How You Post is Who You Are:

Characterizing Google+ Status Updates across Social Groups

Evandro Cunha, Gabriel Magno,
Marcos André Gonçalves, César Cambraia, Virgilio Almeida

Universidade Federal de Minas Gerais (UFMG) – Brazil

Growing interest in *language in social media*

Vast amount of data for linguists

In this specific study

We analyze G+ posts from...

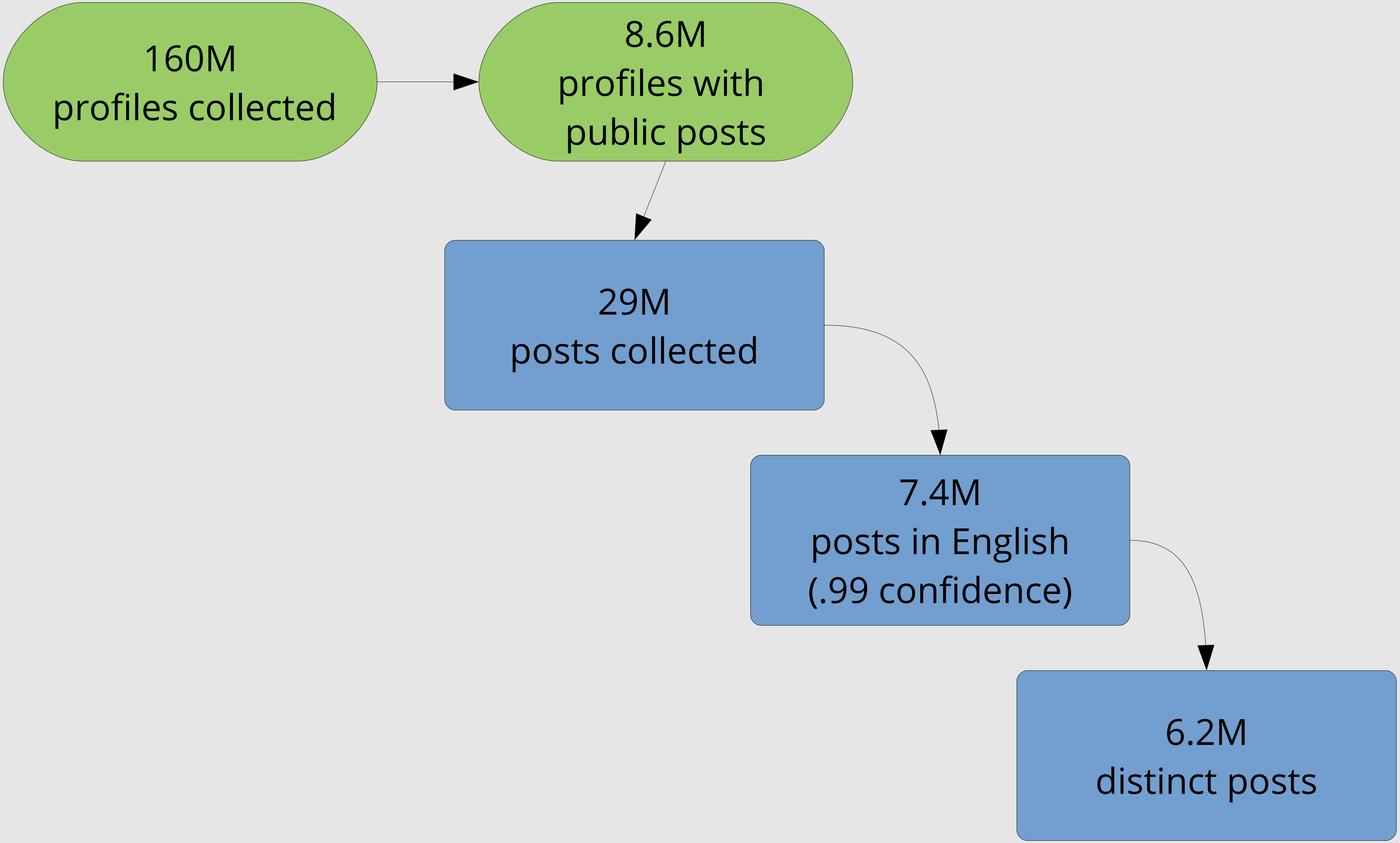
- female and male
- 10 countries
- 15 occupations

To better understand...

- the textual genre 'post'
- collective aspects of OSN users

We all know Google+

Data



160M
profiles collected

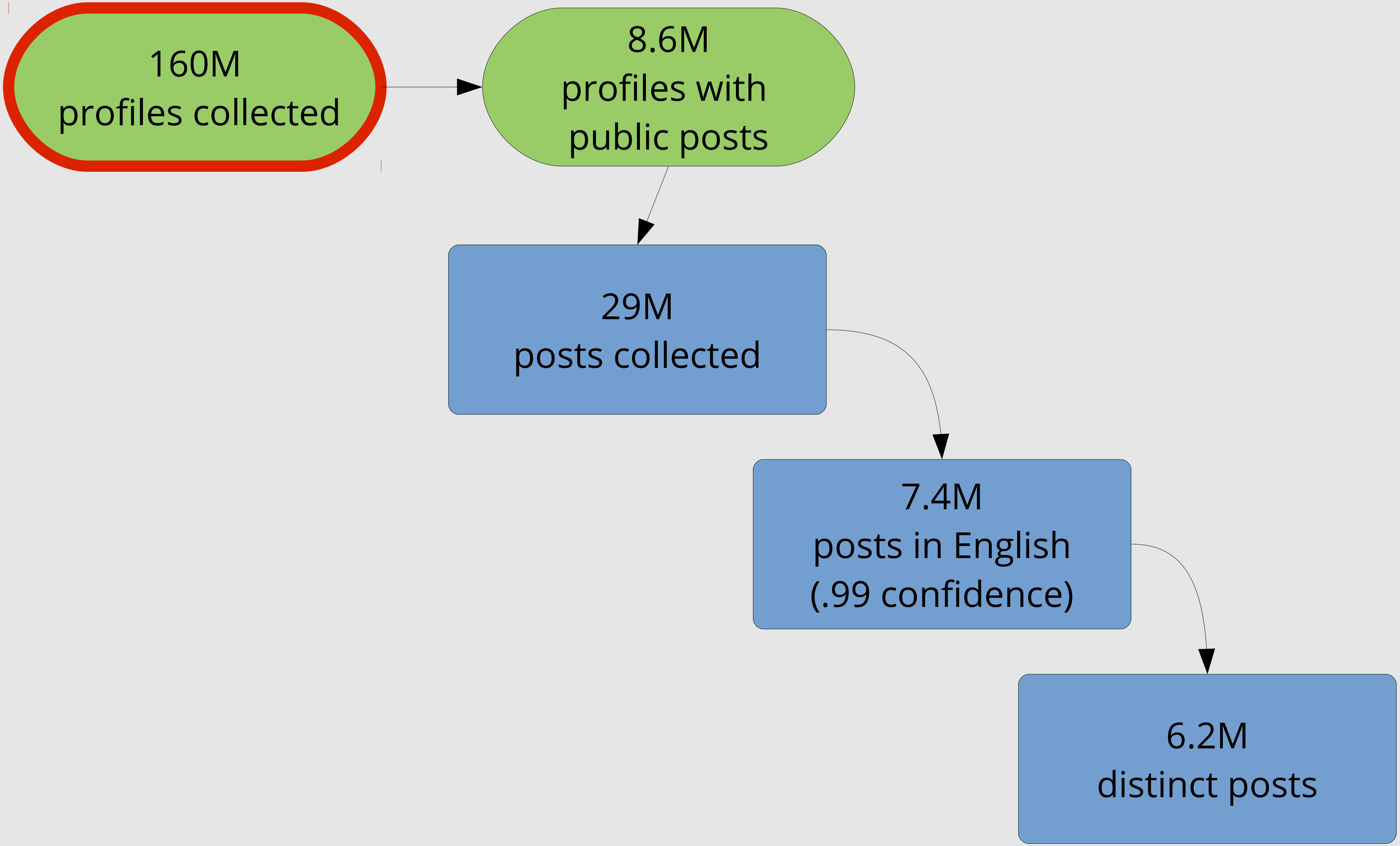
8.6M
profiles with
public posts

29M
posts collected

7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Data



160M
profiles collected

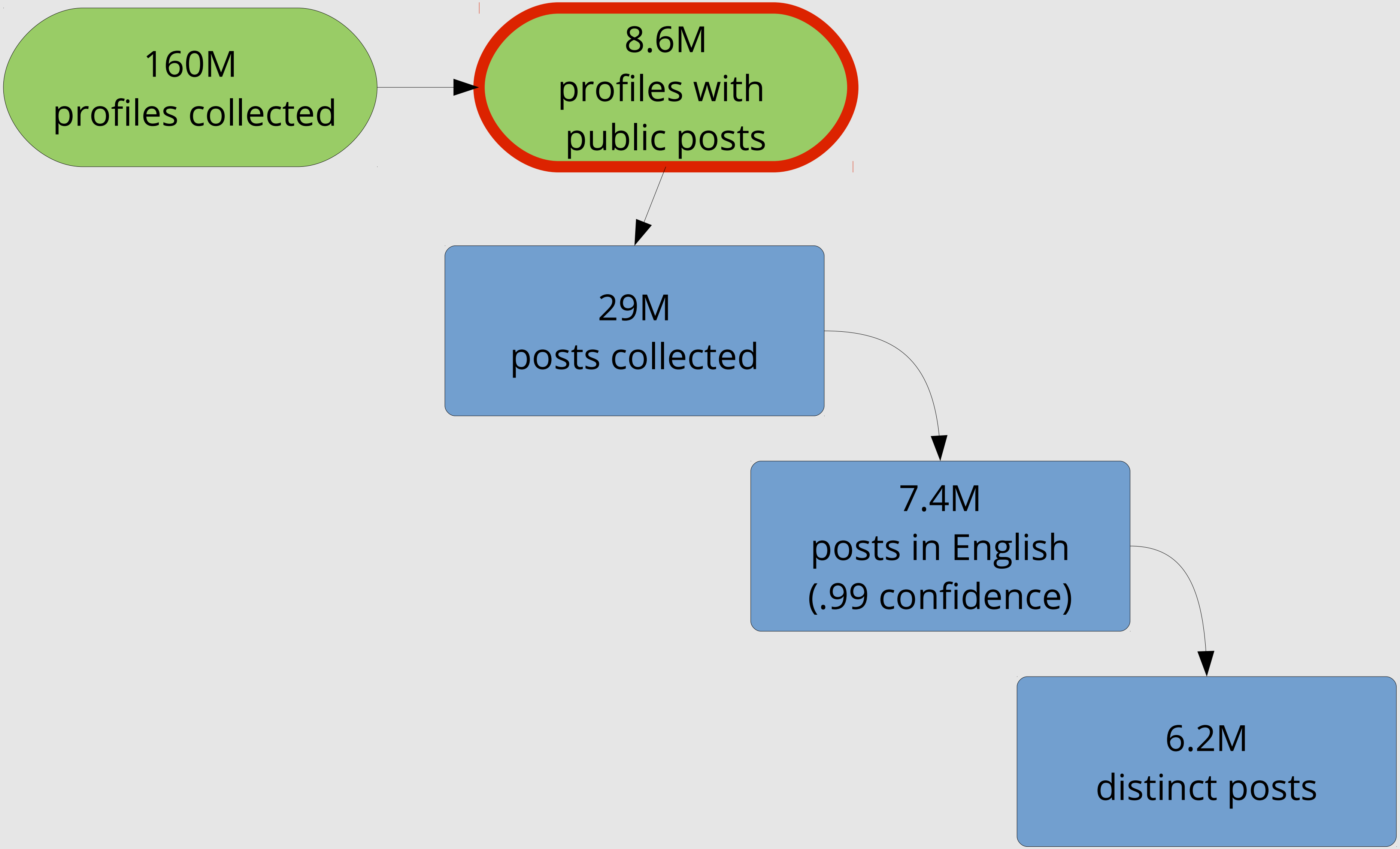
8.6M
profiles with
public posts

29M
posts collected

7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Data



160M
profiles collected

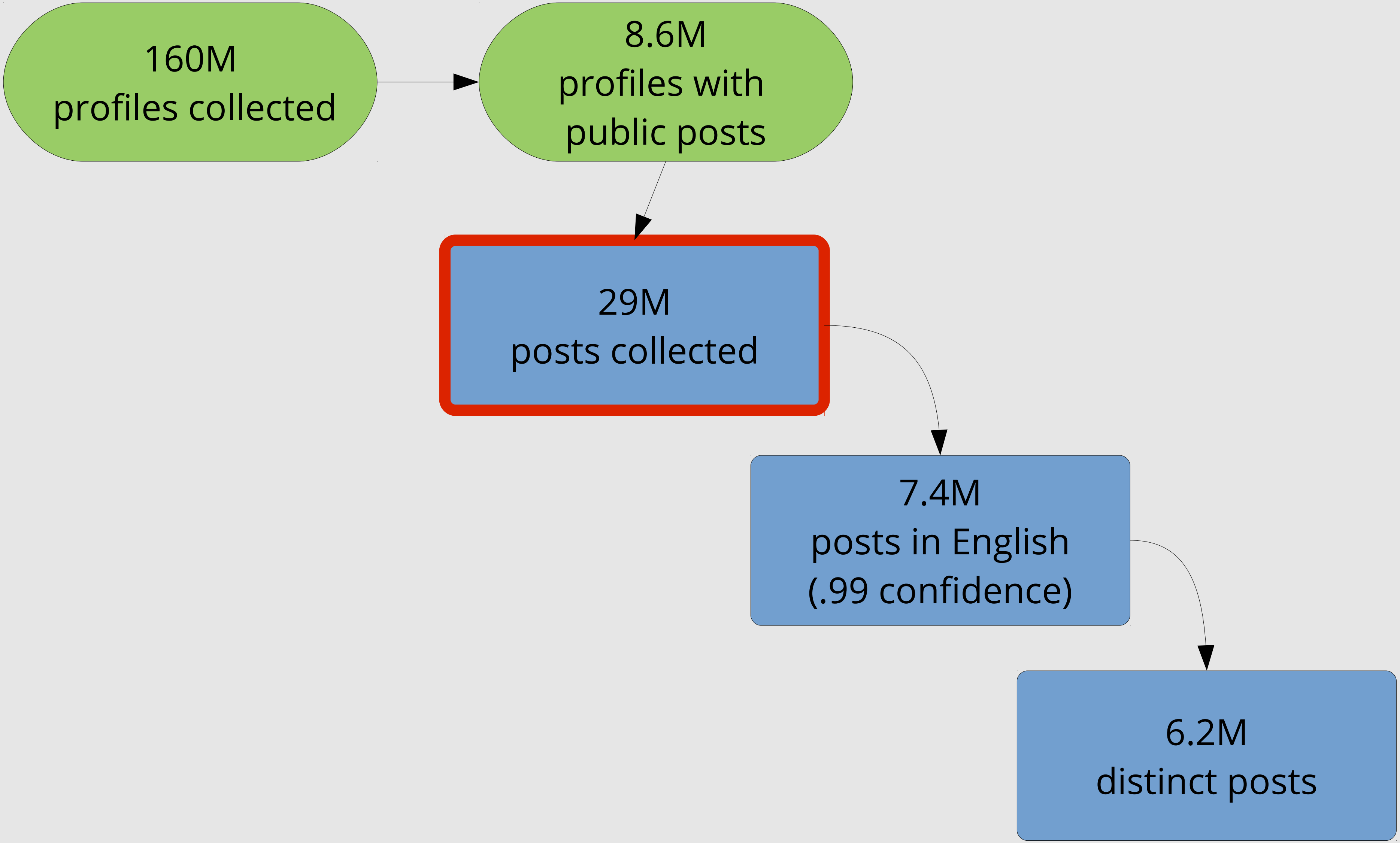
8.6M
profiles with
public posts

29M
posts collected

7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Data



160M
profiles collected

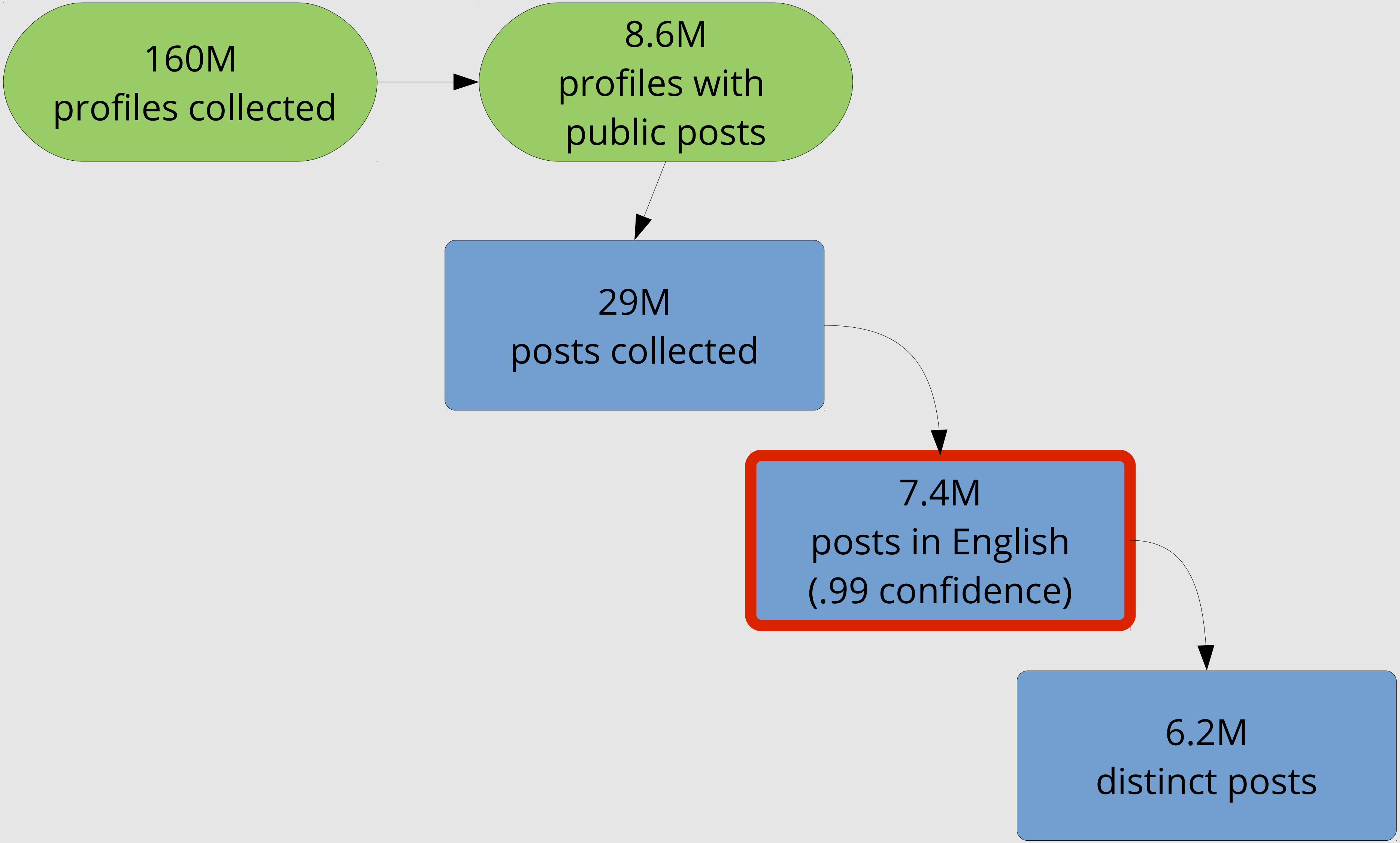
8.6M
profiles with
public posts

29M
posts collected

7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Data



160M
profiles collected

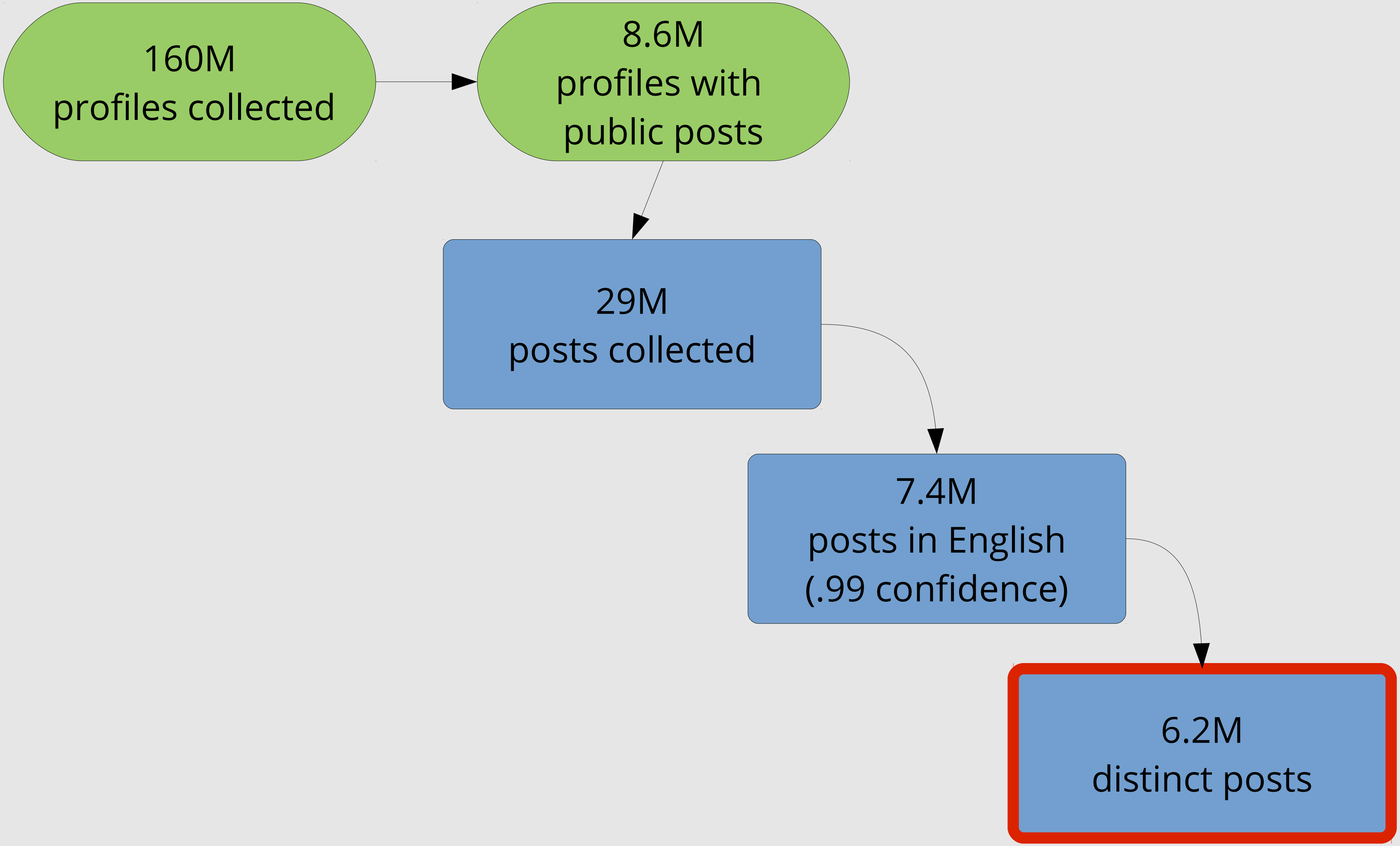
8.6M
profiles with
public posts

29M
posts collected

7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Data



160M
profiles collected

8.6M
profiles with
public posts

29M
posts collected

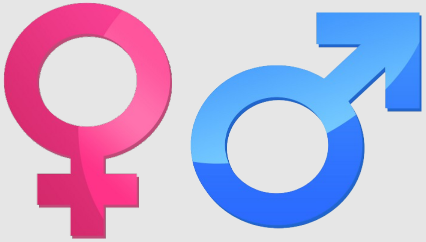
7.4M
posts in English
(.99 confidence)

6.2M
distinct posts

Social groups

{Countries}

{Genders}



{Occupations}

- Architect. and engin.
- Arts and design
- Business and financial
- Computer and math.
- Educ. and library
- Food prep.
- Healthcare
- Legal
- Management
- Media
- Religious
- Retired
- Sales
- Science
- Student

Analyses

Fraction of misspellings

Readability and complexity

Vocabulary variability

Semantic categories of words

Analyses

Fraction of misspellings

Readability and complexity

Vocabulary variability

Semantic categories of words

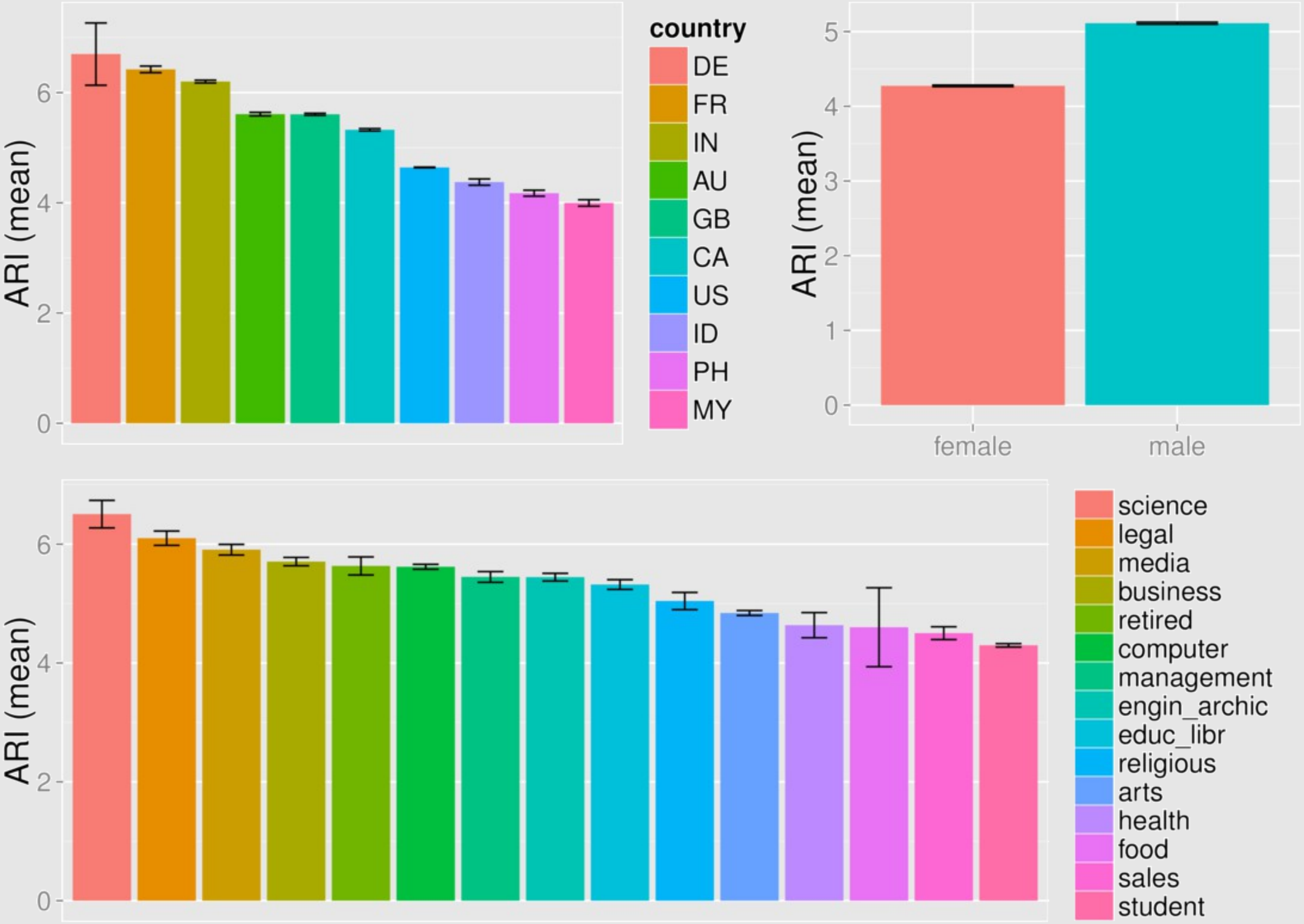
Readability and complexity

Automated Readability Index (ARI)

+ characters per word → + complexity

+ words per sentence → + complexity

Readability and complexity



How you post is who you are: characterizing Google+ status updates across social groups

Readability and complexity

Indo-European speakers → highest ARI
Austronesian speakers → lowest ARI

Male users → highest ARI

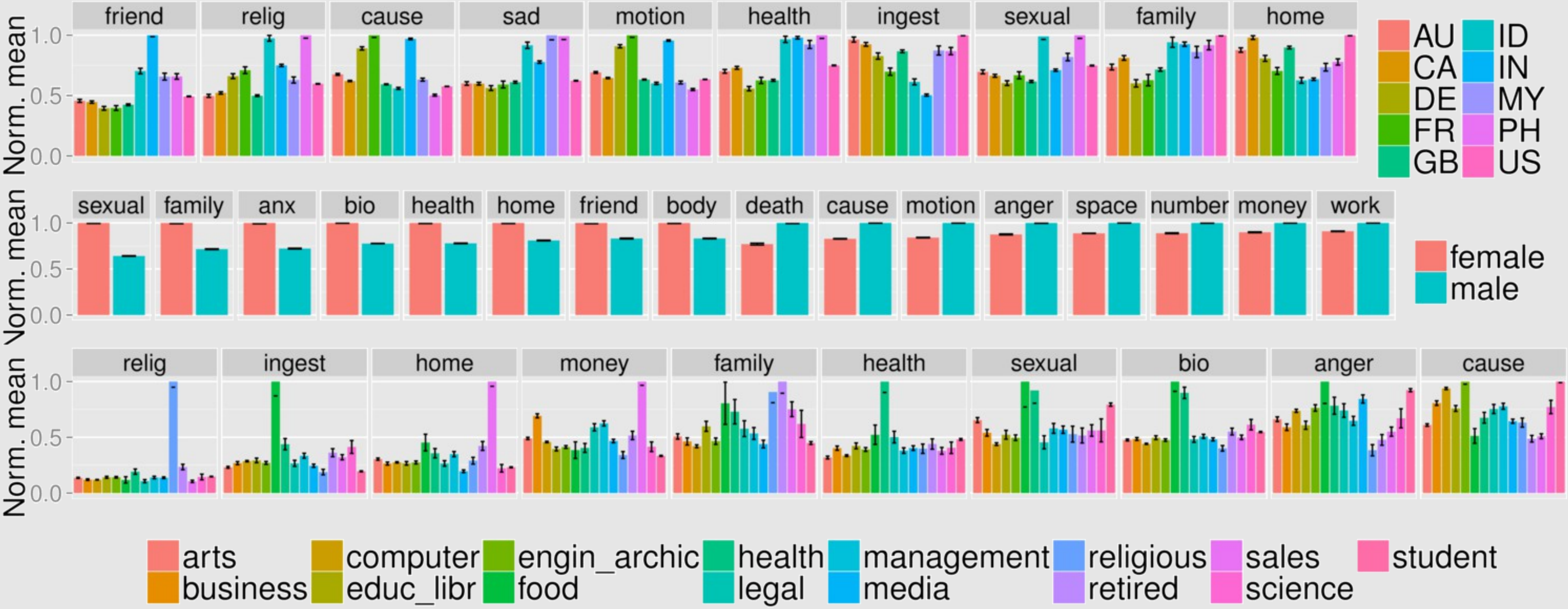
Workers who deal with texts → highest ARI

Semantic categories of words

Analysis of vocabulary

Language Inquiry and Word Count (LIWC)

Semantic categories of words



How you post is who you are: characterizing Google+ status updates across social groups

Semantic categories of words

Indians → +*friend*, +*humans*, +*social*
-*negative emotions*, -*anger*, -*time*

Western countries → +*home*, +*money*, +*work*,
-*health*, -*affection*, -*positive emotions*, -*family*

Semantic categories of words

Women → *+family*, *+home*, *+friend*, *+social*, *+humans*, *+affection*, *+emotions*

Men → *+cause*, *+motion*, *+space*, *+numbers*, *+money*, *+work*

Semantic categories of words

Correlation between words and occupations

Vocabulary related to working activities

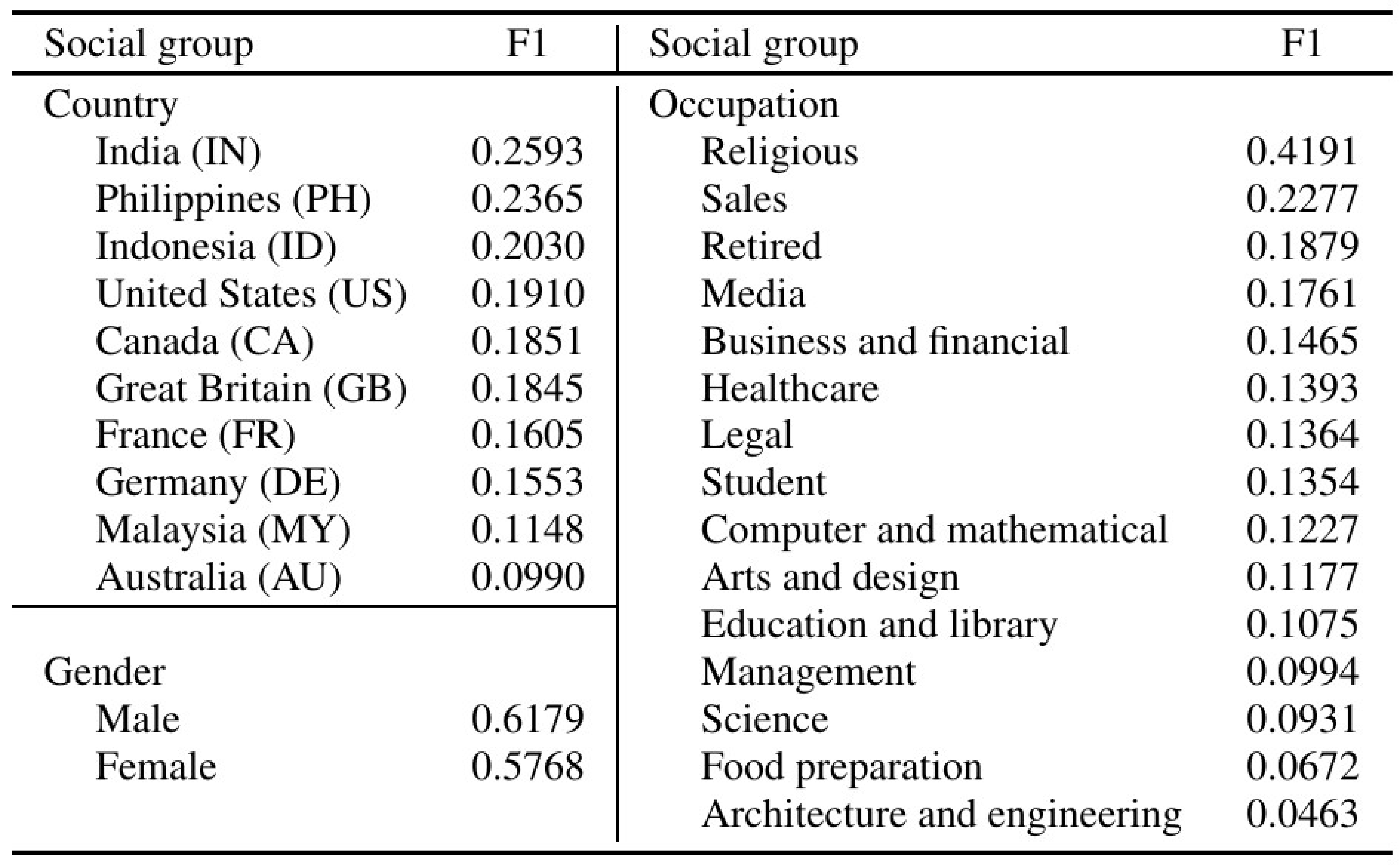
An (preliminary) application

Inference of social groups

Authorship attribution for…

- …personalization of services
- …identification of fake profiles
- …cyber crime investigation

An (preliminary) application



Social group	F1	Social group	F1
Country		Occupation	
India (IN)	0.2593	Religious	0.4191
Philippines (PH)	0.2365	Sales	0.2277
Indonesia (ID)	0.2030	Retired	0.1879
United States (US)	0.1910	Media	0.1761
Canada (CA)	0.1851	Business and financial	0.1465
Great Britain (GB)	0.1845	Healthcare	0.1393
France (FR)	0.1605	Legal	0.1364
Germany (DE)	0.1553	Student	0.1354
Malaysia (MY)	0.1148	Computer and mathematical	0.1227
Australia (AU)	0.0990	Arts and design	0.1177
		Education and library	0.1075
Gender		Management	0.0994
Male	0.6179	Science	0.0931
Female	0.5768	Food preparation	0.0672
		Architecture and engineering	0.0463

How you post is who you are: characterizing Google+ status updates across social groups

Concluding...

Groups hold linguistic particularities

- different groups → different structures/vocabulary
  - women → social and familial relations
  - men → technical topics and achievements
- tendency to keep a professional vocabulary

Open challenges

Other linguistic and social factors
Other popular OSNs
Cross data from different groups

Thank you!

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INCIPIT VITA NOVA
7 DE SETEMBRO DE 1927
Universidade Federal de Minas Gerais

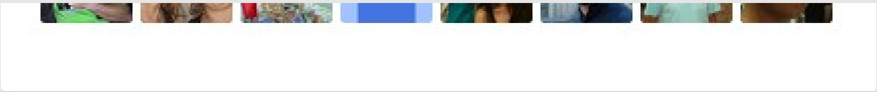
✉ evandrocunha@dcc.ufmg.br

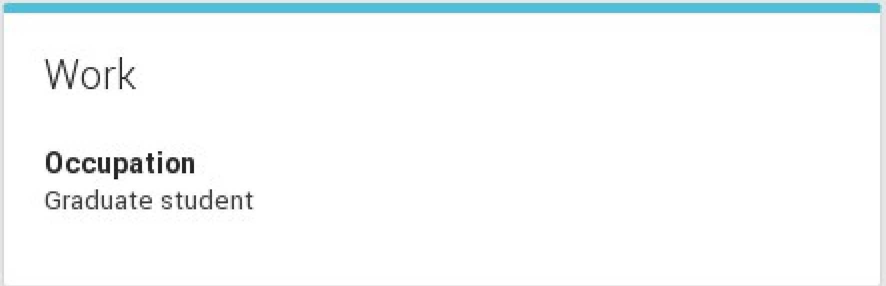
⌂ www.dcc.ufmg.br/~evandrocunha

@Cunha_et_al

Backup slides

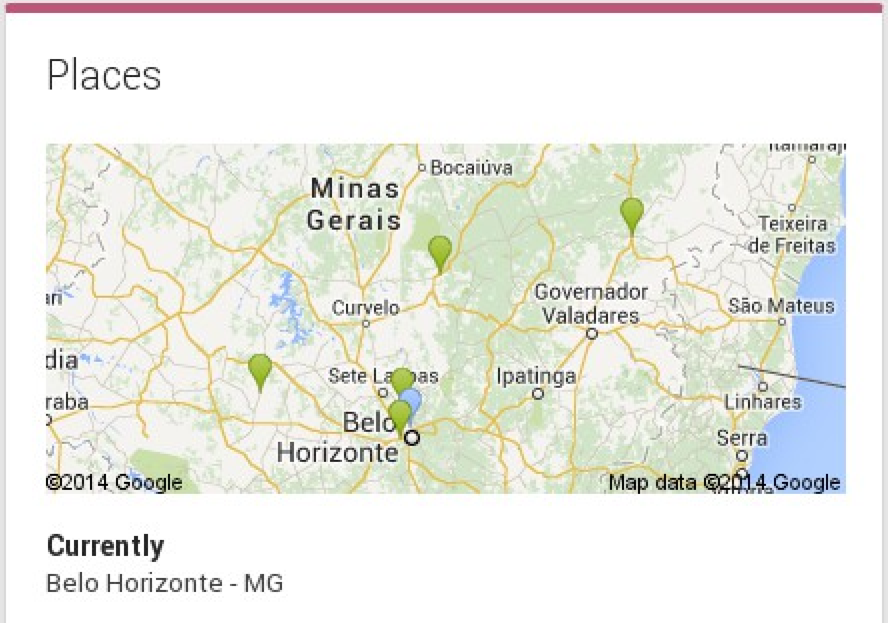
Information fields...





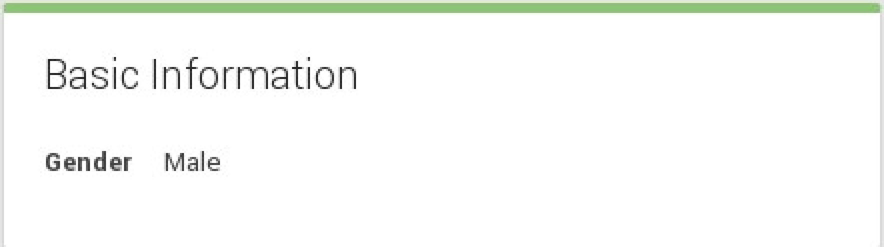
Work

Occupation
Graduate student



Places

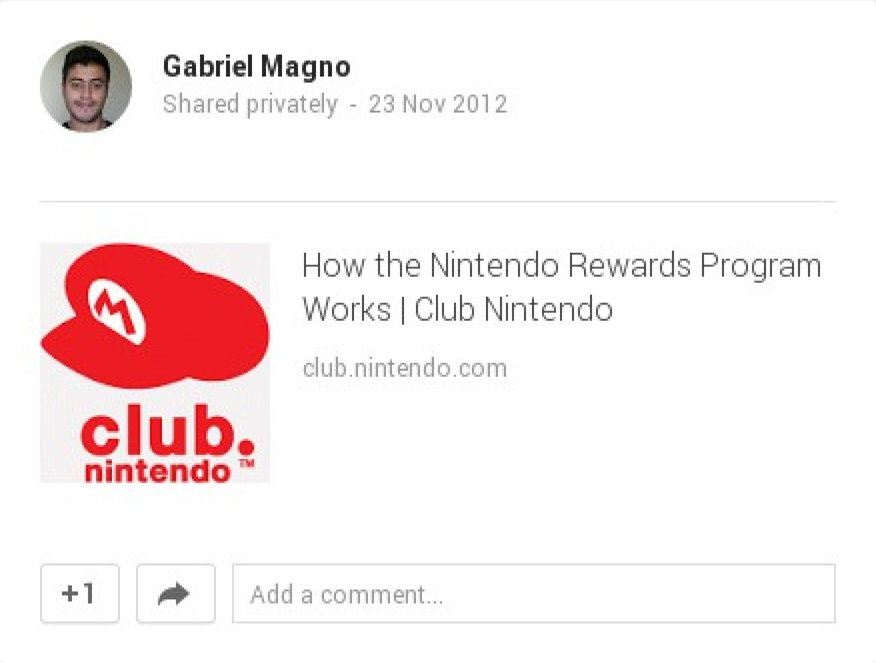
Currently
Belo Horizonte - MG



Basic Information

Gender Male

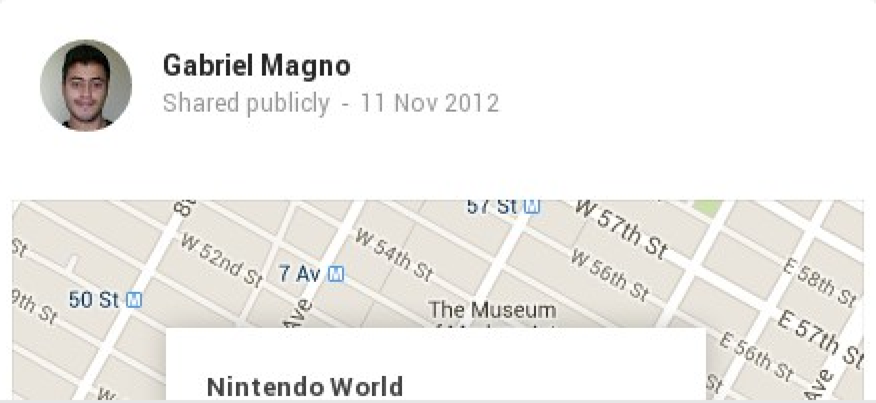
...and shared content



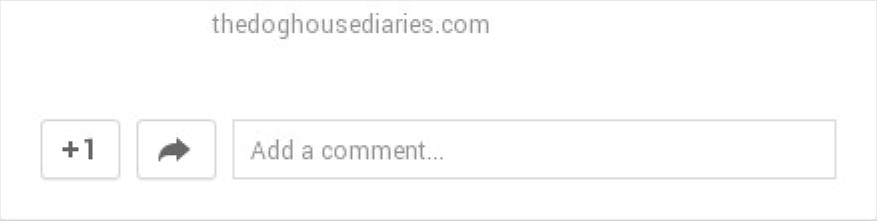
Gabriel Magno
Shared privately - 23 Nov 2012

How the Nintendo Rewards Program Works | Club Nintendo

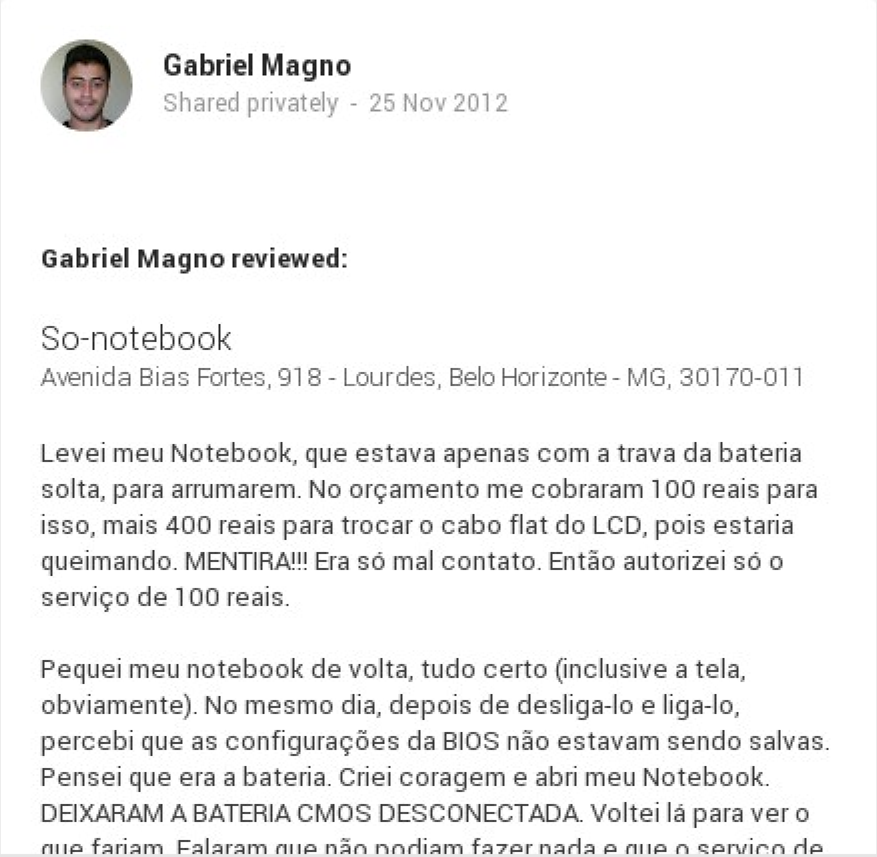
club.nintendo.com



Gabriel Magno
Shared publicly - 11 Nov 2012



thedoghousediaries.com



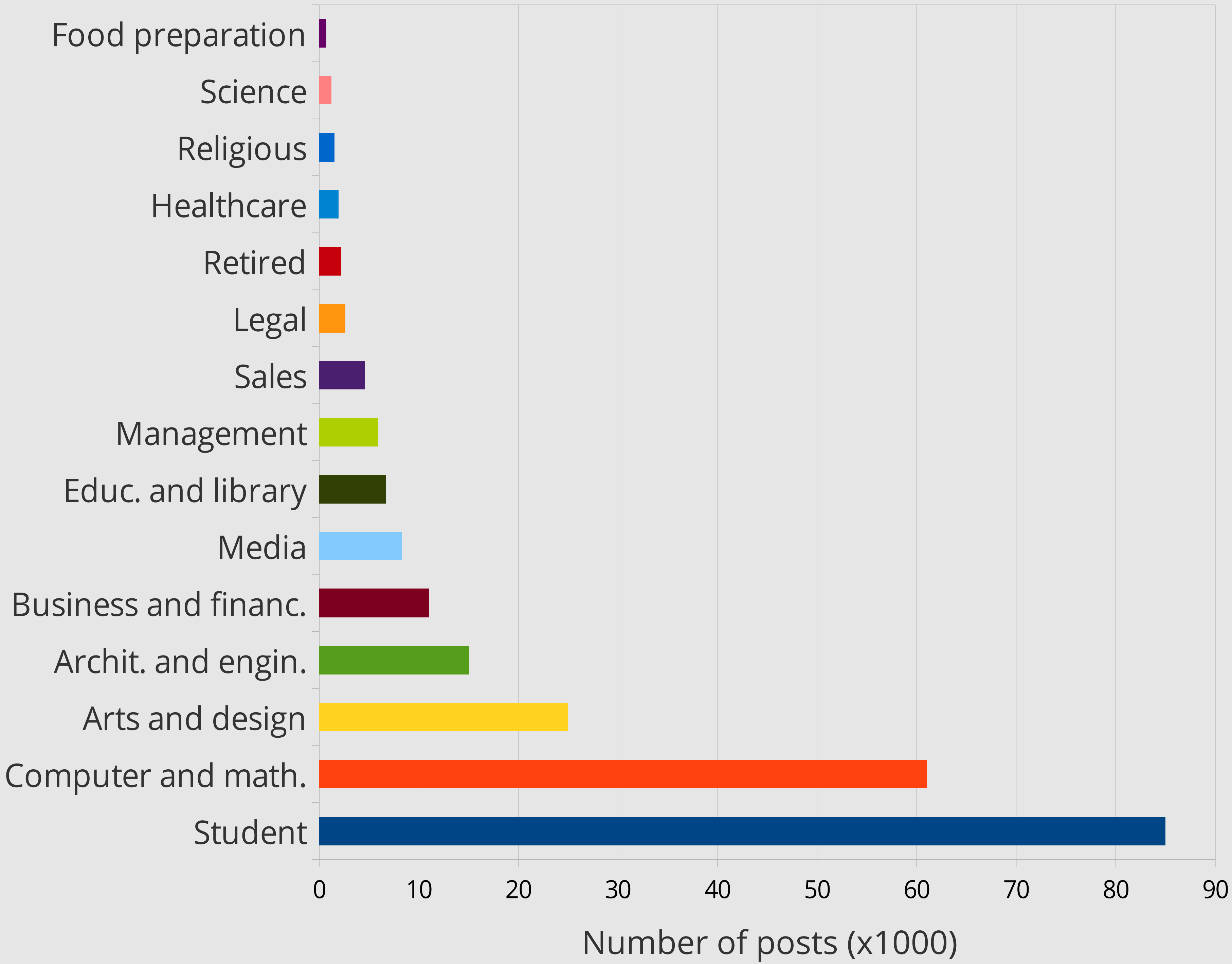
Gabriel Magno
Shared privately - 25 Nov 2012

Gabriel Magno reviewed:

So-notebook
Avenida Bias Fortes, 918 - Lourdes, Belo Horizonte - MG, 30170-011

Levei meu Notebook, que estava apenas com a trava da bateria solta, para arrumarem. No orçamento me cobraram 100 reais para isso, mais 400 reais para trocar o cabo flat do LCD, pois estaria queimando. MENTIRA!!! Era só mal contato. Então autorizei só o serviço de 100 reais.

Pequei meu notebook de volta, tudo certo (inclusive a tela, obviamente). No mesmo dia, depois de desliga-lo e liga-lo, percebi que as configurações da BIOS não estavam sendo salvas. Pensei que era a bateria. Criei coragem e abri meu Notebook. DEIXARAM A BATERIA CMOS DESCONECTADA. Voltei lá para ver o que fariam. Falaram que não podiam fazer nada e que o serviço de


Food preparation
Science
Religious
Healthcare
Retired
Legal
Sales
Management
Educ. and library
Media
Business and financ.
Archit. and engin.
Arts and design
Computer and math.
Student
0
10
20
30
40
50
60
70
80
90
Number of posts (x1000)

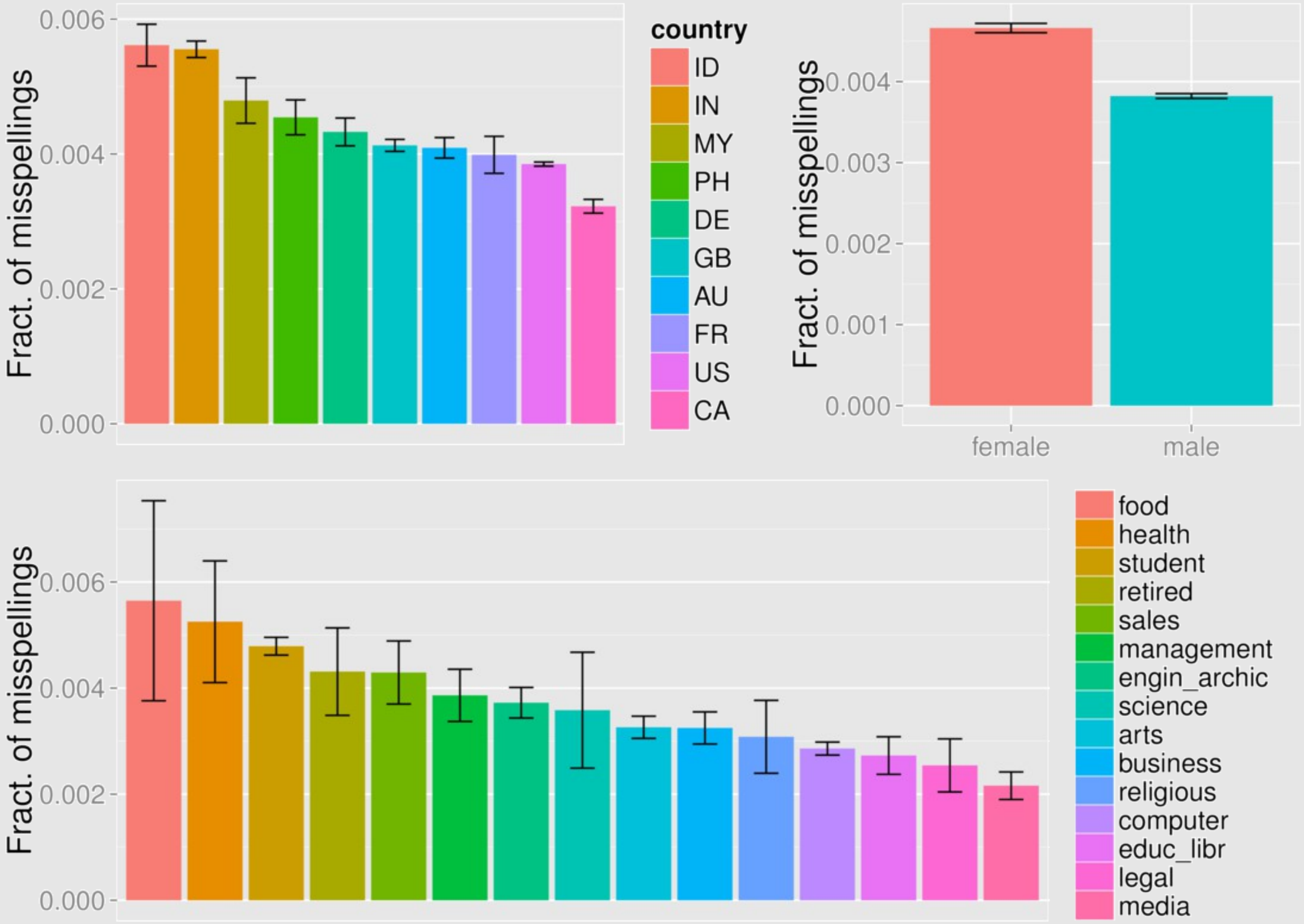
Fraction of misspellings

Wikipedia's list of common misspellings

Gonna, doin', ur = not misspellings

Acident, adn, populer = misspellings

Fraction of misspellings



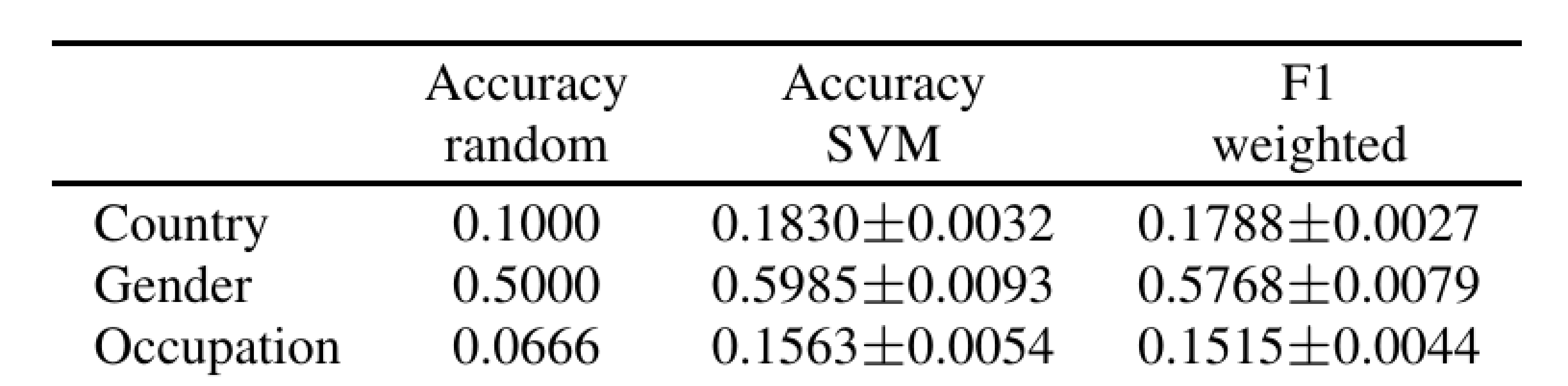
How you post is who you are: characterizing Google+ status updates across social groups

Fraction of misspellings

Non native English speakers → + misspellings

Workers who deal with texts → - misspellings

An (preliminary) application



	Accuracy random	Accuracy SVM	F1 weighted
Country	0.1000	0.1830±0.0032	0.1788±0.0027
Gender	0.5000	0.5985±0.0093	0.5768±0.0079
Occupation	0.0666	0.1563±0.0054	0.1515±0.0044