Collective Intelligence and Crowdsourcing

Maria Maistro
mm@di.ku.dk

Web Science Lecture
22 February 2021

UNIVERSITY OF COPENHAGEN

Credit for many of the slides: Christina Lioma

Previous Lecture

- Using human abilities to resolve tasks;
- Collective intelligence: the group finds better/more answers than the individuals, the law of big numbers is working!
- Crowdsourcing challenges: varying skills, ability, motivation of workers (motivation especially, e.g. less work for individuals -> better quality);
- Remove the noise.

Group Discussion

- Group of 3-4 people;
- 7 minutes discussion;
- Topic: assume you run a crowdsourcing experiment and you collect the data:
  - How can we evaluate the quality of crowdsourcing data?
  - How to detect error/noise/spam in the dataset?

Group Discussion

- Check the correctness of the assessors with the gold questions;
- Trying to estimate the task difficulty;
- Find workers with skillsets that match the task (if possible);
- Assign the same task to different workers: have enough workers (which raised the question: how do you know what number is enough? We want to avoid wasting money but maintain a good level of quality);
- Track worker reputation;
- Majority vote percentage.

Outline

- Evaluation of crowdsourced tasks:
  - Inter Assessors Agreement;
  - Worker consistency;
  - Reproducibility;
  - Error rate;
  - Statistical hypothesis testing.
- Indirect collective intelligence;
- Application of collective intelligence and Crowdsourcing.

EVALUATING CROWDSOURCING TASKS

Inter Assessor's Agreement

Klaus Krippendorff

- Idea: look at the agreement among assessors;

- **Krippendorff's alpha coefficient:**
  - A robust statistic which takes into account the probability that observed variability is due to chance;
  - Does not require that each assessor annotates each item (allows incomplete or missing data);
  - Takes into account the type of data (nominal, ordinal, interval or ratio) being measured;
  - Large and small sample sizes alike, not requiring a minimum.

Krippendorff's alpha

- Krippendorff's alpha range in [-1, 1]:
  - α = 1 perfect agreement between assessors;
  - α = 0 observed agreement is equal to the level of agreement expected by chance;
  - α < 0 disagreements are systematic and exceed what can be expected by chance.

- Reliability Data Matrix:

Units ***u:***	**1**	**2**	.	.	.	***u***	.	.	.	.	.	***N***
Observers: **1**	c_{11}	c_{12}	.	.	.	c_{1u}	.	.	.	.	.	c_{1N}
i	c_{i1}	c_{i2}	.	.	.	c_{iu}	.	.	.	.	.	c_{iN}
j	c_{j1}	c_{j2}	.	.	.	c_{ju}	.	.	.	.	.	c_{jN}
.	.	.				.						.
m	c_{m1}	c_{m2}	.	.	.	c_{mu}	.	.	.	.	.	c_{mN}
Number of observers valuing ***u:***	m_1	m_2	.	.	.	m_u	.	.	.	.	.	m_N

Reliability Data Matrix: Example

- Rows represents the assessors;
- Column represents the items being labelled;
- We can have missing values, i.e. some of the items are not assessed by all assessors.

For example, a 4 observers-by-12 units reliability data matrix:

Units u:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
Observer ***A***:	1	2	3	3	2	1	4	1	2	.	.	.	
Observer ***B***:	1	2	3	3	2	2	4	1	2	5	.	3	
Observer ***C***:	.	3	3	3	2	3	4	2	2	5	1	.	
Observer ***D***:	1	2	3	3	2	4	4	1	2	5	1	.	
Number m_u of values in unit u:	3	4	4	4	4	4	4	4	4	3	2	1	41

Note that 7 out of the 48 possible values in this matrix are missing. m_u varies from 1 to 4.

Compute Krippendorff's alpha

$$\alpha = 1 - \frac{D_0}{D_e}$$

- D_o is the observed disagreement;
- D_e is the disagreement expected by chance;

The actual computation of D_o and D_e depends on:

- Missing data;
- Type of data.

$$\alpha = 1 - (n_{..} - 1)\frac{\sum_u \frac{1}{n_{u.}-1} \sum_c \sum_{k>c} n_{uc} n_{uk} \delta^2_{ck}}{\sum_c \sum_{k>c} n_{.c} n_{.k} \delta^2_{ck}}$$

Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability.

Outline

- Evaluation of crowdsourced tasks:
  - ~~Inter Assessors Agreement;~~
  - Worker consistency;
  - Reproducibility;
  - Error rate;
  - Statistical hypothesis testing.
- Indirect collective intelligence.

Measure Worker Consistency

Measure how consistently workers perform throughout the task:

- not in terms of their answers, but in terms of their modus operandi.

Goal: capture diversity without very extreme outliers

Common practice:

- Mean task completion time (measured in milliseconds);
- Per-worker standard deviation in task completion times;

Challenges

Problem:

- mean & standard deviation are affected by *very* extreme outliers;
- very extreme outliers are rare in general, but not so rare in crowdsourced data.

Common example:

- worker spends 1.5 seconds on trivial task A;
- same worker spends >2 minutes on trivial task B (distracted by external event in the middle of the experiment).

Example

Toy example of mean & standard deviation being affected by *very* extreme outliers

- Input: 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4 (no extreme outlier)
- Mean: 2.45
- Standard deviation: 1.04

- Input: 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, **400** (with extreme outlier)
- Mean: 35.58
- Standard deviation: **114.77**

How to solve this problem?

Solution:

- First, detect and remove *very* extreme outliers;
- Then, compute mean and standard deviation in task completion time.

Detect very extreme outliers based on (either):

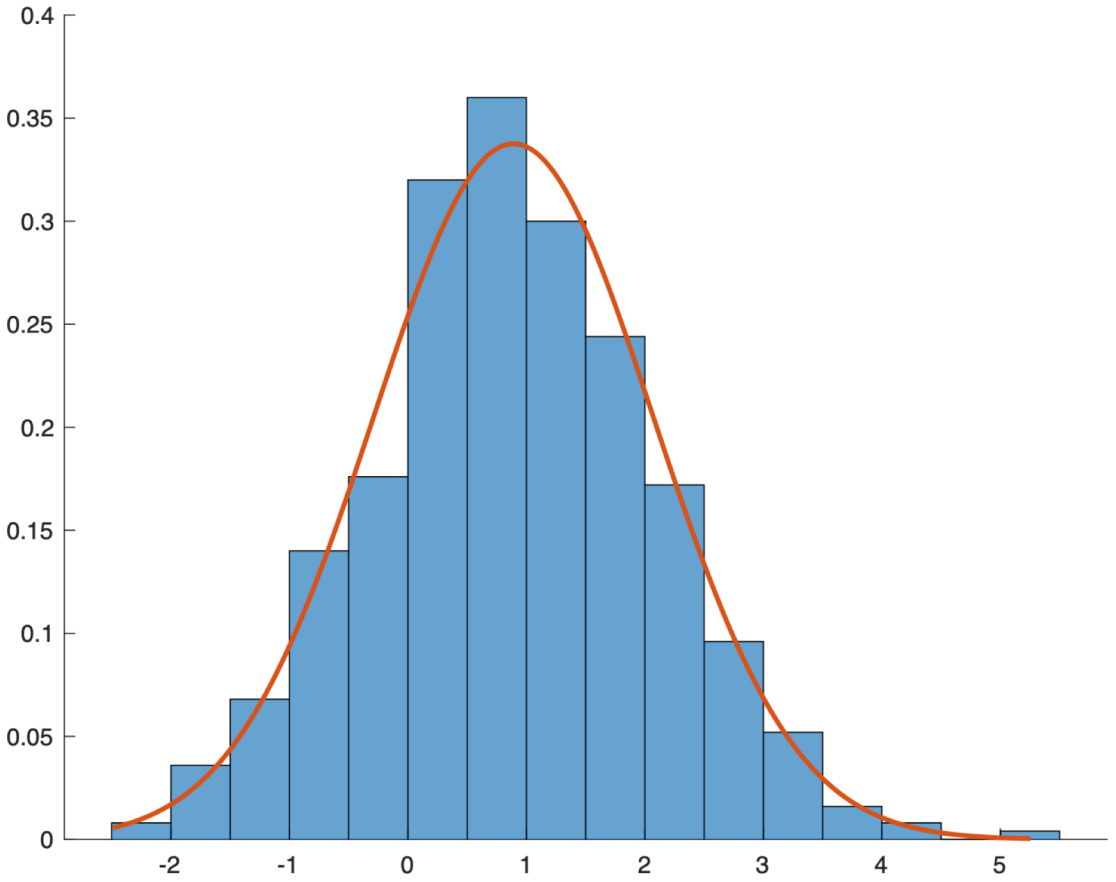
- (log-transformed) per-worker mean time;
- (log-transformed) per-worker maximum time.

Two common ways:

1. Mean and standard deviation
2. Inter-Quartile Range (IQR)

Plot the data!

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix}$$



Method (1)

- Compute mean & standard deviation
- Data that is more than two standard deviations from the mean → *very* extreme outliers
- Remove extreme outliers and re-compute mean & standard deviation
- Assumption: data is normally distributed

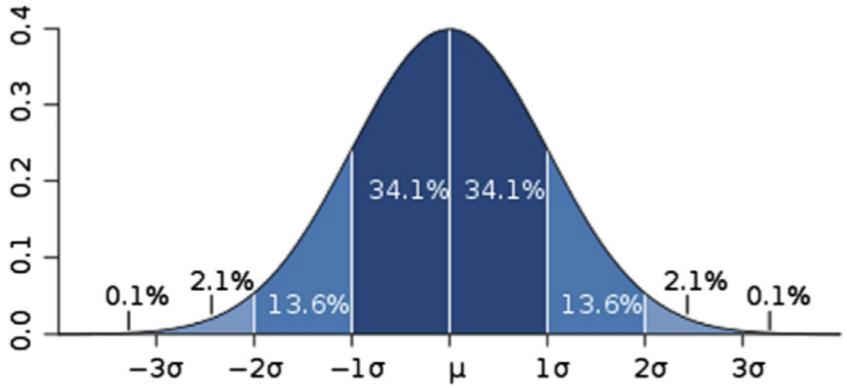


Figure 3 The normal curve.

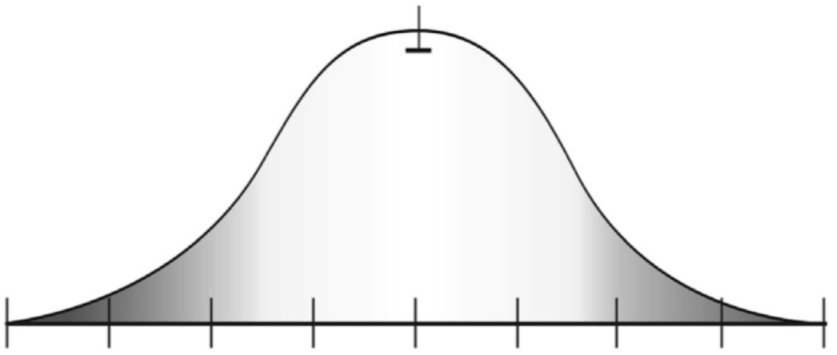


Figure 4 The normal curve (with darker areas at either extreme from the median).

Method (1)

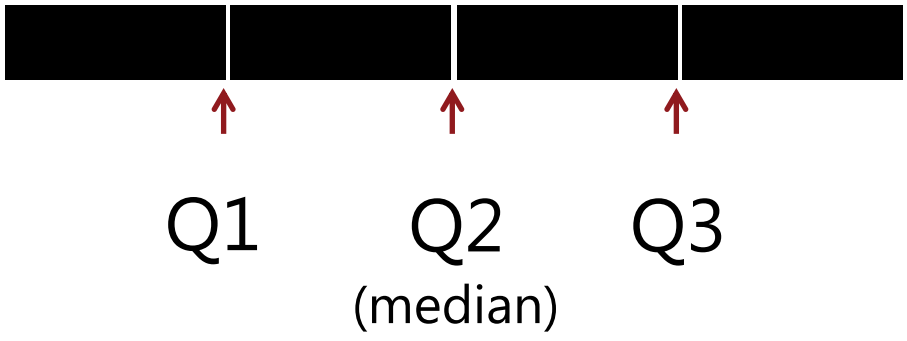
- Compute mean & standard deviation
- Data that is more than two standard deviations from the mean → *very* extreme outliers
- Remove extreme outliers and re-compute mean & standard deviation
- Assumption: data is normally distributed

Problem:

- *Very* extreme outliers bias mean & standard deviation
- Not robust

Method (2)

- Detect outliers with InterQuartile Range (IQR):
- Sort input data;
- Divide sorted data into 4 equal parts;
- IQR = Q3 – Q1;



- Extreme outlier:
- 3 x IQR more than Q3
- 3 x IQR less than Q1

Will this remove more or less outliers than method 1?

Example

Input range: 1, 3, 4, 5, 5, 6, 7, 11

- Method 1: 2 standard deviations from mean
- $\mu = 5.25$, $\sigma = 2.96$
- Extreme outliers: less than -0.67 and more than 11.17
- Removes 4.6% of the data

- Method 2: 3 x IQR from Q1 and Q3
- Q1 = 3.5, Q3 = 6.5, IQR = 6.5 – 3.5 = 3
- Extreme outliers: less than -5.5 and more than 15.5
- Removes 0.0002% of the data (legitimate diversity kept)

Outline

- Evaluation of crowdsourced tasks:
  - ~~Inter Assessors Agreement;~~
  - ~~Worker consistency;~~
  - Reproducibility;
  - Error rate;
  - Statistical hypothesis testing.
- Indirect collective intelligence.

Reproducibility

- Can the same conclusions be drawn if the study is repeated?
- Large variation of worker population over different times of the day → significant impact on performance

	Weekday 11am	Weekday 7pm
Female	74%	57%
Median age	43	32
Use mouse	63%	86%
Use touchpad	27%	14%

Outline

- Evaluation of crowdsourced tasks:
  - ~~Inter Assessors Agreement;~~
  - ~~Worker consistency;~~
  - ~~Reproducibility~~;
  - Error rate;
  - Statistical hypothesis testing.
- Indirect collective intelligence.

Error Rate in Crowdsourcing

- E.g. label pictures showing cats (labels: cat, not-cat)
- How many cats were found out of all cats?
- How many non-cat pictures were mislabelled as cats?
- How many pictures were correctly labelled out of all pictures?

Precision and Recall

		Ground truth	
		true	false
Worker answer	true	true positive (TP)	false positive (FP)
	false	false negative (FN)	true negative (TN)

Precision and Recall

		Ground truth	
		true	false
Worker answer	true	true positive (TP)	false positive (FP) **Type I error**
	false	false negative (FN)	true negative (TN)

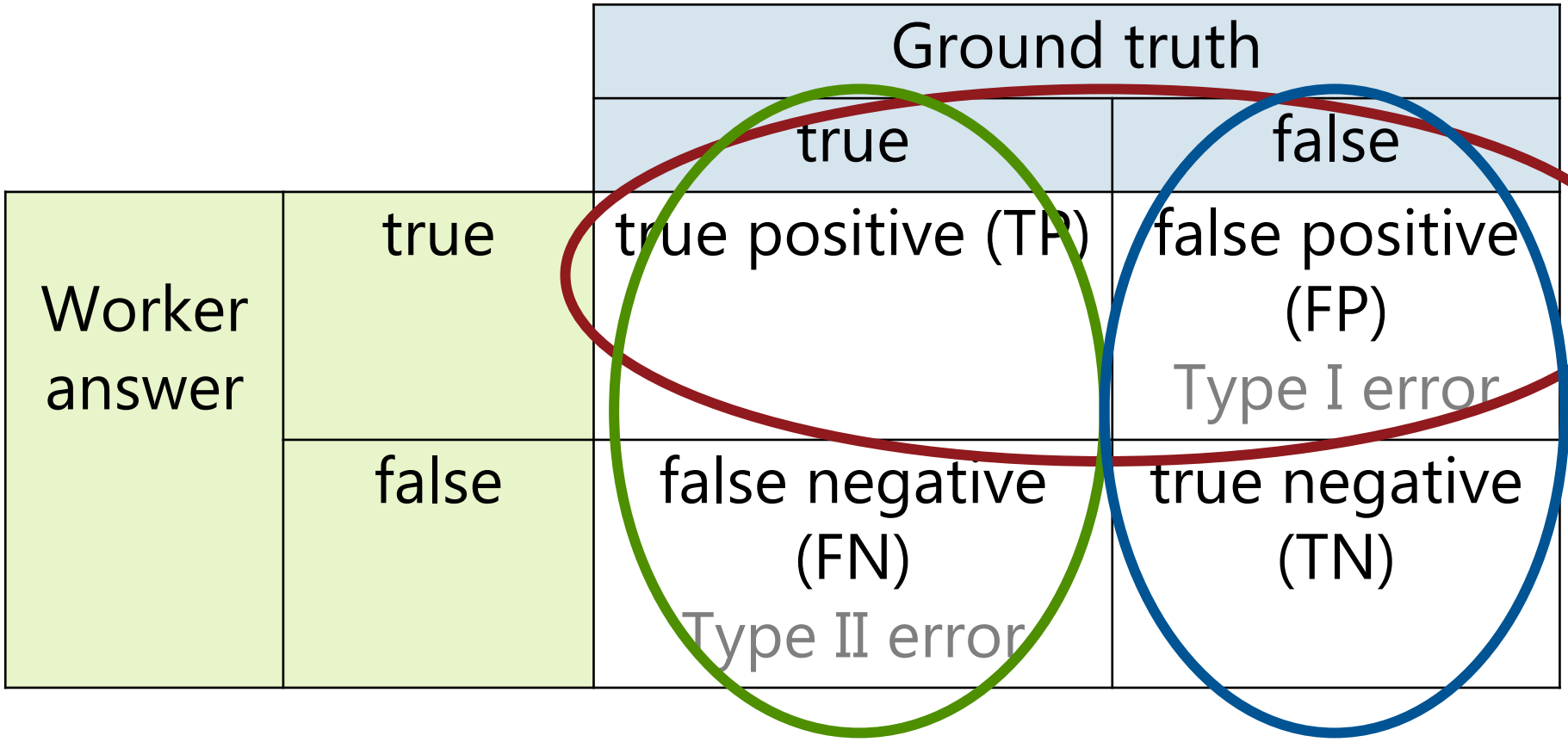
- How many non-cat pictures were mislabelled as cats?

Precision and Recall

		Ground truth	
		true	false
Worker answer	true	true positive (TP)	false positive (FP) Type I error
	false	false negative (FN) **Type II error**	true negative (TN)

- How many cat pictures were mislabelled as not-cats?

Precision and Recall



- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)

Specificity = TN / (TN+FP)

Accuracy

Accuracy: (TP+TN) / (TP+TN+FP+FN)

		Ground truth	
		true	false (95)
Worker answer	true	true positive	false positive
	false	false negative	true negative

Imbalanced dataset (95 false and 5 true labels) → classifying all as false gives 0.95 accuracy

Balanced Accuracy

Balanced accuracy = (TPR + TNR)/2

		Ground truth	
		true (5)	false (95)
Worker answer	true (0)	true positive (0)	false positive (0)
	false (100)	false negative (5)	true negative (95)

- Balanced accuracy = (0/5 + 95/95)/2 = 47.5
- Balanced accuracy = (recall + specificity)/2

Outline

- Evaluation of crowdsourced tasks:
  - ~~Inter Assessors Agreement;~~
  - ~~Worker consistency;~~
  - ~~Reproducibility~~;
  - ~~Error rate;~~
  - Statistical hypothesis testing.
- Indirect collective intelligence.

Are they different?

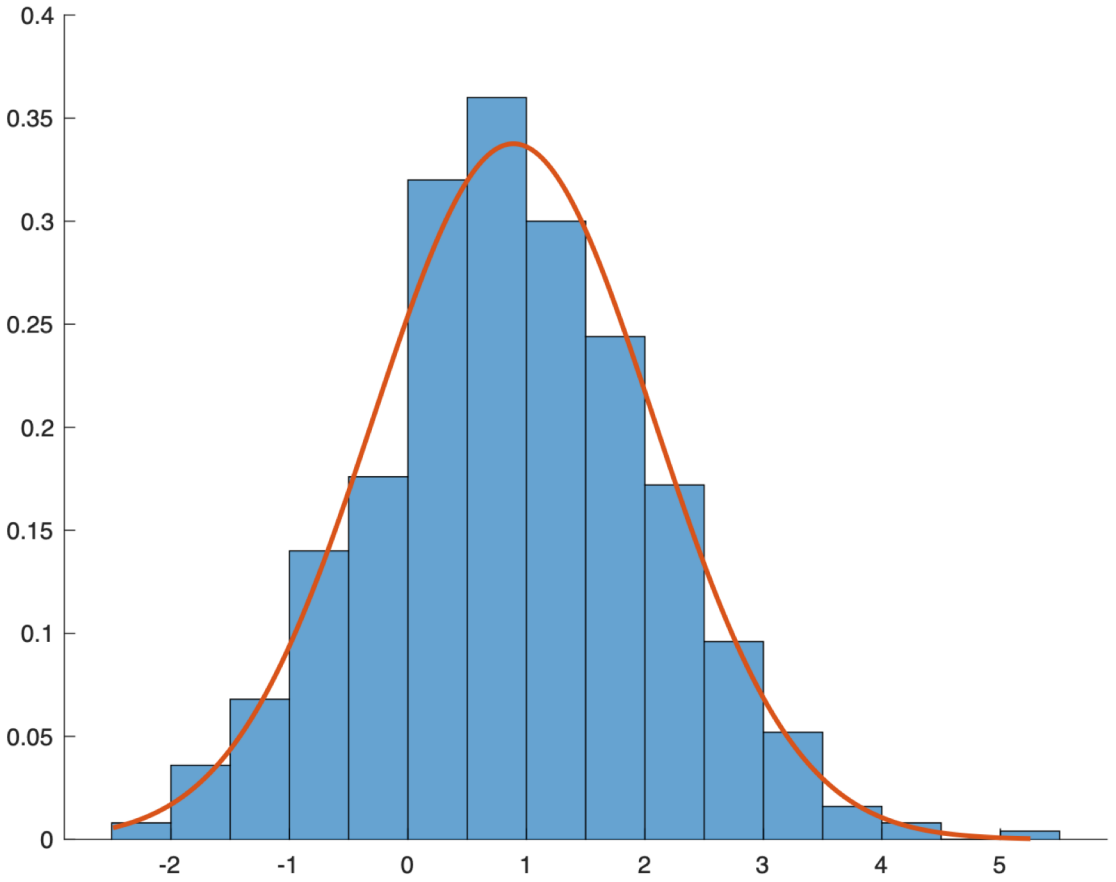
Is the observed difference larger than what you would expect from random fluctuations alone?

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \qquad y = \begin{bmatrix} 0.7471 \\ 1.6357 \\ 1.0390 \\ 1.1421 \\ 1.5731 \\ 3.5389 \\ 1.9563 \\ 0.6754 \\ 2.6305 \\ 1.9505 \end{bmatrix} \qquad z = \begin{bmatrix} 0.6408 \\ 2.6859 \\ 3.0818 \\ 3.1258 \\ 2.9028 \\ 2.4371 \\ 1.7648 \\ 3.2616 \\ 2.0997 \\ 1.0346 \end{bmatrix}$$

$$\hat{\mu}_X = 0.8497 \qquad \hat{\mu}_Y = 1.6889 \qquad \hat{\mu}_Z = 2.3035$$
$$\hat{\sigma}^2_X = 1.5284 \qquad \hat{\sigma}^2_Y = 0.7890 \qquad \hat{\sigma}^2_Z = 0.8255$$

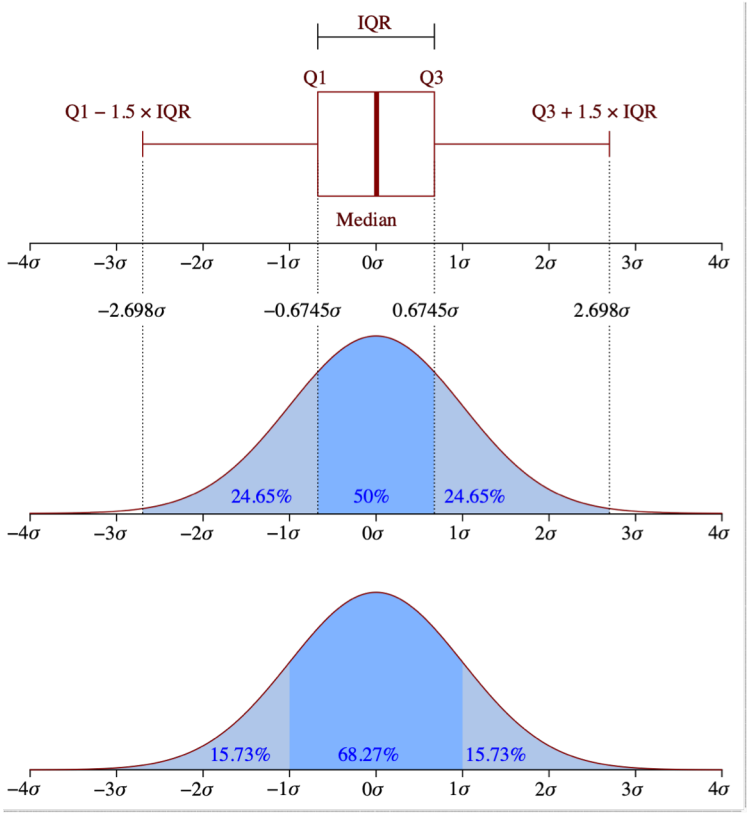
Plot the data!

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix}$$



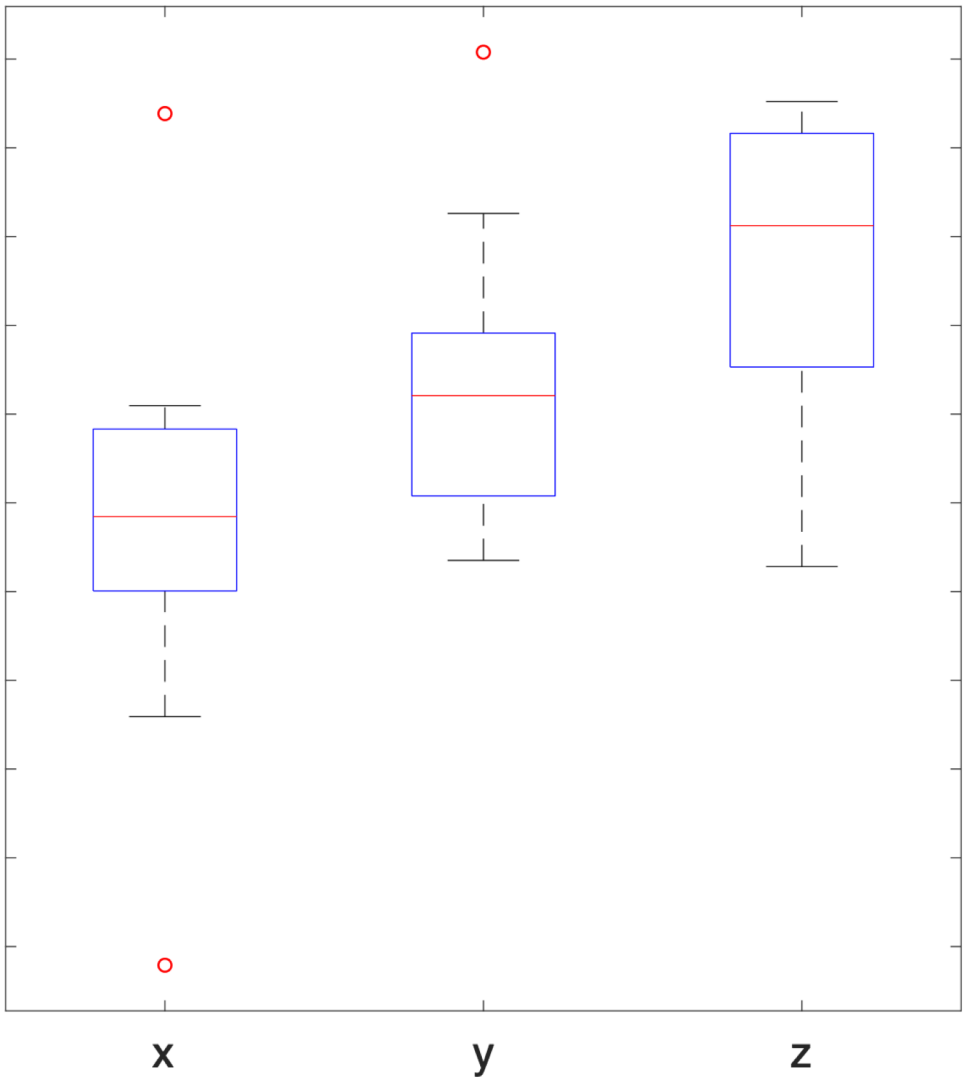
Boxplots

- A **boxplot** is a graphical tool to summarise a distribution of data
- The box shows the first quartile (Q1), the second quartile (Q2, the median) with a line inside the box, and the third quartile (Q3)
- The box represent the Inter-Quartile Range (IQR), i.e. the difference Q3-Q1
- The extension of the whiskers represents 1.5·IQR
- They roughly cover ~99% of the data, assuming a normal distribution
- Any data outside the whiskers is considered an outlier



Are they different?

$$x = \begin{bmatrix} 3.1 \\ 0.6 \\ 1.4 \\ -0.2 \\ -1.6 \\ 1.1 \\ 1.1 \\ 0.7 \\ 0.5 \\ 1.5 \end{bmatrix} = \begin{bmatrix} 0.6408 \\ 2.6859 \\ 3.0818 \\ 3.1258 \\ 2.9028 \\ 2.4371 \\ 1.7648 \\ 3.2616 \\ 2.0997 \\ 1.0346 \end{bmatrix}$$



$\hat{\mu}_X =$

$\hat{\sigma}^2_X =$

$\hat{\mu}_Z = 2.3035$

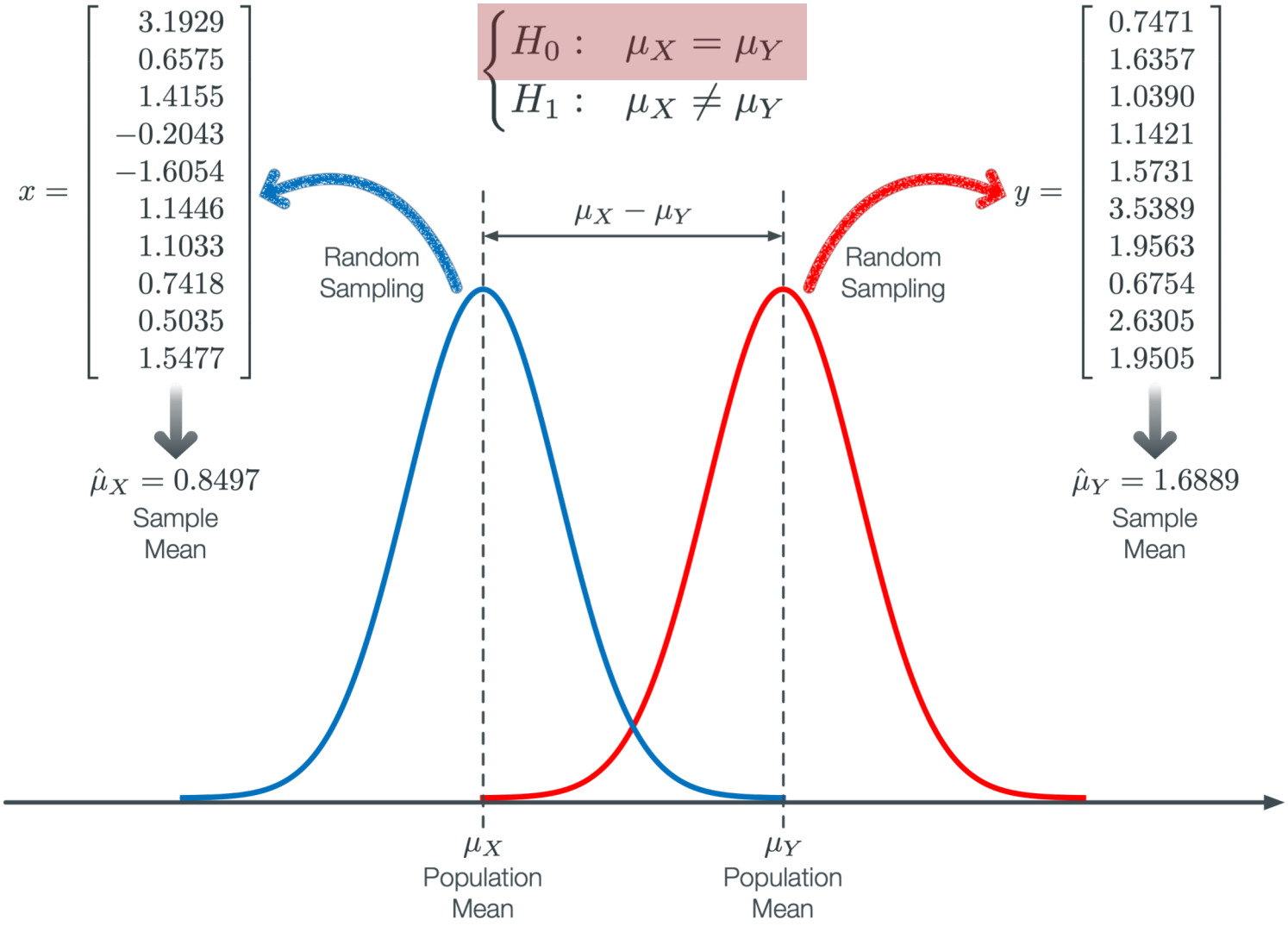
$\hat{\sigma}^2_Z = 0.8255$

Statistical Hypothesis Testing

William Sealy Gosset
"Student"

- Statistical hypothesis testing provides us with a mathematical framework to conduct statistical inference from the data;
- It compares the so-called null hypothesis H_0 against an alternative hypothesis H_1 or H_A;
- The comparison is statistically significant if the data are unlikely to be a realisation of the null hypothesis with respect to a chosen threshold, called significance level α. In this case we reject the null hypothesis; in the opposite case, we fail to reject the null hypothesis.

Formulating the Problem



Significance Level Alpha

- Alpha = the probability of making a wrong decision;
- Interpretation of alpha = 0.05: If we repeat the same experiment multiple times, 95% of the times the data distributions will be considered statistically different, 5% of the times they will not;
- Common values for alpha = 0.05, 0.01.

Basic Idea: Types of Error

	We fail to reject H_0 **[not statistically significant]**	We reject H_0 **[statistically significant]**
H_0 is true **[e.g. distributions are equivalent]**	Correct conclusion Probability 1 - α	Type I error Probability α
H_0 is false **[e.g. distributions are not equivalent]**	Type II Error Probability β	Correct conclusion Probability 1 - β

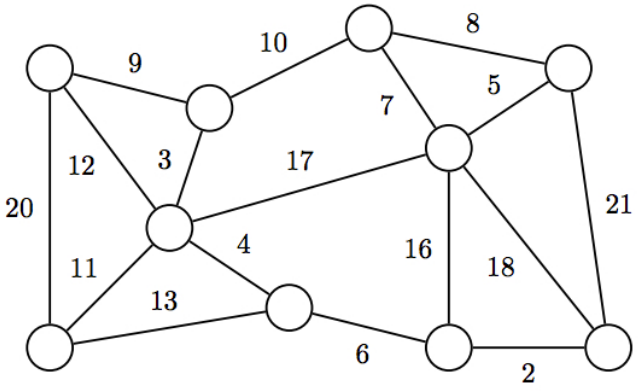
COLLECTIVE INTELLIGENCE

Indirect Collective Intelligence

- Direct: explicit information mining from users.
- Indirect: implicit information mining from user.
  Famous case "find best webpages on the web":
  - "Best" approximated as "most popular";
  - "Popular" approximated as "hyperlinked";
  - Mine hyperlinks given to webpages by large crowds → detect best webpages;

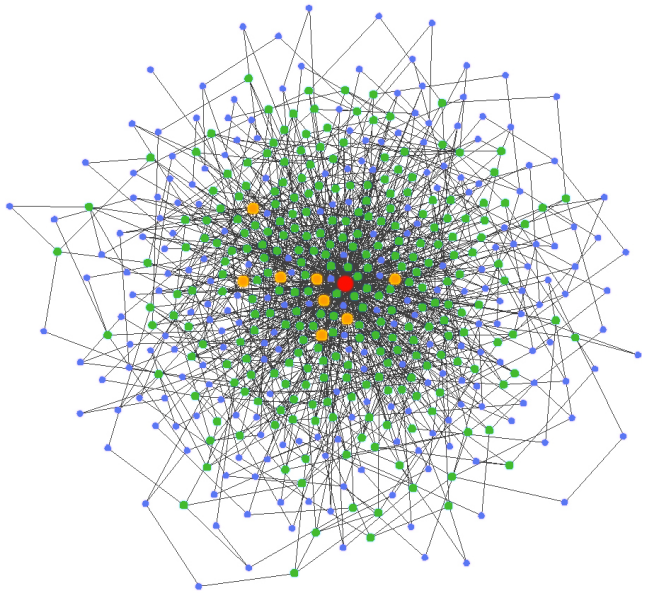
1921: George Polya, Prof. Mathematics

"I did not foresee that he and his fiancee would also set out for a stroll in the woods, and then suddenly I met them there. And then I met them the same morning repeatedly. I don't remember how many times, but certainly much too often and I felt embarrassed: It looked as if I was snooping around which was, I assure you, not the case"



1996: Two CS students, University of Stanford

- Network: 518 million nodes (=webpages);
- Computer simulates Polya's walk (100s of times);
- Finding: most popular webpages;
- PageRank algorithm.



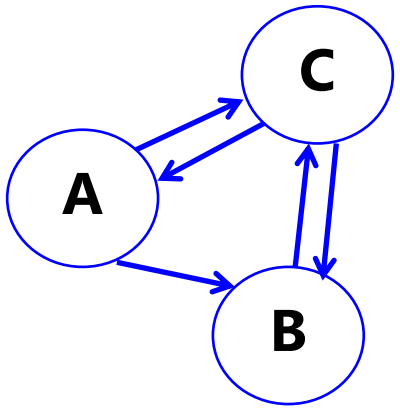
2019: Sergey Brin, Larry Page, Google Founders

Google:

- 150 trillion webpages in 2017
- 3.5 billion searches per day (world population: 7 billion)
- **Indirect collective intelligence of hyperlinks**

https://searchengineland.com/googles-search-indexes-hits-130-trillion-pages-documents-263378

PageRank

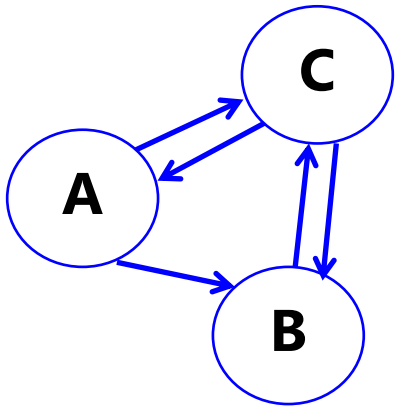


PAGE	IN	OUT
A	1	2
B	2	1
C	2	2

- Assume all pages are equally good.
  StartScore = 0.5. Aim: EndScore = ?
- Take a random page, e.g. A, and ask:
  (i) How many inlinks (IN) does A have?
  1, from C. For each IN:
  (ii) How many outlinks (OUT) does C have? 2
  (iii) What is C's StartScore? 0.5

EndScore (A) = Σ _(A's IN) (StartScore of IN / # OUT of IN)
EndScore(A) = 0.5 / 2 = 0.25

- PageRank(A) = **(1 – d) + d x EndScore(A)**
  d: probability that user jumps between pages (d ∈ [0, 1])



EndScore (A) =
Σ _(A's IN) (StartScore of IN / # OUT of IN)
PageRank(A) = **(1 – d) + d x EndScore(A)**
(let d = 0.85)
1st iteration:
PageRank(A) = (1 – 0.85) + 0.85 x 0.25 = 0.3625
PageRank(B) = (1 – 0.85) + 0.85 x 0.5 = 0.575
PageRank(C) = (1 – 0.85) + 0.85 x 0.75 = 0.7875
2nd iteration: repeat using the PageRank scores of the 1st iteration as StartScores
PageRank(A) = (1 – 0.85) + 0.85 x (0.7875/2) ≈ 0.5
PageRank(B) = (1 – 0.85) + 0.85 x (0.3625/2 + 0.7875/2) ≈ 0.6
PageRank(C) = (1 – 0.85) + 0.85 x (0.3625/2 + 0.575/1) ≈ 0.8
3rd iteration: repeat using the PageRank scores of the 2nd iteration as StartScores

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

$PR(p_x)$: pagerank score of webpage p_x

d: damping factor

N: total number of webpages

$M(p_i)$: set of p_i's inlinks

$L(p_j)$: cardinality of the set of p_j's outlinks

After 100s iterations, score of webpage quality:

- How many webpages point to it;
- How good those webpages are themselves.

Some References

- Komarov, et al. 2013. “Crowdsourcing Performance Evaluations of User Interfaces.” In proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, April 27-May 2, 2013.

- Page et al. 1999. “The PageRank Citation Ranking: Bringing Order to the Web.” Technical report, Stanford Infolab.

Crowdsourcing and collective intelligence

In Internet culture

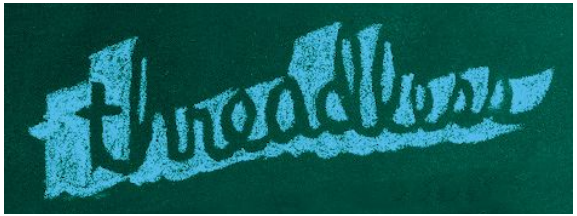
Kacper Lesniak

Web 2.0

Web 2.0 refers to the second generation of the Web, wherein interoperable, user-centered web applications and services promote social connectedness, media and information sharing, user-created content, and collaboration among individuals and organizations. [1]

Crowdsourcing as a business

- Threadless, a t-shirt company, where designs are crowdsourced (2000)
  - Designers receive a cash prize
  - Demand is guaranteed
- Innocentive, an open innovation company, where workers solve scientific and industrial problems for prize money (2001)
  - Over 2,000 challenges
  - 80% of them were awarded

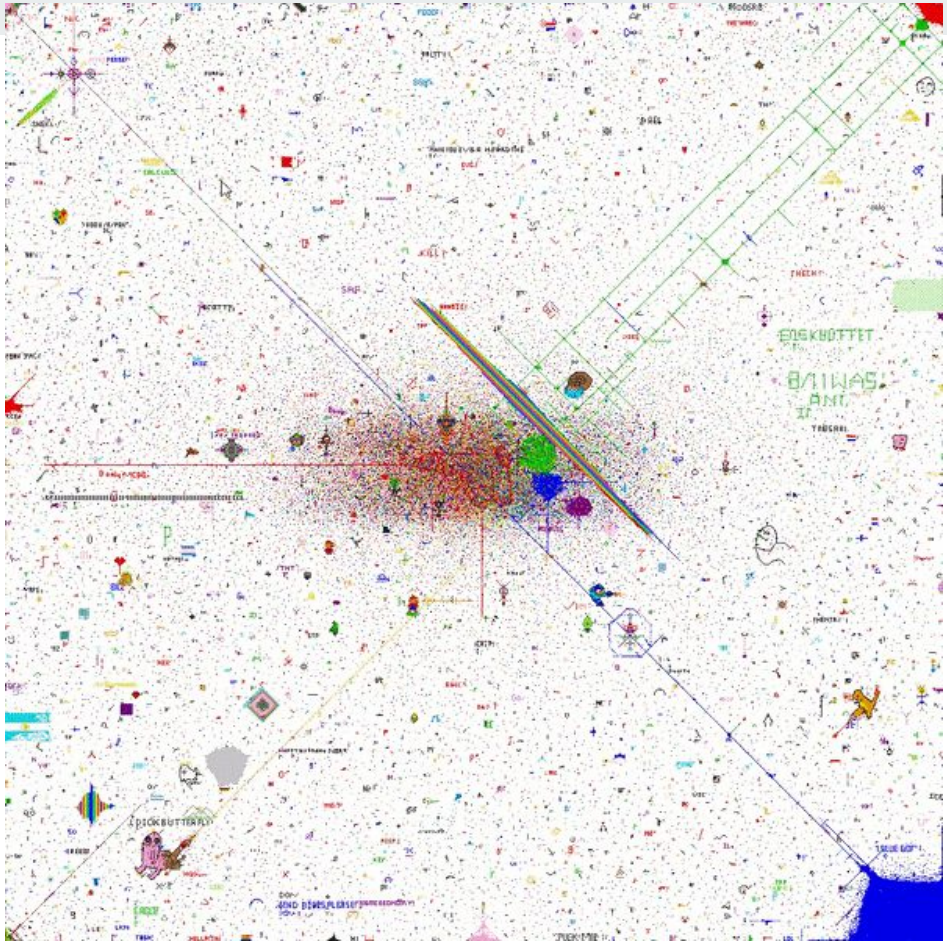


innocentive©
a wazoku brand

Crowdsourcing as a social experiment

Reddit r/place

- April 1st, 2017 - 48 hours
- 1 pixel per one user per 5-20 minutes
- Over 1 million participants
- Self-organising unsupervised communities
- Common goal (?)

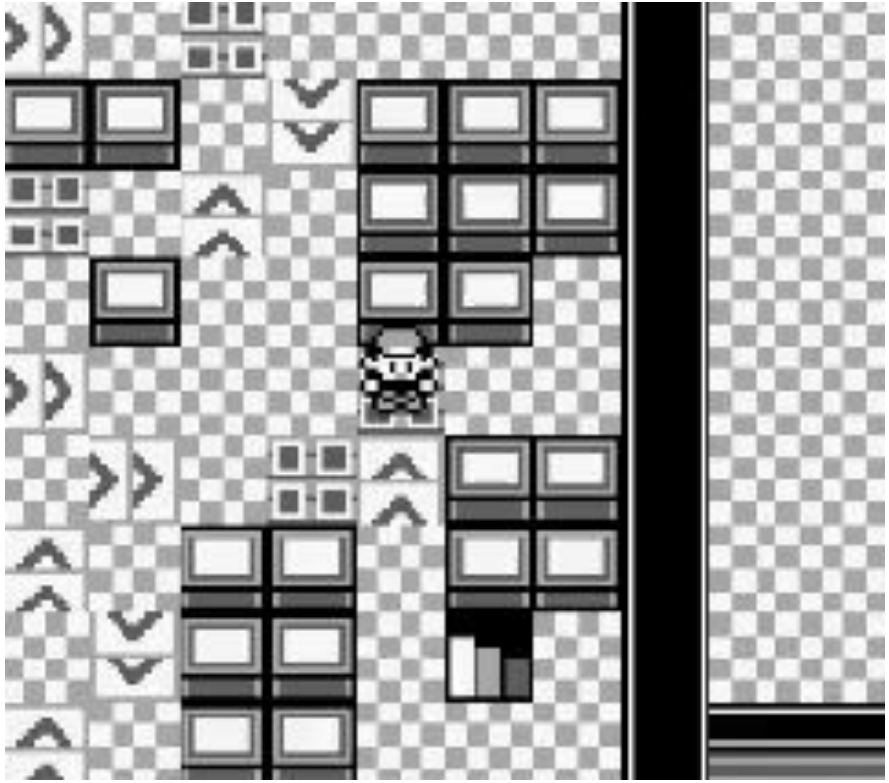


Twitch plays pokemon

- 12th February, 2014
- Peak of 121,000 concurrent viewers
- Total of 1.16 million participants
- Guinness World Record for having "the most participants on a single-player online video game"

Democracy and anarchy

- Some parts of the game were too hard to beat given the chaotic nature of inputs
- Two modes, anarchy, the classic one, and democracy, in which every given interval the game applied most popular input
- Finished after over 16 days of continuous gameplay (for a human player it usually takes 27.5 hours [2])



What if the system was even smarter?

- Could a crowd controlled player be more successful than an average singular player?
- How it would compare with experienced players?

My master thesis

- Real-time crowdsourced controller
- Input coming from network players and AI agents
- Ongoing assessment of player reliabilities based on inter-assessor agreement

User study - I need your help!

- Around beginning of April
- 10 simultaneous players controlling one character
- (hopefully) in DIKU buildings
- Will post more information on Web Science absalon page

Thank you!

Sources

[1] - Wilson, David & Lin, Xiaolin & Longstreet, Phil & Sarker, Saonee. (2011). Web 2.0: A Definition, Literature Review, and Directions for Future Research..

[2] - https://howlongtobeat.com/game.php?id=7169

Daren C. Brabham. 2013. Crowdsourcing. The MIT Press.

https://en.wikipedia.org/wiki/Twitch_Plays_Pok%C3%A9mon

https://en.wikipedia.org/wiki/Place_(Reddit)