




人工智能：模型与算法

逻辑与推理

吴飞

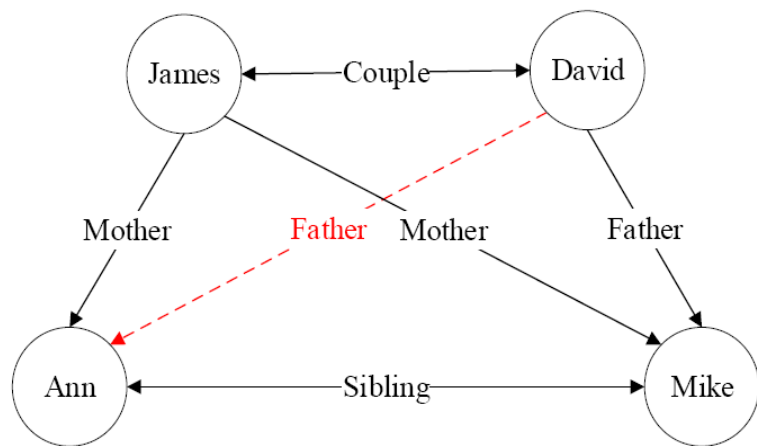
浙江大学计算机学院

提纲

- 1、命题逻辑
 - 2、谓词逻辑
 - 3、知识图谱推理
 - 4、因果推理
- 

知识图谱推理：路径排序

Score(Father(David, Ann))



将实体之间的关联路径作为特征，
来学习目标关系的分类器

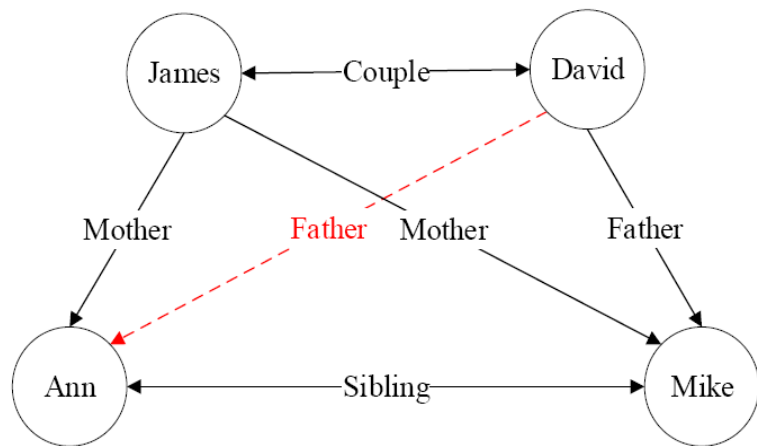
即：判断David和Ann之间的路径关联是否足够支持表述*Father*这一关系。

一个简单的家庭关系知识图谱

Ni Lao, William W. Cohen, Relational retrieval using a combination of path-constrained random walks, *Machine learning*, 2010, 81(1): 53-67, 2010

知识图谱推理：路径排序

$\text{Score}(\text{Father}(\text{David}, \text{Ann}))$



$$\text{score}(s, t) = \sum_{\pi_j \in p_l} \theta_j P(s \rightarrow t; \pi_j)$$

p_l 是链接节点 s 和节点 t 的所有路径集合

θ_j 是某一条路径 π_j 的权重

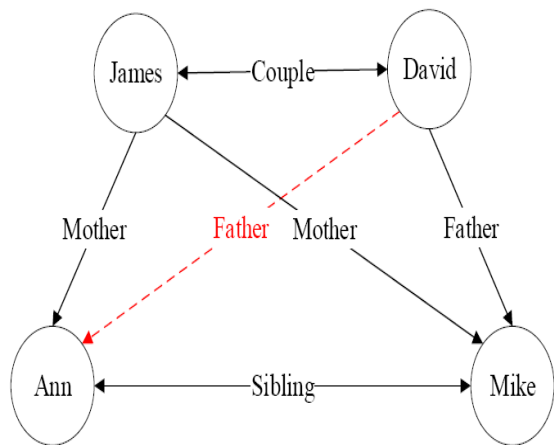
P 是路径 π_j 概率值大小

一个简单的家庭关系知识图谱

知识图谱推理：路径排序

$$\text{score}(s, t) = \sum_{\pi_j \in p_l} \theta_j P(s \rightarrow t; \pi_j)$$

Score(Father(David, Ann))

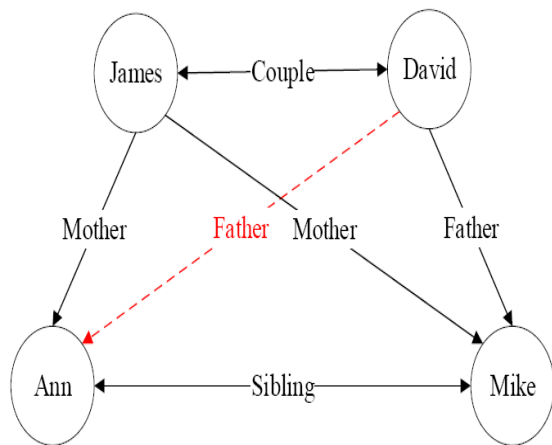


- **特征抽取：**生成并选择路径特征集合。生成路径的方式有随机游走（random walk）、广度优先搜索、深度优先搜索等。
- **特征计算：**计算每个训练样例的特征值 $P(s \rightarrow t; \pi_j)$ 。该特征值可以表示从实体节点 s 出发，通过关系路径 π_j 到达实体节点 t 的概率；也可以表示为布尔值，表示实体 s 到实体 t 之间是否存在路径 π_j ；还可以是实体 s 和实体 t 之间路径出现频次、频率等。
- **分类器训练：**根据训练样例的特征值，为目标关系训练分类器。当训练好分类器后，即可将该分类器用于推理两个实体之间是否存在目标关系。

知识图谱推理：路径排序

$$\text{score}(s, t) = \sum_{\pi_j \in p_l} \theta_j \text{father}(s \rightarrow t; \pi_j)$$

$\text{Score}(\text{Father}(\text{David}, \text{Ann}))$



给定目标关系: $\text{Father}(s, t)$

1. 对于目标关系 Father , 生成四组训练样例, 一个为正例、三个为负例:

正例: (David, Mike)

负例: (David, James), (James, Ann), (James, Mike)

2. 从知识图谱采样得到路径, 每一路径链接上述每个训练样例中两个实体:

(David, Mike)对应路径: $\text{Couple} \rightarrow \text{Mother}$

(David, James)对应路径: $\text{Father} \rightarrow \text{Mother}^{-1}$ (Mother^{-1} 与 Mother 为相反关系)

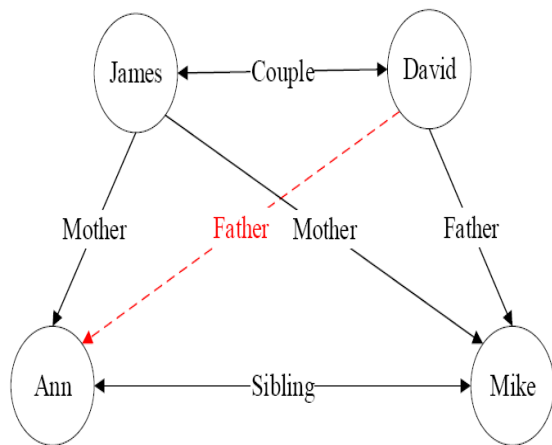
(James, Ann)对应路径: $\text{Mother} \rightarrow \text{Sibling}$

(James, Mike)对应路径: $\text{Couple} \rightarrow \text{Father}$

知识图谱推理：路径排序

$$\text{score}(s, t) = \sum_{\pi_j \in p_l} \theta_j \text{father}(s \rightarrow t; \pi_j)$$

Score(Father(David, Ann))



3. 对于每一个正例/负例，判断上述四条路径可否链接其包含的两个实体，将可链接（记为1）和不可链接（记为0）作为特征，于是每一个正例/负例得到一个四维特征向量：

(David, Mike): $\{[1, 0, 0, 0], 1\}$

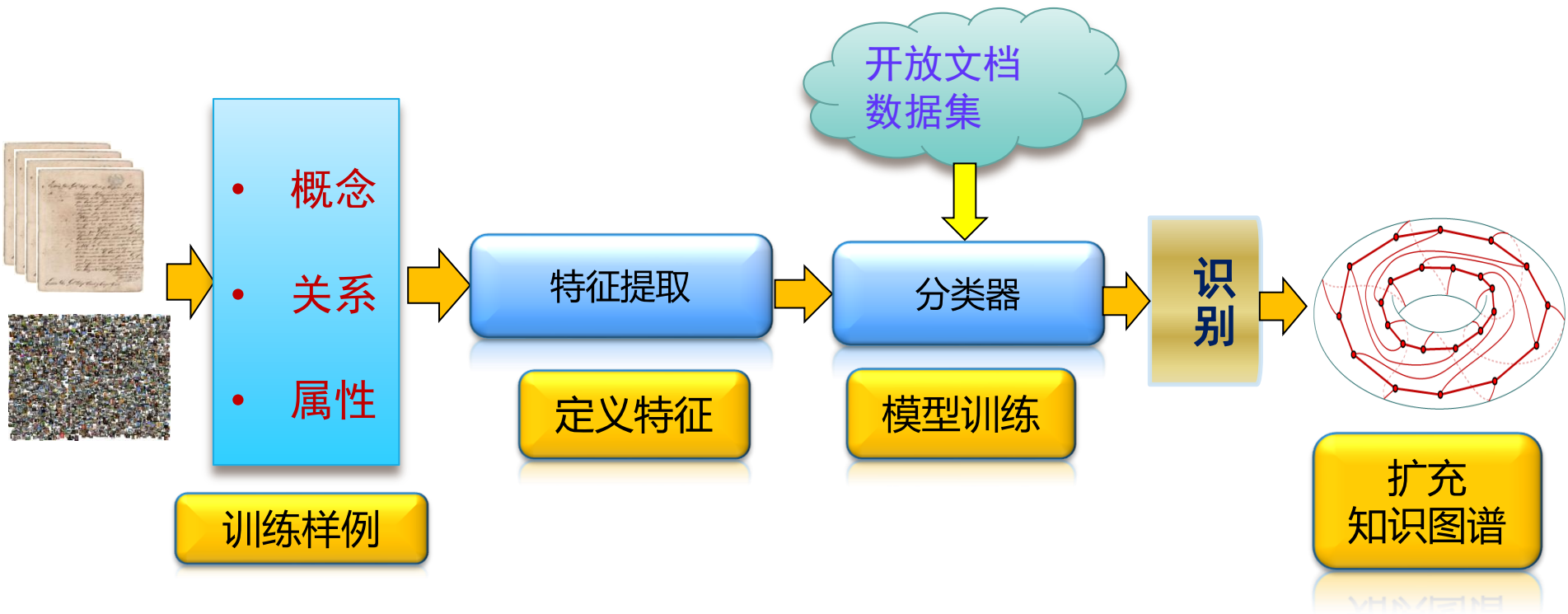
(David, James): $\{[0, 1, 0, 0], -1\}$

(James, Ann): $\{[0, 0, 1, 0], -1\}$

(James, Mike): $\{[0, 0, 1, 1], -1\}$

4. 依据训练样本，训练分类器 M

知识图谱推理: 机器学习



知识图谱构造流程：以Wiki为例子

正文描述

William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, investor, programmer, inventor and philanthropist. [2][3][4] Gates is the former chief executive and chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen.

Gates was born in Seattle, Washington, to William H. Gates, Sr. and Mary Maxwell Gates. His ancestry includes English, German, and Scotts-Irish. [15][16] His father was a prominent lawyer, and his mother served on the board of directors for First Interstate BancSystem and the United Way. Gates's maternal grandfather was JW Maxwell, a national bank president. After being named one of Good Housekeeping's "50 Most Eligible Bachelors" in 1985, [71] Gates married Melinda French on January 1, 1994. They have three children: daughters Jennifer Katharine (b. 1996) and Phoebe Adele (b. 2002), and son Rory John (b. 1999). The family resides in

Bill Gates



Gates in 2013

Born	<u>William Henry Gates III</u> <u>October 28, 1955 (age 58)</u> <u>Seattle, WA, US</u>
Residence	<u>Medina, WA, US</u>
Alma mater	<u>Harvard University</u> (dropped out)
Children	<u>Jennifer, Rory, and Phoebe</u>
Parents	<u>William H. Gates, Sr.</u> <u>Mary Maxwell Gates</u>
Signature	<i>William H. Gates III</i>

属性定义
及描述

44个
类别
标签

Categories: Bill Gates | 1955 births | American billionaires | American chairmen of corporations | American computer businesspeople | American computer programmers | American financiers | American humanitarians | American inventors | American investors | American nonprofit chief executives | American people of English descent | American people of German descent | American people of Scotch-Irish descent | American people of Scottish descent | American philanthropists | American Roman Catholics | American software engineers | American technology chief executives | American technology company founders | American technology writers | Big History | Bill & Melinda Gates Foundation people | Business people from Seattle | Businesspeople in software | Directors of Berkshire Hathaway | Directors of Microsoft | Fellows of the British Computer Society | Gates family | Giving Pledgers | Harvard University people | History of computing | History of Microsoft | Honorary Knights Commander of the Order of the British Empire | Lakeside School alumni | Living people | Members of the United States National Academy of Engineering | Microsoft employees | National Medal of Technology recipients | Personal computing | People from King County, Washington | Placards of the Order of the Aztec Eagle (Mexico) | Windows people | Writers from Seattle, Washington

Wiki中用户对“Bill Gates”这个实例的标注

知识图谱构造流程：以Wiki为例子

- 从概念的专业分类(**taxonomy**)到大众分类(**folksonomy**), 即用户趋向于用自我定义的标签对内容进行组织和分类。
- 44个类别标签: Bill Gates, 1955 births, American billionaires, American chairmen of corporations, American computer, business people, American computer programmers, American financiers, American humanitarians, American inventors, American investors, American nonprofit chief executives, American people of English descent, American people of German descent, American people of Scotch-Irish descent, American people of Scottish descent, American philanthropists, American Roman Catholics, American software engineers, American technology chief executives, American technology company founders, American technology writers, Big History, Bill & Melinda Gates Foundation, people Business, people from Seattle, Business people in software, Directors of Berkshire Hathaway.....

William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, investor, programmer, inventor and philanthropist. [2][3][4] Gates is the former chief executive and chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen.

Gates was born in Seattle, Washington, to William H. Gates, Sr. and Mary Maxwell Gates. His ancestry includes English, German, and Scots-Irish. [15][16] His father was a prominent lawyer, and his mother served on the board of directors for First Interstate BancSystem and the United Way. Gates's maternal grandfather was JW Maxwell, a national bank president. After being named one of *Good Housekeeping's* "50 Most Eligible Bachelors" in 1985, [71] Gates married Melinda French on January 1, 1994. They have three children: daughters Jennifer Katharine (b. 1996) and Phoebe Adele (b. 2002), and son Rory John (b. 1999). The family resides in

Bill Gates

Gates in 2013

Born	William Henry Gates III October 28, 1955 (age 58) Seattle, WA, US
Residence	Medina, WA, US
Alma mater	Harvard University (dropped out)
Children	Jennifer, Rory, and Phoebe
Parents	William H. Gates, Sr. Mary Maxwell Gates
Signature	<i>William H. Gates III</i>

Categories: Bill Gates | 1955 births | American billionaires | American chairmen of corporations | American computer businesspeople | American financiers | American humanitarians | American inventors | American investors | American nonprofit chief executives | American people of English descent | American people of German descent | American people of Scotch-Irish descent | American people of Scottish descent | American philanthropists | American Roman Catholics | American software engineers | American technology chief executives | American technology company founders | American technology writers | Big History | Bill & Melinda Gates Foundation people | Business people from Seattle | Businesspeople in software | Directors of Berkshire Hathaway | Directors of Microsoft | Fellows of the British Computer Society | Gates family | Giving Pledgers | Harvard University people | History of computing | History of Microsoft | Honorary Knights Commander of the Order of the British Empire | Lakeside School alumni | Living people | Members of the United States National Academy of Engineering | Microsoft employees | National Medal of Technology recipients | Personal computing | People from King County, Washington | Placards of the Order of the Aztec Eagle (Mexico) | Windows people | Writers from Seattle, Washington

知识图谱构造流程：以Wiki为例子

William Henry "Bill" Gates III (born October 28, 1955) is an American business magnate, investor, programmer, inventor and philanthropist. [2][3][4] Gates is the former chief executive and chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen.

Gates was born in Seattle, Washington, to William H. Gates, Sr. and Mary Maxwell Gates. His ancestry includes English, German, and Scots-Irish. [15][16] His father was a prominent lawyer, and his mother served on the board of directors for First Interstate BancSystem and the United Way. Gates's maternal grandfather was JW Maxwell, a national bank president. After being named one of Good Housekeeping's "50 Most Eligible Bachelors" in 1985, [71] Gates married Melinda French on January 1, 1994. They have three children: daughters Jennifer Katharine (b. 1996) and Phoebe Adele (b. 2002), and son Rory John (b. 1999). The family resides in



- 手工构造的种子知识(Infobox):
- 比尔盖茨这个实体用了13个属性描述、奥巴马这个实体用了20多个属性描述(由于两者均属于persons这个类别,有些属性是共享的)
- Freebase中定义的cities这个概念时使用了将近200个属性,但是仍然远远不够。

□ 如何发现和学习新的属性?

Categories: Bill Gates | 1955 births | American billionaires | American chairmen of corporations | American computer businesspeople | American computer programmers | American financiers | American humanitarians | American inventors | American investors | American nonprofit chief executives | American people of English descent | American people of German descent | American people of Scotch-Irish descent | American people of Scottish descent | American philanthropists | American Roman Catholics | American software engineers | American technology chief executives | American technology company founders | American technology writers | Big History | Bill & Melinda Gates Foundation people | Business people from Seattle | Businesspeople in software | Directors of Berkshire Hathaway | Directors of Microsoft | Fellows of the British Computer Society | Gates family | Giving Pledgers | Harvard University people | History of computing | History of Microsoft | Honorary Knights Commander of the Order of the British Empire | Lakeside School alumni | Living people | Members of the United States National Academy of Engineering | Microsoft employees | National Medal of Technology recipients | Personal computing | People from King County, Washington | Placards of the Order of the Aztec Eagle (Mexico) | Windows people | Writers from Seattle, Washington

知识图谱构造流程：以Wiki为例子

- 基于机器学习算法进行概念、属性和关系学习，需要大量良好标注数据
- Wikipedia当中，英文文章占总文章数的38.95%，中文文章占4.75%，其他语言文章占56.30%。其中，38.60%的英文文章以Infobox方式被标注(如实体属性或实体之间的关系)，21.43%的中文文章被标注，平均28.57%的其他语言文章被标注。

William Henry "Bill" Gates III
(born October 28, 1955) is an American business magnate, investor, programmer, inventor and philanthropist. [2][3][4]. Gates is the former chief executive and chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen.

Gates was born in Seattle, Washington, to William H. Gates, Sr. and Mary Maxwell Gates. His ancestry includes English, German, and Scots-Irish. [15][16] His father was a prominent lawyer, and his mother served on the board of directors for First Interstate BancSystem and the United Way. Gates's maternal grandfather was JW Maxwell, a national bank president. After being named one of Good Housekeeping's "50 Most Eligible Bachelors" in 1985, [71] Gates married Melinda French on January 1, 1994. They have three children: daughters Jennifer Katharine (b. 1996) and Phoebe Adele (b. 2002), and son Rory John (b. 1999). The family resides in

Bill Gates



Gates in 2013

Born	William Henry Gates III October 28, 1955 (age 58) Seattle, WA, US
Residence	Medina, WA, US
Alma mater	Harvard University (dropped out)
Children	Jennifer, Rory, and Phoebe
Parents	William H. Gates, Sr. Mary Maxwell Gates
Signature	<i>William H. Gates III</i>

用户对wiki中Bill Gates文章所标注的知识

知识图谱构造流程：以Wiki为例子

给定无结构化文档数据，通过机器学习方法对实体描述内容进行分类，同时提取描述实体的属性和对应属性值。

嫦娥三号类别: 2013 in China; Space probes launched in 2013; Chinese Lunar Exploration Program; Chinese space probes; Lunar rovers; Missions to the Moon; Spacecraft that orbited the Moon; Soft landings on the Moon; Space probes



机器学习

种子知识

来自InfoBox

名称	嫦娥三号月球探测器
所属国家	中华人民共和国
构成	着陆器、“玉兔号”月球车
重量	3,750千克
关键技术	7500牛变推力发动机、热控两相流体回路、可 变热导热管等

类别分类、实体识别、属性填充

知识图谱构造流程：属性识别与填充

纳尔逊 罗利赫拉赫拉 曼德拉 (Nelson Rolihlahla Mandela)
1918年7月18日出生于南非特兰斯凯一个大酋长家庭，先后获南非大学文学士和威特沃特斯兰德大学律师资格，当过律师。曼德拉自幼性格刚强，崇敬民族英雄。他是家中长子而被指定为酋长继承人。但他表示：“决不愿以酋长身份统治一个受压迫的部族”，而要“以一个战士的名义投身于民族解放事业”。他毅然走上了追求民族解放的道路。
南非政府2013年12月6日（北京时间）宣布南非前总统曼德拉去世，享年95岁。
曼德拉，南非黑人领袖，因其在废除南非种族歧视政策方面作出了巨大贡献而于1993年荣获诺贝尔和平奖。
1918年7月18日，曼德拉出生于南非特兰斯凯，曼德拉自幼性格刚强，崇敬民族英雄。他是家中长子而被指定为酋长继承人。但他表示：“决不愿以酋长身份统治一个受压迫的部族”，而要“以一个战士的名义投身于民族解放事业”，他毅然……

无结构化文档



人物	
中文名	纳尔逊 罗利赫拉赫拉 曼德拉
外文名	Nelson Rolihlahla Mandela
国籍	南非
出生地	南非特兰斯凯
出生日期	1918年7月18日
职业	政治 南非总统
主要成就	诺贝尔和平奖得主

所提取的实体及其属性值
形式的知识

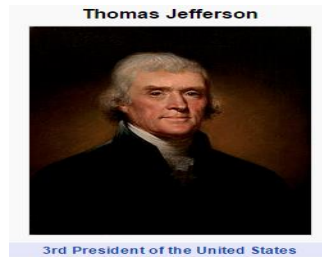
知识图谱推理: 机器学习

infoboxes



Born: February 22, 1732, Westmoreland County, Virginia, United States
Died: December 14, 1799, Mount Vernon, Virginia, United States
Spouse: Martha Washington (m. 1759–1799)
Presidential term: April 30, 1789 – March 4, 1797
Siblings: Lawrence Washington (1718–1752), more

从数据中学习人名
这个类别所定义的
各个属性分类器



Born:
Died: .
Party:
Presidential term:
Children:



知识图谱推理: 机器学习

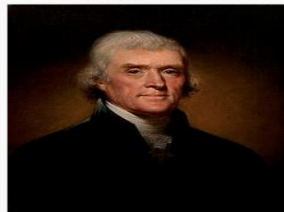
George Washington



George Washington by Gilbert Stuart, 1797

训练得到的
属性分类器

Thomas Jefferson



3rd President of the United States

infoboxes

Born: February 22, 1732, Westmoreland County, Virginia, United States

Died: December 14, 1799, Mount Vernon, Virginia, United States

Spouse: Martha Washington (m. 1759–1799)

Presidential term: April 30, 1789 – March 4, 1797

Siblings: Lawrence Washington (1718–1752), more

Born: April 13, 1743, Shadwell, Virginia, United States

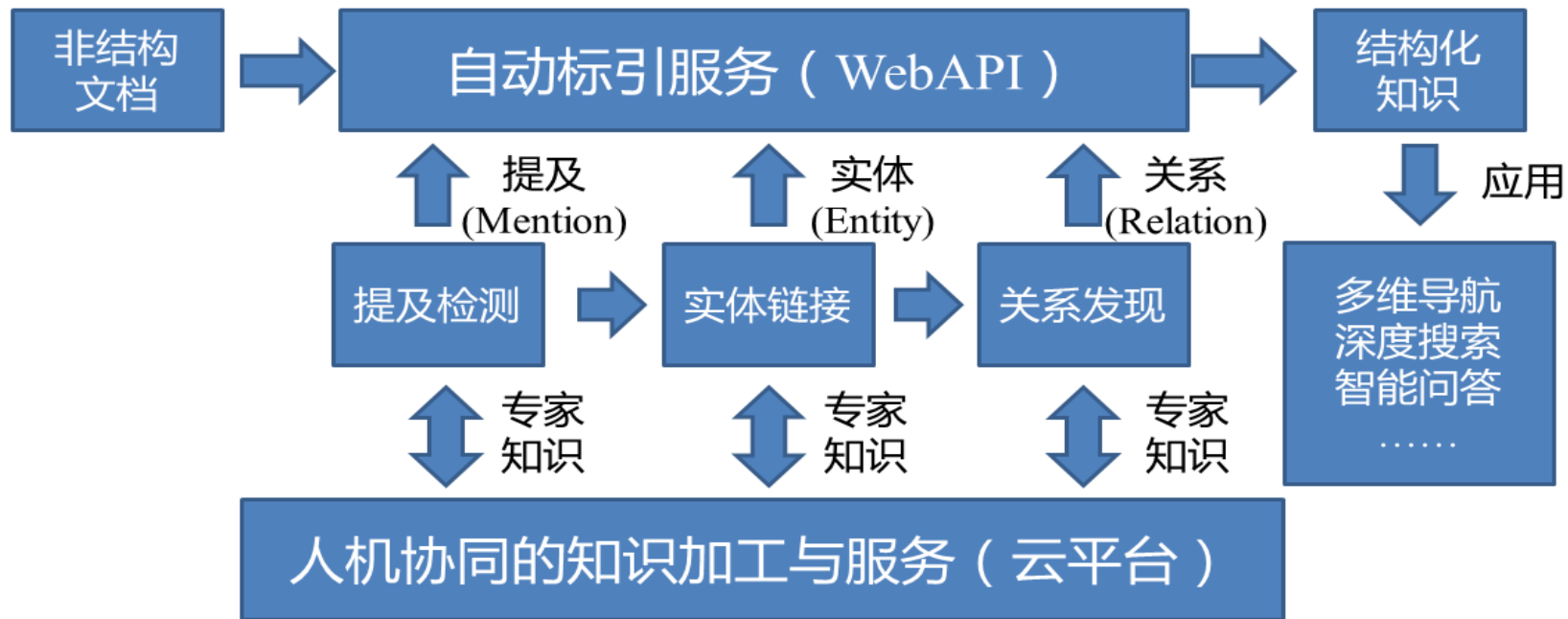
Died: July 4, 1826, Monticello, Virginia, United States

Party: Democratic-Republican Party

Presidential term: March 4, 1801 – March 4, 1809

Children: Martha Jefferson Randolph, Eston Hemings,

知识图谱：从数据到知识、从知识到决策



提纲

- 1、命题逻辑
- 2、谓词逻辑
- 3、知识图谱推理
- 4、因果推理

相关不意味着因果



致酒行（唐-李贺）

零落栖迟一杯酒，主人奉觞客长寿。
主父西游困不归，家人折断门前柳。
吾闻马周昔作新丰客，天荒地老无人识。
空将笺上两行书，直犯龙颜请恩泽。
我有迷魂招不得，**雄鸡一声天下白。**
少年心事当拿云，谁念幽寒坐呜呃。

公鸡打鸣与太阳升起

因果推理 (Causal Inference) : Simpson's Paradox (辛普森悖论)

1973年伯克利本科生录取率

	男生		女生	
	申请数	录取率	申请数	录取率
整体	8442	44%	4321	35%

男生录取率(44%)远高于女生(35%)

学院	男生		女生	
	申请数	录取率	申请数	录取率
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

6个最大的院系中，4个院系女生录取率大于男生。如果按照这样的分类，女生实际上比男生的录取率还高一点点。

女生更愿意申请那些竞争压力很大的院系（比如英语系），但是男生却更愿意申请那些相对容易进的院系（比如工程学系）。

Peter J. Bickel, Eugene A. Hammel, O'Connell, J. W, Sex bias in graduate admissions: Data from Berkeley, *Science*, 187(4175):398-404, 1975


因果推理 (Causal Inference) : Simpson's Paradox (辛普森悖论)

身高(cm)	计算机学院	文学院
矮个人数 (<160)	60	80
高个人数 (≥160)	290	270
高个率 (%)	82.9	77.1

左：计算机学院和文学院学生的身高情况

	计算机学院		文学院	
身高(cm)	男生	女生	男生	女生
矮个人数 (<160)	35	25	10	70
高个人数 (≥160)	235	55	80	190
高个率 (%)	87	68.9	88.9	73.1

右：以性别分组后的计算机学院和文学院的学生身高情况

$$\frac{b}{a} < \frac{d}{c}, \frac{b'}{a'} < \frac{d'}{c'}$$

$$\frac{b+b'}{a+a'} > \frac{d+d'}{c+c'}$$

- 计算机学院学生的高个率高于文学院（左表）。
- 分别比较两所学院男生和女生身高时，却发现计算机学院男生和女生的高个率均低于文学院
- 在总体样本上成立的某种关系却在分组样本里恰好相反。


因果推理 (Causal Inference) : Simpson's Paradox (辛普森悖论)

身高(cm)	计算机学院	文学院
矮个人数 (<160)	60	80
高个人数 (≥160)	290	270
高个率 (%)	82.9	77.1

左：计算机学院和文学院学生的身高情况

	计算机学院		文学院	
身高(cm)	男生	女生	男生	女生
矮个人数 (<160)	35	25	10	70
高个人数 (≥160)	235	55	80	190
高个率 (%)	87	68.9	88.9	73.1

右：以性别分组后的计算机学院和文学院的学生身高情况

$$\frac{b}{a} < \frac{d}{c}, \frac{b'}{a'} < \frac{d'}{c'}$$

$$\frac{b+b'}{a+a'} > \frac{d+d'}{c+c'}$$

- 右表体现了男生比女生个子高这一现象，如计算机学院和文学院男生高个率都比女生高个率要大。
- 性别会影响专业选择，计算机学院招收的男生多，而文学院招收的女生多。因此，当计算机学院的样本中包含更多的男生，文学院的样本中包含更多的女生，就会看到左表所呈现的情况：计算机学院的高个率高于文学院。

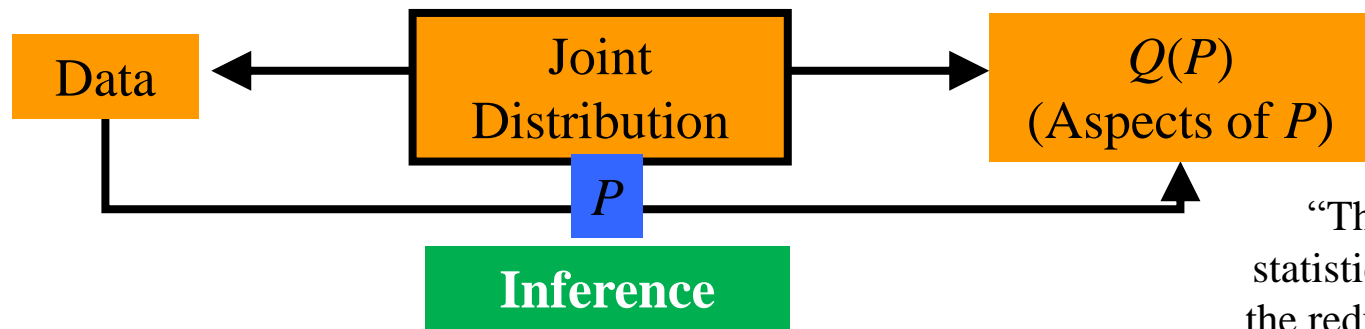
因果推理 (Causal Inference) : Simpson's Paradox (辛普森悖论)

辛普森悖论表明，在某些情况下，忽略潜在的“第三个变量”（如性别就是专业和身高之外的第三个变量），可能会改变已有的结论，而我们常常却一无所知。从观测结果中寻找引发结果的原因，由果溯因，就是本节要介绍的因果推理

不能只满足于数字或图表，必须考虑数据生成过程——因果模型

因果推理 (Causal Inference)

传统以统计建模为核心的推理手段



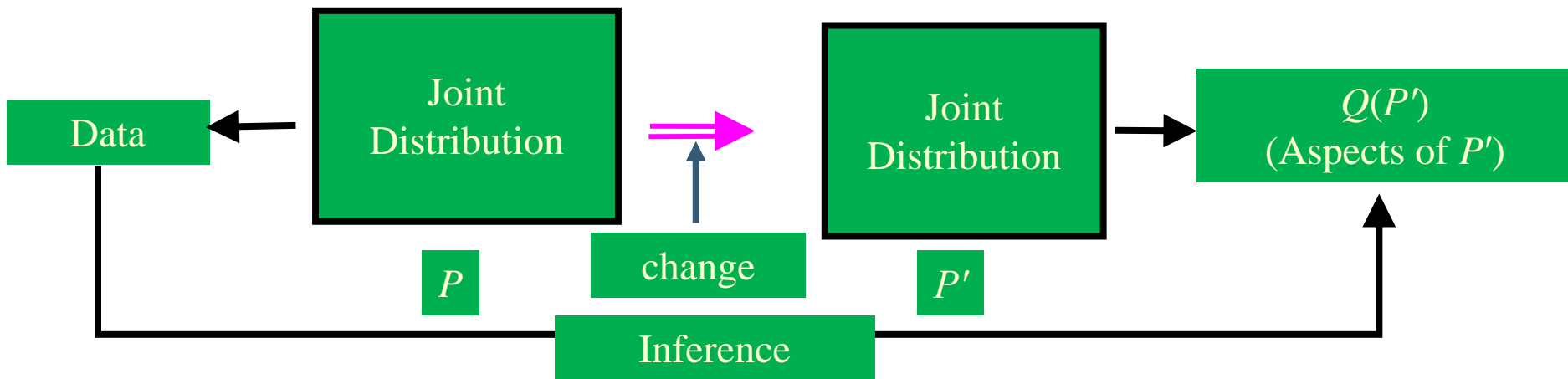
“The object of statistical methods is the reduction of data”
(Fisher 1922).

购买了A商品的顾客是否会购买B商品(对A和B的联合分布建模)

$$Q = P(B | A)$$

因果推理 (Causal Inference)

从统计建模推断到因果推理



数据分布从 P 变换到 P'

- 如果商品价格涨价一倍，预测销售量 P' (sales)的变化
- 如果放弃吸烟，预测癌症 P' (cancer) 的概率

因果推理模型: 结构因果模型和因果图

- 结构因果模型(structural causal model, SCM), 也被称为因果模型 (causal model) 或Neyman–Rubin因果模型。这一模型最早可追溯于Jerzy Neyman在1923年用波兰语所撰写的博士论文中提出的“潜在结果” (potential outcome) 的概念。之后, Donald Rubin发展了“潜在结果”这一概念, 并将其和缺失数据的理论联系在一起。
- 因果图 (causal diagram) 由Judea Pearl于 1995年提出。
- 每个结构因果模型 M 都与一个因果图 G 相对应

因果推理的层级

可观测性问题	What if we see A (what is?)	$P(y A)$
决策行动问题	What if we do A (what if?)	$P(y do(A))$ (如果采取A行为, 则B真)
反事实问题 (Counterfactual)	What if we did things differently	(why?) $P(y' A)$ (如果A为真, 则B将不同)
Options: with what probability		

Actions: B will be true if we do A .

Counterfactuals: B would be different if A were true

因果推理：有向无环图

有向无环图(directed acyclic graphs, DAG): 有向无环图指的是一个无回路的有向图, 即从图中任意一个节点出发经过任意条边, 均无法回到该节点。有向无环图刻画了图中所有节点之间的依赖关系。

DAG 可用于描述数据的生成机制。这样描述变量联合分布或者数据生成机制的模型, 被称为“贝叶斯网络”(Bayesian network)。

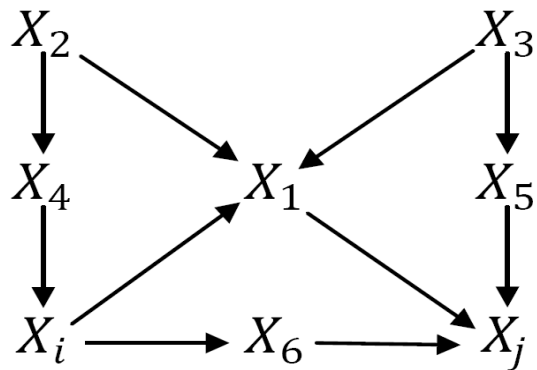
对于任意的有向无环图模型, 模型中 d 个变量的联合概率分布由每个节点与其父节点之间条件概率 $P(child|parents)$ 的乘积给出:

$$P(x_1, x_2, \dots, x_d) = \prod_{j=1}^d P(x_j | x_{pa(j)})$$

其中, $x_{pa(j)}$ 表示节点 x_j 的父节点集合(所有指向 x_j 的节点)。

因果推理：有向无环图

DAG可被视为因果过程：父辈节点“促成”了孩子节点的取值



有向无环图DAG

联合分布可表示为：

$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_i, X_j) \\ &= P(X_2) \times P(X_3) \times P(X_1 | X_2, X_3, X_i) \\ &\quad \times P(X_4 | X_2) \times P(X_5 | X_3) \times P(X_6 | X_i) \times P(X_i | X_4) \\ &\quad \times P(X_j | X_1, X_5, X_6) \end{aligned}$$

显然：一个有向无环图唯一地决定了一个联合分布；反过来，一个联合分布不能唯一地决定有向无环图。反过来的结论不成立。如联合分布 $P(X_1, X_2) = P(x_1)P(x_2|x_1) = P(x_2)P(x_1|x_2)$

因果推理：有向无环图

例 假设某个有向无环图中存在一条依赖路径 $X \rightarrow Y \rightarrow Z$ ，其中 X 表示气候好， Y 表示水果产量高， Z 表示水果价格低，给出 $P(\text{气候好}, \text{水果产量高}, \text{水果价格低})$ 的联合概率。

解 使用乘积分解规则，将 $P(\text{气候好}, \text{水果产量高}, \text{水果价格低})$ 转换为：
 $P(\text{气候好}) \times P(\text{水果产量高} | \text{气候好}) \times P(\text{水果价格低} | \text{水果产量高})$

根据常识，假设：

$$P(\text{气候好}) = 0.5$$

$$P(\text{水果产量高} | \text{气候好}) = 0.8$$

$$P(\text{水果价格低} | \text{水果产量高}) = 0.9$$

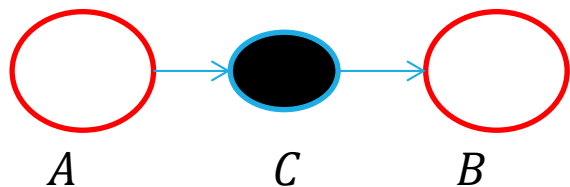
$$\begin{aligned} &P(\text{气候好}, \text{水果产量高}, \text{水果价格低}) \\ &= P(\text{气候好}) \times P(\text{水果产量高} | \text{气候好}) \times P(\text{水果价格低} | \text{水果产量高}) \\ &= 0.5 \times 0.8 \times 0.9 = 0.36 \end{aligned}$$

因果推理：D-分离(directional separation, *d*-separation)

D-分离用于判断集合A中变量是否与集合B中变量相互独立（给定集合C），记为

$$A \perp B \mid C$$

D-分离的例子
(serial connection)



当C取值固定(可观测, observed)，有

$$P(A, B \mid C) = \frac{P(A, B, C)}{P(C)} = \frac{P(A)P(C \mid A)P(B \mid C)}{P(C)} = P(A \mid C)P(B \mid C)$$

可见A和B在C取值固定情况下，是条件独立的。

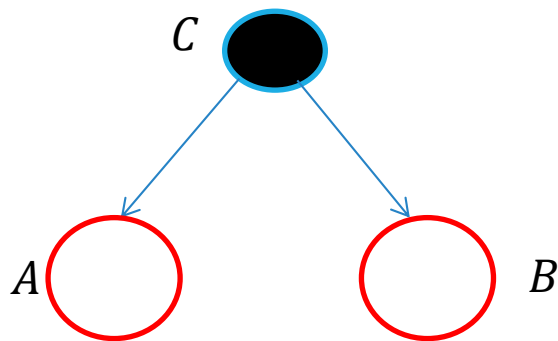
注：上式利用了 $P(A)P(C \mid A) = P(C)P(A \mid C)$

因果推理：D-分离(directional separation, *d*-separation)

D-分离用于判断集合A中变量是否与集合B中变量相互独立（给定集合C），记为

$$A \perp B \mid C$$

D-分离的例子
(diverging connection)



当C取值固定(observed)，有

$$P(A, B \mid C) = \frac{P(A, B, C)}{P(C)} = \frac{P(C)P(A \mid C)P(B \mid C)}{P(C)} = P(A \mid C)P(B \mid C)$$

可见A和B在C取值固定情况下，是条件独立的。

如果C不固定，则有 $P(A, B) = \sum_C P(A \mid B)P(B \mid C)P(C)$
可见，由于 $P(A, B) \neq P(A)P(B)$ ，因此A和B在条件C下不独立的。

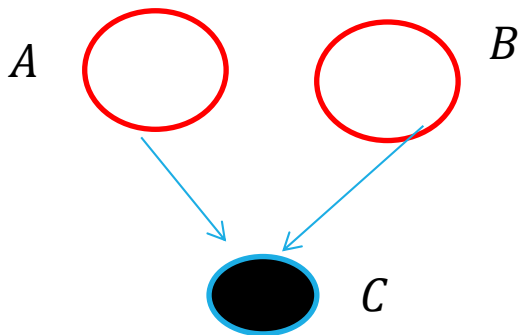
因果推理：D-分离(directional separation, *d*-separation)

D-分离用于判断集合A中变量是否与集合B中变量相互独立（给定集合C），记为

$$A \perp B \mid C$$

$$P(A, B, C) = P(A)P(B)P(C|A, B)$$

D-分离的例子
(V-structure connection)



- 如果C不作为观测点

$$P(A, B, C) = P(A)P(B) \sum_C P(C|A, B) = P(A)P(B)$$

($\sum_C P(C|A, B) = 1$) A和B在条件C下是独立的

- 如果C取值固定(observed)

$$P(A, B|C) = \frac{P(A, B, C)}{P(C)} = \frac{P(A)P(B)P(C|A, B)}{P(C)} \neq P(A)P(B)$$

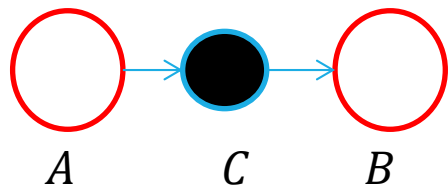
A和B在条件C下是不独立的（是相关的）

因果推理：D-分离(directional separation, d -separation)

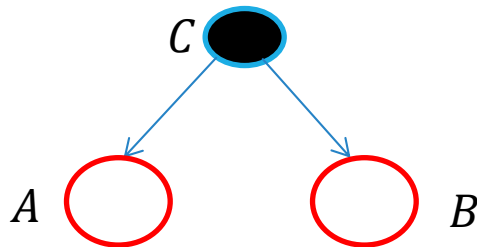
D-分离：对于一个DAG图，如果 A 、 B 、 C 是三个集合（可以是单独的节点或者是节点的集合），为了判断 A 和 B 是否是 C 条件独立的，在DAG图中考虑所有 A 和 B 之间的路径(不管方向)。对于其中的一条路径，如果满足以下两个条件中的任意一条，则称这条路径是阻塞（block）的：

1. 路径中存在某个节点 X 是链结构 $A \rightarrow C \rightarrow B$ 或分连结构 $A \leftarrow C \rightarrow B$ 中的节点、且 X 包含在 C 中
2. 或者路径中存在某个节点 X 是汇连结构 $A \rightarrow C \leftarrow B$ 中节点，并且 X 或 X 后代不包含在 C 中

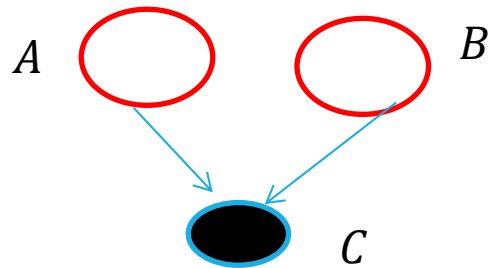
如果 A 和 B 之间所有路径都是阻塞的，那么 A 和 B 就是关于 C 条件独立的；否则 A 和 B 不是关于 C 条件独立



链结构

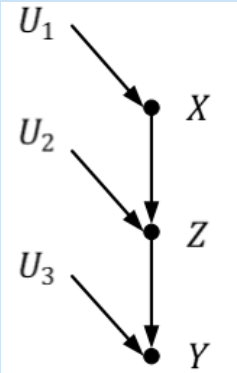
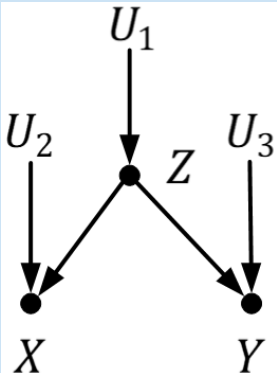
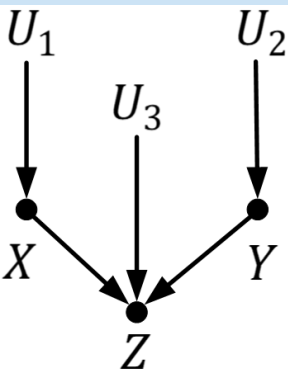


分连结构



汇连结构(也叫V结构或称碰撞点, collider)

因果推理：D-分离(directional separation, d -separation)

链结构(chain)	分连结构(fork)	汇连（或碰撞）结构(collider)
		
Z和X是相关的	X和Z是相关的	Z和X是相关的
Y和Z是相关的	Y和Z是相关的	Z和Y是相关的
Y和X很有可能是相关的	Y和X很有可能是相关的	Y和X是相互独立的
给定Z时，Y和X是条件独立的	给定Z时，Y和X是条件独立的	给定Z时，Y和X是相关的

因果推理：D-分离(directional separation, d -separation)

- D-分离(directional separation, d -separation)方法可用于判断因果图上任意变量间相关性和独立性。
- 在因果图上，若两个节点间存在一条路径将这两个节点连通，则称之为是**有向连接**(d -connected)的；若两个节点不是有向连接的，则称之为是**有向分离**(d -separated)的，即不存在这样的路径将这两个节点连通。当两个节点是有向分离时，意味着这两个节点相互独立。
- 若节点 X 和节点 Y 之间的每一条路径都是阻塞的(blocked)，称节点 X 和节点 Y 是有向分离的；反之，若存在一条路径是非阻塞的(unblocked)，称节点 X 和节点 Y 是有向连接的。

因果推理：干预(intervention)和do 算子(do-calculus)

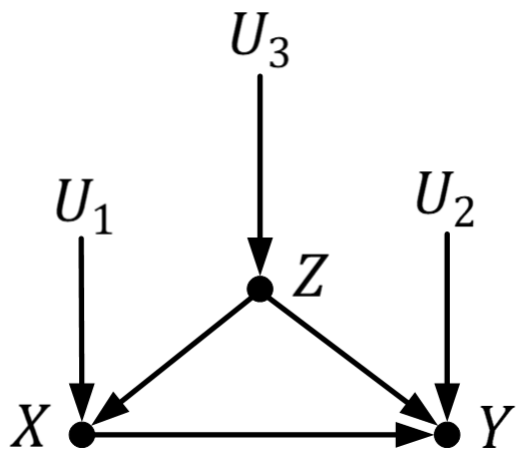
- DAG中具有链接箭头的节点之间存在某种“因果关系”。
- 但是，要在 DAG 上引入“因果”的概念，需要引进do 算子(do-calculus)。do 算子的意思可理解为“干预”(intervention)。
- 在 DAG 中， $do(X_i) = x_i'$ ，表示将DAG中指向节点 X_i 的有向边全部切断，并且将 X_i 的值固定为常数 x_i' 。
- 在这样操作后，所得到新的DAG中变量联合分布为：

$$P(x_1, x_2, \dots, x_d | do(X_i) = x_i')$$

因果推理：干预(intervention)和do 算子(do-calculus)

- **干预** (intervention) 指的是固定 (fix) 系统中某个变量，然后改变系统，观察其他变量的变化。
- 为了与 X 自然取值 x 时进行区分，在对 X 进行干预时，引入“**do算子**”，记作 $do(X = x)$ 。因此， $P(Y = y|X = x)$ 表示当 X 取值为 x 时， $Y = y$ 的概率；而 $P(Y = y|do(X = x))$ 表示对 X 取值进行了干预，固定其取值为 x 时， $Y = y$ 的概率。
- 用统计学的术语来说， $P(Y = y|X = x)$ 反映了在取值为 x 的个体 X 上， Y 的总体分布；而 $P(Y = y|do(X = x))$ 反映的是如果将 X 每一个取值都固定为 x 时， Y 的总体分布。

因果推理：因果效应和因果效应差



两个学院学生身高与学生性别所形成的因果图

二值变量 X ：学院(1表示计算机学院，0表示文学院)

二值变量 Y ：身高情况(1表示高个，0表示矮个)

二值变量 Z ：性别(1表示男生，0表示女生)

U_1, U_2, U_3 ：外生变量

- 为了分析两个学院学生高个率是否存在差别，可对学院这一变量进行干预，即将 X 取值固定为1或0。
- 设 $do(X = 1)$ 和 $do(X = 0)$ 表示这两种干预，如下估计干预后产生的差别：

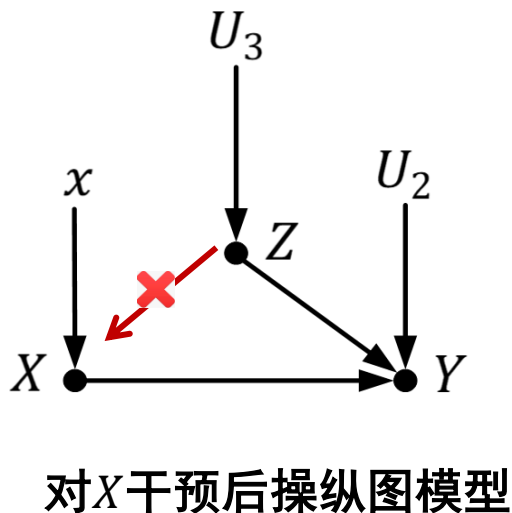
$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$$

该式被称为“**因果效应差**” (causal effect difference)或“平均因果效应” (average causal effect, ACE)

$P(Y = y|do(X = x))$ 被称为**因果效应**(causal effect)

因果推理：操纵图模型和操纵概率

- 对学院变量 X 进行干预并固定其取值为 x 时，可将所有指向 X 的边均移除。
- 因果效应 $P(Y = y|do(X = x))$ 等价于引入干预的**操纵图模型**(manipulated model)中条件概率 $P_m(Y = y|X = x)$ 。



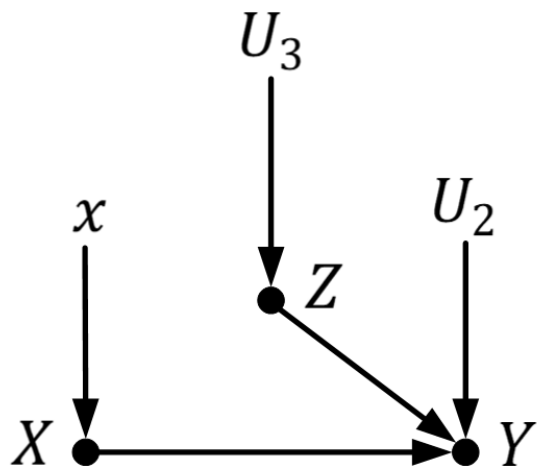
计算因果效应关键在于计算**操纵概率**(manipulated probability) P_m 。 P_m 与正常(无干预)条件下概率 P 在如下两个方面的取值不变：

- 边缘概率 $P(Z = z)$ 不随干预而变化，因为 Z 的取值不会因为去掉从 Z 到 X 的箭头而变化，即：
$$P_m(Z = z) = P(Z = z)$$
- 条件概率 $P(Y = y|X = x, Z = z)$ 不变，因为 Y 关于 X 和 Z 的函数 $f_Y = (X, Z, U_2)$ 并未改变，即：
$$P_m(Y = y|X = x, Z = z) = P(Y = y|X = x, Z = z)$$

因果推理：调整公式

在干预图中， X 和 Z 是 **D -分离**的，因此是彼此独立的，即：

$$P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$$



对 X 干预后操纵图模型
 X 和 Z 相对于 Y 构成了V结构

因果效应 $P(Y = y|do(X = x))$ 有：

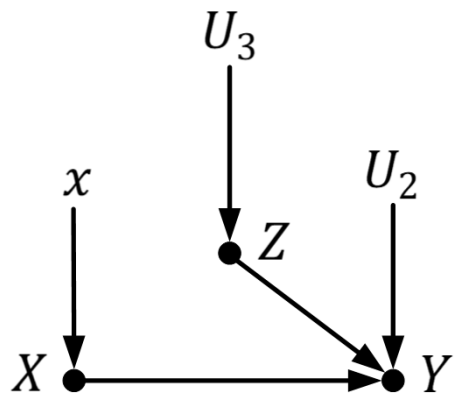
$$\begin{aligned} P(Y = y|do(X = x)) &= P_m(Y = y|X = x) \\ &= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \\ &= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \\ &= \sum_z P(Y = y|X = x, Z = z)P(Z = z) \end{aligned}$$

因果推理：调整公式

因果效应 $P(Y = y|do(X = x))$ 有：

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

该式被称为**调整公式**（adjustment formula），对于 Z 的每一个取值 z ，该式计算 X 和 Y 的条件概率并取均值，这个过程称之为“ Z 调整”（adjusting for Z ）或“ Z 控制”（controlling for Z ）。



对 X 干预后操纵图模型

该式的右端只包含正常(无干预)条件下的概率 P ，即可用**正常(无干预)条件**下的条件概率来计算**干预后**的条件概率。

因果推理：调整公式

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

	计算机学院		文学院	
身高(cm)	男生	女生	男生	女生
矮个人数 (<160)	35	25	10	70
高个人数 (≥ 160)	235	55	80	190
高个率 (%)	87	68.9	88.9	73.1

下面将调整公式用于计算对 X 取值进行干预后计算机学院/文学院高个子率，其中 $X = 1$ 表示计算机学院， $Y = 1$ 表示高个， $Z = 1$ 表示表示男生，则有：

$$\begin{aligned} P(Y = 1|do(X = 1)) \\ &= P(Y = 1|X = 1, Z = 1)P(Z = 1) \\ &+ P(Y = 1|X = 1, Z = 0)P(Z = 0) \end{aligned}$$

以性别分组后的计算机学院和
文学院的学生身高情况

因果推理：调整公式

X干预取值为计算机学院的因果效应为：

$$\begin{aligned} P(Y = 1|do(X = 1)) \\ &= 0.87 \times \frac{(35 + 235 + 10 + 80)}{(350 + 350)} + 0.689 \\ &\times \frac{(25 + 55 + 70 + 190)}{(350 + 350)} = 0.782 \end{aligned}$$

X干预取值为文学院的因果效应为：

$$\begin{aligned} P(Y = 1|do(X = 0)) \\ &= 0.889 \times \frac{(35 + 235 + 10 + 80)}{(350 + 350)} + 0.731 \times \frac{(25 + 55 + 70 + 190)}{(350 + 350)} \\ &= 0.812 \end{aligned}$$

其因果效应差：

$$\begin{aligned} ACE &= P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \\ &= 0.782 - 0.812 = -0.03 \end{aligned}$$

可见计算机学院的高个率不如文学院的高个率。
即学院这一变量影响了高个率。

	计算机学院		文学院	
身高(cm)	男生	女生	男生	女生
矮个人数 (<160)	35	25	10	70
高个人数 (≥160)	235	55	80	190
高个率 (%)	87	68.9	88.9	73.1

以性别分组后的计算机学院和
文学院的学生身高情况

因果推理：因果图的不足

- 实际中难以得到一个完整的DAG，用于阐述变量之间的因果关系或者数据生成机制，使得 DAG 的应用受到的巨大的阻碍。从观测数据学习 DAG 的结构，是充满挑战的问题。
- Pearl 引入 **do 算子**是因果推断领域最主要贡献。所谓 “**do**”，就是 “干预”，即从系统之外人为控制某些变量。但是，这依赖于一个假定：干预某些变量并不会引起 DAG 中其他结构的变化。
- DAG 作为一种简化的模型，在复杂系统中可能不完全适用，需要将其拓展到动态系统（如时间序列），还有待研究。

因果推理：反事实推理 (counterfactual model)

- “反事实”框架是科学哲学家大卫·刘易斯 (David Lewis) 等人提出的推断因果关系的标准。
- 事实是指在某个特定变量 (A) 的影响下可观测到的某种状态或结果 (B)。
“反事实”是指在该特定变量 (A) 取负向值时可观测到的状态或结果 (B')。
- 条件变量对于结果变量的因果性就是 A 成立时 B 的状态与 A 取负向值时“反事实”状态 (B') 之间的差异。
- 如果这种差异存在且在统计上是显著的，说明条件变量与结果变量存在因果关系。

推理总结

推理方法	推理方式	说明
归纳推理	如果 A_i (i 为若干取值), 那么B	从若干事实出发推理出一般性规律
演绎推理	如果A, 那么B	A是B的前提、但不是唯一前提, 因此A是B的充分条件。当然, 在特殊情况下A也可作为B的充分必要条件
因果推理	因为A, 所以B	A是B的唯一前提, 因此“如果没有A, 那么没有B”也成立。