

Data Annotation & Evaluation

Dr. Tirthankar Dasgupta

Recap...

NLP and Text Processing Tasks

- Extracting Drug names, disease names, product mentions, email ids, addresses

Information
Extraction

- Extracting Events / Sentiments Who did / said What, When, How, Why

Linguistic
Analysis

- How many negative reviews – on which aspect of product / service

Statistical
Analysis

- Risk assessment – new policy about environment / health care / education

Drawing
Inference

- What are the causes a disease?
- What is the policy for Paternity leave?

Question
Answering

- Summary of Financial obligations on signing a new contract with an alliance

Summarization

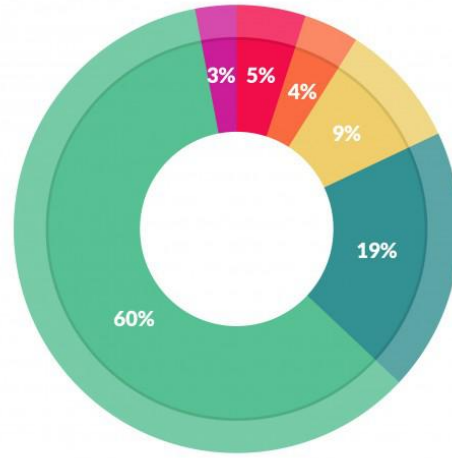
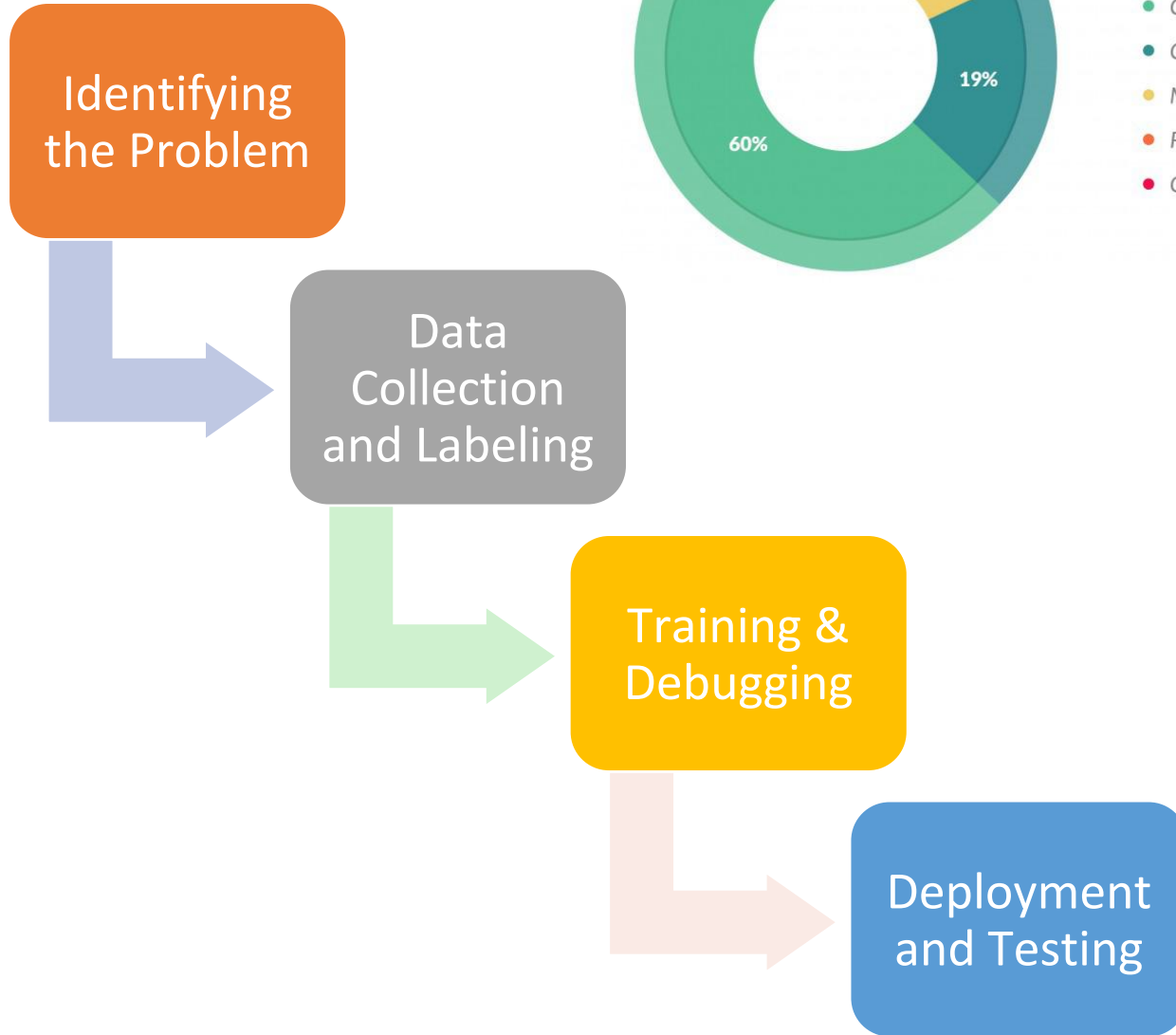
- General Purpose Conversation - Question
- Understand & Empathize
- Answer / Guide / Recommend

Natural
Language
generation

- Any language to any language

Machine
Translation

The Process Flow



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Types of NLP Problem

Core problems

- Phoneme boundary identification (Speech)
- Morphological analyzer
- POS tagging
- Parsing
- ...

Applications

- Named entity
- Sentiment analysis
- Adverse Drug Effects
- Hate Speech
- Document classification
- ...

Types of Annotation

- Phoneme level
- Morpheme level
- Word level
- Phrase level
- Sentence level
- Document level

Applications

- POS tagging
- Sentiment Analysis
- Aspect Extraction

IDENTIFY PARTS OF SPEECH

John saw the Girl

PN V DT N

“saw” is more likely to be a verb V
rather than a noun N

John saw the saw

The second “saw” is a noun N because a noun N
is more likely to follow a determiner.

 NN
 RB
 JJ VB
 VBN VBD TO VB DT NN
PRP

She promised to back the bill

ESTIMATING THE PROBABILITIES

- How can I know $P(V | PN)$, $P(\text{saw} | V)$?
- Obtaining from training data

Training Data:

- (x^1, \hat{y}^1) 1 Pierre/**NNP** Vinken/**NNP** ,/, 61/**CD** years/**NNS** old/**JJ** ,/, will/**MD** join/**VB** the/**DT** board/**NN** as/**IN** a/**DT** nonexecutive/**JJ** director/**NN** Nov./**NNP** 29/**CD** ./.
- (x^2, \hat{y}^2) 2 Mr./**NNP** Vinken/**NNP** is/**VBZ** chairman/**NN** of/**IN** Elsevier/**NNP** N.V./**NNP** ,/, the/**DT** Dutch/**NNP** publishing/**VBG** group/**NN** ./.
- (x^3, \hat{y}^3) 3 Rudolph/**NNP** Agnew/**NNP** ,/, 55/**CD** years/**NNS** old/**JJ** and/**CC** chairman/**NN** of/**IN** Consolidated/**NNP** Gold/**NNP** Fields/**NNP** PLC/**NNP** ,/, was/**VBD** named/**VBN** a/**DT** nonexecutive/**JJ** director/**NN** of/**IN** this/**DT** British/**JJ** industrial/**JJ** conglomerate/**NN** ./.

⋮

Structured prediction (sequence tagging) – label for a word depends on other labels

- Rather than classifying each word independently

Applications

- POS tagging
- Sentiment Analysis
- Aspect Extraction

Sentiment Analysis

This is a good book. 📊 Positive

This book is simply unputdownable! 📊 More Positive

This is a bad book. 📊 Negative

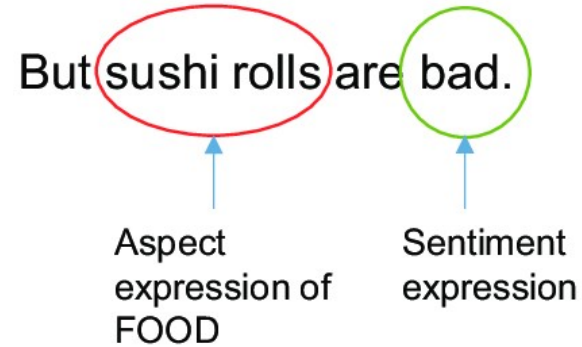
The first chapter of the book is great, but the rest is a junk! 📊 Positive & Negative

Applications

- POS tagging
- Sentiment Analysis
- Aspect Extraction

Aspect Analysis

I was tempted to buy this product as I really like its design, but its price is not very good.

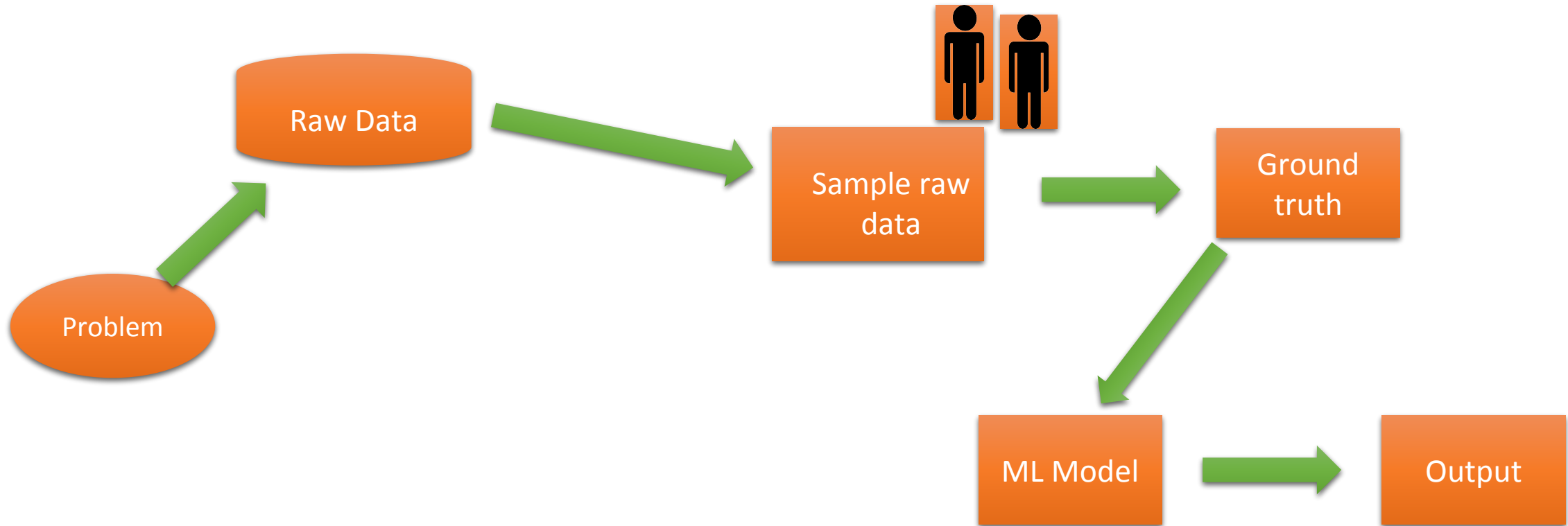


This product has a good price; the one my brother purchased has a good design.

Question 1: How would you rate our service?
Rating Answer: "10"

Question 2: Motivate your answer.
Textual Answer: "Customer care"

Revisiting the process flow



Data Collection

- After we define what we are going to create, baseline, and metrics in the project, the most painful of the step will begin, **data collection and labeling**.
- Most of Deep Learning applications will require a lot of data which need to be labeled.
 - Time will be mostly consumed in this process.
- Although you can also use public dataset, often that labeled dataset needed for our project is not available publicly.

Data Collection



- Strategy:
 - Need to plan how to obtain the complete dataset.
 - Multiple ways to obtain the data
 - The data need to align according to what we want to create in the project.
- Ingest:
 - If the strategy to obtain data is through the internet by scraping and crawling some websites, we need to use some tools to do it.
 - Scrapy is one of the tool that can be helpful for the project

Data Annotation & Evaluation

Part-II

Recap

A NOTE ON DATA COLLECTION

- The raw data needs to fit a certain profile.
 - Will also determine how much data needs to be annotated.
- The ideal training corpus has the following key features:
 - It should be **representative**, covering the **domain vocabulary**, **format**, and **genre** of the text.
 - It should be **balanced**, containing instances of each class type that the system is supposed to extract.
 - For example, a system cannot learn to extract corporate entities if there are not enough mentions of corporate entities in the training data.
 - It should be **clean**.
 - It should be **enough**.
 - This ensures accuracy, and is crucial to the creation of a gold standard that will test your system's performance.

WHAT IS DATA ANNOTATION?

- Annotation is the act of adding vital information to raw data.
- To a supervised learning algorithm, data without tags is simply noise.
 - Through annotation, however, this noise can be turned into a focused training manual that has an impact all the way up.

Here's an example of some raw text data that could be used to train the entity extractor:

1 Johnny Depp has confirmed his return to the Wizarding World in the new film, Fantastic Beasts: The Crimes of Grindelwald. Best known for playing Captain Jack Sparrow in Pirates of the Caribbean, Depp will star as the eponymous character, dark wizard Gellert Grindelwald. He joins an ensemble cast, also including Eddie Redmayne and Katherine Waterston, for the latest instalment of the popular fantasy series.

1 Person Johnny Depp has confirmed his return to the Wizarding World in the new film, Product Fantastic Beasts: The Crimes of Grindelwald. Best known for playing Title Captain Person Jack Sparrow in Product Pirates of the Caribbean, Per Depp will star as the eponymous character, dark wizard Person Gellert Grindelwald. He joins an ensemble cast, also including Person Eddie Redmayne and Person Katherine Waterston, for the latest instalment of the popular fantasy series.

A NOTE ON DATA ANNOTATION

- What is the problem? How complex it is?
- Where is the data coming from?
- Who are the annotators?
- What are the tasks assign to them?

Annotation Tools

- DocAnno
- Brat
- Prodigy
- Tagtog
- DataTurks
- Label Studio
- Stanford Text Annotation Tool

Data Annotation Tool Features

	Pros	Cons
Doccano	web-based, open-source , ability to self-host, Simple, easy-to-use interface, shortcuts for faster labeling, Team collaboration , MIT License	Sometime laggy and unresponsive
Brat	Simple to use, Open-source , easy conversion into other output formats, MIT License, Available API for continuous model training.	Needs to be installed locally, Team collaboration not possible, Outdated UI.
Prodigy	Sleek, modern interface, advance active learning , considerably speeding up the annotation process, Self-hosted, Support for image annotation, Fully scriptable, integrated to spacy , team collaboration .	Expensive
Tagtog	Fast tagging - recognizes all occurrences of an entity once it has been manually labeled and tags them automatically, Support for working with multiple types of data, No installation required, Includes active learning , Team collaboration , Own API for continuous model training.	Expensive , The interface can be a little confusing at first.
DataTurks	Allows Text, image and video labeling, open source , Support for a wide variety of text formats, including PDF, Cloud and local installation, Own API for continuous model training.	Proprietary export format may add friction, support appears to have ceased,
Label Studio	open source , multi-type data, Customizable UI, Quick, easy set up, Mobile friendly	Assets for labeling appear unordered, No simple way of returning to an already labeled asset, No out-of-the-box statistics available

Illustration: Text Classification

- **Supervised Machine Learning**

- A document d
- A fixed set of classes $C = \{c_1, c_2, \dots, c_n\}$
- A training set of m hand-labeled documents $(d_1, c_1), \dots, (d_m, c_k)$.
- Output:
 - A learned Classifier function $\varphi: d \rightarrow c$

- **Classifiers:**

- Naïve Bayes
- Logistic regression
- SVM
- K-NN
- NN
- LSTM
- CNN
- BERT
- ...

Sample Labelled/Annotated Data

- I love the script of the movie.
- The story has satirical humor.
- It's the female character was sweet,
- The dialogues are great!
- The adventure scenes are fun
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre.
- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

Sample Labelled/Annotated Data

- I love the script of the movie. [?] +
- The story has satirical humor. [?] +
- It's the female character was sweet, [?] +
- The dialogues are great! [?] +
- The adventure scenes are fun [?] +
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. [?] +
- unbelievably disappointing [?] -
- Full of zany characters and richly applied satire, and some great plot twists [?] +
- this is the greatest screwball comedy ever filmed [?] +
- It was pathetic. The worst part about it was the boxing scenes. [?] -

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

Sample Labelled/Annotated Data

- I love the script of the movie. [?] +
- The story has satirical humor. [?] +
- It's the female character was sweet, [?] +
- The dialogues are great! [?] +
- The adventure scenes are fun [?] +
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. [?] +
- unbelievably disappointing [?] -
- Full of zany characters and richly applied satire, and some great plot twists [?] +
- this is the greatest screwball comedy ever filmed [?] +
- It was pathetic. The worst part about it was the boxing scenes. [?] -

Sample Labelled/Annotated Data

- I **love** the script of the movie. Ⓢ +
- The story has **satirical humor**. Ⓢ +
- It's the female character was **sweet**, Ⓢ +
- The dialogues are **great**! Ⓢ +
- The adventure scenes are **fun** Ⓢ +
- It manages to be whimsical and **romantic** while **laughing** at the conventions of the fairy tale genre. Ⓢ +
- unbelievably disappointing Ⓢ -
- Full of zany characters and richly applied satire, and some great plot twists Ⓢ +
- this is the greatest screwball comedy ever filmed Ⓢ +
- It was pathetic. The worst part about it was the boxing scenes. Ⓢ -

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words


$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{Love} \\ \text{sweet} \\ \text{satirical} \\ \text{laughing} \\ \text{recommend} \\ \text{several} \\ \text{Times} \\ \text{happy} \\ \text{again} \\ \text{humor} \\ \text{great} \\ \text{adventure} \\ \text{Fun} \\ \text{whimsical} \\ \text{romantic} \end{array} \begin{array}{l} 2 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ \dots \\ \dots \\ \dots \end{array} \right) = \mathcal{C}$$

Frequency/ TF-IDF



Crowdsourcing

Annotation through Crowdsourcing

- Crowdsourcing is an emerging collaborative approach for acquiring annotated corpora and a wide range of other linguistic resources
- Three main kinds of crowdsourcing platforms
 - Paid-for marketplaces such as Amazon Mechanical Turk (AMT) and CrowdFlower (CF)
 - Games with a purpose
 - Volunteer-based platforms such as crowdcrafting

Why Crowdsourcing?

- Paid for crowdsourcing can be 33% **cheaper** than in-house employees when applied to tasks such as tagging and classification (Hoffmann, 2009)
- Games with a purpose can be even cheaper in the long run, since the players are not paid.
- However cost of implementing a game can be higher than AMT/CF costs for smaller projects (Poesio et al, 2012)
- Tap into the **large number of contributors**/players available across the globe, through the internet
- **Easy to reach** native speakers in various languages (but beware Google translate cheaters!)

Genre 1: Mechanized Labor

- Participants (workers) paid a small amount of money to complete easy tasks (HIT = Human Intelligence Task)



Paid for Crowdsourcing

- Contributors are extrinsically motivated through economic incentives
- Carry out micro-tasks in return for micro-payments
- Most NLP projects use crowdsourcing marketplaces: Amazon Mechanical Turk and CrowdFlower
- Requesters post Human Intelligence Tasks (HITs) to a large population of micro-workers (Callison-Burch and Dredze, 2010a)
- Snow et al. (2008) collect event and affect annotations, while Lawson et al. (2010) and Finin et al. (2010) annotate special types of texts such as emails and Twitter feeds, respectively.
- Challenges:
 - **low quality output** due to the workers' purely economic motivation
 - **high costs** for large tasks (Parent and Eskenazi, 2011)
 - **ethical** issues (Fort et al., 2011)

Captcha:




Pick your favorite color:




☒ Red
☐ Green




☐ I'm not a robot  reCAPTCHA




Submit

Select all images with commercial lorries

   [Verify](#)

Report a Problem

Genre 2: Games with a purpose (GWAPs)

facebook Friends Applications Inbox Home Search

US08 Sentiment Quiz
Play Rankings Awards Feedback Help About

Is the following a **negative**, **neutral** or **positive** statement about the candidate?

“ We are headed down a path that is certain to end in the destruction of our experiment in democracy. ”

− − − + +

Status

4
Level

15:4 13 12

Your current score is **65 points**. Invite your friends and earn 10% of the points they make!

Spread the Word

Tell your Friends! You will earn **10% of your friends' points** after they accept your invitation! The calculation is recursive, so if they invite others you will even get more bonus points.

September 2008
843 players See All

1.	Fiorella	2458
2.	Michel	224
3.	Birgit	2139
4.	Rose	1011
5.	Herti	930
...		
11.	Arno	101
12.	Guilherme	77
13.	You	65
14.	Lisa	61

Others currently playing

Election Monitor
Vote for your favorite candidate!

Barack Obama	John McCain	Cynthia McKinney

New Media MBA
www.modul.ac.at/nmt/mba

EDITED BOOK
The Geospatial Web
Geobrowsers, Social Software & the Web 2.0

Page built by Sentiment Quiz (report)

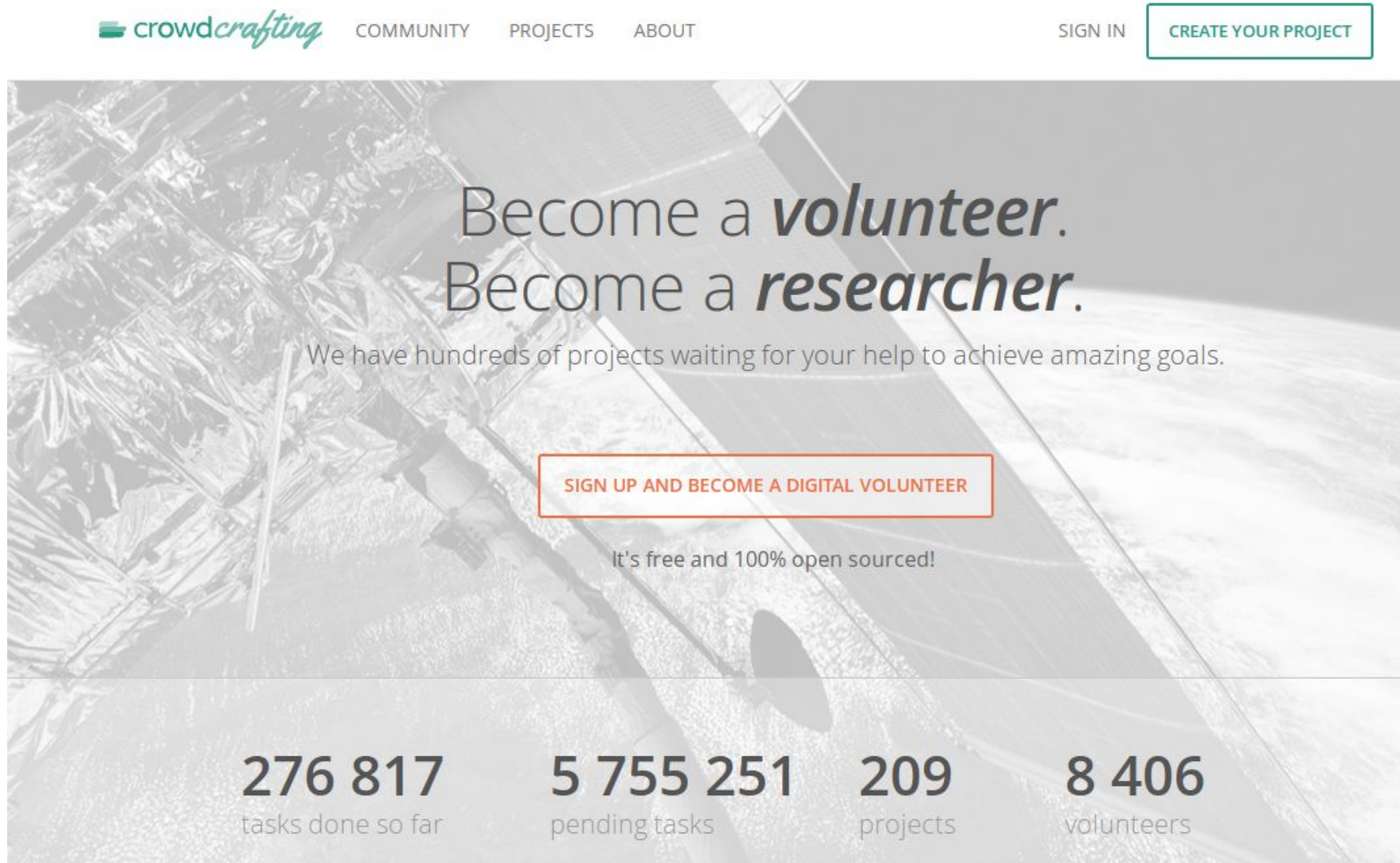
About Find Friends Advertising Developers Terms Privacy Help

?
wordrobe
play what you mean

Ranking (last 50 days)

1		Valerio		32150 points
2		wordrobe		5363 points
3		Aristotle		3998 points
4		sebb		3266 points
5		vincent		3028 points
6		arjanb		2495 points
7		EvaVanmassenhove		1308 points
8		furryfreak		1038 points

Genre 3: Altruistic Crowdsourcing

The image is a screenshot of the crowdcrafting website. The header features the crowdcrafting logo on the left, followed by navigation links for COMMUNITY, PROJECTS, and ABOUT. On the right side of the header, there are links for SIGN IN and a button labeled CREATE YOUR PROJECT. The main content area has a background image of a satellite view of a city. It contains the text 'Become a volunteer.' and 'Become a researcher.' in a large font, followed by the sentence 'We have hundreds of projects waiting for your help to achieve amazing goals.' Below this is a button that says 'SIGN UP AND BECOME A DIGITAL VOLUNTEER'. Underneath the button, it says 'It's free and 100% open sourced!'. At the bottom of the page, there are four statistics: '276 817 tasks done so far', '5 755 251 pending tasks', '209 projects', and '8 406 volunteers'.

COMMUNITY

PROJECTS

ABOUT

SIGN IN

CREATE YOUR PROJECT

Become a ***volunteer.***
Become a ***researcher.***

We have hundreds of projects waiting for your help to achieve amazing goals.

SIGN UP AND BECOME A DIGITAL VOLUNTEER

It's free and 100% open sourced!

276 817

tasks done so far

5 755 251

pending tasks

209

projects

8 406

volunteers

Task Flow

- Data distribution: how “micro” is each microtask?
 - Long paragraphs hard to digest, worker fatigue
 - Single sentences not always appropriate: e.g. for co-ref
- Reward scheme
 - Granularity – per task? Per set of tasks? High scores?
 - What to do with “bad” work
 - How much to reward
 - No clear, repeatable results for quality: reward relation
 - High rewards get it done faster, but not better
 - Pilot task gives timings, so pay at least minimum wage
- Choose the most appropriate genre or mixture of crowdsourcing genres
 - Trade-offs: Cost; Timescale; Worker skills
- Pilot the design, measure performance, try again
 - Simple, clear design important
 - Binary decision tasks get good results

Evaluation and Corpus Delivery

- Evaluate and aggregate contributor inputs to produce final decision
 - Majority vote
 - Discard inputs from low-trusted contributors (e.g. Hsueh et al. (2009))
 - MACE: a) identify which annotators are trustworthy and b) predict the correct underlying labels (Hovy 2013)
- Merge individual units from the microtasks (e.g. sentences) into complete documents, including all crowdsourced markup
- Tune the expert-created “gold” standard based on annotator feedback
 - Gold standard test questions often contain ambiguities and errors
 - Crowd has a broader knowledge-base than a few experts
 - These are a great opportunity to train workers and amend expert data
 - Better gold data means better output quality, for the same cost
- To facilitate reuse, deliver the corpus in a widely used format, such as XCES, CONLL, GATE XML

Example: CF Instructions

Finding location names in text

Instructions ▲

In each sentence below, mark any names that are locations (e.g. **France**). Don't mark locations that don't have their own name.

There may be no locations in the sentence at all - that's OK.

Examples:

"There was a celebration in **London**"
correct - London is a location name

"The **room** is empty"
wrong, because room isn't the name of a particular location

"We traveled to **Spain** and had a great time **there**"
Only mark the location names, not words that just refer to it

"The award went to **Chelsea** Clinton"
wrong, because here Chelsea is a person

Quality of Annotation

Quality of Annotation

Data types:

- Categorical
 - Binary
 - Sentiment +/-
 - Nominal
 - Hepatitis
 - Viral A, B, C,D, E or auto immune
- Continuous
 - Size of tumor
 - Blood Pressure
 - Quality of answers

How Data are repeated?

- Same input
 - Different observers
 - Inter-rater reliability
 - Same observer at different time
 - Intra-rater reliability

Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category.

Quality of Annotation: Some notes

From that measure, you can derive two things:

1. how easy was it to clearly delineate the category: if the annotators make the same decision in almost all cases, then the annotation guidelines, i.e. the definition of the category that needed to be annotated, were very clear, and this implies that it is somehow possible to give the annotator a nicely delineated view on the category.
2. how trustworthy is the annotation: one prefers categories that are firmly delineated — even if that is utopia for linguistics — because they make it easier to perform a quantitative analysis. If the inter-annotator agreement were low, the annotators found it difficult to agree on which items belong to the category, and which didn't. That category might be very very interesting from a qualitative point of view, it is very difficult to incorporate it in a quantitative valorization of the data.
3. The calculation is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone ("expected" agreement). The data layout is shown in Table 1. The observed agreement is simply the percentage of all lectures for which the two residents' evaluations agree, which is the sum of a + d divided by the total n in Table 1. In our example, this is $15+70/100$ or 0.85. We may also want to know how different the observed agreement (0.85) is from the expected agreement (0.65). Kappa is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, ie, potential systematic disagreement between the observers

•

Quality of Annotation : Some notes

- There are basically two ways of calculating inter-annotator agreement. The first approach is nothing more than a percentage of overlapping choices between the annotators.
- This approach is somewhat biased, because it might be sheer luck that there is a high overlap. Indeed, this might be the case if there are only a very limited amount of category levels (only yes versus no, or so), so the chance of having the same annotation is a priori already 1 out of 2. Also, it might be possible that the majority of observations belongs to one of the levels of the category, so that the a priori overlap is already potentially high.

Inter-rater reliability: Cohen's Kappa

		Rater 2	
		Correct	Incorrect
Rater 1	Correct	A	B
	Incorrect	C	D

- **A** => The total number of instances that both raters said were correct. The Raters are in agreement.
- **B** => The total number of instances that Rater 2 said was incorrect, but Rater 1 said were correct. This is a disagreement.
- **C** => The total number of instances that Rater 1 said was incorrect, but Rater 2 said were correct. This is also a disagreement.
- **D** => The total number of instances that both Raters said were incorrect. Raters are in agreement.

Inter-rater reliability: Cohen's Kappa

		Rater 2	
		Correct	Incorrect
Rater 1	Correct	A	B
	Incorrect	C	D

$$K = P_o - P_e / 1 - P_e$$

$$P_o = \text{Number in Agreement} / \text{Total}$$

$$P_{(\text{correct})} = (A + B / A + B + C + D) * (A + C / A + B + C + D)$$

$$P_{(\text{incorrect})} = (C + D / A + B + C + D) * (B + D / A + B + C + D)$$

$$P_e = P_{(\text{correct})} + P_{(\text{incorrect})}$$

p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

$K=1$ □ Full Agreement
 $K=0$ □ Random
 $K<0$ □ No effective agreement