉知识的表达、学习和使用的复杂性

- But interpretation depends on these
- Human languages are ambiguous (unlike programming and other formal languages)
- E.g., “I made her duck.”

a) I cooked waterfowl for her.
b) I cooked waterfowl belonging to her.
c) I created the (toy or sculpture?) duck she owns.
雕像
d) I caused her to quickly lower her head or body.
e) I turned her into waterfowl (using my magic wand?).

E）我把她变成了水禽（用我的魔杖？）。

Ambiguity makes NLP (or Text Mining) hard

模棱两可

- Violinist Linked to JAL Crash Blossoms
- Teacher Strikes Idle Kids
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors
- Juvenile Court to Try Shooting Defendant
- Drunk Gets Nine Months in Violin Case
- The Pope’s baby steps on gays

- 小提琴手涉嫌与日航坠机事件有关
- 教师罢免懒惰的孩子
- 官僚作风阻碍新桥梁
- 7英尺高的医生起诉医院
- 少年法庭将审理枪击被告
- 小提琴案中醉酒者被判9个月监禁
- 教皇对同性恋采取小步骤

Words Properties makes NLP (or Text Mining) hard

- Various relations among word surface forms and their senses:
  - **Homonymy**: same form, but different meaning
    - E.g., bank: river bank and financial institution
  - **Polysemy**: same form, related meaning
    - E.g., man: the human species, male of the human species, and adult males of the human species
  - **Synonymy**: different form, same meaning
    - E.g., singer and vocalist

- 同音异义：形式相同，但含义不同
  - 例如，bank：河岸和金融机构
- 多义：形式相同，含义相关
  - 例如，man：人类、人类男性和人类成年男性
- 同义词：形式不同，含义相同
  - 例如，singer 和声乐家

- **Word-sense disambiguation** is an open problem concerned with identifying which sense of a word is used in a sentence.
  - bank: river bank vs. financial institution

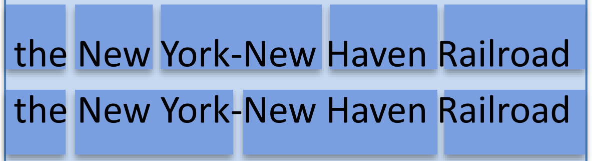
•词义消歧义是一个开放的问题，涉及识别句子中使用的单词的含义。

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues



the New York-New Haven Railroad

the New York-New Haven Railroad

idioms 成语

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

棘手的实体名称

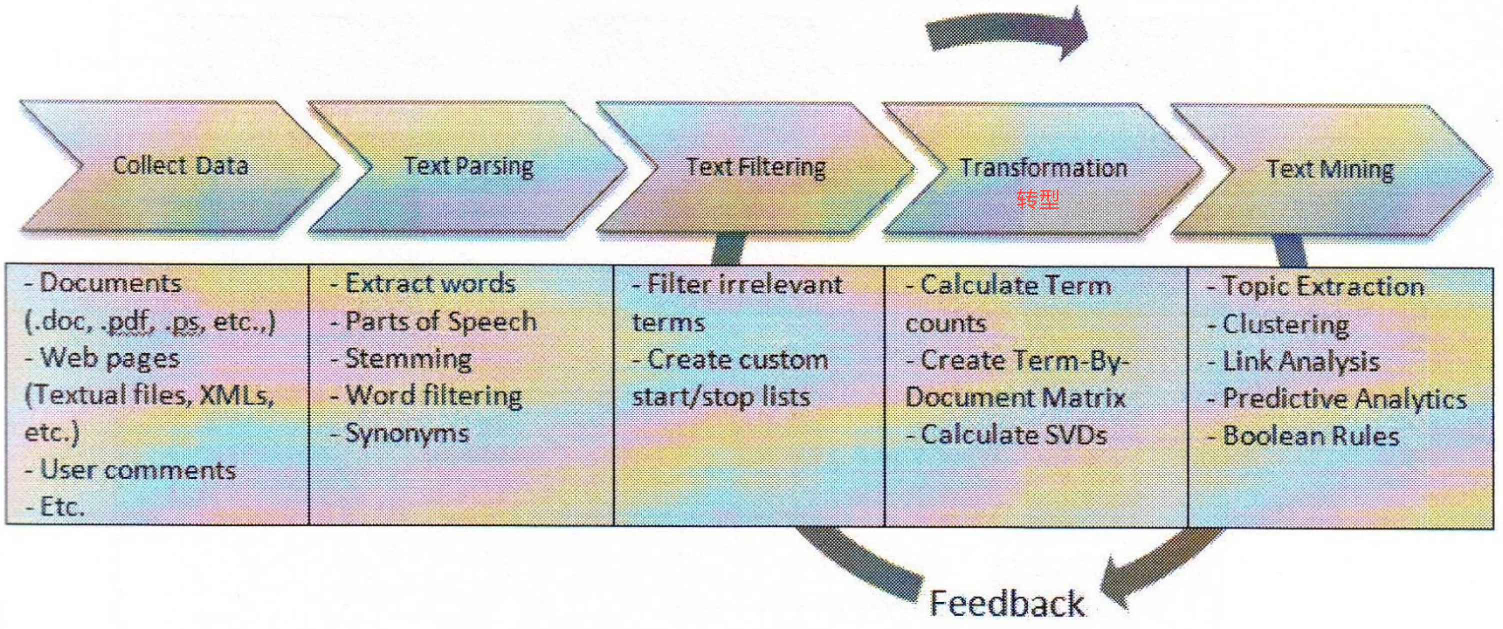
Where is *A Bug's Life* playing …

Let It Be was recorded …

Text Mining Process Flow

A traditional text mining project involves 5 steps.

Display 1.3: Text Mining Process Flow



Collect Data	Text Parsing	Text Filtering	Transformation 转型	Text Mining
- Documents (.doc, .pdf, .ps, etc.,) - Web pages (Textual files, XMLs, etc.) - User comments - Etc.	- Extract words - Parts of Speech - Stemming - Word filtering - Synonyms	- Filter irrelevant terms - Create custom start/stop lists	- Calculate Term counts - Create Term-By-Document Matrix - Calculate SVDs	- Topic Extraction - Clustering - Link Analysis - Predictive Analytics - Boolean Rules

Stemming: e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.
词干提取

Structured or Unstructured Data?

- The **text** is usually a collection of **unstructured documents** with no special requirements for composing the documents.
- In **data mining** applications, the data must be prepared in a very special way (e.g., a spreadsheet format) before any learning methods can be applied.
  - Two types of information are expected: (a) ordered numerical and (b) categorical.

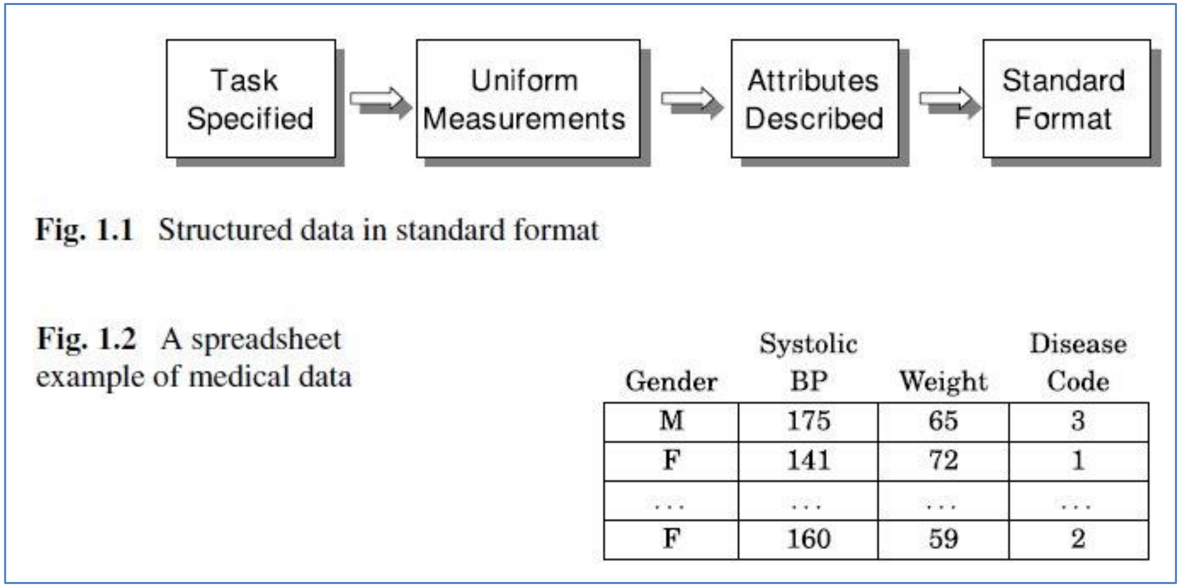


Fig. 1.1 Structured data in standard format

Fig. 1.2 A spreadsheet example of medical data

Gender	Systolic BP	Weight	Disease Code
M	175	65	3
F	141	72	1
...	...	...	...
F	160	59	2

Is Text Different from Numbers?

- One of the main themes supporting text mining is **the transformation of text into numerical data**.
- Although the initial presentation is document format, the data move into a classical data-mining encoding, a spreadsheet format.
- The unstructured data become **structured**.
- Each row represents a document (called a document vector) and each column a word.

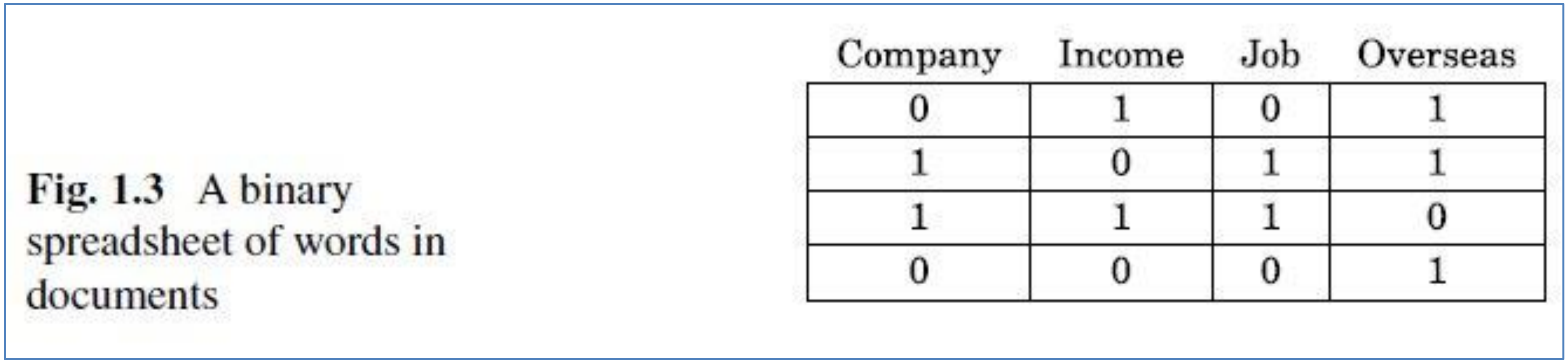
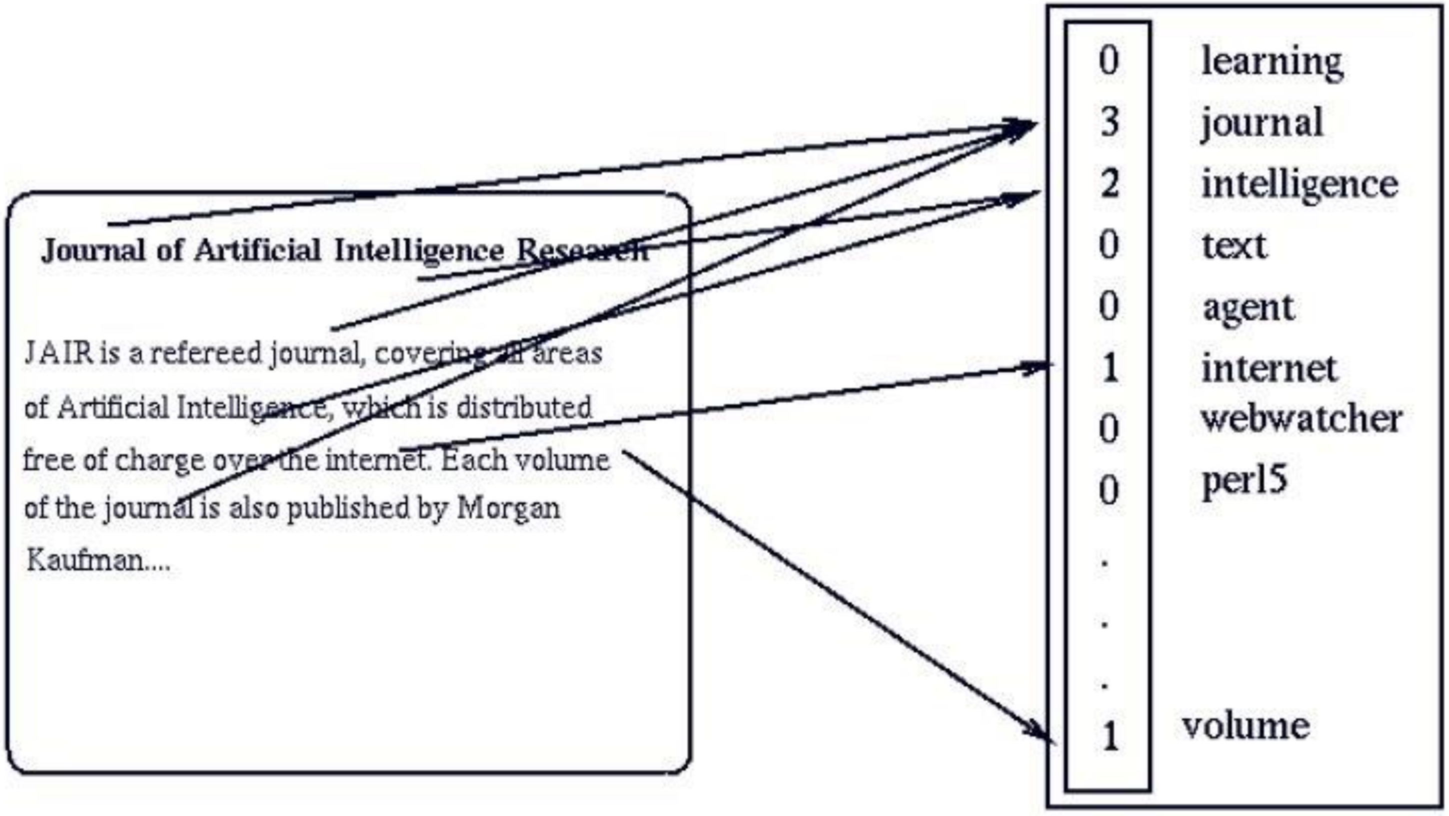


Fig. 1.3 A binary spreadsheet of words in documents

Company	Income	Job	Overseas
0	1	0	1
1	0	1	1
1	1	1	0
0	0	0	1

Bag-of-words document representation 文字包文档表示

A document vector representation of an example document.



Journal of Artificial Intelligence Research
JAIR is a refereed journal, covering all areas
of Artificial Intelligence, which is distributed
free of charge over the internet. Each volume
of the journal is also published by Morgan
Kaufman....
0 learning
3 journal
2 intelligence
0 text
0 agent
1 internet
0 webwatcher
0 perl5
.
.
.
1 volume

Unique words in the collection determine the size of document vectors.

Vector Generation for Prediction

- Document features
  - Presence (0 or 1)
  - Frequencies (0, 1, 2, 3, …..)
  - Thresholding frequencies – three values
    - 0, 1 (occurred once), and 2 (occurred 2 or more times)

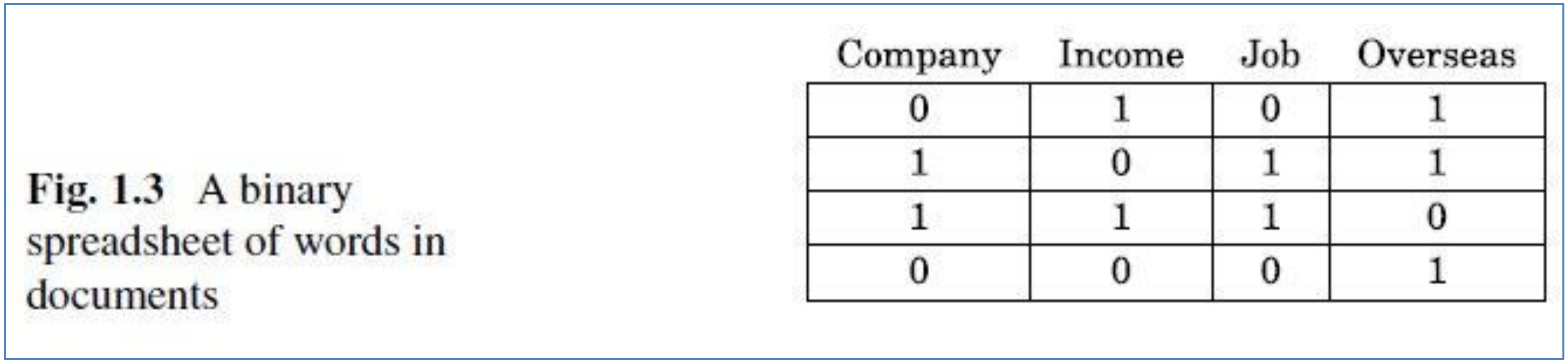


Fig. 1.3 A binary spreadsheet of words in documents

Company	Income	Job	Overseas
0	1	0	1
1	0	1	1
1	1	1	0
0	0	0	1

Vector Generation for Prediction

When *w* appears in 1 doc out of 100 documents.:

$$\log\left(\frac{1+100}{1+1}\right)+1=1.73+1$$

When *w* appears in 100 docs out of 100 documents.:

$$\log\left(\frac{1+100}{1+100}\right)+1=1$$

- Document features (cont.)
- **tf-idf (w)**

$$\Rightarrow tf(w) * idf(w),\ \text{where } idf(w)=\log\left[\frac{(1+N)}{(1+df(w))}\right]+1$$

- The tf-idf weight assigned to word *w* is ***the term frequency*** (i.e., the word count) modified by a scale factor for the importance of the word. 分配给单词w的tf-idf权重是术语频率（即单词数），由单词重要性的尺度因子修改。
- The scale factor is called ***the inverse document frequency***, which checks the number of documents containing word *w* (i.e., *df(w)*) and reverses the scaling. 缩放系数称为反向文档频率，它检查包含word w（即df（w））的文档数量，并反转缩放。
- Thus, when a word appears in many documents, it is considered unimportant and the scale is lowered, perhaps near one, e.g., the, I, on, document, etc.

因此，当一个单词出现在许多文档中时，它被认为是不重要的，规模被降低，可能接近一个，例如，the、I、on、document等。

The matrix is sparse

- An individual document will use only a tiny subset of the potential set of words in **a dictionary**, which is the total set of unique words in the collection.

  单个文档将仅使用字典中潜在单词集的一小部分，即集合中一组独特的单词。
- Text mining methods mostly concentrate on **positive matches**, not worrying whether other words are absent from a document.

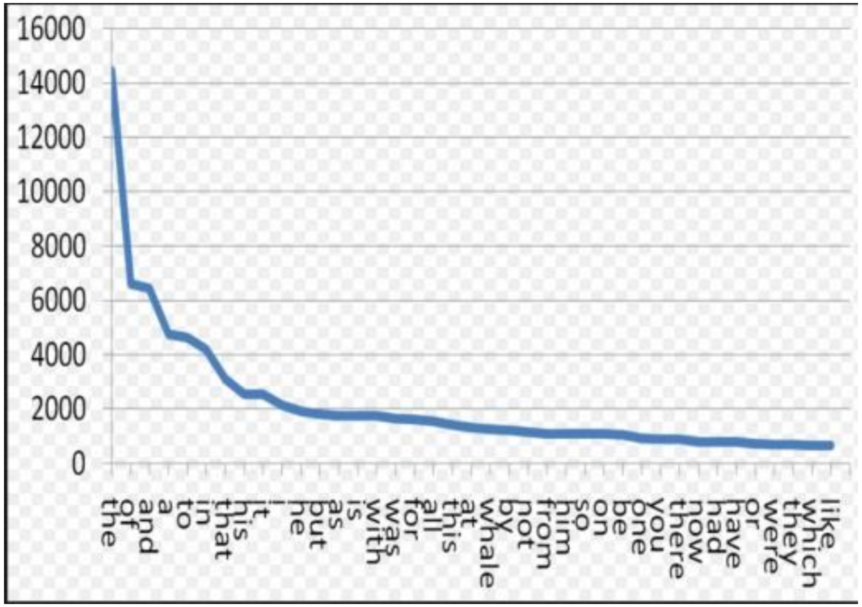
  文本挖掘方法大多专注于正匹配，而不担心文档中是否缺少其他单词。
- For text, **missing values** are a nonissue: words are either present or absent from a document.

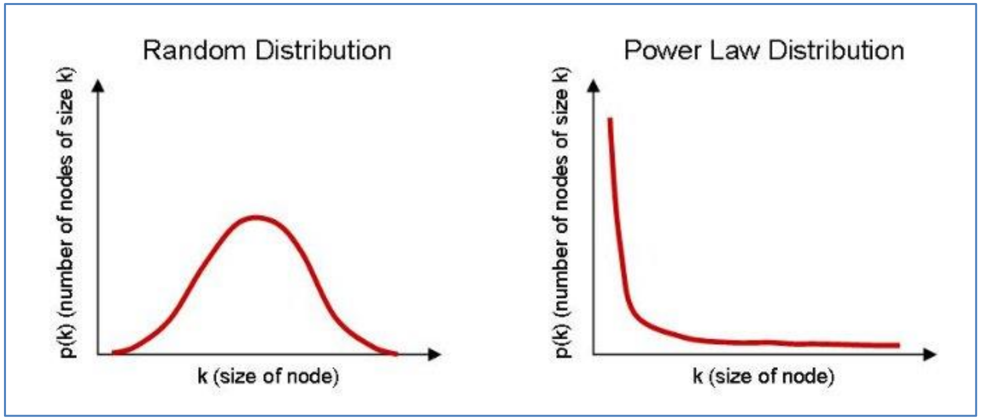
  对于文本来说，缺少值不是问题：文档中的单词要么存在，要么不存在。
- Feature reduction techniques
  - Local dictionary, removing Stopwords, Frequent words, Feature selection, and Token reduction (stemming and synonyms)

DocID	Apple	Bear	Durian	…	…	…	…	…	Zoo	Animal?
1	0	1	0	0	0	0	0	0	1	1
2	1	0	1	0	0	0	0	0	0	0
…										

Words Properties

- Word frequencies in texts have **power law distribution**:
  - small number of very frequent words
  - big number of low frequency words.
  - Also called Zipf’s Law
  - Feature dimension becomes big, and have sparse matrix





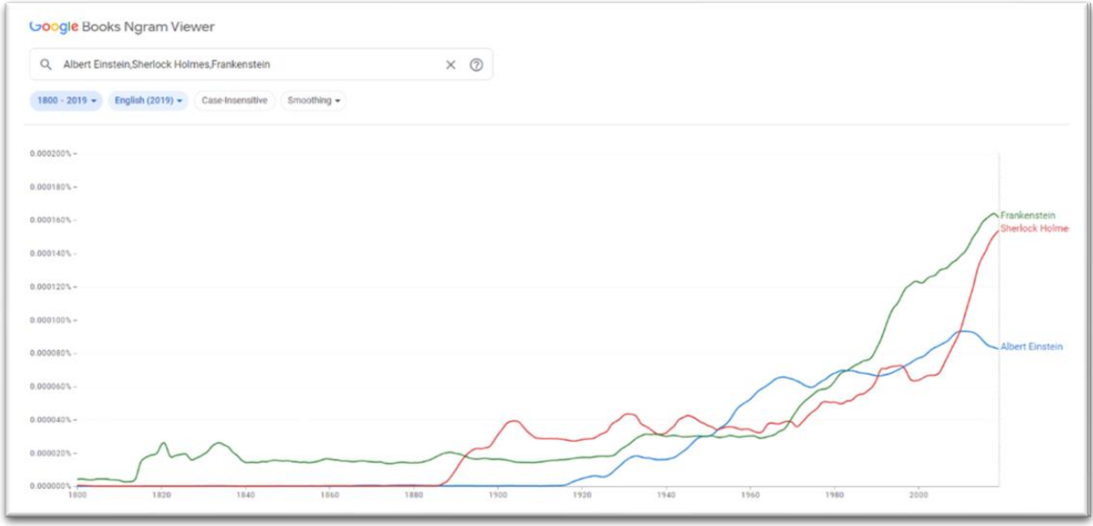
How many words?

N = number of tokens

V = vocabulary = set of types

词汇

|*V*| is the size of the vocabulary



Google Books Ngram Viewer

Albert Einstein,Sherlock Holmes,Frankenstein

1800 - 2019 · English (2019) · Case-Insensitive · Smoothing

https://books.google.com/ngrams

	Tokens = N	Types = |V|
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

E.g. “machine learning techniques”

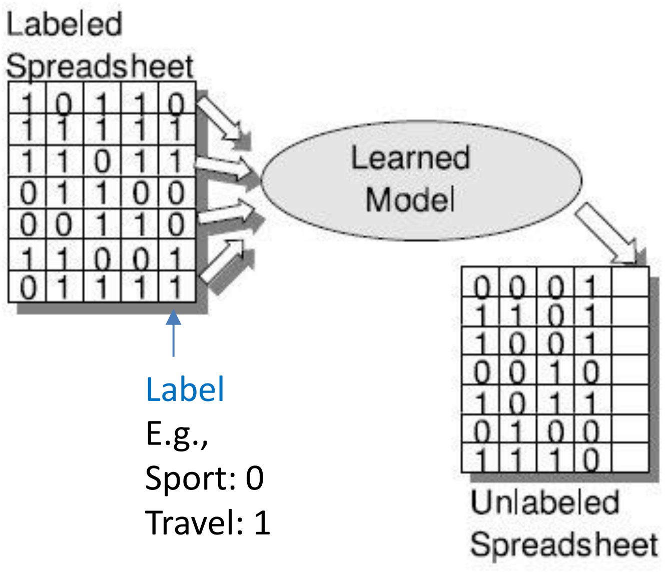
- 1-gram: machine, learning, techniques
- 2-gram: machine learning, learning techniques
- 3-gram: machine learning techniques

The **Google Books Ngram Viewer** is an online viewer, based on **Google Books**, that charts frequencies of any word using yearly count of *n*-grams found in the sources printed between 1800 and 2012 in American English, British English, French, German, Spanish, Russian, Hebrew, and Chinese. The corpora used for the search are composed of total counts, **1-grams, 2-grams, 3-grams, 4-grams, and 5-grams** files for each language.

Document Classification

- In mathematical terms, a solution is a function that maps examples to labels, ***f* : *w* → *L***, where *w* is a vector of attributes and *L* is a label.
- In our case, the attributes are words or tokens.
- The labels can be a goal that is potentially related to the words.

Training Data



1	0	1	1	0
1	1	1	1	1
1	1	0	1	1
0	1	1	0	0
0	0	1	1	0
1	1	0	0	1
0	1	1	1	1

0	0	0	1	
1	1	0	1	
1	0	0	1	
0	0	1	0	
1	0	1	1	
0	1	0	0	
1	1	1	0	

Label
E.g.,
Sport: 0
Travel: 1

Test Data

Fig. 3.6 Learning and applying models

Learning to Predict from Text

- Popular algorithms for text categorization:
  - Naive Bayesian classifier
  - Support Vector Machines
  - Logistic Regression
  - K-Nearest Neighbour
  - Deep Learning Algorithms, such as RNN, LSTM, and BERT
    - We will focus on DL algorithms.

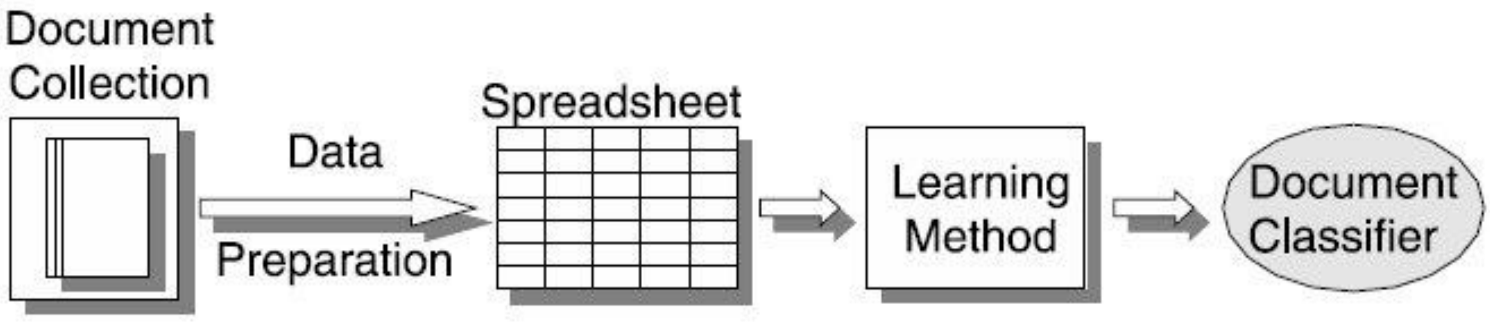


Fig. 3.7 From text to classifiers

Deep Learning, Machine Learning, and AI

- A **Venn diagram** showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI.

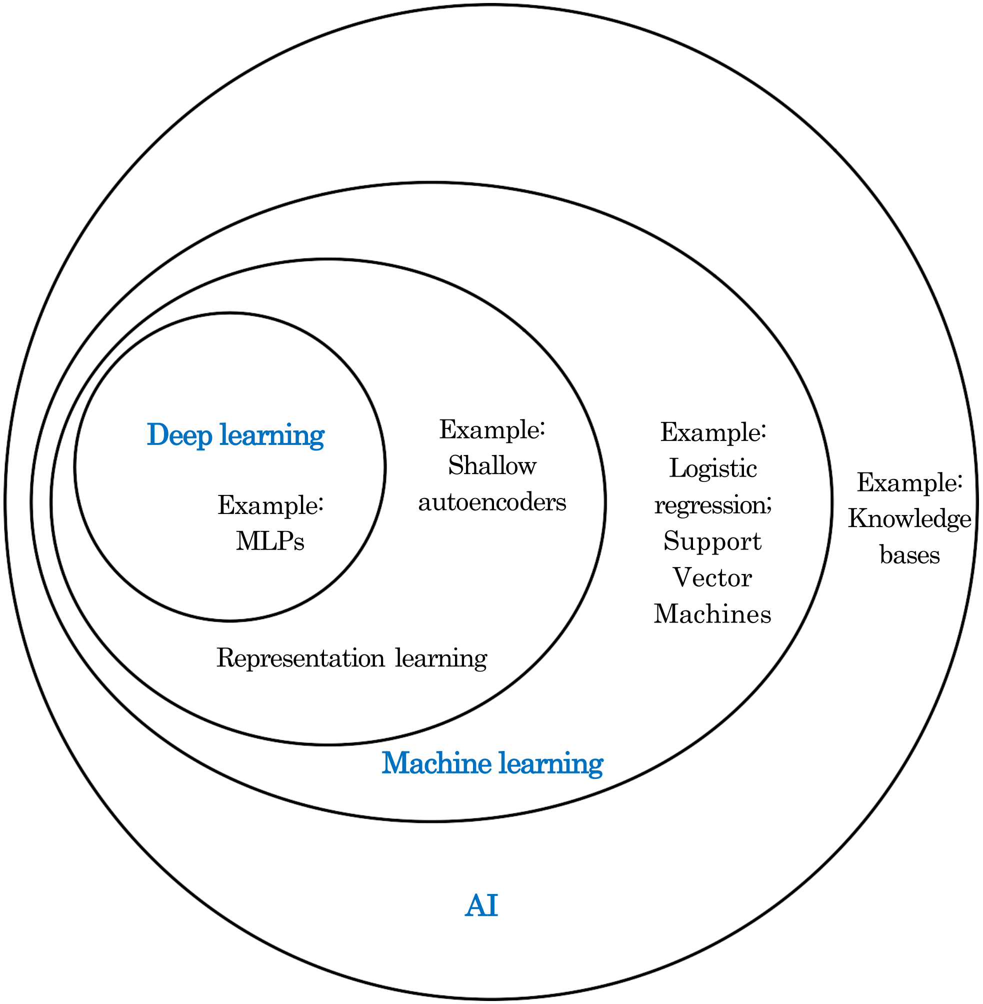


Figure 1.4

(Goodfellow 2016)

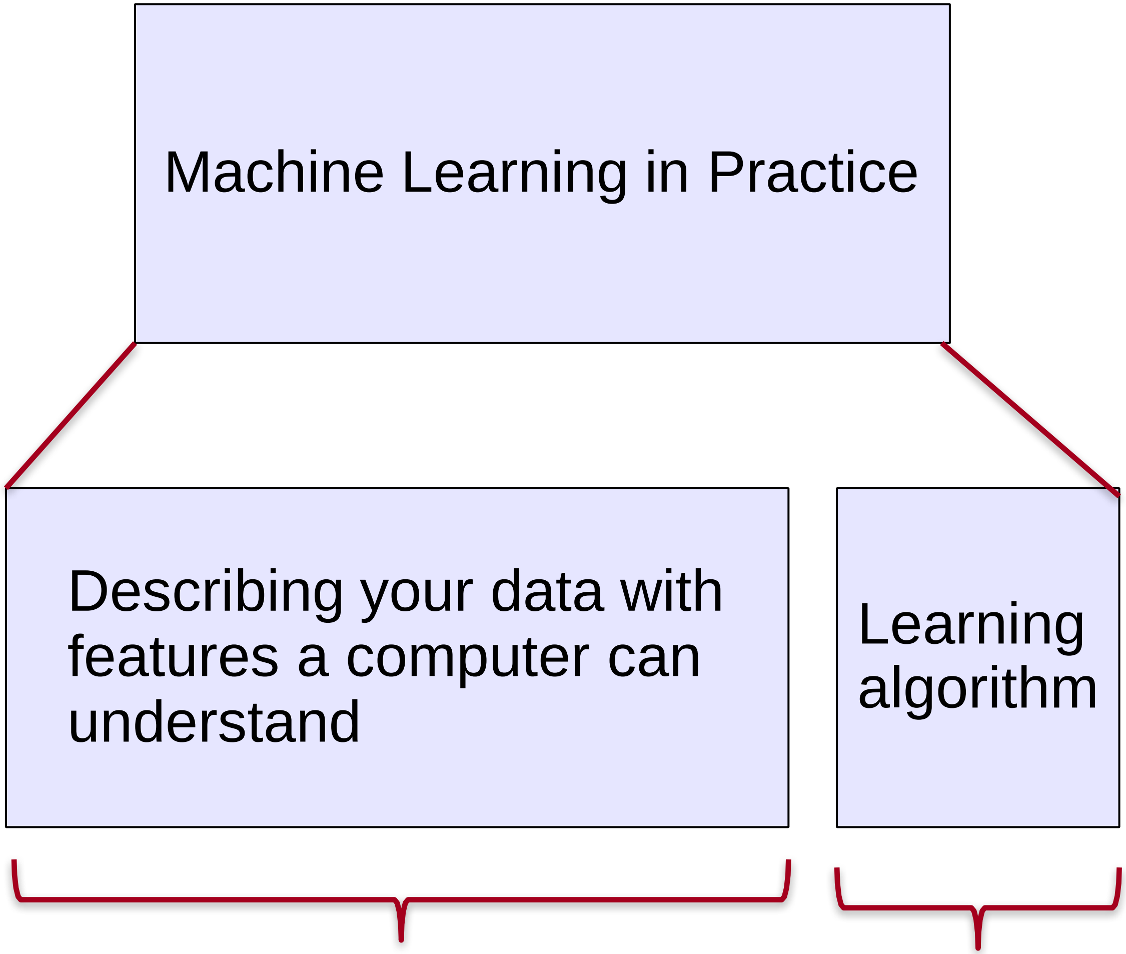
What’s Deep Learning (DL)?

- **Deep learning** is a subfield of **machine learning.**
- Most traditional machine learning methods work well because of **human-designed representations** and **input features** 由于人工设计的表示和输入功能，大多数传统的机器学习方法运行良好
  - For example: feature engineering features for sentiment classification:

	Vocabulary					Added features				
DocID	bad	good	movie	…	zulu	# of pos words	# of neg words	# of pos adjective words	# of neg adjective words	positive?
1	0	1	1	0	0	3	0	1	0	1
2	1	0	1	0	0	0	2	0	1	0
…										

- Machine learning becomes **just** optimizing weights to make a final prediction. 机器学习变得只是优化权重以做出最终预测。

Machine Learning vs. Deep Learning



Machine Learning in Practice

Describing your data with features a computer can understand

Learning algorithm

Domain specific, sometimes requires Ph.D. level expertise

Optimizing the weights on features

What’s Deep Learning (DL)?

- In contrast to standard machine learning,

  与标准机器学习相反，

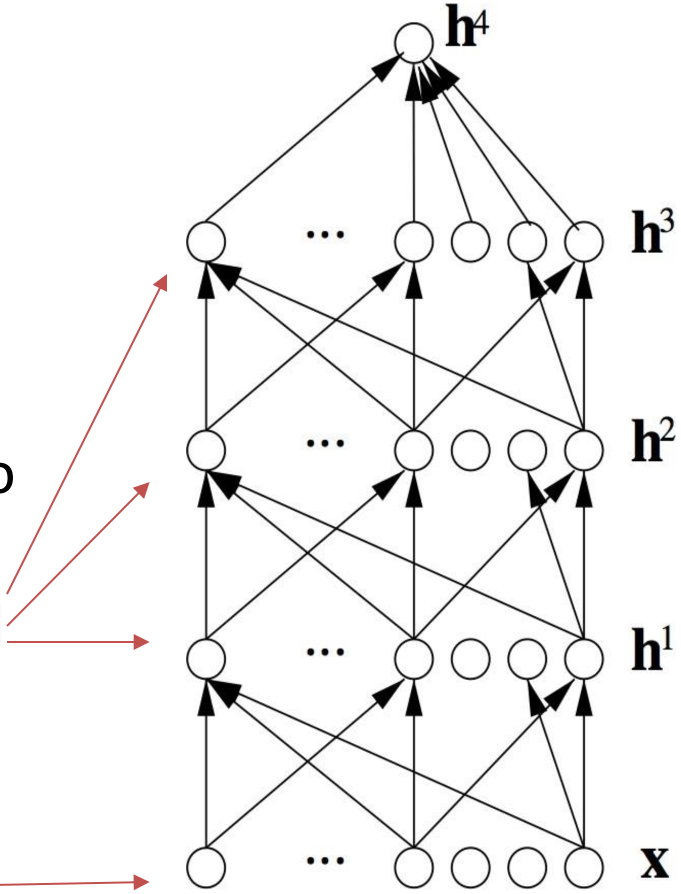
- **Representation learning** attempts to automatically learn good features or representations

  表示学习试图自动学习好的特征或表示

- Deep learning algorithms attempt to **learn (multiple levels of) representations** (here: h^1,h^2,h^3) and an output (h^4)

  深度学习算法试图学习（多个级别的）表示（在：h1，h2，h3）和输出（h4）

- From “raw” inputs **x** (e.g., a document vector)



Multi-layer Neural Networks

Depth: Repeated Composition

- Figure shows how a **deep learning** system can represent the concept of an image of a person by combining simpler concepts, such as corners and contours, which are in turn defined in terms of edges.

图显示了深度学习系统如何通过结合更简单的概念来表示一个人的形象概念，如角落和轮廓，这些概念又以边缘来定义。

- The quintessential example of a deep learning model is the feedforward deep network or **multilayer perceptron** (MLP).

深度学习模型的典型例子是前馈深度网络或多层感知器（MLP）。

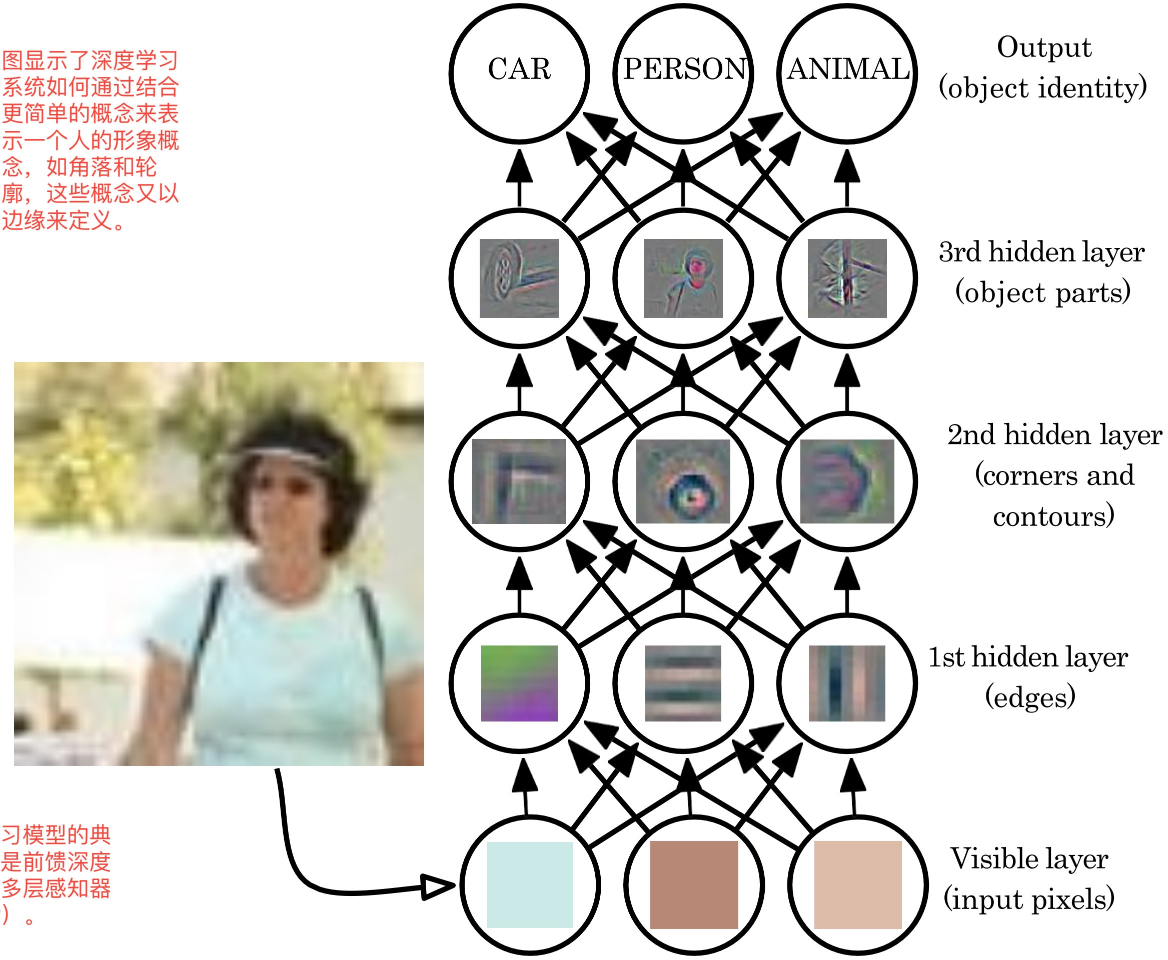


Figure 1.2

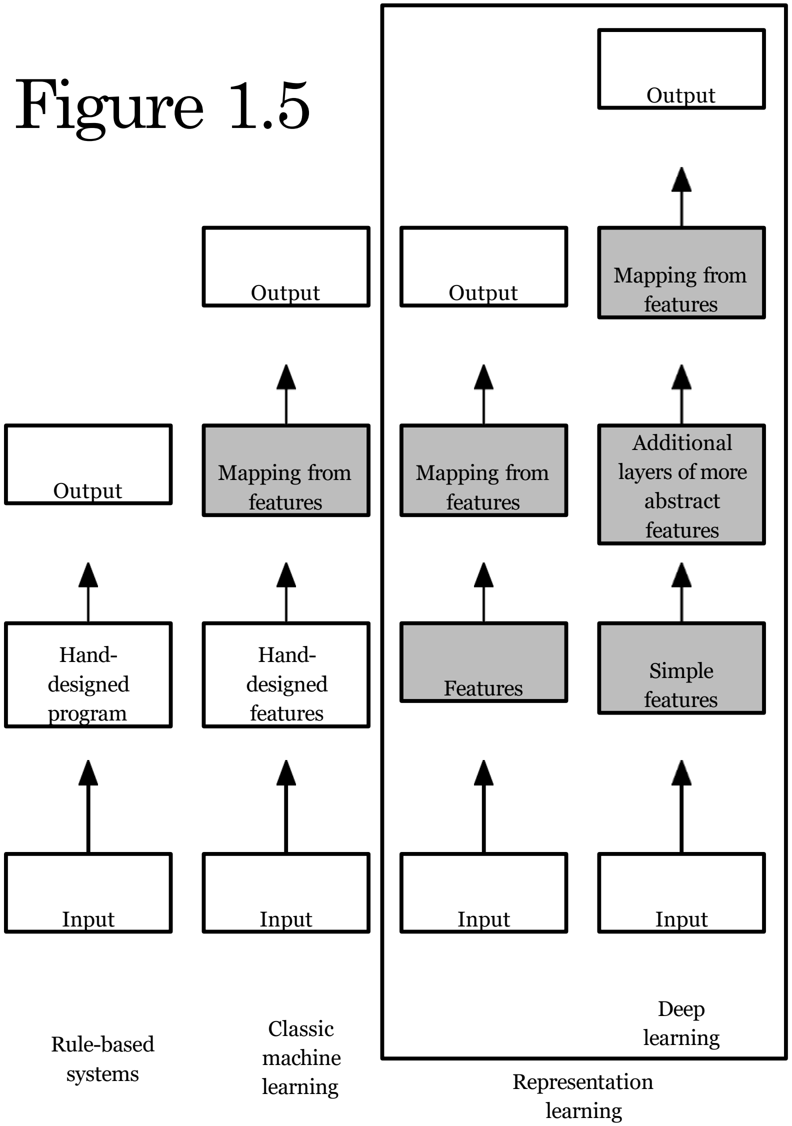
(Goodfellow 2016)

Learning Multiple Components

- Flowcharts showing how the different parts of an AI system relate to each other within different AI disciplines.
- **Shaded boxes** indicate components that are able to learn from data.

阴影框表示能够从数据中学习的组件。

Figure 1.5



(Goodfellow 2016)

Reasons for Exploring Deep Learning

- **Manually designed features** are often over specified, incomplete and take a long time to design and validate. 手动设计的功能通常被过度指定、不完整，并且需要很长时间来设计和验证。
- **Learned Features** are easy to adapt, fast to learn.
- Deep learning provides a very flexible, (almost?) universal, learnable framework for **representing** world, visual and linguistic information. 深度学习提供了非常灵活的，（几乎？）代表世界、视觉和语言信息的通用、可学习的框架。
- Deep learning can learn **unsupervised** (from raw text) and **supervised** (with specific labels like positive/negative). 深度学习可以在无监督（来自原始文本）和监督下（带有正/负等特定标签）学习。

Reasons for Exploring Deep Learning

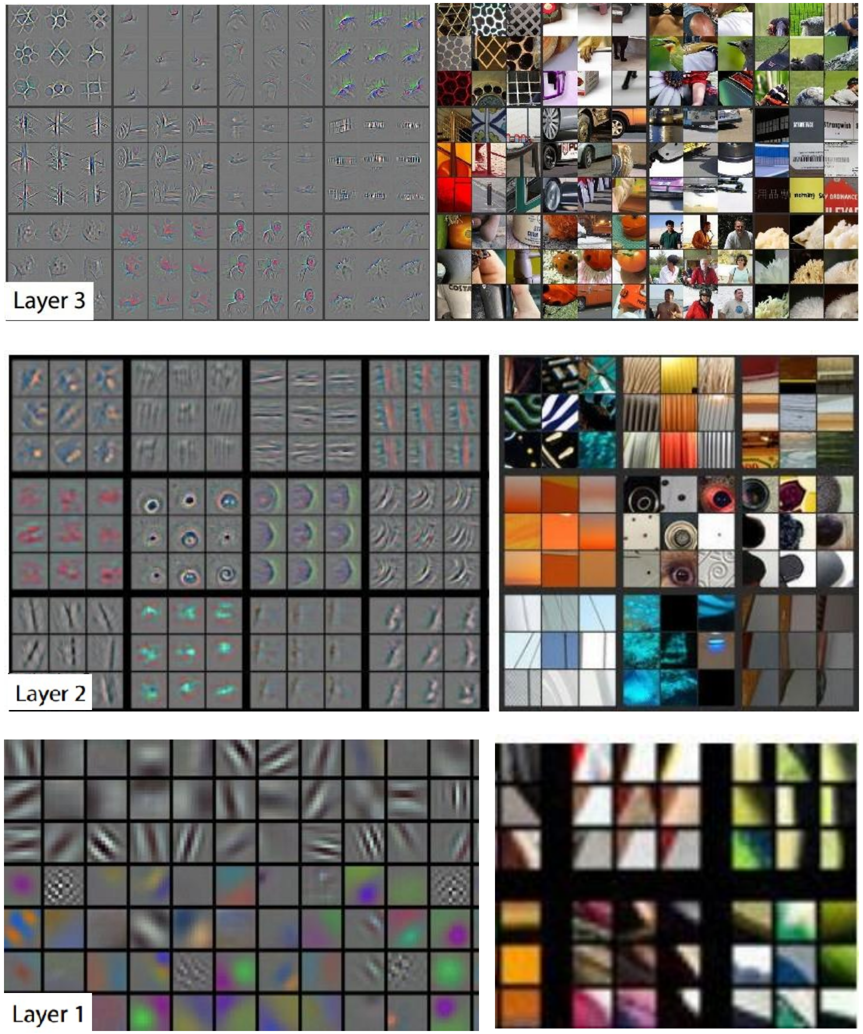
- In 2010, **deep** learning techniques started outperforming other machine learning techniques. Why this decade?

- Large amounts of training data favor deep learning 大量的训练数据有利于深度学习
- Faster machines and multicore CPU/GPUs favor Deep Learning
- New models, algorithms, ideas
  - Better, more flexible learning of intermediate representations
  - Effective end-to-end joint system learning
  - Effective learning methods for using contexts and transferring between tasks
  - Better regularization and optimization methods
- **Improved performance** (first in speech and vision, then NLP)

Deep Learning for Computer Vision

First major focus of deep learning groups was computer vision.

The breakthrough DL paper: ImageNet Classification with Deep Convolutional Neural Networks by Krizhevsky, Sutskever, & Hinton, 2012, U. Toronto. 37% error rate.



Zeiler and Fergus (2013)

Deep NLP = Deep Learning + NLP

Combine ideas and goals of NLP with using representation learning and deep learning methods to solve them.将NLP的想法和目标与使用表示学习和深度学习方法相结合来解决它们。

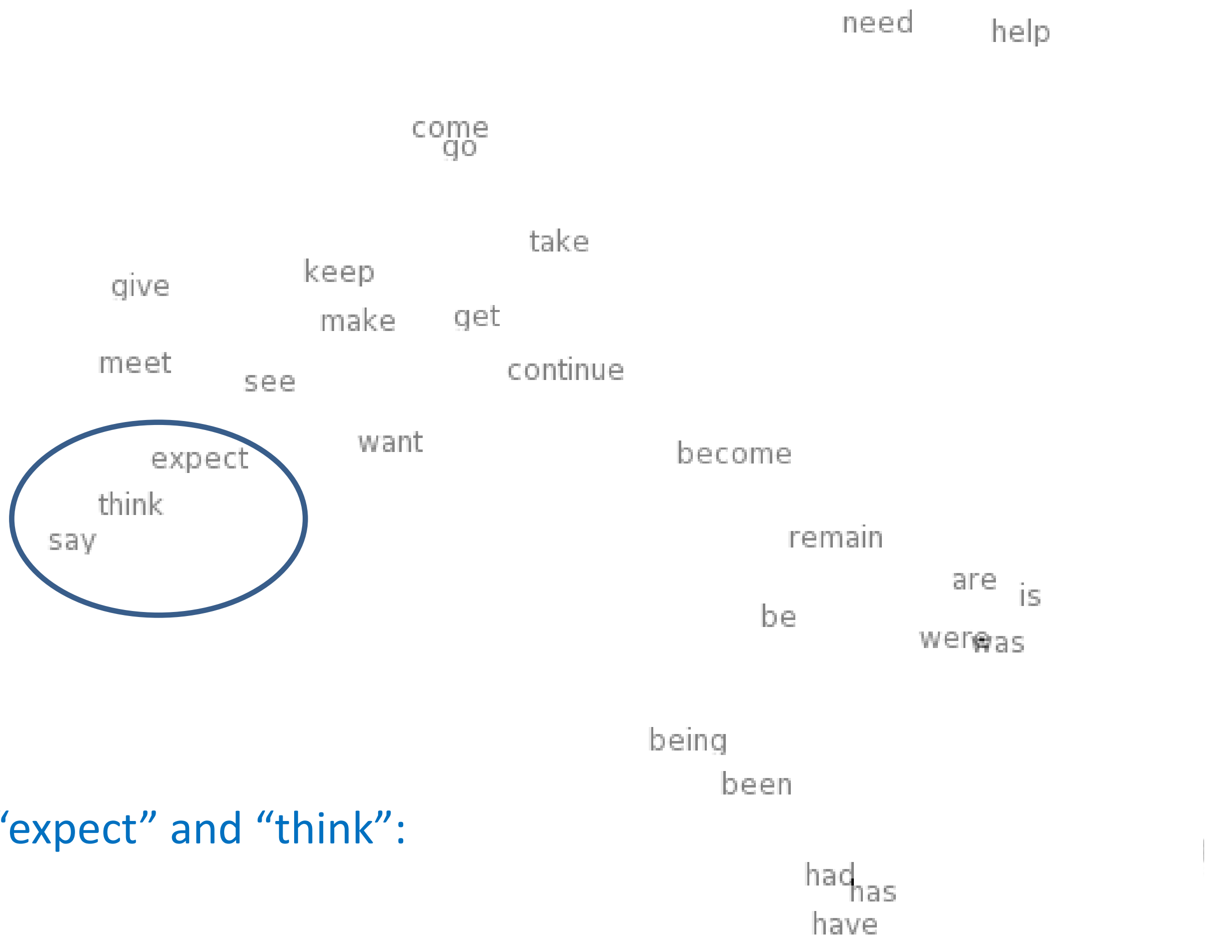
Several big improvements in recent years in NLP

- **Linguistic levels**: (speech), words, semantics
- **Intermediate tasks/tools:** parts-of-speech, entities, parsing
- **Full applications**: sentiment analysis, question answering, dialogue agents, machine translation, etc.

近年来，NLP 领域取得了多项重大进步
- 语言层面：（语音）、单词、语义
- 中级任务/工具：词性、实体、解析
- 完整应用：情感分析、问答、对话代理、机器翻译等。

Word meaning as a neural word vector – visualization

$$\textit{expect} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \\ 0.487 \end{pmatrix}$$



Compared to one-hot encoding of “expect” and “think”:

0 0 0 ……. 0 1 0 ……0 (“expect”)

0 0 0 ……. 0 0 1 ……0 (“think”)

DocID	Apple	…	…	expect	think	…	…	…	Zoo
1	0	0	0	**1**	0	0	0	0	0
2	0	0	0	0	**1**	0	0	0	0
…									

Word similarities

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

rana

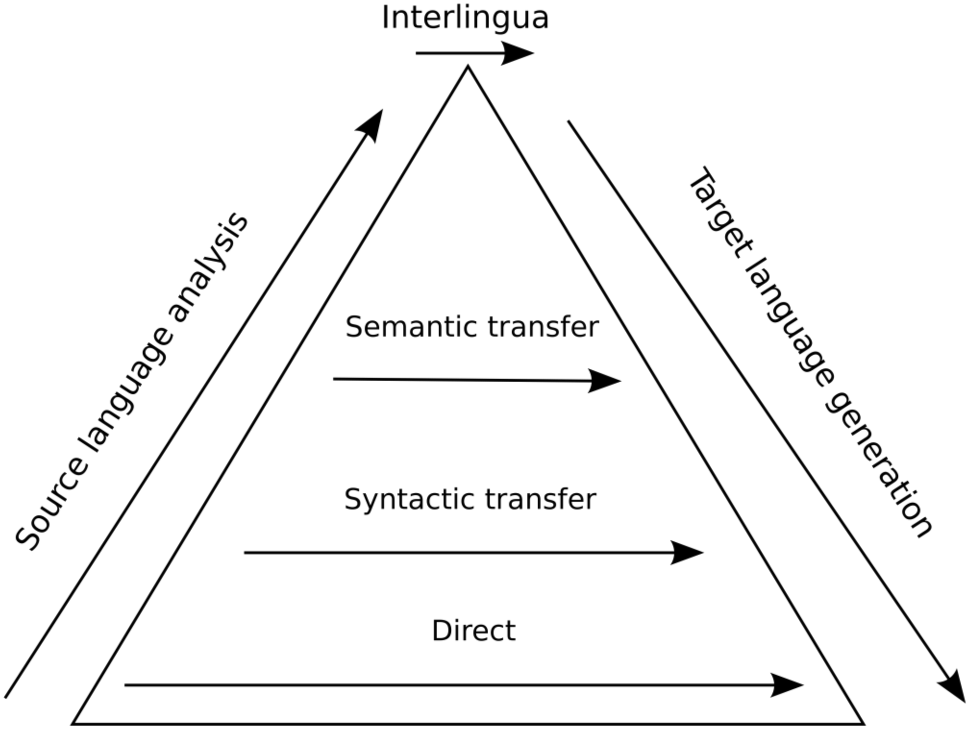
eleutherodactylus

http://nlp.stanford.edu/projects/glove/

Machine Translation

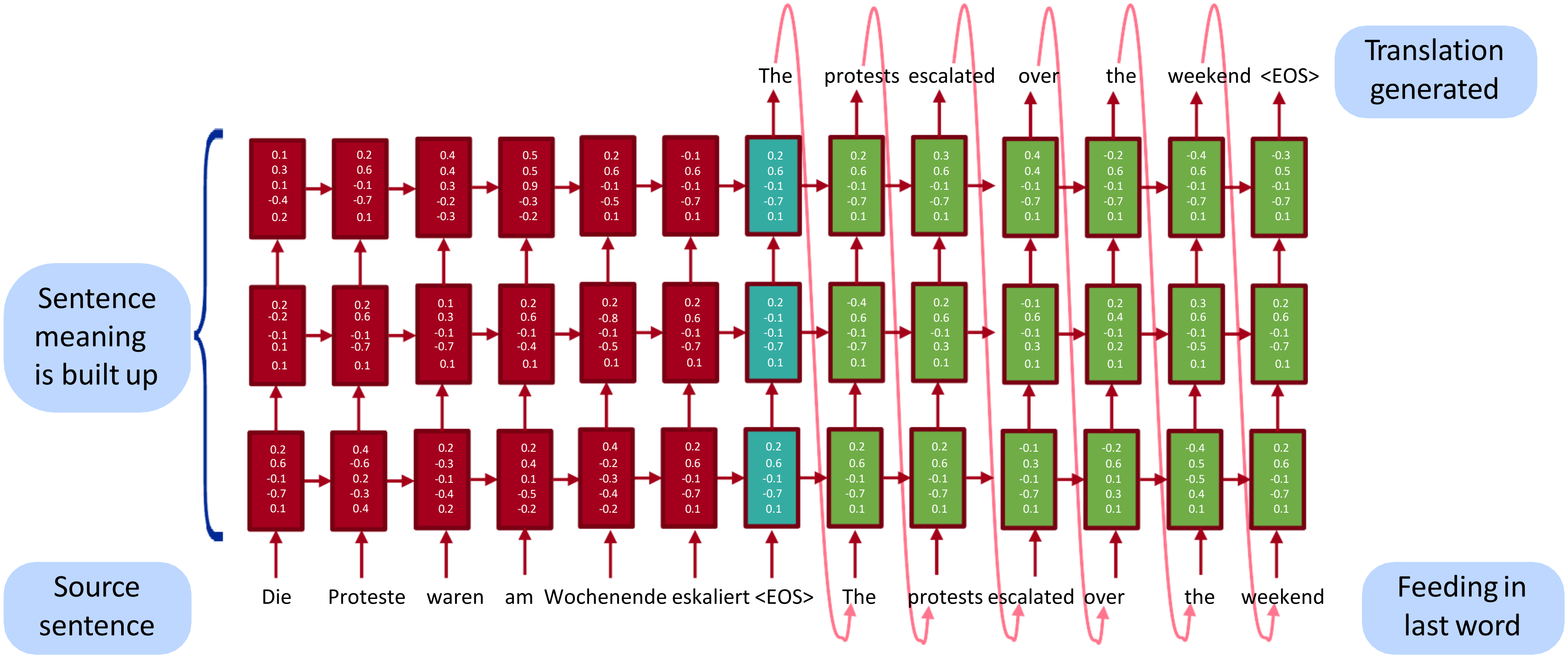
- Many levels of translation have been tried in the past:

- Traditional MT systems are very large complex systems



Neural Machine Translation

Source sentence is mapped to **vector**, then output sentence generated [Sutskever et al. 2014, Bahdanau et al. 2014, Luong and Manning 2016]



Now Google Translate uses NMT with big error reductions!

Topics We Will Cover

- This course will explore text mining and NLP techniques using **deep learning**.
- Intro to text mining (this lecture)
- Preprocessing for Text Mining, and Machine Learning
- Logistic Regression and Neural Networks
- Word Embedding
- Sequence Models
- Convolutional Neural Networks for NLP
- Machine Translation
- Transformers and Self-Attention
- Question Answering and Generative Language Models
- Text Mining Tools
    - Python, PyTorch, nltk, etc.

Prerequisites

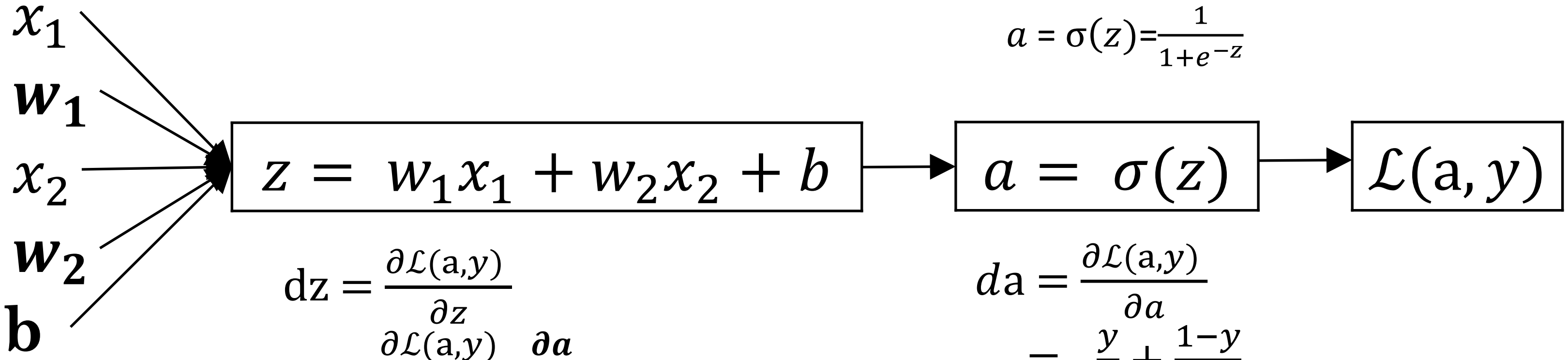
- **Proficiency in Python**
  - All class assignments will be in Python.
  - Pytorch will be used as a deep learning platform.
- Some background knowledge in the following topics would be helpful if you have
  - Statistics and probability
    - Such as standard deviation, variance, and normal distributions.
  - Linear algebra
    - Such as Matrix addition, multiplication, etc.
  - Calculus
    - Such as Partial Derivative
  - Machine learning

Sample Slide:
Logistic regression derivatives

Goal: estimate **W₁** and **W₂** and **b**.

$$\mathcal{L}(a, y) = -(y\log(a) + (1-y)\log(1-a))$$

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



x_1, $\boldsymbol{w_1}$, x_2, $\boldsymbol{w_2}$, **b**

$z = \;w_1 x_1 + w_2 x_2 + b$

$a = \;\sigma(z)$

$\mathcal{L}(\mathrm{a}, y)$

$$\mathrm{dz} = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial z}$$
$$= \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial a} \cdot \frac{\boldsymbol{\partial a}}{\boldsymbol{\partial z}}$$
$$= (-\frac{y}{a} + \frac{1-y}{1-a}) \cdot \boldsymbol{a(1-a)}$$
$$= \mathrm{a} - \mathrm{y}$$

$$\mathrm{d}a = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial a}$$
$$= -\frac{y}{a} + \frac{1-y}{1-a}$$

$$w_1 = \; w_1 - \alpha\mathbf{d}\boldsymbol{w_1}$$
$$w_2 = \; w_2 - \alpha\mathbf{d}\boldsymbol{w_2}$$
$$\mathrm{b} = \; b - \alpha\mathbf{db}$$

$$\frac{\boldsymbol{\partial\mathcal{L}(\mathrm{a},y)}}{\boldsymbol{\partial w_1}} = dw_1 = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial a} \cdot \frac{\boldsymbol{\partial a}}{\boldsymbol{\partial z}} \cdot \frac{\partial z}{\partial w_1} = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial z} \cdot \frac{\partial z}{\partial w_1} = (a-y) \cdot x_1$$

$$\frac{\boldsymbol{\partial\mathcal{L}(\mathrm{a},y)}}{\boldsymbol{\partial w_2}} = dw_2 = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial z} \cdot \frac{\partial z}{\partial w_2} = (a-y) \cdot x_2 \qquad \frac{\boldsymbol{\partial\mathcal{L}(\mathrm{a},y)}}{\boldsymbol{\partial b}} = \mathrm{d}b = \frac{\partial\mathcal{L}(\mathrm{a},y)}{\partial z} \cdot \frac{\partial z}{\partial b} = (a-y) \cdot 1 = (\mathrm{a}\text{-}\mathrm{y})$$

Referenced Materials

- *Fundamentals of Predictive Text Mining*, Sholom M. Weiss, Nitin Indurkhya, and Tong Zhang, Springer.
  - Chapter 1
- *Natural Language Processing with Deep Learning*, Stanford course, http://web.stanford.edu/class/cs224n/
- *Deep Learning,* Ian Goodfellow et al., 2016
  - Chapter 1
- *NLP Introduction*, Dan Jurafsky and Christopher Manning, http://www.stanford.edu/~jurafsky/NLPCourseraSlides.html