# Spring Internship Outcomes

# THE TEAM

Presented by:

Yilin Shen

Rohit Singh

Under the sincere guidance of mentors:

Wonhee Lee

Amy Ng

Mason Jiang

# Project Objective

Understand the crowdsourced data collection process and its application.

Learn about strategies

- Test and ensure data quality
- Select quality workers
- Simulate the experience of MTurk workers
- Analyze data from real MTurk workers which helps to evaluate the quality of their work.

# Table of Contents

- **Introduction to MTurk**
- **Data Extraction**
- **Live Demo - Extracting data**
- **Potential Issues regarding extracting data manually**
- **Analyzing the Data**
- **Errors in the json files**
- **Ideas: data cleaning**
- **How can we figure out the ideal workers for our projects?**
- **Accuracy is essential for us**
- **We pay them hourly, so efficiency matters too!**
- **Averages are more valuable if they are stable**
- **Tasks are different**
- **What we need to do next**

# Introduction to Mturk

Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce that can perform these tasks virtually.

MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce.

- To streamline business processes
- Increase the process of data collection and analysis
- Accelerate machine learning development

Benefits -

- Reduce cost

By leveraging the skills of distributed Workers on a pay-per-task model, you can significantly lower costs while achieving results that might not have been possible with just a dedicated team.

- Optimize efficiency

Using MTurk to outsource micro-tasks ensures that work gets done quickly, while freeing up time and resources for the company – so internal staff can focus on higher value activities.

- Increase flexibility

With access to a global, on-demand, 24x7 workforce, MTurk enables businesses and organizations to get work done easily and quickly when they need it.

# Data Collection

- Manual data collection is a standard operating procedure when you are collecting a new measure.
- We use the Crowdsourced data collection to collect data from SEC as it offers several benefits including access to a large pool of workers and reduced data collection cost.
- Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time-consuming, expensive, difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet.
- Collecting data for the Human Capital Section in SEC Form 10K's on Turkle to support a faculty project.
- This data collection process closely resembles what real MTurk workers have done to collect data.
- And it is used to label data when building machine learning models that often require a large amount of labeled training data.

Turkle    Stats    Help

**Project:** Human Capital Data Collection / **Batch:** hc_batch1

☐ Auto-accept next Task    Return Task    Skip Task    Expires in 23:51

**Human Capital Disclosures Data Collection - Instructions (PLEASE READ VERY CAREFULLY)** (Click to expand)

*** PLEASE BE SURE TO REVIEW THE INSTRUCTIONS ABOVE***

**Name of the Company:**

CATCHMARK TIMBER TRUST

**Direct Link to Financial Statement:**

Click Here for the Financial Statement!

# Section 1

**Section Name:**

**Section Content:**

# Potential issues with manual data extraction

- Manual data entry is expensive and time-consuming.

- Human error leads to inaccurate information.

- Manual data entry creates disengaged employees.

- Manual data entry creates potential financial issues.

- Manual data entry could lead to potential compliance issues.

- No backup records in case of loss or damage.

# Analyzing the data

- **Shift our focus to evaluating the quality of data collected by MTurk workers, as well as their characteristics (e.g., speed, accuracy).**

- **Checking data uses a deep learning model. The learning model is not perfect. We need to check how much accuracy it can achieve.**

- **We reviewed the Human Capital Data in JSON format to detect anomalies in the result testing data.**
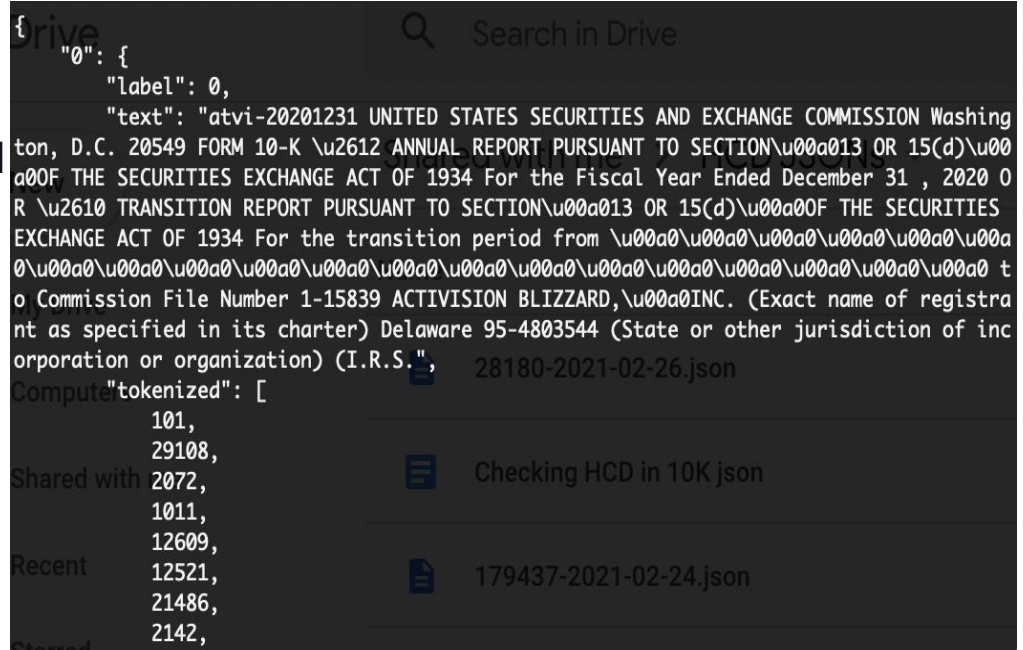
1997,
4579,
1010,
2023,
3189,
2038,
2042,
2772,
2917,
2011,
1996,
2206,
5381,
2006,
6852,
1997,
1996,
20588,
6494,
3372,
1998,
1999,
1996,
1996,
20588,
21157,
1998,
2006,
1996,
5246,
5393,
1012,
102
]
},
"1611": {
    "label": 0,
    "text": "By: /s/\u00a0ROBERT A. KOTICK Director, Chief Executive Officer (Principal Executive Officer) February 23, 2021 (Robert A. Kotick) By: /s/\u00a0DENNIS DURKIN Chief Financial Officer (Principal Financial Officer) February 23, 2021 (Dennis Durkin) By: /s/\u00a0JESSE YANG Chief Accounting Officer (Principal Accounting Officer) February 23, 2021 (Jesse Yang) By: /s/\u00a0REVETA BOWERS Director February 23, 2021 (Reveta Bowers) By: /s/\u00a0ROBERT J. CORTI Director February 23, 2021 (Robert J. Corti) By: /s/\u00a0HENDRIK J. HARTONG III Director February 23, 2021 (Hendrik J. Hartong III) By: /s/\u00a0BRIAN G. KELLY Chairman of the Board and Director February 23, 2021 (Brian G. Kelly) By: /s/\u00a0BARRY MEYER Director February 23, 2021 (Barry Meyer) By: /s/\u00a0ROBERT J. MORGADO Director February 23, 2021 (Robert J. Morgado) By: /s/\u00a0PETER NOLAN Director February 23, 2021 (Peter Nolan) By: /s/\u00a0DAWN OSTROFF Director February 23, 2021 (Dawn Ostroff) By: /s/\u00a0CASEY WASSERMAN Director February 23, 2021 (Casey Wasserman) E-5",
    "tokenized": [
101,
2011,
1024,
1013,
1055,
1013,
1013,
2728,
1037,
1012,
12849,
26348,

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | file name | company name | index of first sentence labeled 1 | number of sentences that were labeled 1 | index of labeled 1 sentence | labeling correct/wrong |
| 307 | | | | | 1092 | correct |
| 308 | | | | | 1093 | correct |
| 309 | | | | | | |
| 310 | 64768-20201231.json | AMAZON.COM | 73 | 6 | 73 | correct |
| 311 | | | | | 74 | correct |
| 312 | | | | | 75 | correct |
| 313 | | | | | 76 | correct |
| 314 | | | | | 77 | correct |
| 315 | | | | | 78 | correct |
| 316 | | | | | | |
| 317 | 180405-20201231.json | ACTIVISION BLIZZARD,INC. | 167 | 35 | 167 | correct |
| 318 | | | | | 168 | correct |
| 319 | | | | | 169 | correct |
| 320 | | | | | 170 | correct |
| 321 | | | | | 171 | correct |
| 322 | | | | | 172 | correct |
| 323 | | | | | 173 | correct |
| 324 | | | | | 174 | correct |
| 325 | | | | | 175 | correct |
| 326 | | | | | 176 | correct |
| 327 | | | | | 177 | correct |
| 328 | | | | | 178 | correct |
| 329 | | | | | 179 | correct |
| 330 | | | | | 180 | correct |
| 331 | | | | | 181 | correct |
| 332 | | | | | 182 | correct |
| 333 | | | | | 183 | correct |
| 334 | | | | | 184 | correct |
| 335 | | | | | 185 | correct |
| 336 | | | | | 186 | correct |
| 337 | | | | | 187 | correct |
| 338 | | | | | 188 | correct |
| 339 | | | | | 189 | correct |
| 340 | | | | | 190 | correct |
| 341 | | | | | 191 | correct |
| 342 | | | | | 192 | correct |
| 343 | | | | | 193 | correct |
| 344 | | | | | 194 | correct |

# Ideas: data cleaning

## Ways to remove Unicode Characters

- Using encode() and decode() method

- Using replace() method to remove Unicode characters

- Using ord() method and for loop to remove Unicode characters

```
{
    "0": {
        "label": 0,
        "text": "atvi-20201231 UNITED STATES SECURITIES AND EXCHANGE COMMISSION Washington, D.C. 20549 FORM 10-K \u2612 ANNUAL REPORT PURSUANT TO SECTION\u00a013 OR 15(d)\u00a0OF THE SECURITIES EXCHANGE ACT OF 1934 For the Fiscal Year Ended December 31 , 2020 OR \u2610 TRANSITION REPORT PURSUANT TO SECTION\u00a013 OR 15(d)\u00a0OF THE SECURITIES EXCHANGE ACT OF 1934 For the transition period from \u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0\u00a0 to Commission File Number 1-15839 ACTIVISION BLIZZARD,\u00a0INC. (Exact name of registrant as specified in its charter) Delaware 95-4803544 (State or other jurisdiction of incorporation or organization) (I.R.S.",
        "tokenized": [
            101,
            29108,
            2072,
            1011,
            12609,
            12521,
            21486,
            2142,
```

# How can we figure out the ideal workers for our projects?

**We care about data accuracy**

- Do workers collect data correctly?

**We prefer to have a larger data library**

- Which worker is more efficient ?

# Accuracy is essential for us

**We need the half of data to train the model and the other half to test the model.**

- Check if there is some data missing

- Check if there are duplications in the collected data

Tool
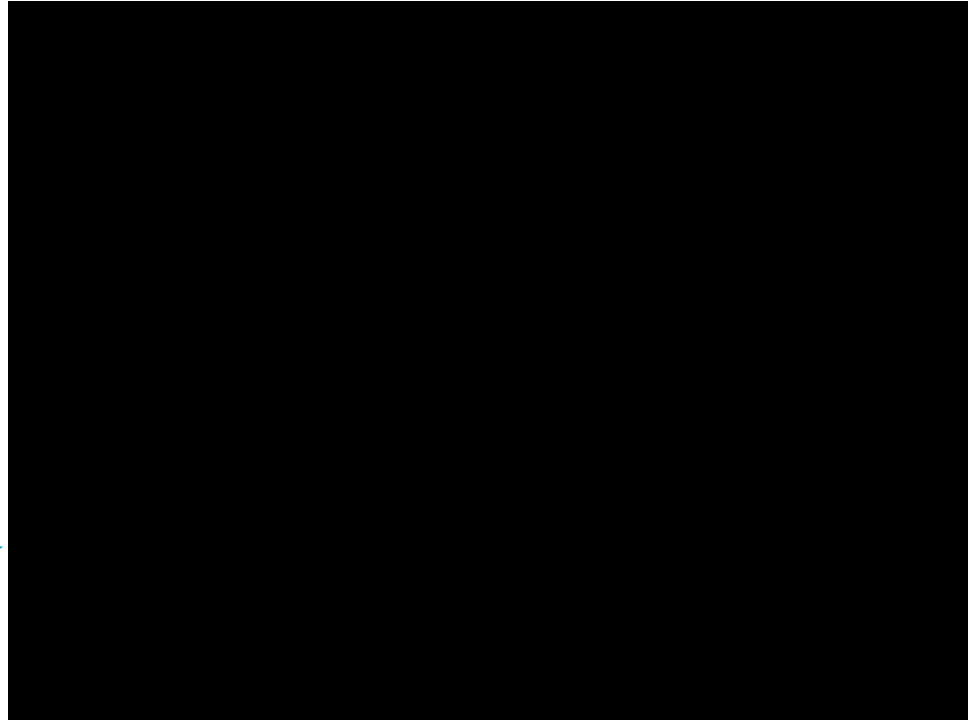
# Tasks are different

- A short task takes about 30s to fetch

- A long task is time-consuming
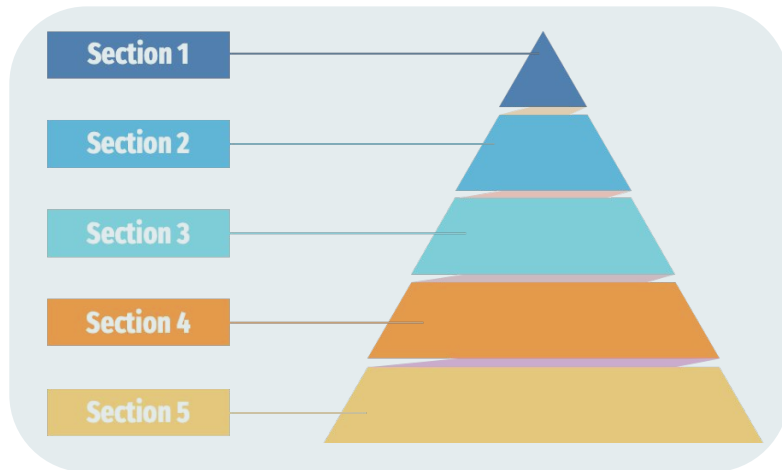
**Business Seasonality and Product Introductions**

The Company has historically experienced higher net sales in its first quarter compared to other quarters in its fiscal year due in part to seasonal holiday demand. Additionally, new product and service introductions can significantly impact net sales, cost of sales and operating expenses. The timing of product introductions can also impact the Company's net sales to its indirect distribution channels as these channels are filled with new inventory following a product launch, and channel inventory of an older product often declines as the launch of a newer product approaches. Net sales can also be affected when consumers and distributors anticipate a product introduction.

**Employees**

As of September 26, 2020, the Company had approximately 147,000 full-time equivalent employees.

**Available Information**

The Company's Annual Reports on Form 10-K, Quarterly Reports on Form 10-Q, Current Reports on Form 8-K, and amendments to reports filed pursuant to Sections 13(a) and 15(d) of the Securities Exchange Act of 1934, as amended (the "Exchange Act"), are filed with the Securities and Exchange Commission (the "SEC"). Such reports and other information filed by the Company with the SEC are available free of charge at investor.apple.com/investor-relations/sec-filings/default.aspx when such reports are available on the SEC's website. The Company periodically provides other information for investors on its corporate website, www.apple.com, and its investor relations website, investor.apple.com. This includes press releases and other information about financial performance, information on corporate governance and details related to the Company's annual meeting of shareholders. The information contained on the websites referenced in this Form 10-K is not incorporated by reference into this filing. Further, the Company's references to website URLs are intended to be inactive textual references only.

## Human Capital

We believe that our continued success and growth is directly related to our ability to attract, retain, and develop top talent. As of December 31, 2020, Activision Blizzard had approximately 9,500 employees, with approximately 65% in North America, approximately 30% in the Europe, Middle East, and Africa ("EMEA") region, and approximately 5% in the Asia Pacific region. Of these employees, approximately 61% are personnel whose primary focus is on game and technology development, which represents an approximate three percentage point increase from 2019. Activision Blizzard takes an active role in the entirety of the employee lifecycle, from candidates to alumni. Recognizing that ours is a rapidly changing industry with constant technological innovation, we remain focused on attracting, recruiting, enabling, developing, and retaining a diverse and innovative employee population.

*Diversity, Equity, and Inclusion ("DE&I")*: We believe that a culture of inclusion and diversity enables us to create, develop, and fully leverage the strengths of our workforce to exceed players' and fans' expectations and meet our growth objectives. We remain committed to building and sustaining a culture of belonging, built on equitable processes and systems, where everyone thrives. By embedding DE&I practices and programs in the full employee lifecycle, we work to recruit, attract, retain, and grow world-class talent. Our employee resource groups play an active role in our DE&I efforts by building community and awareness. We also offer leadership and management development opportunities on the topics of unconscious bias and inclusive leadership and train our recruiting workforce in diverse sourcing strategies.

Our Corporate Governance Principles and Policies provide that the initial list from which any new independent director nominee is chosen includes qualified female and racially/ethnically diverse candidates and, similarly, if we conduct an external search for a new CEO, that the initial list of external candidates includes qualified female and racially/ethnically diverse candidates. As of December 31, 2020, two of our ten directors were women.

Additionally, we have been recognized for our efforts to create an inclusive workplace, including receiving the distinction for two consecutive years as a "Best Place to Work for LGBTQ Equality" by the Human Rights Campaign Foundation's Corporate Equality Index. We are proud of these accolades because we believe that the most innovative work comes from a culture in which all employees can be, and bring, their authentic and best selves.

*Compensation and Benefits*: The main objective of our compensation program is to provide a compensation package that attracts, retains, motivates, and rewards top-performing employees that operate in a highly competitive and technologically challenging environment. We seek to do this by linking compensation (including annual changes in compensation) to overall Company and business unit performance, as well as each individual's contribution to the results achieved. The emphasis on overall Company performance is intended to align our employee's financial interests with the interests of our shareholders. We also seek fairness in total compensation by reference to external comparisons, internal comparisons, and the relationship between development and non-development, as well as management and non-management, remuneration. We believe in equal pay for equal work, and we continue to make efforts across our global organization to promote equal pay practices.

We are committed to providing comprehensive benefit options, and it is our intention to offer benefits that allow our employees and their families to live healthier and more secure lives. Some examples of our wide-ranging benefits offered are: medical insurance, prescription drug benefits, dental insurance, vision insurance, hospital indemnity insurance, accident insurance, critical illness insurance, life insurance, disability insurance, health savings accounts, and flexible spending accounts. We frequently upgrade our benefit portfolio by seeking out pioneer partners that give our employees modern benefit experiences. As an example, at the onset of the COVID-19 pandemic when traditional medical services became under huge demand, in order to help ensure that our employees and their families had access to medical advice, we created an enterprise-wide global network of physicians.

*Talent Assessment and Development*: Recognizing that ours is a rapidly changing industry with constant innovation, developing our diverse and innovative talent base is paramount, imperative, and vital to our business. We intend for our employees to have a clear understanding of their strengths and development opportunities, while fostering a collaborative and productive relationship between employees and their managers. Talent assessment and development are therefore critical aspects of our human capital programs. We employ a broad range of talent processes—for example, talent assessment, succession planning, and performance management. Our performance management process includes the establishment of goals (at the beginning of the year), and throughout the year we encourage regular check-ins on progress and performance so that employees have a clear understanding of their strengths and areas for improvement. We regularly assess employee contributions to our Company results and culture so that we can appropriately recognize and reward performance. Additionally, on an annual basis, we conduct an organizational and performance review process with our CEO and all segment, business unit, and function leaders, focusing on our high-performing and high-potential talent, diverse talent, and the performance and succession for our most critical roles.

*Employee Experience*: We capture and act on the voice of our employees through regular company-wide pulse surveys. We emphasize to employees that this is their chance to "provide honest, candid feedback about their experience working for the company." Our survey participation rates (regularly 75% or higher) demonstrate our collective commitment that Activision Blizzard remains a great place to work. The survey—and other forms of employee feedback—result in actionable steps that lead to positive improvements to the employee experience at the company-wide, business unit, and team levels. Our employee feedback is dynamic and relevant to our employees' immediate needs. For example, most recent surveys focused on whether our employees felt supported as they worked from home during the COVID-19 pandemic, and expectations of the employee population as we anticipate a return to the office.

# Check if there is some data missing

```python
def accuracyTotal(df):
    strName = 'Answer.sectionname_'
    strContect = 'Answer.sectioncontent_'
    n = 0
    for x in range(2,13):
        strSectionName = strName + str(x)
        strSectionContect = strContect + str(x)
        if df[strSectionName].count() != df[strSectionContect].count():
            #print(df[strSectionName].count(),df[strSectionContect].count())
            n += abs(df[strSectionName].count() - df[strSectionContect].count())
            print(f'Section Name/contect {x} probably missing data.')

    if n != 0:
        right = df['HITId'].count() - 1
        rate = right/df['HITId'].count()

        print(f'the accuracy rate is {rate* 100:.2f} %')

accuracyTotal(df)
```
```
Section Name/contect 5 probably missing data.
the accuracy rate is 99.90 %
```

```
41    Answer.sectioncontent_1      898 non-null    object
42    Answer.sectioncontent_10     8 non-null      object
43    Answer.sectioncontent_11     2 non-null      object
44    Answer.sectioncontent_12     1 non-null      object
45    Answer.sectioncontent_2      439 non-null    object
46    Answer.sectioncontent_3      413 non-null    object
47    Answer.sectioncontent_4      388 non-null    object
48    Answer.sectioncontent_5      299 non-null    object
49    Answer.sectioncontent_6      196 non-null    object
50    Answer.sectioncontent_7      91 non-null     object
51    Answer.sectioncontent_8      41 non-null     object
52    Answer.sectioncontent_9      21 non-null     object
53    Answer.sectionname_1         986 non-null    object
54    Answer.sectionname_10        8 non-null      object
55    Answer.sectionname_11        2 non-null      object
56    Answer.sectionname_12        1 non-null      object
57    Answer.sectionname_2         439 non-null    object
58    Answer.sectionname_3         413 non-null    object
59    Answer.sectionname_4         388 non-null    object
60    Answer.sectionname_5         300 non-null    object
61    Answer.sectionname_6         196 non-null    object
62    Answer.sectionname_7         91 non-null     object
63    Answer.sectionname_8         41 non-null     object
64    Answer.sectionname_9         21 non-null     object
```

# Check if there are duplications in the collected data

```python
strContect = 'Answer.sectioncontent_'
for x in range(2,13):

  strSectionContect1 = strContect + str(x)
  for y in range(0,1000):

    strSectionContect2 = strContect + str(x - 1)
    if df[strSectionContect1][y]!= df[strSectionContect2][y]:
      print(df[strSectionContect1][y])
      print("//")
      print(df[strSectionContect2][y])
      print("--------------------")
```

```
Within the United States, approximately 650 were hourly-rated, union
//
nan
--------------------
As of December 31, 2020, we employed approximately 57,900 people in
//
broad-based development and mentorship opportunities, recognize and
--------------------
The health and safety of our employees is one of our highest priorit
//
As of December 31, 2020, we had approximately 2,600 full-time, part-
--------------------
nan
//
The number of Company employees represented by unions, and solely wi
--------------------
•Stewardship: provide safe, reliable and environmentally sound opera
//
Each employee at SunCoke is part of our collaborative and compliment
--------------------
nan
//
None of our employees is represented by a labor union or covered by
```

# We pay them hourly, efficiency matters as well!

- Mean Completion Time per task
- Mean Completion Number per hour
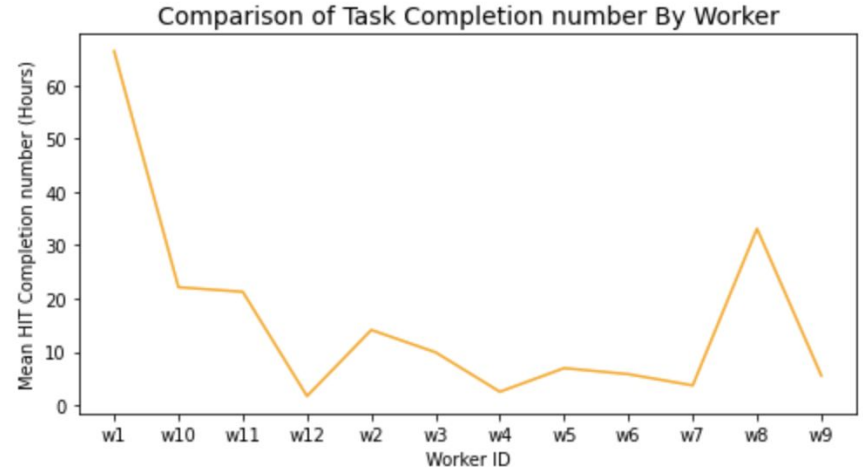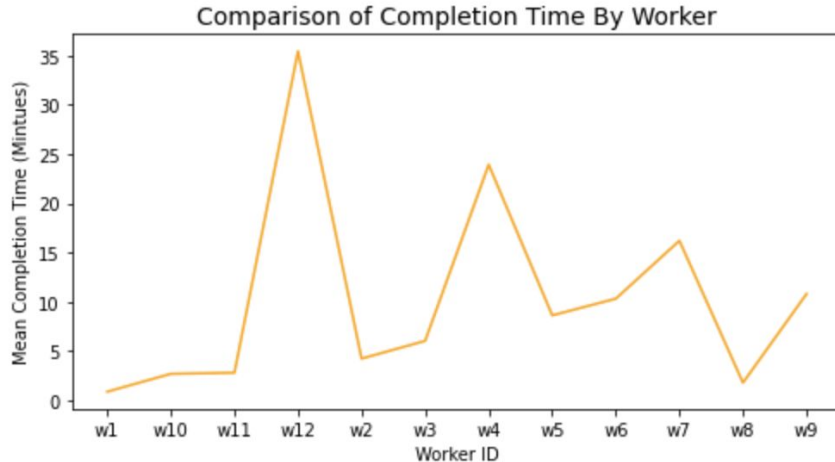
# The workers that we like

- Accuracy times efficiency

98%  100%

```
Section Name/contect 5 probably missing data.
made by worker 3      w8
10      w8
23      w8
34      w8
57      w8
        ..
935     w8
948     w8
980     w8
982     w8
990     w8
Name: WorkerId, Length: 61, dtype: object
the accuracy rate is 98.36 %
```

✕

```
The mean completion numbers (hours) for each MTurker is:

 WorkerId
w1      66.394647
w10     22.119816
w11     21.260889
w12      1.693093
w2      14.107110
w3       9.908984
w4       2.507451
w5       6.943419
w6       5.805821
w7       3.700531
w8      33.102201
w9       5.546009
Name: WorkTimeInSeconds, dtype: float64
```

32.559325

# Averages are more valuable if they are stable

- **We can also check the time variance of each HIT by worker**

- **I feel this can also tell who is the "smart worker"**

```python
# Check the time variance of each HIT by worker
completion_time_variance = df.groupby('WorkerId')['WorkTimeInSeconds'].var()
#df.WorkTimeInSeconds.var()
print(f'The mean completion time (minutes) for each MTurker is:\n\n {completion_time_variance/60}\n\n'+ '-'*70)
```

```
   WorkerId
w1        24.968373
w10      131.108333
w11     2112.390073
w12    72120.715278
w2       229.674365
w3     39192.050286
w4     35244.360658
w5      1959.751172
w6      3877.447335
w7     93323.710300
w8       133.002031
w9     12349.969769
Name: WorkTimeInSeconds, dtype: float64
```
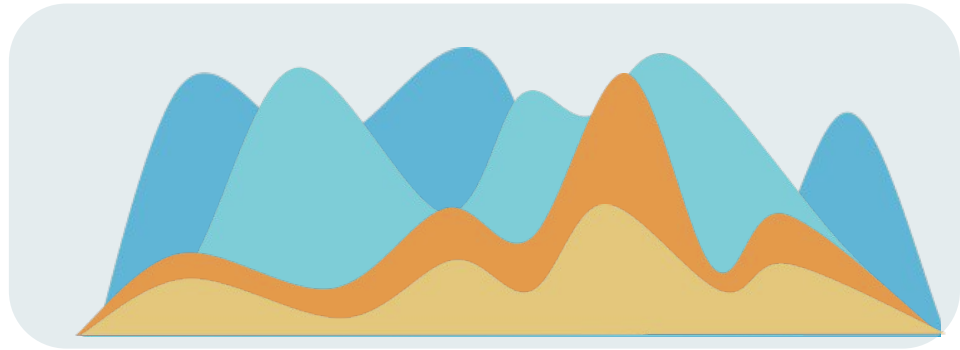
# What we need to do next

- Compare the average section length of the tasks did by each worker

- Use completion time/ section numbers


Big Data

# Thank you!