**Udacity Classifier Variation Analysis**

This 2-pager summarily describes what was done, why it was done, and what I found. What was found takes the form of 5+ talking points.

**I. What Was Done**

A number of data augmentation services exist. This paper investigates a specific set of classifiers used for data augmentation. I use each classifier to obtain the same parameter estimate, and I compare their levels of confidence and cross-correlation. I try to identify a preferred classifier. I also check about the value-added by using a first name vs a full name, a single image vs two images, and a single sample-wide image source vs two or three. I also check whether people with unusable or missing images vary in systematic or important ways, where unusable images include images with multiple people or irrelevant images. I obtained images from Udacity, GitHub, and LinkedIn. Classifiers utilized include Kairos, NamePrism, NamSor, Kairos, Genderize, Survey, and Manual Review. In addition to reviewing for comparative efficacy and cost on commonly estimated variables, I also discuss non-common data associated with some classifiers.

1. Linkedin pooled; sample differences entirely captured in residence effect; unknown residence indicates from a particular sample.
2. In small sample, nameprism correctly identified ethnicity 100% of the time. Only one non-white sample. Includes some interesting findings like a Brazilian and a Middle Eastern person identifying as white. One might have expected an image-based classifier to consider these people likely Hispanic, but NamePrism correctly identified these person’s self-identified ethnicity.
3. Small sample had an awful response rate; 35 responses for how much money? 9/36 had linkedIn with accepted image, and 3 of these were by word of mouth, not FB/surveycircle.
4. Because only 12 with data, I moved to spreadsheet by hand. Kairos gender and ethnicity as plurality of ethnicities.
5. For small sample, image-based was wrong 1/9 times. NamePrism was wrong 0/12 and namsor was wrong 3/12. Namsor appears to provide richer detail, like British or Irish not just white, but it also appears to have a higher error rate.
6. Relationship of interest is classifier error to any factor, but particularly to gender and ethnicity.
   1. In the small sample, only men were incorrectly identified, and only on ethnicity. Twice by name and once by image.
   2. The image error is interesting because the person does look Hispanic. Namsor identified this person as Portuguese based on the name. Portuguese is considered a white European ethnicity, so namsor correctly aligned with the individual’s self-reported ethnicity. However, suppose this individual were Brazillian Portuguese. This would explain this person’s appearance and name. In that case, the person may be genetically close to the ethnicity of Hispanic, latino, or native American, despite preferring to self-identify as white. This appears similar to another case involving a person named Salah Al Din Deban. This is a clearly Middle Eastern name, but Nameprism and Kairos do not offer a Middle Eastern ethnicity. This individual self-reported their best ethnic descriptor as white among the choices. Interestingly, both Kairos and Nameprism identified this person as white. Namsor identified this person as Congolese, which was coerced to black identification. Either of these cases could be simple classifier error or even self-reporting error, but another interesting possibility is the possibility that these individuals may culturally self-identify one way, while classifiers, particularly image-based classifiers, may be identifying in a biologically-close way. On a very different, but related, note, one sample out of the 35 self-reported a gender of other. Clearly this is not chromosomally possible, but this is exactly the sort of survey disagreement we are attempting to quantify. Notice that I refer to this value as a survey disagreement rather than a survey error. Survey disagreement is the sum of survey measurement error plus non-error variance between machine identification and social or self-identification. When Kairos only produces estimates of male or female gender, not allowing an option for other, it is clearly orienting towards a biological gender identification rather than a social gender identification scheme. Among a class of people self-identifying with a gender of other, Kairos will have a 100% variance rate, but it will not have a 100% error rate because it is expected that some of the people self-identifying with a gender of other will also have a biological sex which aligns to the gender estimate produced by Kairos.
   3. All errors occurred among samples who did not report where they live. However, they all provided linkedIn information indicating United States residence.
   4. All findings in the small sample exploration are with low confidence. The small sample data indicates
      1. Nameprism is the most accurate option
      2. Nameprism, namsor, and Kairos provide some data in common, but they also provide some different data. Image-based classifiers also take different inputs. This means some use cases may prefer any of these tools, even if nameprism is the most accurate option for ethnicity in particular.
      3. Image-based classifier error may be orthogonal to name-based classifier error, so the two methods may be complimentary when both types of input data are available.
      4. Namsor overestimates the likelihood of black ethnicity when applied to United States residents
      5. Kairos may overestimate the likelihood of being Hispanic or non-white among people who describe themselves as white.
   5. Describe the nameprism-namsor ethnicity map and why it’s not controversial and it doesn’t give an unfair advantage to nameprism. It’s a two-directional map and namsor ethnicity maps to nameprism ethnicity via third party info like Irish people are white according to third party info.

**II. Why it Was Done**

Certain standard sociological controls exist. These include age, sex, and ethnicity. A number of different data augmentation services exist, which can derive estimates for those data based on other data. One common pattern is to base such estimates on the name of an individual. Another approach is to use images. When a person’s profile on a public site exists, it is sometimes associated with a useable image. It’s also often possible to find additional images about this person by searching online or by following integrated sites and taking those profile images. Obtaining additional images, however, may have a large cost, so it may not be net beneficial to obtain such data if the gains in model usefulness are small. Also, non-common data inputs or outputs, or cost considerations, might contextually optimize a different solution than the simply most accurate solution.

Two use cases are of particular interest:

1. As a researcher, more accurate estimates of ethnicity will reduce noise in regressions of interest, in which ethnicity is a very common correction variable or even a right hand variable of interest.
2. As an employer, to improve estimates of labor productivity for prospective hires.\*
   1. \*Leeson would be interested in rational name-based discrimination

Previous studies compared NamePrism and Namsor using a metric called agreement. Agreement occurs when two models issue the same prediction for the same inputs. Previous studies took high agreement to reflect mutual accuracy, but logically it could just as easily reflect mutual error. The odds that a single tool is correct is structurally unlikely in this situation[[1]](#footnote-1). This paper explores actual accuracy by surveying live subjects. Other considerations play in to the important decision about which classifier to use:

1. Agreement varied from 60-80%. In other words, there was significant disagreement in almost every case. If such disagreement is systematic, then one or the other tool might be systematically better for certain kinds of analysis.
2. NamePrism, Namsor, and other classifiers can sometimes take similar inputs and produce similar outputs, but they can also take different inputs and produce different outputs. This means one or the other tool may be best for specific use cases.
3. NamePrism specifically rejects use in US samples. Do Namsor and others? If they reject such sampling does that make them impotent? Unlikely. So to what degree are they accurate here?
4. There are several relations of interest which other classifiers do not discuss in the literature. I want to check whether those relations exist.
   1. Married females should have ethnicity guessed with reduced accuracy, and thus, females generally. At least, in patrilineal societies, which are basically all societies.
   2. Matrilineally identified cultural ethnicities should be under-identified, such as Judaism.
   3. Some theories exist to expect different directions of error on transgender identification. Are machines easy or difficult to trick along this margin? This is interesting.

**III. What I Found**

1. Classifiers generally have different inputs and outputs. For example, Namsor requires last names. It also optionally accepts country. Namsor provides gender. NamePrism, on the other hand, doesn’t provide gender and won’t accept a country code. It states that it avoids certain countries like the US, but it accepts first-name-only submissions. Kairos, of course, takes an image, which allows for an entirely different dimension of analysis.
2. Different price considerations.
3. Name-based analysis has systematic issues. It will underrepresent minority status in non-homogenous country populations, particularly among women in patrilineal societies. It will miss matrilineal ethnic identities such as Judaism.
4. What did we actually compare?
   1. Kairos
   2. NamePrism (as reported, without suffix, without initials, without initials lowercase\*, first name, first name lowercase)
      1. \*TODO
   3. NameSor (without initials, without initials lowercase)
      1. Using v1.3.2
   4. TODO: genderize
      1. Hi John,
      2. Thanks for your suggestion, what would be the use case ? There is already two good APIs doing that (genderize and gender-api) ... our differentiation is in recognizing automatically the cultural context for improved precision (ex. Karen Smith is likely female, Karen Petrossian is likely male, etc.)
      3. Happy to discuss your specific need,
      4. --
      5. Best Regards,
      6. Elian CARSENAT
      7. +33 6 52 77 99 07
      8. http://namsor.com/
   5. 2-person independent manual
   6. Email survey
5. NamSor Origin vs. NamePrism Nationality?
   1. NamSor ‘Diaspora’ vs NamePrism ‘Ethnicity’?
   2. <https://blog.namsor.com/2017/09/27/visually-comparing-name-nationality-classification-services/>
   3. <https://github.com/namsor/namsor_nameprism>
   4. <https://arxiv.org/abs/1708.07903>
6. Notice than NamSor name parsing api indicates case insensitivity…or does it? If I send “John” or “JOHN” it returns “john.” But if I request ethnicity for /john/smith and /John/Smith I get a different ID

* I thought it was to get the different classifications using different methods (NamePrism, namsor, etc
* yes. Specifically, to find out which method is most accurate
* Now, we can use Udacity data. That's interesting.
* But that's only one sample.
* Bc UDacity has their self reported ethnicity?
* no.
* Emails
* yes
* alternatively, manual review.
* Gotcha.
* but those are two different left hand variables; we can do both

1. Samples: Udacity, GitHub, LinkedIn, Halfaker?\*diff study, Spokane?\* diff study
2. Classifiers may do a better job at identifying biological sex and ethnicity compared to socially identified or self-reported gender or ethnicity.
   1. Kairos varies from self-report in one case by identifying a person as Hispanic while the person self-reports as White, but it is not clearly an error. It may be that Tiago has White Portuguese ancestry through his father and Hispanic ancestry through his mother. It may also be that he has non-White Portuguese, such as Brazilian, ancestry on either side. In either of these cases he might be genetically closer to the average Hispanic or Native American rather than the average White person. This highlights the difference between biological ethnicity and cultural ethnicity. It may be that Kairos is doing a good job of estimating biological ethnicity rather than cultural ethnicity.
   2. Two similar cases exist in the data. In one case, Salah Al Din Deban self-identified as white. In that case Kairos agreed. This is a clearly Middle Eastern individual who selected the best available ethnic descriptor as White when Middle Eastern was not an available choice. Namsor identified this person as Congolese, which was coerced to black and therefore varies with the self-reported identity. It may be the case that Salah is genetically closer to the average black person than the average white. This study includes no genetic information and makes no conclusions, but the point remains important and is suggested for further study.
   3. A second case involves invalid LinkedIn ID cory-siler-ba6784102, called Cory for discussion. Cory self-identified with a gender of Other. When Kairos only produces estimates of male or female gender, not allowing an option for other, it is clearly orienting towards a biological gender identification rather than a social gender identification scheme. Among a class of people self-identifying with a gender of other, Kairos will have a 100% variance rate, but it will not have a 100% error rate because it is expected that some of the people self-identifying with a gender of other will also have a biological sex which aligns to the gender estimate produced by Kairos.
3. TODO: Describe the nameprism-namsor ethnicity map and why it’s not controversial and it doesn’t give an unfair advantage to nameprism. It’s a two-directional map and namsor ethnicity maps to nameprism ethnicity via third party info like Irish people are white according to third party info.
4. TODO: gain more samples for the survey approach and also use the GitHub-scraper-driven approach to gather emails and linkedin profiles.
5. SurveyMonkey self-report, a note on age of sampled respondents: SurveyMonkey obtained this mix by aligning to US census distribution. The mean age is 53, so young people are underrepresented also. This may seem minor, but remember the overarching motivation is online learning, and this mainly affects young people. There is also reason to thinking classifier accuracy will vary with age. If a classifier is built on a large, old samples of name-ethnicity matches, and if names are subject to cultural fashion particularly within ethnicity over time, as they are, then recent trends may not match long term trends and the result is that classifiers built on temporally wide or old samples will have excess error for young people, named according only to recent trends.

**Appendix**

1. Create a simple 2x2 diagram to show game-theoretic possibilities: right/right, right/wrong, wrong/right, wrong/wrong [↑](#footnote-ref-1)