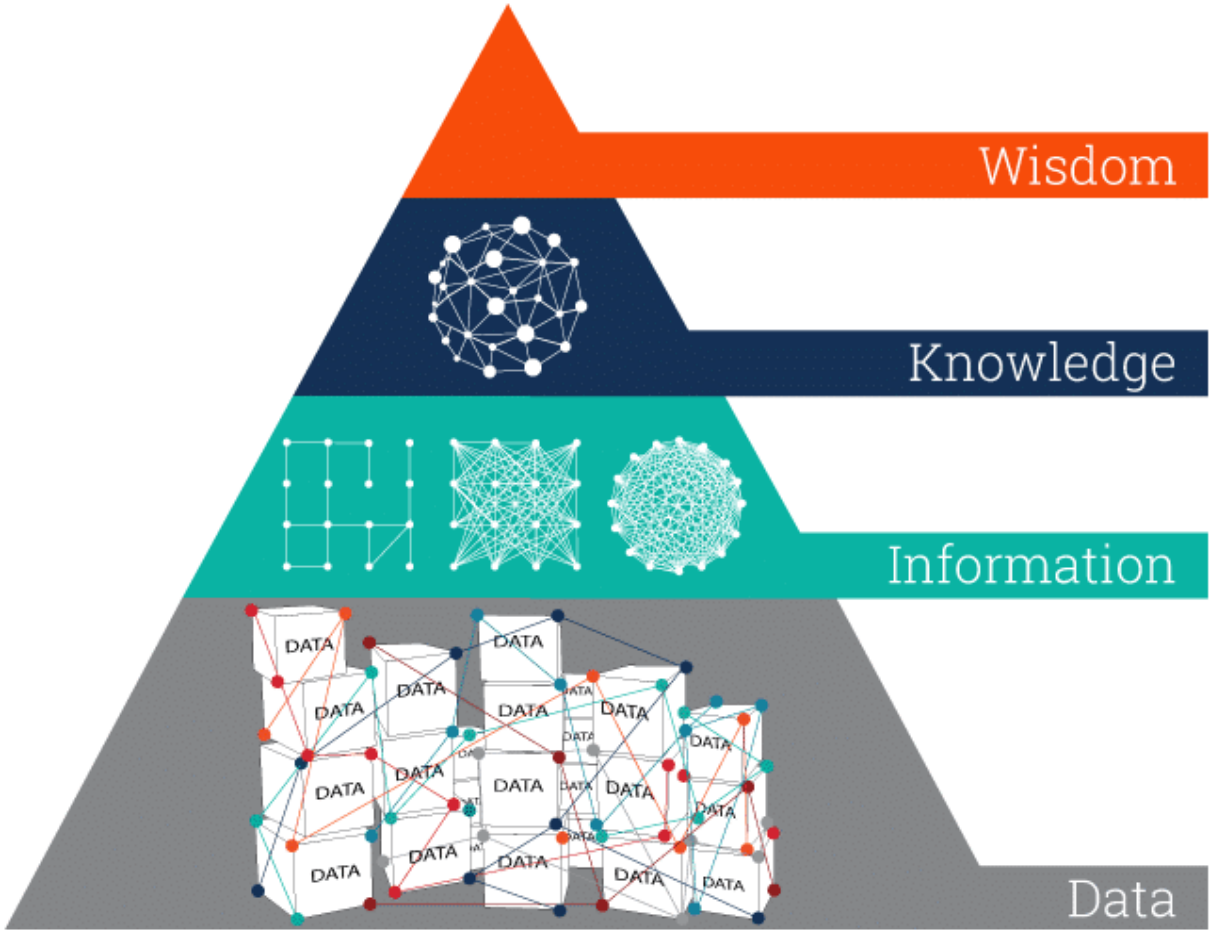
数据科学基础

Foundations of Data Science

1.1 DIKW： 数据、信息、知识和智慧

陈振宇
南京大学智能软件工程实验室
www.iselab.cn



Russell Ackoff “From Data to Wisdom”,
Journal of Applied Systems Analysis, 1989(16):3-9

Data (数据)

220502198602221300130202198602229494411724198602229439511702198602223892
431381198602226595620802198602224158450106198602224559150423198602224 26X
350824198602222462513433198602228458440515198602226016140624198602229904
140601198602223 33X510724198602229730522325198602221374210905198602226960
451023198602228439340501198602222507330101198602223549141126198602222661

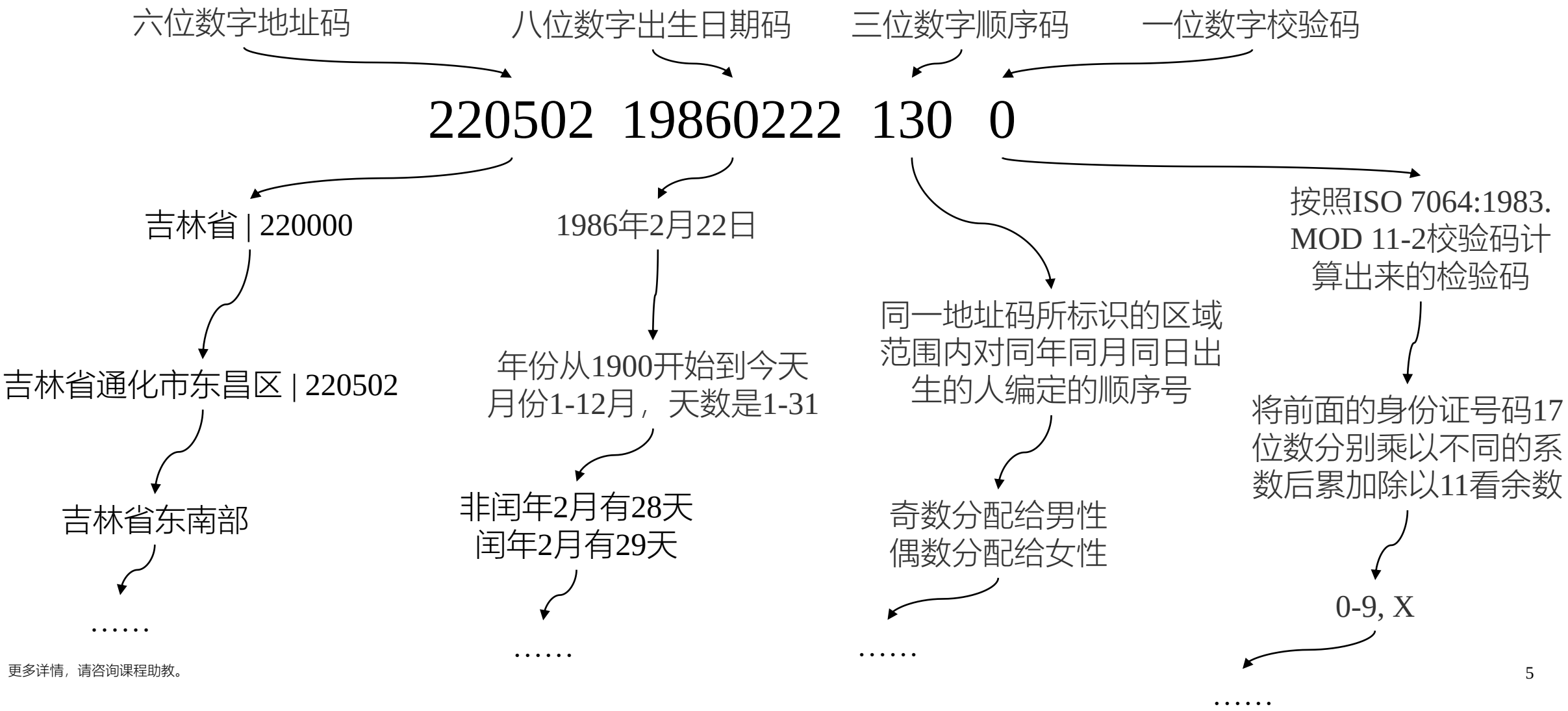
Information (信息)

22050219860221300 13020219860229494 41172419860229439 51170219860223892
43138119860226595 62080219860224158 45010619860224559 15042319860222426X
35082419860222462 51343319860228458 44051519860226016 14062419860229904
14060119860222333X 51072419860229730 52232519860221374 21090519860226960
45102319860228439 34050119860222507 33010119860223549 14112619860222661

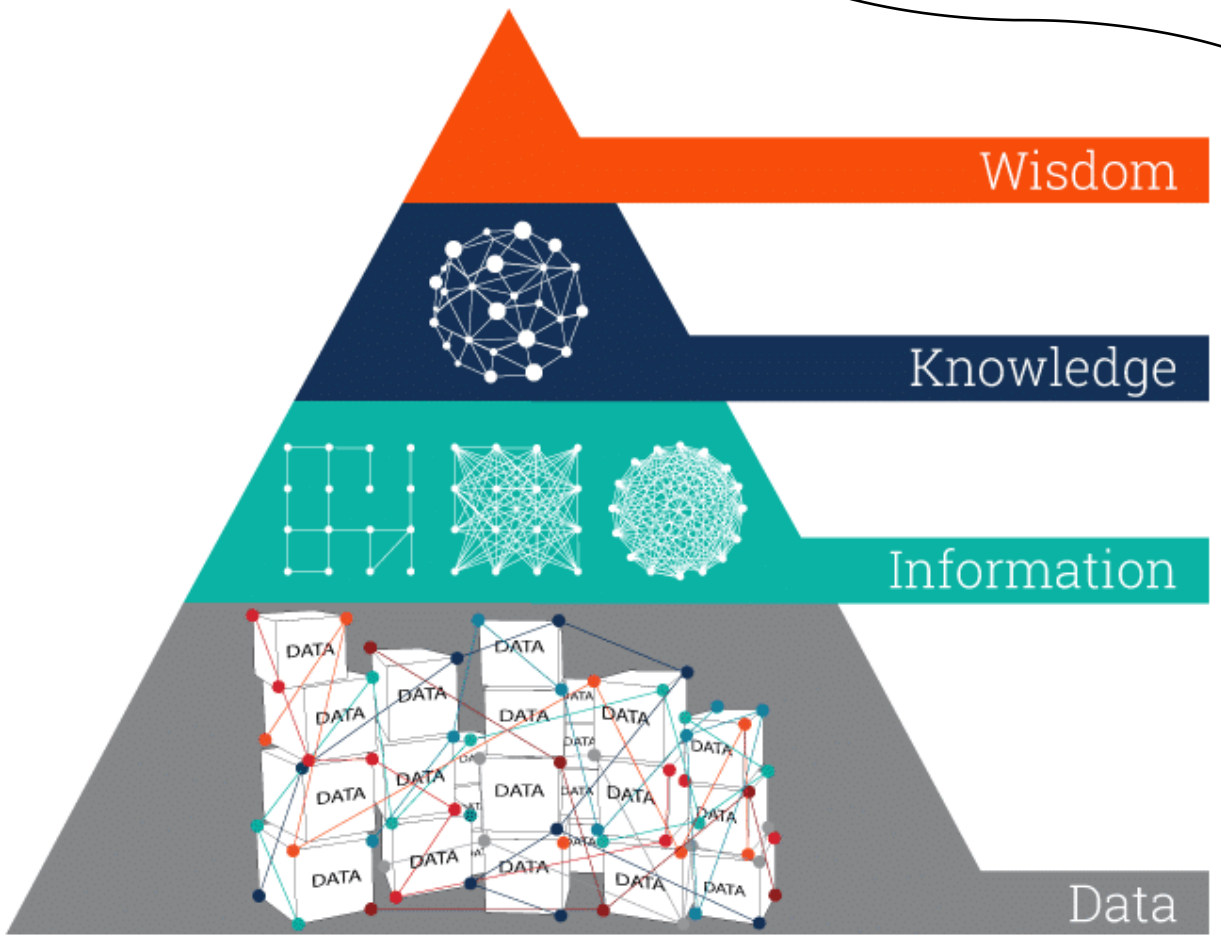
220502 19860222 130 0

六位数字地址码　八位数字出生日期码　三位数字顺序码　一位数字校验码

Knowledge (知识)



"Wisdom is not a product of schooling but of the lifelong attempt to acquire it.“--Albert Einstein



?

Rule

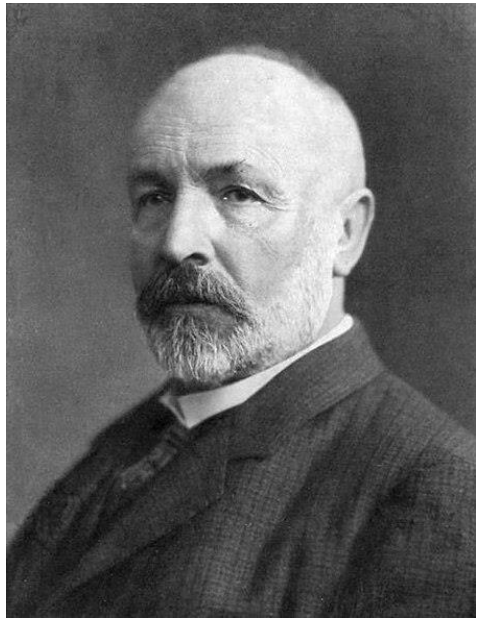
Structure

Set

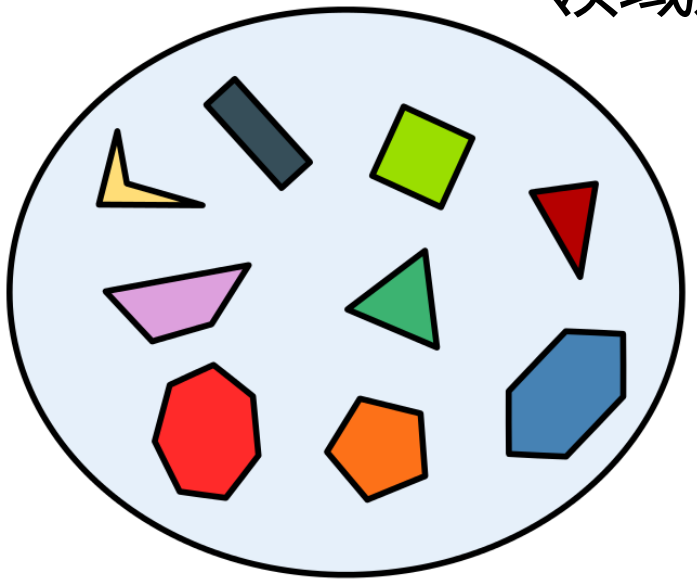
集合论(Set Theory)

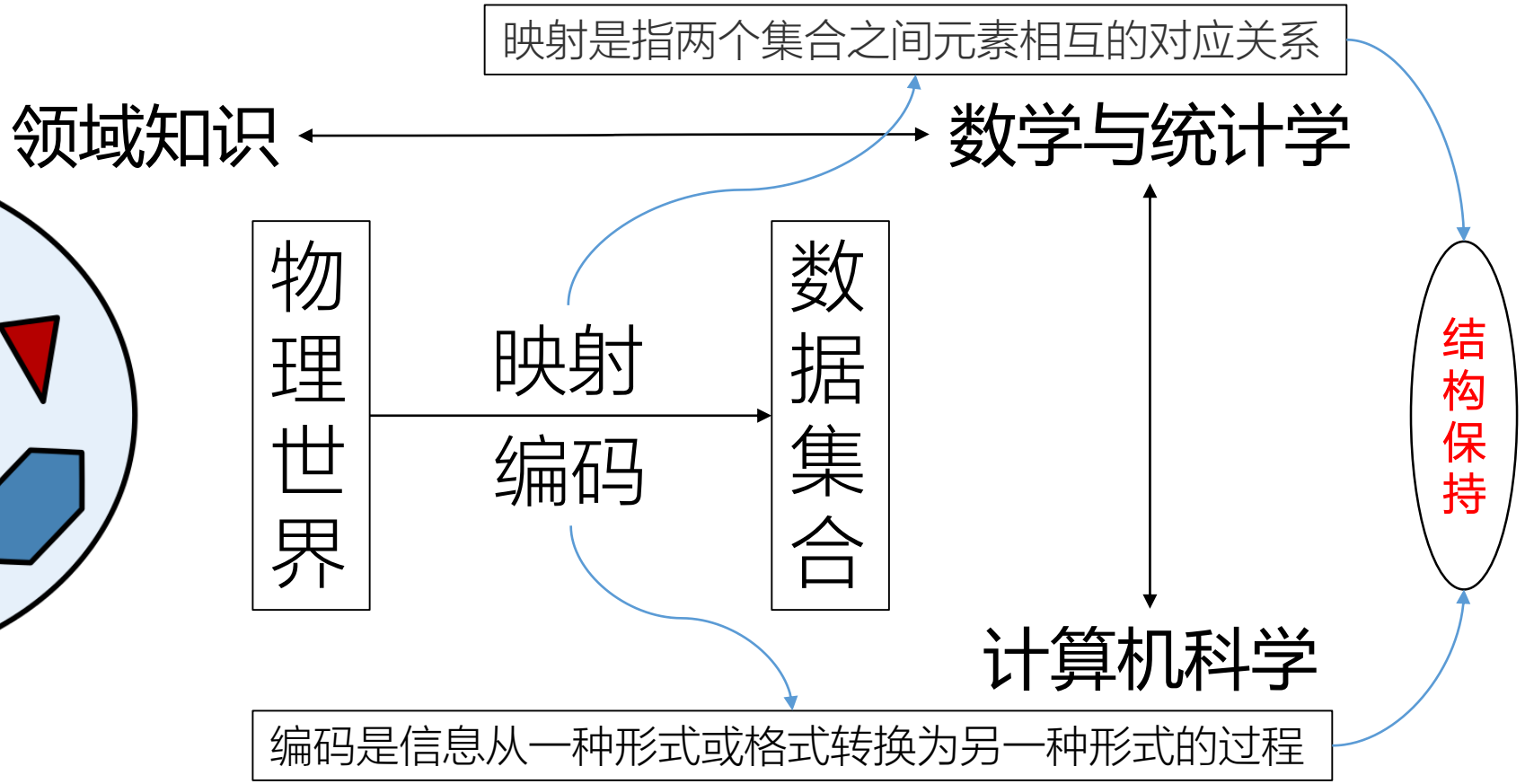
集合论是研究集合（由一堆抽象对象构成的整体）的数学理论，包含集合和元素以及关系等最基本数学概念。

集合论的研究是在十九世纪由俄国数学家康托尔及德国数学家理察·戴德金的朴素集合论开始。集合论被视为现代数学的基础。



数据科学中的集合




映射是指两个集合之间元素相互的对应关系
领域知识
数学与统计学
物理世界
映射
编码
数据集合
结构保持
计算机科学
编码是信息从一种形式或格式转换为另一种形式的过程

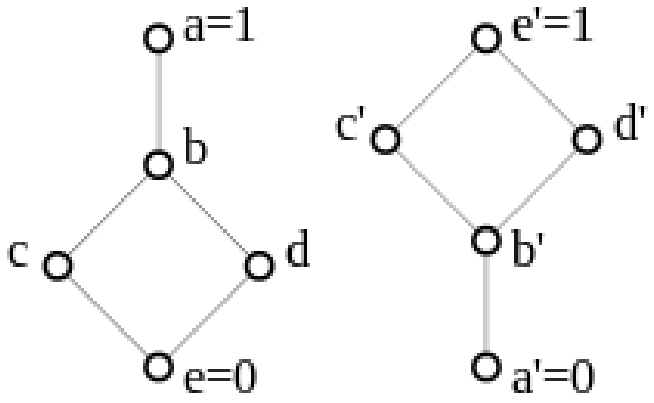
数据科学中的集合

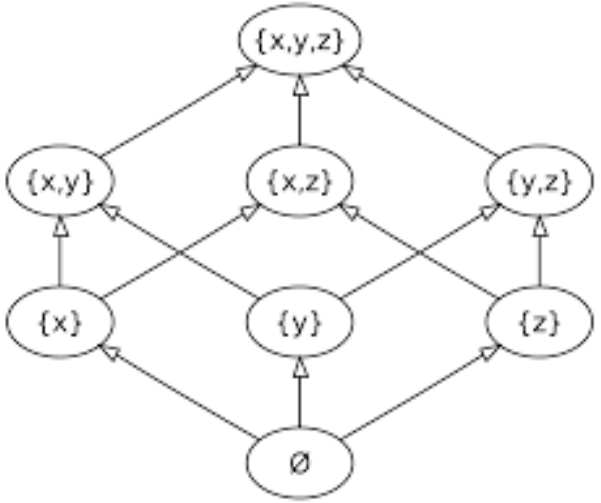
映射
编码

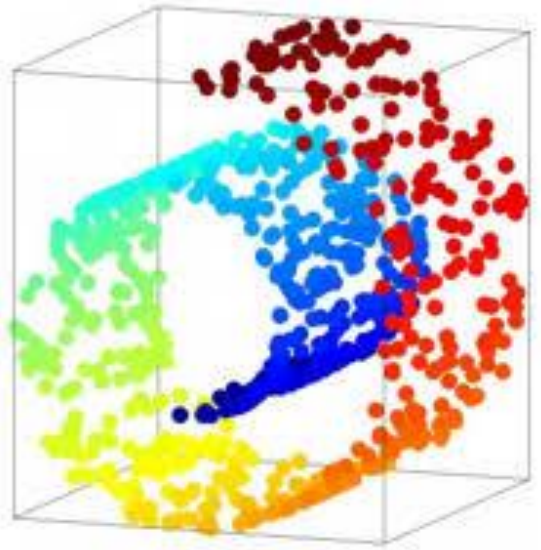
北京市：110000
天津市：120000
河北省：130000
山西省：140000
内蒙古：150000
辽宁省：210000
吉林省：220000
黑龙江：230000
上海市：310000
江苏省：320000
浙江省：330000
……

数学结构 (Mathematical Structure)

在数学中，一个集合上的结构是由附加在该集合上的某种操作和含义。 常见的数学结构包括序结构、代数结构、拓扑结构等等。







序结构

集合S及其关系≤满足以下要求：

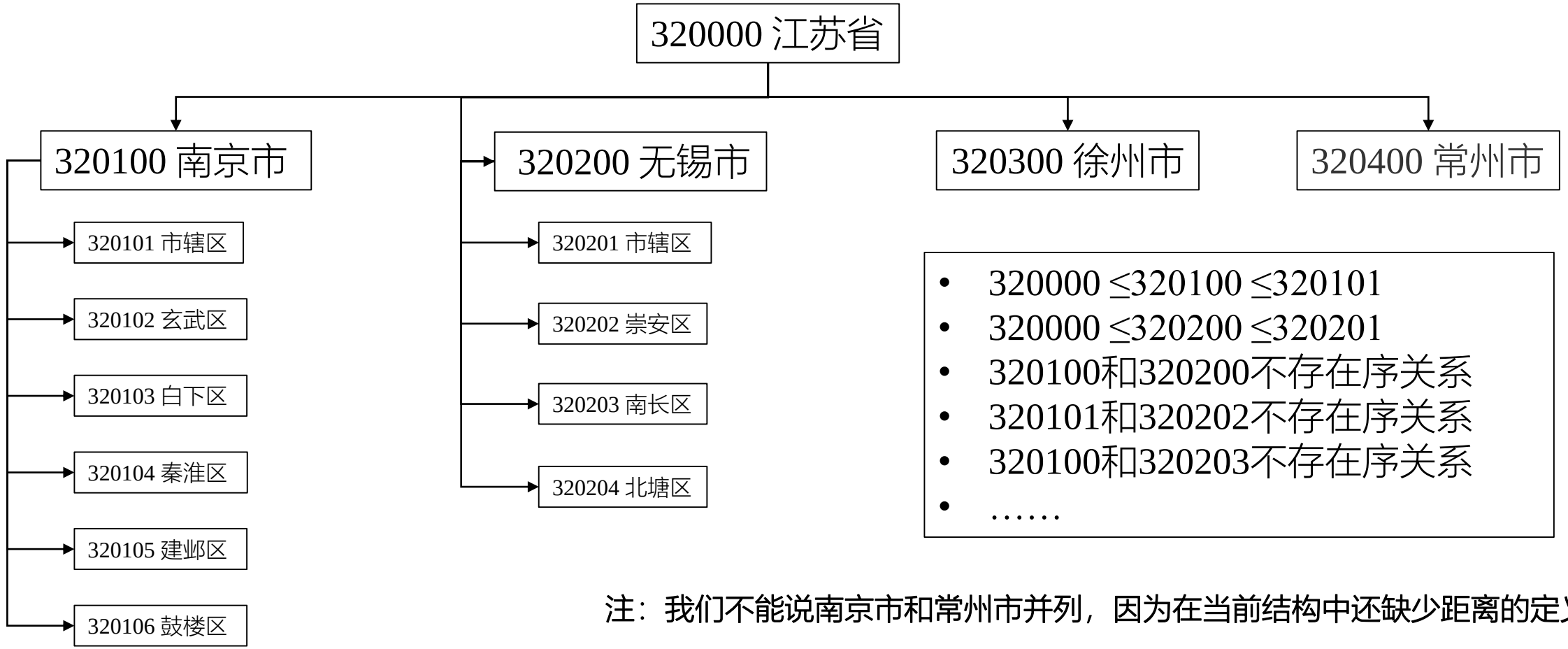
- $a \leq a$ (自反)
- $a \leq b$ 且 $b \leq a$，则有 $a = b$ (反对称)
- $a \leq b$ 且 $b \leq c$，则有 $a \leq c$ (传递)

则≤称为在S的偏序关系，S称之为偏序集。

偏序集的任意两个元素a和b均有a≤b或b≤a，则≤为全序关系。

偏序集的任意非空子集都有最小元素，则≤为良序关系。

通过序结构从数据中获取信息



- 320000 ≤320100 ≤320101
- 320000 ≤320200 ≤320201
- 320100和320200不存在序关系
- 320101和320202不存在序关系
- 320100和320203不存在序关系
- ……

注：我们不能说南京市和常州市并列，因为在当前结构中还缺少距离的定义。

代数结构

群是一种只有一个运算的简单代数结构。

设G为一个非空集合，a、b、c为它的任意元素，如果对G所定义的一种代数运算“$\cdot$”满足：

- 封闭性：$a \cdot b \in G$;
- 结合律：$(a \cdot b)c = a \cdot (b \cdot c)$;
- 唯一性：$\forall a,b \in G \exists$唯一的$x$和$y$：$a \cdot x = b$和$y \cdot a = b$

则称G对于所定义的运算“$\cdot$”构成一个群。

布尔代数

布尔代数是集合A的结构：

- 二元运算∧： A×A→A
- 二元运算∨： A×A→A
- 一元运算 ¬： A→A
- 常数0和1

$a \vee (b \vee c) = (a \vee b) \vee c$	$a \wedge (b \wedge c) = (a \wedge b) \wedge c$
$a \vee b = b \vee a$	$a \wedge b = b \wedge a$
$a \vee (a \wedge b) = a$	$a \wedge (a \vee b) = a$
$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$	$a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$
$a \vee \neg a = 1$	$a \wedge \neg a = 0$

关系代数

关系代数 (L, ∧, ∨, −, 0, 1, •, I, ˘) 是在布尔代数的一个扩展

B1: $A \vee B = B \vee A$

B2: $A \vee (B \vee C) = (A \vee B) \vee C$

B3: $(A^- \vee B)^- \vee (A^- \vee B^-)^- = A$

B4: $A \bullet (B \bullet C) = (A \bullet B) \bullet C$

B5: $A \bullet I = A$

B6: $A^{\breve{}\breve{}} = A$

B7: $(A \bullet B)^{\breve{}} = B^{\breve{}} \bullet A^{\breve{}}$

B8: $(A \vee B)^{\breve{}} = A^{\breve{}} \vee B^{\breve{}}$

B9: $(A \vee B) \bullet C = (A \bullet C) \vee (B \bullet C)$

B10: $(A^{\breve{}} \bullet (A \bullet B)^-)^- \vee B^- = B^-$

A. Tarski (1948) "Abstract: Representation Problems for Relation Algebras," Bulletin of the AMS 54: 80.

1970年E.F. Codd发表数据的关系模型，针对数据进行了关系代数运算的定义，包含了集合运算、投影运算、选择运算、连接预算和重命名。

更多详情，请咨询课程助教。

（推理）规则

规则是建立信息与知识之间的语法关系。推理规则的应用纯粹是语法过程，但某种形式的语义与推理规则有关和推理规则自身的断言是必需的。

数理逻辑的推理规则

$$\begin{array}{l} p \to q \\ p \\ \hline \therefore \quad q \end{array}$$

$$((p \to q) \wedge p) \to q$$

数据-信息-知识

- 对于220502198602221300数据，根据身份证号的分段结构获得了出生地区220502，出生年月19860222等信息。
- 对于出生地区信息220502，根据地区信息的序结构和省市区的推理规则，我们获得了吉林省通化市东昌区的知识，根据地图的结构信息和东西南北的规则，我们获得了吉林省东南部的知识。
- 对于出生年月信息19860222，根据年龄计算规则我们知道他今年33岁，根据出生年月信息的全序结构和年老年少规则，我知道他比我年轻的知识。

开放讨论

陕西省高级人民法院 民事裁定书 （2015）陕民二终字第00122号

上诉人（一审被告、反诉原告）：西安陕鼓动力股份有限公司。住所地：西安市高新区沣惠南路8号。 法定代表人：印建安，该公司董事长。

委托代理人：赵文娟，陕西大秦律师事务所律师。 委托代理人：李成于，陕西大秦律师事务所律师。

被上诉人（一审原告、反诉被告）：贵州盘县紫森源（集团）实业发展投资有限公司盘县仲恒煤矿。住所地：贵州省六盘水市盘县红果镇中沙陀村。 负责人：范斌，该煤矿经理。

委托代理人：米强，北京大成（昆明）律师事务所律师。 委托代理人：程佳庆，北京大成（昆明）律师事务所律师。

上诉人西安陕鼓动力股份有限公司因与被上诉人贵州盘县紫森源（集团）实业发展投资有限公司盘县仲恒煤矿买卖合同纠纷一案，不服西安市中级人民法院（2015）西中民三初字第00069号民事判决，向本院提出上诉。本院在审理过程中，上诉人西安陕鼓动力股份有限公司因考虑到与贵州盘县紫森源（集团）实业发展投资有限公司的后期友好合作关系于2015年10月30日申请撤回上诉。 本院认为，西安陕鼓动力股份有限公司撤回上诉的请求系对该公司民事权利的处置，不违反法律的规定。依照《中华人民共和国民事诉讼法》第一百七十三条的规定，裁定如下： 准许西安陕鼓动力股份有限公司撤回上诉，双方均按原审判决执行。 案件受理费33076元减半收取16538元，由西安陕鼓动力股份有限公司承担。 本裁定为终审裁定。

审判长赵小平

审判员倪健

代理审判员杨晓梅

二〇一五年十一月四日

书记员杨龙龙

更多详情，请咨询课程助教。

开放讨论

- 对于裁判文书数据，讨论能够赋予什么样的结构，获取哪些信息？
- 对于赋予结构获得的信息，结合已有领域知识，我们可以通过哪些规则，获得进一步的知识？

数据来源：中国裁判文书网 http://wenshu.court.gov.cn