大数据分析实践

Data Quality /Data Reduction

Qiong Zeng (曾琼)

qiong.zn@sdu.edu.cn

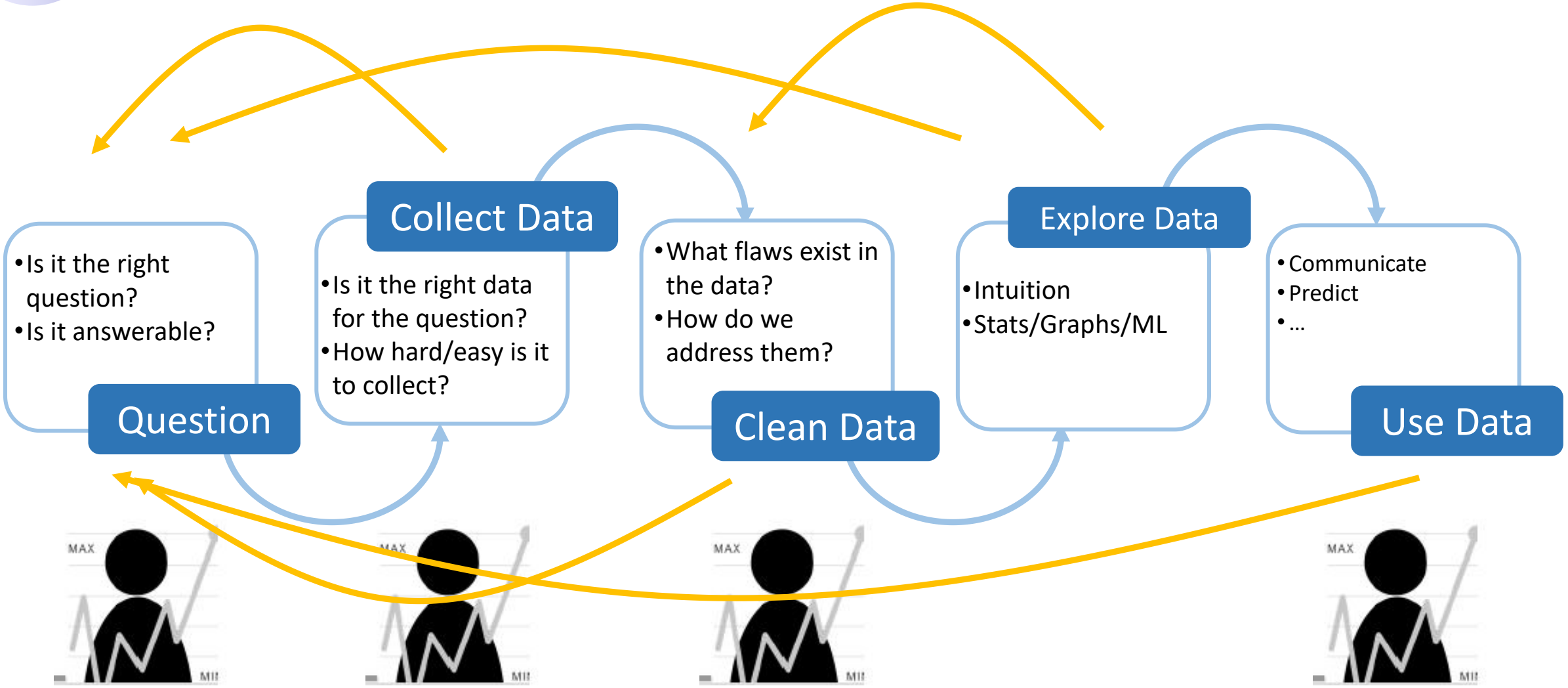
Research *is creative and systematic work undertaken to increase the stock of knowledge.*

Course Outline

6 Personal assignments

6 Group assignments

上课日期	授课内容	实验内容	周次
20250903	课程入门、大数据探索式分析	课程实践项目介绍、项目组队测试	第一周
20250910	项目经验谈、科研实践入门	项目管理工具制定项目计划	第二周
20250917	数据采样与降维	**数据采样实践**	第三周
20250924	**数据质量管理**	数据质量实践	**第四周**
20251001	/	/	第五周
20251008	/	/	第六周
20251015	众包与电子表格	电子表格实践	第七周
20251022	中期汇报（论文+项目进展）1	中期进展报告	第八周
20251029	中期汇报（论文+项目进展）2	BERT实践环境配置	第九周
20251105	统计分析/机器学习方法与工具	BERT实践	第十周
20251112	可视化设计/人机交互方法与工具	Canis/Cast/Libra实践	第十一周
20251119	普适计算（曹烨彤）	手机移动数据采集与分析	第十二周
20251126	大规模数据分析系统（滕德军）	SPARK实践	第十三周
20251202	如何撰写项目论文	大项目收尾	第十四周
20251209	项目结题报告1		第十五周
20251216	项目结题报告2	大项目验收	第十六周


Question
•Is it the right question?
•Is it answerable?
Collect Data
•Is it the right data for the question?
•How hard/easy is it to collect?
Clean Data
•What flaws exist in the data?
•How do we address them?
Explore Data
•Intuition
•Stats/Graphs/ML
Use Data
•Communicate
•Predict
•…
MAX
MIN

Outline

Data Quality

Data Reduction

课程导入：

2008年土耳其石油管道爆炸，3万桶原油散落水中

https://www.aqniu.com/threat-alert/14565.html

系统控制
错误数据
注入

2015年乌克兰断电事件，22.5万居民经历长达数小时的大规模停电

https://www.secrss.com/articles/39853

- 钓鱼邮件获取凭证
- 网络资产探测获取SCADA系统控制能力
- 操纵断路[illegible]电

** 以擦除[illegible]行[illegible]坏组织断电后迅[illegible]中心发起dos攻击阻[illegible]传播

如何提高数据质量?

学习目标

知识目标

学生能够准确描述数据质量评价维度，可以理解数据清洗的常用方法

重点

能力目标

学生能运用数据质量维度定性评估数据质量，并掌握数据清洗方法

重点

难点

素质目标

学生能以科学严谨的态度审查数据，在数据分析过程中遵守职业规范

单选题 1分

设置

小明、小王和小李是好朋友，他们的身材信息如下：

小明（160cm，60kg）

小李（1.7m，61000g）

小王（160cm，59000g）

请问小明跟小王的体型更相似，还是跟小李的更相似？

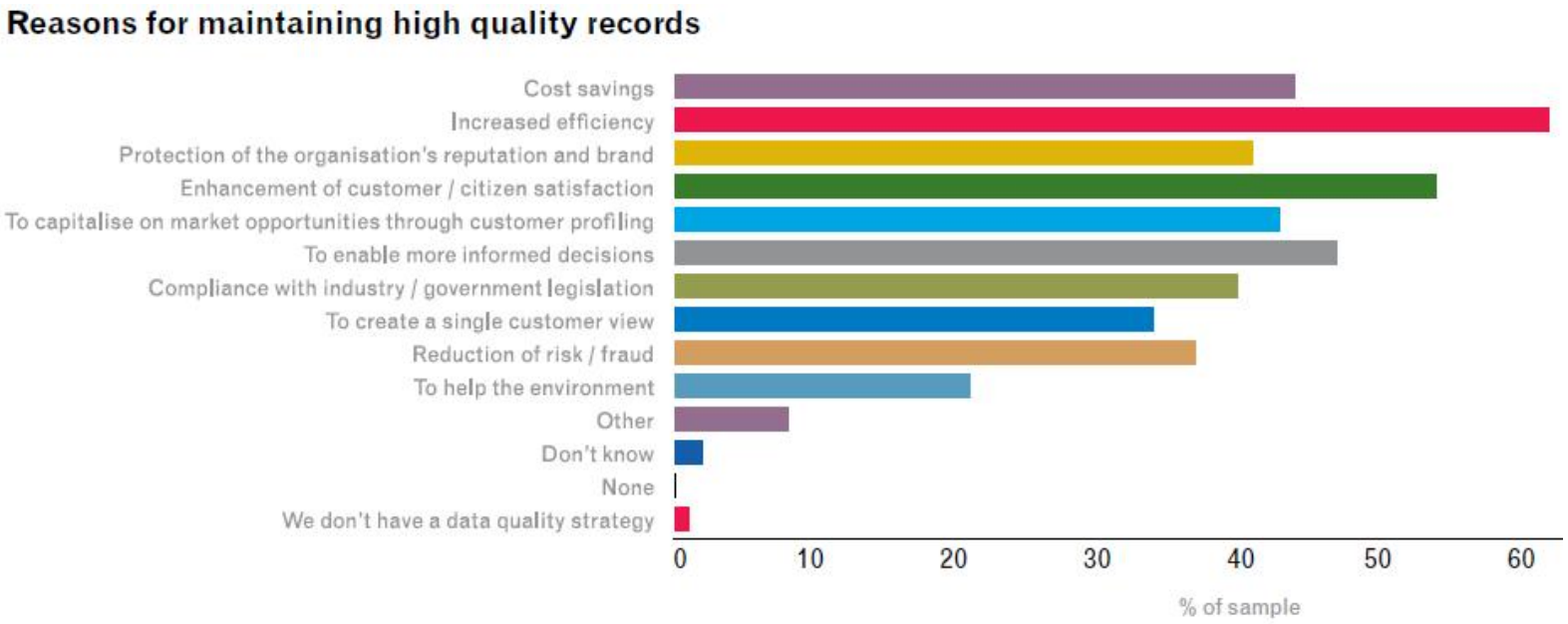
A 小李

B 小王

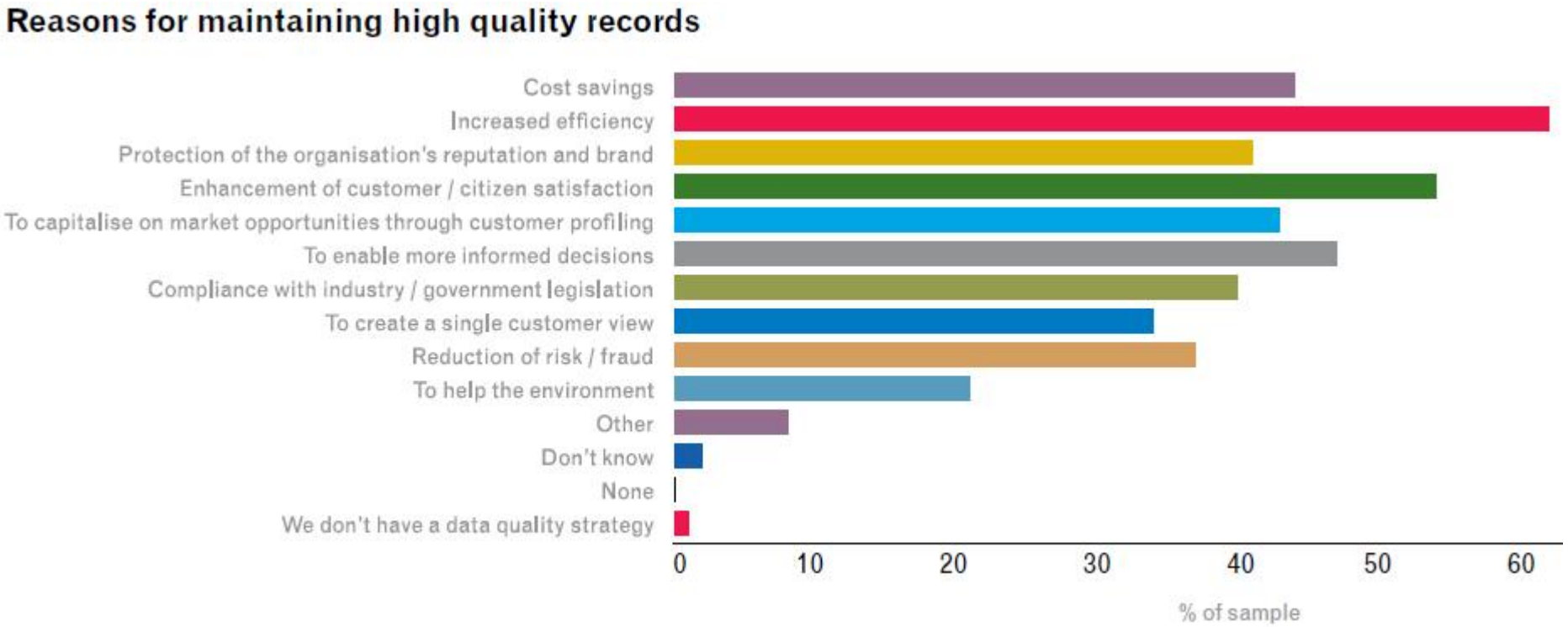
提交

主要内容：数据质量管理

数据质量管理：是指对数据从计划、获取、存储、共享、维护、应用的生命周期每个阶段可能引发的各类数据质量问题，进行识别、度量、监控、预警的一系列管理过程，从而改善和提高数据质量



Data Quality



New research from Experian Data Quality shows that inaccurate data has a direct impact on the bottom line of 88% of companies, with the average company losing 12% of its revenue.

Data Quality

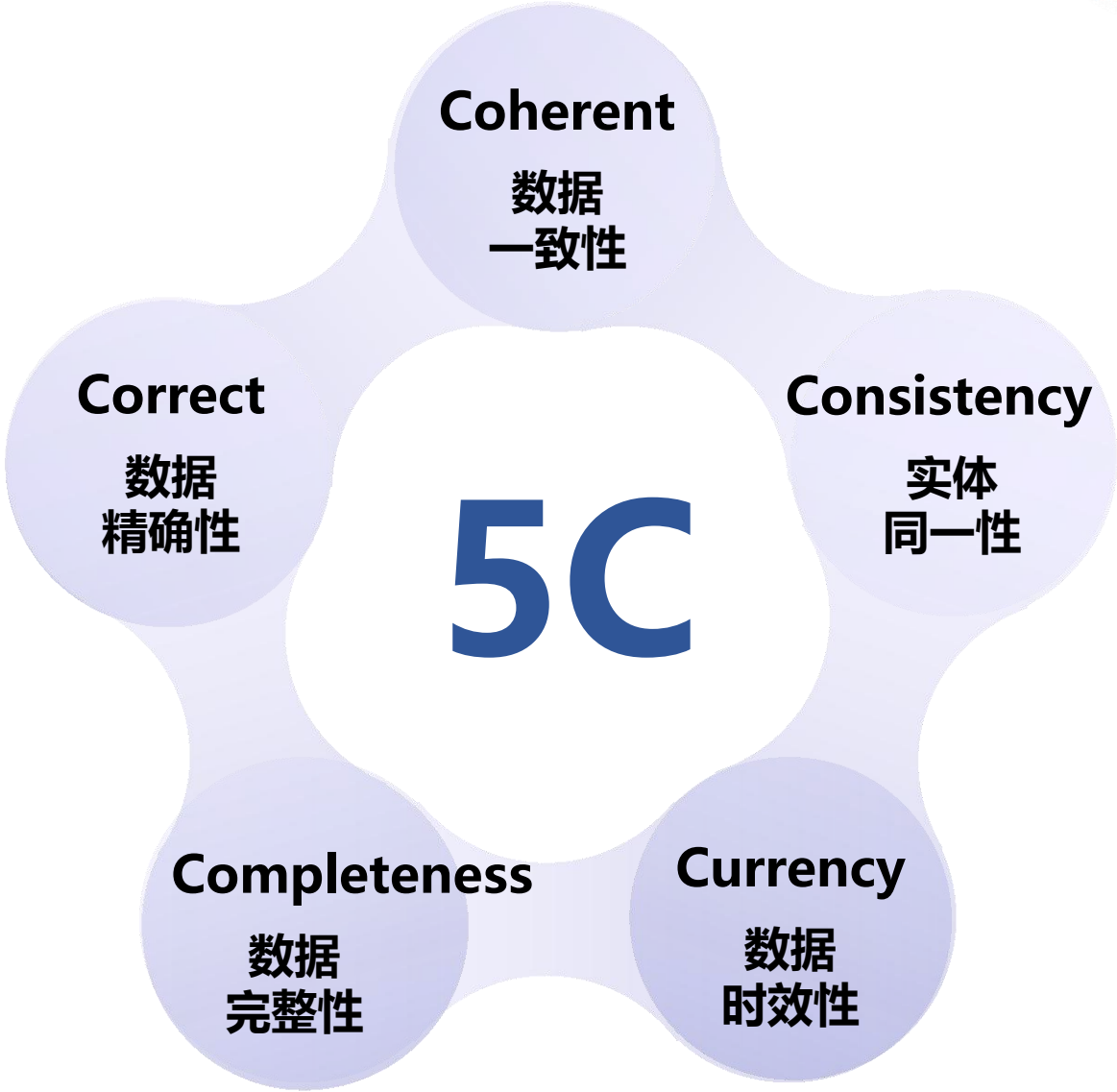
- **Poor data quality negatively affects many data processing efforts**
- **Data mining example: a classification model for detecting people who are loan risks is built using poor data**
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality

Data Quality Management can be defined as a set of practices undertaken by a data manager or a data organization to maintain high quality information.

Data Quality

如何评估数据质量?
数据质量评估的
五个C维度



Data Quality

- Coherent: without semantic errors or contradictory data between attributes of an object **对象属性无语义错误或矛盾**
- Correct: the extent to which data correctly portrays reality **数据反映真实情况**
- Completeness: without missing (null) values in table fields **表字段无缺失值**
- Currency: the degree to which data is up-to-date **数据保持最新状态**
- Consistency: consistent data values for an entity between different tables
  同一实体在不同表之间的一致性

主观题 10分

设置

下面数据是银行用户的贷款信息，用于审核是否为用户继续发放贷款。

请以**小组讨论**一下存在的评估数据的质量情况，并说明原因

ID	是否按时偿还	婚姻状况	税后收入(万)	是否有失信记录	户籍城市	区号
1	是	未婚	12.5	否	上海	21
2	否	已婚	10.0	否	北京	10
3	否	未婚	7.0	否	重庆	23
4	是	已婚	12.0	否	天津	22
5	否	离异	1000.0	是	天津	22
6	否	无	6.0	否	北京	1
7	是	离异	22.0	无	重庆	23
8	否	未婚	8.5	是	上海	10
9	否	已婚	9.0	否	天津	22
9	否	未婚	9.0	否	北京	10

作答

主要内容：数据质量评估维度

分享与小结

一致性　准确性　完整性　时效性　同一性

ID	是否按时偿还	婚姻状况	税后收入(万)	是否有失信记录	户籍城市	区号
1	是	未婚	12.5	否	上海	21
2	否	已婚	10.0	否	北京	10
3	否				重庆	23
4	是				天津	22
5	否				天津	22
6	否	无	6.0	否	北京	1
7	是	离异	22.0	无	重庆	23
8	否	未婚	8.5	是	上海	10
9	否	已婚	9.0	否	天津	22
9	否	未婚	9.0	否	北京	10

问题：如何清洗脏数据?

主要内容：数据清洗

常见的数据质量问题

缺失值　　噪声值　　重复值

ID	是否按时偿还	婚姻状况	税后收入(万)	是否有失信记录	户籍城市	区号
1	是	未婚	12.5	否	上海	21
2	否	已婚	10.0	否	北京	10
3	否	未婚	7.0	否	重庆	23
4	是	已婚	12.0	否	天津	22
5	否	离异	1000.0	是	天津	22
6	否	无	6.0	否	北京	1
7	是	离异	22.0	无	重庆	23
8	否	未婚	8.5	是	上海	10
9	否	已婚	9.0	否	天津	22
9	否	未婚	9.0	否	北京	10

主要内容：数据清洗-缺失值填充

缺失值 **针对不完整数据，主要的清洗方法是缺失值填充，常见方法：**

删除	若一个样本大部分都缺失，可以选择放弃该样本
统计填充	对于数值类型的属性，可根据该维度的统计值进行填充，如平均数、中位数、众数、最大值、最小值等
统一填充	用自定义值统一填充，如“空”、“正无穷”等
预测填充	利用预测模型把数据填充后在做进一步工作，选择的机器学习方法依赖于数据类型以及数据分布

主观题 10分

设置

某公司有一份数据记录了客户是否最终愿意购买他们的产品，年收入上有缺失值，请问可以如何处理?

年收入	性别	年龄	婚姻状况	是否购买
125,000	女	24	未婚	是
385,000	男	35	已婚	是
NULL	男	30	已婚	否
185,000	男	28	未婚	否
205,000	女	26	未婚	是
179,000	女	27	已婚	是
NULL	男	26	未婚	否
265,000	女	30	未婚	是
105,000	男	23	已婚	否
205,000	女	28	未婚	是

正常使用主观题需2.0以上版本雨课堂

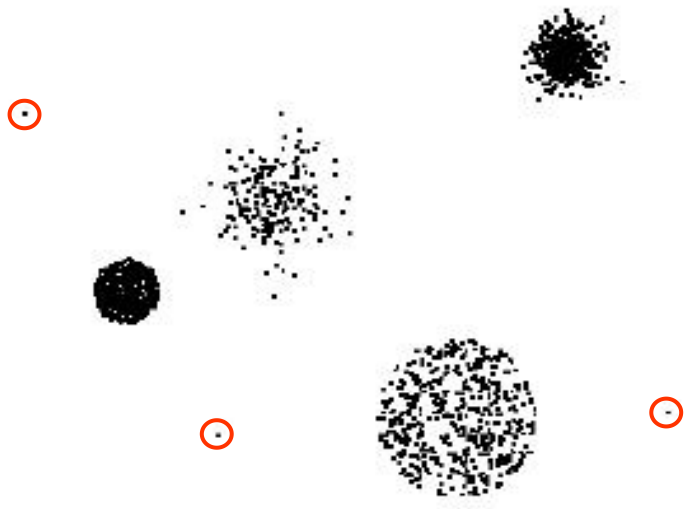
作答

主要内容：数据清洗-异常值处理

异常值 与其他数据在特征上有明显差异的数据值，也称为数据噪声

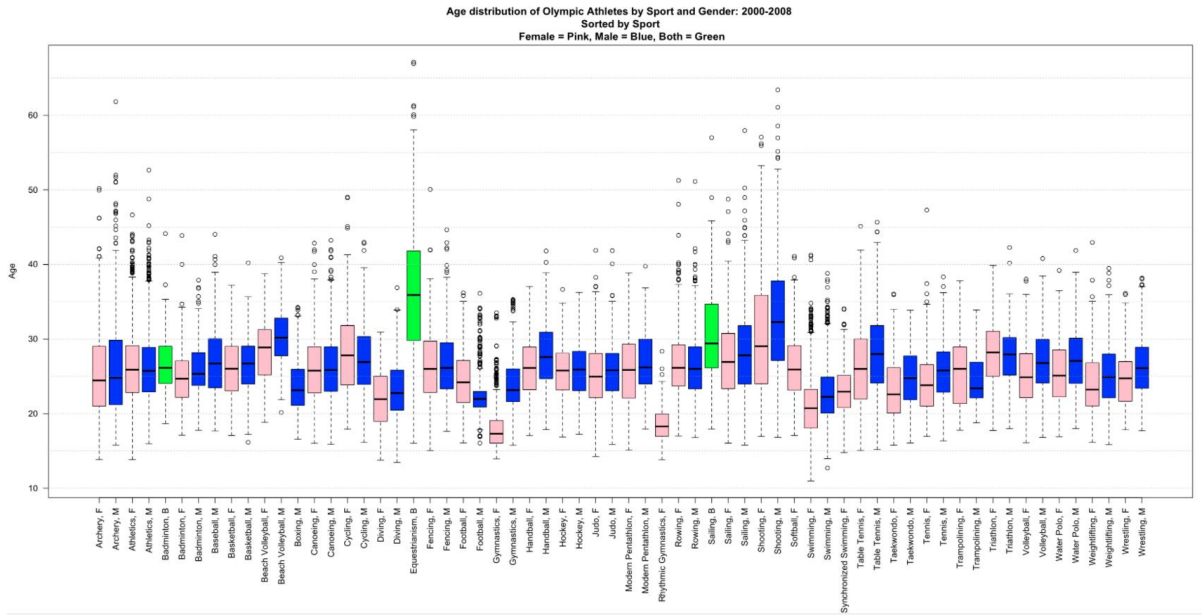
异常值为分析干扰因素

本身对于分析没有帮助，可直接删除或者用其他数值代替



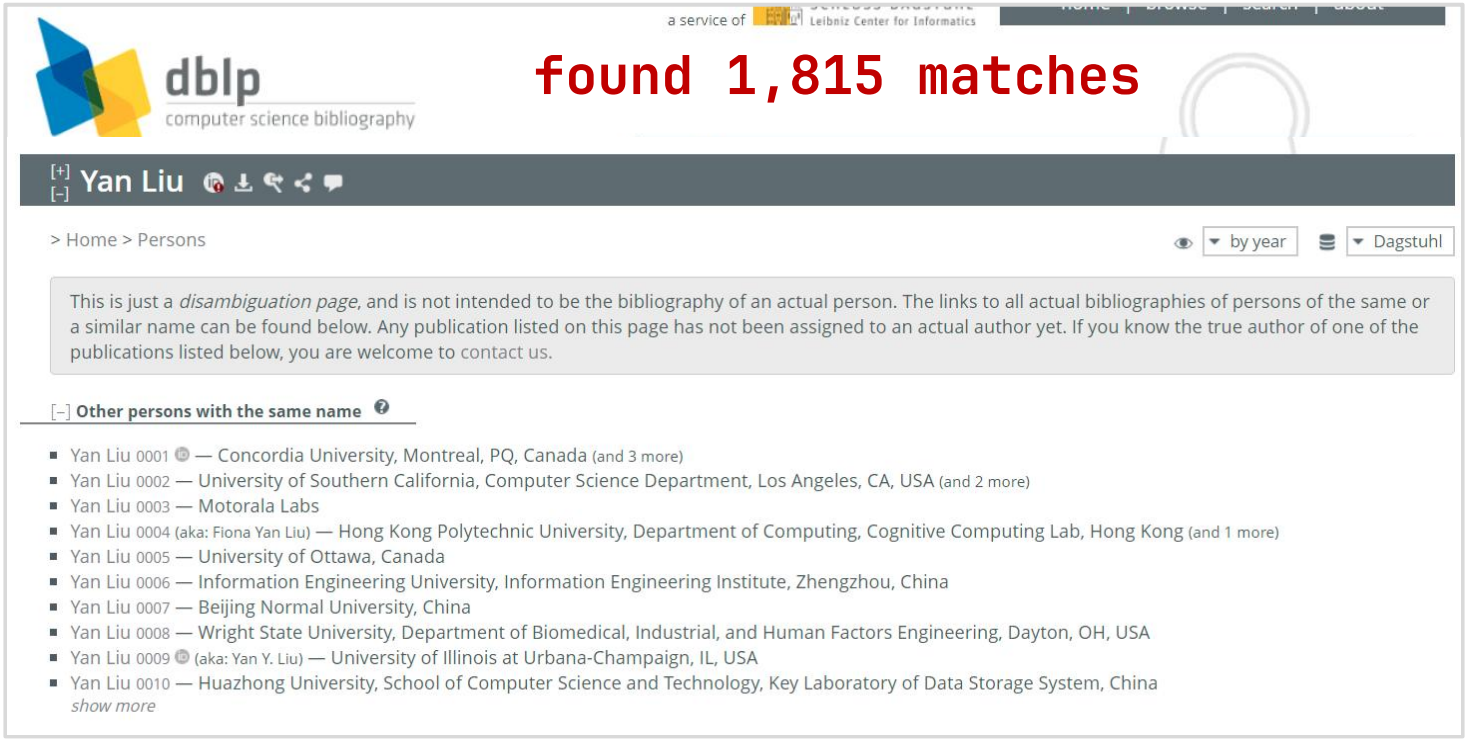
异常值为分析目标

比如金融诈骗分析、错误数据检测



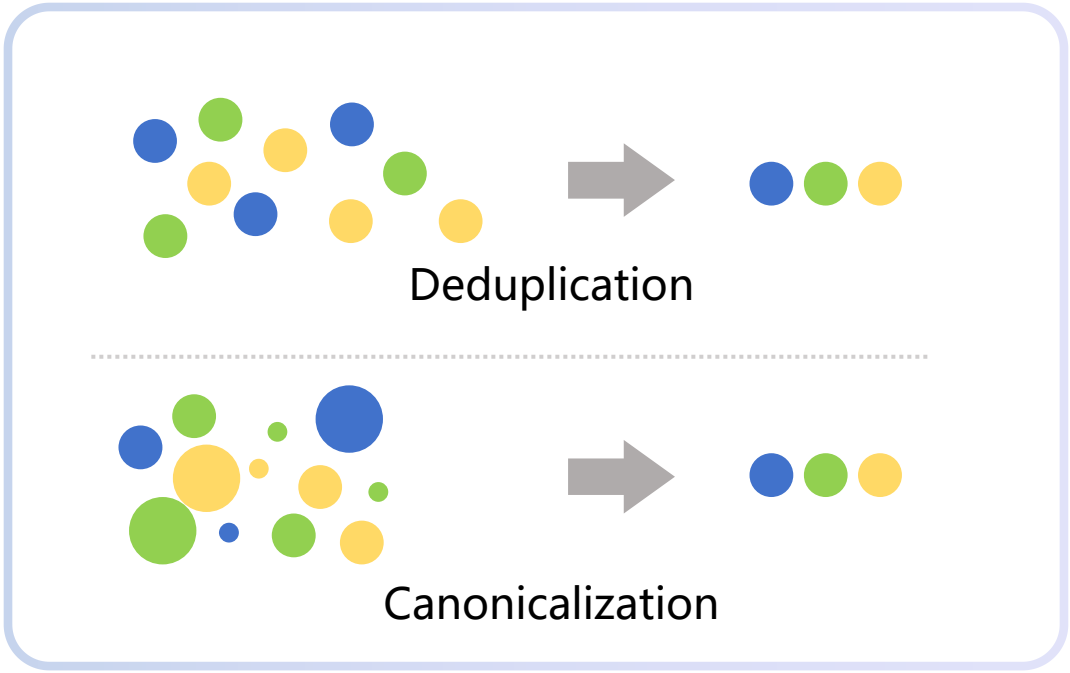
主要内容：数据清洗-重复值处理

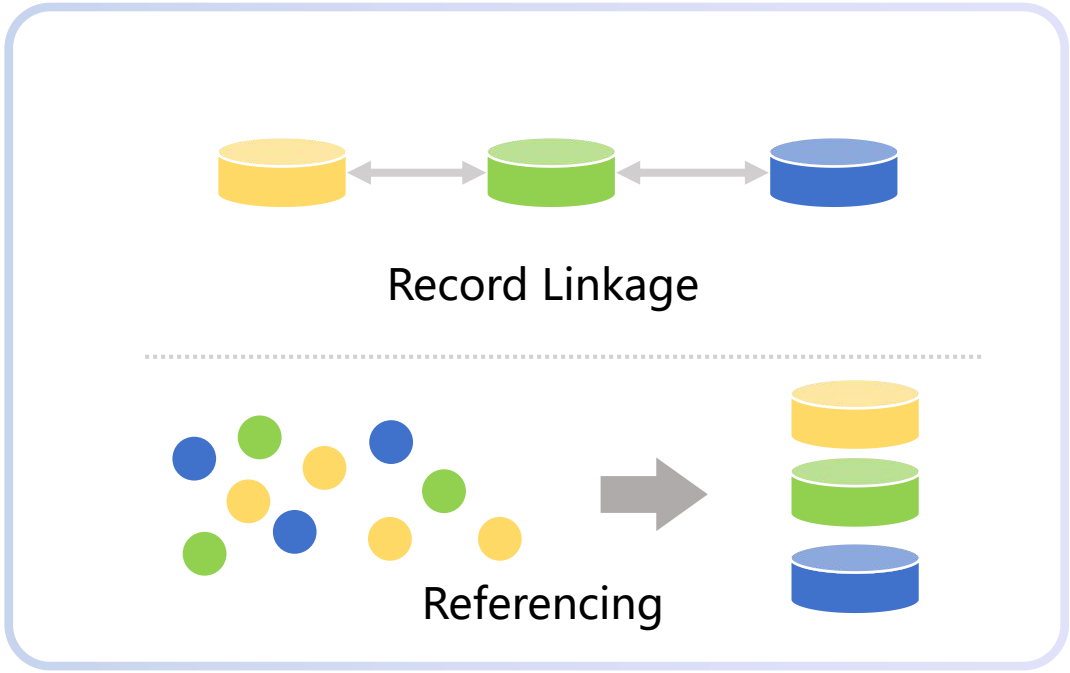
- 重复值可能出现在数据对象层面，也可能出现在属性值层面
- 当重复实体出现在不同数据集中时，将带来大量的干扰信息；还有可能存在重复值有重复数值的情况
- 信息检索中常出现重复值



主要内容：数据清洗-实体识别

实体识别指在**给定的实体对象集合**中发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在现实世界中指代是同一个实体。





【拓展阅读】https://medium.com/@kskaran94/unsupervised-entity-resolution-using-graphs-e12f5abae0ae

主要内容：数据清洗-实体识别

基于相似性函数的实体识别

Id	Name	Coauthors	Title	class
o11	Wei Wang	Zhang	inferring…	e1
o12	Wei Wang	Duncan, Kum, Pei	social...	e1
o13	Wei Wang	Cheng, Li, Kum	measuring…	e1
o21	Wei Wang	Lin, Pei	threshold…	e2
o22	Wei Wang	Lin, Hua, Pei	ranking…	e2
o31	Wei Wang	Shi, Zhang	picturebook…	e3
o32	Wei Wang	Pei, Shi, Xu	utility…	e3

Jaccard Similarity (coefficient)
measures similarities between sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

主要内容：数据清洗-实体识别

基于相似性函数的实体识别

Id	Name	Coauthors	Title	class
o11	Wei Wang	Zhang	inferring…	e1
o12	Wei Wang	Duncan, Kum, Pei	social...	e1
o13	Wei Wang	Cheng, Li, Kum	measuring…	e1
o21	Wei Wang	Lin, Pei	threshold…	e2
o22	Wei Wang	Lin, Hua, Pei	ranking…	e2
o31	Wei Wang	Shi, Zhang	picturebook…	e3
o32	Wei Wang	Pei, Shi, Xu	utility…	e3

$Sim(o11, o12) = 0$
$Sim(o11, o31) = 0.5$

$Sim(o12, o13) = 0.2$
$Sim(o12, o21) = 0.25$

主要内容：数据清洗-实体识别

基于规则的实体识别

Id	Name	Coauthors	Title	class
o11	Wei Wang	Zhang	inferring…	e1
o12	Wei Wang	Duncan, Kum, Pei	social...	e1
o13	Wei Wang	Cheng, Li, Kum	measuring…	e1
o21	Wei Wang	Lin, Pei	threshold…	e2
o22	Wei Wang	Lin, Hua, Pei	ranking…	e2
o31	Wei Wang	Shi, Zhang	picturebook…	e3
o32	Wei Wang	Pei, Shi, Xu	utility…	e3

观察1：某些属性值对的存在对识别元组很有用

主要内容：数据清洗-实体识别

基于规则的实体识别

Id	Name	Coauthors	Title	class
o11	Wei Wang	Zhang	inferring…	e1
o12	Wei Wang	Duncan, Kum, Pei	social...	e1
o13	Wei Wang	Cheng, Li, Kum	measuring…	e1
o21	Wei Wang	Lin, Pei	threshold…	e2
o22	Wei Wang	Lin, Hua, Pei	ranking…	e2
o31	Wei Wang	Shi, Zhang	picturebook…	e3
o32	Wei Wang	Pei, Shi, Xu	utility…	e3

观察2：某些属性值的不存在也能够帮助识别元组

主要内容：数据清洗-实体识别

基于规则的实体识别

- R1: $\forall oi$, 如果oi[name] = “Wei Wang” and $oi[coauthors]$包含”kum”，那么oi指代实体e1
- R2: $\forall oi$, 如果oi[name] = “Wei Wang” and $oi[coauthors]$包含”lin”，那么oi指代实体e2
- R3: $\forall oi$, 如果oi[name] = “Wei Wang” and $oi[coauthors]$包含”shi”，那么oi指代实体e3
- R4: $\forall oi$, 如果oi[name] = “Wei Wang” and $oi[coauthors]$包含”zhang” and 不包含”Shi”，那么oi指代实体e1

主要内容：数据清洗-实体识别

基于规则的实体识别

- r1: $(name = \text{“wei wang”}) \wedge (\text{“}kum\text{”} \in coa) \Rightarrow e1$
- r2: $(name = \text{“wei wang”}) \wedge (\text{“}Lin\text{”} \in coa) \Rightarrow e2$
- r3: $(name = \text{“wei wang”}) \wedge (\text{“}shi\text{”} \in coa) \Rightarrow e3$
- r4: $(name = \text{“wei wang”}) \wedge (\text{“}zhang\text{”} \in coa) \wedge \neg(\text{“}shi\text{”} \in coa) \Rightarrow e1$

主要内容：数据清洗-实体识别

以基于规则的方法讲解实体识别

Input: U, R_E, θ_C
Output: $\mathbb{U}$

Initialize:
for each entity e_j in E **do**
    $U_j \leftarrow \emptyset$;
end for
for each o_i in U **do**
    $R(o_i) \leftarrow$FINDRULES(o_i);
    for each entity e_j in E **do**
        $R(e_j) \leftarrow \{r|\mathsf{RHS}(r) = e_j\}$;
        $C(o_i, e_j) \leftarrow$COMPCONF$(R(o_i) \cap R(e_j))$;
    end for
    SELENTITY(o_i, θ_C);
end for
Return $\mathbb{U} \leftarrow \{U_1, U_2, ..., U_m\}$;

【拓展阅读】Rule-based method for entity resolution.

Open Question

Points: 10

设置

下表为公司统计的培训表信息，请问其中的数据有哪些错误，该如何修复?

姓名	出生日期	联系电话	培训日期	培训内容
元芳	1995.8.29	13278654976	2017.1.5	Hadoop
李连	1994.4.26	13787654938	2016.12.27	Spark
周妍	1995.2.16	166772110291	Hadoop	2017.1.15
李胜	1996.2.30	17667518392	2036.12.27	Spark
赵忠	1993.11.11	17551629384	2017.1.13	HDFS

正常使用主观题需2.0以上版本雨课堂

作答

主观题 10分

设置

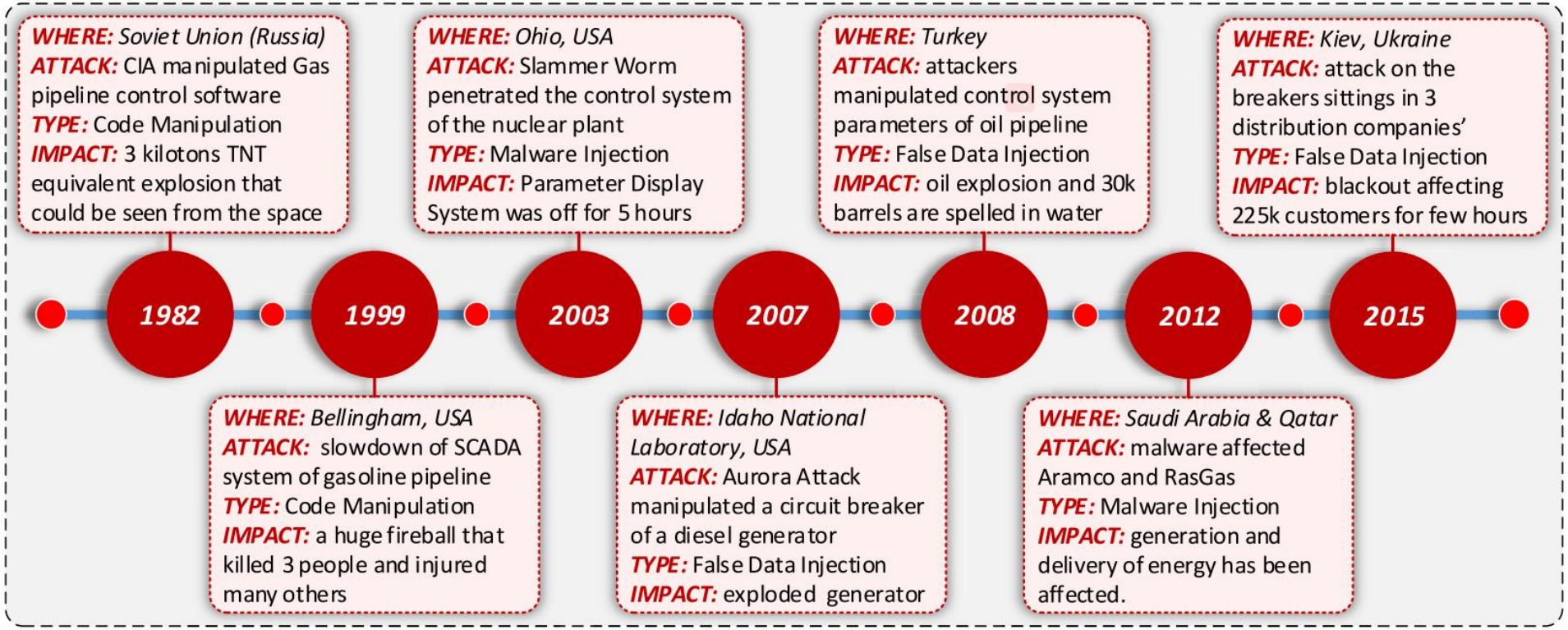
小组总结：

我们在实践过程中遇到了哪些数据质量问题？基本思路是什么？

作答

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen;

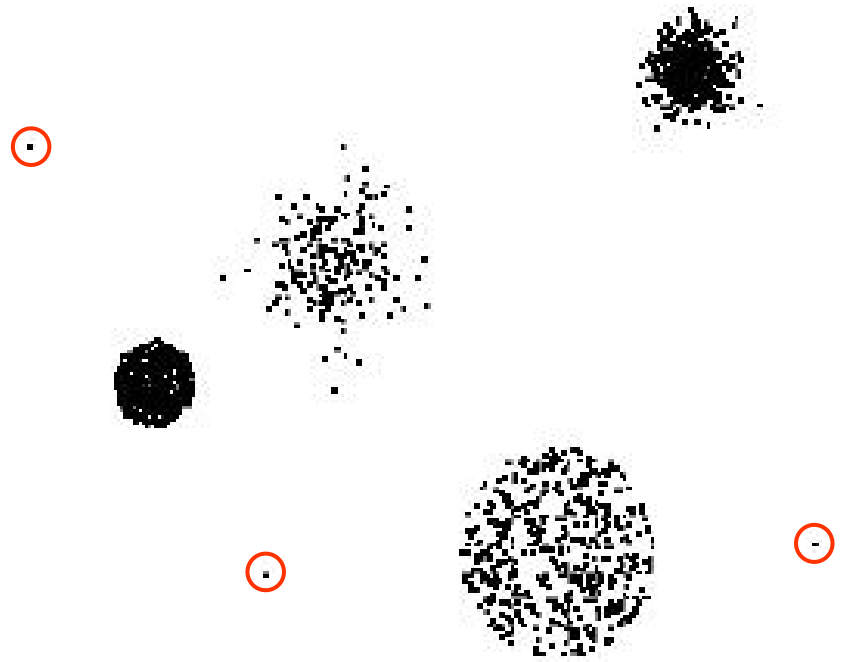


Noise: outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis

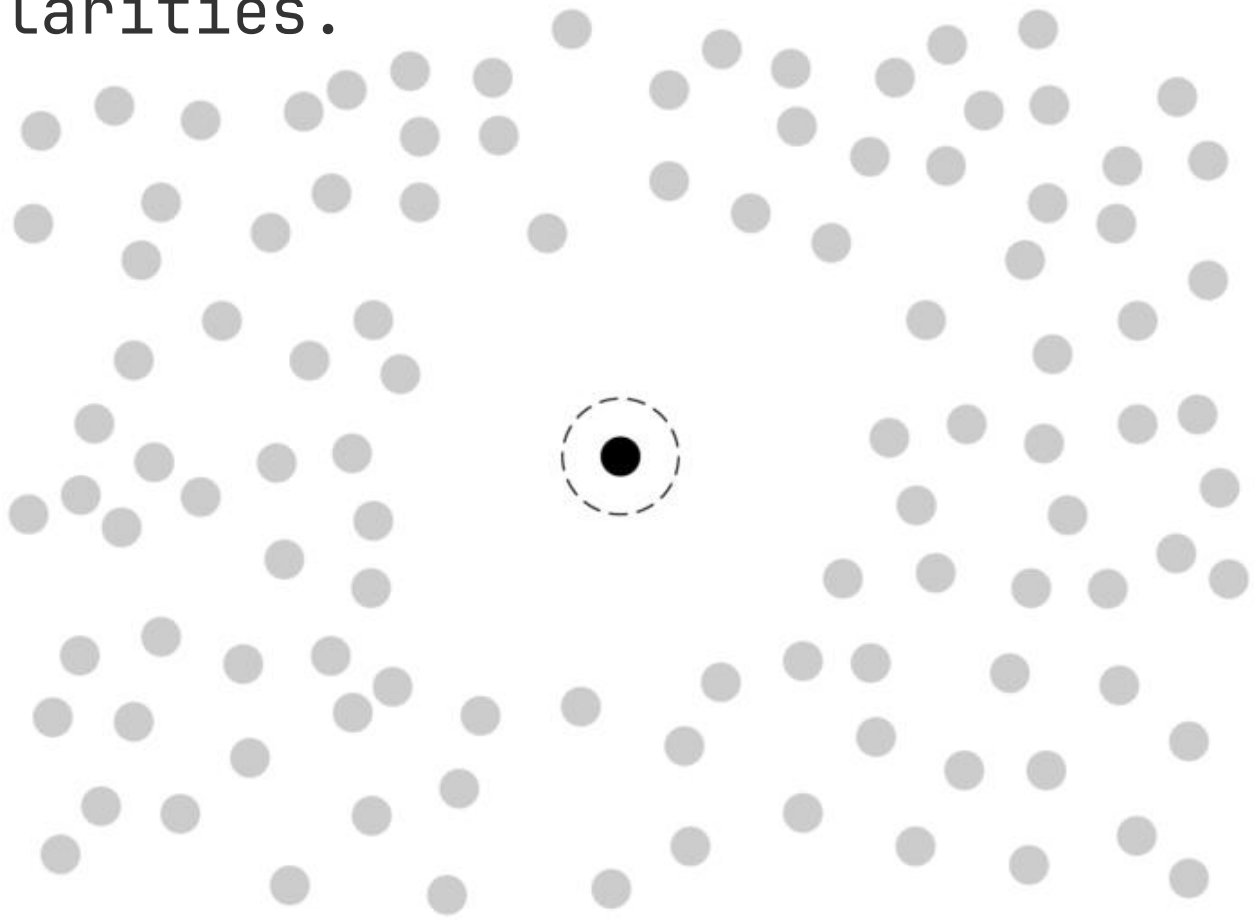
    drop them, replace with a less extreme value, replace them with a default

  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



Visualizing Outliers

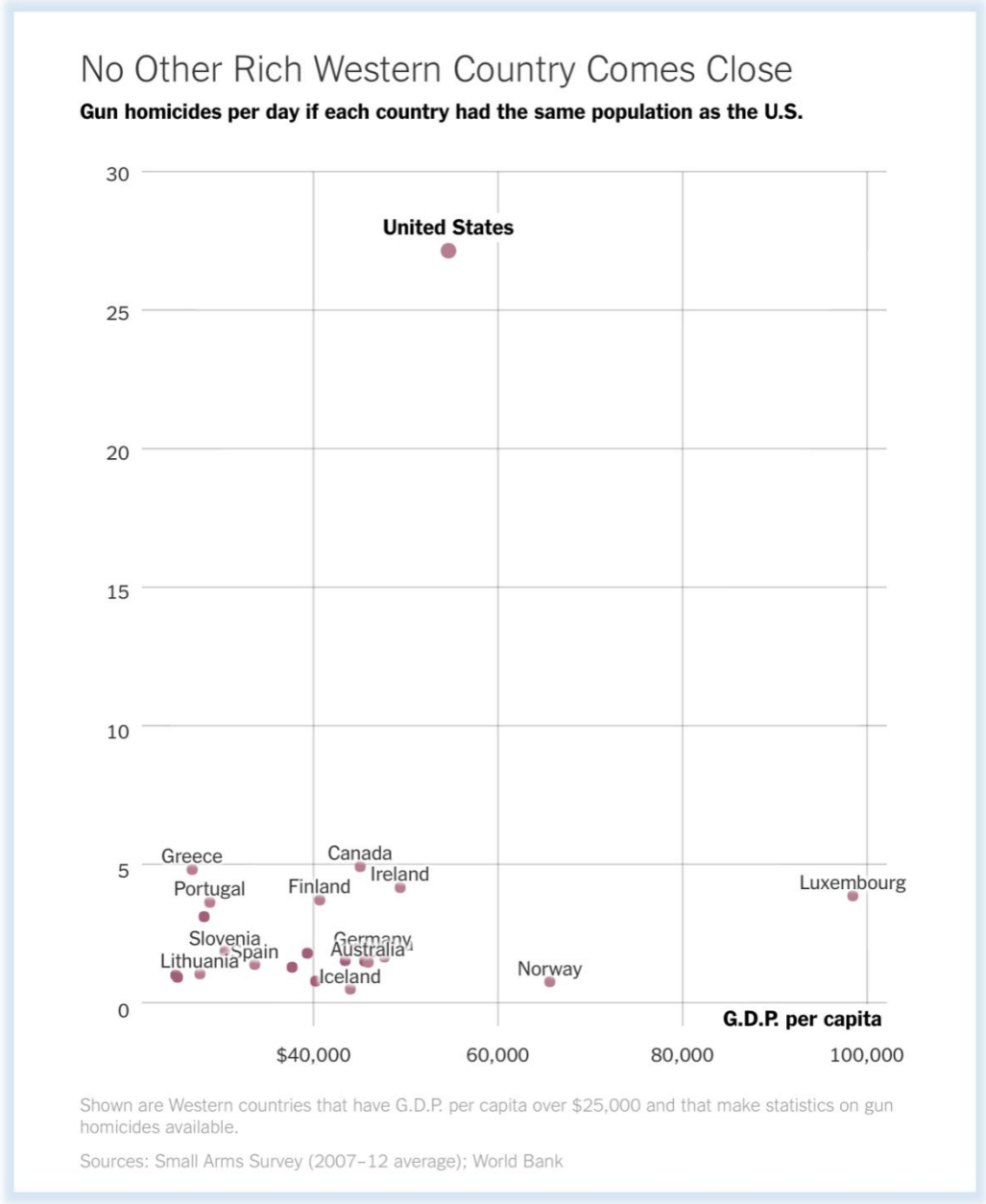
Point of Focus: focus on the outlier directly and show how it stands out from the rest. Visually, differences outweigh similarities.



https://flowingdata.com/2018/03/07/visualizing-outliers/

Visualizing Outliers

Example:
The Upshot highlighted gun homicides per day in the United States, as compared to other Western democracies:

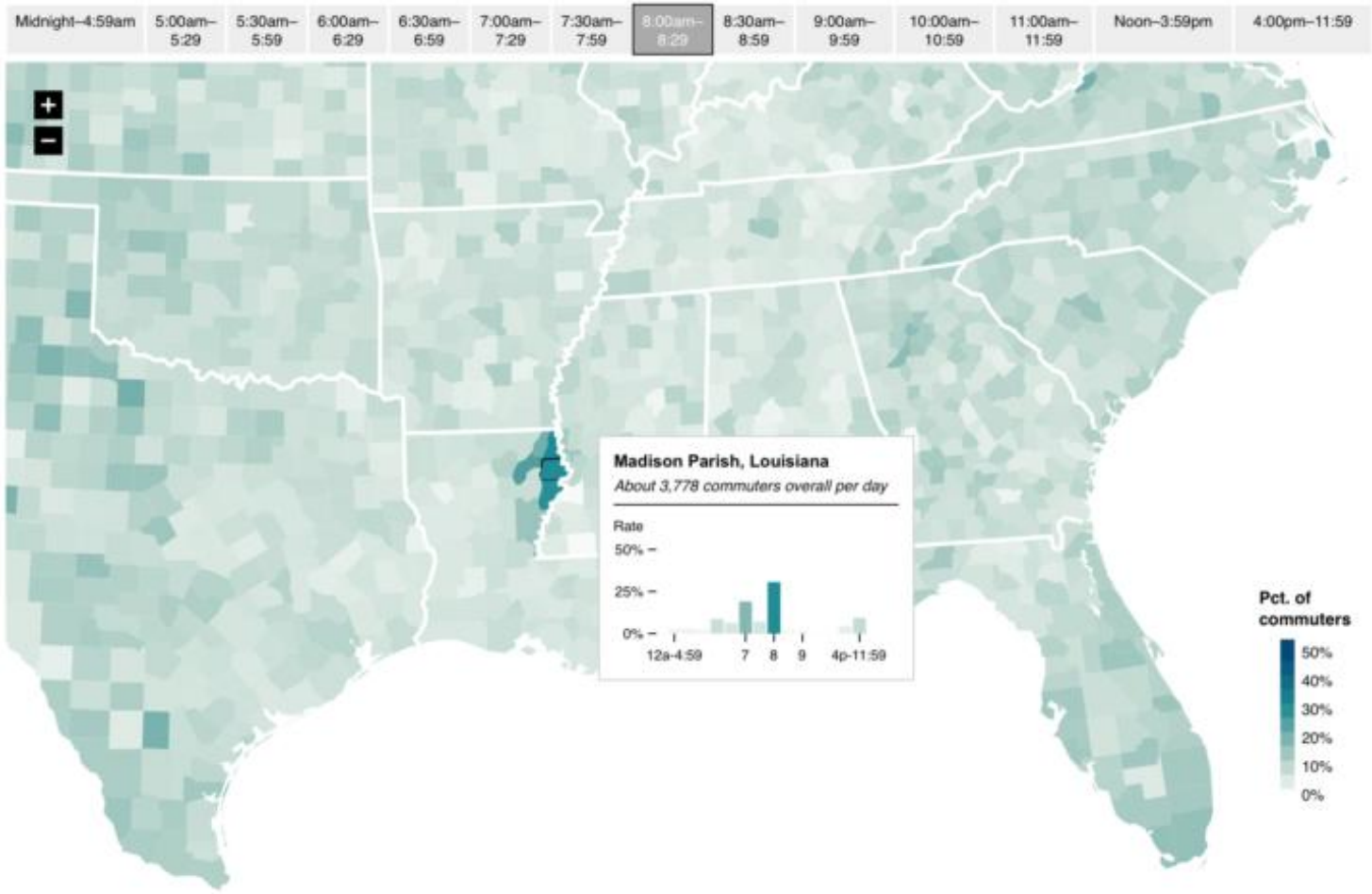


Visualizing Outliers

Breakout: Visualize the data as you normally would for an overview, and then zoom in or highlight outliers to explain

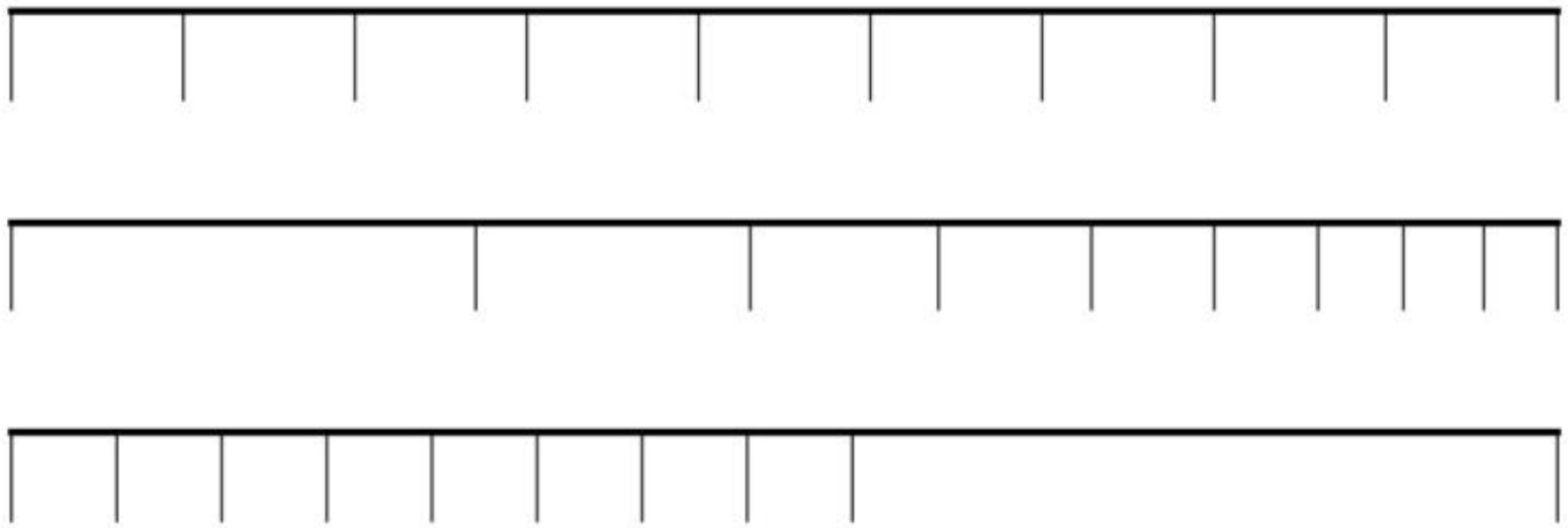
Visualizing Outliers

Breakout: Visualize the data as you normally would for an overview, and then zoom in or highlight outliers to explain



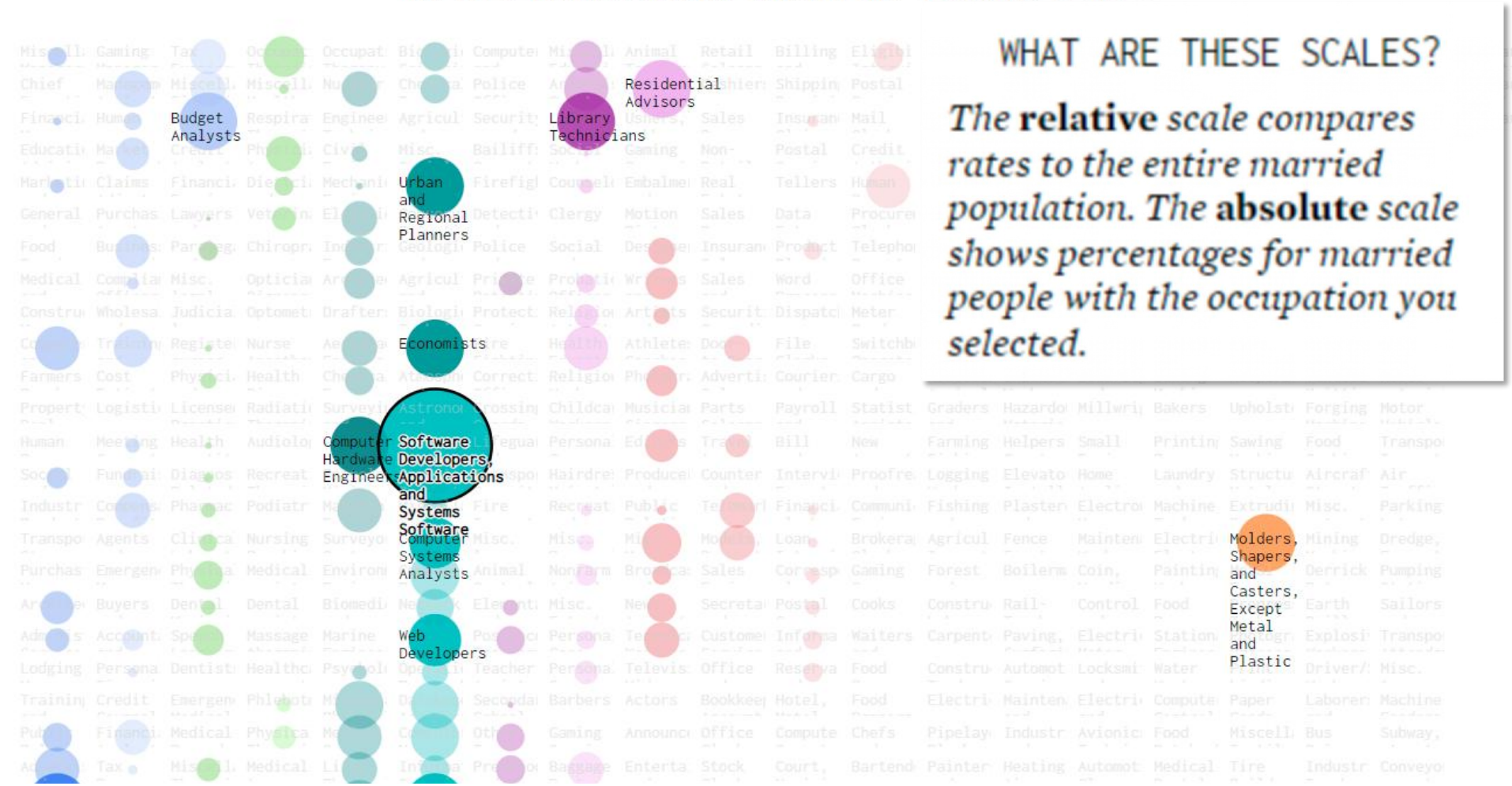
Visualizing Outliers

Scale Adjustment: sometimes outliers are viewed better on a different scale that allows for extremes and averages to display at the same time.



Who do Software Developers, Applicat usually marry?

Or, see a random occupation. Show relative or absolute scale.

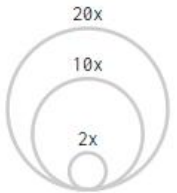


WHAT ARE THESE SCALES?

*The **relative** scale compares rates to the entire married population. The **absolute** scale shows percentages for married people with the occupation you selected.*

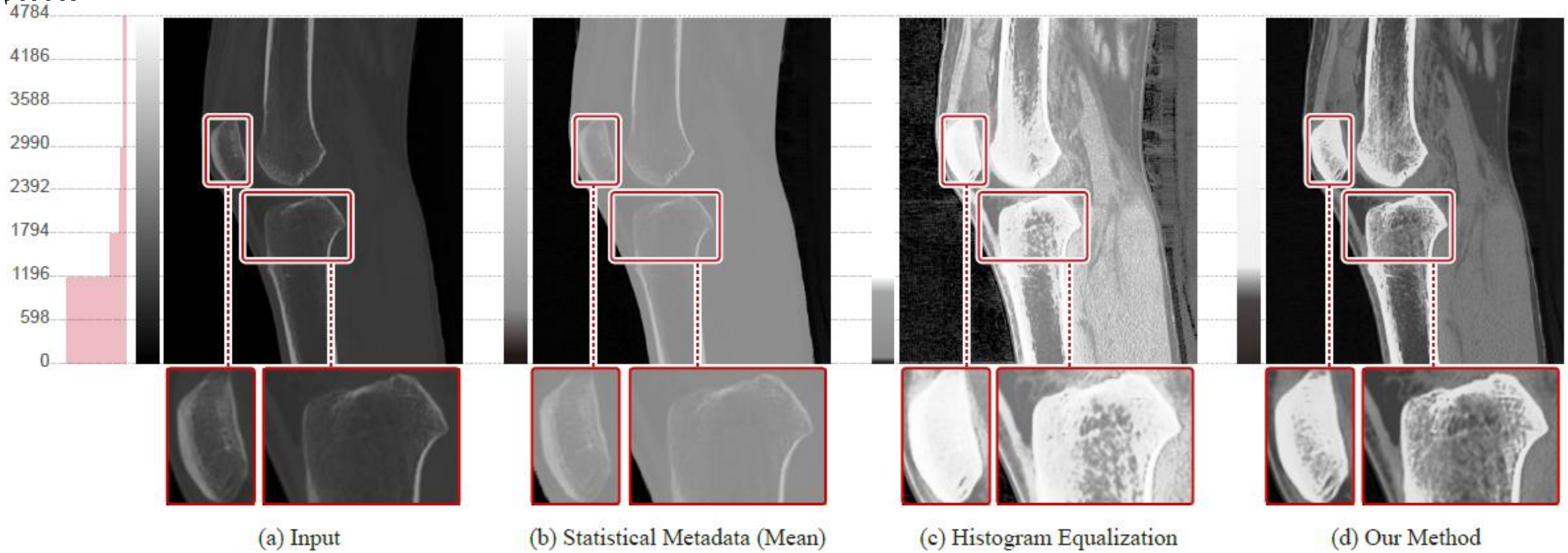
- MANAGEMENT
- BUSINESS OPERATIONS
- FINANCE
- LEGAL
- HEALTHCARE PRACTITIONERS
- HEALTHCARE SUPPORT
- ARCHITECTURE AND ENGINEERING
- LIFE, PHYSICAL, AND SOCIAL SCIENCE
- PROTECTIVE SERVICES
- EDUCATION AND LIBRARY
- COMMUNITY AND SOCIAL SERVICES
- PERSONAL CARE AND SERVICE
- ARTS AND ENTERTAINMENT
- SALES
- OFFICE AND ADMINISTRATIVE SUPPORT
- FOOD PREPARATION AND SERVING
- BUILDING AND GROUNDS CLEANING
- FARMING, FISHING, AND FORESTRY
- CONSTRUCTION
- INSTALLATION AND MAINTENANCE
- PRODUCTION
- EXTRACTION
- TRANSPORTATION AND MATERIAL MOVING
- MILITARY

RELATIVE TO MARRIED POP.



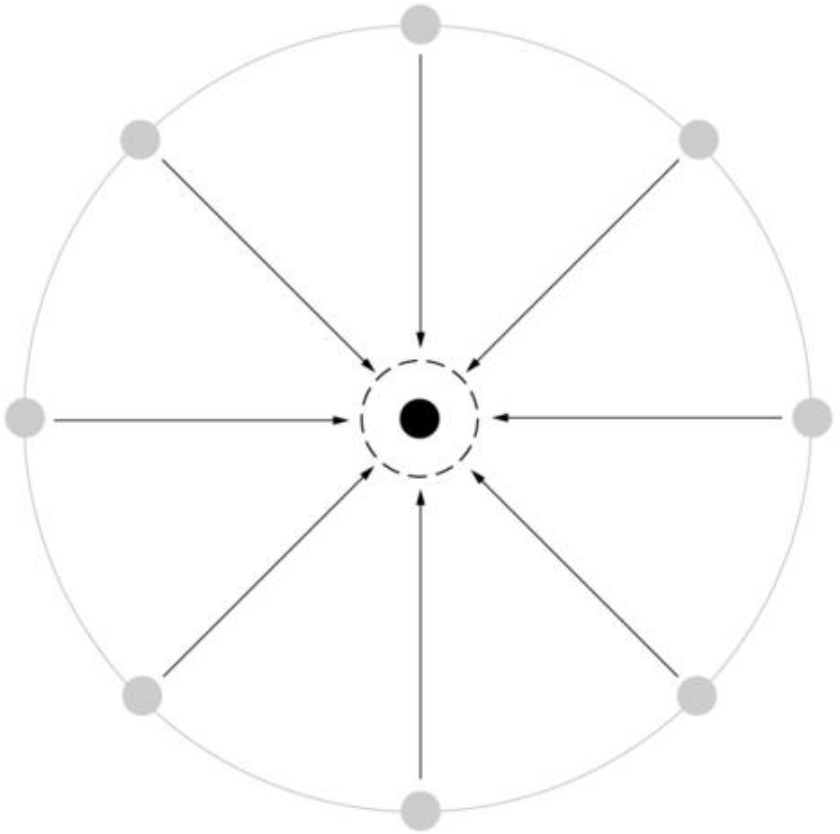
Visualizing Outliers

Scale Adjustment: *sometimes* outliers are viewed better on a different scale that allows for extremes and averages to display at the same time



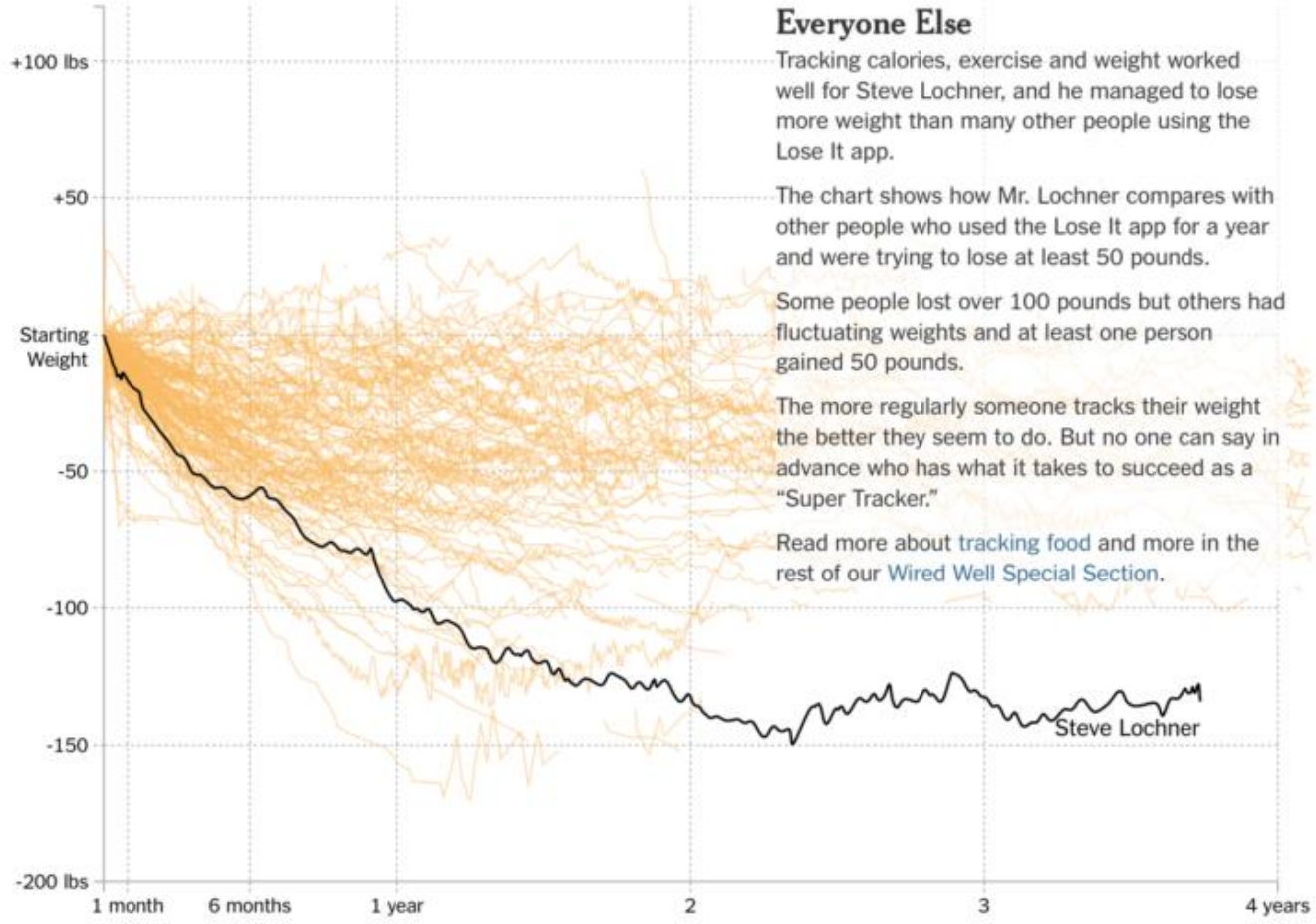
Visualizing Outliers

Reference Point: Use the outlier as a point of comparison for a sense of scale or to make the data more relatable.



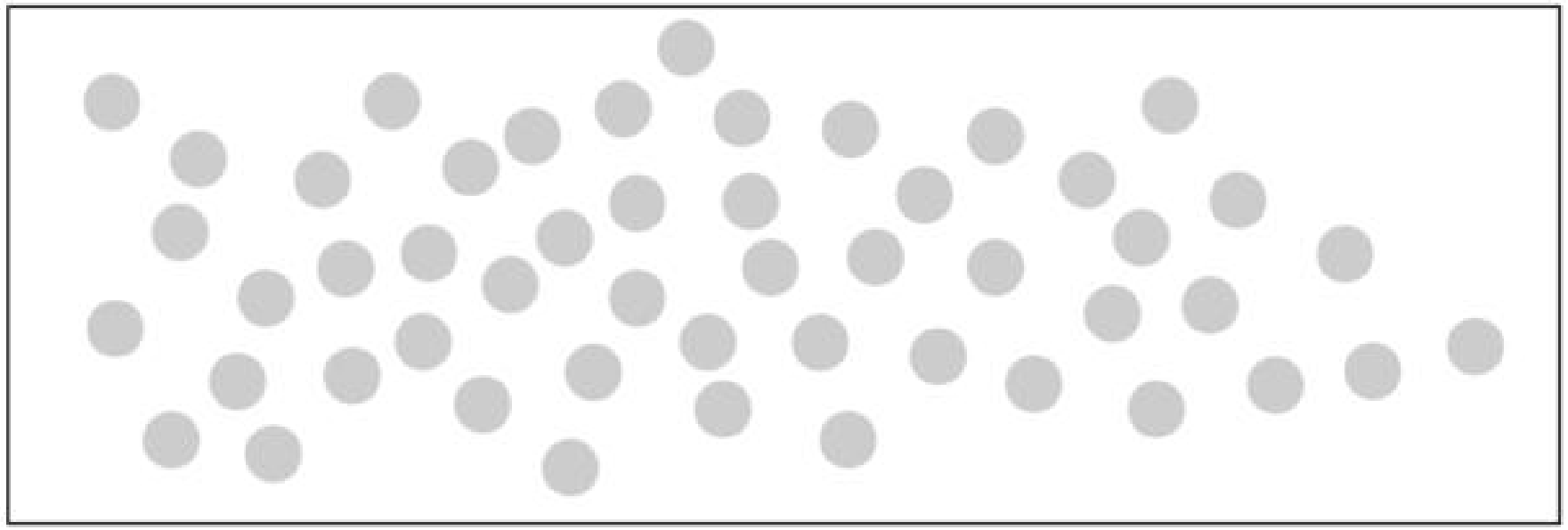
Visualizing Outliers

Reference Point: Use the outlier as a point of comparison for a sense of scale or to make the data more relatable.



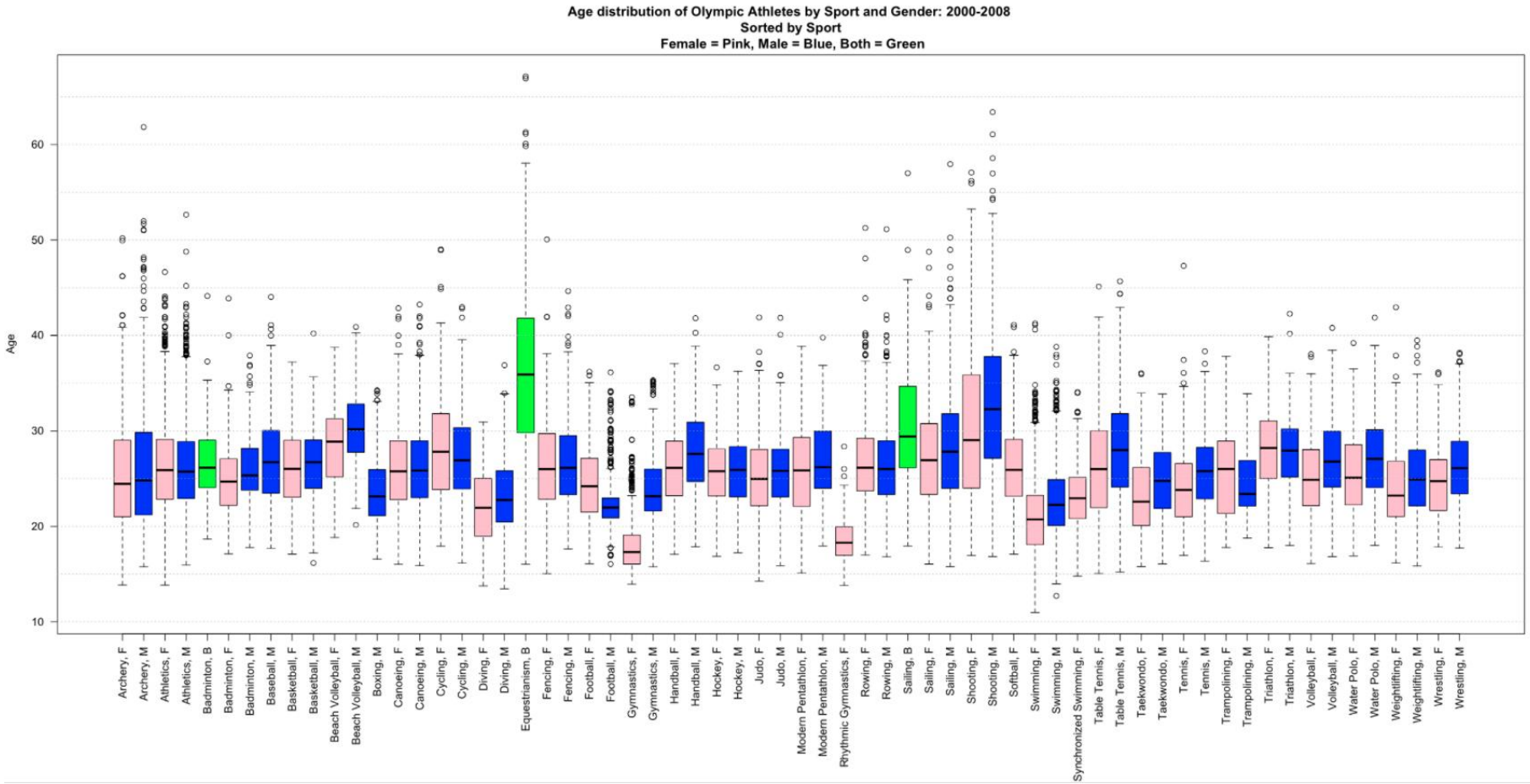
Visualizing Outliers

Providing Context: Maybe you don't want to highlight the outlier. Maybe it's not as important as the rest of the dataset. In this case, use it as context or background.



Visualizing Outliers

Providing Context: use the outlier as a point of comparison for a sense of scale or to make the data more relatable.



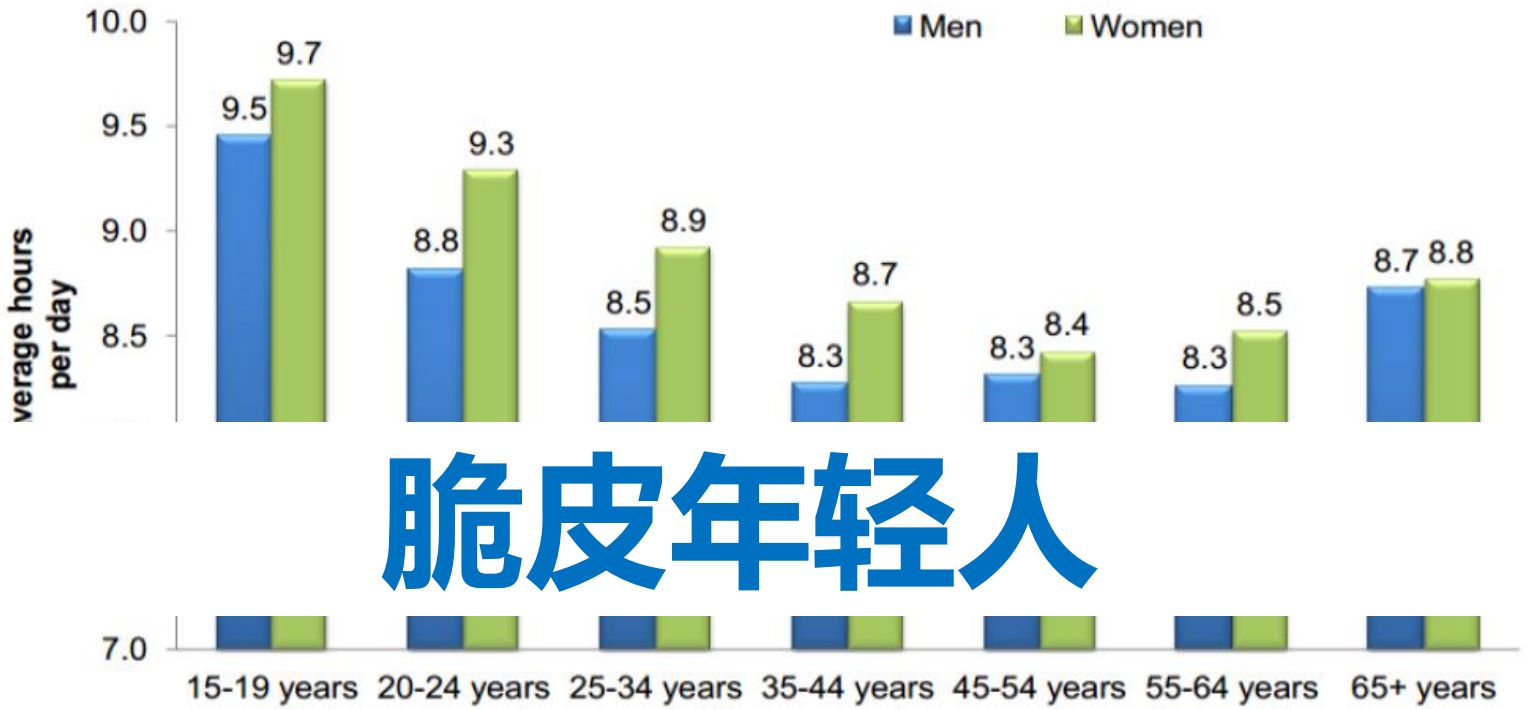
Data Quality Tools

Data Quality Management Comparison Chart

Vendor	Tools	Focus	Key Features
Cloudingo	Cloudingo	Salesforce data	Deduplication; data migration management; spots human and other errors/inconsistencies
Data Ladder	DataMatch Enterprise; ProductMatch	Diverse data sets across numerous applications and formats	Includes more than 300,000 prebuilt rules; templates and connectors for most major applications
IBM	InfoSphere QualityStage	Big data, business intelligence; data warehousing; application migration and master data management	Includes more than 200 built-in data quality rules; strong machine learning and governance tools
Informatica	Data Quality Master Data Management	Accommodates diverse data sets; supports Azure and AWS	Data standardization, validation, enrichment, deduplication, and consolidation
OpenRefine	OpenRefine	Transforms, cleanses and formatts data for analytics and other purposes	Powerful capture and editing functions.
SAS	Data Management	Managing data integration and cleansing for diverse data sources and sets	Strong metadata management; supports 38 languages
Syncsort	Trillium Quality for Dynamics; Trillium Quality for Big Data;	Cleansing, optimizing and integrating data from numerous sources	DQ supports more than 230 countries, regions and territories; works with major architectures, including Hadoop, Spark, SAP and MS Dynamics
Talend	Data Quality	Data integration	Deduplication, validation and standardization using machine learning; templates and reusable elements to aid in data cleansing
TIBCO	Clarity	High volume data analysis and cleansing	Tools for profiling, validating, standardizing, transforming, deduplicating, cleansing and visualizing for all major data sources and file types
Validity	DemandTools	Salesforce data	Handles multi-table mass manipulations and standardizes Salesforce objects and

Survey: how many hours of sleep did you get last night?

Average sleep times per day, by age and sex



脆皮年轻人

NOTE: Data include all persons age 15 and over. Data include all days of the week and are annual averages for 2012.

SOURCE: Bureau of Labor Statistics, American Time Use Survey

Finding the error:

- How was the data collected?
- Who was included?
- How were they selected?

Addressing its impact:

- If it is your sample, pay careful attention to how you select participants
- Is it random or does it introduce bias?
- Is it more likely in some variables than others?
- Is there a specific type of error you expect?

Outline

Data Quality

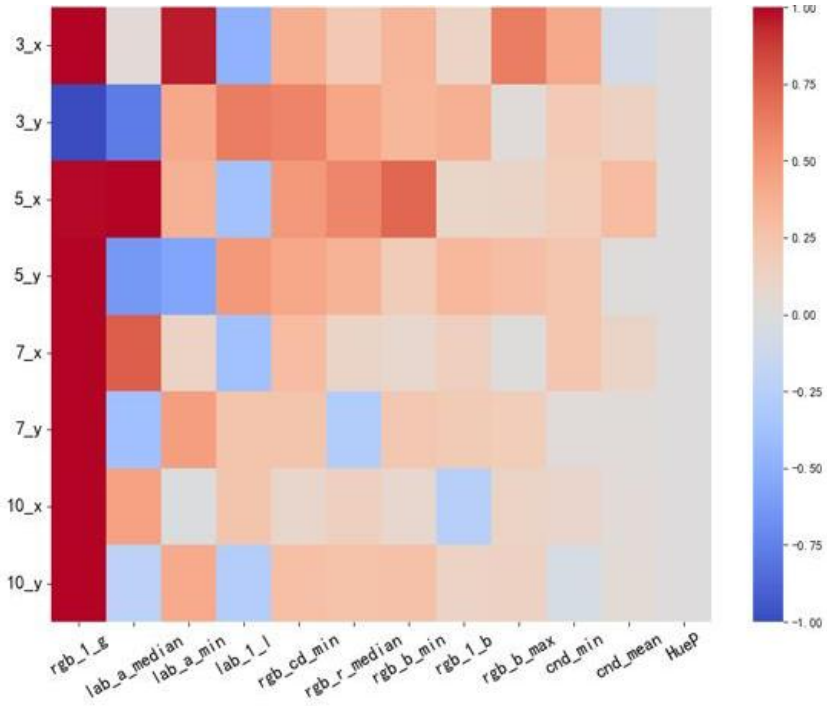
Data Reduction

Data Reduction

- Data is too big to work with
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- **Data reduction strategies**
  - Data Sampling
  - Data Compression
  - Dimensionality reduction — remove unimportant attributes

High-Dimensional Data

- Hard to store and process data (computationally challenging)
- Complexity of decision rule tends to grow with # **features**.
- Hard to learn complex rules as dimension increases (statistically challenging)
- Hard to interpret and visualize

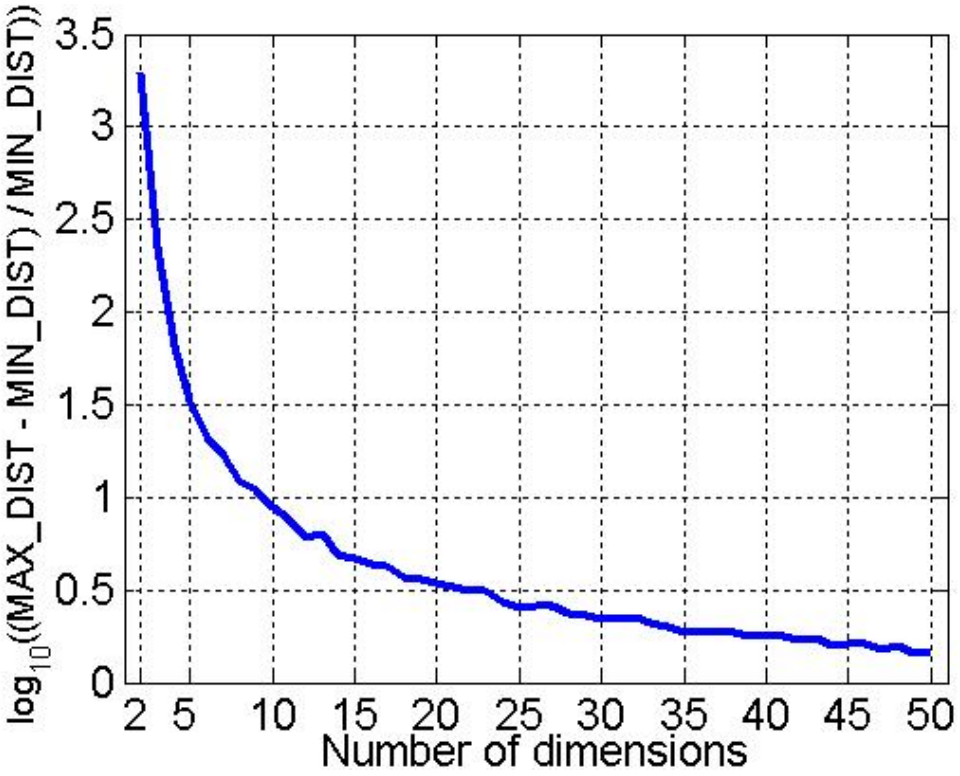


		Basic LMA Features f^i	Derived Features $\hat{f}^i$			
		Description	*max*	*min*	*std*	*mean*
Body	f^1	Left foot-hip distance	$\hat{f}^1$	$\hat{f}^2$	$\hat{f}^3$	$\hat{f}^4$
	f^2	Right foot-hip distance	$\hat{f}^5$	$\hat{f}^6$	$\hat{f}^7$	$\hat{f}^8$
	f^3	Left hand-shoulder distance	$\hat{f}^9$	$\hat{f}^{10}$	$\hat{f}^{11}$	$\hat{f}^{12}$
	f^4	Right hand-shoulder distance	$\hat{f}^{13}$	$\hat{f}^{14}$	$\hat{f}^{15}$	$\hat{f}^{16}$
	f^5	Hands distance	$\hat{f}^{17}$	$\hat{f}^{18}$	$\hat{f}^{19}$	$\hat{f}^{20}$, $\hat{f}^1$
	f^6	Left hand-head distance	$\hat{f}^{21}$	$\hat{f}^{22}$	$\hat{f}^{23}$	$\hat{f}^{24}$
	f^7	Right hand-head distance	$\hat{f}^{25}$	$\hat{f}^{26}$	$\hat{f}^{27}$	$\hat{f}^{28}$
	f^8	Left hand-hip distance	$\hat{f}^{29}$	$\hat{f}^{30}$	$\hat{f}^{31}$	$\hat{f}^{32}$, $\hat{f}^2$
	f^9	Right hand-hip distance	$\hat{f}^{33}$	$\hat{f}^{34}$	$\hat{f}^{35}$	$\hat{f}^{36}$, $\hat{f}^3$
	f^{10}	Hip-ground distance	$\hat{f}^{37}$	$\hat{f}^{38}$	$\hat{f}^{39}$	$\hat{f}^{40}$, $\hat{f}^4$
	f^{11}	Hip-ground minus feet-hip	$\hat{f}^{41}$	$\hat{f}^{42}$	$\hat{f}^{43}$	$\hat{f}^{44}$
	f^{12}	Feet distance	$\hat{f}^{45}$	$\hat{f}^{46}$	$\hat{f}^{47}$	$\hat{f}^{48}$, $\hat{f}^5$
	f^{13}	Left hand and chest	$\hat{f}^{113}$	$\hat{f}^{114}$	$\hat{f}^{115}$	$\hat{f}^{116}$, $\hat{f}^{30}$
	f^{14}	Right hand and chest	$\hat{f}^{117}$	$\hat{f}^{118}$	$\hat{f}^{119}$	$\hat{f}^{120}$, $\hat{f}^{31}$
Effort	f^{15}	Deceleration peaks				$\hat{f}^{49}$, $\hat{f}^6$
	f^{16}	Pelvis velocity	$\hat{f}^{50}$		$\hat{f}^{51}$	$\hat{f}^{52}$, $\hat{f}^7$
	f^{17}	Left hand velocity	$\hat{f}^{53}$		$\hat{f}^{54}$	$\hat{f}^{55}$, $\hat{f}^8$
	f^{18}	Right hand velocity	$\hat{f}^{56}$		$\hat{f}^{57}$	$\hat{f}^{58}$, $\hat{f}^9$
	f^{19}	Left foot velocity	$\hat{f}^{59}$		$\hat{f}^{60}$	$\hat{f}^{61}$, $\hat{f}^{10}$
	f^{20}	Right foot velocity	$\hat{f}^{61}$		$\hat{f}^{61}$	$\hat{f}^{64}$, $\hat{f}^{11}$
	f^{21}	Pelvis acceleration	$\hat{f}^{65}$, $\hat{f}^{12}$		$\hat{f}^{66}$	
	f^{22}	Left hand acceleration	$\hat{f}^{67}$, $\hat{f}^{13}$		$\hat{f}^{68}$	
	f^{23}	Right hand acceleration	$\hat{f}^{69}$, $\hat{f}^{14}$		$\hat{f}^{70}$	
	f^{24}	Left foot acceleration	$\hat{f}^{71}$, $\hat{f}^{15}$		$\hat{f}^{72}$	
	f^{25}	Right foot acceleration	$\hat{f}^{73}$, $\hat{f}^{16}$		$\hat{f}^{74}$	
	f^{26}	Jerk	$\hat{f}^{75}$, $\hat{f}^{17}$		$\hat{f}^{76}$	
Shape	f^{27}	Volume (5 joints)	$\hat{f}^{77}$	$\hat{f}^{78}$	$\hat{f}^{79}$	$\hat{f}^{80}$, $\hat{f}^{18}$
	f^{28}	Volume (All joints)	$\hat{f}^{81}$	$\hat{f}^{82}$	$\hat{f}^{83}$	$\hat{f}^{84}$, $\hat{f}^{19}$
	f^{29}	Torso height	$\hat{f}^{85}$	$\hat{f}^{86}$	$\hat{f}^{87}$	$\hat{f}^{88}$, $\hat{f}^{20}$
	f^{30}	Hands level				$\hat{f}^{89}$, $\hat{f}^{91}$, $\hat{f}^{21}$, $\hat{f}^{22}$
	f^{31}	Volume (upper body)	$\hat{f}^{97}$	$\hat{f}^{98}$	$\hat{f}^{99}$	$\hat{f}^{100}$, $\hat{f}^{26}$
	f^{32}	Volume (lower body)	$\hat{f}^{101}$	$\hat{f}^{102}$	$\hat{f}^{103}$	$\hat{f}^{104}$, $\hat{f}^{27}$
	f^{33}	Volume (right side)	$\hat{f}^{105}$	$\hat{f}^{106}$	$\hat{f}^{107}$	$\hat{f}^{108}$, $\hat{f}^{28}$
	f^{34}	Volume (left side)	$\hat{f}^{109}$	$\hat{f}^{110}$	$\hat{f}^{111}$	$\hat{f}^{112}$, $\hat{f}^{29}$
Space	f^{35}	Total distance				$\hat{f}^{92}$, $\hat{f}^{24}$
	f^{36}	Area per second	$\hat{f}^{93}$	$\hat{f}^{94}$	$\hat{f}^{95}$	$\hat{f}^{96}$, $\hat{f}^{25}$
	f^{37}	Total volume				$\hat{f}^{121}$

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

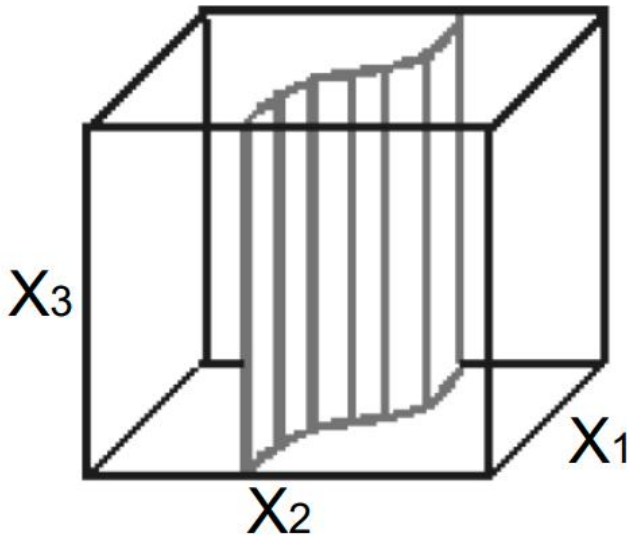
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**

Dimensionality Reduction

- Feature selection :
  - Select a minimum set of attributes (features) that is sufficient for the data analytical task.



X_3 - Irrelevant

  - Filter-based, wrapper-based, embeded-based

主观题 10分

设置

Embeded Method

Wrapper Methods

Filter Methods

1. 每个小组成员根据给定的材料分别学习特征选择方式（5min）
2. 阅读相同部分的同学自动组成专家组讨论不清楚的问题（5min）
3. 同学们把掌握的内容教给其他组员（5min）

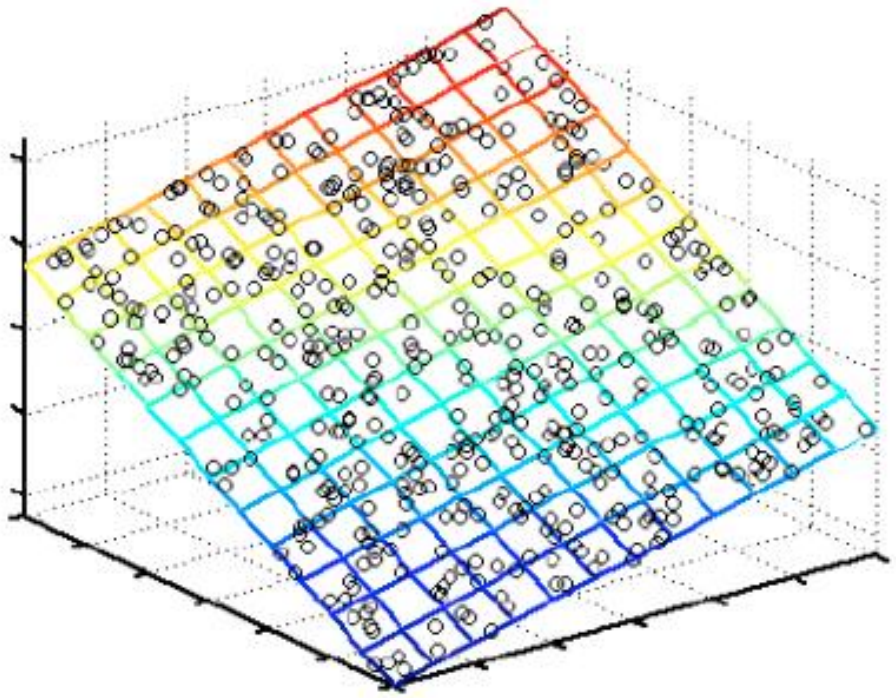
正常使用主观题需2.0以上版本雨课堂

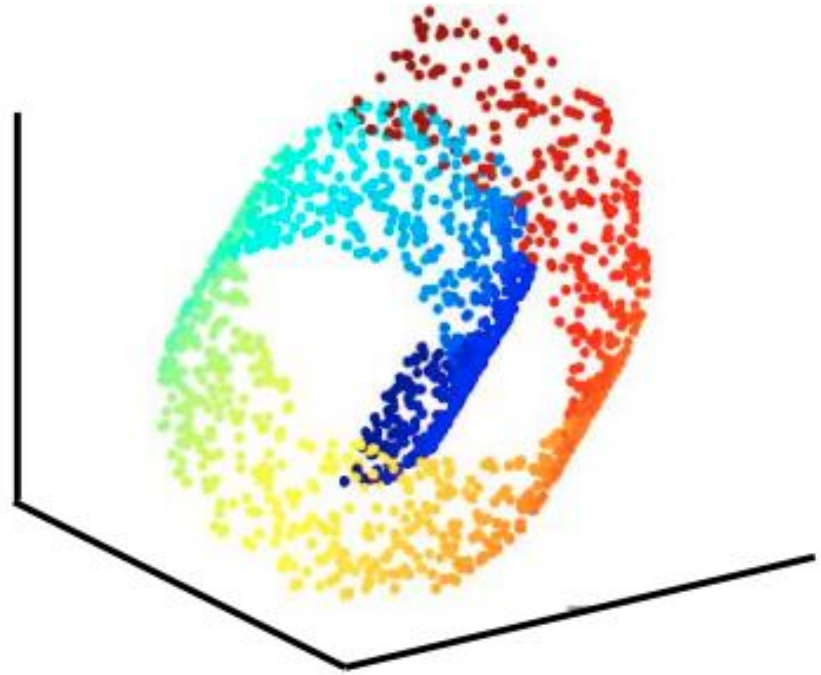
作答

Dimensionality Reduction

Latent Features:

Some linear/nonlinear combination of features provides a more efficient representation than observed features





多选题 1分

设置

此题未设置答案，请点击右侧设置按钮

以下关于 PCA 主成分分析的描述，正确的有（ ）

A PCA 的核心目标是在保留原始数据关键信息的前提下，将高维特征空间映射到低维特征空间，通过构建“主成分”实现降维

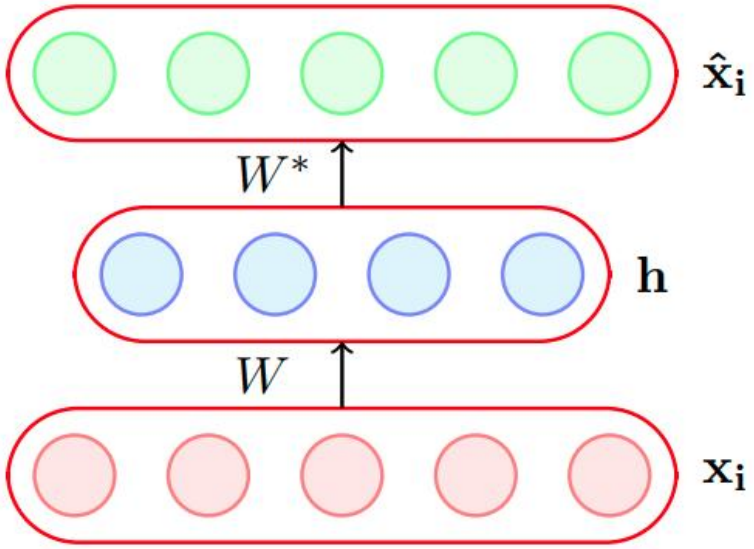
B PCA 中所谓的“关键信息”，本质是数据中能最大化体现样本间差异的信息，该差异程度可通过方差衡量，方差越大意味着该维度承载的信息价值越高

C 协方差矩阵的特征向量对应 PCA 主成分的坐标轴方向，特征值的大小则代表原始数据在对应特征向量方向上的方差大小，特征值越大，对应主成分承载的信息越核心

D 为避免信息冗余，PCA 要求各主成分之间需满足“线性相关”，即后一主成分与前一主成分的协方差需大于 0

提交

Auto-encoder



$$\boldsymbol{h} = g(W\boldsymbol{x_i} + \boldsymbol{b})$$

$$\widehat{\boldsymbol{x_i}} = f(W\boldsymbol{h} + \boldsymbol{c})$$

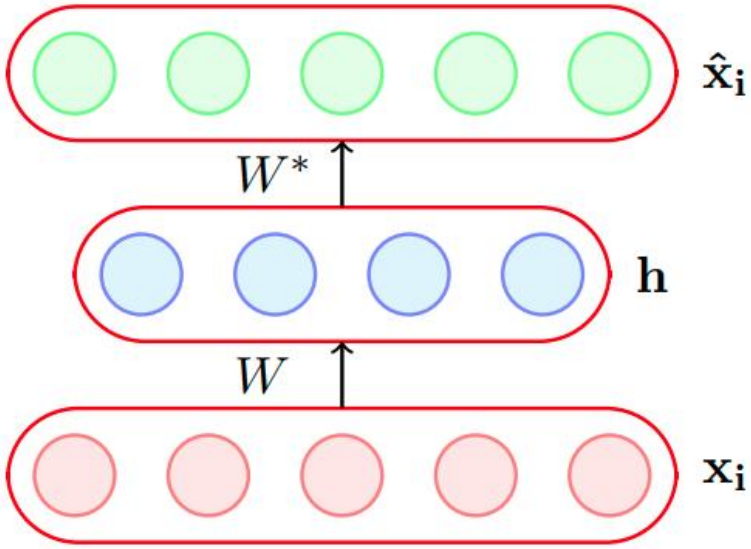
An autoencoder is a special type of feed forward neural network which does the following

- Encodes its input $\mathbf{x_i}$ into a hidden representation **h**
- Decodes the input again from this hidden representation
- The model is trained to minimize a certain loss function which will ensure that $\widehat{\boldsymbol{x_i}}$ is close to $\mathbf{x_i}$

$$\mathcal{L}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2$$

Credit to Mitesh M. Khapra

Auto-encoder



$$\boldsymbol{h} = g(W\boldsymbol{x_i} + \boldsymbol{b})$$

$$\hat{\boldsymbol{x}}_{\boldsymbol{i}} = f(W\boldsymbol{h} + \boldsymbol{c})$$

Let us consider the case where dim(**h**) < dim($\mathbf{x}_i$)

- If we are still able to reconstruct $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$, perfectly from **h**, then what does it say about **h**?
- **h** is a loss-free encoding of $\mathbf{x}_i$. It captures all the important characteristics of $\mathbf{x}_i$
- Do you see an analogy with PCA?

Credit to Mitesh M. Khapra

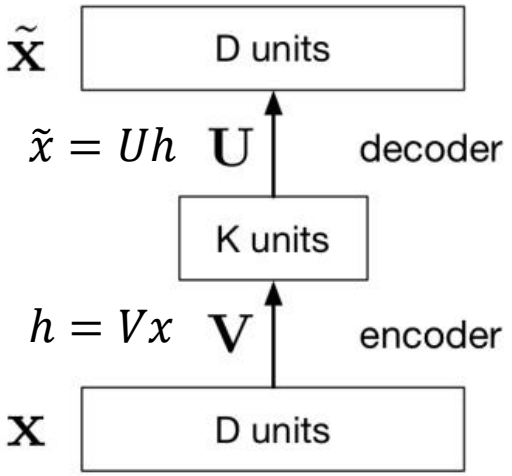
PCA

- The simplest kind of autoencoder has one hidden layer, linear activations, and squared error loss.

$$\mathcal{L}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2$$

$$\boldsymbol{h} = g(W\boldsymbol{x_i} + \boldsymbol{b})$$

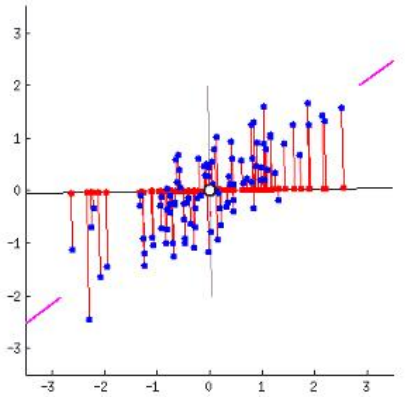
$$\hat{\boldsymbol{x}}_{\boldsymbol{i}} = f(W\boldsymbol{h} + \boldsymbol{c})$$



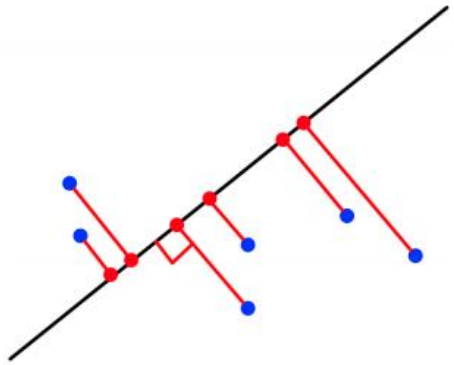
- This network computes $\tilde{\boldsymbol{x}}$ = **UVx**, which is a linear function.
- If K ≥ D, we can choose **U** and **V** such that **UV** is the identity. This isn't very interesting.
- But suppose K < D:
  - **V** maps **x** to a K-dimensional space, so it's doing dimensionality reduction.
  - The output must lie in a K-dimensional subspace, namely the column space of **U**.

PCA

- We just saw that a linear autoencoder has to map D-dimensional inputs to a K-dimensional subspace *S*.

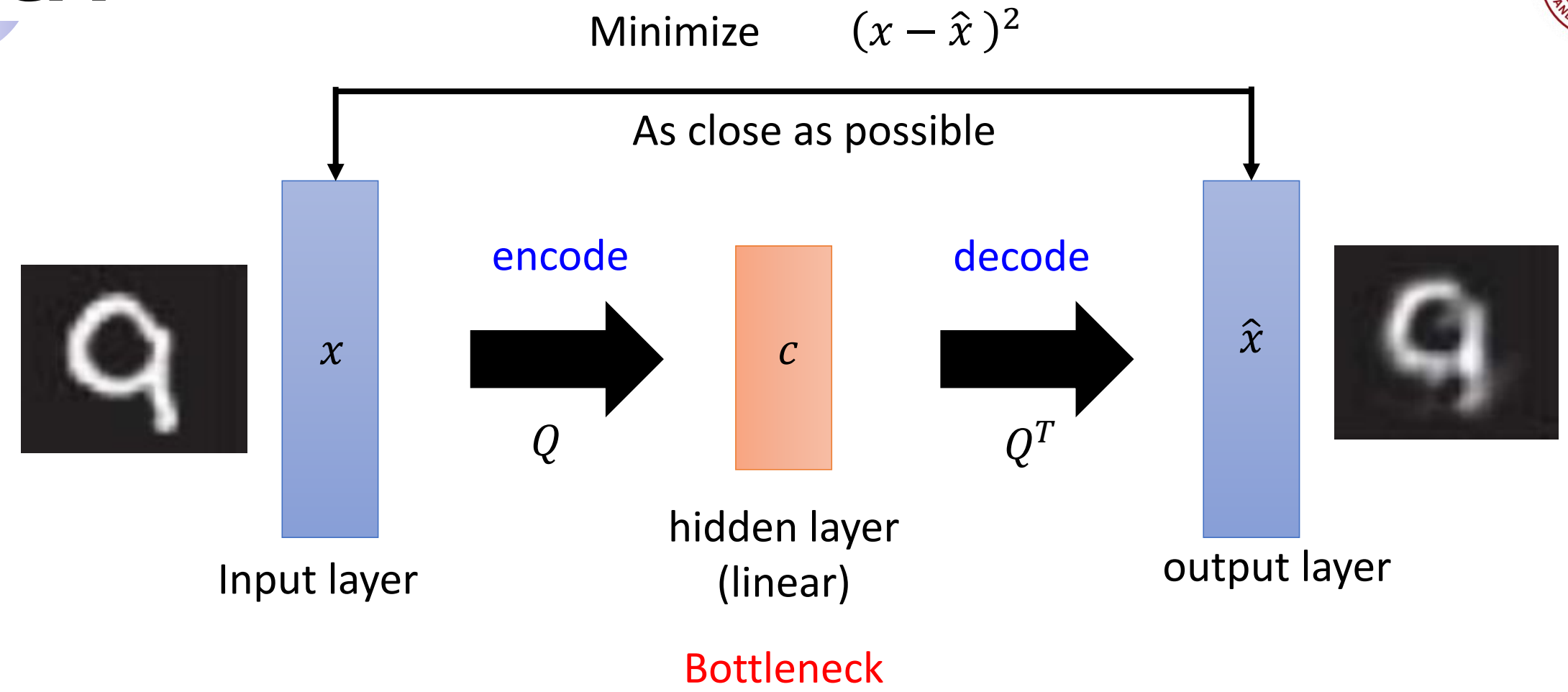


- Knowing this, what is the best possible mapping it can choose?
- By definition, the projection of **x** onto *S* is the point in *S* which minimizes the distance to **x**.



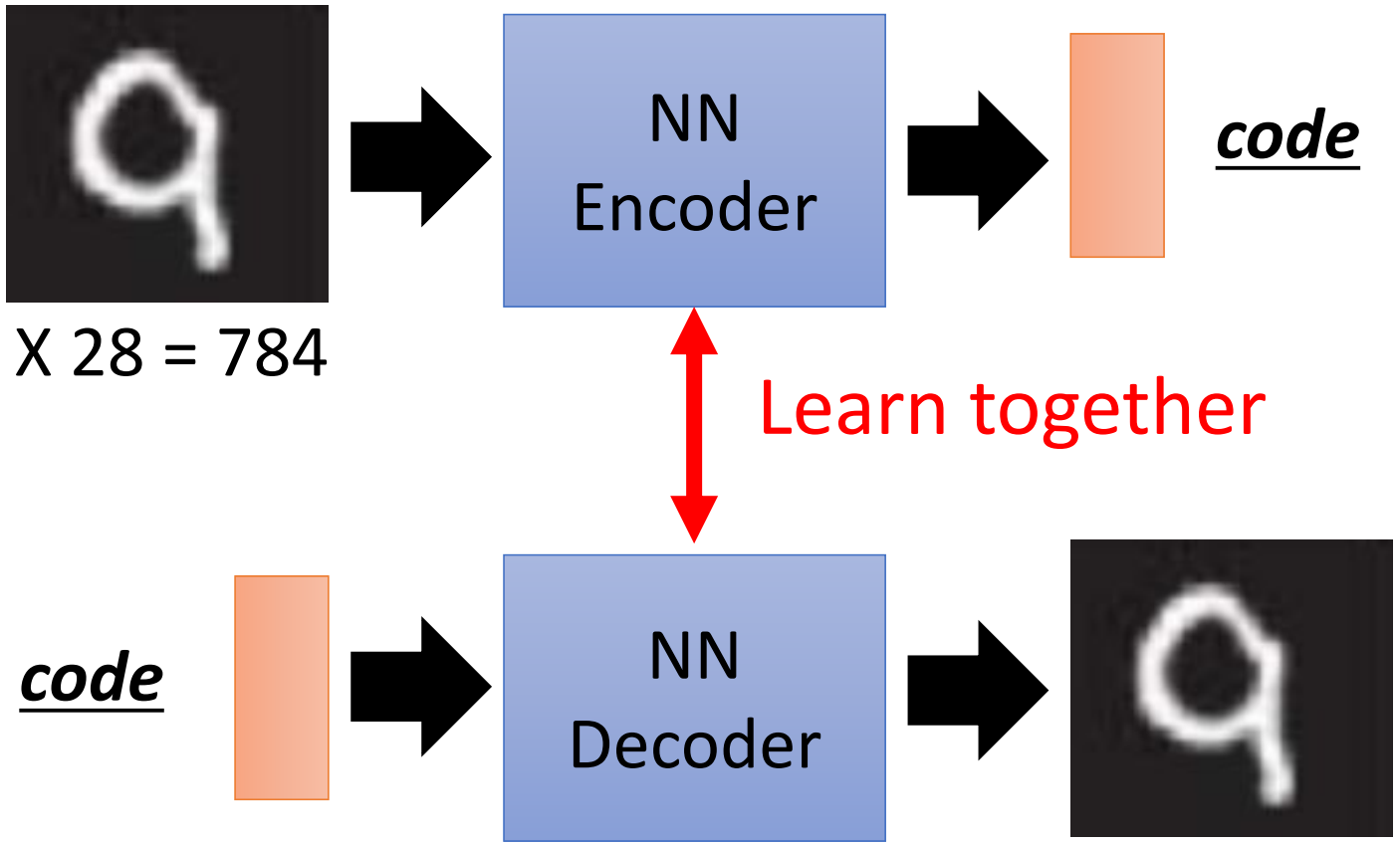
- Fortunately, the linear autoencoder can represent projection onto *S*: pick $\mathbf{U} = \mathbf{Q}$ and $\mathbf{V} = \mathbf{Q}^T$, where **Q** is an orthonormal basis for *S* .

PCA



Output of the hidden layer is the code

Auto-encoder



Compact representation of the input object

Can reconstruct the original object

Auto-encoder

Real data

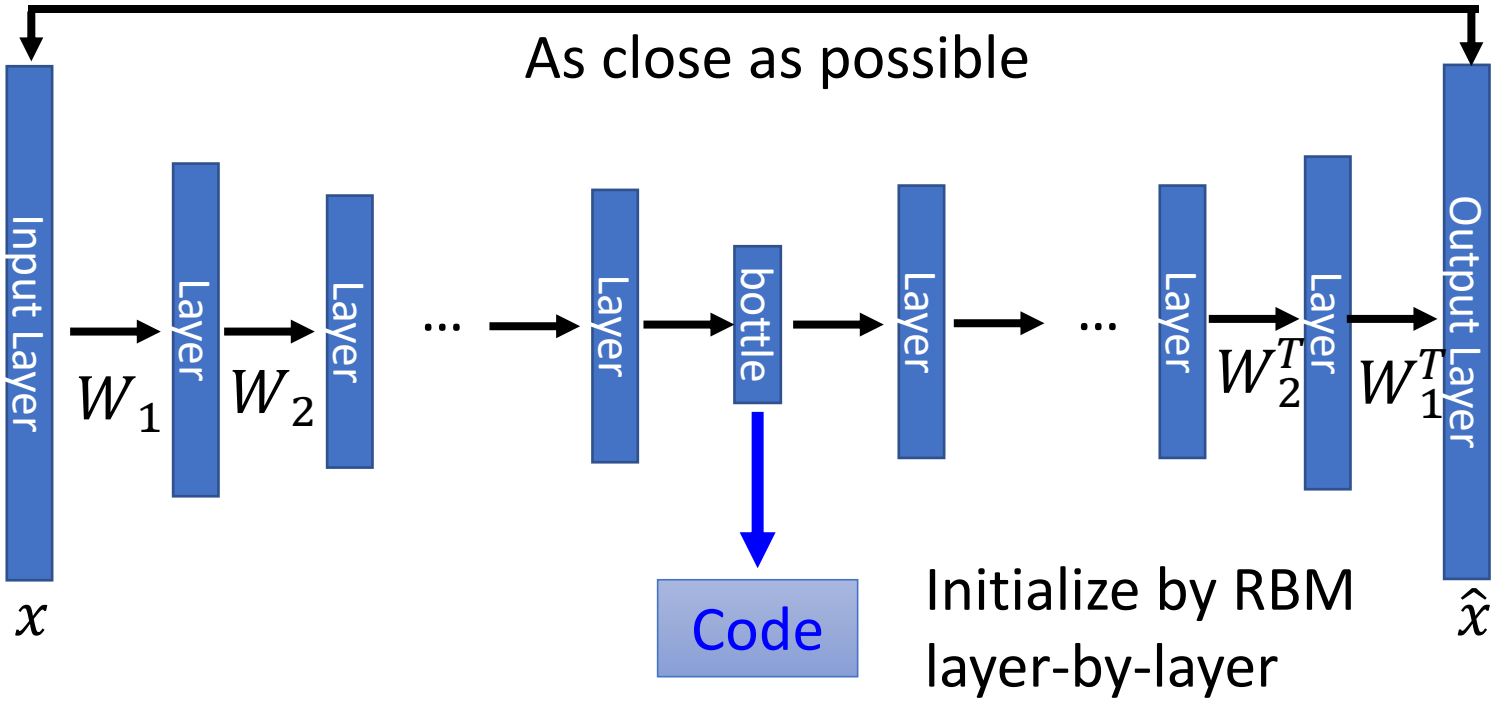
30-D deep auto

30-D PCA

Deep Auto-encoder

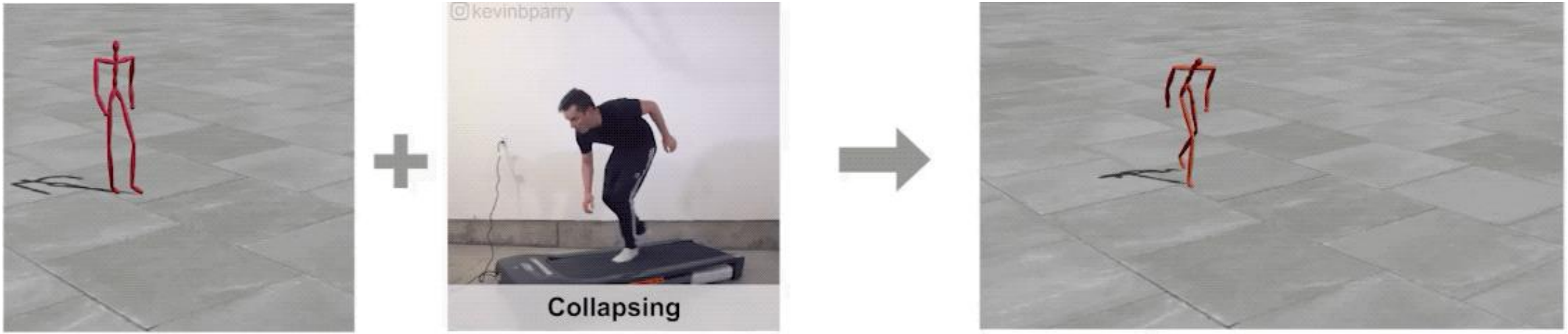
Of course, the auto-encoder can be deep

Symmetric is not necessary.



Reference: Hinton, Geoffrey E., and Ruslan R. Salakhutdinov.
"Reducing the dimensionality of data with neural networks." *Science* 313.5786 (2006): 504-507

Application



[Aberman et al., 2020] Unpaired Motion Style Transfer from Video to Animation

Application

Style Input (proud)

Content Input

Output

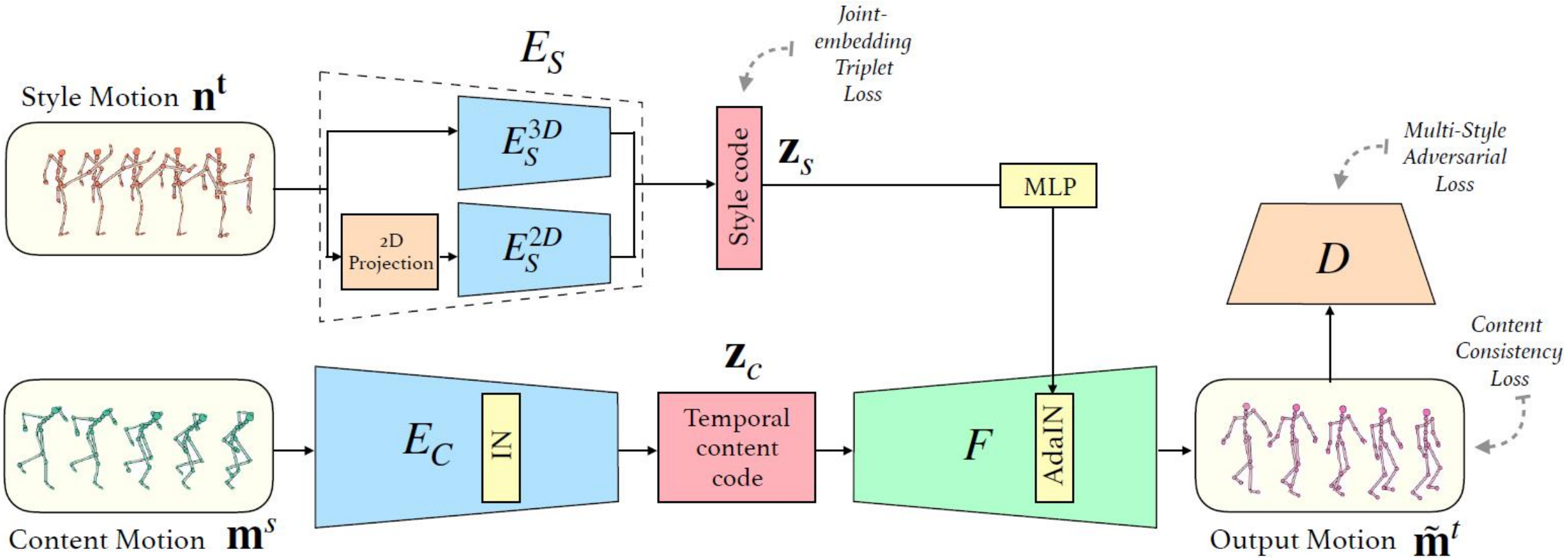
Style Input (crouched)

Content Input

Output

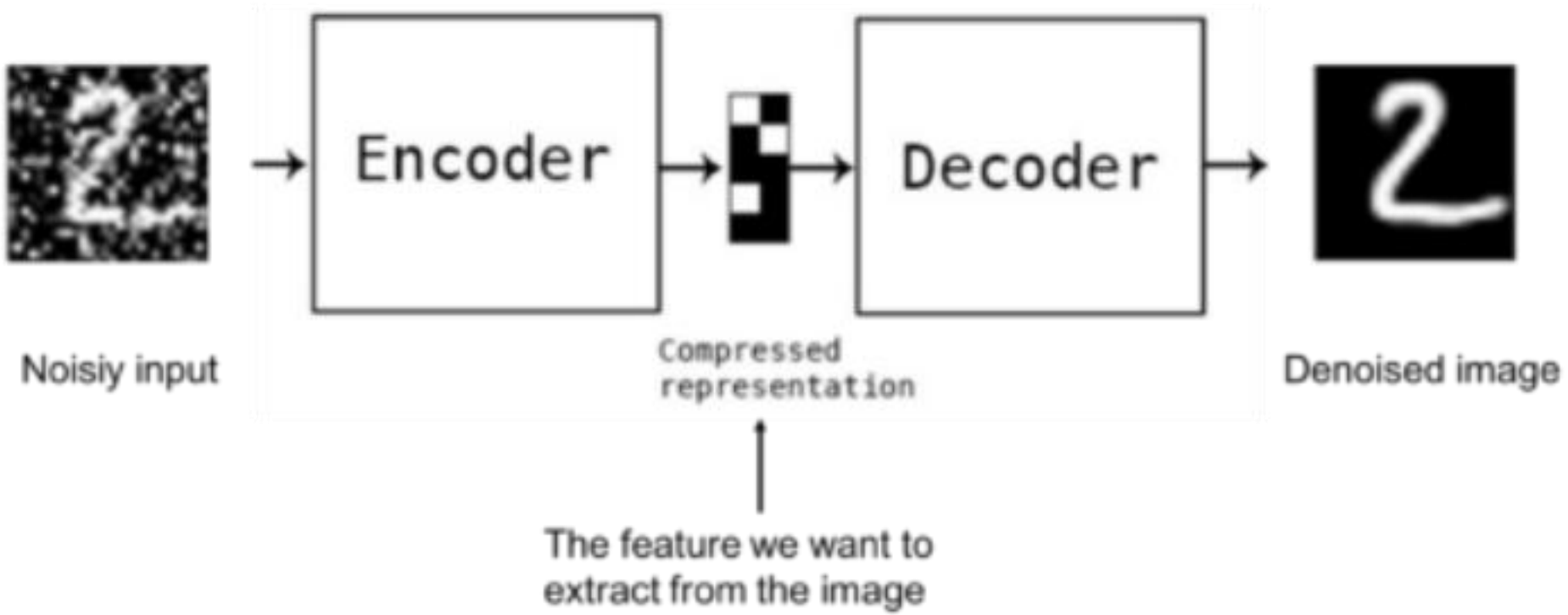
[Aberman et al., 2020] Unpaired Motion Style Transfer from Video to Animation

Application

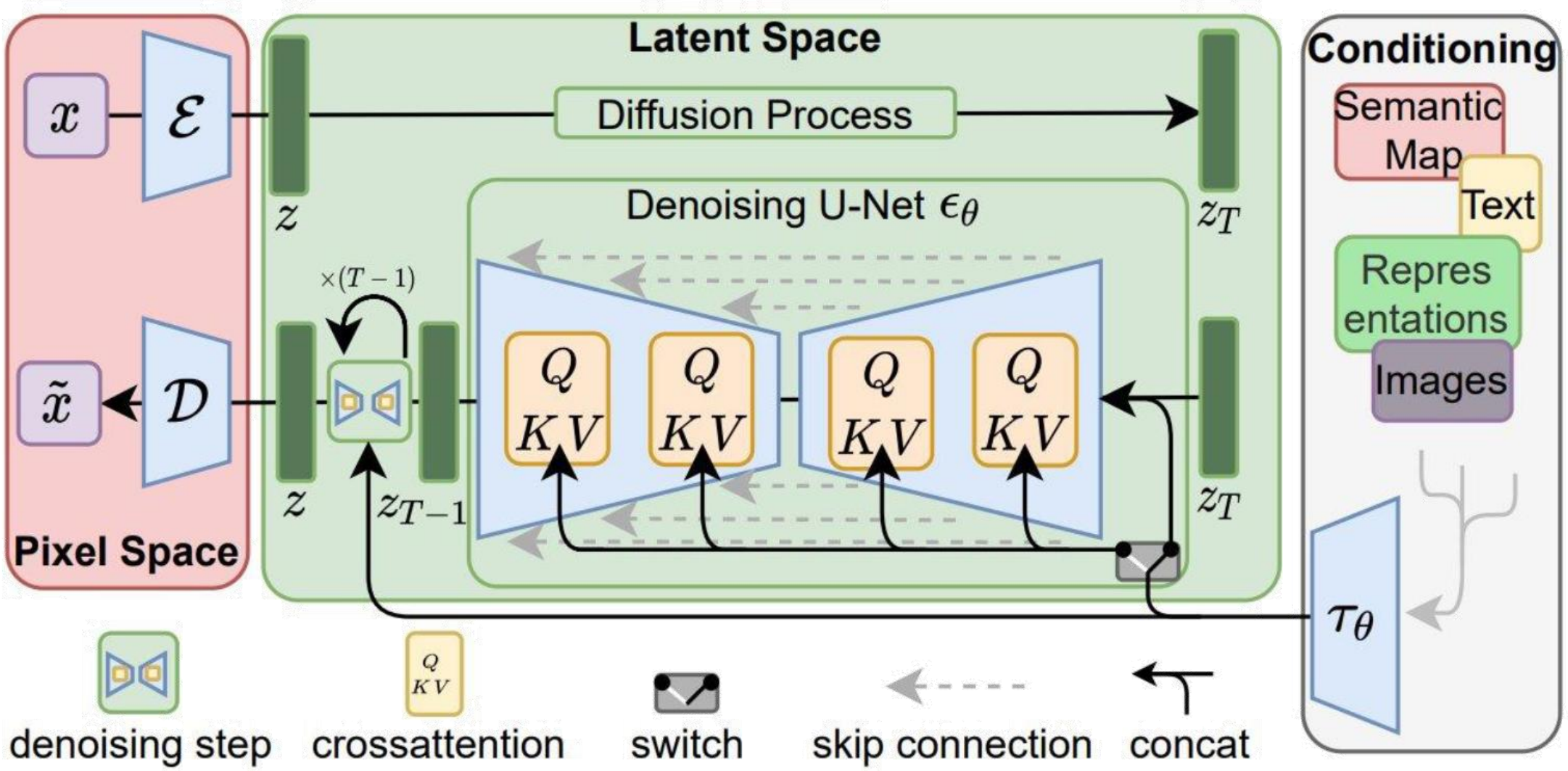


[Aberman et al., 2020] Unpaired Motion Style Transfer from Video to Animation

Application

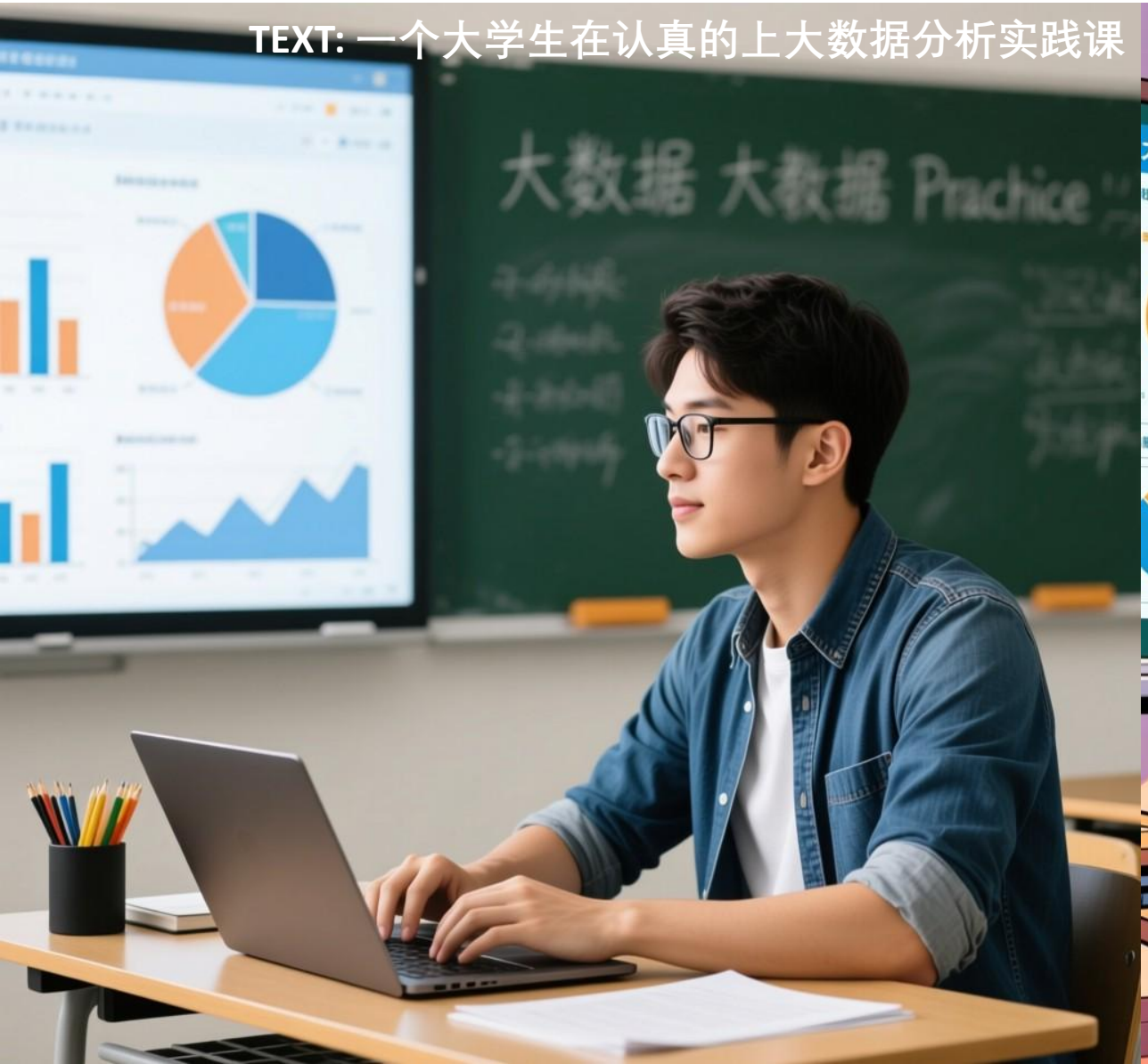


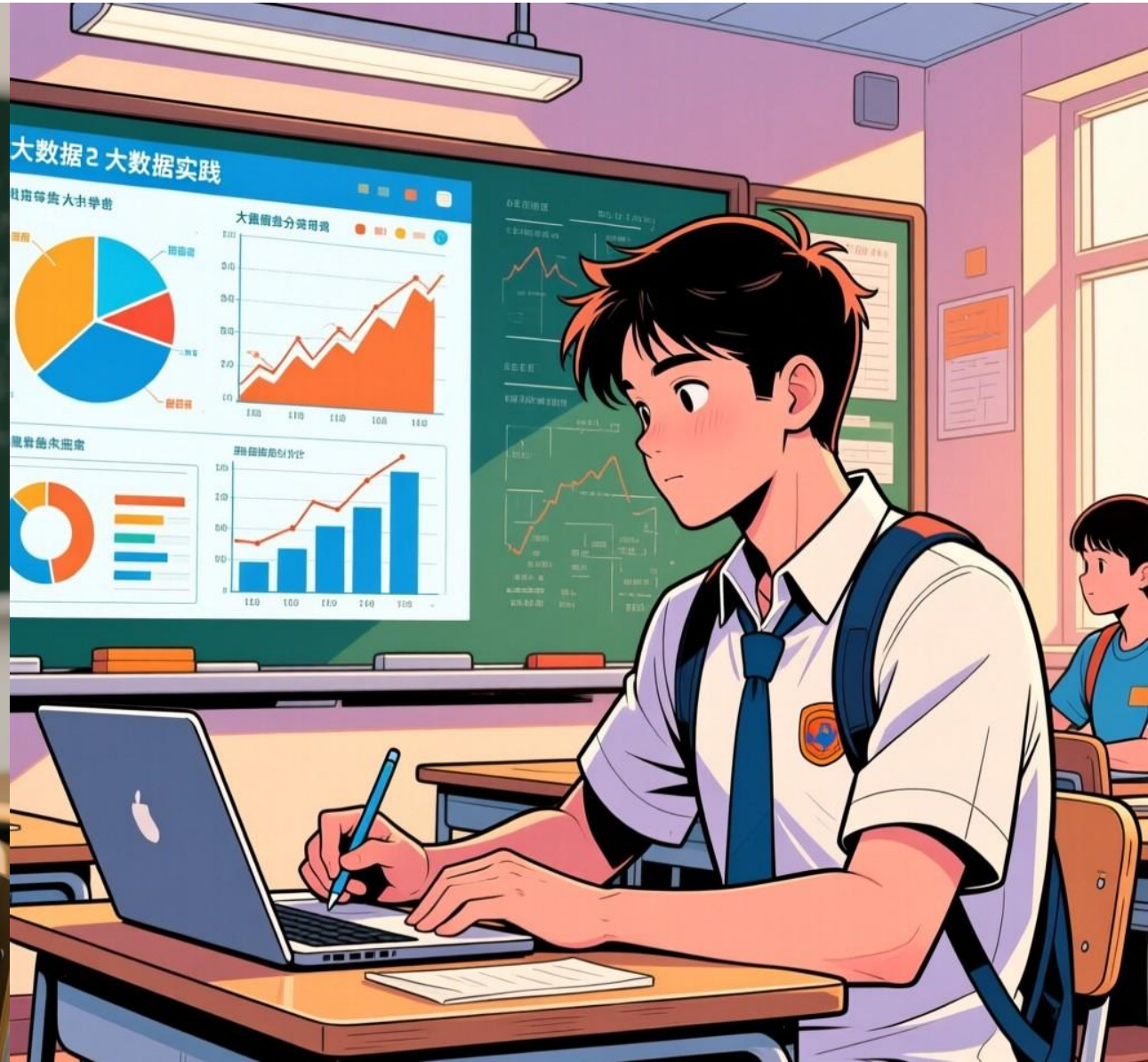
Application



Application

TEXT: 一个大学生在认真的上大数据分析实践课





Similarity & Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity & Dissimilarity Distance

The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

Standardization is necessary, if scales differ.

Minkowski Distance

Minkowski Distance is a generalization of Euclidean Distance

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Minkowski Distance

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

已知：

小明（160cm，60000g）

小王（160cm，59000g）

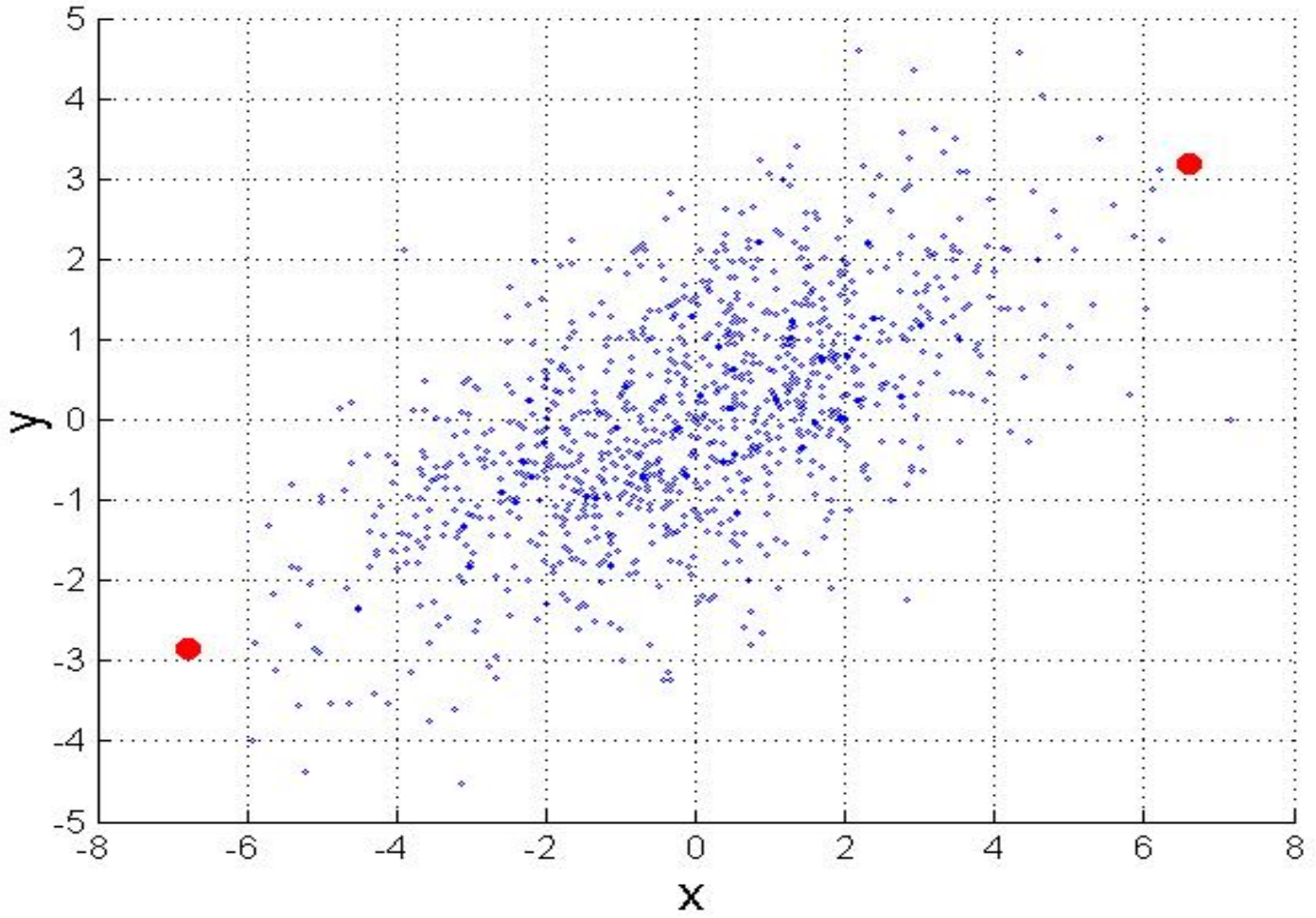
小李（170cm，60000g）

利用欧氏距离计算体型的相似程度

作答

Mahalanobis Distance

$$\mathbf{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})$$



Σ is the covariance matrix

Cosine Similarity

- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\| \mathbf{d} \|$ is the length of vector $\mathbf{d}$.

- Example:

句子A：我喜欢游泳，不喜欢画画。

句子B：我不喜欢游泳，也不喜欢画画。

	我	喜欢	游泳	不	画画	也
$\mathbf{d}_1$ =	1	2	1	1	1	0
$\mathbf{d}_2$ =	1	2	1	2	1	1

$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.9186$

思考：

这种计算方式存在哪些问题？

Cosine Similarity 计算方式存在什么问题?

Example:

句子A：我喜欢游泳，不喜欢画画。
句子B：我不喜欢游泳，也不喜欢画画。

	我	喜欢	游泳	不	画画	也
$\mathbf{d}_1$ =	1	2	1	1	1	0
$\mathbf{d}_2$ =	1	2	1	2	1	1

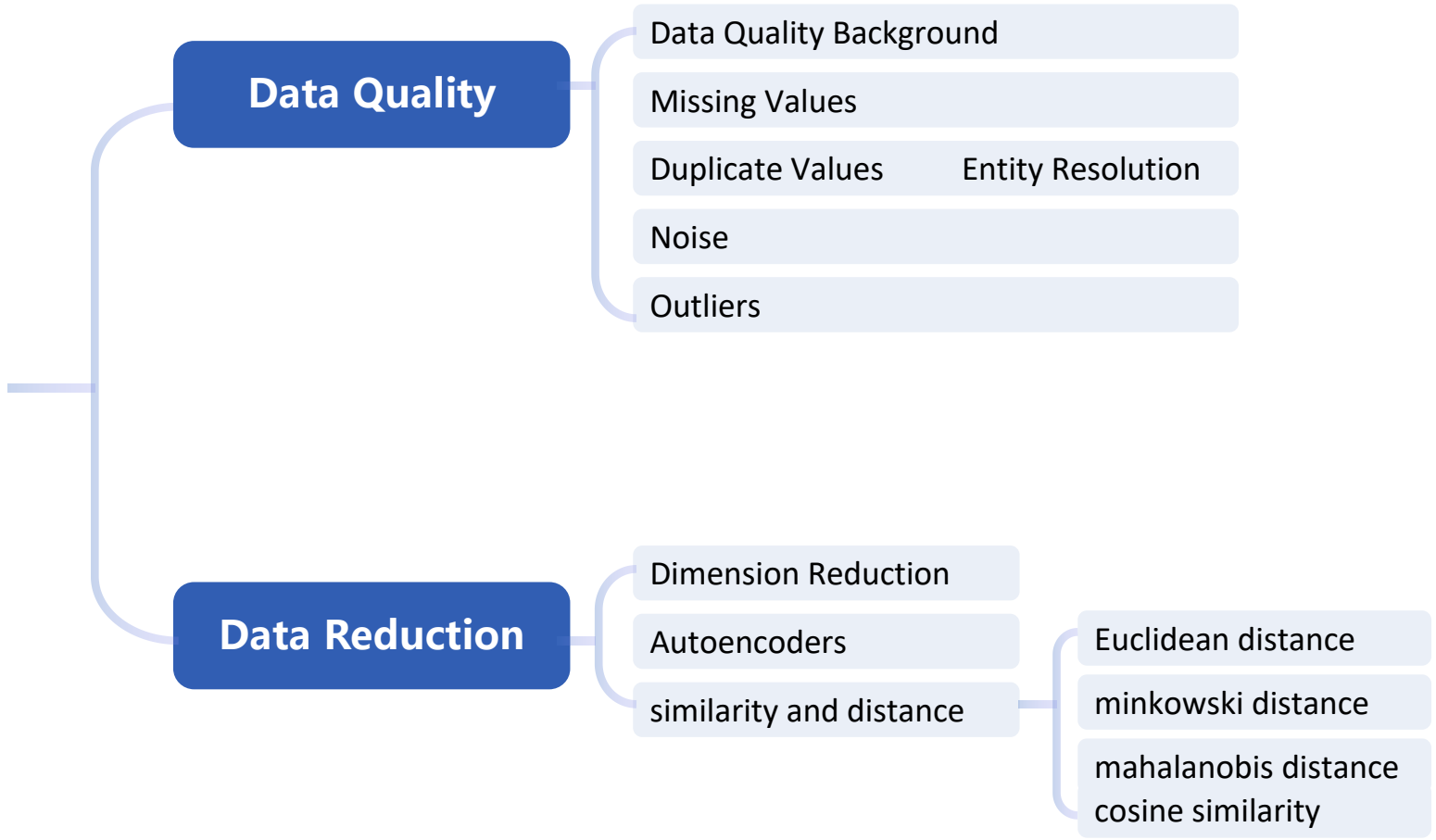
$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.9186$$

作答

Week 4 – Personal Assignments

- Pokemon dataset: 721 Pokemon, including their number, name, first and second type, and basic stats: HP, Attack, Defense, Special Attack, Special Defense, and Speed
- Run an existing data quality tool or Pandas
- Find out data quality issues as many as possible, and summarize your findings

Exploratory DataAnalysis II



Thank You