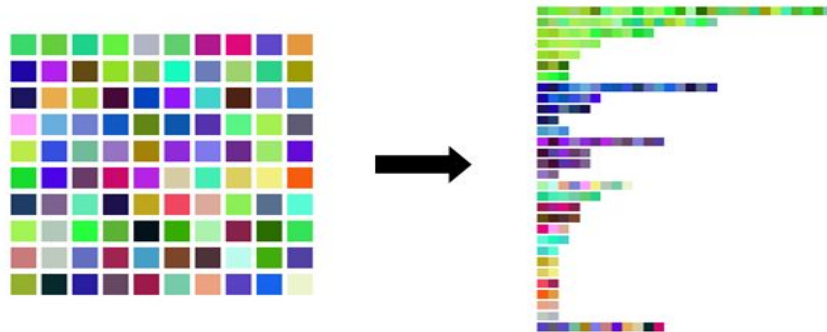


# CASCADE

## Crowdsourcing Taxonomy Creation



# Session Overview

- ❑ The Problem
- ❑ What is Cascade
- ❑ Initial Prototypes
- ❑ Methodology
- ❑ Applications



## *The Problem*



[illegible][illegible]

1. Time consuming
2. Overwhelming

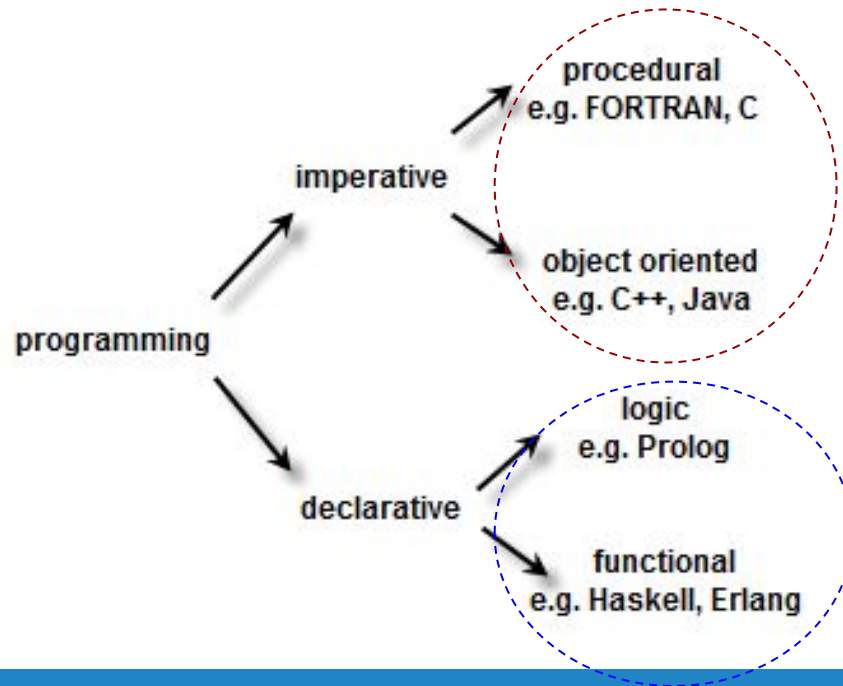


*Solution*



# What is a Taxonomy ?

Hierarchical way of organizing the information

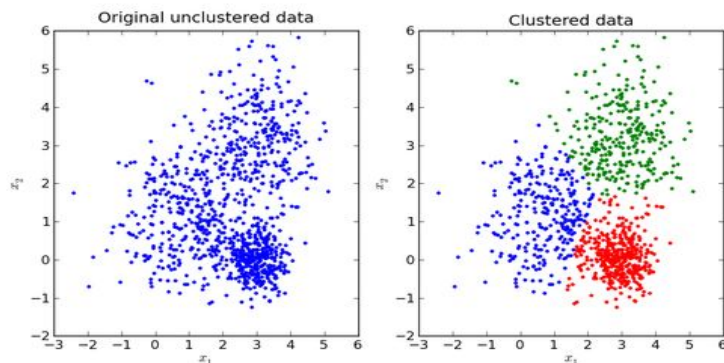




# Taxonomies using AI

## Latent Dirichlet Allocation (Topic Modelling),

- Not sure about the right number of topics
- Produce low quality taxonomies.
  - They lack the common sense and language abilities that come naturally to people



# CASCADE: Crowdsourcing Taxonomy Creation

- A novel crowd algorithm that produces a global understanding of large datasets.
- Cascade is an automated workflow that creates a taxonomy from the collective efforts of crowd workers who spend as little as 20 seconds each.
- None of the workers needs to have a global perspective of the data or the taxonomy under construction.

# Concepts and Evaluation

## 3 HITs

Three HIT primitives, simple task interfaces, used as building blocks by Cascade and useful for future crowd algorithms.

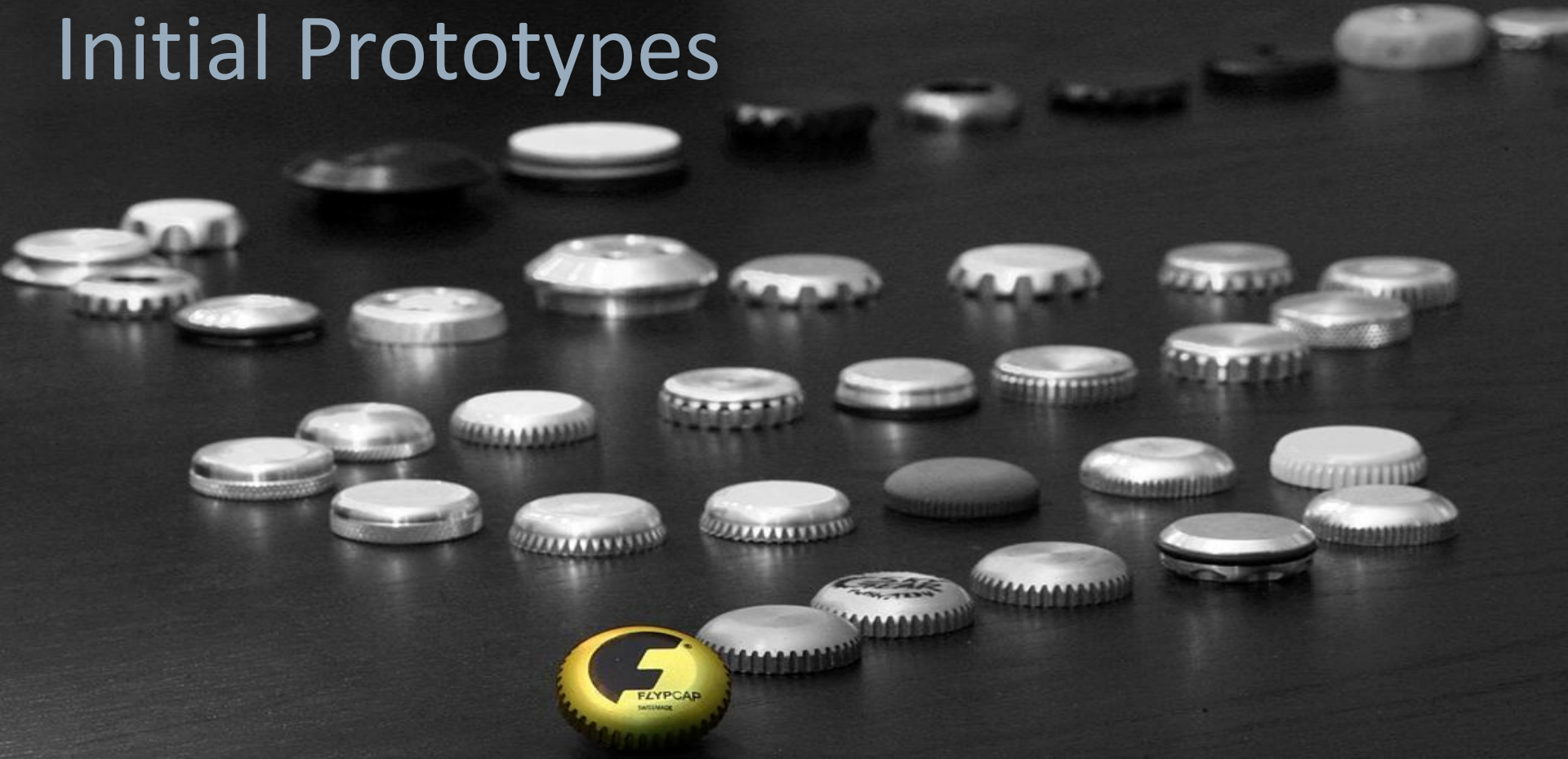
## GSI

We introduce **Global Structure Inference** as a way to combine independently-generated judgments into a cohesive taxonomy.

## Evaluation

We evaluate Cascade on three datasets showing that Cascade can perform close to expert level agreement (80-90% of expert performance) for a competitive cost

# Initial Prototypes



# Initial Prototypes

- Iterative Approach
- Category Comparison
- Clustering

# Initial Prototypes

- Iterative Approach
- Category Comparison
- Clustering

# Approach 1: Iterative Improvement

Task: Add Tips to the Hierarchy of Travel Advice

The screenshot shows a web application interface for categorizing travel tips. On the left, a scrollable list of categories is displayed, including Packing/packing, Clothing, food, flying, carry-on luggage, in flight meals/airline food, airport check in/Check-In, On Board Entertainment, flying on a budget, customs/Immigration, luggage, Communication/communication, long distance calls, Insurance, Security Control/security, Accommodation/lodging, Hostels/the benefits of hostels, Budget Travel/Thrifty Travel Tips, flying on a budget, personal valuables, San Francisco International, tours, and Travel Etiquette. The main area is titled 'Tips that still need categorization' and features navigation buttons for 'Previous 4' and 'Next 4'. It displays a list of tips, each with a number and a description. The tips shown are #100, #101, #102, and #103. A 'Submit' button is located at the bottom left of the interface.

- \* Packing/packing
- \*\* Clothing
- \* food
- \* flying
- \*\* carry-on luggage
- \*\* in flight meals/airline food
- \*\* airport check in/Check-In
- \*\* On Board Entertainment
- \*\* flying on a budget
- \*\* customs/Immigration
- \* luggage
- \* Communication/communication
- \*\* long distance calls
- \* Insurance
- \* Security Control/security
- \* Accommodation/lodging
- \*\* Hostels/the benefits of hostels
- \* Budget Travel/Thrifty Travel Tips
- \*\* flying on a budget
- \* personal valuables
- \* San Francisco International
- \* tours
- \* Travel Etiquette

Previous 4

Tips that still need categorization

Next 4

#100  
"-Luggage with a lifetime guarantee is worth the slight p  
and Riley make a very sturdy bag that's strong enough y  
long pre-boarding wait, and with zippers that rarely brea  
or 10 year"

#101  
"-If you're tall or otherwise picky about airplane seats, u  
understand the seat layout of your flight. Seatguru will v  
boxes under the seat in front of you, cold seats, or seats  
traffic."

#102  
"-From my wife, I learned to \*always\* ask for a better p  
hotel checkin. We stayed 10 nights in a \$2400/night hote  
infinity-edged swimming pool at Jade Mountain in St. L  
the website) fo"

#103  
"-For overnight flights, don't take the sleeping pill until  
the ground. I once had an 11pm redeye with a post-board

Submit

## Problems

- 1.The hierarchy grows and becomes overwhelming
- 2.Workers have to decide what to do

## Lesson

Break up the task more

# Initial Prototypes

- Iterative Approach
- **Category Comparison**
- Clustering



# Approach 2: Category Comparison

Tag #1		Tag #2
"airport security" is:	<input type="radio"/> the same as <input type="radio"/> more general than <input type="radio"/> more specific than <input type="radio"/> other	"Security"
"airport security" is:	<input type="radio"/> the same as <input type="radio"/> more general than <input type="radio"/> more specific than <input type="radio"/> other	"Airport security information"
"flying" is:	<input type="radio"/> the same as <input type="radio"/> more general than <input type="radio"/> more specific than <input type="radio"/> other	"Flights"
"Saving money" is:	<input type="radio"/> the same as <input type="radio"/> more general than <input type="radio"/> more specific than <input type="radio"/> other	"Tips to Save Money"
"Saving money" is:	<input type="radio"/> the same as <input type="radio"/> more general than <input type="radio"/> more specific than <input type="radio"/> other	"SAVINGS"

airline travel	booking airline seats	Seat assignments	Aisle seats	Window seats
Air travel ticket booking	Canceled Flights	Customer Service - canceled flights	flying	airports
tr				

Same Co

## Problem

Without context, it's hard to judge relationships

- **flying** vs. **flights**
- **Packing** vs. **what to bring**

## Lesson

Don't compare abstractions to abstractions

Instead compare data to abstractions

# Initial Prototypes

- Iterative Approach
- Category Comparison
- Clustering

# Approach 3: Clustering

## Help Categorize Travel Tips

### Instructions:

In the table below there are 10 tips for travelers. We want to organize them into categories. Your job is to:

- Read all 10 tips
- Think of 3 categories that at least two of these tips belong to. Write them in the colored boxes as column
- In each column, select **at least two** tips that fit in that category. The more you can select the better. Be
- Mark tips that you think are difficult to categorize as "Singleton Tip." If there aren't any, that's okay.
- Mark tips that don't make sense as "Needs Review." If there aren't any, that's okay.

Travel Tip				
"For packing the trick is BIT: buy it there. Pack the minimum you think you'll need and if you forget something, buy it there. Often I don't end up buying anything, but making this a part of my trip planning helps me relax and pack light."	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
"-Passport, wallet, housekey, phone & charger. That's my checklist when I leave the house on the way to a flight. Anything else is a non vital item I figure I can take care of when I get there. You could buy a new phone charger there but this is such an of"	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
"-They're popular among frequent flyers, but I avoid the Bose noise-canceling headphones because they're too big (and the travel case makes them even bigger). You can get a pair of in-ear noise-isolating headphones that are just as good, half the price, an"	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### Problem

Users often gave categories that fits multiple of the 8 items displayed.

### Lesson

Allow workers to suggest categories that fit one item

# The Right Approach




- *Find the tasks people can do.*
- *Assemble them using complex aggregation techniques.*

# Workflow- PHASE I

## 3 Human Intelligence Tasks (HIT)


What category do you suggest for this color?



Greenish

Generate Labels


What is the best category for this color?



<u>Category</u>	<u>Best?</u>
Aqua	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>
Lime	<input type="checkbox"/>
Pastel	<input type="checkbox"/>

Select Best Labels

What categories does this belong to?



<u>Category</u>	<u>Fits</u>	<u>Doesn't Fit</u>
Green	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Yellow	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Pink	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Categorize

# Workflow: PHASE II

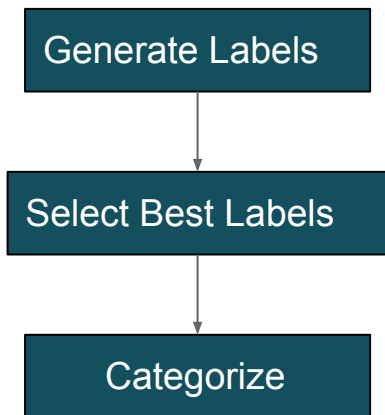
## Global Structure Inference

### **3 Step Process**

- 1) Remove Insignificant Categories
- 2) Remove Duplicate Categories
- 3) Create Nested Categories

# CASCADE Pipeline

## CATEGORY GENERATION



## TAXONOMY GENERATION

(Global Structure Inference)

- **traveling (100)**
  - travel organization and convenience (68)
    - preparing for long flights (15)
      - health tips (5)
        - drinking on flights (3)
  - electronics (14)
    - iphone and ipod (9)
    - airline wifi (3)
  - boarding process (11)
  - airport security (10)
  - entertainment (9)
  - airport transportation (6)
  - laptop (4)
  - laptop power supply (4)
  - website (4)
  - international phone usage (3)
    - international data plans (2)
- insider tips (49)
  - making friends with locals (6)
  - airport amenities (4)
- air travel tips (49)
  - preparation for flying (38)
    - comfortable flying (13)
  - airport tips (26)
    - airport shortcuts (17)
- flight (25)

## REPEAT

Re-run the process on all the sets



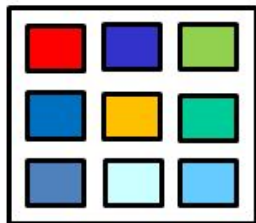
# Cascade Algorithm






# Cascade Algorithm

For a subset of items



Generate Labels


What category do you suggest for this color?



Greenish

Select Best Labels

What is the best category for this color?



Category	Best?
Aqua	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>
Lime	<input type="checkbox"/>
Pastel	<input type="checkbox"/>




{good labels}

Blue  
Light Blue  
Green  
Greenish  
Red  
Gold

Categorize

For all items,  
for all good labels,

What categories does this belong to?

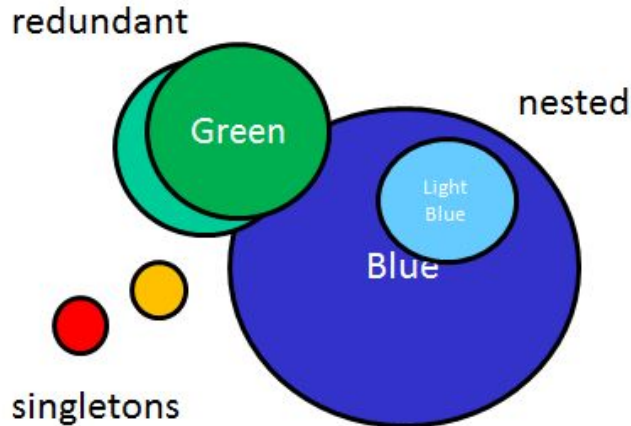
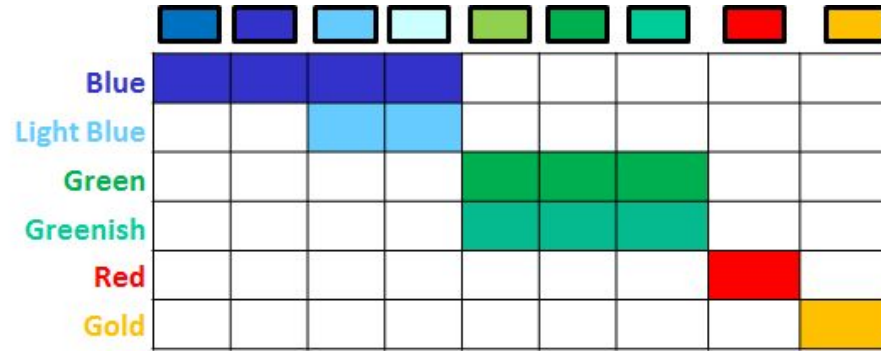


Category	Fits	Doesn't Fit
Green	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Yellow	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Pink	<input type="checkbox"/>	<input checked="" type="checkbox"/>



Recursive approach

# Aggregate Data into Taxonomy



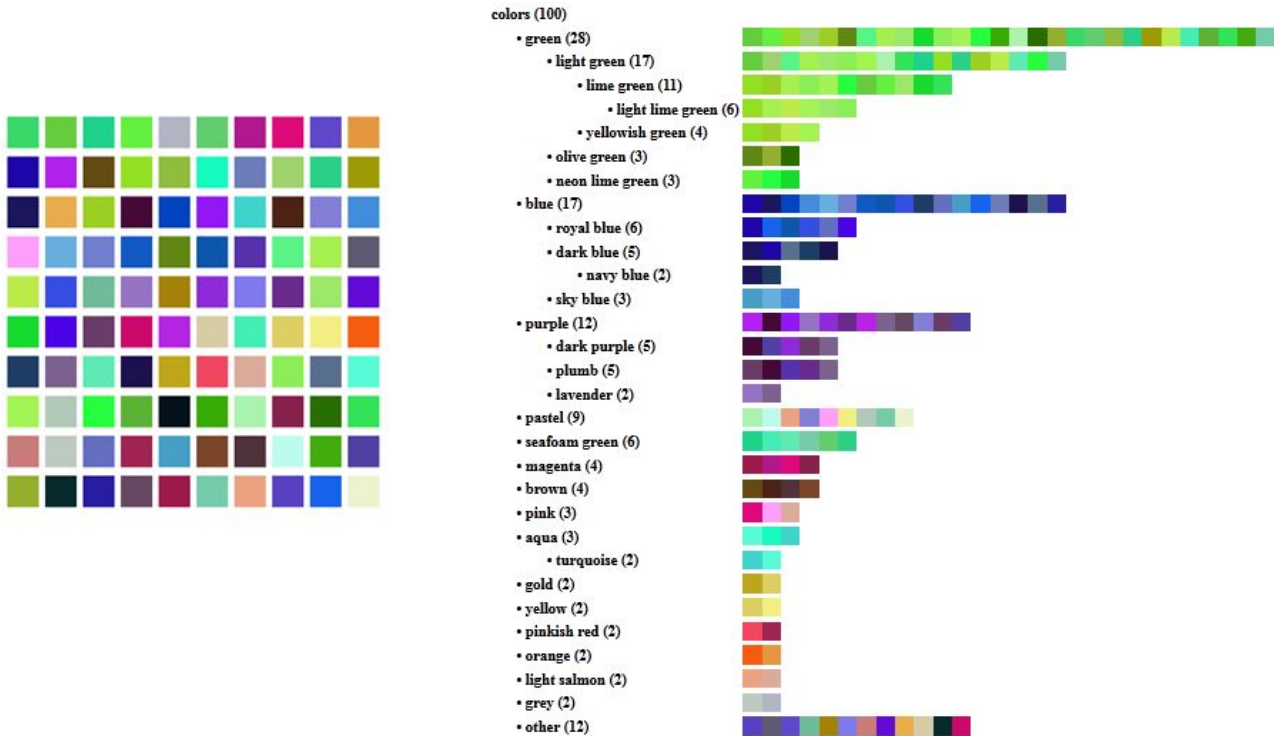
Blue:

Light Blue:

Green:

Other:

# Cascade Results: 100 Colors



# Propose, Vote, Test

1. Let the workers propose categories.
2. Vote on categories to weed out bad ones.
3. Test the heuristics by verifying it on data.

## Propose

What category do you suggest for this color?



(Generate)

## Vote


What is the best category for this color?



Category	Best?
Aqua	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>
Lime	<input type="checkbox"/>
Pastel	<input type="checkbox"/>

(Select Best)

## Test



Greenish

Category	Fits	Doesn't Fit
Green	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Greenish	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Yellow	<input type="checkbox"/>	<input type="checkbox"/>
Pink	<input type="checkbox"/>	<input type="checkbox"/>

(Categorize)

# KEY TAKEAWAY



Propose

(Generate)

Vote

(Select Best)

Test

(Categorize)

# EVALUATION: DataSet

3 Dataset obtained from [Quora.com](https://www.quora.com)

Abbreviation	Topic	# items
editWriting	“What are some tips for editing your own writing?”	22
sideProjects	“How can I increase my productivity on my side projects at the end of the day when I’m tired from work?”	67
travel	“What are your best travel hacks?”	100

# EVALUATION: Metrics

1. Labels Quality
2. Hierarchical structure
3. Cost and Running time

# EVALUATION: Metrics

1. **Labels Quality**
2. Hierarchical structure
3. Cost and Running time



# EVALUATION: Accuracy of Labels

1. What fraction of the Cascade taxonomy categories are also named in at least one expert taxonomy?
2. What fraction of expert categories are named in another expert taxonomy?

	edit- Writing	side- Projects	travel	Avg.
% of Cascade categories shared by at least one expert	47%	50%	53%	<b>50%</b>
avg % of expert categories shared by at least one other expert	32%	70%	64%	<b>55%</b>

# EVALUATION: Metrics

1. Labels Quality
2. **Hierarchical structure**
3. Cost and Running time

# EVALUATION: Appropriate Hierarchical Structure

1. Duplicate categories
2. Missing Parent-Child Relationships
3. Incorrect Parent-Child Relationships

Error Rate:  $\frac{\text{Total Errors}}{\text{Total Categories}}$

	Edit-Writing	Side Projects	Travel: iteration1	Travel: iteration2
# categories	15	18	7	51
Duplicate Categories	2	2	0	2
Missing Nesting	0	0	0	5
incorrect Nesting	0	3	1	3
Correct Nesting	5	3	1	23
total errors	2	5	1	10
<b>Error rate</b>	<b>13%</b>	<b>27%</b>	<b>14%</b>	<b>20%</b>

# EVALUATION: Metrics

1. Labels Quality
2. Hierarchical structure
3. **Cost and Running time**

# EVALUATION: Cost and Running Time

## Cascade versus experts

- ▷ The total time spent on all three datasets by the average expert was 6 hours 50 minutes, and the total time spent by MTurk workers was 43 hours 3 minutes.
- ▷ Cascade amounted to \$ 369 while average experts costed \$ 171

**~Cascade took ~6.5 times longer to complete, and was 2-3 times as expensive**

# EVALUATION: Cost and Running Time

## Contd.

	editWriting	sideProjects	travel
Cascade Time	7 h 56 m	16h 13 m	16h 32m
Avg expert time	1h 23 m	2h 36m	2h 5 m
Cascade Cost	\$35.40	\$109.45	\$224.45
Average Expert Cost	\$34.87	\$65.13	\$71.38

Since the work was replicated 5 times

\$8.39/hour was paid to workers (\$0.05 per HIT and HIT was 21.46 seconds), which is high for MTurk, where \$3-\$4/hour is more typical

# Future work

- ❏ Reduce the cost & Time
- ❏ Adopt machine learning approaches

# Applying Cascade to REAL WORLD Scenarios

## 1. CHI 2013

(Top conference for Human-Computer Interaction)

- Organize 430 accepted papers to help session making

## 2. CrowdCamp Hack-a-thon

- Organize 100 hack-a-thon ideas to help organize teams



# Application: CHI Papers

Patina: Dynamic Heatmaps for Visualizing Application Usage',  
Effects of Visualization and Note-Taking on Sensemaking and Analysis',  
Contextifier: Automatic Generation of Messaged Visualizations',  
Interactive Horizon Graphs: Improving the Compact Visualization of Mu  
Quantity Estimation in Visualizations of Tagged Text',  
Motif Simplification: Improving Network Visualization Readability with  
Evaluation of Alternative Glyph Designs for Time Series Data in a Small  
Individual User Characteristics and Information Visualization: Connecting  
"Without the Clutter of Unimportant Words": Descriptive Keyphrases fo  
Direct Space-Time Trajectory Control for Visual Media Editing  
Your eyes will go out of the face: Adaptation for virtual eyes in video se  
Swifter: Improved Online Video Scrubbing  
Direct Manipulation Video Navigation in 3D  
NoteVideo: Facilitating Navigation of Blackboard-style Lecture Videos  
Ownership and Control of Point of View in Remote Assistance  
EyeContext: Recognition of High-level Contextual Cues from Human Vis  
Your eyes will go out of the face: Adaptation for virtual eyes in video se  
Still Looking: Investigating Seamless Gaze-supported Selection, Positioning, and Manipulation of Distant Targets  
Individual User Characteristics and Information Visualization: Connecting the Dots through Eye Tracking  
Quantity Estimation in Visualizations of Tagged Text

- **Visualization (19)**
  - **evaluating infovis (9)**
    - **text (2)**
  - **video (6)**
  - **visualizing time data (5)**
  - **gaze (4)**
    - **gaze tracking (3)**
  - **user requirements (3)**
  - **color schemes (2)**

Two famous papers for Taxonomy  
creation

**Cascade and Deluge**

# Summary

## → What is Cascade

- ◆ A Crowd-algorithm that generates a taxonomy over a set of independent data items, such as travel items, color blocks etc.

## → Methodology

- ◆ Propose, Vote, Test, Global Structure Inference

## → Applications

- ◆ Q&A sites (Quora), Online Competitions (CrowdCamp)