

# 法律声明

---

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



# 自然语言处理之序列模型

---

主讲人： 史兴

6/28/2017

# 自我介绍

---

- 2008-2012 清华大学计算机系本科
  - 句子压缩
  - 情感分析
  - 爬虫, 社交网络可视化
- 2012-现在 南加州大学计算机系博士在读
  - 机器翻译
  - 英文诗歌生成
  - Seq2seq 的解释, 加速等

# 课程说明

---

## ☐ 积极互动

- 每讲10分钟左右，会回答大家相关的问题(留言区提问)
- 中场休息5分钟

## ☐ 线下提问

- 小象问答社区提问：

- ☐ <http://wenda.chinahadoop.cn/>

- ☐ 每个提问的帖子需要加上标签“自然语言处理”

# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

# 基本概念

---

## ☐ 什么是自然语言处理？

- 使用计算机处理文本以及声音

## ☐ 本质

- I 文本 -> 表示 (representations)

- ☐ 情感分析： 句子 -> -1/0/+1

- ☐ 句法分析： 句子 -> 句法树

- II 表示 -> 文本

- ☐ 诗歌生成： 主题词 -> 诗歌

- III 文本 -> 文本

- ☐ 机器翻译

# 提纲

---

- 基本概念
- 广泛应用
  - 常见应用
  - 前沿应用
  - 行业、技术回顾
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望



# 广泛应用

---

## ☐ 常见应用

### ■ 应用

☐ 输入 --> 输出

☐ 可能涉及到的技术，大概说明一下是怎么做的

### ■ 输入法：

☐ “shurufa” -> “输入法”

☐ 语言模型，自动机

### ■ 自动拼写更正：

☐ “ganrante” -> “guarantee”

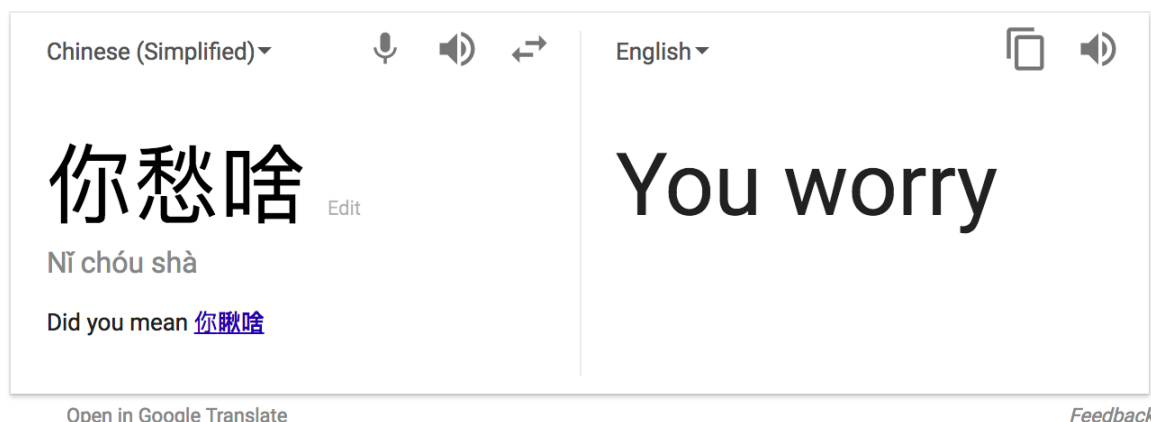
☐ 语言模型，自动机，编辑距离

# 广泛应用

---

## □ 常见应用

### ■ 机器翻译

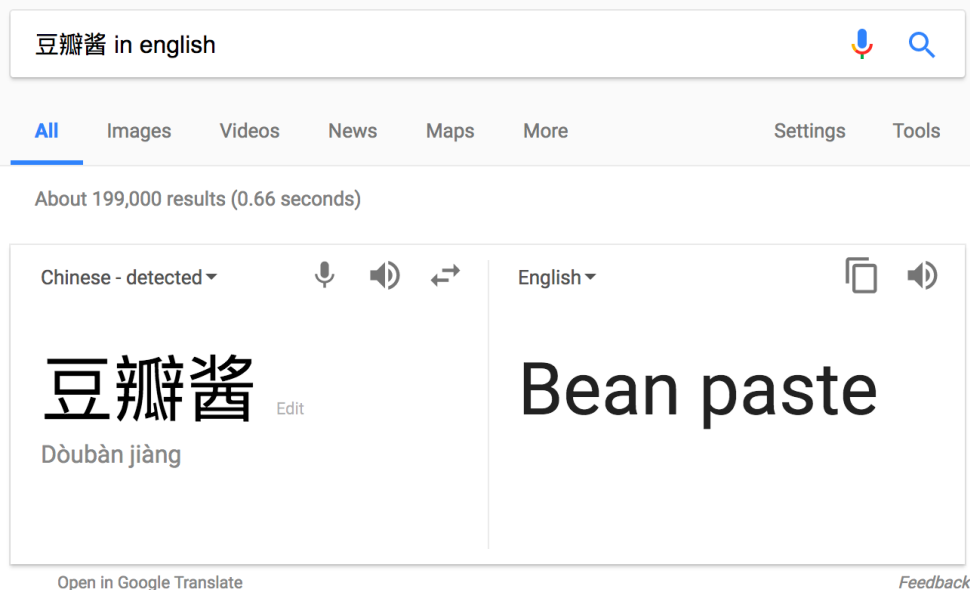


## □ 中文分词，文本对齐，翻译模型，语言模型， Beam Search, ...

# 广泛应用

## □ 常见应用

### ■ Query意图理解



## □ 模板匹配，分类器

# 广泛应用

## □ 常见应用

### ■ Evernote 推荐系统

发挥他们的模范作用、带头作用，通过他们把全连带动起来，通过他们去做政治工作，提高全体指战员的阶级觉悟。有了坚强的党支部的领导，有了坚强的政治工作，就可以做到一呼百应，争先恐后，不怕牺牲，前赴后继。战术、技术也要练好，特别是技术，如果枪打不准，战场上就不能消灭敌人，就不能解决战斗。因此，军事训练不能马虎，党政工作要领导好训练。艺高人胆大，胆大艺更高，部队有了高度的无产阶级觉悟，有了好的战斗作风，再加上过硬的作战本领，就如虎添翼，就可以无敌于天下。

Context

Hide ^

微博 文章 - 傅盛豹变

3/22/15 傅盛豹变 2015年3月  
22日 20:44 . 五年前，傅盛创  
业。天使投资人雷军说，你要...



任正非与华为2012实验室座谈会纪要-IT与  
通讯技术-超级大本营军事论坛-最具影...

9/27/12 任总与2012实验室座谈会纪要 2012年7  
月2日下午，任总与2012实验室干部与专家座  
谈，部分董事会成员、公司各部门领导也应邀...

□ 篇章表示；相似度计算；Local Sensitive Hashing  
文本分类；倒排索引...

# 广泛应用

## □ 常见应用

### ■ 小黄鸡(简易 chatbot)



## □ 关键词匹配, 倒排索引

# 广泛应用

---

## □ 常见应用

### ■ ESLwriter 英文写作助手

▶ \* **(modifies)** impact (collocation)

**significant ... impact (410):** These highly discrepant values had **significant impact** on the mission duration and cost.

**positive ... impact (263):** The game was not very popular, but had a small **positive impact** on reaction times.

**potential ... impact (149):** Following use of the wristband, a number of themes arose with respect to its **potential impact**.

## □ 语法分析，倒排索引，stem (找词根)

# 广泛应用

---

## □ 前沿应用

### ■ Twitter/微博重大事件监测

#### □ 流程：

- 墨西哥湾石油泄漏
- 当地渔民发送Twitter
- 监控系统监测，可靠性分析
- 讲消息提供给签在的股票客户
- 股票客户及时抛出或者做空
- 新闻报道姗姗来迟

#### □ 模板匹配，分类器（哪些是有价值的信息） 社交网络（哪些是值得关注的人） 可信度分析

# 广泛应用

---

## □ 前沿应用

### ■ 医疗诊断书自动生成

#### □ 流程：

■ 医疗图像 + 体检结果 --> 医疗诊断文本

#### □ 规则系统，深度学习



# 广泛应用

---

## □ 前沿应用

### ■ 体育赛事报道自动生成

- “Dylan Tice was hit by a pitch with the bases loaded with one out in the 11th inning, giving the State College Spikes a 9-8 victory over the Brooklyn Cyclones on Wednesday,”
- “Despite the loss, six players for Brooklyn picked up at least a pair of hits. Brosher homered and singled twice, driving home four runs and scoring a couple. The Cyclones also recorded a season-high 14 base hits.”

■ 请关注: <https://automatedinsights.com/>

# 广泛应用

## □ 前沿应用

### ■ 体育赛事报道自动生成



## □ 模板填充，同义词替换，文本对齐

# 广泛应用

---

## □ 前沿应用

### ■ 法律专利生成

□ 律师写一个专利，大概需要3天，~3000美金

□ 使用机器生成相关的专利，小于10分钟，~0 美金

### ■ 模板匹配，分类器

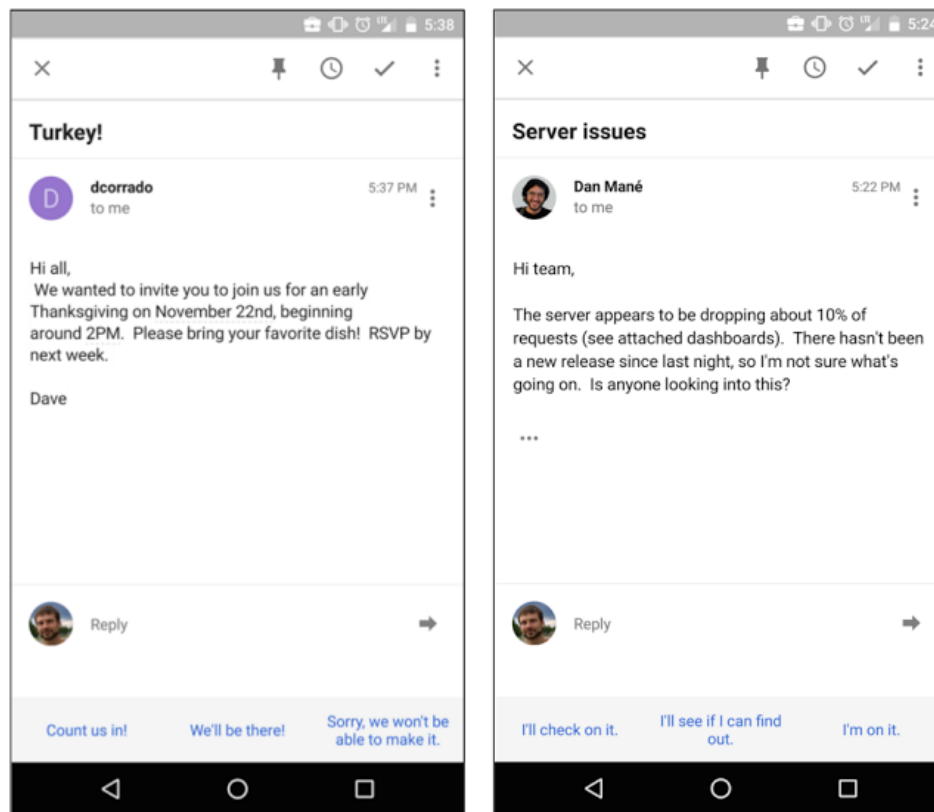
### ■ 模仿Taylor Swift聊天与粉丝互动

### ■ ? 可能是 Seq2Seq 模型

# 广泛应用

## □ 前沿应用

### ■ Gmail自动回复



### ■ Seq2Seq 模型, 语义意图理解

# 广泛应用

## □ 应用行业回顾

行业	难易程度	“钱”途	成熟度
办公自动化	+	++	++++
文体娱乐领域	++	+++	++
财经领域	+++	++++	+++
法律领域	+++	中国++ 美国++++	+
医疗领域	+++++	中国++ 美国++++	+

# 广泛应用

---

## □ 提及技术回顾

语言模型  
自动机  
编辑距离  
文本对齐 (Word Alignment)  
中文分词  
翻译模型  
Beam Search  
模板匹配  
分类器  
相似度计算  
Local Sensitive Hashing  
文本分类

倒排索引  
篇章表示  
关键词匹配  
语法分析  
Stemming  
社交网络  
可信度分析  
规则系统  
同义词识别替换  
Seq2Seq模型  
语义意图理解  
。 。 。

# 广泛应用

---

## □ 提及技术回顾

### 概率模型

语言模型

翻译模型

文本对齐 (Word Alignment)

Seq2Seq模型

### 相似度计算

篇章表示 (feature/embedding)

编辑距离 (Edit Distance)

### Computing Device

自动机

规则系统

分类器

### 搜索技术

关键词匹配

Beam Search

Local Sensitive Hashing

倒排索引

### 语言相关技术

Stemming

同义词识别替换

中文分词

语法分析

语义意图理解

# 广泛应用

---

## □ 提及技术回顾

### 概率模型

语言模型

翻译模型

文本对齐 (Word Alignment)

**Seq2Seq**模型

### 相似度计算

篇章表示 (feature/embedding)

编辑距离 (Edit Distance)

### Computing Device

自动机

规则系统

分类器

### 搜索技术

关键词匹配

**Beam Search**

**Local Sensitive Hashing**

倒排索引

### 语言相关技术

Stemming

同义词识别替换

中文分词

语法分析

语义意图理解



# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

# 挑战何在

---

## □ 人类语言的灵活性

### ■ 一词多义

□ 小红：今晚有NLP的课程，你去么？

□ 小明：我去！！我不去！！

□ 小明是一个老司机了。

□ “又有一大波僵尸出现了” --郭德纲

### ■ 不断更新

□ “十动然拒”

### ■ 人们往往用语不规范

□ 我今天吃了非常好吃的🐔鸭🐟肉！

# 挑战何在

---

## □ 人类语言的灵活性

### ■ 自然语言处理的层次

□ 词法 morphology

□ 语法 syntax

□ 语义 semantics

□ 语用 pragmatics

Stemming

同义词识别替换

中文分词

语法分析

语义意图理解

文章压缩

# 挑战何在

---

## □ 领域隔离

- 语言可以近乎描述领域的事情
- 而各个领域内部的、机器可以理解的“表示”却完全跟不上

### □ 对话系统：

- “Siri, 今天詹姆斯砍下多少分？”=> 篮球领域知识库
- “Siri, 今年有没有厄尔尼诺？”=> 天气领域的知识
- “Siri, 自然语言处理的挑战是什么？”=> 计算机知识
- 任何跨领域的自然语言处理都将困难重重
- 一个新颖的观点：语言自己就是最好的表示

# 挑战何在

---

## ☐ 标注数据获取困难

### ■ 考虑下列问题：

☐ 中文-维吾尔语翻译

☐ 识别句子中被认为和谐的部分：

■ “今天是共（河）产（蟹）党的生日”

### ■ 众包技术，已经大大加快了数据标注的速度

☐ Amazon Mechanical Turk

# 挑战何在

---

## □ 任务评价困难

- 诗歌生成：什么是好的诗歌
- 机器翻译：什么是正确的翻译
- 模范明星说话：怎么才算是模仿像了
- 生成体育赛事报告：怎么才算是个好的报告

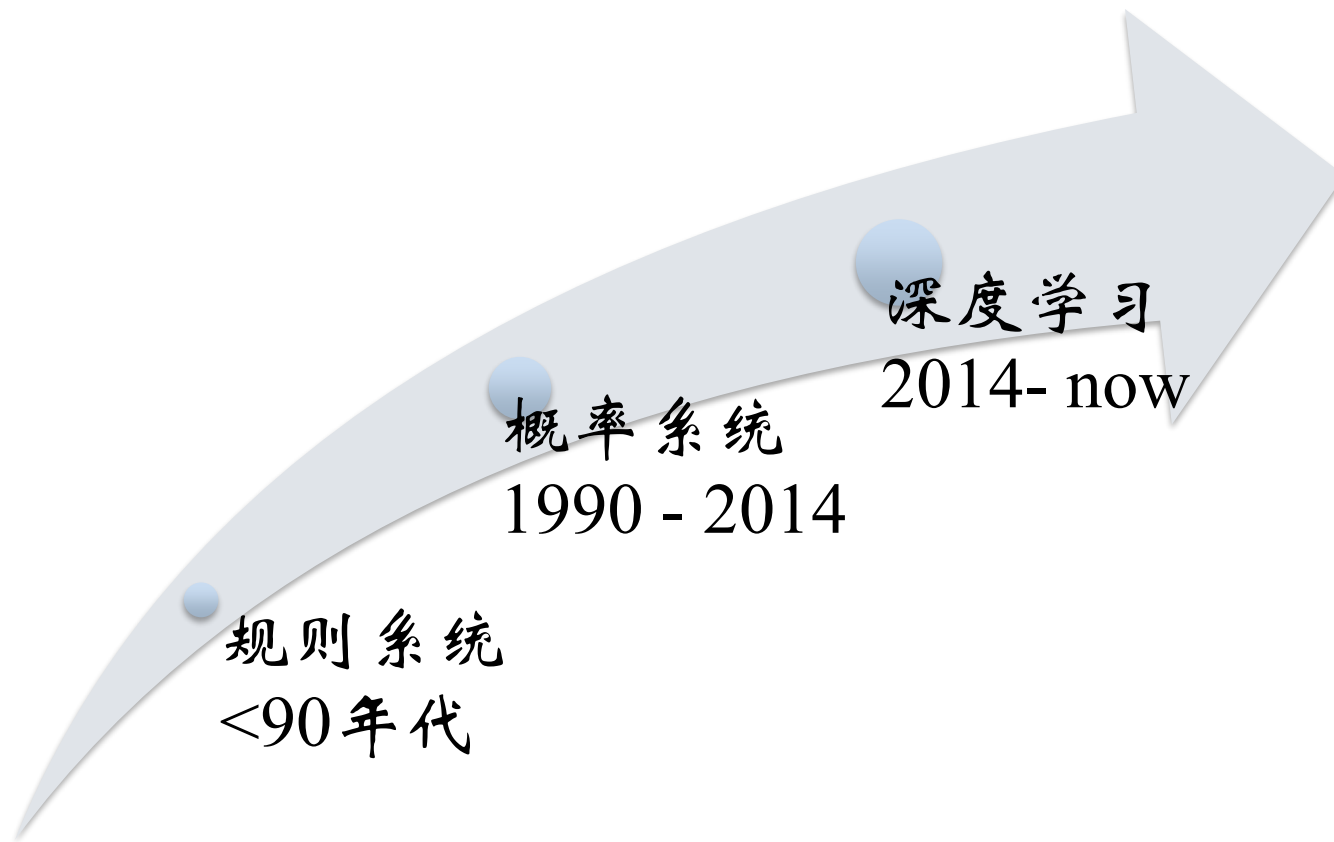
# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

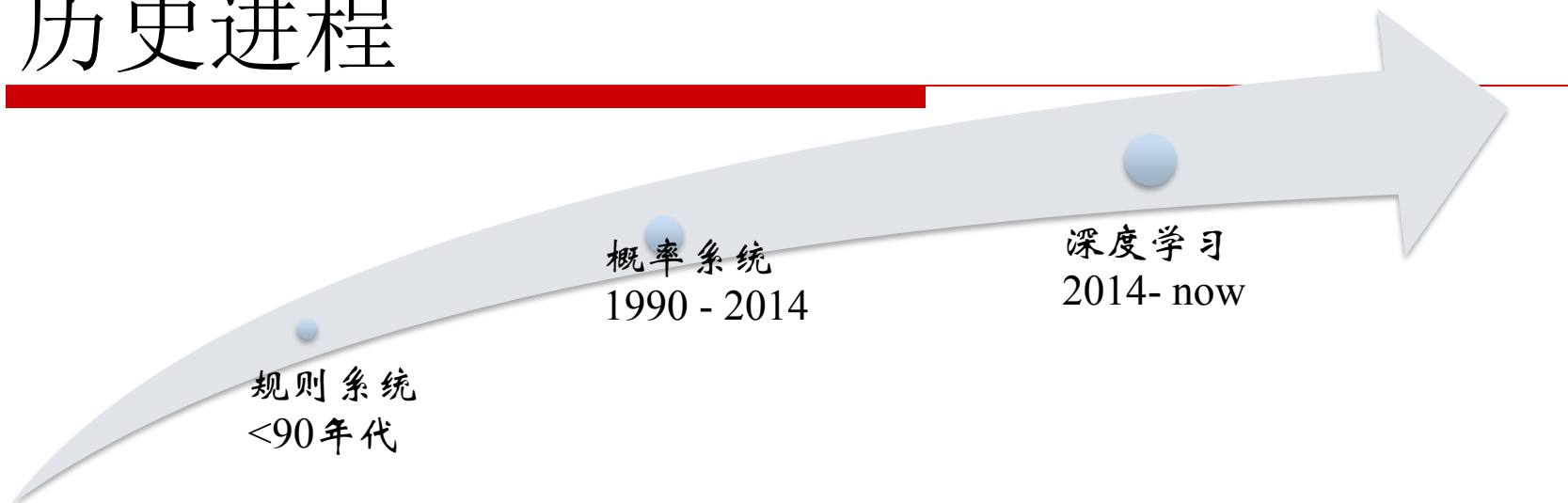
# 历史进程

---





# 历史进程



## 规则系统

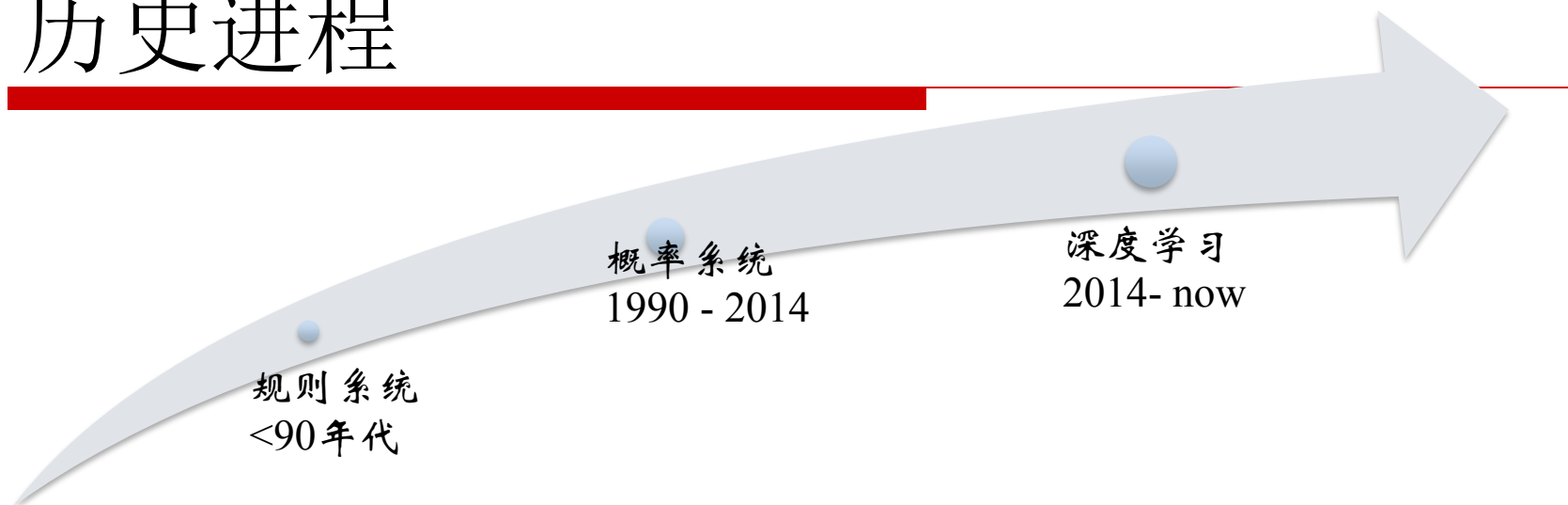
系统的核心，是由**专家**写的一系列规则组成。规则往往都是硬性(Hard)规则，而**非概率性**(soft)规则。

问题：

1. 真实世界的规则远远多于专家制定出来的是规则。
2. 真实世界的规则往往是概率性的规则。

但是，现在在工业界中任然广泛适用：简单，有效，快速。

# 历史进程



## 规则系统

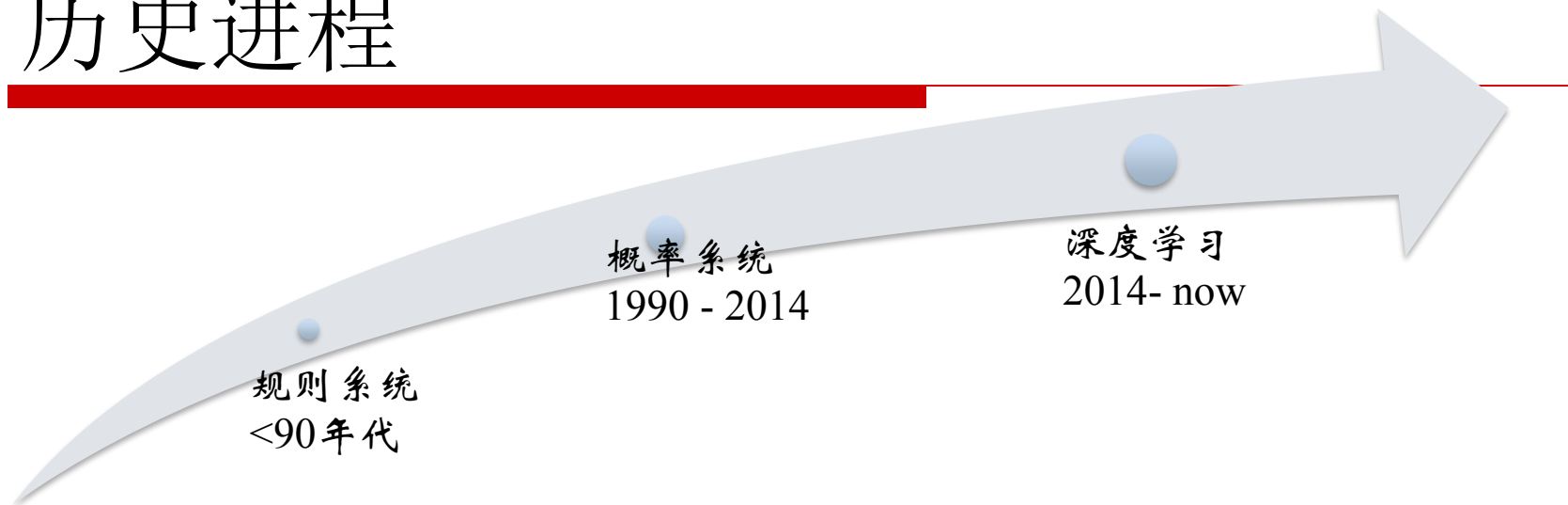
“豆瓣酱用英语怎么说？”

规则：“xx用英语怎么说？” => translate(XX, English)

“我饿了”

规则：“我饿(死)了” => recommend(饭店, 地点)

# 历史进程

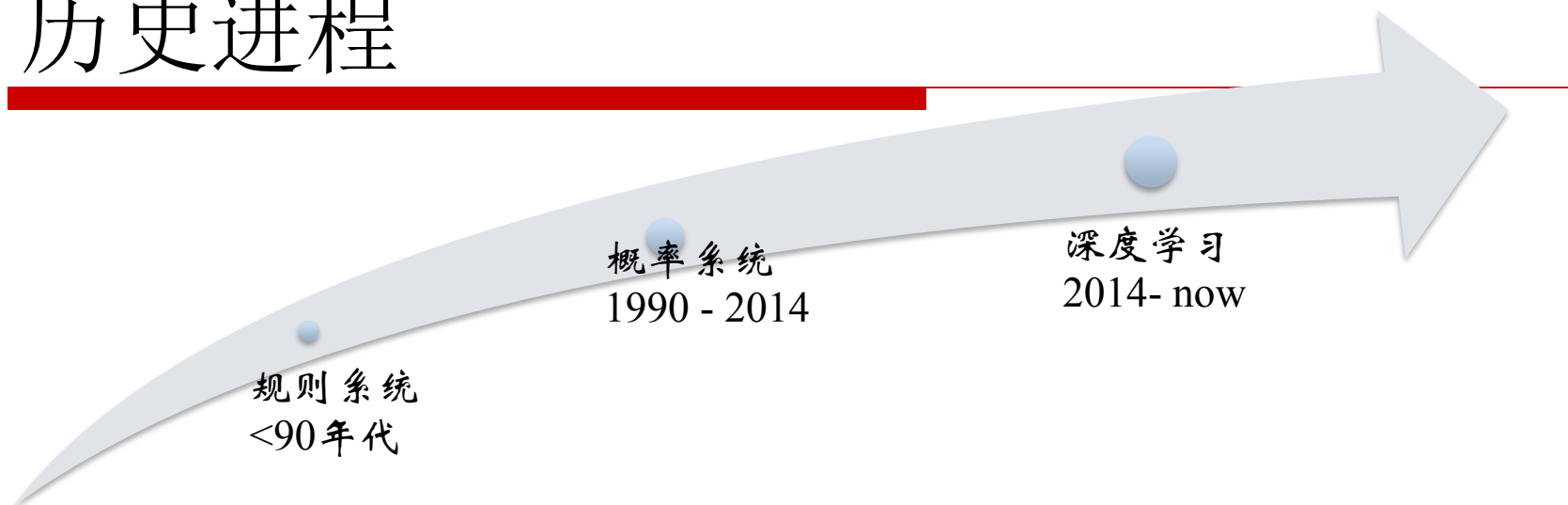


## 概率系统

1. 规则从数据中抽取出来，而不是由专家撰写
2. 规则是有概率的

优点：规则更加贴近于真实事件中的规则，因而效果往往比较好。

# 历史进程



## 概率系统

### A statistical approach to machine translation

**Peter F Brown**, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, **Robert L Mercer**, Paul S Roossin

Computational linguistics 16 (2), 79-85

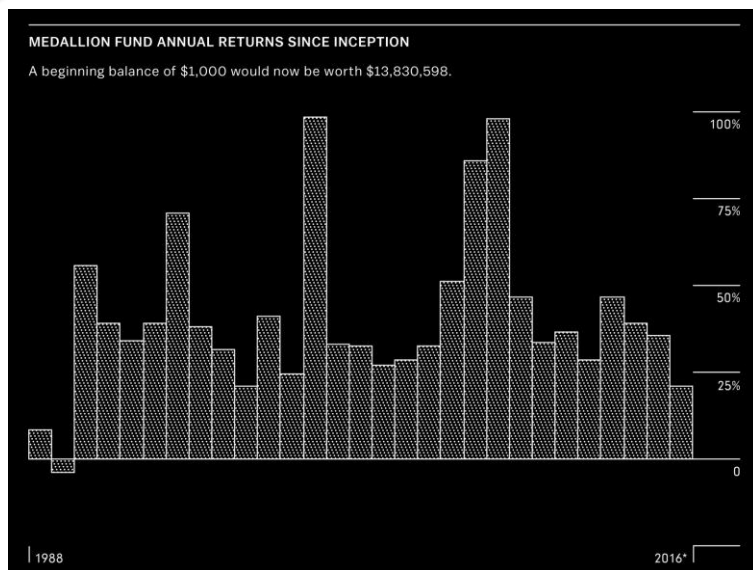
Publication Date: 1990/6/1

# 历史进程

规则系统  
<90年代

概率系统  
1990 - 2014

深度学习  
2014- now



<https://www.bloomberg.com/news/articles/2016-11-21/how-renaissance-s-medallion-fund-became-finance-s-blackest-box>

# 历史进程

---

概率系统的一般工作方式

任务：“豆瓣酱用英语怎么说” => translate(豆瓣酱, Eng)

流程设计

“序列标注问题”

子任务1：找出目标语言 “豆瓣酱用英语怎么说”

子任务2：找出翻译目标 “豆瓣酱用英语怎么说”

收集训练数据

训练数据（子任务1）

“豆瓣酱用英语怎么说”

“茄子用英语怎么说”

“黄瓜怎么翻译成英语”

预处理

分词

“豆瓣酱 用 英语 怎么说”

# 历史进程

---

概率系统的一般工作方式

抽取特征

前后各一个词

0 豆瓣酱： \_ 用

0 用： 豆瓣酱

1 英语： 用 怎么说

0 怎么说： 英语 \_

分类器

“概率规则”

SVM/Neural Network ....

预测

0.1 茄子： \_ 用

0.1 用： 豆瓣酱

0.9 英语： 用 怎么说

0.1 怎么说： 英语 \_

评价

计算准确率

# 历史进程

---

概率系统的一般工作方式

流程设计

收集训练数据

预处理

抽取特征

分类器

预测

评价

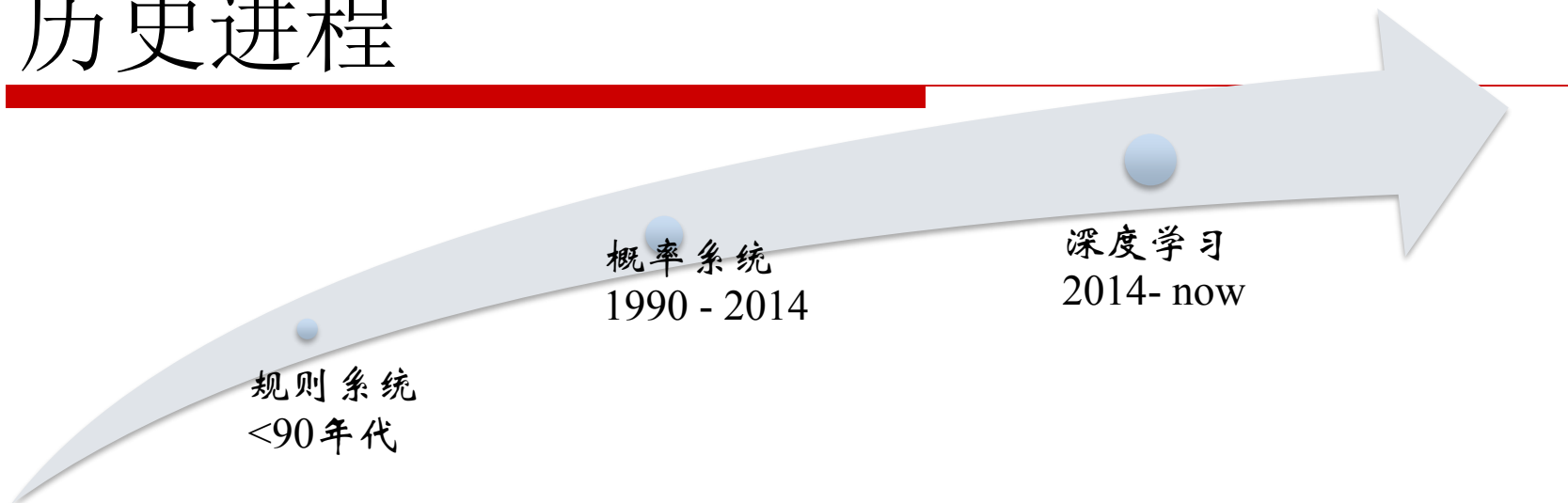


# 历史进程

## 概率系统的一般工作方式

	开发时间	对下限的影响	对上限的影响
流程设计	5%		++
收集训练数据	30%		++++
预处理	20%	+++	
特征抽取	20%	++	++
分类器	15%	+	+
预测	5%		
评价	5%		++

# 历史进程



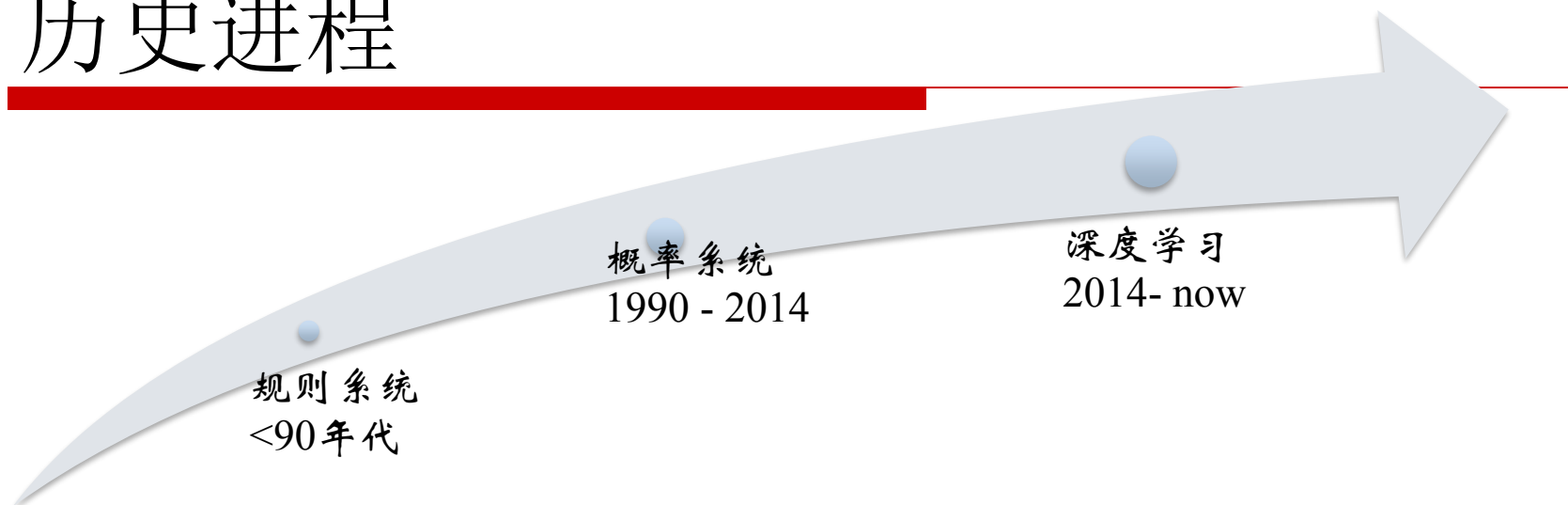
## 概率系统

1. 规则从数据中抽取出来，而不是由专家撰写
2. 规则是有概率的

优点：规则更加贴近于真实事件中的规则，因而效果往往比较好。

缺点：特征是由专家指定的；  
流程是由专家设计的；  
经常存在独立的子任务

# 历史进程



## 深度学习

特征是由专家指定的；      => 特征是由深度学习自己提取的  
流程是由专家设计的；      => 模型结构是由专家设计的  
经常存在独立的子任务      => End-to-End Training: 子任务数量大大减少

# 历史进程

## Sequence 2 Sequence Model

基本分类器

诗歌生成

情感分析

机器翻译

序列标注

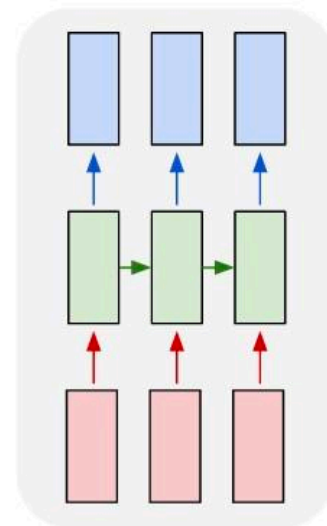
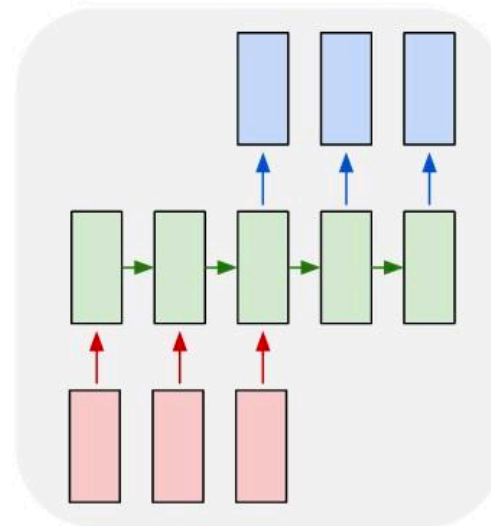
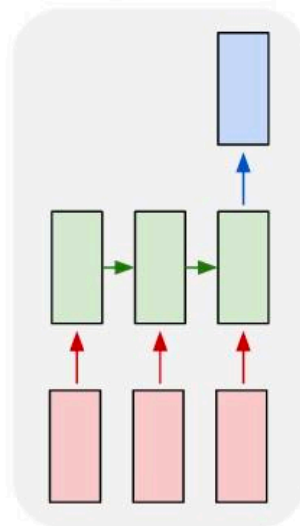
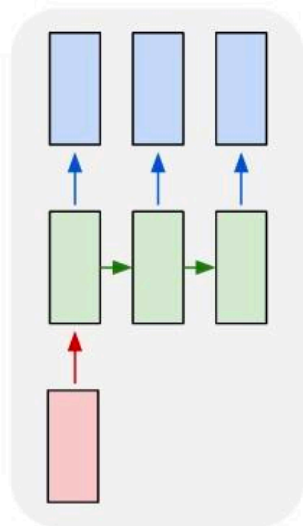
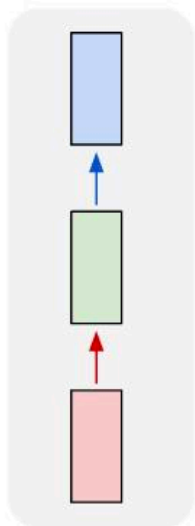
one to one

one to many

many to one

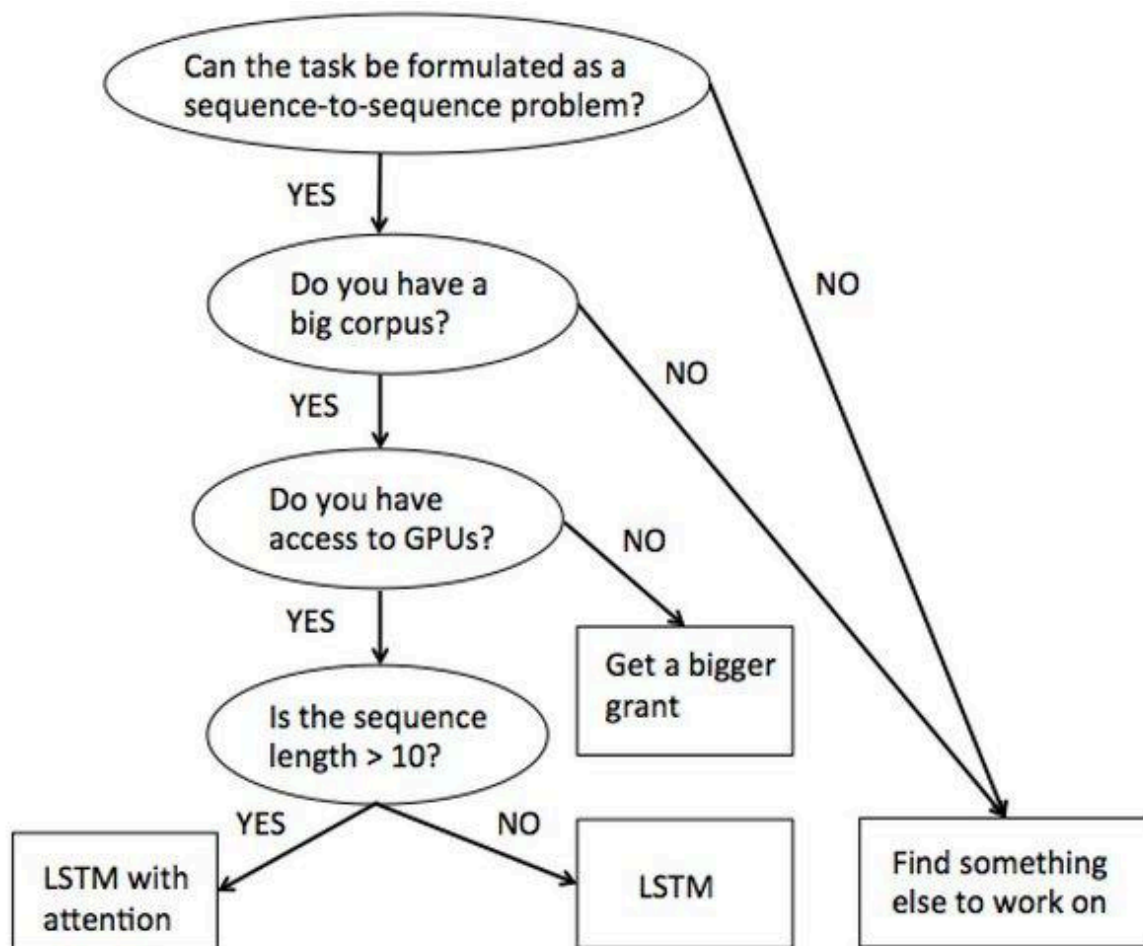
many to many

many to many



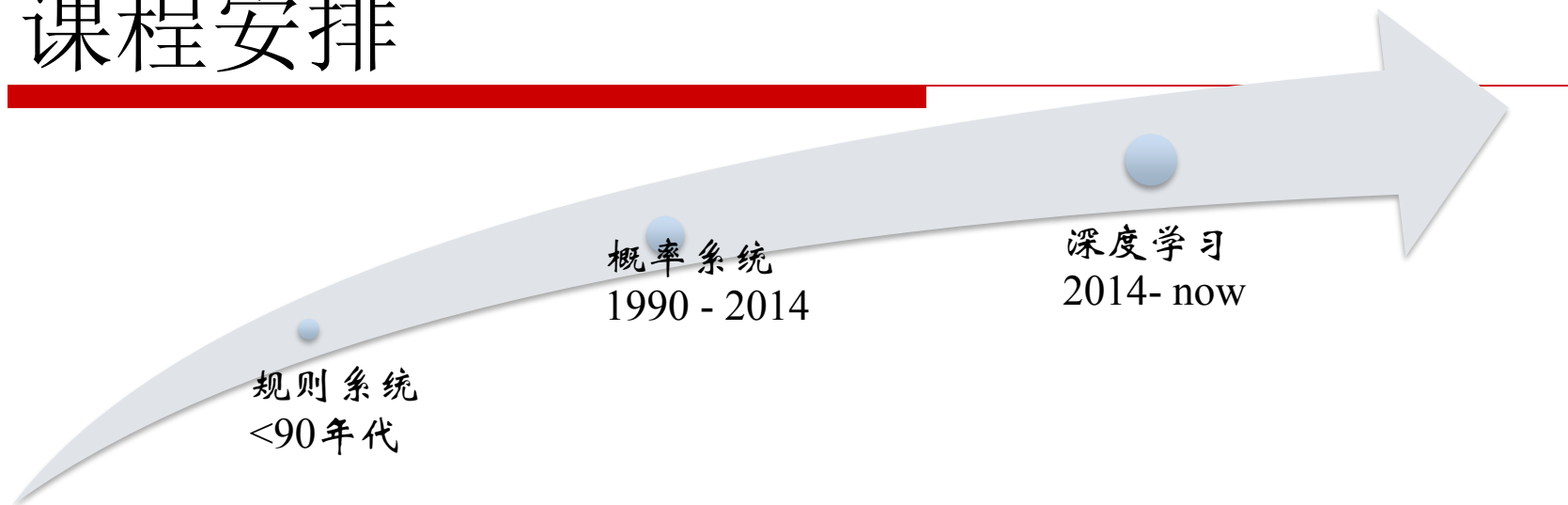
<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

# 历史进程



趋势： 学习的门槛降低了。。。。。

# 课程安排



基本分类器（1课时）

经典序列模型（3课时）

HMM/CRF/EM

自动机

语言模型

神经序列模型（5课时）

概念介绍

Tensorflow实现

机器翻译

提速

其他的应用

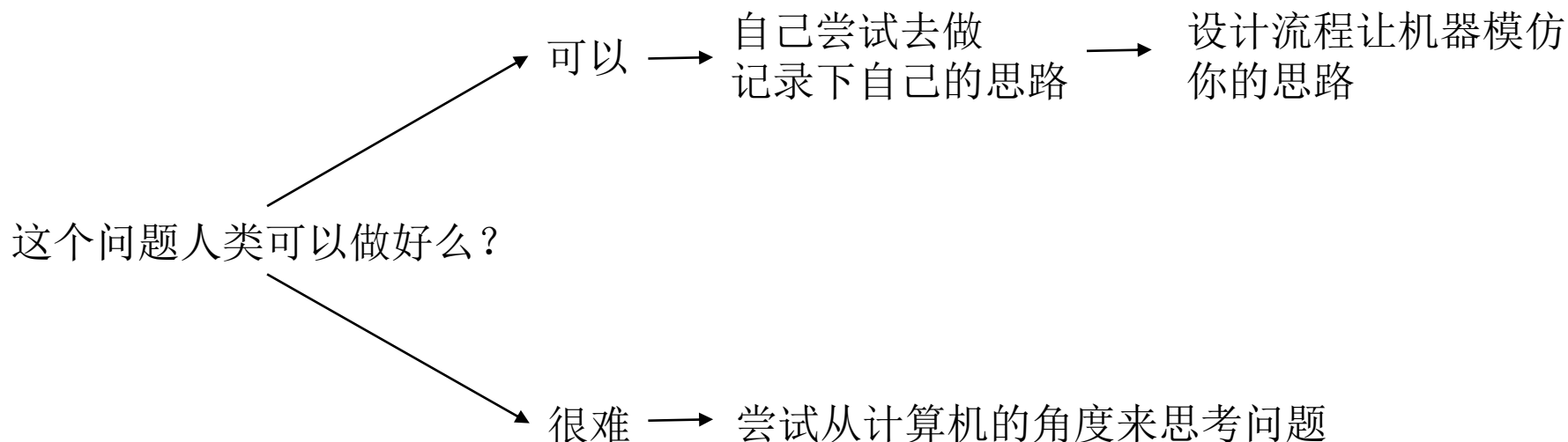
# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

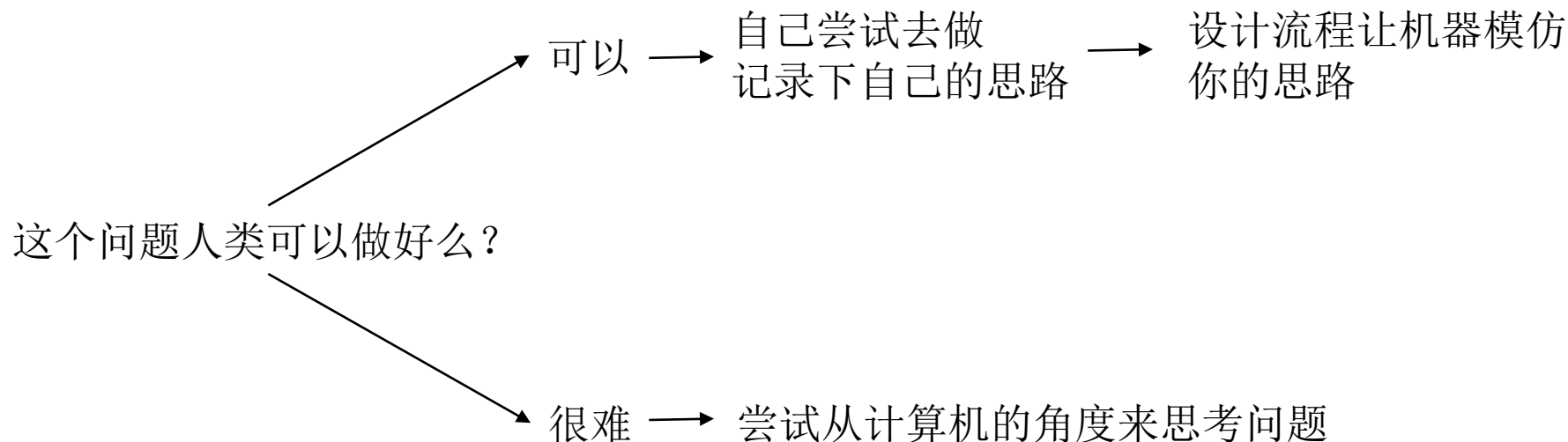
# 解决自然语言处理的一般思路

---





# 解决自然语言处理的一般思路



任务：下节课的PPT我要用电脑自动生成！

我的思路： 写大纲 -> 每个知识要点可能要参考一个Wikipedia -> 举例子，画流程图

机器的模块： 写大纲： “分类器” → 搜索相关论文， 选择前五的不重复的论文  
每个知识点： 相关论文的章节， 第一段， Wikipedia的第一段  
例子和流程图： 直接截取 Wikipedia 的相关图片

# 提纲

---

- 基本概念
- 广泛应用
- 挑战何在
- 历史进程
- 一般思路
- 基本工具
- 一点期望

# 一些常用的工具

---

## ☐ Bash Script

- wc/sed/awk/grep/sort/uniq/paste/cat/head/tail

- ☐ 一个很大的txt, 30s内找出出现次数最多的前10个词汇

- ☐ 查看第30行到第40行的数据

- ☐ ...

## ☐ Python

- 当bash不是那么熟练的时候, 或者处理稍微复杂的问题的时候

# 一些常用的工具

---

## □ Stanford Core NLP

- 语义分析

## □ NLTK

- 句子划分、读取语义树

## □ Tensorflow

- 推荐大家使用最新版的Tensorflow 1.2
- 推荐使用Linux
- python2.7 (3也可以, 感觉没啥差别)

# 对大家的期望

---

□ 每节课只要有一个实实在在的收获就够了

□ 问卷调查



# 联系我们

---

## 小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop

