机器学习程序设计基础
1. 引言

陈湘萍

中山大學
SUN YAT-SEN UNIVERSITY

哪些问题可以设计一个机器学习算法来回答?

哪些问题可以设计一个机器学习算法来回答?

年龄：青年？中年？老年？

授课风格：枯燥？风趣？

对学生要求：严格？宽松

老师

课程

内容：丰富？简单？

是否适合我：是？否

我的学习成绩：优秀？及格？不及格？

哪些问题可以设计一个机器学习算法来回答?

年龄：青年?
中年? 老年?

老师

课程

哪些问题可以设计一个机器学习算法来回答?

年龄：青年?
中年? 老年?

老师

课程

如果只有一张图片....

https://cn.how-old.net/#

机器学习的概念

经验

预测

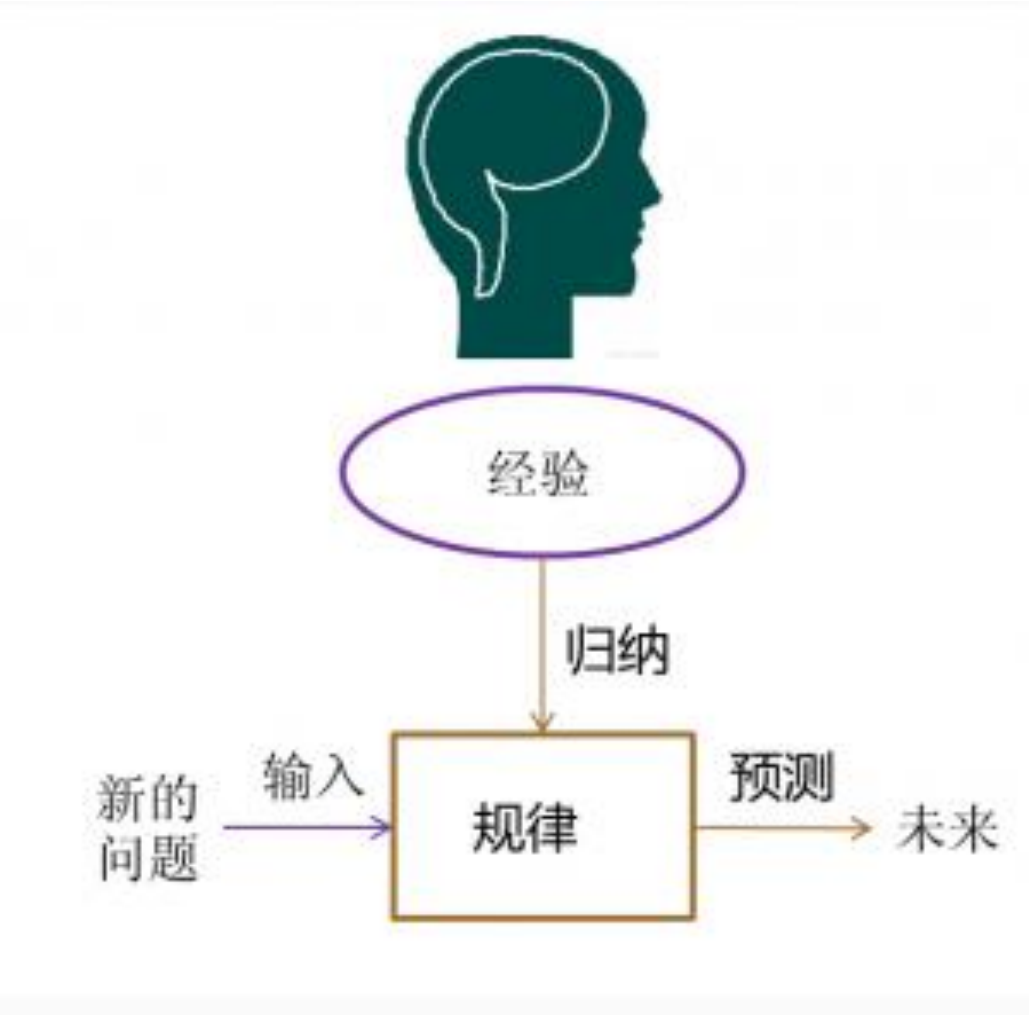
机器学习的概念

预测

模型

数据

机器学习vs.人类学习



机器学习的概念

- 机器学习领域奠基人之一、美国工程院院士T. Mitchell教授在其经典教材《Machine Learning》中所给出的机器学习经典定义为“利用经验来改善计算机系统自身的性能”。
  - 经验对应于历史数据，如互联网数据、科学实验数据等。
  - 系统对应于数据模型，如决策树、支持向量机等。
  - 性能则是模型对新数据的处理能力，如分类和预测性能等。
  - 机器学习的根本任务是数据的建模与智能分析。

哪些问题可以设计一个机器学习算法来回答?

年龄：青年?
中年? 老年?

授课风格:
枯燥? 风趣?

老师

课程

如果还有其他信息....

名字
职业
.....

逻辑清晰的问题能用规则很高效和准确地解决

哪些问题可以设计一个机器学习算法来回答?

老师

课程

内容：丰富？简单？

是否适合我：是？否

对学生要求：严格？宽松

不能用于无法建立统一（或者大多数人认同）的判断标准的问题

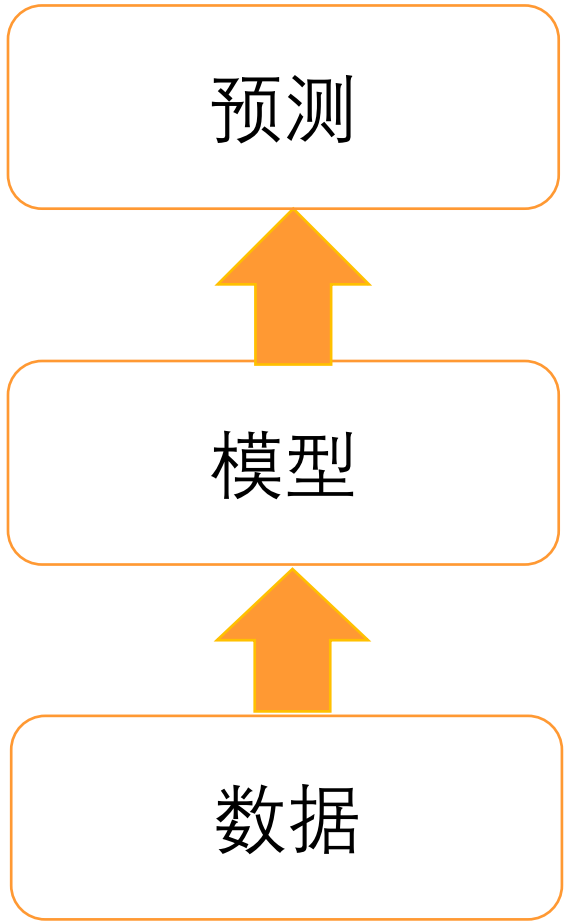
哪些问题可以设计一个机器学习算法来回答?

预测

课程

我的学习成绩：优秀？及格？不及格？

哪些问题可以设计一个机器学习算法来回答?



课程

学号	姓名	学院	年级专业	班号
17305011	李聪彧	地理科学与规划学院	17级城乡规划	201922274
17329038	何选	化学学院	17级高分子材料与工程	201922274
17329134	张梓轩	化学学院	17级化学	201922274
17351070	田彩玉	物理学院	17级光电信息科学与工程（光信息科学与技术方向）	201922274
17358030	赵雅星	哲学系	17级逻辑学	201922274
18306079	吾买尔江·吾布力卡斯木	地理科学与规划学院	18级地理科学类	201922274
18315015	丁杨	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18315046	李宇轩	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18316074	刘韵婷	管理学院（创业学院）	18级会计学（管理学院）	201922274
18316096	舒美琦	管理学院（创业学院）	18级会计学（管理学院）	201922274
18327044	柯乔木	物理学院	19级物理学	201922274
18328059	林禹宏	化学学院	18级化学类	201922274
18328093	覃及佳	化学学院	18级化学类	201922274
18332039	于浩龙	岭南学院	18级管理科学（岭南学院）	201922274
18333243	张梓艳	岭南学院	18级国际商务（岭南学院）	201922274
18333258	周雨乐	岭南学院	18级管理科学（岭南学院）	201922274
18333267	庄子安	岭南学院	18级金融学（岭南学院）	201922274
18335045	杨芬	社会学与人类学学院	18级考古学	201922274
18337057	韩孚嘉	生命科学学院	18级生物技术	201922274
18337111	刘心怡	生命科学学院	18级生物技术	201922274
18339003	陈俊睿	岭南学院	18级金融学（岭南学院）	201922274
18343006	陈佳新	数学学院	8级数学类（广州）	201922274
18343012	陈秋亦	数学学院	8级数学类（广州）	201922274
18343034	管欣怡	数学学院	8级数学类（广州）	201922274
18343070	李洋	数学学院	8级数学类（广州）	201922274
18343123	佟宇博	数学学院	8级数学类（广州）	201922274
18343131	王齐豫	数学学院	8级数学类（广州）	201922274
18344090	孙煜宁	物理学院	19级光电信息科学与工程（光信息科学与技术方向）	201922274
18351075	王晰宁	物理学院	18级光电信息科学与工程（光信息科学与技术方向）	201922274
18352051	昝雪枫	物理学院	18级物理学	201922274
18352060	郑大鹏	物理学院	18级物理学	201922274
18352064	周恩泽	物理学院	18级物理学	201922274
18353015	郭瀚中	岭南学院	18级金融学（岭南学院）	201922274
18358005	冯烨	哲学系	18级逻辑学	201922274
18359028	王煦喆	哲学系	18级哲学（广州）	201922274
18980063	李睿捷	化学学院	18级临床医学（八年制）-化	201922274
19306060	刘旭锋	地理科学与规划学院	19级地理科学类	201922274
19306124	朱方杰	地理科学与规划学院	19级地理科学类	201922274
19313305	周梦霞	管理学院（创业学院）	19级工商管理类（管理学院）	201922274
19323086	梅云皓	化学学院	19级化学类	201922274
19323087	孟子洋	化学学院	19级化学类	201922274

机器学习基本概念介绍

人工智能

人工智能：Artificial Intelligence

英文缩写：AI

机器人仅仅是人工智能的一个分支

定义1 智能机器(intelligent machine)能够在各类环境中自主地或交互地执行各种拟人任务的机器。

定义2 人工智能(学科)：计算机科学中涉及研究、设计和应用智能机器的一个分支。它的近期主要目标在于研究用机器来模仿和执行人脑的某些智力功能，并开发相关理论和技术。

定义3 人工智能(能力)：智能机器所执行的通常与人类智能有关的智能行为，如判断、推理、证明、识别、感知、理解、通信、设计、思考、规划、学习和问题求解等思维活动。

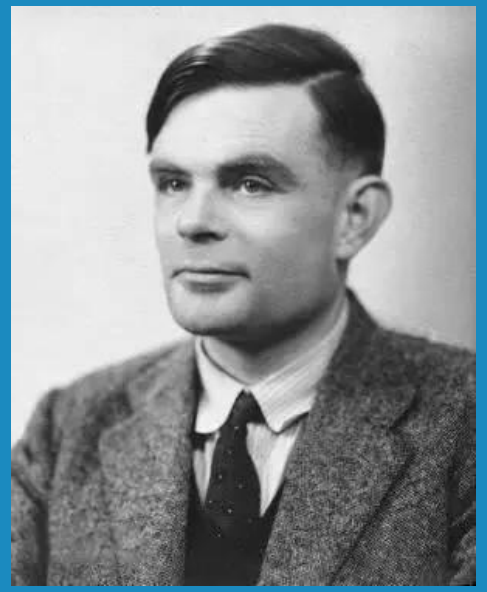
人工智能的基础

- 哲学：标出了AI的大部分重要思想
- 数学：使AI成为一门规范科学 数学形式化
- 神经科学：网络，并行处理
- 心理学：认知理论
- 计算机工程：AI的“载体”
- 语言学：知识表示、语法
- 言学、神经生理学、心理学、数学、哲学

人工智能的评价标准

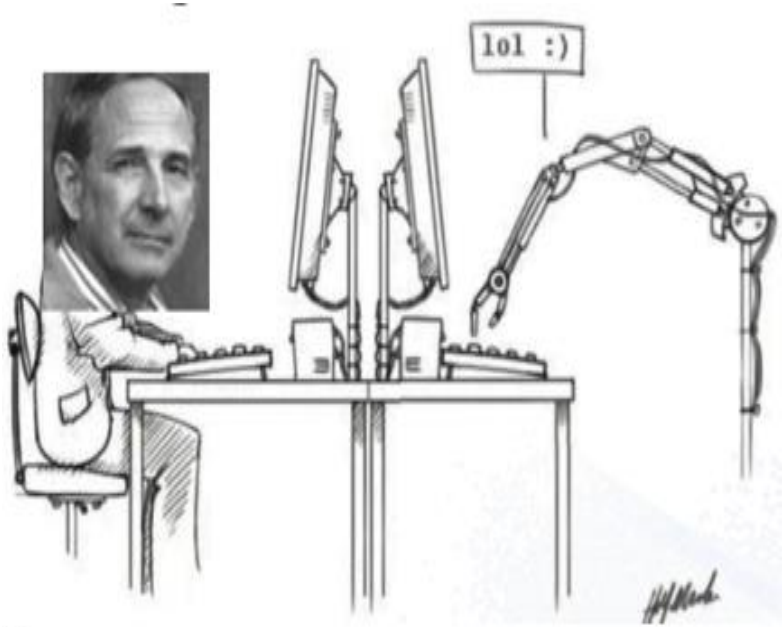
阿兰·图灵

英国数学家、逻辑学家，被视为计算机科学之父。



图灵测试

阿兰·图灵在1950年发表的一篇名为《计算机器与智能》的论文，提出著名的“图灵测试”，测试者在与被测试者（一个人和一台机器）隔开的情况下，通过一些装置（如键盘）向被测试者随意提问。如果机器能够让30%的测试人相信它是人类，那么这台计算机就可以被认为具有人类的思考能力。



怎么测试

- 问：你会下国际象棋吗？
- 答：是的。
- 问：你会下国际象棋吗？
- 答：是的。
- 问：请再次回答，你会下国际象棋吗？
- 答：是的。

怎么测试

- 问： 你会下国际象棋吗?
- 答：是的。
- 问：你会下国际象棋吗?
- 答：是的，我不是已经说过了吗?
- 问：请再次回答，你会下国际象棋吗?
- 答：你烦不烦，干嘛老提同样的问题。

人工智能的发展历程

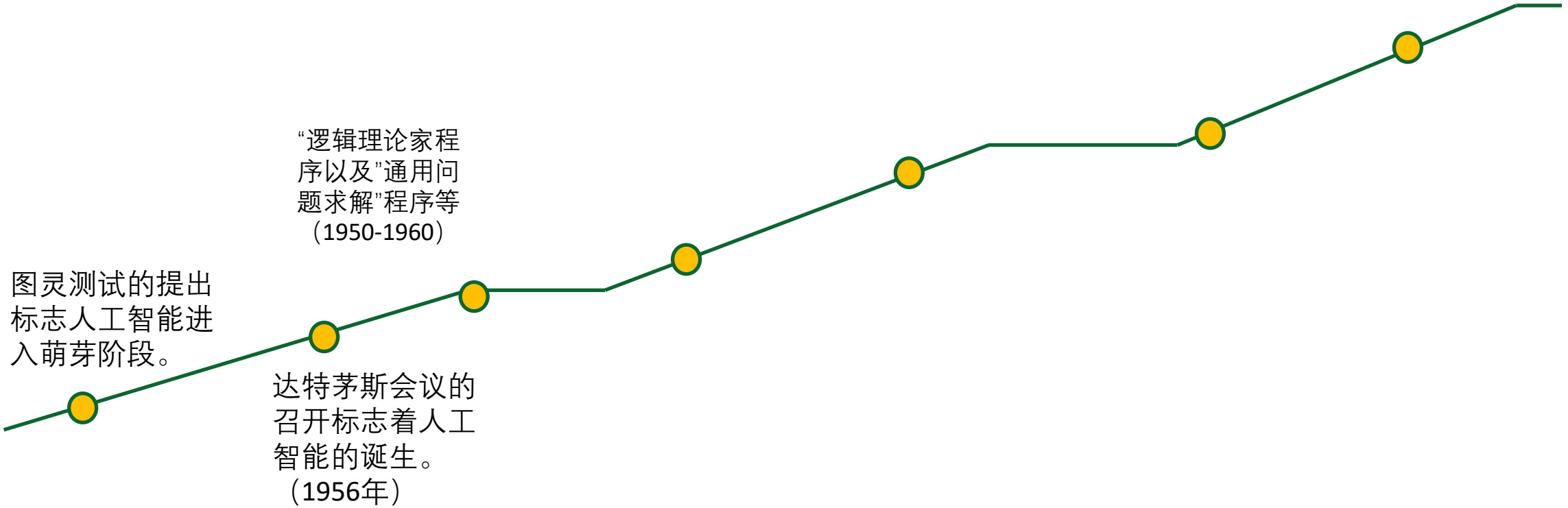
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第一阶段：推理期

1956-1960s: Logic Reasoning

- 出发点：“数学家真聪明！”
- 主要成就: 自动定理证明系统 (例如，西蒙与纽厄尔的“Logic Theorist”系统)

渐渐地，研究者们意识到，仅有逻辑推理能力是不够的 …

赫伯特 西蒙
(1916-2001)
1975年图灵奖

阿伦 纽厄尔
(1927-1992)
1975年图灵奖

人工智能的发展历程

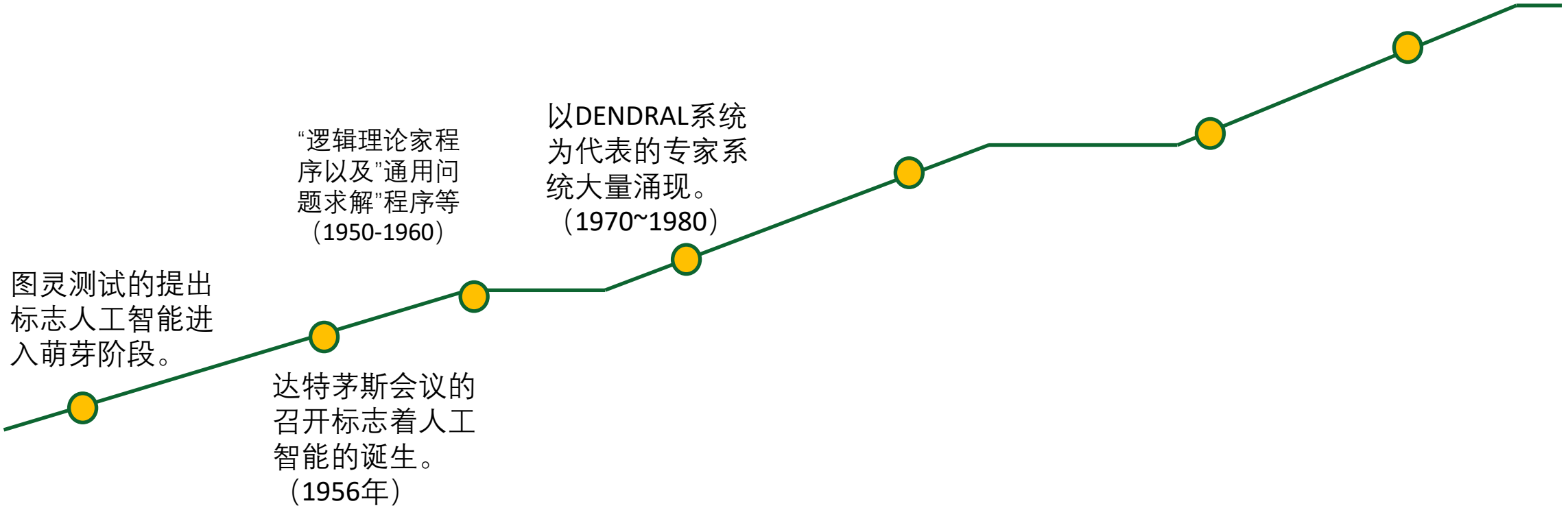
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第二阶段：知识期

1970s -1980s: Knowledge Engineering

- 出发点:
- 主要成就: 专家系统 (例如，费根鲍姆等人的“DENDRAL”系统)

1994年图灵奖
爱德华 费根鲍姆
(1936-)

渐渐地，研究者们发现，要总结出知识再“教”给系统，实在太难了 …

人工智能的发展历程

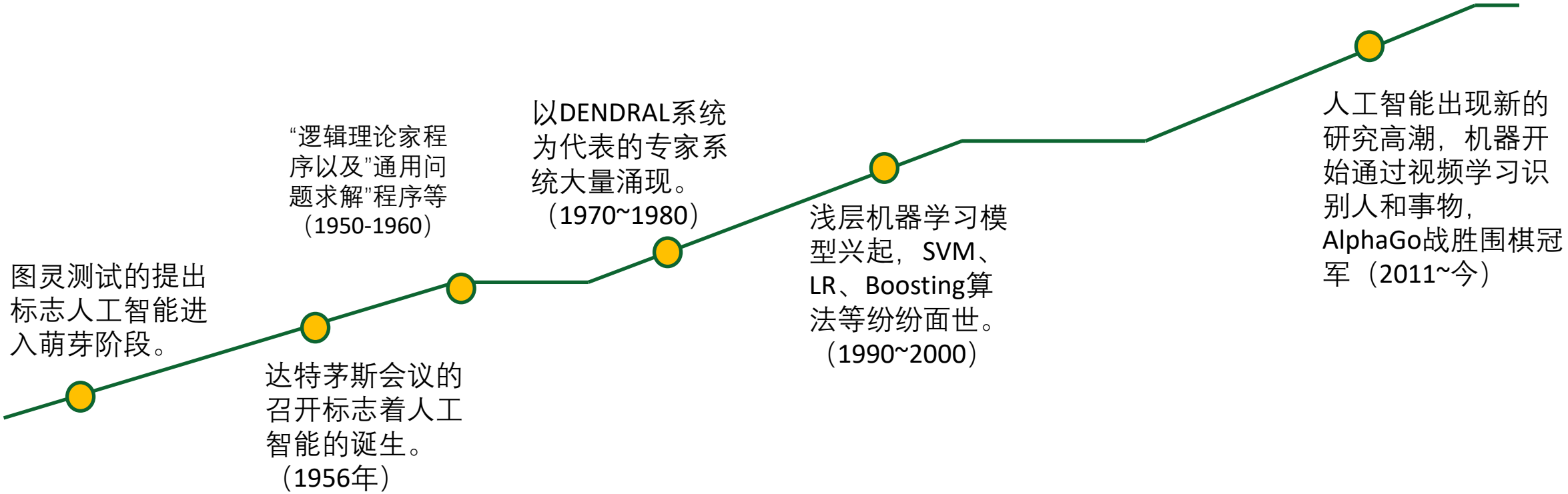
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第三阶段：学习期

1990s -now: Machine Learning

- 出发点: “让系统自己学！”
- 主要成就: ……

机器学习是作为“突破知识工程瓶颈”
之利器而出现的

恰好在20世纪90年代中后期，人类发现自己淹没在数据的汪洋中，对自动数据分析技术——机器学习的需求日益迫切

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

- **1980s：登上历史舞台**
- **机器学习作为一支独立的力量登上了历史舞台。在这之后的10年里出现了一些重要的方法和理论，典型的代表是：**
  - **1980夏-在卡内基梅隆举行第一届机器学习研讨会(IWML)；**
  - **1983第一本机器学习的专著《机器学习-一种人工智能的途径》；**
  - **1984-分类与回归树（CART）**
  - **1986-第一个期刊《Machine Learning》创刊**
  - **1986-反向传播算法**
  - **1989-卷积神经网络**

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

1990-2012：走向成熟和应用

在这20多年里机器学习的理论和方法得到了完善和充实，可谓是百花齐放的年代。代表性的重要成果有：

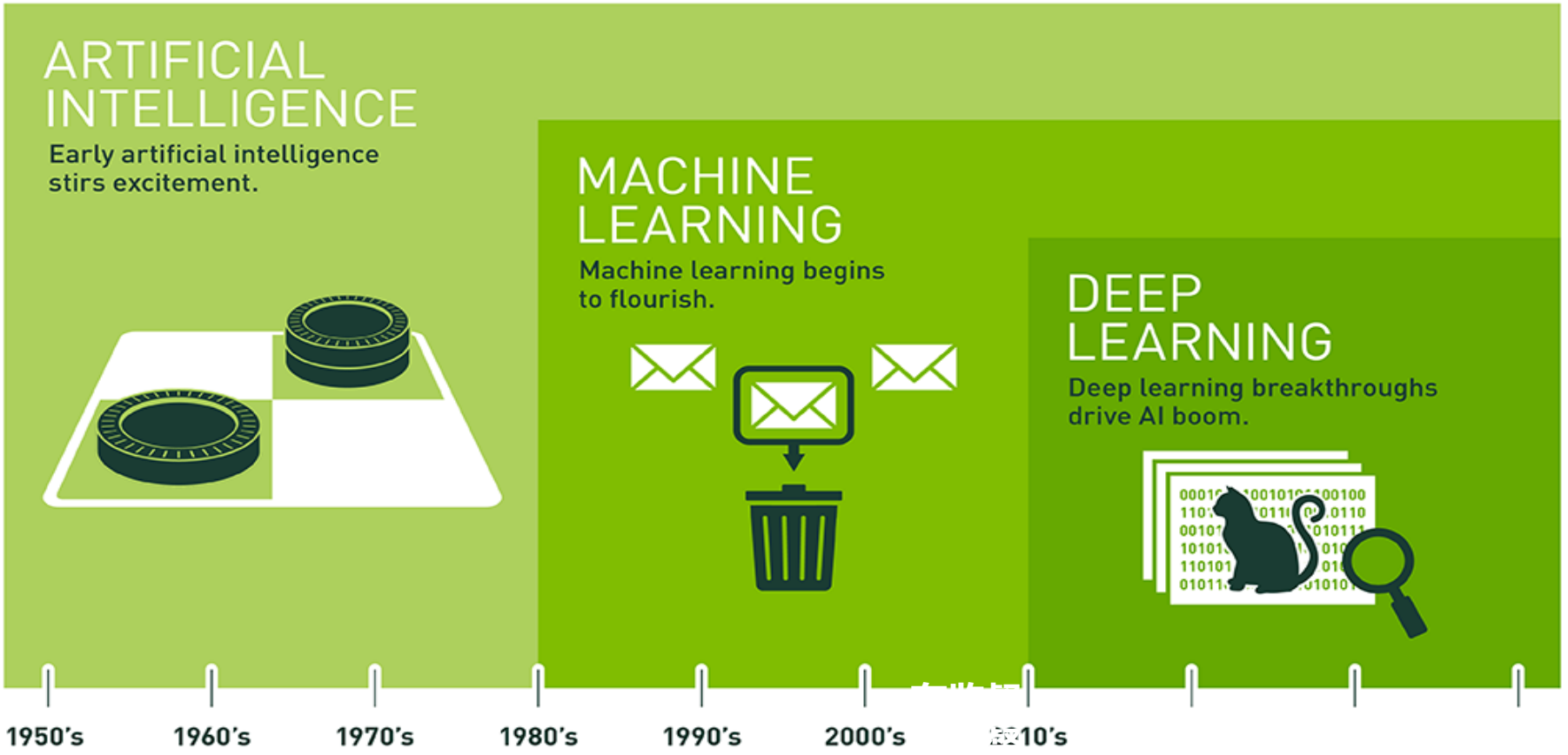
- **1995：支持向量机（SVM）**
- **1997：AdaBoost算法**
- **1997：循环神经网络（RNN）和LSTM**
- **2000：流形学习**
- **2001：随机森林**

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

- **2012：深度学习时代-神经网络卷土重来**
  - **深度卷积神经网络：AlexNet首先在图像分类问题上取代成功，随后被用于机器视觉的各种问题上**
  - **循环神经网络：语音识别，自然语言处理**
  - **深度强化学习：策略、控制类问题，典型的代表是AlphaGo**
  - **……**

几个概念的关系

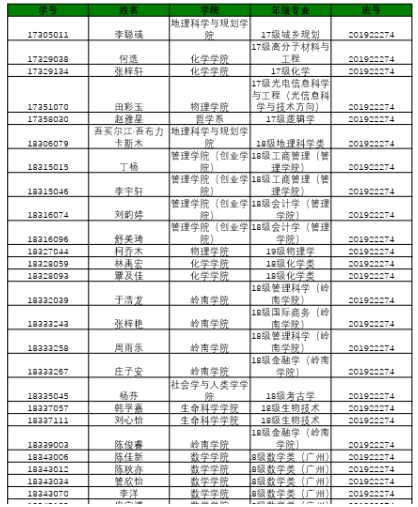


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

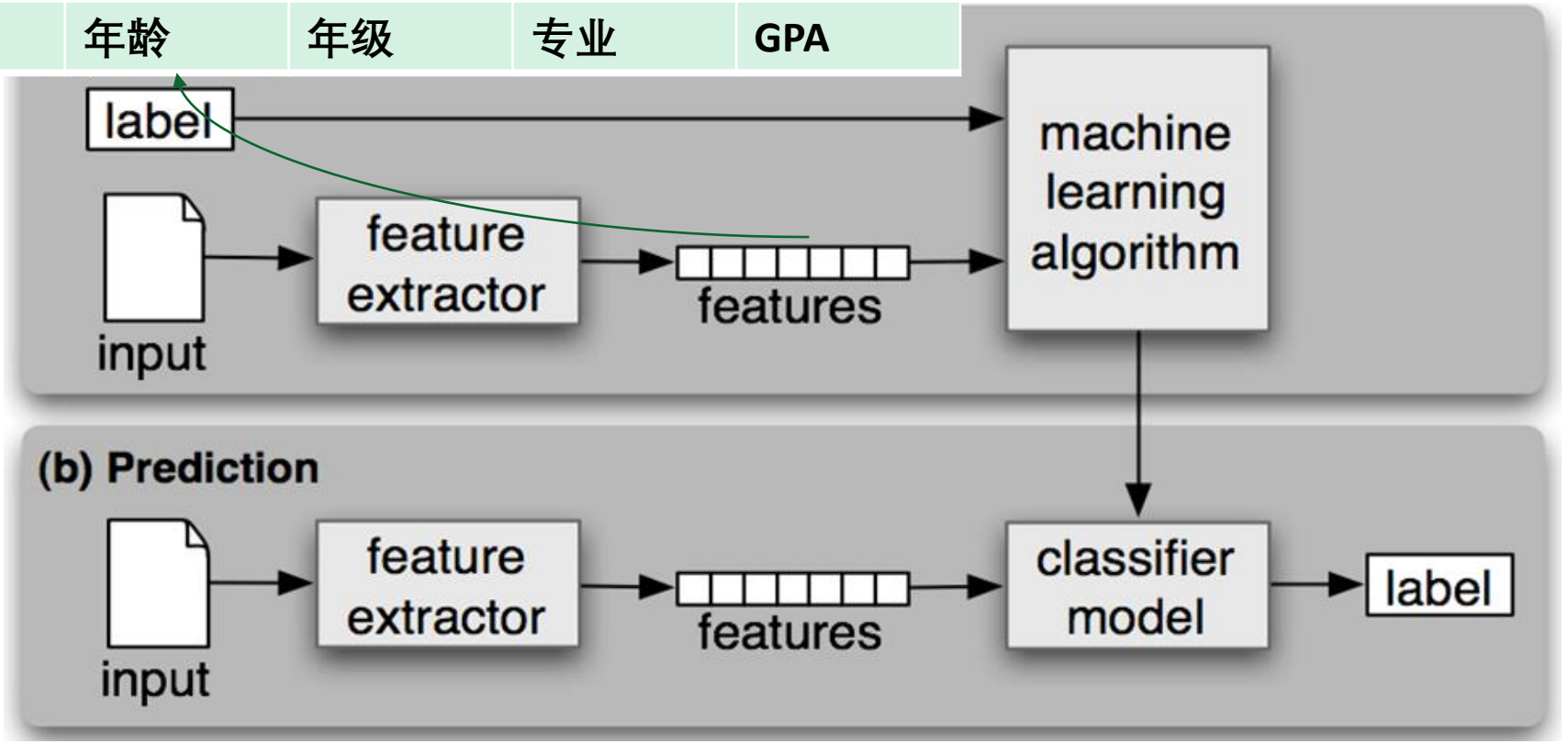
机器学习

- 机器学习主要是设计和分析让计算机可以自动“学习”的算法。学习算法是一类从数据中自动分析获得规律，利用规律对未知数据进行预测的算法。

学号	姓名	性别	年龄	年级	专业	GPA

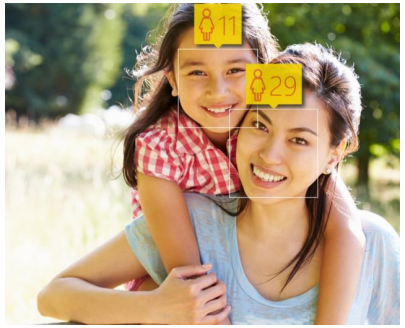


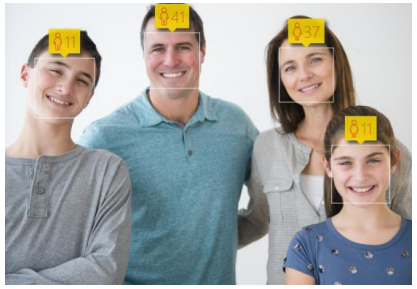
学号	姓名	学院	年级专业	班号
17305011	李聪颖	地理科学与规划学院	17级城乡规划	201922274
17329038	何浩	化学学院	17级高分子材料与工程	201922274
17329134	张梓轩	化学学院	17级化学	201922274
17351070	田彩玉	物理学院	17级光电信息科学与工程（光信息科学与技术方向）	201922274
17358030	赵雅星	药学系	17级医确学	201922274
18306079	喜买尔江·喜布力卡斯木	地理科学与规划学院	18级地理科学类	201922274
18315015	丁杨	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18315046	李宇轩	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18316074	刘韵婷	管理学院（创业学院）	18级会计学（管理学院）	201922274
18316096	舒奚琦	管理学院（创业学院）	18级会计学（管理学院）	201922274
18327044	科乔木	物理学院	19级物理学	201922274
18328059	林禹宏	化学学院	18级化学类	201922274
18328093	覃及佳	化学学院	18级化学类	201922274
18332039	于浩龙	岭南学院	18级管理科学（岭南学院）	201922274
18332243	张梓艳	岭南学院	18级国际商务（岭南学院）	201922274
18332258	周雨乐	岭南学院	18级管理科学（岭南学院）	201922274
18332267	庄子安	岭南学院	18级金融学（岭南学院）	201922274
18335045	杨芬	社会学与人类学学院	18级考古学	201922274
18337057	钟昇嘉	生命科学学院	18级生物技术	201922274
18337111	刘心怡	生命科学学院	18级生物技术	201922274
18339003	陈俊豪	岭南学院	18级金融学（岭南学院）	201922274
18343006	陈佳新	数学学院	8级数学类（广州）	201922274
18343012	陈秋亦	数学学院	8级数学类（广州）	201922274
18343034	管欣怡	数学学院	8级数学类（广州）	201922274
18343070	李洋	数学学院	8级数学类（广州）	201922274

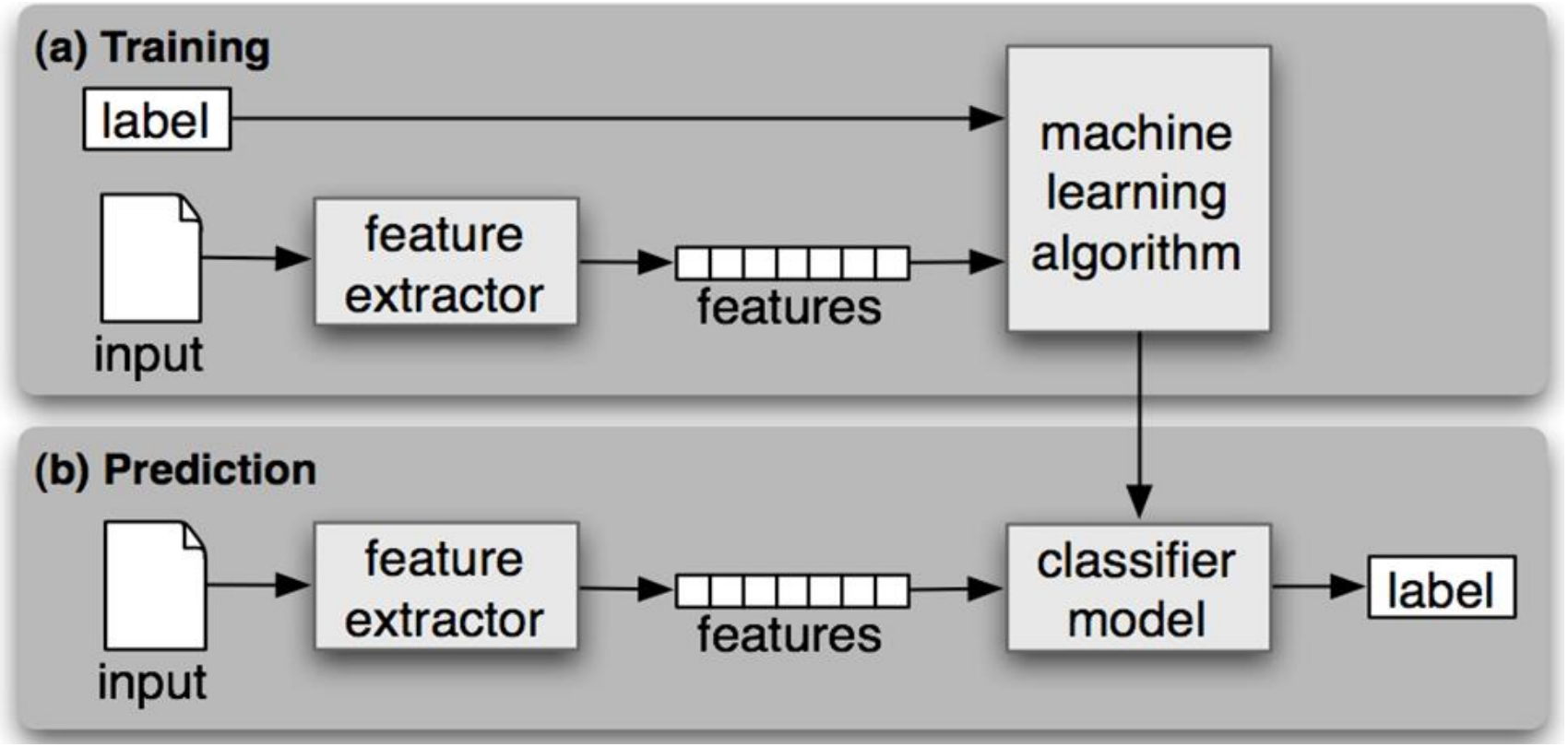


机器学习

- 机器学习主要是设计和分析让计算机可以自动“学习”的算法。学习算法是一类从数据中自动分析获得规律，利用规律对未知数据进行预测的算法。

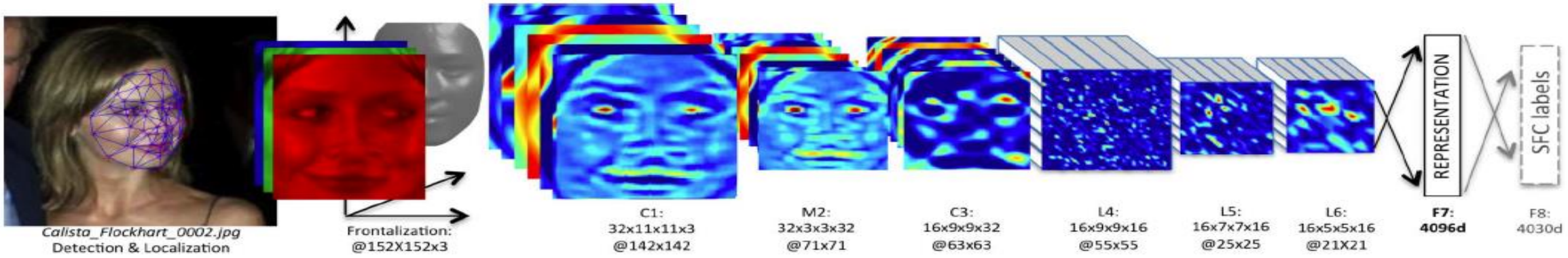






深度学习

- Learn a *feature hierarchy* all the way from pixels to classifier
- Each layer extracts features from the output of previous layer
- Train all layers jointly



例1：帮助奥巴马胜选

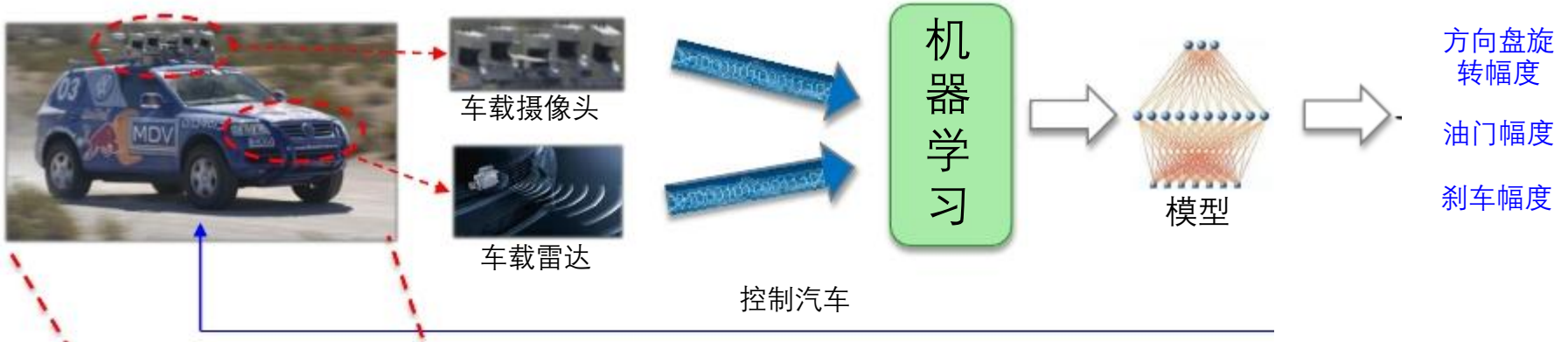
通过机器学习模型：

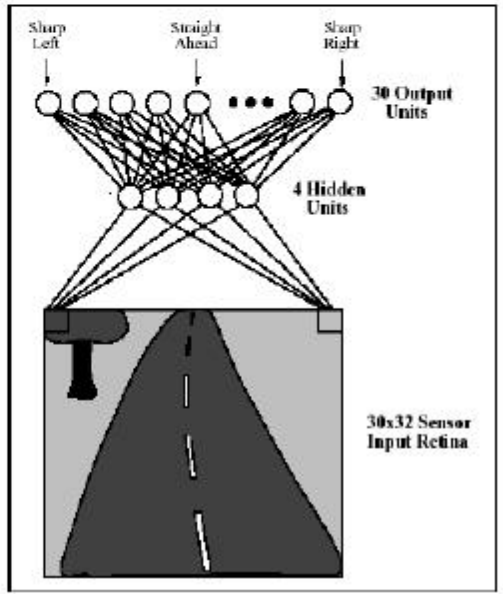
- 在总统候选人第一次辩论后，分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由
- 精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%
- 向奥巴马推荐，竞选后期应当在什么地方展开活动 —— 那里有很多争取对象
- 借助模型帮助奥巴马筹集到创纪录的10亿美元

例如：利用模型分析出，明星乔治克鲁尼（George Clooney）对于年龄在40-49岁的美西地区女性颇具吸引力，而她们恰是最愿意为和克鲁尼/奥巴马共进晚餐而掏钱的人 …… 乔治克鲁尼为奥巴马举办的竞选筹资晚宴成功募集到1500万美元

- … …

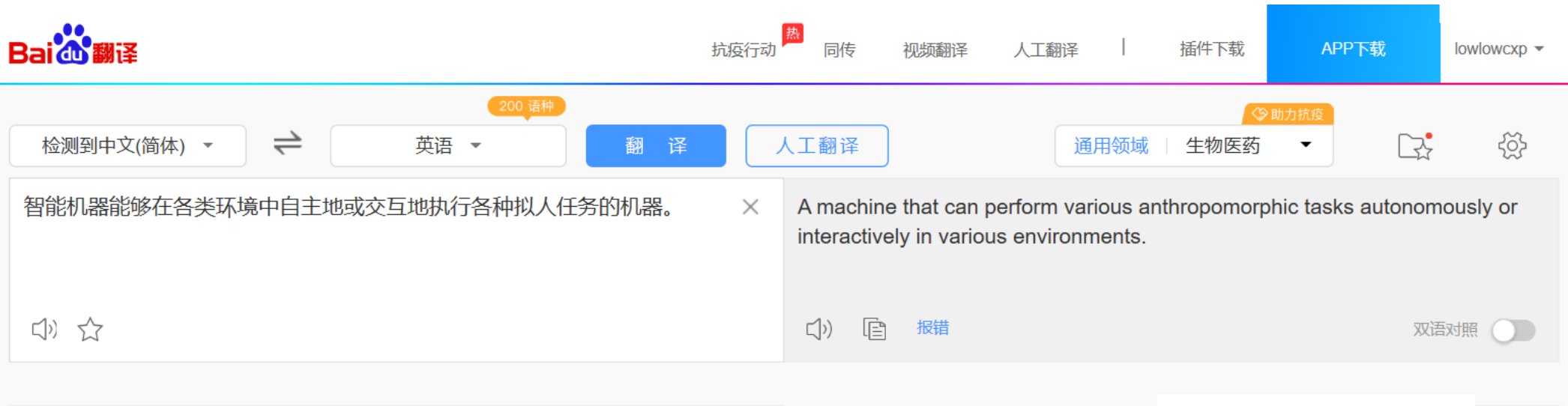
例2: 自动汽车驾驶

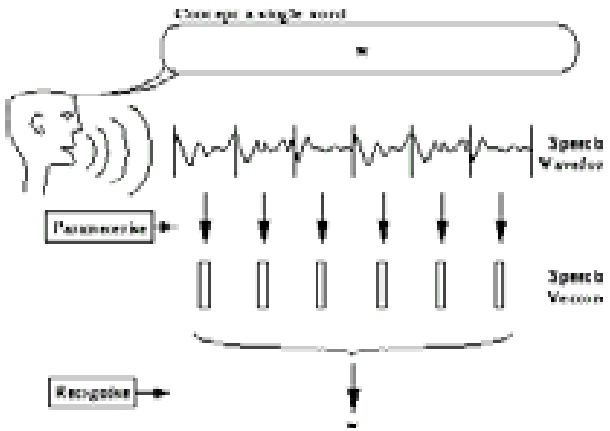




美国在20世纪80年代就开始研究基于机器学习的汽车自动驾驶技术

更多例子





语音识别

人脸识别

机器学习基本概念介绍

人工智能

人工智能：Artificial Intelligence

英文缩写：AI

机器人仅仅是人工智能的一个分支

定义1 智能机器(intelligent machine)能够在各类环境中自主地或交互地执行各种拟人任务的机器。

定义2 人工智能(学科)：计算机科学中涉及研究、设计和应用智能机器的一个分支。它的近期主要目标在于研究用机器来模仿和执行人脑的某些智力功能，并开发相关理论和技术。

定义3 人工智能(能力)：智能机器所执行的通常与人类智能有关的智能行为，如判断、推理、证明、识别、感知、理解、通信、设计、思考、规划、学习和问题求解等思维活动。

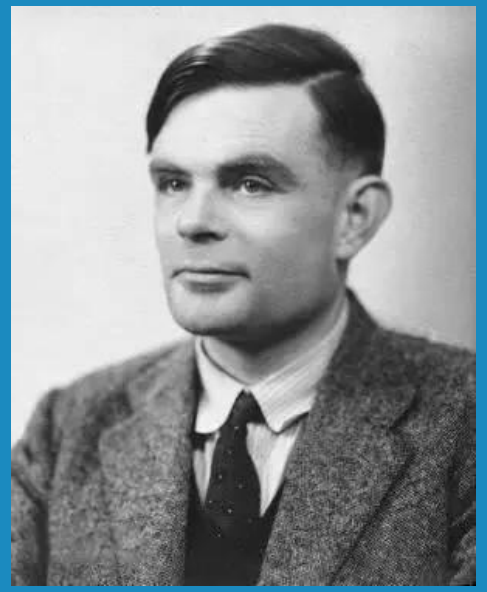
人工智能的基础

- 哲学：标出了AI的大部分重要思想
- 数学：使AI成为一门规范科学 数学形式化
- 神经科学：网络，并行处理
- 心理学：认知理论
- 计算机工程：AI的“载体”
- 语言学：知识表示、语法
- 言学、神经生理学、心理学、数学、哲学

人工智能的评价标准

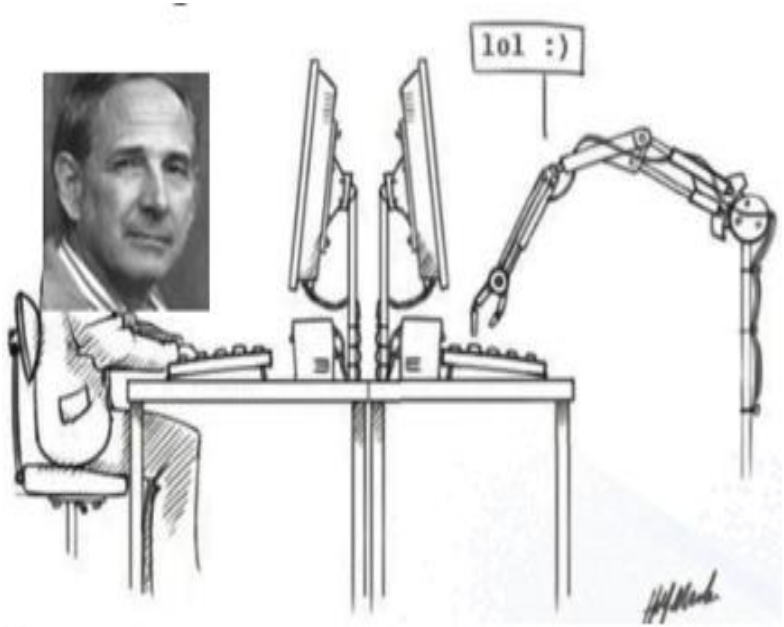
阿兰·图灵

英国数学家、逻辑学家，被视为计算机科学之父。



图灵测试

阿兰·图灵在1950年发表的一篇名为《计算机器与智能》的论文，提出著名的“图灵测试”，测试者在与被测试者（一个人和一台机器）隔开的情况下，通过一些装置（如键盘）向被测试者随意提问。如果机器能够让30%的测试人相信它是人类，那么这台计算机就可以被认为具有人类的思考能力。



怎么测试

- 问：你会下国际象棋吗？
- 答：是的。
- 问：你会下国际象棋吗？
- 答：是的。
- 问：请再次回答，你会下国际象棋吗？
- 答：是的。

怎么测试

- 问： 你会下国际象棋吗?
- 答：是的。
- 问：你会下国际象棋吗?
- 答：是的，我不是已经说过了吗?
- 问：请再次回答，你会下国际象棋吗?
- 答：你烦不烦，干嘛老提同样的问题。

人工智能的发展历程

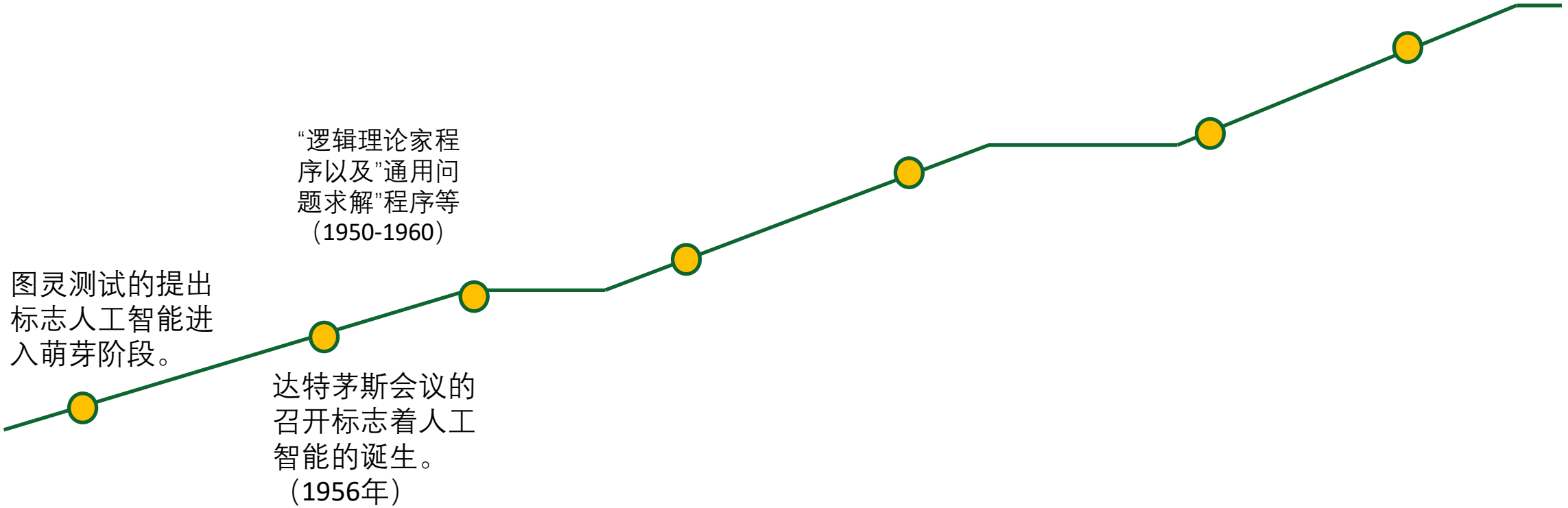
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第一阶段：推理期

1956-1960s: Logic Reasoning

- 出发点：“数学家真聪明！”
- 主要成就: 自动定理证明系统 (例如，西蒙与纽厄尔的“Logic Theorist”系统)

渐渐地，研究者们意识到，仅有逻辑推理能力是不够的 …

赫伯特 西蒙
(1916-2001)
1975年图灵奖

阿伦 纽厄尔
(1927-1992)
1975年图灵奖

人工智能的发展历程

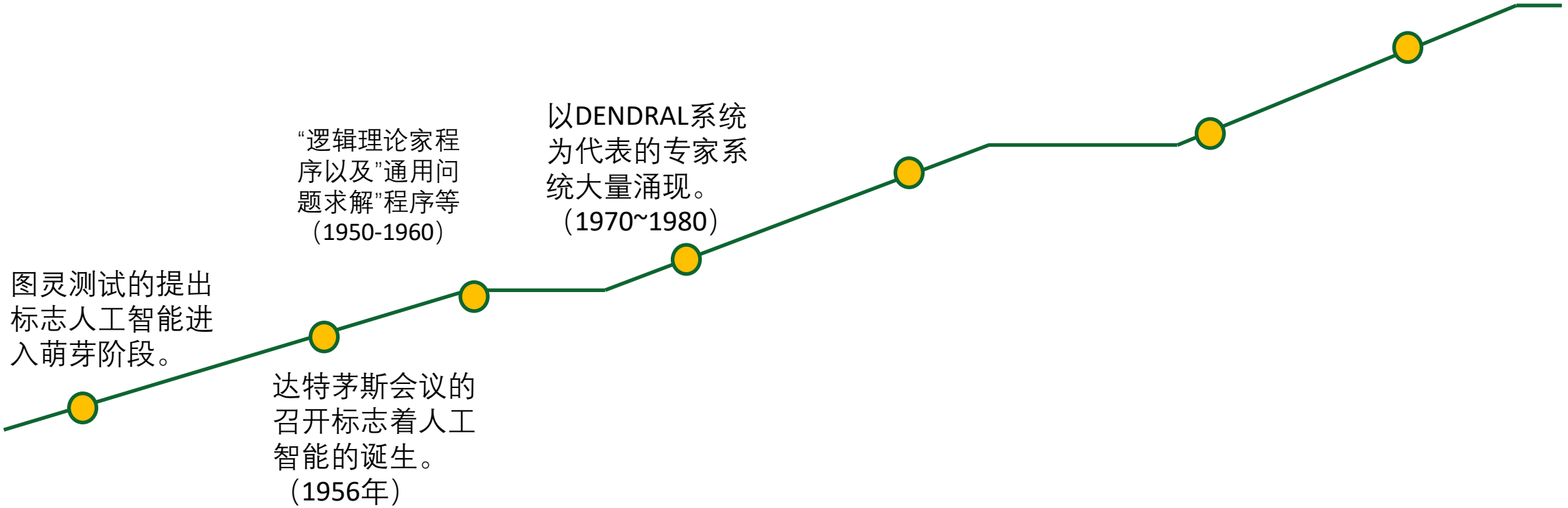
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第二阶段：知识期

1970s -1980s: Knowledge Engineering

- 出发点:
- 主要成就: 专家系统 (例如，费根鲍姆等人的“DENDRAL”系统)

1994年图灵奖
爱德华 费根鲍姆
(1936-)

渐渐地，研究者们发现，要总结出知识再“教”给系统，实在太难了 …

人工智能的发展历程

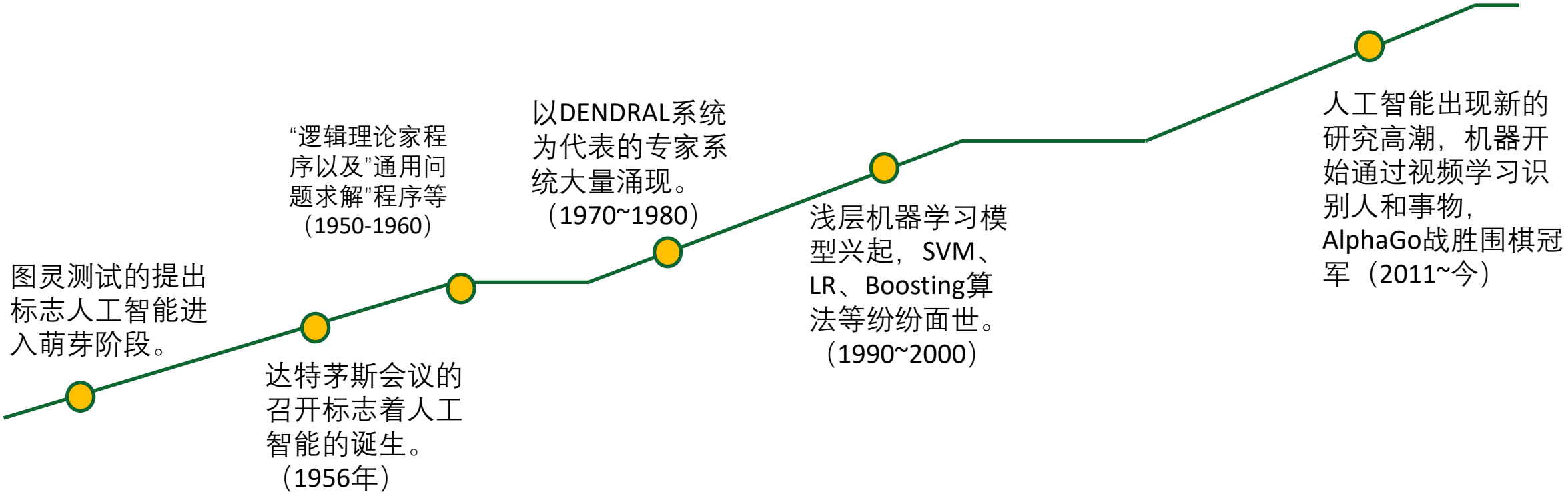
混沌初生 开天辟地

奠定了人工智能的数学基础，出现了人工智能历史上的第一个应用。
-西蒙和纽厄尔提出了“Logic Theorist”自动定理证明系统。

百家争鸣 百花齐放

随着新的算法和模型不断涌现，学科交叉现象日趋明显，人工智能的研究进入了新的阶段。

物竞天择 适者生存



第三阶段：学习期

1990s -now: Machine Learning

- 出发点: “让系统自己学！”
- 主要成就: ……

机器学习是作为“突破知识工程瓶颈”
之利器而出现的

恰好在20世纪90年代中后期，人类发现自己淹没在数据的汪洋中，对自动数据分析技术——机器学习的需求日益迫切

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

- **1980s：登上历史舞台**
- **机器学习作为一支独立的力量登上了历史舞台。在这之后的10年里出现了一些重要的方法和理论，典型的代表是：**
  - **1980夏-在卡内基梅隆举行第一届机器学习研讨会(IWML)；**
  - **1983第一本机器学习的专著《机器学习-一种人工智能的途径》；**
  - **1984-分类与回归树（CART）**
  - **1986-第一个期刊《Machine Learning》创刊**
  - **1986-反向传播算法**
  - **1989-卷积神经网络**

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

1990-2012：走向成熟和应用

在这20多年里机器学习的理论和方法得到了完善和充实，可谓是百花齐放的年代。代表性的重要成果有：

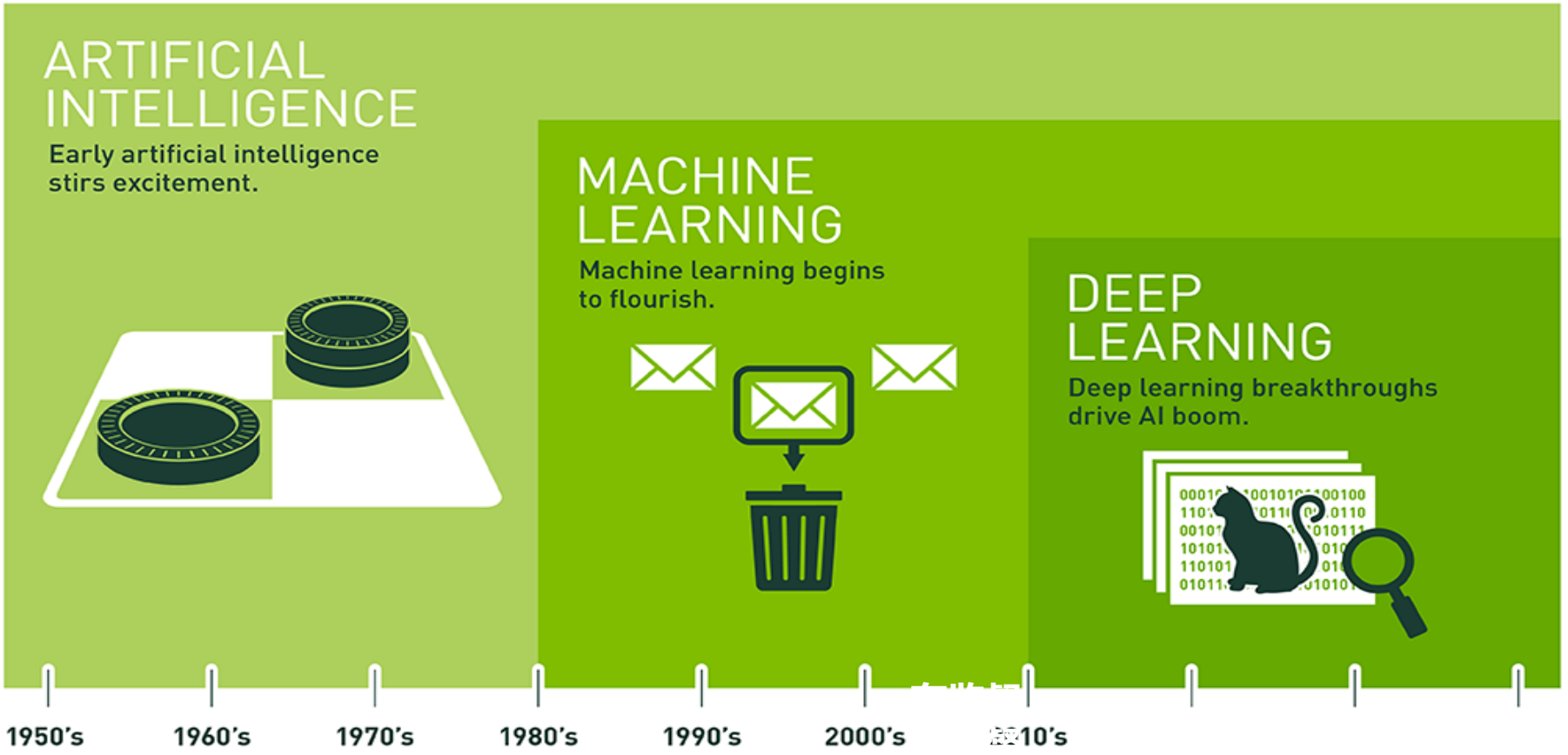
- **1995：支持向量机（SVM）**
- **1997：AdaBoost算法**
- **1997：循环神经网络（RNN）和LSTM**
- **2000：流形学习**
- **2001：随机森林**

机器学习的发展历程

- 机器学习的发展经历了三个阶段
  - 1980年代：成形期
  - 1990-2010年代：蓬勃发展期
  - 2012年之后：深度学习时期

- **2012：深度学习时代-神经网络卷土重来**
  - **深度卷积神经网络：AlexNet首先在图像分类问题上取代成功，随后被用于机器视觉的各种问题上**
  - **循环神经网络：语音识别，自然语言处理**
  - **深度强化学习：策略、控制类问题，典型的代表是AlphaGo**
  - **……**

几个概念的关系

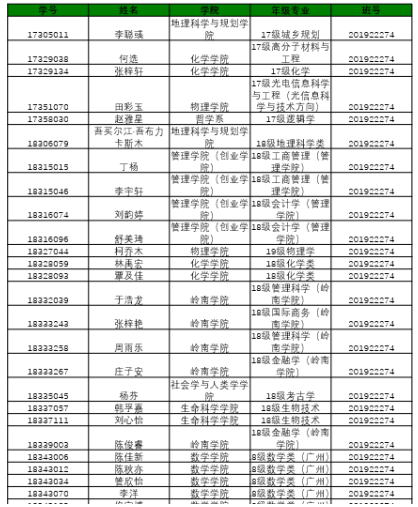


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

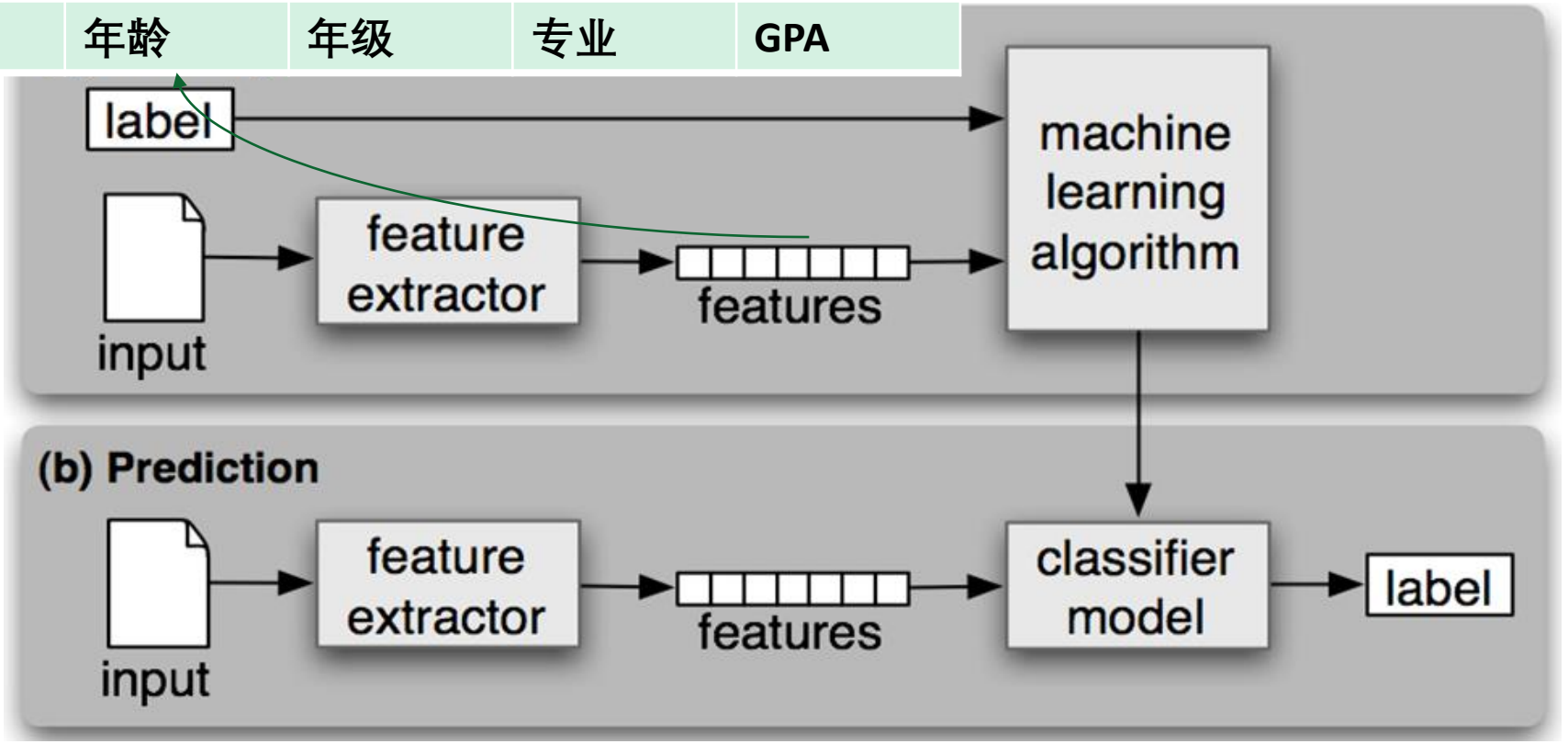
机器学习

- 机器学习主要是设计和分析让计算机可以自动“学习”的算法。学习算法是一类从数据中自动分析获得规律，利用规律对未知数据进行预测的算法。

学号	姓名	性别	年龄	年级	专业	GPA

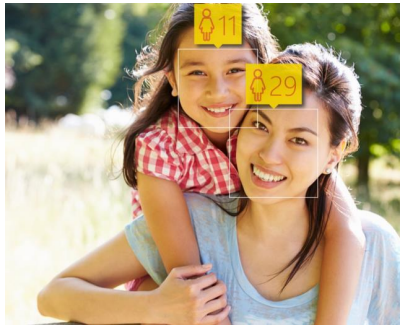


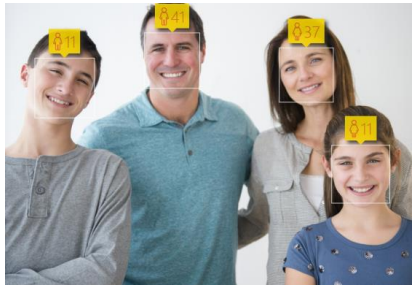
学号	姓名	学院	年级专业	班号
17305011	李聪颖	地理科学与规划学院	17级城乡规划	201922274
17329038	何浩	化学学院	17级高分子材料与工程	201922274
17329134	张梓轩	化学学院	17级化学	201922274
17351070	田彩玉	物理学院	17级光电信息科学与工程（光信息科学与技术方向）	201922274
17358030	赵雅星	药学系	17级医确学	201922274
18306079	喜买尔江·喜布力卡斯木	地理科学与规划学院	18级地理科学类	201922274
18315015	丁杨	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18315046	李宇轩	管理学院（创业学院）	18级工商管理（管理学院）	201922274
18316074	刘韵婷	管理学院（创业学院）	18级会计学（管理学院）	201922274
18316096	舒奚琦	管理学院（创业学院）	18级会计学（管理学院）	201922274
18327044	科乔木	物理学院	19级物理学	201922274
18328059	林禹宏	化学学院	18级化学类	201922274
18328093	覃及佳	化学学院	18级化学类	201922274
18332039	于浩龙	岭南学院	18级管理科学（岭南学院）	201922274
18332243	张梓艳	岭南学院	18级国际商务（岭南学院）	201922274
18332258	周雨乐	岭南学院	18级管理科学（岭南学院）	201922274
18332267	庄子安	岭南学院	18级金融学（岭南学院）	201922274
18335045	杨芬	社会学与人类学学院	18级考古学	201922274
18337057	钟昇嘉	生命科学学院	18级生物技术	201922274
18337111	刘心怡	生命科学学院	18级生物技术	201922274
18339003	陈俊豪	岭南学院	18级金融学（岭南学院）	201922274
18343006	陈佳新	数学学院	8级数学类（广州）	201922274
18343012	陈秋亦	数学学院	8级数学类（广州）	201922274
18343034	管欣怡	数学学院	8级数学类（广州）	201922274
18343070	李洋	数学学院	8级数学类（广州）	201922274

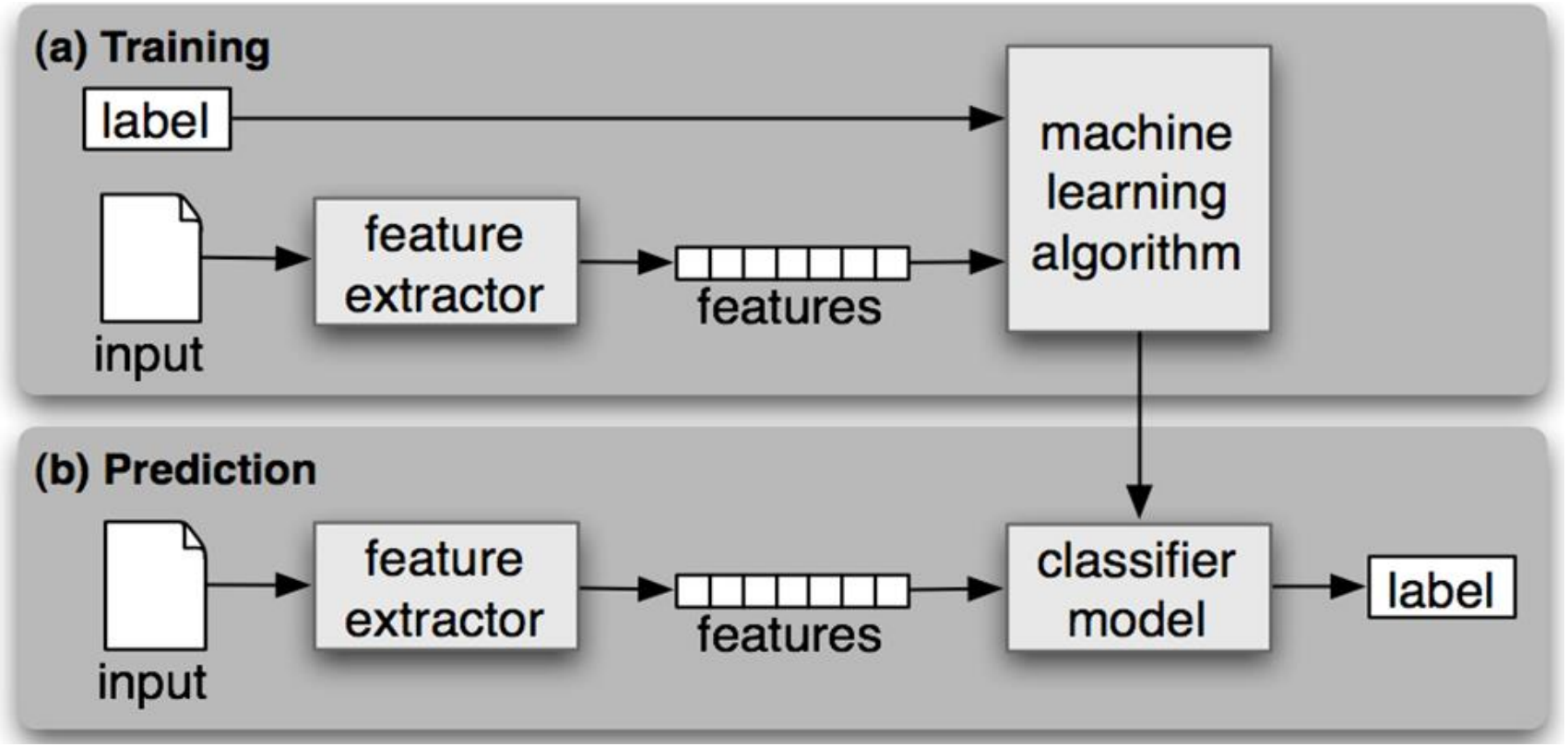


机器学习

- 机器学习主要是设计和分析让计算机可以自动“学习”的算法。学习算法是一类从数据中自动分析获得规律，利用规律对未知数据进行预测的算法。

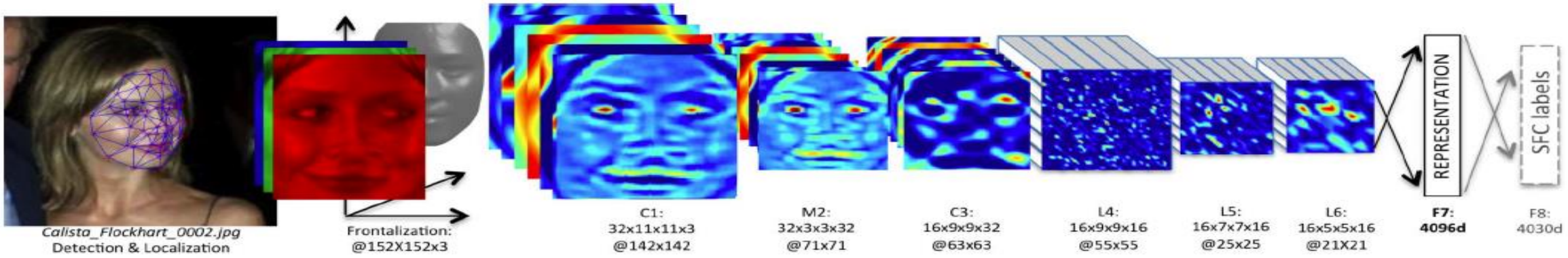






深度学习

- Learn a *feature hierarchy* all the way from pixels to classifier
- Each layer extracts features from the output of previous layer
- Train all layers jointly



例1：帮助奥巴马胜选

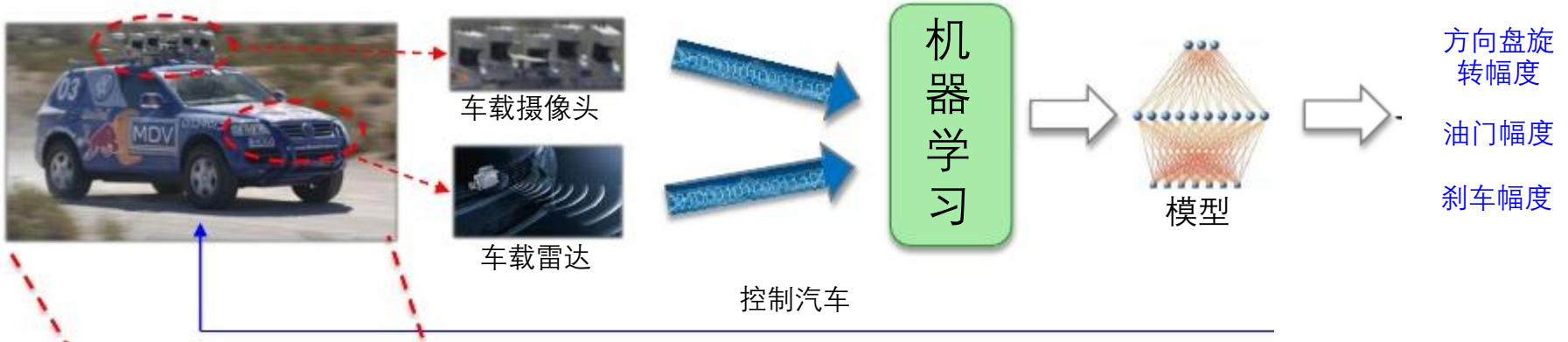
通过机器学习模型：

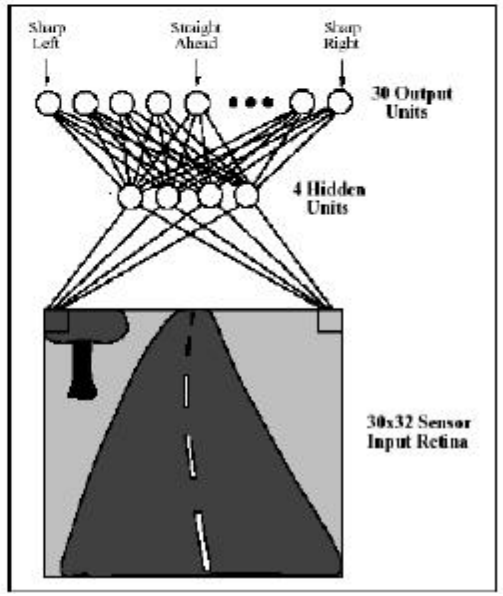
- 在总统候选人第一次辩论后，分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由
- 精准定位不同选民群体，建议购买冷门广告时段，广告资金效率比2008年提高14%
- 向奥巴马推荐，竞选后期应当在什么地方展开活动 —— 那里有很多争取对象
- 借助模型帮助奥巴马筹集到创纪录的10亿美元

例如：利用模型分析出，明星乔治克鲁尼（George Clooney）对于年龄在40-49岁的美西地区女性颇具吸引力，而她们恰是最愿意为和克鲁尼/奥巴马共进晚餐而掏钱的人 …… 乔治克鲁尼为奥巴马举办的竞选筹资晚宴成功募集到1500万美元

- … …

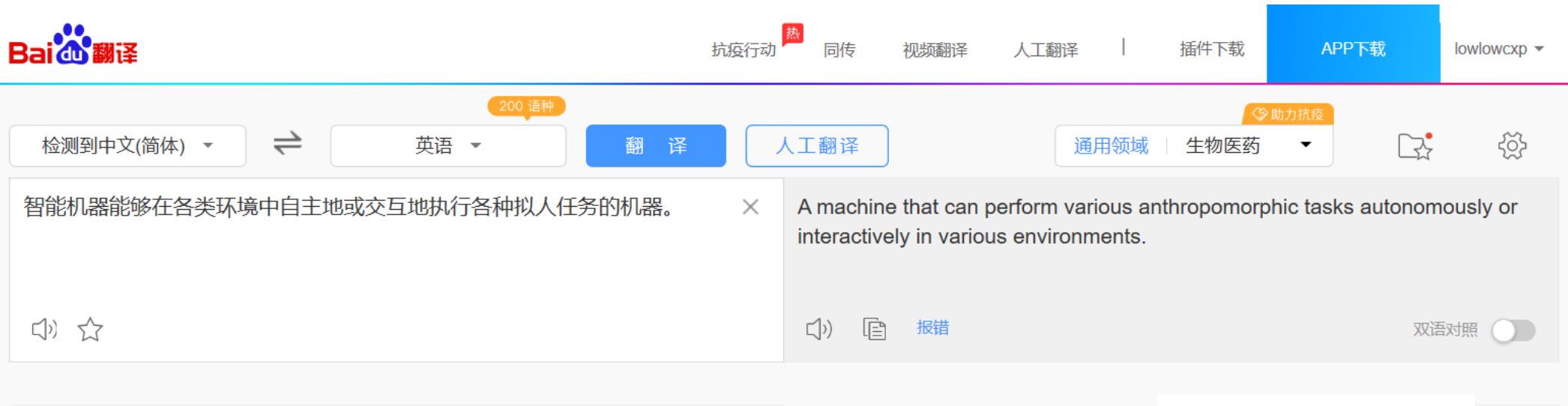
例2: 自动汽车驾驶

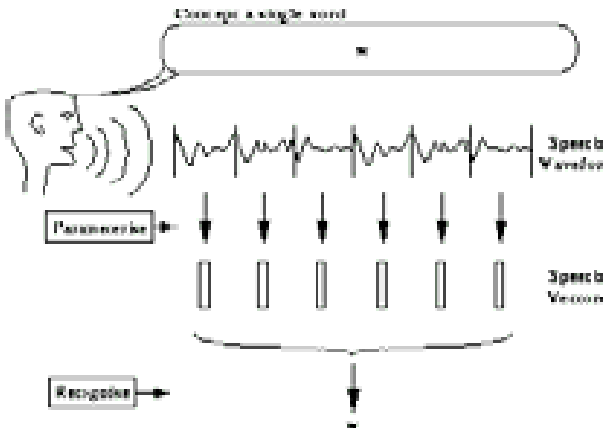




美国在20世纪80年代就开始研究基于机器学习的汽车自动驾驶技术

更多例子





语音识别

人脸识别

机器学习基本概念介绍

机器学习方法分类

- 机械学习（Rote learning）：学习者无需任何推理或其它的知识转换，直接吸取环境所提供的信息。如塞缪尔的跳棋程序。
- 示教学习（Learning from instruction）：学生从环境（教师或其它信息源如教科书等）获取信息，把知识转换成内部可使用的表示形式，并将新的知识和原有知识有机地结合为一体。
- 类比学习（Learning by analogy）：利用二个不同领域（源域、目标域）中的知识相似性，可以通过类比，从源域的知识（包括相似的特征和其它性质）推导出目标域的相应知识，从而实现学习。
- 归纳学习（Learning from induction）：教师或环境提供某概念的一些实例或反例，让学生通过归纳推理得出该概念的一般描述。

归纳学习方法分类

- 归纳与演绎是科学推理的两大基本手段
  - 归纳是从特殊到一般的“泛化”过程，即从具体的事实归结出一般性规律
  - 演绎是从基础原理推理出具体的状况

- 监督学习(Supervised Learning)：监督学习是从标记的训练数据来推断一个功能的机器学习任务。
  - 分类：预测离散值
  - 回归：预测连续值

- 非监督学习(Unsupervised Learning)：无监督学习的问题是，在未标记的数据中，试图找到隐藏的结构。
  - 如**聚类**、密度估计

例子

- 基于学生的资料 （◯ 回归，□分类）

问题	监督	非监督
预测学生期末成绩		
预测分数在10%的同学		
预测不及格人数		
预测中途退课人数		

- 监督vs非监督
- 回归vs分类

机器学习基本过程

表示

(Representation)

将数据对象进行特征(feature)化表示

训练

(Training/Learning)

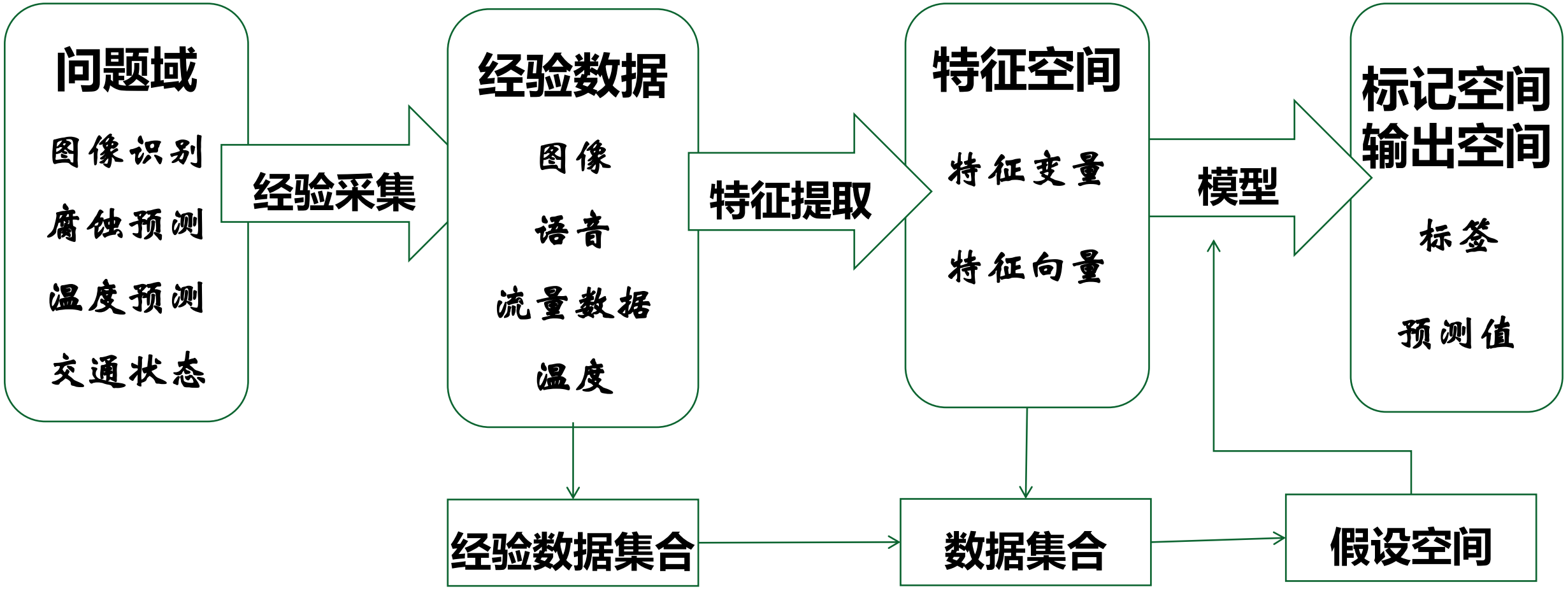
给定一个数据样本集，从中学习出规律（模型）
目标：该规律不仅适用于训练数据，也适用于未知数据

测试

(Testing/Predicting/Inference)

对于一个新的数据样本，利用学到的模型进行预测

机器学习的基本过程


问题域
图像识别
腐蚀预测
温度预测
交通状态
经验采集
经验数据
图像
语音
流量数据
温度
特征提取
特征空间
特征变量
特征向量
模型
标记空间
输出空间
标签
预测值
经验数据集合
数据集合
假设空间

机器学习的基本概念:特征空间

选择一组变量描述问题性质，称为特征变量（属性），特征变量组成的向量称为特征向量，变量张成的空间称为特征空间（样本空间），变量的取值称为属性值。

特征变量（属性）记为：x_i，i=1、…、d

特征空间记为：G

男、女

性别　女

年级

特征向量记为： $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

机器学习的基本概念:数据集合

样本：特征空间（样本空间）中的一组示例。记为： $D=\{x_1、x_2、\ldots、x_m\}$

标记空间：标签变量或预测变量的取值集合，记为： Y

样例集合：特征向量与标签变量对集合，记为：

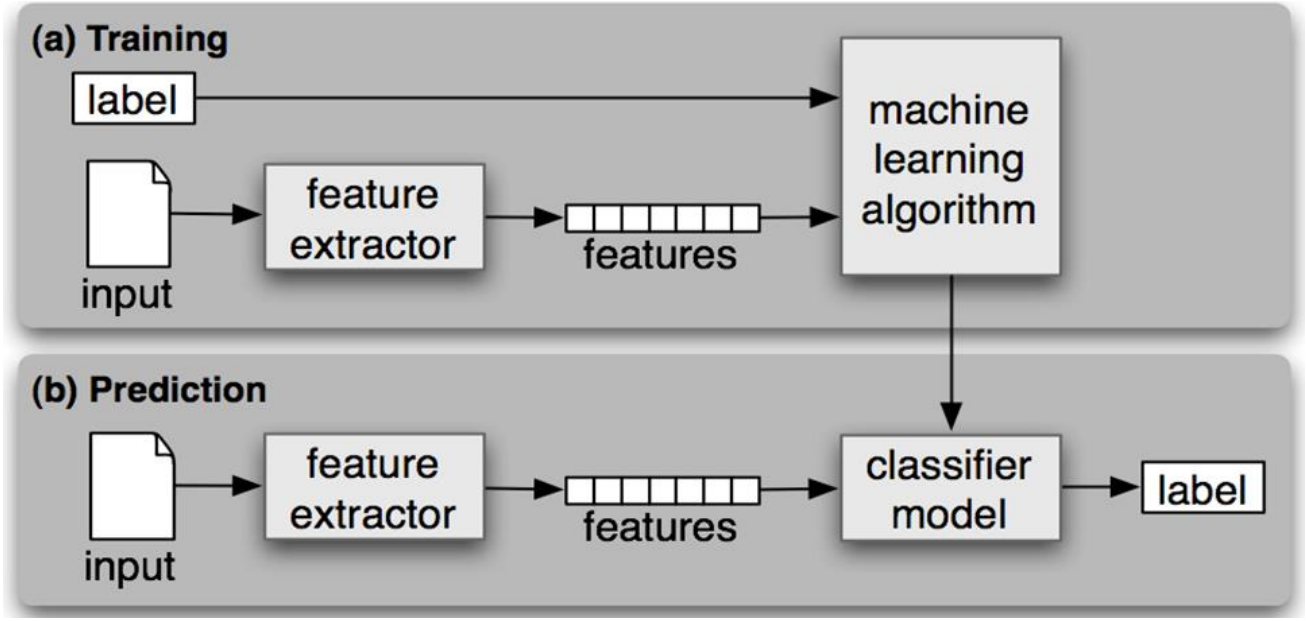
$$D=\{(x_1、y_1)、(x_2、y_2)、\ldots、(x_m、y_m)\}$$

{女，二年级，3.9}

机器学习的基本概念:数据集合

学习(训练)数据：在训练过程中使用的数据称为训练数据，每一个样例称为训练样本，全体训练样本集合称为训练集(*training set*)。

测试数据(*testing data*)：用于检测学习得到模型的数据称为检测数据，每一个样例称为检测样本，全体检测样本集合称为检测集(*testing set*)。



机器学习的基本概念：学习的任务

$y=f(x)$

分类：$Y=\{1，2，3，\ldots\}$，是离散值集合。二分类、多分类。

回归：$Y \in (0 \quad 1)$，是连续值集合，预测。

有监督学习

聚类：没有Y的信息。 → **无监督学习**

泛化能力：学习的结果对新样本的适应能力，对样本空间的描述能力。

机器学习的基本概念:假设空间

机器学习是通过数据集学得规律，是一个典型的归纳推理的过程，学习的结果是从样本空间到标记空间的一个映射，所有可能的映射的集合我们称为假设空间。

机器学习的任务：求$f \in H$：　$f: G \to Y$，满足数据集合

例：年龄预测问题：假设皱纹、发色和皮肤完全决定人的年龄，我们可以用布尔表达式表达年轻和高龄的概念。

年轻 ⟷ (皱纹=？)∧ (发色=？) ∧ (皮肤=？)

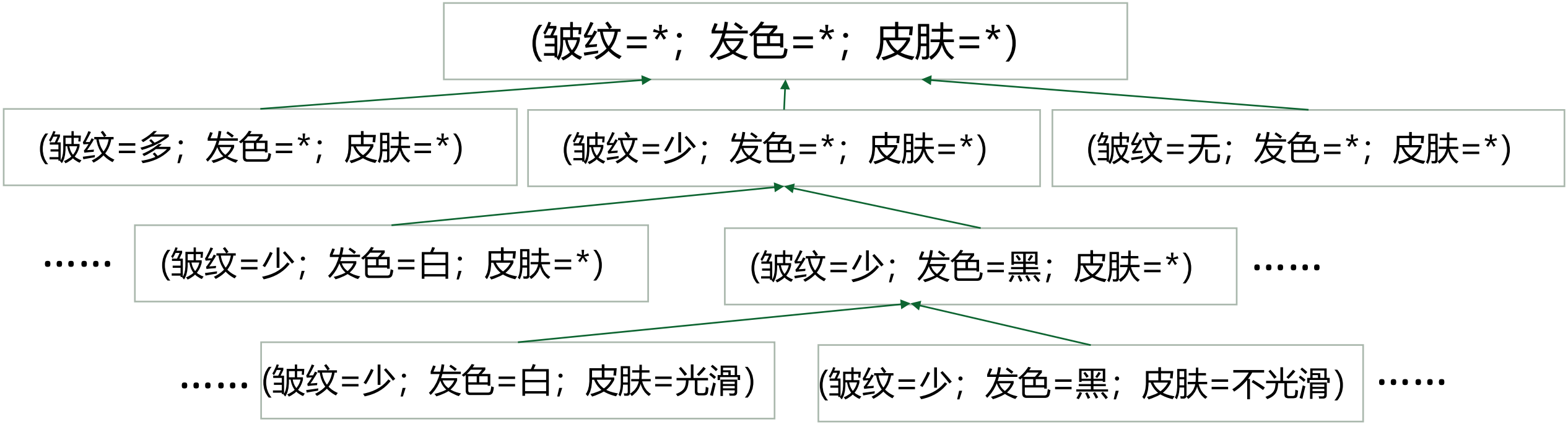
年轻 ⟷ (皱纹=少/无)∧ (发色=黑) ∧ (皮肤=光滑)

f(x)

((皱纹=无)∨(皱纹=少))∧ (发色=黑) ∧ (皮肤=光滑)

机器学习的基本概念:假设空间

年龄问题的所有布尔表达式表达：假设空间



版本空间：假设空间的一个子集，与训练样例一致的所有假设的集合。

机器学习的基本概念:归纳偏好

- 版本空间≠假设空间

  (皱纹=？)∧(发色=？)∧(皮肤=？)

- 如何在版本空间获得模型?
- 在机器学习算法学习的过程中对某种假设的偏好称为**归纳偏好**。

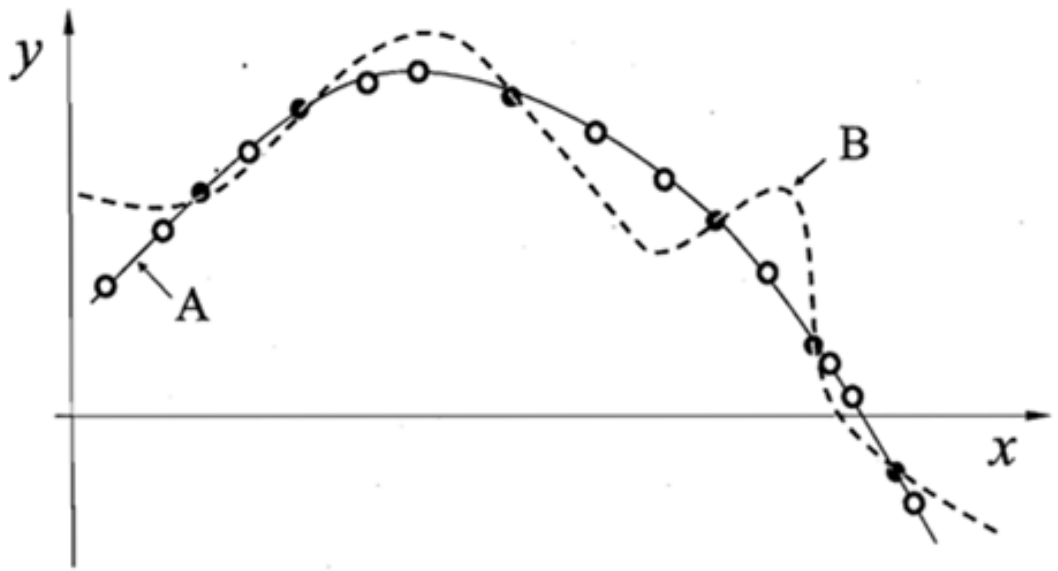
机器学习的基本概念

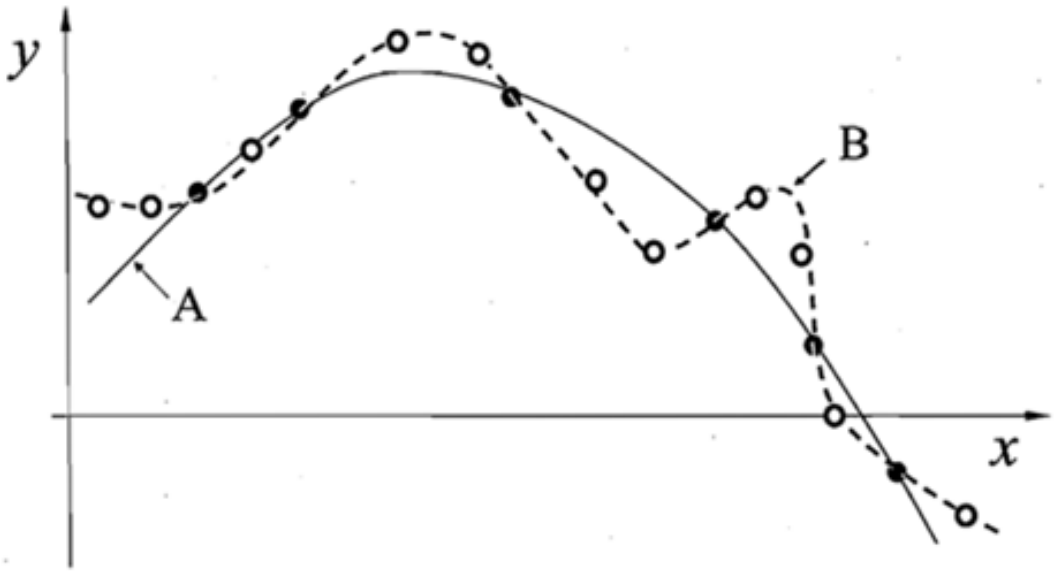
奥卡姆剃刀(Occam’s razor)：
若多个假设与观察一致，选择最简单的那个

没有免费午餐定理(NFL No Free Lunch Theorem)：总误差与算法无关

$$\sum_f E_{ote}\left(L_a \middle| X, f\right) = \sum_f E_{ote}\left(L_b \middle| X, f\right)$$

具体问题具体分析





黑点
训练
样本

白点
测试
样本

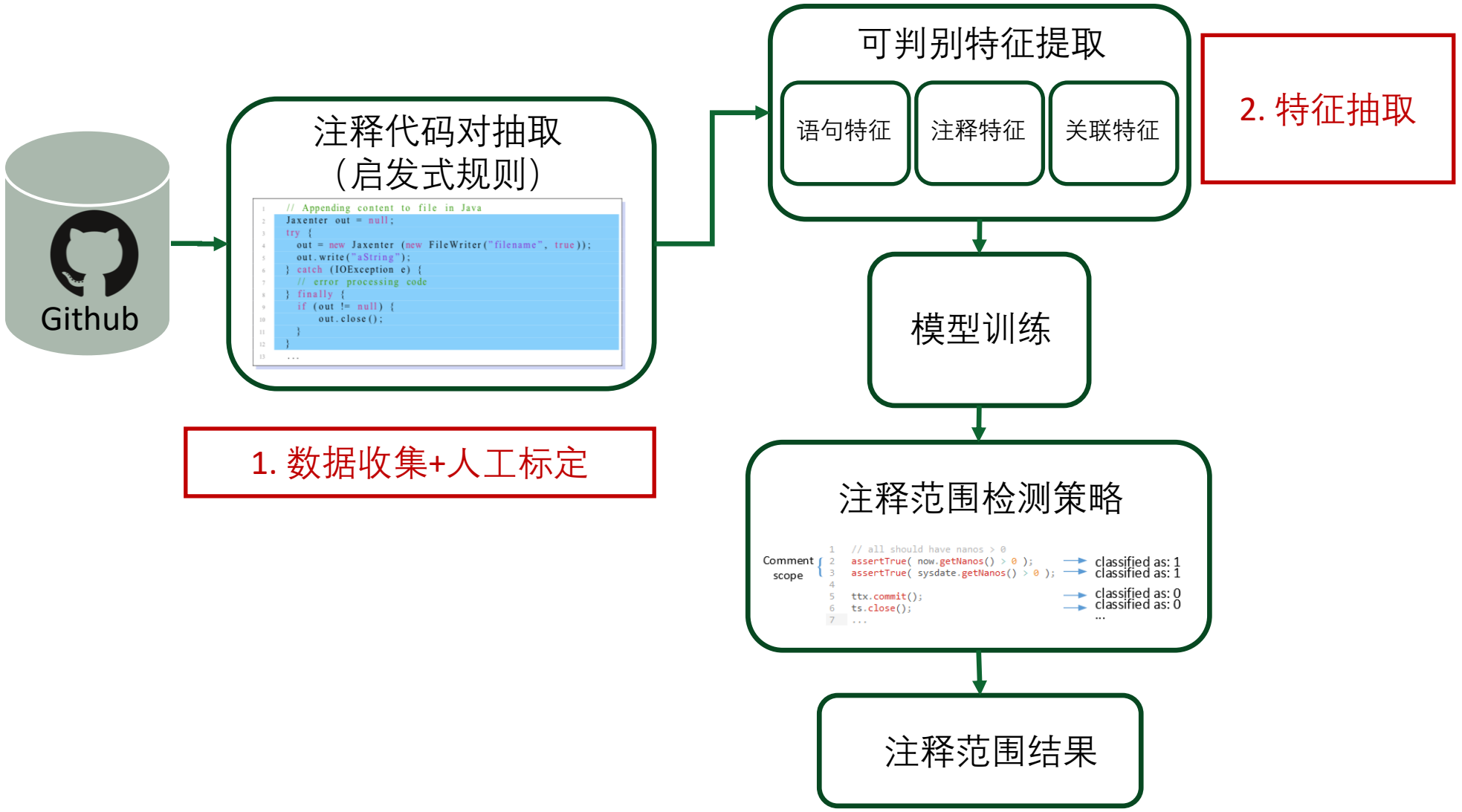
机器学习实例

问题描述

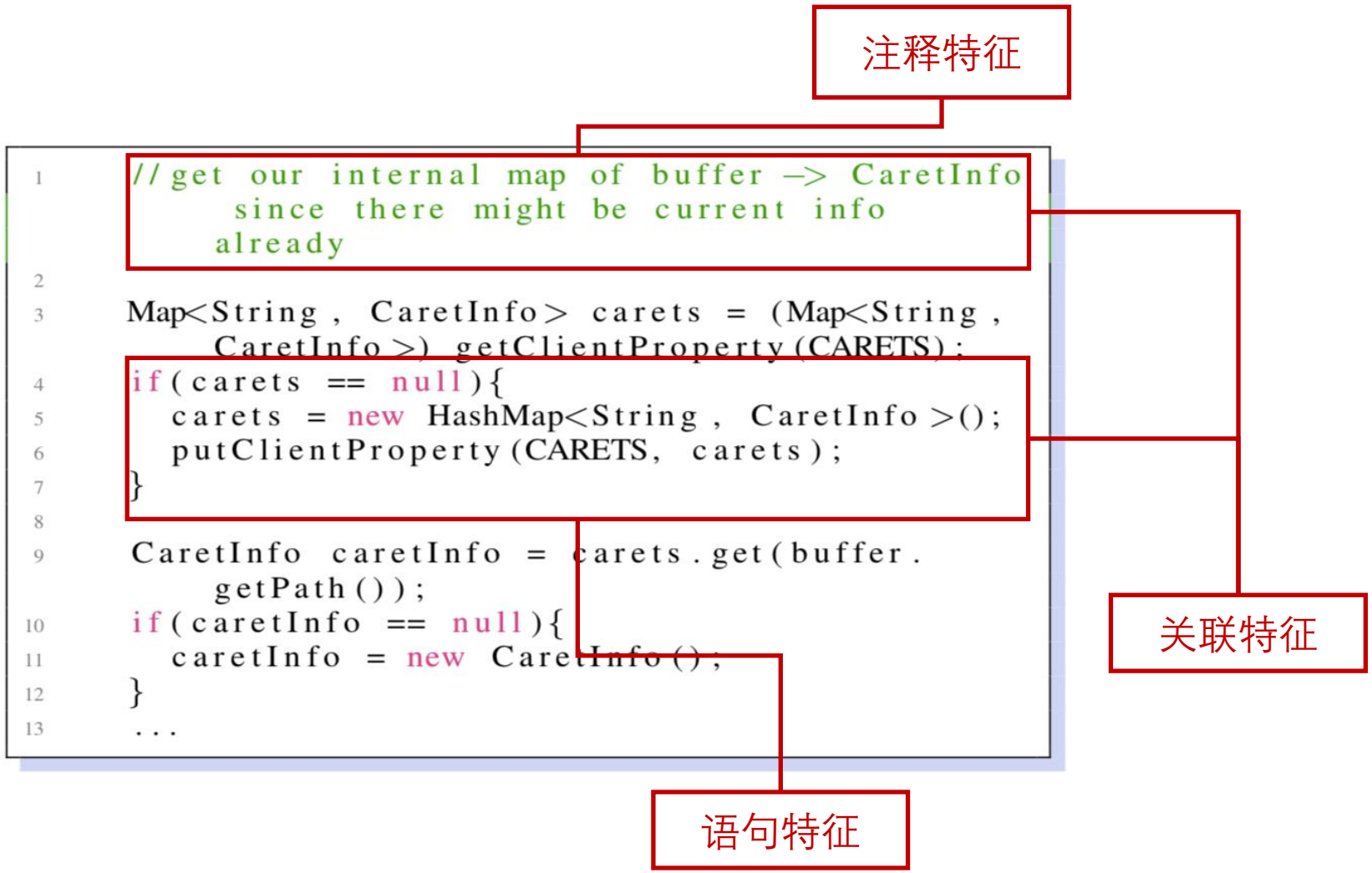
- 如何识别出注释描述的是哪些代码

```
1     //get our internal map of buffer -> CaretInfo
          since there might be current info
          already
2
3     Map<String, CaretInfo> carets = (Map<String,
          CaretInfo>) getClientProperty(CARETS);
4     if(carets == null){
5       carets = new HashMap<String, CaretInfo>();
6       putClientProperty(CARETS, carets);
7     }
8
9     CaretInfo caretInfo = carets.get(buffer.
          getPath());
10    if(caretInfo == null){
11      caretInfo = new CaretInfo();
12    }
13    ...
```

方法设计

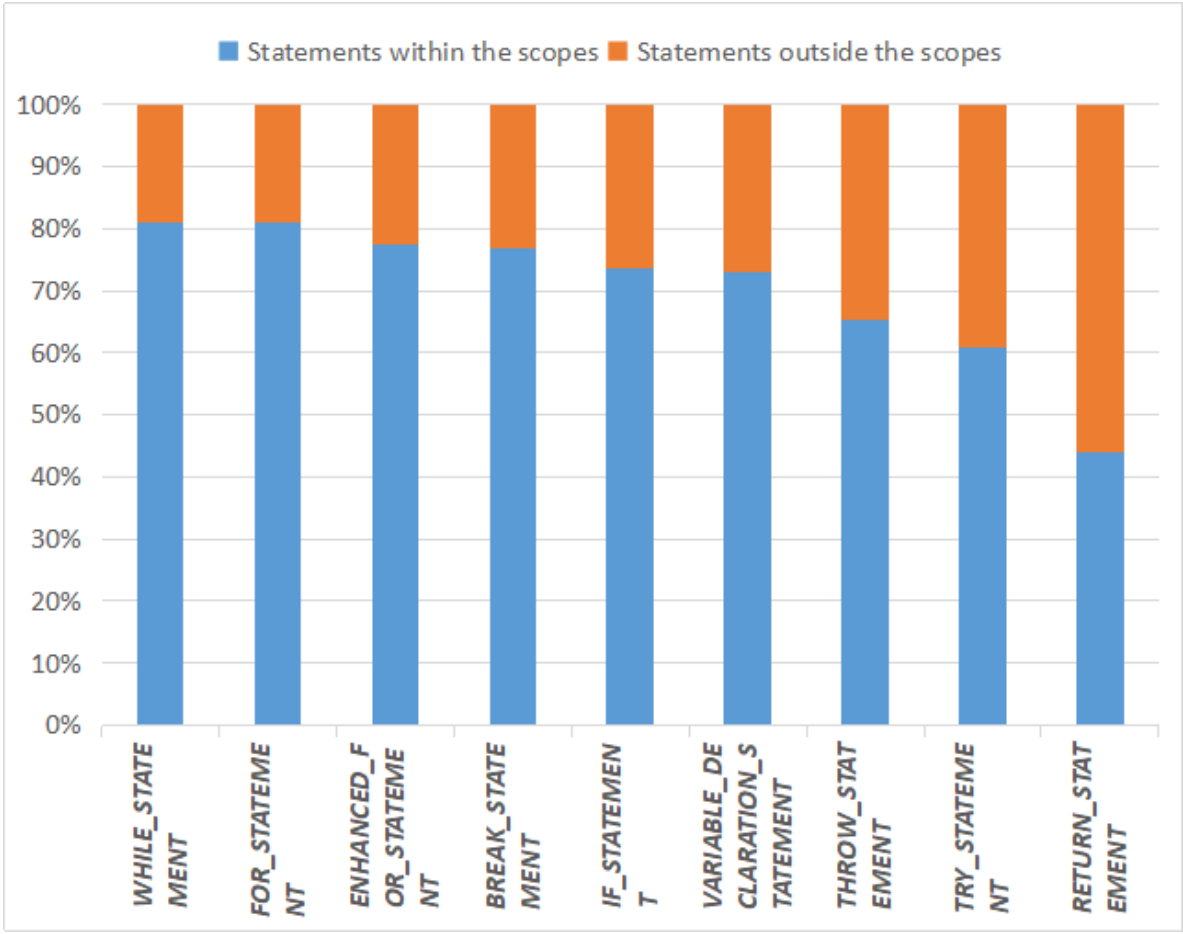

Github
注释代码对抽取
（启发式规则）
// Appending content to file in Java
Jaxenter out = null;
try {
out = new Jaxenter (new FileWriter("filename", true));
out.write("aString");
} catch (IOException e) {
// error processing code
} finally {
if (out != null) {
out.close();
}
}
...
1. 数据收集+人工标定
可判别特征提取
语句特征
注释特征
关联特征
2. 特征抽取
模型训练
注释范围检测策略
Comment scope
// all should have nanos > 0
assertTrue(now.getNanos() > 0);
assertTrue(sysdate.getNanos() > 0);
ttx.commit();
ts.close();
...
classified as: 1
classified as: 1
classified as: 0
classified as: 0
...
注释范围结果

特征提取


注释特征
1 //get our internal map of buffer -> CaretInfo since there might be current info already
2
3 Map<String, CaretInfo> carets = (Map<String, CaretInfo>) getClientProperty(CARETS);
4 if(carets == null){
5 carets = new HashMap<String, CaretInfo>();
6 putClientProperty(CARETS, carets);
7 }
8
9 CaretInfo caretInfo = carets.get(buffer.getPath());
10 if(caretInfo == null){
11 caretInfo = new CaretInfo();
12 }
13 ...
关联特征
语句特征

特征提取

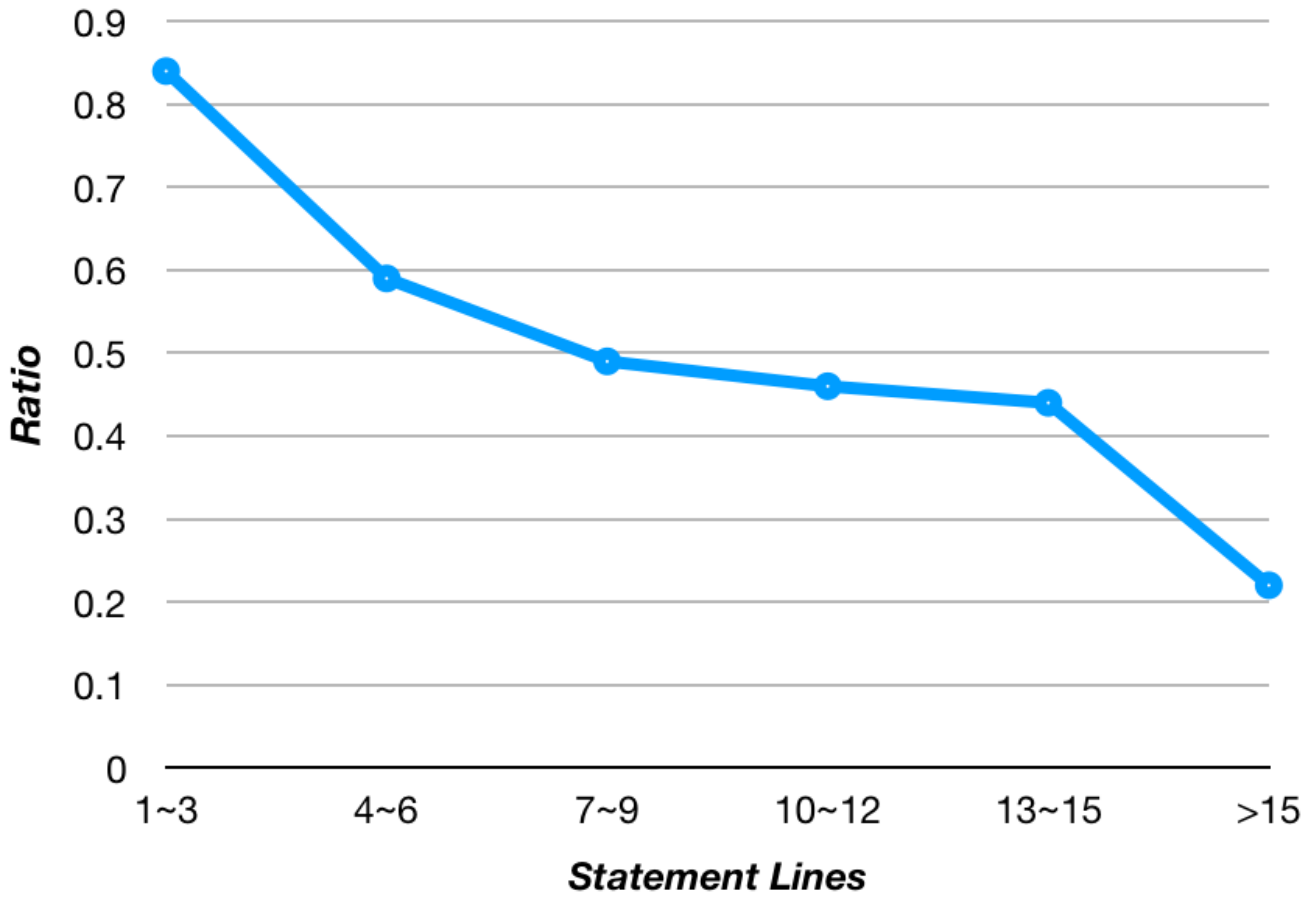
- 语句特征提取
  - 语句类型
  - 语句组成
  - 语句上下文
- 注释特征提取
  - 注释长度
  - 注释文本
  - 注释上下文
- 语句和注释的关联特征提取
  - 文本相似度
  - 语义相似度:
  - 语句和注释之间的距离



注释范围内外不同语句类型的分布

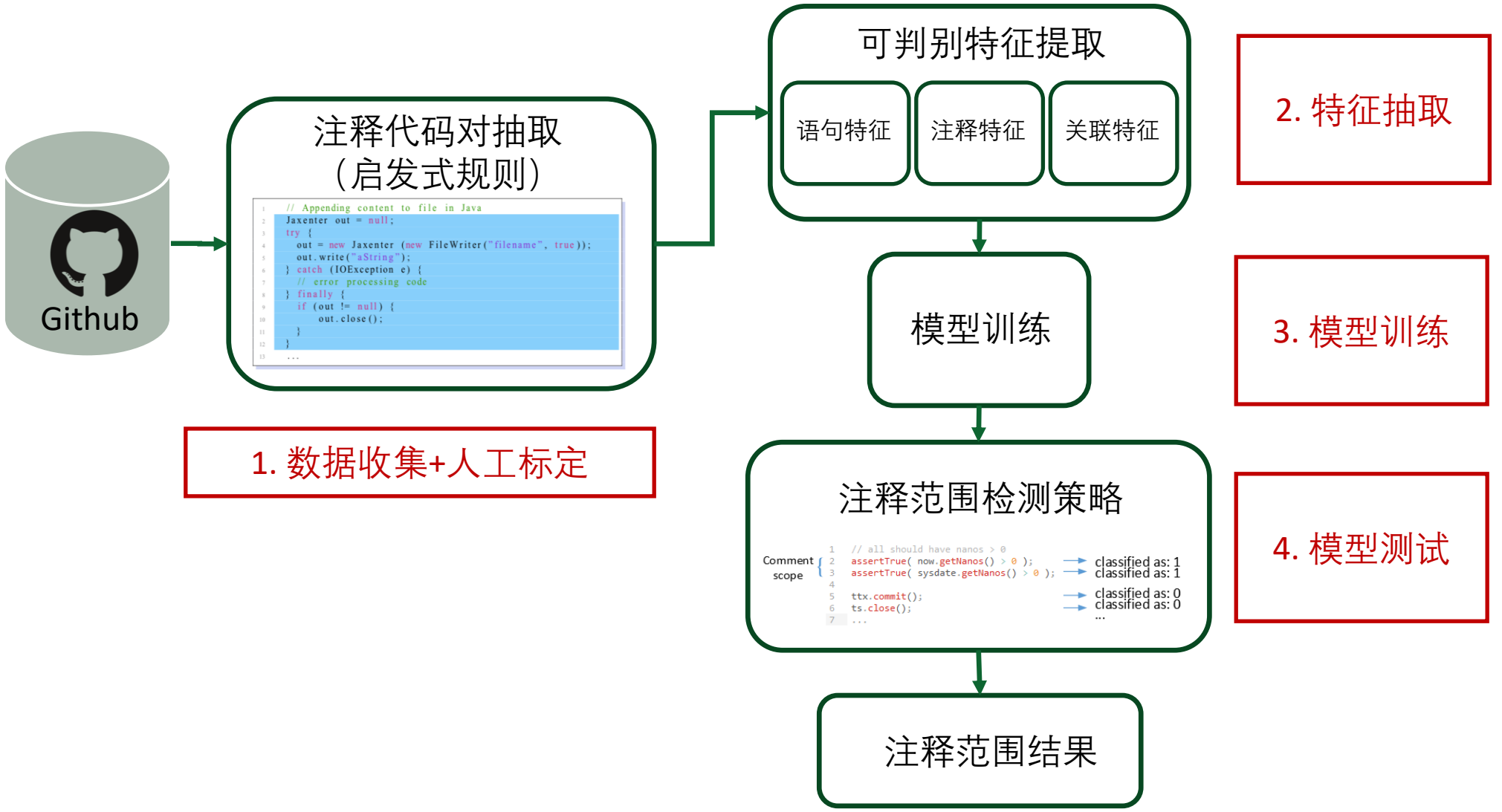
特征提取

- 语句特征提取
  - 语句类型
  - 语句组成
  - 语句上下文
- 注释特征提取
  - 注释长度
  - 注释文本
  - 注释上下文
- 语句和注释的关联特征提取
  - 文本相似度
  - 语义相似度:
  - 语句和注释之间的距离



不同的注释到语句的距离下，语句属于注释范围内的比例曲线

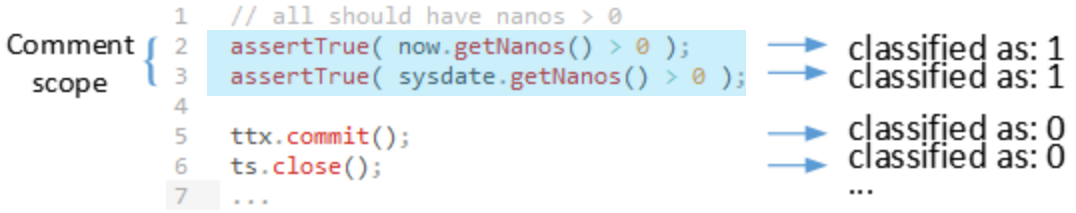
方法设计


Github
注释代码对抽取
（启发式规则）
// Appending content to file in Java
Jaxenter out = null;
try {
out = new Jaxenter (new FileWriter("filename", true));
out.write("aString");
} catch (IOException e) {
// error processing code
} finally {
if (out != null) {
out.close();
}
}
...
1. 数据收集+人工标定
可判别特征提取
语句特征
注释特征
关联特征
2. 特征抽取
模型训练
3. 模型训练
注释范围检测策略
Comment scope
// all should have nanos > 0
assertTrue(now.getNanos() > 0);
assertTrue(sysdate.getNanos() > 0);
ttx.commit();
ts.close();
...
classified as: 1
classified as: 1
classified as: 0
classified as: 0
...
4. 模型测试
注释范围结果

模型训练

- 模型训练
  - 数据集：注释范围内的语句视为正样本，注释范围外的语句视为负样本
- 注释范围检测策略
  - 对语句进行分类之后，第一个分类为注释范围外的语句视为注释范围的界定点

项目	注释-代码对个数	语句数		
		注释范围内	注释范围外	总计
Hadoop	1,000	2,825	1,211	4,036
Hibernate	1,000	3,473	1,299	4,772
JDK	1,000	2,066	677	2,743
jEdit	1,000	2,500	2,214	4,714
总计	**4,000**	**10,864**	**5,401**	**16,265**



实验结果

1. 代码注释范围检测的准确率能达到多少?

- 不同机器学习算法在语句分类上的表现
  - 所有实验都利用十折交叉验证进行评估
  - 结果表明随机森林表现最优

- 最终注释范围检测准确率对比
  - 我们方法的准确率为**81.45%**，相比启发式规则提高约**17%**

$$accuracy = \frac{|Correct\ Comments|}{|Total\ Comments|} \times 100\%$$

算法	正样本			负样本			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
J48	86.57%	87.78%	87.12%	71.92%	70.37%	70.97%	76.25%
SMO	83.53%	90.46%	86.83%	73.47%	60.13%	65.99%	79.95%
Naive Bayes	78.24%	89.28%	83.32%	62.48%	40.51%	47.84%	64.93%
Logit Boost	84.52%	**90.56%**	87.40%	74.20%	62.51%	67.74%	76.85%
Adaboost	**92.34%**	84.55%	88.19%	71.05%	**84.47%**	76.96%	81.35%
Random Forests	90.49%	88.20%	**89.29%**	**74.97%**	79.51%	**77.01%**	**81.45%**

方法	范围检测正确的注释数	范围检测的注释数	Accuracy
启发式规则	2,684	4,000	64.25%
注释范围检测方法	3,251	4,000	81.45%

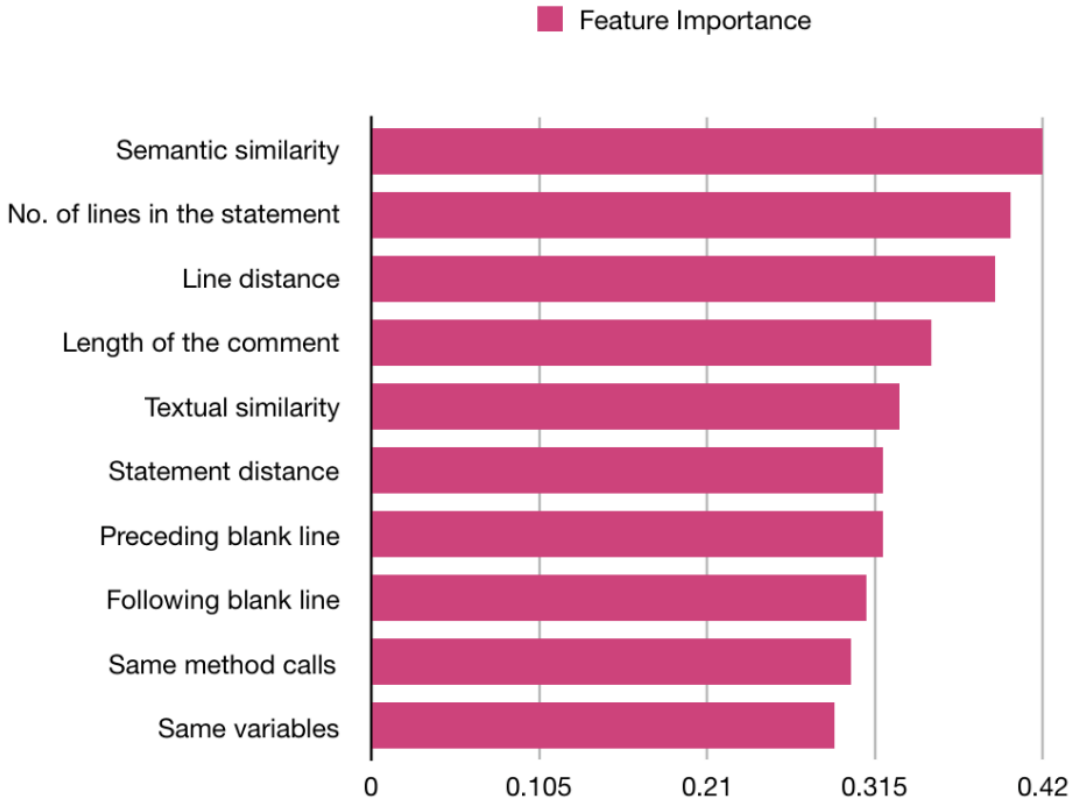
实验结果

2. 三个维度的特征对代码注释范围检测模型有影响吗?

- 不同维度的特征以及不同特征组合对注释范围检测模型的影响
  - 每一维度的特征都有用
  - 最有效的特征是语句特征

特征	正样本			负样本			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
语句	89.32%	86.63%	87.89%	72.14%	77.29%	74.46%	78.55%
注释	74.45%	35.30%	47.59%	36.33%	74.57%	48.72%	63.35%
关联	87.09%	85.92%	86.47%	68.68%	70.33%	69.38%	70.63%
语句 + 注释	89.44%	87.21%	88.25%	73.05%	77.66%	75.10%	78.95%
语句 + 关联	89.83%	88.21%	88.96%	74.46%	77.63%	75.85%	80.32%
注释 + 关联	75.74%	85.66%	80.37%	60.19%	44.30%	50.90%	64.30%
语句 + 注释 + 关联	90.49%	88.20%	89.29%	74.97%	79.51%	77.01%	81.45%

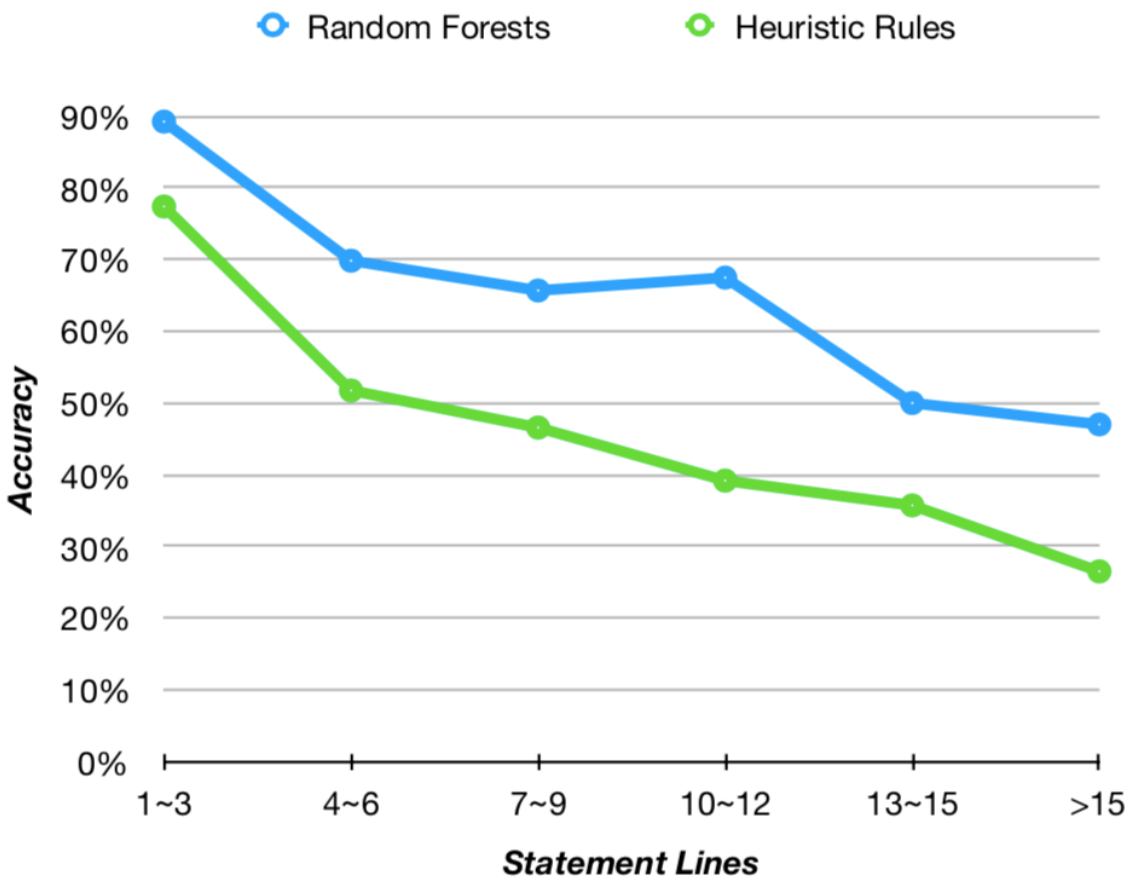
- 特征的重要性排名
  - 排名第一的特征是语句和注释之间的语义相似度



实验结果

3. 代码注释范围的大小对代码注释范围检测模型的准确率有影响吗?

- 比较我们的方法和启发式规则方法在不同大小的注释范围下的准确率变化情况
  - 当注释范围的大小为1-3条语句时，我们方法的准确率最高可达**90%**左右
  - 当注释范围大小增加时，我们方法的准确率下降，但仍优于启发式规则



实验结果

4. 代码注释范围检测模型在项目内和跨项目场景下的表现如何?

- 项目内评估
  - 训练集和测试集都来自同一项目
  - JDK项目的准确率最高，达到**82.20%**
- 跨项目评估
  - 其中三个项目作为训练集，余下的一个项目作为测试集
  - 平均准确率达到**80.73%**

项目	正样本			负样本			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Hadoop	90.58%	90.94%	90.73%	78.12%	77.57%	77.71%	80.50%
Hibernate	87.94%	88.44%	87.97%	69.22%	68.44%	67.66%	79.50%
JDK	89.37%	89.79%	89.39%	68.18%	67.20%	66.21%	82.20%
jEdit	92.13%	85.51%	88.60%	77.93%	86.86%	81.93%	80.30%
平均	90.01%	88.67%	89.17%	73.36%	75.02%	73.38%	80.63%

项目	注释范围内			注释范围外			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Hadoop	88.52%	92.78%	90.60%	81.02%	71.92%	76.20%	79.80%
Hibernate	89.30%	85.55%	87.38%	65.26%	72.59%	68.73%	79.50%
JDK	90.87%	90.13%	90.50%	70.61%	72.38%	71.48%	82.30%
jEdit	94.31%	80.92%	87.10%	81.43%	94.49%	87.48%	81.30%
平均	90.75%	87.35%	88.90%	74.58%	77.85%	75.97%	80.73%

小结

- 选定问题
- 收集数据
  - 如何标记
  - 数据质量
  - 数据量
  - ...
- 训练模型
- 实验结果
  - 基本
  - 问题特色

Q&A

谢谢大家!