CS 8803-MDS
Human-in-the-loop Data Analytics

Lecture 3

08/29/22

Logistics

Paper review 1 is available on gradescope

due Tuesday midnight

Office hour change

Kexin: 11AM-12PM Friday

Paper presenters

send slides to Kaushik by Wednesday noon

archeologist role still open

Today’s class

Introduction to approximate query processing

Measuring user experiences

Help us learn your names!

Data is growing exponentially

IDC survey

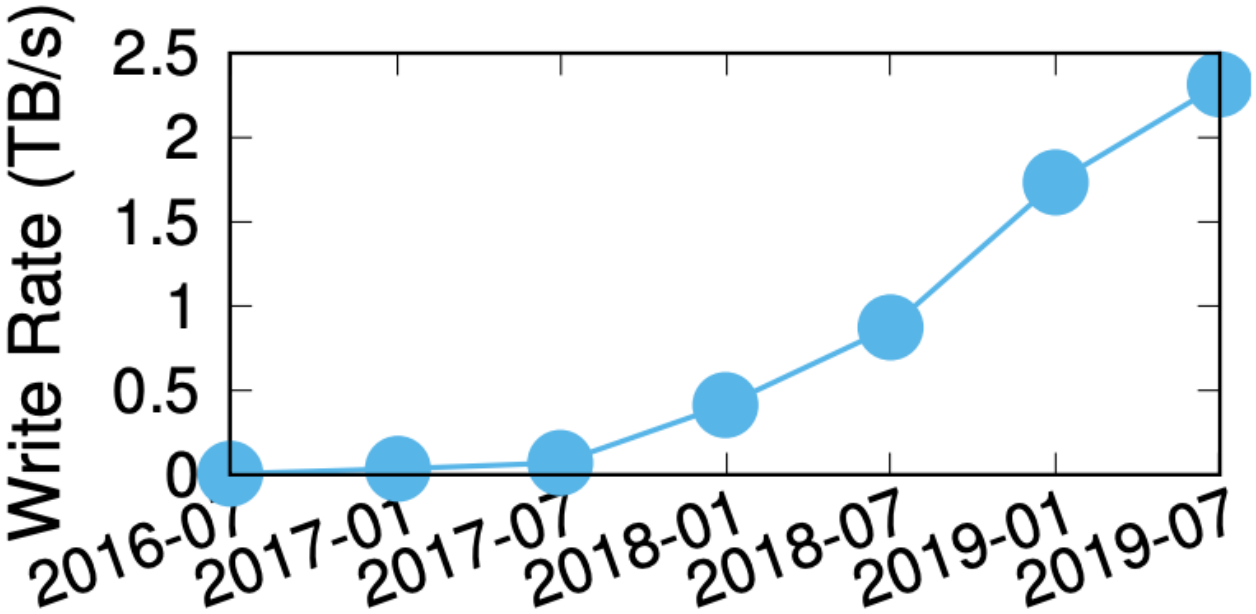
- >70ZB data created worldwide in 2021

Machine generated data

- Server event logs
- Sensor readings
- …

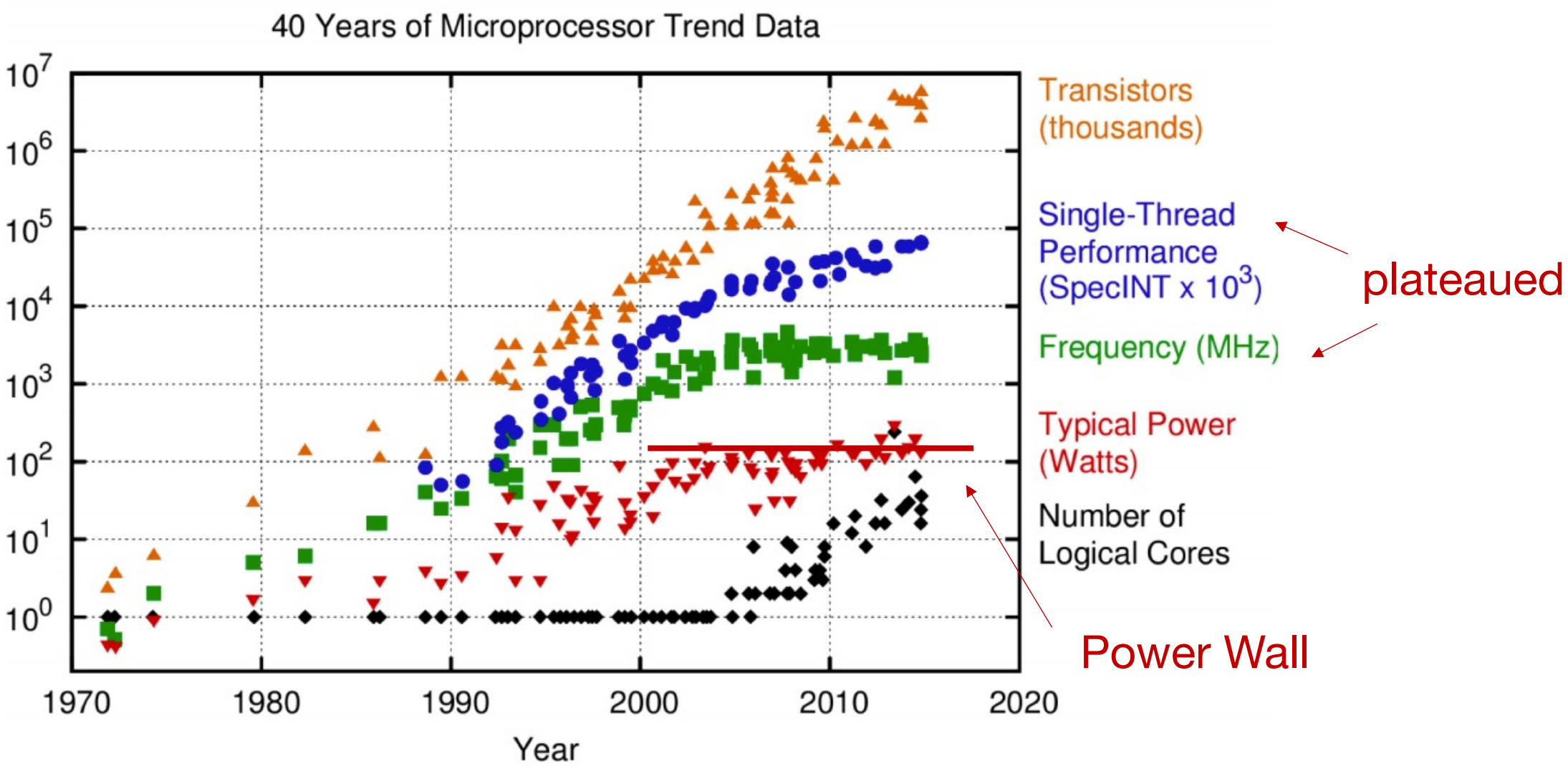
Google

- 2.5TB of monitoring data collected per second [1]



[1] C. Adams, et al.. Monarch: Google's Planet-Scale In-Memory Time Series Database. PVLDB, 13(12): 3181-3194, 2020.

Data outpaces compute



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Today: a lot of data remain unused

Example: infrastructure monitoring

Infeasible to inspect all data

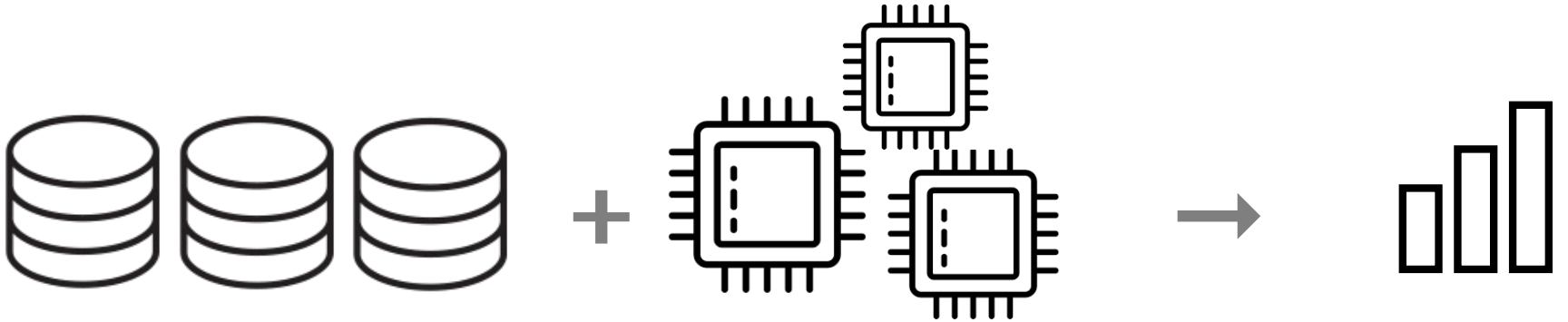
In practice:

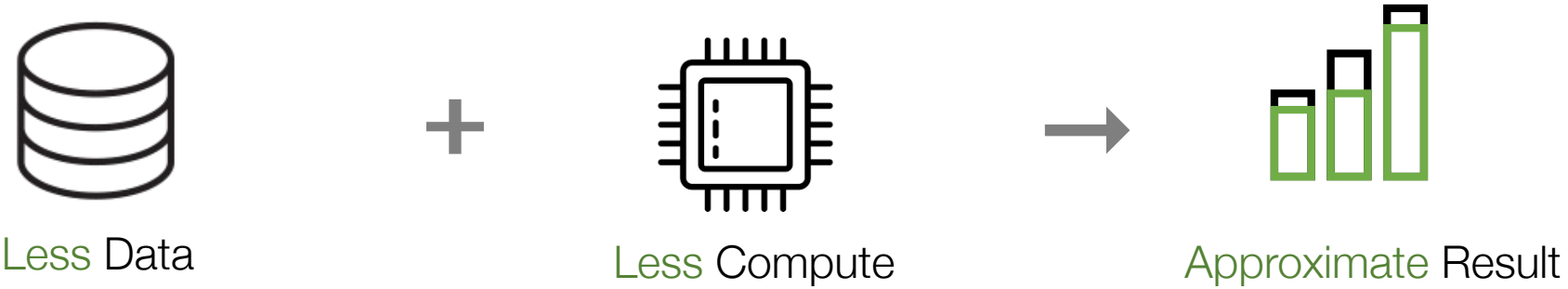
- data only accessed for post-hoc root cause analyses

*Top SV orgs say: **< 6%** data read*

Future: impracticable to process all data

Literature: Approximate Query Processing





Approximate Query Processing: Use Cases

Data exploration: exact answers NOT always required

Goal is to quickly report the leading digits of answers

e.g., avg salary $59,000±$500 (with 95% confidence) in 10 sec
vs. $59,152.25 in 10 min

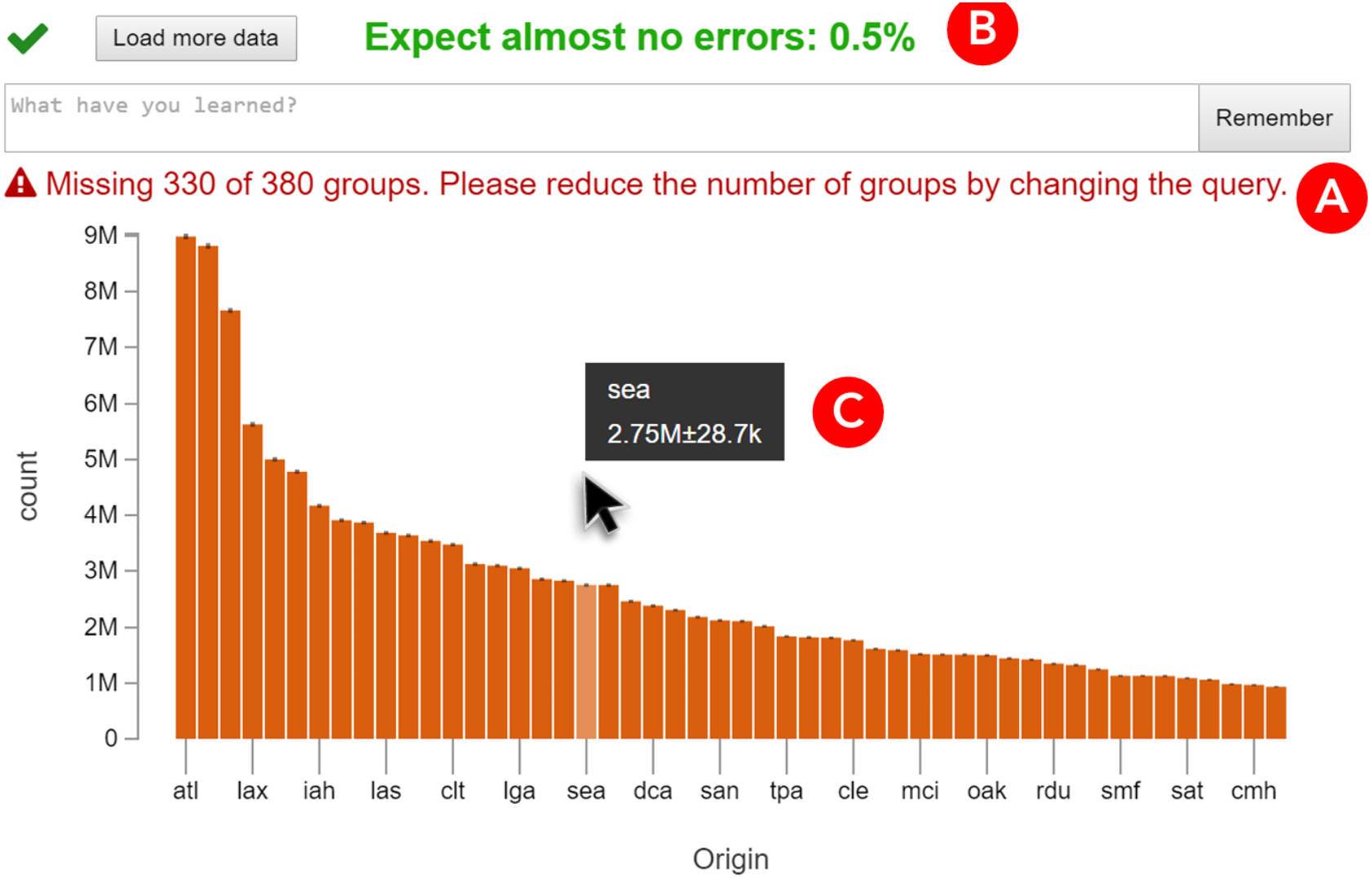
Approximate Query Processing: Use Cases

Dat

Goa

e

0 sec
0 min



Approximate Query Processing: Use Cases

OLAP: online analytical processing

OLTP: online transaction processing

HTAP: hybrid transaction/analytical processing

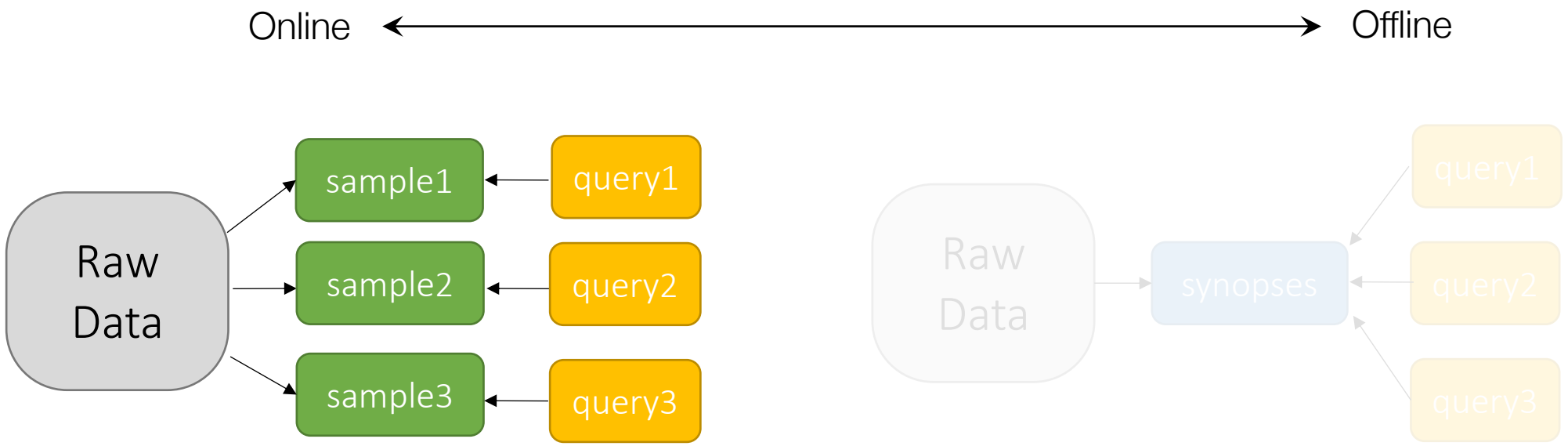
Aggregate queries (COUNT, SUM, AVG)

```
SELECT agg FROM table
WHERE condition
GROUP BY dimensions
```

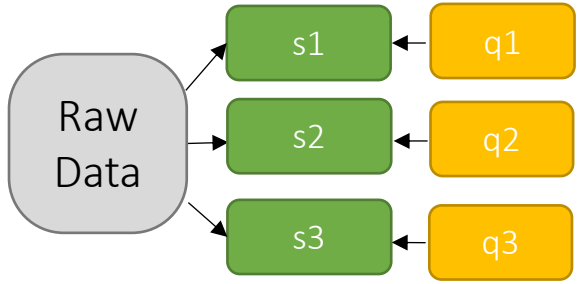
AQP: Basic approaches

Online vs. offline

- Online: sample data at query time
- Offline: precompute synopses to use in place of data



Online: Query-time Sampling



Table

X	Y
ant	10
ant	12
bee	1
cat	40
…	

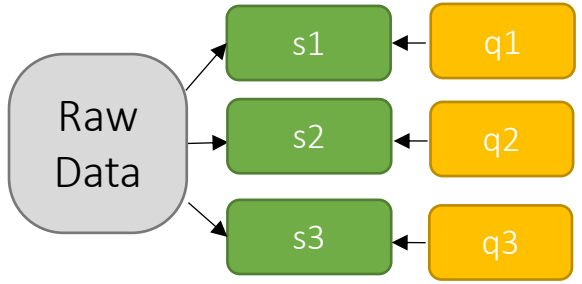
Query

```
SELECT SUM(Y)
```

Idea

Estimate query answer using a random sample of the rows

Sampling: Uniform samples



Uniform sample

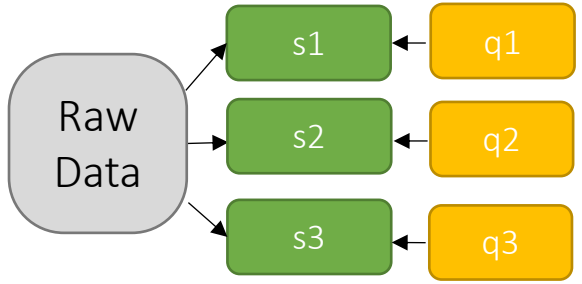
tuples are sampled with the same probability

How to implement

Fixed-sized sampling (e.g., simple random sampling with or without replacement): assign each tuple a number from 1 and N and select samples with a random number generator

Bernoulli sampling: each tuple has probability p of being in the sample more efficient

Sampling: Confidence Interval

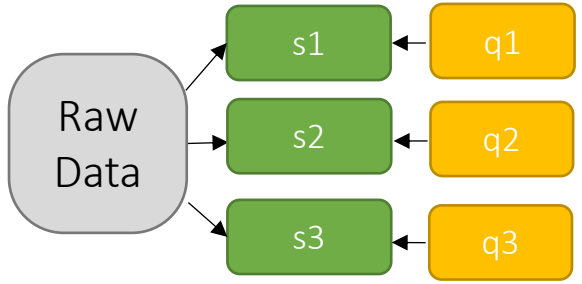


95% CI of $x \pm \alpha$ means that the true value of the query answers is within $\pm \alpha$ of the current estimate x with 95% probability.

True or false:

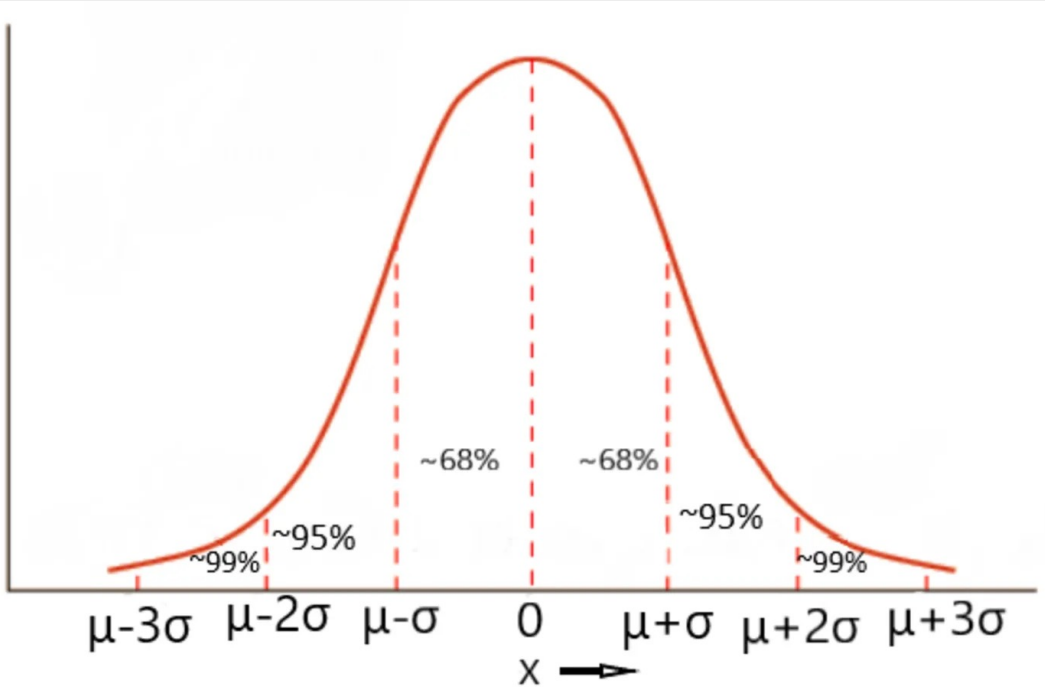
- *90% CI is wider than a 95% CI for the same data.* F
- *A 95% confidence level means that 95% of the sample data lie within the confidence interval.* F
- *There is a 95% probability that the 95% CI calculated from a given sample will cover the true value of the population parameter.* T

Sampling: Confidence Interval

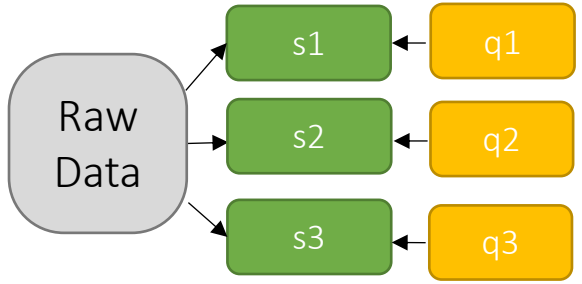


S: samples, n = |S|, σ: std, μ: sample mean

Method	90% CI ($\pm$)	Assumption
Central Limit Theorem	$z^* * \frac{\sigma(S)}{\sqrt{n}}$ (z^*=1.65)	as $n \to \infty$



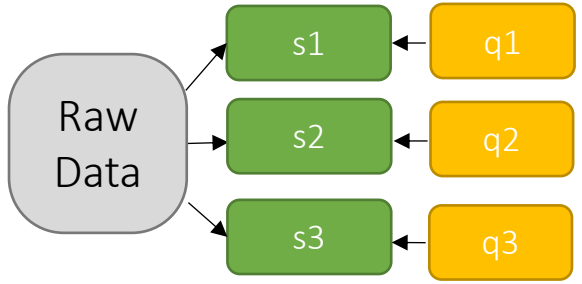
Sampling: Confidence Interval



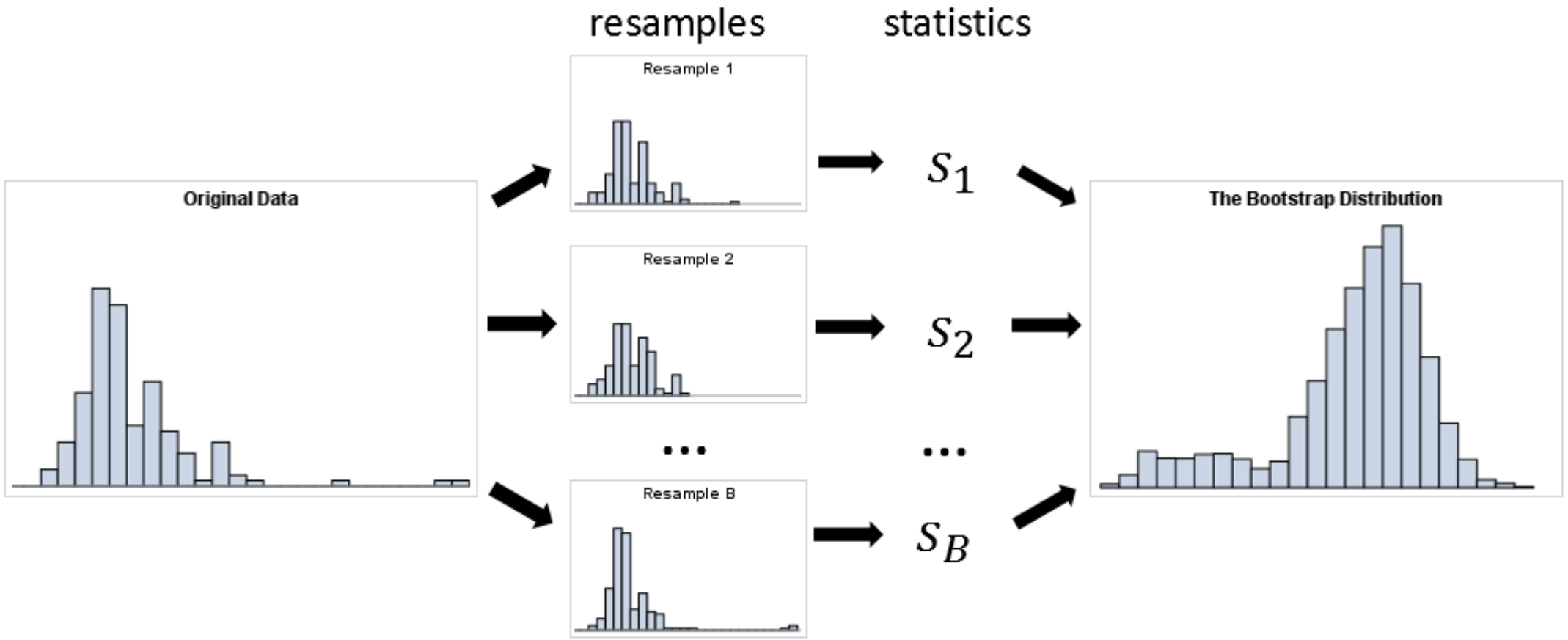
S: samples, n = |S|, σ: std, μ: sample mean

Method	90% CI ($\pm$)	Assumption
Central Limit Theorem	$z^* * \frac{\sigma(S)}{\sqrt{n}}$ (z^*=1.65)	as $n \to \infty$
Chebychev	$P(\lvert X-\mu\rvert < k\sigma) \geq 1-\frac{1}{k^2}$ $(k = 3.16)$	Finite non-zero σ
Hoeffding (tighter)	$P(\lvert X-\mu\rvert \geq t) \leq 2\exp(-\frac{2nt^2}{(b-a)^2})$	a < X < b
Bootstrap	via resampling	

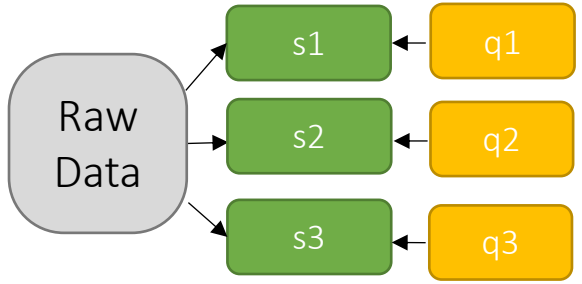
Sampling: Confidence Interval



Bootstrap: resample data with replacement

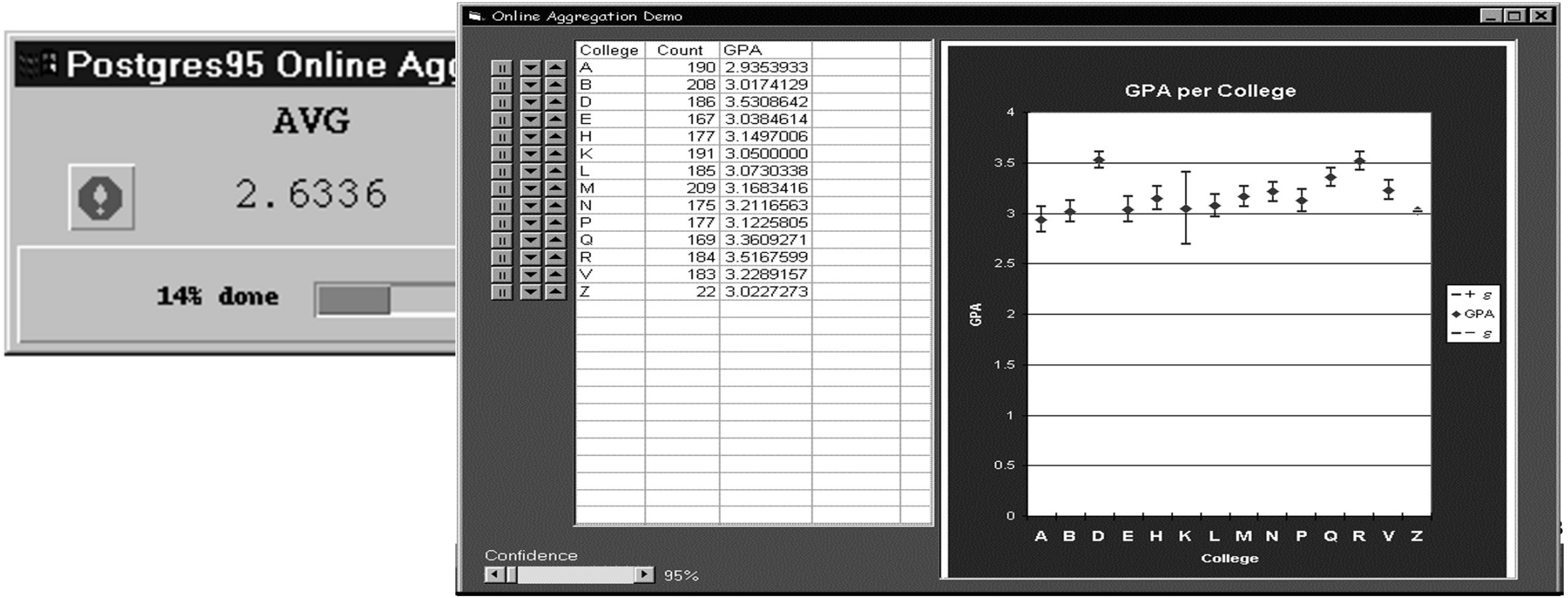


Online Aggregation [HHW'97]

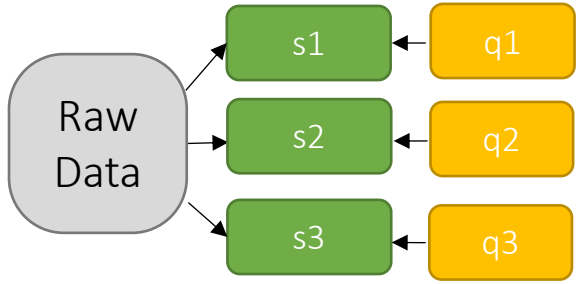


Sampling at query time

Answers continually improve, under user control



More readings on error bars



Presenting uncertainty to users:

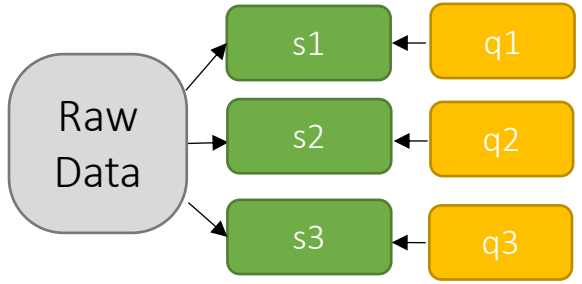
How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results

How Progressive Visualizations Affect Exploratory Analysis

Producing reliable error bars:

Knowing When You're Wrong: Building Fast and Reliable Approximate Query Processing Systems

Sampling: Predicates, GROUP BY



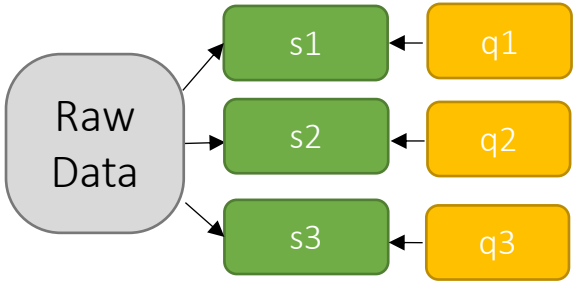

```
SELECT X, SUM(Y) … GROUP BY X
SELECT SUM(Y) … WHERE Z > 10
```

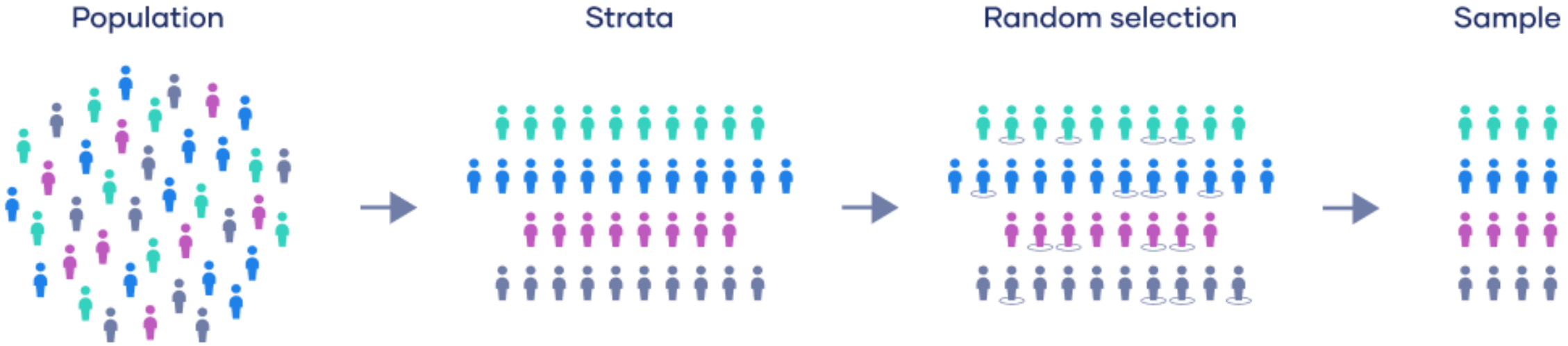
What can go wrong with uniform samples?

selective predicates and small groups

=> very few samples / inaccurate estimates

Sampling: Stratified samples

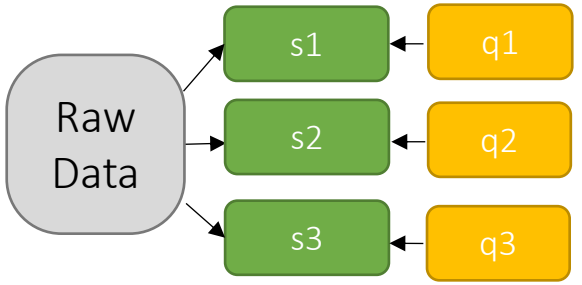




Every member of the population is in **exactly one** strata

Used to highlight differences between groups in a population

Congressional samples[AGP'00]



House

uniform random sample of the entire relation

Senate

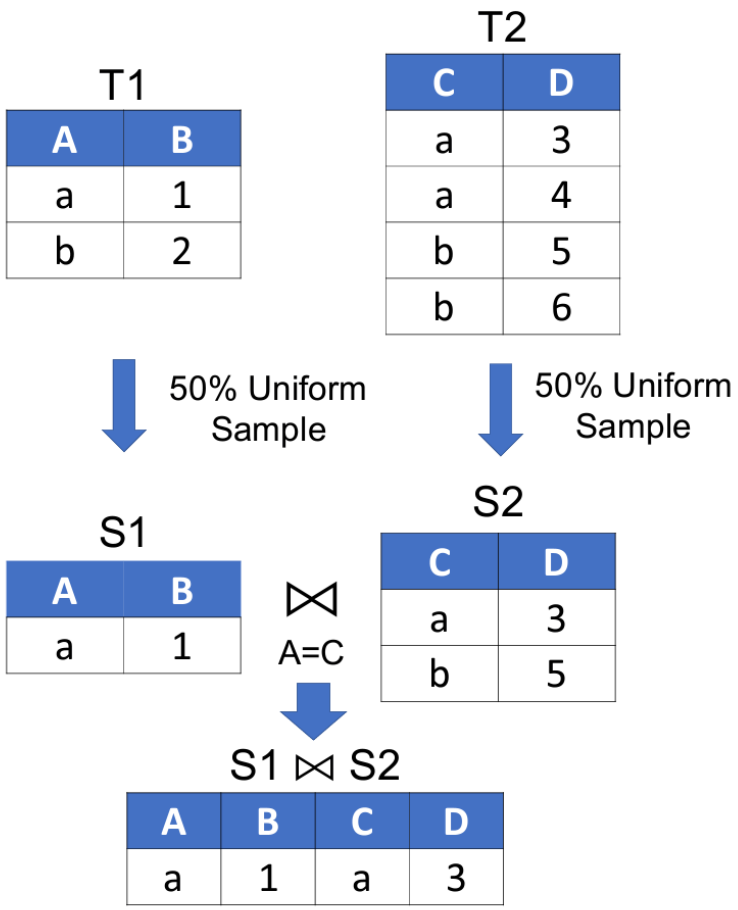
uniform random sample from each group

e.g., grouping attribute `T = state`, N/50 samples from each state

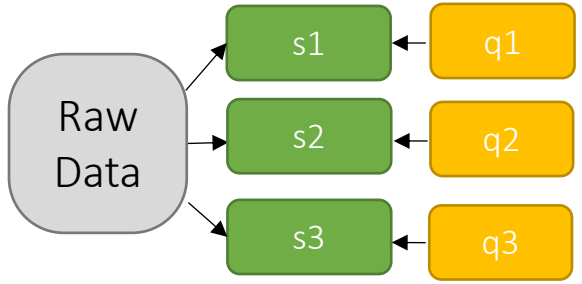
Congress

split the budget between House and Senate

Sampling: JOIN



Sample before join



T1 ⋈ T2

A	B	C	D
a	1	a	3
a	1	a	4
b	2	b	5
b	2	b	6

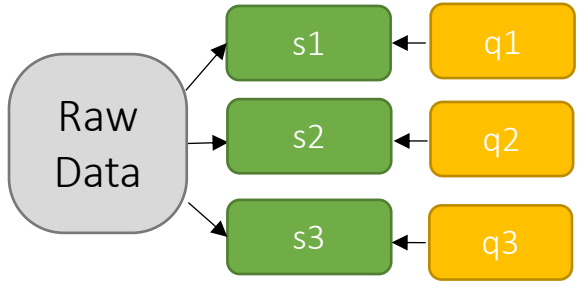
50% Uniform Sample

S(T1 ⋈ T2)

A	B	C	D
a	1	a	3
b	2	b	5

Join before sample

Sampling: JOIN



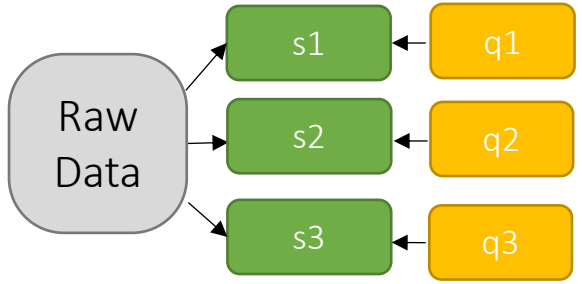
Join before sample

- Joins are expensive!

Sample before join

- joining two uniform samples that are each p fraction ($0 \leq p < 1$) of the original tables will only produce p^2 of the output tuples of the original join
- joining two i.i.d. samples is an identical, but not independent, sample of the join

More readings on sampling



Smart stratified samples:

Learning to sample: counting with complex queries

Handling JOINs:

Experiences with Approximating Queries in Microsoft's Production Big-Data Clusters

Joins on Samples: A Theoretical Guide for Practitioners.

Online: small overheads and small gains

Example:

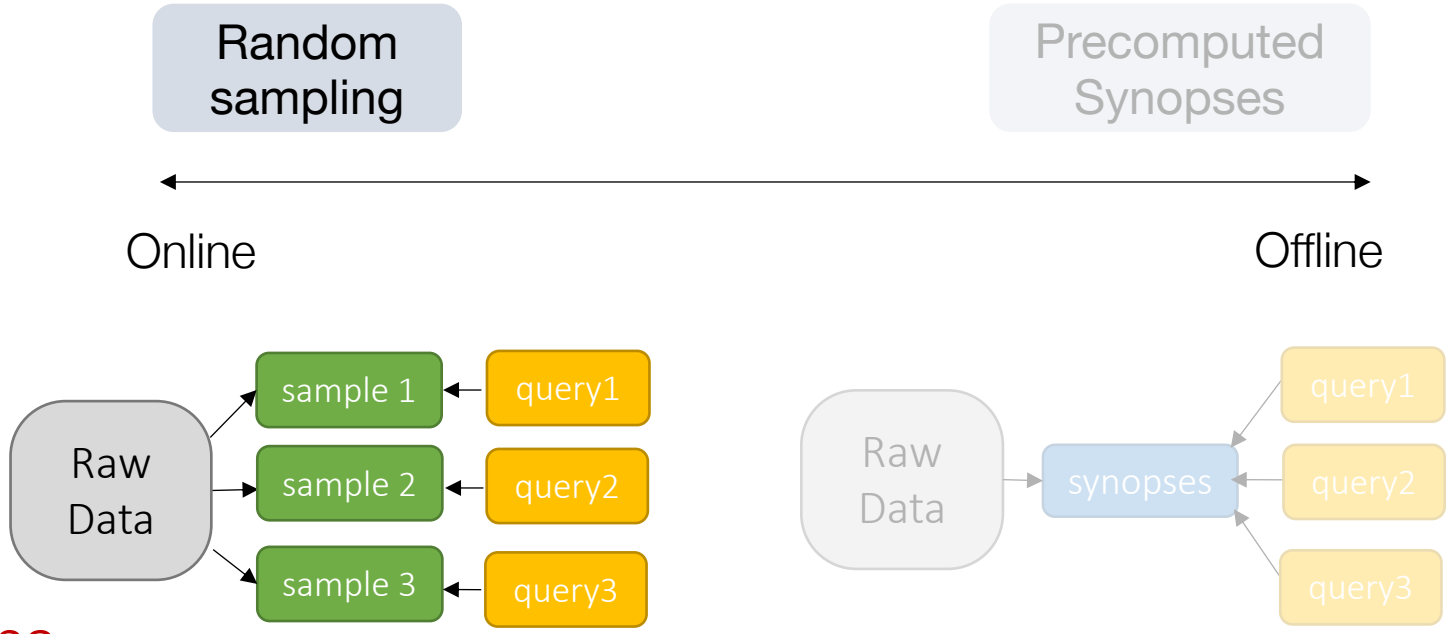
- Online aggregation [1]
- Query-time sampling [2]

Pros

- No precomputation
- General

Cons

- Not accurate => limited performance gains

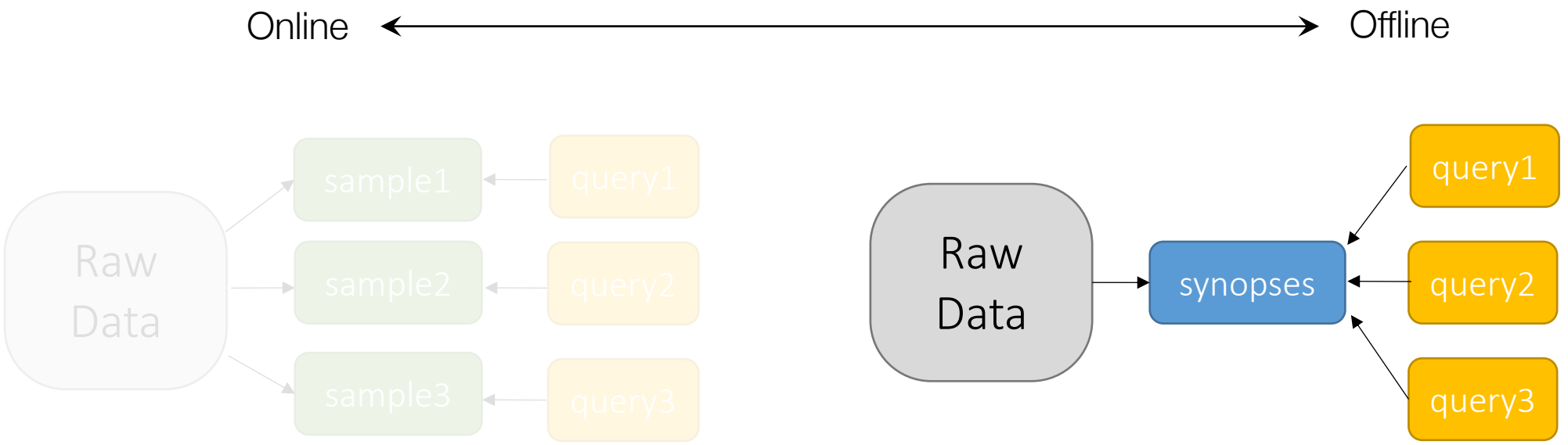


[1] J. Hellerstein, P. Haas, H. Wang. Online aggregation. In SIGMOD, pages 171–182, 1997.
[2] S. Kandula, et al. Quickr: Lazily approximating complex ad-hoc queries in bigdata clusters. In SIGMOD, 2016.

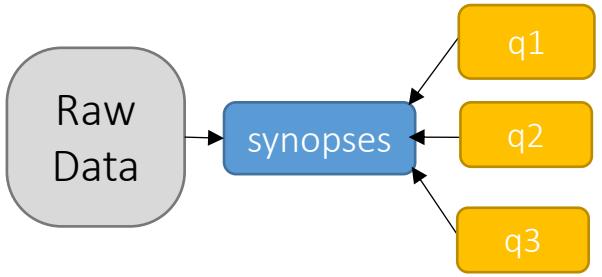
AQP: Basic approaches

Online vs. offline

- Online: sample data at query time
- Offline: precompute synopses to use in place of data



Offline: Materialized Views



What it is: caching results of a query (exact)

Order table

Partition key	Row key	Order date	Shipping address	Total invoice	Order status
001 (Customer ID)	1 (Order ID)	11082013	One Microsoft way Redmond, WA 98052	$400	In process
005	2	11082013	One Microsoft way Redmond, WA 98052	$200	Shipped

OrderItem table

Partition key	Row key	Product	Unit Price	Amount	Total
1 (Order ID)	001_1 (OrderItem ID)	XX	$100	2	$200
1	001_2	YY	$40	5	$200
2	002_1	ZZ	$200	1	$200

Customer table

Partition key	Row key	Billing Information	Shipping address	Gender	Age
US East (region)	001 (Customer ID)	*****0001	One Microsoft way Redmond, WA 98052	Female	30
US East	002	*****2006	One Microsoft way Redmond, WA 98052	Male	40

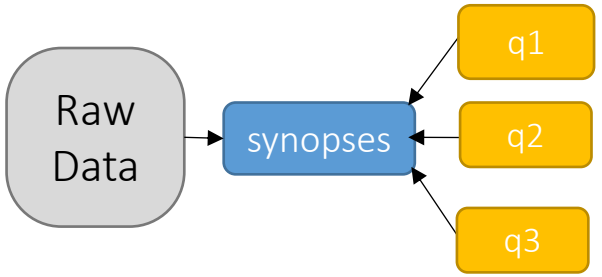
Materialized View

Partition key	Row key	Product Name	Total sold	Number of customers
Electronics (Product category)	001 (Product ID)	XX	$30,000	500
Electronics	002	YY	$100,000	400

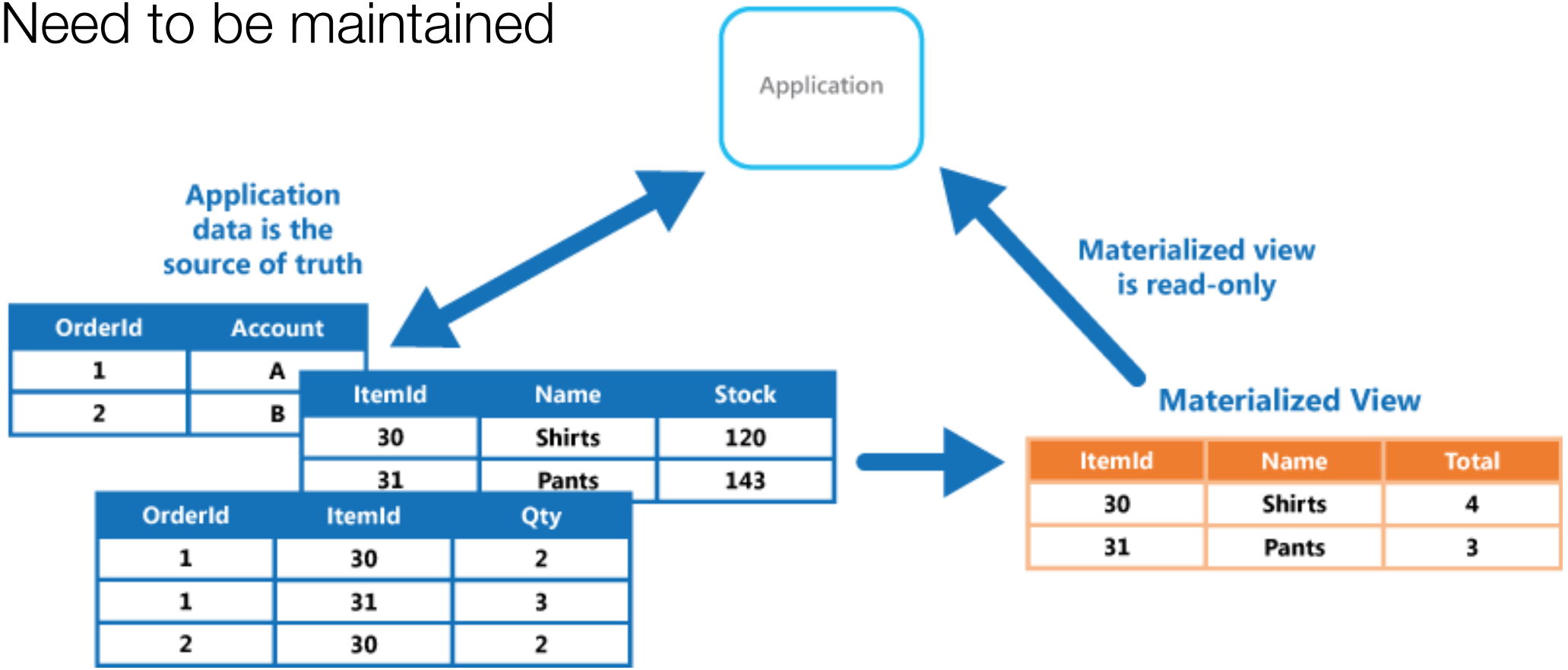
total sales value for each product

Source: https://docs.microsoft.com/en-us/azure/architecture/patterns/materialized-view

Offline: Materialized Views

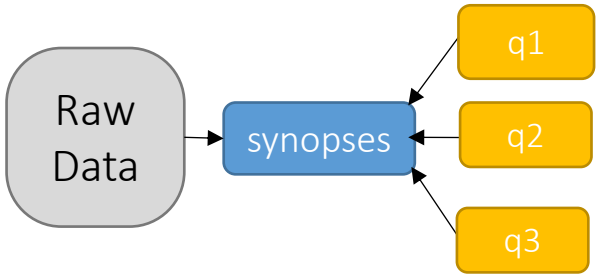


Need to be maintained



Source: https://docs.microsoft.com/en-us/azure/architecture/patterns/materialized-view

Offline: Materialized Views



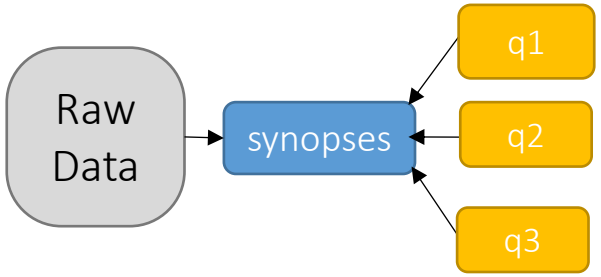
Useful when

- Data is difficult to query directly
- Creating temporary views can dramatically improve query performance
- Can act directly as source views for the UI, for reporting, or for display

Not useful when

- The source data is simple and easy to query.
- The source data changes very quickly

Offline: precomputed samples



Input: storage budget

(optional) information about the query workload

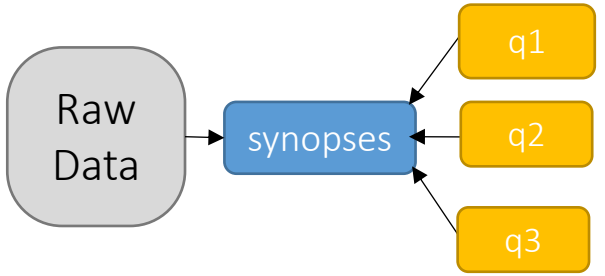
Output: a sample of the dataset

How is this different from query-time sampling?

Samples are precomputed and do not change from query to query

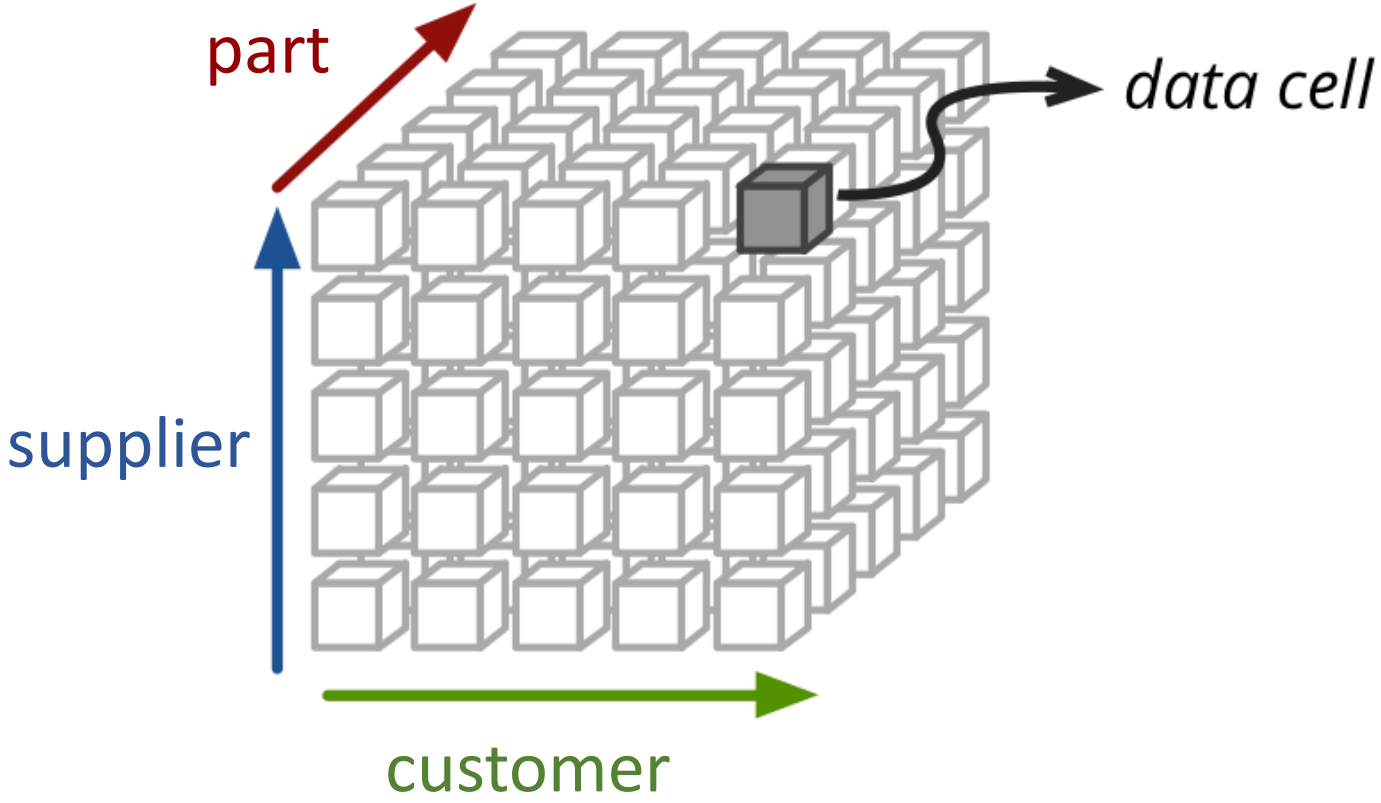
[BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data]

Offline: Data Cubes



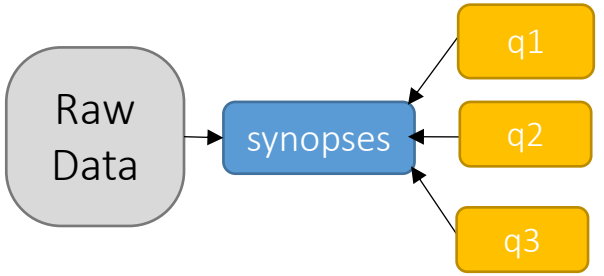
Dimensions
part
supplier
customer

Measure
sales




```
SELECT SUM(sales) …
GROUP BY part, supplier, customer
```

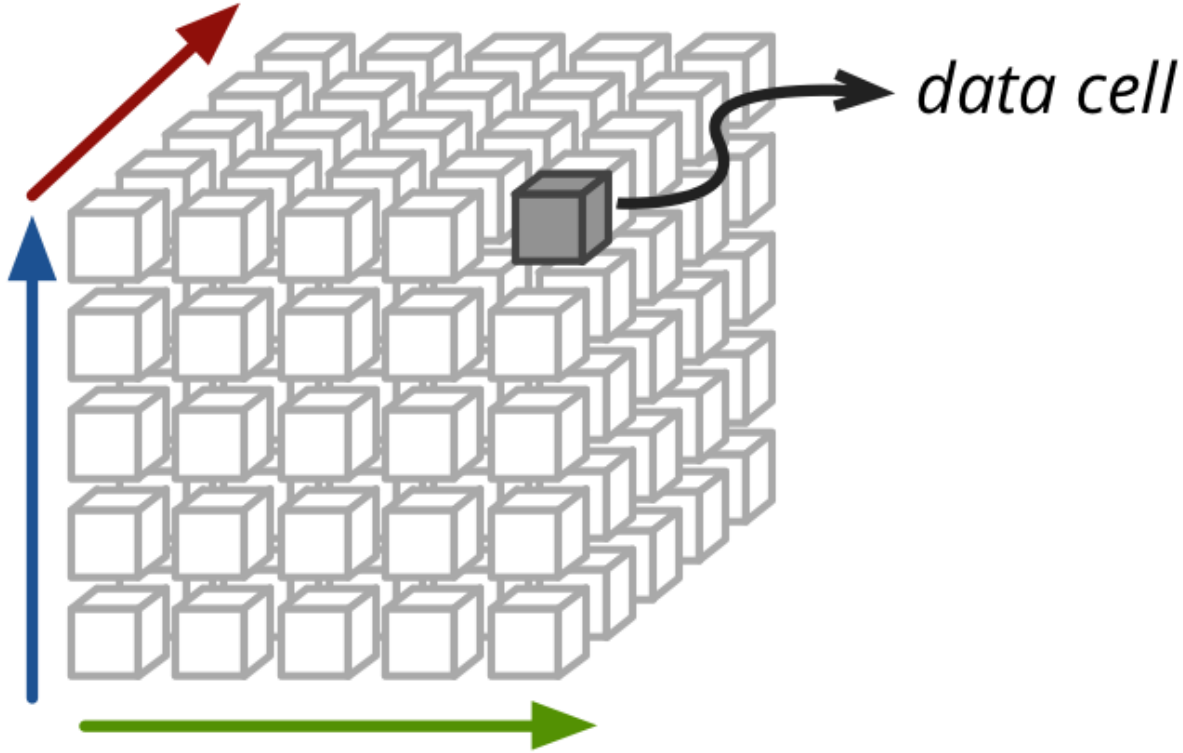
Offline: Data Cubes



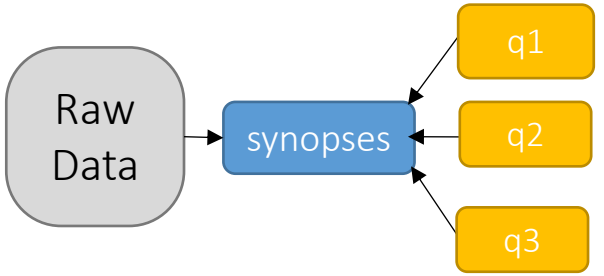
Do not want to recreate the cube from scratch each time

Solution
Materialized views

Problem
Which views to store?

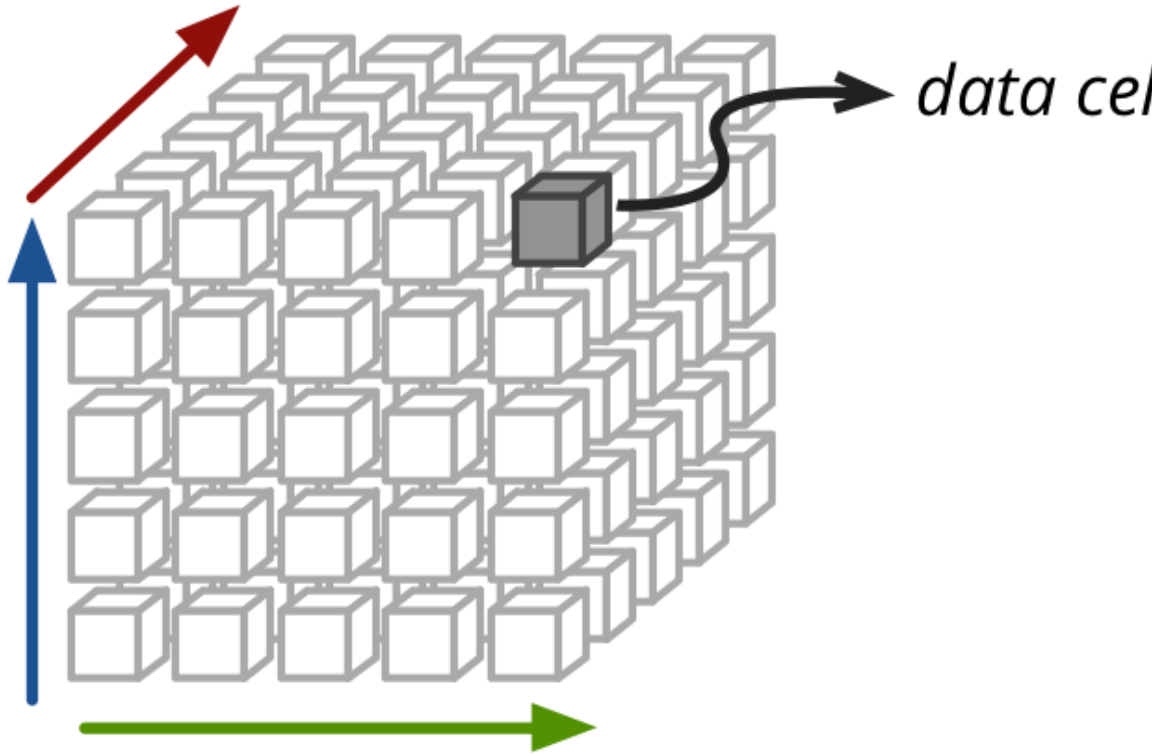


Offline: Data Cubes



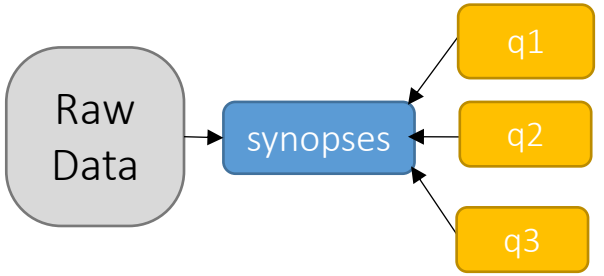
Problem

Which views to store?




```
part, supplier, customer (6M)
part, supplier (0.9M)
part, customer (6M)
supplier, customer (6M)
supplier (0.01M)
part (0.2M)
Customer (0.1M)
none (1)
```

Offline: Data Cubes



Problem

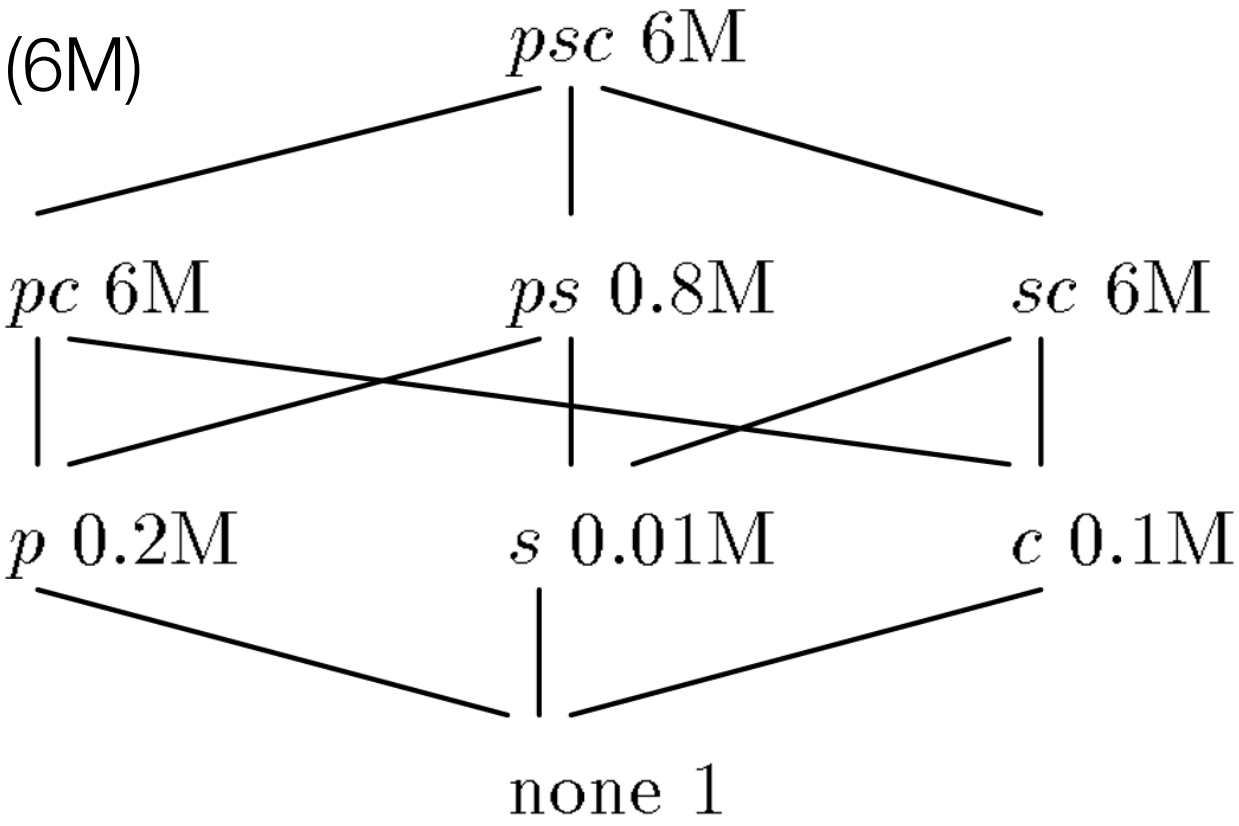
Which views to store?

Given fixed storage budget, how many and which GROUP BYs should we materialize to get reasonable performance and minimum average query cost?

Implementing data cubes efficiently [HRU'96]

The lattice structure

part, supplier, customer (6M)
part, supplier (0.9M)
part, customer (6M)
supplier, customer (6M)
supplier (0.01M)
part (0.2M)
Customer (0.1M)
none (1)



Implementing data cubes efficiently [HRU'96]

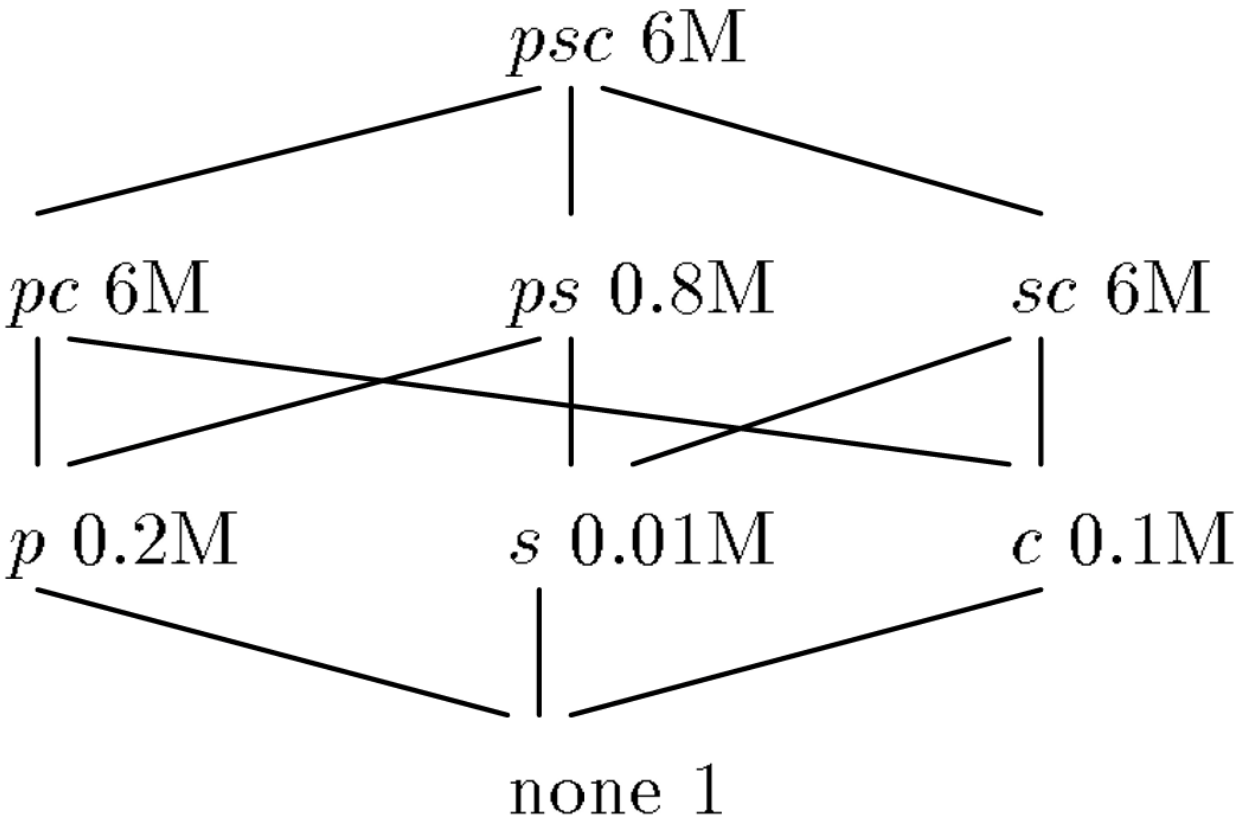
The lattice structure shows dependency:

(p) ≼ (p, c)

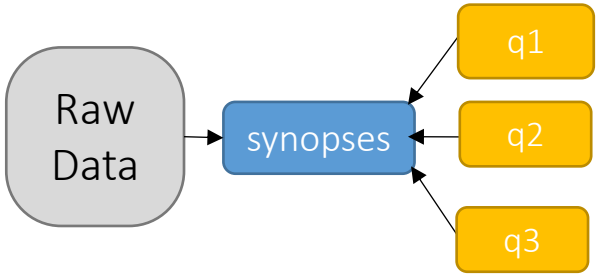
(s, c) ≼ (p, s, c)

Query cost ∝ #rows processed

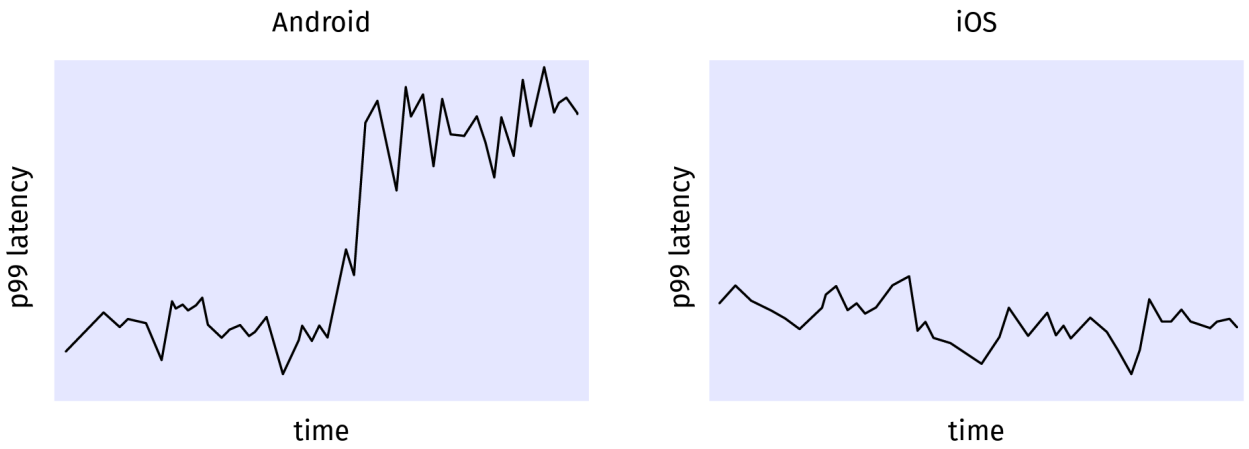
Greedy algorithm to select the best query group to materialize



Offline: mergeable summaries



Billions of events/day of mobile app telemetry data



Query for 99-th **percentile**

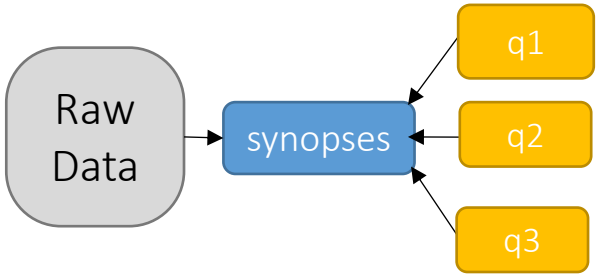
Group By Operating system

Where Location = USA

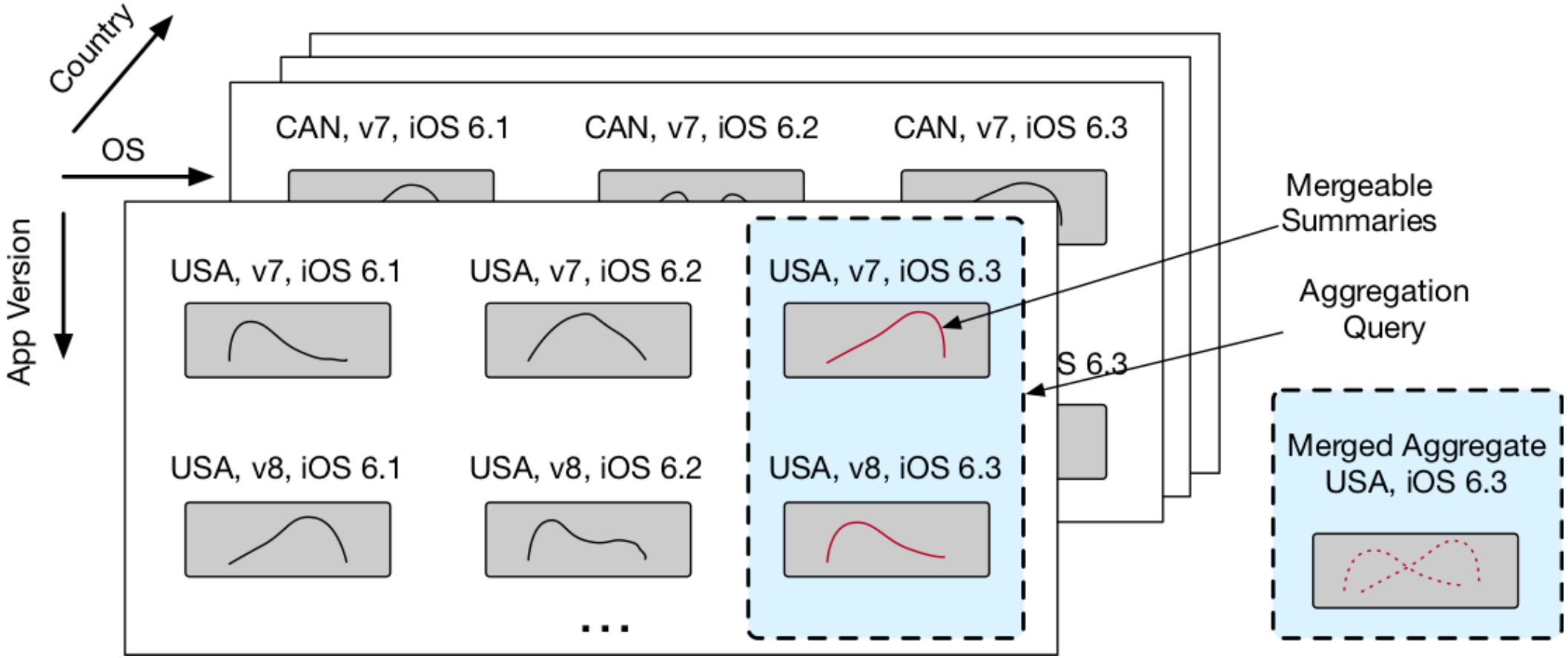
Quantile Query

Use case: compare the 99th percentile response latency across different operation systems

Offline: mergeable summaries

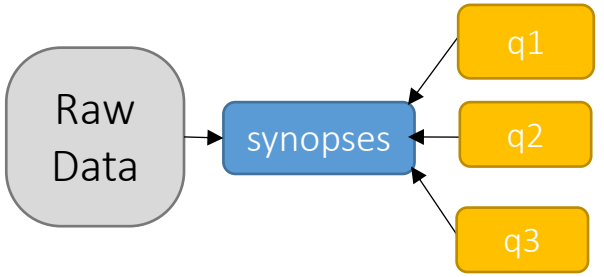


Can be merged without loss of accuracy



Source: https://dawn.cs.stanford.edu/2018/08/29/moments/

Offline: mergeable summaries



Summary size related to error guarantee

Type	Examples
Quantiles	Q-digest, GK sketch[1]
Distribution	Histogram
Distinct value	HyperLogLog, K minimum value sketch, count–min sketch[4]
Heavy hitter	SpaceSaving[2], MG[3]

[1] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In Proc. ACM SIGMOD International Conference on Management of Data, 2001.
[2] A. Metwally, D. Agrawal, and A. Abbadi. An integrated efficient solution for computing frequent and top-k elements in data streams. ACM Transactions on Database Systems, 31(3):1095–1133, 2006.
[3] J. Misra and D. Gries. Finding repeated elements. Science of Computer Programming, 2:143–152, 1982.
[4] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. Journal of Algorithms, 55(1):58–75, 2005.

Offline: large overheads and large gains

Example:

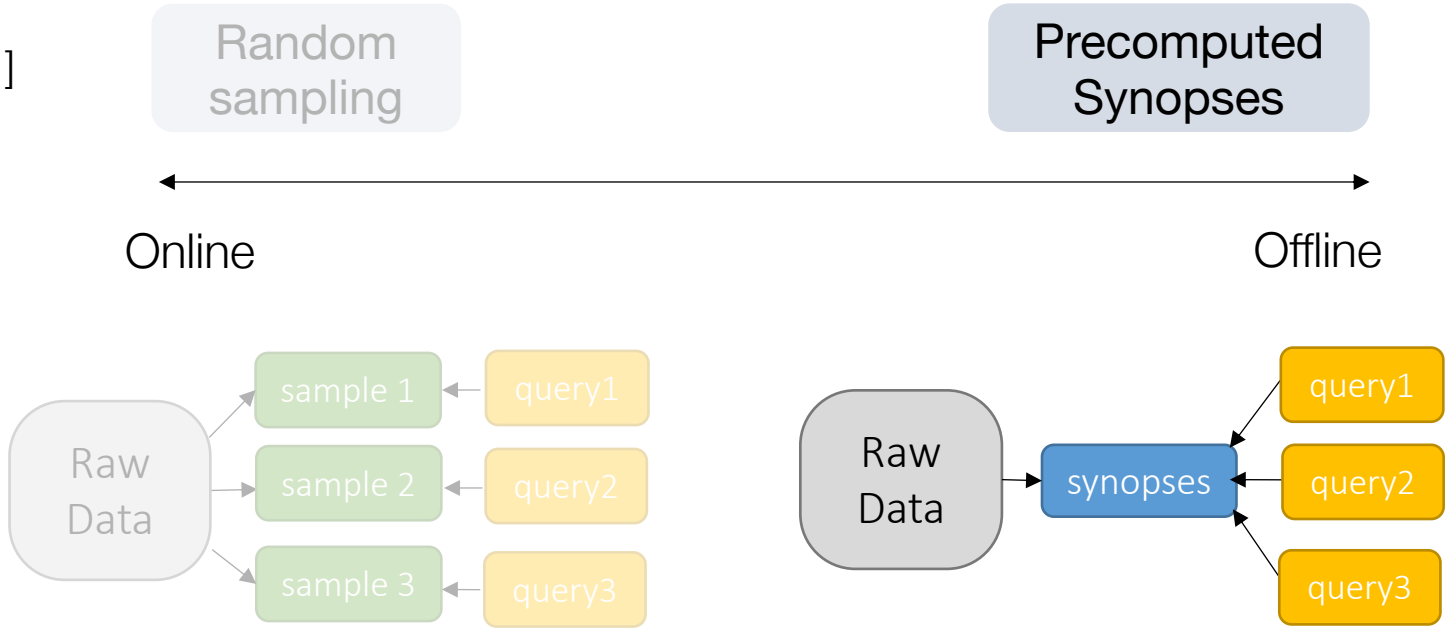
- precomputed samples [1]
- materialized views [2]

Pros

- Fast and accurate

Cons

- Require large storage
- Or not general

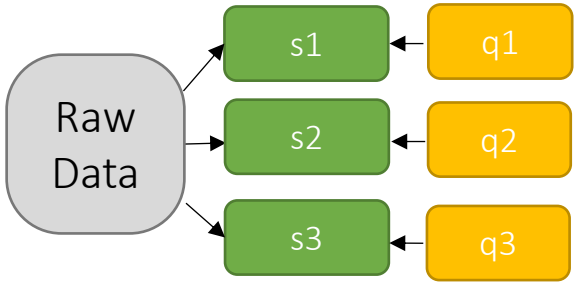


[1] S. Agarwal, B. Mozafari, et al. BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data. In Eurosys, pages 29-42, 2013.
[2] V. Harinarayan, et al. Implementing Data Cubes Efficiently. In SIGMOD, 1996

AQP summary

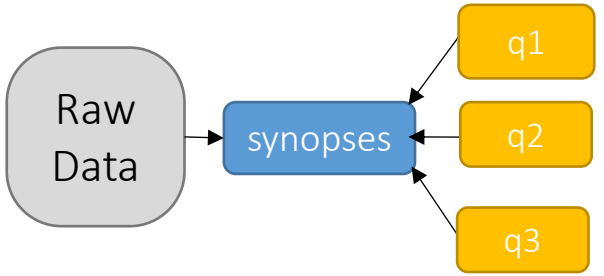
Online (at query time)

- Uniform sampling
- Confidence interval
- Stratified sampling
- Sampling for JOINs

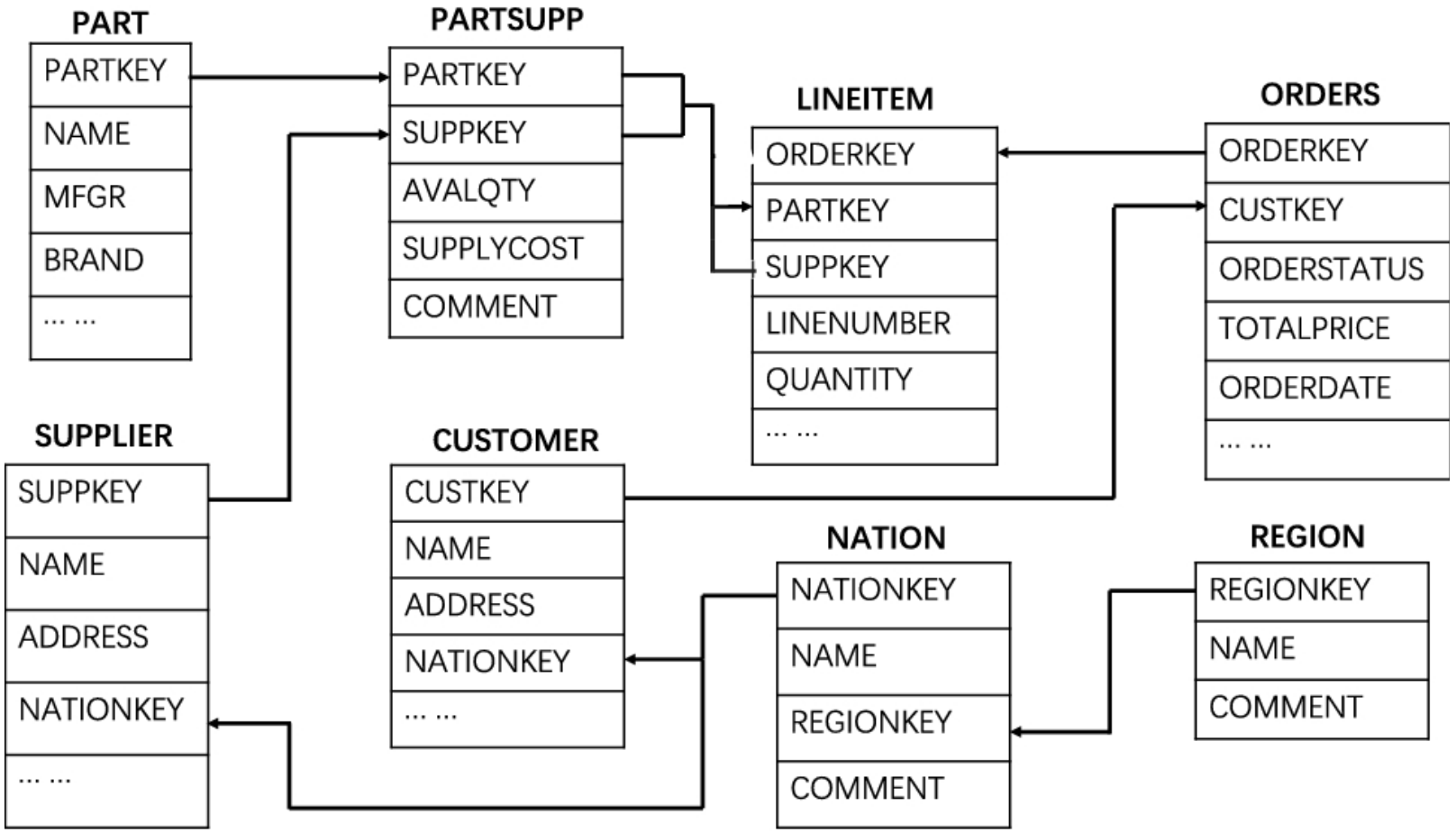


Offline (precomputation)

- Materialized views
- Precomputed samples
- Data cubes
- Mergeable summaries

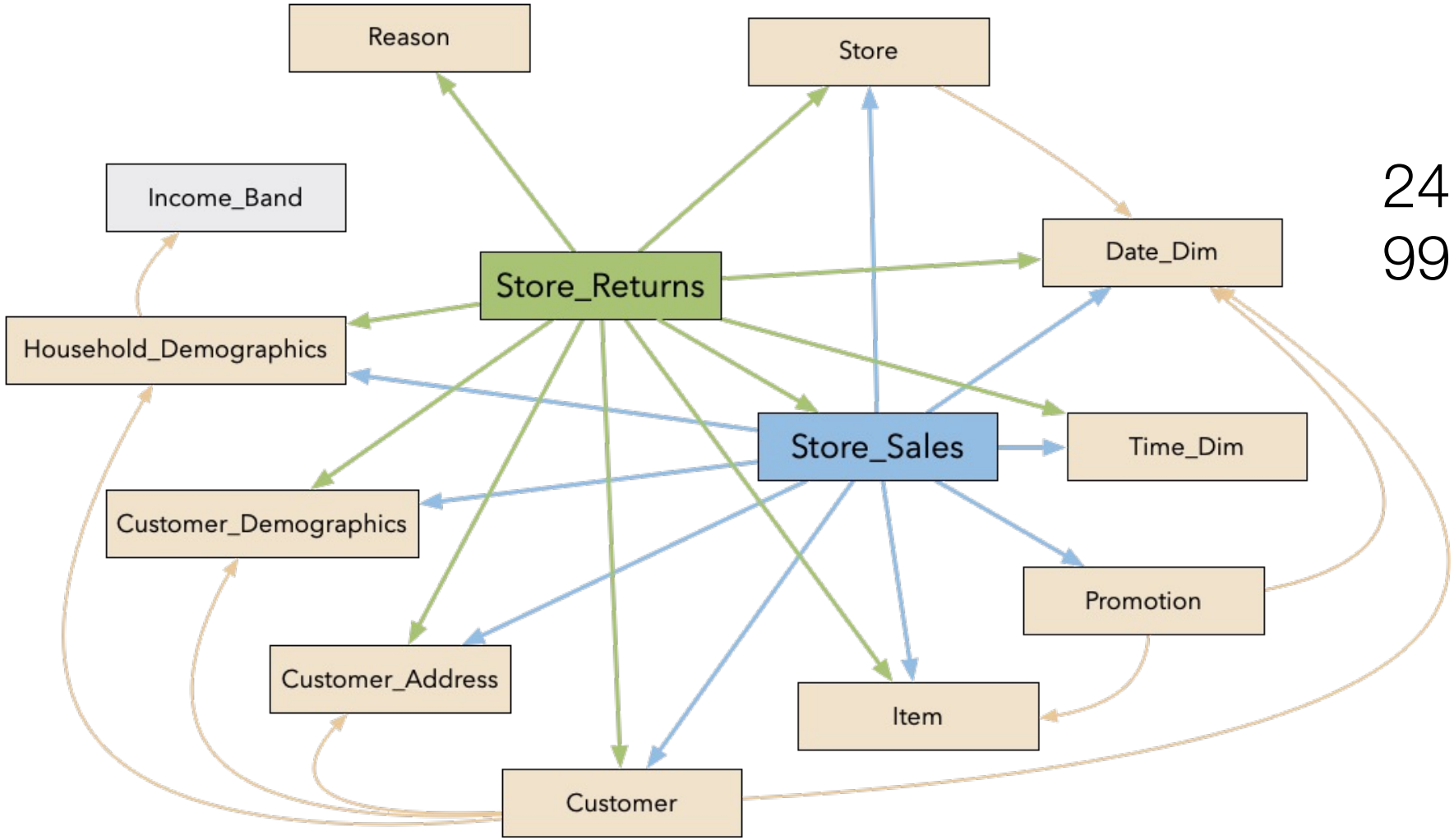


Benchmarks: TPC-H



8 tables
22 query templates

Benchmarks: TPC-DS



24 tables
99 query templates

Measuring the user experience

Types of Evaluation: what's the goal?

Usability Evaluations/Testing

- Identify good and bad aspects of an interface
- Characterize how a user interacts with a system
- Formative vs. summative evaluations

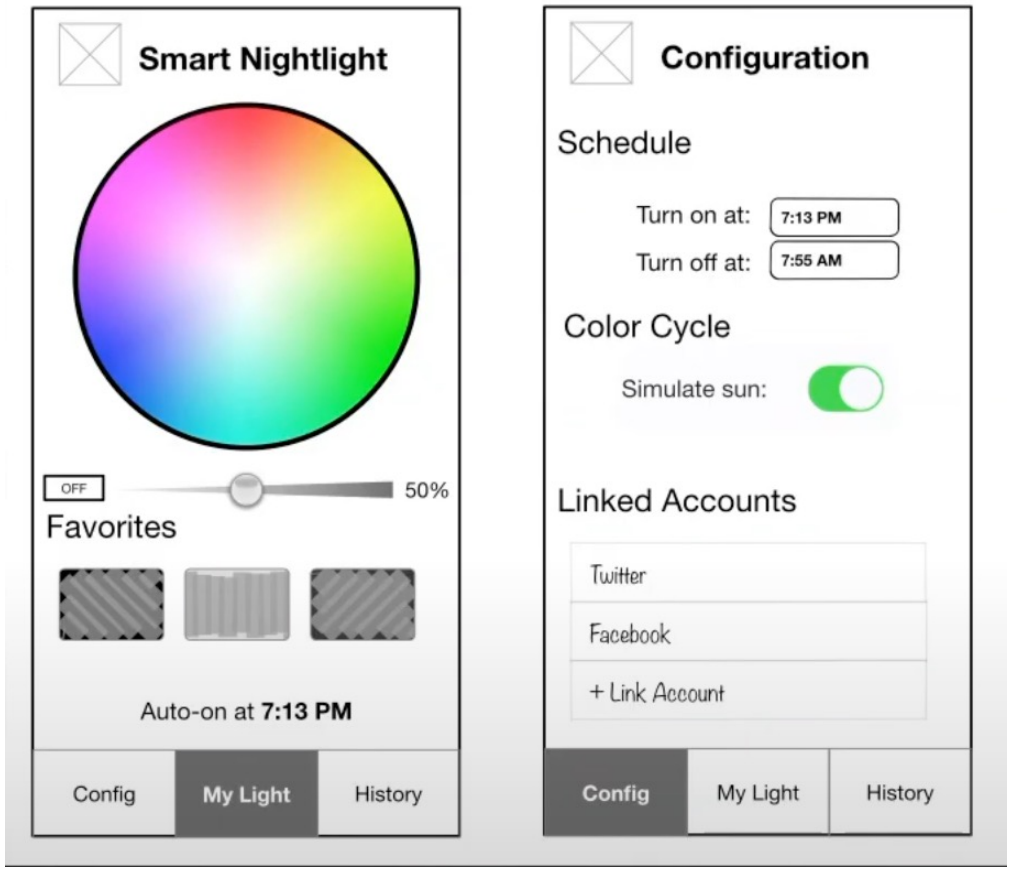
exploration validation

User Studies

- Identify and characterize difference between conditions
- Observe how manipulations affect user interactions

Adapted from MSU CSE 6663

Example



Usability Evaluations

Can the user turn the light on/off?

Can the user set the light's color?

User Studies

Does the amount of time it takes a user to adjust brightness vary between V1 and V2 of the interface?

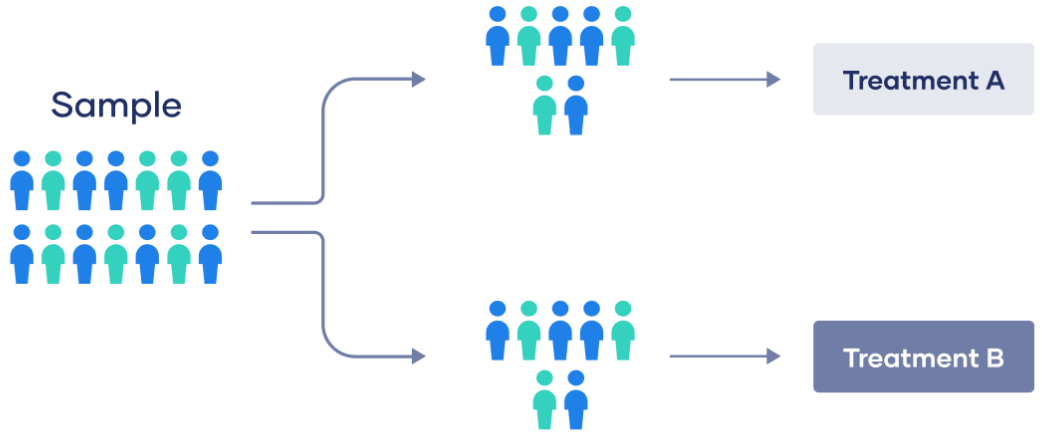
Adapted from MSU CSE 6663

Study Design

Assigning participants to conditions

Between subjects

- Each participant evaluates 1 condition
- You compare these groups
- Groups should be similar (verify!)



Within subjects

- Every participant tests everything
- Very important to randomize order!

Study Design Example

Between subjects

Within subjects

Independent vs dependent variables

Independent variables (cause)

- Manipulated by researcher (e.g., blue or orange)
- Participant characteristics (e.g., demographics)

Dependent variables (effect)

- Measured through study instruments
- Not controlled by researcher

Data to collect

Qualitative

- collected via observation, interview etc.
- collection itself does not constrain data
- usually non-numeric (e.g., language, videos)

Quantitative

- collected through some form of direct measurement
- summary of what happened (e.g., success, time, error)
- usually numeric or can be compared on numeric scales
- i.e., the dependent variables

Tip: start with qualitative data

Qualitative data gives good overview of where problems are

Quantitative data tells you something is wrong but not where to fix

The “Thinking Aloud” method (ask users to talk while performing tasks)

- tell us *what they are thinking*
- tell us *what they are trying to do*
- tell us *questions that arise as they work*
- tell us *things they read*

How to measure

Self-report

- reported directly by a participant
- subject to many biases
- common, useful and easy to administer

Behavioral

- measured through observation
- may not capture users experience fully

Physiological

- heart rate, skin conductance, etc.

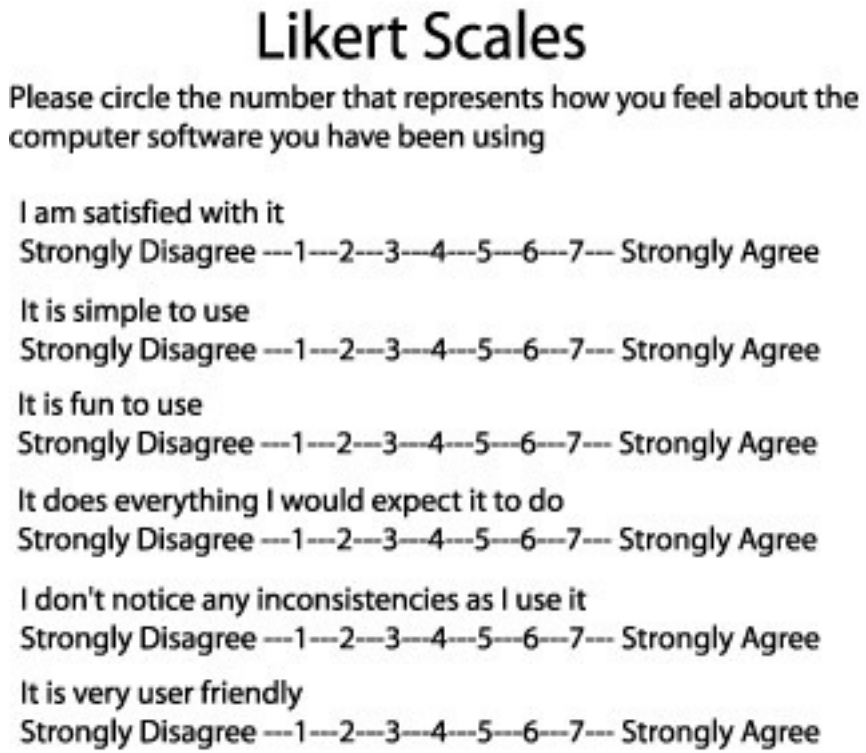
How to measure: self-report

Questionnaire:

- forced choices: yes/no
- scale responses
- open-ended responses

Designing questionnaires is an art in itself

- Odd or even number of values
- How many points on a scale?



Likert Scales

Please circle the number that represents how you feel about the computer software you have been using

I am satisfied with it
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is simple to use
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is fun to use
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It does everything I would expect it to do
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

I don't notice any inconsistencies as I use it
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is very user friendly
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

Use validated, commonly used self-report questionnaires:

- System Usability Scale (SUS)

How to measure: behavioral

Direct

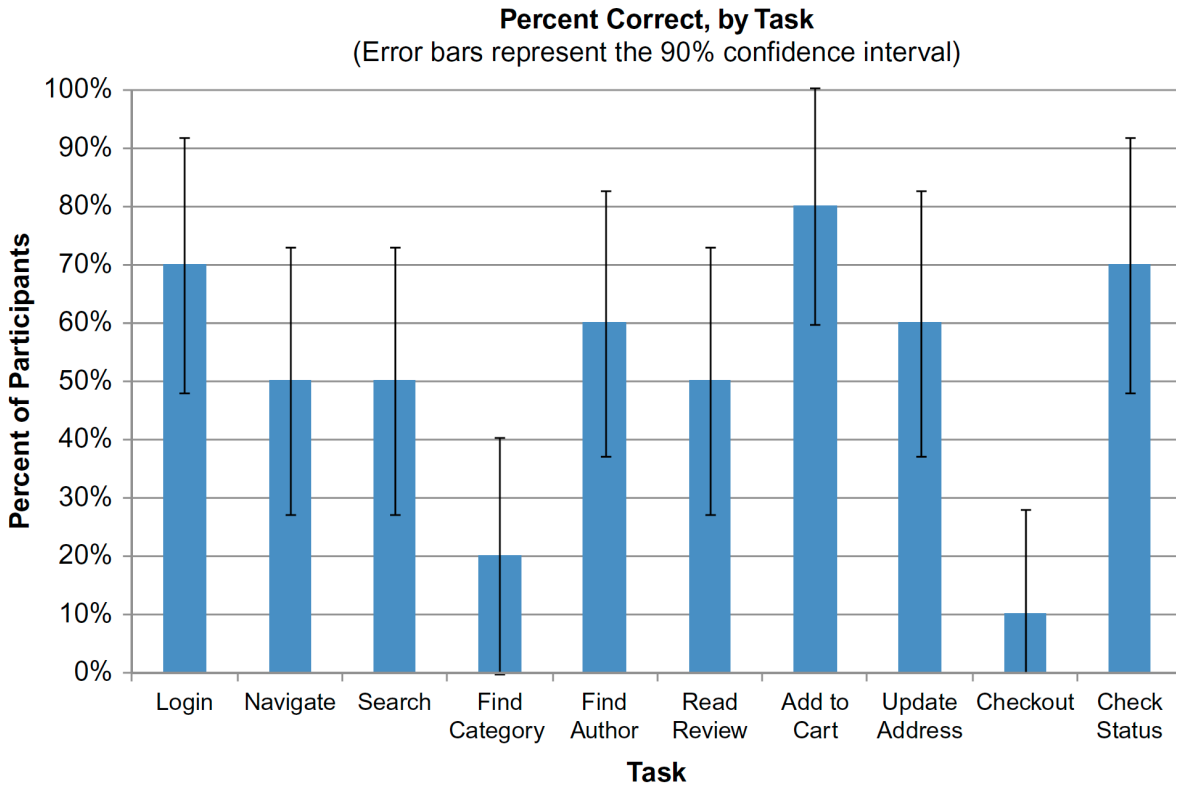
researcher observes and records participant activity

Indirect

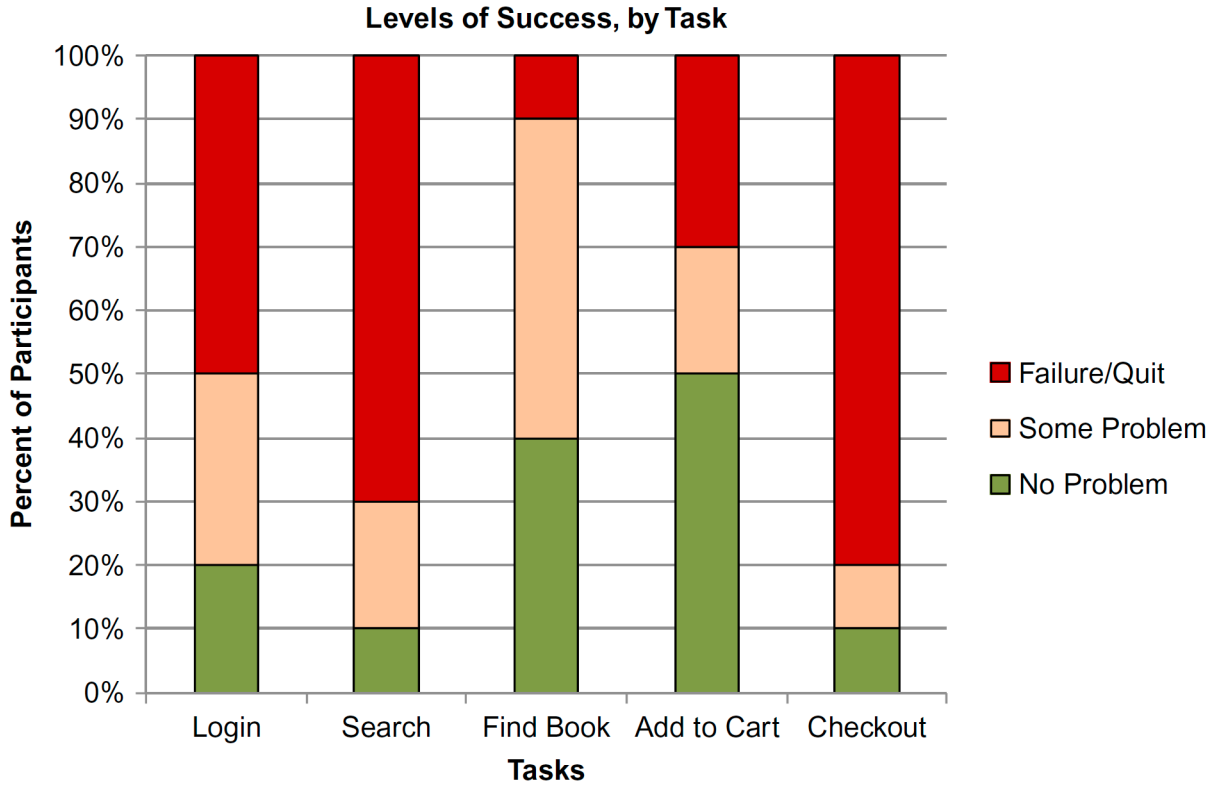
video recorded and analyzed later

participant records activity (e.g., diary entries)

How to measure: behavioral



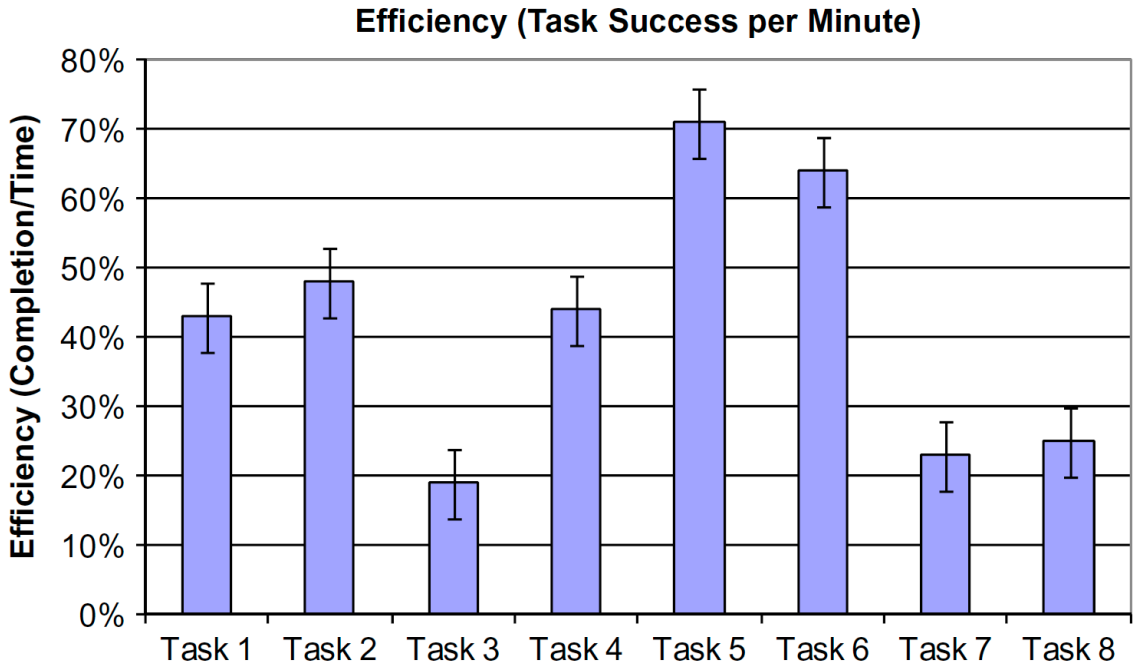
Binary success



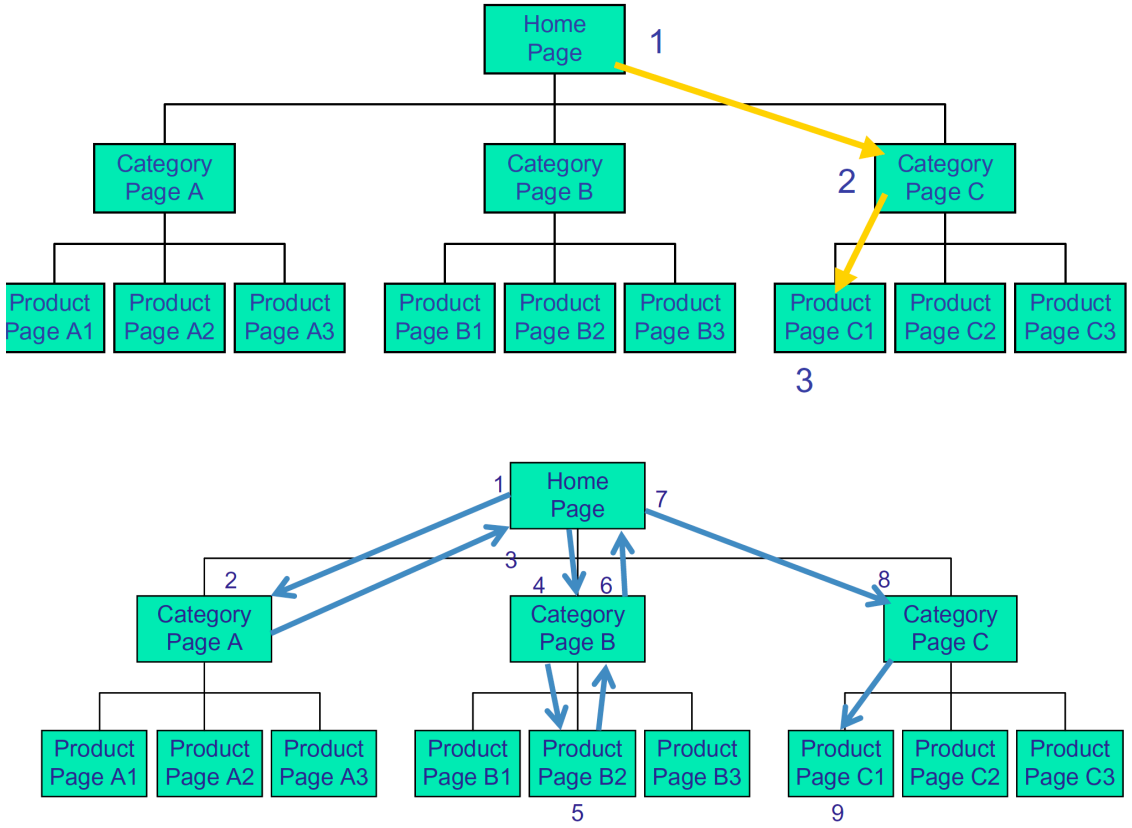
Levels of success

Source: Albert, Bill, and Tom Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

How to measure: behavioral



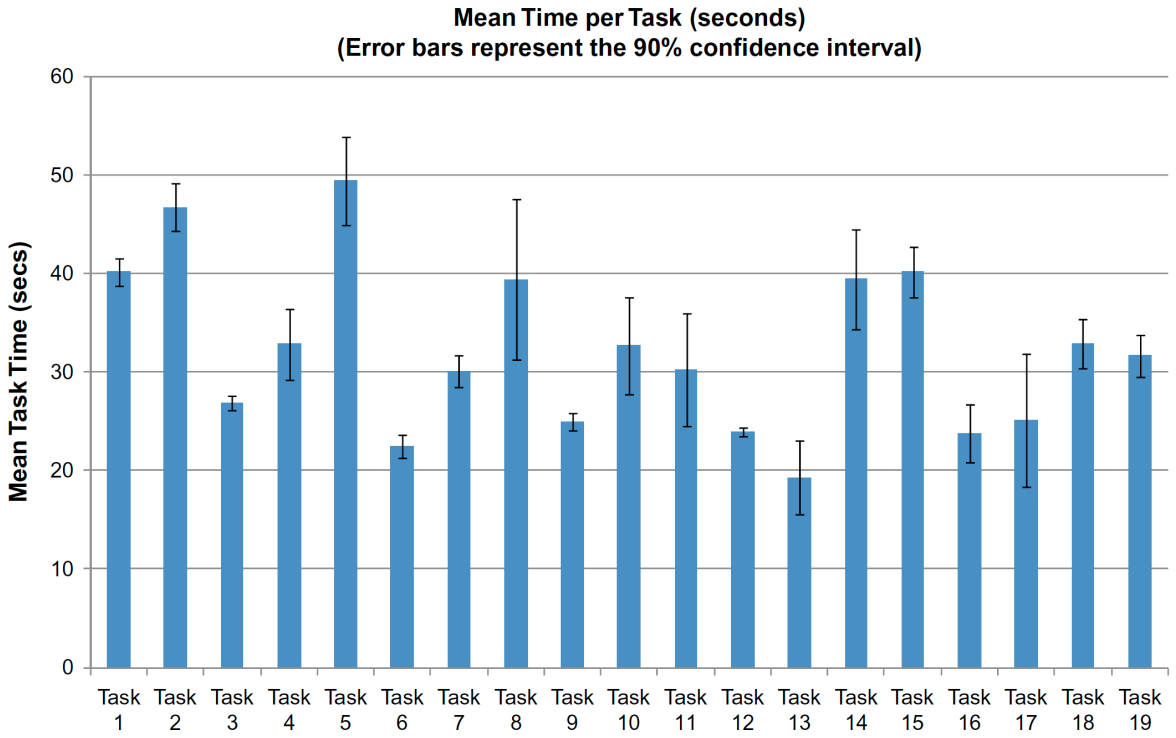
Efficiency



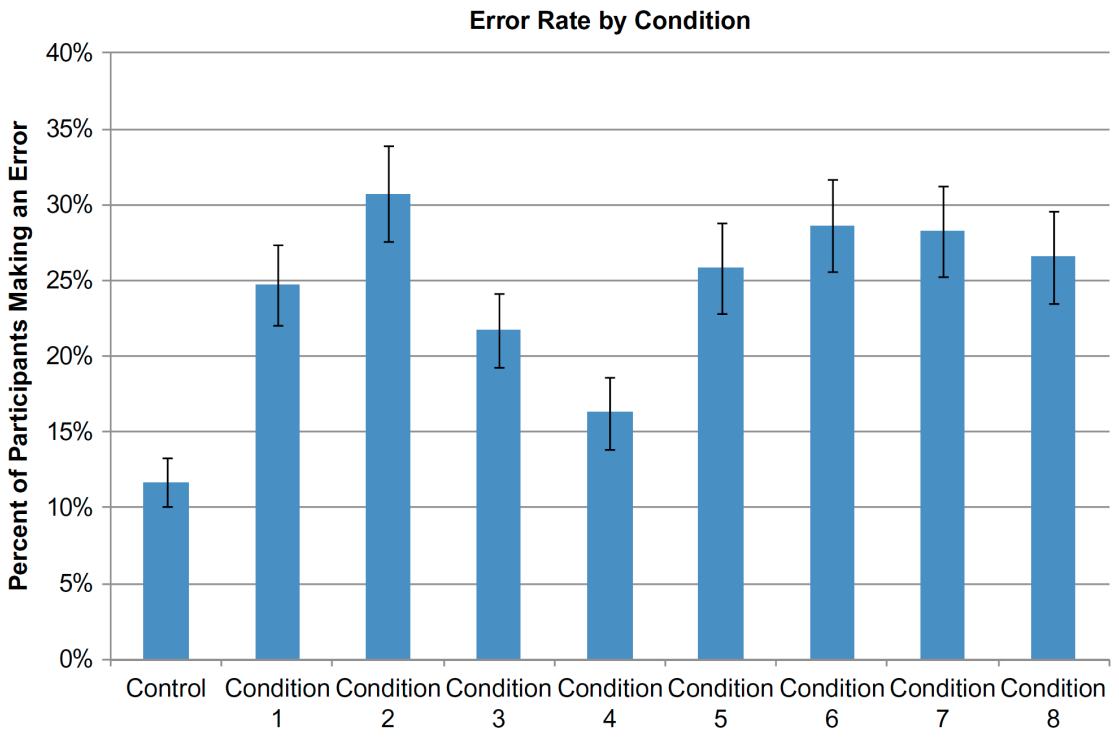
actions

Source: Albert, Bill, and Tom Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

How to measure: behavioral



Time on Task



Errors

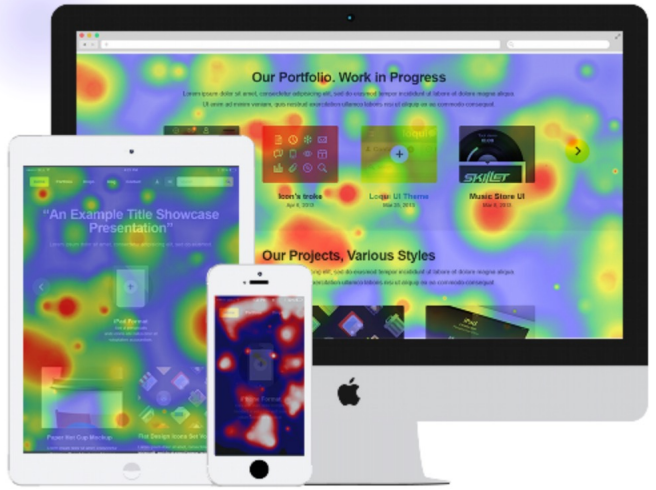
Source: Albert, Bill, and Tom Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

How to measure: behavioral

Engagement patterns

heatmap.js

Dynamic Heatmaps for the Web



heatmap.js is a lightweight, easy to use JavaScript library to help you visualize your three dimensional data!

https://www.patrick-wied.at/static/heatmapjs/

IRB: Institutional Review Board

https://oria.gatech.edu/irb

Experiments conducted at universities require ethical oversight

Consider: risk to participants, data privacy etc.

Protocols must be reviewed and approved by IRB

Status updates must be submitted to IRB

Study might quality for Exempt Review (but still need to apply)

In summary

Articulate the questions you want your evaluation to answer

Select the instruments you will use

Develop a clear protocol and adhere to it

Have a data analysis plan before conducting your study

Start early with IRB

Your tasks for next class

Paper review 1 is available on gradescope

- due Tuesday midnight

Paper presenters

- send slides to Kaushik by Wednesday noon
- archeologist role still open

Post on Piazza for teammates