sd01332210
Practices on Big Data Analytics

大数据分析实践

Asking Questions / Data / Data Sampling

Qiong Zeng (曾琼)
qiong.zn@sdu.edu.cn

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

课程导入：学生身体素质调查

某市的教育部门为了解该地区学生的整体身体状况，决定开展中小学生身体素质调查。假设你是一名数据分析师：

课程导入：美国大选民意调查

1948年美国总统大选中，多数民调机构预测杜威会击败杜鲁门，结果是杜鲁门胜了。

2016年，多数民意调查预测希拉里会赢得选举，但最终特朗普却获得了更多的选票。

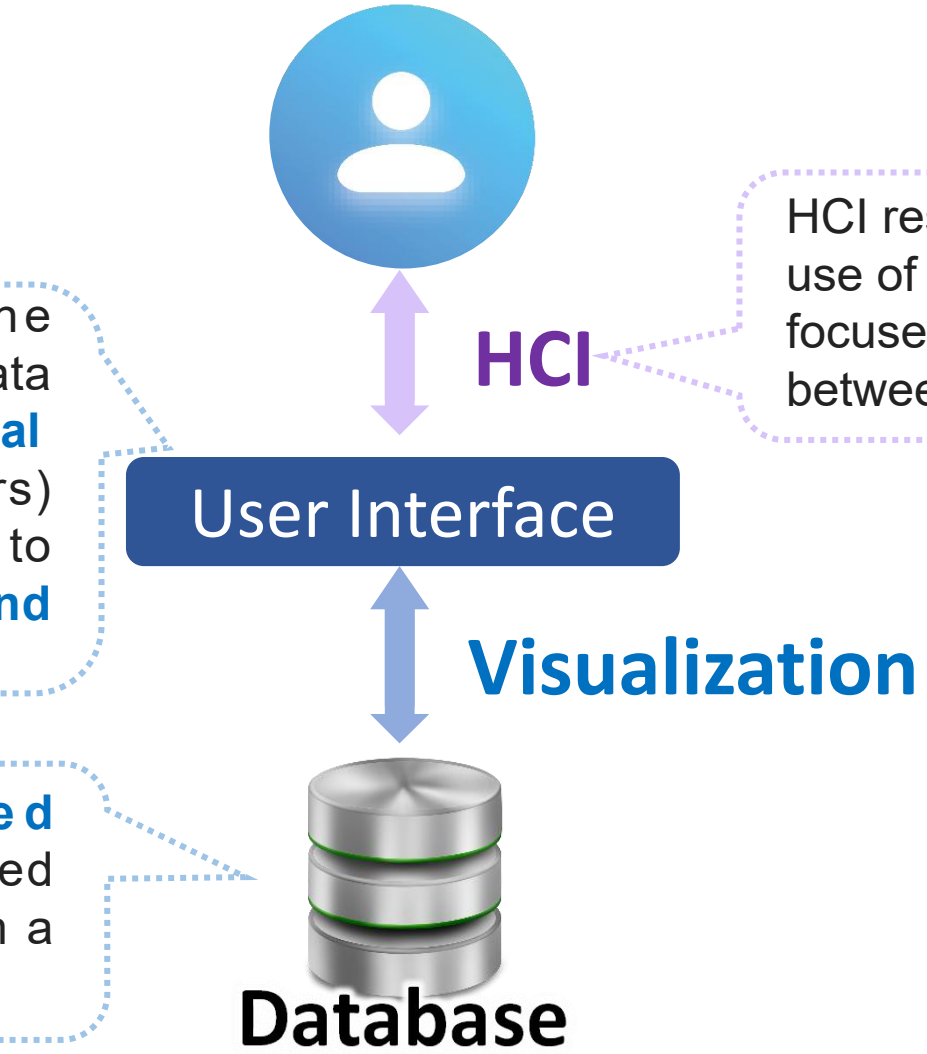
采样不当导致发生偏差

学习目标

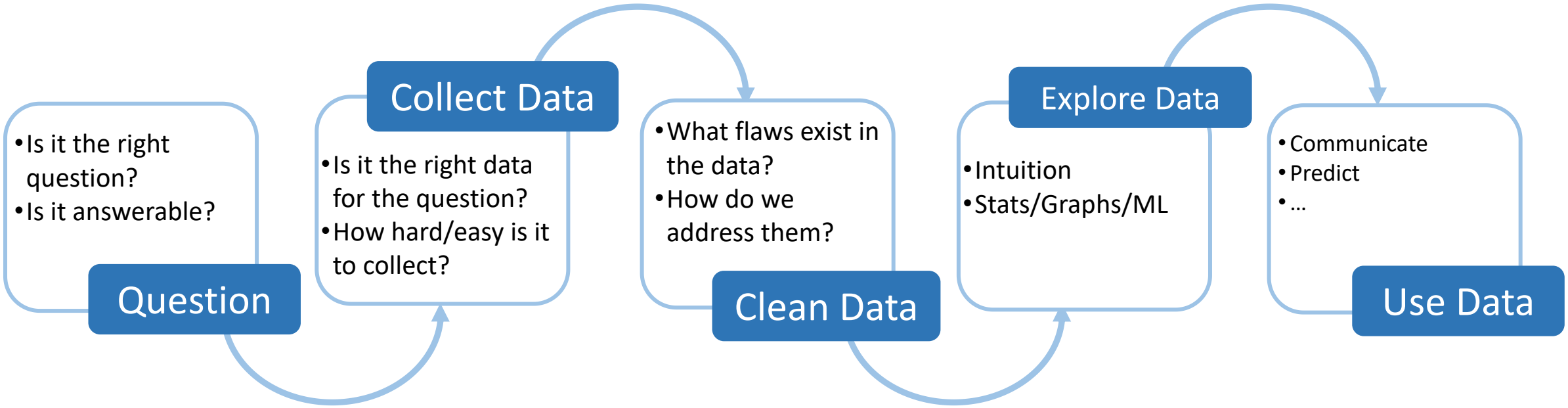
- 能够辨别questions对于分析某个数据集的作用以及如何影响后面的分析过程
- 了解数据及其属性的类型
- 能够根据数据和分析任务选择合适的数据采样方法
- 理解在数据归一化对于后续数据处理的重要

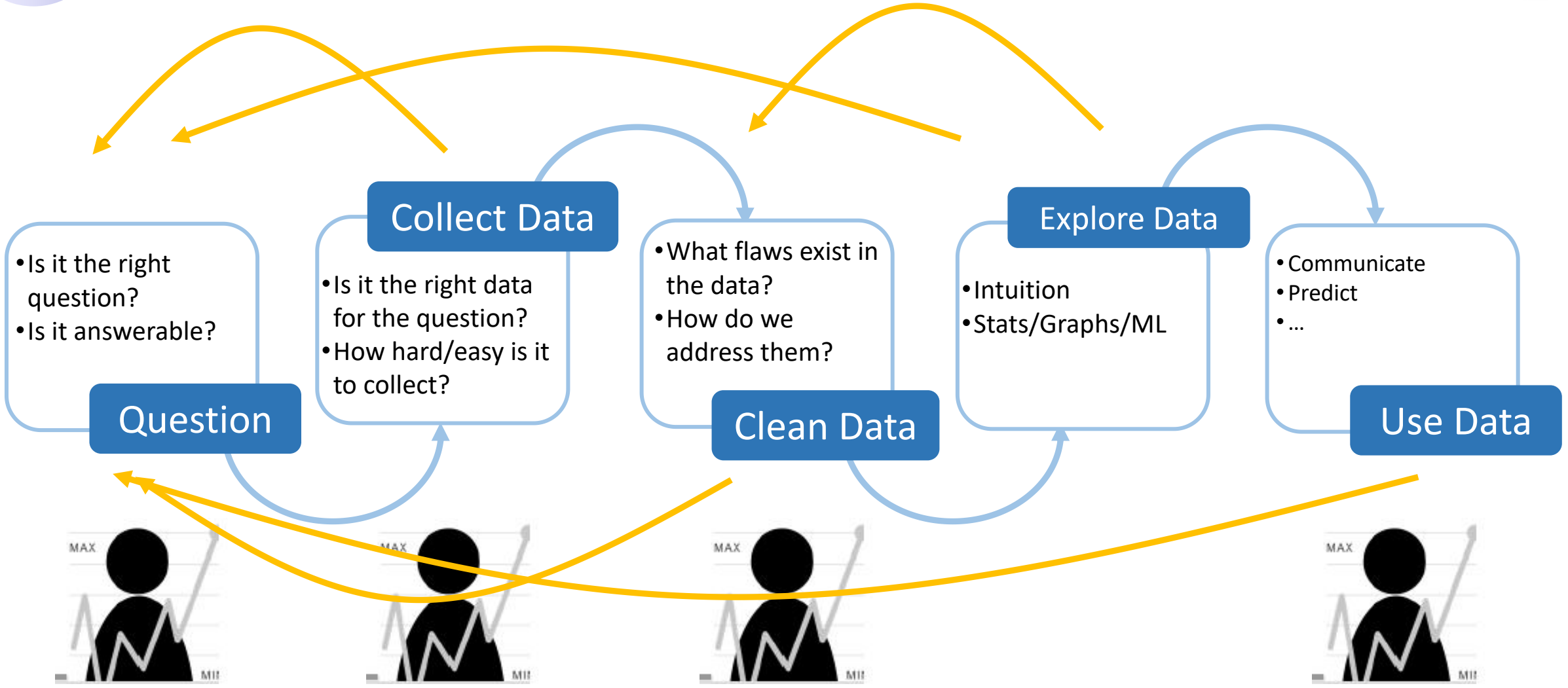
Data visualization refers to the techniques used to communicate data or information by encoding it as **visual objects** (e.g., points, lines or bars) contained in graphics. The goal is to communicate information **clearly and efficiently** to users.

A database is an **organized collection** of data generally stored and accessed electronically from a computer system.



HCI researches the design and use of computer technology, focused on the **interfaces** between people and computers.


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
•…

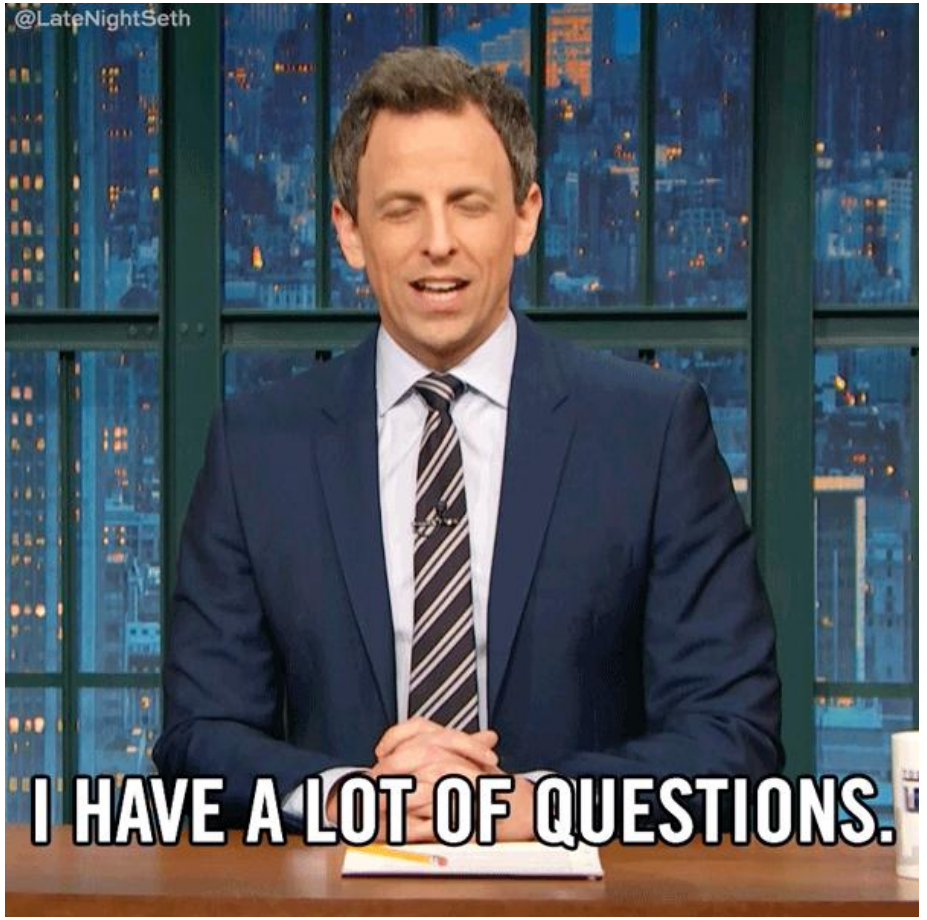

Question
•Is it the right question?
•Is it answerable?
Collect Data
•Is it the right data for the question?
•How hard/easy is it to collect?
Clean Data
•What flaws exist in the data?
•How do we address them?
Explore Data
•Intuition
•Stats/Graphs/ML
Use Data
•Communicate
•Predict
•…
MAX
MIN

Outline

Asking Questions

Data

Data Sampling



投票(匿名)

最多可选1项

设置

Which is the best question for data analysis?

A *What can my data tell me about my business?*

B *What should I do?*

C *How can I increase my profits?*

D *How many eggs will I sell in Qingdao during the third quarter?*

提交

What is a good question?

- A good question has an answer
- Should be interesting, provides information
- Should be at the right granularity

Predicting

What is going to win the election?

Explaining

What blocks are voting for each candidate?

What areas have most undecided voters?

What ideas are resonating with which voters?

主观题 10分

设置

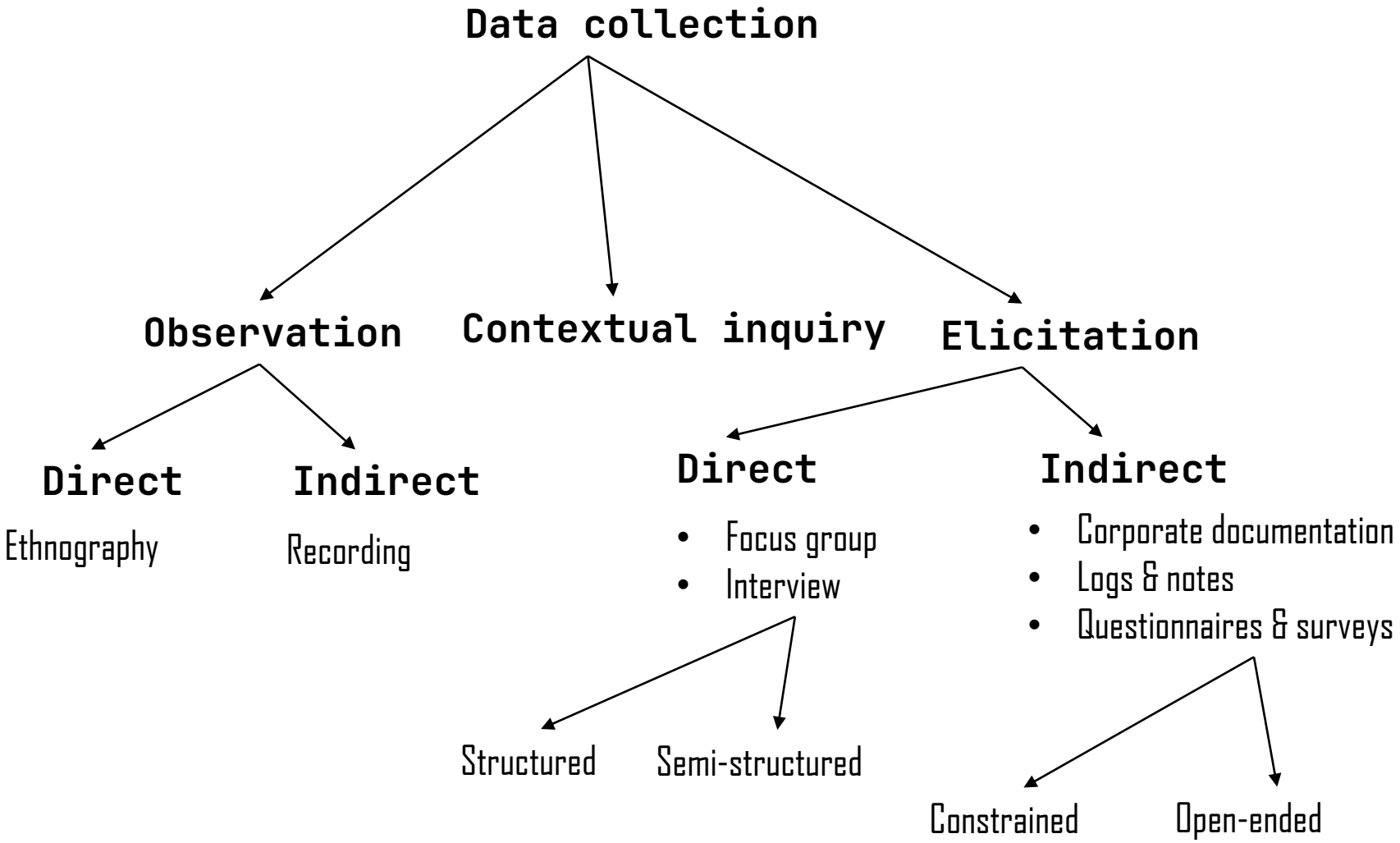
Given driver/owner, pickup/dropoff location, and fare data for every taxi trip taken, can you pose an interesting question?

4													
5	Trip data, 2013 ->												
6													
7	medallion	hack_license	vendor_id	rate_code	pickup_datetime	dropoff_datetim	passenger_c	trip_time_	trip_distance	pickup_longitud	pickup_latitude	dropoff_longitude	dropoff_latitude
8	89D227B655E5C82AEC	BA96DE419E7116	CMT	1	1/1/13 15:11	1/1/13 15:18	4	382	1	-73.978165	40.757977	-73.989838	40.751171
9	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1	1/6/13 0:18	1/6/13 0:22	1	259	1.5	-74.006683	40.731781	-73.994499	40.75066
10	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1	1/5/13 18:49	1/5/13 18:54	1	282	1.1	-74.004707	40.73777	-74.009834	40.726002
11	...												
12													
13													
14	Fare data, 2013 ->												
15													
16	medallion	hack_license	vendor_id	pickup_datetime	fare_amount	surcharge	mta_tax	tip_amou	tolls_amount	total_amount			
17	89D227B655E5C82AEC	BA96DE419E7116	CMT	1/1/13 15:11	6.5	0	0.5	0	0	7			
18	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1/6/13 0:18	6	0.5	0.5	0	0	7			
19	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1/5/13 18:49	5.5	1	0.5	0	0	7			

正常使用主观题需2.0以上版本雨课堂

作答

Data Collection



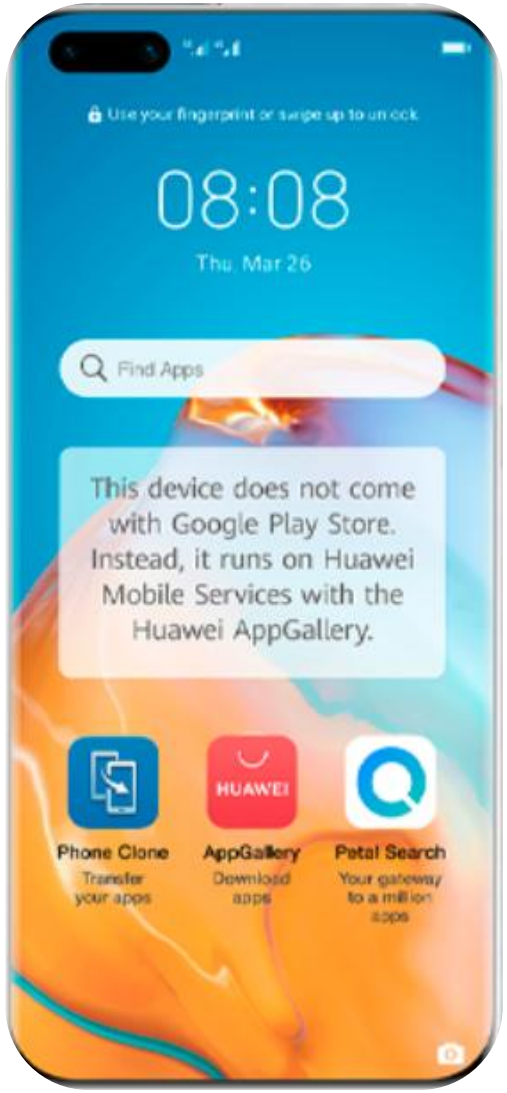
ProactiveTasks: the Short of Mobile Device Use Sessions

Nikola Banovic, Christina Brant, Jennifer Mankoff, Anind K. Dey
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
{nbanovic, jmankoff, anind}@cs.cmu.edu, cbrant@andrew.cmu.edu

Mobile Data

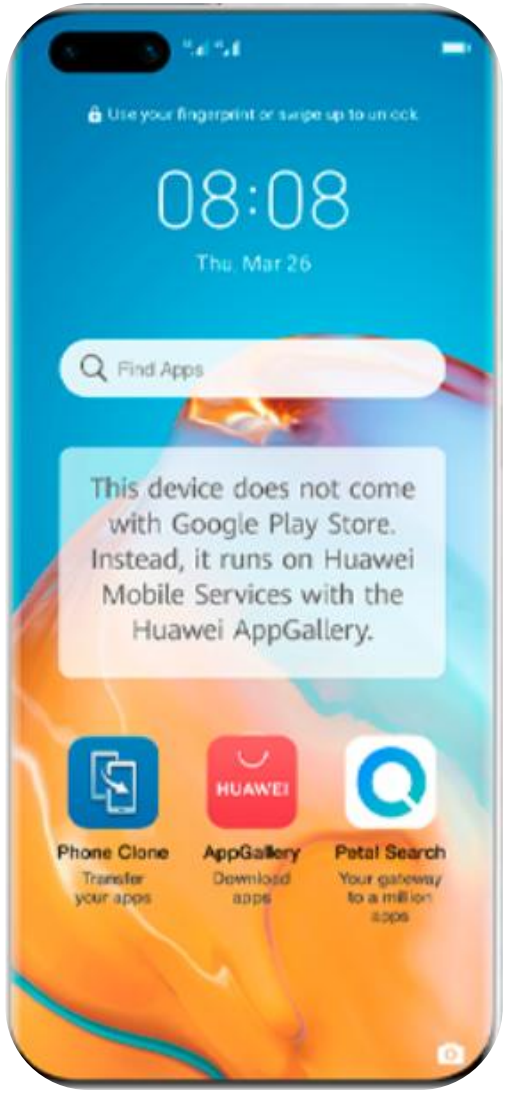
People carry phones everyday, and plenty of information can be collected:

- Interactions with virtual information
- Social engagement
- Loads of sensors about physical actions



Mobile Data

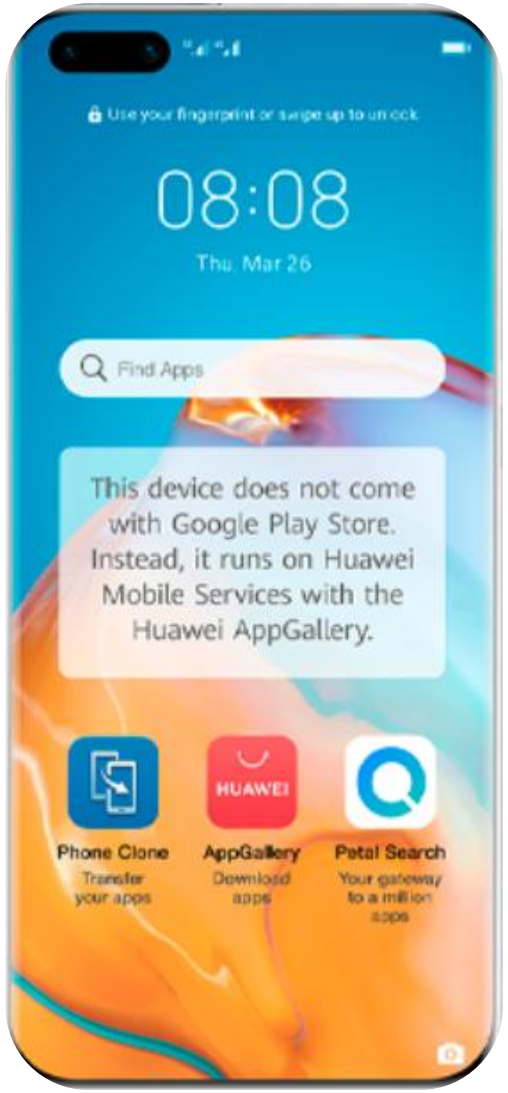
- Accelerometer
- Applications
- Battery
- Bluetooth
- Calls
- Messaging
- Gravity
- Gyroscope
- Light
- Linear Accelerometer
- Location
- Magnetometer
- Network Usage
- Orientation
- Pressure
- Processor
- Proximity
- Rotation
- Screen
- Telephony
- Temperature
- Traffic
- Wi-Fi
- ... many more



Build compelling and useful apps that provide value in everyday life and compelling user experience

Mobile Usage

What do we know about how people use their mobile devices?



Iterative Process

Questions

How much users use their phones?

How do users interact with their phones?

Can we predict when users use their phones?

Data Sets

Analysis & etc (not today's focus)

New Questions

[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions

How much users use their phones?

How do users interact with their phones?

Can we predict when users use their phones?

Data Sets

A formative study with 10 participants for 1 month

Analysis & etc (not today's focus)

New Questions

[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions

How much users use their phones?

How do users interact with their phones?

Can we predict when users use their phones?

_id	timestamp	device_id	double_latitude	double_longitude	do	dou...	dou...	provider	accuracy	label	state
147	1389022525388	1fc44d2d-a832-420d-b249-24e	40.4436079	-79.9455869		0	0	fused	51.1	LOC_87	arrived
148	1389022533705	1fc44d2d-a832-420d-b249-24e	40.443612	-79.9455925		0	0	fused	39.9	LOC_87	arrived
149	1389022537590	1fc44d2d-a832-420d-b249-24e	40.4436122	-79.9455924		0	0	fused	33.9	LOC_87	arrived
150	1389022581525	1fc44d2d-a832-420d-b249-24e	40.4435916	-79.9455911		0	0	fused	45.4	LOC_87	arrived
151	1389022609907	1fc44d2d-a832-420d-b249-24e	40.4435916	-79.9455911		0	0	fused	45.4	LOC_87	arrived
152	1389022646266	1fc44d2d-a832-420d-b249-24e	40.4436045	-79.9455827		0	0	fused	44.2	LOC_87	arrived
153	1389022706306	1fc44d2d-a832-420d-b249-24e	40.443592	-79.9455864		0	0	fused	45.1	LOC_87	arrived
154	1389022776324	1fc44d2d-a832-420d-b249-24e	40.4436054	-79.945588		0	0	fused	44.4	LOC_87	arrived
155	1389022836102	1fc44d2d-a832-420d-b249-24e	40.4435869	-79.9455903		0	0	fused	45.6	LOC_87	visiting
156	1389022899878	1fc44d2d-a832-420d-b249-24e	40.443621	-79.9455951		0	0	fused	44.0	LOC_87	visiting
157	1389022904756	1fc44d2d-a832-420d-b249-24e	40.4436181	-79.945601		0	0	fused	31.4	LOC_87	visiting
158	1389022965977	1fc44d2d-a832-420d-b249-24e	40.4436163	-79.9456055		0	0	fused	44.8	LOC_87	visiting
159	1389023026441	1fc44d2d-a832-420d-b249-24e	40.4436247	-79.9455487		0	0	fused	41.1	LOC_87	visiting
160	1389023086134	1fc44d2d-a832-420d-b249-24e	40.4436078	-79.9455671		0	0	fused	43.1	LOC_87	visiting
161	1389023105908	1fc44d2d-a832-420d-b249-24e54dd46d44	35905	-79.9456086		0	0	fused	62	LOC_87	visiting
162	1389023111346	1fc44d2d-a832-420d-b249-24e	40.4436033	-79.9456097		0	0	fused	44.1	LOC_87	visiting
163	1389023135555	1fc44d2d-a832-420d-b249-24e	40.4436232	-79.945589		0	0	fused	43	LOC_87	visiting
164	1389023140778	1fc44d2d-a832-420d-b249-24e	40.4436129	-79.9455979		0	0	fused	35.3	LOC_87	visiting
165	1389023141684	1fc44d2d-a832-420d-b249-24e	40.4436066	-79.9456007		0	0	fused	29.1	LOC_87	visiting
166	1389023174414	1fc44d2d-a832-420d-b249-24e	40.4435837	-79.9455703		0	0	fused	64	LOC_87	visiting
167	1389023178990	1fc44d2d-a832-420d-b249-24e	40.4435959	-79.9455486		0	0	fused	19.0	LOC_87	visiting
168	1389023202868	1fc44d2d-a832-420d-b249-24e	40.4432896	-79.9451562		0	0	fused	52.2	LOC_87	visiting
169	1389023205137	1fc44d2d-a832-420d-b249-24e	40.4432896	-79.9451562		0	0	fused	52.2	LOC_87	visiting
170	1389023268446	1fc44d2d-a832-420d-b249-24e	40.4432896	-79.9451562		0	0	fused	52.2	LOC_87	visiting
171	1389023307902	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
172	1389023312886	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
173	1389023322688	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
174	1389023325474	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
175	1389023337465	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
176	1389023338878	1fc44d2d-a832-420d-b249-24e	40.4432978	-79.945201		0	0	fused	85	LOC_87	visiting
177	1389023345169	1fc44d2d-a832-420d-b249-24e	40.4437775	-79.9454877		0	0	fused	44	LOC_87	visiting
178	1389023355968	1fc44d2d-a832-420d-b249-24e	40.4437775	-79.9454877		0	0	fused	44	LOC_87	visiting
179	1389023384880	1fc44d2d-a832-420d-b249-24e	40.4439232	-79.945263		0	0	fused	71.8	LOC_87	visiting
180	1389023409327	1fc44d2d-a832-420d-b249-24e	40.4439232	-79.945263		0	0	fused	71.8	LOC_87	visiting
181	1389023410739	1fc44d2d-a832-420d-b249-24e	40.4444575	-79.9451064		0	0	fused	30.4	LOC_87	visiting
182	1389023414851	1fc44d2d-a832-420d-b249-24e	40.4444585	-79.9451024		0	0	fused	22.2	LOC_87	visiting
183	1389023424216	1fc44d2d-a832-420d-b249-24e	40.4444398	-79.9451177		0	0	fused	21.4	LOC_87	visiting
184	1389023429401	1fc44d2d-a832-420d-b249-24e	40.4444398	-79.9451177		0	0	fused	26.1	LOC_87	visiting
185	1389023434818	1fc44d2d-a832-420d-b249-24e	40.4444398	-79.9451177		0	0	fused	30.1	LOC_87	visiting
186	1389023440237	1fc44d2d-a832-420d-b249-24e				0	0	fused	28.2		

[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions

How much users use their phones?

How do users interact with their phones?

Can we predict when users use their phones?

Data Sets

Analysis & etc (not today's focus)

New Questions

- A usage session is defined from the time the user turns the screen on, until the time the screen is off.
- What about automatic screen timeouts?
- Simple heuristic: group all sessions that are less than 5 seconds apart into the same session.
- Can we classify which interactions belong in the same session?

[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions

How much users use their phones?

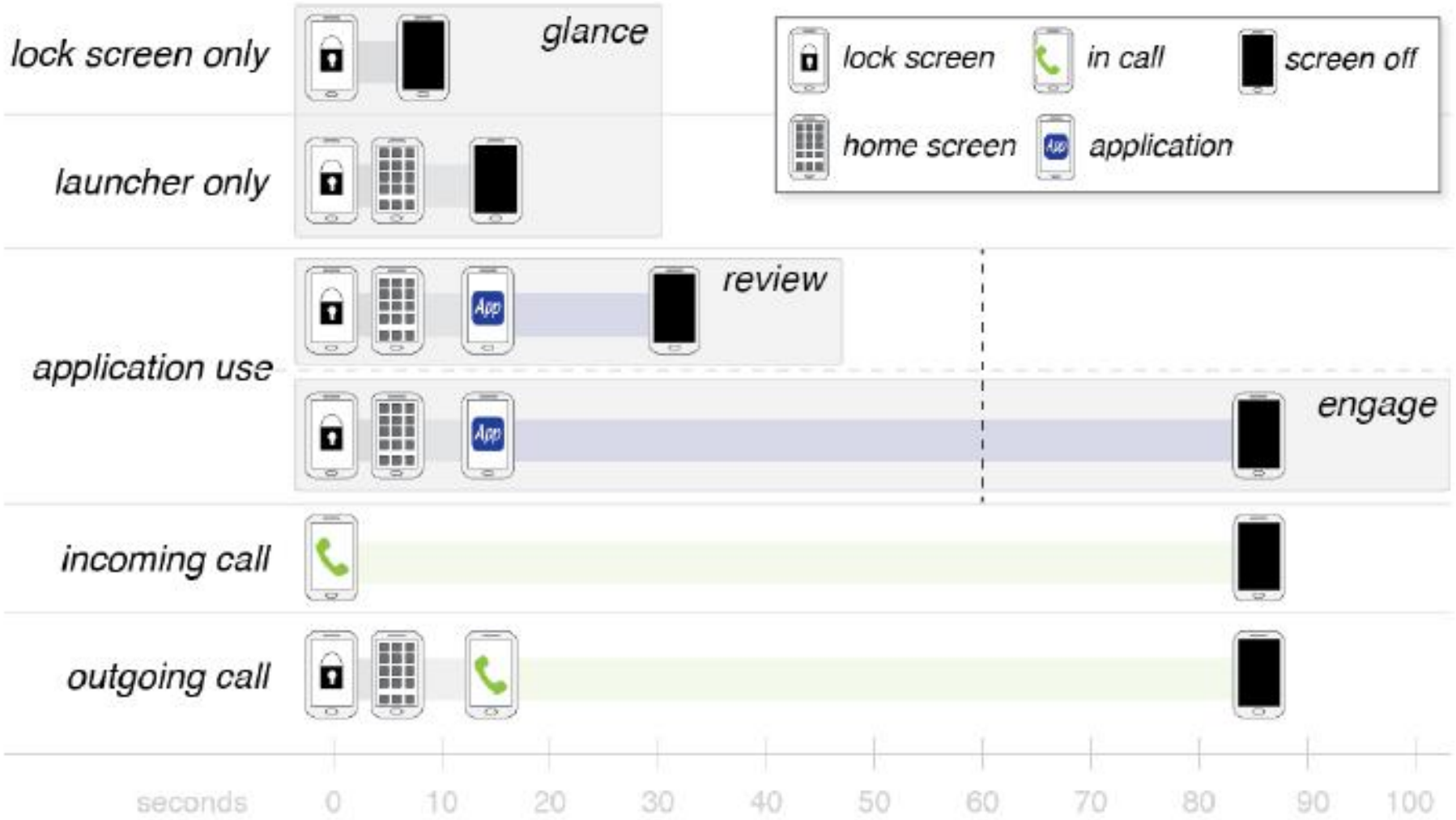
How do users interact with their phones?

Can we predict when users use their phones?

Data Sets

Analysis & etc (not today's focus)

New Questions



[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions | **Data Sets** | **Analysis & etc (not today's focus)** | **New Questions**

How much users use their phones?

How do users interact with their phones?

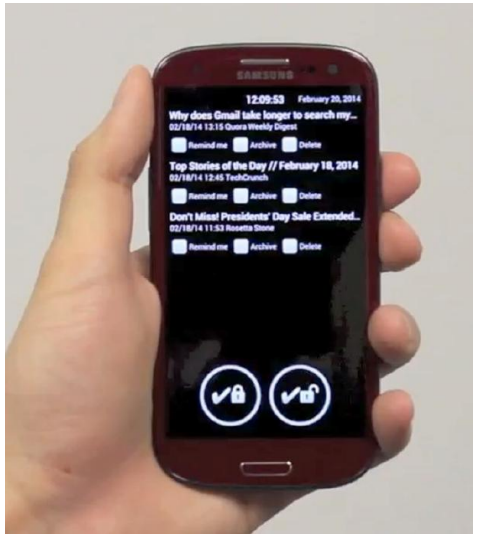
Can we predict when users use their phones?

- 95% of sessions shorter than 5 minutes; most < 60 secs
- Notifications lead to engagement... only 25% of the time!
- Self-interruption therefore more common than we would think
- Notifications prevent unnecessary engages
- No good support for reviews

[CHI14] Proactive Tasks: the Short of Mobile Device Use Sessions

Iterative Process

Questions	Data Sets	New Questions	Datasets
How much users use their phones?	A formative study with 10 participants for 1 month	How can we streamline mobile device use?	A filed user study with 30 participants for 1 month
How do users interact with their phones?			
Can we predict when users use their phones?			



Iterative Process

Questions

How much users

Data Sets

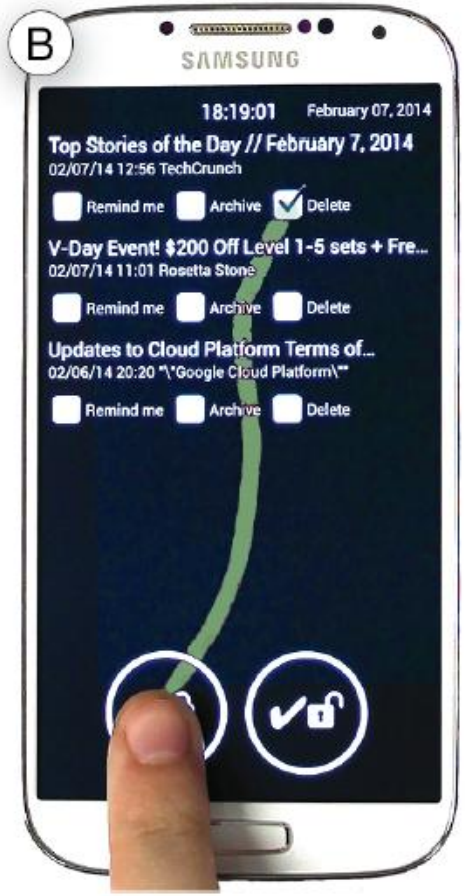
A formative study

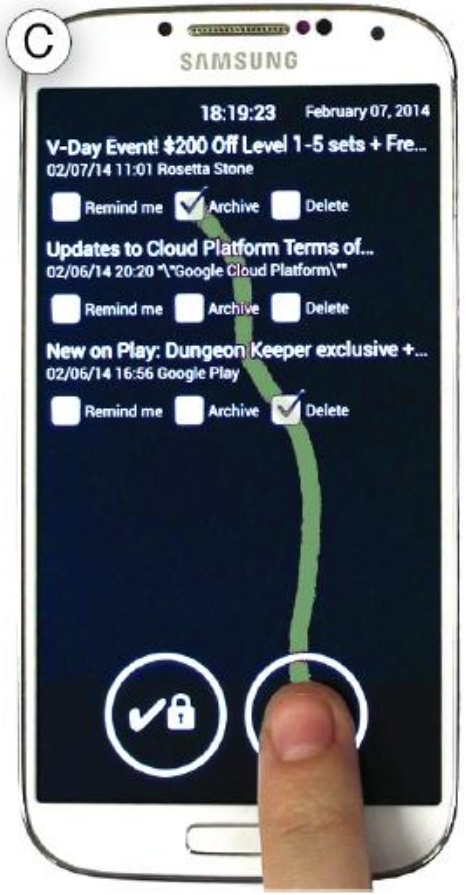
New Questions

How can we

Datasets

A filed user study





Iterative Process

Questions	Data Sets	New Questions	
How much users use their phones?	A formative study with 10 participants for 1 month	Are there different user types?	Can we predict when users are interruptible?
How do users interact with their phones?		Can we predict intent?	Can we predict next task?
How can we streamline mobile device use?	A filed user study with 30 participants for 1 month	How do users interact with wearables?	Can we model mobile usage routines?

Iterative Process

Questions	Data Sets	New Questions	
How much users use their phones?	A formative study with 10 participants for 1 month	Are there different user types?	Can we predict when users are interruptible?
How do users interact with their phones?		Can we predict intent?	Can we predict next task?
How can we streamline mobile device use?	A filed user study with 30 participants for 1 month	How do users interact with wearables?	Can we model mobile usage routines?

主观题 10分

设置

You are a data analyst for a professional basketball team. You have been asked which players you should draft.

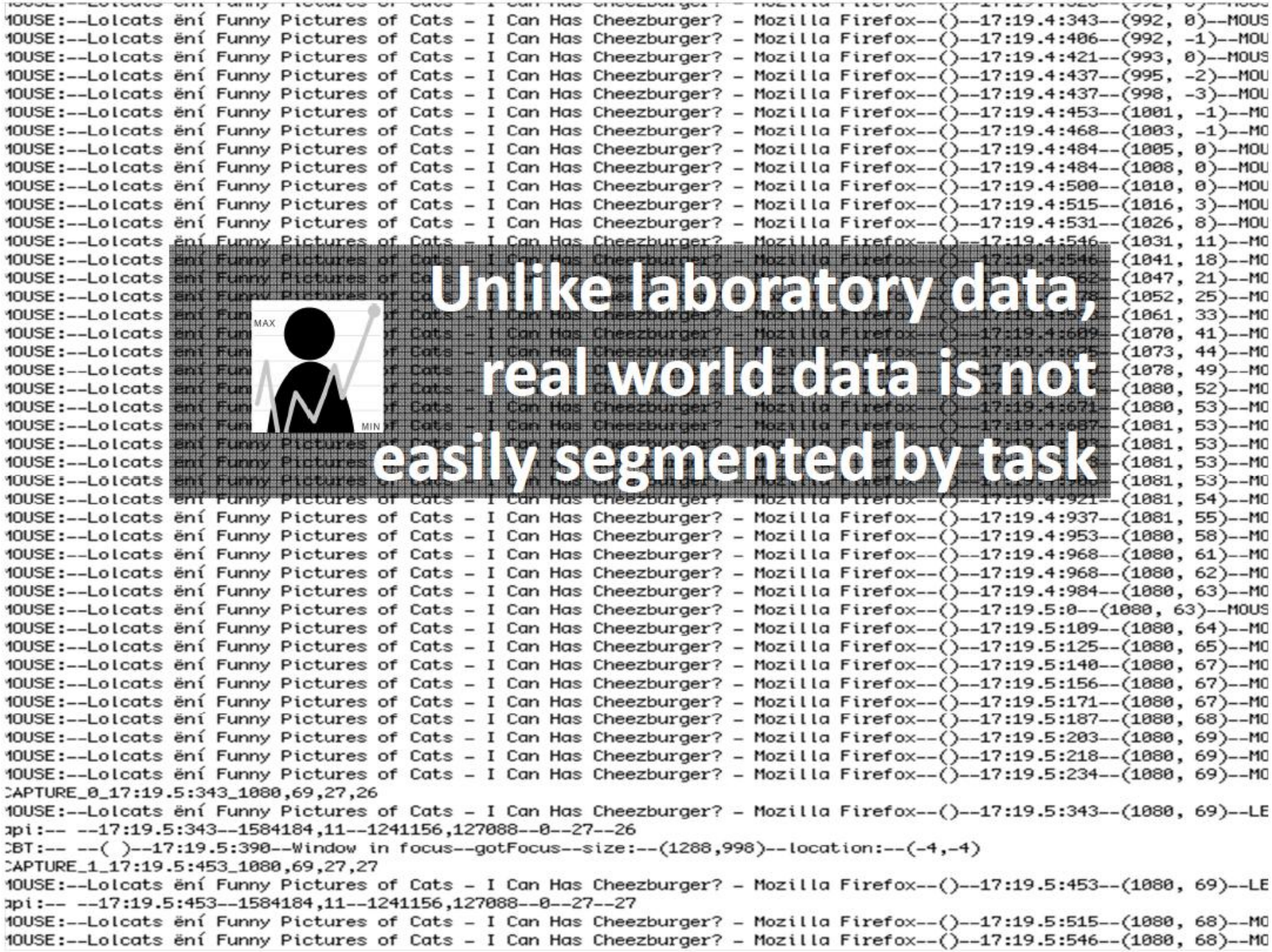
- What questions would you need to decide the players?
- What data would you need to answer the question?

以小组为单位讨论，每个小组上传一个回答

10 min

正常使用主观题需2.0以上版本雨课堂

作答


Unlike laboratory data,
real world data is not
easily segmented by task
MAX
MIN

Outline

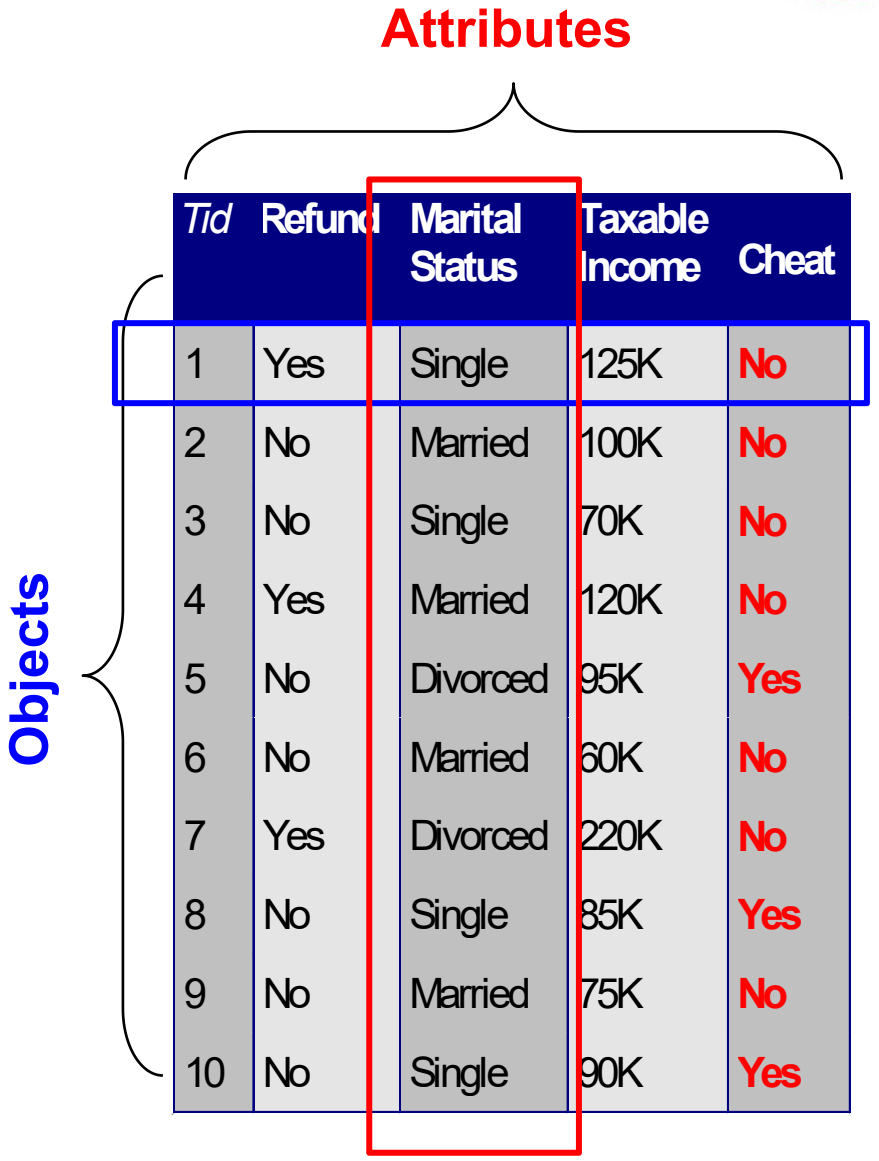
Asking Questions

Data

Data Sampling

What is Data?

- Collection of ***data objects*** and their ***attributes***
- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

Measurement of Length

The way you measure an attribute may not match the attributes properties.

This scale preserves only the ordering property of length.

This scale preserves the ordering and additvity properties of length.

Types of Attributes

There are different types of attributes

- Nominal（标称）：A type of **data** that is used to label variables without providing any quantitative value
- Ordinal（序数）：A categorical, statistical **data** type where the variables have natural, ordered categories and the distances between the categories is not known
- Interval（区间）：Measured along a scale, in which each point is placed at equal distance from one another
- Ratio（比率）：A quantitative **data**, having the same properties as interval **data**, with an equal and definitive **ratio** between each **data** and absolute “zero”

Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Types of Data Sets

Record

- Data Matrix
- Document Data
- Transaction Data

Graph

- World Wide Web
- Molecular Structures

Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same **fixed set of numeric attributes**, then the data objects can be thought of as points in a multi-dimensional space, where **each dimension represents a distinct attribute**

- Such data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

Each document becomes a ‘term’ vector

- Each term is a component (attribute) of the vector
- The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

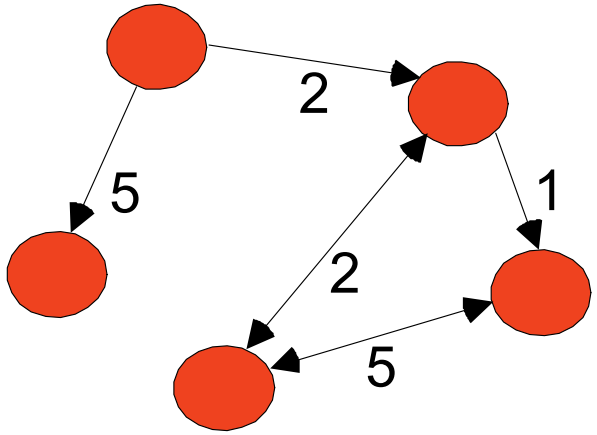
A special type of record data, where

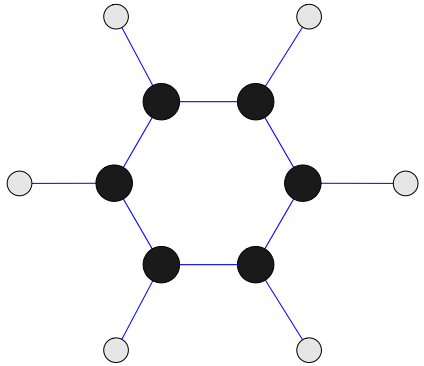
- Each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

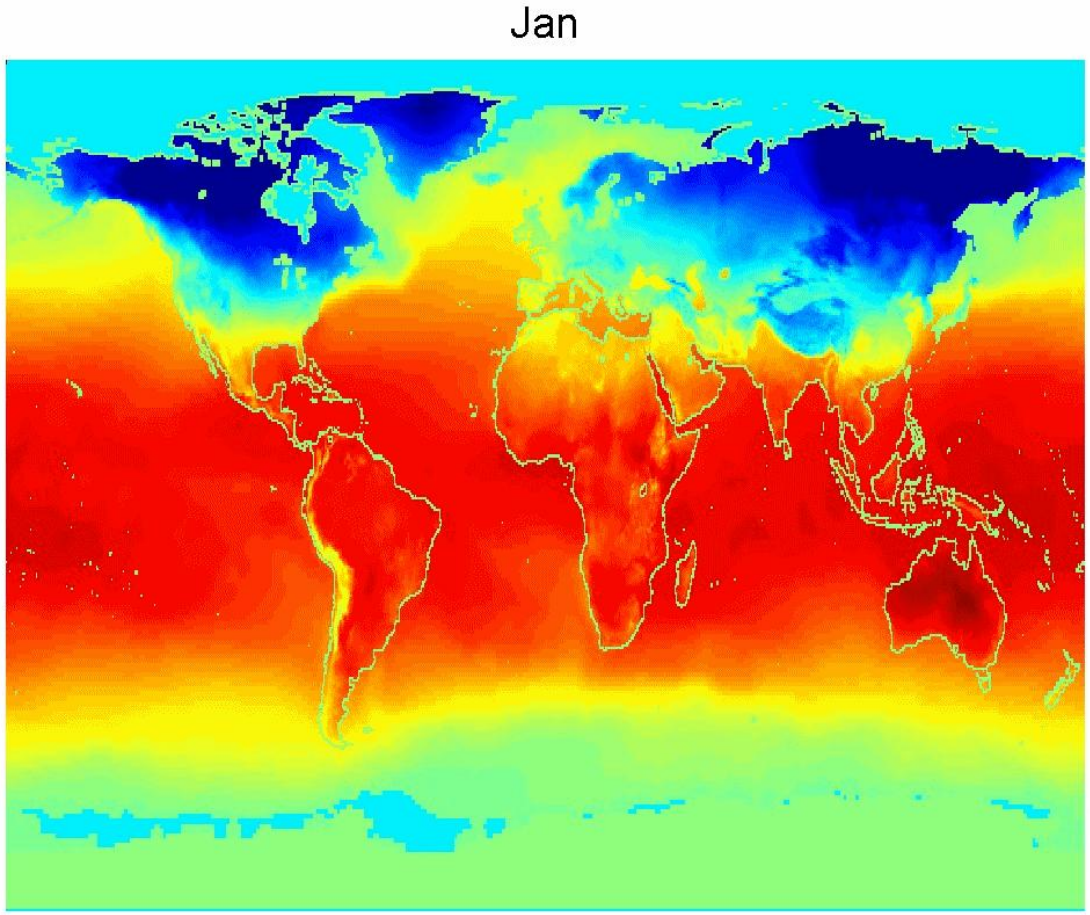
Ordered Data

Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Temporal Data

Average Monthly Temperature of land and ocean



Type of data

- **Unstructured, *e.g.*, text, html, images**
- **Semi-Structured, *e.g.*, XML, json, email**
- **Structured, *e.g.*, a MySQL database**

```
<CATALOG>
<CD>
<TITLE>Empire Burlesque</TITLE>
<ARTIST>Bob Dylan</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>Columbia</COMPANY>
<PRICE>10.90</PRICE>
<YEAR>1985</YEAR>
</CD>
<CD>
<TITLE>Hide your heart</TITLE>
<ARTIST>Bonnie Tyler</ARTIST>
<COUNTRY>UK</COUNTRY>
<COMPANY>CBS Records</COMPAN
<PRICE>9.90</PRICE>
<YEAR>1988</YEAR>
</CD>
<CD>
<TITLE>Greatest Hits</TITLE>
<ARTIST>Dolly Parton</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>RCA</COMPANY>
<PRICE>9.90</PRICE>
<YEAR>1982</YEAR>
</CD>
```

13

9/13/2018

What is a data model?

- What data is stored and how it is organized

Key benefits:

- Leverage (big impact on programming; cost; etc)
- Conciseness (implies an interaction model)
- Can encode integrity constraints

Relational Data Model

- Represent data as a **table** (or *relation*)
- Each **row** (or *tuple*) represents a record
- Each record is a fixed-length tuple
- Each **column** (or *field*) represents a variable
- Each field has a *name* and a *data type*
- A table's **schema** is the set of names and types
- A **database** is a collection of tables (relations)

Relational Algebra [Codd' 70] / SQL

Operations on Data Tables: table(s) in, table out

Projection (`select`): select a set of columns

Selection (`where`): filter rows

Sorting (`order by`): order records

Aggregation (`group by`, `sum`, `min`, `max`, ...):

partition rows into groups + summarize

Combination (`join`, `union`, ...):

integrate data from multiple tables

Relational Algebra [Codd' 70] / SQL

Projection (select): select a set of columns

select day, stock

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT

Relational Algebra [Codd' 70] / SQL

Sorting (order by): order records

select * order by stock

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45
10/3	MSFT	74.26
10/4	MSFT	74.69

Relational Algebra [Codd' 70] / SQL

Selection (where): filter rows

select * where price > 100

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45

Relational Algebra [Codd' 70] / SQL

Aggregation (group by, sum, min, max, …):

select stock, min(price) group by stock

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

stock	min(price)
AMZN	965.45
MSFT	74.26

Relational Algebra [Codd' 70] / SQL

Join (join) multiple tables together

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

stock	min
AMZN	965.45
MSFT	74.26

select t.day,t.stock,t.price,a.min from table as t, aggregate as a where t.stock = a.stock

day	stock	price	min
10/3	AMZN	957.10	965.45
10/3	MSFT	74.26	74.26
10/4	AMZN	965.45	965.45
10/4	MSFT	74.69	74.26

Roll-Up and Drill-Down

Want to examine population by year and age?

Roll-up the data along the desired dimensions


```
SELECT year, age, sum(people)
FROM census
GROUP BY year, age
```

Roll-Up and Drill-Down

Want to see the breakdown by marital status?

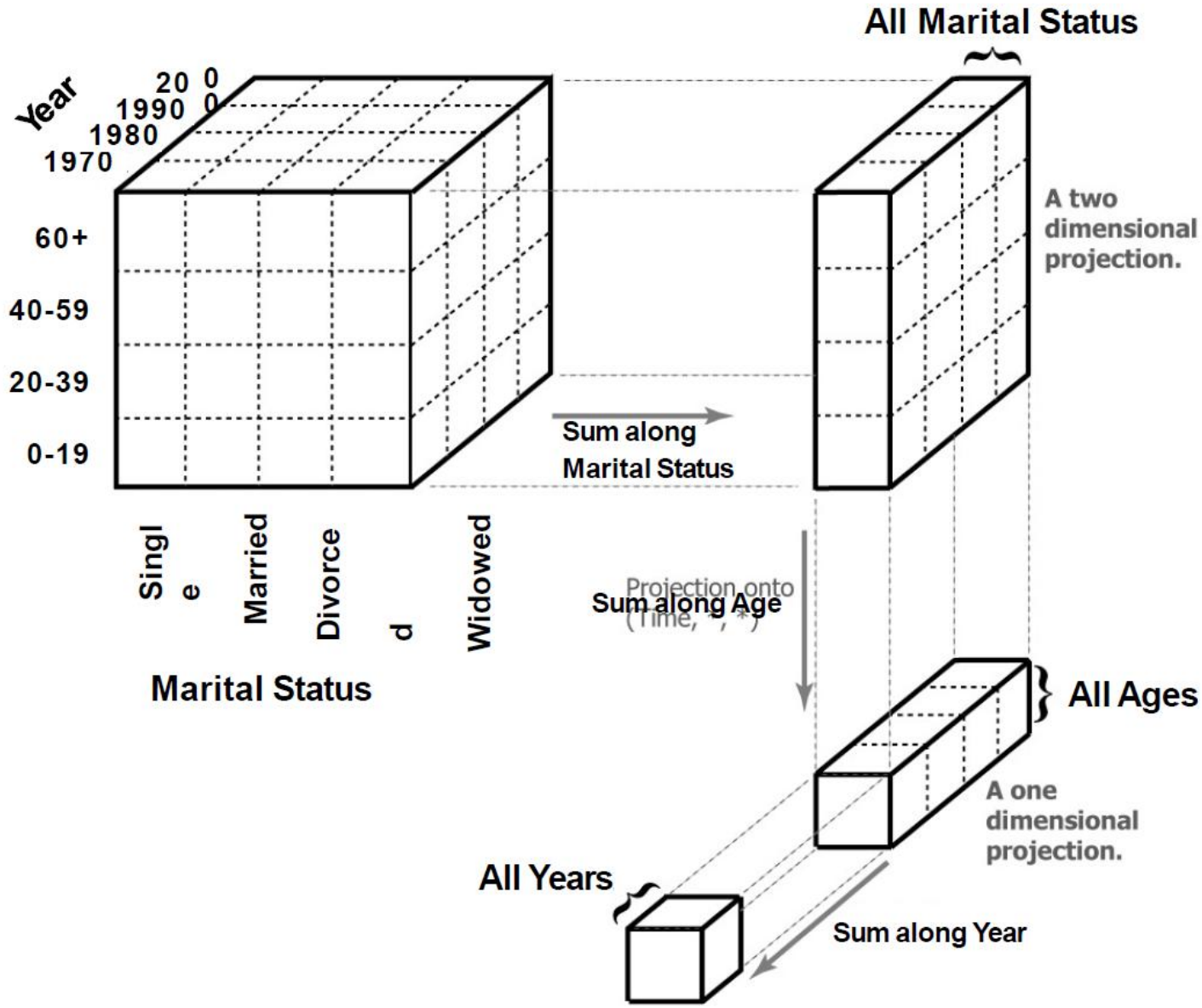
Drill-down into additional dimensions

SELECT year, age, marst, sum(people)

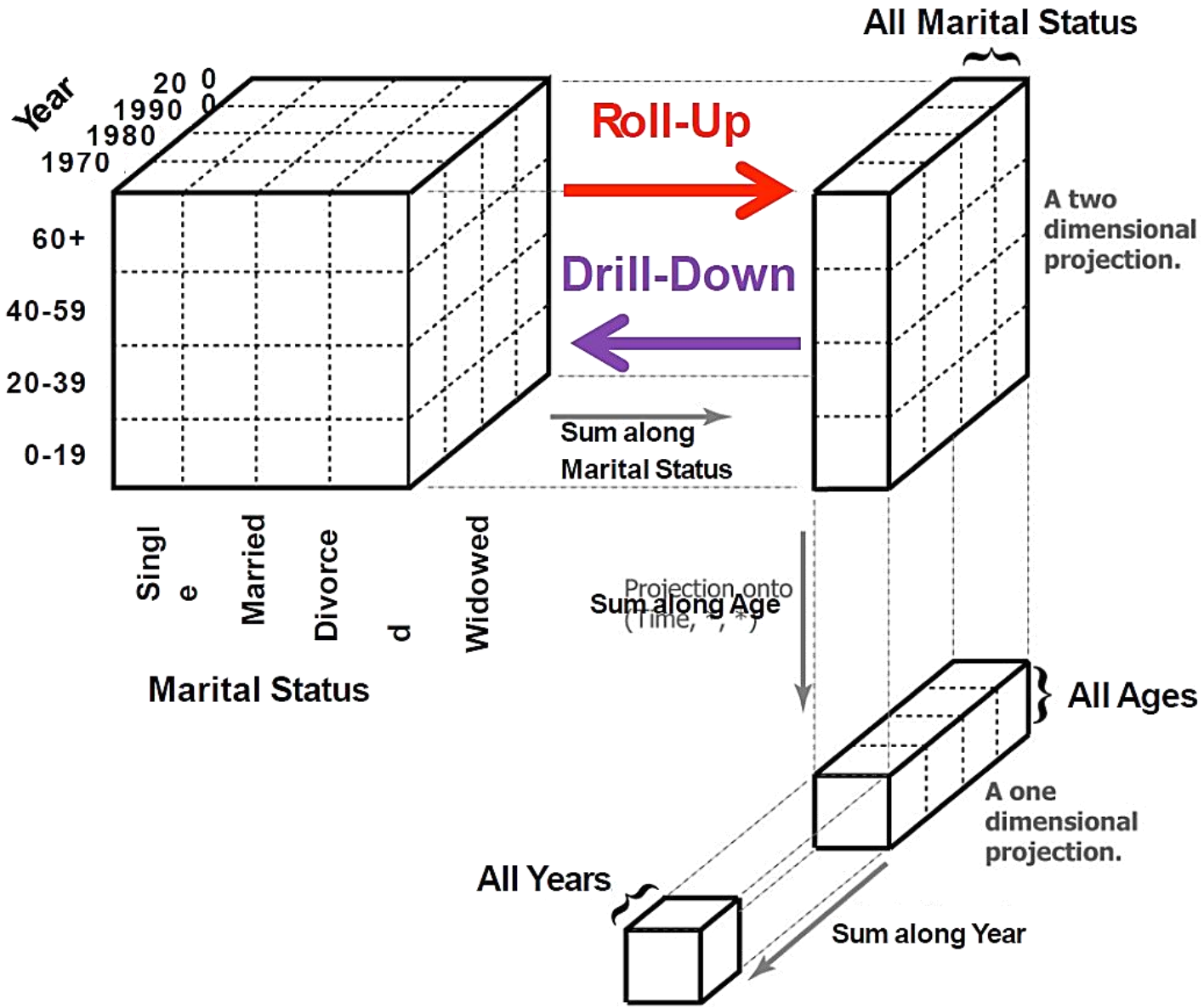
FROM census

GROUP BY year, age, marst

Roll-Up and Drill-Down



Roll-Up and Drill-Down



Outline

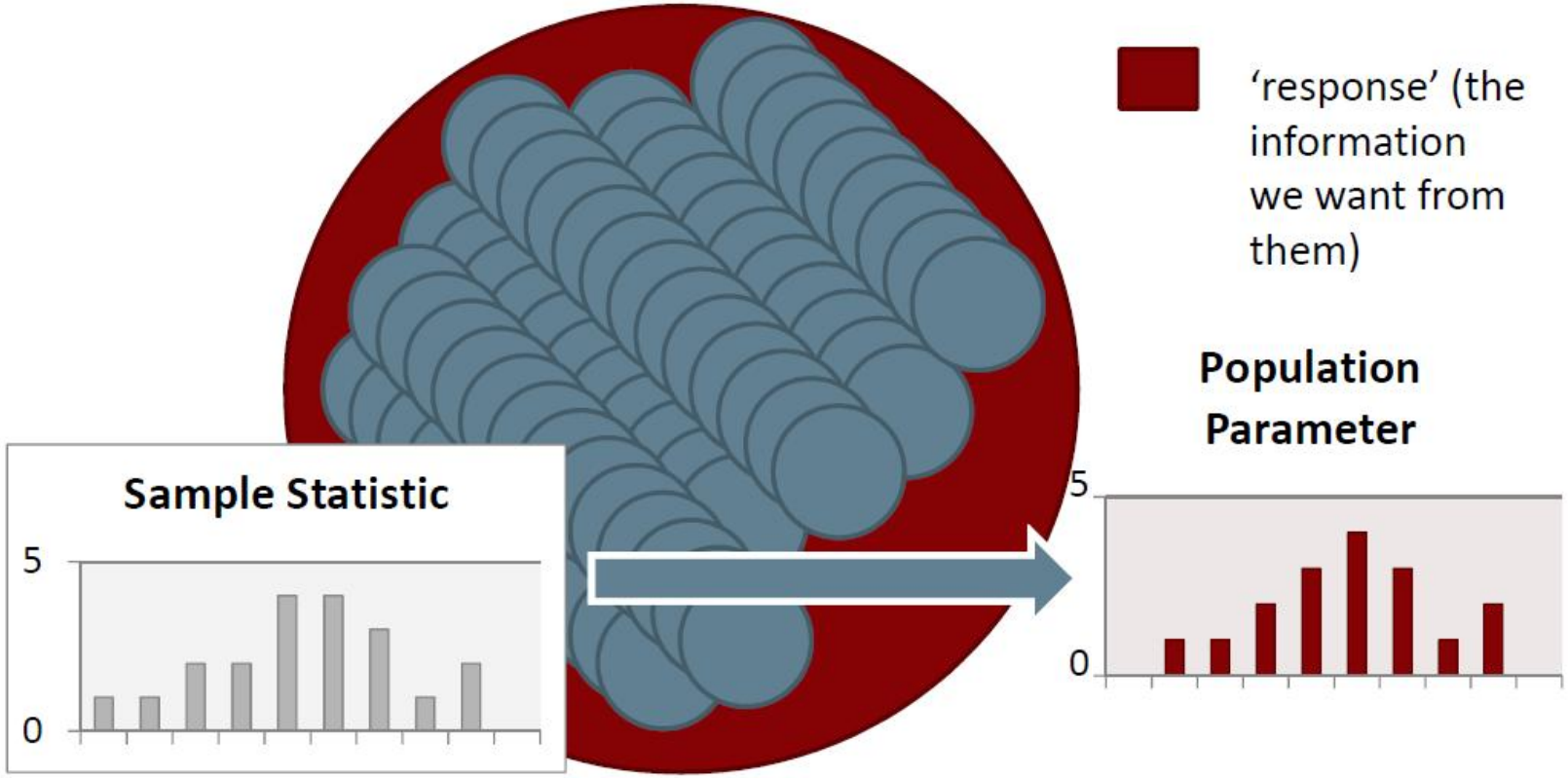
Asking Questions

Data

Data Sampling

Data Sampling

Independent Samples (infinite)



主观题 10分

设置

大数据和大数据中的抽样是否矛盾?

作答

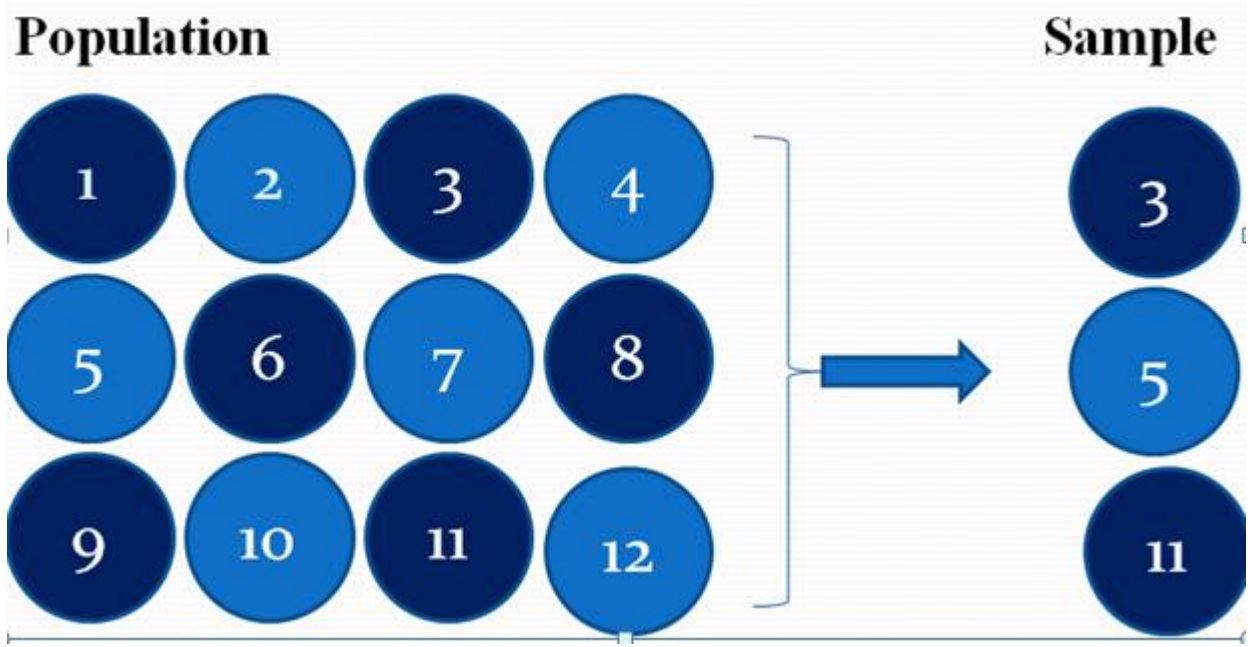
Data Sampling

- Sampling is the main technique employed for data reduction. It is often used for both the preliminary investigation of the data and the final data analysis.

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

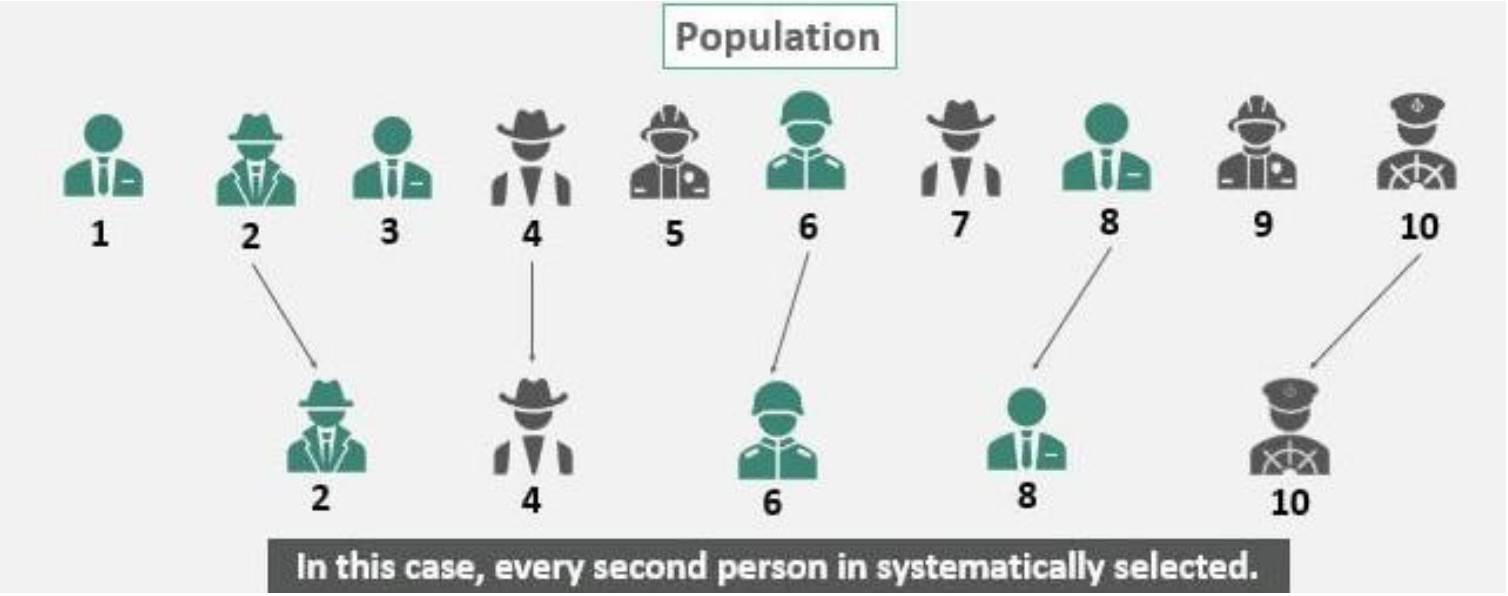
Types of Data Sampling

- Random Sampling
  - There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population



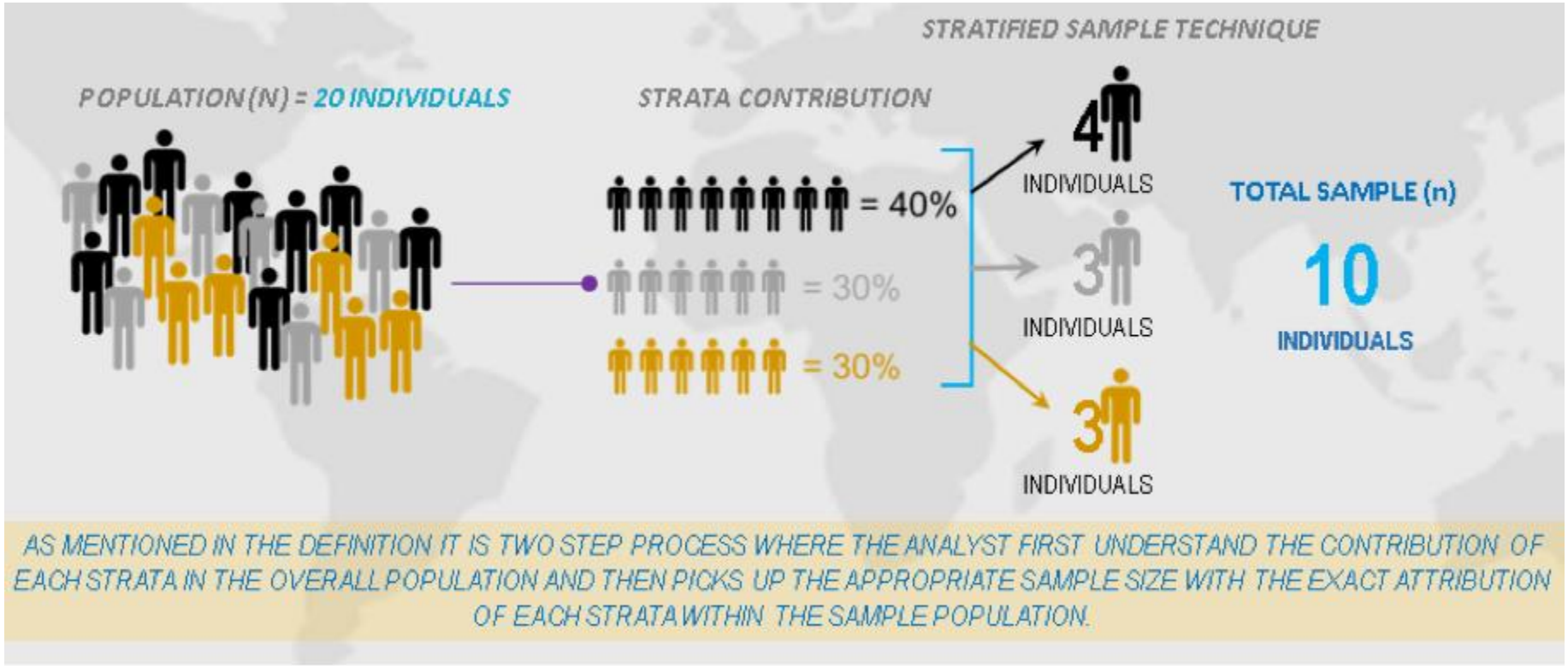
Types of Data Sampling

- Systematic Sampling
  - a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval.
  - This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.



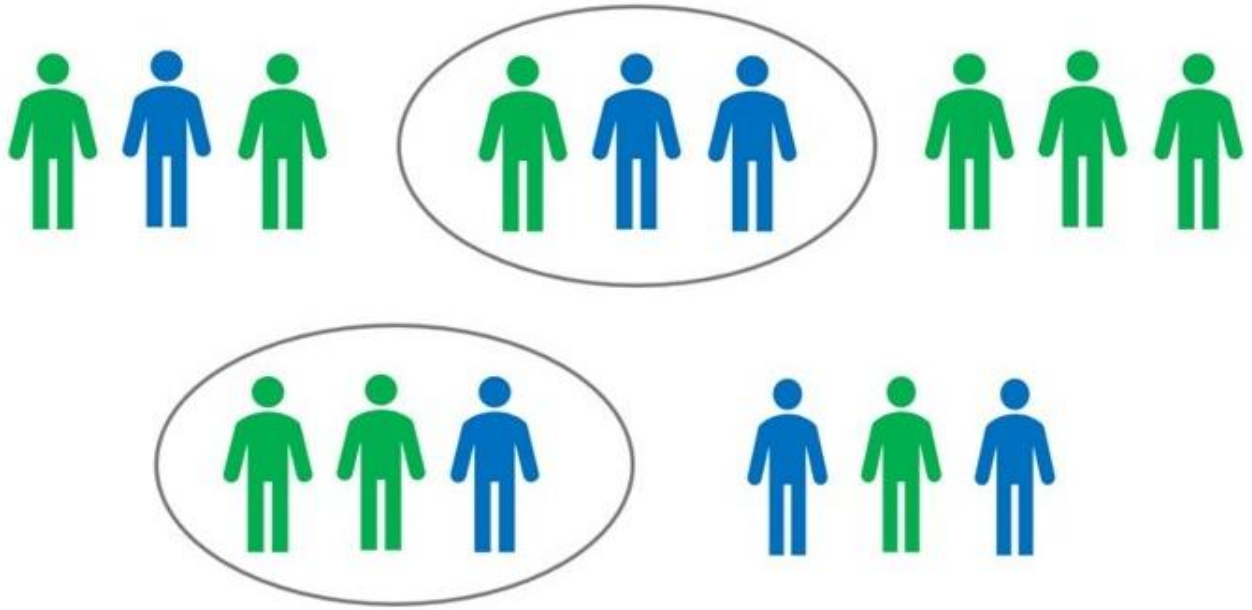
Types of Data Sampling

- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition



Types of Data Sampling

- Cluster sampling
  - a probability sampling technique where researchers divide the population into multiple groups (clusters) for research.
  - then select random groups with a simple random or systematic random sampling technique for data collection and data analysis.



投票(匿名) 最多可选1项

设置

一个公司有500人，其中小于30岁的有125人，30-40岁的有280人，40岁以上的有95人。为了解职工的血压情况，要抽样一个容量为100人的样本进行分析。**采用什么方法抽样更为合适?**

A System Sampling

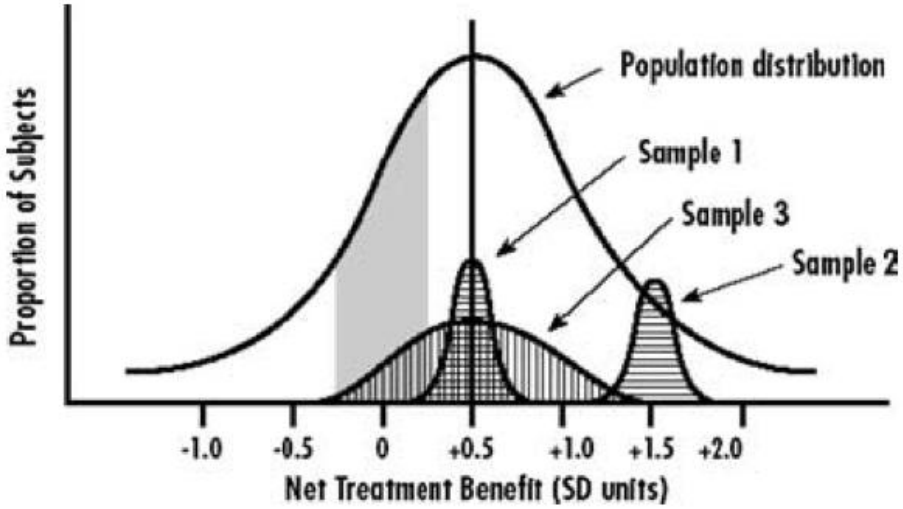
B Stratified Sampling

C Cluster Sampling

提交

Sampling

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative



A sample is representative if it has approximately the same properties (of interest) as the original set of data

Sampling

Use secondary data for verification:

- Compare demographics to ‘known, good’ samples (e.g., census data)
- Needs to relate to variables of interest

8000 points | 2000 Points | 500 Points

Sampling: two errors

Random Sampling Error:

- Generally decreases as the sample size increases (but not proportionally)
- Depends on the variability of the characteristic of interest in the population

Systematic Bias in sampling:

- Does not decreases as the sample size increases (but not proportionally)
- Depends on assumptions made by the experimenter about the population

Filtering

Data filtering: the process of choosing a smaller part of your data set and using that subset for viewing or analysis.

Unfiltered

	Average	Sample size
Coca-Cola	3.73	15
Diet Coke	2.93	15
Coke Zero	3.07	15
Pepsi	3.60	15
Diet Pepsi	2.73	15
Pepsi Max	2.80	15

Filtered to Males

	Average	Sample size
Coca-Cola	4.50	6
Diet Coke	2.17	6
Coke Zero	3.00	6
Pepsi	4.17	6
Diet Pepsi	2.67	6
Pepsi Max	2.83	6

主观题 10分

设置

假设服务器端具有2000个时间片段的三维密度场模拟数据（可以理解成三维中的每一个位置都有数值，每个时间步的数据大小是1GB ）

请问如何才能让用户在本地能更快的访问这个数据?

作答

主观题 10分

设置

假设S中包含10亿个允许的邮件地址（非垃圾邮件），邮件地址通常包含至少20个字节的信息，内存大小为 1 GB。

对于包含邮件地址和邮件内容的流数据，如何以S为依据过滤掉垃圾邮件?

正常使用主观题需2.0以上版本雨课堂

作答

Sampling

Rapid Sampling for Visualizations with Ordering Guarantees

Albert Kim
MIT
alkim@csail.mit.edu

Eric Blais
MIT and University of Waterloo
eblais@uwaterloo.ca

Aditya Parameswaran
MIT and Illinois (UIUC)
adityagp@illinois.edu

Piotr Indyk
MIT
indyk@mit.edu

Sam Madden
MIT
madden@csail.mit.edu

Ronitt Rubinfeld
MIT and Tel Aviv University
ronitt@csail.mit.edu

Sampling

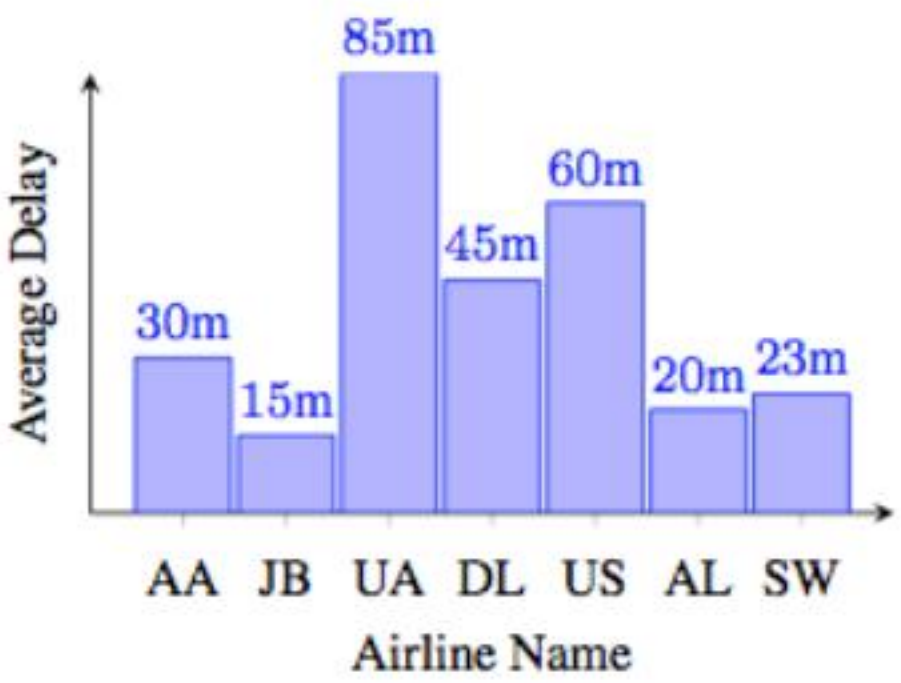
Analysts may want answers to:

Which airline has the largest delay?

Is delay of US > DL?

How much worse is UA than DL? 10X? 2X?

Q: SELECT name, AVG (delay)
FROM FLT
GROUP BY name

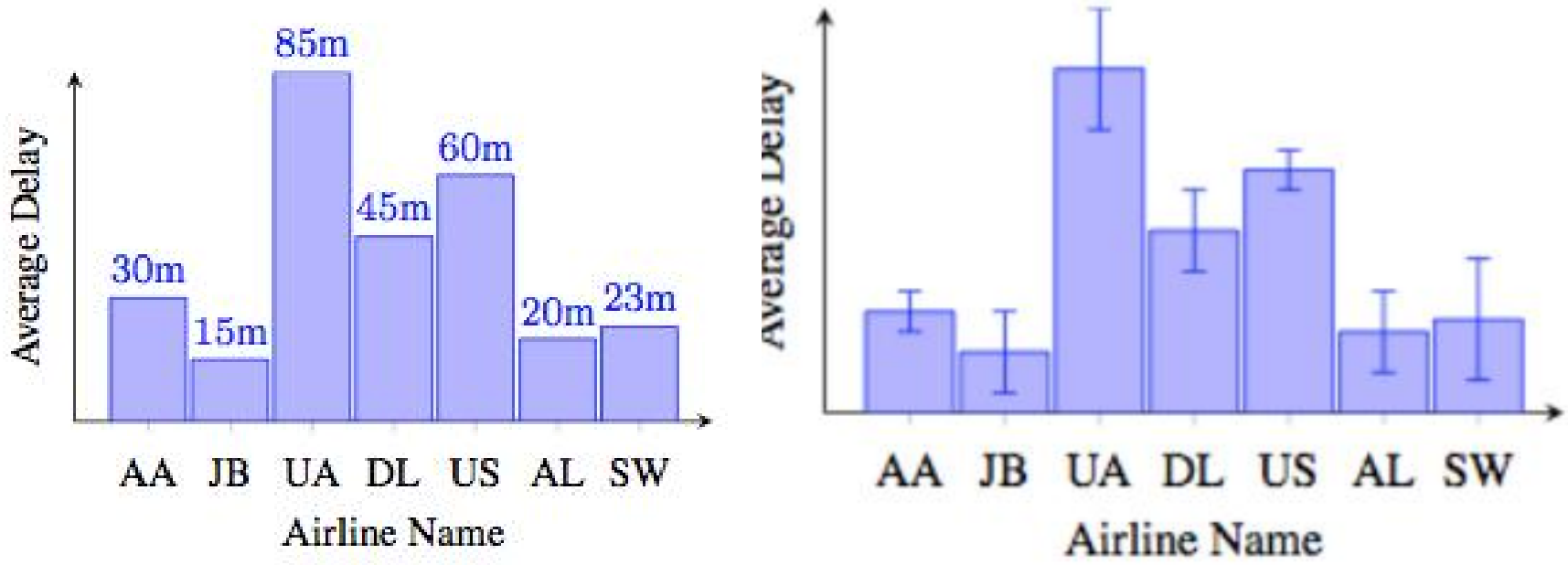


Too Long!!

Sampling

Insight: these questions are related to trends and comparisons, as opposed to actual values

Can we generate approximate visualizations that look like visualizations on the entire data but are computed on much less?

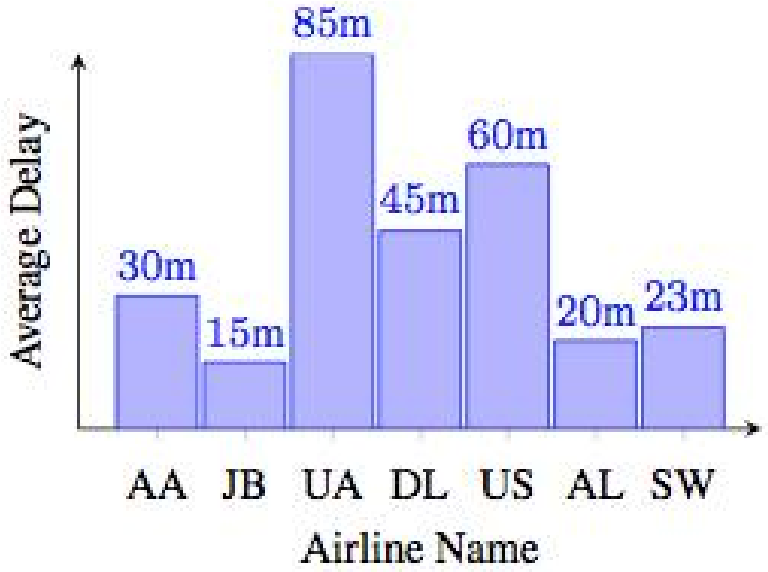


Sampling

Can we generate approximate visualizations that look like visualizations on the entire data but are computed on much less data?

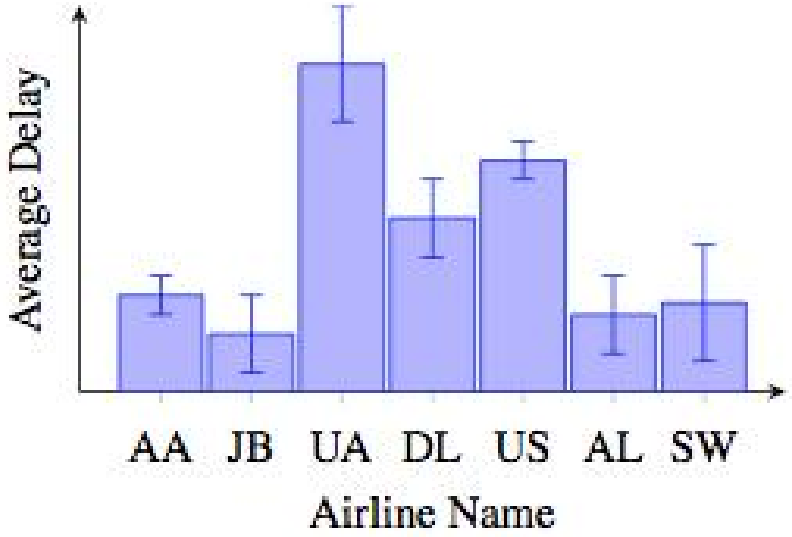
Our definition of “look like”

correct ordering property



if AVG(UA) > AVG(US) in the data, then it must be so in the visualization

Sampling



Round-Robin Stratified Sampling

How do I get the minimum number of samples?

- In what order should I sample? How much?
- Should I sample from UA (larger CI, fewer conflicts) or AL (smaller CI, more conflicts)?
- When do I stop?

Sampling

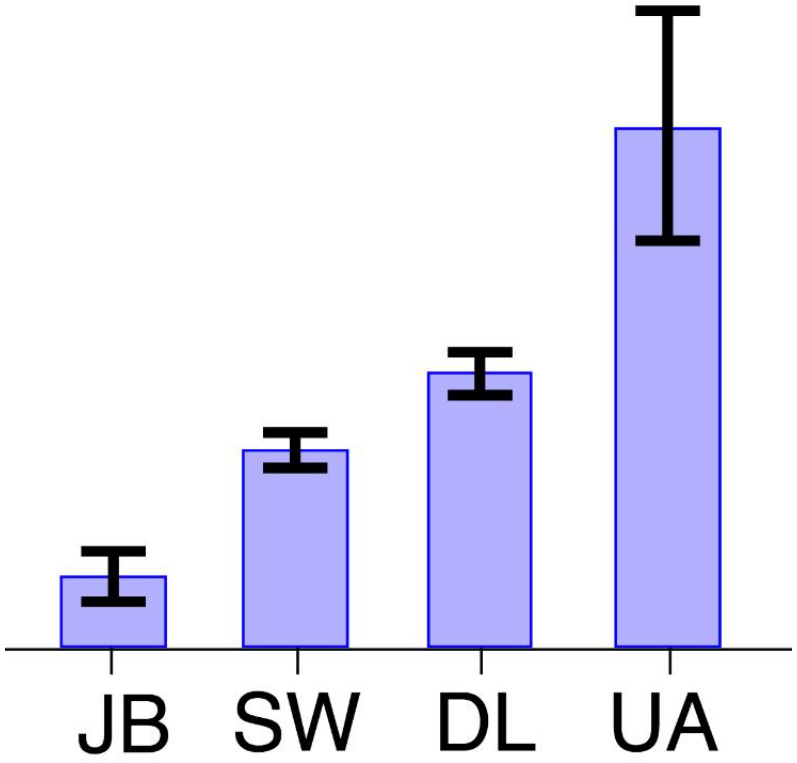
Set all groups as active - whose CI intervals overlap with other groups, a single additional sample is taken

- While any active groups remain
- Sample from all active groups
- Recompute CI
- Mark groups whose CI don’t overlap as inactive

	Group 1		Group 2		Group 3		Group 4	
1	[60, 90]	A	[20, 50]	A	[10, 40]	A	[40, 70]	A
				...				
20	[64, 84]	A	[28, 48]	A	[15, 35]	A	[45, 65]	A
21	[66, 84]	I	[30, 48]	A	[17, 35]	A	[46, 64]	A
				...				
57	[66, 84]	I	[32, 48]	A	[17, 33]	A	[46, 62]	A
58	[66, 84]	I	[32, 47]	A	[17, 32]	I	[46, 61]	A
				...				
70	[66, 84]	I	[40, 47]	A	[17, 32]	I	[46, 53]	A
71	[66, 84]	I	[40, 46]	I	[17, 32]	I	[47, 53]	I

Table 1: Example execution trace: active groups are denoted using the letter A, while inactive groups are denoted as I

Sampling



- Take samples from active groups
- Update CI
- Update active groups

Recent Cited Papers

- Making data visualization more efficient and effective: a survey
- Secure Sampling for Approximate Multi-party Query Processing
- Generalized Measure-Biased Sampling and Priority Sampling
- Cache-Efficient Top-k Aggregation over High Cardinality Large Datasets
- OM3: An Ordered Multi-level Min-Max Representation for Interactive Progressive Visualization of Time Series
- LAQy: Efficient and Reusable Query Approximations via Lazy Sampling
- A structured review of data management technology for interactive visualization and analysis
- Impact of cognitive biases on progressive visualization

Data Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

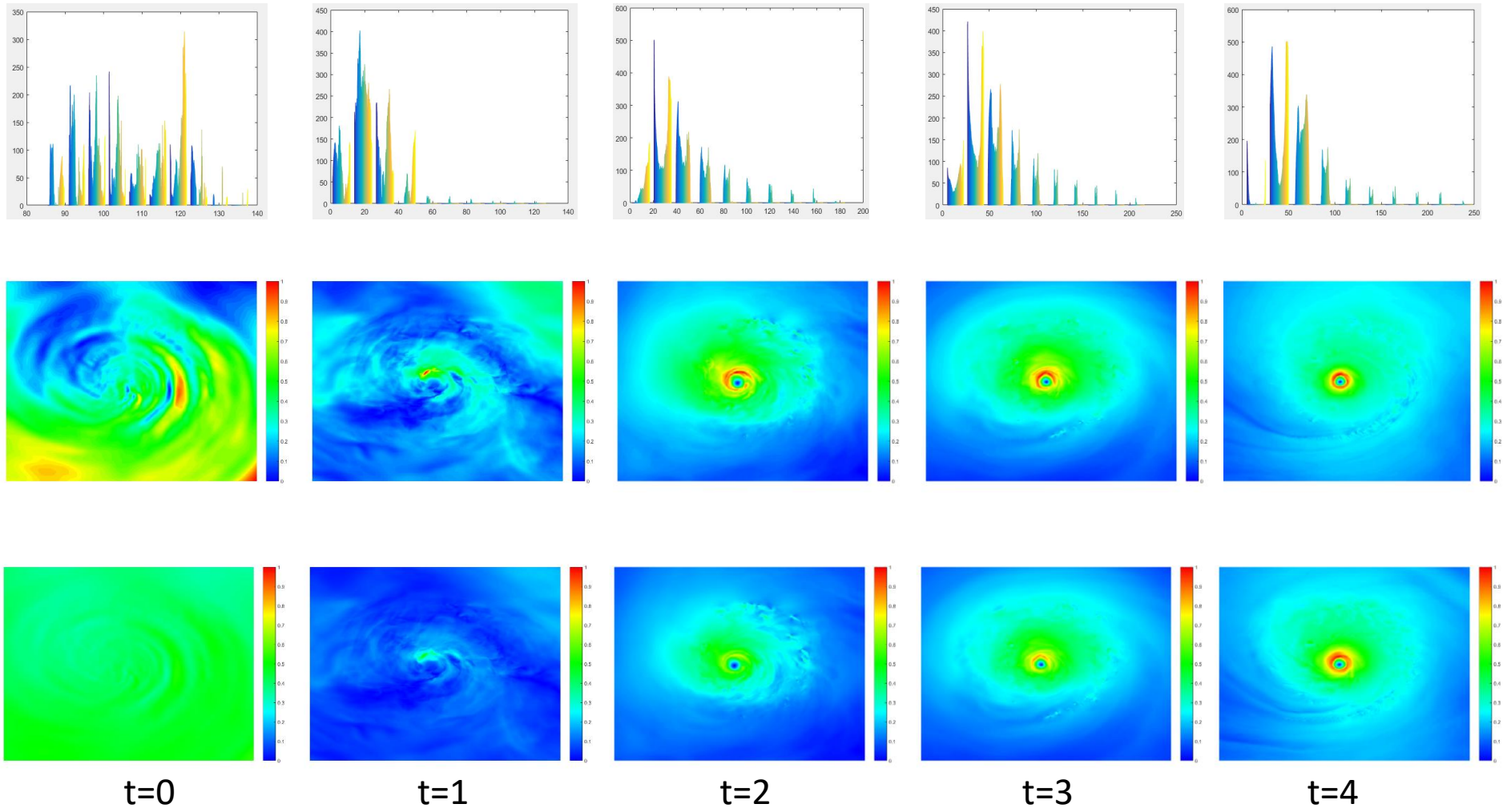
$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1

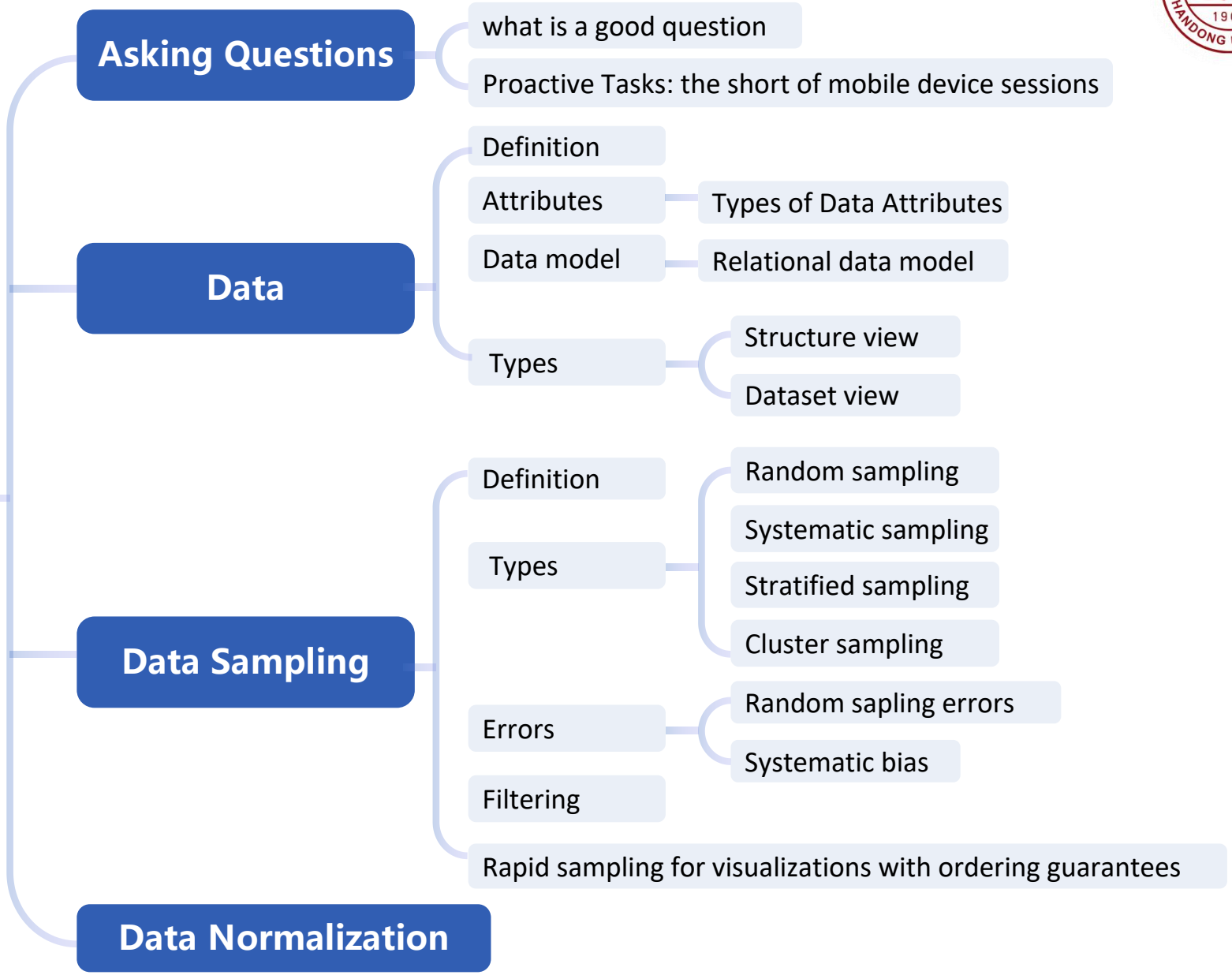
- Other nonlinear scaling: log, tan

Data Normalization



Summary

Exploratory Data Analysis



Week-3 Assignment

Personal:

Test different data sampling / filtering methods with PANDAS

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	众包与电子表格	电子表格实践	第五周
20251008	统计分析方法与工具	统计方法实践	第六周
20251015	可视化设计	可视化设计实践	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	机器学习方法与工具	BERT实践	第十周
20251112	人机交互方法与工具	Canis/Cast/Libra实践	第十一周
20251119	普适计算	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	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

Thank You