大数据分析实践III

Crowdsourcing & Spreadsheet

Qiong Zeng (曾琼)

qiong.zn@sdu.edu.cn

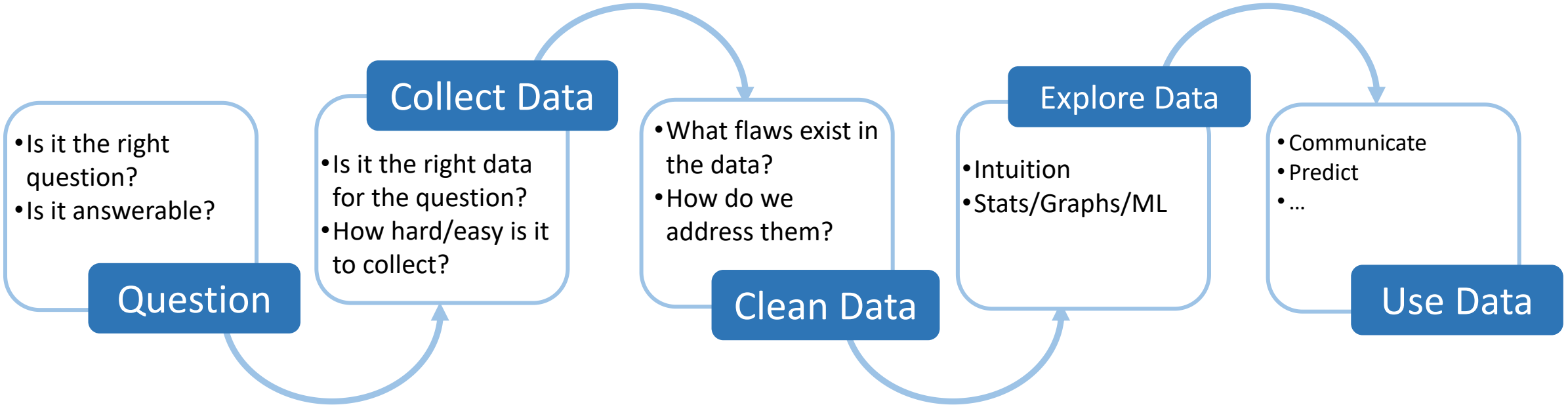
Research *is creative and systematic work undertaken to increase the stock of knowledge.*

Course Outline

6 Personal assignments

6 Group assignments

上课日期	授课内容	实验内容	周次
20240905	课程入门、大数据探索式分析	/	第一周
20240912	课程实践项目介绍、项目组队测试、项目经验谈	项目成员集结	第二周
20240919	科研实践入门、数据采样与降维	项目管理工具制定项目计划、**Pandas数据采样实践**	第三周
2024926	数据质量管理	Pandas数据质量实践	第四周
20241003	/	/	第五周
20241010	众包与电子表格	电子表格实践	第六周
20241017	可视化设计	可视化设计实践	第七周
20241024	统计分析方法与工具	统计方法实践	第八周
20241031	中期汇报（论文+项目进展）1	中期进展报告	第九周
20241107	中期汇报（论文+项目进展）2	BERT实践环境配置	第十周
20241114	机器学习方法与工具	BERT实践	第十一周
20241121	人机交互方法与工具	Canis/Cast/Libra实践	第十二周
20241128	普适计算	手机移动数据采集与分析	第十三周
20241205	大规模数据分析系统	SPARK实践	第十四周
20241212	如何撰写项目论文	大项目收尾	第十五周
20241219	项目结题报告1	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
•…

学习目标

可复述Crowdsourcing的基本概念、主要难点、以及应用场景，了解crowdDB基本思想

了解Spreadsheet的发展背景、传统spreadsheet的缺陷以及dataspread

Crowdsourcing

- Crowdsourcing is a sourcing model in which individuals or organizations obtain goods or services including ideas, voting, micro-tasks, and finances from a large, relatively open, and often rapidly evolving group of participants.
- At a livestock fair in 1906, villagers were invited to guess the weight of a ox. Galton observed that none of the almost 800 observers guessed the correct weight (1,178 pounds). Yet the average guess was amazingly close: 1,179 pounds!

Open Question

Points: 10

设置

Games: How many pennies are in this jar?

正常使用主观题需2.0以上版本雨课堂

作答

Open Question

Points: 10

设置

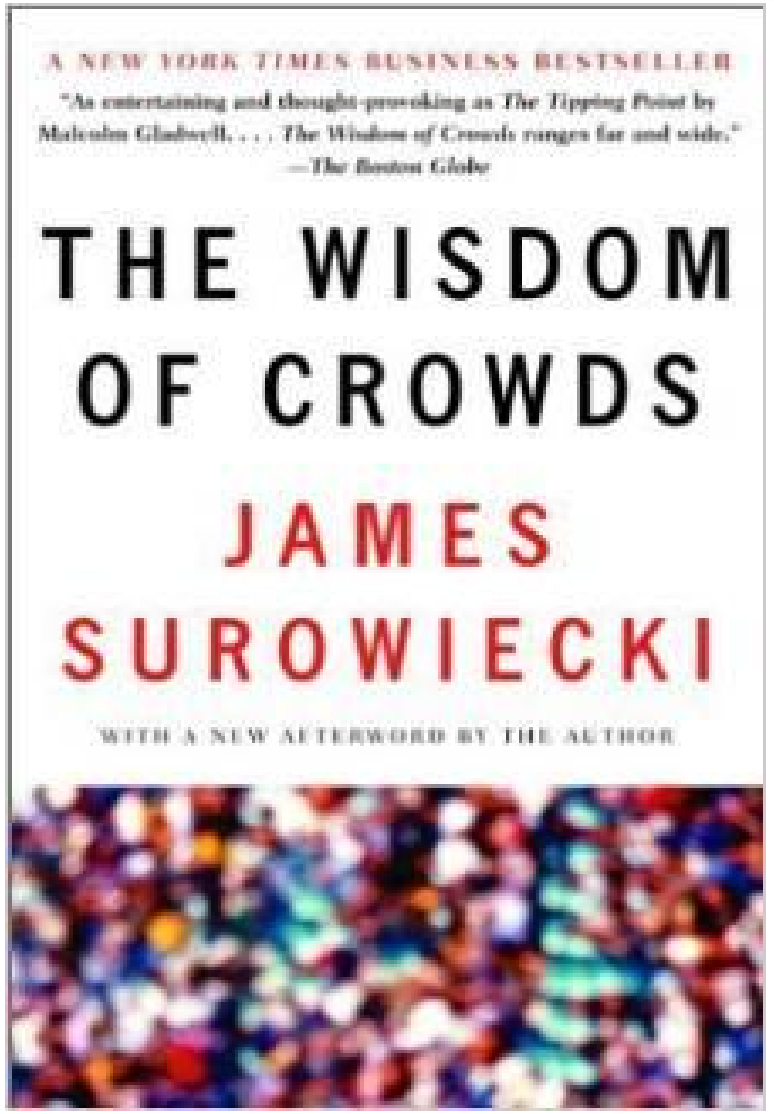
Games: How many pennies are in this jar?

正常使用主观题需2.0以上版本雨课堂

作答

Wisdom of Crowds

- When the opinions are independent: this avoids groupthink.
- When the crowd consists of people with diverse knowledge / methods.
- When the problem is in a domain where the crowd does not need specialized knowledge.
- Opinions can be fairly aggregated.



Crowdsourcing

Crowdsourcing is an important source of building analytic models, especially for the human perceptive tasks.

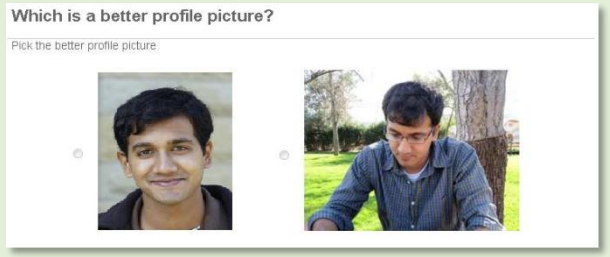
Why? Many tasks done better by humans

Pick the “cuter” cat


Is this a photo of a car?

How? We use an internet marketplace

Requester: Aditya Reward: 1$ Time: 1 day



- ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), having an average of over five hundred images per node.
- More than 14 million images have been hand-annotated by the project to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided. Amazon Mechanical Turk
- **CVPR 2019 Longuet-Higgins奖:** fundamental contributions in cv that have withstood the test of time

Crowdsourcing

Several types

- Microwork-small usually quick low paid tasks that require minimal skill to complete

Amazon Turk is the most common example; people can join or exit marketplace easily; low wages

- Macrowork-larger more complex tasks that require specific expertise

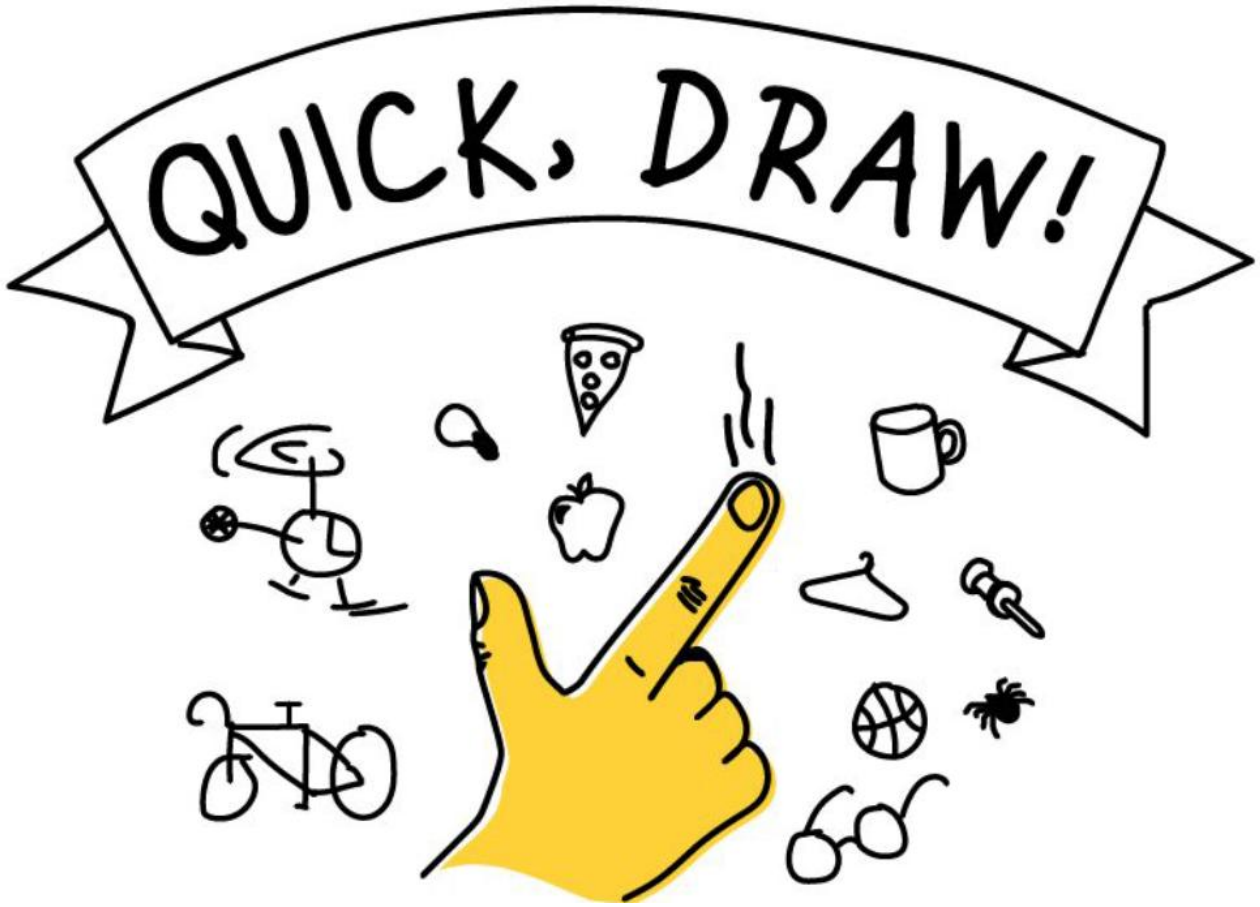
More specialized and longer term engagements; Upwork, kuaima, Johnny Cash Project

- Implicit-tasks that are not clearly described ascrowd work or have dual application(recaptcha)

Crowdsourced games, reCaptcha, GPS aggregation, Quick Draw

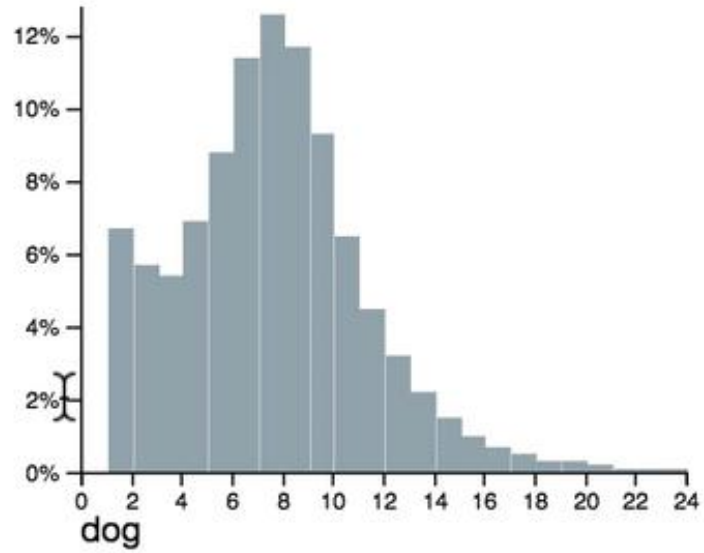
Solving hard problems for Computers

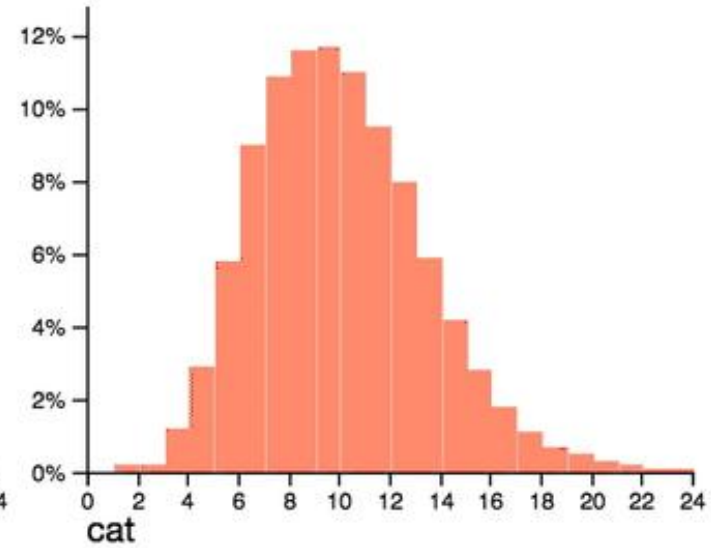
Crowdsourcing

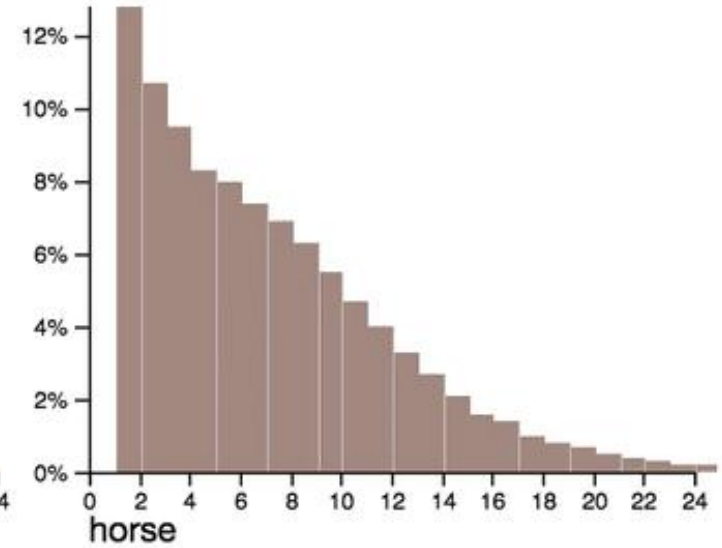


Google Creative Lab
https://quickdraw.withgoogle.com/


```
word: "cat"
countrycode: "JP"
timestamp: "2017-03-15 21:41:50.79245 UTC"
recognized: true
key_id: "6429928587264000"
strokes: ▾Array(21) [
 0: ▾Array(8) [
   0: ▸Object {x: 348.75, y: 185.25}
   1: ▸Object {x: 344.25, y: 168.75}
   2: ▸Object {x: 350.25, y: 155.25}
   3: ▸Object {x: 384.75, y: 99.75000000000001}
   4: ▸Object {x: 399.75, y: 83.25000000000001}
   5: ▸Object {x: 407.25, y: 87.75000000000001}
   6: ▸Object {x: 414.75, y: 120.75000000000001}
   7: ▸Object {x: 423.75, y: 182.25}
 ]
 1: ▾Array(3) [
   0: ▸Object {x: 531.75, y: 89.25000000000001}
   1: ▸Object {x: 512.25, y: 164.25}
   2: ▸Object {x: 510.75, y: 183.75}
 ]
 2: ▾Array(6) [
   0: ▸Object {x: 540.75, y: 83.25000000000001}
   1: ▸Object {x: 546.75, y: 81.75000000000001}
   2: ▸Object {x: 554.25, y: 92.25000000000001}
   3: ▸Object {x: 561.75, y: 134.25}
   4: ▸Object {x: 566.25, y: 185.25}
   5: ▸Object {x: 561.75, y: 207.75}
 ]
 3: ▾Array(2) [
   0: ▸Object {x: 537.75, y: 174.75}
   1: ▸Object {x: 554.25, y: 176.25}
 ]
 4: ▾Array(5) [
   0: ▸Object {x: 378.75, y: 174.75}
```

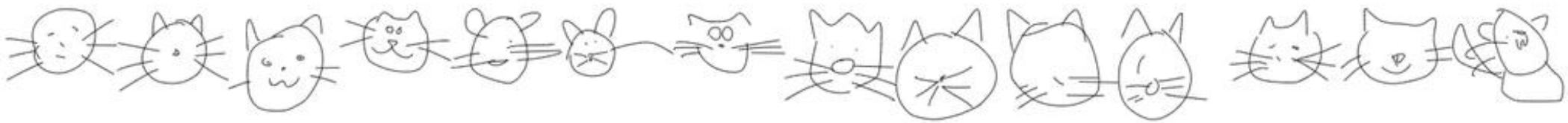






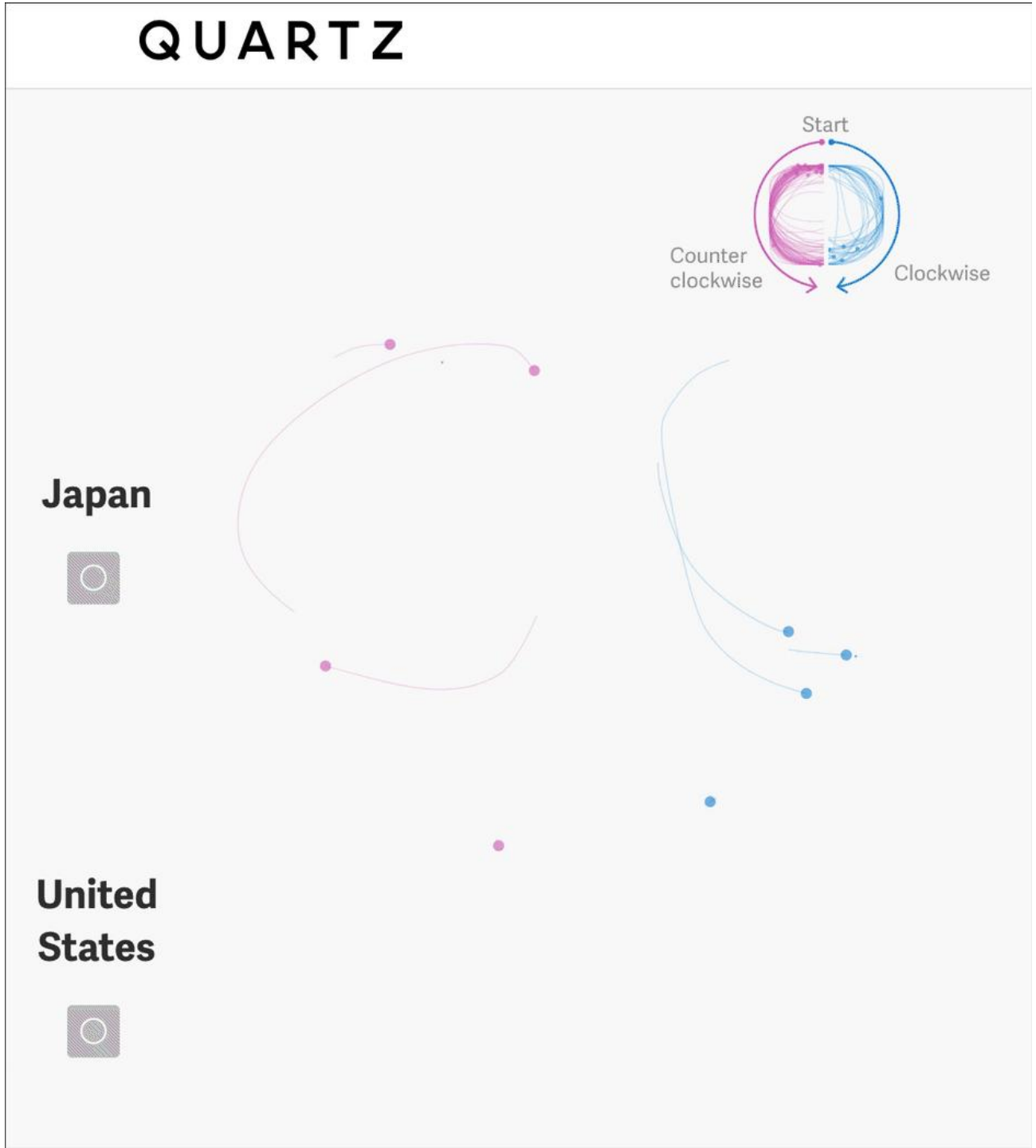
Dog drawn with 11 strokes

Cat drawn with 11 strokes



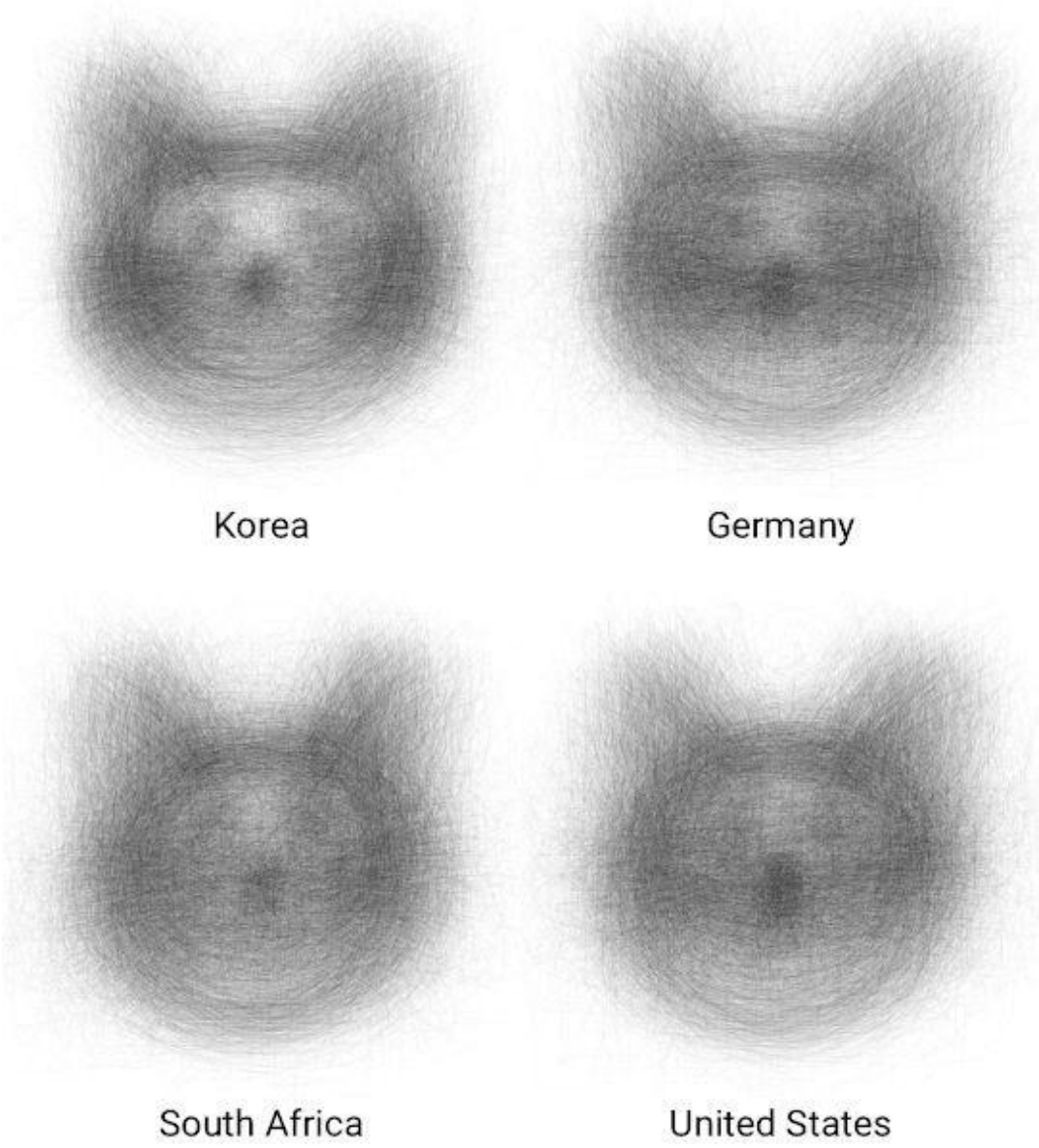
Horse drawn with 11 strokes

http://vallandingham.me/quickdraw/



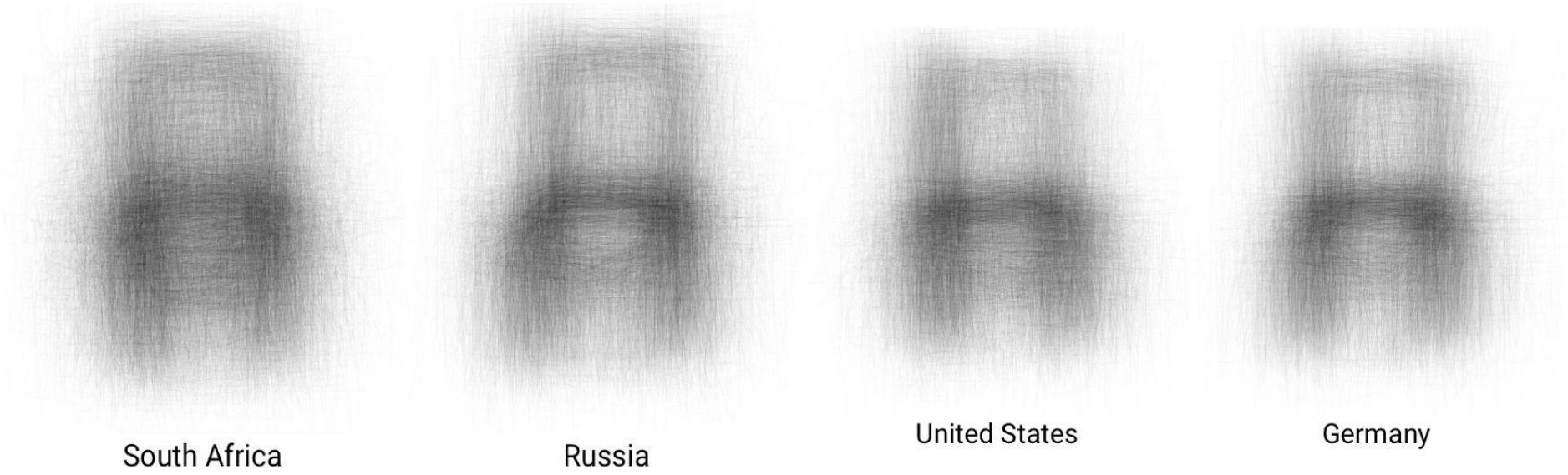
https://qz.com/994486/the-way-you-draw-circles-says-a-lot-about-you/

When things look alike across cultures



Machine Learning for Visualization
Let's Explore the Cutest Big Dataset
Ian Johnson

And when they don’t



Visual Averages by Country
Kyle McDonald

Crowdsourcing Platform

- Crowsourcing marketplaces: web service for connecting requesters with workers -- access to on-demand workforce (requesters), new work style unbound by time or place (workers)

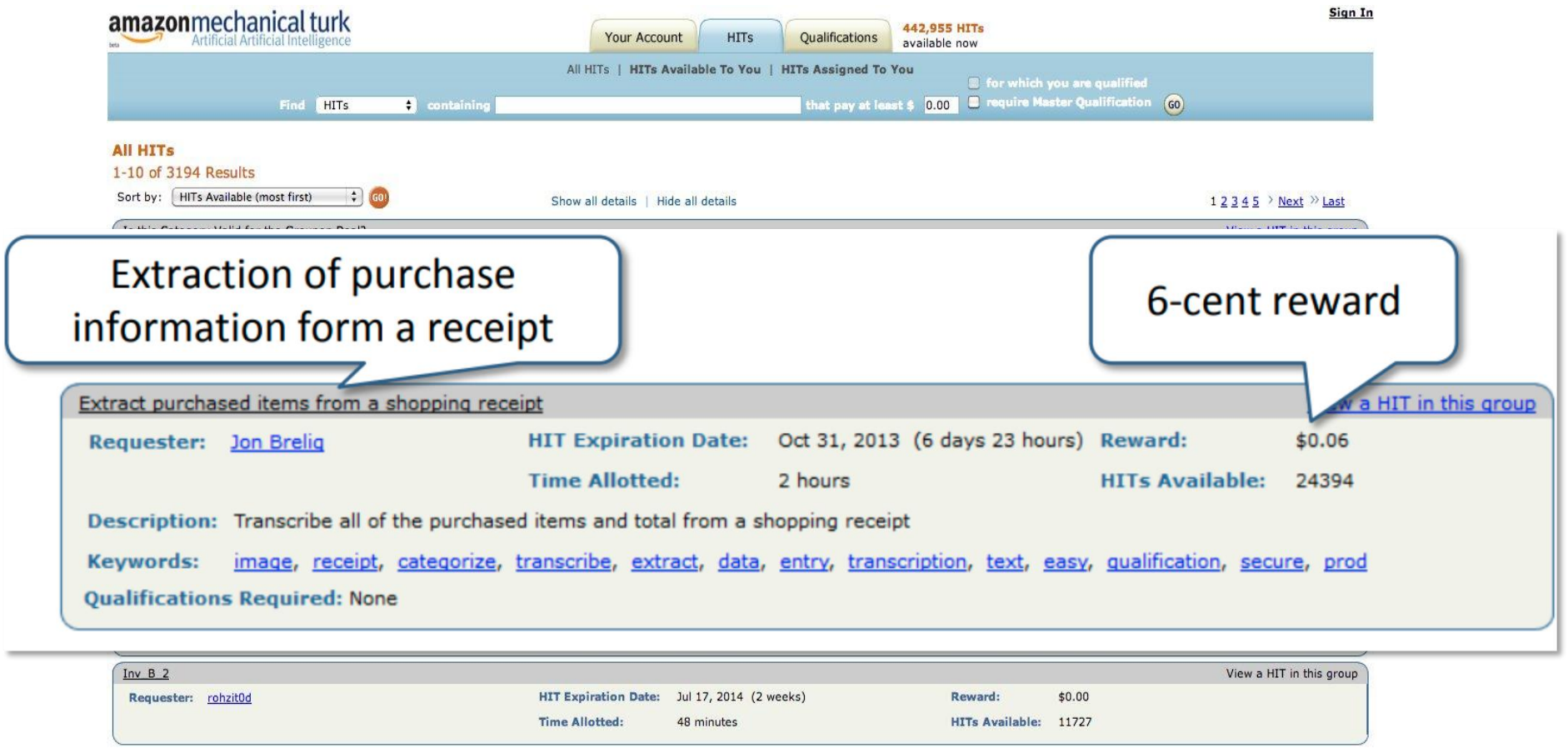
- Emergence of online crowd-labor marketplaces (AMT, Upwork, Clickworker)

- Signed up to work for AMT: 500,000
- 75% of AMT workers are American
- 50% of workers earn below U.S. federal minimum wage
- Typical weekly wage: 79$
- Minimum payment for a task: 0.01$
- Estimated annual gross revenue: > 120 millions
- > 1,200 people or organizations were posting jobs to AMT in 2015

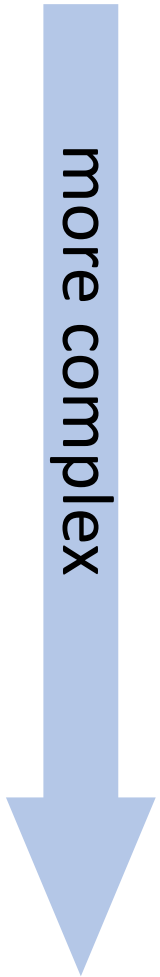
Image: Erik Underwood / TechRepublic

Crowdsourcing Platform

Amazon Mechanical Turk (AMT).



Crowdsourcing – Task Types



- **Very easy tasks**
  Image labeling, entity resolution, format checking..

Output format: Yes/No

- **Easy tasks**
  Restaurant reviews, web service tests, …

Output format: Multiple-choice or short sentences

Microtasks: Popular in computer science

- **Tasks requiring some expertise**
  Logo design, report writing,…

Output format: Images or sentences

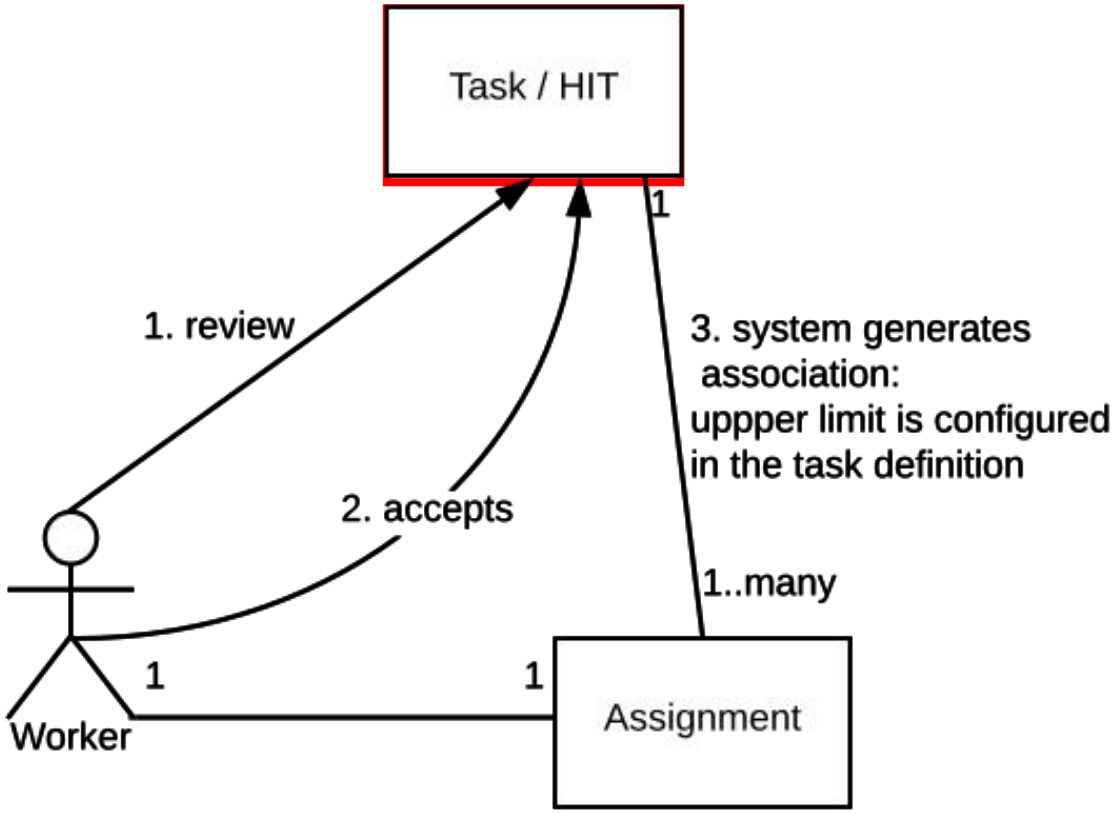
- **Complex tasks requiring high level of expertise**
  Web/software developments, professional work, ..

Output format: system, codes, documents

Crowdsourcing – AMT Basics

- HIT (human intelligent task): the smallest entity of work that could be accepted by a worker

- Assignment: HIT can be replicated into multiple assignments. A worker can process at most a single assignment per HIT.



Crowdsourcing – AMT Basics

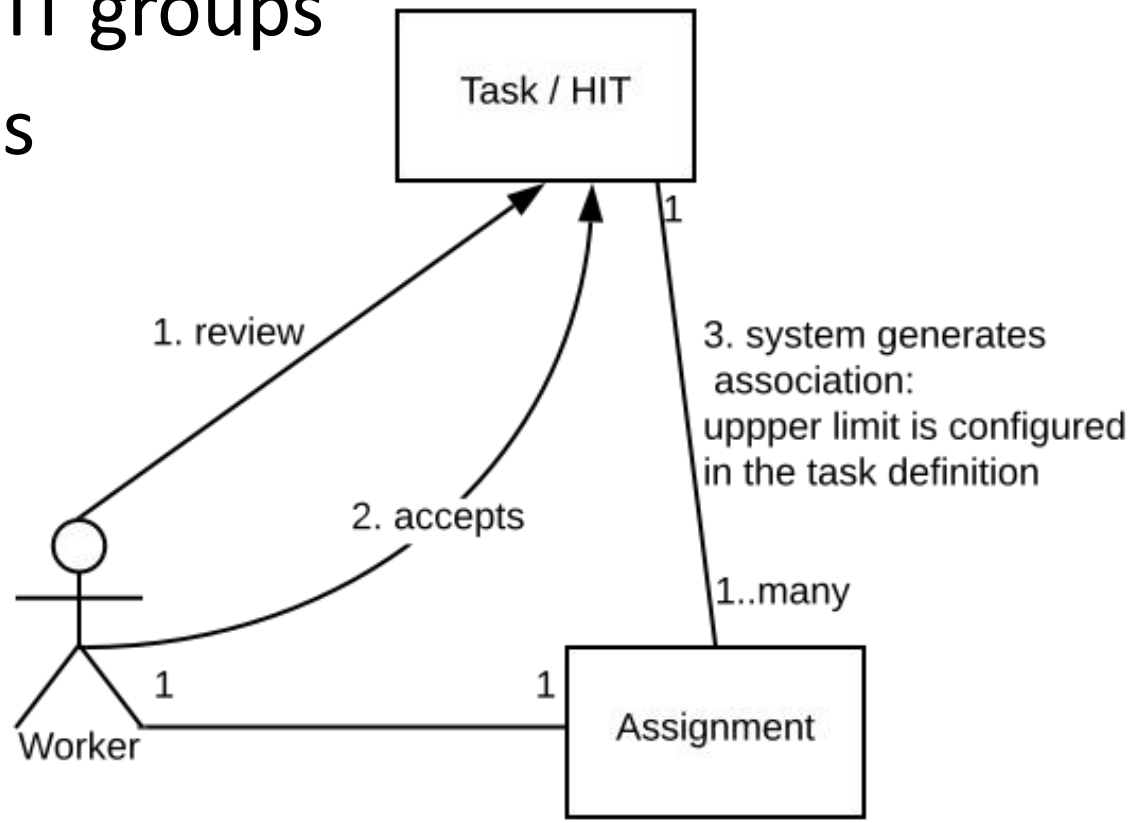
```
HIT:{
        HITId:"123RVWYBAZW00EXAMPLE",
        HITTypeId:"T100CN9P324W00EXAMPLE",
        HITTypeId:"2005-06-30T23:59:59",
        HITStatus:"Assignable",
        MaxAssignments:"5",
        AutoApprovalDelayInSeconds:"86400",
        LifetimeInSeconds:"86400",
        AssignmentDurationInSeconds:"300",
        Reward:{
                Amount:"25"
                CurrencyCode:"USD"
                FormattedPrice:"$0.25"
        },
        Title:"Location and Photograph Identification",
        Description:"Select the image that best represents...",
        Keywords:"location, photograph, image, identification, opinion",
        Question:{
                QuestionForm:[XML-encoded Question data]
        },
        QualificationRequirement:{
                QualificationTypeId:"789RVWYBAZW00EXAMPLE",
                Comparator:"GreaterThan",
                Value:"18"
        },
        HITReviewStatus:"NotReviewed"
}
```

```
Assignment:{
        AssignmentId: "123RVWYBAZW00EXAMPLE456RVWYBAZW00EXAMPLE",
        WorkerId:"AZ3456EXAMPLE",
        HITId:"123RVWYBAZW00EXAMPLE",
        AssignmentStatus:"Submitted",
        Deadline: "2005-12-01T23:59:59Z",
        AcceptTime: "2005-12-01T12:00:00Z",
        SubmitTime: "2005-12-07T23:59:59Z",
        Answer:{
                QuestionFormAnswers:[XML-encoded Answer data]
        }
  }
```

Crowdsourcing – AMT Basics

- Requester post HITs
- AMT post them into compatible HIT groups
- Worker search, accept and process
- Requester approve or reject

For each task completed requester pay the predefined reward, bonus and commission to Amazon.



Crowdsourcing Platform

- Amazon Mechanical Turk (AMT).
- AMT APIs:
  - **createHIT**(title, description, question, keywords, reward, duration, maxAssignments, lifetime)→ **HitID**
  - **getAssignmentsForHIT**(HitID)→**list**(asnId, workerId , answer)
  - **approveAssignment**(asnID)
  - **rejectAssignment**(asnID)
  - **forceExpireHIT**(HitID)

https://workersandbox.mturk.com/

https://github.com/jcjohnson/simple-amt/blob/master/README.md

At a high level…

- I (and other requesters) post my tasks to a marketplace
  - one such marketplace is Mechanical Turk, but there are 30+ marketplaces
- Workers pick tasks that appeal to them
- Work on them
- I pay them for their work

Pay anywhere from a few cents to dollars for each task; get the tasks done in any time from a few seconds to minutes

Why should we care?

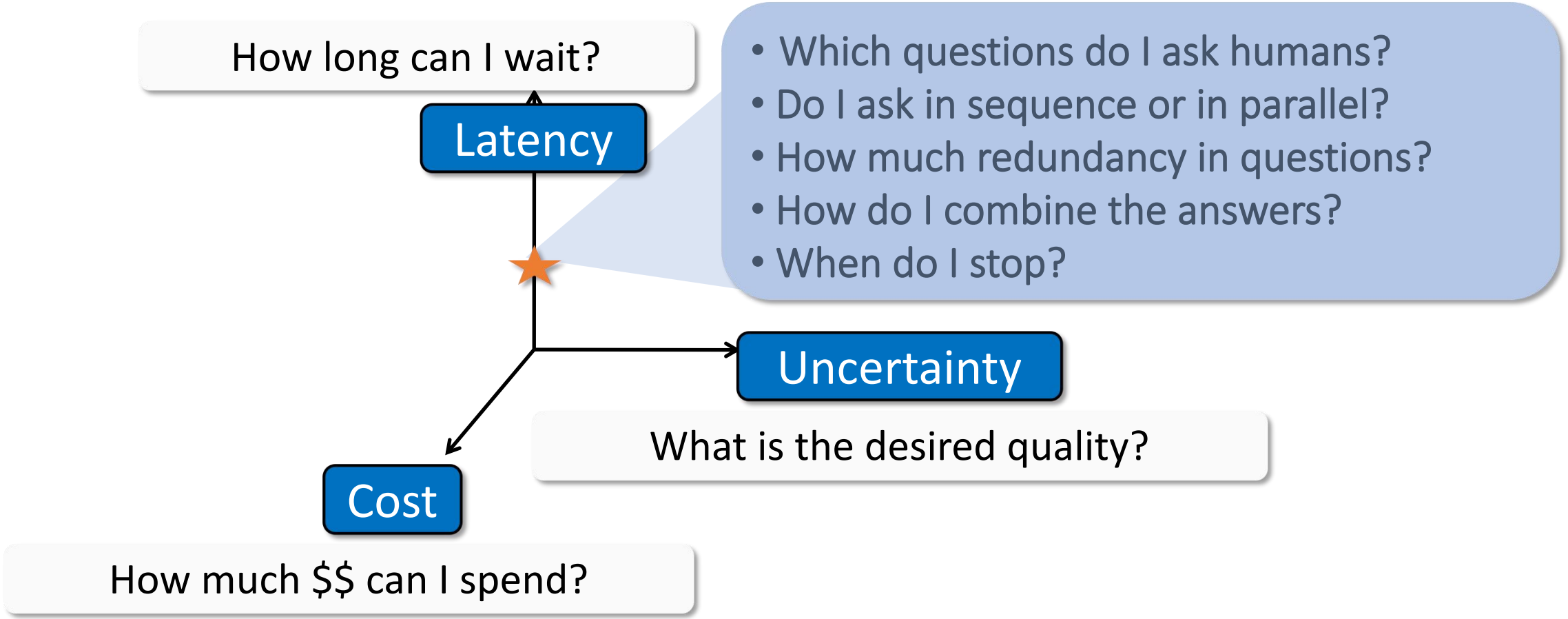
- Most major companies spend millions of $$$ on crowdsourcing every year
  - This includes Google, MS, Facebook, Amazon
- Represents our only viable option for understanding **unstructured data**
- Represents our only viable option for generating training data @ scale

OK, so why is this hard?

- People need to be **paid**
- People take **time**
- People make **mistakes**

- And these three issues are correlated:
  - If you have more money, you can hire more workers, and thereby increase accuracy
  - If you have more time, you can pay less/hire more workers, and thereby increase accuracy/reduce costs
  - …

Fundamental Tradeoffs

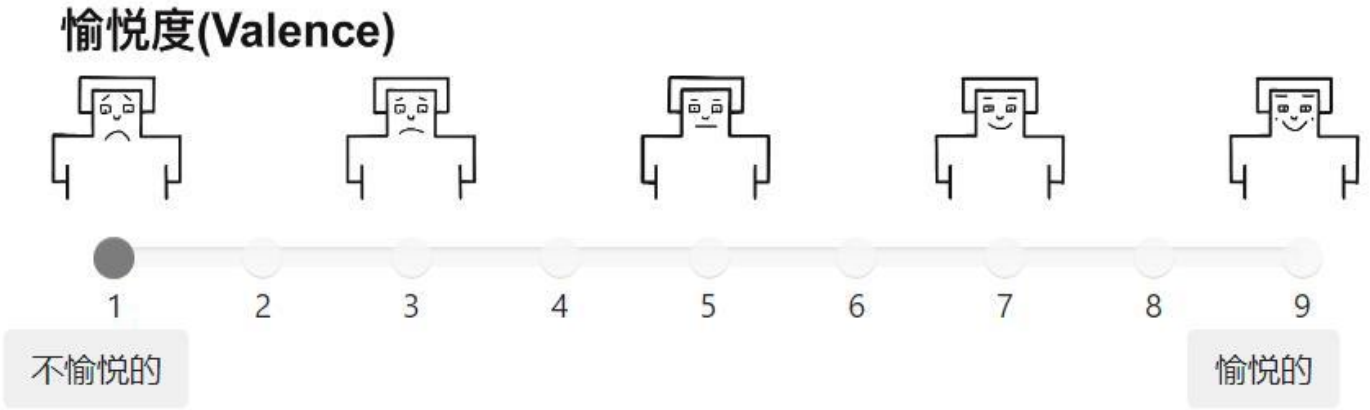


Open Question

Points: 10

设置

假设你在AMT平台邀请工人阅读图片，并标注出图片中所包含的基本情感。这是一个主观感知的任务，我们如何才能通过算法判断工作是否随意作答、没有认真对待这项任务呢?



正常使用主观题需2.0以上版本雨课堂

作 答

Need to Revisit Basic Algorithms

- Given that humans are complicated unlike computer processors, we need to revisit even **basic data processing algorithms where humans are "processing data"**
  - Max: e.g., find the best image out of a set of 1000 images
  - Filter: e.g., find all images that are appropriate to all
  - Categorize: e.g., find the category for this image/product
  - Cluster: e.g., cluster these images
  - Search: e.g., find an image meeting certain criteria
  - Sort: e.g., sort these images in terms of desirability
- Using human unit operations:
  - Comparisons, Ranking, Rating

Goal: Design **efficient** crowd algorithms

Carnegie Mellon University
https://15799.courses.cs.cmu.edu › static › papers PDF

CrowdDB: Answering Queries with Crowdsourcing

by MJ Franklin · 2011 · Cited by 877 — ABSTRACT. Some **queries** cannot be **answered** by machines only. Processing such **queries** requires human input for providing information that is.
12 pages

CrowdDB: Answering Queries with Crowdsourcing

Michael J. Franklin
AMPLab, UC Berkeley
franklin@cs.berkeley.edu

Donald Kossmann
Systems Group, ETH Zurich
donaldk@inf.ethz.ch

Tim Kraska
AMPLab, UC Berkeley
kraska@cs.berkeley.edu

Sukriti Ramesh
Systems Group, ETH Zurich
ramess@student.ethz.ch

Reynold Xin
AMPLab, UC Berkeley
rxin@cs.berkeley.edu

on courtesy to https://cs598.github.io/index.html#papers

Motivation of CrowdDB

- Two reasons why present DB systems won't do:
  - **Closed world assumption**
    - Get human help for finding new data
  - Very literal in processing data
    - SELECT marketcap FROM company
      WHERE name = "IBM"
    - SELECT title FROM paper
      ORDER BY novel_idea LIMIT 10

Get the best of both worlds:

human power for processing and getting data

traditional systems for heavy lifting/data manipulation

Goals

Develop CrowdDB:

- A relational query processing system ➔ maintain SQL semantics.
- Rely on traditional RDBS to do the heavy lifting data manipulation.
- Extend SQL to enable queries that involve human computation.

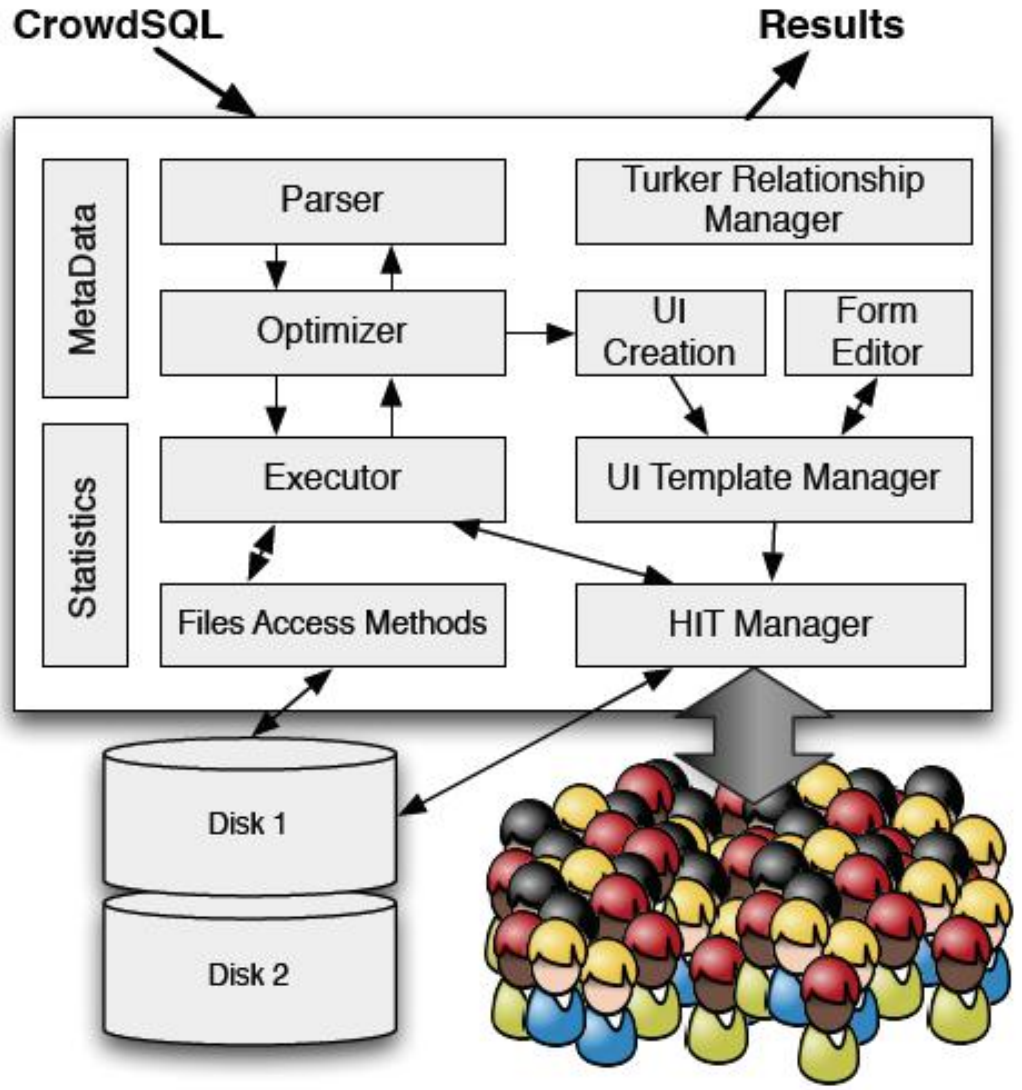
Issues in building CrowdDB

- Performance and variability:
  - Humans are slow, costly, variable, inaccurate
- Task design and ambiguity:
  - Challenging to get people to do what you want
- Affinity / Learning
  - Workers develop relationships with requesters, skills
- Open world
  - Possibly unbounded answers

Using human unit operations: Comparisons, Ranking, Rating

Overview of CrowdDB

- An application issue requests using CrowdSQL.
- The complexities of dealing with the crowd are encapsulated by CrowdDB.
- Results obtained from the crowd can be stored in the database for future use.



At a High Level

- Modifications to QL: CrowdSQL
- Automatic UI generation
- Automatic interaction with marketplace
- Storing data for future use

Modifications to SQL

- Special keyword: CROWD
- Used in two ways
- First: crowdsourced columns
  - CREATE TABLE Department (
    university STRING,
    name STRING,
    url CROWD String,
    phone STRING,
    primary key (university, name));

CROWD attribute cannot be PK

Modifications to SQL

- Crowdsourced Tables
  - CREATE CROWD TABLE Profs (
    name STRING PRIMARY KEY,
    email STRING UNIQUE,
    university STRING,
    department STRING,
    FOREIGN KEY (university, department)
    REF Department (university, name));

  Still need a PK

CrowdSQL – Incomplete Data

- Use new value type CNULL to indicates that a value should be crowdsourced when it is first used .
- CNULL is the default value of any CROWD column.
- CNULL values are generated as a side-effect of INSERT statements:

```
INSERT INTO Department(university, name)
VALUES ("UC Berkeley", "EECS");
```

university	name	url	phone
UC Berkeley	EECS	CNULL	NULL

Allow crowdsourcing as a side-effect of query processing:

```
SELECT url FROM Department
WHERE name = "Math";
```

CrowdSQL – Subjective Comparisons

- Use two new built in functions: CROWDEQUAL,CROWDORDER.

1. CROWDEQUAL (~=)

```
SELECT profile FROM department
WHERE name ~= "CS";
```

2. CROWDORDER

```
CREATE TABLE picture (
        p IMAGE,
        subject STRING);

SELECT p FROM picture
WHERE subject = "Golden Gate Bridge"
ORDER BY CROWDORDER(p,
"Which picture visualizes better %subject");
```

CrowdSQL in Practice

Practical issues that limit the usage of CrowdSQL:

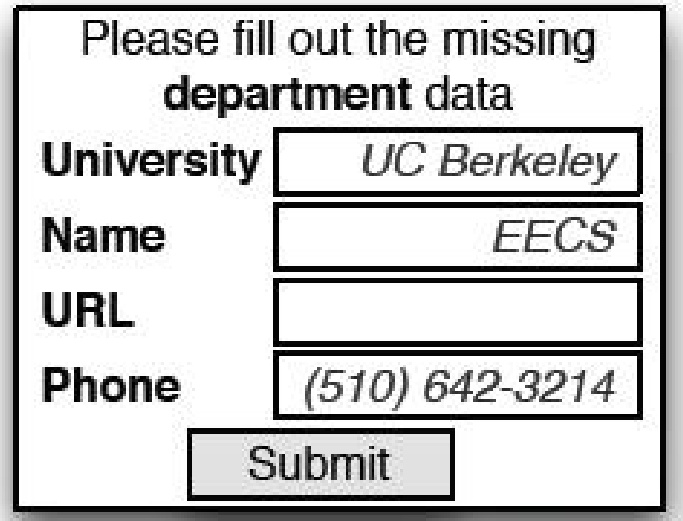
1. Cost and response time of queries can be unbounded.
2. Lineage: track source of data to take actions.
3. Cleansing of crowdsourced data → entity resolution.

User Interface Generation

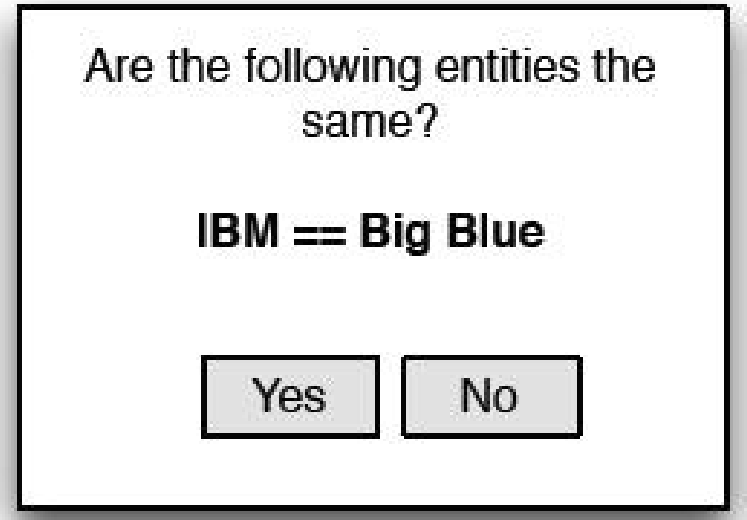
- Automatically generates user interfaces for incomplete information and subjective comparisons.
- Create templates at a compile-time.
- Templates are instantiated at a run-time for each tuple.
- Templates can be edited for customized instruction.

User Interface Generation

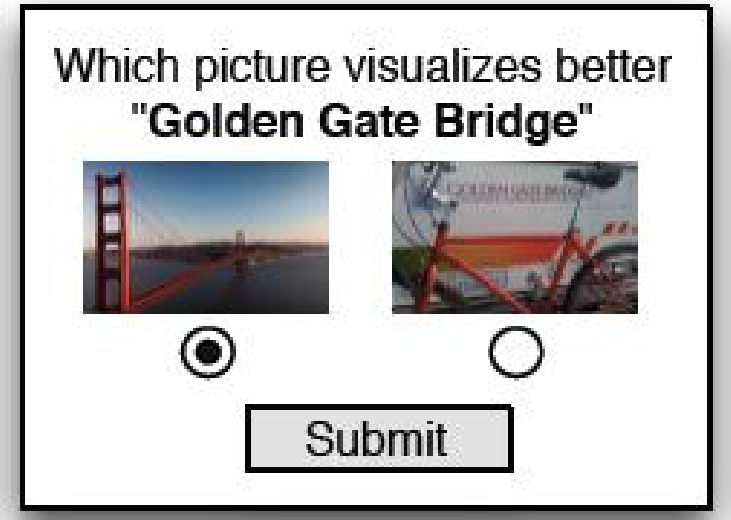
Basic Interface:



(a) Crowd Column & Crowd Tables w/o Foreign Keys



(b) CROWDEQUAL



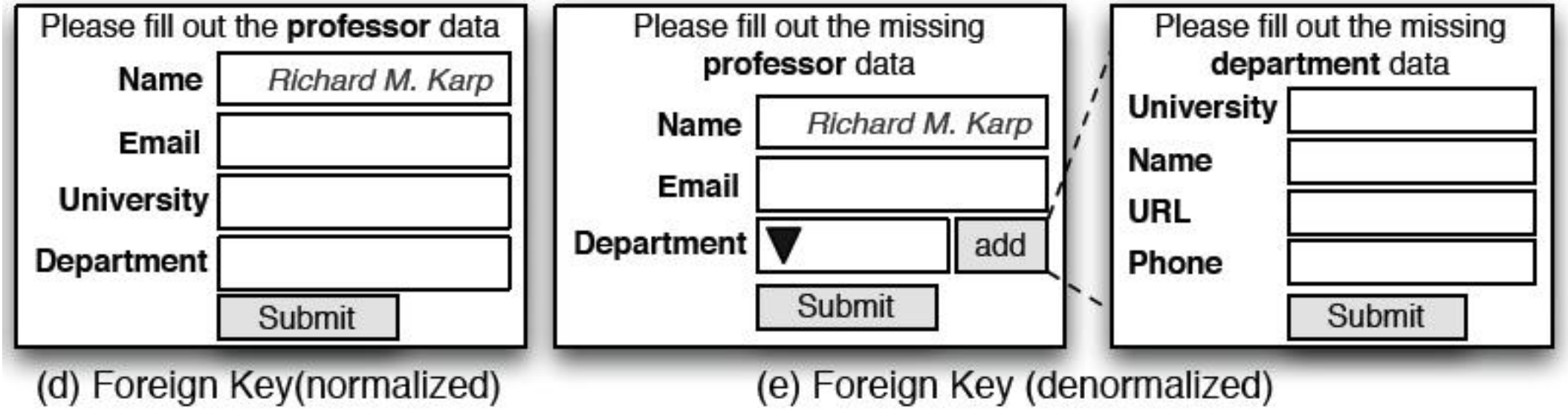
(c) CROWDORDER

- Two types of optimization:
  1. Batch several tuples.
  2. Prefetching of attributes of the same tuple.

User Interface Generation

Multi-relational interface:

- Foreign-key references a non-crowdsourced table:
  - A drop-down box of possible foreign keys.
  - Ajax-based “suggest” function.
- Foreign-key references a crowdsourced table:
  - Normalized interface ➔ suggest function can be used to avoid entity resolution problem.
  - Denormalized interface.



(d) Foreign Key(normalized)

(e) Foreign Key (denormalized)

Query Processing – Crowd Operators

- Three crowd operators:
  1. **CrowdProbe:** Crowdsources missing information of CROWD and new tuples.
  2. **CrowdJoin:** At least one table is a crowdsourced table.
  3. **CrowdCompare:** Implement the CROWDEQUAL and CROWDORDER function.
- Quality control is carried out by a majority vote.

Query Processing – Physical Plan Generation

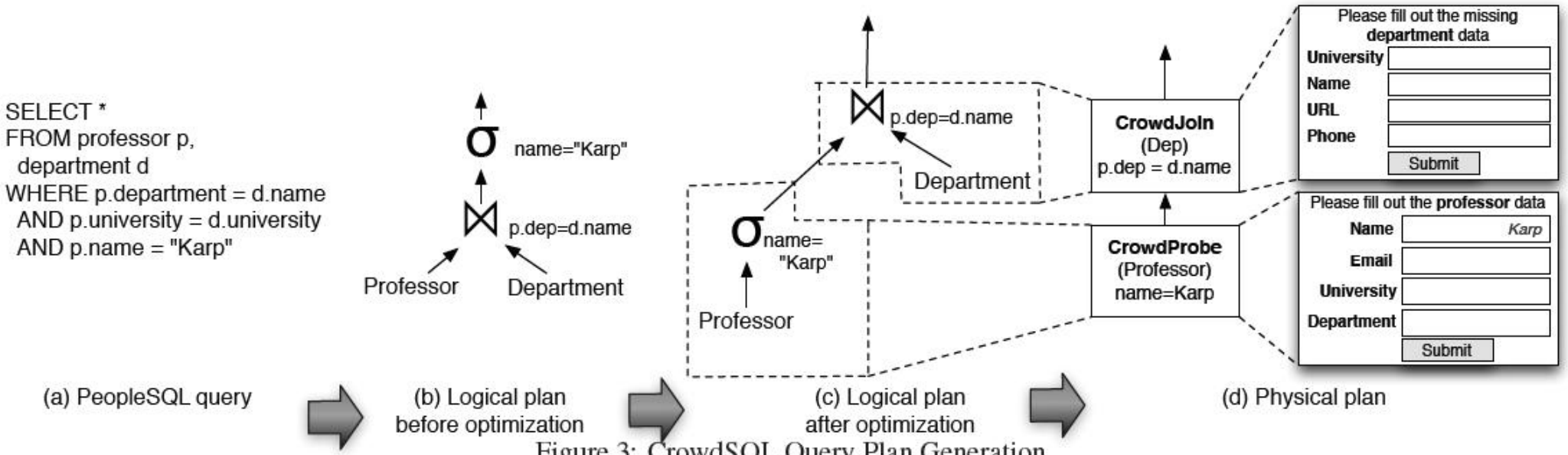


Figure 3: CrowdSQL Query Plan Generation

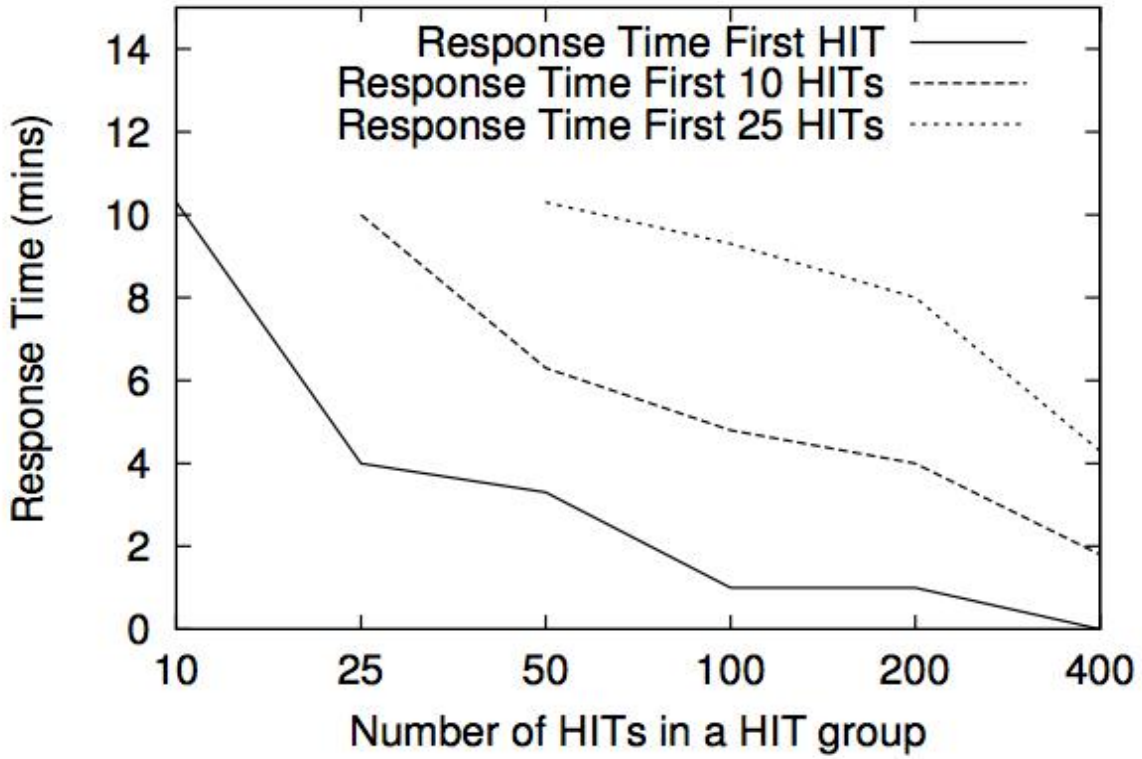
Heuristics:

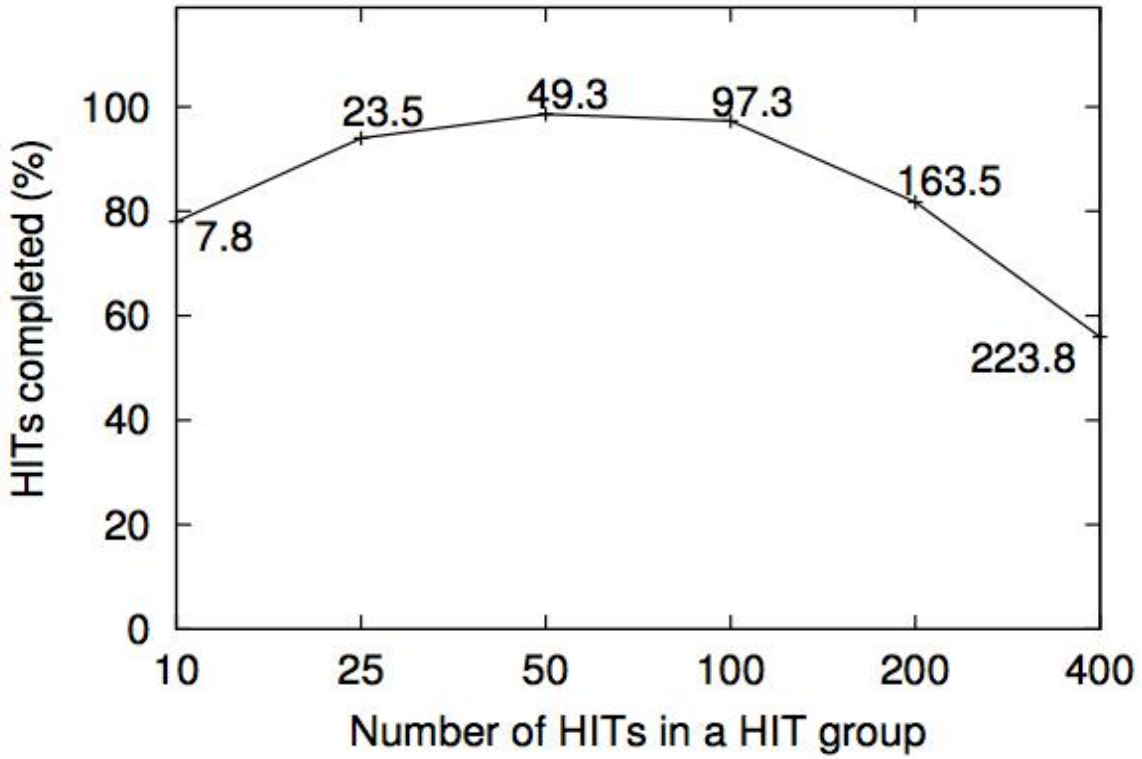
- Simple rule-based optimizer: e.g. predicate push-down.
- Crowdsourcing rules:
  - Set the basic crowdsourcing parameters (price, batching-size).
  - Select the user interface (normalized vs. denormalized).

A cost-based optimize that considers the changing conditions on AMT, remains future work.

Results on benchmarks

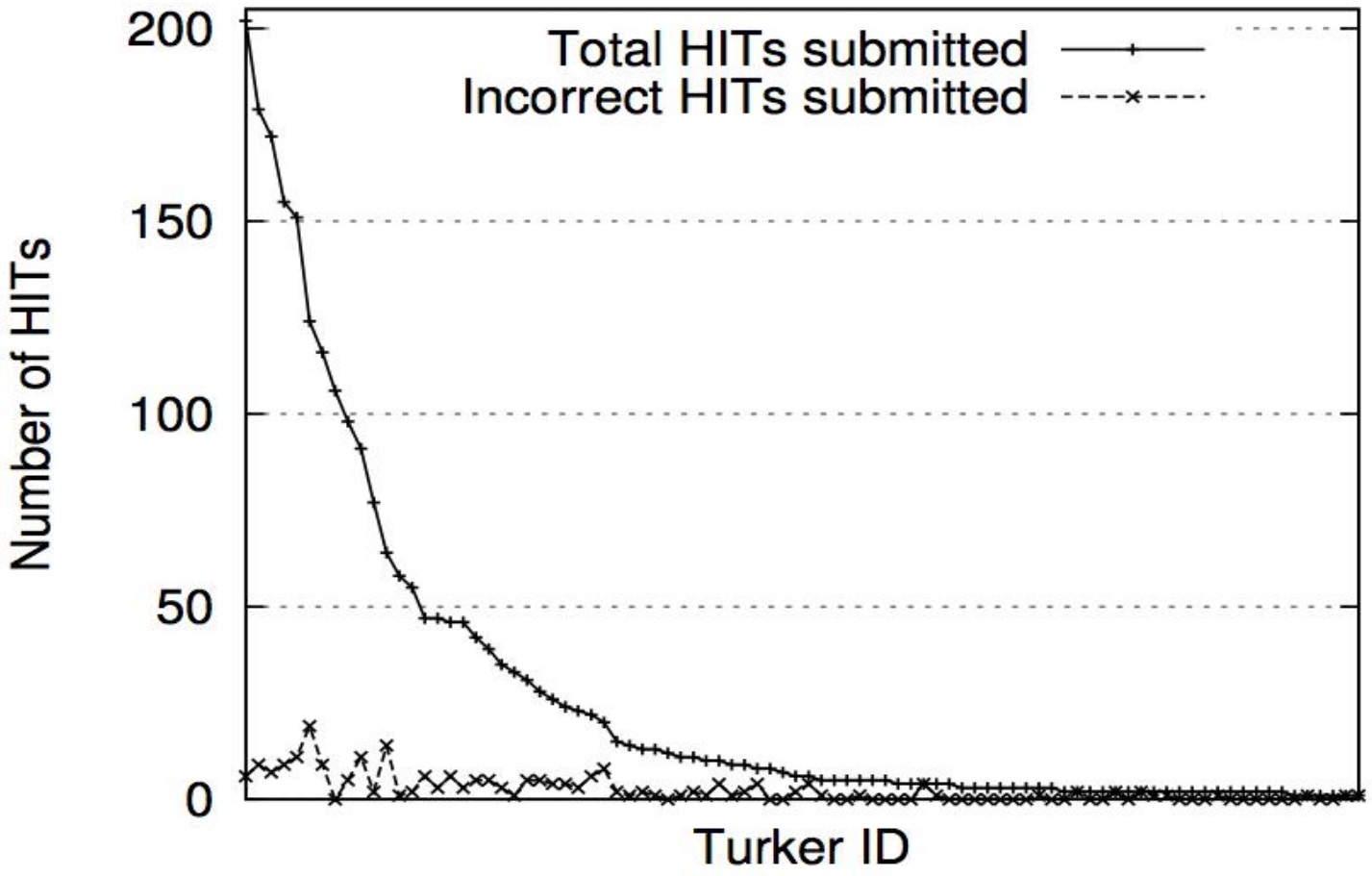
HIT Size vs. responsiveness





Tradeoff between HITS completed/time and % completion of HITS

Completion across workers



- Skewed distribution
- No variation in error rate between high freq workers and others

Complex Queries

Ordering Pictures

(a) 15, 1, 1

(b) 15, 1, 2

(c) 14, 3, 4

(d) 13, 4, 5

(e) 10, 5, 6

(f) 9, 6, 3

(g) 4, 7, 7

(h) 4, 7, 8

{# of votes by the workers, picture rank based on workers votes, picture rank ordered by experts}

Complex Queries

- Joining professors and departments
  - SELECT p.name, p.email, d.name, d.phone
  - FROM Professor p, Department d
  - WHERE p.department = d.name AND
  - p.university = d.university AND
  - p.name = "[name of a professor]"
- Method 1: first prof details collected, then dep details
- Method 2: prof and dep details collected together via a denormalized interface

Method 2 is cheaper, but Method 1 outperforms Method 2 in accuracy:

- Instructions for denormalized interface unclear

Other observations

- Relationship with workers is long-term
  - Keep workers happy
  - Implement less stringent approval mechanisms
- Good interface design and instructions matter
  - Simple choices like “none of the above” improve quality dramatically

History Lesson

Even now, there is no real complete, fully-functional crowd-powered database

- No one understands the crowds (EVEN NOW)
  - We were all naïve in thinking that we could treat crowds as just another data source.
- People don’t seem to want to use crowds within databases
  - Crowdsourcing is a one-off task
- Crowds have very different characteristics than other data

Still…

The ideas are very powerful and applicable everywhere you want data to be extracted

Very common use-case of crowds

Open Question

Points: 10

设置

CrowdDB only records either CNULL or the final outcome.

Why might this be a bad idea?

作答

Crowdsourcing Graphical Perception:
Using MechanicalTurk to Assess Visualization Design

Jeffrey Heer and Michael Bostock
Computer Science Department, Stanford University

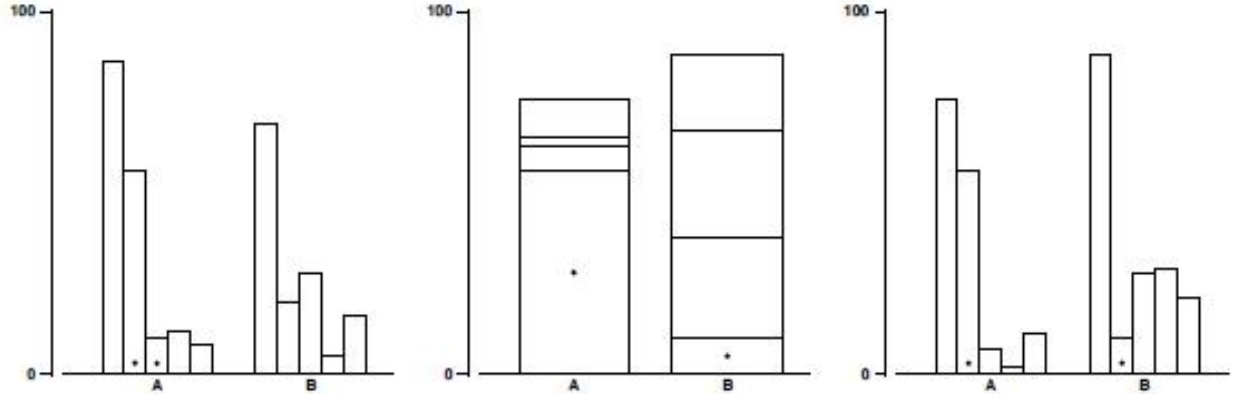


Figure 1: Stimuli for judgment tasks T1, T2 & T3. Subjects estimated percent differences between elements.

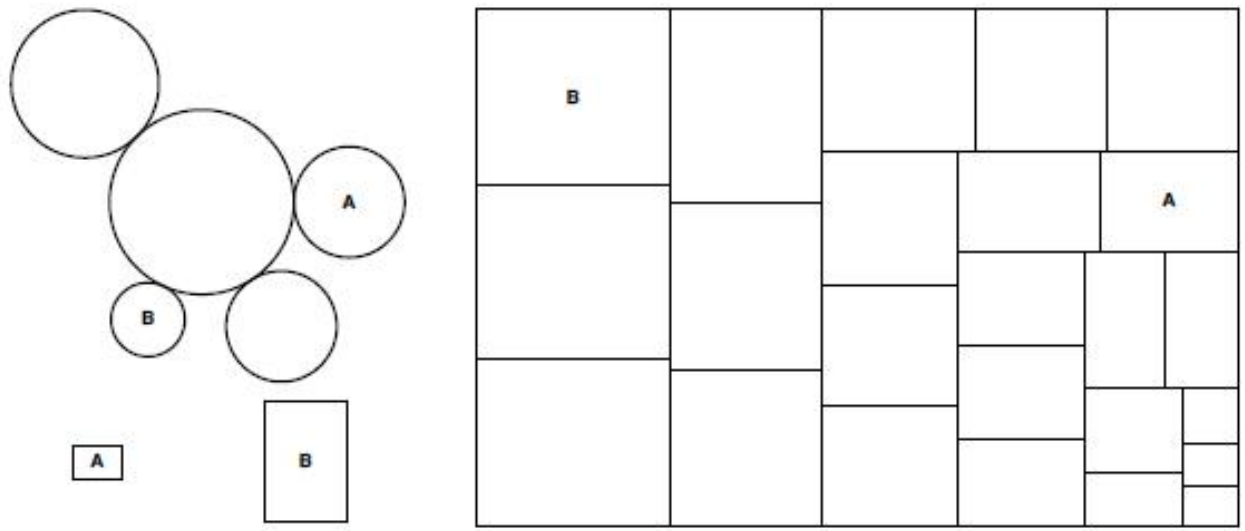


Figure 2: Area judgment stimuli. Top left: Bubble chart (T7), Bottom left: Center-aligned rectangles (T8), Right: Treemap (T9).

SpreadSheets are everywhere!

Excel has a global user base with estimates ranging from 1.1 billion to 1.5 billion people(2023)

Spreadsheets have continued to support increasingly large scale data (google sheets supports two million cells)

What is an electronic spreadsheet?

A **spreadsheet** is a computer application for organization, analysis, and storage of data in tabular form

It is the electronic equivalent of an accounting worksheet, comprised of rows and columns to allow you to do many tasks in the organization of numbers in a clear, easy to understand format

What is an electronic spreadsheet?

- It is a tool to help you calculate budgets, do economic analysis, statistics, planning, engineering calculations, …
- Replaces pen, paper and pocket calculator
- Can show diagrams and graphs
- Can input data from other programs
- Can output data to other programs

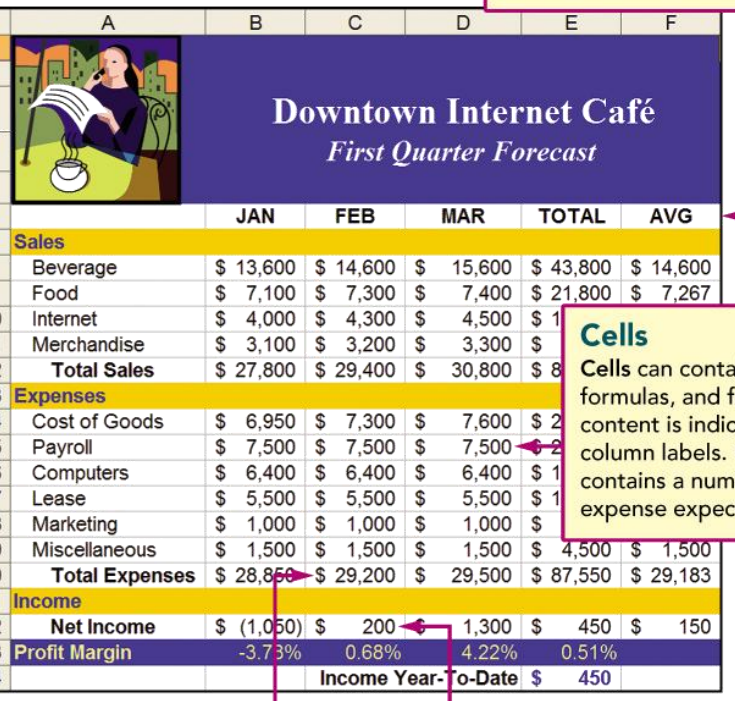
What is an electronic spreadsheet?

Worksheets

Worksheets are used for a wide range of different applications. One of the most common is to create, analyze, and forecast budgets.

Text Entries

Text entries provide meaning to the values in the worksheet. The rows are labeled to identify the various sales and expense items. The columns are labeled to specify the months.



	A	B	C	D	E	F
1						
2						
3		Downtown Internet Café				
4		*First Quarter Forecast*				
5						
6		JAN	FEB	MAR	TOTAL	AVG
7	Sales					
8	Beverage	$ 13,600	$ 14,600	$ 15,600	$ 43,800	$ 14,600
9	Food	$ 7,100	$ 7,300	$ 7,400	$ 21,800	$ 7,267
10	Internet	$ 4,000	$ 4,300	$ 4,500	$ 1	
11	Merchandise	$ 3,100	$ 3,200	$ 3,300	$	
12	Total Sales	$ 27,800	$ 29,400	$ 30,800	$ 8	
13	Expenses					
14	Cost of Goods	$ 6,950	$ 7,300	$ 7,600	$ 2	
15	Payroll	$ 7,500	$ 7,500	$ 7,500	$ 2	
16	Computers	$ 6,400	$ 6,400	$ 6,400	$ 1	
17	Lease	$ 5,500	$ 5,500	$ 5,500	$ 1	
18	Marketing	$ 1,000	$ 1,000	$ 1,000	$	
19	Miscellaneous	$ 1,500	$ 1,500	$ 1,500	$ 4,500	$ 1,500
20	Total Expenses	$ 28,850	$ 29,200	$ 29,500	$ 87,550	$ 29,183
21	Income					
22	Net Income	$ (1,050)	$ 200	$ 1,300	$ 450	$ 150
23	Profit Margin	-3.78%	0.68%	4.22%	0.51%	
24			Income Year-To-Date	$ 450		

Cells

Cells can contain labels, numbers, formulas, and functions. A cell's content is indicated by the row and column labels. For example, cell D15 contains a number for the Payroll expense expected for March.

Functions

One advantage of using **functions** rather than entering formulas is that they are easier to enter. In this case, cell C20 (Total Expenses for February) contains the function SUM(C14:C19) rather than the formula =C14+C15+C16+C17+C18+C17.

Formulas

Formulas provide a way to perform calculations in the worksheet. In this case Cell C22 (Net Income for February) contains the formula = C12 (Total Sales for February) −C20 (Total Expenses for February).

Bob Frankston (standing) and Dan Bricklin, 1982.

VisiCalc,1979

The first spreadsheet: VisiCalc

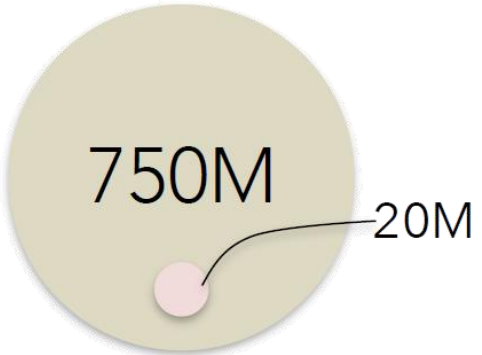
Some Advantages of Spreadsheets

- Spreadsheets are capable of exploring “what-if”scenarios (e.g. budgets, submitting bids)

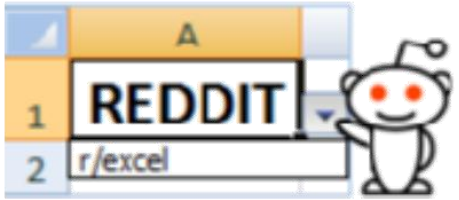
- Once it is set up properly, the user can save time by never having to set up the spreadsheet again
  - Blank spreadsheets are called templates.
  - Monthly salaries, grade sheets

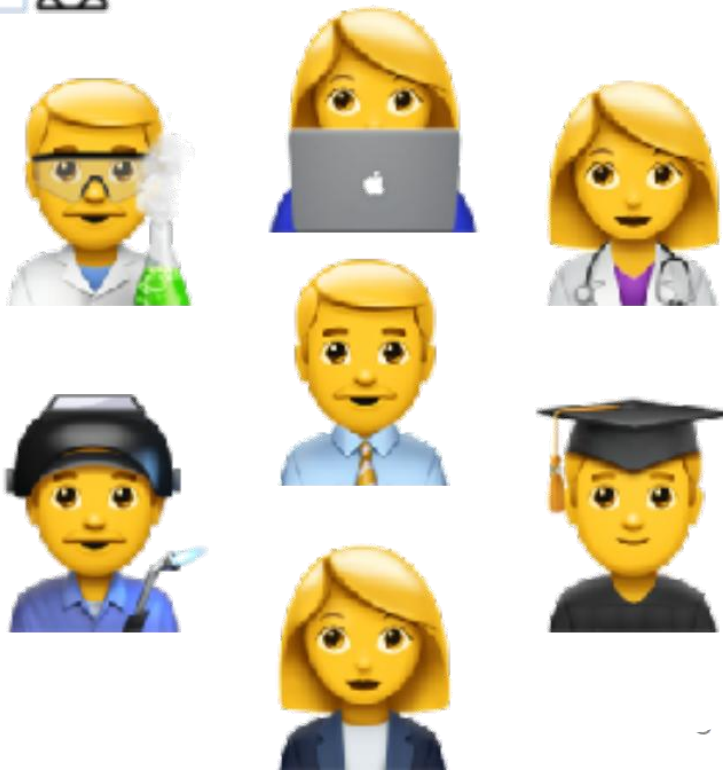
The most popular data management tool

- 10% of the word uses spreadsheets (750 M)
  - Programmers a small fraction (20M)



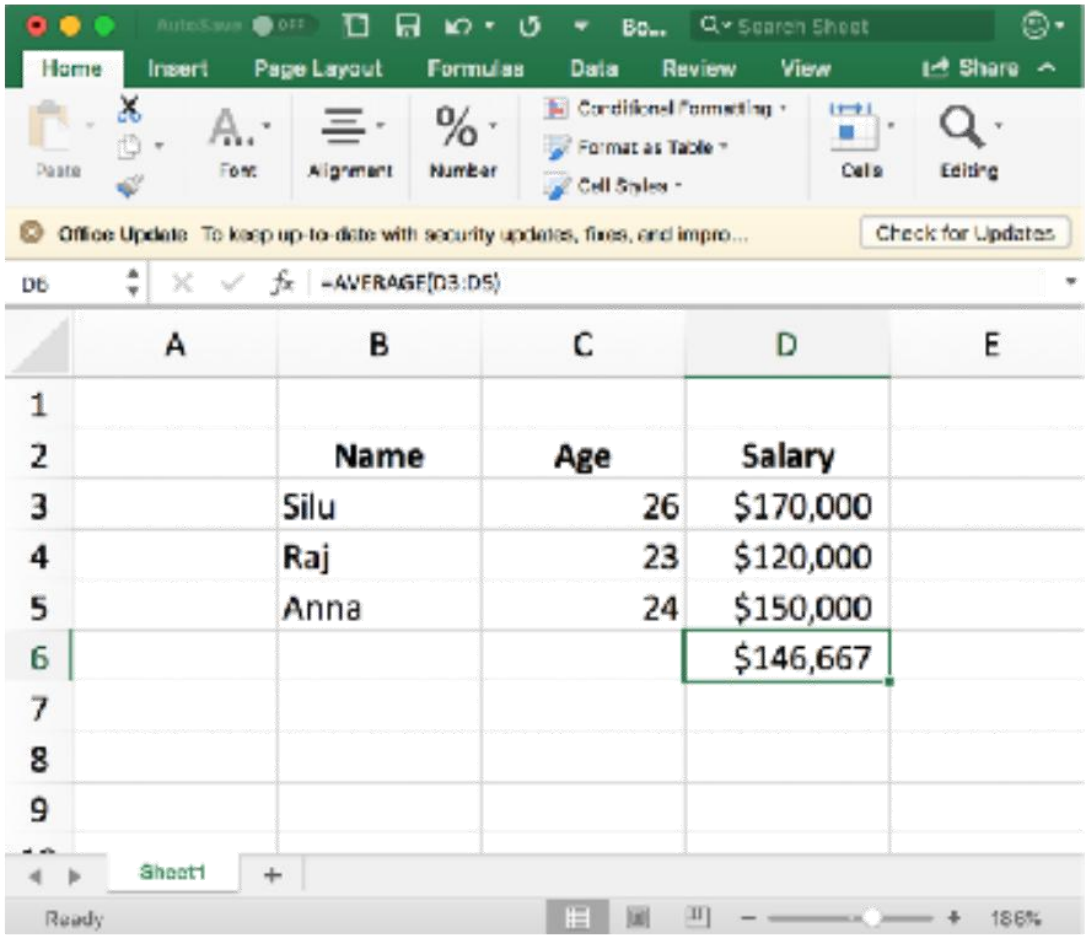
- Use cases from /r/Excel [Mack, …, P., CHI'18]
  - Professional:
    - stock tracking
    - finance data
    - inventory tracking
    - real-estate & manuf.
    - scientific exp. data
    - accounting info
    - patient info
  - Personal:
    - health & quantified self
    - sports
    - personal finance
    - …





Why is it so popular?

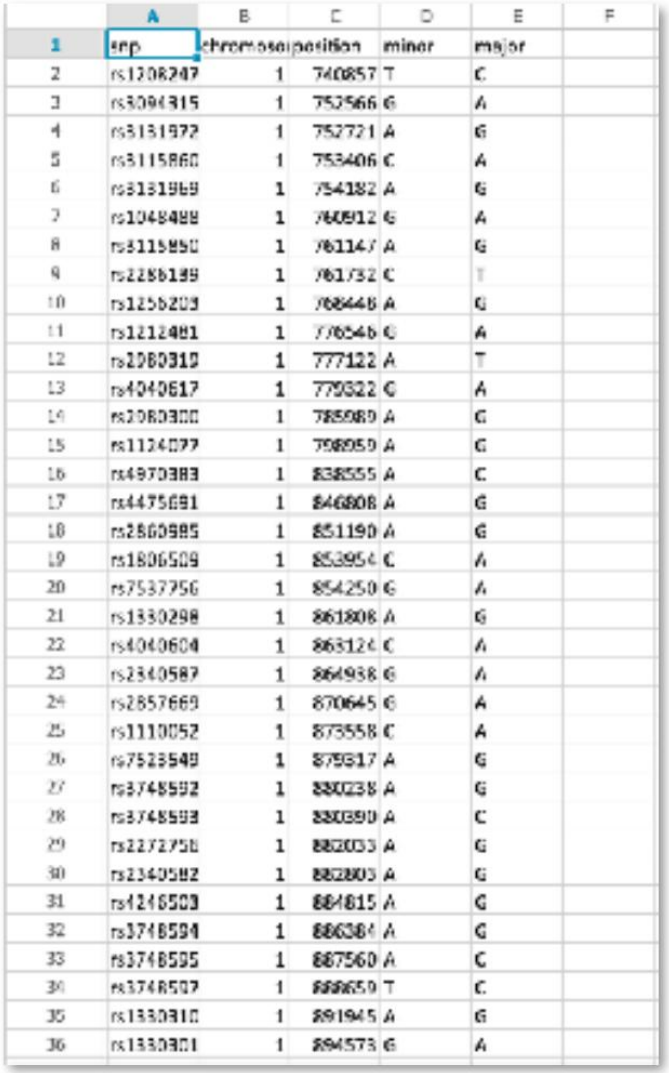
- Easy-to-use and flexible
- Can get started immediately
- Easy to see what is going onand get feedback
- Comes bundled together withmost office software
- “Export to excel”



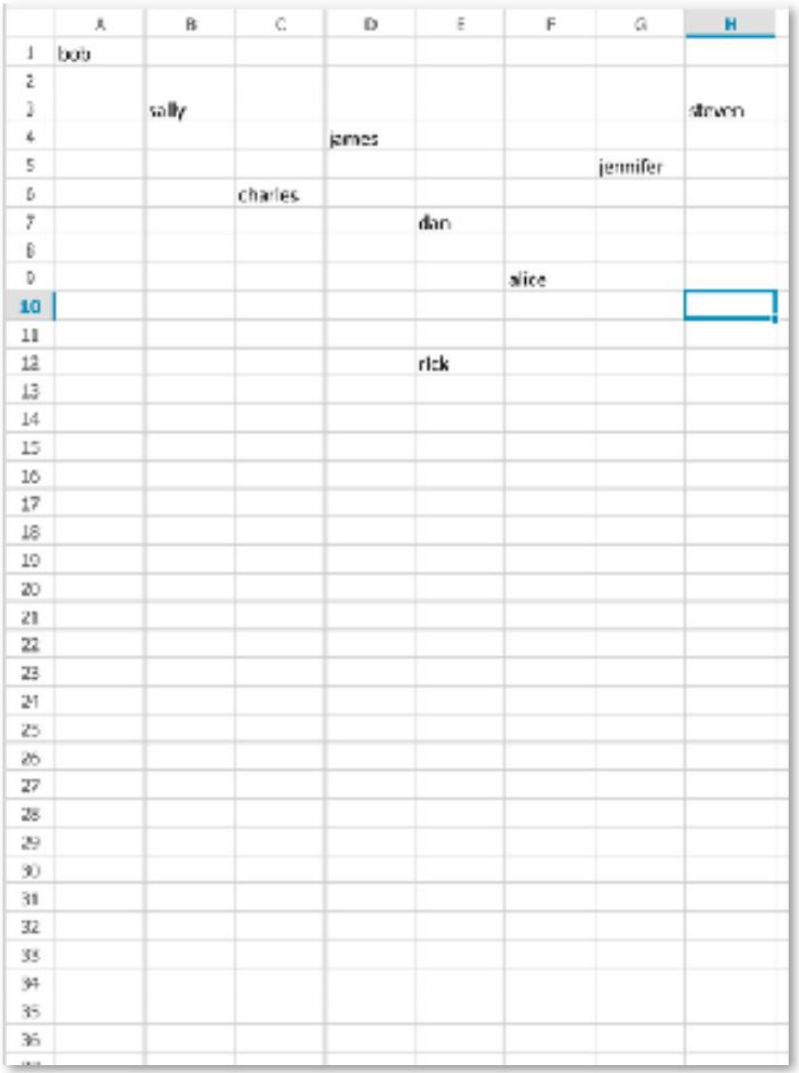
Spreadsheet Concept

- A Spreadsheet Workbook comprises many sheets
- Each sheet has *cells* — thus, a spreadsheet is structured around cells
  - Cells contain
    - *values*, e.g., numbers, strings, date/time; or
    - *formulae*, indicated by a "=<Expression>"
      - formula expressions can involve arithmetic +/-
        - e.g., =A1+B1
      - or special *functions*
        - e.g., =AVERAGE(B1, D1)

Ad-hoc data layout, from dense to sparse



	A	B	C	D	E	F
1	snp	chromosome	position	minor	major	
2	rs1308247	1	740857	T	C	
3	rs3094315	1	752566	G	A	
4	rs3131972	1	752721	A	G	
5	rs3115860	1	753406	C	A	
6	rs3131969	1	754182	A	G	
7	rs1048488	1	760912	G	A	
8	rs3115850	1	761147	A	G	
9	rs2286139	1	761732	C	T	
10	rs1256203	1	768448	A	G	
11	rs1212481	1	776546	G	A	
12	rs2980319	1	777122	A	T	
13	rs4040617	1	779322	G	A	
14	rs2980300	1	785989	A	G	
15	rs1134077	1	798959	A	G	
16	rs4970383	1	838555	A	C	
17	rs4475691	1	846808	A	G	
18	rs2860985	1	851190	A	G	
19	rs1806509	1	853954	C	A	
20	rs7537756	1	854250	G	A	
21	rs1330298	1	861808	A	G	
22	rs4040604	1	863124	C	A	
23	rs2340587	1	864938	G	A	
24	rs2857669	1	870645	G	A	
25	rs1110052	1	873558	C	A	
26	rs7523549	1	879317	A	G	
27	rs3748592	1	880238	A	G	
28	rs3748593	1	880390	A	C	
29	rs2272756	1	882033	A	G	
30	rs2340582	1	882800	A	G	
31	rs1246503	1	884615	A	G	
32	rs3748594	1	886384	A	G	
33	rs3748595	1	887560	A	C	
34	rs3748597	1	888659	T	C	
35	rs1330810	1	891945	A	G	
36	rs1330801	1	894573	G	A	



	A	B	C	D	E	F	G	H
1	bob							
2								
3		sally						drawn
4				james				
5							jennifer	
6			charles					
7					dan			
8								
9						alice		
10								
11								
12					rick			
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								
36								

Spreadsheet: formula functions

- The arguments to formulae are other cells, which can contain formulae or values.
  - Can be tedious when referring to lots of cells, e.g., cell A1 to A1000
- Shortcuts:
  - rectangular ranges of cells
    - e.g., B1:C3 = B1, B2, B3, C1, C2, C3
  - entire column
    - e.g., F:F = F1, F2, ….
- Standard statistical functions
  - AVERAGE (B1:C3), SUM (F:F), MIN (A1:A100)
  - Relational mapping: Like the aggregation functions in a group by query

Spreadsheet: statistical functions

- COUNTIF, AVERAGEIF, SUMIF

- Two arguments: list of cells, followed by a condition
  - e.g., = COUNTIF (F:F, "*HURRICANE*")
    - counts number of values in text field that contain HURRICANE

- Q: What does this map to from a relational database perspective?

Spreadsheet: lookup

- VLOOKUP or Value Lookup
  - VLOOKUP (value v, tabular range R, col index i, approximate = FALSE)
  - Look for value v in first column of R, if matched, fetch the value in the ith column of R on the same row, and return it
- Demo: VLOOKUP of states
- Q: What does this remind you of from a relational perspective?

Online Spreadsheet -- Airtable

DATASPREAD: Unifying Databases and Spreadsheets

Mangesh Bendre, Bofan Sun, Ding Zhang, Xinyan Zhou
Kevin Chen-Chuan Chang, Aditya Parameswaran
University of Illinois at Urbana-Champaign (UIUC)
{bendre1 | bsun6 | dzhang13 | xzhou14 | kcchang | adityagp}@illinois.edu

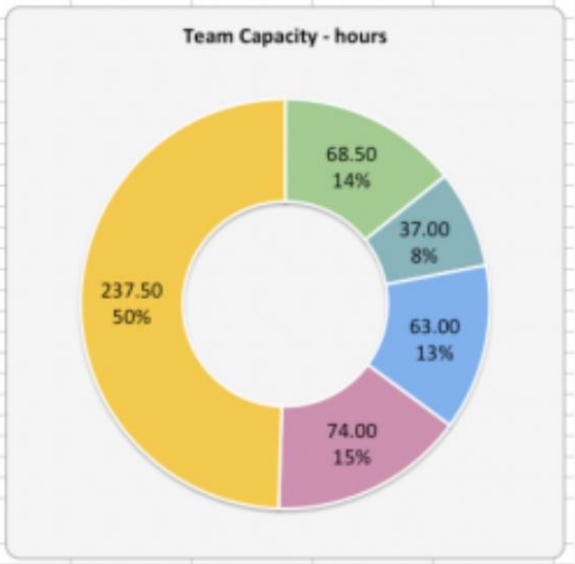
What is DataSpread

- A spreadsheet-database hybrid
- Marrying the flexibility and ease of use of spreadsheets with the scalability and power of databases

Motivation

Most of the people doing ad-hoc data manipulation and analysis use spreadsheets.

	A	C	D	E	F	G	H	J
1		FE1	TechLead	Dev2	Dev3	Dev4	QE	TOTAL
2	sprint planning	3.50	3.50	3.50	3.50	3.50	3.50	21.00
3	stand ups	2.25	2.25	2.25	2.25	2.25	2.25	13.50
4	pre-BLG		2.00				2.00	4.00
5	backlog grooming	1.00	1.00	1.00	1.00	1.00	1.00	6.00
6	sprint demo	1.00	1.00	1.00	1.00	1.00	1.00	6.00
7	sprint retrospective	1.00	1.00	1.00	1.00	1.00	1.00	6.00
8	other teams' demos	2.00	2.00	2.00	2.00	2.00	2.00	12.00
9	*SubTotal (repeating Scrum)*	*10.75*	*12.75*	*10.75*	*10.75*	*10.75*	*12.75*	*68.50*
10	Tech Leadership mtg		3.00					3.00
11	CCB		2.00					2.00
12	Guilds		2.00	2.00	2.00	2.00		8.00
13	Company-wide All-hands	1.00	1.00	1.00	1.00	1.00	1.00	6.00
14	Engineering dept All-hands	2.00	2.00	2.00	2.00	2.00	2.00	12.00
15	Engineering dept Tech Talks	1.00	1.00	1.00	1.00	1.00	1.00	6.00
16	*SubTotal (repeating non-Scrum)*	*4.00*	*11.00*	*6.00*	*6.00*	*6.00*	*4.00*	*37.00*
17	training / orientation					6.00		6.00
18	interviews		3.00					3.00
19	holiday							0.00
20	PTO	8.00	8.00		8.00		6.00	30.00
21	travel / conferences / etc			12.00	12.00			24.00
22	*SubTotal (non-repeating)*	*8.00*	*11.00*	*12.00*	*20.00*	*6.00*	*6.00*	*63.00*
23	*subtotal of three above*	22.75	34.75	28.75	36.75	22.75	22.75	168.50
24	*available less planned non-working*	57.25	45.25	51.25	43.25	57.25	57.25	311.50
25	*individual focus factor (higher is more distracted)*	20%	33%	20%	20%	25%	25%	23.8%
26	*Other non-task time*	*11.50*	*14.90*	*10.30*	*8.70*	*14.30*	*14.30*	*74.00*
27	**Total non-working**	**34.25**	**49.65**	**39.05**	**45.45**	**37.05**	**37.05**	**242.50**
28	sprint size: 10 days x 8 hours	80.00	80.00	80.00	80.00	80.00	80.00	480.00
29	**hours available for sprint**	**45.75**	**30.35**	**40.95**	**34.55**	**42.95**	**42.95**	**237.50**
30								
31	NOTES:							
32	TechLead is also Dev1							
33	Dev4 is new to team							



Easy to use, built-in visualization capabilities, flexible (schema-free)

Spreadsheets are terrible!

- *Slow*
  - single change ➔ wait minutes on a 10,000 x 10 spreadsheet
  - can't even open a spreadsheet with >1M cells
  - speed by itself can prevent analysis
- *Tedious + not Powerful*
  - filters via copy-paste
  - only FK joins via VLOOKUPs; others impossible
  - even simple operations are cumbersome
- *Brittle*
  - sharing excel sheets around, no collab/recovery
  - using spreadsheets for collaboration is painful and error-prone

Spreadsheets are terrible!

How do I know the impact of your experimental scores on your final score?

Grade point > 80

学生学号	学生姓名	班级	实验成绩	总成绩
202000202204	谢灿	数据20	100	100
202022161228	刘奕彤	数据20	100	99
202000161244	方博文	数据20	90	87
202000130231	胡彦蓉	数据20	80	96
202000150254	刘铭心	数据20	70	65
202000202092	谢斌	数据20	90	75
202000130052	臧传超	数据20	90	86
202000150069	陈炜麟	数据20	100	100
202000130203	颜恺楠	数据20	100	99
201900180075	陈曾喆	数据20	90	87
202000130074	方鹏贺	数据20	80	96
201900202045	苏兆义	数据20	70	65
202000130155	王雅轩	数据20	90	75
202000130214	赵瑞坤	数据20	90	86
202000150007	王乾润	数据20	100	100
202000130215	苗琳瑜	数据20	100	99
202000130010	刘博	数据20	90	87
202000130061	宋家庆	数据20	80	96
201900202024	史惠敏	数据20	70	65
202000130093	何文鑫	数据20	90	75
202000120166	孙留羿	数据20	90	86
202022161218	王子畅	数据20	100	100
202000130152	师语	数据20	100	99
202000130051	董凯	数据20	90	87
202018130179	袁海峰	数据20	80	96
202018150189	申彤	数据20	70	65

学号	姓名	班级	联系方式	家乡
202000130010	刘博	数据20	18364823089	湖南
202000130093	何文鑫	数据20	13011691780	河北
202000161244	方博文	数据20	18152767038	陕西
202000130231	胡彦蓉	数据20	17860708869	山西
202000150254	刘铭心	数据20	15610519182	湖南
202000120166	孙留羿	数据20	19948788911	陕西
202022161218	王子畅	数据20	15986188808	山西
202000150069	陈炜麟	数据20	18419339896	河北
202000130203	颜恺楠	数据20	17838588928	陕西
202000130214	赵瑞坤	数据20	18633693205	河北
202000150007	王乾润	数据20	13847777210	陕西
201900180075	陈曾喆	数据20	18907409259	山西
202000130074	方鹏贺	数据20	17866833125	湖南
202000130214	赵瑞坤	数据20	18633693205	河北
202000150007	王乾润	数据20	13847777210	陕西
202000130215	苗琳瑜	数据20	15020010469	山西
202000202204	谢灿	数据20	17783868971	山东
202000130061	宋家庆	数据20	17720439452	福建
201900202024	史惠敏	数据20	13111753933	山东
202022161228	刘奕彤	数据20	15626660219	河北
202000202092	谢斌	数据20	13953552926	福建
202000130052	臧传超	数据20	15085461401	山东
202000130152	师语	数据20	18973630888	湖南
202000130051	董凯	数据20	15858473273	福建
202000120166	孙留羿	数据20	19948788911	陕西
202022161218	王子畅	数据20	15986188808	山西

Spreadsheets are terrible!

How to plot the average grade point by your hometown group?

学生学号	学生姓名	班级	实验成绩	总成绩
202000202204	谢灿	数据20	100	100
202022161228	刘奕彤	数据20	100	99
202000161244	方博文	数据20	90	87
202000130231	胡彦蓉	数据20	80	96
202000150254	刘铭心	数据20	70	65
202000202092	谢斌	数据20	90	75
202000130052	臧传超	数据20	90	86
202000150069	陈炜麟	数据20	100	100
202000130203	颜恺楠	数据20	100	99
201900180075	陈曾喆	数据20	90	87
202000130074	方鹏贺	数据20	80	96
201900202045	苏兆义	数据20	70	65
202000130155	王雅轩	数据20	90	75
202000130214	赵瑞坤	数据20	90	86
202000150007	王乾润	数据20	100	100
202000130215	苗琳瑜	数据20	100	99
202000130010	刘博	数据20	90	87
202000130061	宋家庆	数据20	80	96
201900202024	史惠敏	数据20	70	65
202000130093	何文鑫	数据20	90	75
202000120166	孙留羿	数据20	90	86
202022161218	王子畅	数据20	100	100
202000130152	师语	数据20	100	99
202000130051	董凯	数据20	90	87
202018130179	袁海峰	数据20	80	96
202018150189	申彤	数据20	70	65

学号	姓名	班级	联系方式	家乡
202000130010	刘博	数据20	18364823089	湖南
202000130093	何文鑫	数据20	13011691780	河北
202000161244	方博文	数据20	18152767038	陕西
202000130231	胡彦蓉	数据20	17860708869	山西
202000150254	刘铭心	数据20	15610519182	湖南
202000120166	孙留羿	数据20	19948788911	陕西
202022161218	王子畅	数据20	15986188808	山西
202000150069	陈炜麟	数据20	18419339896	河北
202000130203	颜恺楠	数据20	17838588928	陕西
202000130214	赵瑞坤	数据20	18633693205	河北
202000150007	王乾润	数据20	13847777210	陕西
201900180075	陈曾喆	数据20	18907409259	山西
202000130074	方鹏贺	数据20	17866833125	湖南
202000130214	赵瑞坤	数据20	18633693205	河北
202000150007	王乾润	数据20	13847777210	陕西
202000130215	苗琳瑜	数据20	15020010469	山西
202000202204	谢灿	数据20	17783868971	山东
202000130061	宋家庆	数据20	17720439452	福建
201900202024	史惠敏	数据20	13111753933	山东
202022161228	刘奕彤	数据20	15626660219	河北
202000202092	谢斌	数据20	13953552926	福建
202000130052	臧传超	数据20	15085461401	山东
202000130152	师语	数据20	18973630888	湖南
202000130051	董凯	数据20	15858473273	福建
202000120166	孙留羿	数据20	19948788911	陕西
202022161218	王子畅	数据20	15986188808	山西

How about databases?

- ~~*Slow*~~ Scalable
- ~~*Tedious + not Powerful*~~ Powerful and expressive (SQL)
- ~~*Brittle*~~ Collaboration, recovery, succinct

So why not use databases?

Well, for the same reason why spreadsheets are so useful:

- ~~*Easy to use*~~ Not easy to use
- ~~*Built-in visualization*~~ No built-in visualization
- ~~*Flexible*~~ Not flexible

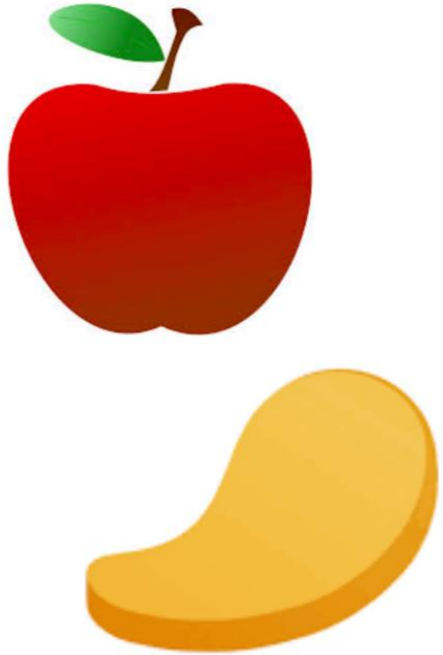
Spreadsheets + Databases

- Spreadsheet as a front-end interface
- Databases as a backend engine

Is it so simple?

Challenges

Feature	Databases	Spreadsheets
Data Model	Schema-first	Dynamic/No Schema
Addressing	Tuples with PK	Cells, using Row/Col
Presentation	Set-oriented, no such notion	Notion of current window, order
Modifications	Must equal queries	Can be done at any granularity
Computation	Query at a time	Value at a time



Semantics and Syntax

Support for dynamic schema:

allow users to select an arbitrary range on the spreadsheet and use it to define the structure and the data for a table within the database.

Semantics and Syntax

- **Support for dynamic schema**: allow users to select an arbitrary range on the spreadsheet and use it to define the structure and the data for a table within the database.

- **Make databases interface aware**: make databases aware of implicit data layout

Semantics and Syntax

- **Support for dynamic schema**: allow users to select an arbitrary range on the spreadsheet and use it to define the structure and the data for a table within the database.
- **Make databases interface aware**: make databases aware of implicit data layout
- **Novel spreadsheet constructs**: DBSQL (RangeValue, RangeTable), DBTable

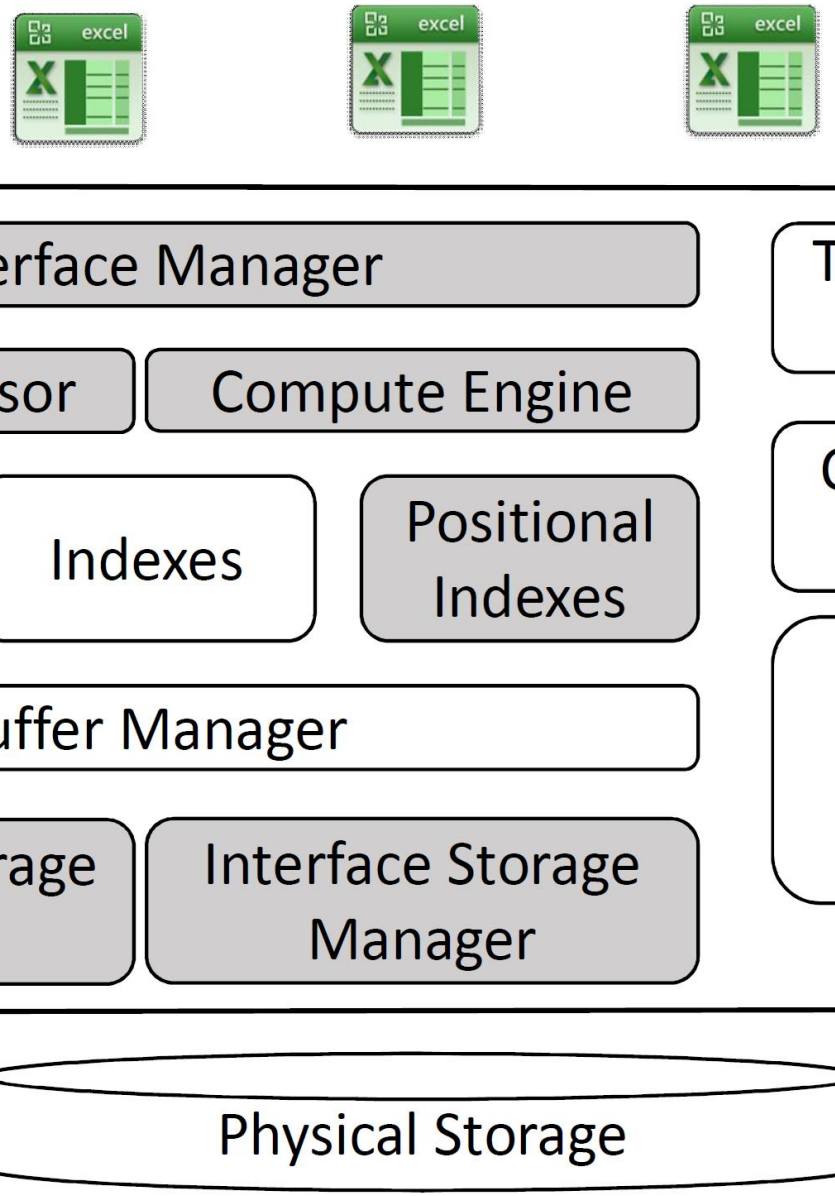
Select from Actors

Where ActorId=RangeValue(A1)

Other Semantics

- **SQL support on spreadsheets**: support both spreadsheet and SQL, and give flexibility to the user to interchangeably use either.
- **Real-time sync**: a real time two way synchronization of the displayed on the spreadsheet with the underlying database
- **Data typing**: automatically assigning data types within the databases based on the tuples
- **Computational optimization**: prioritize computations for the data that is displayed
- **Lazy computation**: calculations of the visible cells should be prioritized and the remaining long running computations should be performed in the background

Architecture



Architecture

Making databases interface-aware

Support and optimize the execution for positional addressing

Interface Manager

Query Processor

Compute Engine

Records

Indexes

Positional Indexes

Buffer Manager

Relational Storage Manager

Interface Storage Manager

Transactional Manager

Concurrency Control

Main Memory

Buffers: data, index, log, etc.

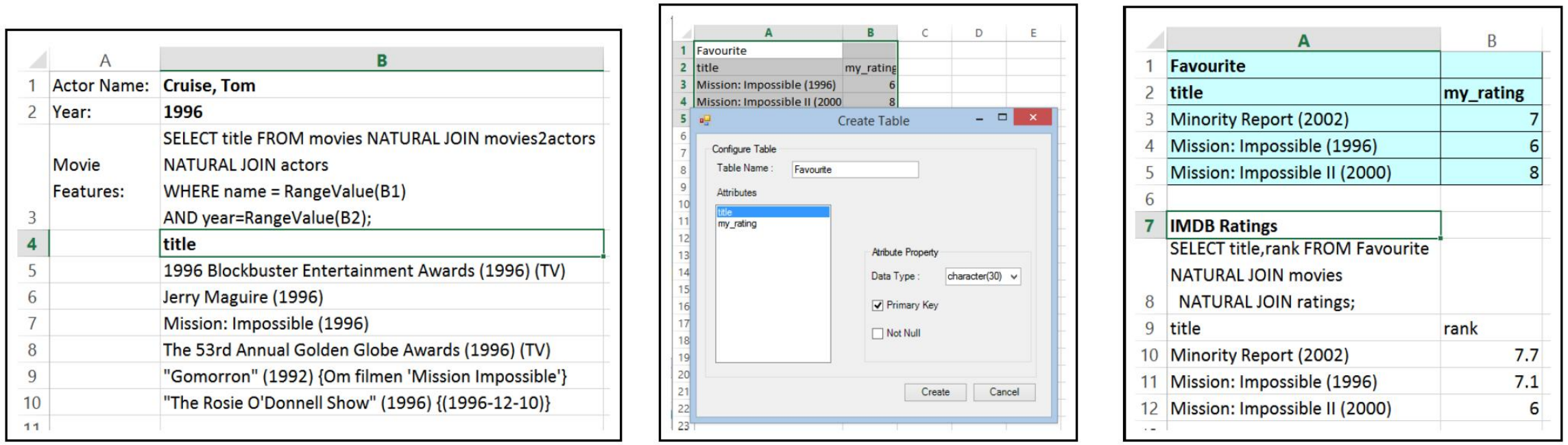
Leverages interface storage aspects to optimize execution

Store data in the interface but not a relational table

Support interface related operations such as schema changes

Physical Storage

Demonstration



	A	B
1	Actor Name:	**Cruise, Tom**
2	Year:	**1996**
		SELECT title FROM movies NATURAL JOIN movies2actors
	Movie	NATURAL JOIN actors
	Features:	WHERE name = RangeValue(B1)
3		AND year=RangeValue(B2);
4		**title**
5		1996 Blockbuster Entertainment Awards (1996) (TV)
6		Jerry Maguire (1996)
7		Mission: Impossible (1996)
8		The 53rd Annual Golden Globe Awards (1996) (TV)
9		"Gomorron" (1992) {Om filmen 'Mission Impossible'}
10		"The Rosie O'Donnell Show" (1996) {(1996-12-10)}

	A	B
1	**Favourite**	
2	**title**	**my_rating**
3	Minority Report (2002)	7
4	Mission: Impossible (1996)	6
5	Mission: Impossible II (2000)	8
6		
7	**IMDB Ratings**	
	SELECT title,rank FROM Favourite	
	NATURAL JOIN movies	
8	NATURAL JOIN ratings;	
9	title	rank
10	Minority Report (2002)	7.7
11	Mission: Impossible (1996)	7.1
12	Mission: Impossible II (2000)	6

Figure 2: (a) Executing SQL with relative referencing. (b) Table creation. (c) Two-way table sync.

Excel + PostgreSQL

https://dataspread.github.io/
https://github.com/dataspread/dataspread-web
https://github.com/myliang/x-spreadsheet

Ability to load and explore large datasets spanning billions of rows

Week 6 – Assignments

- 以项目小组为单位，设计系统界面原型
- **10月14日前**将界面设计初稿提交至**Canvas**平台
- **10月15日**请各小组对其他小组的作品提修改意见
- **10月15-16日**，针对反馈修改设计

** **注意**：**本阶段仅需设计系统界面原型，无需代码实现**

Reminder

- **Student representation will begin on Oct. 31**
- **Send your slides to me before Oct. 29**

Oct. 31	Group Name	Nov. 7	Group Name
1	你们无敌了队	6	我们做的都队
2	这是我们队	7	星穹列车队
3	strong队	8	汪汪队
4	花开复队	9	发际线总和我作队
5	古灵井盖队	10	你不对我队
		11	你知道的大数据分析实践是一门好课

Course Outline

6 Personal assignments

6 Group assignments

上课日期	授课内容	实验内容	周次
20240905	课程入门、大数据探索式分析	/	第一周
20240912	课程实践项目介绍、项目组队测试、项目经验谈	项目成员集结	第二周
20240919	科研实践入门、数据采样与降维	项目管理工具制定项目计划、**Pandas数据采样实践**	第三周
2024926	数据质量管理	Pandas数据质量实践	第四周
20241003	/	/	第五周
20241010	众包与电子表格	电子表格实践	第六周
20241017	**可视化设计**	**可视化设计实践**	**第七周**
20241024	统计分析方法与工具	统计方法实践	第八周
20241031	中期汇报（论文+项目进展）1	中期进展报告	第九周
20241107	中期汇报（论文+项目进展）2	BERT实践环境配置	第十周
20241114	机器学习方法与工具	BERT实践	第十一周
20241121	人机交互方法与工具	Canis/Cast/Libra实践	第十二周
20241128	普适计算	手机移动数据采集与分析	第十三周
20241205	大规模数据分析系统	SPARK实践	第十四周
20241212	如何撰写项目论文	大项目收尾	第十五周
20241219	项目结题报告1	大项目验收	第十六周

Thank You