



In the Interviews & Surveys lecture we have discussed how to ask about gender. However, asking people does not scale up. Hence, today we illustrate the notion of advanced repository mining by discussing how gender (and national cultural background) can be “guessed” based on the repository data.



Created by RF_Design
from Noun Project

REMINDER: gender is a complex social construct and any kind of automated detection will necessarily make simplifying assumptions. We do not discuss biological sex.



Whatever technique we use, we should keep in mind that gender is privacy sensitive and should be treated as such. Open source contributors might be hiding their gender on purpose, e.g., many women-developers prefer not to disclose their gender due to safety concerns. Some open source projects do not necessarily want us to know the genders of their members (but some do!) and companies might be sensitive to this topic as well.

Where do you identify on the gender spectrum?

Your answer _____

<https://www.r>

Do you have experience in Java programming?

Yes No

How many years of Java programming experience do you have?

9

How much experience do you have in Java programming?

I got my first Java certificate in 2009, and I have been working with Java at companies since 2011. I am familiar with Struts and Hibernate but not with Google web toolkit.

Inspired by <https://www.slideshare.net/mendezfe/surveys-in-software-engineering>

You might remember that when discussing interviews and surveys we have discussed how to ask about gender and experience.



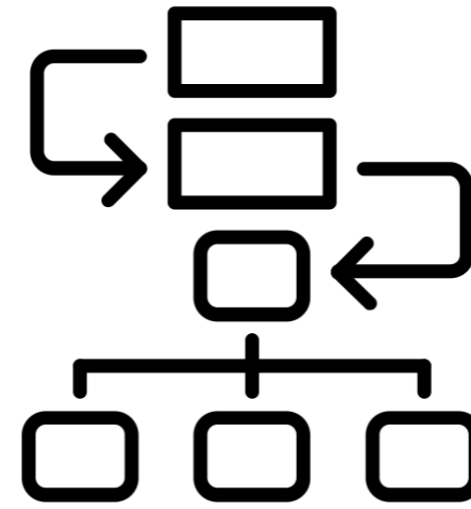
Created by Gan Khoo Lay
from Noun Project

10-20%

However, whatever survey techniques we use and however we ask the questions, two open problems remain: scale of the data and lack of response. This is a problem if we want to perform a large scale data analysis to tease out minor effects using traditional statistical techniques, since to apply these techniques we need a lot of data to ensure the power of statistical tests. Our study in 2019 has involved ~60K individuals. To get this number we will need to survey ~300K-600K developers; if everyone spams 600K respondents the respondents will be even more fed up with us and will not answer our questions...



Created by Gan Khoo Lay
from Noun Project



Created by QualityIcons
from Noun Project


Enter automatic gender detection mechanisms





All these tools are based on the main assumption, namely, that gender can be inferred from the way developers present themselves (username, name, avatar) or artefacts they produce (code, comments, etc.)

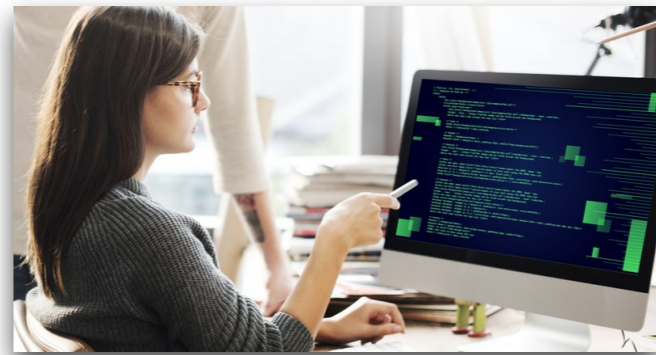


YOUR RESULT


Man


From 31 to 38 years old


Redo



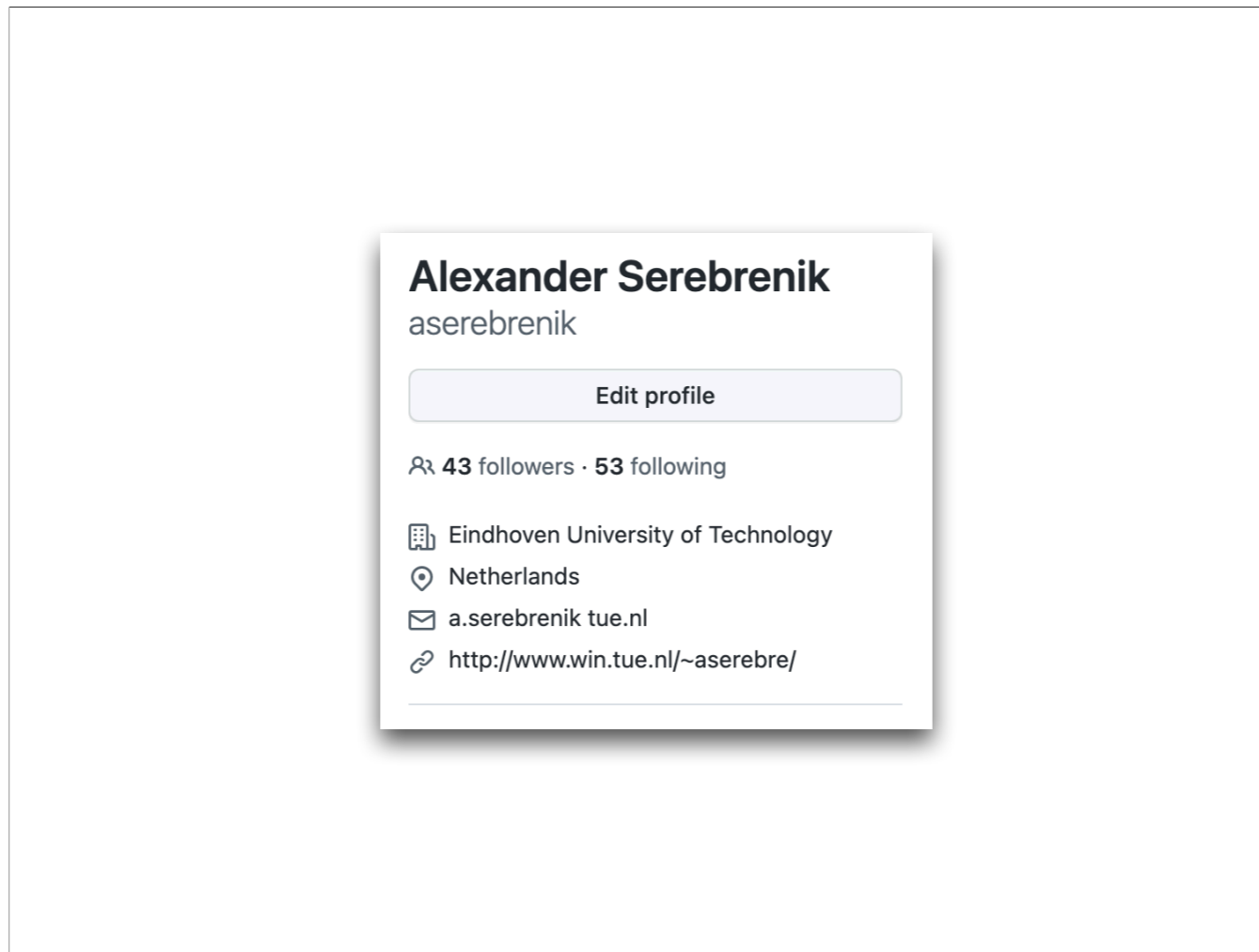
Names, profile pictures, artefacts



The practice of associating certain names with people of a certain gender is well established and in some countries it is even recorded in laws and administrative procedures. This is the case for Belgium, where by law the first name should not be confusing. Many local administrations interpret it as “no girls’ names for boys, no boys’ names for girls”.



However, the data we analyse comes from a mix of different countries, and certain names are more commonly associated with men in some countries and with women in other countries. Andrea: IT vs DE. Karen: Armenia vs USA. Etc etc. We will discuss inference of national culture based on names in a moment.



The first proxy that is often used in the literature is location as indicated by the respondent. I am showing my profile since I am the only person who has provided consent to share their data :)

However, on StackOverflow ca 25% users provide location and this location is not always reliable (/dev/null). Moreover, location is where I work, not where I live, what nationality I have or where I have been born!

Country of residence



Establish a country of residence based on a last name or a full name.

Script: LATIN

Score: 0.6128163899275012

Country or residence: RU Russia

Alternative country of residence: UA
Ukraine

Region: Europe

Country of origin



Determine a country of origin based on a first name and a last name.

Script: LATIN

FirstName: Alexander

Last name: Serebrenik

Country of origin: UA Ukraine

Alternative country of origin: RU Russia

Top countries of origin: UA Ukraine, RU

Russia, SK Slovakia,

Poland, SI Slovenia,

Czechia, DE Germany

Score: 0.946926760

Region of origin: Eur

Top region of origin:

Sub region of origin:

Probability: 0.42360

Probability for alter

0.4262431971384

Identify the diaspora of a name. Country code defined by default: US.

Script: LATIN

FirstName: Alexander

Last name: Serebrenik

Score: 19.69730530161225

Alternative ethnicity: Russian

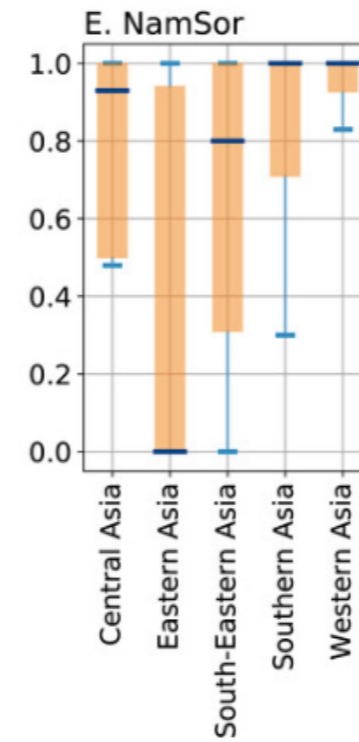
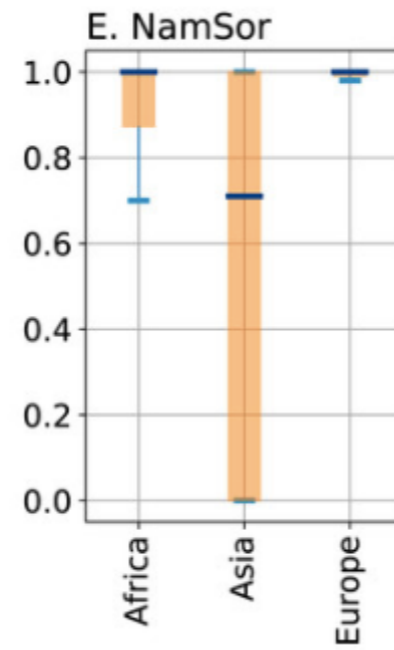
Ethnicity: Jewish

Lifted: false

This is why more advanced tools such as NamSor are trying to split this into several elements. The answers are again about me.

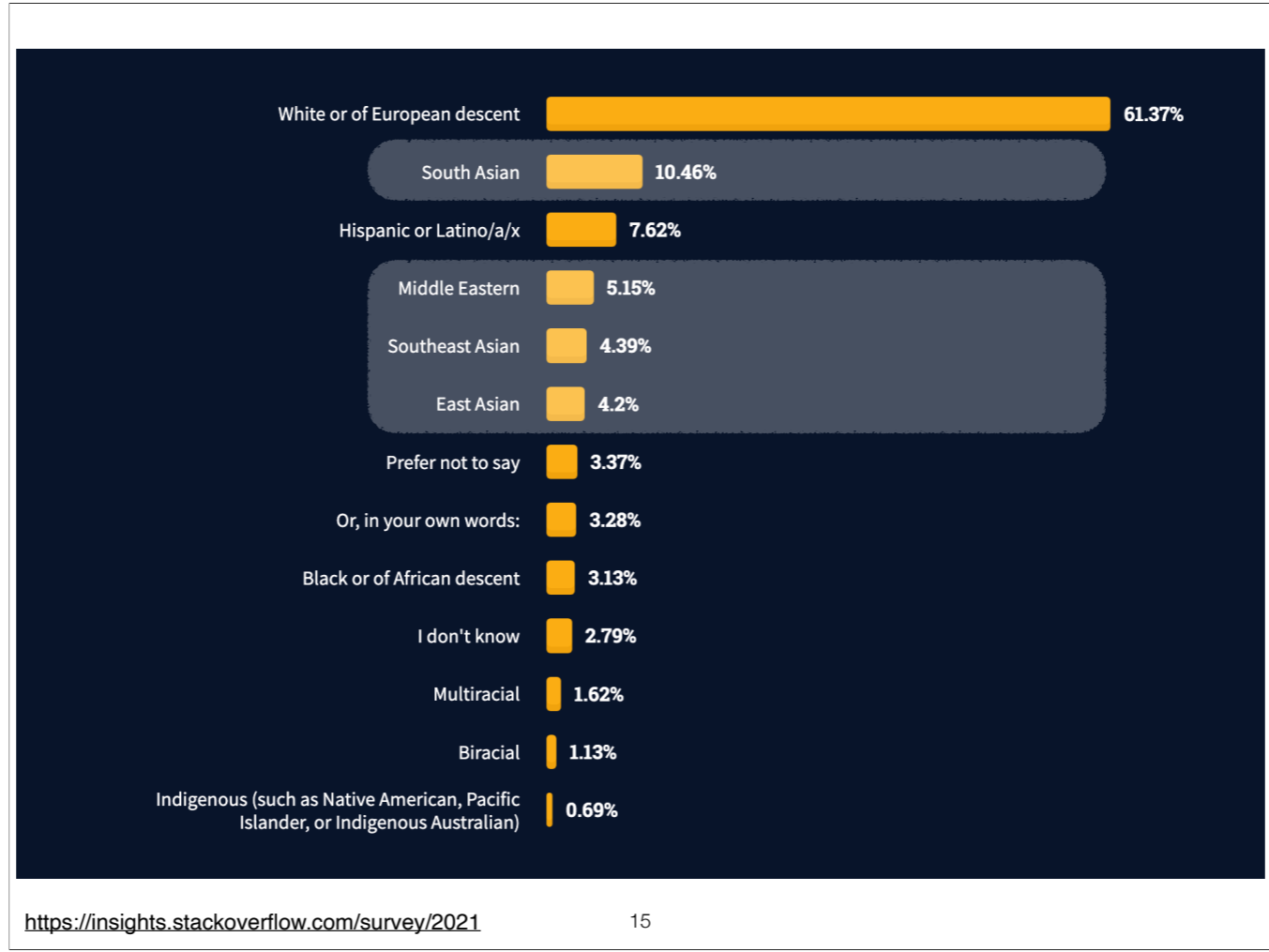


And these tools change! Most recent version of NamSor

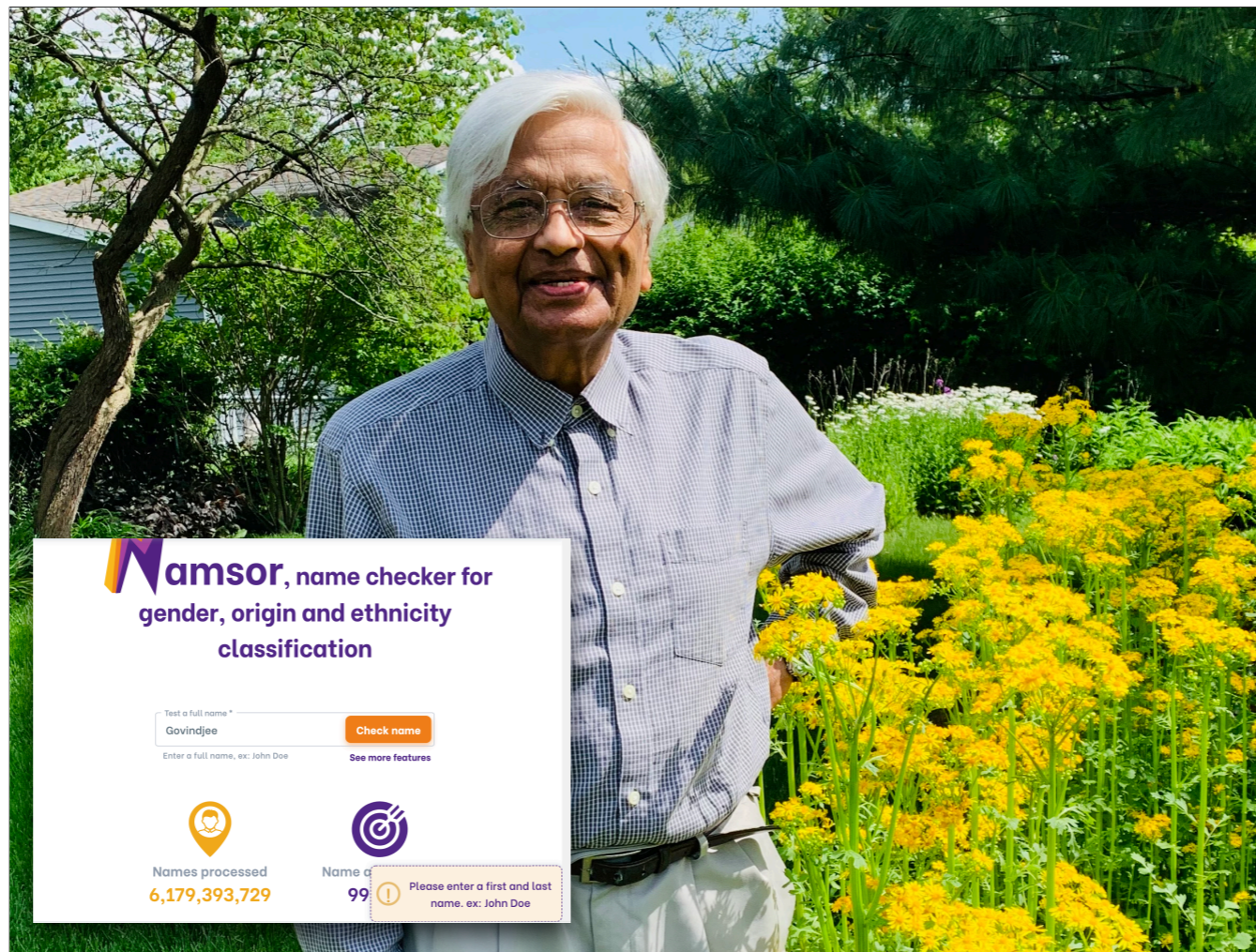


Lucia Santamaría and Helena Mihaljević (2018), Comparison and benchmark of name-to-gender inference services. PeerJ Comput. Sci. 4:e156; DOI 10.7717/peerj-cs.156

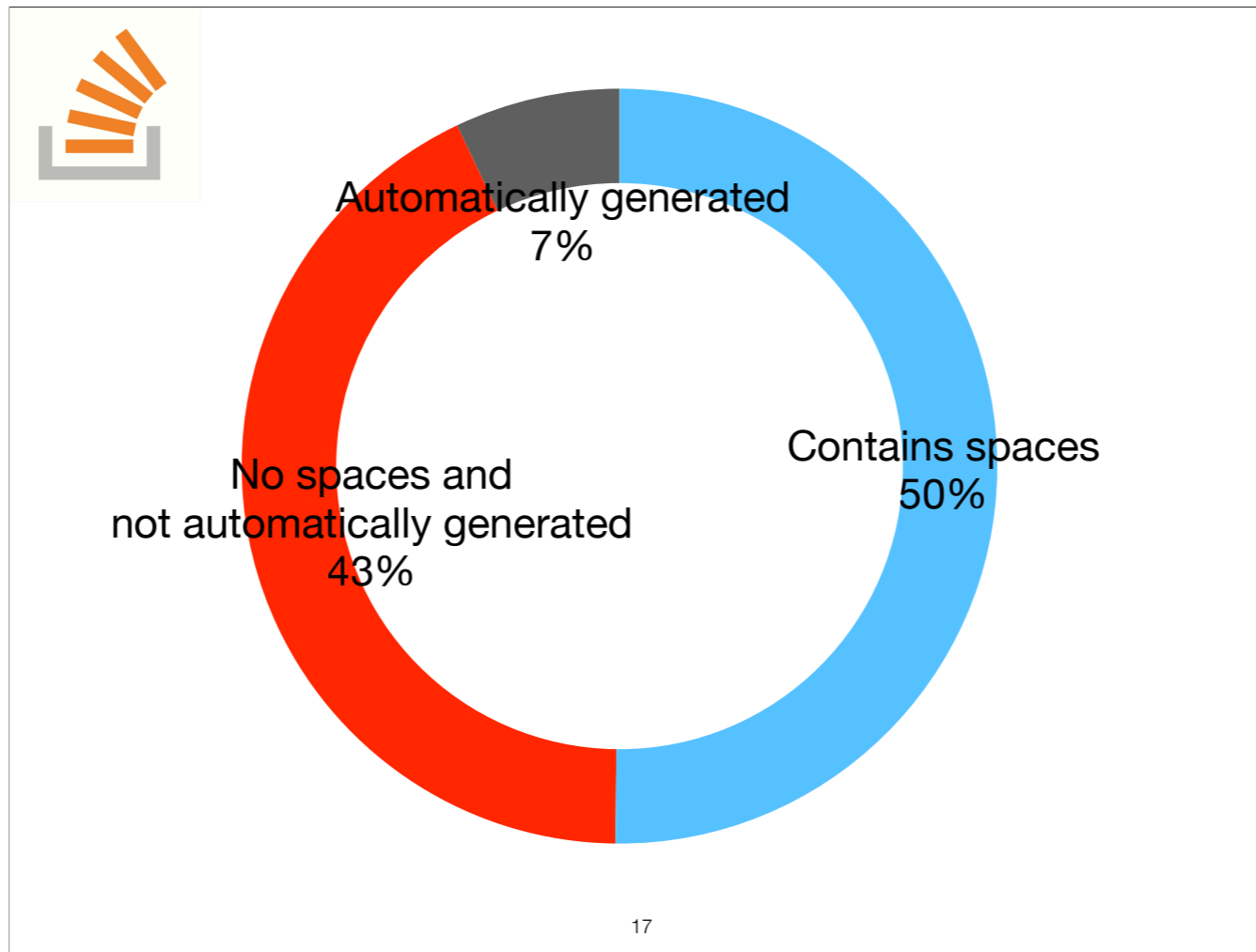
Closer inspection reveals a different story, however. Confidence of NamSor (version 2018) drops the we move to Asian names and particularly Easter and South-Eastern Asian names. Half of the East-Asian names have a confidence score of 0!



And this is indeed deeply problematic when trying to apply automatic gender inference techniques to software developers: looking at the recent Stack Overflow survey we see that almost one out of four software developers have indicated different Asian regions as their ethnic background.

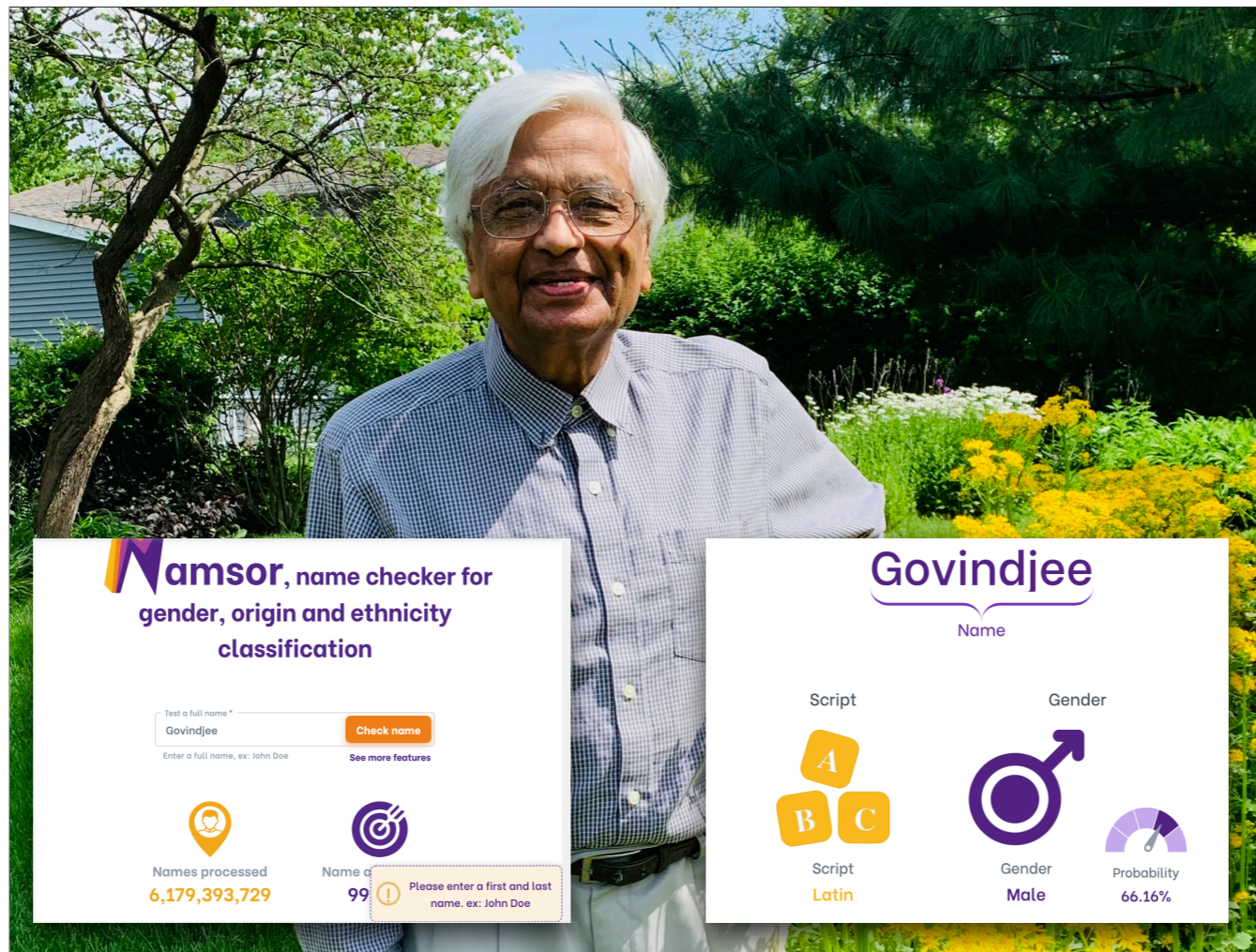


Moreover, this is of course only working if the name has first name and last name which is by far not always the case. This is Govindjee, an Indian-American professor emeritus of Biochemistry, Biophysics and Plant Biology. He is recognized internationally as a leading expert on photosynthesis. Indeed, many tools fail to handle mononymic names!

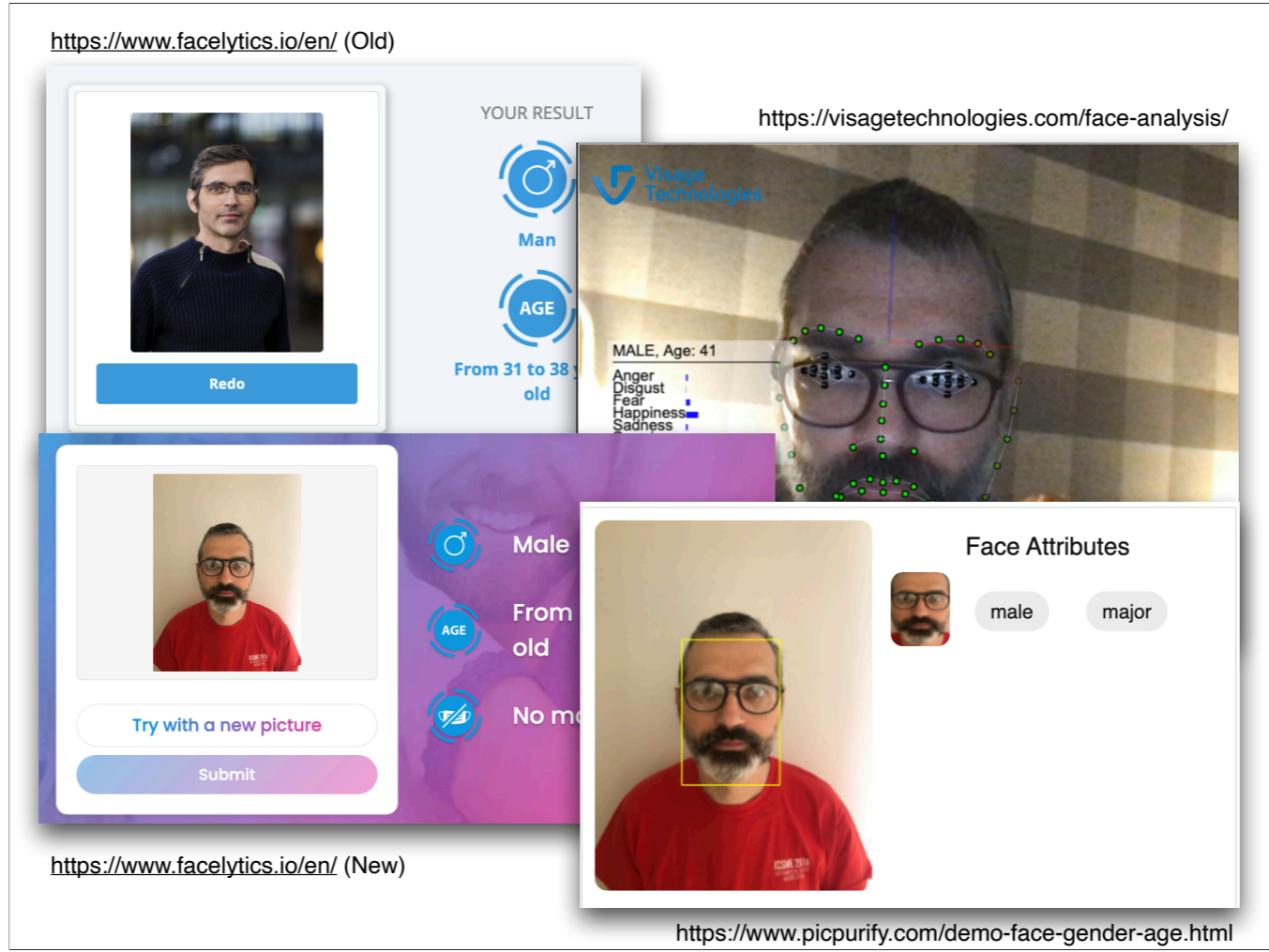


However, we do not know how often software developers use mononymic names. The grey segment indicates percentage of the SO contributors with automatically generated usernames such as user12345. For these contributors no inference technique can be successful; both the red and the blue segments can be analysed by techniques such as genderComputer and NamSor; the red ones only by genderComputer

```
select count(Id)
from Users
where DisplayName LIKE CONCAT('user',LTRIM(STR(Id)))
```



Luckily a more recent version of Namer seems to do something meaningful here



Another way developers present themselves on social platforms is by using face recognition techniques; here we see that all tools have correctly identified my gender. Age-wise they are off scoring me from 12 years younger to 10 years older than my actual age.



~30%

autogenerated profile images

Another example: not everybody has a meaningful profile picture. For instance, ca. 30% of the Stack Overflow users only have a default profile picture automatically generated based on the MD5 hash of the users' mail

	Age not indicated	15-25	26-31	≥32
Reputation 1-199	150	50	50	50
Reputation 200-999	150	50	50	50
Reputation ≥1000	150	50	50	50

Bin Lin, Alexander Serebrenik: Recognizing gender of stack overflow users. MSR 2016: 425-429

Moreover not all profile images represent faces (rather than logos or cat pictures). This is why we have carefully selected 900 non-generated profile images and classified them manually. Reputation classes are related to different privileges associated with these classes; age intervals to the general distribution of the ages on SO


53% (479/900)



Bin Lin, Alexander Serebrenik: Recognizing gender of stack overflow users. MSR 2016: 425-429

Moreover not all profile images represent faces (rather than logos or cat pictures). This is why we have carefully selected 900 non-generated profile images of users of different ages and reputations, and classified them manually. Reputation classes are related to different privileges associated with these classes; age intervals to the general distribution of the ages on SO

YOUR RESULT



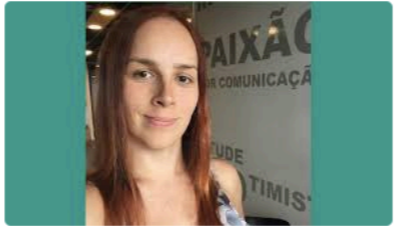
Redo

Man

AGE

From 33 to 40 years old

YOUR RESULT



Redo

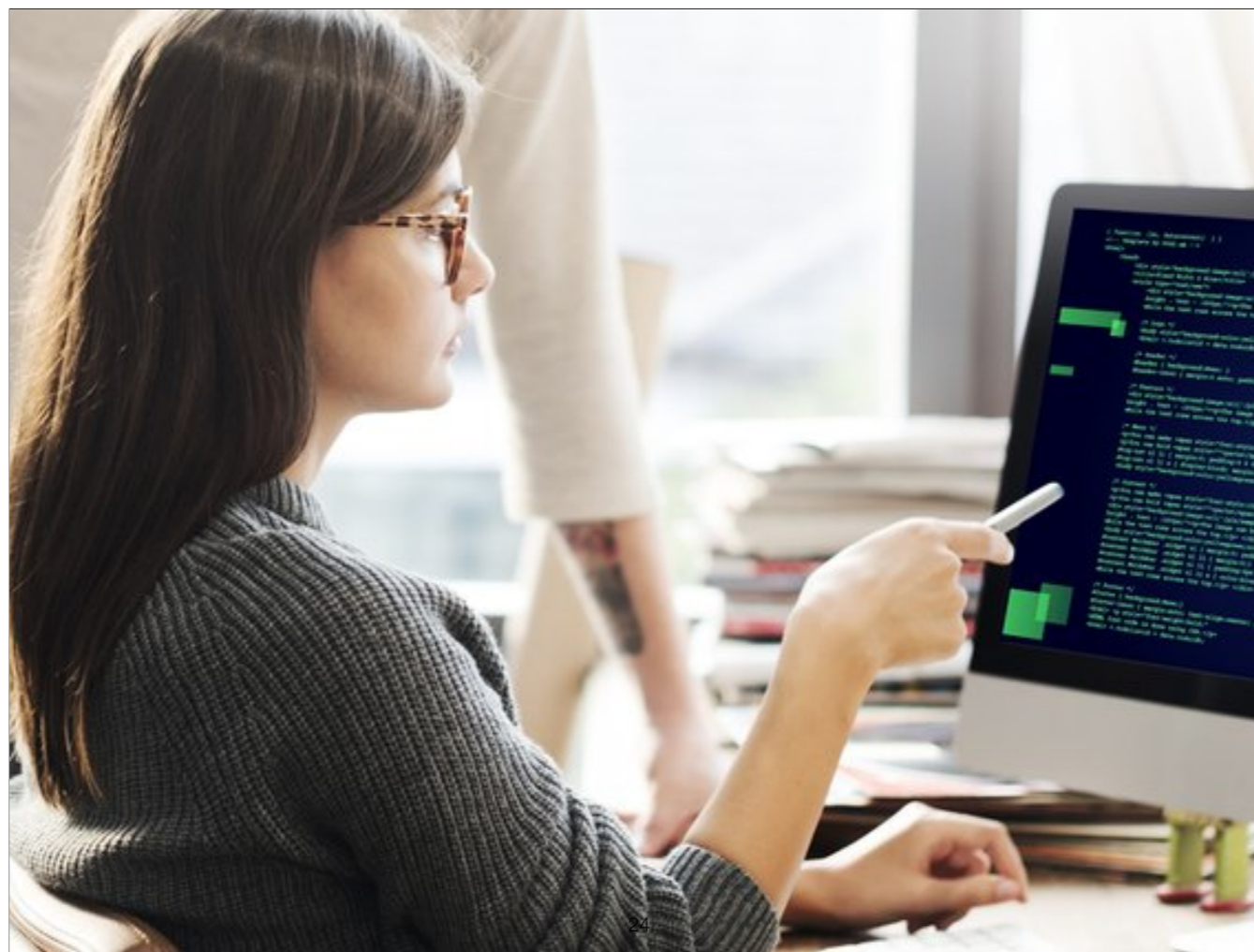
Woman

AGE

From 27 to 34 years old

23

Example of inaccurate detection



Finally, let us move to the discussion of artefacts created by software developers

Table 1: Classification of Automated Gender Identification Approaches 2017 - 2019

Approach	Training Data	Test Data	Best Reported Accuracy
Team nissim17 at PAN 2017	240,000 Arabic Tweets	160,000 Arabic Tweets	68.31%
	360,000 English Tweets	240,000 English Tweets	74.3%
	420,000 Spanish Tweets	280,000 Spanish Tweets	80.4%
Team miura17 at PAN 2017	120,000 Portuguese Tweets	80,000 Portuguese Tweets	85.75%
Tellez et al. [25] at PAN 2018	150,000 Arabic Tweets	100,000 Arabic Tweets	81.7%
Daneshvar and Inkpen [6] at PAN 2018	300,000 English Tweets	190,000 English Tweets	82.21%
	300,000 Spanish Tweets	220,000 Spanish Tweets	82.0%
Team deepCybErnet at IberEval 2017	4,319 Catalan Tweets	1,081 Catalan Tweets	48.6%
González et al. [7] at IberEval 2017	4,319 Spanish Tweets	1,081 Spanish Tweets	68.6%
Markov et al. [17] at RusProfiling 2017	Tweets from 300 Users	Tweets from 200 Users	68.3%
	Facebook Posts from 228 users		93.4%
	Reviews from 776 authors		61.9%
Bhargava et al. [2] at RusProfiling 2017	Essays from 370 authors		78.4%
	Texts from 94 authors in Gender Imitation Corpus		66%
Bsir and Zrigui [3]	240,000 Arabic Tweets	160,000 Arabic Tweets	79.2%
Bsir and Zrigui [4]	240,000 Arabic Tweets	160,000 Arabic Tweets	79%
	Facebook Posts from 4,444 users		62.1%
	Sanchez-Perez et al. [20]	5,187 Spanish news articles (using cross-validation)	
Company and Wanner [5]	4,284 Journalistic Posts		89.97%
	48 Novels		91.78%

Stefan Krüger, Ben Hermann. Can an Online Service Predict Gender? - On the State-of-the-Art in Gender Identification from Texts. Gender Equality Workshop ICSE 2019

When it comes to gender recognition based on the artefacts created most of the approaches consider blog posts and Twitter data. For example the work of Company & Wanner has been designed in the first place for authorship attribution; similar authorship attribution techniques have been designed for the source code, usually with a limited number of prospective candidate authors (up to 60, but in one of the studies more than 200). Still, code is a much more constrained use of language...

TABLE III
LIST OF ATTRIBUTES

Type of Attributes	Fifty Attributes
Keywords	#include, #define, using, void, cout, cerr, cin, return, exit, int, float, char, const, double, bool, new, break, public, private
Operators	<, ->, >, &, &&, +, ++, !, !=, ==, =, -, --, *, /, , , /=, +=, -=, *=, <=, >=
Comments	//*, //, /*
Brackets	{}, ()
Loops	for, while, switch

TABLE V
K* CONFUSION MATRIX

Gender	Female	Male
Female	39	11
Male	17	33

Fariha Naz, Jacqueline E. Rice: Sociolinguistics and programming. PACRIM 2015: 74-79

This is why Naz and Rice have collected code samples from students of different genders, removed whitespaces and comment texts and used the attributes from the table to predict gender of the authors. The results are not spectacular but the accuracy of 72% is not far behind the techniques designed for natural language texts. And yes, a more extensive and more careful study would be required to confirm these findings.



So yes, the automated genderization works and produces (more or less) reasonable results meaning that gender information can be inferred from such information as names, profile pictures and code/text written



However, we need to remember that these methods are far from being perfect and one has to be very careful when applying them



The first group of concerns are ethical. They are mostly raised in relation to face-to-gender techniques but similar concerns can be raised for any genderization method and is related to assigning any kind of categories to human beings without their explicit consent. Of course, as humans we might be doing it continuously, e.g., when we are describing people, this kind of automation might be dangerous, e.g., what if we recognise woman driving a car in a country where women are not allowed to drive cars? Uygur example. In fact, Nature has surveyed ~500 researchers in facial recognition, CS and AI and about 2/3 believe that application of facial-recognition methods to recognize or predict personal characteristics (such as gender, sexual identity, age or ethnicity) from appearance, should be done only with the informed consent of those whose faces were used, or after discussion with representatives of groups that might be affected. Getting an informed consent from all GitHub or StackOverflow developers is not realistic.

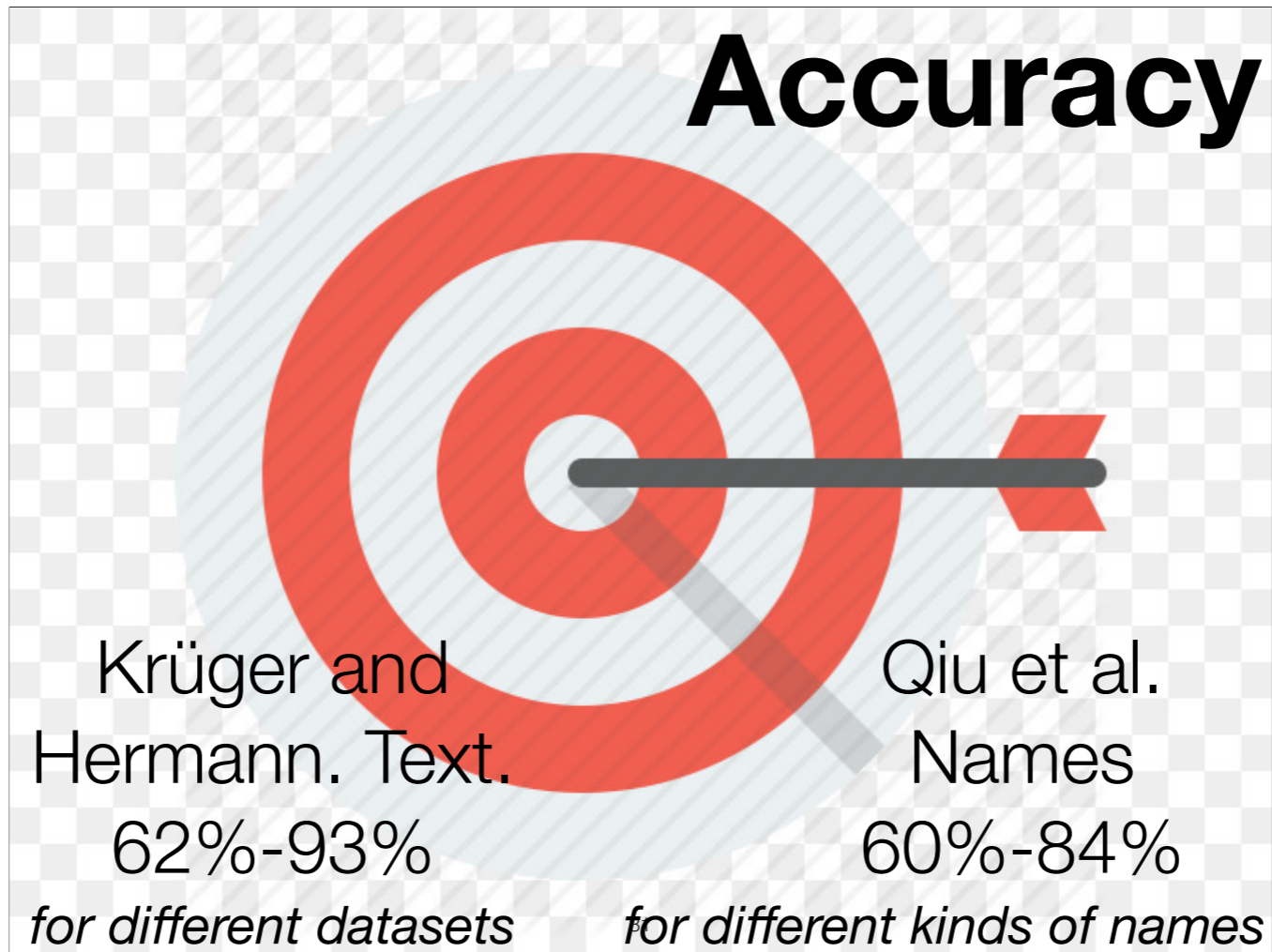
Reliability

“I have used a **fake GitHub handle** (my normal GitHub handle is my first name, which is a distinctly female name) so that **people would assume I was male**”

Bogdan Vasilescu, Vladimir Filkov, Alexander Serebrenik:
Perceptions of Diversity on Git Hub: A User Survey. CHASE@ICSE 2015: 50-56

And again, this leads to the situations when machines are “smarter” than humans and capable of breaking the illusion that the person has consciously created.

Accuracy



The accuracy of our techniques is not perfect. It can be even lower for some subcommunities, e.g., for Chinese names, when some of the gender-specific information is lost during the romanization.

Gender binary

Krüger and Hermann. Text. 100%

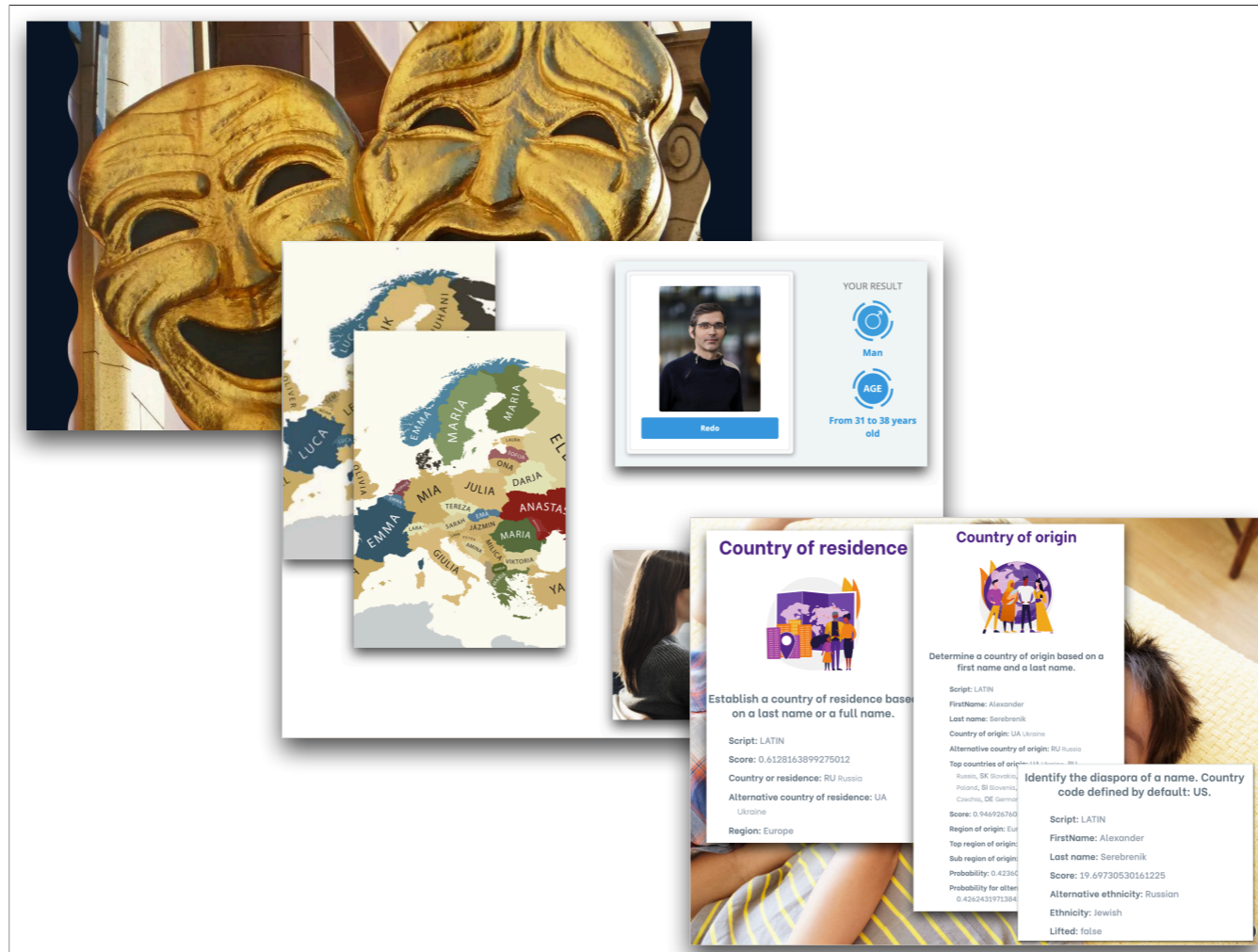
Keyes. Face. 92.9-96.7%

Santamaría and Mihaljević. Names.

Stefan Krüger, Ben Hermann. Can an Online Service Predict Gender? - On the State-of-the-Art in Gender Identification from Texts. Gender Equality Workshop ICSE 2019

Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. CSCW 2018

Most automatic techniques we discuss assume gender binary. These are percentages of papers reviewed in two meta-studies. Keyes: the first number corresponds to the % in papers that introduce automatic gender recognition and the second one - to papers that use automatic gender recognition. The situation with names is a bit better since the tools tend to be probabilistic and at least recognise their own lack of confidence.



To summarise: we have seen how automatic tools can be used to enrich the data to augment with automatically inferred information about emotion or demographics. In its turn this information can be used for follow up analysis, quantitative or qualitative. This being said one has to remember that automatic tools might have limited applicability or accuracy, and their application requires utmost care.