From Big Data to Small Languages: communication and information processing in a newly connected world

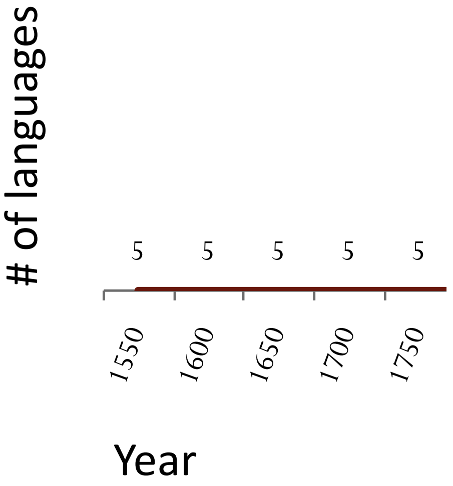
Robert Munro

2012

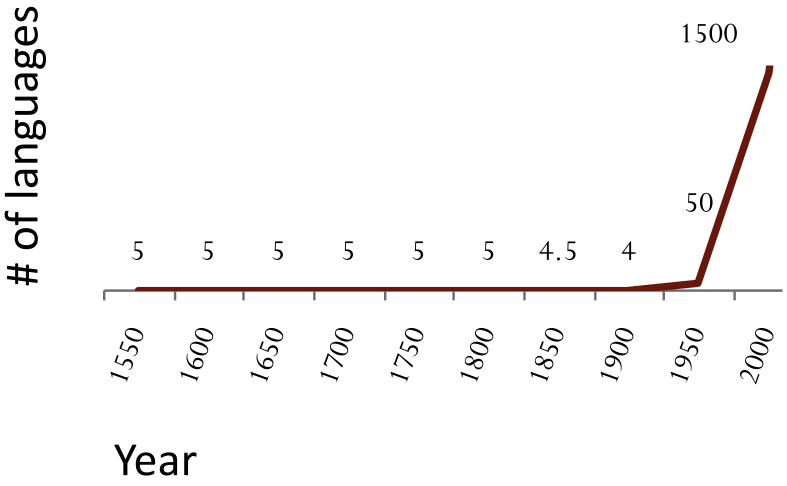
Daily potential language exposure

- On a given day, what is the average number of languages that someone could potentially hear?
- How has this changed?

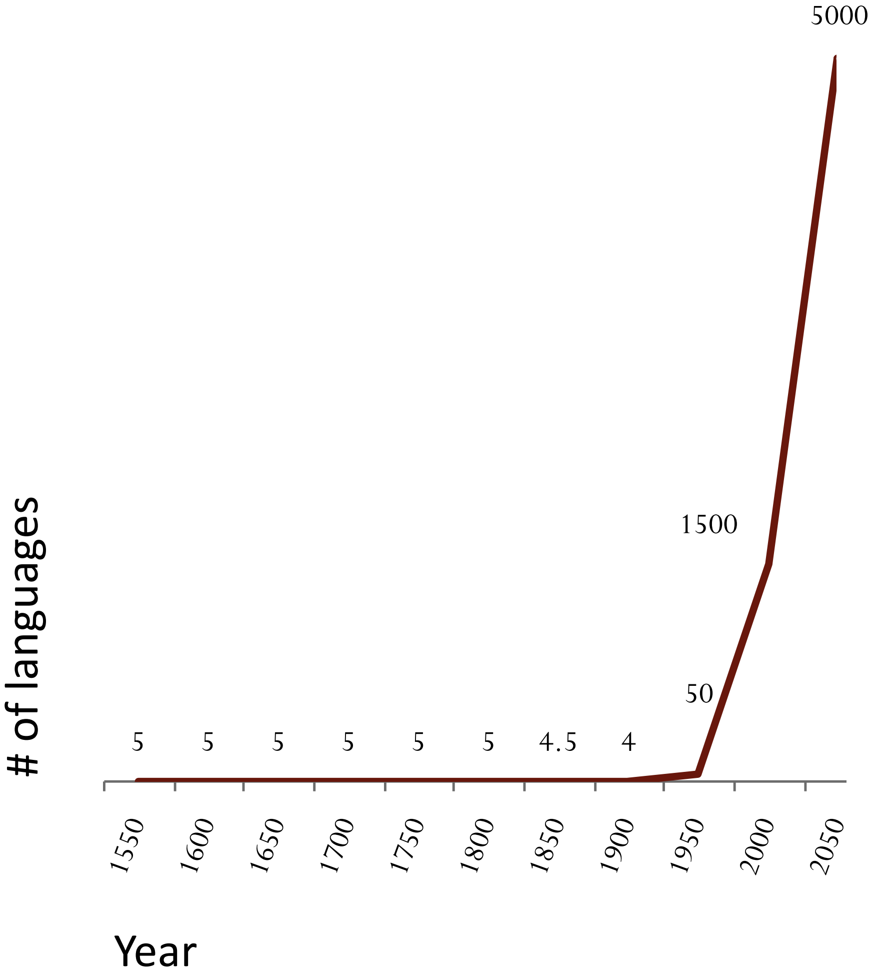
Daily potential language exposure



Daily potential language exposure

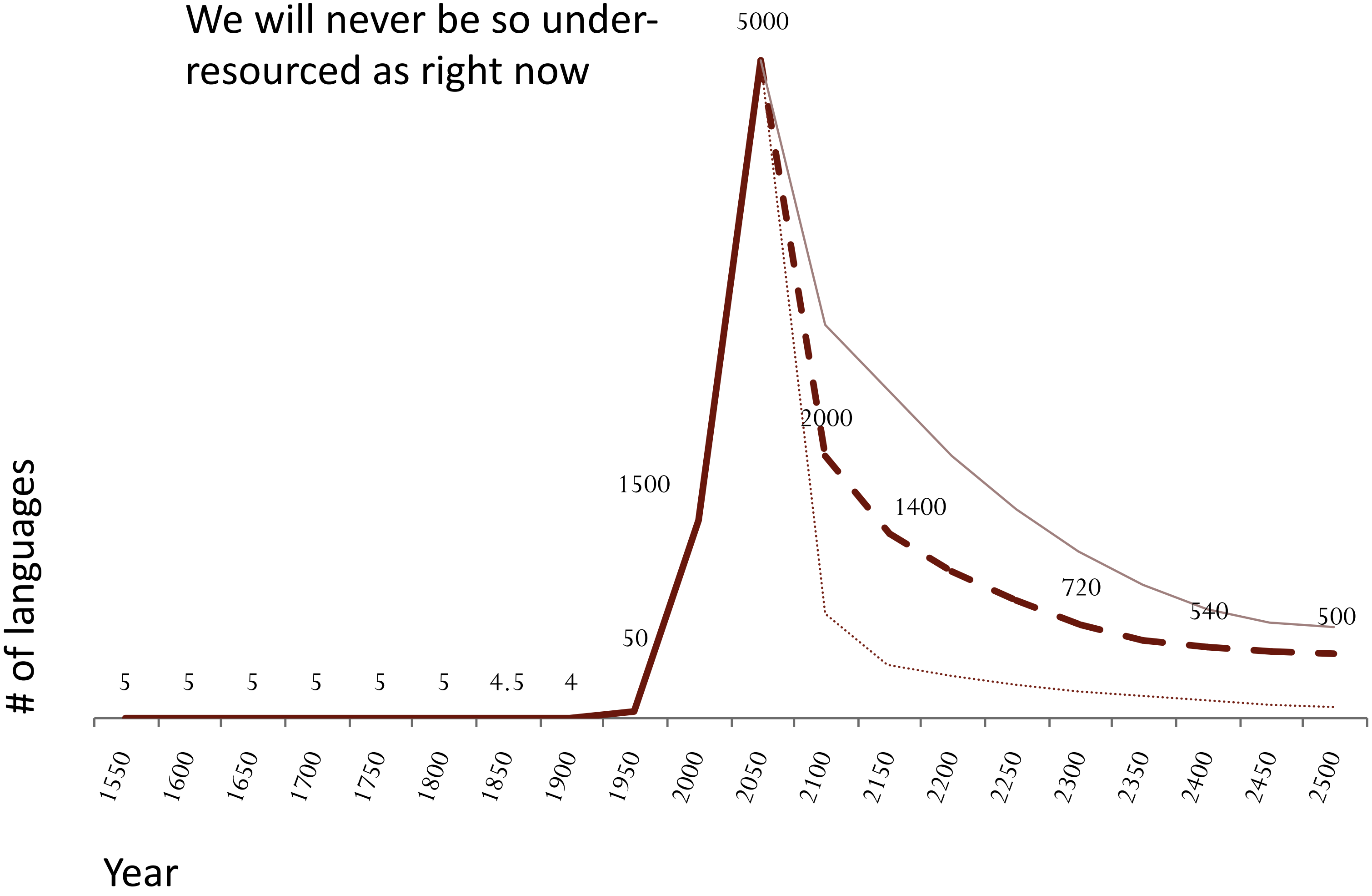


Daily potential language exposure

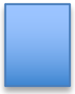


Daily potential language exposure

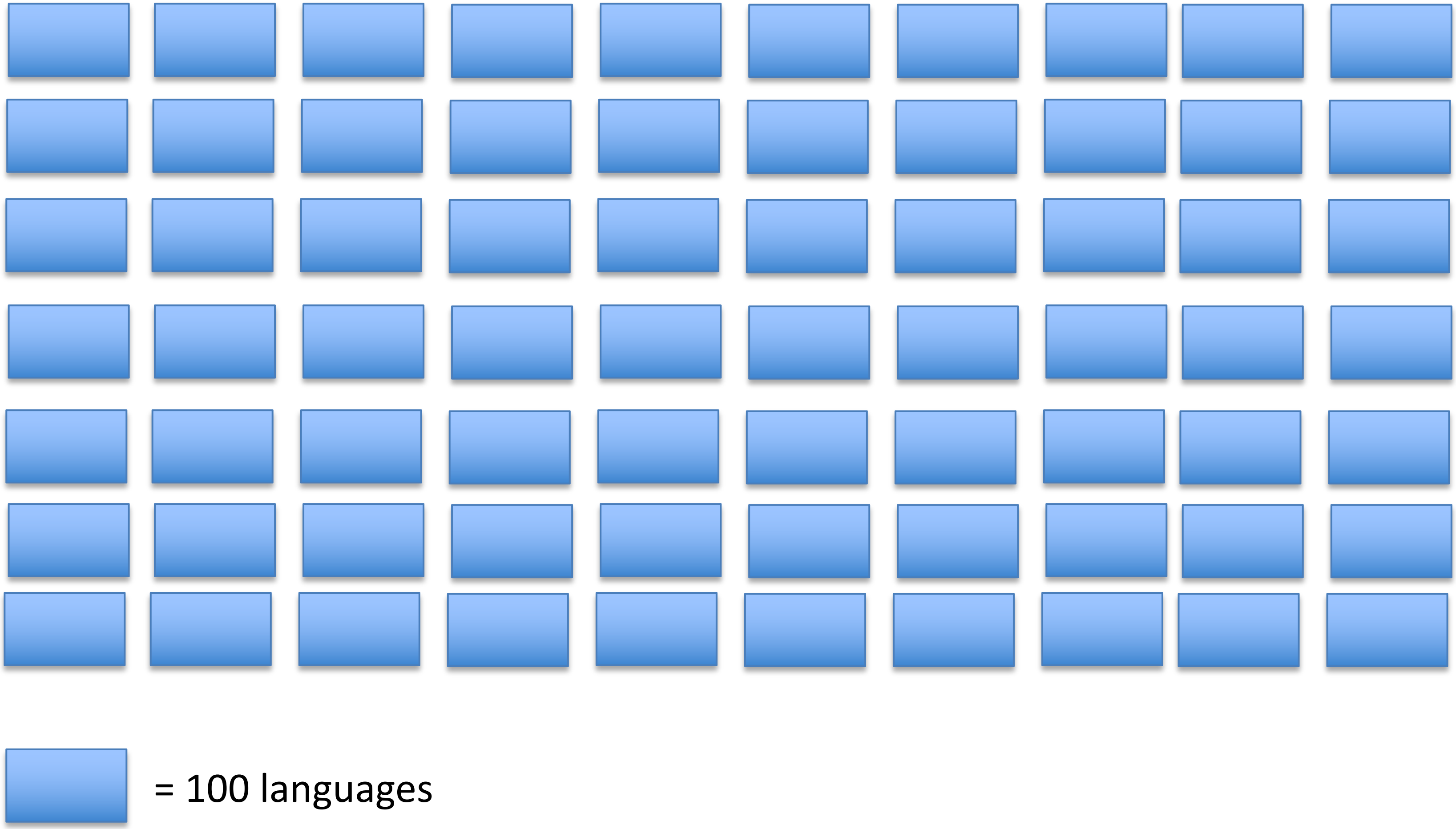
We will never be so under-resourced as right now



Languages shared by the world’s doctors

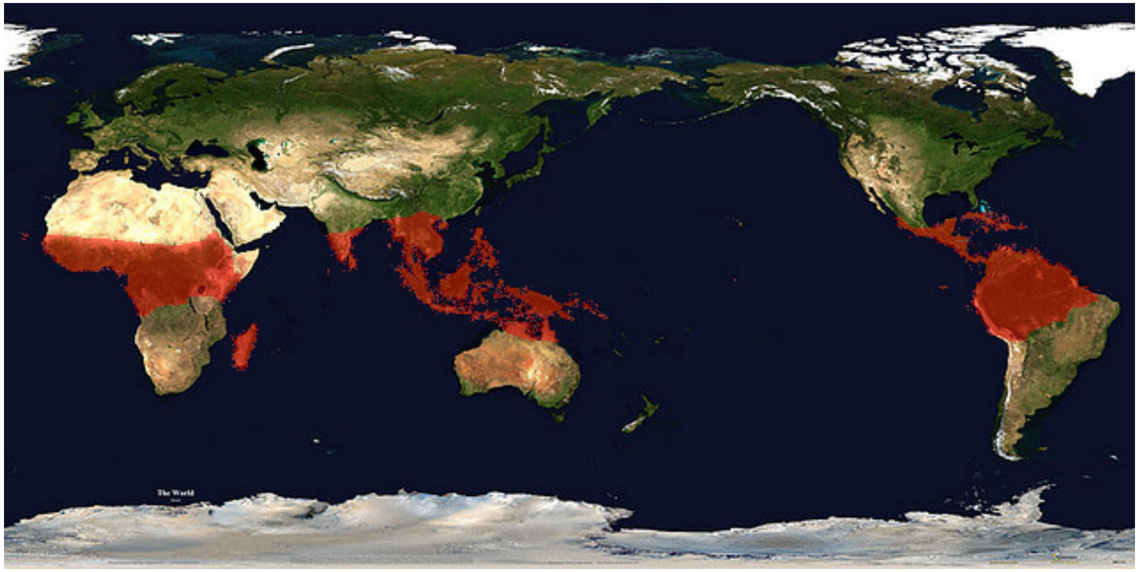


Languages spoken by the world’s patients

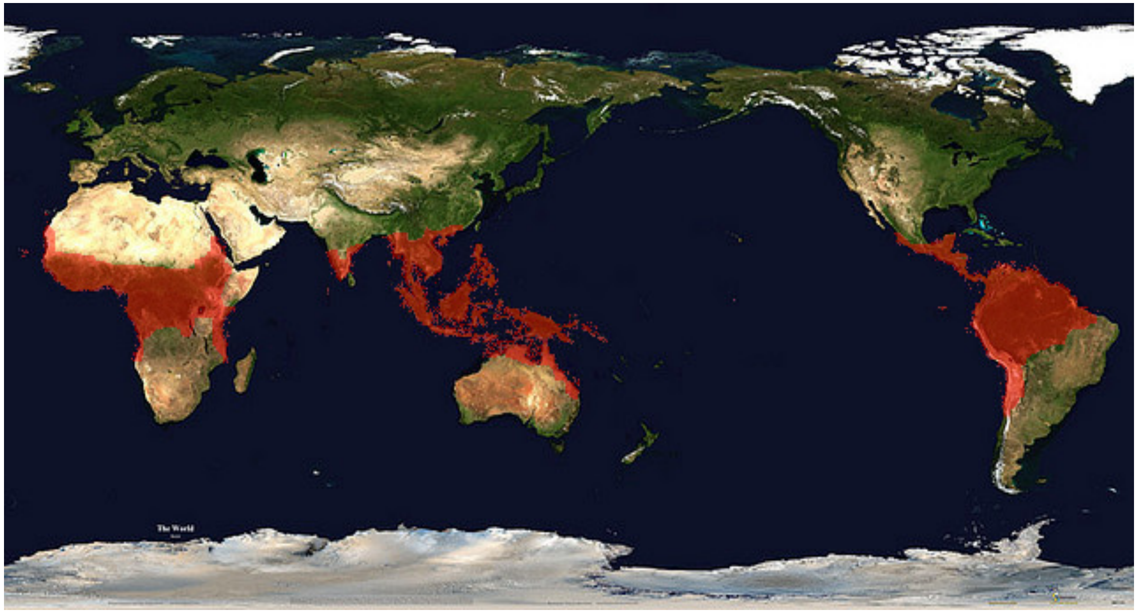


= 100 languages

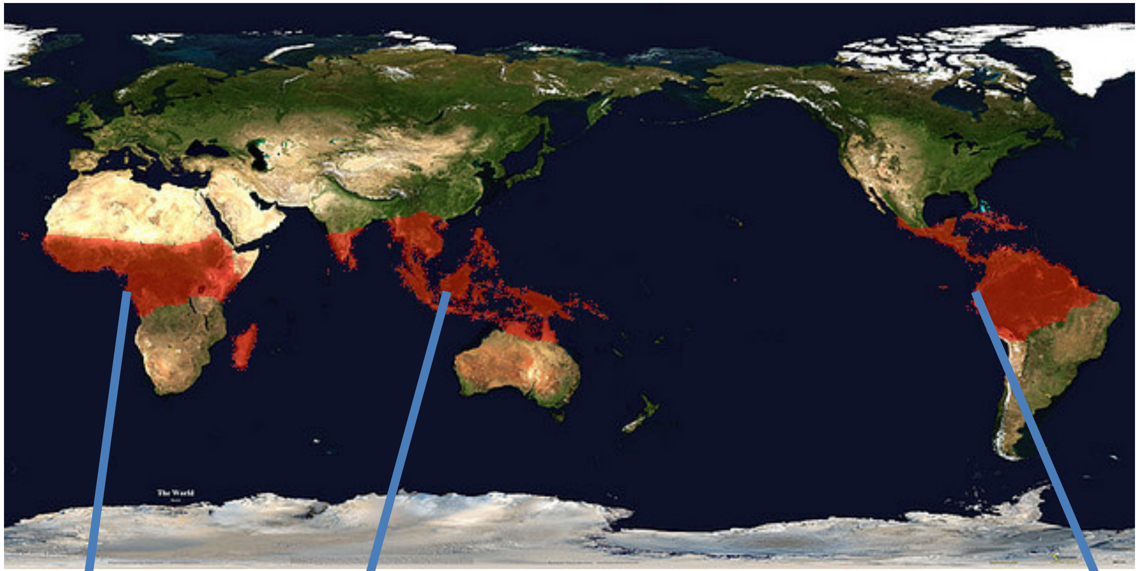
Connecting everyone was the easy part



90% of the world's ecological diversity



90% of the world's linguistic diversity



Reported locally before identification

H1N5 (Bird Flu) – weeks

(>50% fatal)

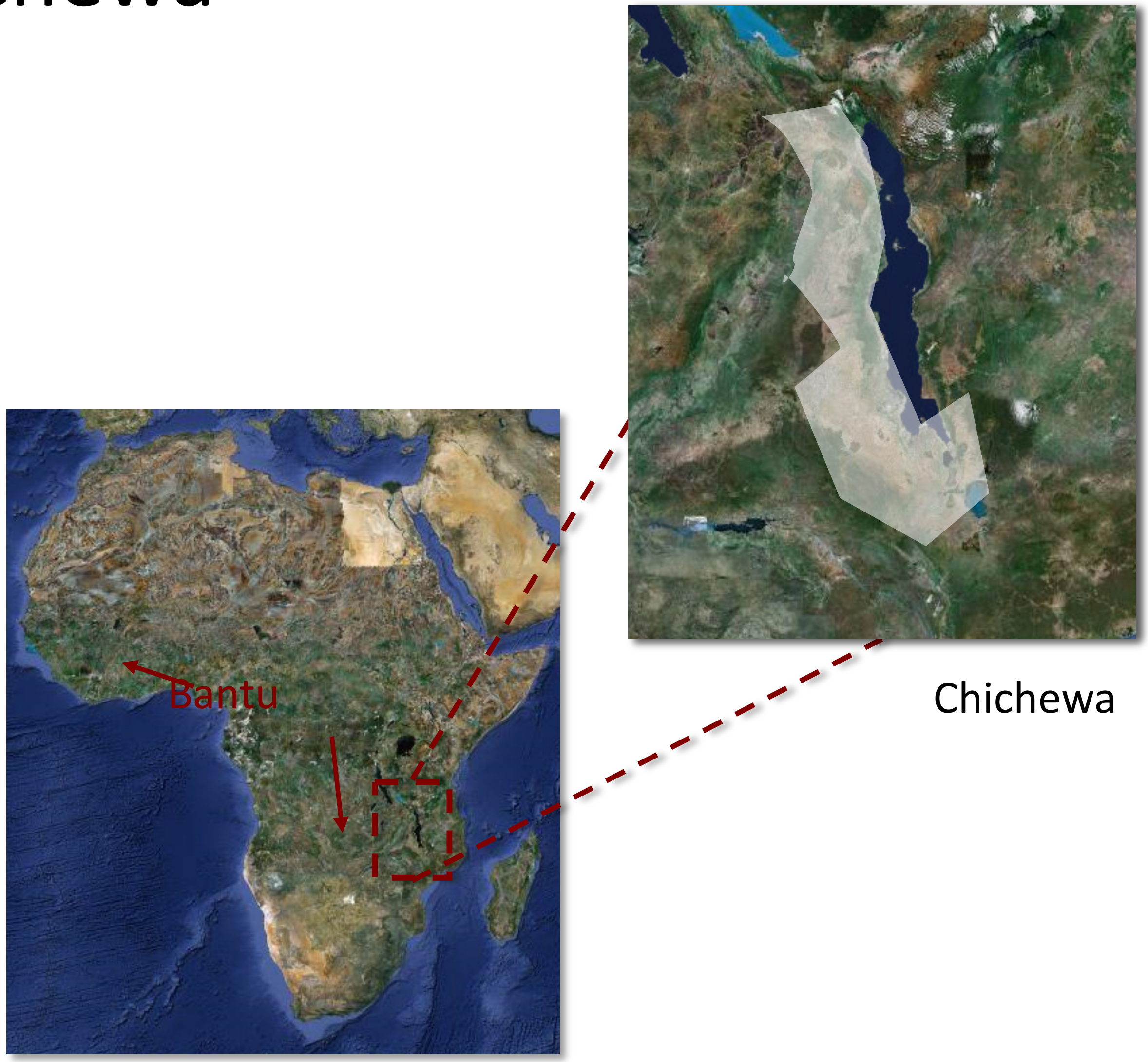
H1N1 (Swine Flu) – months

(10% of world infected)

HIV – decades

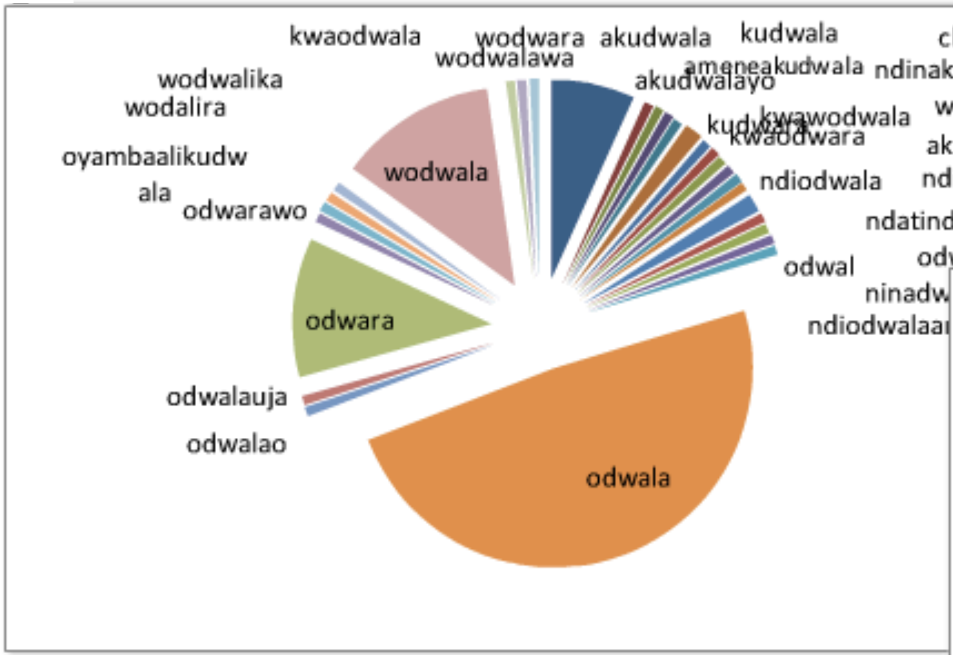
(35 million infected)

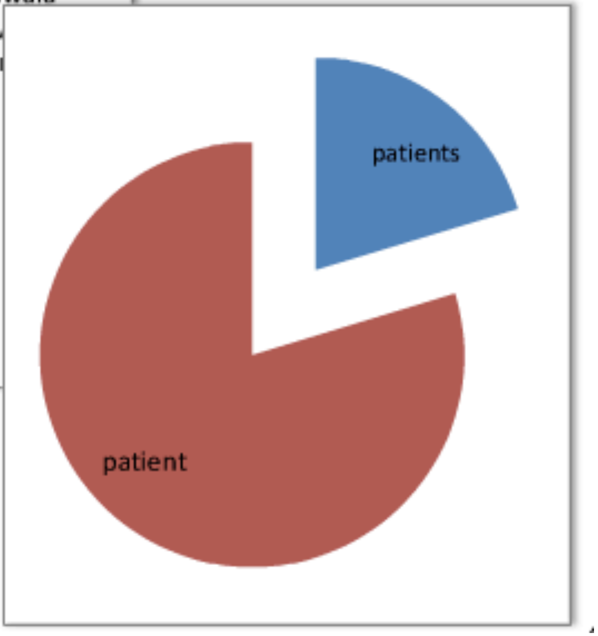
Chichewa



Chichewa

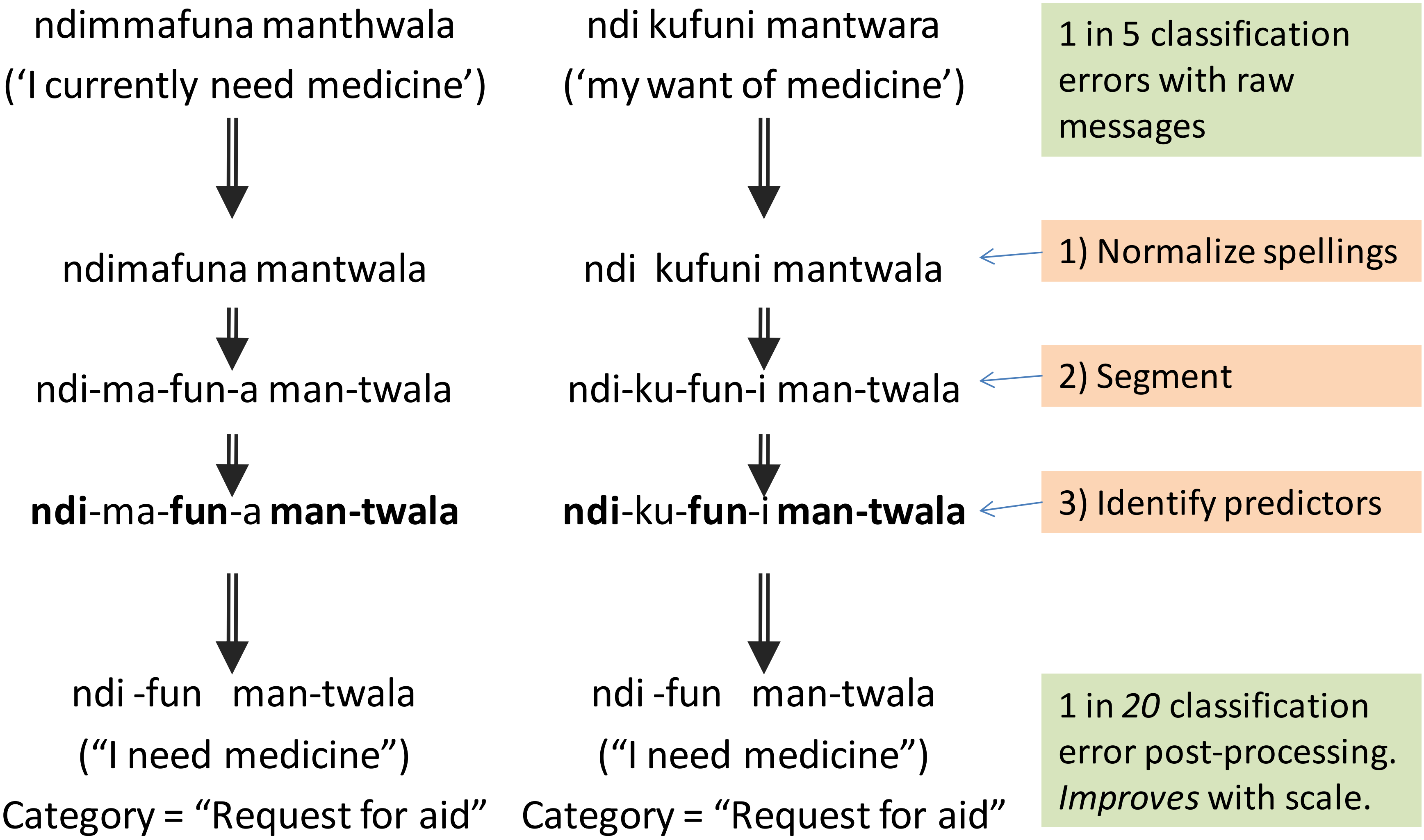
What does language look like?





The word *odwala* (‘patient’) in 500 text-messages in Chichewa (Malawi), and the English translations

Smarter language processing



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

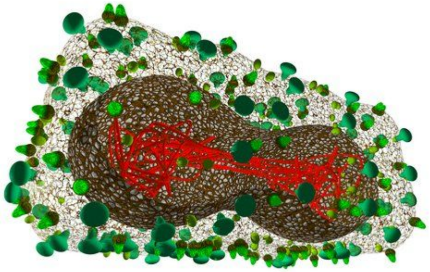
1) Normalize spellings

2) Segment

3) Identify predictors

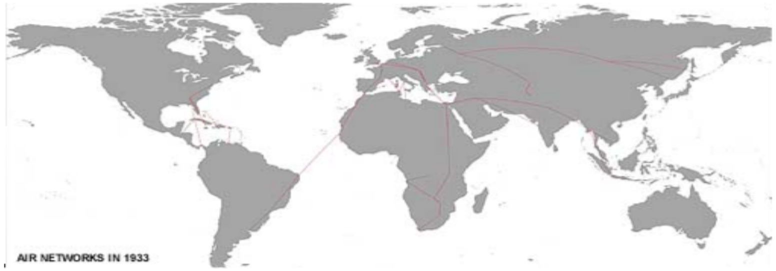
1 in *20* classification error post-processing. *Improves* with scale.

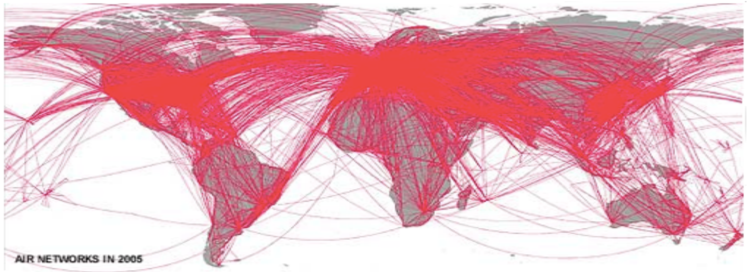
Diseases eradicated in the last 75 years:



smallpox

Increase in air travel in the last 75 years:

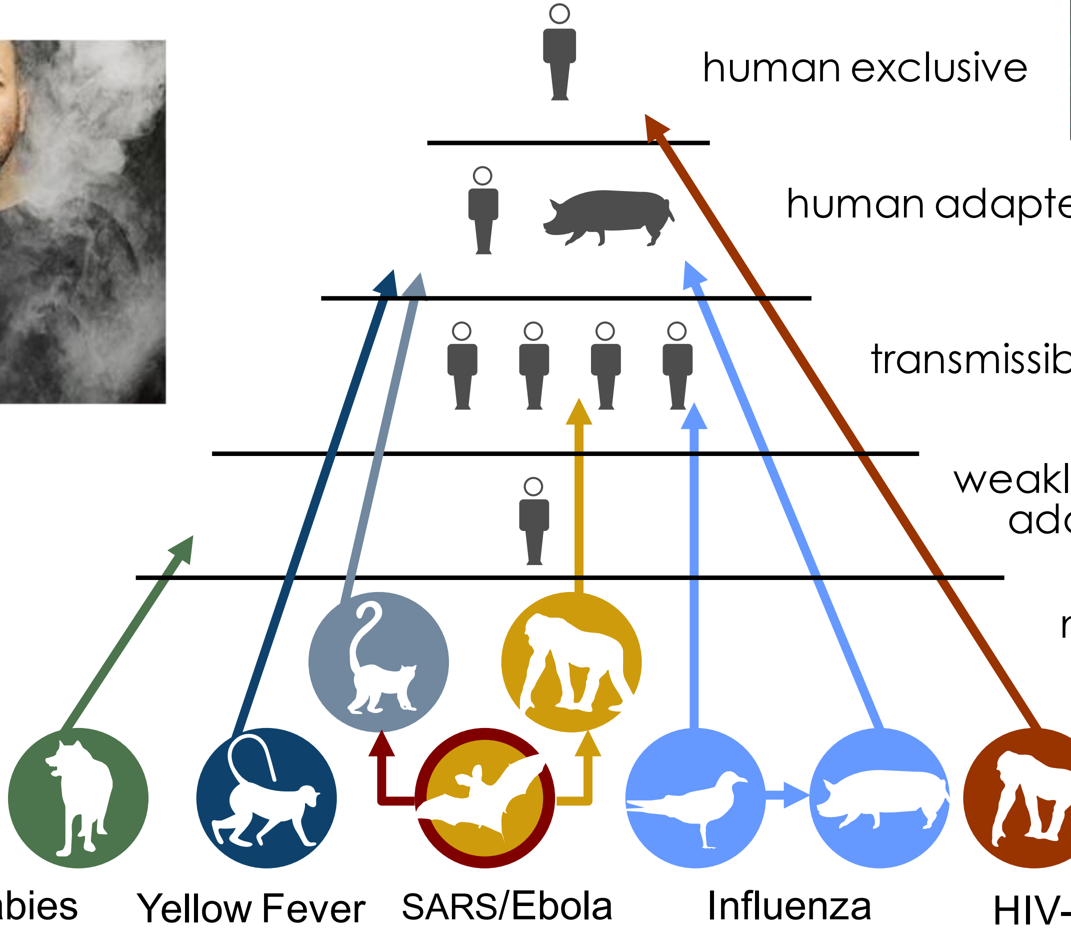




No one is tracking all the world’s outbreaks

- NASA is tracking thousands of potentially dangerous near-Earth objects (NASA 2011).
- National security agencies are tracking tens of thousands of suspected terrorists daily (Chertoff 2008).
- A deadly microbe is far more likely to sneak onto a plane undetected.

Global Viral Forecasting



human exclusive

human adapted

transmissible

weakly human adapted

not human adapted

Rabies

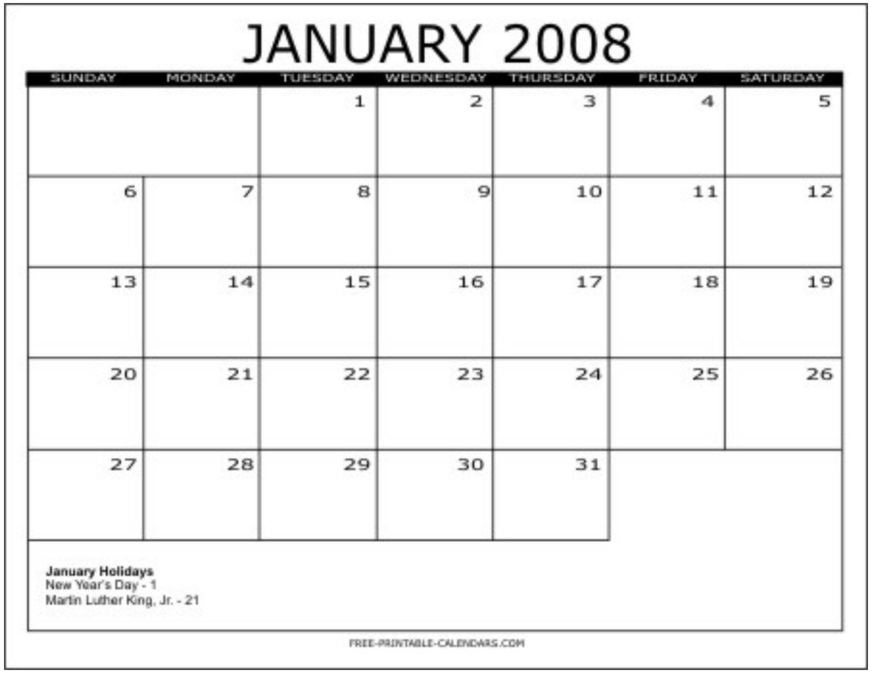
Yellow Fever

SARS/Ebola

Influenza

HIV-1

CDC vs Google Flu Trends?

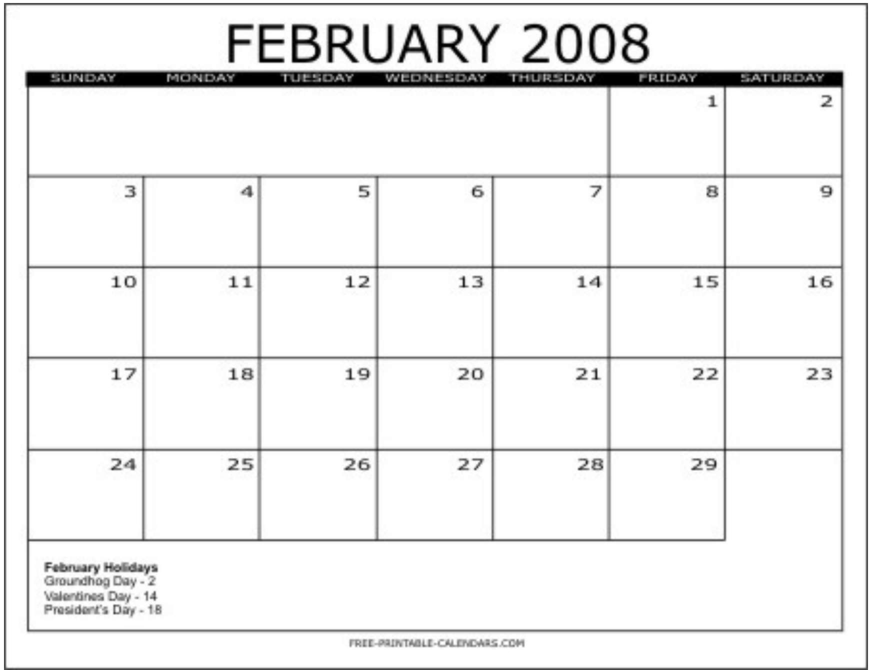


JANUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

January Holidays
New Year's Day - 1
Martin Luther King, Jr. - 21

FREE-PRINTABLE-CALENDARS.COM



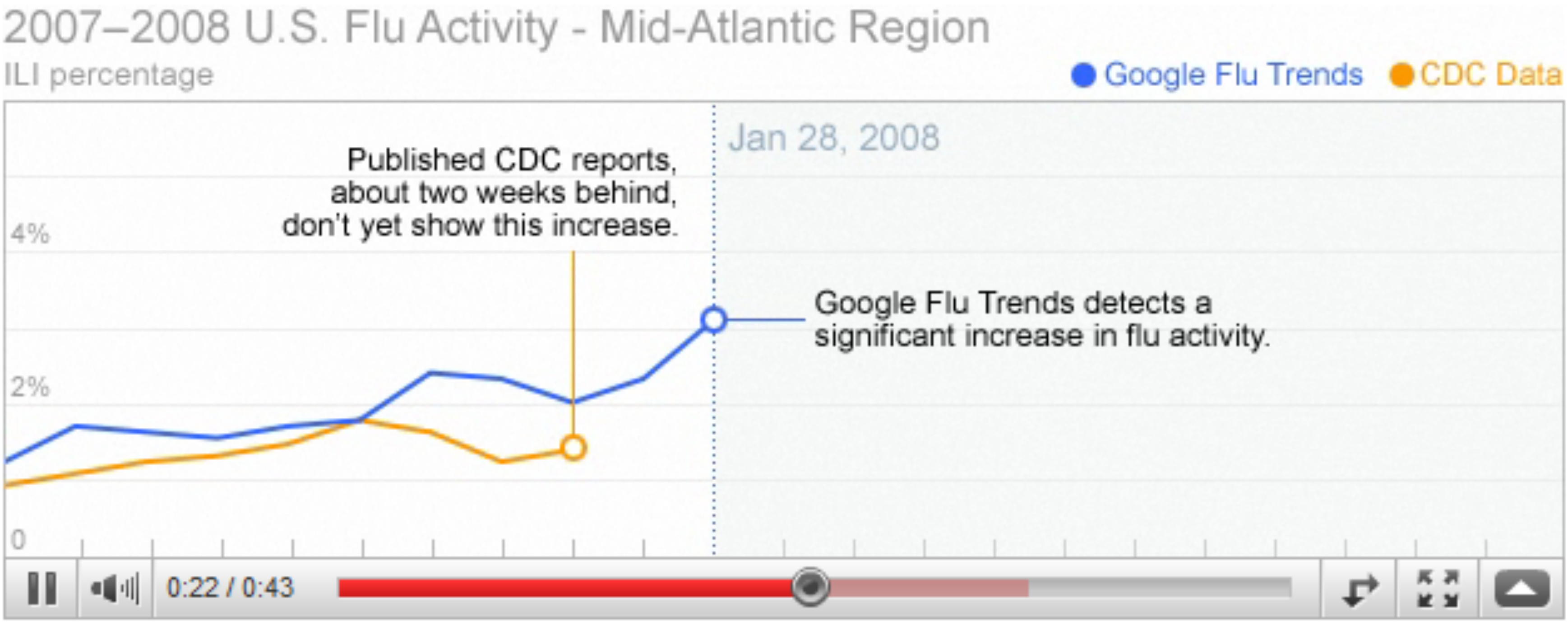
FEBRUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	

February Holidays
Groundhog Day - 2
Valentines Day - 14
President's Day - 18

FREE-PRINTABLE-CALENDARS.COM

CDC vs Google Flu Trends?



2007–2008 U.S. Flu Activity - Mid-Atlantic Region

Published CDC reports, about two weeks behind, don't yet show this increase.

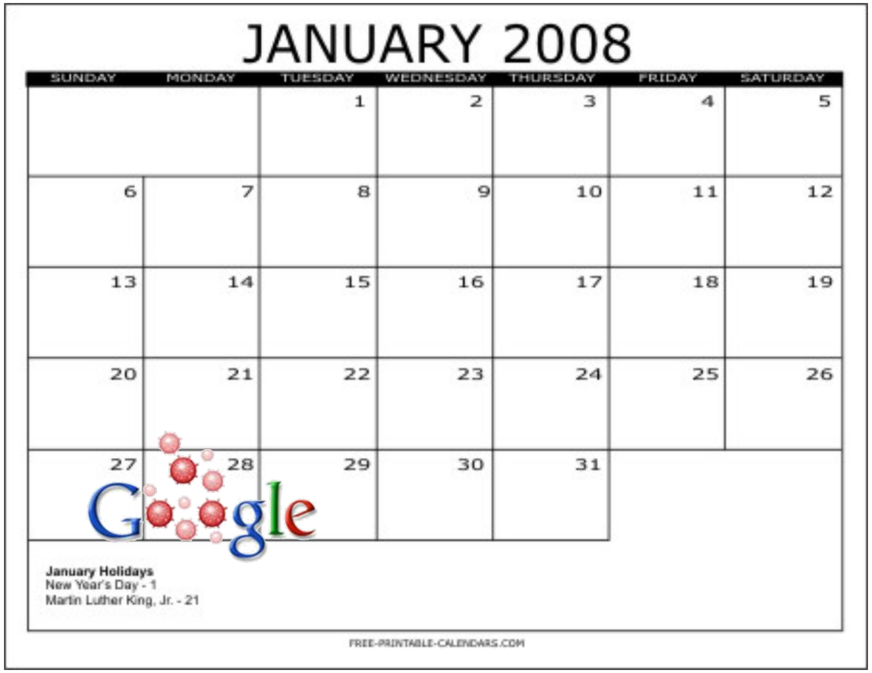
Jan 28, 2008

Google Flu Trends detects a significant increase in flu activity.

Source: http://www.google.org/flutrends/

CDC vs Google Flu Trends?

"I'm Jacqui Jeras with today's cold and flu report ... across the mid- Atlantic states, a little bit of an increase here" **Jan 4th**

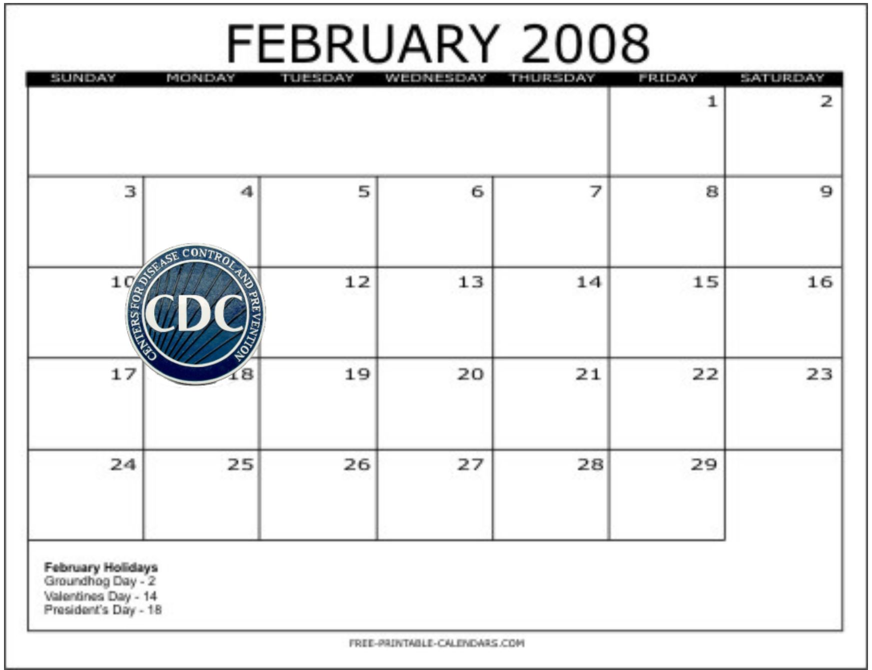


JANUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

January Holidays
New Year's Day - 1
Martin Luther King, Jr. - 21

FREE-PRINTABLE-CALENDARS.COM

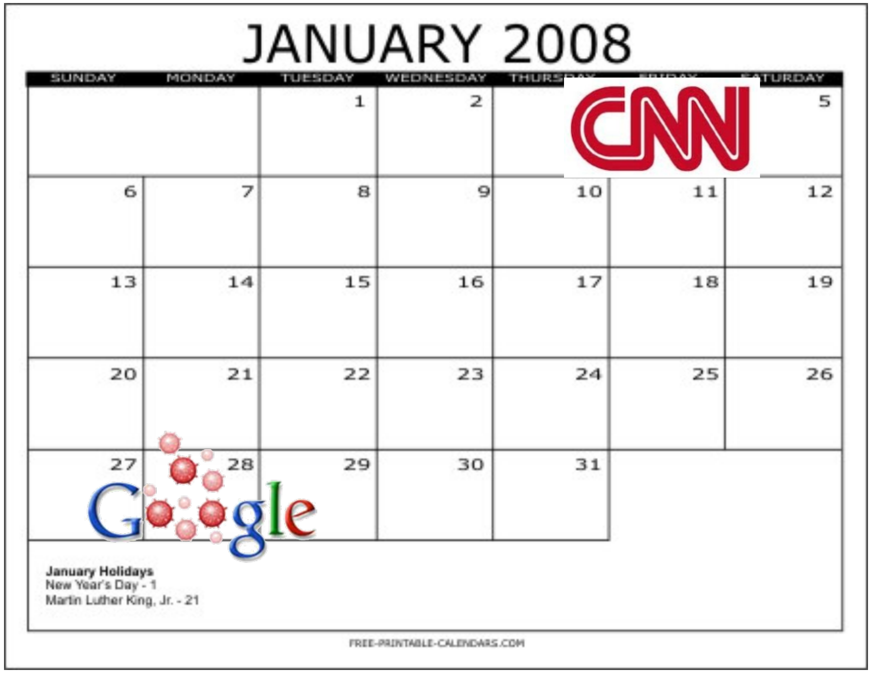

CENTERS FOR DISEASE CONTROL AND PREVENTION
CDC

FEBRUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
					1	2
3	4	5	6	7	8	9
10		12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	

February Holidays
Groundhog Day - 2
Valentines Day - 14
President's Day - 18

FREE-PRINTABLE-CALENDARS.COM

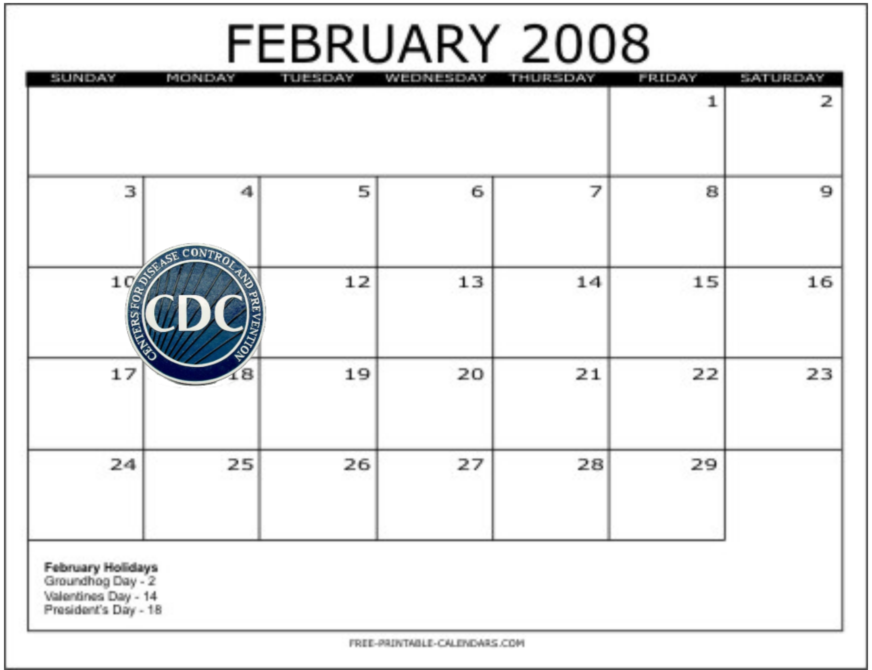


JANUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
		1	2			5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

January Holidays
New Year's Day - 1
Martin Luther King, Jr. - 21

FREE-PRINTABLE-CALENDARS.COM


CENTERS FOR DISEASE CONTROL AND PREVENTION
CDC

FEBRUARY 2008

SUNDAY	MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY	SATURDAY
					1	2
3	4	5	6	7	8	9
10		12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	

February Holidays
Groundhog Day - 2
Valentines Day - 14
President's Day - 18

FREE-PRINTABLE-CALENDARS.COM

The first signal is plain language

“today's cold and flu report … across the mid-Atlantic states, a little bit of an increase” CNN
Jan 4, 2008

Google Flu Trends
+ 3 weeks

CDC
+ 5 weeks

… but buried in plain view

“today's cold and flu report … across the mid-Atlantic states, a little bit of an increase”

“We're worried about the markets.”

“We're going to take you to Kenya where the U.S. has dispatched some diplomatic help to try to get the country back on political balance.”

“A spunky boy reels in a 550-pound shark.”

“Is individualism an endangered concept in Saudi Arabia?”

“Well, in St. John's County, one man lost his home trying to keep his pig warm.”

“The pig did not make it.”

“He had everything but the cape. A good samaritan in Ohio saved a family from this ferocious house fire.”

… in 1000s of languages

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

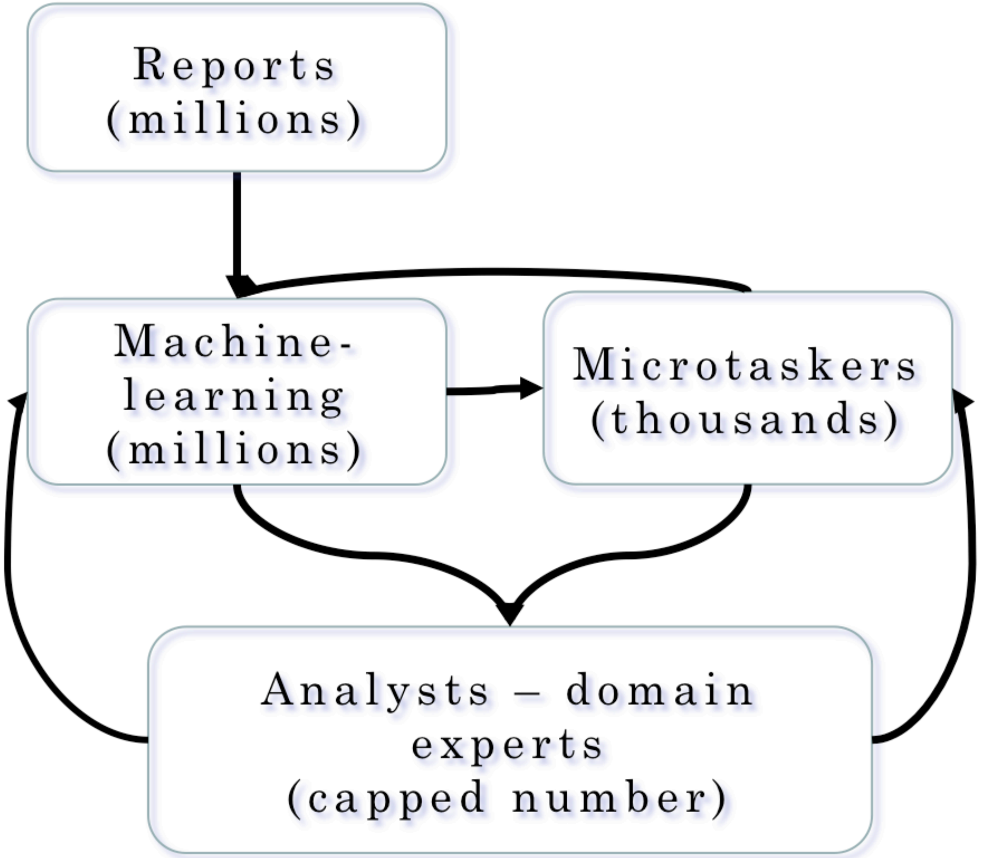
(2 flu outbreaks predicted for the Ukraine)

مزيد من انفلونزا الطيور في مصر

(more flu in Egypt)

香港现1例H5N1禽流感病例曾游上海南京等地

(Hong Kong had a case of avian influenza that traveled to Shanghai and Nanjing)



Reports
(millions)

Machine-learning
(millions)

Microtaskers
(thousands)

Analysts – domain experts
(capped number)

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

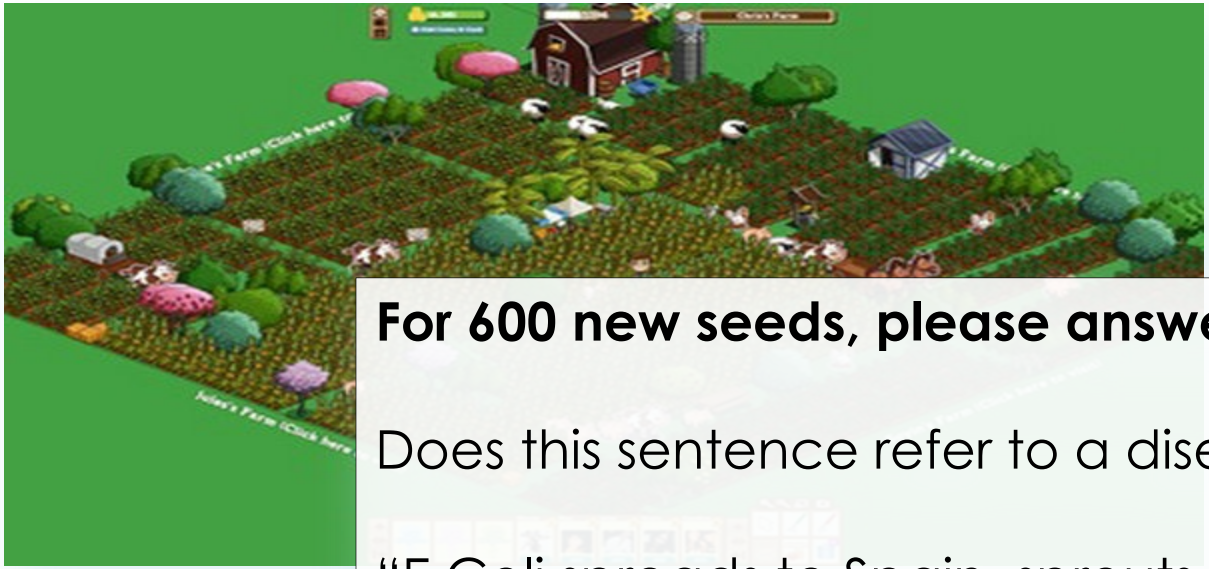
مزيد من انفلونزا الطيور في
مصر

香港现1例H5N1禽流感病例
游上海南京等地

Data structuring

- Disease (if known)
- Case counts / demographics
- Location
- Responding organizations
- Transport used
- Quotes from officials
- Changing conditions (spreading / ending)
- Public reaction

Motivations



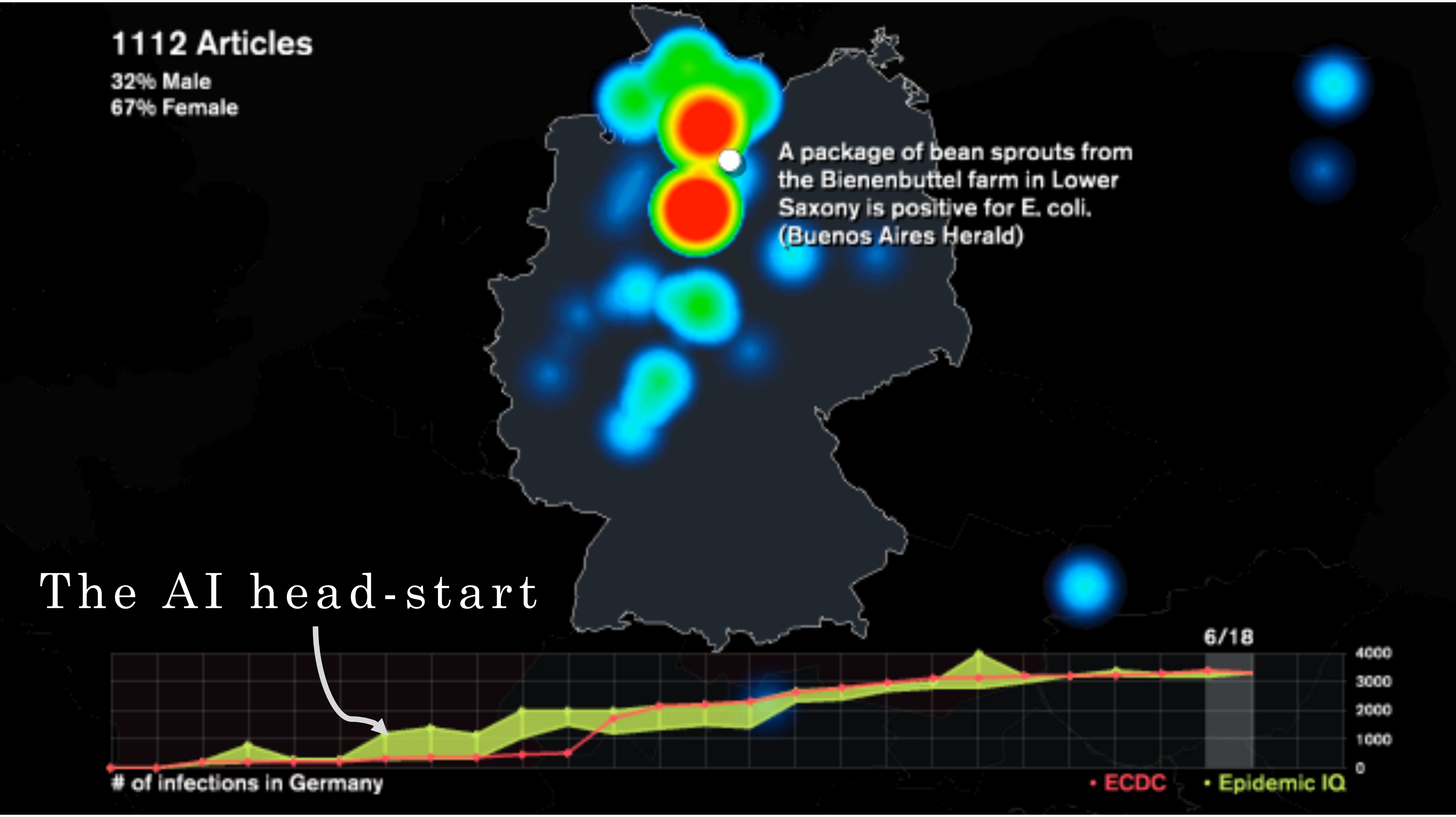
For 600 new seeds, please answer this question:

Does this sentence refer to a disease outbreak:

“E Coli spreads to Spain, sprouts suspected”

Yes/no: __
What disease: _______
What location: _______

The AI head-start



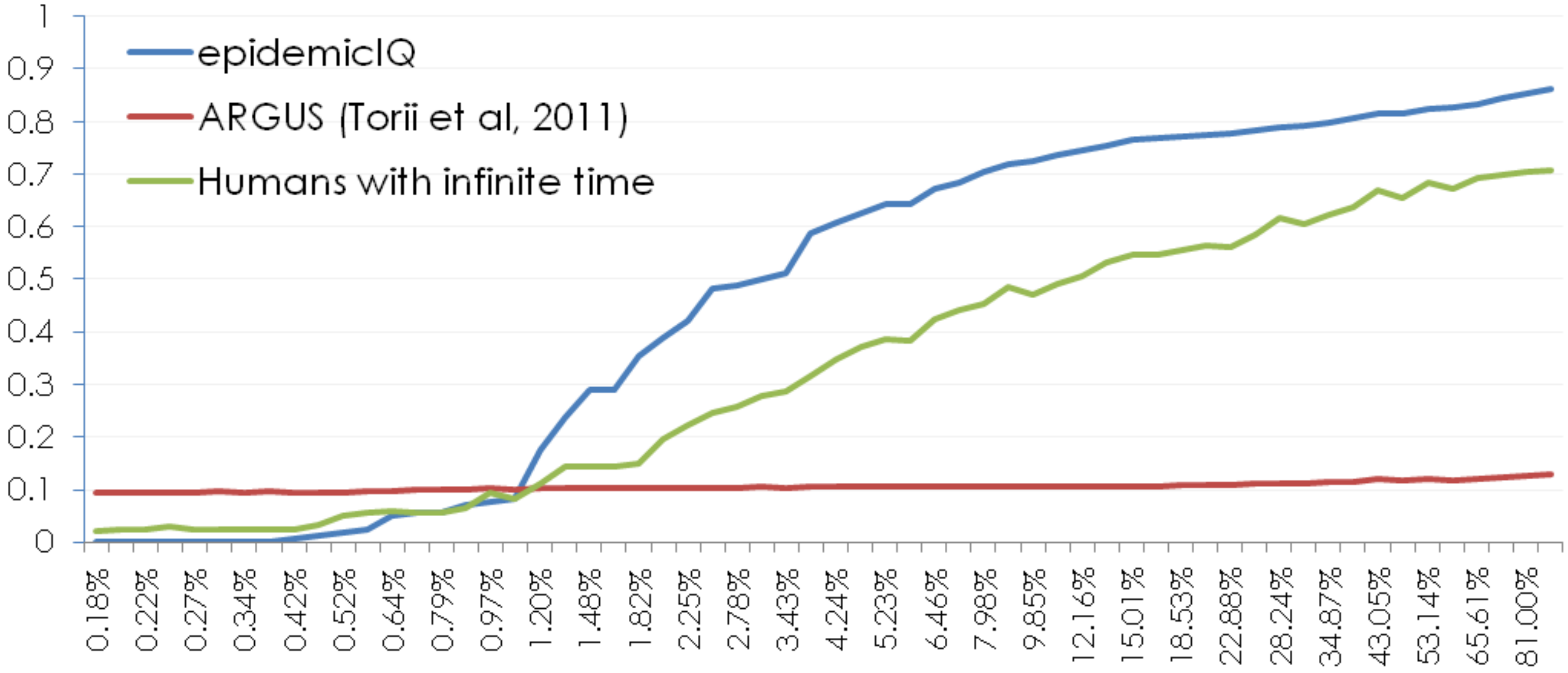
1112 Articles

32% Male

67% Female

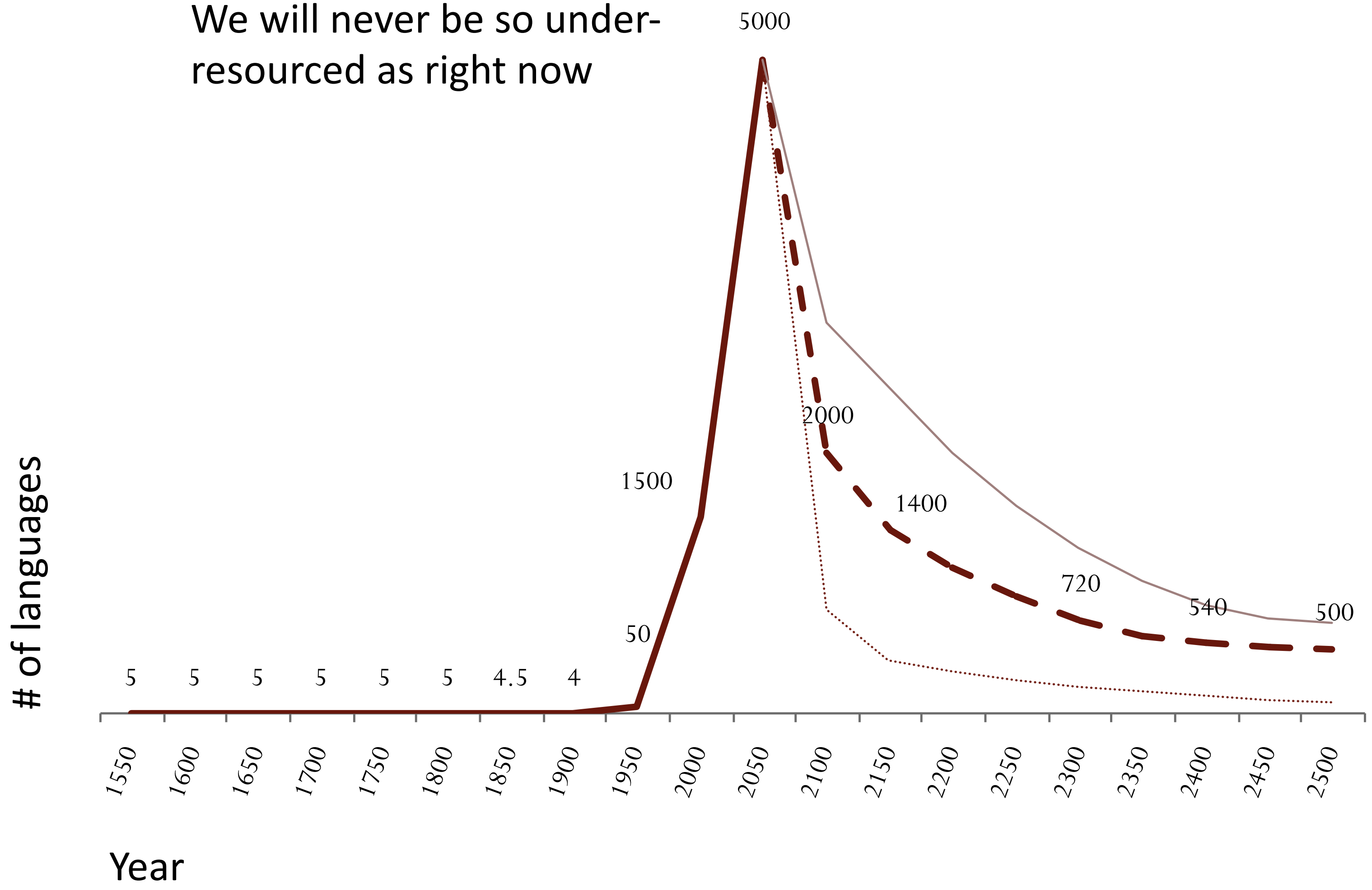
A package of bean sprouts from the Bienenbuttel farm in Lower Saxony is positive for E. coli. (Buenos Aires Herald)

Predicting epidemics, 100K training items



Daily potential language exposure

We will never be so under-resourced as right now


# of languages
Year