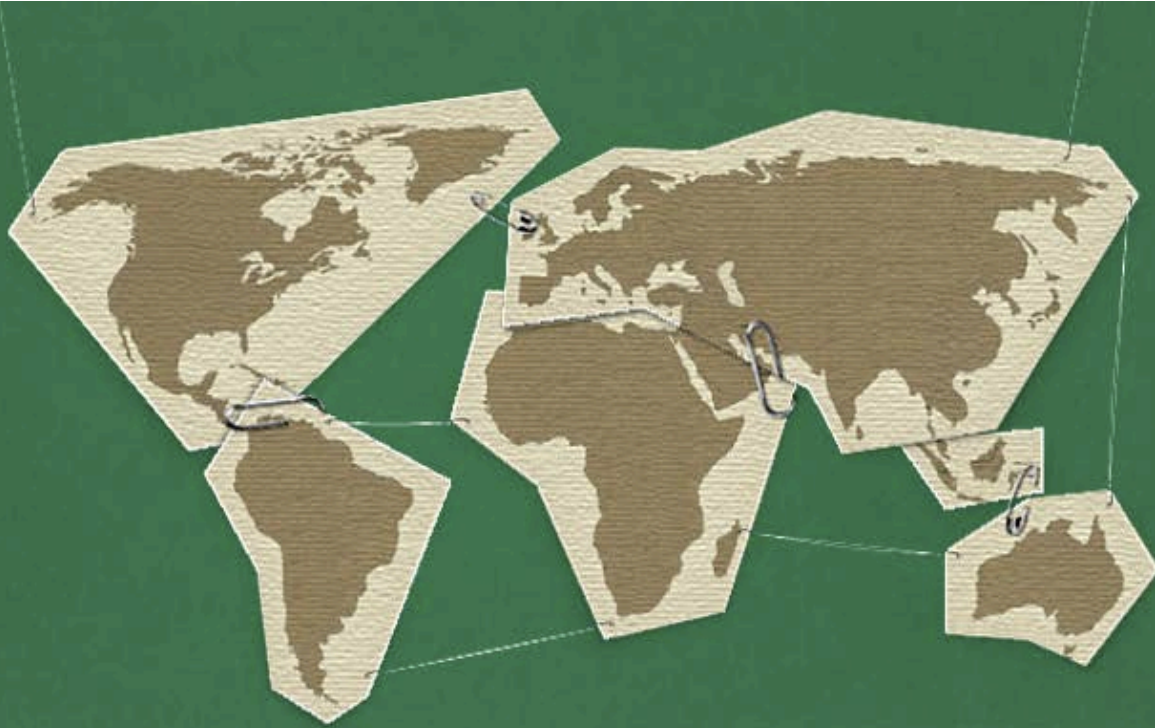


Bringing Data Science to the Speakers of Every Language



LANGUAGE
TECHNOLOGIES FOR A
CONNECTED
WORLD

Robert Munro, PhD
CEO, Idibon

About me: technology and global development

CEO, Idibon

Global text analytics in
50+ languages

Working with leaders in
industry & social good



MIT Humanitarian
Response Lab



Industry: CTO / CIO

Energy infrastructure in Liberia and
Sierra Leone

Global epidemic tracking

Crowdsourcing and natural language
processing for disaster response



Other

Ph.D. in NLP from Stanford
Bicycled 20+ countries



idibon

Recommendations for language processing for social good

Look beyond English

Inherent benefit understanding and support speakers of every language

Employ people in those languages

Crowdsourced workers speak 100s of languages, and want to use them

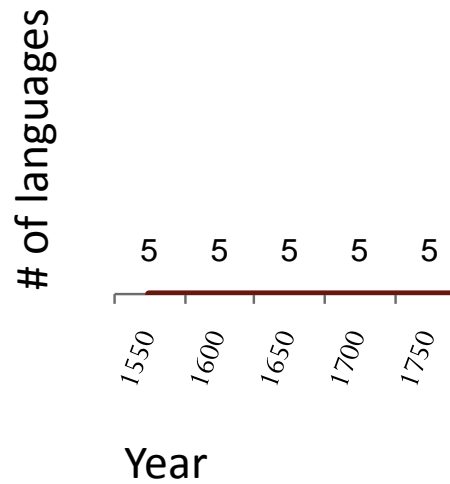
Embrace the variation

You can't rely on consistent spellings, but you can learn to model the diversity



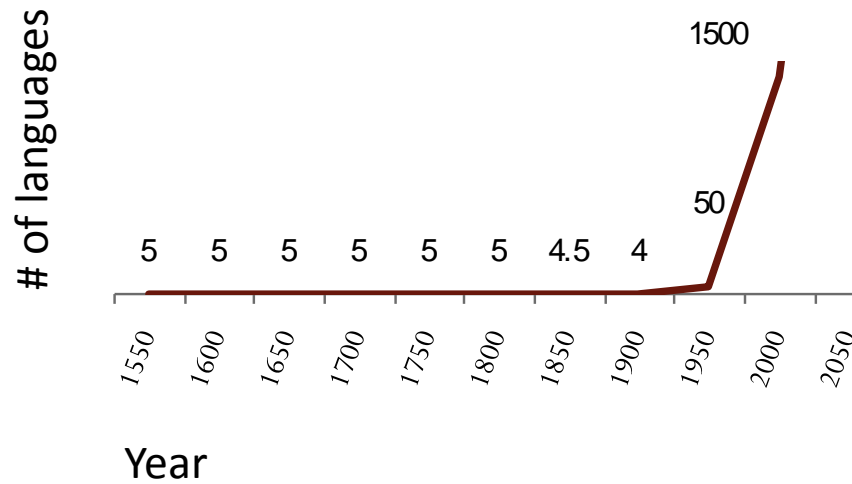
How many languages are in the connected world?

How has this changed?

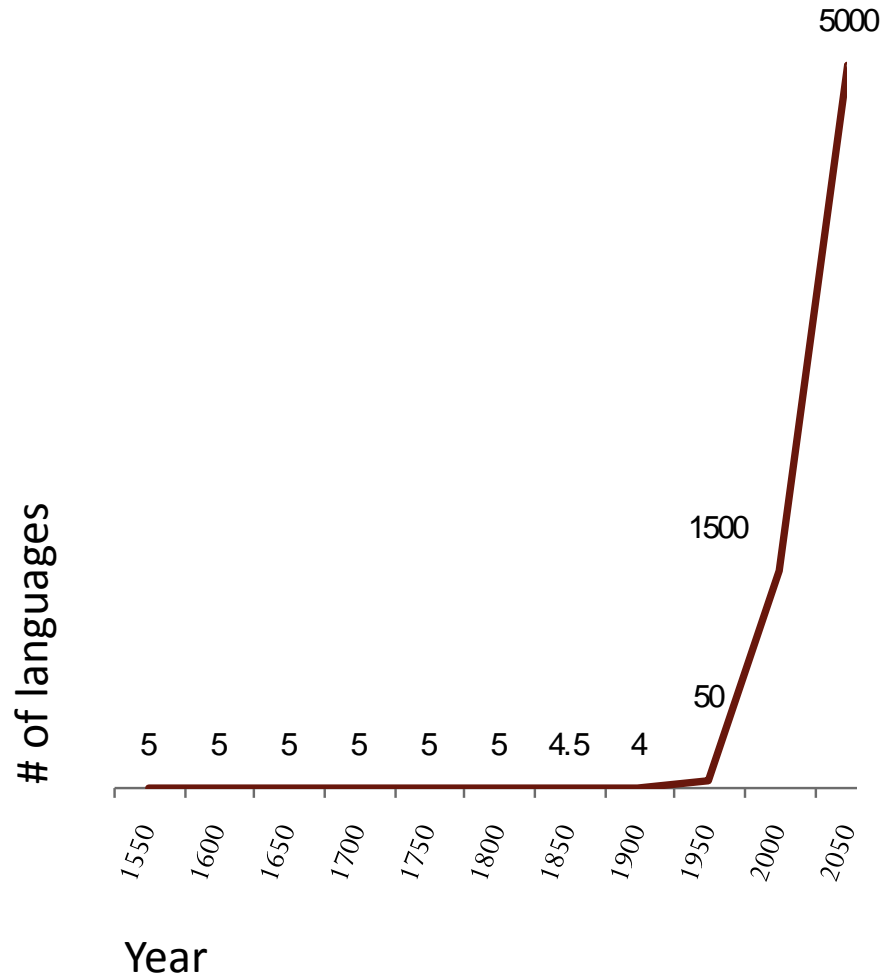


How many languages are in the connected world?

How has this changed?



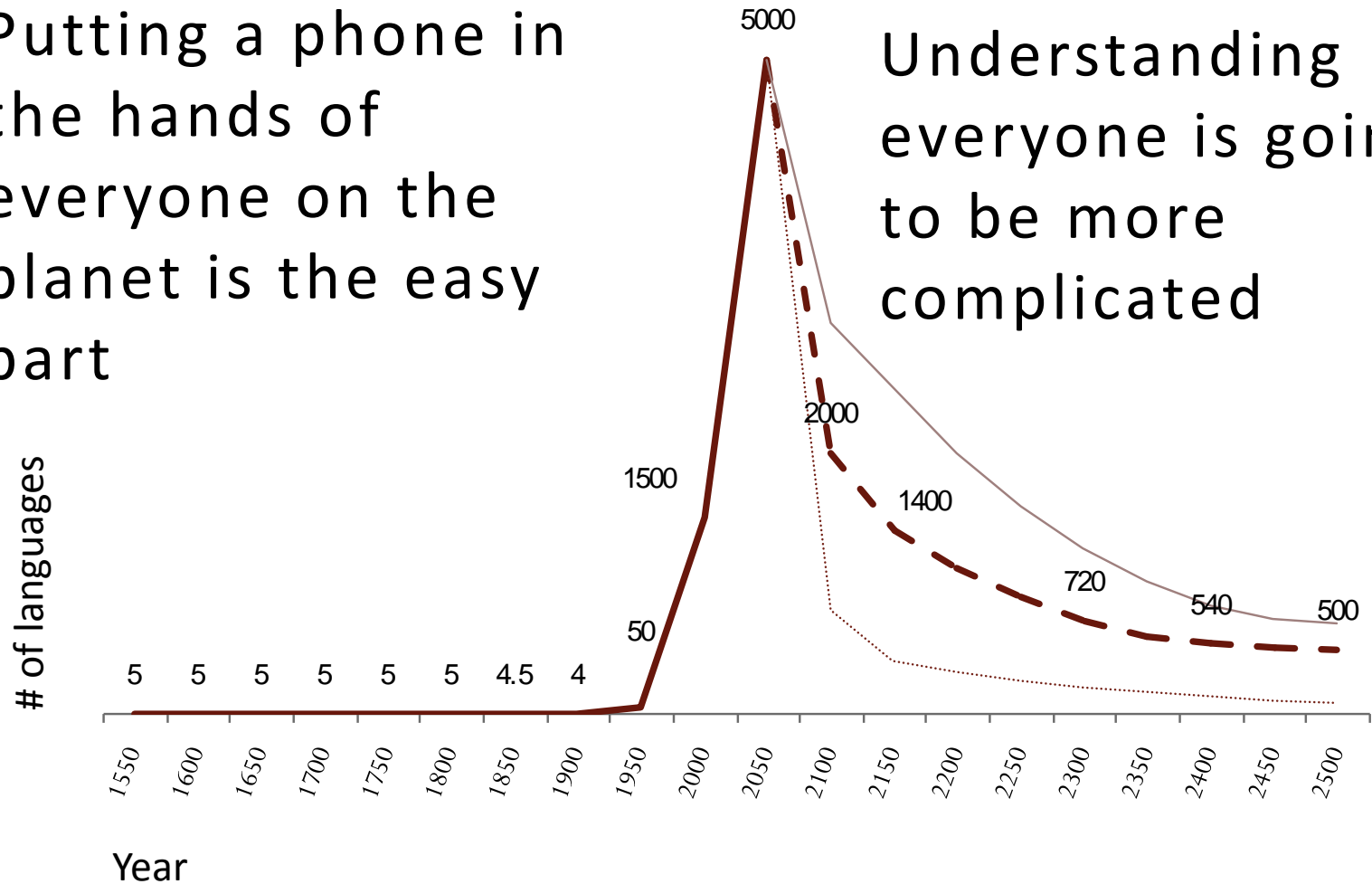
How many languages are in the connected world?



How many languages are in the connected world?

Putting a phone in the hands of everyone on the planet is the easy part

Understanding everyone is going to be more complicated



Every human communication this year



Source: Ethnologue, Nationalencyklopedin

7% of our communications are digital, most is still direct spoken language



If every online picture is worth a thousand words, it would double social media.

Every picture →



idibon

Every 3 months, the world's text messages exceed the word count of every book.

Every book. Ever.



Source: Google Books

Print communication is smaller than anything shown.

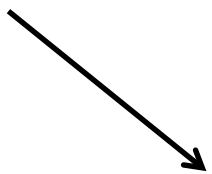


Print communication is smaller than anything shown.
Ditto any one social network.





**There are more than 6,000 other languages.
Only the top 1% are shown.**



No language from the Americas made the cut.



Quechua



español	官话	
हन्दी		
English		
العَرَبِيَّةُ	русский язык	বাংলা
português	日本語	اُردُو
		Deutsch
		ਪੰਜਾਬੀ
吴语	தமிழ்	한국어
		తెలుగు
Bahasa Indonesia / Melayu	Türkçe	தமிழ் (தமிழ்)
		français
मराठी	閩南語	فارسی
		język polski
		粵語
		tiếng Việt
italiano	پښتو	ඊලිට්ටු
		هۆس
		湘語
		українська мова
		سرائیکی
		晉語
ಕನ್ನಡ	客家話	सिंधي
		فارسی
		Fulani
		தமிழ்
		Cebuano
		অসমীয়া
		Oromo
		Tagalog
èdè Yorùbá	Akan	Af-Soomaalí
		Mãiwêri
		Nederlands
		ភាសាខ្មែរ
		Asuu Igbo
		română
മലയാളം	தமிழ்	湘語
		ελληνικά
		Štokavian
తెలుగు	यऱ्बेक तिलि	भोजपुरी
		Kreyòl ayisyen
		ಕನ್ನಡ
		Kurdi
Malagasy	magyar nyelv	Chizewa
		සිංහල
		Madhura
		অসমীয়া
		sms
		email
		social networks
		phone calls
		instant messaging
		W
		W
		W
other languages		

Email spam would be larger than every block except spoken Mandarin (官话).



Source: Mashable

Short messages (SMS and IM) make up 2% of the world's communications.

The largest and most linguistically diverse form of written communication that has ever existed.

PhDs focused on processing large volumes of short messages in low resource languages?

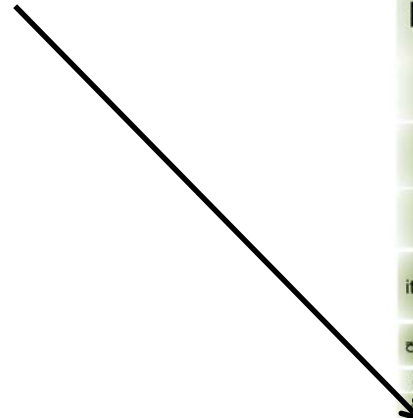
1



If the Facebook “like” is a one-word language it is in the top 5% of languages by word count.



Your browser probably won't show Sundanese script



Sundanese speakers outnumber the populations New York, London, Tokyo and Moscow.

Combined.



You misread "Sundanese" as "Sudanese" which is a variety of Arabic



We have a blind spot for knowing about the existence of languages.



This is the breakdown of languages that most of our data is moving towards



An earthquake struck Haiti on January 12, 2010

Most local services failed, but most cell-towers remained functional.





Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31



Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31
- My family in Carrefour, 24 Cote Plage, 41A needs food and water
- People trapped in Sacred Heart Church, PauP
- General Hospital has less than 24 hrs. supplies
- Undergoing children delivery Delmas 31

Mission 4636



“Fanm gen tranche pou fè yon pitit nan Delmas 31”



Message translated, categorized & geolocated

“Fanm gen tranche pou fè yon pitit nan Delmas 31”
Undergoing children delivery
Delmas 31
18.495746829274168,
72.31849193572998
Emergency

Location is refined & actionable items are identified

Global collaboration

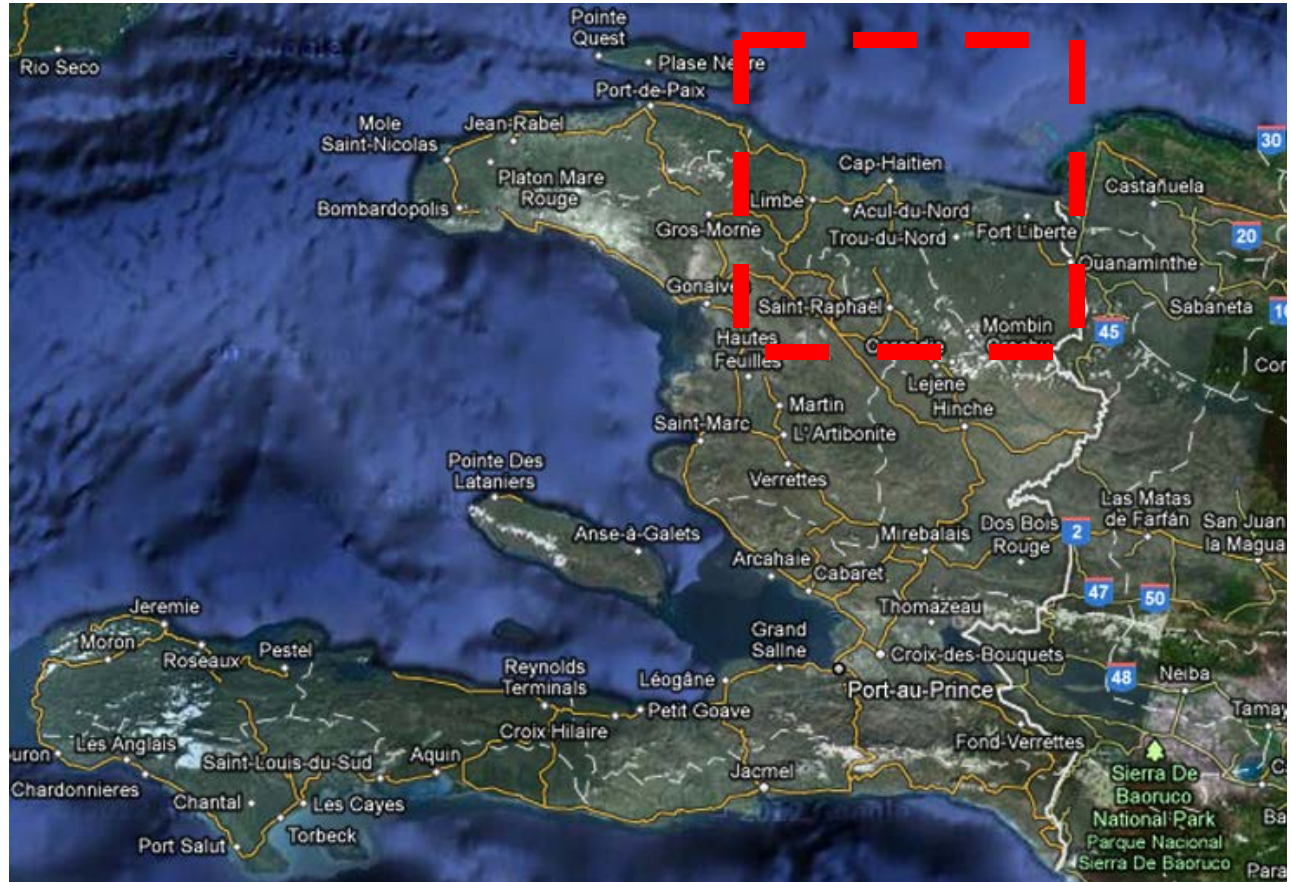
2,000 volunteers, transferred to paid workers in Haiti



idibon

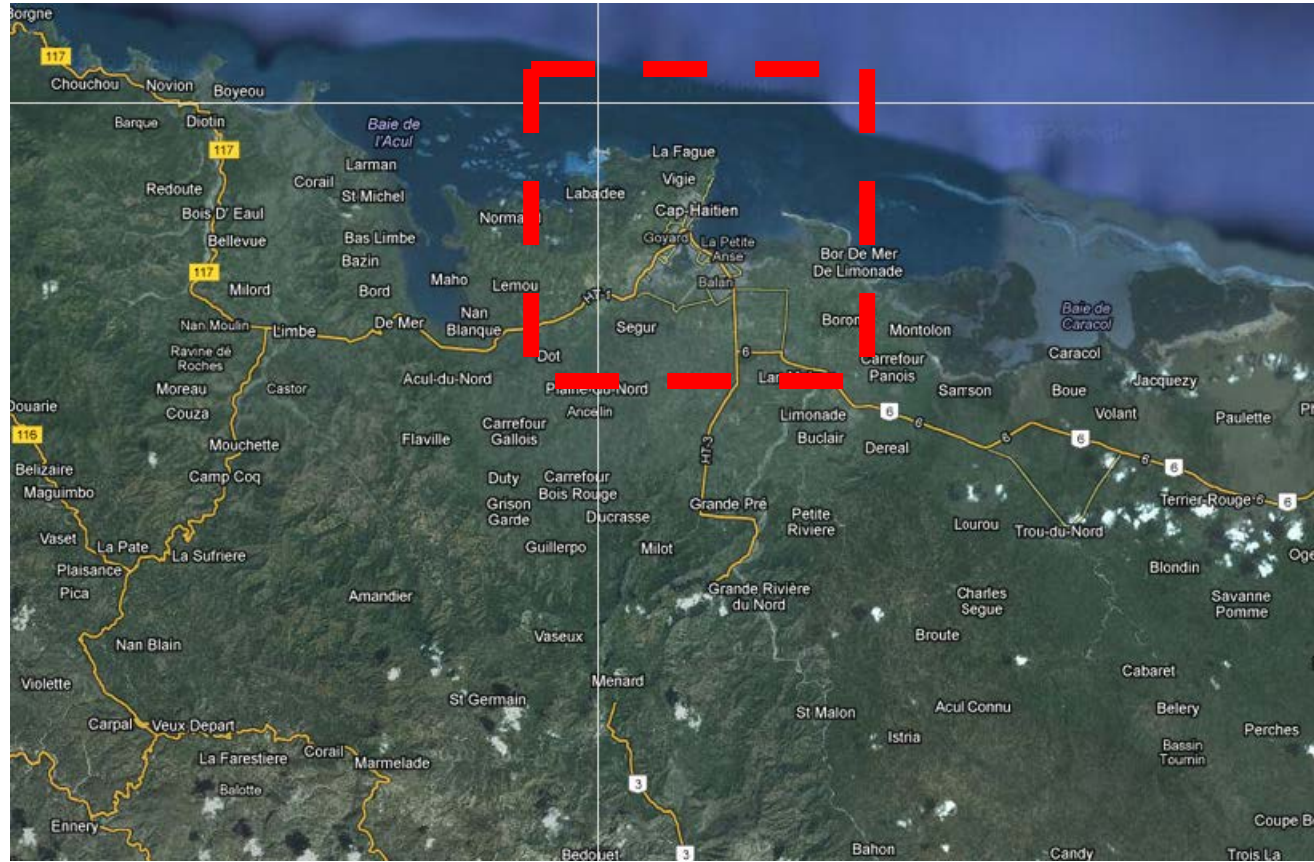
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of **Okap** is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”



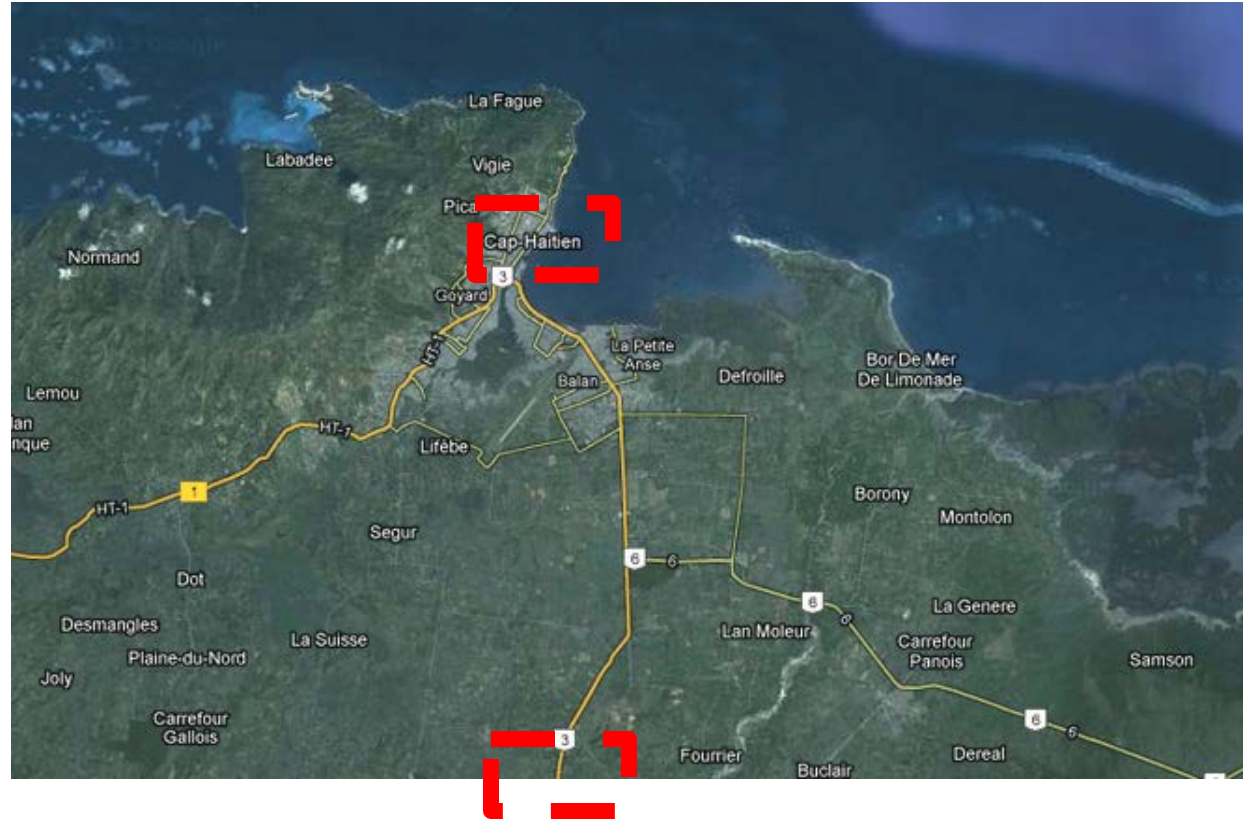
Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of **Okap** is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”



Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

“Sacre-Coeur Hospital which located in this village of **Okap** is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital.”



Local knowledge

Workers collaborating to find locations:

Dalila: I need Thomassin Apo please

Apo: Kenscoff Route: Lat: 18.495746829274168, Long:-72.31849193572998

Apo: This Area after Petion-Ville and Pelerin 5 **is not on Google Map. We have no streets name**

Feedback from responders:

"just got emergency SMS, child delivery, USCG are acting, and, the GPS coordinates of the location we got from someone of your team were **100% accurate!**"

The ability for someone to make a real-time difference at any other place in the world:

Apo: I know this place like my pocket

Dalila: thank God u was **here**

'here' = anywhere





**How do we automate processing the
world's data?**

Generations of standardization in spelling and simple morphology

Whole words suitable as features for NLP systems

Most other languages

Relatively complex morphology

Less (observed) standardized spellings

More dialectal variation

No standard (wide-spread) spellings

- More or less French spellings

- More or less phonetic spellings

Frequent words (esp pronouns) are shortened and compounded

- Regional slang / abbreviations

The extent of the subword variation

>30 spellings of *odwala* ('patient') in Chichewa

>50% variants of 'odwala' occur only once in the data used here:

Affixes and incorporation

'kwaodwala' -> 'kwa + odwala'

'ndiodwala' -> 'ndi odwala' (official 'ngodwala' not present)

Phonological/Orthographic

'odwara' -> 'odwala'

'ndiwodwala' -> 'ndi (w) odwala'

Modeling the variation gives accurate results

ndimmafuna manthwala
(‘I currently need medicine’)



ndimafuna mantwala



ndi-ma-fun-a man-twala



ndi-ma-fun-a man-twala



ndi -fun man-twala
(“I need medicine”)

Category = “Request for aid”

ndi kufuni mantwara
(‘my want of medicine’)



ndi kufuni mantwala



ndi-ku-fun-i man-twala



ndi-ku-fun-i man-twala



ndi -fun man-twala
(“I need medicine”)

Category = “Request for aid”

1 in 5 classification errors with raw messages

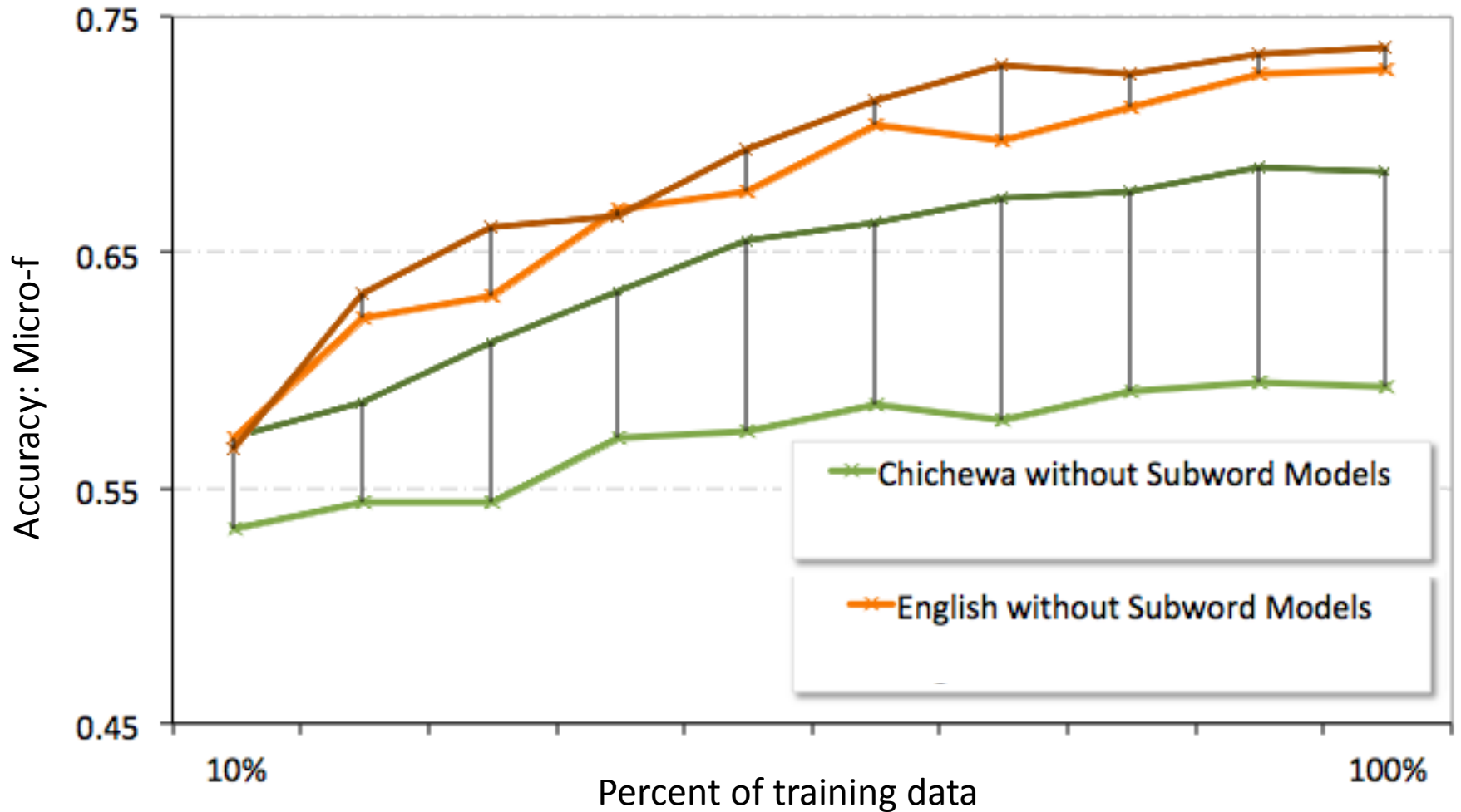
1) Normalize spellings

2) Segment

3) Identify predictors

1 in 20 classification error post-processing.
Improves with scale.

Comparison with English

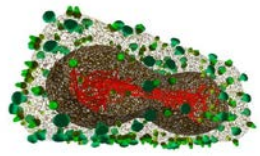




Taking it to the world

The benefits of understanding everyone

Human diseases eradicated in the last 75 years:



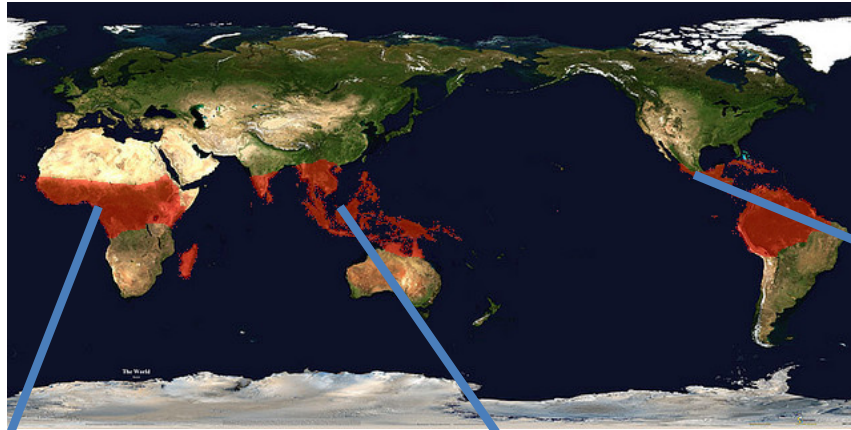
smallpox

Increase in air travel in the last 75 years:



idibon

Reports of 'strange new illnesses' pre-date official records

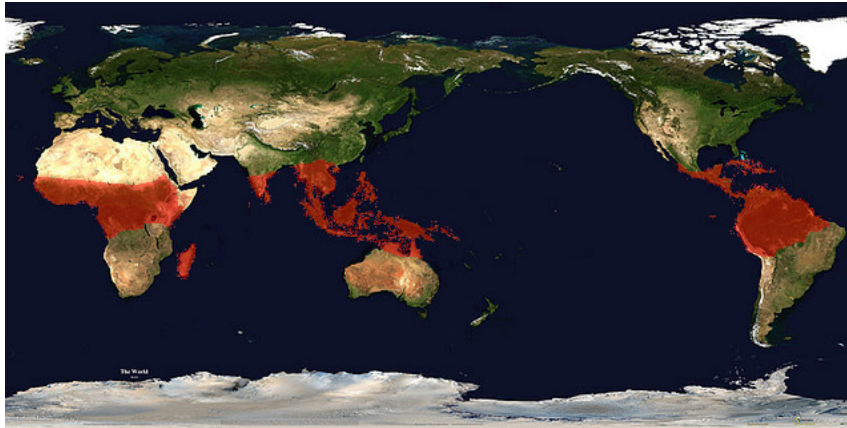


HIV
decades
(35 million
infected)

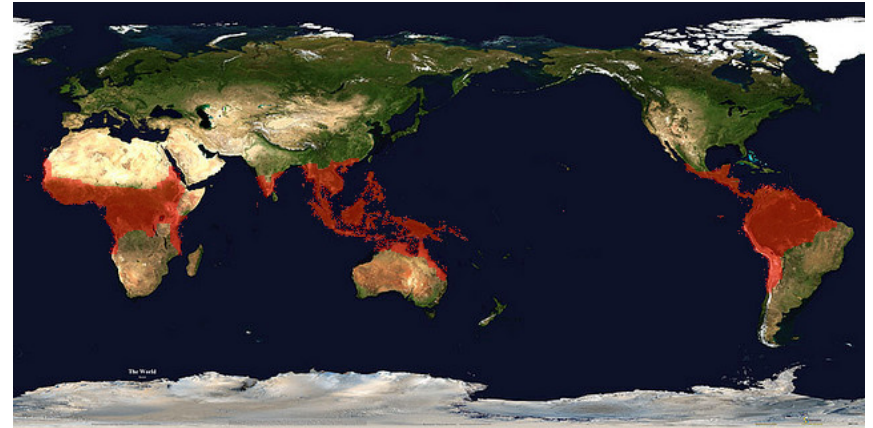
H1N5 (Bird Flu)
weeks
(>50% fatal)

H1N1 (Swine Flu)
months
(10% of world
infected)

...but the reports are in 1000s of languages



90% of ecological diversity



90% of linguistic diversity

в предстоящий осенне-зимний период в Украине
ожидаются две эпидемии гриппа

مزید من انفلونزا الطيور في مصر

香港现1例H5N1禽流感病例曾游上海南京等地

Crowdsourcing, big data, and expert analysts

Most information is in plain language:
Multiple skill and processing strategies required.





в предстоящий осенне-зимний период в Украине
ожидаются две эпидемии гриппа


مزید من انفلونزا الطيور في مصر

香港现1例H5N1禽流感病例曾游上海南京等地

Digital Disease Discovery


Reports
millions per day:
many languages,
much noise


Big Data
machine learning:
extraction, filtering
& prioritization


Crowdsourcing
thousands of
native-language
speakers


Analysts
several
domain
experts

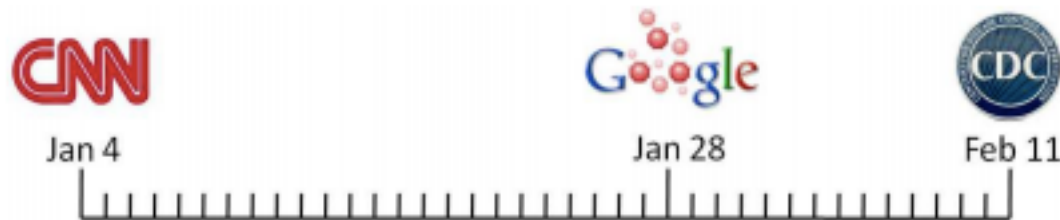

Global
monitoring
Safer world

The impact of scalable monitoring

Found historical signals that pre-dated *Google Flu Trends* by 3 weeks, CDC by 5 ... on CNN

How can we filter/model media-driven amplification?

“I’m Jacqui Jeras with today’s cold and flu report ...
across the mid-Atlantic states, a little bit of an increase”
January 4, 2008 CNN Weather





The impact of scalable monitoring

Tracked Ebola in Uganda 5 days before World Health Organization.

“we were able to pull in much richer data from a larger number of sources, so we knew not just how many people were infected, but what kind of transport they took when they went from their village to the hospital in the nearest main town.”

Robert Munro, UN General Assembly on “Big Data and Global Development”

Age + gender + village + went to hospital.

This is personally identifying & could lead to persecution.

Open data would have a negative impact.

The impact of scalable monitoring

Tracked E-Coli Outbreak in Germany 2 days before ECDC.

How do we motivate information processing?



Margins are small: only for-profit big data and crowdsourcing can have a sustained impact

Idibon's current work

Hurricane Sandy

Idibon's CTO ran FEMA's Aerial Damage Assessments.

We have >1,000,000 manual tags on communications.



MIT Humanitarian Response Lab

Identifying reports about supply-line interruptions.

Research data from a combination of crowdsourcing and natural language processing

Recommendations for language processing for social good

Look beyond English

Inherent benefit understanding and support speakers of every language

Employ people in those languages

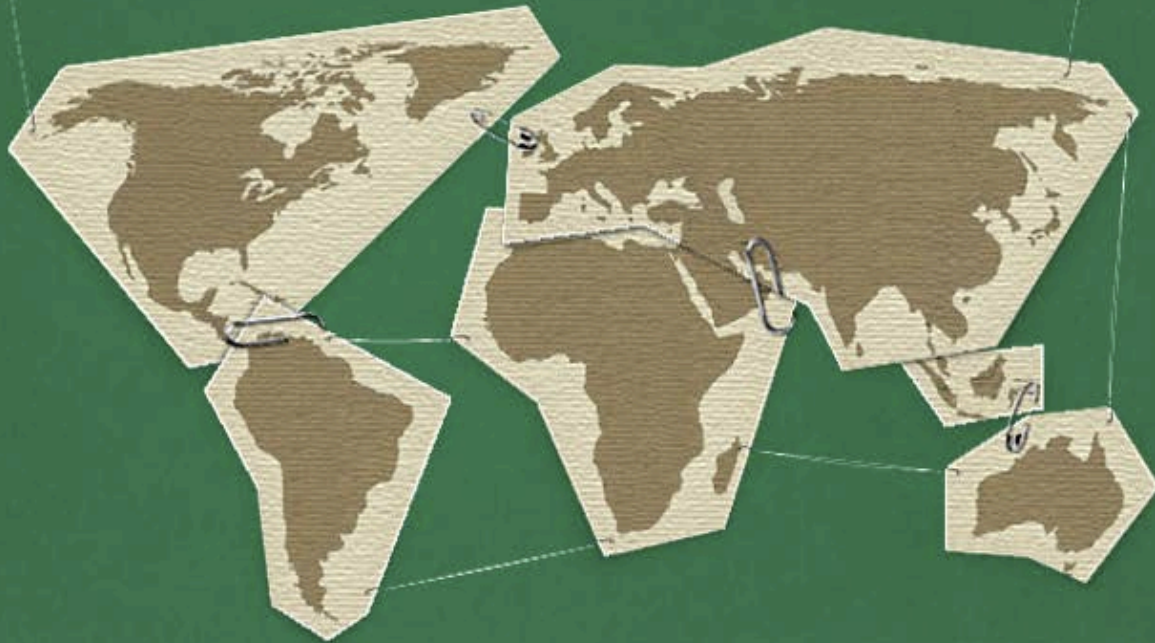
Crowdsourced workers speak 100s of languages, and want to use them

Embrace the variation

You can't rely on consistent spellings, but you can learn to model the diversity



Thank you



LANGUAGE
TECHNOLOGIES FOR A
CONNECTED
WORLD

Robert Munro, PhD
CEO, Idibon