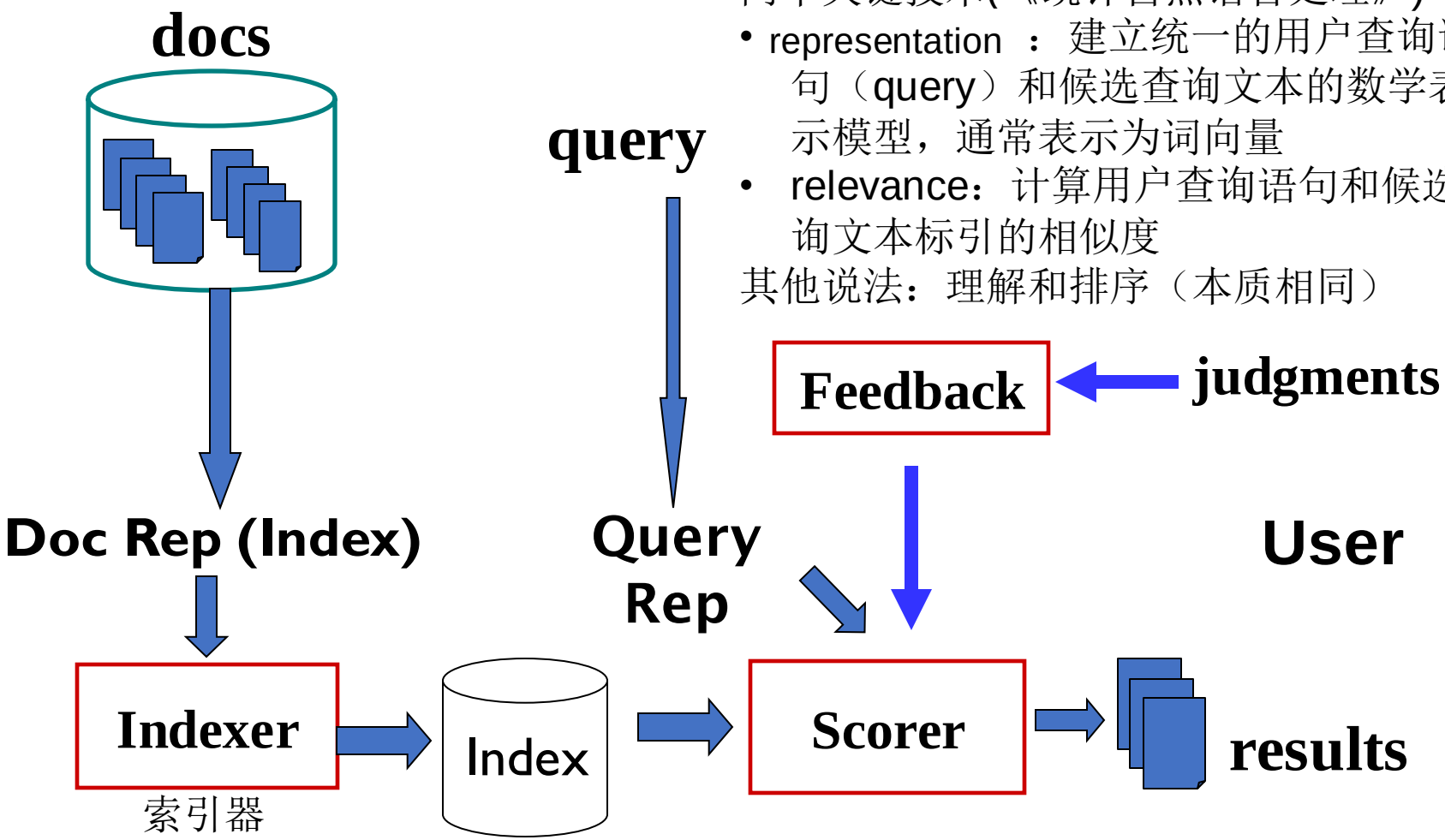
# 应用任务

- 信息检索 **(Information Retrieval, IR)**
  - 信息检索也称文本检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息→有用的文档
  - Text Retrieval is defined as the matching of some stated user query against useful parts of free-text records.
  - 起源：图书馆的参考咨询和文摘索引
  - 精确匹配模型（主要用于内部文本库） **vs** 文档相关匹配模型（主要用于网络搜索）
  - 应用：搜索引擎；邮件搜索；电脑文件搜索；法律知识检索…

# 应用任务

- IR系统的一般模式




两个关键技术(《统计自然语言处理》)

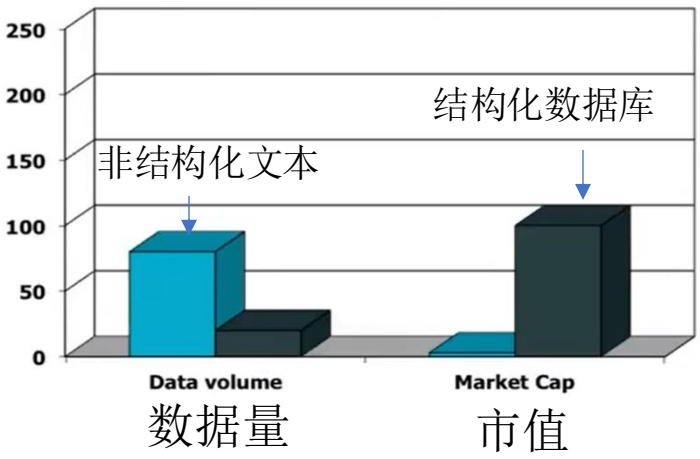
- representation ： 建立统一的用户查询语句（query）和候选查询文本的数学表示模型，通常表示为词向量
- relevance：计算用户查询语句和候选查询文本标引的相似度

其他说法：理解和排序（本质相同）

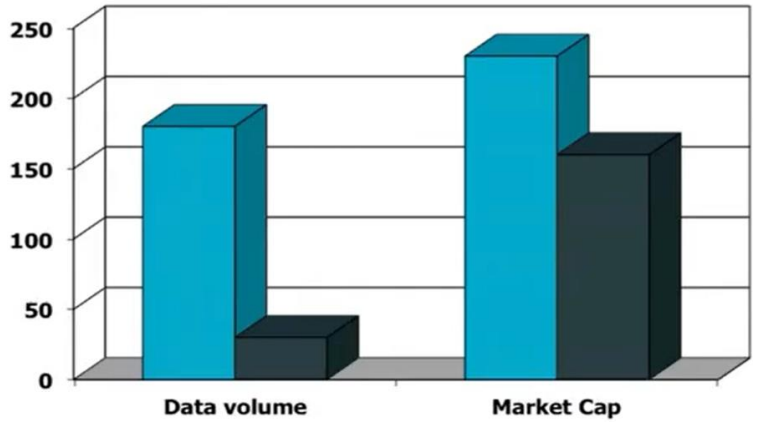
indexing：建立倒排索引的过程(《信息检索导论》)

# 应用任务

- **信息检索 (Information Retrieval)**




90年代中期




约2014年

# 应用任务

- 信息过滤 (Information Filtering)
  - 通过计算机系统自动识别和过滤那些满足特定条件的文档信息。
  - IR相关
  - 应用：网页收集与过滤、安全监控软件、个性化搜索引擎的设计

# 应用任务

- 信息抽取 (Information Extraction, IE)
  - 从文本中确定和抽取出特定的事实信息，包括实体、关系、事件等，然后进行结构化处理，输出结构化的表示。
  - 应用：与信息检索系统和Web搜索引擎结合使用、促进语义Web的实现 → IR从文档库中检索相关文档，IE是从文档中取出相关信息
  - 开放域信息抽取（open information extraction, OIE）：文本领域开放

# 应用任务

- **信息抽取 (Information Extraction, IE)**

- 子任务：关系抽取

  抽取(主体，关系，客体)三元组

  e.g. 刘翔生于上海 → (刘翔，出生地，上海)

  难点：关系重叠

- 子任务：事件抽取 **e**vent **e**xtraction：

  抽取事件相关的信息 如 触发词(主语，宾语，时间，地点)；

  触发词可以是谓词，也可以是其他（和SRL的区别）

  根据数据集范围：限定域**vs**开放域

  经典数据集：ACE2005数据集

- 子任务：命名实体识别

# 应用任务

- 信息抽取-例子

新华社北京3月8日电（记者李术峰）：中国农工民主党第十二届中央常务委员会第一次会议今天在北京召开。

会议研究通过了贯彻落实“两会”精神的有关决定，审议通过了中国农工民主党中央1998年工作要点（草案），并任命了中央副秘书长。

农工民主党中央主席蒋正华主持了会议，他说，农工民主党有100多名党员作为代表和委员参加了今年的“两会”，各位党员要认真履行代表和委员的职责，开好会，在1998年的工作中认真贯彻“两会”精神，加强农工民主党的自身建设，推动事业进一步发展，为建设有中国特色社会主义事业作出新的贡献。

会前，农工民主党中央邀请参加“两会”的来自全国各省、自治区、直辖市的农工民主党党员进行了联谊活动。

# 应用任务

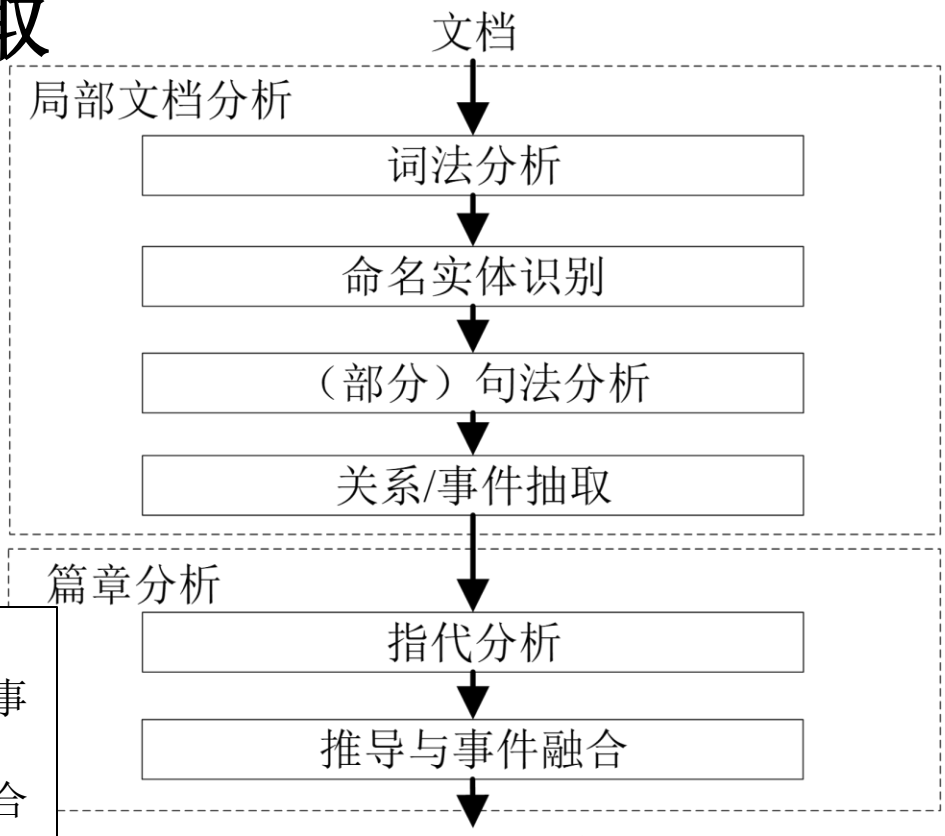
- 信息抽取-例子

| | | |
|---|---|---|
| 会 议 时 间 Time | 1998年3月9日 | |
| 会 议 地 点 Spot | 北京 | |
| 会议召集者/主 持 人 Convener | 个人姓名/团体名称 Name | 蒋正华 |
| | 机 构 、 职 位 Org/Post | 主席，农工民主党中央 |
| 会议名/标题 Conf-Title | 中国农工民主党第十二届中央常务委员会第一次会议 | |

和**自动文摘**相比，信息抽取一般是有目的地从文本中寻找想要的信息；信息抽取一般采用预定义的格式化表示。

# 应用任务

- 信息抽取




信息抽取系统的处理流程

事件融合：
- 认知语义层面：子事件通过逻辑关联融合
- 句法层面：小句融合成一个单句表征

传统方法

# 应用任务

- 问答系统 (Question-Answering system)
  - 通过计算机系统对人提出的问题的理解，利用自动推理等手段，自动求解答案并做出相应的回答。问答技术有时与语音技术和多模态输入/输出技术，以及人机交互技术等相结合，构成人机对话系统。
  - 核心： Question-Answering (QA)
  - IR的高级形式
  - 单轮 vs 多轮（对上文提及的利用，和上文的一致性）
  - 限定域 vs 开放域
  - 任务型（自动客服）vs 闲聊型（chatbot, e.g.chatGPT）

# 应用任务

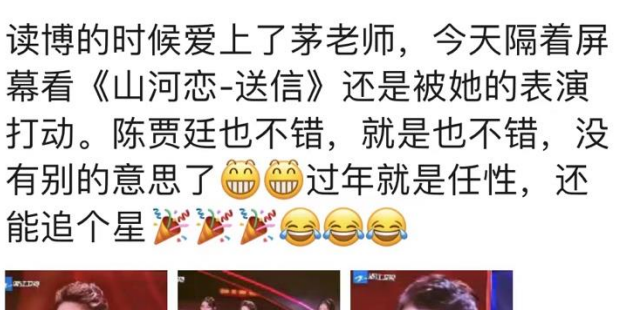
- 问答系统 (Question-Answering system)


例子：客服机器人


例子：聊天机器人

# 应用任务

- 意见/观点挖掘 (Opinion Mining)
  - 对主观性文本，例如断言或评论文本的处理
  - 包含对个人、群体、组织等的意见(opinion) 、情感(sentiment) 和态度
  - 实体属性的提取、观点内容提取、观点情感极性分析、观点的总结
  - 篇章级、语句级、属性/方面级
  - 应用：产品用户意见调查、舆情监督



# 应用任务

- **情感分析 (Sentiment analysis)**
  - 可以看成观点挖掘的一个子任务
  - 机器对于情感的要求就是机器情感计算，也就是机器理解人类的情感和生成情感的能力。
  - 目标：粗/细粒度的分类、元素抽取、生成


e.g. 这家餐厅服务不错，但位置太偏了，去一趟倒车几次，以后不会再去了

相关属性：服务，位置
属性情感：好，差
句子好/差情感分类：差
句子细粒度情感：e.g. 失望

# 应用任务

- **文本自动校对 (Automatic proofreading)**
  - 对文字拼写错误、语法错误、语义错误等进行自动检查、校对和编排。
  - 应用：排版、印刷和书籍编撰等。

| 错误类型 | | 错误 | 正确 |
|---|---|---|---|
| 形似字错误 | | 诞续 | 延续 |
| 音似字错误 | 同音同调 | 火势向四周漫(man4)延 | 火势向四周蔓(man4)延 |
| | 同音异调 | 但是不行(xing2)还是发生了 | 但是不幸(xing4)还是发生了 |
| | 相似音 | 词青(qing1)标注 | 词性(xing2)标注 |
| 知识型错误 | | 埃及有金子塔 | 埃及有金字塔 |
| 推断型错误 | | 他的求胜欲很强，为了越狱在挖洞 | 他的求生欲很强，为了越狱在挖洞 |

拼写错误举例

# 应用任务

- 文本自动校对 (Automatic proofreading)

| 错误类型 | 错误 | 正确 |
|---|---|---|
| 字词冗余 | 我根本不能理解这妇女辞职回家的现象。 | 我根本不能理解妇女辞职回家的现象。 |
| 字词缺失 | 我河边散步的时候。 | 我在河边散步的时候。 |
| 搭配不当 | 还有其他的人也受被害。 | 还有其他的人也受伤害。 |
| 字词乱序 | 世界上每天由于饥饿很多人死亡。 | 世界上每天很多人由于饥饿死亡。 |

语法错误举例

| 错误类型 | 错误 | 正确 |
|---|---|---|
| 知识错误 | 中国的首都是南京 | 中国的首都是北京 |
| 搭配错误 | 他戴着帽子和皮靴就出门了 | 他戴着帽子穿着皮靴就出门了 |

语义错误举例

- 流程：

1) 错误识别：判断文本是否存在错误以及错误位置
2) 生成纠正候选：构建错误字符的纠正候选
3) 评估纠正候选：利用某种评分函数对纠正候选排序

# 绪论

- 自然语言处理的概念
- 自然语言处理的主要研究内容
- 自然语言处理研究的困难
- 自然语言处理的研究方法

# 主要难点

- 困难1：歧义(ambiguity)现象
  - 词法歧义

    **Mary's →Mary's？ Mary is？ Mary has？**

    庸医治病害死人 **Q: 怎么分词？**

    - 庸医/ 治/ 病害/ 死人
    - 庸医/ 治/ 病害/ 死/ 人
    - 庸医/ 治/ 病/ 害/ 死人
    - 庸医/ 治/ 病/ 害/ 死/ 人 (√)
    - 庸/ 医治/ 病害/ 死/ 人
    - 庸/ 医治/ 病/ 害/ 死/ 人
    - 庸/ 医/ 治/ 病害/ 死/ 人
    - 庸/ 医/ 治/ 病/ 害/ 死/ 人

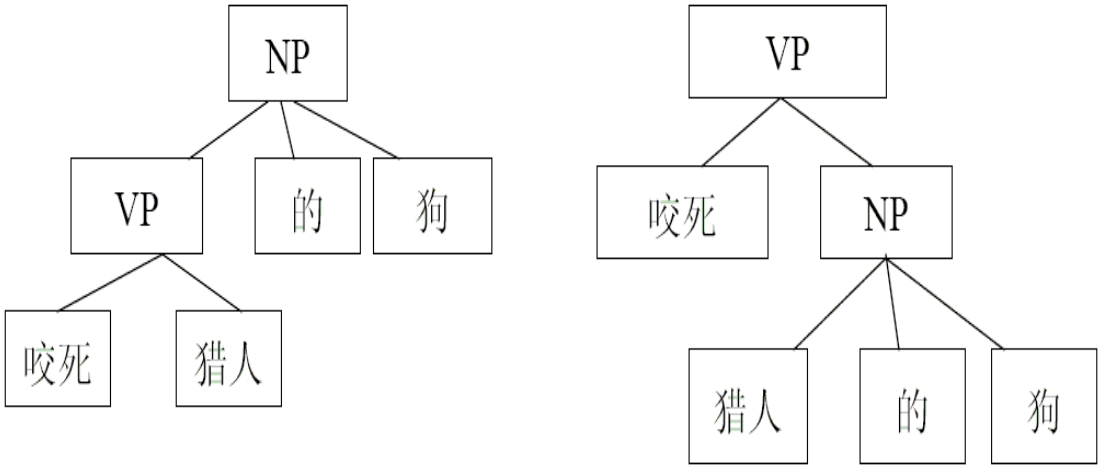
# 主要难点

- 词性标注歧义
  - 把/q-p-v-n 这/r 篇/q 报道/v-n 编辑/v-n 一/m-c 下/f-q-v
- 缩写歧义
  - 南大：南京大学？南昌大学？
  - ABC: 中国农业银行？。。。

# 主要难点

- 句法结构歧义
  - 咬死猎人的狗

VP →VP + NP | v
NP →NP + NP| NP + 的+ NP | VP + 的+ NP |n



# 主要难点

- 结构歧义的爆炸性
- 句子中的歧义的组合能够产生大量的可能解释.
- A sentence ending in *n* prepositional phrases has *over* $2^n$ syntactic interpretations (cf. Catalan numbers).
  - **“I saw the man with the telescope”: 2 parses**
  - **“I saw the man on the hill with the telescope.”: 5 parses**
  - **“I saw the man on the hill in Texas with the telescope”: 14 parses**
  - **“I saw the man on the hill in Texas with the telescope at noon.”: 42 parses**
  - **“I saw the man on the hill in Texas with the telescope at noon on Monday” 132 parses**

# 主要难点

- 语义歧义

配钥匙师傅：你配吗？

食堂阿姨：你要饭吗？

算命先生：你算什么东西？

快递小哥：你是什么东西？

上海垃圾分拣阿姨：你是什么垃圾？

滴滴司机：你搞清楚你自己的定位了么？

理发师傅：你自己照照镜子看看你自己，觉得还行么？

小区保安：你是谁？你从哪里来？要到哪里去？

# 主要难点

- 篇章歧义
  - 指代/共指消解 **(anaphora/coreference resolution)**
    - 张三看到了李四，当时他在公共汽车上。由于车子开得太快，没看清和他在一起的那位女孩子是谁。
    - The [Programmer i] successfully combined [Prolog j] with C, but [he i] had combined [it j] with Pascal last time.
    - The [Programmer i] successfully combined Prolog with [C j], but [he i] had combined Pascal with [it j] last time.

# 主要难点

- 困难2：大量未知语言现象
  - 新的词汇、人名、地名、术语等
    - **out-of-vocabulary (OOV) 未登陆词**
  - 新的含义
    - 如：川普？

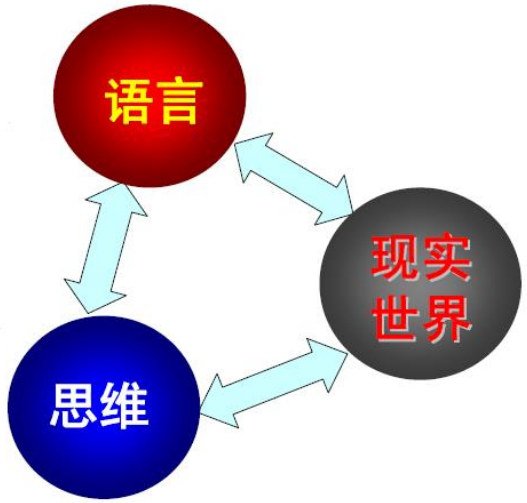

  - 新的用法和语句结构等
    - 尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构：被平均，这很××

# 主要难点

- 归纳起来，**NLU** 所面临的挑战：
  - 普遍存在的不确定性：词法、句法、语义、语用和语音各个层面
  - 未知语言现象的不可预测性：新的词汇、新的术语、新的语义和语法无处不在
  - 始终面临的数据不充分性：有限的语言集合永远无法涵盖开放的语言现象
  - 语言知识表达的复杂性：语义知识的模糊性和错综复杂的关联性难以用常规方法有效地描述，为语义计算带来了极大的困难

# 主要难点

- 理解语言是一个复杂的思维过程
  - — 语言学、心理学
  - — 逻辑学、认知科学
  - — 计算机科学
  - — 统计学
  - — 背景知识、常识等



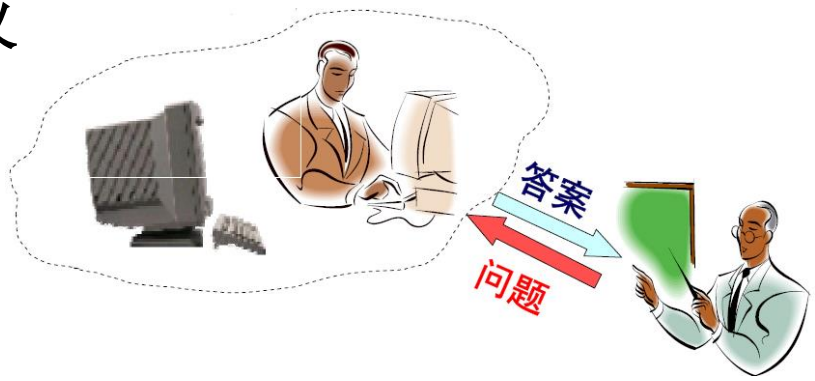

- **1956年 达特茅斯会议提出人工智能技术成熟的两个标志性目标**
  1. 在国际象棋上可以战胜人类 √
  2. 在机器翻译上能够超越人类

# 绪论

- 自然语言处理的概念
- 自然语言处理的主要研究内容
- 自然语言处理研究的困难
- 自然语言处理的研究方法

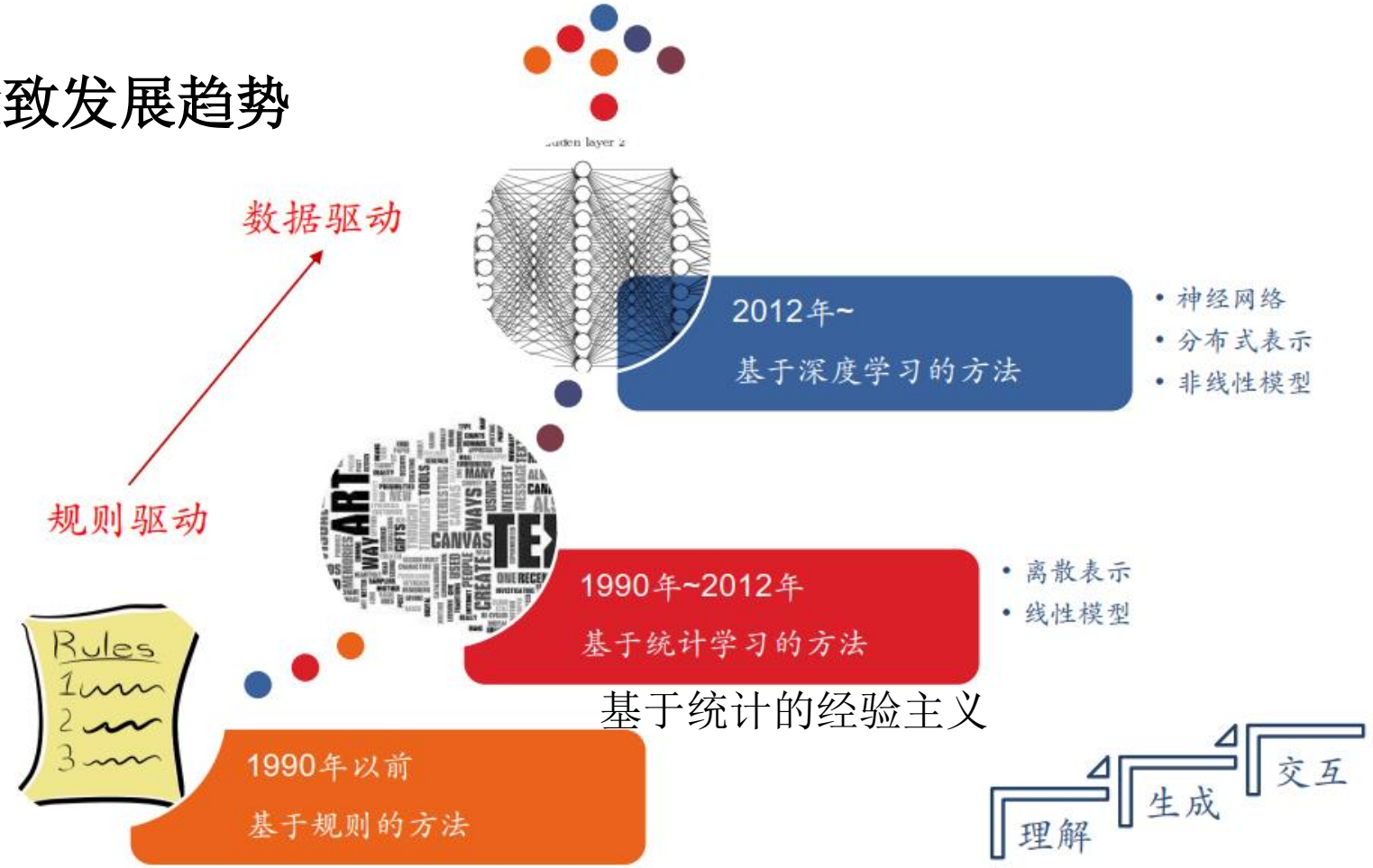
# 机器能够理解人的语言吗？

- 很难，但是没有证据表明不行
- 什么是理解？
  - 理性派的结构主义：机器的理解机制与人相同
    - 问题在于人类也说不清自己理解语言的步骤
  - 经验派的功能主义：机器的表现与人相同
    - 图灵测试：如果通过自然语言的问答，一个人无法识别和他对话的是人还是机器，那么就应该承认机器具有智能
    - 深度学习属于经验主义

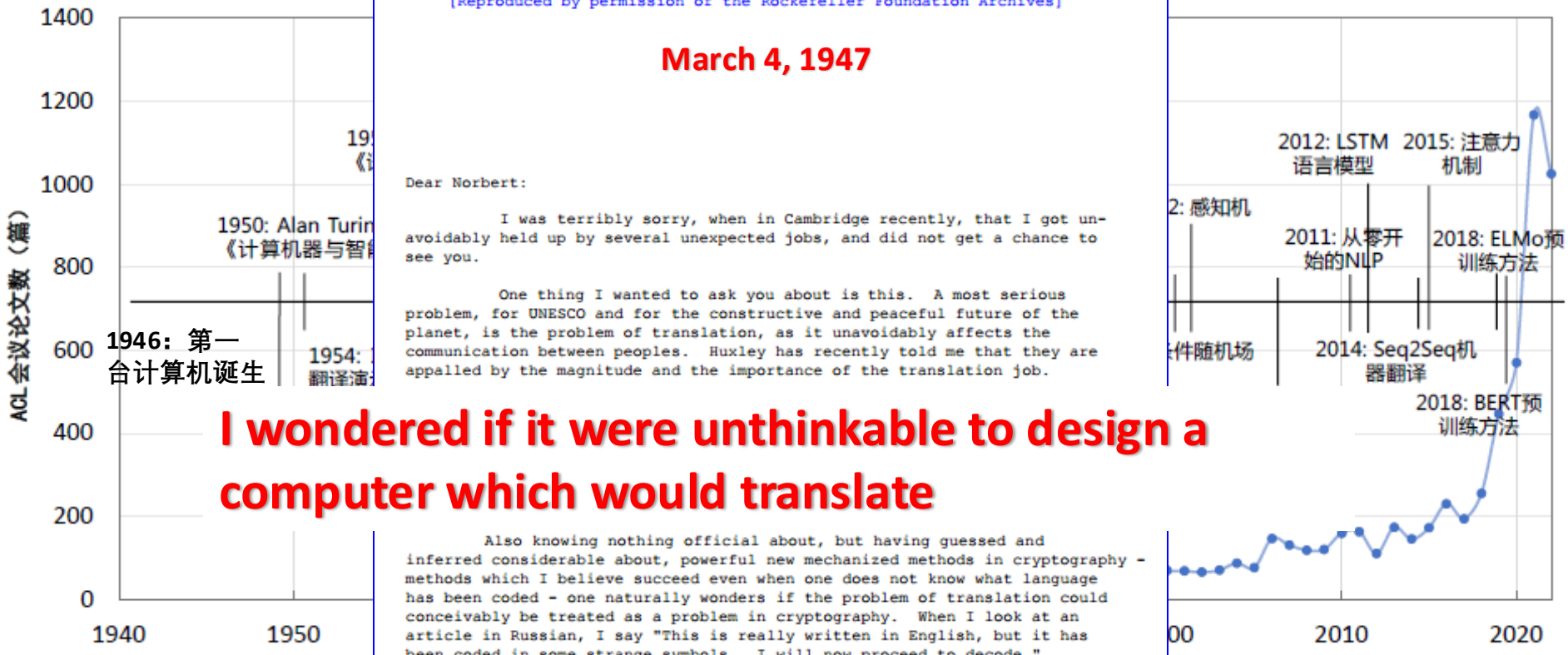


# NLP的研究历史

## 大致发展趋势



# NLP的研究历史




[Reproduced by permission of the Rockefeller Foundation Archives]

**March 4, 1947**

Dear Norbert:

I was terribly sorry, when in Cambridge recently, that I got unavoidably held up by several unexpected jobs, and did not get a chance to see you.

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

**I wondered if it were unthinkable to design a computer which would translate**

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography - methods which I believe succeed even when one does not know what language has been coded - one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Cordially,

Warren Weaver.

Professor Norbert Wiener
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

WW:AEB

20 世纪50 年代末至
20 世纪70 年代到8
20 世纪90 年代到2
2006 年至今的深度

吾言处理导论》

# 理性主义

- 基于规则/符号主义
- 理论基础：Chomsky的文法理论

## 人物简介

播报 编辑

艾弗拉姆·诺姆·乔姆斯基博士（Avram Noam Chomsky，1928年12月7日-）是麻省理工学院语言学的荣誉退休教授。乔姆斯基的《生成语法》被认为是20世纪理论语言学研究上最伟大的贡献。他还通过对伯尔赫斯·弗雷德里克·斯金纳的《口头行为》的评论，发动了心理学的认知革命，挑战在1950年代占主导地位的行为主义者学习精神和语言的方式。他那自然的学习语言的方法也对语言和精神的哲学起了很大的影响。他的另一大成就是建立了乔姆斯基层级：根据文法生成力不同而对正则语言做的分类。

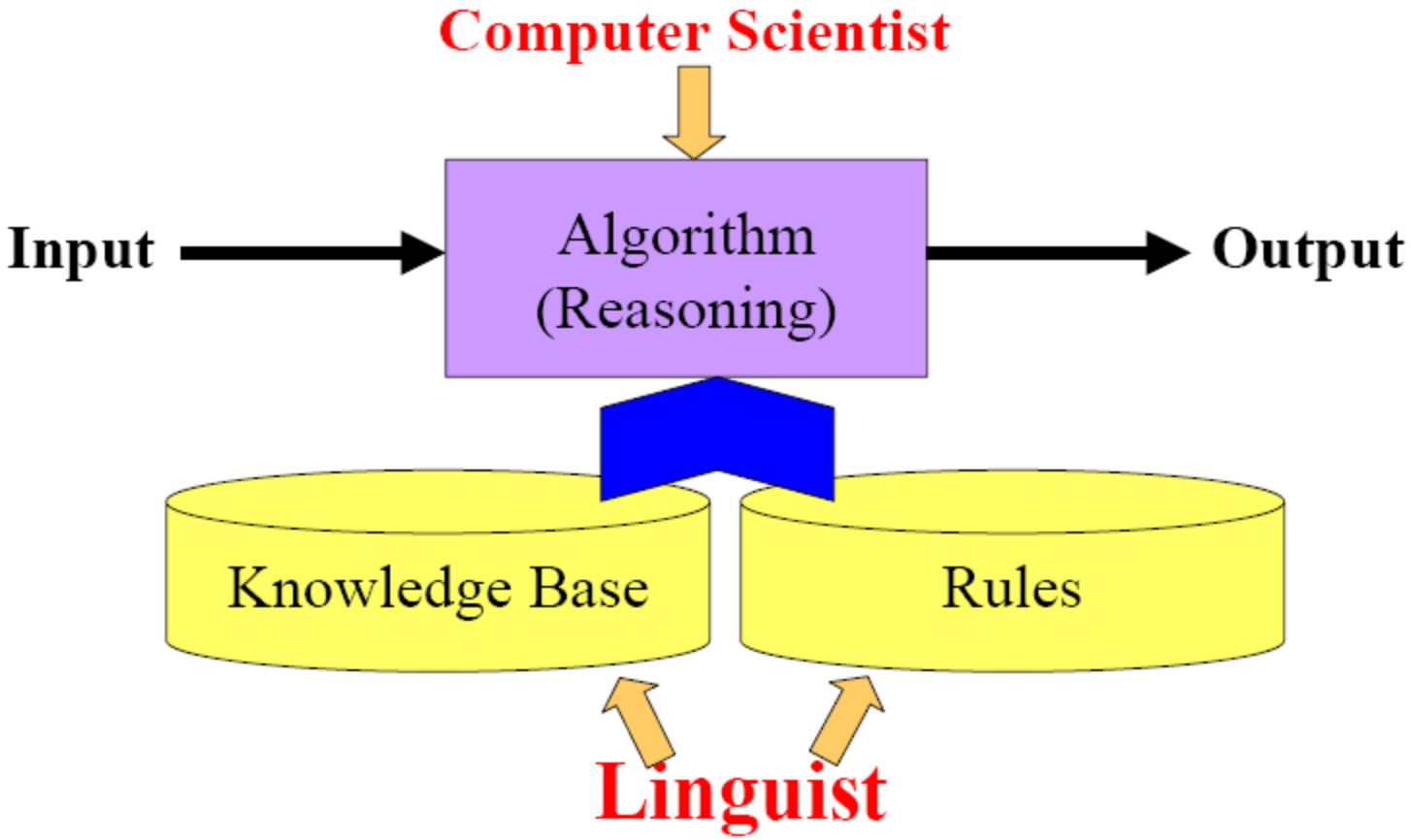
诺姆·乔姆斯基

乔姆斯基还因他对政治的热忱，尤其是他对美国和其它国家政府的批评而著名。乔姆斯基把自己归为自由社会主义者，并且是无政府工团主义的同情者。一般认为他是活跃在美国左翼政坛的著名主要知识分子。据艺术和人文引文索引说，在1980年到1992年，乔姆斯基是被文献引用数最多的健在学者，并是有史以来被引用数第8多的。

# 理性主义

- 基于规则/符号主义
- 理论基础：Chomsky的文法理论
- 通常基于Chomsky的语言原则(principles)，通过语言所必须遵守的一系列原则来描述语言。
  - 规则库开发：N + N → NP
  - 词典标注: N；V；
  - 推导算法设计：归约？推导？歧义消解方法？……
- 知识库＋ 推理系统 → NLP 系统
- 问题：语言的变化性；规则的无穷性、复杂性

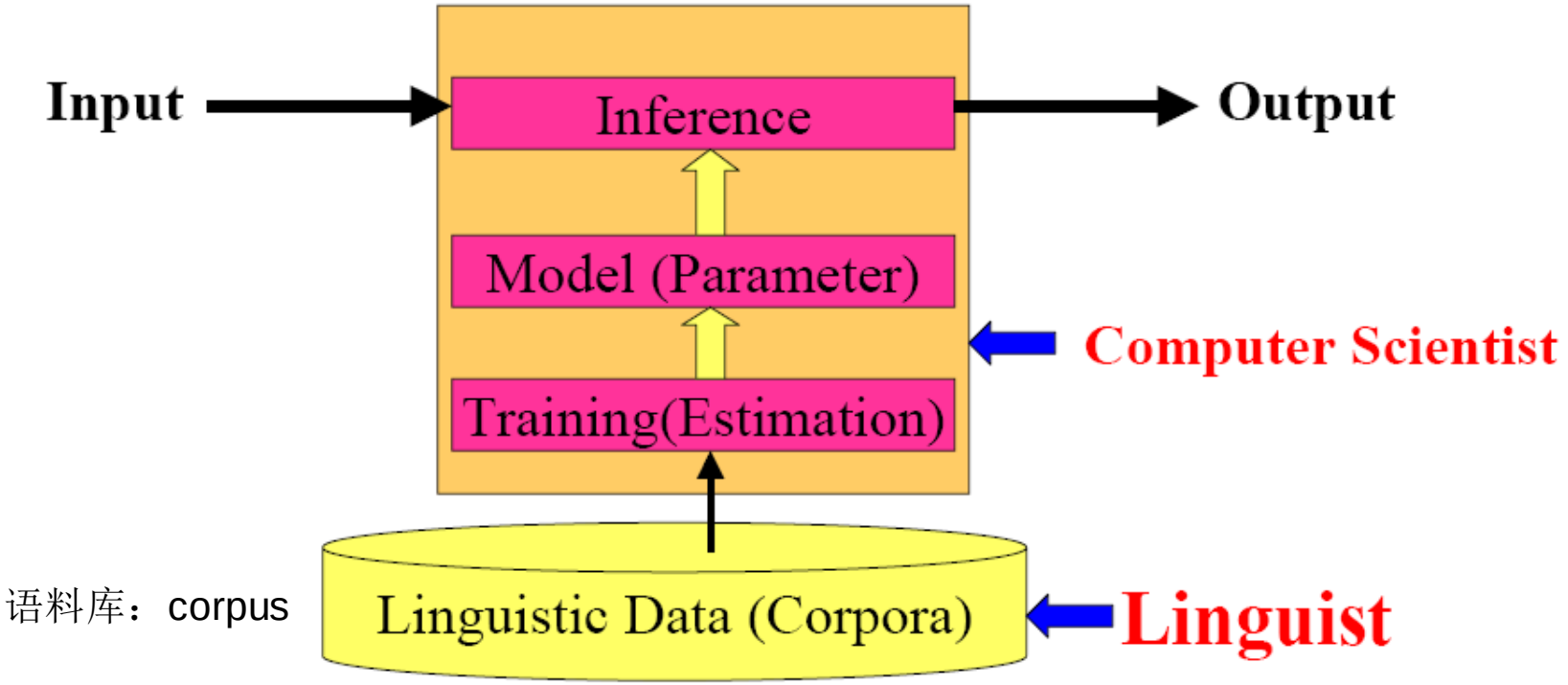
# 理性主义


Computer Scientist
Input
Algorithm
(Reasoning)
Output
Knowledge Base
Rules
Linguist

# 经验主义

- 基于统计
  - 人的语言知识通过感观输入，经过一些简单的联想(association)与泛化(generalization)的操作而得到
  - 从大量的语言数据中获得语言的知识结构
- 设定一个语言学习模型，推导出参数值
  - 形成基于统计的语言处理技术
  - 对每一种语言现象均给出统计量化指标
- 语料库＋ 统计模型 → NLP 系统
- 理论基础：统计学、信息论、机器学习

# 经验主义




## Optimization method

- 近似
- Eg. MEM, SVM

## Statistical method

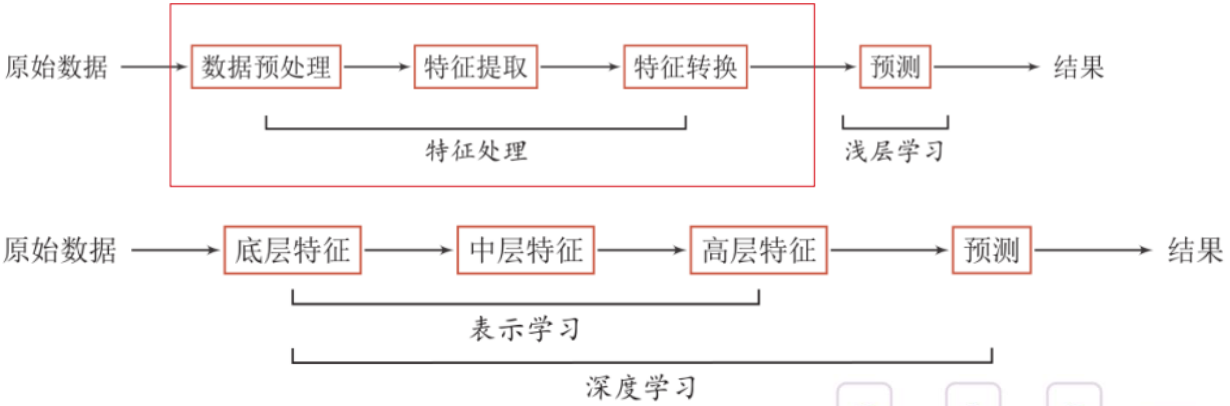
- 最大似然学习
- 最大后验
- 贝叶斯学习

# 深度学习

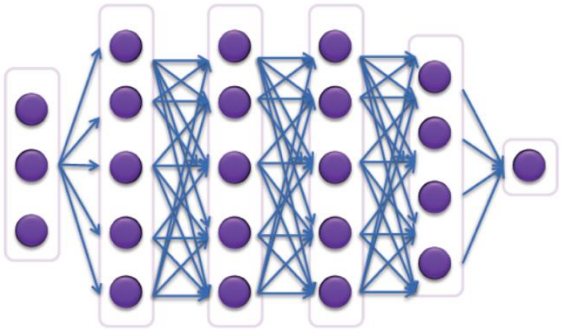
- 机器学习的发展 + 海量数据的易得
- 人工神经网络通过模拟生物大脑结构和功能，由大量节点（神经元）相互联接构成，对数据之间的复杂关系进行建模
- 深度学习=表示学习+浅层学习

表示学习：算法自动地从数据中学习数据的表示

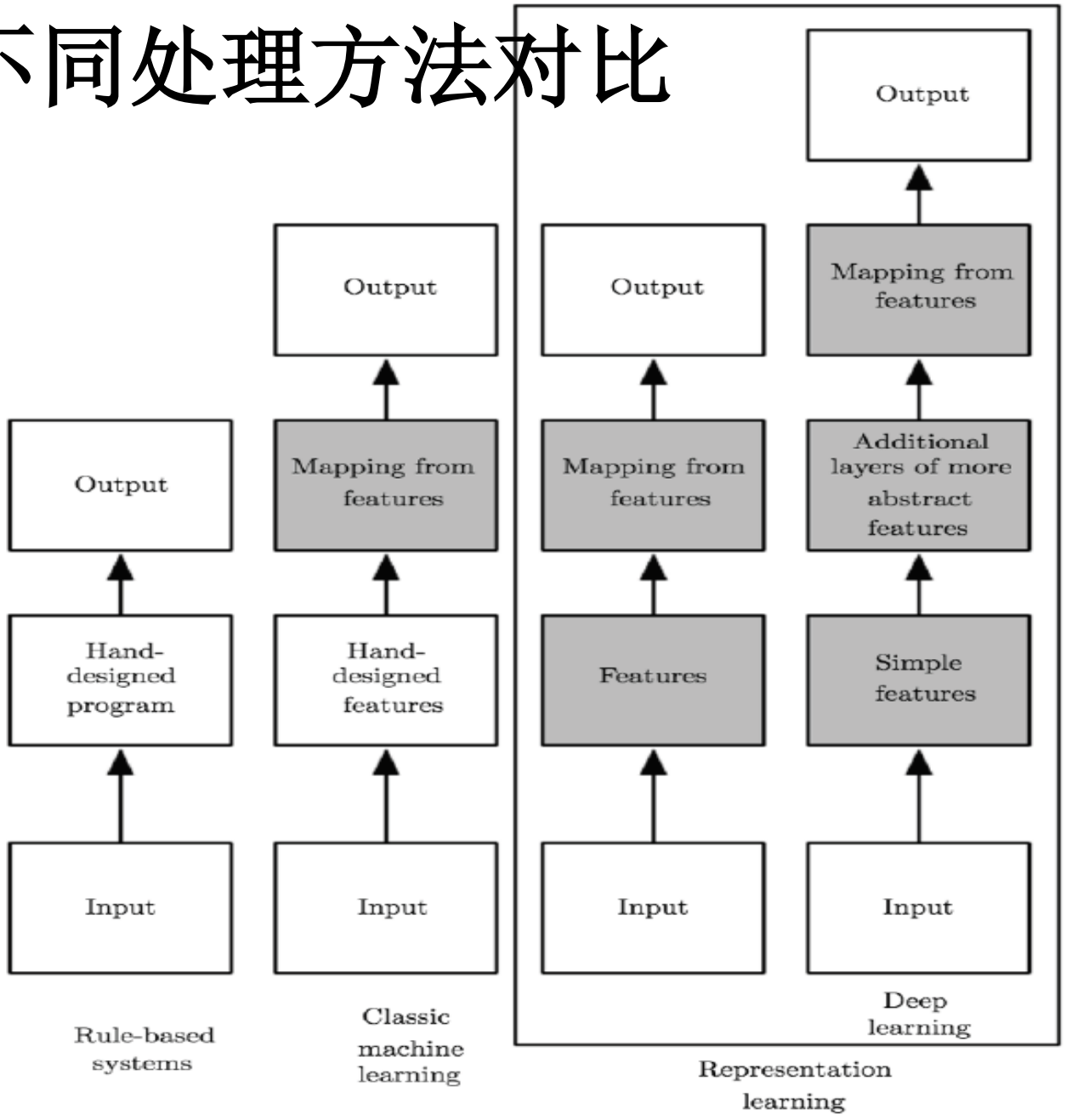
- 数据清洗：无关数据，脏数据
- 分句、中文分词、英文词干化、去除停用词




- 自然语言处理应用： 分布式表示，神经机器翻译，自动问答……几乎你能想到的任务，都可以用深度学习解决！



# NLP不同处理方法对比



# 更多…

- **NLP与其他领域研究的结合**
- 计算机领域：语音（语音识别），视觉（图像说明生成），软件工程（代码改错）…
- 生物/化学领域：蛋白质结构预测，新药结构生成
- 音乐领域：乐谱生成
- 教育领域：学生学习态度分析。。。

- 趋势/热点：
  - ✓ 多模态
  - ✓ 新领域
  - ✓ 数据受限条件
  - ✓ 个性化需求
  - ✓ 大模型…

# 结束语

- 自然语言处理是一个非常重要的、具有广泛应用前景的研究领域。
- 自然语言处理的研究目前还处于发展阶段。
- 基于统计学习和神经网络的自然语言处理研究方法是目前的主流方法。
- 希望同学们今后能在**NLP**研究领域中做出成绩，早日实现让机器自由理解和处理语言和文本的梦想！