DESIGN PROBLEMS IN CROWDSOURCING: IMPROVING THE QUALITY OF CROWD-BASED DATA COLLECTION

BY

PIOTR ORGANISCIAK

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Michael Twidale, Chair
Associate Professor Miles Efron
Professor J. Stephen Downie
Jaime Teevan, Microsoft Research

Abstract

Text, images, video and other types of information objects can be described in many ways. Having detailed metadata and a various people's interpretations of the object helps in providing better access and use. While collecting novel descriptions is challenging, crowdsourcing is presenting new opportunities to do so.

Large-scale human contributions open the door to latent information, subjective judgments, and other encoding of data that is otherwise difficult to infer algorithmically. However, such contributions are also subject to variance from the inconsistencies of human interpretation.

This dissertation studies the problem of variance in crowdsourcing and investigates how it can be controlled both through post-collection modeling and better collection-time design decisions.

Crowd-contributed data is affected by many inconsistencies that differ from automated processes: differences in attention, interpretation, skill, and engagement. The types of tasks that we require of humans are also more inherently abstract and more difficult to agree on. Particularly, qualitative or judgment-based tasks may be subjective, affected by contributer opinions and tastes.

Approaches to understanding contribution variance and improve data quality are studied in three spaces.

First, post-collection modeling is pursued as a way of improving crowdsourced data quality, looking at whether factors including time, experience, and agreement with others provide indicators of contributions quality. Secondly, collection-time design problems are studied, comparing design manipulations for a controlled set of tasks. Since crowdsourcing is borne out of an interaction, not all crowdsourcing data corrections are posterior: it also matters how you collect that data. Finally, designing for subjective contexts is studied. Crowds are well-positioned to teach us about how information can be adapted to different person-specific needs, but treating subjective tasks similarly to other tasks results in unnecessary error.

The primary contribution of this work is an understanding of crowd data quality improvements from non-adversarial perspectives: that is, focusing on sources of variance or errors beyond poor contributors. This includes findings that:

1. Collection interface design has a vital influence on the quality of collected data, and better guiding contributors can improve crowdsourced contribution quality without greatly raising the cost of collection nor impeding other quality control strategies.
2. Different interpretations of instructions threaten reliability and accuracy in crowdsourcing. This source of problems even affects trustworthy, attentive contributors. However, contributor quality can be inferred very early in an interaction for possible interventions.
3. Certain design choices improve the quality of contributions in tasks that call for them. Anchoring reduces contributor-specific error, training affirms or corrects contributors' understanding of the task, and performance feedback can motivate middling contributors to exercise more care. Particularly notable due to its simplicity, an intervention that forefronts instructions behind an explicitly dismissable window improves contribution quality greatly.
4. Paid crowdsourcing, often used for tasks with an assumed ground truth, can be also be applied in subjective contexts. It is promising for on-demand personalization contexts, such as recommendation without prior data for training.
5. Two approaches are found to improve the quality of tasks for subjective crowdsourcing. Matching contributors to a target person based on similarity is good for long-term interactions or for bootstrapping multi-target systems. Alternately, explicitly asking contributors to make sense of a target person and customize work for them is especially good for tasks with broad decision spaces and is more enjoyable to perform.

The findings in this dissertation contribute to the crowdsourcing research space as well as providing practical improvements to crowd collection best practices.

Contents

Introduction

The internet is growing increasingly interactive as it matures. Rather than merely transmitting information to readers, web pages allow their audience to react and interact with their information. The products of these interactions are a trove of qualitative judgments, valuable to modeling information objects. In recent years, this form of creation-through-collaboration has been studied as *crowdsourcing*.

This work in a sentence: *Maximizing data quality in using paid crowds for objective and subjective encoding tasks, leveraging post-collection and collection-time strategies.*

There are many circumstances where access to human encoding and human judgments is invaluable to information science, whether it is in transcribing scanned material, organizing or judging the quality of documents within a collection, building evaluation datasets for information retrieval, or preparing training data for better inferential algorithms. People can provide latent information about documents that would not be possible to ascertain computationally, such as quality judgments or higher-level thematic description. They are also adept at critical actions such as correcting, describing in different language, or inferring relationships with other documents. Most importantly, crowdsourcing looks at human contribution at scales that are difficult to attain in alternate ways.

However, humans have predictable and unpredictable biases that make it difficult to systematically adopt their contributions in an information system. *How do we control and interpret qualitative user contributions in a quantified system?*

This work focuses on understanding the characteristics of data

collected through crowdsourcing, toward two ends: awareness of potentially unanticipated biases in crowd data collection, and subsequent strategies to improve the quality of crowd-collected data. As will be demonstrated, crowdsourcing research is sensitive to various circumstances of instrumentation, context, and community. This work seeks to understand the intricacies of these biases: looking at how tasks are completed when they are more or less engaging, restrictive, or subjective. Valid research needs to be aware of how circumstance affects crowds, as well as know what information is important to report for reproducibility. Understanding leads to practical recommendations for maximizing data quality in crowdsourced data, and this study focuses on a priori instrumentation choices and posterior data normalization for improving how both subjective and objective tasks are collected.

This work is scoped to a particular type of crowd production – metadata about existing information objects – and a particular form of collection: microtasks in paid crowd platforms. These are viewed in the space of subjective and objective types of tasks.

It is important to stay aware of the broader space of crowdsourcing and how characteristics of paid microtasks generalize to it. However, as the chapter *A Typology for Crowdsourcing* makes clear, crowdsourcing is a broad expanse; the treatment here is controlled to a subset pertinent to information science research.

In the interest of not obscuring the details, the specifics of this scoping will be presented after first drawing out the problem and this dissertation's approach to tackling it.

The goal of this work is to leave the reader with an understanding of how online crowds can reliably generate metadata about information system objects, both for subjective or objective ends. The main contribution of this study is *methodological*: understanding issues related to proper – or improper – crowdsourcing in information sciences. It is written in the service of uncovering issues and answering them thoroughly, where a reader may develop realistic expectations or hypotheses for tasks beyond the

tasks used for this study's experiments.

A reader of this work will understand: the issues related to using crowdsourcing contributions for improving document metadata, particularly for information retrieval indexing and evaluation, and user-based filtering or recommendation; the effect of different designs of crowdsourcing collection tasks on the resulting reliability and consistency of the collected data, particularly designs that train workers, give them feedback, or hurry them; sources of contributor-specific error in information retrieval evaluation tasks; and how these findings may assist future working in information science and cultural heritage spaces.

Among the most valuable or promising outcomes, this study includes the findings that:

- Collection interface design is a vital influence on the quality of collected data, and strategies to better guide workers can improve crowdsourced contribution quality without greatly raising the cost of collection nor impeding other quality control strategies.
- Varying interpretations of instructions are an important threat to reliability and accuracy in crowdsourcing, a source of problems that even affects trustworthy, attentive workers.
- The accuracy of a worker on the first task in a task set is a significant indicator of their future performance, which can be used to intervene early on expected poor workers.
- Interventions such as anchoring, training, and performance feedback improve the quality of contributions. Anchoring reduces user-specific bias in scaled forms by tying the interface to more explicit benchmarks. Training helps affirm or correct workers' understanding of the task, particularly in cases where the task stays constant throughout multiple interactions. Performance feedback presents to workers an estimate of their performance, effective for less abstract tasks, except for the absolute worst workers.
- An intervention that forefronts instructions behind an explicitly dismissable window improves contribution quality great

in the studied context. This finding is promising for future work because of the simplicity of the design change.

- Paid crowdsourcing, often used for tasks with an assumed ground truth, can be also be applied in subjective contexts. This is particularly promising for on-demand personalization contexts, such as recommendation without prior data to train on.
- Taste-matching and taste-grokking, introduced as two approaches to crowdsourcing subjective information, are both found to be promising, with strengths in different areas. Matching, where crowd workers are matched to the target person[1] based on their similarity, is good for long-term interactions or for bootstrapping multi-target systems. Grokking, where crowd workers make sense of the target person and customize their contributions based on an intuited understanding of the target, is especially good for tasks with broad decision spaces and is more enjoyable to perform.

[1] In subjective contexts, the concept of a 'good' contribution depends on the person or situation calling for the contribution. The person being tailored for is referred to as the target.

Problem

The growth of digital collections has outpaced the ability to comprehensively clean, transcribe, and annotate them. Similar roadblocks are affecting born-digital information, where the rapid creation of documents often follows from passive or unrestricted forms of production. The lack of strong descriptive metadata poses an obstacle for information retrieval, which must infer the aboutness of a document in order to surface it for an interested user. Crowdsourcing is increasingly being used to address this problem.

Crowdsourcing is the distributed, large-scale collaboration of users contributing to a common product. It describes the act[2] of a system opening up for contributions from distributed users. Users do not necessarily collaborate directly with each other – though they can – so the crowd in the term refers broadly to the collective users of the system.

[2] Crowdsourcing as a verb is a technical point worth noting. Sourcing describes the act of soliciting user contribution, regardless of whether it is successfully executed or not.

It is an umbrella term preceded by a number of more narrowly

scoped concepts, such as *commons-based peer production* (Benkler 2006), *free and open source* software development (Raymond 1999; Lakhani and E. v. Hippel 2003), and *human computation* (Ahn 2006; Law and Ahn 2011). Surowiecki discussed aggregate crowd intelligence as the *wisdom of the crowds* (2004); one way to interpret crowdsourcing is as the process of trying to utilize that wisdom.

Many of the benefits of crowdsourcing follow from the fact that humans approach tasks in qualitative and abstract ways that are difficult to emulate algorithmically. A human can respond to complex questions on a Q&A website, judge the quality of a restaurant/product/film, or decipher a sloppy piece of handwriting.

Since many information systems are intended to serve an information-seeking user, the information that crowdsourcing collects can better reflect the needs of users. For example, a user-tagged image in a museum collection can fill in terms that are more colloquial than the formal vocabulary employed by a cataloguer (Springer et al. 2008; Trant and Wyman 2006). Such information is invaluable in indexing items for information retrieval, where the goal is commonly to infer what a user is searching from their textual attempt to describe it in a query.

Similarly, other uses of crowdsourcing capitalize on humans' abilities to spot when algorithmic attempts at understanding an information object have failed. ReCaptcha uses human contributions to transcribe transcriptions of OCR problem text from Google Books and the New York Times (*What is reCAPTCHA?* 2008) The National Library of Australia's Trove also crowdsources corrections of scanned text, by allowing readers of their scanned newspapers to edit transcribed text when they come across problems (Holley 2009).

Humans are also being used to encode parsable text descriptions for non-text materials or higher-level latent concepts. In libraries, this approach is being adopted with crowd transcription of materials which are too difficult for computer vision, such as digitized letters. For example, the Bentham Project at Univer-

sity College London has a pilot project for crowdsourcing the transcription of Jeremy Bentham's letters (Moyle, J. Tonra, and V. Wallace 2010; Causer, Justin Tonra, and Valerie Wallace 2012).

More than simply describing works, addition useful information can be in people's reactions or critical interpretations. Indexing human judgments of a document's quality, for example, can enable an information retrieval system to rank the best version of multiple similar documents.

While the complex qualitative actions of human contributions are the cornerstone of such contributions' usefulness, they present a challenge for algorithmic use because they can be highly variable.

A task becomes more open to interpretation the more complex it becomes. Some projects revel in the broad interpretive nature of complex tasks, like collaborative art projects reimagining movies (Star Wars Uncut) or music videos (Johnny Cash Project) through a hodgepodge of styles, or coding challenges (TopCoder) that benefit from alternative approaches to a problem.

However, in cases where there is a goal to find either an objective truth, manifest or latent, or to gauge the subjective approaches and opinions of people comparably, the breadth of interpretations possible for a task presents a problem for reliably understanding it in aggregate.

The variability seen in human interpretations of complex tasks is not a novel issue. It is a problem that we call low *intercoder reliability*[3], and can result from a variety of issues. Four 'threats to reliability' that Neuendorf (2002) lists echo issues in crowdsourcing document description: an insufficient coding scheme, inadequate training, fatigue, and problem coders.

[3] "the extent to which two or more independent coders agree on the coding of the content of interest with an application of the same coding scheme" (Cho 2008).

Seeking to account for the variability of human contributions in leveraging online crowds, this study looks to understand the threats to reliability in three spaces:

- error introduced by contributors (e.g., problem coders),
- error owing to the external factors (e.g., an insufficient coding

scheme, inadequate training),
- and error owing to the task (i.e., subjective tasks treated objectively).

Overview

What are the properties of data collected from crowds for objective and subjective information system tasks, and how can the quality of data – in terms of consistency and variance – be optimized?

Each research chapter turns the lens on a piece of this question. The broad research questions informing the chapters are as follows:

- **Broad Research Question 1 (RQ1)**: What are the *post-collection* indicators of quality in worker-contributed objective task data, and can these be leveraged for improved data modeling? [4]
- **Broad RQ 2**: What are the biases inherent to the task design for *objective* tasks (i.e., the data collection instrument), and can design manipulations correct for them at *collection time*? [5]
- **Broad RQ 3**: What are the quality losses when treating *subjective* tasks in objective ways, and can *collection-time* framing or *post-collection* modeling approaches reduce these? [6]

[4] Reported in *Interpreting Tasks for Objective Needs*, with an additional approach reported in the second part of *Designing Tasks for Objective Needs*.

[5] Reported in *Designing Tasks for Objective Needs*.

[6] Reported in *Designing Tasks for Subjective Needs*.

Though in each chapter there are concrete solutions proposed and evaluated, the first step of each research question is to understand the scope of the problem. Regardless of implementation, this dissertation's *pertinent and valuable contribution* is in understanding some ways that crowdsourced data may have unexpected and perhaps overlooked variance, bias, and low-consistency.

Before conducting our own experiments, the next two chapters present an in-depth look at crowdsourcing.

Introduction to Crowdsourcing presents a brief overview of crowdsourcing. One can consider this chapter the seed of what might be taught in the first two weeks of a course on crowdsourcing.

A Typology of Crowdsourcing then presents an information-science typology of crowdsourcing, a necessity for appreciating the expansive area of crowdsourcing and this study's particular scoping. Both of these chapters are general, and literature reviews pertinent to the experiments in this study are reported in the relevant research chapters.

A post-collection lens is applied to crowdsourcing error in *Interpreting Tasks for Objective Needs*, looking to identify and promote high-quality contributions from strong contributors, while adjusting for poor work.

The chapter is largely analytical, hoping to understand what we can infer from crowd behaviors about the strength of their contributions and evaluating strategies for better paid crowdsourcing. By taking this approach, this chapter seeds some the expected outcomes driving later chapters.

The questions this chapter asks are the following:

- **RQ 1.1**: Does the length of time that a worker spends on a question reflect the quality of their rating?
- **RQ 1.2**: Do worker contributions improve predictably with experience?
- **RQ 1.3**: Does a worker's agreement or disagreement with other workers reflect their overall quality as a worker?
- **RQ 1.4**: If so, can disagreement be used for data improvements?

In many circumstances, contributions are not simply a hallowed set of data bestowed upon a researcher or practitioner to work with. Rather, contributions are collected, and as such the *way they are collected* can change what they look like at the end. The next chapter turns our attention toward this less-explored corollary of post-collection data modeling: the effect of the collection instrument on the resultant contributions, toward understanding and potentially optimizing the contribution collection process. This is about how you ask, and how it affects what you are told.

Designing Tasks for Objective Needs is presented through two studies.

The first part selects two control tasks, and measures the effect of three different design manipulations on the makeup of the data – consistency and quality, but also contribution patterns. Looking at interfaces that give users training, performance feedback, or timer-driven nudges, it asks:

- **RQ 2.1**: Which approaches to collection interface design are worth pursuing as alternatives to the basic designs commonly employed in crowdsourcing?
- **RQ 2.2**: Is there a significant difference in the quality, reliability, and consistency of crowd contributions for the same task collected through different collection interfaces?
- **RQ 2.3**: Is there a qualitative difference in contributor satisfaction across different interfaces for the same task?

The second part of Designing Tasks for Objective Needs bridges the studied strategies in a real world setting, applying post-collection corrections as well as collection-time task manipulations to the human judgments used in evaluating audio similarity for the Music Information Retrieval Exchange (MIREX). Finding the intercoder consistency to be very low, this small chapters asks:

- **RQ 2.4**: Are coder differences responsible for low intercoder consistency in MIREX judgments?
- **RQ 2.5**: Are problem coders responsible for low intercoder consistency?
- **RQ 2.6**: Is subjectivity or disagreement of the grading task responsible for low intercoder consistency?
- **RQ 2.7**: Does the task design affect the quality of contributions?

Moving beyond objective contexts, *Designing Tasks for Subjective Needs* again focuses on maximizing quality through a priori design and instrumentation choices, but for a different class of

task. Subjective tasks are rarely done in paid contexts, so *personalized crowdsourcing* is introduced as a way to formalize and argue for the approach. Two protocols for personalized crowdsourcing are then presented, referred to as *taste-matching*[7] and *taste-grokking*[8], and compared.

- **RQ 3.1**: Is it feasible to apply paid crowdsourcing to subjective problems?
- **RQ 3.2**: Does the taste-matching protocol reduce the amount of error in personalized crowdsourcing?
- **RQ 3.3**: Does the taste-grokking protocol reduce the amount of error in personalized crowdsourcing?
- **RQ 3.4**: How do different types of subjective tasks affect the efficacy of personalized crowdsourcing approaches?

[7] *Taste-matching* seeks to find crowd workers that are similar to a target person, using their future work as a proxy for the target person's opinions or style.
[8] *Taste-grokking* focuses on communicating a target person's opinions or style to workers, who then perform work specific to their impression of that person.

Scope

As mentioned at the outset, this work focuses on a particular, but pertinent, corner of crowdsourcing. The form of crowd production studied is metadata about existing work, and the type of collection is microtasks in paid crowd platforms. Both subjective and objective types of tasks are considered, however, given that they manifest very distinctly. Let's consider each of these parts in order.

Metadata about existing work: an important albeit rarely formalized distinction in crowdsourcing contributions is whether the crowd *creates* new intellectual works, or whether they *react* to existing information objections. Generally, the uses of crowdsourcing of interest to information scientists, librarians, and information retrieval researchers are in the latter category.

Paid crowdsourcing: Paid crowdsourcing platforms are markets for on-demand online labour. They reduce much of the overhead seen in volunteer crowdsourcing related to attracting and motivating users, replacing intrinsic motivation with financial incentive. The most popular paid platform is Amazon's Mechanical Turk, which also happens to be one of the first such platforms and the one used to run experiments in this study.

Microtasks: Microtasking refers to the common practice of breaking tasks down to small practical units, which both simplifies the task distribution process in a modularized style and accommodates the short interaction style that is common online. For example, consider a task where you are transcribing and annotating the themes in scanned correspondence: rather than asking workers to do everything in one task, there may be a set of tasks to transcribe the text, another set of tasks to annotate the themes of the text, and a final set of tasks to check for errors. Breaking a task into microtasks prevents workers from too much context switching (*Guidelines for Academic Requesters* 2014), improving their capacity for short, on-demand interactions and making it easier to find errors.

Microtasks are often associated with paid platforms, where interactions are generally shorter than in volunteer crowdsourcing contexts. In fact, paid platforms are sometimes called "microtask markets" (e.g., Kittur, Chi, and Suh 2008; Ambati, Vogel, and J. G. Carbonell 2011), although this is a misnomer given that they are not inherently or necessarily based on microtasks, nor is microtasking unique to paid contexts.

Objective-Subjective contexts: Objective tasks assume the existence of a universal ground truth, or at least an agreed-upon truth, while subjective tasks have truth relative to different individuals. This work starts by looking at objective contexts, which are less complicated to study. In the latter part of *Designing Tasks for Objective Needs*, a study of poor intercoder reliability in music information retrieval evaluation is found to be due, at least partially, to the task being quite subjective. Following this, *Designing Tasks for Subjective Needs* looks at how tasks that are known to be subjective can be performed on paid platforms.

The experiments in this work general follow this scoping. While I will aim to discuss broader generalizations to other forms of crowdsourcing, like volunteer-driven crowdsourcing (e.g., Wikipedia), this will follow from secondary sources and not original research.

A Typology for Crowdsourcing provides a more thorough language for understanding these subclasses.

Relevance to Information Science

The contribution of this work is in the application of corrective techniques to the crowd-based encoding of metadata about existing information objects, and the broader understanding of the nature of such contributions.

There are many ways to apply a lens to such research. This study reflects my own field of information retrieval, and more broadly in information science.

Information science deals with representation of information objects, an area where crowdsourcing holds tremendous potential as a tool for item description.

By way of example, consider crowd curation. In the presence of large collections of information objects, information-seeking and discovery can be aided by user-curated lists of thematically-similar objects. Sites like Amazon[9], LibraryThing[10] and the Pinterest[11] let people create lists of products, books, and images, respectively.

The themes binding the lists are also user-defined, so a list can be about quality (e.g., "favorites", "worst of"), thematic (e.g., "teen vampire romance novels"), or administrative (e.g., "to buy", "read this year"). This crowdsourced information is useful to users directly, but it also provides high-quality information for understanding the content in a collection and its relationship to other materials.

Inversely, a well-designed system can make use of the additional user-supplied information on co-occurring objects. This in turn can return value to users curating the content themselves: consider a system that can discover further items for a user that are thematically in line with a group that they have compiled.

New Online Public Access Catalogues (OPACs) are also giving users the ability to classify and curate content, connecting to user habits that are commonly associated with public libraries. For

[9] https://www.amazon.com. The online store includes a curated feature called "Listmania Lists", one of a series of crowdsourcing features they refer to as the "Amazon Community". Others include customer reviews, customer communities, a pre-release review program, customer images, and the similar "So You'd Like to…" guides.

[10] https://www.librarything.com/. A community for book lovers that includes a curated 'Lists' feature for books. Other crowdsourcing features include member recommendations, tagging, and rating.

[11] http://www.pinterest.com. A social visual bookmarking website. Images from Pinterest are used as a dataset in a later chapter.

example, BiblioCommons – deployed at many library systems in North America, including the New York Public Library – positions list-making as a "curated topic guide," a way to "share your expertise with others" (*Lists* 2011). According to one study of social OPACs, the list feature in BiblioCommons is heavily used, many times greater than commenting and more than ratings (Spiteri 2011).

Similarly, cultural heritage collections have reported past success in using crowd contributions for increasing discoverability to content, improving metadata quality, or even contributing to item description. For example, after a pilot partnership with Flickr, the Library of Congress implemented a workflow for reviewing public comments on images for research or information to integrate back into item records (Springer et al. 2008).

Crowd curation is just one example of a use of crowdsourcing to create information. Table 1 shows a number of different actions that have been observed for collecting metadata.

Table 1: Types of metadata contribution activities often seen in crowdsourcing.

Action	Examples
Rating	Rating helpfulness of online comments or reviews (e.g., Amazon), rating the quality of online content (e.g., items on Youtube, Netflix, LibraryThing)
Classification / Curation	tagging (e.g., Delicious), labeling, adding to lists
Saving / Recommending	Starring, liking/recommending (i.e., Facebook), adding to favourites (e.g., Flickr)
Editing	Translations (e.g., Facebook), Corrections (e.g., National Library of Australia)
Feedback	Marking online comments as inappropriate (e.g., ABC News), "Did you find this helpful?" (e.g., Edmunds)
Other	Commenting, sharing, encoding

The desires of the contributor do not necessarily have to align with the needs of the system or the requester of the contributions. At the most basic level, a contributor may be a paid worker, where their motivation is simply to earn some money or pass the time. Other times, a contributer may contribute because of an interest in the topic, some form or personal benefit, or even as an altruistic time-killer. Table 2 shows how some commonly observed forms of contributions may mean different things to the contributor or the collecting system.

Table 2: Chart comparing contributor and system uses for a selection of crowdsourcing actions.

Action	Contributor Use	System Use
Tagging a photo / bookmark	Easy personal retrieval, appeal of collecting, item grouping for easy sharing	Improved search, improved browsing
Rating a product	Sharing opinion	improved recommendations, prioritize good values
Rating a digitally digested item i.e., video, Comment	sharing opinion, communicating approval	Identifying and promoting quality
Flagging content	cleaning windows for the community, catharsis	Higher signal-to-noise in editorial maintenance
Starring	communicating approval, saving for future reference	Identifying quality content
Sharing	showing items to friends, referring or curating content	Identifying popular/interesting content

Action	Contributor Use	System Use
Feedback	sharing personal knowledge and opinions, altruism	Correct problem data, discover system issues

While crowdsourcing has shown itself as a useful method for enriching information objects, there remains the question of how the method of collection affects the way the data can be used. Contributors are self-selected and often without verified reliability, training or expertise. Agreement is a useful metric for collecting and reconciling objective information, but sometimes there is value in disagreement, such as in collaborative filtering.

Variance that exists between different contributors adds noise both to tasks that make a subjective assumption and tasks that make an objective assumption.

In subjective tasks, it is assumed that there is no universally correct form of contribution. For example, when crowd contributions are used to inform recommendations, such as for music or film, it often assumed that different types of people enjoy different products. We thus see approaches to recommendation such as collaborative filtering, where users are matched to similar users based on the overlap between their tastes rather than a global definition of 'good' or 'bad' products. In such a case, intercoder consistency is still important, to make it possible to identify similar users. Modern approaches to collaborative filtering commonly normalize ratings against a user-specific bias (i.e., "how does this rating compare this user's average rating") and sometimes against an item-specific bias (i.e., "how does this rating compare to what the rest of the community thinks about the item").

For objective tasks, Neuendorf (2002) differentiates between two types: manifest and latent.

In a simplified comparison, tasks with manifest content are ones where there is a clear correct contribution. Transcribing text from a scanned image would be grouped in the category: the 'right

answer' is there in the image.

In contrast, latent tasks are assumed to have a theoretical truth, but one that is not outwardly stated. When a person tags a photograph with a free-text label or a worker classifies the sentiment of an opinionated tweet, they are interpreting the content: a much more abstract action. As Neuendorf (2002) notes, "objectivity is a much tougher criterion to achieve with latent than with manifest variables".

Key concepts

Before proceeding, the terminology of this study should be established. As this work spans multiple domains, and makes reference to recently introduced concepts, it is important to establish a shared understanding of language within these pages.

Note that the treatment here is cursory; a more in-depth look is available in chapters 2 and 3.

Descriptive crowdsourcing is shorthand used in this study to refer to crowdsourcing applied to descriptive metadata.

The distinction here is that the human contributions are reactive. There is an information object that already exists, and crowdsourcing workers add information about it. The response can be subjective, such as ratings or interpretations, or objective, such as descriptions or corrections.

Crowdsourcing descriptive metadata stands in contrast to crowdsourcing that *creates*, introducing new information objects into the world. One example of this is T-shirt design contests on Threadless[12].

[12] http://www.threadless.com

This approach to crowdsourcing was looked at in Organisciak (2013) when defining the concept of *incidental crowdsourcing*. Incidental crowdsourcing is an approach to crowdsourcing that is unobtrusive and non-critical. This form of peripheral collection of data was noted to favour descriptive activities.

Human computation is a separate but closely related concept

to crowdsourcing. It refers to activities where humans perform work in a paradigm reminiscent to computing, and which could conceivably one day be done by computers (Law and Ahn 2011; Quinn and Bederson 2011). Human computation does not need to be crowdsourced, but many such tasks benefit from crowdsourcing. Likewise, while there are many creative crowdsourcing tasks, such as writing or commenting, human computation represents a large portion of the types of crowdsourcing seen in the wild.

Most of the experiments in this study fall into the paradigm of human computation: collecting relevance judgments for information retrieval research, collecting descriptive labels (tags) of images on image-sharing social network Pinterest, collecting judgments of how similar songs are for music information retrieval evaluation, and collecting opinion judgments of products and food for the purpose or recommendation.

Worker, volunteer, contributor: there are many labels for people within the crowd. The space of crowdsourcing is large and the incentives for contributors are varied. The most significant distinction within crowdsourcing is in comparing uses that pay their contributors and those that do not. It's valuable to make this distinction because paying a person changes they way that they perform, while also simplifying some concerns of incentives that are necessary in retaining volunteers.

In general, crowd individuals are referred to here as *contributors*. When the distinction is necessary, paid contributors are referred to as *workers*, while elective contributions are made by *volunteers*. The former is used more commonly because more of the work in this study is paid.

In discussing 'data quality', intercoder reliability, consistency, and variance are the primary measures used.

Intercoder reliability refers to the "extent to which two or more independent coders agree" (Cho 2008), and usually is used to refer to the ability of a collection method to measure what needs to be measured. An example of low reliability would be if two raters

have the same opinion for a question, for example "is this a good tag for this image", but they choose different values on a five-point scale to register that opinion.

It is important to consider the trade-offs of intercoder reliability. In crowdsourcing, increasing intercoder reliability is sometimes at odds with the collection strategy. The most effective crowdsourcing deals with large numbers of people, and part of maximizing the involvement of contributors, especially those which are volunteers or self-selected workers, is to minimize the restrictions on a contribution. Whereas reliability can be increase by strictly enforcing a strong coding scheme or vigorously training contributors, it is also likely to reduce the number of individuals willing to perform the task. Whether the improvements in quality are worth the losses in contributions or not will be considered during this study.

Other times, controlling the circumstances under which the contribution is created is not possible, such as in information retrieval over web documents. For tasks where the contribution is numeric and ordinally or continuously coded, methods exist for interpreting when coders are similar but operating with different frame. These include using covariation instead of agreement (Neuendorf 2002), and normalizing by a user mean (Hofmann 2004; Bell, Koren, and Volinsky 2008).

A related concept is that of variance, which refers to how greatly measurements deviate. High variance means that many measurements of the same thing will vary quite a bit. Variance has this conceptual meaning, and it has a statistical meaning. Generally in this study, variance will not be used in the statistical sense; in the statistical sense, the *standard deviation* will be used (root of the variance) or root-mean-squared-error (similar to standard deviation in most circumstances). Variance is used in this study to refer broadly to varying measurements, including circumstances that do not fit into the statistical definition; e.g., "how much or how little the tagging vocabulary expands when new workers tag an image."

Chapter Outline

This dissertation is organized into seven chapters: three chapters contextualizing this dissertation and crowdsourcing in general, three chapters contributing original research, and a concluding chapter to tie it all together.

The next chapter, *Introduction to Crowdsourcing* (Chapter 2), provides a general overview of crowdsourcing. Here, a reader less familiar with the history and significant general research in the area will be introduced to them. *Design Facets of Crowdsourcing* (Chapter 3) subsequently provides a typology of crowdsourcing, tailored to understanding the breadth of online crowd systems through an information science lens. As in the previous chapter, the typology is general, intended to provide a language for speaking about crowdsourcing in the reset of the dissertation.

Interpreting Objective Tasks for Paid Crowdsourcing (Chapter 4) looks into the interpretation of already collected objective data from paid crowd tasks. Particularly, this chapter focuses on methods to remove data variance and user noise. Post-hoc data corrections and problem contributor identification has been studied from numerous angles, so Chapter 4 is careful to present past work. In addition, a study on the sources of error in crowdsourced information retrieval relevance judgments is presented, looking at the problem from the contexts of agreement, experience, and temporality.

Designing Tasks for Objective Needs (Chapter 5) delves into the design of objective tasks for paid crowdsourcing. This is one of the most common uses of crowds, to collect or encode information with a ground truth or deriving a consensus. Designing tasks that adequately motivate contributors and which collect the information that a requester thinks that are collecting is an important but often overlooked part of crowdsourcing.

Presented in this chapter are two studies that ask, *how does crowdsourcing task design affect the resulting data?*

First, a new set of experiments directly compares the effect of

design manipulations in a paid crowdsourcing platform. The same two tasks - an image retrieval relevance task and an image tagging task - are presented in drastically different ways, the designs motivated by incomplete or peripheral observations of past studies.

Secondly, a study of paid music similarity judgments is presented, which finds systematic problems in the consistency of ground truth for a task of the Music Information Retrieval Exchange attributable to task design concerns. Because the finding of this study bridge well into the later look at subjective crowdsourcing, this study is presented as a standalone half-chapter.

The final research chapter, *Designing Tasks for Subjective Needs* (Chapter 6), shifts the focus to subjective crowdsourcing. While paid crowdsourcing is often applied to objective goals, this chapter asks how collection-time strategies can improve the quality of contributions where the task goals are conditioned on a specific person's tastes or needs. Building on work developed by Organisciak et al. (2013), methods are presented to perform subjective crowdsourcing for on-demand personalization, showing it to be feasible for our evaluated settings. Following from the earlier study on the effect of design manipulations for objective tasks, this chapter also studies the influence of task design changes in how crowds contribute using one of our subjective crowdsourcing protocols, taste-grokking.

Introduction to Crowdsourcing

Crowdsourcing is a conceptually simple idea that has received considerable research attention in the past few years, alongside a realization of the power of the internet for effectively connecting people in large numbers.

The language of crowdsourcing has developed fairly recently, but the ideas it represents have been practiced and studied in various forms prior. Perhaps it is not surprising then that research in crowdsourcing has been uneven and discussion scattered. After all, this is a term that seemingly sprung from a very specific place, on a specific date, and yet what *is* crowdsourcing has been largely appropriated and defined by collective imagination.

As an introduction to key concepts of crowdsourcing, this chapter provides an overview and the notable research that has stemmed from it. The purpose of this chapter is as an interstitial of sorts, providing background information which will be helpful in grounding an understanding of the rest of this dissertation.

Crowdsourcing broadly describes the use of distributed crowds to complete a task that would otherwise be done by one or a few people. It broadly captures the abilities of the internet as a communication medium in efficiently connecting people. Many concepts exist within or overlapping with this broad mandate.

Nothing about crowdsourcing is fundamentally tied to the internet, however. It is entirely possible to bring together large groups of people in different ways, but the access and efficiency of the internet is both what makes the concept seem so novel and what makes it valuable in the various realms where it is applied.

Whereas crowds have long been noted for their collective simplicity (Le Bon 1896) or irrationality (Mackay 1852), through the internet one can perform human-specific tasks at a scale usually only seen for computational tasks.

The term *crowdsourcing* comes from a 2006 Wired article by Jeff Howe (2006c). While the word is recent and has an unambiguous source, immediately upon its introduction it was adopted and expanded on through public discourse. Howe was writing from a labor perspective, looking at online marketplaces for people to solve problems and create content. His focus was on systems like InnoCentive[13], a site for companies to outsource research and development problems for a bounty, and iStockPhoto[14], a website that allowed amateur photographers to sell their images as stock photos. The article briefly looked at user-generated online content, though in the context of television programs that use online video as content, rather than the bottom-up style of content creation associated with the first two decades of the internet. Despite the narrowness and brevity of the initial definition, the term *crowdsourcing* struck a chord more broadly and was culturally co-opted. The definitional appropriation happened very quickly: within nine days Howe noted a jump from three Google results to 189,000 (Howe 2006a). Within a month, Howe addressed the co-opting of the term, "noticing that the word is being used somewhat interchangeably with Yochai Benkler's concept of commons-based peer production" (Howe 2006b). He gives his definition[15], but also notes that language is slippery, and he is "content to allow the crowd define the term for itself (in no small part because [he is] powerless to stop it.)."

[13] http://www.innocentive.com/

[14] http://www.istockphoto.com/

Thus, crowdsourcing was adopted to refer broadly to a series of related concepts, all related to people being connected online. These concepts included free and open-source development (Lakhani and E. v. Hippel 2003; Raymond 1999), the 'wisdom of the crowds' (Surowiecki 2004), human computation (Ahn and Dabbish 2004), and commons-based peer production (Benkler 2006). Further, it overlaps with the content of user-generated content, at

[15] "For the purposes of the article... we would only look at case studies involving big established companies. For the purposes of [Howe's crowdsourcing blog] ... I interpret crowdsourcing to be taking place any time a company makes a choice to employ the crowd to perform labor that could alternatively be performed by an assigned group of employees or contractors". To not leave the journalist with his 2006 definition, Howe's definition expanded further as time went on, away from top-down companies doing the outsourcing, and eventually to "content created by amateurs", a movement influenced by free and open source software (Howe 2008).

least to the extent that user-generated content is used toward a common production or purpose. Each of these are discussed in greater detail below.

Related Concepts

Free and Open Source Software (FOSS). The FOSS movement started with the sharing of software source code for interested parties. Distributed collaboration was not initially a tenet of this openness, but it followed as a consequence. Open-source development began to adopt some unique properties: users and distributed developers could jump into the code to fix a bug, or add a feature that they wanted to see.

The significance of this became apparent when Linus Torvalds released Linux in 1992 with a development model that accepted external code contributions heartily, released early and often, and followed the pulse of users' needs. Raymond compared this form of software development to a bazaar, "open to the point of promiscuity", and contrasted it to the traditionally managed 'cathedral' style seen in the commercial world and earlier open source projects (Raymond 1999).

The many hands approach to open-source demonstrated that technologically-connected crowds can coherently delegate and create works. Like with crowdsourcing, open source software development often does not discriminate on credentials or background; if a contributor can make an adequate contribution, it can be used.

The roots of crowdsourcing in open source are credited in Howe (2008) and also are on display in Howe's "soundbyte" definition: "the application of Open Source principles to fields outside of software" (sidebar, www.crowdsourcing.com).

Wisdom of the crowds. *The Wisdom of the Crowds* (Surowiecki 2004) observed the collected effectiveness of crowds when properly aggregated. Building from Francis Galton's *Vox Populi* (1907), where Galton aggregated guesses at a steer weight guessing competition and found that the median guess was more accurate that

any individual guess, Surowiecki argues that the ability of many autonomous people to aggregate into a product comparable to something an expert would produce has important ramifications on the internet.

The term 'wisdom of the crowds' has survived the book to refer to the strength of human decision-making in aggregate, and design patterns that make use of that strength.

For example, the wisdom of the crowds is utilized in crowd-sourcing opinions (e.g. product reviews on Amazon, film reviews on Netflix) and in filtering (e.g. liking of starring posts on a social network).

Part of the wisdom of the crowd is simply statistical. In one of Surowiecki's examples, he points to the quiz-based game show *Who Wants to be a Millionaire?* On the show, contestants unsure about their response can poll the audience. The audience poll turned out to be remarkably effective, but not surprising: even if most of the audience does not know the answer and the probability of choosing one of the four choices in ignorance is roughly equal, then only a few people that are informed of the answer can sway the "crowd" response in the right direction.

This is a fitting anecdote, given that many crowdsourcing efforts do come down to connecting to the right individual from the mass of candidates. It is seen most clearly in cases such as question and answer websites (e.g., Stack Overflow, Ask Metafilter, Quora). However, increasing the pool applies in much more than cases of 'wisdom': many successful websites receive the bulk of their contributions from a small core group of contributors (e.g., Wikipedia - Muchnik et al. 2013; Transcribe Bentham - Causer, Justin Tonra, and Valerie Wallace 2012; The Commons - Springer et al. 2008) and the benefit of opening up their projects to public contributions is in increasingly the likelihood of a "power user" (Springer et al. 2008).

The second lesson of the wisdom of the crowds that permeates crowdsourcing is the idea of aggregation that results in a product better than the sum of its parts. Grand projects like Wikipedia and FoldIt[16] (Khatib et al. 2011) allow contributions to build on the

[16] An online research-supporting game that looks for the most efficient ways to fold proteins.

work of past contributors.

Human computation. Human computation was introduced in the doctoral work of von Ahn, accompanying work on the *ESP Game*, a game where the players tag online images during play (Ahn and Dabbish 2004; Ahn 2006). If the wisdom of the crowds refers to the unique abilities of human intelligence in aggregate, human computation focuses on human abilities as distinct from computational methods – for so long as they are distinct – and aims to formalize methods to organize humans in manners akin to automation. It refers to the process of computation – the "mapping of some input representation to some output representation using an explicit, finite set of instructions" (Law and Ahn 2011) – performed by humans.

Quinn and Bederson (2011) offer a taxonomy of human computation, classifying along dimensions of motivation, quality control, aggregation, human skill, process order, and task-request cardinality. In synthesizing the various definitions of human computation in relation to crowdsourcing, collection intelligence, and social computing, Quinn and Bederson (2011) note two characteristics of consensus in the definition: that "the problems fit the general paradigm of computation, and as such might someday be solvable by computers", and that "the human participation is directed by the computational system or process".

As noted by Law and Ahn (2011), Turing defined the purpose of computers as carrying out operations that humans would normally do. Human computation, is humans performing work that computers would normally do, but are not yet able to.

By this definition, much human computation aligns with crowdsourcing, but large swaths of crowdsourcing are not relevant to human computation. For example, creative crowdsourcing projects like T-shirt design website Threadless are not human computation. Inversely, human computation does not have to be sustained by an open call; a more traditionally employed closed system can suffice (Law and Ahn 2011).

The paradigm of computation in human computation is just a

subset of ways that crowds can collaborate in crowdsourcing, and human computation can be performed without the modality of multiple collaborators seen in crowdsourcing.

Commons-based peer production and **user innovation**.

Recent cultural observers have noted the behaviours seen in crowdsourcing through various lenses. Crowdsourcing emerges from various affordances – both technical and social (Wellman et al. 2003) – of modern information networks. Such as was seen with open-source software development, networked society encourages new forms of cultural creation, not by intention but by consequence of the type of connectedness it allows.

As networked society has developed and the internet has grown ubiquitous, numerous scholars have noted the cascading consequences in how individuals interact with culture and participate in the creation of cultural objects. Two such streams of study are von Hippel's work on *user innovation* and Benkler's study of the networked information economy, including his concept of *commons-based peer production*. Both of these borrow from economic and market-driven theory rather than sociological theory, but they offer valuable language for understanding crowdsourcing as a cultural phenomenon.[17]

[17] One might argue for the term *consequence* rather than *phenomenon*, because it positions crowdsourcing as neither an accident nor a product of intention, but acknowledges a history for it where it is a side effect of external influences.

If crowdsourcing is a generalized version of open source principles, von Hippel's work on user innovation (E. v. Hippel 1988; E. v. Hippel 2006) was an early observation of the trend toward a greater user focus in computer tools and services.

With user innovation, new information products or physical products are generated by users – those that benefit from using rather than selling the product. Notably, von Hippel focuses on 'lead users,' users with specific needs that precede broader trends. These users either develop new products to fill their needs or modify existing products.

Not all crowdsourcing creation is user innovation, though there are echoes of von Hippel's work in companies that turn to the Internet for help in conducting their business, whether it is soliciting feedback and suggestions (e.g., MyStarbucksIdea [18]), bug reports,

[18] http://mystarbucksidea.force.com/

or even work at a bounty (e.g., 99Designs[19]). User sharing of work performed for themselves is another similar area: for example, when a music service allows users to share their playlists publicly, their realization of a personal need has potential value to other users.

[19] http://99designs.com

Benkler's work takes a political economy view on what he calls the 'networked information economy', but arrives at a very similar place to von Hippel. He argues that the unique landscape of the 'networked information economy' empowers individuals to do more for themselves and in collaborative groups outside of established economic spheres (Benkler 2006). This agency allows commons-based peer-production: for innovation and creation to rise out of the commons rather than from firms.

Benkler (2006) singles out two user behaviors borne out of access to information networks, which in turn underlie the rise of crowdsourcing. First, individuals are more empowered to operate autonomously, for themselves and with less reliance on mass-market goods. At the same time, loose collaborations are easier to organize, allowing the pursuit of individual needs at scales beyond the capabilities of a single person.

Citizen science. Citizen science refers to collaboration between scientific communities and members of the public on research. Early crowdsourcing projects, such as galaxy annotation site *Galaxy Zoo* (Lintott et al. 2008)[20] and protein-folding competition *FoldIt*(Khatib et al. 2011)[21], were noted as a form of citizen science, and crowdsourcing has been used for numerous successful results in the field.

[20] http://www.galaxyzoo.org

[21] https://fold.it

Wiggins and Crowston (2012) present a typology of citizen science projects, organizing them into action-oriented, conservation-focused, investigative, wholly-virtual, and educational projects.

Crowdsourcing in Practice

There is a great deal of crowdsourcing "in the wild", including notable successes and failures. The successful projects are partic-

ularly worth looking at for clues as to what distinguishes them in the face of less successful or failed sites. Below is a selection of projects that have lasted. This small list is chosen in the service of a few points.

First, many of these projects are approaching or have surpassed a decade of existence: an eon for the networked age. The age shows, however, which certainly adds dimension: novelty wears off and communities gentrify. A project such as Wikipedia or LibraryThing has a very different makeup than a new and novel project as the British Library's LibCrowds (Chiesura et al. 2015).

Additionally, the small selection of examples below is chosen for breadth. This dissertation focuses on a small corner of crowdsourcing, but there are many models for online contribution that have been tried, so it is good to have concrete anchors to go by.

Still, looking at crowdsourcing *web sites* misses part of the legacy of crowdsourcing. Adopting a speculative position for a moment, it appears that many of the design patterns that will survive from the past decade of experimentation with crowdsourcing will be in the augmentative, supportive roles it can play: community-contributed translations or subtitles; qualitative contributions like flagging, rating, or 'likes'; casual filtering activities like up/down voting. Likewise, the best new projects are ephemeral: they are not intended to last by design. The point has been made that the amount of human effort and leisure-time labour on the Internet is endless (Shirky 2009; McGonigal 2011); however, attention is scare.

Projects like those from citizen science exemplar *Zooniverse*[22], discussed below, or *LibCrowds*[23], or the crowdsourcing projects from NYPL Labs [24]: they develop single-serving projects to symbiotically engage communities with their collections in focused, short-term ways, rather than grandiose 'digitize all of history' projects. In my past work on motivations of crowds (Organisciak 2010), revisiting in the next chapter's crowdsourcing typology, I avoid discussing novelty given the expectation that it was unsubstantial. Novelty is indeed ephemeral, but this does not par-

[22] *Zooniverse* is a collection of citizen science crowdsourcing projects. https://www.zooniverse.org

[23] *LibCrowds* is the space tying together the British Library's crowdsourcing initiatives. http://www.libcrowds.com

[24] One example of a single-serving project from NYPL labs is *What's on the Menu*, a transcription effort for restaurant menus. http://menus.nypl.org

ticularly detract: projects with short-term design may make that ephemerality acceptable while capitalizing on the public's initial excitement at a new project.

With that in mind, below are some notable examples in the wild and as a whole, while the next chapter's typology provides a contrasting view of crowdsourcing in its parts.

Wikipedia is a collaboratively-written encyclopedia, where the majority of contributors are volunteers. Wikipedia, formed in 2001 and now containing 4,579,708 articles (as August 2014: *Wikipedia* 2014), has an open editing policy that allows anonymous contributions and only restricts who can edit a page for few special cases where vandalism is likely. The policy also ensures that readers are latent editors (Shirky 2009), helping police, correct, and improve poor quality content.

Despite being a notable success, the maturing of the community and the increased difficulty of contribution that comes with more community rules has been blamed for falling numbers of new users (Angwin and Fowler 2009). Wikipedia also has a high gender bias, and it has been argued that the exclusionary effects of the increasingly strict community (or at the least the perception of such) disproportionally turn away women contributors (Gardner 2011).

Threadless is a community of artists that design and vote on T-shirt designs. Winning designs are licensed by Threadless to print and sell, providing a commission to the designer and additional profit for subsequent shirt reprintings.

Threadless was one of the examples discussed in the initial treatment of crowdsourcing by Howe (2006c), and it has stayed remarkably similar in the ensuing nine years. Despite also becoming a platform that commissions designs from professionals, the central model still hinges on anybody-can-contribute, anybody-can-vote design contests.

The **Netflix Prize** was a competition run by film rental (and now streaming) company Netflix, offering a million dollar bounty to the person or team that could improve film recommendation

by 10% over the root-mean-squared-error performance of Netflix's own system. Claiming the prize required the winner to publish their results but did not require transfer of intellectual property, only a license for Netflix. A 2008 New York Times article about the prize noted that the community of participants were notably open in sharing their insights (Thompson 2008).

Underlying the Netflix Prize's open call for expert contributions was another type of crowdsourcing: modeling the quality of Netflix's collection through user-contributed rating. This use of user-generated content for prediction and recommendation is an area known as collaborative filtering (Resnick et al. 1994; Hofmann 2004).

Kickstarter is a microfunding platform that enables patronage of artists and creators in their project through small but plentiful contributions. A project creator on Kickstarter proposes a project and offers tiers of rewards for backers that contribute varying amounts. When researched in Organisciak (2010), the balance between the altruistic support-based motivation and opportunistic reward-based incentives seemed to weigh slightly more toward the former, though I expect this has changed in recent years as more products have been offered on the site. Regardless, the model of small contributions from many has been seen in many other so-called crowdfunding contexts, including charity, politics (Fung n.d.), and small business (Cortese 2011; Cortese 2013).

Zooniverse is a series of crowdsourcing projects that started with Galaxy Zoo. Galaxy Zoo allowed the public to classify images of galaxies from the Sloan Digital Sky Survey, many being seen for the first time, at a pace much quicker than any one human could perform. Another popular project, Old Weather (Eveleigh et al. 2013), transcribes weather logs from old ship's journals. In Snapshot Serengeti (Swanson et al. 2015), participants classify animals photographed in camera traps. Many of the Zooniverse projects follow a similar pattern: encoding of curious, novel, or interesting images while contributing to real research.

FoldIt is a game where users try to develop the most efficient

folding of a protein (Khatib et al. 2011). Folds are scored and placed on a leaderboard, adding a competitive edge. FoldIt shows that, when well matched to competitive impulses, complex problems can be tackled through semi-anonymous online workers.

ReCaptcha (Ahn, Maurer, et al. 2008) cleverly took a system intended to distinguish humans from bots – obfuscated text transcription with Captchas – and combined it with a problem that by definition only humans can do: fixing scanned text that computational techniques failed at. With ReCaptcha, online visitors prove they are human and help digitize scanned archives at the same time.

Crowdsourcing in Information Science

In information retrieval, the focus on crowdsourcing has been predominantly in the use of paid crowds for generating evaluation datasets, though there have been efforts to use crowds to improve document representation or even query-specific ranking.

The benefit of paid crowds for relevance judgments is that it allows for on-demand evaluation datasets (Alonso, Rose, and Stewart 2008). This has been a costly and exhausting process in the past, making it difficult to perform IR research on more novel datasets than the judged sets available from TREC. Relevance judgments benefit from the agreement among multiple humans, since the concept of 'relevance' is not clear-cut but rather negotiated and agreed upon. The ability to attract a breadth of rater types also positions paid crowdsourcing as an effective means to collecting evaluation data.

TREC itself ran a crowdsourcing track for three years, the primary task a competition to improve relevance judgment quality (Lease and Kazai 2011; Smucker, Kazai, and Lease 2012).

Another common use of crowdsourcing is for information retrieval correction of results. Manual tweaking of results is not a scholarly activity, but there is evidence that it is done often in practice, by companies such as Twitter (E. Chen and Jain 2013),

Using crowdsourcing in the machine, as evidence for search engine algorithms rather than evaluation, is less common. PageRank is one such effort, utilizing the linking habits of web page authors as a proxy for authoritativeness and quality (Page et al. 1999). Recently, crowdsourcing has proven useful for time-sensitive queries, and has been used by Twitter to model searches that may have never been seen before (E. Chen and Jain 2013).

One of the better explored spaces of retrieval over or incorporating crowdsourced information is in folksonomies. Folksonomies refer to free-text labelling (i.e., 'tagging') by non-professionals. A popular resource for folksonomies over general web documents is the older incarnation of bookmarking website del.icio.us. In folksonomies such as on del.icio.us, over 50% of tags contribute information that was not contained in the document; for music tags (on the website Last.fm), over 98% of tags provide text information not previously held in the record (Bischoff et al. 2008). Information retrieval can benefit for this extra information, and a comparison of web query logs to folksonomies from del.icio.us, Flickr, and Last.fm shows that 58.43-71.22% of queries overlap at least partially with tags in those systems (ibid).

Studying ways to retrieve saved bookmarks on del.icio.us, (Hotho et al. 2006) present *FolkRank*, a technique to adjust authority of authors and importance of tags in order to find important resources. While their approach has limited success as a generalized retrieval approach, they find that it holds value in identifying communities of interest within the community.

(Zhou et al. 2008) present a generalized framework for dealing with social annotations within the language modeling approach. Their model categorizes users by expertise domain and builds domain topics from related annotations. These are linearly smoothed with document and query language models. In the context of del.icio.us, their approach improves over traditional unigram models over the document text.

Finally, Harris and Srinivasan (2012) provide a comprehensive overview of ways that crowdsourcing and *games with a purpose*[25] can be incorporated in the information retrieval workflow. While crowdsourcing is noted as highly feasible for evaluation, it is also noted as an approach which can help in building document collections, identifying information needs, and query refinement.

Discussion of crowdsourcing in information science continues generally in the next chapter, an IS-centric typology of crowdsourcing, then in the context of paid crowdsourcing in the subsequent two chapters.

[25] *Games with a purpose* was introduced by von Ahn (2006) to describe online games as a mechanism to collect information from crowds. A popular example was the *ESP Game*, where paired players competed with the clock to independently agree on a tag for an image.

Summary

Crowdsourcing is a phenomenon with a wide umbrella and a broad range of parameterizations. For information science, it is potentially very valuable for its ability to efficiently gather extra-textual information about existing objects. The next chapter presents a typology, again focused on broad crowdsourcing, before turning back to focus on information science crowdsourcing with the subsequently original research chapters.

A Typology for Crowdsourcing

The scope of crowdsourcing is broad and the myriad approaches to collaboration among distributed crowds lend a lack of coherence which may intimidate a practitioner. To address the sprawl and provide a structure for the rest of this work, this chapter presents a typology of crowdsourcing for information science.[26]

Crowdsourcing, the collaboration of distributed contributors on a common product, promises value to library and information science in a variety of ways. Information systems and digital repositories deal with overwhelming amounts of materials that can be annotated with help from many hands, and the relationship that cultural heritage collections hold with their audience can potentially be strengthened by pursuing meaningful collaboration between the two. Holley (2010) notes some potential uses to crowdsourcing, including tapping into the expertise of the community, building loyalty of users while tapping into their altruistic tendencies, adding value to data such as with quality ratings, and improving information access to materials.

There have been earlier attempts at crowd taxonomies (e.g. Geiger et al. 2011; Vukovic and Bartolini 2010; Schenk and Guittard 2009; Rouse 2010). However, these have primarily emerged from other domains, with a focus on economic or quantitative variables. Perhaps the most valuable prior work is in Quinn and Bederson's taxonomy of human computation (2011), a field focusing on humans performed work in the mode of computing. Human computation often overlaps with crowdsourcing but focuses

[26] A version of this chapter was previously presented at iConference 2015 with co-author Michael B. Twidale (Organisciak and Twidale 2015). Co-authorship notes in appendix. Copyright retained by authors.

on a more narrow type of labor and is not necessarily performed by distributed crowds. Wiggins and Crowston (2012) also offer a typology of a useful related concept, citizen science.

Geiger et al. (2011) identify crowdsourcing processes by four defining characteristics: the pre-selection process for contributors, the accessibility of peer contributions, the aggregation of contributions, and the form of remuneration for contributors. While these are all valid ways of viewing crowdsourcing, more qualitative or naturalistic models are also necessary in order to understand crowdsourcing websites, such as motivation or centrality.

Schenk and Guittard (2009) provide a management science view on crowdsourcing, with a typology along two dimensions. First, crowdsourcing is distinguished by how work is collaborated on: in an integrative or selective manner. Secondly, the type of work that is performed is faceted into routine, complex, and creative tasks. Vukovic and Bartolini (2010) take yet another frame, of business-centric crowdsourcing uses. Crowd type, incentives, quality assurance, government and legal, and social factors play into a parsing of crowdsourcing in this scope. Finally, Rouse (2010) propose a taxonomy based on the nature of the crowdsourcing, focusing on capabilities (simple, moderate, sophisticated), benefits (community, individual, or mixed), and motivation. Their hierarchical taxonomy notes motivations relative to the other two conditions.

Each of these taxonomies has features to inform our understanding of crowdsourcing. However, they are generally grounded in different domains than information science, focusing more on crowds as labour and missing ways of conceptualizing the product or the volunteer contributor that are useful for our purposes.

An Information Science Typology of Crowdsourcing

The space of crowdsourcing is large, and there have been a number of attempts to organize the sub-concepts within it or to reconcile it in a space alongside other areas of research. Some of the

most important questions in differentiating crowdsourcing include:

- Motivation: How are contributors motivated? Are they paid or do they volunteer for other incentives?
- Crowd type: Who are the contributors? What are their skills?
- Contribution type: Are contributions new, or do they react to existing documents or entities? What do the contributions look like? Are they subjective (involving opinions or ranking) or objective (there is an agreed best response)?
- Aggregation and style of collaboration: Are contributions preserved in the form they are submitted, or are they combined into a larger contribution? Is the collaboration indirect (i.e. contributors work on parts independently) or truly collaborative? How is quality controlled for?
- Beneficiary / Director: Who is asking for the contributions? Who is benefiting?
- Centrality: Is the crowdsourcing central to the system?

Table 1 provides an overview of this crowdsourcing typology, including references when the dimensions are influenced closely by prior work. In the next section, we consider existing work more thoroughly, adapting it into our typology, explain how we reinterpret it, and argue for new facets not present in non-IS taxonomies or classifications.

Table 3: Overview of facets in this study's crowdsourcing typology.

Category	Description	Sub-categories
Motivation	How are contributors incentivized?	Primary/Secondary (Organisciak 2010), Contribution/commitment (Kraut and Resnick 2011) Extrinsic/Intrinsic
Type of Crowd	What are the dimensions of the crowd and how they are expected to perform?	Unskilled, locally trained, specialized heterogeneous / diverse
Type of Contribution	What is the nature of the work?	Human computation / Creative Generative / Reactive Subjective / Objective
Aggregation	How are diverse contributions reconciled into a common product?	Selective /Integrative (Geiger et al. 2011; Schenk and Guittard 2009) Summative / Iterative / Averaged
Beneficiary / Director	Who benefits? What is their relationship to contributors?	Autonomous / sponsored (Zwass 2010) Crowd / individual
Centrality	How central is the crowdsourcing to the overall project?	Core / Peripheral (Organisciak 2013)

Motivation

The incentives for contributors to participate in crowdsourcing are complex and not always consistent from contributor to contributor.

Motivation in crowdsourcing follows related work in the motivations of humans in general (Maslow 1943; Alderfer 1969; Ryan and Deci 2000). While a review of that work is beyond the scope of this paper, many views of crowdsourcing motivation adopt the lens of motivation as a mixture of *intrinsic* factors and *extrinsic* factors (Ryan and Deci 2000). In the former, fulfillment is internal to the contributor, psychologically motivated, while in the latter the rewards are external.

The spectrum of intrinsic to extrinsic motivators is commonly paralleled in crowdsourcing literature through a dichotomy of paid and volunteer crowdsourcing (Rouse 2010; Geiger et al. 2011;

Kraut and Resnick 2011; Schenk and Guittard 2009).

Paid and volunteer crowdsourcing are not exclusive, and there are extrinsic motivators beyond money. However, this separation is common because it accounts for some of the starkest differences between how crowdsourcing is implemented and motivated. There are differing design implications around people being paid and performing work for other reasons: money is a direct currency for obtaining labor, while convincing volunteers to contribute requires a greater sensitivity of their needs and ultimately more complexity in engineering the crowdsourcing system.

It has been shown that intrinsic motivation still plays a part in paid crowdsourcing (Mason and Watts 2010), and some systems mix intrinsically motivated tasks with payment or the chance at remuneration. For example, some contest-based marketplaces are popular among users looking to practice their skills, such 99Designs for designers or Quirky for aspiring inventors.

Some taxonomies make a distinction between forms of payment. Geiger et al. (2011) makes the distinction between fixed remuneration, with a pre-agreed fee, and success-based remuneration, such as contest winnings or bonus.

Taxonomies of specific motivators seen in crowdsourcing have been previously attempted, with varying results that touch on similar issues. Organisciak (2010) identified a series of primary and secondary motivators from a diverse set of crowdsourcing websites. We adopt the categories from that study for our typology, as related work is accommodated well.

Primary motivators are those that are considered critical parts of a system's interaction. Systems do not require all of them, but to attract and retain contributions, they need one or more of them. In contrast, *secondary motivators* are system mechanics that generally were not observed as necessary components of a system, but were elements that encourage increased interaction by people that are already contributors. (Kraut and Resnick 2011) parallel the primary/secondary split by differentiating between encouraging

contributions and encouraging commitment.

The motivators in (Organisciak 2010) were observed from a content analysis of 13 crowdsourcing websites and subsequent user interviews. For sampling, 300 websites most commonly described as 'crowdsourcing' in online bookmarks were classified with a bottom-up ontology, then the 13 final sites were selected through purposive stratified sampling, to represent the breadth of the types of crowdsourcing seen. These cases were studied in case studies followed by user studies.

Below is a list of primary motivators seen in Organisciak (2010), but also paralleled and supported by the similar broad view social study published by Kraut and Resnick (2011).

- **Money and extrinsic reward**. Paying people is the most reliable approach for collecting contributions, and is an option in the absence of other motivators or where certainty is required. However, it also introduces bottlenecks of scale, and negates some benefits of intrinsic motivation. Mason and Watts (2010) note that, while intrinsic motivation still exists on paid crowdsourcing platforms, it is overwhelmed when tasks are too closely tied to reimbursement, resulting in contributions that are done minimally, briskly, and with less enjoyment. Kraut and Resnick (2011) point to psychology research that shows the ability of reward in other settings to subvert intrinsic motivation, leading to less interested contributors.

- **Interest in the Topic**. Projects catering to people that have a pre-existing interest in their subject matter or outcomes tend to get longer, more consistent engagement. For example, the Australian Newspaper Digitisation Project (now part of a larger project called *Trove*) found that that amateur genealogists, with pre-existing communities and a willingness to learn new technologies, took "to text correction like ducks to water" (Holley 2009). Similarly, Galaxy Zoo found similar success with amateur astronomers helping annotate galaxies. Kraut and Resnick likewise argue that asking people to

perform tasks that interest them results in more engagement than asking people at random.

- **Ease of entry and ease of participation**. Low barriers to entry and participation were cited by every user interviewed in Organisciak (2010). Wikipedia has a low barrier to entry but its interface and demanding community standards have been criticized in recent years for raising the barrier to participation Angwin and Fowler (2009) and Sanger (2009). "Simple requests" generally lead to more productive contributions, according to Kraut and Resnick (2011).

- **Altruism and Meaningful contribution**. People like to help if they believe in what they're helping. Writing about Flickr Commons, Library of Congress noted that they "appear to have tapped into the Web community's altruistic substratum by asking people for help. People wanted to participate and liked being asked to contribute" (Springer et al. 2008). With Galaxy Zoo, the appeal for many contributors that it offers a tangible way to contribute to real science. Rouse (2010) also argues for altruism's place in a taxonomy of crowd motivation. Kraut and Resnick (2011) argue that appeals to the value of a contribution are more effective for people that care about the domain.

- **Sincerity**. "People are more likely to comply with requests the more they like the requester," Kraut and Resnick (2011) note. A recurring theme among interview participants in Organisciak (2010) was whether a project seems sincere or exploitative. Since crowd contributions often exist as a parallel to labour, crowds are often weary of anything that smells like them being taken advantage of.

- **Appeal to knowledge and opinions**. One curious source of motivation is simply asking the right people. Online visitors presented with a question are often compelled to answer it simply because they know the response, be it part of their knowledge, skills, circumstance, or opinions. The 'appeal'

itself can be explicit or implicit. Kraut and Resnick (2011) refer to this sort of appeal as "Ask and Ye Shall Receive", asserting that online communities stand to benefit from easily accessible lists of what work needs to be done. They also assert that direct requests for contribution are better than broadcast.

One motivator overlooked in Organisciak (2010) is *novelty*. Novelty or curiosity is ephemeral and unsustainable, but nonetheless a unique idea can attract contributions for a short amount of time. Kraut and Resnick (2011) also note structure, goals, and deadlines as incentives. Such an effect is strongly felt on Kickstarter, where the tenor of crowdfunding for projects changes relative to the funding end date.

The supplemental secondary motivators, based on Organisciak (2010), which encourage more engagement but not initial contribution, are:

- **External indicators of progress and reputation**. Using games, badges, or leaderboards encourages more contribution among certain people. An important caveat is that this form of performance feedback needs to be perceived as sincere (Kraut and Resnick 2011).
- **Feedback and impression of change**. Showing the contribution in the system or conveying how it fits into the whole. Kraut and Resnick (2011) tie feedback to goals, emphasizing the importance of showing progress relative to personal or site-wide goals.
- **Recommendations and the social**. Prodding by friends, colleagues, and like-minded individuals. Simply seeing that other people have contributed makes a person more likely to contribute (Kraut and Resnick 2011). This motivator factors into the taxonomy by as *social status*.
- **Window fixing**. Nurturing a well-maintained community where the members feel compelled to support its health.

Type of Crowd

Vukovic and Bartolini (2010) define two extremes of crowd types: *internal* and *external*. Internal crowds are composed solely of contributors from the organization that is crowdsourcing, if it is thus centralized. External crowds are members outside of the institution. Vukovic and Bartolini (2010) also note that *mixed* crowds are observable.

A point of separation between crowd methods is the skills required to perform the work. *Unskilled*, *locally training*, and *specialized* are all seen among crowdsourcing systems. Where unskilled labour encourages contributions from anybody at anytime, systems that use methods for authority control leave certain tasks to long-term, involved contributors. For example, on question and answer service *Stack Overflow*, a user's administrative ability grows more open as they contribute more to the management of the system, a way of ensuring that those users have learned the proper management of the site.

In additional to what the crowd is, there is a distinction to be made on what the crowd is desired to be. Here, it is helpful to think of a spectrum between *diverse* and *homogeneous* crowds. In some cases, the crowdsourcing task benefits from multiple unique viewpoints. When online players compete to fold proteins in the most efficient way possible for *FoldIt*, the project's success is predicated on the ability of people to problem-solve in variable ways. In contrast, for a project like *Building Inspector* where participants outline building boundaries from scanned survey records, the desire is for the participants to perform in a standard way. Here, reliability and consistency are important traits.

Type of Contribution

The type of work performed by crowds can vary greatly in its complexity and style.

One notable form of crowdsourced work is represented by the concept of human computation, where "the problems fit the general paradigm of computation, and as such might someday be solvable by computers" (Quinn and Bederson 2011). Understanding that crowdsourcing is not solely human computation tasks, the inferred corollary to these types of tasks are those that are expected to be too complex for computers: creative, judgment-based, or requiring critical thinking. Creative crowdsourcing might take the form of artistic human expression, such as online contributors collectively animating a music video (*Johnny Cash Project*) or the sum of YouTube. Opinion or judgment-based crowdsourcing often does not have a definitive answer, and is seen in areas such as movie reviews or product ratings. More complex critical thinking tasks do not fit the paradigm of computation and are much more complex, such as Wikipedia or protein-folding project FoldIt.

Schenk and Guittard (2009) have previously distinguished between three types of crowdsourcing. First are routine tasks, such as crowdsourcing of OCR text correction with ReCaptcha. The majority of human computation tasks would likely fall within this category of rote tasks. Second are complex tasks, such are open-source software development. Finally, they suggest creative tasks, with a slightly different meaning than our typology's usage as a disjunct to human computation. An example of their final category would be a system like *MyStarbucksIdea*, a space where people suggest changes they would like to see at the coffee chain Starbucks. Since Schenk and Guittard (2009) focus on crowdsourcing when there is a client, usually a corporate client, they do not consider the wider space of creative crowdsourcing tasks.

Another view that touches on the nature of the contribution is *generative* versus *reactive*. In the former, new intellectual products are created. With reactive work, the work is a reaction or interpretation of an existing information object: reviews, ratings, encoding.

Such a distinction is neglected in most views of crowdsourc-

ing, but important in information science. At the heart of many projects by libraries, museums, and cultural heritage institutions, is a focus on information objects. There is much effort expended in archiving, enriching, appreciating, and sharing works, and a reactive view of crowdsourcing products places the public within this tradition.

A FINAL VIEW OF TYPES OF WORK, one adopted strongly in this dissertation, is the spectrum between *objective* and *subjective* tasks.

Objective tasks are assumed to have an authoritative truth, even if it is unknown. For example, in transcribing scanned texts, it is assumed that there is a 'correct' transcription in the work that has been scanned.

In contrast, subjective tasks have a variable concept of correctness, as they are not expected to be consistent between contributors.

Human computation undertakings are commonly objective tasks, and taxonomic efforts for human computation – such as Schenk and Guittard's split of routine, complex, and creative (2009) – do not touch on the subjective/objective separation directly.

The subjective-objective distinction has consequences for training and quality control. Objective tasks lead to a training approach where the ideal result is that everyone performs the task in the same one right way. Quality control on those tasks can employ approaches such as intercoder reliability, since it can be assumed that there is an objective set of results that raters are striving for. Subjective tasks can still need training and quality control, but it will necessarily be of a different kind. For example, certain subjective tasks want to take advantage of the diversity of human activity and so explicitly do not want everyone to do the same thing in the same way.

This distinction is still present with different forms of aggregation. Multiple contributions can be aggregated with an objective assumption, expecting a truth and deviations from it as bad work

or data. Other systems try to aggregate a normative opinion or judgment of subjective contributions. This latter assumption is seen often in opinion ratings, such as film or restaurant ratings: just because there is an aggregated rating presented, there is an understanding that some people might disagree and that they are not incorrect for doing so.

Aggregation

Schenk and Guittard (2009) and Geiger et al. (2011) discuss two types of aggregation: *integrative* and *selective*. Integrative aggregation pools contributions into a common product, like a wiki, while selective aggregation tries to choose the best contributions, such as in contests.

This simple separation hides some complexity seen in aggregation approaches. Reconciling multiple different contributions can be difficult, and integrative aggregation can be approached in a number of ways. We argue the following finer views on integrative aggregation are useful:

- **Summative**. In summative aggregation, people contribute to an ever-expanding base of information. Contributions are clearly part of a bigger whole, but their individual form is retained. For example, with online reviews, each individual contributor writes their own review with their own interpretation of the given product, movie, travel destination; at the same time, the collection of reviews forms a more comprehensive document of people's attitudes.

- **Iterative**. In versioned aggregation, multiple contributions are used toward a larger product, but the contributions are permutations of a common work. For example, with collaboratively written wikis, such as Wikipedia, each user's iterates on the work of all the previous writers of the page.

- **Averaged**. In averaged aggregation, contributions are still pooled, but a consensus-seeking process tries to reconcile

them. Our use of *averaged* here alludes to quantified consensus seeking, even when it is not simply a case of derive a statistical mean. With contributions such as opinion ratings of information objects the process might be to average; with multiple-keyed classification, the aggregation process may be a vote majority, where the most popular option is retained; with starring (sometimes referred to as favoriting, liking, or recommending), the averaged aggregation may simply show the number of people that have performed the action.

A consideration related to aggregation is that of quality control, something other typologies have considered as a top-level dimension in its own right. Quinn and Bederson (2011) consider how the system protects against poor contributions, such as reputation systems, input or output agreement, multi-contribution redundancy, a crowd review workflow, expert review, and designs that disincentive poor quality or obstruct the ability to do so. Quinn and Bederson (2011) likewise look at quality assurance, noting the large focus on improving quality for quantifiable contributions.

In our typology, we consider quality as a best practices issue that follows from how users are aggregated. With summative aggregation, for example, quality is often pursued by a separate crowdsourcing step: allowing online visitors to flag low-quality or otherwise problematic contributions. Other times, such as with question and answer websites *Stack Overflow* or *Quora*, visitors vote on the quality of answers to surface the best ones. With iterative contributions, peer review is sometimes used, as in the versioned workflow of many open-source projects or with the concept of watching pages and reversions on Wikipedia. As noted, averaged aggregation receives a lot of focus because it lends itself to quantification, and numerous studies focus on the quality increases of adding redundant contributors or methods to identify low-quality contributors (Sheng, Provost, and Ipeirotis 2008; Snow et al. 2008; Wei and Croft 2006; B. Wallace et al. 2011).

Beneficiary / Director

Who directs the crowdsourcing activities and who benefits from the contributions?

Considering the director of a crowdsourcing task, Zwass (2010) distinguishes between *autonomous* and *sponsored* forms of crowdsourcing. *Sponsored* crowdsourcing is when there is an entity at the top soliciting the contributions: a client of sorts. In contrast, *autonomous* crowdsourcing serves the community itself. Autonomous crowdsourcing can be in a centralized location, like a community-written wiki or video-sharing website, or exist loosely, as in blogs. Zwass (2010) explains: "Marketable value is not necessarily consigned to the market—it may be placed in the commons, as is the case with Wikipedia."

Considering the soliciting party as a case of sponsorship or autonomy is useful, though a further distinction should be made between the collective (the *crowd*) and the individual (the *contributors*). Crowds collaborate toward a shared goal, as with Wikipedia or certain kinds of open-source software development, while individuals are more self-motivated. For example, in citation analysis through web links, as was done with PageRank (Page et al. 1999), the large-scale benefits of the crowds are unrelated to what the individuals creating the links are thinking. Rouse (2010) offers a similar designation in the beneficiary, between individual, crowd, and a mix of the two.

One way to view this relationship between contributor and director is in light of effort against benefit. Do both director and contributor benefit (symbiosis)? Does one benefit at the expense of the other (parasitism)? Or is it a case of commensalism, where both benefit but in mutually exclusive ways.

Centrality

How central, or necessary, is the crowdsourcing to the task at hand? Is it *peripheral*, or *core*?

The work in Organisciak (2013) tried to counterbalance a per-

ceived focus on whole-hog crowdsourcing – the large, highly novel initiatives like Wikipedia – by introducing *incidental crowdsourcing*. Incidental crowdsourcing focused on types of crowdsourcing – like rating, commenting, or tagging – that are peripheral and non-critical. The shift to an incidental mode brings with it its own design tendencies, such as lower bandwidth forms of contribution and fallback strategies for low engagement cases.

This distinction between peripheral and core is important to an information science treatment of crowdsourcing. It shows that the benefits of crowdsourcing are not only attainable by those with the infrastructure and resources to commit to a new large system. It can be an augmentative feature, that engages with users and accepts useful feedback from them in addition to a non-crowdsourcing primary objective. Peripheral crowdsourcing also often accompanies a pattern of reacting to existing information objects, pertinent to those that deal with museum repositories or digital libraries.

Common Design Patterns

A number of design patterns have been established and repeated in crowdsourcing, some organically and some, like the ESP Game, carefully engineered. These include:

Microtasking. The concept of splitting a large task into many smaller parts improves the ability for that task to be worked on by different people. Microtasking was an important tide change in the history of open-source software (Raymond 1999), and the same model has been often adopted in crowdsourcing. With so-called 'microtasks', the overhead to participation is low, and the pressure or dependence on any one contributor is low.

Gamification. Gamification is predicated on a reframing of what would traditionally be labour into game-like or leisurely tasks. Gamification follows in the philosophy, as with Tom Sawyer re-contextualizing a fence painting chore into a game, "that work consists of whatever a body is obliged to do, and that play consists

of whatever a body is not obliged to do" (Twain 1920).

Of course, Tom Sawyer used his fence painting game as a manipulation, intended to trick other children to do his work for him: an apt comparison to ethical concerns about gamification. Those defending the ethics of gamification have argued for it as an extension of contributors' desire to perform meaningful work. Shirky, for example, argues that people have a 'cognitive surplus' to give during their leisure time, a desire to spend their free time doing useful, creative or stimulating tasks (2009). Gamification is an extension of serious games – games meant to do more than simply entertain (Abt 1987; Michael and S. L. Chen 2005; Ritterfeld, Cody, and Vorderer 2010). In areas of crowdsourcing and human computation, Games with a Purpose (Ahn 2006) is an extension of serious games in the context of distributed, collaborative crowds. Harris and Srinivasan (2012) consider the applicability of applying games with a purpose to various facets of information retrieval, concluding it is a feasible approach for tasks such as term resolution, document classification, and relevance judgment. Eickhoff, Harris, et al. (2012) have investigated the gamification of relevance judgements further, augmenting the financial incentive on paid crowdsourcing platforms.

Opinion Ratings. A standard and highly familiar activity online is soliciting qualitative judgments from visitors. These ratings have different granularities, often 5-level (e.g. 1 to 5 stars) or binary (e.g. thumbs up/thumbs down). Unary judgments have grown in popularity as ways of showing support with minimal effort. Their popularity seems to stem from when social network *Friendfeed* implement a unary voting button labelled, succinctly, "I like this" (Taylor 2007) and subsequently when similar wording was adopted by Facebook after acquiring Friendfeed.

Platforms. There is a cottage industry of services that offer the infrastructure for requesters to crowdsourcing, using in domain-specific ways. For example, *Kickstarter* and *Indiegogo* ease crowdfunding, *99Designs* enables contest-based design tasks, and Mechanical Turk offers the tools and people for microtasks.

Contests. In the contest design pattern, a requester offers a bounty to the best solution to a problem or task of their choosing, such as in design (e.g. 99Designs), coding (e.g. *TopCoder*), and research and development (e.g. Innocentive). Here the "crowdsourcing" is simply using internet to connect to many potentially talented individuals, though contests have been integrated into more collaborative workflows. For example, with the collaborative product incubator Quirky, the community votes on the best ideas to develop into products, discussing how to improve the ideas openly.

Wisdom of crowds. Wisdom of the crowds, in addition to the principle referring to the effectiveness of human judgment in aggregate, also refers to a design pattern which uses that principle (Surowiecki 2004). This is embodied by multiple-keying for tasks which are expected to have a real answer, such as classifying galaxies, or averaging opinions for subjective tasks to derive a normative judgment.

Practitioner's Questions

To conclude, I offer some practical examples to provide a template for crowdsourcing planning using this typology.

Q: *Are you augmenting existing data, which already exists in an online system or repository or which is appropriate for presentation already?*

- *Yes.* Peripheral crowdsourcing is an option to consider, because it collects information from people that are already interested in the content and consuming it. *Trove* does this with newspaper scans: visitors can read the scans and the poor computer transcription, but are also given an option to fix the transcriptions.

- *No.* Core crowdsourcing requires more technical overhead, but results in some of the more interesting examples of crowdsourcing. *Old Weather* or *Transcribe Bentham* show how archives can engage with interested members of the public,

while arguably providing a strong form of material appreciation than passive reading would offer.

Q: *Does your data compile, iterate, or combine?*

- *Compile.* Summative aggregation is seen in digital history projects like *Make History*, a 9/11 Memorial Museum project compiling people's photos and stories of the 9/11 terrorist attacks. However, simpler crowdsourcing mechanics, such as commenting and tagging on Flickr's *The Commons*, also follow this pattern.

- *Iterate.* Digital archive transcription projects such as *Transcribe Bentham* work with the model of iteration. One concern with these forms of projects is that contributors sometimes do not want to conflict with a previous author; a way to encourage iteration is to mark unfinished pages and discourage single edit perfection, as is done on Wikipedia.

- *Combine.* Information science has a tradition of considering averaged aggregation in the context of multiple-coder classification. For an example of a novel, notably low-tech version of this pattern, Simon (2010) writes of voting bins at the exit of the Minnesota Historical Society's History Center. Visitors, who are given pins to show they have paid admission, can vote on their favourite exhibits by disposing of the pins in one of a set of labeled containers.

Q: *Can your data be collected while contributors work for themselves?*

- *Yes.* Social OPACs like Bibliocommons collect various user-generated metadata about materials, such as tags or comments. A study into two such systems found that the features are generally underutilized, but are most popular in cases where participants are creating things for themselves: compiling list bibliographies, personal collection bibliographies, or use a "save for later" feature (Spiteri 2011).

Q: *Does your project have any primary motivators to incentivize contributions, such as an existing community of interest or a compelling, easy to answer question?*

- *Have primary motivators.* Systems such as Galaxy Zoo or Trove provide examples of how a system can emphasize the incentives they offer to potential volunteers. Most of the successful projects noted in this study offer some primary motivators.
- *Don't have motivators.* For trickier or less intrinsically interesting data, it is possible to hire on-demand workers through a platform like *Mechanical Turk*. Examples of efficient routing on these sorts of systems include *Soylent* – crowdsourced writing assistance (Bernstein, Little, et al. 2010) – and *VizWiz*, an accessibility application that allows visually impaired users to receive transcribed descriptions of photos that they take (Bigham et al. 2010)[27].

[27] VizWiz also outsources some tasks to social networks.

Conclusion

Crowdsourcing offers potential in information science for involving the public and improving data in digital libraries and cultural heritage repositories. However, the scope of crowdsourcing is so large and the implementation possibilities so varied that it can seem rather daunting to pursue it.

This chapter attempted to provide a way of making a bit more sense of the patterns that emerge when considering these projects not so much from the perspective of what they are for (e.g., rating books, movies or restaurants versus citizen science or digital humanities) but rather in terms of how they were designed to achieve particular ends.

The typology presented consolidated a number of past taxonomies of crowdsourcing and project examples toward a view of crowdsourcing appropriate for information science. In addition to modifications on previously studied dimensions such as motivation, aggregation, and beneficiary, new dimensions were argued for, regarding centrality of crowdsourcing, the diversity needs of

the crowd, and the dichotomy of generative or reactive types of work. This typology offers a framework for making sense of the differences between crowdsourcing projects and thinking through practical possibilities for implementing crowdsourcing mechanics in new projects.

The design of a crowdsourcing activity, like any design activity is an exploration of a design space navigating goals (often multiple goals, some of which may be contradictory), and constraints, while exploiting technological and social opportunities, and taking account of certain issues such as privacy, security. For any given product, there are many experiences that could be constructed. The dimensions provided offer help in comprehending the alternatives and how they are practiced.

Within the framework introduced, the rest of this dissertation pursues crowdsourcing in the following space:

- Crowds that are paid, rather than motivated to volunteer;
- Tasks that are reactive, positioning contributions relative to existing documents, rather than generative;
- Both objective and subjective contexts, considered distinctly.

The data quality of contributions is considered in this context, starting with the next chapter: a treatment of post-collection data modeling of contributions for objective tasks.

Interpreting Tasks for Objective Needs

This chapter approaches a crucial problem: disambiguating the influence of unreliable annotators from natural uncertainty in multi-worker aggregation.[28] The accessibility of large groups of contributors online allows for large-scale annotation tasks to be completed quickly. However, it also introduces new problems of reliability by problematizing assumptions about expertise and work quality. The actual workers in these tasks are generally self-selected and unvetted, making it difficult to ascertain the reliability of the ratings.

[28] A version of this chapter was previously published at ASIS&T 2012, with co-authors Miles Efron, Katrina Fenlon, and Megan Senseney (Organisciak, Efron, et al. 2012). Copyright retained by authors.

Online annotations need to be both collected and interpreted. Where later chapters focus on issues in collection, here the post-collection interpretative stage is considered. In the absence of traditional measures of reliability, how do we know what online contributions can be trusted, and is it possible to improve their signal?

This goal is pursued for tasks with an expected truth – that is, objective tasks. However, a key assumption is made: that of a negotiated "ground truth" over an objective one. By assuming that the truth-value is a negotiated truth, worker disagreement is not in itself a sign of bad workers, but should be considered in light of the agreement among workers.

This chapter makes the following contributions:

- Description of the problem of reconciling annotation contributions or work by non-expert, semi-anonymous workers.
- Evaluation of a number of approaches for separating worker quality from rating difficulty, including dwell time, worker

experience, task difficulty, and agreement with other workers.
- Introduction of an iterative algorithm that allows task difficulty (inherent disagreement) to be disambiguated from worker reliability (i.e., synthetic disagreement).

The scope of this study is in relevance assessment for information retrieval related to a cultural heritage aggregation. Relevance assessments are a vital part of information retrieval evaluation and help in addressing the unique search challenges faced by large aggregators of cultural heritage content.

Problem

Online annotation generally takes the form of short, fragmented tasks. To capitalize on the scale and ephemerality of online users, services such as Amazon's Mechanical Turk (AMT) have emerged, which encourage the short task model as a form of on-demand human work.

AMT has shown itself useful in information retrieval, where many individual human tasks benefit from parallelized contributions. The workers are non-expert workers. However, their lack of domain or even task expertise is not inherently a problem: past studies have found that only a few parallel annotations are required to reach expert quality (Snow et al. 2008) and that increasing the amount of parallel labor per item offers diminishing returns (Novotney and Callison-Burch 2010). Training is possible on AMT, but the large workforce and transience of individual workers means that training conflicts with the cost and speed benefits of micropayment-based labor.[29]

As AMT has grown, however, the appeal of cheating has also grown.[30] The workforce, who was originally a mix between those looking to pass time and those looking to earn money, has been shifting primarily to the latter(Eickhoff and Vries 2012). Since reimbursement is done per task rather than per hour, contributors have a monetary incentive to complete tasks as quickly as possible. The site's continued growth may attract more cheaters in the

[29] The next chapter considers whether a small localized training can be effective on AMT, and whether it can be cost effective.
[30] Anecdotal impressions by Mitra, Hutto, and Gilbert (2015) suggest that this is reversing.

future, making it more important to be able to properly identify them within classification data.

Even among non-malicious workers, there is still the potential problem of varying expertise. Workers begin a task with no prior experience and grow more experienced over time. When there may be hundreds or thousands of workers, each one potentially following a learning curve, the effect of inexperience should be taken more seriously than in traditional settings with only one or a few workers. Making decisions from majority voting is quite robust in many cases. However, to safeguard against the presence of cheaters and their strengthened influence in low-consensus tasks, a less naive decision-making process may be valuable.

The problem of reconciling ground truth votes from unvetted and potentially unreliable workers is not limited to the use of Mechanical Turk. Digital libraries now have the ability to interact with their users in ways that crowdsource the creation of new content or metadata. Volunteer contributions may provide entirely new content – such as suggested labels or corrections – or feedback on existing content – such as rating the quality of an item's metadata. While unpaid engagement does not have the same financial motivation for malicious workers, contributions that are open to the public are still susceptible to low-quality results: whether through recklessness, misunderstanding, mischief, or simply spam. Furthermore, even when the ratings or annotations from unvetted semi-anonymous online workers are of a high quality, there is nonetheless a need to justify the quality of those ratings.

Related Work

As non-expert classification has become more common, there have been a number of studies into the quality of its workers. Generally, such studies have found that, while a single worker does not match the quality of an expert, aggregating votes from multiple earnest workers can match the quality of an expert.

Snow et al. (2008), found that for natural language processing

tasks, only a few redundant classifications are necessary to emulate expert quality – their task required an average of four labels. Similarly, Novotney and Callison-Burch (2010), looking at online transcription, found that the increase in quality from adding redundant annotations was small, and recommended allocation resources to collecting new data. Additionally, they noted that disagreement measures are more effective for identifying and correcting for bad workers than they are for finding good workers, due to false positives among highly ranked workers.

In understanding the role of non-expert workers, a number of studies have taken differing approaches to ranking worker reliability and dealing with noise. Some have attempted to model worker noise against gold standard labels (Hsueh, Melville, and Sindhwani 2009; Eickhoff and Vries 2012). However, more commonly, researchers look at ways to understand worker quality without the presence of ground truth data. One common approach to separate the latent variable of worker quality from task difficulty enlists the Expectation Maximization (EM) algorithm, weighing worker judgments based on past performance (Whitehill et al. 2009; Welinder and Perona 2010; Wang, Ipeirotis, and Provost 2011). The approach taken in this study is similar in principle to the EM algorithm.

Raters have been treated as a mix of good or bad, where the nature of the problem is to identify the latter for removal (Dekel and Shamir 2009). Other work has treated reliability not as an issue of replacing users, but rather of identifying low quality ratings to reinforce with additional ratings (Sheng, Provost, and Ipeirotis 2008).

One notably unique concept of user quality was the assumption by Donmez, J. Carbonell, and Schneider (2010) that the quality of workers changes over time. In other words, worker quality was considered a distribution over time, rather than an overall score. Notable about this approach is that there are no assumptions about the direction of quality change by workers, allowing them to account not only for inexperience but also for occasional patches of low quality ratings by a worker.

Alongside prior work in representing non-expert workers, research has also considered using the information for deciding on future actions. This has been considered both as an act of choosing the next tasks for a worker (B. Wallace et al. 2011; Welinder, Branson, et al. 2010), and alternately an act of choosing the next workers for a task (Donmez, J. Carbonell, and Schneider 2010).

In 2011, the Text Retrieval Conference (TREC) held a Crowdsourcing track for the first time, which dealt directly with the evaluation of search engines by non-expert workers hired through micropayment services. Teams looked at one or both of two tasks. The first task was to effectively collect high-quality relevance judgments. The second task, in line with the goals of this study, was to "compute consensus (aka 'label aggregation') over a set of individual worker labels' (Lease and Kazai 2011).

There were two evaluation sets used with the second task of the TREC Crowdsourcing Track: one of consensus labels from among all the participating teams and one of ground truth gold labels done by professional assessors. Accuracy rates – the number of properly labeled ratings divided by all ratings – spanned from 0.35 to 0.94 with a median of 0.835 against the crowdsourced consensus labels, while the runs against the gold labels spanned from 0.44 to 0.70 with a median of 0.66. In achieving these results, the ten teams used a variety of approaches, including the EM algorithm, rules-based learning models, and topic-conditional naive Bayes modeling (ibid).

When measured by accuracy, the EM algorithm was among the most prominent. The best performing team against each evaluation set – BUPT-WILDCAT and uc3m, respectively – both had an EM implementation in their submission, though the latter was paired with a number of additional rules. However, uc3m's second, non-official run slightly outperformed the accuracy of their official submission with an approach using support vector machines (SVM) (Urbano, Marrero, et al. 2011).

Data

Post-collection looks at the data quality of descriptive crowdsourcing is studied for contributions over two datasets.

In the dataset of primary focus, workers contributed judgements of the relevance of cultural heritage documents to a given query. This data was rated with three-annotator redundancy, which means that for each document, there were three workers that completed the rating task. There were three label options available to workers: *relevant*, *non-relevant*, and *I don't know*. The unknown option was considered a skipped option and the data was removed from the final dataset.

Annotations were collected through Amazon's Mechanical Turk service, using a custom rating interface. When a worker accepted a judgment task, they were shown a page with a query, description of the task, description of the coding manual (i.e., what types of documents should be rated as relevant), and up to ten ribbons of documents to rate (see Figure 1).

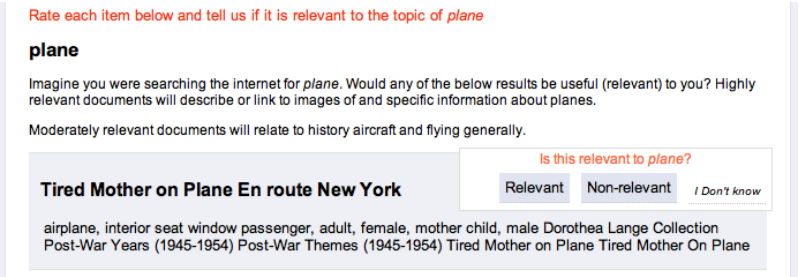

Figure 1: The rating interface.

The structured form of digital item records lends itself well to such tasks, which we represented through the title, description, and related image thumbnail. To aid the task of scrolling through ratings and decrease the time spent on tasks, our interface automatically scrolled to the next tasks once the previous one was rated.

The impetus for the chapter was improving the effectiveness of an information retrieval system for the Institute of Museum and Library Services Digital Collections and Content project (IMLS

DCC). The IMLS DCC is a large aggregation of digital cultural heritage materials from museums, libraries, and archives across the country. Originally launched in 2002 as a point of access to digital collections supported by the IMLS through National Leadership Grants and LSTA funding, it has since expanded its scope to provide more inclusive coverage of American history collections, regardless of funding source. As a result of its position among the largest cultural heritage aggregations in the US, research through the IMLS DCC looks at the problems associated with reconciling content from thousands of different providers, including metadata interoperability, collection-item relationships, and access to materials. One of the difficulties that IMLS DCC must address is information retrieval when the metadata records in its aggregation are of inconsistent length, style, and informativeness. Overcoming these types of problems in order to improve subject access to the breadth of materials is an active problem (e.g., Efron, Organisciak, and Fenlon 2011; Efron, Organisciak, and Fenlon 2012). In doing so, human relevance ratings are an invaluable resource for evaluating document relevance in a given query.

Most of the evaluations are measured through accuracy, which is percentage of correct classifications that are made:

$$accuracy = \frac{\#of correct classifications}{total \#of classifications}$$

There are two comparison sets of data by which 'correct' classifications were taken. The first was against consensus labels, which were simply generated by taking the majority vote for a given task. Since these are derived from the actual dataset, they may not be completely reliable. However, for comparative purposes, they offer us a metric by which to see trends in the data.

The cleaner set of ground truth data is a set of oracle ratings done by myself and the authors of Organisciak, Efron, et al. (2012). Since the authors are of known reliability and have a close understanding of both the rating task and the data being rated, the oracle judgments serve as an effective measure for evaluating the accuracy of the majority votes themselves.

Research Questions

In the context of paid crowdsourcing, this study looks to simultaneously interrogate worker quality and task difficulty, allowing the estimates of one to inform the estimates of the other.

The intention is to better understand post-collection strategies to improve data quality, pursued through three areas:

- *Temporality*.

  RQ 1.1: Does the length of time that a worker spends on a question reflect the quality of their rating?

- *Experience*.

  RQ 1.2: Do workers grow more reliable over time?

  Can you account for the rating distribution given the worker's experience? In this study, tasks are grouped topically, by "queries". Workers were asked 'is this metadata record relevant to Query X' or 'what is the tone of Query X?' Subsequently, how a worker's experience with a query affects their performance was looked at.

- *Agreement*.

  RQ 1.3: Does a worker's agreement or disagreement with other workers reflect their overall quality as a worker?

  RQ 1.4: If so, can disagreement be used for data improvements?

Approach

The documents in the rating tasks were brief collection metadata documents, which workers annotated according to their relevance to a given query. Workers contributed ratings ten items at a time. The task set size was chosen for two reasons. First, this allowed for less time loading and adjusting to new tasks. If there was a learning curve for each query – as this study finds to be present, albeit minor – it seemed sensible to allow workers some time to rate

once they grew comfortable with a task. The second reason was to create a minimum usable profile of a worker's performance, which would have been difficult with fewer tasks. Note that not all sets of tasks had ten items, as our system would track tasks that were completed or in progress, serving fewer than ten when ten were not available.

Originally 17700 data points were collected, though this was later increased to just under 23000. The average amount of time spent on each individual item was 4.8 seconds, with half of all ratings being done in less than 1.8 seconds and full rating sets being completed in an average time of 37.3 seconds.

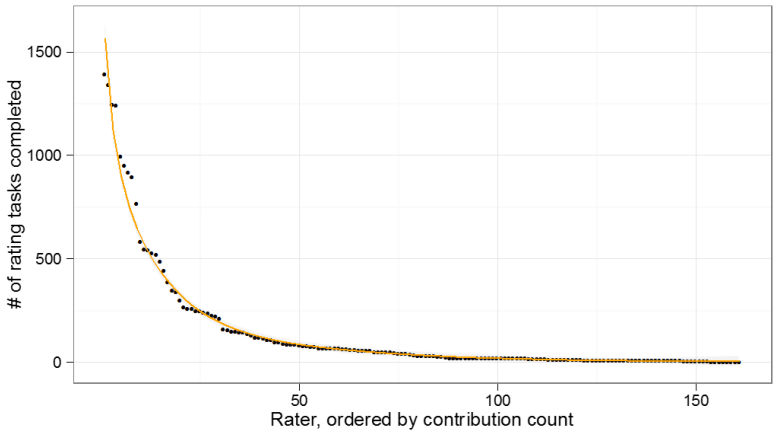

Figure 2: The number of ratings contributed per rater, roughly following a power-law distribution.

There were 157 unique workers that contributed ratings, rating an average of 141.9 tasks. The most dedicated worker completed a total of 1404 ratings. The distribution for contribution count resembles an inverse power law, a distribution commonly seen among contributions from online users (see Figure 2).

For comparison with other tasks, a second dataset was also analyzed, in which workers classified the tone of a number of political tweets. This Twitter sentiment dataset included more classification options - workers rated the tweet as having positive, negative, or neutral tone or whether it was incoherent or spam.

For both the primary and secondary datasets, there was an accompanying set of ground truth oracle judgments. These were

used for evaluation.

Temporality

> **RQ 1.1**: Does the length of time that a worker spends on a question reflect the quality of their rating?

Among the statistics collected for the relevance judgment dataset was *dwell time*: the time spent on each rating. The hypothesis motivating this metric was that dwell time was not significant when understood independently, but might indicate the quality of workers when taking into account the order in which tasks were completed. Since tasks were done in sets of ten, the order referred to where in this set they occurred.

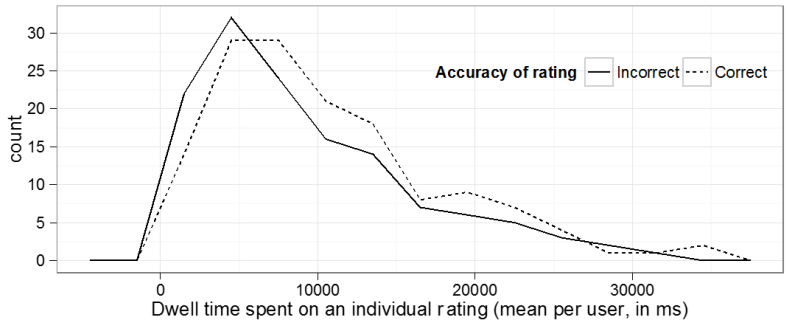

Figure 3: Frequency distribution of the average amount of time that users spent on the tasks they rated incorrectly and those they rated correctly.

Order served as a useful grouping factor because the time spent on the first rating is confounded with the time spent reading the rating instructions, which is to say that the two are inseparable. Figure 3 shows the distribution of worker performance by dwell time alone. As expected, a correct classification does tend to take slightly more time, but there is not enough evidence to reject the null hypothesis of equal distributions. Thus, for this the setting of cultural heritage retrieval relevance judgments, dwell time alone is insignificant to performance (Wilcoxon rank sum $p = 0.064$; $p = 0.154$ when excluding extreme outliers).

However, dwell time considered alongside the order or task completion (i.e., how much time was spent on the first task? On the second?) tells a more complete story.

Consider first the amount of time that is spent on each n^{th} task. Pairwise Wilcoxon Rank Sum tests show that the amount of time spent on the first rating in a set is significantly different from all other ratings ($p < 0.001$, with Bonferroni adjustment), as were all pairwise comparisons with the second rating in a set ($p = 0.02$ vs order 3, $p < 0.001$ vs all others; Bonferroni adjustment). Notably, however, we fail to reject the null hypothesis for all other ratings in a set.

This means that there is extremely little difference in time spent between a worker's third and tenth ratings, as well as all comparisons in between. This is more abrupt than the gradual decline that was expected, and suggests that the learning curve for a worker to rate comfortably is only the first two ratings.

Comparing the accuracy of ratings by dwell time, the time spent on the first rating of a set is significantly higher for ratings that are correct than those that are incorrect (Wilcoxon Rank Sum one-sided $p = 0.01$). This stands in contrast to every rating after the first one, none of which show significant difference in time spent on ratings that are true and ones that are false.

The measurement of dwell time for the first item in an item set is confounding with the readying of instructions[31] The fact that a worker spending more time on the first rating indicates a higher likelihood of correctness suggests that there is a latent effect in how closely people read the description,

[31] Which is to say, we had no measurement for when a worker's attentions turn away from the background material at the start of a task toward the first task in the set of ten.

If this is in fact what accounts for the significant different, it should be an effect that lingers across the full set of data.

Figure 4 shows this to be the case, with workers that make a correct rating on the first item are much more reliable in the rest of the rating set.

As part of the rating instructions, workers were presented with a description of what types of results are relevant to the given query (see screenshot in Figure 2). If a worker does not read this section carefully, their ratings would be more interpretive, possibly resulting in inconsistencies with workers that followed the instructions more carefully.

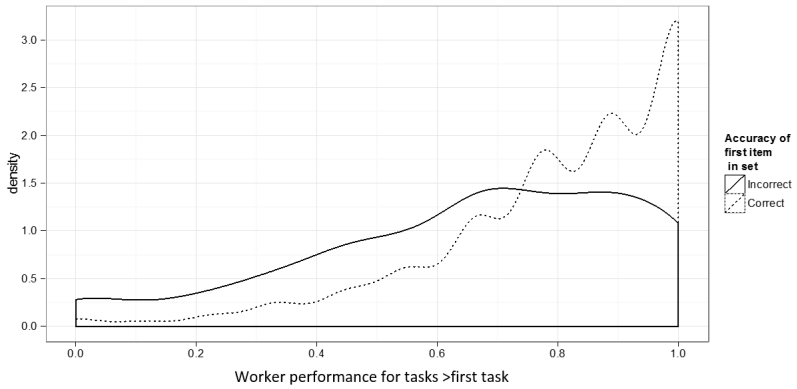

Figure 4: Accuracy of workers on ratings after the first one, shown as density distribution of all workers. Dotted line shows workers that are correct on the first task, solid line shows workers that are incorrect.

Answer: *RQ 1.1*

Answer (RQ1.1): the amount of time that a user spends on each task is not, by itself, an indicator of a quality contribution. However, workers that spend more time at the start of a task set, particularly before their first contribution, *are shown to perform better.*

Experience

RQ 1.2: Do workers grow more reliable over time?

An extension of the order grouping factor, the next factor considered was the long-term experience of a worker. Experience was looked at in two forms: lifetime experience and query experience.

Lifetime experience is the overall number of tasks that a worker has completed.[32] Is a worker's 100th task more likely to be correct than their first task? The hypothesis motivating this was that over time workers would grow more reliable. However, this hypothesis proved to be incorrect.

[32] *Lifetime* refers to the task type, such as *all relevance judgments*, not all tasks completed on the platform.

Lifetime experience was not an indicator of contribution quality. Plotting makes the case emphatically: Figure 5 shows the distribution of ratings across lifetime experience. Each point represents the percentage of the n^{th} ratings that were correctly rated. If a point at position x shows an accuracy of 0.80, this means that 80% of tasks which were workers' x^{th} rating agreed with the majority, our estimated value for the correct label. As is apparent, there is no trend

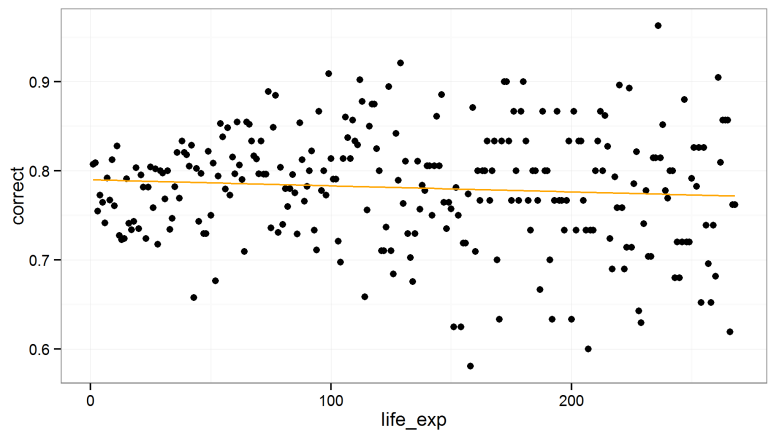

Figure 5: Average accuracy of workers' nth contribution overall.

with increased lifetime experience, or it is confounded by other things.

The second measure of experience, query experience, refers to the number of tasks that a worker has completed within a single topical domain. In information retrieval relevance judgments, workers are asked to judge whether a document is relevant to a given query; thus, the query experience. Similarly, in the secondary dataset of Twitter sentiment ratings, workers were asked to annotate the opinion of the tweet regarding a given topic; i.e., what is the sentiment toward entity Q.

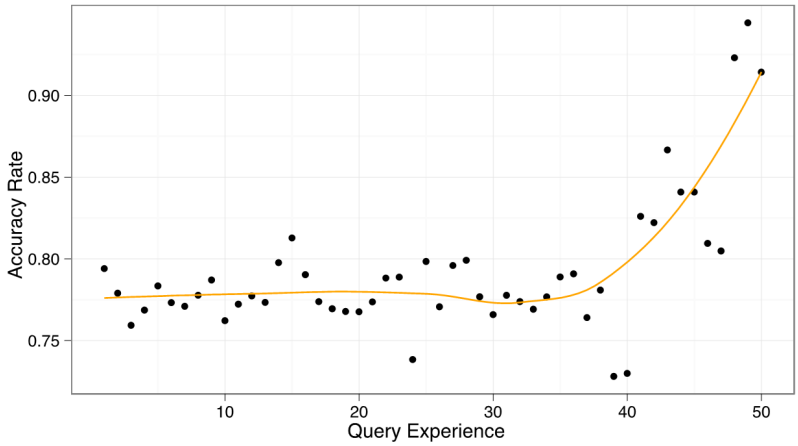

Figure 6: Average accuracy of workers' nth contribution with a query. Only points aggregating 20 or more workers are shown.

Query experience proved to be an indicator of worker qual-

ity among the most experienced users, but not notable otherwise (Figure 6). For approximately the first thirty tasks which workers completed with a single query, they did not demonstrate any meaningful difference in quality. However, ratings beyond that point showed a sharp increase in quality. What is unclear, is whether this is an effect of improvement through experience, or self-selection by better workers. Regardless, for the purposes of determining what information to trust from a data perspective, this distinction is not greatly important.

Answer: *RQ 1.2*

> **Answer (RQ 1.2)**: Workers do not appear to improve with practice for the type of task studied. The sole exception is the most experienced experienced workers, which may be a factor related to the self-selection of which workers stay around for that long. As noted above, it only took about two ratings for most workers to get into the groove of relevance judgments.

Worker Agreement

> **RQ 1.3**: Does a worker's agreement or disagreement with other workers reflect their overall quality as a worker?
>
> **RQ 1.4**: Can disagreement be used for data improvements?

Finally, in addition to worker experience and time spent per tasks, this chapter looks at the ability of worker agreement and task difficulty to discern the accuracy of ratings. The reason that these were considered together is that they are invariably confounded: a task has as little as three ratings informing any estimates of the quality, and those ratings are each biased by the quality of the specific workers involved. There were two approaches looked at: identifying and replacing low quality workers, and an iterative algorithm for weighing workers and tasks.

Replacing Problem Workers

One of the immediate problems with our primary data was a low-worker agreement (Fleiss' Kappa = 0.264). In our first attempt to improve the agreement between workers, we identified low-quality workers and replaced their contributions. First, a confusion

matrix was calculated for all workers and an accuracy rate was taken as a measure of a worker's reliability. Raters below a certain threshold were removed and new workers replaced their ratings. The threshold chosen was 0.67, meaning workers whose ratings agreed with their co-annotators on a task less than two-thirds of the time were removed.

The threshold for removing workers was supported by a simulation where an undiscerning worker was emulated, replacing randomly selected classifications in the data with its own random ratings. While a worker in an environment completely populated by random workers would be in the majority two-thirds of the time, inserting random workers alongside the real workers in the data provides a more realistic estimate. Across 100 runs, the mean accuracy rate of the random worker was 0.680, with a median of 0.677 and standard deviation of 0.080. In other words, the workers whose data was removed – with an accuracy less than 67% – were less likely to be in the majority opinion on a rating than a randomized bot. This accuracy rate also puts our data in perspective, falling somewhere between the 0.75 agreement that would be expected of a random worker in a completely random triple-redundancy labeling system and the 0.50 agreement expected of a random worker in an ideal human setting with all other workers agreeing on ratings.

There were 23 workers below or at the threshold that were removed, accounting for 2377 ratings (17.7% of the data). Notably, there were 10 workers with a total of 1069 ratings that had accuracy rates right at the threshold, meaning that nearly half of removed ratings would not have been taken out with a slightly lower threshold.

After removing problem workers, there was an increase in kappa score from 0.264 to 0.358. The increase in intercoder agreement is expected, given that our metric for problematic workers is how much they agreed with other workers. However, since these workers were by definition already in the minority much of the time, their influence on actual votes was not high. Thus, the as-

sumption to test is whether, when low-agreement workers do end up in the majority, they cause damage by influencing votes in the wrong direction.

In fact, the negative quality impact of problem workers proved to be very small. The accuracy rate of final votes after replacing them increased from 0.856 to 0.862.

An alternative to selective replacement of problem workers is selective redundancy. Rather than removing data, one can take the approach of adding more labels, as encouraged by Sheng, Provost, and Ipeirotis (2008). This approach resulted in an increase to 0.859, a smaller increase than that of removing problem workers. In other words, majority rating proved fairly efficient at smoothing over individual bad workers, limiting their influence.

In order to further increase worker agreement, one could presumably run the replacement process again. However, when non-expert labels are being paid for, removing problematic workers can grow quite costly – especially given the low payoff in accuracy. A cheating or sloppy worker can also rate many ratings quickly, making the potential lost profit even higher. However, the removal and blocking of low-agreement workers can be automated fairly easily, making it possible to incorporate in real time within a rating interface.

Why were some workers correct – or at least in the majority opinion of what a correct rating is – less than chance? One possibility is sincere workers misunderstanding the task. Wang, Ipeirotis, and Provost (2011) refer to such situations as recoverable error and offer a method for identifying consistently incorrect workers and correcting their votes. In the case of binary data such as our relevance judgments, this would simply mean inverting relevant votes to non-relevant, and vice-versa. However, none of the workers in our data would improve with such an approach, and it seems like an unlikely occurrence for a worker to make such a drastic mistake systematically. However, it is possible that less drastic misinterpretations can lead to problems with difficult tasks due to misunderstanding the delineation between categories. As

we found in our tests on dwell time, workers that appear to spend less time on instructions tend to make more errors: perhaps subtle misunderstandings can lead to consistently poor performance.

Iterative Optimization

While removing workers based on their error rate has a positive effect on the data, it does not take into account the difficulty of the task that is being completed by the worker. If a worker has the misfortune of being assigned a particularly difficult or debatable set of results to rate, their error rate may prove to be quite high. More unfortunate still, a worker may be rating alongside numerous low quality workers. If two low quality workers disagree with one higher quality worker, the dissenting worker's reliability should reflect the circumstances. There may be latent variables that are not accounted by our system which adversely affect the data.

To account for unknown variables and separate out signal from noise, an iterative algorithm was developed to simultaneously weigh worker votes and the difficulty of the task. In line with the purpose of this study, this approach allows one to not only evaluate workers, but to separate out the effect of the task itself.

The algorithm iterates over two steps.

In the first step, an expected truth for each document is calculated, given the information that is available about that document, the possible labels for that document, and the workers evaluating that document. Early on, that information is limited: while it is known how often each option was chosen for each document rating and how often each worker used each option, there is no information about the quality of the ratings or the workers making them.

In the second stage of the algorithm, the assigned labels of the expected votes are used to update the parameters that are used in step one. This involves assigning values of confidence for the results and determining worker quality based on that confidence

value. After this stage, the algorithm was iterated again, returning to the first stage for voting on the expected true value.

This algorithm converges or approaches a convergence limit after a number of iterations. The number of iterations that are required before the data converges varies, but only a few are generally needed for relatively simple data such as information retrieval relevance judgments.

The anticipated benefit to this approach is that worker quality is less dependent on circumstance. Consider the following scenarios:

- A worker makes a dissenting rating on a difficult task. To form an opinion only on whether they agreed or disagreed with other workers would be unfair to this worker and possibly remove authority from a good worker. For example, in an instance with five workers, there is a difference in whether a worker is the lone dissenter against a majority of four or one of two dissenters against a majority of three. In the latter case, there is a more uncertainty in what the truth value really is. Unfortunately, this scenario is limited for instances with only two categories and three workers, such as a large portion of this study's relevance judgment dataset.
- A cheating worker is correct by chance. As the earlier simulation found, a random voting worker will be correct 67% of the time in the relevance judgment dataset. By weighing this worker's vote according to their overall reliability, their votes, even if correct, will hold less sway. By setting their reliability score based on the confidence in their ratings, their influence will be even lower in subsequent iterations.

For confidence scores $C_i \in C_{i1}, C_{i2}, \ldots C_{il}$ where l is a set of all possible labels – $L \in 0, 1$ for the cultural heritage relevance judgements and $L \in 0, 1, 2, 3, 4$ for the Twitter sentiment ratings – the truth value vote is always chosen as the highest confidence label:

$V_i = \max jC_i$

As the vote can change in subsequent iterations, it is a *soft label*.

Since voting is always done on the highest confidence label, a number of methods were evaluated for assigning a confidence value to a rating. For calculating vote confidence, we looked at the following approaches:

- Probability of worker agreement for label j of rating task i. This approach represents simple majority voting and was used for comparison. It counts the number of i category labels, $|l_i|$, and divides it by the total number of labels received by the task:

  $C_{ij} = \frac{|l_{ij}|}{|l_i|}$

  Due to the lack of worker influence in the expression, this does not require iteration, as it will not change.

- Probability of worker agreement for task i given a worker of quality U. This approach, taken before in Sheng, Provost, and Ipeirotis (2008), weighs confidence C according to the mean worker reliability scores of the workers choosing each label:

  $C_{ij} = \sum_j U_{ij} \frac{|l_{ij}|}{|l_i|}$

- A weighted ranking function previously described in Organisciak (2012). This heuristically-determined approach accounts for higher numbers of redundant workers, while also offering diminishing returns on each worker added.

  $C_{ij} = log(1 + |l_{ij}| * \prod_{k=1}^{|l_i|} \frac{|l_i|}{|l_i| + |l_{ik}| * U_{ik}})$

In addition to task confidence, numerous approaches were evaluated for weighing worker scores. The basic approach is to use the mean confidence for every single rating that a worker has made before. However, there are two problems with doing so. First, since task confidence is bound between zero and one, setting workers' scores based on confidence alone will result in continuously declining worker reliability scores between iterations, without any sort of convergence. Such an inequality would also be unevenly distributed, algorithmically punishing workers with more completed tasks. Secondly, since a random worker has an average accuracy of 0.67 in our dataset, the range between good and

bad workers is small and skewed upward, making it ineffective for weighing votes. Ideally, on a task where two theoretical cheaters disagree with a near-perfect worker, an early iteration should flip the vote in favor of the better voter.

Rater quality was weighed in the following ways:

- Exponential decay. Reliability scores are calculated by the mean confidence of a worker's tasks and then raised exponential, to the power of two or three, depending on how aggressively the algorithm's confidence weighting is. A decay function can disrupt an algorithm's convergence and requires lower boundaries.

- Reliability score normalization. The mean of all reliability scores is normalized to a value of 1.0. This weighting is calculated as the sum of all reliability scores divided by the number of workers:

$$U_i = U_i \frac{1}{|U|} \sum_j U_j$$

- Relative scoring. Reliability scores are calculated on the confidence of their ratings relative to the highest rating in each set.

For comparison, we also ran a worker reliability scoring function as described in Wang, Ipeirotis, and Provost (2011), which is based on the accuracy rate of the workers (i.e., how many they rated correctly compared to incorrectly) without any weight given to the confidence in the tasks that they completed. The various techniques for calculating confidence and setting worker reliability scores were combined in sensible ways and evaluated.

Accuracy rates were recorded for the number of correct labels applied at the fifth iteration of each algorithm.

Robustness was also tested, by emulating malicious workers. Bots replaced random workers' ratings with their own undiscerning ratings. The false ratings consisted of 5% of the data and were used to see whether there were differences in how the algorithms handled clear cheaters.

The algorithm combinations were as follows:

Majority: The baseline vote based on majority labels.

Basic Algorithm: Described by Wang, Ipeirotis, and Provost (2011). Confidence is weighed by worker reliability, and worker reliability is dependent on basic accuracy rate.

Basic with Reliability Decay: Modification of basic algorithm, with exponential worker reliability score decay.

Regular with Reliability Decay / Normalized / Relative Scoring: Confidence is weighed by worker reliability, and worker reliability is weighed in one of the ways introduced above.

Alternate Algorithm: Confidence is calculated using the approach previously described in Organisciak (2012).

Table 4: Accuracy rates of iterative algorithms on different datasets. All iterated data shown at 5th iteration.

	IMLS DCC Relevance Ratings	Twitter Sentiment	IMLS DCC Relevance Ratings w/ cheater
Majority Vote (baseline, no iteration)	0.8573	0.5618	0.8479
Basic Algorithm (Sheng, Provost, and Ipeirotis 2008)	0.8590	0.5876	0.8494
Basic w/ Reliability Decay	0.8669	0.6082	0.8605
Regular w/ Reliability Decay	0.8590	0.5979	0.8557
Regular w/ Reliability Normalization	0.8590	0.5876	0.8494
Regular w/ Relative Reliability	0.8621	0.5825	0.8479
Alt. Algorithm	0.8637	0.5928	0.8510

Table 4 displays the accuracy ratings for the various runs. This can inform a number of observations.

Again, the majority voting appears to be quite effective. Con-

sider the baseline majority accuracy of 0.8573 in comparison to the similar task of relevance judgment in the TREC Crowdsourcing Track, where the best algorithms peaked at 0.70 accuracy (Lease and Kazai 2011) of the gold label set, and it becomes clear that our dataset is fairly clean from the start. The effectiveness of the baseline majority vote for the primary data is also accentuated by the relatively small gains in accuracy that is gained by the algorithm combinations.

In contrast, the Twitter sentiment dataset has a much lower baseline. The bandwidth of contribution with this data is considerably more spread out — where with the binary categories the worst case scenario for three workers is agreement between only two, the five-category Twitter data can result in no agreement. With the Twitter data, workers also showed an aversion to administrative categories: when the oracle worker would rate a message as "spam" or "incoherent", the online workers avoided doing so. In our IMLS DCC data, this worker coyness was seen with the "I don't know" ratings, but those were treated as missing data and removed.

For its lower starting point accuracy, the Twitter data showed greater improvements in accuracy with the iterative algorithms than the relevance ratings. Similarly, the iterative algorithms proved more robust against the cheaters that were artificially inserted into the data. This seems to point to their usefulness with particularly problematic data.

The iterative algorithms did not have the same effects, however. Notably, the basic algorithm with an exponential decay performed better than expected. This algorithm weighs voting according to worker reliability scores, but rather than weighing worker reliability by the confidence in the rating that the worker makes, it simply uses the worker's accuracy rate. By applying an exponential decay to the worker reliability scores, it gives the generally conservative algorithm more power to pull down low quality workers. Still, one possibility for this surprising result is that it is not as aggressive in separating out workers as the other versions. A future direction

worth exploring would be a deeper look into the individual votes that flip or do not flip with these algorithms, and how often good votes are accidentally overturned.

Investigating an iterative algorithm for optimizing worker quality and task difficulty, we found that it held limited usefulness for three-annotator two-category annotation. This is likely due to the limited amount of variance allowed by the structure. There are only two states of majority opinion –three-annotator consensus or a two agree/one disagree– meaning that when a worker disagrees there is little information on whether it is because they are a bad worker or because it is inherently a difficulty to agree-upon tasks. More information can become available by including more categories or increasing the number of workers. However, including more workers also has a positive effect on quality. Thus, the experience of this study is that for binary labels, majority rating is generally robust enough.

Answer: *RQ1.3, RQ1.4*

> **Answer (RQ1.3, RQ1.4)**: Agreement does appear to indicate quality for objective tasks. However, while removing high disagreement workers improves, well, measures of agreement, for low granularity tasks like relevance judgments it more fruitful to collect multiple independent contributions rather that seeking to punish the black sheep workers. Still, for more complex data like the Twitter sentiment ratings, correcting judgments based on measures of a worker's quality (by proxy of agreement) is effective.

Conclusion

This study looked at the growth of online annotation microtasks and the problems of non-expert workers, looking at indicators of performance among non-expert performance.

Most significantly, it was found that workers who spend more time on the first rating of a task set are significantly better performers on that task. This points to a latent variable in the instructions of the task. Indeed, the effect of extra time on the first rating seems to follow throughout a task, and *annotators that are correct on*

the first task are more likely to be correct on subsequent tasks in a set.

We also looked at the effect of experience on a worker. Generally, the amount of overall rating experience a worker had at the point of a rating did not reflect on their quality. However, a worker's *query* experience does result in better performance, though after some time.

Finally, this study looked at agreement as an indicator of worker quality. For simple tasks, there is a notable robustness in the basic agreement measure of whether a worker is in the majority opinion of a multi-annotator annotation. For more complex tasks or noisier data, an iterative algorithm can offer slight improvements on majority opinion.

Just because there is disagreement does not mean that the data is problematic, however. It was found that high disagreement among non-expert workers is not necessarily indicative of problematic results. Low intercoder agreement may indicate a difficult task or individual rogue workers. While intercoder agreement can be increased significantly by replacing the work of low quality workers, the improvement in accuracy is less defined.

These results shed light on the characteristics of workers on a simple relevance judgment tasks. However, the most interesting finding seems to suggest the importance of a worker's time spent internalizing the codebook. Are there ways to encourage this sort of behavior? What other changes can we make through collection-time tweaks? The next chapter shifts focus to these issues.

Designing Tasks for Objective Needs

This chapter investigates how the design of crowdsourcing tasks for collecting useful metadata for information retrieval metadata affects the quality of the content.

Humans don't operate with the formality of computers. Many of the benefits of crowdsourcing follow from that fact: human contributions are valuable specifically because they are not easily automated. However, when using crowd contributions to inform an algorithmic system, as in information retrieval, the inconsistencies of human work present a challenge.

In a controlled set up, crowdsourcing usually follows a common design: a task, description, and a set of one or more documents that are reacted to. This type of design is common for creating custom evaluation datasets through relevance judgments (Alonso, Rose, and Stewart 2008), but has been used for encoding and verifying indexing information (e.g., E. Chen and Jain 2013).

Evidence suggests that the design of a data collection interface affects the quality and distribution of user contributions (Alonso, Rose, and Stewart 2008; Howe 2008; Mason and Watts 2010; Mitra, Hutto, and Gilbert 2015). However, the manner to improve on a basic task/description/items interface design is not immediately clear, a problem that this chapter seeks to address.

If we consider crowdsourcing data quality as something that can be addressed not only through post-collection modeling but through the choices made in designing the collection task, the latter approach is surely the lesser studied problem. However, in

cost-time considerations, design promises more efficient improvements. A design that is more interesting to workers or less prone to error may result in better contributions at no extra cost, while designs that offer bonuses or training include short-term costs. For example, Mason and Watts (2010) found that a small change in instrumentation – changing remuneration to less tightly govern the task – resulted in more work contributed with happier contributors.

This chapter looks at collection-time task design manipulations or interventions for collecting objective data through paid crowdsourcing. Multiple interfaces for encouraging less deviation between contributors are evaluated against identical controlled tasks. Two of these design manipulations are intended to slow down workers and make them aware of how their perception of the task deviates from the standard. A third design encourages quicker responses.[33] These are compared to a realistic baseline interface which follows Mechanical Turk conventions and best practices. These design manipulations are measured against two control tasks: image tagging and image-based information retrieval relevance judgments.

[33] Why are these design manipulations chosen? Later in this chapter, various possibilities for design manipulation are considered, and compared to the existing literature.

It is found that training interventions improve collected contribution quality, and performance feedback improves quality in certain circumstances.

Afterward, an applied experiment is presented, where both a priori and posterior collection quality optimization methods are applied to a music information retrieval evaluation. The design changes that are made improve the quality of results drastically, with negligible cost differences.

Related Work

Grady and Lease previously explored the effect of changing human factors on information retrieval relevance judging through Mechanical Turk (2010). They considered four factors: terminology, base pay, offered bonus, and query wording. Their findings

were inconclusive; however their study provides guidance on the issues related to this form of study.

The effect of wording and terminology, one of Grady and Lease's focal points, has often been alluded to as a factor in crowdsourcing, including in Library and Information Science work. In writing about The Commons, a successful museum crowdsourcing project with Flickr, the Library of Congress reported that the "text announcing the Commons ('This is for the good of humanity, dude!!') struck just the right chord" (Springer et al. 2008).

Grady and Lease's work is in the space of parameterization studies, that look at how changes to the parameters of typical tasks – title, description, payment – affect the outcome. Another notable parameterization study, by Mason and Watts (2010), contributed insights on the relationship between payment and worker satisfaction. In one experiment, they found that increased payment does not improve the quality of results, only the duration of engagement. This was attributed to an anchoring effect where, in paying workers more, a worker's perception of the task's value increased with the payment. In a second experiment, they found that tying payment too closely to a task, in this case paying by word in a word search, lowered the intrinsic motivation and satisfaction of workers. This chapter builds upon parameterization studies to evaluate slightly more drastic deviations for the typical structure of a paid crowdsourcing task.

Alonso and Baeza-Yates have also written about the effect of different parameterizations of paid crowdsourcing tasks, considering the quality of relevance judgments with varying numbers of contributors evaluation each task, topics per task, and documents per query. In doing so, they cite interface design as the most important part of experimental design on Mechanical Turk and recommend following survey design guidelines and provided clear, colloquial instructions (Alonso and Baeza-Yates 2011). This study agrees with their sentiment, and strives to formally understand and articulate the differences that interface design influences in crowdsourcing.

In the TREC crowdsourcing track (Lease and Kazai 2011;

Smucker, Kazai, and Lease 2012), much much of the focus was on identifying and accounting for lower quality workers. However, there were also some efforts which built novel interfaces to try to streamline contributions or increase reliability. For example, the Glasgow team encourage fast turnaround, reducing rating click counts, pre-loading pages, and floating the assessment question (McCreadie, Macdonald, Santos, et al. 2011). Earlier, the same team crowdsourced judgments for the TREC Blog track with a design that color coded completed tasks based on whether they matched other raters and a gold standard (McCreadie, Macdonald, Santos, et al. 2011).

One novel approach to information retrieval evaluation was performed by Eickhoff, Harris, et al. (2012). They created a game for collecting labels, finding that it resulted in more contributions for substantially less cost. Akin to Mason and Watts (2010)'s word search in the paid-by-task condition, Eickhoff, Harris, et al. (2012) found that workers continued playing even after they the completed the required portion of the task. This chapter does not study the use of games, but uses elements of gamification indirectly in studying the effect of communication quantified worker performance.

The collection-time design problem has been previously pursued by Mitra, Hutto, and Gilbert (2015), who looked at "person-oriented strategies" over "process-oriented strategies". Their study is a unique precedent for a controlled experiment of different collection-time strategies for paid crowdsourcing contributions. They consider the following strategies: (1) screening workers, (2) providing examples and training workers, (3) offering financial incentives for improved quality (bonuses), and (4) aggregating or filtering multiple independent workers.

Screening (1) and aggregation (4) are strategies discussed further and employed respectively in the previous and next chapter, and performance bonuses (3) are a parameterization manipulation that has been studied before[34]. Pertinent to this study, however, Mitra, Hutto, and Gilbert (2015) found that training contributors on task expectations improved contribution quality on nearly all tasks, generally compounding improvements on top of other conditions. Similar to this study, Mitra, Hutto, and Gilbert (2015) compare interaction against a set of tasks that range in their subjectivity.

[34] We also used a performance-based bonus in the 'taste-grokking' personalization approach detailed in a later chapter (*Designing Tasks for Subjective Needs*). Though we hypothesized it may have a self-competitive effect, it was not the focus of that study and a controlled comparison was not performed to see if it was exerted an inordinate bias on the results. Mitra, Hutto, and Gilbert (2015) did not find this type of incentive to improve quality.

Finally, Kazai et al. (2011) approach the problem of HIT design quality improvement by inputting various trap mechanisms for inattentive workers. Tasks were completed by a survey flow, where the set of questions to be answered depended on the answers, allowing peculiar flows to be filtered. Captchas were also used to confirm human input and, perhaps most amusing, questions were planted that had workers check a box if they "did not pay attention" or "did not read the instructions". This is a direct solution to some problems this chapter looks at. Kazai et al. (2011) tested a number of confounded features, but these quality control metrics appeared to improve worker agreement with gold standard data.

Research Questions

This chapter compares the effect of task design on the quality of crowdsourced objective data. Scoped to a reasonable parameterization of crowdsourcing as it is commonly practiced in information science – a typical encoding task performed by paid crowds – the following questions are pursued:

- **RQ 2.1**: Which approaches to collection interface design are worth pursuing as alternatives to the basic designs commonly employed in paid crowdsourcing?[35]
- **RQ 2.2**: Is there a significant difference in the quality of crowd contributions for the same task collected through different collection interfaces?[36]

[35] The *design space* question.

[36] The primary *data quality* question.

- **RQ 2.3**: Is there a qualitative difference in contributor satisfaction across different interfaces for the same task?[37]

[37] The secondary *satisfaction* question.

RQ 2.1 is the question of design, on synthesizing prior work and brainstorming directions to explore. It is a partially subjective question, but one still worth pursuing with diligence. As research by Komarov, Reinecke, and Gajos (2013) found, the effects seen in traditional user studies are still present in online crowd markets. Their finding suggests that non-crowdsourcing research in human-computer interaction is informative for our purposes. This chapter explores some possible design decisions and argues why they should be studied.

RQ 2.2 and RQ 2.3 are the primary questions being explored in this chapter of the proposed dissertation, on quality for computational use and on satisfaction. While this dissertation is explicitly pursuing the former question, collecting computationally useful contributions needs to be understood in the context of contributor satisfaction. The trade-off between contributions that crowds want to make and the reliability of the data is a central consideration for fostering sustainable, or alternately affordable, crowdsourcing.

Design Space

Commonly, a paid crowdsourcing worker goes through the following steps:

1. Worker w arrives at task page
2. w is shown a preview of task t
3. Worker w accepts the task t
4. Work performs task t and submits
5. A new task t' is chosen and, worker is taken back to *step 2* or *step 3*

The above steps are the model used by Amazon Mechanical Turk when a task is followed through to completion. Workers are also given escape options, to skip, reject or return tasks.

Metadata encoding tasks generally consist of the following parts:

- **Goal** statement/question (e.g., "Is this page relevant to query q?", "Find the topic of a tweet.")
- **Instructions** for performing the task.
- One or more **Items** that worker responds to (e.g., webpage snippets, microblogging messages).
- **Action**, one per item: the data collection mechanism.

Within this framework, a number of factors are observable that may potentially affect how our microblog encoding task is completed. First are the parameterizations of the task within its existing structure.

The *task* can be modified, changing parts such as payment (e.g., Mason and Watts 2010), bonuses (e.g., Grady and Lease 2010), or number of tasks available.

The *goal* can also be modified; as discussed in the introductory chapter, worker and system goals can and often do vary.

The *instructions* can change: changes can be made to the clarity, the restrictiveness or open-endedness, or the length.

The *item* can change, such as modifying the presentation or the size of the single assignment set.

Even the *contribution* action can change: for example, the granularity of the contribution mechanism (e.g., free text, multiple choice, single button).

There are also harder to qualify elements such as the appeal of the topic and the visual layout.

Of course, there is no constraint insisting on the task structure provided above. We can add elements to the task design before the task is accepted, at the start of the task, during or in response to individual interactions, or after the task is completed. Taking away elements might also be possible, such as the instructions, though it is hard to imagine that doing so would have a positive effect on the reliability or variance of the data.

There are countless possibilities for adding parts to the basic task. To inspire useful ones, it is helpful to consider one final, naturalistic set of factors that may affect the outcome of a paid

crowdsourcing task: worker behaviours.

A worker's contribution may be affected by factors such as experience, skill, time spent per task, and attentiveness. Which of these can be influenced by external factors?

- *Experience*. Experience is a product of sustained interaction with the current type of task. It can be affected indirectly by focusing on methods to extend the length of a user's interaction, such as bonus payments for staying around.
- *Skill*. Skill is developed over time and is mostly affected by factors internal to the worker. To the extent that we could affect it, most functionality would encourage greater experience. Teaching workers by reinforcing their successes and failures might also have an effect.
- *Self-confidence and decisiveness*. Contributors or workers that second-guess themselves more often may be less internally consistent.
- *Attentiveness and fatigue*. Environmental distractions or fatigue can change how consistently a task is completed. The microtasking design pattern in paid crowdsourcing is meant to negate some of the fatigue seen in traditional classification labour, but there is no way to anticipate other outside factors, such has how many tasks from other directors were completed. It is possible to affect attentiveness and fatigue within a task, however, with higher- or lower-effort tasks.
- *Perceived importance of task*. The perceived importance of a task might affect some other factors, such as attentiveness or self-confidence.
- *Time spent on each task*. The time spent on a task does not always translate to an indicator or quality, but might encourage greater numbers of contributions or more decisive contributions when controlled.

With these in mind, consider this study's image tagging task. How would the contribution change if:

- Tasks were 100 items long? 200? 1000? Only 1?

- Instructions were written very tersely? Verbosely, with many examples?
- Contributors were tested on the instructions at the beginning of the task? If there were known (gold label) items throughout the task? If everything had a known answer and workers were inconvenienced (e.g., with a time delay) when they got an answer wrong?
- Contributors were asked to volunteer their time? Were paid 1c per task? Were paid 10c per task? Were paid by the hour?
- Contributors were paid bonuses for performance against a ground truth or internal consistency? For continued task completion? For difficulty of their classification?
- Contributors were shown their performance (or estimated performance)? What if they were ranked against other workers? What if they gained levels or earned badges for performance?
- Contributors had tasks/time quotas to meet for bonuses? What if they were forced into these quotas (with tasks automatically moving forward)? What if a timer ticked away until their task disappear?
- Contributors were told when they got something wrong? What if you lie to them?

Some of these ideas of exciting, others are unfeasible.

Designs to encourage longer-term engagement from individuals do not appear to be a promising direction. As the previous chapter found with regard to relevance judgments, worker experience was not found to be significant. It is unclear whether this relates to the relative simplicity of the task or if it is indicative of a broader rule, though pushing against the on-demand nature of Mechanical Turk would likely be more effective for significantly more complex tasks than the basic information science ones considered here.

Other areas are already well-tread. The effect of incentive structures – payment and bonuses – has been studied frequently, notably by Mason and Watts (2010).

With regards to designs that mislead workers about their performance, there are ethical and trust issues that limit such an approach, in addition to the warning by Kraut and Resnick (2011) that feedback is only effective when contributors believe it is sincere.

Summary (*RQ 2.1*)

> **Summary (RQ 2.1)**: Different human factors may affect how people perform microtasks. This RQ formalized the design space for paid task design and explored possible manipulations that might change worker behaviors. Following from findings in previous work as well as promising areas that have not been previously well studied, this chapter will focus on interventions rather than parameterizations, designing manipulations around attentiveness and worker confidence.

Approach

While parameterization studies have compared how shifts in description or payment structure affect contribution, very few studies looking at more drastic design manipulations have been performed on a controlled task (A few of the exceptions have already been notes, such as Mitra, Hutto, and Gilbert 2015; Eickhoff, Harris, et al. 2012).

Still, some unresolved questions in the area are necessary to understand in the pursuit of quality crowd contributions. For example, it is still unclear whether simple encoding tasks benefit more from workers' gut instincts or careful consideration. Designs that can change a worker's attentiveness may help – or perhaps hinder – the quality of contributions.

Having found in the previous chapter that reading instructions slowly is important for properly performing work, it should be seen whether a task can push a worker into internalizing the codebook rather than interpreting it. Understanding that many reliability errors are introduced by honest workers that intend to do well, it may also be important to keep workers informed of their performance, at least when they are not performing well. Other work, to be reported in later chapters, finds that a task redesign for

an evaluation task can improve collected data immensely at little extra cost, while in the case of simple item ratings, over-thinking the task is actually detrimental.

With those considerations in mind, this study compares data collected through three interfaces for crowdsourced data collection: a training interface, a feedback interface, and a time-limited interface.

The training interface takes more care to slow down the task and walk new workers through the codebook and the style of a good contribution.

The feedback interface tries to reflect performance back to workers, to check their understanding of the codebook.

Finally, the time-limited interface contrasts the introspective approach of the other interface by encouraging quicker contributions.

For comparison, a carefully designed baseline follows conventions and best practices for paid microtasking platforms.

Basic interface (Baseline)

The basic interface resembles an archetypal task, following conventions seen in Mechanical Turk usage. It shows workers a task with a goal, description, and ten items to perform actions on. Prior to submission, there is also an optional feedback form. Though it is a baseline, it is not a hobbled baseline, designed around recommended practices.

The *goal* is the summarized statement for the task requirement, such as 'Tag images with descriptive words' or 'judge the relevance of documents in a search'.

The *instructions* describe, clearly but succinctly, the parameters of the task and any necessary details about completing the task. Part of doing so is explaining what a good contribution is: that is, delineating between good and bad tags in the tagging task, or explaining what a relevant or non-relevant document is for the relevance judgment task. The reason that instructions are intended to be succinct is again by convention. Amazon's advice for designing good tasks states that the task should not require scrolling to

start (*Requester Best Practices* 2011). In addition to conciseness, the instructions for this study's basic interface strive to follow other recommendations in a conservative and uncontroversial manner: specificity, examples, and clarity about poor work (*Requester Best Practices* 2011; *Guidelines for Academic Requesters* 2014).

It is difficult to balance the various needs of a good instruction set. With concern to succinctness and ease of readability, key information was italicized, examples were added as mouse-over popups, and secondary information (e.g., 'Tips') was hidden behind a tab. Another tab held a reference copy of the IRB disclosure (which, for this condition and all others, was shown fully when a worker was previewing the task before acceptance). Finally, an empty 'tab' to collapse the instructions completely was added. This is not a common feature of archetypal tasks, but given the difficulty of scrolling in the embedded window on Mechanical Turk, was deemed a humane addition. Figure 7 shows the limited task space when it is embedded within the Mechanical Turk interface.[38] Collapsible instructions have been recommended previously (E. Chen 2012a; E. Chen 2012b).

[38] Incidentally, Figure 7 also shows of the more challenging images to tag. How would you tag them?

Following the advice laid out by a notable set of best practices (*Guidelines for Academic Requesters* 2014)[39], a time estimate for task completion was also provided. The time estimate was determined based on testing and updated following an initial batch of tasks.

[39] These unofficial guidelines, on the Dynamo Wiki, were written collaboratively by academic researchers and Mechanical Turk workers.

It is recommended to be clear about what work is rejected (*Requester Best Practices* 2011; *Guidelines for Academic Requesters* 2014). Given that an underlying premise of this chapter is investigating whether the work director is sometimes to blame for poor work, it would be a troublesome foregone conclusion to actually reject work, so no work was rejected. Instead, even for the basic interface, improper work that would have been rejected in other settings was validated by the system when possible. For example, workers were asked for a minimum of two tags in the tagging task; as shown in Figure 8, they could not submit before entering two tags. To account for instances where a second tag was too difficult to create, workers could also add a placeholder 'TOOHARD'

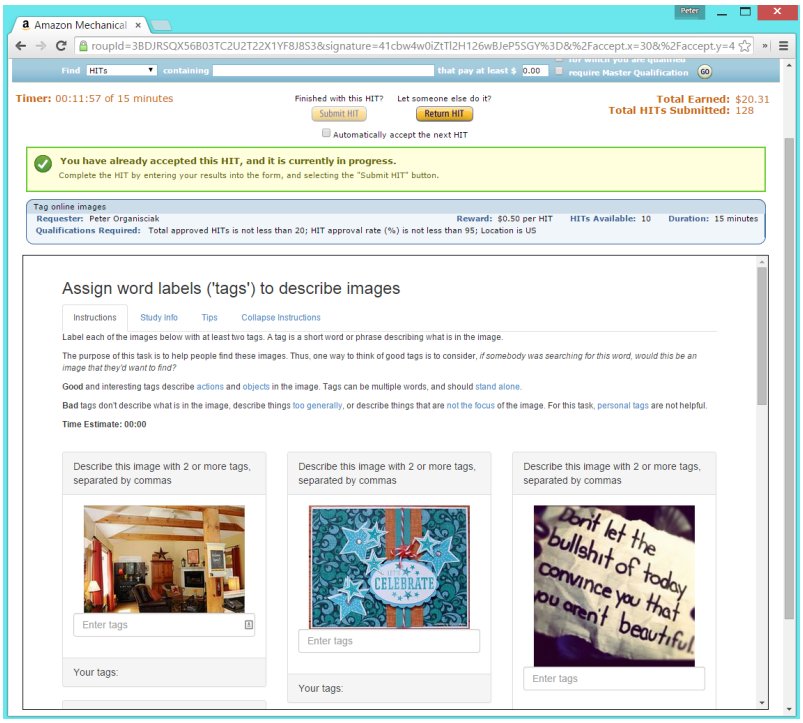

Figure 7: Task within the Mechanical Turk interface.

tag, which the interface alerted a worker to if their cursor was inactive for a few seconds (Figure 9).

The *task set* of items to perform work on, again followed a basic archetypal pattern, listing each item in a grid. The tasks themselves were small and did not require context shifting, as Amazon recommends (*Requester Best Practices* 2011).

Finally, the basic interface included an open-ended feedback form at the end. While this is far from a standard convention, many have recommended it as a standard element in task design (E. Chen 2012b; *Guidelines for Academic Requesters* 2014). The ability to respond easily provides valuable information on worker satisfaction and task problems.[40]

Training interface (TRAIN)

In the training interface, the worker is walked through their first task slowly. As they complete the tasks, their answers are evaluated against a gold standard and they are informed if they com-

[40] Most feedback form practice is anecdotal, because its value is qualitatively palpable but quantitatively intangible. Feedback forms provide a space for critical information – such as broken tasks – and qualitative information – such as worker satisfaction. Most importantly, they provide a more human interface between workers and directors. This runs contrary to Amazon's purposes as 'artificial artificial intelligence' but encourages directors to respect their workers.

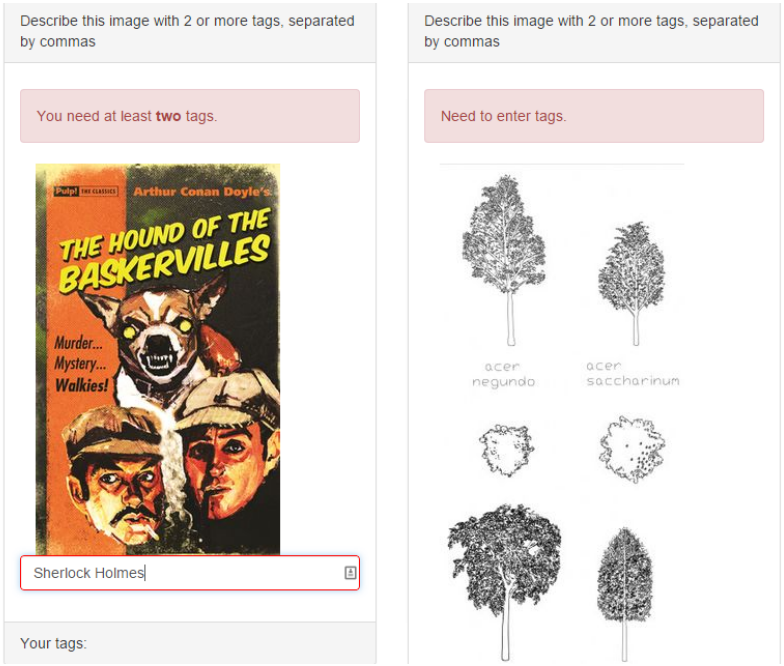

Figure 8: Contribution validation, which informs contributors of issues before submission.

Figure 9: An example of the additional help message in the basic condition of the tagging task, which appears after the input field is active for a small period.

pleted it correctly or incorrectly. Accompanying the 'evaluation', incorrect answers are also given an explanation of why the actual answer is correct.

The training tasks were hand-designed, based on a random sample of items. One can imagine an optimal sample, where the training set starts with easy tasks and quickly turns focus to the items that are most difficult. However, removing the most difficult items from the post-training pool would unfairly bias the training condition: this is why a random sample for training was selected.

During the training interface, workers are greeted with a message noting that their first task will be atypical, in that answers will be provided. The tasks set itself appears similar to a basic interface taskset, except that the individual items have a 'Check your Answer' button (Figure 10).

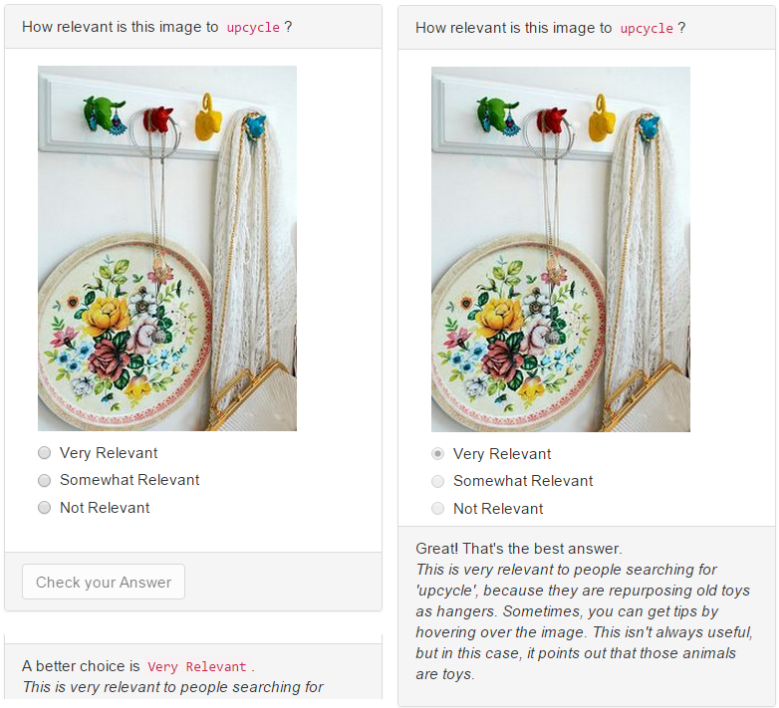

Figure 10: Item in a training condition, before and after checking the answer.

To better guide workers and for a clearer understanding of how the worker is performing, the 'Check Answer' button is disabled until a submission is made, and the submission interface is dis-

abled after the answer is checked. This helps explain to workers the intended order of contribution with minimal text: we want workers to try a contribution before checking their answer, and we don't want them changing the answer afterward.

The two task settings looked at here, image tagging and relevance judgments, have different levels. While the tagging task stands alone, each image functioning independently of the others, relevance judgments are grouped within queries. This means an initial interaction training task focuses on teaching specific to a query, teaching in-depth the mode of thinking associated with conducting the judgments, while not particularly training workers for the specific queries they will see.[41]

[41] In this study, workers had a 1/19 chance of getting a task set where they judged documents for the same query as they saw in the training task.

To also measure query-specific training, albeit at a smaller scale, the relevance judgment experiment is also evaluated against a training intervention (INSTRUCT). This intervention amounts to a full screen window with the task instructions as well as visual examples of very relevant, somewhat relevant, and non-relevant images (Figure 11). Contrary to the main training condition, workers do not have their own choices evaluated; instead, INSTRUCT focuses on recontextualizing the standard training instructions in a direct manner, one that has to be explicitly dismissed. However, it is also applied alongside each task set, rather than existing solely as a first-interaction task.

To summarize, the TRAIN condition:

- Introduces a training set on a worker's first task set.
- Pursues an answer-checking mechanism, where worker make a contribution and have it verified.
- Walks the worker through a set of tips and examples.

Additionally, the INSTRUCT condition:

- Uses a start of task set intervention, for each task set.
- Forefronts instructions, with demonstrative examples.
- Requires user input to continue past the instructions.

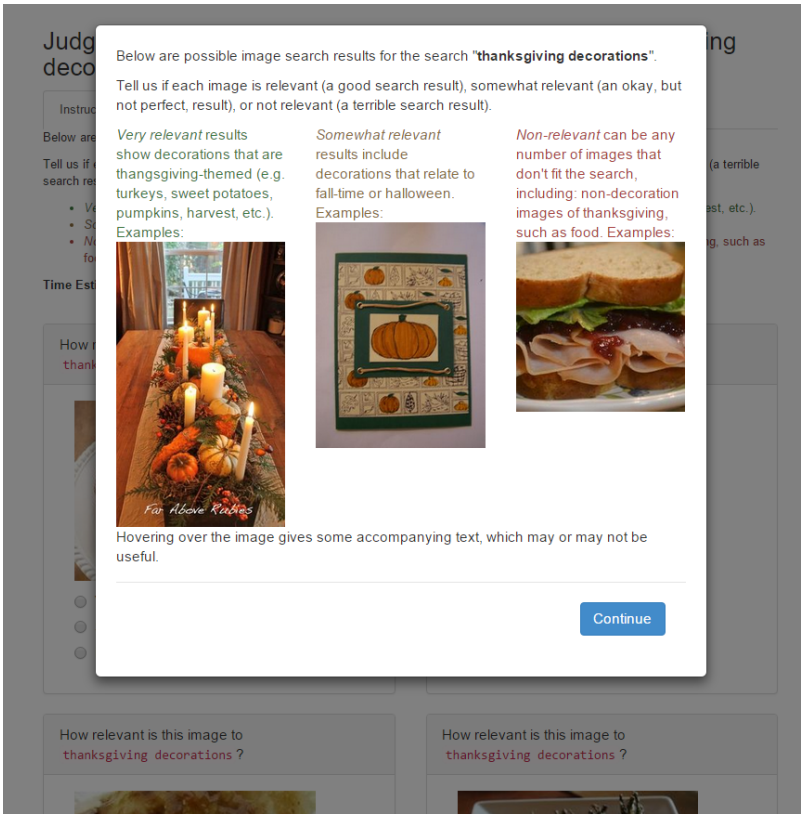

Figure 11: Intervention in INSTRUCT condition.

Feedback Interface (FDBK)

In the previous chapter, it was found that workers that did not spend enough time reading instructions did not perform as well overall, even when their typical task completion time looked the same as well-performing workers. Since in this case the poorer workers did not exhibit any time-optimization wage-maximizing behaviours, a possible reason is that they were performing honestly but simply did not internalize the codebook adequately.

If this is the case, is it possible to improve the performance of poor workers but simply letting them know of issues with their work? The feedback condition of this study attempted to do just that, with an intervention at the start of tasks, after a worker's first task, estimating the worker's performance.

In the feedback interface, a worker is shown feedback about their estimated performance on past tasks. The first task that they complete is identical to the basic interface. Starting with the second task, however, the interface gives them a window with their estimated performance, relative to other workers.

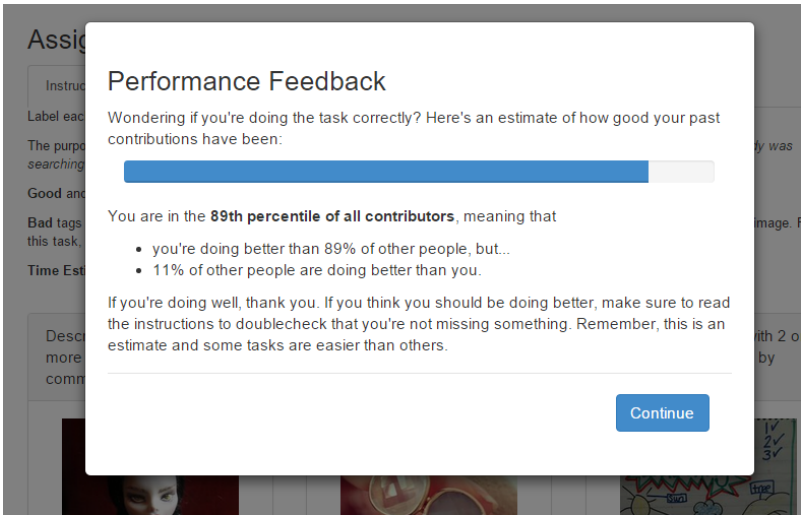

Figure 12: Example of showing feedback shown to workers.

The estimate of performance was determined differently for the different task types, image tagging and image relevance judgments, and is described in those respective sections.

As seen in Figure 12, feedback was given in the form of ranked percentile information relative to actual workers. The underlying measures or the exact estimate were not revealed to workers. Unlike the training interface, it did not provide any feedback on what was done wrong and what was performed correctly. This black box approach was because this interface was not intended to train, simply to inform and – it was hypothesized – encourage a return to the task instructions if a worker needed to recalibrate their understanding of the task.

Though it would have been trivial to adapt the text relative to performance, it was deliberately decided that the written copy stay the same, and that this was clear to workers. The intention was to avoid a perception of scolding, leaving the *interpretation* of the performance feedback to the worker; "*If you're doing well...*", or "*If you think you should be doing better*".

The expectation was that showing feedback may trigger an external motivation, simply in seeing that these statistics are kept, as well as intrinsic motivation, trying to perform better for self-competitive reasons. The former cannot be discounted, but the design tried to encourage more of the latter.

McCreadie, Macdonald, and Ounis (2011) attempted a similar approach, where contributors were shown a sidebar color-coding all their contributions based on their agreement with other raters and the authors. Showing this information with such granularity encourages workers to go back to reconsider debated answers, whereas this study's take tries to encourage more care and competition moving forward.

As with the training condition, workers are encouraged to continue performing tasks with a plea and a bonus (Figure 13). This is because feedback is only applied starting with the second task. Base payment was $0.05 lower than was provided for the basic interface, while the continued engagement bonus was $0.10 per task.

In sum, of the ample ways to design a feedback mechanism, this study's condition:

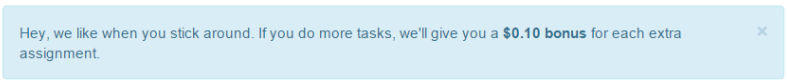

Figure 13: Message encouraging repeat work.

- Positions workers relative to the workforce, rather than absolute measures.
- Is designed with focus on worker self-motivation more than the observer effect.
- Focuses on intervention at a start of a new task.

Time-Limited Interface (FAST)

Not all crowdsourcing contribution cases may require more focus; sometimes a worker in a quicker mode of thinking contributes more consistent and reliable work. This was the case in an incidental finding reported in Chapter 6, where asking workers performing subjective opinion-based tasks to explain their judgment seemed to change the judgment habits themselves.

In contrast to the training and feedback interfaces, which serve in a way to slow down workers and make them more focused on their contributions, the final data collection interface pursues the opposite approach. The time-limited interface encourages quicker interactions by giving users a timer to complete all tasks. Providing a time-limit is expected to encourage less second-guessing of the contribution.

The goal and instructions do not differ greatly from that of the basic interface, except for text explaining the limited amount of time that a worker has.

The amount of time workers actually had differed between task types, because relevance judgments are completed more quickly than item tags. The former task lasted for 90 seconds, while the latter lasted for 60 seconds.

It is important not to distress the worker when trying to push them into a visceral form of task completion, as this might have the opposite effect. Instead, this design seeks to encourage *flow* (Csikszentmihalyi 1991), where a user is in an uninterrupted state

on concentration on the task. To avoid the potential distress of thinking about what is to come, this interface does not show a list of tasks to complete (e.g., "complete these ten tasks in a minute"). Instead, tasks are shown one at a time (e.g., "See how many tasks you can complete in a minute"), with bonuses paid for each complete task and increased for correct answers. Figure 14 shows an example of the interface.

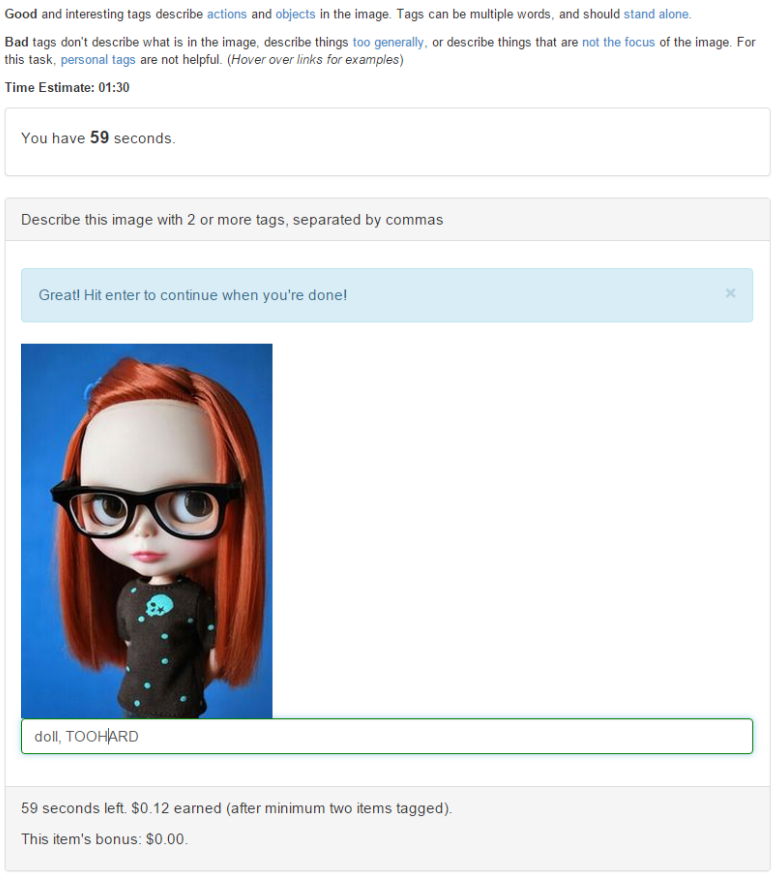

Figure 14: Task in FAST design.

Determining a payment is nuanced for this condition. Bonuses are bound in promises: there is no system restraint to guarantee payment other than the director's word. To assure workers that they will be paid, it is important to still have a notable base payment. At the same time, an effort-optimizing worker might realize that it is favourable to avoid the per-contribution bonuses, and keep completing 'do as many as you can in X seconds' tasks with

only one contribution. To counter such a possibility, the per-item bonuses ramp up; e.g., no bonus for the first task, $0.01 for the second, $0.02 for the third.[42] This provides incentive to actually try to maximize time.

[42] The screenshot in figure 14 was taken for the first item in a set, so it shows a zero-sum bonus.

Evaluation

The experiments in this study were run in a naturalistic setting, running directly on a paid crowdsourcing platform, Amazon Mechanical Turk, with real workers.

There are trade-offs to this setting. It is easy to instrumentalize and properly capture the actual skills and attentiveness of paid crowd workers. However, working within the conventions of the system means that some parts cannot be controlled. For example, workers cannot be forced to perform multiple tasks, simply encouraged to do so. Also, while sampled from the sample pool, the actual workers testing the different interfaces are not necessarily the same individuals, given that it is a workers choice whether to complete a task (or even when to be on the platform performing tasks). Thus, it is important that the users are similarly representational: it would be problematic if one interface was used mainly by Indian residents while another was performed mainly by American residents (the second and first largest nationalities on Mechanical Turk, respectively).

For this reason, each interface was evaluated with temporal and geographic restrictions. Workers were restricted to American workers, and most tasks were during the American work day, with only slight deviations.

Finally, restrictions implemented into the system restricted workers from participating in multiple conditions.

Implementation

The experiments were performed on Amazon's Mechanical Turk, using an API that allows external pages to be hosted within the Mechanical Turk interface.

Experiments were run using a custom system called Crowdy.

Crowdy was developed using JavaScript on the front end, built on top of the AngularJS library[43]. The software is released with an open-source MIT license[44].

The back-end of the stack also runs on JavaScript, with a Node.js[45] server run on the Express[46] web application framework. Because of the complexity of the logic in serving tasks, this code was optimized toward asynchronous operations whenever possible. Data storage uses the MongoDB[47] database. The task serving code is also released online[48].

[43] https://angularjs.org/

[44] https://github.com/organisciak/crowdy

[45] https://nodejs.org/

[46] http://expressjs.com

[47] https://www.mongodb.com

[48] https://github.com/organisciak/crowdy-backend

Experiment #1: Relevance Judgments

Lowering the barrier to custom evaluation is one of the most notable contributions of crowdsourcing to information retrieval research. While production systems benefit from actual humans in the machine to identify topics and correct algorithmic quirks[49], research aiming to improve pure information retrieval performance still needs ways to appropriately evaluate different models and approaches. Paid crowdsourcing platforms offer a way to tap into large and diverse groups of people for relevance judgments, making custom evaluation datasets – and subsequently research over novel corpora – greatly more accessible.[50]

Continuing from the previous chapter's look at contribution behaviours and post-collection indicators of quality for relevance judgments, this chapter judges the effects of collection-time design manipulations[51], starting again with information retrieval relevance judgments as a experimental setting on which to compare these manipulations. In this experiment, judgments are collected for image retrieval results.

[49] Discussion of crowdsourcing in information retrieval beyond evaluation uses is in *Introduction to Crowdsourcing* (Chapter 5).

[50] The role and value of crowdsourcing for information retrieval evaluation was discussed at length in the previous chapter. For brevity, it is only lightly recalled here.

[51] One may wonder about the order of chapters, from focusing on post-collection to collection-time. The reason is that the preceding post-collection modeling work motivates approaches seen in this chapter.

Data

The dataset being evaluated consists of 389 query–image document pairs, evaluating 30 results each for 13 queries[52], against a corpus of 185.6*k* documents from image-sharing social network Pinterest. All the data, control and experiments, were collected

[52] With one result removed for sensitive content.

specifically for this study, allowing for a fair comparison of design manipulations against a baseline that was competently implemented.

This section details the process of developing the test corpus. The process was as follows:

- Collecting a large randomized sample of image documents from Pinterest (*pins*).
- Sampling realistic queries, collected from Pinterest's query auto-suggestions.
- Implementing a retrieval system with the sampled documents, and retrieving results against the sampled queries.

The retrieval results are used for measuring the efficacy of different designs for collecting relevance judgments. As such, there was a desire for some heterogeneity in the results, which motivated the creation of a custom sample. With over 200 million documents, it was worried that the alternate approach of scraping the results from Pinterest's own search system would results in a set highly skewed toward very relevant documents, making it difficult to separate a signal in the experiment.

Pinterest is a social network centered on the saving, sharing, and curation of online images It is built entirely on crowd contributions. On Pinterest, the document unit is a 'pin': an image, associated with a web URL and page title, and a required text description provided by the user. Though the most common type of pin is saved from an external website, it is also possible to upload personal content. The 'descriptions' are required but free-text, meaning they do not necessarily *describe* the image.

Pins are sorted into curated lists, referred to as 'boards'. Like pins, classification into boards is not controlled. While adding a pin to a board is an act of classification, the classes are user-defined and can be created for various reason, such as quality judgments (e.g., "Neat stuff"), thematically descriptive (e.g., "dream wedding"), or miscellany of various sorts (e.g., "inspiration", "funny"). Boards are user-specific, created by a user with

a title, description, category, and optional map.

In the words of the company materials, Pinterest features three primary purposes: saving (as pins), organizing (into boards), and discovery (*About Pinterest* 2014). In this way, it is organized in a way familiar to library and archival communities, distributing online images with an eye toward discoverability and curation.

It is also a large-scale site of descriptive crowdsourcing, recalling past trends in social bookmarking (i.e., the eventually doomed *del.icio.us*) but with a visual spin to the bookmarking activity. Users describe pins and categorize them into boards; describe, title, and categorize boards; and contribute various social information, such as comments, repins[53], and voting (in the form of 'heart'-ing).

[53] To *repin* is to save a new pin from an existing pin, using the same source URL and image, but applying a new description and saving to a new board. A document's repin count can be interpreted as a measure of a document's internal influence among the Pinterest community.

Pinterest is a novel site for studying crowdsourcing in the context of retrieval. This study, concerned with the methodology of crowdsourcing, is not dependent on Pinterest, but Pinterest is nonetheless an appropriate site to underpin it. Structurally, it resembles the archival form of many library and museum systems, albeit at a larger scale, it is itself a product of [volunteer] crowdsourcing, and it deals in the type of sparse, simple content that crowdsourcing is appropriate for.

- The organizational form of Pinterest, grouping documents into curated lists called 'boards', is an interface pattern that is relevant to many forms of information repository. Social OPACs, for example, allow library patrons to collect books into similarly uncontrolled lists.
- Pinterest contains very little information about the source web document. It is feasible to crawl the full text of the source, but as it stands, a Pinterest 'pin' alone offers a record of a *single person's interpretation* of the source.
- Since the primary form of Pinterest document is a human reaction to a web document, the user contributions on the site may have possible future use for web retrieval.
- For the purpose of this study, collecting relevance judgments

for retrieved pins, the image-centric format is the type of task that crowds should be adept at. Good tasks should focus on one thing, with little context switching (*Guidelines for Academic Requesters* 2014), and understanding an image is a quick, natural activity for people (Ahn 2006).

Finally, Pinterest is an interesting but understudied website. Demographically there is a female skew, interesting precisely it counter-balances the typical male-heavy community demographic.

Sampling was done against Pinterest's provided sitemap. An initial survey (August 2014) suggested that Pinterest had approximately 107.5 million users, with 207.5 million pins and 572 million boards containing those pins.

This is a very large amount of data, and only a sample was needed for this study. For the sample, random 25*k* pin sitemap listings were downloaded, a process randomly pulled out approximate 1% pin listings, the collected pins were randomly ordered, and the full metadata of pins was collected against this master list.

For the information retrieval system underlying this experiment's relevance judgments, a sample of 195k pins was collected and indexed.[54]

[54] This sample was collected in 40k pin batches, and not all at once. As a result, the final number of pins successfully downloaded was lower, at 184583 documents. The time difference between batches provides a sense of attrition on Pinterest. The first two batches of pins were collected against a five-month old sitemap, and 1.4-1.5k pins were no longer accessible per batch (~3.4% attrition). Another batch was scraped when the sample list was 9 months old, and 6.4% of the links were no longer active, the final two batches were collected 2 months later, with 7.2% links no long online.

Queries were sampled from auto-complete suggestions on Pinterest. When a user starts to type in a query, five suggestions appear. For example, typing 'r' will suggest 'recipes', 'red hair', 'rings', 'relationship quotes', and 'rustic wedding'. These appear to be the most probable queries starting with the provided string.

The top queries for each letter of the alphabet were collected for the sampling pool. To shift the sample list away from the most-popular queries, the sampling frame also included 500 queries derived from auto-complete suggestions based on two character strings: specifically, the one hundred most common two-character pairs occurring at the start of the English language (via Norvig 2014).

Thirteen queries were sampled. For each query, a description

of what constitute the different levels of relevance was written by myself. Three point categorical relevance was used, with 'not', 'somewhat', and 'very' relevant as the options.

Results to judge were generated using a Dirichlet-smoothed language modeling system. A basic form of query expansion was used wherein the original query was run, and a word list consisting of the top results was resubmitted as a secondary query. Given the short nature of these documents, document expansion in this style would have been appropriate (Efron, Organisciak, and Fenlon 2012), but the query expansion sufficed for widening the net for results.

Measurements

Across all conditions, 12037 relevance judgments were collected, approximately 30 redundant judgments per document. For every condition except the time-limited (FAST) condition, these were completed in randomly selected batches of up to ten.

The ground truth data is constructed from the majority vote label for the documents; that is, the most common judgment made by the 30+ workers that have judged each document.

Additional information was gathered on satisfaction and time spent.

Regarding feedback, workers had an optional free text response form, an optional five-point colloquially-worded 'pay satisfaction' input, and a similar optional 'task satisfaction' input.

Time spent was gathered in seconds. It should be noted that, unlike the later tagging experiment, where a worker's time spent focused on an input box could be measured, there is no easy proxy for measuring per-task time in a set. Since the relevance judgment options are a set of radio buttons, we do not capture the *start* of a worker's attention, just the moment that they actually make their contribution. As a proxy, a measure was taking of the amount of time that the *previous* item was in focus; i.e., worker clicks item A, and while they think about item B, A is still in focus. This provides a rough estimate of the time spent, good enough for broad

comparisons, though not robust enough to tie time to a specific item.

Results

This section present the results for performance, time, and satisfaction. Analysis will follow after.

The performance of workers across conditions is shown in Table 5, with the primary statistics for the likelihood of a document's relevance being correctly judged in each condition. All the judgments were collated by query-document pair, so each datum represents that condition's mean 'correctness' on a given query-document pair, for all 389 pairs. Significance tests are also shown for the condition's equality with the baseline, which is rejected at 0.01 for FAST and INSTRUCT, at 0.05 for FDBK, and not rejected for TRAIN.

The distribution of this data is shown as violin plots in Figure 15, with the lower quartile and median marked. The upper quartile is at 100%, meaning that even for the worst condition, FAST, at least one-quarter of documents were always correctly judged.

condition	mean	median	std.dev	sig
BASE	0.734	0.800	0.269	/
FAST	0.693	0.800	0.267	**
FDBK	0.796	0.875	0.216	*
INSTRUCT	0.791	0.857	0.259	**
TRAIN	0.780	0.800	0.214	

Table 5: Statistics for the likelihood of a document's relevance being correctly judged, by condition. Significance marks rejection of equal distribution to the baseline (Mann-Whitney U, Bonferroni-adjusted significance at 0.05 - *, 0.01 - **, and 0.001 - ***).

How good were the workers on average? Table 6 shows the median and mean quality of worker, scored by their accuracy rate. This does not take into account whether workers were given easy or difficult tasks to perform or if some documents were judged more than others, but it reflects the same order of INSTRUCT, FDBK, TRAIN, BASE, as was seen above.

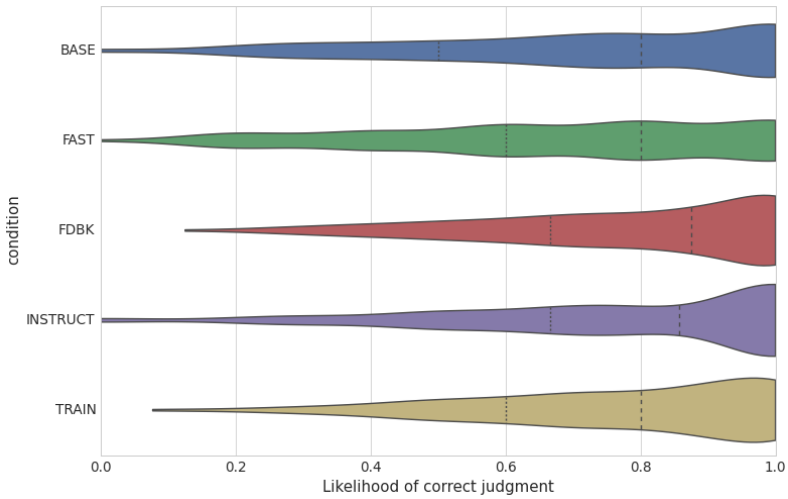

Figure 15: Distribution of correct judgments by item, shown by condition. The Lower quartile is marked by dotted lines and the median is marked by a dashed line.

condition	mean	median	std
BASE	0.750	0.759	0.179
FAST	0.685	0.716	0.197
FDBK	0.783	0.800	0.119
INSTRUCT	0.782	0.817	0.172
TRAIN	0.775	0.791	0.101

Table 6: Mean and median quality of workers in each condition, by accuracy rate – the proportion of all judgments performed correctly.

A MEASUREMENT OF TIME SPENT on each task was taken, in seconds. Figure 16 shows the distributions of time spent per task, faceted by the experimental condition. Interestingly, the time-limited condition, FAST, was found to be *slower* than the baseline (Table 7) This finding stands in stark contrast to what is seen later for the tagging experiment. There are some possible reasons for this, which are discussed later.

condition	mean	median	std	N	sig
BASE	2.98	1.79	3.92	1894	/
FAST	3.59	2.91	2.30	1921	***
FDBK	2.90	1.79	3.15	3318	**
INSTRUCT	2.90	1.76	3.53	1629	
TRAIN	4.06	2.58	5.72	3906	***

Table 7: Time spent per relevance judgment in each condition. Significance marks rejection of equal distribution to the baseline (Mann-Whitney U, Bonferroni-adjusted significance at 0.05 - *, 0.01 - **, and 0.001 - ***).

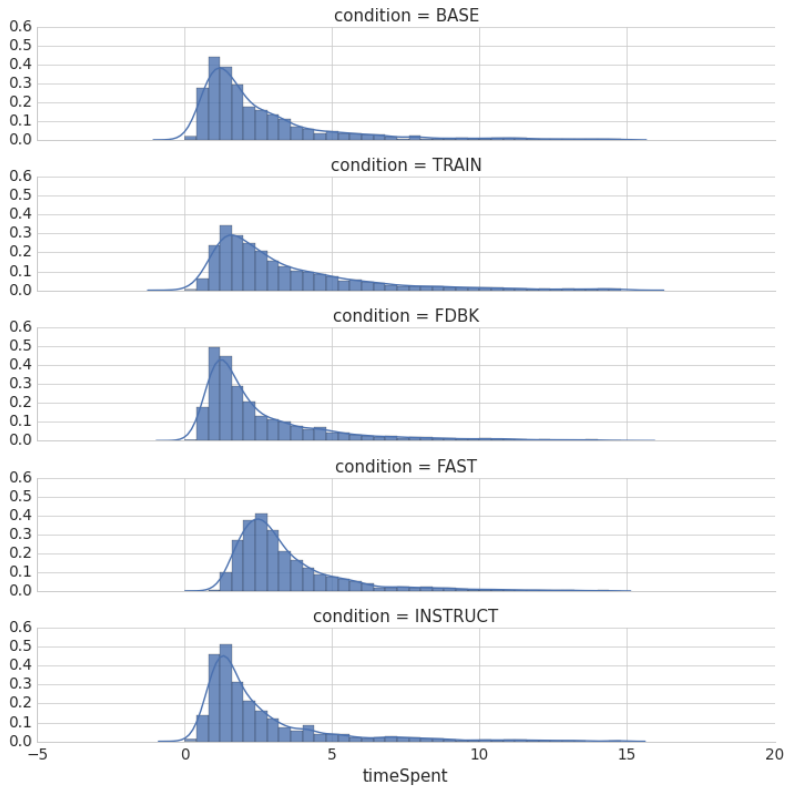

Figure 16: Comparison of time spent per task, in seconds. N=12667.

Considering this data by mean user time, to reduce the influence of outliers, tells a comparable story. Viewed in this manner, the time-limited interface and the baseline were comparable, while feedback and training shortened the mean time of the average worker.

While this data tells us how quick the actual tasks in a task set are completed, there is also the time spent in-between contributions, which can be assumed to primarily time spent on instructions. Table 8 shows the time spent in these moments.

Notably, FDBK appears to be considered in much less time than other conditions. Recalling that FDBK and TRAIN are never a worker's first interaction, Table 8 also shows values excluding first-time interactions, for a better comparison. The results do not change much, other than showing that the baseline tasks are completed quicker in later interactions. In contrast to FDBK, training intervention INSTRUCT compelled people to spend the most time

on the instructions.

condition	mean	median	std	N
BASE	30.79 (22.93)	18.03 (12.66)	36.09	200
FAST	34.86 (31.54)	23.45 (21.09)	35.86	132
FDBK	19.58 (–)	13.17 (–)	19.35	359
INSTRUCT	48.64 (48.87)	22.01 (20.49)	74.65	210
TRAIN	38.89 (–)	14.12 (–)	73.90	378

Table 8: Time spent on non-contribution parts of a task, excluding tasks where feedback form was completed. Parenthetical values show information only when worker's Nth task is 2 or more.

Finally, Workers were given the option to rate the task and their satisfaction with the payment, on a scale from 1-5.

Figure 17 shows the distribution of task satisfaction scores for each condition, and Figure 18 shows the payment satisfaction scores. In all cases, they were skewed toward the upper end – the median is 5 for each condition – as may not be surprising. The main point to note is that none of the conditions are troublesome for workers, and that workers in the pseudo-competitive conditions (FAST and FDBK) seem to enjoy the tasks slightly more.

Figure 17: Relevance judgment task satisfaction scores, by condition.

Figure 18: Relevance judgment payment satisfaction scores, by condition.

A related question is whether a person's rank – as given in the feedback condition – affected their satisfaction. There were not enough measurements for a non-parametric comparison, but it

did not appear to be a notable factor. Given a larger sample, one interesting quirk that was observed is that the lowest pay satisfaction was among the best contributors ($mean = 4.375$, $median = 4$, $N = 16$).

Analysis

The best work was contributed by workers in the training intervention (INSTRUCT) condition. Their contributions were significantly more accurate at the same cost and with no discernible change in time per task. However, supporting the interpreted finding in the previous chapter, accompanying the improved performance was more time spend reading instructions. In the previous chapter this measurement was confounded with the completion of the first task, here we confirm it.

INSTRUCT is an easy condition to parameterize, only requiring a one-time cost from the director to collect training examples and perhaps a slight development cost to implement instructions as a dismissable, up-front modal window.

The intervention in INSTRUCT was in part motivated by Shu et al. (2012), who found that for reporting forms requiring a signature to confirm honesty, such as tax or insurance forms, asking people to sign at the top led to more honest reporting. By foregrounding the instructions, this condition seems to encourage workers to be more honest about the codebook. An unknown caveat is whether this is a persistent effect or, if after enough tasks with this condition, workers start to dismiss the information sooner. With the rotating workforce on Mechanical Turk, this would not be a concern for contexts similar to this study, but would be worth studying if applied in alternate contexts.

The quality of contributions in the feedback condition also improves on the baseline condition. Like with INSTRUCT, three-quarters of query-document pairs were corrected classified by two-thirds of contributors. This means that with as few as three redundant judgments, most of the consensus votes would be

correct.

A surprise with the feedback task was that workers did not slow down, but indeed performed the tasks quicker. The reason for this behavior is unclear. Figure 19 shows the correlation between time spent the ranked percentile that the worker was given, which does not show any clear linear pattern.

One possible explanation is that, while workers in the 60th percentile slowed down to try to improve, the best workers were validated to trust their instincts while the worst workers did not care. The goal of the performance feedback was to give honest but poorly performing workers an indicator of lower quality contributions. This somewhat performed its function in the middle of the pack, but didn't compel the poorest workers.

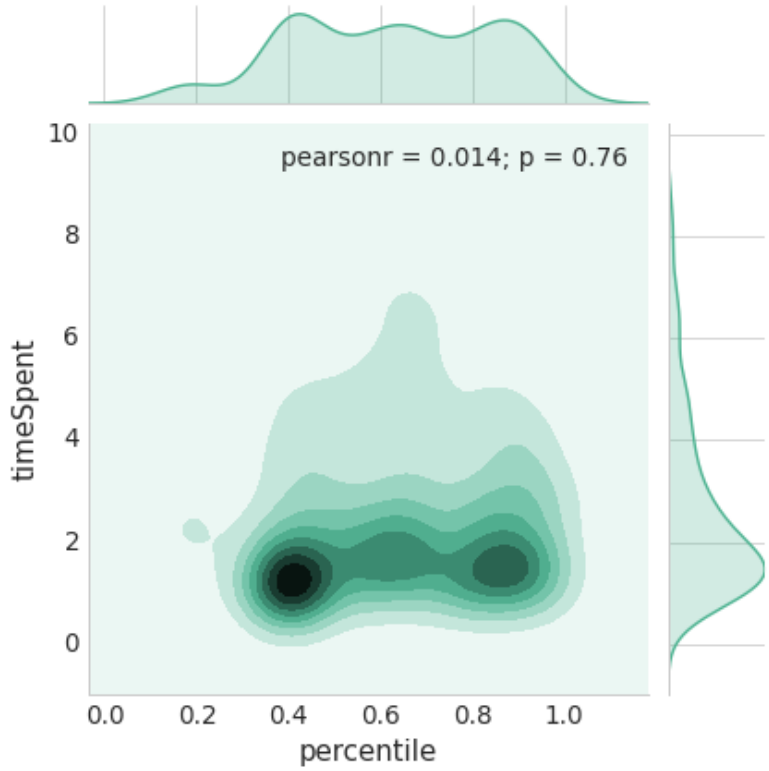

Figure 19: Comparison of the time workers in the Feedback condition spent on relevance judgments against their percentile rank, as provided to them.

Work in the first-task training condition, TRAIN, was not significantly different in quality from the baseline, which ap-

pears to confirm that training on one query does not assist in completing relevance judgments for other queries.

Notably, TRAIN slowed down workers, as if the close training with one query made then sensitive to nuances that other queries may have. This did not translate to performance improvements, however.

As expected, time-limited workers did not perform as strongly as expected. However, the accompanying expectation was not seen in the data: that for the loss in quality, FAST would increase the speed on contributions and potential improve the capacity to collect redundant judgments.

To the contrary, FAST was not fast. A possible explanation is that image relevance judgments are already a very quick interaction, only a few seconds. The one-at-a-time fast interface may have stood in the way of workers' comprehension of the entire task set: where in a traditional setting they can click on their answer and already be thinking about the next one, here they had to click again to move to the next task. It is possible that for relevance judgments over more complex types of documents, the fast interface would perform differently, akin to what will be seen in the next experiment.

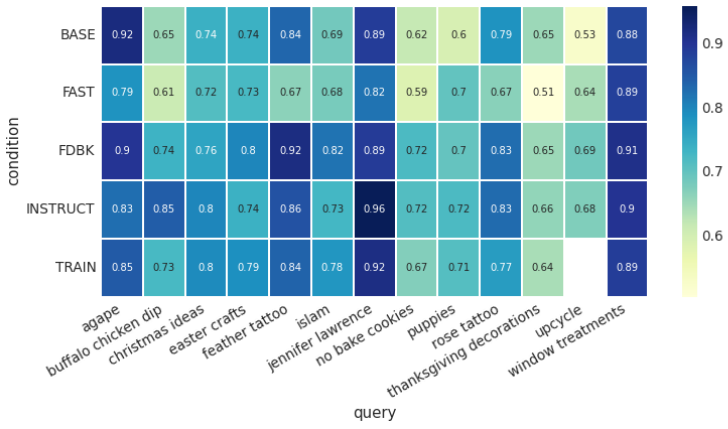

Figure 20: Likelihood of correct relevance judgment shown by query.

A comparison of performance per query (Figure 20) offers

some insight as to the relative strengths and weaknesses of the various conditions.

The thirteen queries tested are shown as columns in Figure 20. Despite across-the-board worse results, the places where FAST appears to stumble most is where there was nuance to the instructions, in edge cases where one might consult the instructions again. For example, are other tattoos relevant to *rose tattoo*? (No.) Are other flowers relevant? (Somewhat.) The instructions for what was very relevant, somewhat relevant, or not relevant were in a way subjective, in that they were interpreted by myself when written, but this simply grounded the correct answer in a single interpretation.

The query analysis also shows places where the hand-chosen examples for INSTRUCT seemed to mislead. '*agape*', a Christian word relating to love, was extremely underrepresented in our sample, and all the results were non-relevant. This was a tricky inclusion, and by showing workers in INSTRUCT examples of relevant or somewhat relevant, it may have additionally misled workers to expect that there *are* relevant results. Likewise, it seems that the manual choice of examples for 'Islam' and 'Easter crafts' also misled to a certain extent.

Summary

In sum, for relevance judgments of image documents, a per-task training intervention and performance feedback improved the quality of judgments, while a training set on an unrelated query had no significant effect, and a time-limited interface actively disrupted workers.

Experiment #2: Tagging

To study design manipulations in a more inherently interpretative, difficult task, we turn to tagging. Tagging is a type of free-text labeling often applied in online social contexts. While it has the potential for generating useful descriptions, it is difficult to moti-

vate volunteers to provide tags uniformly across a collection, and the style of tags contributed is not always the most useful form of contribution. Paid crowdsourcing may resolve these issues, by exerting a stronger codebook and directing the attention of workers.

Tagging allows a system to collect more metadata about its records than it may have, as well as representing different *types* of description. Such open-ended contribution can grow unwieldy and hard to protect against vandalism, but public good institutions such as libraries and museums can use them for a sense of how the people they serve are interpreting the materials of their collections. Tags are also useful for augmenting large encoding efforts. For example, on business recommendation system Yelp, tags allow users to contribute data about the type of business[55].

[55] A competing service, Foursquare, also uses tags, but in a more structured way.

Trant and Wyman argue that tagging from online users "appears to fill gaps in current documentation practice" (2006). Following from this, tagging is particularly helpful for difficult to model formats (i.e., non-text) and when corpus sizes surpass the ability to formally classify works. Tagging has been used to encode scans of text (Ahn, Maurer, et al. 2008), improve information retrieval document modeling (Lamere 2008; Bao et al. 2007), augment personalized search (Lerman, Plangprasopchok, and Wong 2007; Noll and Meinel 2007).

Tagging also promises, in theory, a break from the Vocabulary Problem (Furnas et al. 1987). Furnas et al. performed a set of term generation experiments in 1987 where they asked participants to describe functions or objects. They found that the amount of spontaneous consensus was very low, arguing that this problem of vocabulary becomes a user issue because different users expect different vocabularies. However, it's the designers who get to choose the primary vocabulary and, "as heavy users, grow to find [their] terms obvious and natural" (ibid).

The proposed solution to the vocabulary problem was to allow 'essentially unlimited numbers of aliases'[56] (ibid) – something that tagging functionally allows. With tags, the descriptors for

[56] At least when those aliases are not more popular for other functions.

an information object are not one centralized viewpoint, but an amalgam of different viewpoints and different vocabularies.

This is the ideal setting.

However, there are numerous difficulties with tagging, especially when reliant on volunteers.

While tagging seems to address the vocabulary problem in its potential, the multiple alias approach does not guarantee that they will be approached reliably. In instances of low contributor engagement and sparse tags, the fact that people's vocabularies differ can be aggravated by the loose contribution style of tagging.

Additionally, volunteer tagging practices often do not match the needs of system designers. Of all the contribution approaches enabled by The Commons on Flickr, for example, Springer et al. (2008) found that tags were the least fulfilling type of information contributed to the Library of Congress account on Flickr. Likewise, in a look at social features used in library online catalogues, Spiteri (2011) finds tagging to be among the least used. That finding does suggest that features which are arguably more self-serving, such as curating lists of library materials and starring liked works, are easier to collect than more pragmatic features for item description.

Finally, while it has been suggested that tags with the most general usefulness also tend to be those that are applied by the most *different* users (Sen, Harper, et al. 2007), volunteer tagging does not necessarily reflect that diversity. On The Commons, it was noted that 40% of the tags were contributed by 10 'power taggers'; nonetheless, with over 59 thousand tags contributed, this still meant that other users still contributed a notable amount (Springer et al. 2008).

Due to the discord between the promise of tagging and difficulties of volunteer-based implementation, paid crowdsourcing is an appealing approach to collecting tags. While it does not include some tertiary benefits related to community engagement, it allows us to control for quality by enforcing a codebook and to make use of many diverse viewpoints in practice.

Following from the experiments on information retrieval rele-

vance judgments, this section evaluates our design manipulations over an image tagging task. Image tagging is an interesting departure from relevance judgments in that there is no *clear* correct contribution. A good contribution is dependent on the need. For this reason, designs that help guide a worker in relevance judgments may mislead for tagging.

There are different interpretations on tagging, what types are desired, how much variance is acceptable, and what the role of tagging is.

With tagging, some degree of variability is desired, because that diversity is central to many benefits of tags. However, it would be ill-advised to view tagging completely as a relativist activity. It has been found that tagging begins to converge on a set of popular, common tags (S. A. Golder and B. A. Huberman 2006) and there certainly are notions of 'good' or 'bad' tags (Sen, Harper, et al. 2007).

For museums, tagging is considered not only in pragmatic subject access terms, but as a medium for critical value and engagement, according to Trant and Wyman (2006). Indeed, they note the value in tagging for understanding the tagger: a way to understand how patrons react and interact with museum collections.[57]

[57] "A tag is a user's assertion that a work of art is about something." - Trant and Wyman (2006)

Sen, Harper, et al. (2007) study the quality of community tags in the MovieLens film recommendation system, toward methods to prioritize tags in the interface. They find that high-quality tags, as determined by survey, are not necessarily the most-applied tags, likely because the most common tags are locally useful 'personal' tags. However, tags that are applied by many *unique* users are more likely to be useful, as are tags that are clicked by many unique users. While this form of usage-based quality indicator does not help in collecting good tags, it does affect how to determine quality tags for ground truth in this study.

When Sen, Lam, et al. (2006) compared different approaches to collecting tags – an interface where prior tags are seen, an interface where only popular prior tags are seen, an interface that

shows recommended prior tags – the interface that did not show prior tags had a much larger proportion of never before seen tags. This is an unsurprising phenomenon, given that tagging habits appear to be influenced by the community (Golder and Huberman 2007; Sen, Lam, et al. 2006); however, it is a factor influencing our approach to tag collection through paid crowdsourcing.

A crowd marketplace is emphatically not a community, at least not in the service of a director's particular task, and generally the pragmatic system-oriented uses are underlaid by a desire for convergence and minimal redundancy. This is to say, a director looking to pay for image tags may not want a vocabulary explosion, but also may be looking to avoid the added complexity and cost of collecting prior tags to show to workers.

If the goal is to collect high quality tags, it must first be clear what a 'good' or 'bad' tag is. When studying tag quality in a film recommendation system, Sen, Harper, et al. (2007) found that only 21% of tags are worthy of display to other users.

This study looks to augment image record data and metadata with additional information that cannot be trivially inferred without human contribution. Particularly, we look to information retrieval uses, to help in findability, filtering, and organization.

One typology for types of tags was offered for tagged bookmarks by S. A. Golder and B. A. Huberman (2006). They present seven kinds of bookmarking tags: those for identifying what the item is about (i.e., topical), for identifying what the item is (e.g., blog), for identifying the creator of the item, for qualifying or refining other tags, for labeling subjective characteristics of the content, for establishing a relationship to the tagger (e.g., 'my post'), and for organization.

Sen, Lam, et al. (2006) collapse the seven classes from Golder and Huberman (2007) into three: *factual* tags conveying objective information, *subjective* tags conveying opinions, and *personal* tags that are intended only for the tagger. As expected, factual tags were found to be most generally useful, particularly for learning

and finding, although personal and subjective tags were useful for self-expression and organizing, respectively.

These are uses that matter to users of online communities, though for the organizational purposes of this study's image tagging task, and perhaps more generally for the controlled paid setting, factual tags are the most desired.

Finally, Springer et al. (2008) analyzed a sample of tags and derived a number of non-exclusive categories for image tags, including tags derived from the description, new descriptive tags, new subject words, commentary, emotional/aesthetic responses, personal knowledge or research, machine tags, variant forms, foreign language tags, and miscellaneous tags (Springer et al. 2008). Of these, description-derived and image subject tags were the most common.

These studies informed how this study presented the tagging task. First, the underlying criteria motivating a good or bad tag was directly stated, that the purpose is to help people find images. "If somebody was searching for this tag, would this be an image that they would want to find?"

Of the typologies studied, the types of tags that would inform this task are factual and descriptive (new descriptive). Machine tags, variant forms, these can be generated without crowdsourcing; on the other end of the spectrum, personal tags are not useful here.

The examples of good and bad tags to anchor workers were as follows:

> **Good** and interesting tags describe *actions* and *objects* in the image. Tags can be multiple words, and should *stand alone*.
> **Bad** tags don't describe what is in the image, describe things *too generally*, or describe things that are *not the focus* of the image. For this task, *personal tags* are not helpful.

The words italicized above were presented as links, showing a pop-up box with examples on hover (a note made this fact clear to workers). Emotional/aesthetic tags were neither explicitly encouraged nor discouraged.

Data

The images gathered for this experiment were again textually-sparse documents from Pinterest, collected as described earlier. From the 185k document corpus, 100 images were randomly sampled from among pins with a minimum one 'like' rating and one repin. The sample size was contained to 100 items to focus resources on maximizing the tags per item, under the expectation the vocabularies can grow large. 6201 tags were collected, comprising just under 2000 unique tags.

Tags were collected on Mechanical Turk in task sets of 10 images, except in the FAST condition, where the number of images tagged was dependent on time. To add variety and move away from the single most obvious tag, workers were required to contribute two tags – though they were given an escape path in the form of a 'TOOHARD' tag.

In the interface, only the images were shown. The image's original title and description from Pinterest could still be seen, behind a popup activated by mouse hover over the image. Workers were warned that the text may be useful, but may also be misleading.

The training condition, TRAIN, trained workers on 10 images, providing feedback on how good their tag attempts were and showing examples of other tags, organized by 'poor', 'OK', 'good', and 'poor'. There was no INSTRUCT training intervention tested for this experiment.

In running the FAST condition, the payment structure was developed to approximate the payment of the basic interface if completion behaviors were equal. That is, since the first batch of basic tagging contributions averaged 23 seconds each for 10 tasks at $0.50, the payment for the timed interface was intended to match that reimbursement rate at 4 tasks.

Measurements

As with the relevance judgment experiment, time and feedback information was collected. The data on time spent per task was

more specific, because it could be collected more directly than in the relevance judgment condition.

Despite prior work suggesting that the most popular tags are also the most useful in some contexts, a metadata enrichment setting did not seem like one of them. The concern was that the most popular tags would be those lacking in adequate specificity. Consider that a tag for 'German Shepherd' or 'Calico' is more useful than 'dog' or 'cat'. However, the latter are easy tags to converge on.

Accordingly, a large manual evaluation of the tags was performed by myself. 1976 tags were judged on a four-point scale consisting of the following ordinal categories: *poor*, *OK*, *good*, and *great*. The 'TOOHARD' tag was removed. These were evaluated on their own merits against the goals stated in the task – *how good would this image be for a person searching for the tag* – and with no information regarding how often they were applied or in which condition.

A relationship between good tags and how often they were applied if weak at best, only for tags used very often, eight times or more. Tags applied a handful of times were not any better than tags applied once or twice.

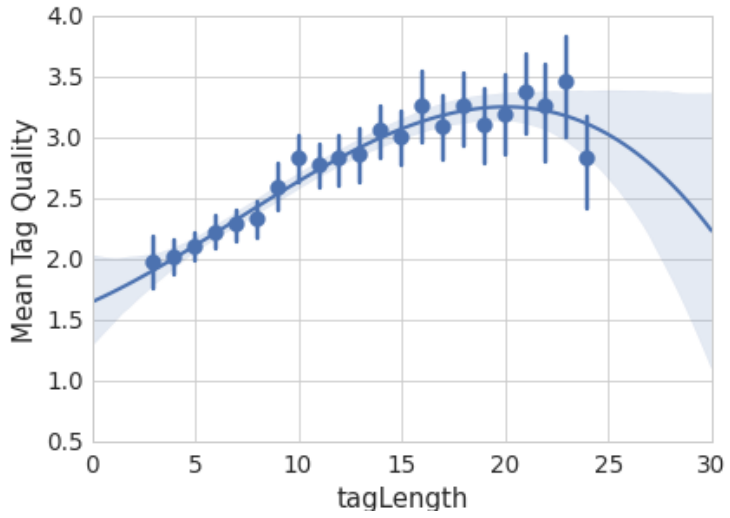

Figure 21: Relationship between the tag length (by character count) and tag quality. N=1976.

There does appear to be a relationship between strong tags and tag length, in that moderate to long length tags are better than short tags and – to the extent that there was data – very long tags. Figure **fig:tagLength2** demonstrates this relationship clearly.

Results

Figure 22 shows the quality distribution of tags applied in each condition, not factoring in user or item effects. Performing a Kruskal-Wallis ranked test comparing the conditions to the baseline, we find that TRAIN and FAST are different distributions than BASE, at $\alpha = 0.001$, but we fail to reject the null hypothesis of equal medians for FDBK.

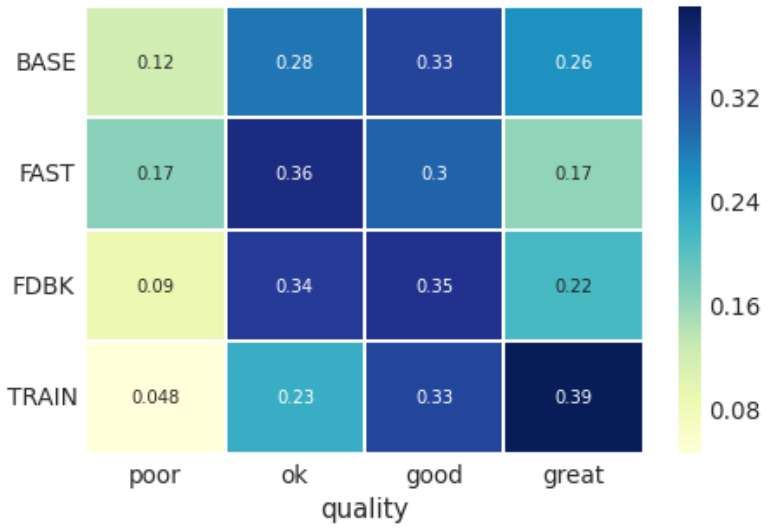

Figure 22: Proportions of tags which were poor, okay, good, or great, by condition.

Another approach to understanding the quality of tags is in averaging the ratings from 1 ('poor') to 4 ('great'). This is an imperfect assumption, given that distances between ordinal categories are not perceived in a perfectly linear manner, but it provides a general sense of each condition's performance. Controlling by item, Figure 23 shows the distribution of average qualities for each image, with accompanying statistics in Table 9. Each datum represents the average quality for one of the 100 tags studies; for example, a median of 3.00 for the TRAIN condition means that for

half of the tags, you are likely to get a tag that is at the 'good' end of the scale.

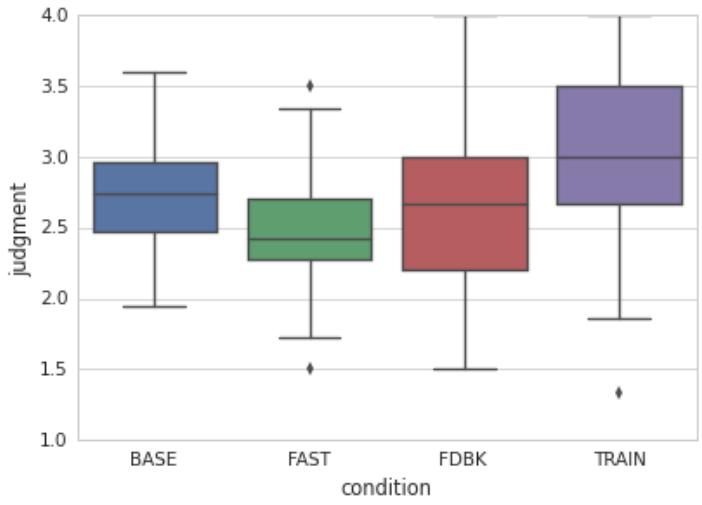

Figure 23: Distribution of each item's average tag quality.

condition	mean	median	std
BASE	2.739	2.738	0.353
FAST	2.457	2.420	0.355
FDBK	2.648	2.667	0.580
TRAIN	3.071	3.000	0.591

Table 9: Statistics for average item tag quality.

In completion time, the most drastic difference was in the fast condition, with a median tagging time of under 7 seconds. This contrasts with the earlier relevance judgment experiment. Like the earlier experiment, however, the feedback condition also shortened workers' per-task completion time.

The relative difference in per-item tagging tasks is shown in Figure 24. To avoid misleading graphics due to difficulties in controlling for N, the kernel density estimates are shown rather than counts. Additionally, the associated measures for each distribution are shown in Table 10.

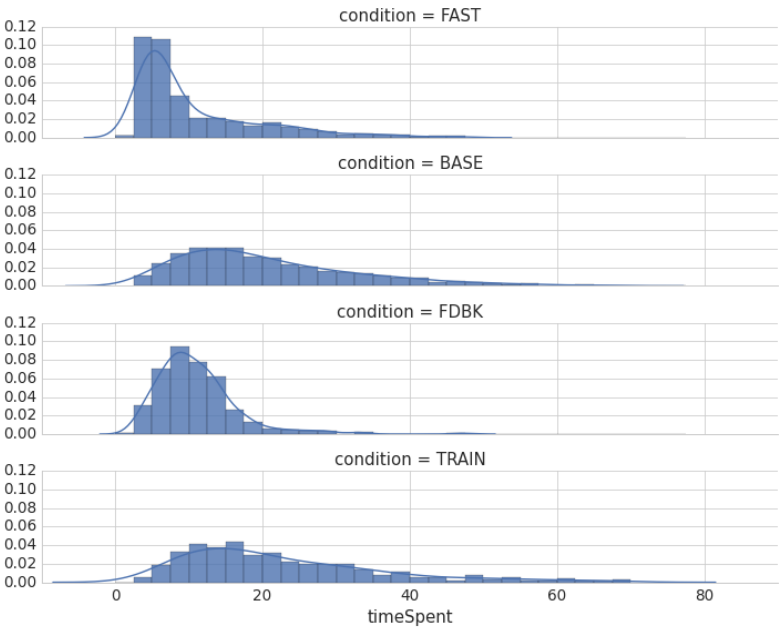

Figure 24: Distribution of time spent, per tagged item, for each condition. Kernel density shown rather than nominal counts, to account for variations in N.

condition	mean	median	std	N	sig
BASE	22.669	18.270	17.523	866	/
FAST	11.459	6.940	11.461	500	***
FDBK	11.019	10.137	5.566	308	***
TRAIN	25.027	19.846	19.875	490	*

Table 10: Metrics for time spent per task, in seconds. Significance marks rejection of equal distribution to the baseline (Mann-Whitney U, Bonferroni-adjusted significance at 0.05 - *, 0.01 - **, and 0.001 - ***).

Comparisons to the baseline rejected equality with the baseline, though for the slightly slower TRAIN, this was only at $\alpha = 0.05$.

The rated satisfaction of the task by workers is shown in Figure 25, and their satisfaction with the payment are shown in Figure 26.

Figure 25: Task satisfaction ratings for tagging task.

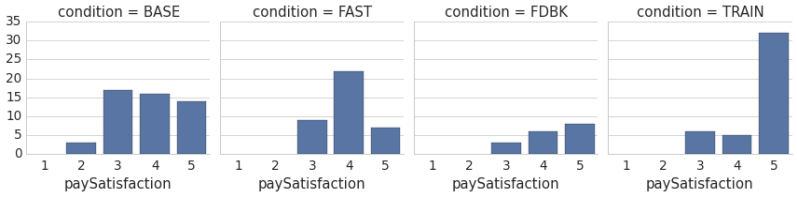

Figure 26: Pay satisfaction ratings for tagging task.

Satisfaction with both the task and pay are similar. The TRAIN condition is heavily skewed upward, while the basic interface received notably poor feedback.

Analysis

The training condition was very appropriate for this type of task, clearly improving on the quality of tags received as well as worker satisfaction. These finding are promising because the implemented approach to training only requires extra effort in designing the training task set, and extra costs for the one task set, per worker, where they are not contributing new information on unknown images. However, beyond this, there are no ongoing costs. An alternative to implementing the training set as its own task set is implementing it as a 'Qualification Test' rather than a paid task – a function of Mechanical Turk (and similarly implemented on other platforms) that allows a test to be used in order to assign a custom qualification, against which task sets can be restricted.

The disproportionally high satisfaction scores for TRAIN are notable, because they lend insight on worker needs. Note that

the measurement of TRAIN is through the post-training task sets, which are identical to the baseline. The only different is the priming task that preceded them. In context, it seems that workers prefer the close guidance, to confirm for them how the task should be completed.

Unlike the relevance judgment task type, performance feedback was inefficient for tagging. While the satisfaction response rates were low for this condition, they did not indicate any discontent with the task.

Instead, it is likely that this is not a task where unguided feedback is the intervention that is necessary. By design, the feedback condition does not tell a worker the reasons for their rank, instead trying to coax them to read the instructions closer if they are unsatisfied with their rank. However, the tagging task was less structured, and workers might need inspiration more than feedback. It would be interesting to conduct the feedback task in concert with training.

The FAST condition resulted in contributions received twice as quickly, though at a quality loss. Recalling Figure **fig:tagLength2** and the relationship between good tags and length, the types of tags contributed by the time-limited condition skewed toward the short end (Table 11).

Table 11: Mean tag character and word length of tags in each condition.

condition	Mean Length	Median Length	Mean Word Count	Median Word Count
BASE	10.223	8	1.682	1
FAST	9.014	7	1.524	1
FDBK	7.798	7	1.260	1
TRAIN	13.596	13	2.143	2

In a circumstance where shorter tags are useful, the time-limited condition might be preferable for its ability to collect contributions twice as fast.

Discussion

Answer: *RQ 2.1*

> **RQ 2.1**: Which approaches to collection interface design are worth pursuing as alternatives to the basic designs commonly employed in paid crowdsourcing?

With regard to the design space question, the start of this chapter explored the possibilities for collection task design, and looked at promising past works such as games-based tasks (Eickhoff and Vries 2012; Ahn and Dabbish 2004) and bonus-based payment manipulations (Mason and Watts 2010).

The approaches that were chosen to be pursued focused on attention and awareness, measuring how design can refocus attention on instructions, inform workers of problems, or push workers to more instinctual forms of contribution. The lower performance of contributions in the time-limited interface without sufficiently compelling improvements in time (except perhaps in the tagging condition) and satisfaction suggests to future work is better focused on the former types of task design.

Answer: *RQ 2.2*

> **RQ 2.2**: Is there a significant difference in the quality of crowd contributions for the same task collected through different collection interfaces?

With regard to the primary data question, this chapter shows that a notable degree of data influence exists in the design of the task. The type of design manipulations that are fruitful vary based on the needs of the workers, and are worthy of discussion here.

When collecting relevance judgments, per-task training interventions and performance feedback significantly improved the quality of judgments, while pressuring workers forward with a timer lowered quality. When collecting image tags, first-task training tasks significantly improved tag quality, while the time-limited interface again resulted in lower quality tags.

In developing and testing the tagging experiment, it became apparent how uncertain one feels when asked to be creative on demand, at least compared to the structured nature of relevance

judgments. Despite careful effort in describing what a good or bad tag is, workers may have been similarly less confident about their contributions. The drastic improvement in satisfaction in the training task certainly seems to suggest that guidance is desired, and the increase in mean tag length (Table 11) shows that trained workers grasped latent indicators of quality.

Confidence in the tasks seems to point to part of the difference between relevance judgments and tags. People liked the relevance judgments, in that they were quick and with few hiccups. It is not a particularly fun task, but it is dependable - at least in the image-centric context presented. The relevance judgment form was already naturally optimized, to the extent that the time-limited interface measure actually slowed down workers.

Training was useful in both conditions, but in different implementations. The first-taskset training condition was helpful in assisting tagging workers in understanding the requirements because all the tasks followed the same convention. For relevance judgments, a smaller training condition at the start of each task was more effective, likely because the context – i.e., the query – would change between task sets. This may be a differentiating factor for applying training to future conditions: can all task sets be described with the same instructions?

Performance feedback was effective for image relevance judgments but not for image tagging. With image tagging, the failure may have been in not provided the right type of support: whereas the goal of the feedback interface was in measuring awareness, workers seemed to need guidance. This was by design for this experiment, but it remains to be seen whether a paired approach applying both training and feedback would improve over the gains seen in the training interface.

In a finding contrary to the expectation at the outset of this chapter, performance feedback seemed to encourage quicker contribution. The reason for this is not clear, but one reading of the data suggests that better workers were encouraged to perform fast by the validation, enough to counteract workers that slowed down.

Since the start, the time-limited condition was an outlier, to compare a drastically different approach to design than the other manipulations. Though it was not expected to improve performance, the gains in efficiency and worker interest may have balanced that out. However, for both tasks, the quality of contributions fell. In different contexts, one can imagine this interface being helpful: in cases where the faster contribution style did not lower collection quality. For example, a less restrained tagging task might be better suited; indeed, this is similar to the fast tagging mode of the ESP Game (Ahn and Dabbish 2004).

Answer: *RQ 2.3*

> **RQ 2.3**: Is there a qualitative difference in contributor satisfaction across different interfaces for the same task?

Workers did express different satisfaction with different interfaces. The purpose of formalizing this secondary research question was to stay sensitive of possible negative effects in worker experience accompanying the design manipulations. This was not found to be the case, and most shifts in satisfaction improved on the basic interface. This many suggest that workers find variety refreshing, but it can also simply be a self-selection effect, where the only respondents were the ones that took note of the shift away from an archetypal task.

For relevance judgments, none of the conditions adversely affected worker satisfaction, with the time-limited and performance feedback improving satisfaction. For image-tagging, the most notable change in task satisfaction was in the tasks following training, meaning workers were happier in addition to better performers.

For both tasks, workers that were told they were better workers in the feedback condition also exhibited lower pay satisfaction. Mason and Watts (2010) have previous found that the perceptions of a task's value are elastic, and this quirk is worthy of future study.

Conclusion

This chapter measured three design manipulations, one in two variants, against two control tasks appropriate for information science. Notably, it was found that training interventions improved data quality at little extra cost, while performance feedback improved over a baseline in circumstances where workers were capable of self-correction.

These findings can support future work on crowdsourcing design, as well as informing practical applications for tasks where there is an objective truth or a reasonable expectation of agreement. The next chapter presents one real-world application of crowdsourcing, in an experiment that applies both post-collection corrections and collection-time design changes.

Designing Tasks for Objective Needs 2

The next chapter will turn to more subjective settings. Before continuing, however, it is worth reported on one more study of objective task design, one which has been extracted to its own section because the study looks at both posterior data correction and task design corrections, as discussed in the two chapters prior, while the findings bridge the shift in focus from objective task design to subjective.

Judging the similarity of audio is a difficult and time-consuming task. Since 2006, the Music Information Retrieval Evaluation eXchange (MIREX) has been using volunteer human workers for evaluating the performance of music systems submitted to the Audio Music Similarity and Retrieval (AMS) task.[58]

After analyzing four years of crowd judgments from AMS, finding that the consistency across different raters and years is remarkably poor, this chapter looks at the role of crowdsourcing design and modeling choices in this data variable. Following from the previous chapter and especially the first half of this chapter, the low intercoder consistency is tackled from both a collection approach perspective and a post-collection perspective. Specifically, user normalization, collection instrument design changes, and multiple independent judgments are pursued.

The primary contribution here is a better understanding of data issues that stem from crowdsourced music evaluation datasets, and methods to avoid data quality pitfalls. Particularly, our case study of music information retrieval judgments generalizes to a class of evaluation tasks that are subjective-biased.

[58] A version of this work was previously presented at JCDL 2015, with co-author J. Stephen Downie (Organisciak and Downie 2015). Copyright held by ACM, permission provided for dissertation reuse.

Music similarity is desired by music digital library users (Lee and Downie 2004), and other digital libraries deal with a comparable form of *normative* task where there is no absolutely correct ground truth but a desire to reach a consensus or a generally agreeable classification; e.g., item similarity ratings, information quality judgments, and information retrieval relevance judgments. The findings are also important to understanding the reliability of Audio Music Similarity evaluation, and we provide recommendations to improve future tasks.

Problem

MIREX is an annual evaluation event where techniques tailored to a variety of Music Digital Library (MDL) and Music Information Retrieval (MIR) tasks are submitted by research laboratories from all over the world.

The Audio Music Similarity and Retrieval (AMS) task was started in 2006. AMS resembles a classic information retrieval scenario, whereby the systems being evaluated are expected to return a ranked list of audio items that are considered similar to a given query (Downie 2003). It is also desired by digital library users: in a survey of MDL users, 54% said they were likely to use music similarity functions (Lee and Downie 2004). AMS relies on human judgments for evaluation, recruiting volunteers each year to judge the similarity of song "candidates" to randomly selected queries.

For each query song, each retrieval system under evaluation gives MIREX a list of candidate similar songs. These query–candidate sets are presented randomly to evaluators in a judging system called 'Evalutron 6000' (E6K) (Downie 2006; Gruzd et al. 2007). To avoid exhaustion, E6K saves judgments continuously, so that workers can step away and return without losing data.

The problem looked at in this chapter is that there is a lack of agreement between workers in song pairs judged across multiple years. Since our best prediction for the true simi-

larity of two songs is the mean of both judgments, we can measure the deviation from the expected value as Root Mean Squared Error (RMSE). In this case, $RMSE = 16.58$ against a prediction assuming symmetric similarity.

RMSE is generally meaningful in comparison, but for a sense of the variance with an RMSE of 16.58, consider that it is in the same unit as the scale, which only has a max range of 101 points[59]. Alternately, the FINE scale judgements are plotted in Figure 27, which shows this variance clearly. The slope shows the expected relationship if similarity was an agreeable metric independent of "which song is listened to first" order effects – an assumption implicitly made in treating similarity as something that can be evaluated.

[59] It may be helpful in approximating the severity of the problem to remember that RMSE aligns with sample standard deviation in a normal distribution. No assumptions are made about distribution in this case, but in a normal distribution, a range of about 53 points on the scale would be required to represent 95% of contributions. Alternately, the RMSE of uniformly distributed random judgments would be approximately 40-41.

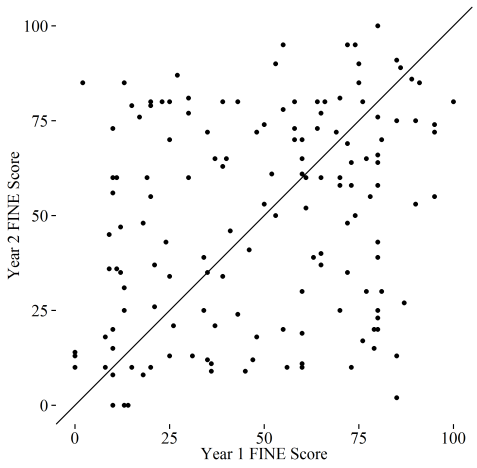

Figure 27: Audio similarity judgments for (Song x, Song y) pairs judged in multiple years.

The noise presented here suggests a great deal of circumstance and randomness in evaluating music similarity algorithms for MIREX.

Comparing the BROAD category of reciprocal pairs tells a similar story (Table 12): only 35% of workers agreed on the category and nearly half was agreement on "somewhat similar" item. While some of this is to be expected, it also suggests that *SS* functions as a catch-all category where workers hedge their bets. This is supported by its much wider range (Figure 28).

	NS	SS	VS
Not Similar (NS)	5	—	—
Somewhat Similar (SS)	20	14	—
Very Similar (VS)	10	21	8

Table 12: Relationship of categorical judgments for pairs of songs that have been judged twice over a four-year span of AMS.

The weak correlation in re-judging makes it difficult to assess the extent to which the evaluation is actually reflecting the 'truth' of what songs are similar.

What are the reasons for this weak correlation? This study considers this question in the context of crowdsourcing choices, looking at collection format and data treatment as possible sources for the variance. First, lets consider some possible explanations.

- **Order and priming effects**. Perhaps there is an order effect based on either which song a worker listens to first, or a priming effect caused by a worker listening consecutively to a set of song pairs with the same query. Research in other contexts has noted the possibility of asymmetrical effects (Tversky 1977; Polk et al. 2002; Hiatt et al. 2013).
- **Different interpretations of the scale**. Do different people treat the rating scale differently? This would be a user bias, but a predictable one.
- **Bad intercoder reliability due to task design**. Perhaps the E6K system does a poor job controlling for consistency?
- **Bad workers**. Much crowd research looks at malicious or unreliable workers. This is possible, but less likely to happen systematically since the volunteers are trusted members of the MIR community.
- **An inherently subjective task**. Does this task present challenges to agreement?

It is likely that the noisy, high-variance MIREX music similarity judgments stem from multiple sources. In line with the thrust of this dissertation, I focus on measuring how much of that is recoverable: what can be improved by changes to practice. The rest of this chapter will at consider 1) corrections for user-specific biases, 2) multiple-keyed judgments, and 3) a task design. While order effect are not focused on, partially because their measurement is possibly confounded by the other issues, this chapter's positive results – showing improved judgment consistency – provide a better sense of the magnitude at which such effects might exist.

Related Work

The feasibility of scoring melodic similarity has been challenged by Marsden (2012), who noted high variation in MIREX 2005 similarity judgments. Though on different MIREX data, our study is

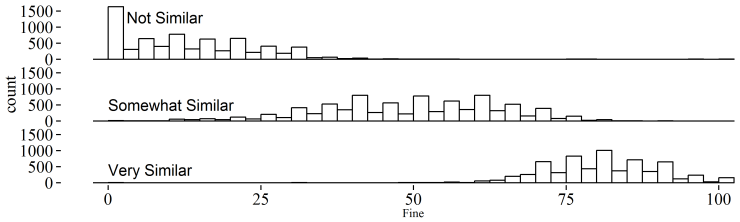

Figure 28: Distribution of FINE scores relative to BROAD categories.

able to identify collection instrument design as one such factor source of variance.

An alternative to the form of graded similarity judgment that MIREX uses is partially ordered lists, introduced by Typke et al. (2005). This form of judgment has been shown to be effective for judging the relative similarity of candidates to a query. However, it is more time-consuming to create, a factor in the decision to use a graded scale for MIREX. Also, it suffers from similar inconsistency problems to what we observe in this paper (Urbano, Morato, et al. 2010).

Despite the inconsistency observed in this study, research into the power of AMS evaluation for a year not overlapping with our study concluded that the relative rankings of AMS systems in MIREX are sound, with contention on about 4% of pairwise system comparisons (Urbano, Martín, et al. 2011). This chapter turn to Mechanical Turk for additional judgments, an option shown as a viable approach to music similarity judgments in multiple studies (Urbano, Morato, et al. 2010; Lee 2010).

Urbano, Morato, et al. (2010) looked to paid crowdsourcing for lowering the difficulty of finding human labor for *ranked* similarity judgments. They use an approach similar to ordered lists, inferring an order through pairwise preference judgments, whereby workers choose the more similar option between two candidates. Additional research has also looked at *graded* similarity judgments in the context of the AMS task (Lee 2010), finding that the MIREX style of evaluation does not suffer significant drops in quality with paid workers.

Data

26024 human judgments of audio similarity were compiled, comprising four years of MIREX's AMS judgments. The candidate songs were selected for judgment by 8 submitted systems in 2010, 18 in 2011, 10 in 2012, and 8 in 2013. Until 2011, 100 queries were evaluated each year, after which MIREX shifted to 50 queries per year.

All the candidates for a query were graded on two scales of similarity:

- The **BROAD** scale is a categorical ranking from three choices: "not similar", "somewhat similar", and "very similar".
- The **FINE** scale is a 101-point numerical rating, from 0-100.

The workers were generally trusted volunteers from the MIR community, and multiple keying was not done.

To understand the consistency of judgments across years of MIREX, we need to look at song pairs that have recurred in judging. Since AMS evaluation queries are randomly selected each year, there are only two instances where a query has recurred. However, 80% of queries have also occurred as candidates for other songs. As a result, there are 156 judgments of the same song pairs across the years, with the caveat that the query-candidate relationship is inverted.[60]

[60] Indeed, the initial spark that led to this study was a curiosity in whether the query-candidate assignment of a song pair – i.e., which song is presented as the query, which song is presented as the candidate – was meaningful. By studying other possible sources for the variance in the data, as will be seen, a significant portion of the error was accounted for, not precluding but certainly limiting the potential magnitude of a query order effect.

Approach

To address possible sources of the error in MIREX's crowdsourced relevance judgments, four approaches are taken:

1. Normalizing workers by their personal habits;
2. Asking new, different workers for judgments;
3. Adding additional redundant workers;
4. Testing an alternate interface that gives workers more guidance on what rating is appropriate.

Within the larger dissertation, these correspond respectively to the following research questions:

- *RQ 2.4*: Are intra-worker inconsistencies responsible for the lack of reciprocation in AMS similarity judgments?
- *RQ 2.5*: Are problem workers responsible for inconsistent reciprocal ratings?
- *RQ 2.6*: Is subjectivity or disagreement of the grading task responsible for inconsistent reciprocal ratings?
- *RQ 2.7*: Does the task design affect the quality of judgments?

Normalizing for Grader-Specific Effects

> *RQ 2.4*: Are intra-worker inconsistencies responsible for the lack of reciprocation in AMS similarity judgments?

The human workers are given a large amount of leeway regarding how they perform a task. We set out to see if this contributes to superficial variance, and whether correcting for it can address the poor reciprocation in AMS. While the BROAD categories are fairly clear, the FINE scale does not constrain workers to follow a specific codebook. This is appears to be done by design: workers are told,

> You have the freedom to make whatever associations you desire between a particular BROAD Category score and its related FINE Score. In fact, we expect to see variations across evaluators with regard to the relationships between BROAD Categories and FINE Scores as this is a normal part of human subjectivity.

Instructions continue to suggest that workers apply a level of 'reasonableness' regarding what is intuitively sensible. For example, a low FINE score when the BROAD category is 'very similar' is not reasonable.

This type of error is commonly seen in collaborative filtering for recommendation, where users' opinions are often treated as a mixture of their nominal rating, adjusted by user-specific and item-specific biases (Koren 2009). To normalize workers against

their specific biases, FINE judgments were translated to z-score values, represented as standard deviations from the worker's mean rating habit. This approach was previously seen in Hofmann (2004); in our case, adjusted ratings were blocked by a worker's BROAD score, resulting in three values for each worker: deviation from their typical FINE score for "not similar", "somewhat similar", and "very similar" candidates.

The adjusted rating $r'_{u,b}$ for worker u and BROAD category b was calculated in the following way:

$$r'_{u,b} = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} r_{u,b} - \mu_{u,b}}$$

where $b \in B$ and $B = \{"NS", "SS", "VS"\}$. Since this normalization provides ratings against three different scales, we mapped it back into a new FINE score by assuming a normal distribution for each category. With this mapping, 95% of not similar ratings occur between $FINE = 0 - 27.63$; somewhat similar ratings between $30.21 - 67.80$; and very similar ratings between $68.74 - 92.17$.

Results

Answer: *RQ 2.4*

Normalizing user FINE judgments weighted against their BROAD judgments resulted in variance of $RMSE = 16.15$, a non-significant change. Thus, there is no evidence that greatly different internal scales by workers were the reason for the low consistency. In other words, the notion that workers were internally consistent in a way that can be normalized globally is not tractable.

Verifying Judgments with New Graders

> *RQ 2.5*: Are problem workers responsible for inconsistent reciprocal ratings?
>
> *RQ 2.6*: Is subjectivity or disagreement of the grading task responsible for inconsistent reciprocal ratings?

Would the same low consistency be seen if new workers were asked? Getting a second opinion addresses two possibilities: ex-

pected error (good workers, biased task) and unexpected error (agreeable task, bad workers).

To answer these two research questions, 156 tasks were posted on Mechanical Turk. Asking paid workers *individually* provides an insight into MIREX worker quality, while asking *multiple* workers helps to see if it is simply a task that is not easily agreed upon, regardless of how well-intentioned a worker is.

In parameterizing the task for this study, worker workers were presented with a query and a single candidate. The audio files were the same clips used in MIREX.

Restrictions were not placed how fully the clips were listened to, and in fact the average task time was lower than the length of the clips. The task was carefully designed to mimic the question phrasing and level of guidance from the original task. As a result, Turk workers are potentially less fatigued (Lee 2010), but may also be less experienced. This was done both due to the conventions of Mechanical Turk and because our MIREX data was not rich enough to emulate the order or continuity of task sets. Thus, any priming effects from the series of songs would not translate here.

Results

Answer: *RQ 2.5*

Asking individual paid amateur workers to provide judgments yielded an average $RMSE = 15.53$, a comparable level of inconsistency. With regards to *RQ 2.5*, the low consistency when asking a new group of workers for judgments suggests that the MIREX volunteers are not unreliable compared to other workers.

Answer: *RQ 2.6*

In contrast, aggregating multiple worker judgments toward a normative opinion results in drastic improvements: aggregating two workers by mean judgment improved the RMSE to 9.72 (41.4% improvement), while three worker judgments improved the RMSE to 7.45 (55.1% improvement). This means that, as asked in *RQ 2.6*, the task is too subjective to trust a single worker and has a high natural variance in judgment.

Approach	RMSE
Baseline (AMS Graders)	16.58
Normalizing Graders	16.15 (-0.03%)
Second-opinion (Individual turk workers)	15.53 (-6.3%)
Aggregating workers: 2 votes/judgment	9.72 (-41.4%)
Aggregating workers: 3 votes/judgment	7.45 (-55.1%)
Alternate design (individual judgments)	11.44 (-31.0%)
Alternate design (2 votes/judgments)	7.55 (-54.5%)
Alternate design (3 votes/judgments)	5.40 (-66.1%)

Table 13: Deviation (in RMSE) of similarity judgments from expectation.

Improving Task Guidance

RQ 2.7: Does the task design affect the quality of judgments?

One of the threats to grading reliability is a hard to understand or poorly defined coding scheme (Neuendorf 2002). Following from earlier discussion, we turn to the effect of a task's design on the consistency of judgments by evaluating a different collection interface.

New judgments are again collected on Mechanical Turk. In contrast to the previous evaluation's fidelity to the original collection interface, here the task design is changed to more carefully guide workers.

Previous literature notes that the similarity ratings can be biased because the perceived distance between points in a rating scale is not linear, and word choice can affect interpretation of the task (Katter 1968; Eisenberg 1988). This motivated us to measure

some changes to the rating scale: BROAD scores were no longer collected, and FINE scores gave textual descriptions for ranges of the 0-100 scale, serving as anchors. We also tested this interface with colloquial language to make the instructions more broadly accessible, with the wording shown in Table 14.

Table 14: The colloquial wording presented to workers in the alternative task interface.

Range	Description
0-20	The candidate couldn't be more different from the query.
20-40	The candidate is not really similar to the query song.
40-60	The candidate doesn't sound like the query too much, but shares some themes
60-80	The candidate has a similar sound or feel to the query song
80-99	The candidate sounds like the query song.
100	They are the same song!

Results

Answer: *RQ 2.7*

When workers used the modified FINE rating scale, they averaged an RMSE of 11.44. In light of the gains observed earlier with multi-worker aggregation, this interface was also looked at in conjunction with 2- and 3-worker judgments, which yielded additional improvements still: respectively RMSE=7.55 and 5.40. As seen in Table 13, this means that the alternate design offered consistent per-worker improvement without increasing cost.

Discussion

The poor consistency in crowdsourced similarity judgments in MIREX results can be greatly attributed to difficulties inherent to the task of grading music. This set of experiments show that MIREX does not have a problem with poor or misguided workers.

However, notable improvements to the evaluation data quality can be made by changes to the collection and treatment of judgments. For AMS and similarly semi-subjective tasks, there are two changes that can be implemented to greatly improve the evaluation quality:

Collecting multiple judgments. Despite the added complexity or cost of collecting multiple judgments for each query-candidate pair, it is an important step toward collecting consistent results. While finding enough volunteer workers in the MIR community is a restricting factor, amateur paid crowds offer similar performance (Lee 2010; Urbano, Morato, et al. 2010) and may be one way to augment the volunteer judgments.

Providing a more specific codebook. While it is important to acknowledge the subjectivity of similarity judging, providing structure for workers to anchor their interpretations into a score improves the reliability of their contributions.[61] Unlike multiple judgments, these sorts of task design changes do not add to the cost of evaluation.

For the benefit of further study, it would be also beneficial for MIREX to retain information about judgment order and time taken for each judgment. While the poor consistency is improved through multiple judgments and stronger instructions, an outstanding question is whether a worker's approach to a task evolves over time.

Normalizing for systematic user-specific biases did not improve the consistency of the data. However, when workers were provided a rating scale that gave them more guidance, they performed better. Why did the former not improve consistency, while the latter did? One possibility is that, in addition to intra-coder differences in interpreting the FINE scale, workers were also internally less strict, something that the task design might have corrected.

[61] There was purportedly great discussion at the conception of the AMS task around the expected subjectivity, which may have motivated the loose instructions stating that "we expect to see variations across evaluators... as this is a normal part of human subjectivity." In light of the results here, I would argue that there is a confounding between natural, expected subjectivity of the task, and artificial variance stemming from the treatment of the task itself.

Conclusion

Finding human workers for a time-consuming task is difficult. However, since music similarity tries to derive a consensus for a quality that people do not always agree on, it is imperative to collect multiple judgments for reliable evaluation. Judging music similarity is normative: it does not have a clear truth but it is possible to strive for a rough consensus that strives to satisfy most opinions. This type of task is important to building better information systems: it can apply to certain contexts of information retrieval relevance, or ratings of item quality in online collections, or even in crowd-curated lists, but as we found with audio similarity, it is important to treat it as such.

The next chapter follows this thread further, to highly subjective contexts. It considers how to collect subjective data on Mechanical Turk and introduces two protocols for designing such tasks.

Designing Tasks for Subjective Needs

Not all crowdsourcing uses have a common goal or objective. There are many needs that differ from person to person, and access to large crowd of diverse people can help us in parsing the variant needs of an individual. In recent years, information systems have figured out how to successfully incorporate large-scale feedback from others for purposes such as good movie or product recommendations (Koren 2009; Linden, Smith, and York 2003), to personalize web search (Noll and Meinel 2007), or even to support specific needs in crisis situations (Vieweg et al. 2010). To do this successfully, these systems must account for the fact that not all people want the same thing. Given enough behavioral data, systems like Netflix and Amazon have been able to successfully personalize their content to individual users by identifying other related users and showing them what those users have consumed.[62]

While large-scale approaches to personalization have been successful, they can only be applied to cases where significant behavioral data already exists. For example, Netflix can do a good job recommending popular publicly released movies, but would have a much harder time recommending content from a small private collection.

With growing access to real-time human workers through paid crowd markets, a new opportunity to use information from others to address personal information needs is becoming feasible: *personalized crowdsourcing*.

[62] This chapter is a new reporting of work previously presented at HCOMP 2014, with co-authors Jaime Teevan, Susan Dumais, Robert C. Miller, and Adam Tauman Kalai (Organisciak, Teevan, Dumais, et al. 2014). Research was performed for Microsoft Research. This treatment includes additional data reporting, including discussion of costs and qualitative feedback, as well as an additional set of experiments around handwriting emulation. Co-authorship notes in appendix.

Problem

As was established in earlier chapters, crowdsourcing is commonly used for presumed objective tasks, such as for evaluation (e.g., Kiritchenko, Zhu, and Mohammad 2014; Radinsky, Davidovich, and Markovitch 2012). This chapter turns focus to its potential role for subjective uses, and examines how we may organize paid crowds for such purposes.

Recent work has begun to exploring crowdsourcing to solve person-specific problems, such as in travel planning (H. Zhang et al. 2012), document editing (Bernstein, Little, et al. 2010), and email management (Kokkalis et al. 2013). The approach taken by this chapter introduces personalized crowdsourcing as a general solution to this class of problem, applying human computation through paid crowdsourcing for on-demand personalization.

To explore how paid online crowds can be leveraged to personalize for individuals in sparse data settings, two protocols are presented: *taste-matching*, where workers are matched in similarity to the target, and *taste-grokking*, where unfiltered crowd workers are asked to perform a task as if they were the target. It is shown that personalized crowdsourcing is feasible, within the scope of a number of evaluated task types and domains. By studying personalized crowdsourcing for image recommendation, text summarization, and handwriting emulation, this chapter offers insight into the relative strengths and weaknesses of each.

These protocol are introduced in more detail later, but the fundamental difference is that in taste-matching, the system finds people with the same opinions and tastes as the target and asks them for their opinions as a proxy for the target, while taste-grokking asks *any* worker, similar or not, to make an educated guess about what the target would like[63].

[63] *grok*: "Understand (something) intuitively or by empathy" (OED)

The primary goal in evaluating taste-matching and taste-grokking for various problems is to compare the space of possibility for personalized crowdsourcing: what works, what does not, and when. Beyond the underlying philosophies of 'matching'

or 'grokking' being compared in the two protocols, this chapter touches on training set size and selection, tasks of different complexity and in different domains, contribution granularity, and inherent worker skill.

Finally, the details of when each is appropriate are discussed, including task complexity, profiling issues, and the amount of possible subjectivity.

Related Work

The work presented here builds on existing crowdsourcing research, leading to a generalized treatment of previous approaches that have been taken to support subjective crowdsourcing. This section provides an overview of relevant approaches in crowdsourcing, highlighting approaches that are particularly similar to the taste-matching and taste-grokking protocols introduced in this chapter.

Crowdsourcing for subjective tasks is common in volunteer crowdsourcing settings. Some projects indulge in the variability of human contributions for artistic effect, such as the crowdsourcing fan film *Star Wars Uncut* (Pugh 2009) and crowdsourcing music video *The Johnny Cash Project*. More generally, certain patterns of reactive user-generated content are a familiar part of everyday information system use, such as rating, and commenting. Casual online users contribute subjective opinions based on reactions to the content as well as the contributions of other people (Dellarocas and Narayan 2006). This parallels user-generated content contribution in general (Daugherty, Eastin, and Bright 2008), though a complication of volunteered subjective information is that it is biased. For example, many online ratings exhibit a bimodal distribution, seeming to suggest that self-selected contributors tend to be either very negative or positive, with moderate contributors less like to contribute (Hu, Pavlou, and J. Zhang 2006; Dellarocas and Narayan 2006). Similarly, early contributors of opinion ratings or reviews tend to affect later opinions (Li, Yang, and Xue 2009).

Paying workers may lower self-selection biases for subjective tasks. However, the most common uses of paid crowds are in the style of human computation (Quinn and Bederson 2011; Law and Ahn 2011): tasks such as evaluation dataset creation (e.g., Snow et al. 2008; Novotney and Callison-Burch 2010; Alonso, Rose, and Stewart 2008). As a result, much literature focuses on issues of reconciling multiple contributions into a trustworthy output (Sheng, Provost, and Ipeirotis 2008; B. Wallace et al. 2011; Eickhoff and Vries 2012);

Though personalized crowdsourcing can be applied in numerous contexts, it is particularly valuable in highly-specific on-demand settings: where a person might not have the time to spend on completing a task themselves, but the subjectivity of their needs alongside the specificity of the task means that there are few alternative options. Some people do not find optimal completion of such tasks to be worthwhile, a factor influenced by the perceived value of their time and their enjoyment of the task (Marmorstein, Grewal, and Fishe 1992). This trade-off is present in areas such as price comparison shopping (ibid) and travel-planning (Gursoy and McCleary 2004). This study does not make any assumptions about where the target person's preferences lie in balancing the quality cost of not personalizing, time cost of completing the task themselves, or monetary cost of personalization.

The underlying assumption in taste-matching is that you can personalize for a person by finding similar people or groups of people, and using them as a proxy for the target person. This mirrors the approach seen in collaborative filtering (Hofmann 2004), one of the most common forms of recommendation. Collaborative filtering is also similarly motivated at a higher-level, by the difficulty to predict people's subjective desires and needs purely by analyzing the content. Where taste-matching differs is that workers contribute data on demand, sidestepping the common collaborative filtering problem of sparse data (Konstan et al. 1997).

The taste-grokking approach pursued in this study looks to

generate personalized content by asking workers to understand the target and guess at their tastes and needs; E.g., guessing a target's opinion on a rating scale. A similar approach was explored by Krishnan et al. (2008), where the MovieLens collaborative filtering system was compared to human recommenders. MovieLens, which functions like a more mature, higher n version of taste-matching, was found to perform better. Where humans did excel was in recommending for targets with eclectic or novel tastes.

Recent work has warned about expected ground truth tasks having subjective components, biasing work around them (Alonso, C. C. Marshall, and Najork 2013). Tasks such as selecting the best frame of a video (Bernstein, Brandt, et al. 2011) or rating the similarity between objects (Tamuz et al. 2011) can be argued to contain person- or worker-specific biases. Some efforts have identified the need to subjective affordances, such as crowdsourced email assistant EmailValet (Kokkalis et al. 2013), travel-planning system Mobi (H. Zhang et al. 2012), and parts of document-editing system Soylent (Bernstein, Little, et al. 2010). In all three of these cases, the systems allow directors to communicate their particular tastes and needs with a natural language description. In the taste-grokking protocol of this study, communication-by-example is used rather than free-text communication, but this explicit approach constitutes another approach to personalized crowdsourcing. While paid crowdsourcing has dealt with subjective tasks before, a generalized approach has not been previously defined.

Personalized crowdsourcing has been pursued by prior projects, but not explicitly framed as such. The work presented here differs from earlier work in providing a generalized treatment of personalized crowdsourcing, comparing different approaches in different domains and subsequently presenting problems affecting general use of paid crowdsourcing. Two protocols are presented for comparison, taste-matching and taste-grokking. Taste-matching has precedent in other areas, though this chapter's implementation differs in applying the concepts to on-demand needs. Taste-

grokking offers a more novel approach relying on critical thinking by workers, capitalizing on the human core underlying crowdsourcing.

Research Questions

This chapter looks at approaches for conducting subjective tasks through paid crowdsourcing, and evaluates both both a collection-time task re-framing approach, taste-grokking, and a post-collection modeling approach, taste-matching.

Formally, the following questions are asked:

> **RQ 3.1**: Is it feasible to apply paid crowdsourcing to subjective problems?
>
> **RQ 3.2**: Does the taste-matching protocol reduce the amount of error in personalized crowdsourcing?
>
> **RQ 3.3**: Does the taste-grokking protocol reduce the amount of error in personalized crowdsourcing?
>
> **RQ 3.4**: How do different types of subjective tasks affect the efficacy of personalized crowdsourcing approaches?

Approach

Two protocols are developed for personalized crowdsourcing: *taste-matching* and *taste-grokking*. Most basically, matching finds similar workers to the target and uses their work as a proxy for the target, while grokking asks workers to build a mental model of the target and manually personalize against it.

Both protocols begin with a minimal profile of the target person.

The profiling set is a set of tasks in the style that the workers will be expected to complete. For example, for one of this paper's settings, where personalized crowdsourcing is applied to predicting a target person's opinion of salt and pepper shaker products on a five-point scale, a set of rated salt and pepper shakers comprise the profiling set. – the target of personalization. In the context of taste-matching, the target profile is compared to worker

profiles constructed against the same items. For taste-grokking, the target profile is used to visually communicate the target person's tastes and needs to workers.

Profile construction is subject to variation from a number of parameters: the profiling set size, the profiling set selection, and the domain. While information content increases with larger profiling sets, working with humans restricts size based on considerations of attention, time, and exhaustion (Rzeszotarski et al. 2013). For the target particularly, the time cost of a large profiling set also works against the time and effort saving goals of many personalization settings.

The items acted upon in the profiling set are another profiling consideration. If a film recommender only asks for a person's opinions of different action movies, for example, the information value of each response will be much lower than if they were to ask about a genre-spanning set of films. Most basically, the profiling set may be selected randomly from the larger task set, as is done in this study, though more complex selection strategies may be done. Later, purposively sampled profiling sets are also considered.

Finally, the domain of the profiling set may differ from the eventual work. In this study, targets perform profiling work that is identical to the eventual personalization work: e.g., rating products or highlighting texts. However, it is conceivable that cross-domain profiling can be used; e.g., taking a target person's book and film 'interests' from a social network and using that to aid taste-grokkers in recommending television shows.

In taste-matching, subjective tasks are performed by human workers that complete those tasks similarly to the target person.

The underlying intuition pursued by taste-matching is that people who perform similarly or express similar preferences on seen tasks will continue to be similar on future, unseen tasks. Taste-matching pursues this notion by asking the target person being personalized for to complete a small task in the same manner as the workers will be expected to complete it. The results of this task comprise a profile of the target person.

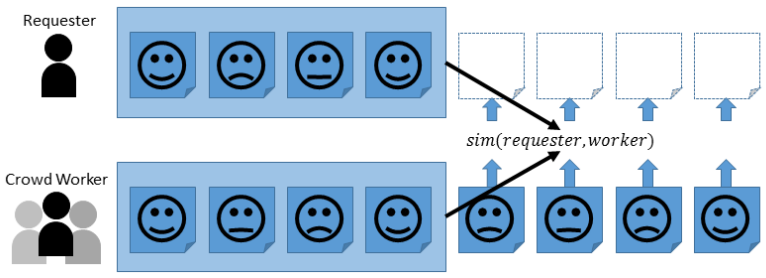

Figure 29: Simplified example of Taste-matching protocol.

Figure 29 illustrates a simple taste-matching setting. As new workers arrive, they complete the same task set as the target person, and a similarity process is used to measure how similar their work is to the target profile. For example, in this study's product image recommendation task, the target and workers both rate their preferences for a set of online shopping results on a five point scale, then matched using Root-Mean-Squared-Error (RMSE) on a normalized version of their rating.

Depending on the similarity of the individual worker to the target person – the taste-match – the system may choose to keep or reject the worker for further contributions.

Even if the intuition of continued similarity is true, there is the issue of adequately capturing the tastes of the target person and workers, and adequately measuring their similarity. This requires proper parameterization and user modeling, and can result in variation between taste-matching implementations. In the settings evaluated for this chapter, efforts were made to base these decisions on precedents and realistic settings, but these are decisions rather than rules. Two such decisions in taste-matching are the method for measuring similarity, and the method for representing worker contributions as personalized content.

Taste-grokking considers a different intuition than taste-matching: that workers can be adept at understanding – or grokking – the needs of a target, even if they are not similar to the target.

'Grokking', a term referring to interpreted understanding, belies the human activity underlying this protocol. Whereas taste-matching performs an algorithmic similarity matching based on

target and worker profiles, taste-grokking leaves the personalization logic to worker interpretation.

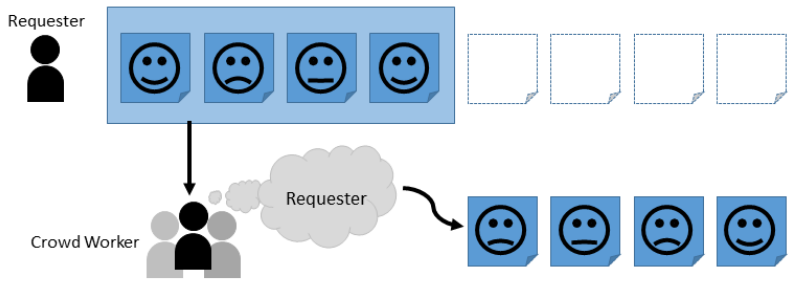

Figure 30: Simplified example of Taste-grokking protocol.

As illustrated in Figure 30, taste-grokking workers are shown a target person's profiling set and asked a variant of the question, 'how do you think the target person would perform the next tasks?' For example, with the product image rating task, workers were shown a target's ratings for a few items, and asked to rate what the target's opinion would be for additional products.

As defined here, taste-grokking uses a 'train-by-example' approach to communicate a target's taste to workers. Other crowdsourcing studies that have personalized by communication have had the target articulate their needs in written instructions, e.g., explaining email priorities (Kokkalis et al. 2013) or travel preferences (H. Zhang et al. 2012). There are potential benefits and difficulties to this approach. Taste-grokking trains by example due on presumed consistency, attempts to minimize target effort, and concerns about the technique's sensitivity to hard to articulate decision factors. While written requests may be dependent on the target's skill, having a target simply perform a small amount of the work does not have this confounding factor. Still, a potential variant of taste-grokking could combine the train-by-example approach with an optional written articulation, as the latter is more explicit about what a target cares or does not care about when effectively executed.

A benefit of taste-grokking is that it realigns a subjective task to a ground truth: all the workers are trying their best to make

sense of the target. Since there is an assumed correct answer, taste-grokking is well-suited to error correction and quality metrics performed in more traditional, non-personalized paid crowd contexts. For example, whereas with matching it is hard to differentiate between a cheating or poor worker and one with eclectic tastes, with taste-grokking a worker that is deviating from the consensus can more confidently be understood as a malicious or poor worker. Additionally, taste-grokking is well suited to multi-contribution aggregation. For example, in a simple aggregation case when predicting opinion ratings, suggestions from multiple workers may be collected for each rating, and the personalized prediction can be a voted rating (mode) or a mean. Doing so smooths over individual errors, although at extra cost.

Before continuing, it is important to make some language clarifications regarding domain-specific words in this study.

- *Target*: In the typology chapter, the beneficiary and director were introduced, to refer to the person benefiting from the work and the person running the data collection, respectively. This chapter will continue referring to the director (called a requester on Amazon's Mechanical Turk) as the person running the experiment. However, *beneficiary* is too generalized. To protect against potential confusion in reporting, a more context-specific term is used to refer to the beneficiary of personalization: a *target person*.

- *Worker*: Though personalized crowdsourcing and the protocols introduced are not necessarily specific to paid crowd contexts, this study focuses on paid crowds in order to control for motivation. For this reason, the crowd contributors are *workers*.

- *Profiling set*: In both protocols, targets are profiled by performing work on a subset of items. This is used generally as a training set, though one evaluation also collects data for cross-validation. To stay consistent and because the 'training' is done in notably different ways in taste-matching and

taste-grokking, this data is referred to as a *profiling set*.

Taste-matching and taste-grokking are used as two possible protocols for personalized crowdsourcing and parameterized in a subset of possible ways, but beyond the question of feasibility, we contribute insights on the broader space, exploring how to maximize the effectiveness on personalized crowdsourcing, how task contexts affects the efficacy of the methods, and the concerns or consequences that follow.

Taste-matching and taste-grokking are evaluated over three problems.

- *Image-based recommendation:* For a familiar, common context, image recommendation is performed. The purpose is to guess a target's opinions on images of

  a) online shopping results (specifically, salt-shakers), and
  b) restaurant meal offerings.

- *Text-highlighting*: Measuring the protocols in a more difficult and complex setting, personalized crowdsourcing is evaluated for text summarized via highlighting. Here, workers highlight film reviews for a target person.

- *Handwriting imitation*: To measure a subjective context that is skill-based rather than opinion- or preference-based, a handwriting mimicry task is performed.

Each of these tasks sees whether an on-demand personalization approach is feasible, through either taste-matching or taste-grokking, for the particular problem.

Experiment #1: Image Recommendation

A popular type of personalization is recommendation, commonly performed in large and subjective topic spaces such as film, music, and literature. This type of task attempts to predict what a target

person will like or prefer. For personalized crowdsourcing, taste-matching and grokking are evaluated in two domains with limited preference data.

In the first domain, online products – specifically, salt and pepper shakers – are recommended based on photographs. The space of salt and pepper shakers is highly variable and expected to be subjective, but prior preference data would likely be sparse, because buying one is likely too trivial for much comparison shopping and few customers would reasonably buy more than one set. For this task, 100 salt and pepper shaker images were used from the US version of the Amazon online store.

In the second evaluated domain, captioned images of restaurant food are recommended. Whereas restaurant recommendations are common through services such as Yelp, Google Maps, or Foursquare, recommendation based on the actual offerings in those restaurants is more difficult, again because it is difficult to otherwise collect enough preference data. Two datasets of 100 cuisine images and names each were collected from Foodspotting.com, one with food from Boston and the other from Seattle.

Worker contributions were collected through Amazon's Mechanical Turk paid crowd marketplace. Workers were asked to rate 100 images in each task set.

For each task set of 100 ratings, workers were paid $1.50. To incentivize taste-grokking workers, who had a performance goal in trying to 'grok' the target person, the taste-grokking remuneration included bonuses paid against ranked performance.

Requesters are first profiled. The profiling set size has to compromise between the system-end value of maximum data and considerations of target effort expenditure and tolerance. For image recommendation, the profiling set was set to 20 randomly selected images. This value was chosen instinctively, though this size-effort compromise is worthy of future study.

The targets were simulated from paid crowd members. That is, a crowd worker contributed 100 ratings, 20 of which were used

as their profiling set, and the remainder used for measuring the effectiveness of recommendations.

Workers in the taste-matching group were asked to rate images based on their own opinion of how much they like the salt-shaker or how appetizing they find the pictured food.

As not all people have the same mental concept of the rating scale (e.g., how much one has to like the item to give it five stars rather than four), taste-matching ratings were normalized ($r \rightarrow r^{'}$) as the deviation from each user's mean rating (Hofmann 2004):

$$r^{'} = \frac{r - \mu_{\text{rater}}}{\sigma_{\text{rater}}} \tag{1}$$

.

Normalization was not necessary for grokking because workers were performing against a target user's world-view rather than their own.

Taste-grokking workers were shown a profiling set of *10* items with the target's opinions. They were told that a single target had contributed the judgments, and were asked to 'guess' what the same target person would think for the subsequent 90 items.

The fact that taste-matching trains a model while taste-grokking trains a human worker means that a realistic parameterization differs between the two. Note that although *20* items had been collected for profiling, the choice of a smaller training set for grokking was motivated by an expectation that it would be too difficult for a person to make sense of too many examples. The other *10* items were kept for cross-validation, to measure whether workers that were notably strong or weak at grokking.

Measurements

Root-Mean-Squared Error (RMSE) was used both for a measuring similarity in the profiling set and for evaluating the quality of recommendations, with a smaller RMSE representing better similarity or a better recommendation.

RMSE offers a measure of how much the predicted opinions for a target deviate from the actual opinions. It is calculated as,

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t-p)^2}$,

where t denotes the target person's true opinion, and p denotes the recommended opinion.

A SYSTEM THAT DOES NOT ASSUME VARIATION across individuals would not personalize, and instead might use the opinions of any worker to make recommendations for the target. This is what comprises the baseline measure, performing at an average RMSE of 1.64 for the salt-shaker recommendation task, and 1.51 and 1.58 for the cuisine recommendation tasks in Seattle and Boston, respectively. An error of 1.51-1.64 on a five-point scale is fairly high and shows that the tasks are notably subjective to begin with.

While considerable variation is seen across individuals, similarities between individuals are found to be consistent over time. This can be seen in the correlation between worker-target similarity on the profiling data and worker-target similarity on the testing data. Where a Pearson correlation coefficient would be 1 if the assumption is correct and the profiling set is perfect at finding people that are similar, the actual collected preference data has a Pearson correlation of 0.73 for the salt and pepper shaker task, and 0.67 and 0.71 respectively for the cuisine in Seattle and Boston tasks. This shows an imperfect but strong correlation. Such a correlation could still exist if the task was not subjective, and is only insightful in combination with variation seen in the baseline RMSE.

Taste-Matching

Table 15: Taste-matching performance for recommendation task.

	Products - Salt shakers	Food - Boston	Food - Seattle
Baseline: Prediction by any worker	1.64	1.51	1.58

	Products - Salt shakers	Food - Boston	Food - Seattle
Best 3 workers overall (top 10%)	0.89 (-46%)	1.02 (-32%)	1.19 (-25%)
Best matched worker from random 5	1.43 (-13%)	1.19 (-22%)	1.26 (-20%)
Best matched worker from random 10	1.35 (-18%)	1.08 (-29%)	1.08 (-31%)

Table 15 shows the performance of recommendations predicted by taste-matching. In both task types, taste-matching improved over the baseline, with stronger gains against the cuisine recommendation tasks.

The parameterizations were selected based on an expectation of a realistic task setting: the best worker from random five and ten. In this setting, a target starts a personalized crowdsourcing task, and n workers are profiled. Based on the profiling 'match', the best of these workers is retained to perform more work as a surrogate for the target. The amount of workers to profile is dependent on a task director's quality-cost trade-off. Though there is no formal expectation of constant improvement, profiling additional workers nonetheless keeps improving performance.

With the parameters used in this study – payment of $1.50 per 100 ratings and a profiling set of 20 items – profiling each worker comes to 30 cents, followed by 1.5 cents for every predicted rating by the matched worker.

Also shown is the average performance of the top three workers overall for each of the thirty target people. This value is included for comparison, of the best possible improvements if the matching process were to successfully identify these workers. For the cuisine recommendation tasks, the matching does in fact work well, given that the best matched workers perform comparable to the ideal. For the salt and pepper shaker recommendation task, the taste-matching improvements are not as strong as the ideal, suggesting

that while good workers are being found by matching, they are not always the *best* workers.

Taste-Grokking

Table 16: Taste-grokking performance.

	Products - Salt shakers	Food - Boston	Food - Seattle
Baseline: Prediction by any worker	1.64	1.51	1.58
Best 3 workers overall	0.87 (-47%)	0.78 (-48%)	0.79 (-50%)
Average individual	1.29 (-21%)	1.53 (+1.3%)	1.57 (-0.5%)
Aggregated prediction (mean, 5 random workers)	1.07 (-34%)	1.38 (-9%)	1.28 (-19%)
Aggregated prediction (mean, 5 top workers)	1.02 (-34%)	1.22 (-19%)	1.13 (-28%)

Taste-grokking improves over the baseline in many settings, but not all. In all cases, it works better for product recommendation than it does for cuisine recommendation, a reversal of what was seen with taste-matching. Table 16 shows the average performance of taste-grokking when performed by any given worker, when aggregated from five worker contributions, and when aggregated from five workers that had been cross-validated as 'good grokkers' from a pool of thirty.

The performance of any single worker's grokking prediction averages an RMSE of 1.29 for salt and pepper shaker recommendation, and 1.53 and 1.57 for the cuisine recommendation. The performance for cuisine recommendation shows no improvement over the baseline; thus, it is tricky to trust only one grokking worker for

these tasks.

More effective than asking one grokking worker is to ask multiple workers and aggregate their predicted ratings. This is sensible because all the workers are striving for the same ground truth, to understand the target, but individuals vary in their grokking proficiency or make occasional errors. For this chapter, workers' recommended ratings were aggregated with a simple mean. Aggregating through the mean of five workers' predictions provides improvements of 34% for salt and pepper shaker taste prediction, and 9% and 19% for the cuisine tasks. The choice to aggregate 5 workers is motivated by a recommendation in Novotney and Callison-Burch (2010) for a different task type but with similar complexity.

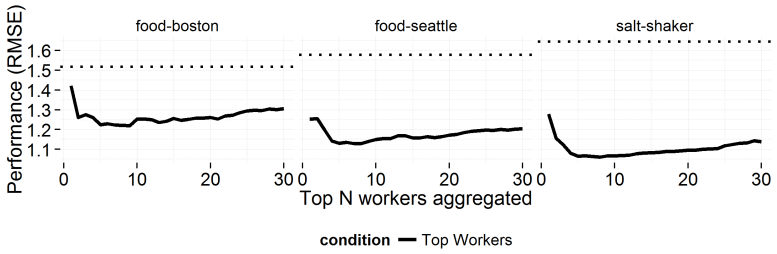

Figure 31: Performance of opinion predictions aggregated from top workers, as determined by a held-out cross-validation set.

Using a cross-validation set of ten ratings to identify and aggregate top workers improves the quality of predictions further, though at the cost and time of collecting additional recommendations. Figure 31 demonstrates the improvement in performance as more of the top workers (e.g., 2^{nd}, 3^{rd}, ..., 30^{th} best cross-validated worker included in prediction). For each task domain there is an increase in performance followed by a gradual decrease, suggesting that even with good workers, aggregation is an important approach for improving quality.

Taste-grokking does not include any fixed costs for profiling as with taste-matching. With the intuitively chosen payment parameters in this study, taste-grokking remains more affordable with a single worker per recommendation. However, while the salt and pepper shaker product recommendation task was well suited for

single-worker taste-grokking, the much improved performance with aggregation means that an ideal taste-grokking setting is more expensive after a few rating predictions.

Worker Behavior

Given that in one protocol workers shared their own opinions of food and products while in the other they had to interpret another person, it was expected the time spent per contribution to be higher for taste-grokking tasks. This was indeed the case, though the median time per grokked rating prediction was not drastically higher: 4.6 seconds (grokking) compared to 3.3 (matching). The range of time spent per item is much larger for taste-grokking, as shown in the box plots in Figure 33. Among crowd workers, it is common to find high-end time outliers due to casual workers that multi-task, but the length of the tail for grokking seems to suggest an additional effect. Though it difficult to know with certainty, it is possible that some workers have particular difficulty with the task or some workers perform the work very carefully, with much cross-checking with the profiling set. Based on suggestions from voluntary feedback that taste-grokking is more interesting to some workers, the possibility of a subset of workers stepping back from time- and profit- maximization to taste-grok is possible.

The time is measured from the start to end of the worker interaction, for a set of 100 ratings for taste-matching or 90 rating predictions for taste-grokking. As such, part of the time spent might be related to the reading of instructions, which may also contribute to differences between the two protocols.

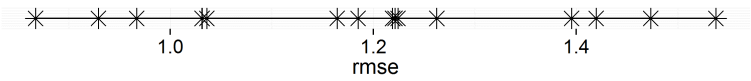

Figure 32: Taste-grokking performance for each individual target x profiling set experiment.

When collecting taste-matching data, a tertiary evaluation was done where workers were not only asked to provide a rating of their opinion, but were also asked 'what is your reason for this

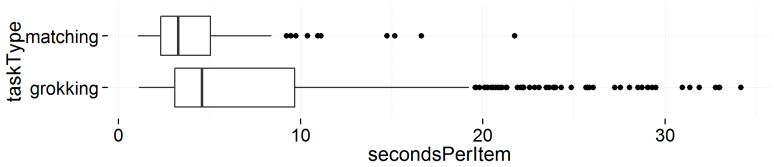

Figure 33: Comparison of time per item rating spent by workers in taste-grokking and taste-matching.

rating?' It was found that by asking workers to contemplate and explicate their reasoning for judgments, their behaviors changed. The mean worker ratings (i.e., each worker's average opinion) were more measured, with workers that were overall consistently negative or positive not represented. This data was not used for the main evaluation, but serves to emphasize that unexpected variance when working with online crowds not only stems from their tastes and needs, but also the contexts in which they contribute data.

For all tasks, an optional feedback form was provided to allow for workers to communicate with us. These provided some qualitative insights into worker satisfaction and task issues.

In general, taste-grokking received more affirmative responses in the style of 'fun!' and 'that was really interesting'. However, it also frustrated workers when the profiling set failed. Particularly for the cuisine task, some workers lamented that the target person's opinions that were shown did not communicate enough about the target. For example, one worker reported, "I think a few more rated pictures would have been helpful in helping me to decide some of the choices, because there wasn't anything that similar in the rated items to those items such as the cappuccino, burgers, hot chocolate." The taste profiling sets were selected randomly for each target person, and taste-grokking was tried with two different set per target, but the salt and pepper shaker tasks appeared less likely to settle with a poor profiling set. This feedback seems to align with the poorer grokking results for cuisine, which will be discussed below.

Figure 32 shows the performance of an all-worker aggregation

for each individual taste-grokking experiment, in order to demonstrate the range of overall prediction quality related to different profiling sets. For comparison, the worst optimized profiling set (described in next section) performance was RMSE=1.05. Figure 34 offers an example of a successful profiling set alongside one where workers performed poorly. The reason the bottom set in Figure 34 performed poorly is only speculative, though it is notable that it did not capture the workers' opinion of 'cute figurines', as the one above did with the cuddling bird and the cupcake salt and pepper shakers. The taste cluster examples in Figure 35 show that this is a large facet of the space. However, of the nine workers that left feedback for the poorly recommended task, none expressed concerns about the training set.

Figure 34: An example of a taste-grokking communication set where workers' subsequent predictions were very strong (top, RMSE=0.87) and poorly (bottom, RMSE=1.47). The examples do not correspond to the same targets.

In the results presented, the items used for profiling and subsequently for profiling in taste-grokking were selected randomly. Random selection can potentially fail when there are many decision dimensions to communicate, as was observed for taste-grokking over cuisine. To measure how robust randomness is compared to alternative selection strategies, an evaluation was completed on taste-grokking over a more purposively sampled profiling set.

The alternative set used using stratified random sampling, from items clustered against opinions contributed by taste-matching

workers. K-means clusters was used, where the number of clusters k was equal to the profiling set size (i.e., k=10). The clusters used as strata for sampling are partially shown in Figure 35.

For the profiling set, one item was randomly chosen from each strata. The intention was to capture the breadth of tastes.

Using an optimized selection of items in the profiling set improved performance greatly over the salt and pepper shaker prediction task. Figure 36 shows the quality of aggregating 1-30 workers; for comparison to Table 16, aggregating 5 random workers gave an RMSE of 1.04.

Figure 35: Example of salt-and-pepper clusters, k=10.

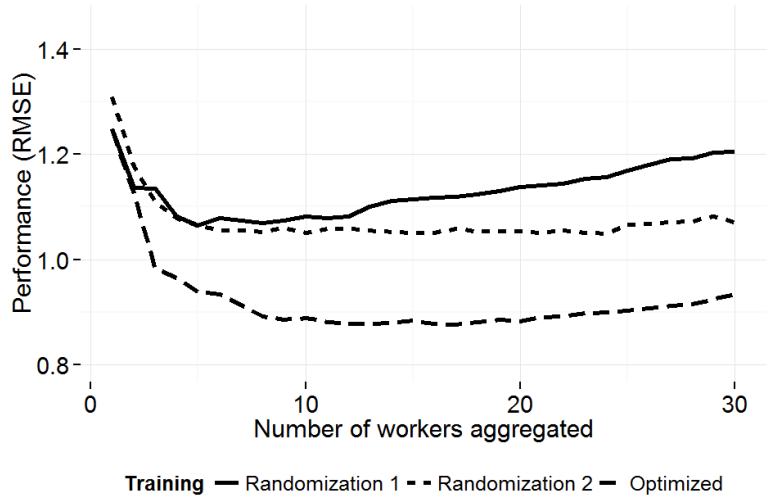

Figure 36: Performance of optimized training set for predicting product opinions, aggregating multiple workers.

Predicting a target person's rating was a suitable application of

personalized crowdsourcing for two image recommendation tasks. Both taste-grokking and taste-matching improved over the baseline. However, the task domains mattered, and taste-grokking was stronger for salt and pepper shaker recommendation while taste-matching was strong for cuisine recommendation. The complexity of the tasks seemed to have contributed to this disparity, where the richer decision space of food resulted in harder-to-understand profiling sets with taste-grokking.

Experiment #2: Text Highlighting

Rating prediction against a five-point scale is an easily controlled task, making it well suited for personalized crowdsourcing. To observe personalized crowdsourcing in a more complex setting, a new task was developed: text highlighting to make film reviews easy to skim. Highlighting texts has much more possible variation (C. Marshall 2000), and involved target-specific *needs* in addition to target-specific *opinions*.

Many settings call for people to digest large amounts of texts, such as in academia, medicine, law, and business. To varying degrees, different individuals may look for different information in the same texts. Can online crowds be leveraged to highlight texts for target persons, for the purpose of summarization?

Film reviews were chosen as a generally useful domain where people having varying information needs and opinions, as well as one that may be interesting to workers. Thus, it is expected to be an easy domain for a difficult task, providing insight into the tractability of text highlighting.

The texts for highlighting were six professional film reviews from *The A.V. Club*, averaging 456 words each.

Workers were recruited on Amazon Mechanical Turk and contributed highlights through a custom web interface. As with the item recommendation task, crowd workers stood in as target persons.

Measurement

The F1 measure is used to measure how similar a worker's highlights are to the targets'.

F1 is the harmonic mean between precision and recall, measured by word overlap. Precision is the proportion of a worker's highlighted words that overlap with the target person's highlights, and recall is the proportion of the target person's highlights that are highlighted in the worker's highlights.

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

A worker that highlights everything will receive perfect recall but poor precision, while a single word highlight that happens to overlap with a target's highlights will be very precise but have poor recall. The goal of F1 is to balance these two measures.

The motivation in using F1 is to capture not only when the same passages were highlighted, but also similarity in brevity: how much is highlighted. Highlighted text that overlaps with the target highlights is rewarded, while irrelevant text is punished.

As the baseline, the average performance of a predicted highlight without matching or grokking was used. The performance of a non-personalized highlighting, measured from 200 highlighted reviews, was $F1 = 0.32$.

Profiling

In a realistic setting, the purpose of personalized crowdsourcing for text highlighting is to lower the effort necessary for a target person to seek a desired set of information. Given such a purpose, profiling is difficult in this instance because highlighting even a single review takes a fair amount of effort. Thus, the size of the profiling set was 2 highlighted film reviews.

Taste-Matching

For taste-matching, workers were asked, "if somebody gave you a summary of this review, what would you like to know to help you

decide if it is movie worth seeing?" Highlights were collected for 50 target persons.

Workers were matched to targets based on the overlap of their highlights with the target's highlights, using F1 as a measure of similarity. One concern with this approach is the conceivable setting where two lines of the review have practically the same information, but the target highlights one line and a worker highlights the other. The relatively short average length of reviews may limit this effect, though a real-world setting would require a more robust similarity metric.

Taste-Grokking

Table 17: Taste-matching text highlighting results.

	Performance (F1)
Baseline	0.32
Best-matched workers	0.39 (+20%)
5 best-matched workers	0.38 (+17%)

Table 18: Taste-grokking text highlighting results.

	Performance (F1)
Baseline	0.32
Any worker	0.30 (-7%)
Best workers (pool of 5)	0.52 (+62%)

Taste-grokking workers were shown a single film review highlighted by a target person and asked to highlight additional reviews for that person. The review that was shown was randomly selected from the two profiling examples that targets had provided, though data was collected for each profiling example.

The broad space, highlighting a custom set of words from approximately 456 words, did not lend itself easily to aggregation. For example, majority voting among multiple workers would be expected to shrink the predicted highlights to less than any in-

dividual's highlights, hurting recall and measured performance. More complex possibilities would require study beyond the scope of this paper, so no aggregation was performed.

Results

Table 17 shows the performance of the best-matched and average performance of the five best-matched workers in each condition. Taste-matching workers that matched well also performed well for the matching.

In contrast, taste-grokking suffered from the difficulty of the task, and the lack of aggregation. The typical recommended highlight was actually *worse* than the baseline. Grokking workers, it seems, over-fit their mental models of the target person, providing worse highlights than if they had simply highlighted of their own accord. Though the improvements from taste-matching show some degree of subjectivity, the poor performance of grokking may also be due to less than anticipated target-specific variance.

For comparison, the best taste-grokking workers from random sets of five are shown. This uses posterior information that was not known at collection, but serves to emphasize the very high theoretical performance of workers with a mean 62% improvement. It seems, while most taste-grokking workers were poor at the task, some very effective 'super-grokkers' were observed.

Experiment #3: Handwriting Imitation

Despite adopting the term 'taste' in taste-matching and taste-grokking, there are many subjective or context-specific tasks that do not refer to taste but nonetheless may be suited for personalization. Two additional areas may be characterized as those affected by *style* and *biases*. By way of example, a small handwriting imitation study was performed, looking at the ability of strangers to personalize text in a person's handwriting. This is more a question of style than of taste.

Handwriting samples and imitations were collected from targets

and workers in person, rather than online, to avoid differences in pen and paper type biasing the evaluation. The hypothetical goal of such a system would be to produce a sample of arbitrary text in something that looks like the target's handwriting. For simplicity, a single phrase was focused on. The *training phrase*, "The quick brown fox jumps over the lazy dog," was used to communicate the target's handwriting style. The *target phrase*, "Wizard's hex," represents an example of arbitrary text one may desire.

Nine targets each provided a profiling phrase sample in their own handwriting, as well as a target phrase sample for evaluation.

Measurement

Similarity evaluation here raises two interesting points. First, unlike ratings, where similarity differences can easily be evaluated numerically, similarity in handwriting samples is more difficult to judge automatically. Hence, paid crowd workers were recruited to judge handwriting similarity.

Second, similarity can be judged across examples and from a single sample pair, even if the text is different. This obviates the need to rank overall similarity between users as one can simply use similarity between sample pairs.

To evaluate similarity between any given sample pair of training and target phrases, one hundred workers on Mechanical Turk were shown the two samples. Regardless of whether the test sample was written by the target (i.e., the person who also wrote the training sample), an imitator, or an independent person in their own handwriting, the worker was asked, "Do you think these two samples were written by the same person? (Y/N)". The proportion of one-hundred evaluators that answered "Yes" to the question is referred to as the *score*.

Style-Matching

Handwriting was assumed to be too varied for workers to match well to targets within a reasonable worker pool size. Matching was nonetheless measured for comparison: how similar any given

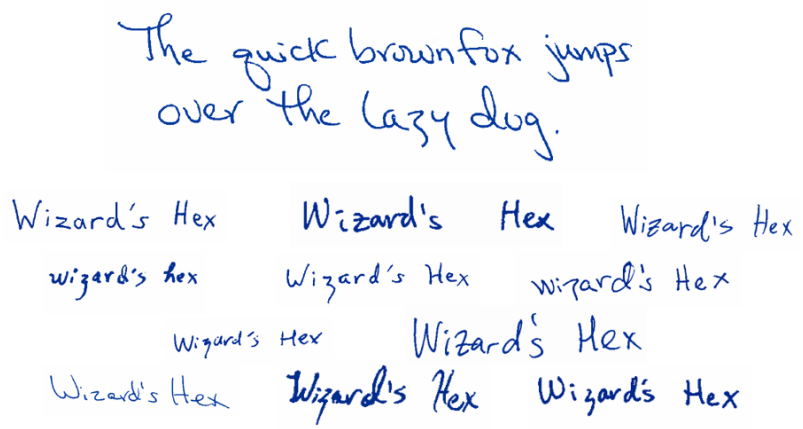

Figure 37: Handwriting imitation example, showing imitators, true sample, and non-imitated distractors.

sample is to the target's, and whether more similar handwriting on the target phrase predicts similarity on the testing phrase. For taste (style) matching, the similarity of the true target sample was evaluated against 13 samples in other people's handwriting.

Style-Grokking

For taste (style) grokking, five grokking workers attempted to imitate the target's handwriting, for each of the nine target people. They were shown the target's writing of the profiling phrase, and imitated that style for the target phrase. Prior to performing these imitations, the five imitators wrote the target phrase in their own handwriting. This gave us a total of 14 target phrase samples in people's own handwriting.

While handwriting similarity evaluation is subjective in itself, the question phrasing makes it clear that there is an objective ground truth correct answer. This means that one could potentially evaluate workers on their ability to correctly distinguish authentic from forged handwriting. Furthermore, it is possible that some workers will try harder simply knowing that their work may be objectively scrutinized.

As a baseline, the non-imitated handwriting of a random person scored 0.17 - i.e., 17% of the time workers thought that the evaluated sample was written by the target person. This was low, but

expected, given how different handwriting may be. On the upper end, the target's true handwriting scored 0.83 providing an insight into the cautiousness of evaluators.

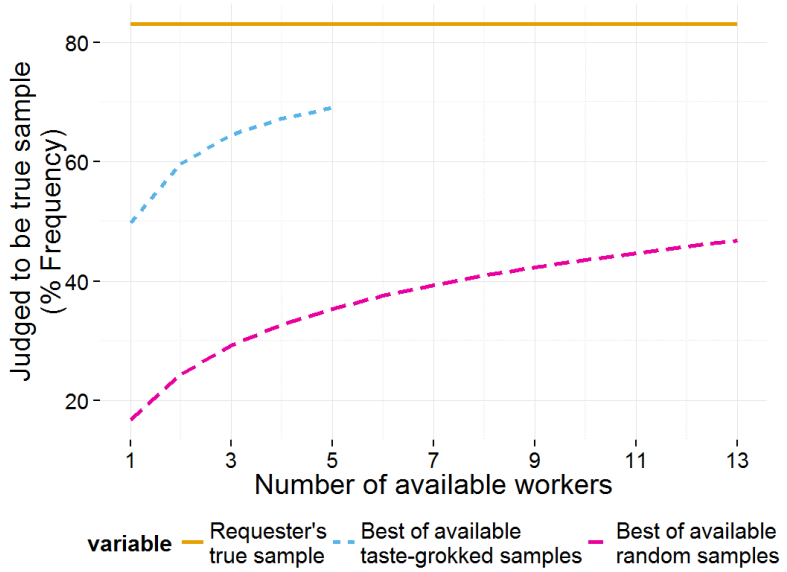

Figure 38: Success of handwriting samples in passing for the target's handwriting, in percentage. The performance of the best grokked sample from a set of 1…5 shown, and the best non-grokked sample from 1…13.

Matching ranked as poorly as expected. Figure 38 shows the score of the best of k random samples in people's own handwriting. Even for k=13, this number is below 0.50. For this reason, matching was determined to be intractable.

Interestingly, there were cases where a grokking worker's target sample was judged to be at least as similar to the training sample as that of the target person themselves. This is possibly due to internal inconsistencies present in a target's casual handwriting, whereas a methodical worker would not include such quirks. For example, in the most drastic example, it was found that a target wrote the profiling phrase with a crossed 'z', but the test phrase without the line through it. An evaluator would likely be looking for a crossed 'z' each time, while grokkers are likely to mimic that.

On average, individual grokkers scored 0.50, well below the 0.83, average score of a target on their own sample pair. There was a large variation in imitation ability among grokking workers, with average scores ranging from 0.38 to 0.67 for the five individual

grokkers. For comparison, the best grokked sample for each condition achieved a score of 0.69. Figure 38 shows how this varies across k, with $k = 1$ being a random single imitator and $k = 5$ being the best of the five.

Discussion

Answer: *RQ 3.1*

> **Answer (RQ 3.1)**: As seen through two possible approaches, well-organized paid crowds can adeptly perform subjective tasks. There may or may not be better approaches than the ones introduced and evaluated, but the efficacy of the ones tested shows that subjective crowdsourcing is tractable.

Answer: *RQ 3.2*

> **Answer (RQ 3.2)**: Taste-matching was effective in reducing the error in personalized crowdsourcing, particularly for complex tasks without latent decision factors and long-term interactions.

Answer: *RQ 3.3*

> **Answer (RQ 3.3)**: Taste-grokking was effective in reducing the error in personalized crowdsourcing, particularly for manifest (i.e. easily 'grokkable') tasks and instances where finding a good matched pair is unlikely.

Taste-matching and taste-grokking provide insights on personalized crowdsourcing for on-demand subjective needs. They generally provided improvements over a non-personalized baseline, but the appropriateness and effectiveness of the protocols was subject to considerations such as the number of decision-making dimensions, task and domain enjoyability, how manifest and easily communicated the factors of a target's needs were, and the cost or time needs.

Answer: *RQ 3.4*

> **Answer (RQ 3.4)**: Different types of tasks benefited from different approaches. When tasks could be aggregated sensibly, taste-grokking benefited from the approach. Some tasks, like text highlighting, require more work for profiling and may benefit from alternate approaches without a profiling step. Evaluating two protocols provided a comparison of these issues; this section generalizes from what was observed.

One way to consider personalization tasks is as a mixture of numerous latent and manifest decision factors, and tempered by the bandwidth of the contribution metric. Handwriting is a highly granular task, where every angle of the pen can change the output, while rating cuisine on a five-point scale has low granularity: five different possibilities. However, what motivates the style of work, whether an opinion, highlight, or a handwritten phrase, can be highly explicit (manifest) or powered by unseen latent factors. Consider a highlighted passage that a film is "a great family film" and the difficulties of parsing the highlighter's intention: do they highlight because they find it generally informative, or do they prefer information of that particular length, or is the family-context very specifically what they were looking for? In contrast, the salt and pepper shaker recommendation seemed motivated by manifest factors related to its appearance. This can be seen when clustering items by opinions: the clusters are qualitatively visually coherent.

A consideration of decision factors and contribution bandwidth helps anticipate the choice of taste-grokking or taste-matching. Taste-grokking worked best for tasks that were manifest, and poorly when there are difficult-to-parse decision factors. When performing the same clustering by opinion as above against opinions in the cuisine dataset, the relationships are more unexpected. A taste-grokker might not expect the correlation between liking shawarma and those that enjoy wheat beer, while a well-matched worker would represent it naturally. It is only at around 14 clusters that the food clusters start to look qualitatively coherent.

While taste-matching is better suited for tasks where the subjective component is more latent, the ability to match is complicated by high-bandwidth tasks. The handwriting imitation task demonstrates that high-bandwidth tasks can be handled by taste-grokking workers, but the chance of a good match increases when the space of possible approaches to a task increases. However, while task bandwidth seems to be a consideration conceptually, this study did not compare bandwidth for the same task (e.g.,

matching vs. grokking for opinions on 10-point vs. 2-point rating scales).

In settings with lower-quality workers, taste-grokking allows for more straightforward quality control metrics. While taste-matching has no conceptual ground truth to judge worker quality, only suspicious behaviors, taste-grokking takes subjective tasks and grounds them in the specific needs of a single target person. Since all workers are striving to do the same thing, traditional quality control metrics can be used more easily, such as aggregation (by voting, average, or other method), worker weighting, and cross-validation.

Taste-matching is more robust to scaling on-demand tasks to large numbers of target people or many workers. This is because taste-grokking contributions are explicitly conditioned on the target, while workers contributing tasks in their own style or opinions can be reused for new targets, and for modeling person-types. Scaling worker counts to large numbers of workers provided better improvements with taste-matching, almost monotonically improving with each additional worker.

Taste-matching at scale bridges the protocol to its conceptual antecedent: collaborative filtering. Taste-matching follows the same intuitions as collaborative filtering and can be interpreted as on-demand collaborative filtering for contexts where no pre-existing data exists. Mature collaborative filtering systems, like Netflix, make sense of large numbers of opinion data from their users; for a context like filtering one's vacation photos or buying a salt shaker, this rich data does not exist and it is hard to imagine a robust system around it.

One challenge in collaborative filtering is data sparsity when a new user joins, when new items are added, or both, when the system is brand new. On-demand personalization through taste-matching can be used to bootstrap a system as it grows.

Personalized crowdsourcing is a new area of focus which will benefit from further study. This work has contributed an initial set of observations, which lead to new questions. One such set of

questions regards optimization strategies around taste-grokking, to further understand good workers and what makes them strong. How do the personal preferences of workers affect their grokking ability? Do better matched workers taste-grok better? How does one identify and nurture 'super-grokkers'?

Selecting the profiling set size and items was found to affect the quality of personalization. An initial study of optimized item selection shows that purposive strategies can improve personalization, though the methods presented here require prior data for clustering; content-based strategies (e.g., analyzing images or metadata) may yield similar improvements. A poor profiling set was found to particularly frustrate grokking workers, who at times asked for a large set. Deciding on the balance between enough and too many examples – for both protocols – was outside of the current study's scope, but is an important need at the nexus of qualitative considerations (time, stamina, interest) and modeling needs.

In all three task types, the best taste-grokking workers performed remarkably well compared to the average. One worker's handwriting imitations fooled evaluators two-thirds of the time, while the best text-highlighting grokkers far surpassed the ineffective random worker. The difficulty with 'super-grokkers' is that while they may be observed after collection, it is difficult to anticipate them a priori without ground truth.

Finally, the application of personalized crowdsourcing in different contexts, different domains, and for different tasks remains to be seen. This chapter focused on personalized item recommendation, with additional consideration paid to a more complex task (text highlighting), and a style-based personalization task (handwriting imitation).

Conclusion

The ability to reach large online crowds and efficiently manage them in a common task has impacted the scale of problems that

can be solved without automation and made possible for on-demand human-in-the-loop systems. While crowdsourcing has been notably applied to ground-truth tasks, this study turned the lens on a less-studied type of problem, one which human workers are especially well-positioned to address: subjective, person-specific tasks. *Personalized crowdsourcing* is discussed as an additional facet of crowdsourcing, one that is well-suited to on-demand personalization.

In the space of on-demand personalization, two notably different approaches were contrasted – *taste-grokking* and *taste-matching* – over three task types: opinion prediction for the purpose of recommendation, text highlighting for the purpose of summarization, and handwriting imitation to view style-based subjectivity. It was found that within these spaces, personalized crowdsourcing is feasible. Understanding that, the question of which technique is better was less consistently answered. Taste-matching proved to be capable for tasks with less manifest decision factors and, as implemented here, more cost-effective. Taste-grokking worked particularly well for product recommendation and usually was responsible for the best single-worker recommendations, albeit with difficulties in anticipating this a priori. Taste-grokking is a promising but less-explored protocol, and while the experiments presented here provided initial insights on how to communicate to and organize grokking workers, additional questions remain about sufficient communication sets and on making use of the task's greater engagement.

We demonstrate that the area of personalized crowdsourcing is promising for on-demand personalization, able to provide person-specific work such as recommendations and filtering without prior data.

Summary and Conclusion

Crowdsourcing presents a great deal of potential value to information science, in its ability to supplement existing metadata objects with new descriptive information, qualitative reactions, and different perspectives. However, the benefits of contributions from self-selected amateur humans are also potential pitfalls.

To efficiently and reliably crowdsource descriptive metadata, one has to account for economies of attention, motivational concerns, subjective variations between contributors, misinterpretations and lack of expertise, and differing contributor contexts.

This dissertation focuses on particular area of crowdsourcing – paid labor though crowd platforms – and studies these issues as they relate to the quality of data contributed. This is primary a study of data quality maximization: in what ways can crowdsourcing data be optimized, both before and during collection time, and in both objective and subjective contexts. *How do we control and interpret qualitative user contributions in a quantified system?*

There were three stages to this dissertation: better making sense of collected data, collecting better quality data, and collecting better quality subjective data.

First, a post-collection approach was taken to interpreting objective tasks: what indicators exist that help us identify and use good contributions while excluding poor ones? The way you ask affects what you receive, so next this study looked at objective tasks at the collection stage. How does the implementation of the collection instrument improve or otherwise bias the collected contributions? Following in this direction, this study finally shifted focus to the implementation of subjective tasks, ones that do not have a con-

cept of correctness except in relation to who they are collected for.

What are the properties of data collected from paid crowds for objective and subjective information system tasks, and how can the quality of data – in terms of consistency and variance – be optimized? We addressed this motivating question through three broad research questions.

> **Broad Research Question 1**: What are the *post-collection* indicators of quality in worker-contributed objective task data, and can these be leveraged for improved data modeling?

In *Interpreting Objective Tasks for Paid Crowdsourcing*, time, experience, and agreement were studied as potential indicators of quality in already collected data.

One of the primary findings of this chapter was that for low-granularity tasks such as information retrieval relevance judgments over short documents, crowd quality is not of much concern. Worker agreement was found to be a notable indicator of quality; however, using this to score workers and weigh their contributions upward or downward was overengineering: a simple majority voting approach worked nearly as well.

Experience was not found to be an indicator of quality for information retrieval judgments. Since this task appeared to have a two-item learning curve, at least in time spent, it would seem that continued engagement with the task did not belie any improved understanding of performing it well. Though this is a null finding, it does mean that study of similarly constructed tasks does not not need to block by number of tasks completed in analyzing results.

Finally, the amount of time taken by workers was not found to be significant, except for the first task item of a task set. Exploring the possibility that this is due to time spent in reading instructions, the measurement of which was confounded with the first task, it was found the otherwise identical workers in their contribution habits could have their eventual performance predicted simply based on the time they spend on the outset of the first task and whether they classify it correctly.

Since poor workers in this class were not exhibiting profit-optimizing behaviours, this finding suggests that interventions during collection time might assist in course-correcting workers – leading to the next research question.

> **Broad RQ 2**: What are the biases inherent to the task design for objective tasks (i.e., the data collection instrument), and can design manipulations correct for them at *collection time*?

In *Designing Tasks for Objective Needs*, this work moved past an immutable treatment of contributions and looks at how contribution quality can be influenced at collection-time.

This chapter explored the design space, exploding the different task parameterization possibilities and considering possible design manipulations. Three manipulations were then pursued in practice: a training interface (in both close interaction first-taskset parameterizations and a less involved per-taskset approach), a performance feedback intervention, and a speed-encouraging time-limited interface. Of these, training and performance feedback improved the quality of contributions over the best practices baseline.

Initial interaction training was found to be effective for a task where the best practices are not conditioned on an extra variable. Whereas the results were not significant for relevance judgments, where subsequent tasks may be for different queries than the training set, it showed considerable promise for tagging tasks.

An alternate training condition was attempted for relevance judgments, which fore-fronted the task instructions with a modal window, one requiring an explicit input to move past, and presented the instructions with strong examples of what each type of judgment would look like. This condition resulted in more time spent on understanding the codebook and, confirming the suggestion from earlier, subsequently strong gains in performance.

These findings support, to the extent of the types of tasks collected, a position that crowd contribution quality is as much a responsibility of requesters as it is of the contributors. More to the point, they show that there are ways of improving results at no

extra collection cost. This makes it an asset for practical implementations, and exploring alternative forms of design manipulations and interventions is worthy of further study.

Tying together the preceded two chapters, a small study was next conducted on the low intercoder reliability of audio similarity judgments for music information retrieval. Correcting the existing data was not found to be effective, but increasing the redundant contribution count and redesigning the collection interface to anchor the categories was able to account for a significant portion of the intercoder error.

Part of this chapter focused on the problem of contributor subjectivity. Despite being performed by trusted contributions, the subjectivity of audio similarity judgment tasks confounded their normative use for system evaluation.

> **Broad RQ 3**: What are the quality losses when treating subjective tasks in objective ways, and can collection-time framing or post-collection modeling approaches reduce these?

Designing Tasks for Subjective Needs again focused on maximizing quality through a priori design and instrumentation choices, this time pursued in the context of subjective tasks. Focusing on a setting of on-demand personalization, it introduced two approaches. The first was a collection-time design choice, *taste-grokking*, where workers were asked to infer the needs of a target user. The second was a post-collection modeling choice, *taste-matching*, where workers were profiled on their preferences, matched to users, and their contributions used to personalize for matched users.

Both approaches improved over an unpersonalized baseline. Taste-matching was strong in contexts whether the factors affecting a person's tasks were more complex or latent, but in simple contexts taste-grokking performed better and resulted in more satisfied workers.

Some Answers for Better Crowdsourcing

Q. How do I start?

The first step of designing a crowdsourcing task is determining the nature of the task. This work's crowdsourcing typology and related practitioner questions can help in understanding the nature of your task.

One of the first questions to answer is whether a task is subjective or objective.

Tasks with a subjective lean – when a 'correct' answer is person-specific or context-specific – require special design. This much is perhaps apparent, but it is not always clear that a task is subjective. It is useful to pilot a task with multiple trusted, careful contributors: how well do they agree? Internally consistent disagreement may be a sign of subjectivity, in addition to other problems like varying interpretations of the codebook.

Collecting subjective information can be aided by taste-matching or taste-grokking, as introduced in this dissertation. A reader in a hurry can consult the discussion section of *Designing Tasks for Subjective Needs* for advice on which approach is more appropriate. Another possible approach is one where the target person explicitly articulates their needs (e.g., Kokkalis et al. 2013; H. Zhang et al. 2012).

Q. My task looks to have some subjectivity, but I want a single output. What do I do?

Though it is often used for quality control to find errant workers, multi-worker aggregation or consensus voting is also useful in deriving a 'normative' objective answer when there is no universal answer. This was necessary in one of this dissertation's studies, where we wanted to use crowd judgments of music similarity for evaluating music algorithms, even though people themselves often disagree.

As shown in comparison with taste-matching and taste-grokking, however, the overall system performance is worse when taking this approach. Furthermore, it needs to be tested how many redundant

workers need to be aggregated: past studies find that this number changes depending on the type of task.

Q. I have a task with a clear concept of a correct contribution. How do I collect the contributions on a paid platform?

The pattern that has been repeatedly found to be effective in crowdsourcing is microtasking, which involves breaking down a task into the smallest possible unit of contribution, preferably so that each microtask does not require context-shifting (e.g. writing and editing would be two different tasks in a composition task).

Much focus in this dissertation was on the recoverable error around collection. Some basic rules can be inferred:

- Design task instructions to be short, with the key points highlighted: people skim and overlook details. Show examples of good or bad contributions.
- Detailed codebooks should be taught, not simply shown.
- If possible, forefront instructions with a dismissable window.
- Optional free-text feedback should be included to allow a manner for workers to communicate problems.

A number of practices can be also be recommended based on context:

- Collection mechanisms that may collect at different granularities are more reliable with less choices. For example, a rating scale can be unary (e.g. star, like), binary (e.g. thumbs up/thumps down), five-point, maybe even 100 points: each step up in complexity lowers intercoder reliability.
- Scales should be anchored with text descriptions: labeling what each choice means.
- If the task is straightforward, performance feedback helps motivate middling workers. If it does not help, there may be instruction issues (i.e., a worker trying to do better cannot figure what they were not doing well).

Q. I started collecting some test data and it doesn't look right (low agreement, doesn't match what I know is true, etc.). What's wrong?

You can try to identify poor workers, as described in the chapter *Interpreting Tasks for Objective Needs*, and weigh them down or remove them altogether. As that chapter notes however, in many cases simply using consensus voting from multiple redundant workers smooths over poor workers. For tasks that do not lend themselves to majority selection, it is possible to conduct a second set of microtasks where workers explicitly choose the best option (see verify step of Find-Fix-Verify pattern in Bernstein, Little, et al. 2010).

Before assuming poor workers it is important to consider other possibilities. Are instructions clear enough? Some more testing may be necessary. Are there any bugs in the interface (e.g. are some images failing to load), or are there outlier tasks that cannot be encoded (e.g. not providing a 'spam' or 'broken' option)? Worker feedback forms should be reviewed. Are there multiple possible ways to perform a task and you want a specific approach? A training task can help.

Finally, workers that are confused about a task or simply bad at it can be identified early; for relevance judgments, this study could identify a poor worker after the first item in a task set. This allows you to focus interventions where they are needed.

Future Directions

An array of new questions follow from those answered in this dissertation. Some relate to different contexts, practical implementations, or directions grazed but not directly measured. Other new questions arise as next steps, now that we know more about the makeup of paid crowds and how their contributions may be guided through the collection interface implementation.

Different contexts. Applying the methods of this study to different contexts is the most pertinent direction forward. How amenable is personalized crowdsourcing to other personalization tasks, like film recommendation, personal photo collection filtering, or comparison shopping? How effective are training interven-

tion or performance feedback for activities such as transcription of historical letters, or other types of encoding beyond tagging? Where possible, this work tried to study multiple task types to get a better sense of how they react to different collection methods. Now that there is a sense of what works when, more focused research can follow in alternate contexts. For example, in the earlier example of transcribing historical correspondence, a first-task training condition would be promising for a collection of a single author's letters, but in a collection with a mixed of different people's writing, resources would be better spent on per-task training interventions.

Cost. One question that was not thoroughly explored in this dissertation is that of cost. This is because the practical floor for payment is lower than what might be ethical to pay: you are usually paying more than you actually need to. Some paid workers rely on the money from Mechanical Turk, so studies that push against the low end of payment in order to compare what methods are less costly are difficult to measure directly. This dissertation looked at cost indirectly, focusing on indicators such a time spent on contribution and, in the case of personalized crowdsourcing, how the number of contributions needed changes in different contexts. Future work might consider cost not in terms of how low contributions are, but how motivated workers are to continue after they have already been paid (similar to Harris and Srinivasan 2012; Mason and Watts 2010).

Targeted interventions. This study found evidence of intervention efficacy and also noted that early success predicts long-term good workers. The natural next step is to study targeted interventions, specifically targeting workers that need them.

Typology of contributors. Studying post-collection indicators of worker quality hinted at different styles of contributor. Bernstein, Little, et al. (2010) speaks of lazy turkers and eager beavers, but there seem to be many different classes on contributor; e.g., those who skim over instructions[64], those who stick around as long as tasks are available, those who work slower and speed up, those

[64] *rtfmers*, as suggested by my advisor in earlier drafts of this document.

whose attention drops over time. A survey of different types of individuals was afield from this dissertation's focus, but would be a valuable future contribution.

Contribution distributions. The quantity that people contribute is another behavior worth studying. In crowdsourcing, the contribution quantity usually follows a curve similar to the inverse power law (i.e., the second most active contributor provides half as many contributions of the most active, the third most active contributor provides 1/4 of the contributions, and so on). However, while the drop-off is non-linear, the steepness of the curve varies between systems. A large survey of contribution distributions and a comparison of how different approaches are able to extend the long tail or soften the curve is work that would further contextualize the experiments in this study.

Long-term effects. Finally, some of the more novel implementations – such as the feedback design manipulation or the taste-grokking personalized crowdsourcing protocol – should be studied in the future within the context of their novelty. Are some of their effects propped up by the attention associated with 'something different', or do they continue with prolonged interactions?

Conclusion

Crowdsourcing is a promising approach for teaching us more about the data in our information systems. Volunteer crowdsourcing inherits various incentivization complexities, something that paid crowdsourcing is able to sidestep. However, as this study shows, crowds still exhibit biases and economies of attention that can influence their contributions in unexpected ways and – particularly concerning for practitioners – unseen ways.

This is part of the territory for crowdsourcing - you benefit from the dynamism of actual humans, but gain it by exchanging some predictability. Aggregating and cross-checking contributions helps in controlling against such issues, but there is also much to be done at little or minimal extra cost, both financial and human cost.

We discover some of these, including modified implementations of collection instruments and more thoughtful treatment of subjective content.

The primary contribution of this dissertation is in understanding when crowd collected data may be biased, and how to improve upon it. Particularly, a non-adversarial approach is taken, for the most part, focusing on how changes come from the circumstances *around* the contributor's context. This does not mean that there are not good or bad contributions – rather, by shifting focus to a parallel track, many of the methods describe here can be implemented alongside worker-centric quality control research.

As crowdsourcing matures as a concept and as a focus of research, it is important to remember that well-organized online crowds are individuals, operating differently from the single-minded, simplified 'crowds' described by Le Bon (1896). Crowd individuals are capable and intelligent, but subject to the whims of attention and influence that we all are.

Jesse Shera once scolded that the computer "should neither be feared as a competitor nor condemned and ridiculed because it has not yet achieved the intellectual capabilities of the human being" (Shera 1967). With crowdsourcing, we see an embrace of the computer as a collaborator, borrowing its efficiencies while turning to humans to assuage its intellectual faults. The development of crowdsourcing has been fundamentally about the pairing of people with machines, insofar as the machines guide the connection between worker and worker, as well as worker and director. It is a story of efficiencies: quicker connections between people, greater access through ubiquitous computing, improvements in modularizing tasks. This work, concerned with raising the value of the individual contribution, contributes to this story: with improvements in designing crowdsourcing tasks and organizing special cases, the abstract and interpretive benefits of crowds can be tapped with fewer people, less uncertainty, and stronger outcomes.

Appendix: Co-authorship Notes

A number of studies within this dissertation are the product of collaborations.

In *Design Facets of Crowdsourcing*, work was completed with Michael B. Twidale. Twidale provided advising and editing, and his ideas co-mingle throughout the paper or have prodded my own contributions in unquantifiable ways.

This study was completed for this dissertation, originally drafted for the preceding proposal, and was updated for Organisciak and Twidale (2015). The version presented here iterates on that presentation. Copyright to the text presented is retained by the authors.

In *Interpreting Objective Tasks for Paid Crowdsourcing*, work was completed alongside Miles Efron, Katrina Fenlon, and Megan Senseney. Work was advised by Efron, and the study was an outgrowth of Efron, Organisciak, and Fenlon (2011). All authors contributed editing support. Efron and Fenlon contributed oracle judgments, making up part of the evaluation dataset. Fenlon contributed the initial text describing the IMLS DCC.

This work was initially presented at the annual meeting of the Association of Information Science and Technology (ASIS&T 2012 - Organisciak, Efron, et al. 2012). Copyright to the text is retained by the authors.

In *Designing Task for Objective Needs*, work on part 2 was completed alongside Stephen Downie. Downie advised on the work.

This work was completed for this dissertation alongside work on the Music Information Retrieval Exchange, but published beforehand as a standalone study (Organisciak and Downie 2015) at

the ACM/IEEE Join Conference on Digital Libraries (JCDL 2015). Permission for dissertation re-use is provided by the ACM alongside reference to the original work.

In *Designing Tasks for Subjective Needs*, work was completed with Jaime Teevan, Susan Dumais, Adam Tauman Kalai, and Robert C. Miller.

The co-authors advised greatly on this study. Since this chapter is a new presentation of previously reported results, much of the writing is new. However, Teevan contributed significant editing and advising support, and parts of the introduction and related work section include text attributable to Teevan. Additionally, data in the section *Handwriting Imitation; Style-Grokking* was collated by co-author Kalai. Earlier publication of this work (Organisciak, Teevan, Dumais, et al. 2014; Organisciak et al. 2015) has signification editing contributions from the coauthors. The research underlying this chapter was completed for Microsoft Research.

Bibliography

About Pinterest (2014). Pinterest. URL: http://about.pinterest.com/en (visited on 08/15/2014).

Abt, Clark C. (1987). *Serious Games*. University Press of America. 200 pp. ISBN: 978-0-8191-6148-2.

Ahn, Luis von (2006). "Games with a purpose". In: *Computer* 39.6, pp. 96–98. ISSN: 0018-9162.

Ahn, Luis von and Laura Dabbish (2004). "Labeling images with a computer game". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. Vienna, Austria: ACM, pp. 319–326. ISBN: 1-58113-702-8.

Ahn, Luis von, Benjamin Maurer, et al. (2008). "recaptcha: Human-based character recognition via web security measures". In: *Science* 321.5895, pp. 1465–1468.

Alderfer, Clayton P. (1969). "An empirical test of a new theory of human needs". In: *Organizational behavior and human performance* 4.2, pp. 142–175.

Alonso, Omar and Ricardo Baeza-Yates (2011). "Design and Implementation of Relevance Assessments Using Crowdsourcing". In: *Advances in Information Retrieval*. Ed. by Paul Clough et al. Lecture Notes in Computer Science 6611. Springer Berlin Heidelberg, pp. 153–164. ISBN: 978-3-642-20160-8 978-3-642-20161-5.

Alonso, Omar, Catherine C. Marshall, and Marc Najork (2013). "Are Some Tweets More Interesting Than Others? #HardQuestion". In: *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*. HCIR '13. New York, NY, USA: ACM, 2:1–2:10. ISBN: 978-1-4503-2570-7. DOI: 10.1145/2528394.2528396.

Alonso, Omar, Daniel E. Rose, and Benjamin Stewart (2008). "Crowdsourcing for relevance evaluation". In: *SIGIR Forum* 42.2, pp. 9–15. ISSN: 0163-5840. DOI: 10.1145/1480506.1480508.

Ambati, Vamshi, Stephan Vogel, and Jaime G. Carbonell (2011). "Towards Task Recommendation in Micro-Task Markets." In: *Human computation*. Citeseer, pp. 1–4.

Angwin, Julia and Geoffrey A. Fowler (2009). "Volunteers log off as Wikipedia ages". In: *Wall Street Journal* 23.

Bao, Shenghua et al. (2007). "Optimizing Web Search Using Social Annotations". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. New York, NY, USA: ACM, pp. 501–510. ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242640.

Bell, Robert M., Yehuda Koren, and Chris Volinsky (2008). "The BellKor 2008 Solution to the Netflix Prize". In: *Statistics Research Department at AT&T Research*.

Benkler, Yochai (2006). *Wealth of Networks*. New Haven: Yale University Press.

Bernstein, Michael S., Joel Brandt, et al. (2011). "Crowds in two seconds: enabling realtime crowd-powered interfaces". In: *Proceedings of the 24th annual ACM symposium on User interface software and technology*. UIST '11. New York, NY, USA: ACM, pp. 33–42. ISBN: 978-1-4503-0716-1. DOI: 10.1145/2047196.2047201.

Bernstein, Michael S., Greg Little, et al. (2010). "Soylent: a word processor with a crowd inside." In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. UIST '10. New York, NY: ACM Press, pp. 313–322. ISBN: 978-1-4503-0271-5. DOI: 10.1145/1866029.1866078.

Bigham, Jeffrey P. et al. (2010). "VizWiz: nearly real-time answers to visual questions". In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. UIST '10. New York, NY, USA: ACM, pp. 333–342. ISBN: 978-1-4503-0271-5. DOI: 10.1145/1866029.1866080.

Bischoff, Kerstin et al. (2008). "Can All Tags Be Used for Search?" In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. New York, NY, USA: ACM, pp. 193–202. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458112.

Causer, Tim, Justin Tonra, and Valerie Wallace (2012). "Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham". In: *Literary and Linguistic Computing* 27.2, pp. 119–137. ISSN: 0268-1145, 1477-4615. DOI: 10.1093/llc/fqs004.

Chen, Edwin (2012a). *Making the Most of Mechanical Turk: Tips and Best Practices*. URL: http://blog.echen.me/2012/04/25/making-the-most-of-mechanical-turk-tips-and-best-practices/.

– (2012b). *Mechanical Turk Best Practices*. URL: Mechanical%20Turk%20Best%20Practices.

Chen, Edwin and Alpa Jain (2013). *Improving Twitter search with real-time human computation*. Twitter Engineering Blog. URL: https://blog.twitter.com/2013/improving-twitter-search-real-time-human-computation.

Chiesura, Sara et al. (2015). "Introducing LibCrowds: a crowdsourcing platform aimed at enhancing access to British Library collections". In:

Cho, Young Ik (2008). "Intercoder Reliability". In: Lavrakas, Paul. *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA, USA: SAGE Publications, Inc. ISBN: 978-1-4129-1808-4 978-1-4129-6394-7.

Cortese, Amy (2011). "A Proposal to Allow Small Private Companies to Get Investors Online". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2011/09/26/opinion/a-proposal-to-allow-small-private-companies-to-get-investors-online.html.

– (2013). "Crowdfunding for Small Business Is Still an Unclear Path". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2013/01/06/business/crowdfunding-for-small-business-is-still-an-unclear-path.html.

Csikszentmihalyi, Mihaly (1991). *Flow: The psychology of optimal experience*. Vol. 41. HarperPerennial New York.

Daugherty, Terry, Matthew S. Eastin, and Laura Bright (2008). "Exploring Consumer Motivations for Creating User-Generated Content". In: *Journal of Interactive Advertising* 8.2, pp. 16–25. ISSN: null. DOI: 10.1080/15252019.2008.10722139.

Dekel, Ofer and Ohad Shamir (2009). "Vox Populi: Collecting High-Quality Labels from a Crowd". In: COLT 2009.

Dellarocas, Chrysanthos and Ritu Narayan (2006). "What motivates consumers to review a product online? A study of the product-specific antecedents of online movie reviews". In: *WISE*.

Donmez, Pinar, Jaime Carbonell, and Jeff Schneider (2010). "A probabilistic framework to learn from multiple annotators with time-varying accuracy". In: *SIAM International Conference on Data Mining (SDM)*, pp. 826–837.

Downie, J. Stephen (2003). "Music information retrieval". In: *Annual Review of Information Science and Technology* 37.1, pp. 295–340. ISSN: 1550-8382. DOI: 10.1002/aris.1440370108.

– (2006). "The Music Information Retrieval Evaluation eXchange (MIREX)". In: *D-Lib Magazine* 12.12, pp. 795–825.

Efron, Miles, Peter Organisciak, and Katrina Fenlon (2011). "Building Topic Models in a Federated Digital Library Through Selective Document Exclusion". In: *Proceedings of the American Society for Information Science and Technology*. ASIS&T Annual Meeting. ASIS&T '11. New Orleans, USA.

Efron, Miles, Peter Organisciak, and Katrina Fenlon (2012). "Improving Retrieval of Short Texts Through Document Expansion". In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. New York, NY, USA: ACM, pp. 911–920. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348405.

Eickhoff, Carsten, Christopher G. Harris, et al. (2012). "Quality Through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments". In: *Proceedings of the 35th International ACM*

SIGIR Conference on Research and Development in Information Retrieval. SIGIR '12. New York, NY, USA: ACM, pp. 871–880. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348400.

Eickhoff, Carsten and Arjen P. Vries (2012). "Increasing cheat robustness of crowdsourcing tasks". In: *Information Retrieval*. ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-011-9181-9.

Eisenberg, Michael B. (1988). "Measuring relevance judgments". In: *Information Processing & Management* 24.4, pp. 373–389. ISSN: 0306-4573. DOI: 10.1016/0306-4573(88)90042-8.

Eveleigh, Alexandra et al. (2013). "I want to be a captain! i want to be a captain!: gamification in the old weather citizen science project". In: *Proceedings of the First International Conference on Gameful Design, Research, and Applications*. ACM, pp. 79–82.

Fung, Brian. "Larry Lessig's super PAC to end super PACs raised $2.5 million in just 2 days. Here's what comes next." In: *The Washington Post*. ISSN: 0190-8286.

Furnas, G. W. et al. (1987). "The vocabulary problem in human-system communication". In: *Communications of the ACM* 30.11, pp. 964–971. ISSN: 00010782. DOI: 10.1145/32206.32212.

Galton, Francis (1907). "Vox populi". In: *Nature* 75, pp. 450–451.

Gardner, Sue (2011). *Nine Reasons Women Don't Edit Wikipedia (in their own words)*. Sue Gardner's Blog. (Visited on 08/07/2015).

Geiger, David et al. (2011). "Managing the crowd: towards a taxonomy of crowdsourcing processes". In: *Proceedings of the seventeenth Americas conference on information systems, Detroit, Michigan*, pp. 1–15.

Golder, Scott A. and Bernardo A. Huberman (2006). "Usage patterns of collaborative tagging systems". In: *Journal of Information Science* 32.2, pp. 198–208. DOI: 10.1177/0165551506062337.

Golder and Huberman (2007). "The Structure of Collaborative Tagging Systems". In:

Grady, Catherine and Matthew Lease (2010). "Crowdsourcing document relevance assessment with Mechanical Turk". In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. CSLDAMT

'10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 172–179.

Gruzd, Anatoliy A. et al. (2007). "Evalutron 6000: Collecting Music Relevance Judgments". In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '07. New York, NY, USA: ACM, pp. 507–507. ISBN: 978-1-59593-644-8. DOI: 10.1145/1255175.1255307.

Guidelines for Academic Requesters (2014). URL: http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters.

Gursoy, Dogan and Ken W. McCleary (2004). "AN INTEGRATIVE MODEL OF TOURISTS' INFORMATION SEARCH BEHAVIOR". In: *Annals of Tourism Research* 31.2, pp. 353–373. ISSN: 0160-7383. DOI: 10.1016/j.annals.2003.12.004.

Harris, Christopher G. and Padmini Srinivasan (2012). "Applying human computation mechanisms to information retrieval". In: *Proceedings of the American Society for Information Science and Technology* 49.1, pp. 1–10. ISSN: 1550-8390. DOI: 10.1002/meet.14504901050.

Hiatt, Laura et al. (2013). *The Role of Familiarity, Priming and Perception in Similarity Judgments*.

Hippel, Eric von (1988). *The Sources of Innovation*. SSRN Scholarly Paper ID 1496218. Rochester, NY: Social Science Research Network.

– (2006). "Democratizing Innovation". In:

Hofmann, Thomas (2004). "Latent Semantic Models for Collaborative Filtering". In: *ACM Transactions on Information Systems* 22.1, pp. 89–115. ISSN: 1046-8188. DOI: 10.1145/963770.963774.

Holley, Rose (2009). *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*. National Library of Australia.

Holley, Rose (2010). "Crowdsourcing: How and Why Should Libraries Do It?" In: *D-Lib Magazine* 16.3. ISSN: 1082-9873. DOI: 10.1045/march2010-holley.

Hotho, Andreas et al. (2006). *Information retrieval in folksonomies: Search and ranking*. Springer.

Howe, Jeff (2006a). *Birth of a Meme*. Crowdsourcing. URL: http://www.crowdsourcing.com/cs/2006/05/birth_of_a_meme.html (visited on 04/26/2014).

– (2006b). *Crowdsourcing: A Definition*. Crowdsourcing: Tracking the rise of the amateur. URL: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html.

– (2006c). "The rise of crowdsourcing". In: *Wired Magazine* 14.6.

– (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. 1st ed. Crown Business. 320 pp. ISBN: 0-307-39620-7.

Hsueh, Pei-Yun, Prem Melville, and Vikas Sindhwani (2009). "Data quality from crowdsourcing: a study of annotation selection criteria". In: *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. HLT '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 27–35.

Hu, Nan, Paul A. Pavlou, and Jennifer Zhang (2006). "Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-mouth Communication". In: *Proceedings of the 7th ACM Conference on Electronic Commerce*. EC '06. New York, NY, USA: ACM, pp. 324–330. ISBN: 1-59593-236-4. DOI: 10.1145/1134707.1134743.

Katter, R. V. (1968). "The influence of scale form on relevance judgments". In: *Information Storage and Retrieval* 4.1, pp. 1–11. ISSN: 0020-0271. DOI: 10.1016/0020-0271(68)90002-8.

Kazai, Gabriella et al. (2011). "Crowdsourcing for Book Search Evaluation: Impact of Hit Design on Comparative System Ranking". In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '11. New York, NY, USA: ACM, pp. 205–214. ISBN: 978-1-4503-0757-4. DOI: 10.1145/2009916.2009947.

Khatib, Firas et al. (2011). "Algorithm discovery by protein folding game players". In: *Proceedings of the National Academy of Sciences* 108.47, pp. 18949–18953.

Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad (2014). "Sentiment Analysis of Short Informal Texts". In: *J. Artif. Int. Res.* 50.1, pp. 723–762. ISSN: 1076-9757.

Kittur, Aniket, E. Chi, and Bongwon Suh (2008). "Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements". In: *Proc. CHI 2008*.

Kokkalis, Nicolas et al. (2013). "EmailValet: managing email overload through private, accountable crowdsourcing". In: *Proceedings of the 2013 conference on Computer supported cooperative work*. CSCW '13. New York, NY, USA: ACM, pp. 1291–1300. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441922.

Komarov, Steven, Katharina Reinecke, and Krzysztof Z. Gajos (2013). "Crowdsourcing Performance Evaluations of User Interfaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: ACM, pp. 207–216. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2470684.

Konstan, Joseph A. et al. (1997). "GroupLens: Applying Collaborative Filtering to Usenet News". In: *Commun. ACM* 40.3, pp. 77–87. ISSN: 0001-0782. DOI: 10.1145/245108.245126.

Koren, Yehuda (2009). "The bellkor solution to the netflix grand prize". In: *Netflix prize documentation*.

Kraut, Robert E. and Paul Resnick (2011). *Building Successful Online Communities*. Cambridge, MA: MIT Press.

Krishnan, Vinod et al. (2008). "Who Predicts Better?: Results from an Online Study Comparing Humans and an Online Recommender System". In: *Proceedings of the 2008 ACM Conference on Recommender Systems*. RecSys '08. New York, NY, USA: ACM, pp. 211–218. ISBN: 978-1-60558-093-7. DOI: 10.1145/1454008.1454042.

Lakhani, Karim R. and Eric von Hippel (2003). "How open source software works: "free" user-to-user assistance". In: *Research Policy* 32.6, pp. 923–943. DOI: 10.1016/S0048-7333(02)00095-1.

Lamere, Paul (2008). "Social tagging and music information retrieval". In: *Journal of New Music Research* 37.2, pp. 101–114.

Law, Edith and Luis von Ahn (2011). "Human Computation". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5.3, pp. 1–121. ISSN: 1939-4608, 1939-4616. DOI: 10.2200/S00371ED1V01Y201107AIM013.

Le Bon, Gustav (1896). *The Crowd: A Study of the Popular Mind*.

Lease, Matthew and Gabriella Kazai (2011). "Overview of the TREC 2011 Crowdsourcing Track (Conference Notebook)". In: *Text Retrieval Conference Notebook*.

Lee, Jin Ha (2010). "Crowdsourcing Music Similarity Judgments using Mechanical Turk." In: *ISMIR*, pp. 183–188.

Lee, Jin Ha and J. Stephen Downie (2004). "Survey Of Music Information Needs, Uses, And Seeking Behaviours: Preliminary Findings." In: *ISMIR*. Vol. 2004. Citeseer, 5th.

Lerman, Kristina, Anon Plangprasopchok, and Chio Wong (2007). "Personalizing image search results on flickr". In: *Intelligent Information Personalization*.

Li, Bin, Qiang Yang, and Xiangyang Xue (2009). "Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction". In: *Proceedings of the 21st international jont conference on Artifical intelligence*, pp. 2052–2057.

Linden, G., B. Smith, and J. York (2003). "Amazon.com recommendations: item-to-item collaborative filtering". In: *IEEE Internet Computing* 7.1, pp. 76–80. ISSN: 1089-7801. DOI: 10.1109/MIC.2003.1167344.

Lintott, Chris J et al. (2008). "Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey". In: *0804.4483*.

Lists (2011). Bibliocommons. (Visited on 10/11/2011).

Mackay, Charles (1852). *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds*. 503 pp.

Marmorstein, Howard, Dhruv Grewal, and Raymond P. H. Fishe (1992). "The Value of Time Spent in Price-Comparison Shopping: Survey and Experimental Evidence". In: *Journal of Consumer Research* 19.1, pp. 52–61. ISSN: 0093-5301.

Marsden, Alan (2012). "Interrogating Melodic Similarity: A Definitive Phenomenon or the Product of Interpretation?" In: *Journal of New Music Research* 41.4, pp. 323–335. ISSN: 0929-8215. DOI: 10.1080/09298215.2012.740051.

Marshall, C. (2000). *The Future of Annotation in a Digital (Paper) World*. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. ISBN: 0-87845-107-2.

Maslow, A.H. (1943). "A theory of human motivation". In: *Psychological Review* 50.4, pp. 370–396. ISSN: 1939-1471(Electronic);0033-295X(Print). DOI: 10.1037/h0054346.

Mason, Winter and Duncan J. Watts (2010). "Financial incentives and the "performance of crowds"". In: *SIGKDD Explor. Newsl.* 11.2, pp. 100–108. ISSN: 1931-0145. DOI: 10.1145/1809400.1809422.

McCreadie, Richard, Craig Macdonald, and Iadh Ounis (2011). "Crowdsourcing blog track top news judgments at TREC". In: *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*, pp. 23–26.

McCreadie, Richard, Craig Macdonald, Rodrygo LT Santos, et al. (2011). "University of Glasgow at TREC 2011: Experiments with Terrier in Crowdsourcing, Microblog, and Web Tracks." In: *TREC*.

McGonigal, Jane (2011). *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. Reprint. Penguin Books. 416 pp. ISBN: 0-14-312061-1.

Michael, David R. and Sandra L. Chen (2005). *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade. ISBN: 1-59200-622-1.

Mitra, Tanushree, C. J. Hutto, and Eric Gilbert (2015). "Comparing Person-and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, pp. 1345–1354.

Moyle, M., J. Tonra, and V. Wallace (2010). "Manuscript transcription by crowdsourcing: Transcribe Bentham". In: *LIBER Quarterly* 20.3.

Muchnik, Lev et al. (2013). "Origins of power-law degree distribution in the heterogeneity of human activity in social networks". In: *Scientific Reports* 3. DOI: 10.1038/srep01783.

Neuendorf, Kimberly A. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA, USA: Sage Publications. 301 pp.

Noll, Michael G. and Christoph Meinel (2007). "Web Search Personalization via Social Bookmarking and Tagging". In: *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag, pp. 367–380. ISBN: 3-540-76297-3 978-3-540-76297-3.

Norvig, Peter (2014). *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*. URL: http://norvig.com/mayzner.html (visited on 05/20/2014).

Novotney, S. and C. Callison-Burch (2010). "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Stroudsburg, PA, USA, pp. 207–215.

Organisciak, Peter (2010). "Why Bother? Examining the motivations of users in large-scale crowd-powered online initiatives". Thesis. Edmonton, Alberta: University of Alberta. 167 pp. URL: http://hdl.handle.net/10048/1370.

– (2012). "An Iterative Reliability Measure for Semi-anonymous Annotators". In: Joint Conference on Digital Libraries. Washington DC, USA.

– (2013). "Incidental Crowdsourcing: Crowdsourcing in the Periphery". In: Digital Humanities 2013. Lincoln, Nebraska.

Organisciak, Peter and J. Stephen Downie (2015). "Improving Consistency of Crowdsourced Multimedia Similarity for Evalu-

ation". In: Joint Conference on Digital Libraries 2015. JCDL '15. Knoxville, TN: ACM.

Organisciak, Peter, Miles Efron, et al. (2012). "Evaluating rater quality and rating difficulty in online annotation activities". In: *Proceedings of the American Society for Information Science and Technology*. ASIS&T. Vol. 49. ASIS&T '12. Baltimore, MD, pp. 1–10. DOI: 10.1002/meet.14504901166.

Organisciak, Peter, Jaime Teevan, Susan T. Dumais, et al. (2014). "A Crowd of Your Own: Crowdsourcing for On-Demand Personalization". In: *Proceedings of the Second AAAI Conference on Human Computation & Crowdsourcing*. HCOMP 2014. Pittsburgh, USA: AAAI.

Organisciak, Peter and Michael B. Twidale (2015). "Design Facets of Crowdsourcing". In: *Proceedings of the 2015 iConference*. iConference '15. Newport Beach, CA.

Organisciak, Peter et al. (2013). "Personalized Human Computation". In: HCOMP 2013. Palm Spring, CA.

– (2015). "Matching and Grokking: Approaches to Personalized Crowdsourcing". In: International Joint Conferences on Artificial Intelligence. IJCAI '15 (Best Papers from Sister Conferences track). Buenos Aires, Argentina: AAAI.

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab.

Polk, Thad A. et al. (2002). "Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency". In: *Cognition* 82.3, B75–88. ISSN: 0010-0277.

Pugh, Casey (2009). *Star Wars Uncut: Director's Cut*. URL: www.starwarsuncut.com.

Quinn, Alexander J. and Benjamin B. Bederson (2011). "Human computation". In: ACM Press, p. 1403. ISBN: 978-1-4503-0228-9. DOI: 10.1145/1978942.1979148.

Radinsky, Kira, Sagie Davidovich, and Shaul Markovitch (2012). "Learning to Predict from Textual Data". In: *J. Artif. Int. Res.* 45.1, pp. 641–684. ISSN: 1076-9757.

Raymond, Eric S. (1999). *The Cathedral and the Bazaar*. O'Reilly Media. 241 pp.

Requester Best Practices (2011). URL: http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (visited on 08/08/2014).

Resnick, Paul et al. (1994). "GroupLens". In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. CSCW '94. ACM Press, pp. 175–186. ISBN: 0-89791-689-1. DOI: 10.1145/192844.192905.

Ritterfeld, Ute, Michael Cody, and Peter Vorderer (2010). *Serious Games: Mechanisms and Effects*. Routledge. 553 pp. ISBN: 978-1-135-84891-0.

Rouse, Anne (2010). "A Preliminary Taxonomy of Crowdsourcing". In: *ACIS 2010 Proceedings*.

Ryan, Richard M. and Edward L. Deci (2000). "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions". In: *Contemporary Educational Psychology* 25.1, pp. 54–67. ISSN: 0361-476X. DOI: 10.1006/ceps.1999.1020.

Rzeszotarski, Jeffrey M. et al. (2013). "Inserting Micro-Breaks into Crowdsourcing Workflows". In: *First AAAI Conference on Human Computation and Crowdsourcing*. First AAAI Conference on Human Computation and Crowdsourcing.

Sanger, Lawrence M. (2009). "The Fate of Expertise after Wikipedia". In: *Episteme* 6.1, pp. 52–73. DOI: 10.3366/E1742360008000543.

Schenk, Eric and Claude Guittard (2009). "Crowdsourcing: What can be Outsourced to the Crowd, and Why?" In: *Workshop on Open Source Innovation, Strasbourg, France*.

Sen, Shilad, F. Maxwell Harper, et al. (2007). "The Quest for Quality Tags". In: *Proceedings of the 2007 International ACM Conference on Supporting Group Work*. GROUP '07. New York, NY, USA: ACM, pp. 361–370. ISBN: 978-1-59593-845-9. DOI: 10.1145/1316624.1316678.

Sen, Shilad, Shyong K. Lam, et al. (2006). "tagging, communities, vocabulary, evolution". In: *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. CSCW '06.

New York, NY, USA: ACM, pp. 181–190. ISBN: 1-59593-249-6. DOI: 10.1145/1180875.1180904.

Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis (2008). "Get another label? improving data quality and data mining using multiple, noisy labelers". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '08. New York, NY, USA: ACM, pp. 614–622. ISBN: 978-1-60558-193-4. DOI: 10.1145/1401890.1401965.

Shera, Jesse H. (1967). "Librarians against Machines". In: *Science* 156.3776, pp. 746–750. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.156.3776.746.

Shirky, C. (2009). *Here comes everybody*. Penguin Books.

Shu, Lisa L. et al. (2012). "Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end". In: *Proceedings of the National Academy of Sciences* 109.38, pp. 15197–15200. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1209746109.

Simon, Nina (2010). *The participatory museum*. Museum 2.0.

Smucker, Mark D., Gabriella Kazai, and Matthew Lease (2012). *Overview of the trec 2012 crowdsourcing track*. DTIC Document.

Snow, R. et al. (2008). "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 254–263.

Spiteri, Louise F. (2011). "Social discovery tools: Cataloguing meets user convenience". In: *Proceedings from North American Symposium on Knowledge Organization*. Vol. 3.

Springer, Michelle et al. (2008). "For the common good: The Library of Congress Flickr pilot project". In:

Surowiecki, James (2004). *The Wisdom of Crowds*. Doubleday.

Swanson, Alexandra et al. (2015). "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna". In: *Scientific Data* 2. ISSN: 2052-4463. DOI: 10.1038/sdata.2015.26.

Tamuz, Omer et al. (2011). "Adaptively Learning the Crowd Kernel". In: *Proceedings of the International Conference on Machine Learning*.

Taylor, Bret (2007). *FriendFeed Blog: I like it, I like it*. friendblog. URL: http://blog.friendfeed.com/2007/10/i-like-it-i-like-it.html (visited on 08/09/2014).

Thompson, Clive (2008). "If You Liked This, You're Sure to Love That". In: *The New York Times*. ISSN: 0362-4331. URL: http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html.

Trant, Jennifer and Bruce Wyman (2006). "Investigating social tagging and folksonomy in art museums with steve. museum". In: *Proceedings of the WWW'06 Collaborative Web Tagging Workshop*.

Tversky, Amos (1977). "Features of similarity". In: *Psychological Review* 84.4, pp. 327–352. ISSN: 1939-1471(Electronic);0033-295X(Print). DOI: 10.1037/0033-295X.84.4.327.

Twain, Mark (1920). *The Adventures of Tom Sawyer*. Harper & brothers. 330 pp.

Typke, Rainer et al. (2005). "A Ground Truth For Half A Million Musical Incipits." In: *JDIM* 3.1, pp. 34–38.

Urbano, Julián, Mónica Marrero, et al. (2011). *The University Carlos III of Madrid at TREC 2011 Crowdsourcing Track*.

Urbano, Julián, Diego Martín, et al. (2011). "Audio Music Similarity and Retrieval: Evaluation Power and Stability." In: *ISMIR*, pp. 597–602.

Urbano, Julián, Jorge Morato, et al. (2010). "Crowdsourcing preference judgments for evaluation of music similarity tasks". In: *ACM SIGIR workshop on crowdsourcing for search evaluation*, pp. 9–16.

Vieweg, Sarah et al. (2010). "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp. 1079–1088.

Vukovic, Maja and Claudio Bartolini (2010). "Towards a Research Agenda for Enterprise Crowdsourcing". In: *Leveraging Applications of Formal Methods, Verification, and Validation*. Ed. by Tiziana

Margaria and Bernhard Steffen. Lecture Notes in Computer Science 6415. Springer Berlin Heidelberg, pp. 425–434. ISBN: 978-3-642-16557-3 978-3-642-16558-0.

Wallace, B. et al. (2011). "Who should label what? Instance allocation in multiple expert active learning". In: *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

Wang, Jing, Panagiotis G. Ipeirotis, and Foster Provost (2011). "Managing Crowdsourcing Workers". In: Winter Conference on Business Intelligence. Utah.

Wei, Xing and W. Bruce Croft (2006). "LDA-based document models for ad-hoc retrieval". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '06. New York, NY, USA: ACM, pp. 178–185. ISBN: 1-59593-369-7. DOI: 10.1145/1148170.1148204.

Welinder, P., S. Branson, et al. (2010). "The multidimensional wisdom of crowds". In: *Neural Information Processing Systems Conference (NIPS)*. Vol. 6, p. 8.

Welinder, P. and P. Perona (2010). "Online crowdsourcing: Rating annotators and obtaining cost-effective labels". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 25–32. ISBN: 978-1-4244-7029-7. DOI: 10.1109/CVPRW.2010.5543189.

Wellman, Barry et al. (2003). "The Social Affordances of the Internet for Networked Individualism". In: *Journal of Computer-Mediated Communication* 8.3, pp. –. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2003.tb00216.x.

What is reCAPTCHA? (2008). Recaptcha. URL: http://recaptcha.net/learnmore.html (visited on 09/27/2008).

Whitehill, J. et al. (2009). "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise". In: *Advances in Neural Information Processing Systems* 22, pp. 2035–2043.

Wiggins, A and Kevin Crowston (2012). "Goals and Tasks: Two Typologies of Citizen Science Projects". In: *2012 45th Hawaii International Conference on System Science (HICSS)*. 2012 45th Hawaii International Conference on System Science (HICSS), pp. 3426–3435. DOI: 10.1109/HICSS.2012.295.

Wikipedia (2014). *Wikipedia:Size of Wikipedia*. In: *Wikipedia, the free encyclopedia*. Page Version ID: 615924147. URL: http://en.wikipedia.org/w/index.php?title=Wikipedia:Size_of_Wikipedia&oldid=615924147.

Zhang, Haoqi et al. (2012). "Human computation tasks with global constraints". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. New York, NY, USA: ACM, pp. 217–226. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207708.

Zhou, Ding et al. (2008). "Exploring Social Annotations for Information Retrieval". In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. New York, NY, USA: ACM, pp. 715–724. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367594.

Zwass, Vladamir (2010). "Co-Creation: Toward a Taxonomy and an Integrated Research Perspective". In: *International Journal of Electronic Commerce* 15.1, pp. 11–48. DOI: 10.2753/JEC1086-4415150101.