ASdb: A System for Classifying Owners of Autonomous Systems

Maya Ziv
Stanford University
mziv@cs.stanford.edu

Liz Izhikevich
Stanford University
lizhikev@stanford.edu

Kimberly Ruth
Stanford University
kcruth@cs.stanford.edu

Katherine Izhikevich
UC San Diego
kizhikev@ucsd.edu

Zakir Durumeric
Stanford University
zakir@cs.stanford.edu

ABSTRACT

While Autonomous Systems (ASes) are crucial for routing Internet traffic, organizations that own them are little understood. Regional Internet Registries (RIRs) inconsistently collect, release, and update basic AS organization information (e.g., website), and prior work provides only coarse-grained classification. Bootstrapping from RIR WHOIS data, we build ASdb, a system that uses data from established business intelligence databases and machine learning to accurately categorize ASes at scale. ASdb achieves 96% coverage of ASes, and 93% and 75% accuracy on 17 industry categories and 95 sub-categories, respectively. ASdb creates a more rich, accurate, comprehensive, and maintainable dataset cataloging AS-owning organizations. This system, and resulting dataset, will allow researchers to better understand who owns the Internet, and perform new forms of meaningful analysis and interpretation at scale.

ACM Reference Format:
Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A System for Classifying Owners of Autonomous Systems. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, Virtual, 17 pages. https://doi.org/10.1145/3487552.3487853

1 INTRODUCTION

To make sense of the prohibitively large number of Internet hosts and services, operators and researchers frequently aggregate hosts and networks by their origin Autonomous System (AS). ASes are a natural aggregation level—they are typically owned and controlled by a single organization. Current AS classification systems provide only coarse categorization of common industries and topological roles (e.g., ISPs), which fundamentally limits the types of questions we can ask about hosts. For example, today, it is nearly impossible to comprehensively ask seemingly simple questions like "Which utility companies have vulnerable Internet-facing services?" and "Which industries display the most BGP instability?"

In this work, we introduce ASdb, a system that classifies organizations into 17 industry categories and 95 sub-categories, by strategically combining data from external business databases (e.g., Dun & Bradstreet), website classifiers (e.g., Zvelo), crowdwork (e.g., Amazon Mechanical Turk), and our own machine learning classifiers. Even with the increased granularity, ASdb achieves both higher coverage and accuracy than prior work with 96% coverage of all registered ASes and 93% and 75% accuracy on 17 categories and 95 sub-categories, respectively (Section 2).

ASdb builds on two key observations. First, while there are no data sources that provide sufficient data about ASes, nearly all ASes belong to identifiable organizations, and there exists an established industry that maintains and provides access to business records. Second, nearly 90% of ASes have associated domains that host websites with descriptive text that can be used for classification. We start our study by evaluating popular business databases, website classifiers, and existing AS classification datasets against a "gold standard" dataset curated by a team of expert researchers (Section 3). We find that in aggregate, business data sources and website classifiers provide accurate category labels for up to 89% of non-technology companies, but fail to accurately categorize the two largest classes of ASes: ISPs and hosting providers.

To fill gaps and arbitrate disagreement between external data sources, we explore building our own machine learning classifiers and using crowdwork to categorize ASes (Section 4). We show that machine learning can correctly classify ISPs and cloud/hosting providers with 94% and 90% accuracy, respectively. We find that crowdworkers can both catch ML failures and resolve data source disagreements with a 98.7% and 94% accuracy, respectively. However, the monetary cost required to incentivize crowdworker accuracy introduces a barrier that ultimately makes crowdwork impractical for our system.

Building on the strengths of business databases, website classifiers, existing AS databases and our new machine learning classifiers, we introduce ASdb, a system that continuously maintains a dataset of Autonomous Systems, their owners, and their industry types (Section 5). ASdb uses a configurable internal matching algorithm to unify all components, handling data source inconsistencies and missing information gracefully. We evaluate ASdb against 620 manually labeled ASes. ASdb provides multi-layer classification for 96% of all ASes and achieves 93% accuracy for top-level categories and 75% accuracy for sub-categories.

We hope that a high-fidelity AS classification dataset will enable the research community to answer new research questions. We will continually release the up-to-date ASdb dataset at https://asdb.stanford.edu for research use.

IMC '21, November 2–4, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-9129-0/21/11…$15.00
https://doi.org/10.1145/3487552.3487853

2 BACKGROUND AND RELATED WORK

Regional Internet Registries (RIRs) like ARIN and RIPE maintain basic AS ownership information (e.g., business name, address, website, and abuse contacts), which they publish through WHOIS. Unfortunately, WHOIS data is only semi-structured, and, in many cases, outdated or incomplete. More critically, RIRs do not publish AS owners' industry sector, age, revenue, or other firmographic details, thus obscuring even basic information about whether an AS is used by an ISP, a cloud hosting provider, or a non-technology company. While it is typically possible to manually research individual ASes, it remains an open problem to label ASes at scale.

AS Classification. There have been several attempts to comprehensively categorize the organizations that control ASes. Dimitropolous et al. employed text classification on AS WHOIS data to categorize ASes into six categories (large and small ISP, IXP, customer, university, network information centers) with a reported 95% coverage and 78% accuracy [33]. Until January 2021, CAIDA provided a dataset based on Dimitropolous et al.'s methodology, *CAIDA UCSD AS Classification Dataset*, which coarsely categorized ASes as "transit/access," "enterprise," or "content" [5]. Due to declining dataset accuracy over the past 15 years, CAIDA recently phased out the dataset. We confirmed this finding by manually classifying 150 ASes (using the methodology detailed in Section 3), and found that the December 2020 CAIDA dataset achieved 72% coverage and 58%, 75%, and 0% accuracy for each category, respectively.

More recently, Baumann and Fabian [27] performed a keyword analysis of WHOIS data to classify ASes into 10 categories (communication, construction, consulting, education, entertainment, finance, healthcare, transport, travel, and utilities) with 57% coverage. They augment their keyword analysis by matching AS names to U.S. Securities and Exchange Commission (SEC) records and extracting industry classification codes. This analysis is restricted to publicly traded companies in the U.S., and they furthermore omit all SEC search results with multiple matches for any AS, limiting the augmentation to 469 ASes.

Routing Topology. Cai et al. [31] clustered RIR records belonging to the same organization; CAIDA publishes a dataset based on the methodology [12]. However, the dataset does not classify the organizations it identifies. There is also a large body of work on AS peering relationships and Internet topology (e.g., [34, 46, 48, 57, 59]). Most relevant, Dhamdhere and Dovrolis [32] use topological properties of ASes to infer broad AS types (enterprise customers, small and large transit providers, access/hosting providers, and content providers) with an accuracy of 76–82%.

Non-Academic Work. PeeringDB [6] is a crowd-sourced database where operators can voluntarily register ASes as one of six categories: "Cable/DSL/ISP", "Network Service Provider", "Content", "Education/Research", "Enterprise", and "Non-profit." As we describe in Section 3, PeeringDB contains only 15% of ASes but has a 95% recall. IPinfo.io [13] uses a black-box methodology to provide the organization name and domain of many ASes as well as a broad classification into one of 4 categories: ISP, hosting, education, and business. In Section 3 we show that it has a 30% coverage and 96% recall, making it one of the most accurate datasets.

Website and Business Classification. Our work draws on both web classification systems and existing business databases. Prior work has examined mechanisms for classifying web domains [24, 30, 47, 53] as well as the difficulty (and sometimes subjectivity) of website classification [51, 60]. Another line of work has looked at the origins, development, and research impact of business classification systems [45, 49, 58], as well as biases and disagreement of business databases that use them [35, 42]. We particularly draw from Phillips and Ormsby [52] in shaping our classification approach.

Source	Searchable	Name	Industry	Domain	Bulk
Business DB					
D&B	N, W, L	✓	NAICS	✓	Paid
Crunchbase	N, W	✓	Custom	✓	Free
ZoomInfo	N, W, L	✓	NAICS	✓	Paid
Clearbit	W	✓	NAICS*	✓	Paid
Networking					
PeeringDB	A	✓	Custom	✓	Free
IPinfo	A	✓	Custom	✓	Paid
Website Class					
Zvelo	W	-	Custom	✓	Paid

Table 1: Candidate Data Sources—We catalogue the attributes of business datasources. Sources are searchable by different metadata (N = Name, W = Website, L = Location, A = ASN). Only three sources overlap in their classification system, utilizing NAICS. *Clearbit provides 2-digit NAICS prefixes and their own custom system. Based on our Section 3 evaluation, ASdb uses D&B, Crunchbase, PeeringDB, IPinfo, and Zvelo.

3 EVALUATING EXTERNAL DATA SOURCES

While there are no data sources that describe ASes at the granularity, coverage, and accuracy we seek, we observe that nearly all ASes belong to identifiable organizations, and there exists an established industry that maintains and sells access to business records. Often advertised to sales teams for researching prospective customers, companies like Dun & Bradstreet [10] and Crunchbase [9] offer products that allow looking up companies by name, address, and domain, and, in turn, provide details like business sector, financial health, and employee count.

In this section, we analyze popular business data sources in depth under a new standard evaluation framework. We find that external business data sources are weak at differentiating technology companies—the most common AS organizational category—but strong for non-technology entities. To address these data sources' weaknesses, we build a machine learning framework in Section 4, which, in combination with the external data sources analysis, lays the foundation for our overall system design (Section 5).

3.1 Potential Data Sources

While many datasets provide business data, not all are suitable for classifying ASes at scale. Some, like LinkedIn [15], do not have accessible APIs for bulk lookups. Others, like Wikipedia [21], provide only loosely formatted text and are difficult to parse in an automated manner. Nonetheless, we find a handful of popular databases that appear reliable, easily queryable, and allow for bulk access, which we further investigate (Table 1).

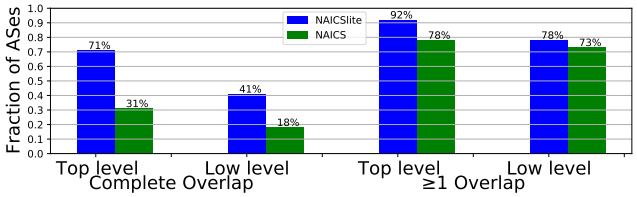


Figure 1: Comparison of Classification Frameworks—The NAICSlite classification system allows for higher agreement among human labelers than using NAICS due to less redundancy and greater specificity in technology related categories. We define complete overlap to mean thats both labels have the exact same set of codes, while ≥ 1 overlap is defined as having one shared label from both labelers.

Dun & Bradstreet, Crunchbase, ZoomInfo, and Clearbit provide firmographic details (e.g., business sector) about organizations. Unfortunately, they cannot be directly queried by ASN, but rather are queried by organization name, address, phone number, and/or domain, which are variably present in bulk WHOIS data. (100% of RIR records have some form of name, 99.7% have a country, 61.7% have a physical address, 45% have a phone number, and 87.1% contain some kind of domain.)

We additionally evaluate two networking-oriented datasets—IPinfo [13] and PeeringDB [6]. These two sources are directly queryable by ASN, but offer limited information about the owning organization. Last, we assess the applicability of the website classifiers mentioned in Vallina et al. [60] for our use case. We find Zvelo to be the most promising option for classifying ASes using their associated domains.

3.2 Evaluation Methodology

We manually build a "Gold Standard" dataset to have a baseline against which to compare external data sources (Table 2). Starting with 150 randomly selected ASes, we assign 60 ASes to each of five computer-networking researchers each such that each AS is independently classified by two researchers. We provide researchers with parsed WHOIS data (Appendix A) and ask them to identify the owning organization's name, website, and to classify organizations using NAICS (North American Industry Classification System), the de facto U.S. federal standard for classifying industries. We ask researchers to manually look up ASes in each candidate data source/service as opposed to performing automated look-ups to ensure that the correct data source entry is found. Researchers then meet in pairs to resolve any labeling discrepancies.

Unfortunately, we find that data sources differ in the classification systems they use, thereby requiring them to be translated into a common classification system. For example, Dun & Bradstreet and ZoomInfo [23] provide the exact NAICS (North American Industry Classification System) [17] code for an organization, while Clearbit [7], Crunchbase, PeeringDB, and Zvelo provide their own organization classification systems that describe business type (e.g., "bank" or "financial industry").

NAICS appears to be a potential option, but during our own classification process, we found that NAICS has several drawbacks. First, NAICS is exceptionally complex, defined across a 517 page manual [1] that describes the hierarchical classification system of over 2,000 categories. Our team found the framework unnecessarily complicated for what we need as a network community (e.g., there are 132 different classifications for industries in agriculture and mining alone). NAICS frequently hampers consensus: 34% of ASes classified contain no overlap in labelers' NAICS codes despite researchers sharing semantic agreement on the type of organization. For instance, AS56885 (SUMIDA Romania SRL) was labeled 335911 (Storage Battery Manufacturing) and 334416 (Capacitor, Resistor, Coil, Transformer, and Other Inductor Manufacturing) by each respective labeler.

In addition, NAICS is not well suited to categorize technology organizations, making idiosyncratic choices about what to distinguish in the computer technology category (e.g., "data processing" has the same NAICS code as "hosting provider" while "software publishers" and "custom computer programming services" are separate codes). Further, NAICS omits categories important to the research community (e.g., NAICS combines ISPs and phone providers in one code, and has no code for computer security organizations).

NAICSlite Translation Layer. To provide a translation between classification systems, while compensating for NAICS' shortcomings, we introduce a simplified version of NAICS:NAICSlite (Appendix C). We build NAICSlite by both collapsing and expanding NAICS categories as appropriate for Internet Measurement. For example, NAICSlite collapses 163 NAICS retail categories into 3 NAICSlite categories and it expands the NAICS information technology category to more clearly distinguish between ISPs, software companies, cloud and hosting providers, and other kinds of technology companies. NAICSlite eschews NAICS' 6-digit hierarchical system for a simpler two-layered approach that offers 17 top-level ("layer 1") categories (e.g., "Computer and Information Technology", "Education and Research", "Finance and Insurance") and up to 9 lower-layer ("layer 2") categories per top level. NAICSlite has a total of 95 layer 2 categories; this is tenfold more categories than in prior AS classification work [27], but an order of magnitude less than NAICS.

We translate all NAICS categories to NAICSlite and find that NAICSlite decreases disagreement amongst researchers categorizing ASes by a factor of two (Figure 1), while still maintaining a rich suite of 95 categories. We note that although the Gold Standard was labeled using NAICS, researchers constructed additional codes to capture finer granularity than NAICS supported during the labeling process, and as such this translation can be done automatically. Researchers do one additional review pass to ensure the resulting categories are accurate and fully descriptive.

We translate other data sources' custom classification schemes into NAICSlite using a manual process, with each mapping reviewed by at least two researchers.

3.3 Data Source Evaluation

Our translation layer provides us with a common denominator against which to formally evaluate business databases, website classification services, and existing AS datasets. We evaluate the data sources in Table 1 across three metrics: coverage, recall, and precision. We show that while existing data sources are able to

Name of Dataset	Number of ASes	Sampling Process	Use of Dataset
Gold Standard	150	Random	To provide a ground-truth for evaluating external datasources and ASdb (Section 3.2)
Uniform Gold Standard	320	Uniformly sub-sampled across all 16 NAICSlite Layer 1 categories	To uniformly evaluate each data source across all NAICSlite categories (Section 3.3)
ML training set	225	150 random, 75 D&B-labeled hosting providers	To provide sufficient hosting-class balance to train a machine learning classifier (Section 4.1)
New test set	150	Random	To provide a fairer evaluation of how ASdb performs when deployed at scale (Section 5.2)

Table 2: Labeled Ground Truth—We use four unique sets of labeled autonomous systems to evaluate external data sources and ASdb.

Source	Coverage	Tech	Non-Tech
D&B	122/148 (82%)	73/96 (76%)	49/52 (94%)
Crunchbase	55/148 (37%)	28/96 (29%)	27/52 (52%)
ZoomInfo	101/148 (68%)	55/96 (57%)	46/52 (88%)
Clearbit	91/148 (61%)	77/96 (80%)	57/52 (90%)
Zvelo	138/148 (93%)	86/96 (90%)	52/52 (100%)
PeeringDB	22/148 (15%)	21/96 (22%)	1/52 (2%)
IPinfo	45/148 (30%)	37/96 (39%)	8/52 (15%)
All - ZI, CL	148/148 (100%)	96/96 (100%)	52/52 (100%)

Table 3: External Data Source Coverage—Zvelo and D&B achieve the highest gold standard coverage. We include in the coverage count only database entries with classification metadata from each datasource. Percents are given out of the 148 gold standard ASes that researchers were able to assign a NAICSlite label to.

achieve promising coverage and precision when categorizing non-technology organizations, they are significantly worse at differentiating the most common AS-owning technology organizations: ISPs and hosting providers. We address these weaknesses by building a machine learning framework in Section 4.

Coverage. Dun & Bradstreet (a business database) and Zvelo (a website classifier) have the highest coverage on our Gold Standard ASes, labeling 82% and 93% of ASes, respectively (Table 3). Zvelo and D&B also provide the most unique coverage, each being the sole providers of coverage for 7/150 and 2/150 ASes, respectively. Neither result is inherently surprising. D&B is one of the oldest, most well known, and most respected business databases. Zvelo's unique coverage is likely because it operates a real-time website classifier. Crunchbase focuses more on startups and specifically US companies and has the lowest coverage of any business database at 37%. The two networking databases, IPinfo and PeeringDB, have by far the worst coverage at 30% and 15% respectively.

All business data sources consistently provide higher coverage for non-technology entities. As shown in Table 3, while PeeringDB and IPinfo classify a maximum of 15% of all non-technology entities, all other data sources classify at least 52%. On the other hand, networking data sources (i.e., PeeringDB and IPinfo) provide 2–8 times more coverage for technology entities, but provide far less overall coverage. No other data source provides significant additional unique coverage or significantly better coverage of any specific regions or categories (per a two-sided hypothesis test with a Bonferroni correction) when compared to the union of Zvelo and D&B.

Recall and Precision. We evaluate each data source's recall and precision: recall to understand if datasets are capable of providing "accurate coverage" of different AS industry sectors, and precision to understand the trustworthiness of the labels applied by data sources. We map each data source's classification system to NAICSlite as described in Section 3.2 and define a match to be accurate if there exists at least one NAICSlite category overlap between the Gold Standard and data source. While this metric does not account for false positives, we note that 80% of data source matches assign only one category and a maximum of seven categories are assigned to a single AS.

The data sources with the highest overall layer 1 recall (96%) are D&B and IPinfo (Table 4), with PeeringDB coming in a close third at 95%. IPinfo and PeeringDB also have the highest precision at 96% and 95%, respectively. However, we emphasize that PeeringDB and IPinfo provide coverage for a very small subset (< 5) categories, limiting their applicability to industry classification. The data sources with the worst recall are Clearbit (34%) and ZoomInfo (70%), which also exhibit the worst precision: 55% and 66%, respectively.

For 99% of ASes in our Gold Standard, at least one data source accurately categorizes the AS. However, given that AS categories are not uniformly distributed, with 64% of ASes being owned by technology-related entities, we separately evaluate technology and non-technology ASes.

Technology Companies. About two thirds of ASes belong to technology companies. The majority of data sources do well at accurately distinguishing tech vs. non-tech organizations, with the union of all data sources accurately providing coverage for 99% and 99% of tech and non-tech organizations, respectively. However, the majority of data sources are nearly two times worse at differentiating between the types of tech organizations (e.g., ISP, hosting provider) than non-tech organizations (e.g., banks, insurance providers), as can be seen in Table 4.

Zvelo and D&B provide weak accurate coverage, but a noticeably higher precision; while Zvelo and D&B achieve a recall rate of 25% (±7% margin of error[1]) and 45% (±9%), they achieve a precision of 86% and 78%, respectively. Sources generally more accurately classify ISPs—PeeringDB reliably classifies ISPs with a 100% true positive rate—but the majority are far from perfect—D&B achieves a recall of 70% (±8%) and precision of 89%. They are, however, far worse at classifying all hosting providers; D&B and Zvelo achieve a recall of 45% (±9%) and 25% (±7%), respectively. We more generally investigate D&B's and Zvelo's inaccurate matches and

[1]We report the margin of error with a 5% α for sample sizes less than n=30.

Source	Overall Layer1	Tech	Non-tech	Overall Layer2	Tech	Non-tech	Hosting	ISP
D&B	116/122 (96%)	70/73 (96%)	46/49 (94%)	93/121 (77%)	39/62 (63%)	51/59 (86%)	5/11 (45%)	28/40 (70%)
Crunchbase	44/55 (80%)	24/28 (86%)	20/27 (74%)	28/53 (53%)	13/24 (54%)	14/15 (93%)	2/5 (40%)	8/13 (62%)
ZoomInfo	71/101 (70%)	39/55 (71%)	32/46 (70%)	84/138 (61%)	23/37 (62%)	34/46 (74%)	5/8 (63%)	14/23 (61%)
Clearbit	31/91 (34%)	3/49 (6%)	32/42 (76%)	–	–	–	–	–
Zvelo	119/138 (86%)	78/86 (91%)	41/52 (79%)	84/138 (61%)	46/74 (62%)	26/64 (41%)	4/16 (25%)	38/47 (81%)
PeeringDB	21/22 (95%)	20/21 (97%)	1/1 (100%)	18/22 (82%)	18/19 (95%)	0/3 (0%)	0/1 (0%)	18/18 (100%)
IPinfo	43/45 (96%)	37/37 (100%)	6/8 (75%)	34/45 (76%)	26/32 (81%)	14/19 (74%)	5/6 (83%)	21/26 (81%)
Union of Best	146/148 (99%)	95/96 (99%)	51/52 (98%)	126/147 (86%)	69/83 (83%)	57/64 (89%)	9/17 (53%)	49/54 (91%)

Table 4: External Data Source Correctness—All data sources, except IPinfo, do poorly when classifying hosting providers, exhibiting a correctness (i.e., the fraction of correctly labeled ASes out of all ASes that are labeled by that data source) of less than 63%. These numbers are based on the 148 Gold Standard entries that labelers could classify. For layer-2 numbers, we also drop the 6 data points that researchers could only assign a layer-1 NAICSlite categorization in the Gold Standard.

find that 67% and 58%, respectively, are due to their ambiguous and inconsistent categorization, preventing a reliable translation to NAICSlite categories. For example, D&B uses three different NAICS codes interchangeably to classify both ISPs and hosting providers: 517911 ("Telecommunications Resellers"), 541512 ("Computer Systems Design Services"), and 519190 ("All Other Information Services").

Beyond Technology. All data sources achieve impressive recall (96–100%) and precision (89–100%) on the two largest non-technology NAICSlite categories: education and finance. Evaluating dataset performance for other categories is complicated by the low number of data points to evaluate against in the long tail of the Gold Standard. To accurately evaluate each data source across all NAICSlite categories, we compile a "Uniform Gold Standard" dataset of 320 registered ASes that are uniformly sampled across all 16 NAICSlite Layer 1 categories (Table 2). We calculate in Table 11, located in the appendix, the precision of data sources across all NAICSlite layer 1 categories. Although individual sources are flawed, in aggregate these sources are a promising resource for categorizing ASes by business sector. At least one data source achieves a 100% precision for 11 of 16 NAICSlite categories. D&B and Zvelo have the best coverage on the Uniform Gold Standard dataset, and Crunchbase achieves at least a 90% precision across half of the NAICSlite categories. Nonetheless, all data sources fail to accurately distinguish between types of technology companies.

3.4 Data Source (Dis)agreement

When using the union of categories applied by at least two data sources that agree on classification nearly all NAICSlite categories achieve an impressive 100% precision (Table 11, located in the appendix). However, this occurs for only 33% and 60% of ASes in the Uniform Gold Standard and Gold Standard set, respectively.

Data sources frequently disagree on the category of an AS. Data sources had zero overlap in the categories they applied for 40% and 13% of ASes in the Uniform Gold Standard and Gold Standard set, respectively. We uncover three categories of disagreement: (1) nuanced disagreement (i.e., both categories applied accurately describe the entity) (2) blatant disagreement (i.e., one of the categories applied is incorrect) and (3) entity disagreement (i.e., the entity being matched to is different), a problem that detail in Section 3.5.

Nuanced disagreement affects 6% of Gold Standard ASes. For example, AS 32169 is an online learning service, and thus gets labeled as "education" by Zvelo, "media" by Crunchbase, and "information technology" by D&B. At the layer 2 level, nuanced disagreement most often occurs when technology companies offer multiple services (e.g., ISP, Hosting, Cell), and data sources match to different services. We observe nuanced disagreement even amongst researchers labeling the Gold Standard set, where 13% of ASes had each researcher label with disagreeing, yet accurate, categories. Vallina et al. [60] attributes the complexity of classifying organizations to the subjective perceptions and priorities of labelers.

Nonetheless, 7% of ASes face disagreeing data sources in which all but one data source are incorrect (e.g., Zvelo labels AS 23414, the Panama Canal, as "Finance and Insurance"). Ultimately, while performance of individual datasets varies widely across ASes, increasing the number of overall data sources also increases the probability of data source agreement and thus the likelihood of an accurate classification.

3.5 Scaling Entity Resolution

Our goal is to build a database that characterizes *all* ASes. Scaling requires both access to the full business datasets and developing an automated method for looking up organizations. While we previously evaluated data source coverage, precision, and recall in Section 3.4 using manually matched and verified data source entries, providing a theoretical upper bound on those metrics must also account for losses brought about by incorrect automated matching. We purchase data for our full dataset from D&B, and Zvelo due to their high coverage and accuracy (Section 3.3). We additionally utilize IPinfo, Crunchbase, and PeeringDB using their provided free or inexpensive ($50/mo) research access. We choose to drop Clearbit and ZoomInfo as neither data source markets full data access to academic researchers.

In this section, we describe and evaluate the automated process we designed to look up ASes.

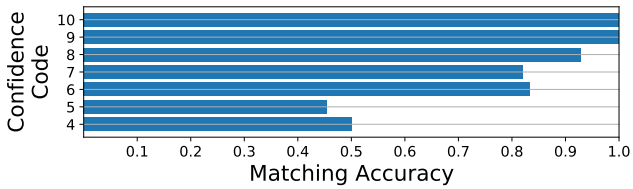


Figure 2: Distribution of D&B Confidence Codes—D&B accurately matches fewer than 50% of ASes when returning a confidence level below 6, but accurately matches at least 80% of ASes when returning a confidence level at or above 6.

Matching Target	Algorithm	Match Accuracy	Correct Matches	Incorrect Matches	Missing Matches
D&B	Conf. ≥1	83%	73%	15%	11%
	Conf. ≥6	89%	67%	8%	25%
Crunchbase	Domain	100%	12%	0%	88%
	Name	95%	14%	1%	85%
Domain	Random	70%	60%	8%	15%
	Least Common	90%	77%	8%	15%
	Most Similar	91%	78%	7%	15%
	IPinfo	86%	82%	14%	4%

Table 5: Accuracy of Automated Entity Resolution—The accuracy of entity resolution (i.e., finding the organization within a data source that corresponds to a given AS) affects overall accuracy and coverage. Entity resolution often involves choosing an organization's corresponding domain. We find that, within the set of domains present in the RIR records for the AS, the domain whose homepage title (or, for unreachable sites, the domain itself) is most similar to the AS name yields greatest accuracy.

Website Identification. D&B, Zvelo, and Crunchbase all, or in part, rely on being provided with the correct domain of an AS-owning organization as a unique identifier. While RIRs do not directly provide the domain of the AS-owning organization, the correct organization domain is often present within multiple abuse contact emails for 85% of ASes. We explore two selection heuristics for selecting the correct organization domain: (1) "least common domain" (i.e., choosing the domain that appears in the fewest WHOIS organization records), and (2) "most similar domain" (i.e., choosing a name with the highest similarity between the website's homepage title and the registered AS name). Using "least common domain" selection eliminates common third-party providers like Gmail and achieves 90% accuracy. Using "most similar domain" selection, achieves a 91% accuracy (Table 5).

Dun & Bradstreet. D&B allows searching for companies by name, address, phone, and domain. In response, their service returns a single company's information (e.g., DUNS#, a unique company identifier) and a 1–10 confidence score. For bulk access, there is no control over which company is chosen if multiple companies share the same name or address.

To evaluate the accuracy of D&B's matching algorithm, we manually verify the returned DUNS# against our hand-identified matches in the Gold Standard and find that across all confidence levels (Figure 2), D&B returns a correct DUNS number 83% and 89% when confidence scores greater than 0, and greater than 5, respectively (Table 5).

Crunchbase. Crunchbase provides a bulk dataset that can be queried by name and/or domain. For all ASes with an available domain, Crunchbase achieves a 100% matching accuracy and 12% coverage when we query all Gold Standard ASes (Table 5). To query ASes with no available domains, we search Crunchbase using a tokenized version of the AS name; Crunchbase achieves 95% matching accuracy on the Gold Standard ASes, while providing 15% coverage.

Zvelo. Zvelo can only be queried by a working domain; thus, Zvelo's coverage is directly dependent on the identification of the correct domain associated with each AS.

With access to all five datasets in full, we query all datasets using all available RIR information and the "most similar" domain matching strategy. We find that all data sources exhibit a relatively similar coverage between the Gold Standard and complete set of registered ASes. Combining all five data sources allows us to reach 99% coverage of all registered ASes. However, non-perfect matching accuracy also adds the possibility of entity disagreement. We discover that when automatically queried, 14% of Gold Standard ASes are matched to at least two data sources that disagree on the entity.

3.6 Summary

Existing data sources are successful at categorizing businesses at the NAICSlite layer 1 granularity, achieving an overall recall on the Gold Standard between 83% and 96%. Furthermore, data source precision increases dramatically when multiple data sources agree on the classification for an AS and is always nearly at 100%. Nonetheless, existing data sources have two primary drawbacks that do not allow them to be directly used to classify ASes: (1) Data sources are not accurate at differentiating the types of technology subcategories—which make up 64% of all ASes (2) Multiple data sources disagree for 21% of ASes, making it unclear when to trust a particular data source. In Section 4 we address these deficiencies by developing additional techniques for classifying ASes.

4 EMPLOYING ARTIFICIAL AND HUMAN INTELLIGENCE TO CLASSIFY ASES

The data sources we evaluated in Section 3 are promising building blocks for classifying ASes globally. However, they have two major shortcomings. First, more than half of ASes belong to technology companies, which business datasets struggle to identify correctly (e.g., Zvelo and D&B achieve a 25% and 45% recall when identifying hosting providers). Second, for 21% of ASes, multiple data sources do not agree. In this section, we show how machine learning and crowdwork can close the gap by correctly classifying ISPs and cloud/hosting providers with 98.7% accuracy and arbitrating disagreements between sources.

4.1 Machine Learning

Despite data sources' difficulty in differentiating between types of technology companies, we note that the two largest categories of

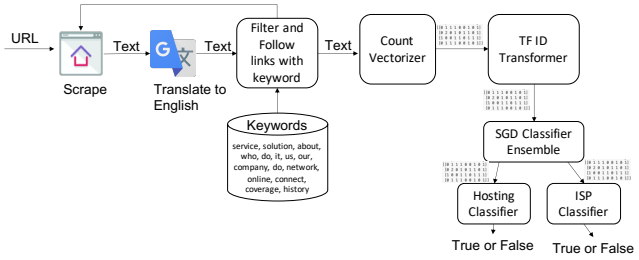


Figure 3: Classification Pipeline— ASdb uses an ML classification pipeline to help identify ISPs and hosting providers (Section 5). ASdb's ML classifiers use stochastic gradient descent to classify a website's scraped and featurized text.

Truth	Prediction	
	Hosting	~Hosting
Hosting	97 (79%)	8 (7%)
~Hosting	4 (3%)	14 (11%)

Truth	Prediction	
	ISP	~ISP
ISP	67 (54%)	6 (5%)
~ISP	1 (1%)	49 (40%)

Table 6: Classifier Evaluation— We introduce two binary classifiers trained to identify hosting provider and ISP websites. The classifiers achieve high test accuracy (90% and 94%, respectively) and minimize false positives (3% and 1%, respectively).

ASes in our Gold Standard dataset—ISPs and hosting providers—use common language and have common descriptors in their websites, which allows humans to quickly identify them. Building on this observation, we hypothesize that an ML classifier will perform well when specifically investigating these two classes. We introduce a classification pipeline that uses web scraping and machine learning classifiers to classify technology ASes (Figure 3).

Pipeline Design. Our ML pipeline accepts a single domain as input and scrapes the text from the root page of the website hosted at the domain. Since 49% of Gold Standard AS websites are not in English, we translate scraped text to English using Chrome's Google Translate [11]. We find that many pages include service descriptions on inner pages rather than the homepage. Using the Gold Standard as a guide, we compile a list of keywords that most frequently appear in the page titles of internal pages containing organization information (see Figure 3). We configure our scraper to visit up to five internal pages whose link titles contain a list of these keywords.

Once relevant text is collected and translated, our pipeline converts the text into a vector of word counts, and uses a TF IDF (Term Frequency Inverse Document Frequency) transformer [55]—used in a majority of text-based recommender systems [28]—to convert the text into features by computing the relative importance of each word found in the text. The features are then used as inputs into two Stochastic Gradient Descent classifiers—often used in text classification due to their scalability [41]. Each is trained to classify whether the organization is a hosting provider or ISP.

Evaluation. To train the pipeline, we compiled a labeled training set of 225 ASes, of which 150 ASes are random and 75 Ases are sampled from D&B-labeled hosting providers to provide sufficient hosting-class balance to train the model (Table 2). We evaluate our pipeline by using the Gold Standard (Section 3.2) as our test set. Each AS takes 5–30 seconds to scrape, depending on load time and number of internal pages. Our model uses 6 CPU cores and 5 seconds to train, and it requires about 1 second to classify 150 domains.

The ISP and hosting classifiers exhibit a test AUC score of .94 and .80, respectively. The ISP classifier achieves an accuracy of 94% and a 1% false positive rate; the hosting classifier achieves a 90% accuracy, with a 3% FP rate (Table 6). We investigate the false positives and find that all are attributed to sites that contain misleading keywords likely to appear on an ISP or hosting website. For example ASN 133002 is owned by the Indian Institute of Tropical Meteorology, whose home page discusses using high performance computing and data analytics to study (nature's) clouds. Its homepage is dominated by keywords like "cloud," "computing," and "performance".

The ISP and hosting classifiers are more likely to produce false negatives than false positives, producing 5% and 7% of false negatives, respectively. 67% of failure cases are due to the initial scraper not having found an internal page that would likely have provided better textual information. These internal pages are often either not linked from the home page or are found in a unique website structure unsuitable for easy scraping (e.g. much of the text is contained in images).

Arbitrating Disagreements. While our pipeline helps differentiate between the two most common types of technology companies, it cannot help arbitrate data source disagreements pertaining to non-ISP and hosting providers. We argue that implementing a ML pipeline to resolve data source disagreements amongst all NAICSlite categories is not the best solution; (1) not enough data is available to train a classifier to distinguish all 17 layer 1 NAICSlite categories (2) Zvelo runs an existing production-grade machine learning classifier whose goal is to differentiate between over 100 business categories; there is no reason to reinvent the wheel.

In the next section, we investigate how crowdwork can be leveraged to supplement existing ML solutions.

4.2 Crowdwork

While our machine learning algorithm accurately classifies ISPs and hosting providers with more than a 90% accuracy, many of the remaining inaccurate classifications appear "easy" to guess from the point of view of a trained human—humans can easily interpret images and navigate through websites without relying on a preset list of keywords. We thus hypothesize and test whether human crowdworkers are effective at classifying ASes that automated solutions miss, regardless of organization type.

We detail our experiments and results in Appendix B. Leveraging Amazon Mechanical Turk (MTurk) workers, we explore two concrete applications of crowdwork to ASdb:

Catching cases where our ML classifiers fail. Our classifiers' main source of error is false negatives (5% and 7% for ISPs and hosting – Section 4.1). Human crowdworkers are effective at catching these errors: for each misclassified AS in our test set, we pay 5 MTurks 30 cents to assign a correct NAICSlite category, and achieve 100% correctness. However, the raw volume of candidate false negatives is too high for this to be cost-effective: we estimate that about

20.7K registered ASes would need crowdworker review, costing at least $31,000. This is untenable for our research budget.

Resolving cases where external data sources disagree or have incomplete coverage. For Gold Standard ASes with conflicting labels (Section 3), we pay 3 MTurks 10 cents to choose among the union of category labels from external data sources; crowdworkers converge on at least one correct category in 94% of cases. Using crowdwork to resolve data source disagreements is much more affordable, costing an estimated $6,000. However, in Section 5.1 we develop an automated heuristic that resolves conflicting labels with an accuracy comparable to crowdwork. Adding crowdwork to the pipeline leads to an overall accuracy improvement of up to 3%.

For ASdb, the accuracy gain from crowdwork is not worth the cost, and we omit crowdwork from our final system design.

5 ASDB: A SYSTEM TO CLASSIFY ASES

In this section we introduce ASdb, a system that uses existing data sources (Section 3) and machine learning (Section 4.1) to create and maintain a dataset of autonomous systems, their owners, and their industries. We also introduce a new heuristic for classifying ASes when data sources disagree. **ASdb is able to classify the type of organization for 96% of ASes with 93% accuracy.**

5.1 System Architecture

ASdb combines data from our classifiers and business datasets using a tuned matching algorithm (Figure 4). ASdb is a modular framework that allows for adding new data sources and changes to the internal matching algorithm.

ASdb's pipeline begins upon the receipt of WHOIS data for an AS (e.g., ASN, AS name, organization name, address, abuse contacts). ASdb checks if the owning organization has previously been classified (e.g., because another AS belonging to the same organization was previously classified), and, if so, ASdb returns the cached data. Otherwise, ASdb begins the classification process by querying data sources that index by ASN (PeeringDB and IPinfo). If a high confidence match occurs (i.e., only if PeeringDB returns an ISP label), ASdb translates the existing data source's categorization system to NAICSlite, stores, and returns the AS's classification.

If there isn't a high confidence match in the first stage, ASdb uses PeeringDB and IPinfo to help determine the most likely domain for the organization. Leveraging the domain extraction analysis in Section 3.3 (Table 5), we use the following algorithm for domain extraction: (1) pool domains from RIR metadata and ASN-queryable data source matches; (2) remove all domains that belong to a hand-curated list of the top 10 email domains (e.g., Gmail); (3) if at least one provided domain appears in < 100 ASes, filter out domains that appear in ≥ 100 ASes; (4) choose from the remaining pool of domains using "most similar" domain matching (91% accuracy, 85% coverage). If a "most likely domain" exists, ASdb then feeds it into our ML classification pipeline (Section 4.1). Indexing by the most likely domain and by WHOIS data, ASdb then uses the AS name, domain, and address to match into other external data sources: D&B, Crunchbase, and Zvelo (Section 3.5). To reduce entity disagreement, ASdb rejects matches where the data source provides a domain that does not match ASdb's chosen domain. ASdb also translates all data source categories into NAICSlite.

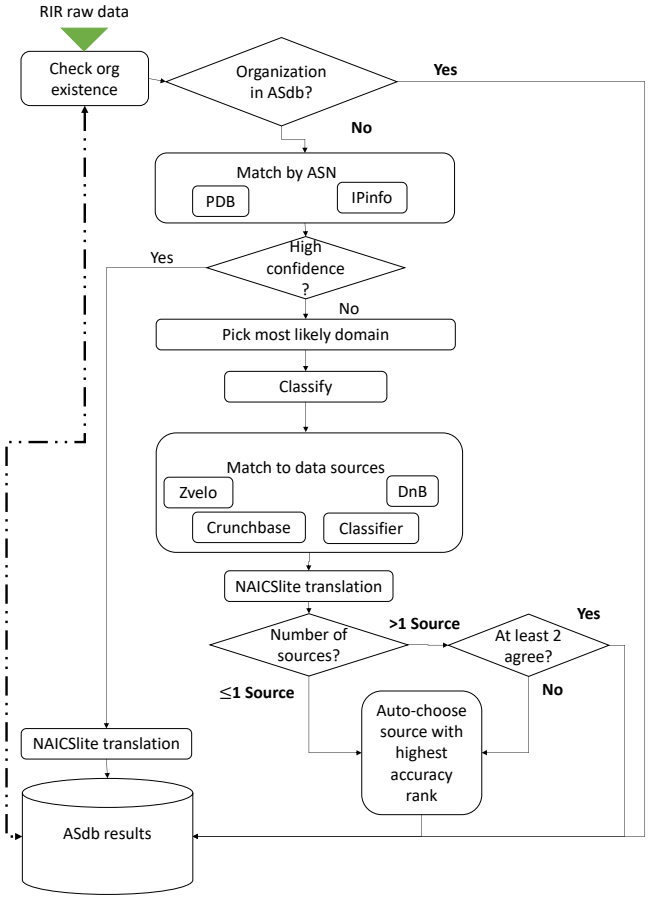


Figure 4: ASdb Design—ASdb uses external data sources and machine learning to identify and classify owners of ASes. ASdb achieves 96% coverage of ASes, and 93% and 75% accuracy on 17 business category labels and 95 sub-labels, respectively.

After data source matching, ASdb begins its consensus phase. If more than one source has information about the AS, and if any overlap exists between data sources' categories, ASdb labels both data sources as trustworthy, returns the union of the overlapping data sources' categories, and exits. Otherwise, we select the category from the source with the best overall accuracy (Section 3.3): IPinfo (96% accuracy), DnB (96%), PeeringDB (95%), Zvelo (88%), Crunchbase (83%).

In the failure case where no data source matches, ASdb returns no category.

5.2 Evaluation

We evaluate ASdb's performance on three data sets: Gold Standard (Section 3.2), Uniform Gold Standard (Section 3.3), and a new test set, which is built using the same methodology as the Gold Standard (Section 3.3). Whereas the Gold Standard was used to evaluate existing data sources and iterate on the ASdb system design, the test set acts as a fresh, random sample of ASes that provides a fairer evaluation of how ASdb performs when deployed at scale (Table

	Gold Standard			Test Set		
Category	ASdb	IPinfo	PeeringDB	ASdb	IPinfo	PeeringDB
Business (N=55)	0.86	0.62	0.07	0.79	0.61	0.0
ISP (N=66)	0.90	0.58	0.36	0.81	0.61	0.47
Hosting (N=13)	0.76	0.30	0.13	0.65	0.24	0.0
Education (N=14)	0.88	0.60	0.13	0.94	0.88	0.19

Table 7: F1-scores for ASdb, IPinfo, and PeeringDB — ASdb is 2.5–6 times more accurate classifying hosting providers, 1.3–2.5 times for ISPs, 1.1–5 times for education entities, and 1.3–12 times for business entities in the gold standard and test set than both prior works. ASdb is 2.5–6 times more accurate classifying hosting providers, 1.3–2.5 times for ISPs, 1.1–5 times for education entities, and 1.3–12 times for business entities in the gold standard and test set than both prior works.

	Gold Standard		Test Set		Uniform Gold Standard	
Stage	Coverage	Accuracy	Coverage	Accuracy	Coverage	Accuracy
Matched By ASN	14%	100%	15%	100%	5%	94%
Classifier	29%	98%	20%	97%	12%	90%
0 Sources Matched	3%	–	4%	–	5%	–
1 Sources Matched	18%	92%	14%	80%	13%	78%
≥2 Sources Matched — ≥ 2 Agree	30%	100%	40%	100%	54%	96%
≥2 Sources Matched — None Agree	5%	89%	7%	60%	12%	68%
Overall Layer 1	97%	97%	96%	93%	95%	89%
Layer 2 — Tech	95%	89%	98%	74%	99%	89%
Layer 2 — Not Tech	89%	82%	92%	80%	97%	76%
Overall Layer 2	93%	87%	96%	75%	98%	82%

Table 8: Evaluation of ASdb Stages — ASdb provides a layer 1 and layer 2 classification for at least 93% of all ASes across all three data sets and achieves a 93% Layer 1 accuracy on the test set. We note that NAICSlite layer 2 coverage can be greater than NAICSlite layer 1 coverage, as only the ASes with a labeler-assigned NAICSlite layer 2 category (142, 141, and 189 for the three data sets, respectively) are evaluated in NAICSlite layer 2 metrics. Recall that our sample size is 150 ASes for the Gold Standard and test set, and 320 for the Uniform Gold Standard set.

2). We include the Uniform Gold Standard data set to demonstrate ASdb's performance at classifying non-technology ASes.

Performance Breakdown. ASdb provides a layer 1 and layer 2 classification for at least 93% of ASes across all three data sets (Table 8). This is at least 25% better coverage than any individual data source (Table 5). ASdb also achieves accuracy on par with or better than external data sources (Table 4) at the layer-1 granularity — achieving a 93%, 97% and 89% accuracy for the test, Gold Standard, and Uniform Gold Standard sets respectively — and layer 2 granularity — achieving a 75%, 87% and 82% accuracy for the test, Gold Standard, and Uniform Gold Standard sets respectively. The weakest points in the ASdb pipeline correspond to cases where there is no multi-source agreement (60% accuracy on the test set where ≥2 sources matched but none agree, 80% accuracy where only 1 source matched). Furthermore, without additional manual review, ASdb cannot classify the 4% of test-set ASes where no data sources match.

Layer 1 Precision and Coverage. To assess ASdb's coverage and accuracy across the long tail of NAICSlite layer-1 categories, we perform a per-category analysis using the Uniform Gold Standard dataset (Table 10). Predictably, ASdb's coverage and accuracy is dependent on external data sources' coverage and accuracy. ASdb consistently achieves nearly identical coverage compared to the data source with the best coverage in the same NAICSlite layer 1 category, while achieving equivalent or better accuracy across 9/16 of categories. The lower precision ASdb achieves in certain categories, as compared to the most accurate data source, is due in all but one case to the most accurate data source — Crunchbase — exhibiting coverage up to 5 times worse.

Number of Applied NAICSlite Categories. We confirm that ASdb does not achieve its measured accuracy by inflating the number of categories it assigns to each AS. Of the 142 test ASes that ASdb successfully classifies, 84 (59%) are assigned only 1 layer-2 NAICSlite category, 16 (11%) are assigned 2 categories, and the maximum number of assigned categories is 10. At the layer-1 level, 104 (73%) are assigned 1 NAICSlite category, 20 (14%) are assigned 2 categories, and the maximum number of layer-1 categories is 4. For both layer-1 and layer-2, the long tail is even sparser for the Gold Standard than it is for the test set.

Comparison With Prior Works. ASdb offers at least 89 additional categories compared to the most popular AS classification databases: IPinfo, which offers 4 categories, and PeeringDB, which offers 6. To compare ASdb's performance with IPinfo, we map IPinfo and NAICSlite's hosting, ISP, and education categories to each other, and also map all other 92 NAICSlite categories to IPinfo's "business." To compare with PeeringDB, we map PeeringDB's content,

enterprise and non-profit, education, and all remaining categories to IPinfo's hosting, business, education, and ISP categories, respectively. We note that ASdb and IPinfo/PeeringDB are not mutually independent, as ASdb relies on both as data sources, but they still serve as a useful benchmark.

ASdb is able to categorize 3 times and 7 times more ASes than IPinfo and PeeringDB, respectively. We compute the F1 metric as a proxy for accuracy, and discover that ASdb always performs better (Table 7). Nonetheless, we notice it only achieves an F1-score of 65% for hosting providers in the test set (albeit still 2.7 times more accurate than IPinfo). Upon further investigation, 17% of all hosting providers do not have domains (i.e., are exceptionally hard to categorize), 9% have no data source matches, and another 9% were marked as non-hosting by at least two data sources, even when our classifier classified the AS as hosting.

Overall, we find that ASdb is 2.5–6 times more accurate when classifying hosting providers, 1.3–2.5 times more accurate when classifying ISPs, 1.1–5 times more accurate when classifying education entities, and 1.3–12 times more accurate when classifying business entities than in prior works.

5.3 Maintaining ASdb

Between October 2020 and February 2021, an average 21 ASes were registered every day, belonging to an average 19 new organizations. Furthermore, 4% of all registered ASes changed their ownership metadata at least once during that period. It is crucial that ASdb is easily updated, as we estimate an average of 140 ASes will need to be updated every week.

ASdb will be primarily maintained by automatically querying data sources available to our research group. We have also integrated a simple way for the research community to submit AS classification corrections; submitted corrections will be verified by a human prior to ASdb integration. For all system components that require human intervention, we plan to devise a community program that requests users of ASdb to periodically complete a human-maintenance task (e.g., review corrections, fetch Zvelo data).

6 CONCLUSION

In this paper, we introduced ASdb, a system that classifies 96% of ASes with a 93% and 75% accuracy on 17 industry categories and 95 sub-categories, respectively. ASdb allows the research community to understand the largely overlooked long-tail of industry sectors that run the Internet. ASdb adds a novel perspective to even the most well-studied research questions: for example, we join ASdb's dataset with an Internet Telnet scan (using a 1% IPv4 LZR [38] scan conducted in March 2021 across 65,535 ports), and alarmingly find that critical-infrastructure organizations like electric utility companies, government organizations, and financial institutions are *more* likely to host Telnet than technology companies.

The process of building ASdb in and of itself also offers insights for the Internet measurement community. We show that business-oriented databases can be applied to networking-specific problems. Nonetheless, data sources not tailored towards the technology community (e.g., business databases) should not be solely relied upon, as they consistently provide worse coverage and accuracy for data pertaining to technology entities. We learn that crowdwork is not the most promising solution; machine learning and simple heuristics perform with nearly the same accuracy at a fraction of the cost. Nonetheless, aggregating existing data sources — no matter their coverage or accuracy — and different classification solutions (e.g. machine learning, simple heuristics), helps build the best-performing classification system.

We designed ASdb to be extendable and maintainable, and we plan to release it and the resulting dataset at asdb.stanford.edu.

ACKNOWLEDGEMENTS

We thank Natasha Sharp, Julie Plummer, and Casey Mullins for support with project logistics and data labeling. We thank David Adrian, Fengchen Gong, Catherine Han, Hans Hanley, Tatyana Izhikevich, Deepak Kumar, and Gerry Wan for providing feedback on the paper, and members of the Stanford Empirical Security Research Group for valuable discussions. We thank the anonymous reviewers and shepherd Romain Fontugne for their helpful comments. This work was supported in part by the National Science Foundation under award CNS-1823192, two NSF Graduate Fellowships DGE-1656518, a Stanford Graduate Fellowship, and gifts from Google, Inc., Cisco Systems, Inc., and Comcast Corporation.

REFERENCES

[1] 2017 NAICS definitions. https://www.census.gov/eos/www/naics/2017NAICS/2017_Definition_File.pdf.
[2] About Amazon Mechanical Turk. https://www.mturk.com/worker/help.
[3] Amazon Mechanical Turk. https://www.mturk.com.
[4] Appen. https://appen.com.
[5] The CAIDA UCSD AS classification dataset. https://www.caida.org/data/as-classification/.
[6] The CAIDA UCSD PeeringDB dataset. https://www.caida.org/data/peeringdb.xml.
[7] Clearbit. https://clearbit.com.
[8] ClickWorker. https://www.clickworker.com.
[9] Crunchbase. http://crunchbase.com.
[10] Dun & Bradstreet. https://www.dnb.com/.
[11] Google Translate. https://translate.google.com/.
[12] Inferred AS to organization mapping dataset. https://www.caida.org/catalog/datasets/as-organizations.
[13] IPinfo.io. https://ipinfo.io/.
[14] Lab In The Wild. https://www.labinthewild.org.
[15] LinkedIn. http://linkedin.com.
[16] Minimum wage. https://www.dol.gov/general/topic/wages/minimumwage.
[17] NAICS Association. https://www.naics.com.
[18] Prolific. https://www.prolific.co.
[19] Unit testing pages or components. https://tapestry.apache.org/unit-testing-pages-or-components.html.
[20] Upwork. https://www.upwork.com.
[21] Wikipedia. http://wikipedia.org.
[22] WorkFusion. https://www.workfusion.com.
[23] ZoomInfo. http://zoominfo.com.
[24] A. Akusok, Y. Miche, J. Karhunen, K.-M. Björk, R. Nian, and A. Lendasse. Arbitrary category classification of websites based on image content. *IEEE Computational Intelligence Magazine*, 10(2), 2015.
[25] S. Albakry, K. Vaniea, and M. K. Wolters. What is this URL's destination? empirical evaluation of users' URL reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[26] N. M. Barbosa and M. Chen. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *2019 CHI Conference on Human Factors in Computing Systems*, 2019.
[27] A. Baumann and B. Fabian. Who runs the Internet? Classifying Autonomous Systems into industries. In *Proceedings of the 2014 International Conference on Web Information Systems and Technologies*, 2014.
[28] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Paper recommender systems: a literature survey. *Intl. Journal on Digital Libraries*, 2016.
[29] D. Bounov, A. DeRossi, M. Menarini, W. G. Griswold, and S. Lerner. Inferring loop invariants through gamification. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.
[30] R. Bruni and G. Bianchi. Website categorization: A formal approach and robustness analysis in the case of e-commerce detection. *Expert Systems with Applications*, 142(113001), 2020.
[31] X. Cai, J. Heidemann, B. Krishnamurthy, and W. Willinger. Towards an AS-to-organization map. In *ACM Internet Measurement Conference*, 2010.
[32] A. Dhamdhere and C. Dovrolis. Twelve years in the evolution of the Internet ecosystem. *IEEE/ACM Transactions on Networking*, 19(5), 2011.
[33] X. Dimitropoulos, D. Krioukov, G. Riley, and k. claffy. Revealing the Autonomous System taxonomy: The machine learning approach. In *Passive and Active Network Measurement Workshop (PAM)*, Adelaide, Australia, Mar 2006. PAM 2006.
[34] V. Giotsas, C. Dietzel, G. Smaragdakis, A. Feldmann, A. Berger, and E. Aben. Detecting peering infrastructure outages in the wild. In *ACM SIGCOMM*, 2017.
[35] E. Han, L. M. Powell, S. N. Zenk, L. Rimkus, P. Ohri-Vachaspati, and F. J. Chaloupka. Classification bias in commercial business lists for retail food stores in the U.S. *International Journal of Behavioral Nutrition and Physical Activity*, 9(46), 2012.
[36] K. Hara, A. Adams, K. Milland, S. Savage, C. Callison-Burch, and J. P. Bigham. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.
[37] J. Huh, H. S. Heo, J. Kang, S. Watanabe, and J. S. Chung. Augmentation adversarial training for self-supervised speaker recognition. In *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.
[38] L. Izhikevich, R. Teixeira, and Z. Durumeric. LZR: Identifying unexpected Internet services. In *30th USENIX Security Symposium*, 2021.
[39] G. Jäger, L. Zilian, C. Hofer, and M. Füllsack. Crowdworking: working with or against the crowd? *Journal of Economic Interaction and Coordination*, 14:761–788, 2019.
[40] H. Jhamtani and T. Berg-Kirkpatrick. Modeling self-repetition in music generation using generative adversarial networks. In *36th International Conference on Machine Learning*, 2019.
[41] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, pages 137–142. Springer, 1998.
[42] K. M. Kahle and R. A. Walkling. The impact of industry classifications on financial research. *The Journal of Financial and Quantitative Analysis*, 31(3):309–335, 1996.
[43] G. Y. Kebe, P. Higgins, P. Jenkins, K. Darvish, R. Sachdeva, R. Barron, J. Winder, D. Engel, E. Raff, and C. M. Francis Ferraro. A spoken language dataset of descriptions for speech-based grounded language learning. In *NeurIPS*, 2021.
[44] V. Koshy, J. S. Park, T.-C. Cheng, and K. Karahalios. "We just use what they give us": Understanding passenger user perspectives in smart homes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.
[45] J. Krishnan and E. Press. The North American Industry Classification System and its implications for accounting research. *Contemporary Accounting Research*, 20(4):685–717, 2003.
[46] E. Liu, G. Akiwate, M. Jonker, A. Mirian, S. Savage, and G. M. Voelker. Who's got your mail? Characterizing mail service provider usage. In *ACM Internet Measurement Conference*, November 2021.
[47] D. López-Sánchez, A. G. Arrieta, and J. M. Corchado. Visual content-based web page categorization with deep transfer learning and metric learning. *Neurocomputing*, 338:418–431, 2019.
[48] R. Motamedi, R. Rejaie, and W. Willinger. A survey of techniques for Internet topology discovery. *IEEE Communications Surveys & Tutorials*, 17(2):1044–1065, 2015.
[49] L. O'Connor. Approaching the challenges and costs of the North American Industrial Classification System (NAICS). *The Bottom Line*, 2000.
[50] E. Peer, J. Vosgerau, and A. Acquisti. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods*, 46(4):1023–1031, 2014.
[51] P. Peng, L. Yang, L. Song, and G. Wang. Opening the blackbox of VirusTotal: Analyzing online phishing scan engines. In *ACM Internet Measurement Conference*, 2019.
[52] R. L. Phillips and R. Ormsby. Industry classification schemes: An analysis and review. *Journal of Business & Finance Librarianship*, 2016.
[53] X. Qi and B. D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2):Article 12, 2009.
[54] S. Rajpal, K. Goel, and Mausam. POMDP-based worker pool selection for crowdsourcing. In *32nd Intl. Conference on Machine Learning*, 2015.
[55] J. Ramos. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, volume 242, pages 29–48, 2003.
[56] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *36th Intl. Conf. on Machine Learning*, 2019.
[57] M. Roughan, W. Willinger, O. Maeneel, D. Perouli, and R. Bush. 10 lessons from 10 years of measuring and modeling the Internet's Autonomous Systems. *IEEE Selected Areas in Communications*, 2011.
[58] M. Ruef and K. Patterson. Credit and classification: The impact of industry boundaries in nineteenth-century America. *Administrative Science Quarterly*, 54(3):486–520, 2009.
[59] S. Shakkottai, M. Fomenkov, R. Koga, D. Krioukov, and K. C. Claffy. Evolution of the Internet AS-level ecosystem. *The European Physical Journal B*, 74:271–278, 2010.
[60] P. Vallina, V. Le Pochat, Á. Feal, M. Paraschiv, J. Gamba, T. Burke, O. Hohlfeld, J. Tapiador, and N. Vallina-Rodriguez. Mis-shapes, mistakes, misfits: An analysis of domain classification services. In *ACM Internet Measurement Conference*, 2020.
[61] M. E. Whiting, G. Hugh, and M. S. Bernstein. Fair work: Crowd work minimum wage with one line of code. In *AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 2019.
[62] T. Ye, Y. Sangseof, and L. Robert, Jr. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *International AAAI Conference on Web and Social Media*, volume 11, 2017.

APPENDIX

A RIR DATA EXTRACTION

All RIRs release their own subset of information in a unique format. We detail our specific data extraction method for each relevant AS field below.

Name. Baumann and Fabian [27] found that provided RIR provided AS names are often uninformative. Thus, across all RIRs, we

extract names using the following fields and order of preference: organization name (provided for 80.19% ASes), description (provided for 24.81% ASes) and AS name (provided for 100% of ASes).

Street Address. Our street address extraction method varies by RIR, as detailed below.

- RIPE: We use the description field; RIPE has no address field.
- APNIC: We use the address field (99.98% of entries contain an address field).
- AFRINIC: We use the address field (90.01% of entries contain an address field). Note that 92% of entries obfuscate their address with "*" characters and only reveal the city, state/province, and country; we remove all obfuscated parts of the address.
- LACNIC: We use the provided city and country fields, as no other address data is available.
- ARIN: We use the address field (100% of entries contain the entire street address).

Phone. APNIC and ARIN provide contact phone numbers for 100% of their ASes. No other RIRs provide phone numbers.

Country. We use CAIDA's AS2org dataset [12] to get country information for 32% of ASes.

Domain. For all RIRs except LACNIC, we extract candidate domains by using the provided emails in the aut-num objects and connected org and contact objects, in addition to a regex match to find all URLs in the "remarks" field. LACNIC does not provide domains or contact emails.

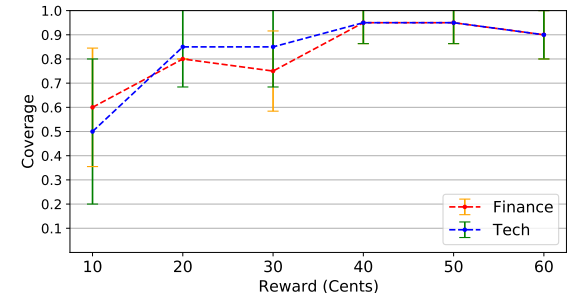
B CROWDWORK EXPLORATION

While building ASdb, we observed that our automated system struggles with some classification tasks that appear "easy" from a human's perspective, as humans have additional context and can more skillfully infer what information is relevant to a given question. We therefore tested whether human crowdworkers are effective at classifying ASes that automated solutions fail to correctly categorize.

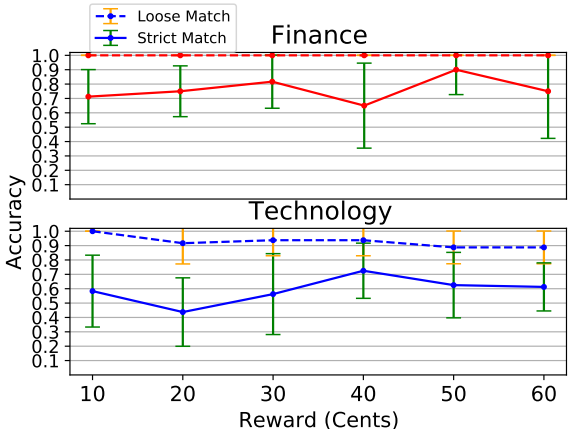
In our experiments we find that crowdwork is impractical for our use case. Here we describe our experiments and surface several lessons for large-scale labeling of networking data.

Ethics. We submitted an IRB protocol for institutional review. The Stanford IRB ruled that our study does not constitute human subjects research and does not require IRB approval, as we are studying only the quality of crowdworker-generated labeled data and not identifiable individuals or their behavior. Nevertheless, compensation and fair treatment of crowdworkers require careful consideration, and we detail the steps we take to interact with crowdworkers ethically.

Platform choice. We explore seven candidate crowdwork platforms. Six are poor fits: Workfusion [22] does not guarantee that "labelers" will be human, Appen [4] markets to companies with bigger projects, Clickworker [8] is notably more expensive, Lab In The Wild [14] requires tasks to be "fun," and Upwork [20] and Prolific [18] require a unique survey per task, therefore not scaling to creating hundreds of labeling tasks. We thus choose Amazon Mechanical Turk (AMT) [3], which offers an easily scalable framework to deploy labeling tasks and custom pricing. AMT also allows specifying worker qualifications (e.g., IT employment industry) at a premium price, but requires that at least 10 "qualified" workers be assigned to a single unbatched classification task, drastically increasing the cost of labeling a single AS (≥ $7). AMT also provides a separate "master" qualification for select workers who "consistently submit a lot of high quality work" [2]. Master MTurks cost 5% more than regular MTurks (based on the offered reward) and can be individually assigned to a single un-batched task. Prior work demonstrates that Master MTurks provide higher-quality results [50], and we therefore hire only Master MTurks for the duration of our experiments.



(a) Coverage—The number of ASes classified increases as the reward offered increases.



(b) Accuracy—ASes are not classified appreciably more accurately when MTurks are offered increased rewards.

Figure 5: Evaluating Amazon Mechanical Turk

Crowdworker Wages At Scale

We support the research community's push for fair crowdworker compensation. MTurks often make well below the US federal minimum wage ($3 per hour on average [39]), and we strive to do better. However, setting a fair crowdworker wage for micro-tasks at scale is not a straightforward task. Amazon Mechanical Turk does not monitor or enforce hourly wages, nor does it provide a way to set MTurk pay as a function of minimum wage in the MTurk's jurisdiction. Setting MTurk payments per task can only be done in advance of releasing the task, therefore requiring advance knowledge of how much time a task should take, which can be difficult to estimate.

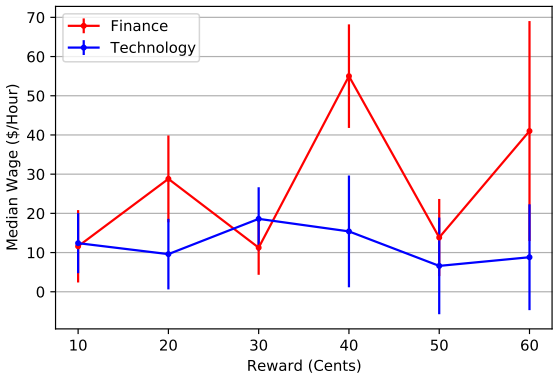


Figure 6: Amazon Mechanical Turk Wages—Reward-per-task and median hourly wage is not directly correlated.

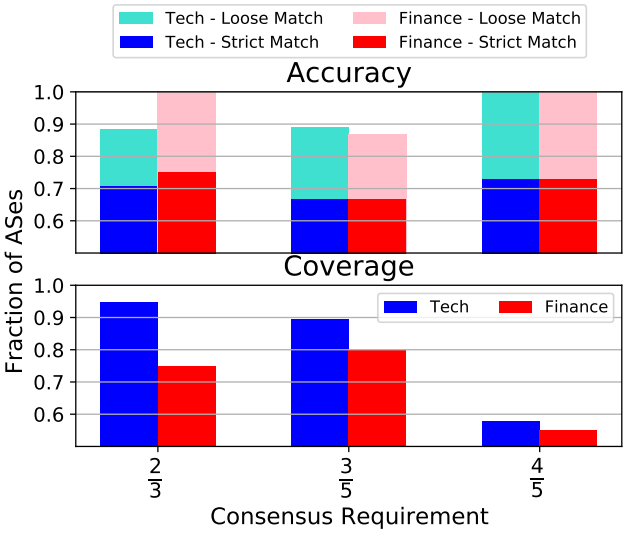


Figure 7: Amazon Mechanical Turk Consensus—Increasing the number of MTurks required for consensus (.e.g., requiring 4/5 MTurks to agree instead of 2/3) increases accuracy by up to 12% and decreases coverage by up to 35%.

Research sub-communities have different approaches to these hurdles: the ML community routinely labels large datasets using a small or unstated flat fee per task (e.g., [37, 40, 43, 54, 56]). The HCI community often sets a living-wage objective for task compensation [26, 36, 61] and regularly reports crowdworker wages for research tasks (e.g., [25, 29, 44]). In addition, beyond fair-wage ethics, MTurk compensation can also affect the accuracy of tasks [62].

To understand what compensation per task must be offered to an MTurk in order to ensure fair wages, accurate results, and compatibility with our research budget, we conduct an experiment to quantify MTurk performance depending upon the offered reward.

Concretely, we select a group of 20 technology and 20 finance ASes and ask 3 MTurks to choose one or more NAICSlite layer 2 Technology category for each AS. We set the consensus requirement to be at least two out of three MTurks assigning an AS the same category label. We replicate this setup 6 times, varying only the amount paid to each MTurk (between 10–60 cents in 10 cent increments), and ensure that no MTurks overlap between assignments.

Based on the average amount of time spent by each MTurk for each task, and the reward given for each task the average hourly wage is $19.41/hour across all tasks. However, we discover that **reward-per-task and overall hourly wage is not directly correlated**. We calculate the median hourly wage per task in Figure 6 and find that the median wage ranges extensively, between $55/hour – $6.60/hour, due to MTurks spending a variable amount of time across tasks. Only in one experiment did the median hourly wage fall below the US federal minimum wage [16], due to MTurks unforeseeably taking longer than expected compared to experiments above and below the offered reward. Thus, increasing the reward-per-task does not necessarily increase the overall median wage.

The *number* of ASes classified increases as the reward offered increases (Figure 5a) due to increase in consensus (i.e., agreement amongst MTurk labels). For example, offering 50 cents per task leads to 95% of both tech and financial ASes being classified, which is a 10–20% increase compared to offering 30 cents.

To understand how wages affect the accuracy of a task, we use the same experiment and define accuracy using a strict-match criterion (i.e., all consensus-backed crowdworker categories match all Gold Standard categories) and a loose-match criterion (i.e., at least one consensus-backed crowdworker category is contained in the set of Gold Standard categories). If no consensus among the MTurks is reached for a particular AS, we exclude it from our accuracy count because there is no reliable label.

We find that **increased rewards and classification accuracy are not directly correlated** (Figure 5b). When evaluating loose-match accuracy, all MTurks, no matter the reward, achieve a 100% and 90%–100% accuracy when classifying finance and technology ASes, respectively. As the reward decreases, there is a slight increase in loose-match accuracy for technology ASes, which we attribute to a decrease in coverage (as crowdworkers may not be willing to spend as much time generating accurate answers for the "hard" cases); as consensus decreases amongst MTurks, only the "easy" cases achieve consensus, leading to higher accuracy. Strict match accuracy is not correlated with reward; the difference in average accuracy between rewarding 10 and 60 cents is less than 5% when classifying both technology and finance ASes. Across all rewards, MTurks perform consistently worse at accurately labeling technology categories compared to financial categories, even when technology category labels are accompanied with definitions within the task interface.

Crowdworker Consensus

Another factor affecting total cost of crowdwork in addition to accuracy and coverage is the crowdworker consensus requirement. To test if increasing the number of MTurks per assignment increases accuracy and coverage, we run an experiment in which we select the same technology and financial ASes, fix the reward at 30 cents,

		Gold Standard		Test Set		Uniform Gold Standard	
System	Stage	Coverage	Accuracy	Coverage	Accuracy	Coverage	Accuracy
ASdb (AMT = True)	0 Sources Matched	3%	33%	4%	0%	5%	20%
	1 Source Matched	18%	92%	14%	81%	13%	80%
	≥2 Sources Matched - None Agree	5%	100%	7%	74%	12%	93%
	Overall Layer 1	97%	98% (+1%)	94%	95% (+2%)	95%	93% (+4%)
	Overall Layer 2	91%	87% (+0%)	93%	78% (+3%)	97%	83% (+1%)

Table 9: Evaluation of ASdb supplemented with crowdwork — Adding crowdwork to help categorize ASes affects coverage and accuracy negligibly. We note that NAICSlite layer 2 coverage can be greater than NAICSlite layer 1 coverage, as only the ASes with a labeler-assigned NAICSlite layer 2 category (142, 141, and 189 for the three data sets, respectively) are evaluated in NAICSlite layer 2 metrics. Recall that our sample size is 150 ASes for the Gold Standard and test set, and 320 for the Uniform Gold Standard set.

and assign either 3 or 5 MTurks to each task. We vary the consensus requirement to require 2/3, 3/5, or 4/5 MTurks to agree on a category.

Strengthening the consensus requirement from 2/3 to 4/5 leads to a 100% loose matching accuracy (Figure 7). However, this comes at the expense of coverage, which drops by 35% when classifying tech ASes due to the lack of consensus. Crowdworker fair pay and cost-effectiveness of research are not at odds for this crowdwork parameter: equivalent overall accuracy and better coverage is achieved when paying 40 cents for 3 MTurks (in the previous experiment) compared to 30 cents for 5 MTurks.

Applying Crowdwork To ASdb

Using our analysis of how offered reward and consensus requirement affect coverage and accuracy (Appendix B), we evaluate the potential for crowdwork to address two concrete failure modes of ASdb's automated analysis: catching ML false negatives and choosing between disagreeing data sources. We also evaluate the cost efficiency of applying crowdwork to each of these use cases. In all experiments, we define coverage to be the percentage of ASes for which the Amazon Mechanical Turk crowdworkers ("MTurks") reached consensus for at least one category. We evaluate accuracy using the Gold Standard and Uniform Gold Standard datasets.

Catching ML failure cases. We ask crowdworkers to classify the ASes that our ML classifiers (Section 4.1) incorrectly classify, at 30 cents per task. We find that a 2/3 MTurk consensus ratio correctly classifies 60% of the tech ASes that were misclassified in our experiment set, and a 3/5 consensus correctly classifies 100% of the misclassified ASes. Thus, with a sufficiently forgiving consensus requirement, MTurks are successful at catching ML failures.

Our classification pipeline achieves a low rate of false positives (1% and 3% when classifying ISPs and hosting providers, respectively). Given that crowdworkers are capable of catching a classifier's false negatives with high accuracy, we consider the possibility that they could serve as an additional review stage for potential false negatives. We identify the class of potential false negatives to be any AS which is classified as Technology by a data source, but not flagged by either of our classifiers. 23% of Gold Standard ASes fall into this category (i.e., roughly 20.7K of all registered ASes), thereby implying that it would cost at least $31,000 to complete this task with the necessary pay and consensus requirements to achieve high accuracy. We find no practical way to more granularly filter out which ASes need human review beyond this heuristic, so we rule out this application of crowdwork as too expensive.

Resolving data source disagreements. We test whether MTurks can effectively determine the correct category in the presence of conflicting labels from multiple data sources. We select 35 random ASes from the Gold Standard with conflicting category labels, along with their manually identified working websites, and ask MTurks to select all applicable layer 2 NAICSlite categories (or "none of the above") from the union of all NAICSlite categories provided by the matched data sources. Requiring a 2/3 MTurk consensus ratio at 10 cents per task, we see that MTurks achieve consensus in 92% of cases and achieve a 94% and 50% loose-match and strict-match layer 2 accuracy, respectively. Thus, MTurks can be leveraged to resolve disagreement between data sources.

By contrast to catching ML failures, using crowdworkers to choose the best NAICSlite category is more cost-efficient. MTurks can accurately select an NAICSlite layer 2 category for 86% of all provided ASes when each AS is labeled by 3 MTurks at a rate of 10 cents per task. Roughly 22% of all registered ASes could be sent to MTurks to select the best NAICSlite category: 4% of all Gold Standard ASes are only matched to one data source, 17% are matched to multiple disagreeing data sources, and an estimated 1% of ASes have a working domain, but match to zero sources. In total, applying crowdwork to these cases would cost an estimated $6,000.

We evaluate the potential impact of crowdwork-based data source disagreement resolution on the overall ASdb system, compared to an automated "auto-choose source" heuristic that we develop in Section 5.1. Surprisingly, we find that **crowdwork adds little to the system overall**: while crowdworkers inexplicably misclassify 9% of ASes they are given, they achieve roughly the same accuracy (80–92%) as our "auto-choose source" heuristic. Table 9 shows the final accuracy of ASdb with crowdwork integrated; using crowdworkers instead of "auto-choose source" leads to an accuracy improvement of up to 3% and coverage decrease of up to 3%. We attribute this result to the fact that unlike ML failure cases, data source disagreements are typically difficult corner cases: of the ASes that we sent to crowdworkers when evaluating our final ASdb system, 31% do not have a working website, 11% have an uninformative website (e.g., an Apache test page [19]), and 49% of organizations achieve no consensus amongst data sources or crowdworkers. This analysis calls into question whether the cost of resolving data source disagreements is worth the small accuracy gain. For our application, we conclude that it is not.

Source	Overall	Agriculture	Nonprofits	Tech	Construction	Education	Finance
D&B	229 / 341 (67%)	13 / 17 (76%)	6 / 7 (85%)	58 / 117 (49%)	8 / 10 (80%)	18 / 22 (81%)	15 / 17 (88%)
Zvelo	199 / 262 (75%)	4 / 7 (57%)	2 / 5 (40%)	73 / 91 (80%)	6 / 9 (66%)	20 / 21 (95%)	9 / 11 (81%)
Crunchbase	108 / 125 (86%)	2 / 3 (66%)	3 / 3 (100%)	27 / 30 (90%)	3 / 4 (75%)	4 / 4 (100%)	9 / 10 (90%)
ASdb	283 / 326 (86%)	14 / 15 (93%)	6 / 6 (100%)	96 / 112 (85%)	8 / 9 (88%)	21 / 22 (95%)	15 / 16 (93%)

Source		Shipping	Government	Health Care	Manufacturing	Media	Entertainment
D&B		15 / 17 (88%)	17 / 23 (73%)	12 / 14 (85%)	8 / 11 (72%)	12 / 19 (63%)	5 / 9 (55%)
Zvelo		11 / 16 (68%)	15 / 19 (78%)	13 / 13 (100%)	4 / 8 (50%)	15 / 18 (83%)	6 / 7 (85%)
Crunchbase		6 / 7 (85%)	1 / 3 (33%)	4 / 4 (100%)	6 / 10 (60%)	9 / 10 (90%)	2 / 2 (100%)
ASdb		16 / 17 (94%)	16 / 23 (69%)	13 / 14 (92%)	8 / 11 (72%)	17 / 19 (89%)	6 / 7 (85%)

Source		Retail	Service	Travel	Utilities
D&B		15 / 17 (88%)	12 / 16 (75%)	5 / 10 (50%)	10 / 14 (71%)
Zvelo		5 / 10 (50%)	11 / 14 (78%)	5 / 7 (71%)	0 / 6 (0%)
Crunchbase		10 / 11 (90%)	9 / 11 (81%)	7 / 7 (100%)	6 / 6 (100%)
ASdb		15 / 16 (93%)	14 / 16 (87%)	7 / 9 (77%)	11 / 13 (84%)

Table 10: Category-based (layer 1) Accuracy and Coverage with Matching —ASdb consistently achieves very similar coverage when compared to the data source with the best coverage in the same NAICSlite layer 1 category, while achieving an equivalent or better accuracy across 50% of categories.

Source	Overall	Agriculture	Nonprofits	Tech	Construction	Education	Finance
D&B	259/307 (84%)	10/13 (77%)	12/16 (75%)	25/32 (78%)	15/19 (79%)	17/20 (85%)	18/19 (95%)
Zvelo	200/253 (79%)	5/5 (100%)	3/5 (60%)	49/57 (86%)	12/12 (100%)	19/20 (95%)	16/18 (89%)
Crunchbase	109/125 (87%)	–	3/3 (100%)	25/27 (93%)	1/2 (50%)	6/6 (100%)	12/12 (100%)
DB + ZV	112/115 (97%)	3/3 (100%)	2/3 (67%)	18/18 (100%)	6/6 (100%)	14/15 (93%)	11/11 (100%)
DB + CB	62/64 (97%)	–	3/3 (100%)	9/11 (82%)	1/1 (100%)	2/2 (100%)	9/9 (100%)
ZV + CB	56/57 (98%)	–	1/1 (100%)	18/18 (100%)	2/2 (100%)	4/4 (100%)	6/6 (100%)
All 3	32/32 (100%)	–	1/1 (100%)	8/8 (100%)	1/1 (100%)	2/2 (100%)	4/4 (100%)

Source		Shipping	Government	Health Care	Manufacturing	Media	Entertainment
D&B		17/20 (85%)	16/17 (94%)	17/18 (94%)	17/19 (89%)	19/20 (95%)	8/19 (42%)
Zvelo		6/6 (100%)	14/15 (93%)	13/14 (93%)	7/8 (88%)	17/18 (94%)	7/7 (100%)
Crunchbase		7/8 (88%)	2/2 (100%)	5/8 (63%)	7/10 (70%)	10/10 (100%)	2/2 (100%)
DB + ZV		6/6 (100%)	8/8 (100%)	11/11 (100%)	4/4 (100%)	7/7 (100%)	4/4 (100%)
DB + CB		5/5 (100%)	–	3/3 (100%)	6/6 (100%)	7/7 (100%)	1/1 (100%)
ZV + CB		1/1 (100%)	1/1 (100%)	1/2 (50%)	2/2 (100%)	8/8 (100%)	–
All 3		1/1 (100%)	–	1/1 (100%)	2/2 (100%)	5/5 (100%)	1/1 (100%)

Source		Retail	Service	Travel	Utilities
D&B		17/20 (85%)	17/18 (94%)	17/17 (100%)	17/20 (85%)
Zvelo		4/4 (100%)	15/51 (29%)	13/13 (100%)	–
Crunchbase		7/10 (70%)	5/6 (83%)	10/11 (91%)	7/8 (88%)
DB + ZV		3/3 (100%)	7/8 (88%)	8/8 (100%)	–
DB + CB		2/2 (100%)	3/3 (100%)	5/5 (100%)	6/6 (100%)
ZV + CB		2/2 (100%)	–	4/4 (100%)	6/6 (100%)
All 3		–	2/2 (100%)	4/4 (100%)	–

Table 11: Category-based (layer 1) Precision and Coverage for External Data Sources—At least one data source achieves a 100% precision on 11 out of 16 NAICSlite categories on the Uniform Gold Standard set. When using the intersection of at least two data sources that agree on classification—occurring in only 33% of ASes in the Uniform Gold Standard—nearly all NAICSlite categories achieve 100% precision. In the Gold Standard set, 60% of ASes have two sources which agree on a classification, and overall precision is 96%. The denominators of the provided fractions denote coverage. Note that given their relatively poor coverage and performance in comparison to other data sources (as well as prohibitive cost) we drop ZoomInfo and Clearbit from our evaluation.

C NAICSLITE CATEGORIZATION SYSTEM

Here we describe in full the NAICSlite categorization system.

- Computer and Information Technology:
  - Internet Service Provider (ISP)
  - Phone Provider
  - Hosting, Cloud Provider, Data Center, Server Colocation
  - Computer and Network Security
  - Software Development
  - Technology Consulting Services
  - Satellite Communication
  - Search Engine
  - Internet Exchange Point (IXP)
  - Other
- Media, Publishing, and Broadcasting:
  - Online Music and Video Streaming Services
  - Online Informational Content
  - Print Media (Newspapers, Magazines, Books)
  - Music and Video Industry
  - Radio and Television Providers
  - Other
- Finance and Insurance:
  - Banks, Credit Card Companies, Mortgage Providers
  - Insurance Carriers and Agencies
  - Accountants, Tax Preparers, Payroll Services
  - Investment, Portfolio Management, Pensions and Funds
  - Other
- Education and Research:
  - Elementary and Secondary Schools
  - Colleges, Universities, and Professional Schools
  - Other Schools, Instruction, and Exam Preparation (Trade Schools, Art Schools, Driving Instruction, etc.)
  - Research and Development Organizations
  - Education Software
  - Other
- Service:
  - Law, Business, and Consulting Services
  - Buildings, Repair, Maintenance (Pest Control, Landscaping, Cleaning, Locksmiths, Car Washes, etc)
  - Personal Care and Lifestyle (Barber Shops, Nail Salons, Diet Centers, Laundry, etc)
  - Social Assistance (Temporary Shelters, Emergency Relief, Child Day Care, etc)
  - Other
- Agriculture, Mining, and Refineries (Farming, Greenhouses, Mining, Forestry, and Animal Farming)
- Community Groups and Nonprofits
  - Churches and Religious Organizations
  - Human Rights and Social Advocacy (Human Rights, Environment and Wildlife Conservation, Other)
  - Other
- Construction and Real Estate:
  - Buildings (Residential or Commercial)
  - Civil Eng. Construction (Utility Lines, Roads and Bridges)
  - Real Estate (Residential and/or Commercial)
  - Other
- Museums, Libraries, and Entertainment:
  - Libraries and Archives
  - Recreation, Sports, and Performing Arts
  - Amusement Parks, Arcades, Fitness Centers, Other
  - Museums, Historical Sites, Zoos, Nature Parks
  - Casinos and Gambling
  - Tours and Sightseeing
  - Other
- Utilities (Excluding Internet Service):
  - Electric Power Generation, Transmission, Distribution
  - Natural Gas Distribution
  - Water Supply and Irrigation
  - Sewage Treatment
  - Steam and Air-Conditioning Supply
  - Other
- Health Care Services:
  - Hospitals and Medical Centers
  - Medical Laboratories and Diagnostic Centers
  - Nursing, Residential Care Facilities, Assisted Living, and Home Health Care
  - Other
- Travel and Accommodation:
  - Air Travel
  - Railroad Travel
  - Water Travel
  - Hotels, Motels, Inns, Other Traveler Accommodation
  - Recreational Vehicle Parks and Campgrounds
  - Boarding Houses, Dormitories, Workers' Camps
  - Food Services and Drinking Places
  - Other
- Freight, Shipment, and Postal Services:
  - Postal Services and Couriers
  - Air Transportation
  - Railroad Transportation
  - Water Transportation
  - Trucking
  - Space, Satellites
  - Passenger Transit (Car, Bus, Taxi, Subway)
  - Other
- Government and Public Administration:
  - Military, Defense, National Security, and Intl. Affairs
  - Law Enforcement, Public Safety, and Justice
  - Government and Regulatory Agencies, Administrations, Departments, and Services
- Retail Stores, Wholesale, and E-commerce Sites:
  - Food, Grocery, Beverages
  - Clothing, Fashion, Luggage
  - Other
- Manufacturing:
  - Automotive and Transportation
  - Food, Beverage, and Tobacco
  - Clothing and Textiles
  - Machinery
  - Chemical and Pharmaceutical Manufacturing
  - Electronics and Computer Components
  - Other
- Other:
  - Individually Owned