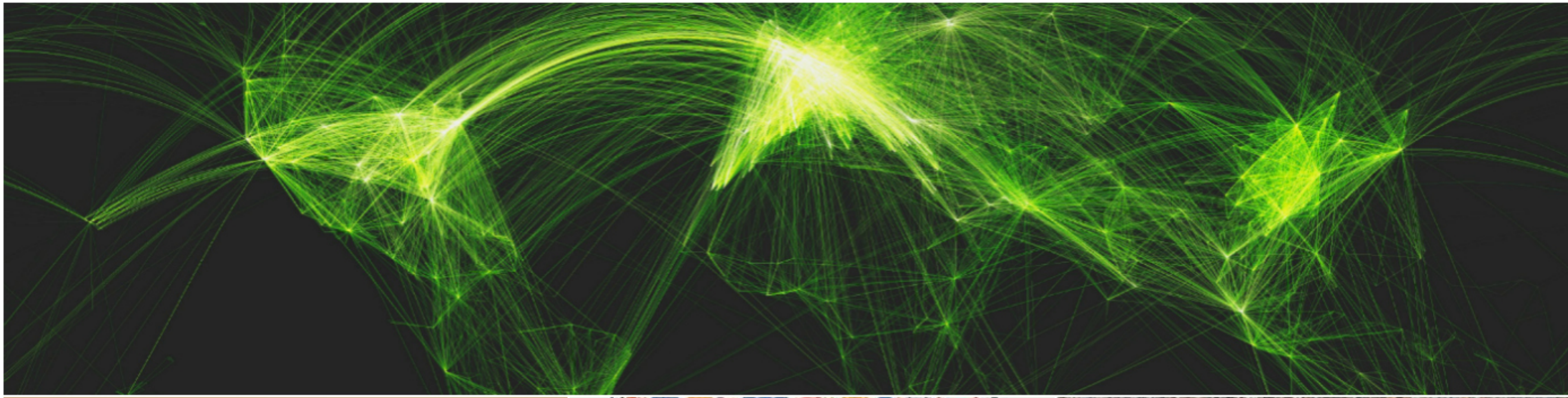
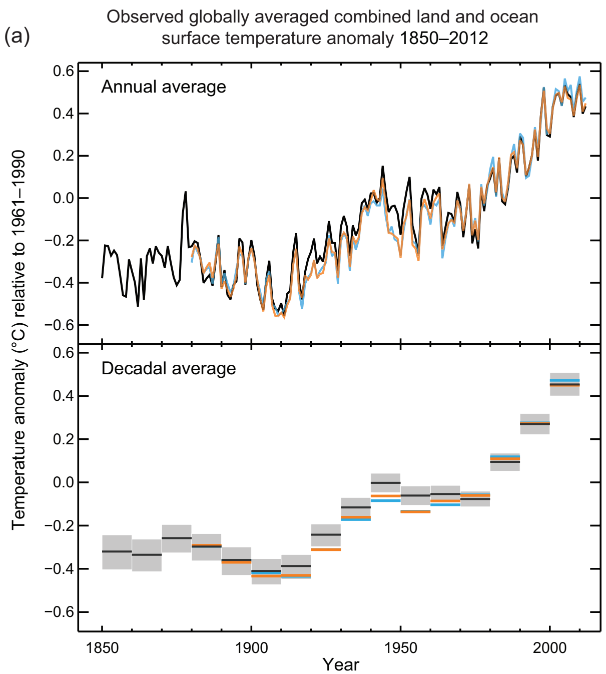
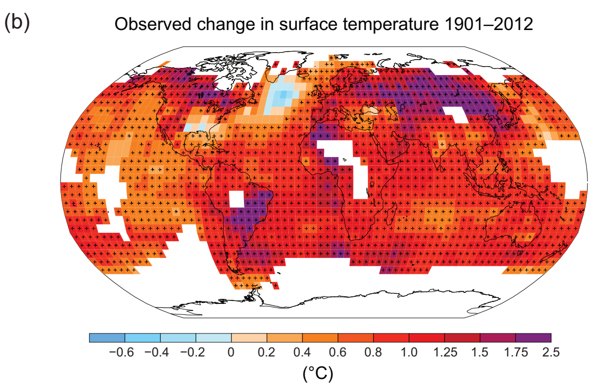
Big data and the rise and spread of infectious disease

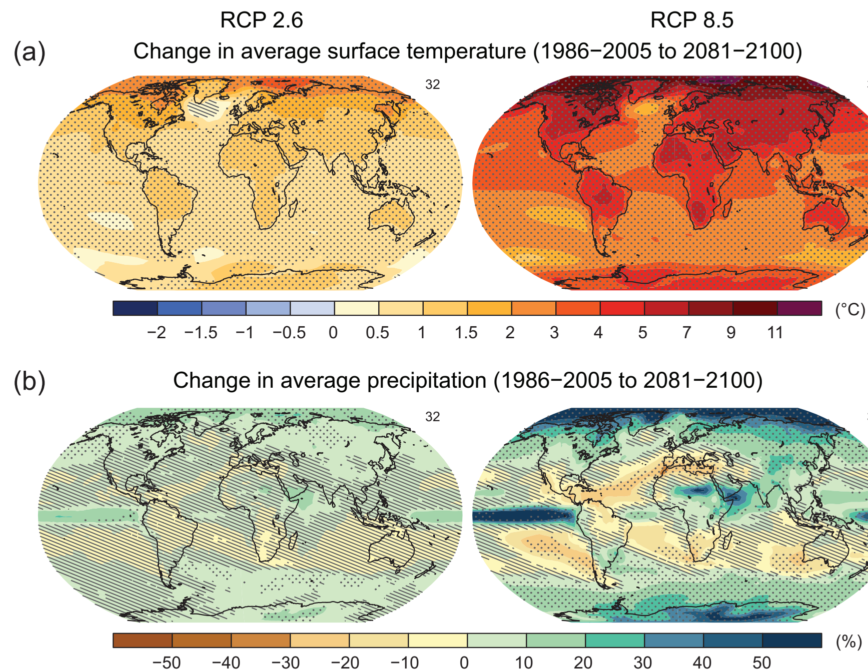


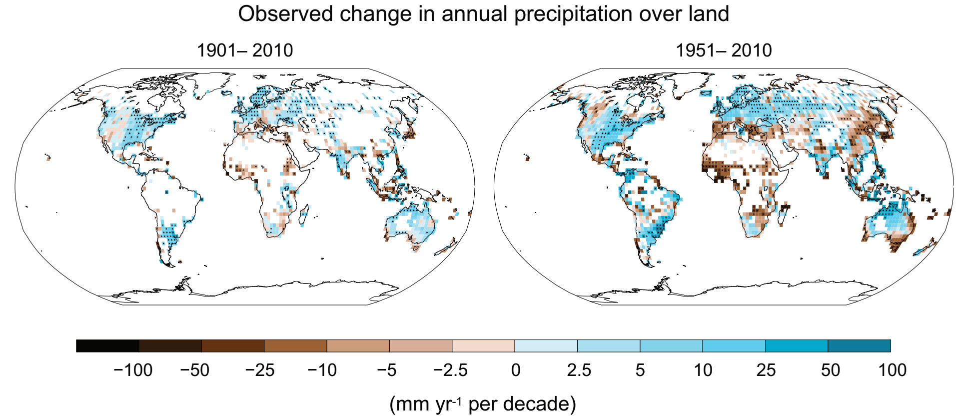
David Taylor, Professor of Tropical Environmental Change
Department of Geography,
National University of Singapore

Climates have changed, are changing & will continue to change ……





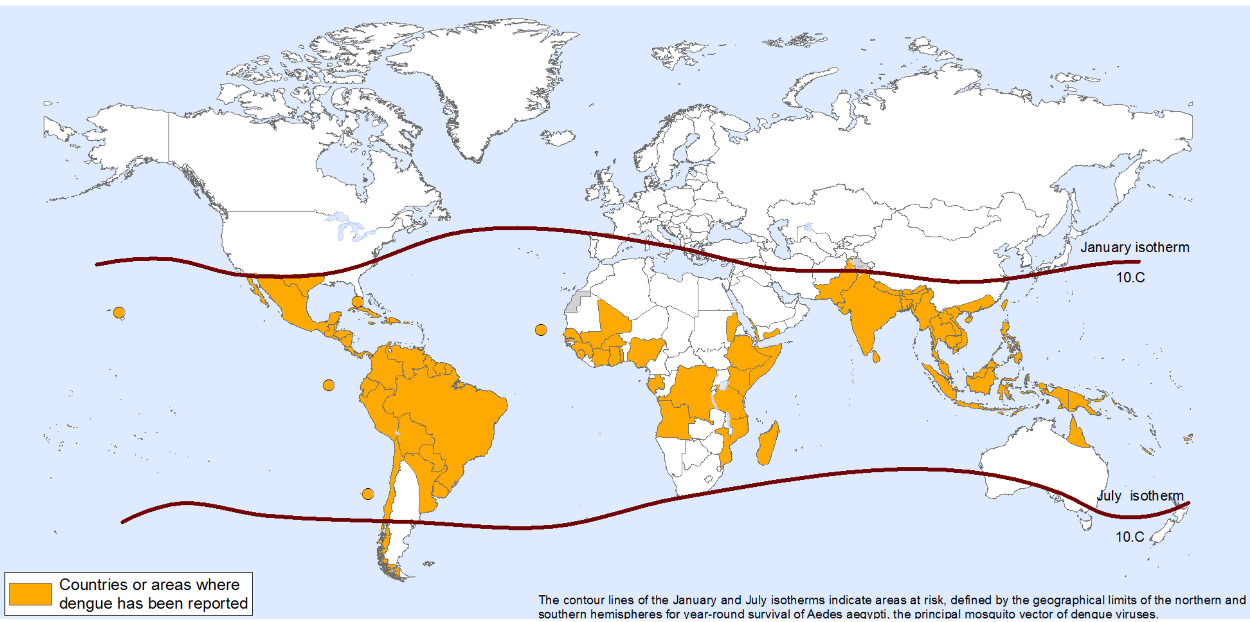




IPCC A5(2013) Summary for Policy Makers – WG1 report

Can expect distribution and epidemic potential of climate sensitive infectious diseases to shift ….

Dengue ~ climatically-sensitive vector-borne virus



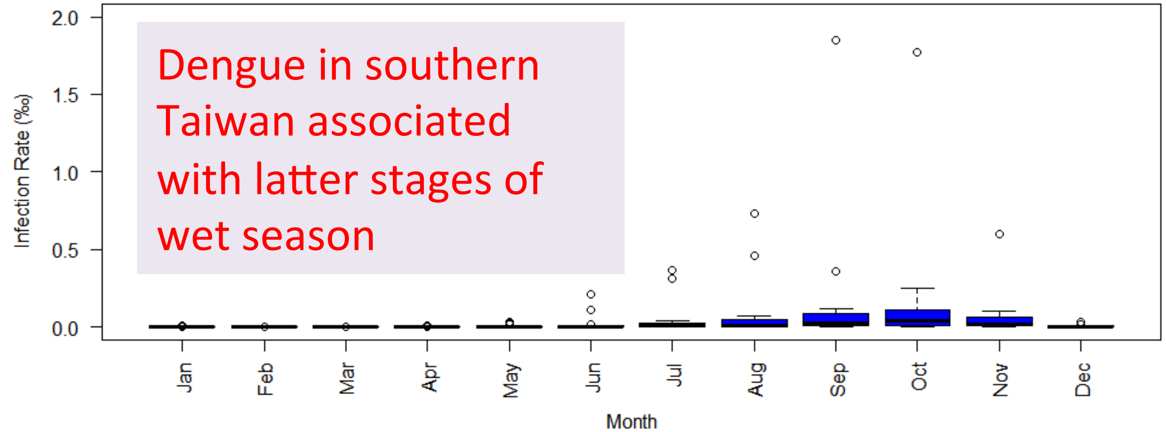
Dengue distribution (2013) From WHO

E.g. Dengue …

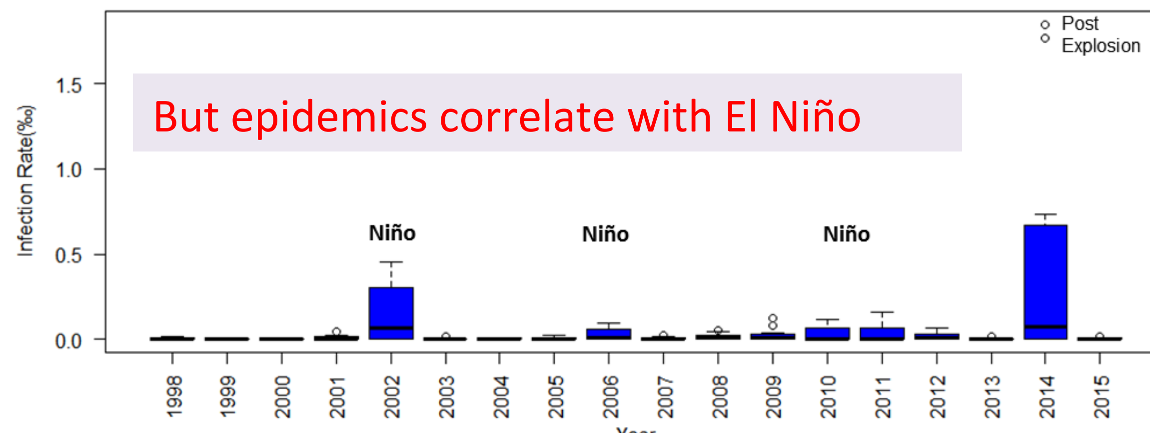
- transmitted by **Aedes** mosquitoes *(Ae. aegypti* and *Ae. Albopictus*)
- Highly **anthrophyllic** vectors
- Same vectors transmit **chikungunya**, **yellow fever** and **zika** (and **Rift Valley fever virus** in Africa!)
- Vectors and diseases are expanding range & severity of infections is increasing
- Not everyone infected shows symptoms!

Recent spread of dengue to higher altitude parts of Asia (e.g. Nepal) suggests that climate change may be a factor

Dengue epidemics in southern Taiwan: a possible link to climate?

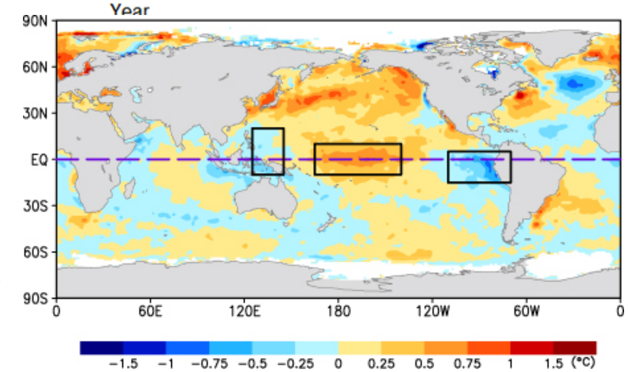


Dengue in southern Taiwan associated with latter stages of wet season

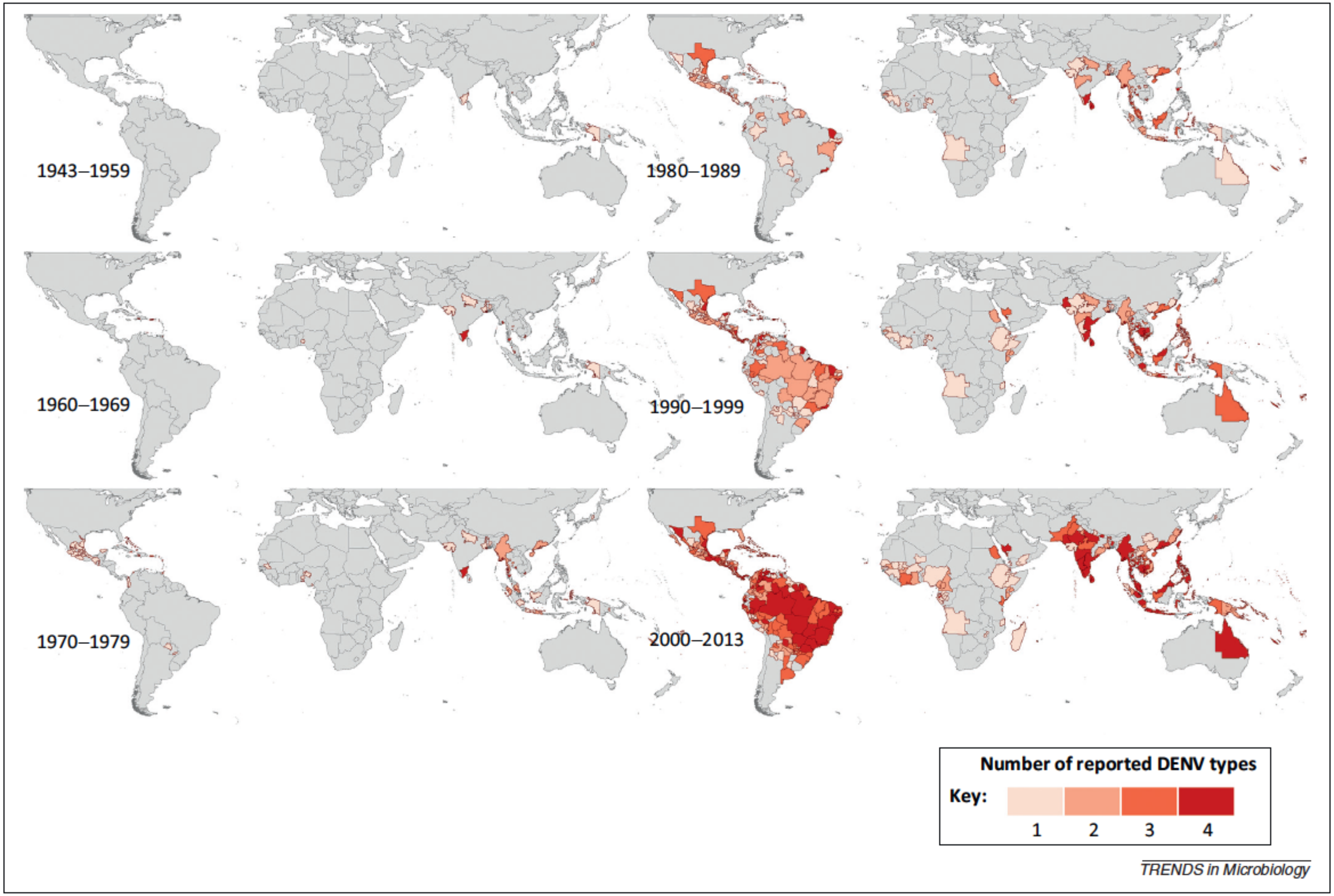


But epidemics correlate with El Niño

Especially with **modoki form of El Niño** – modoki El Niño forces south the main typhoon track bringing more rain to Taiwan

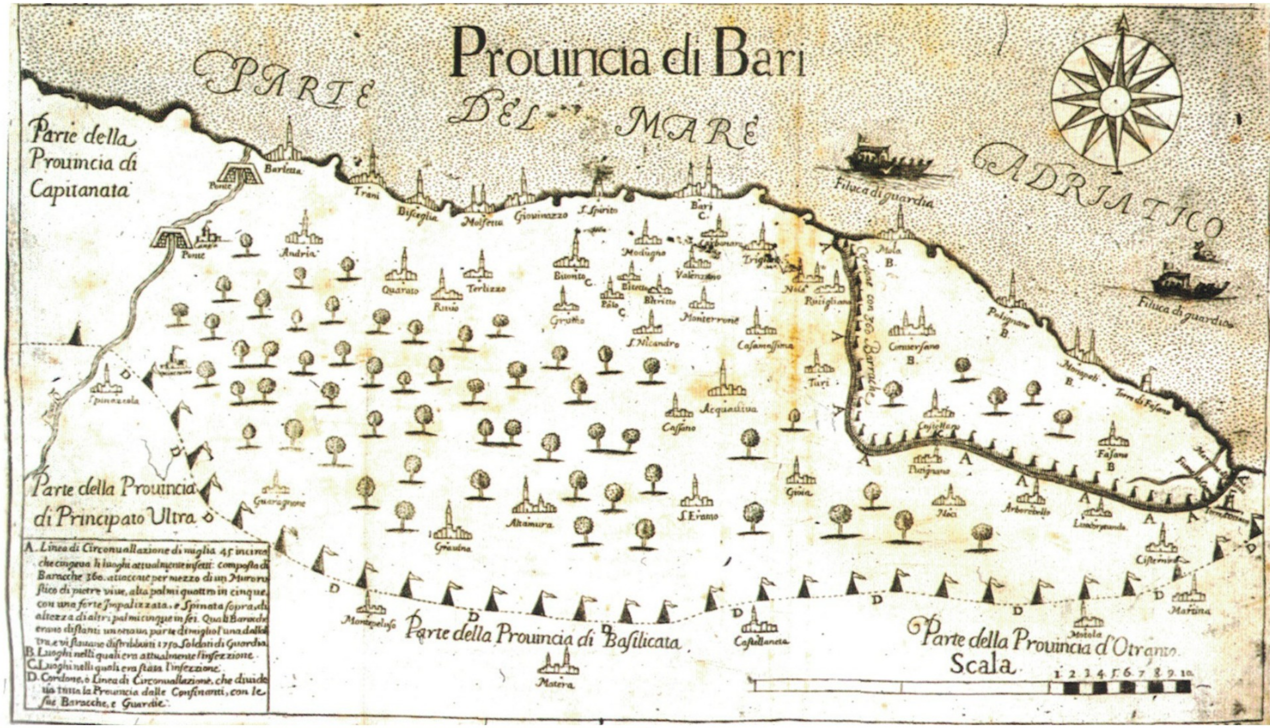


Spread of dengue from late 1970s suggests more than climate change



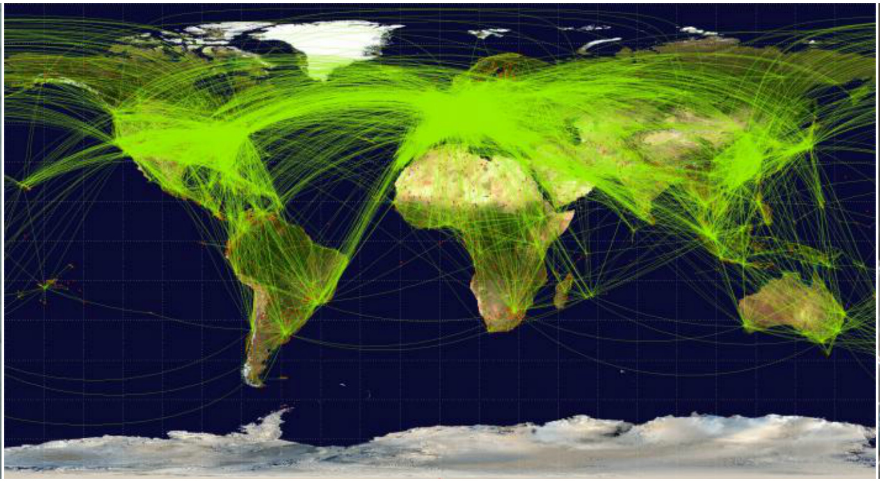
From Messina et al. (2014) Global spread of dengue virus types: mapping the 70 year history, *Trends in Microbiology* 22: 138-146

Rate of spread of new and existing infectious diseases raises questions about conventional disease surveillance and containment techniques …

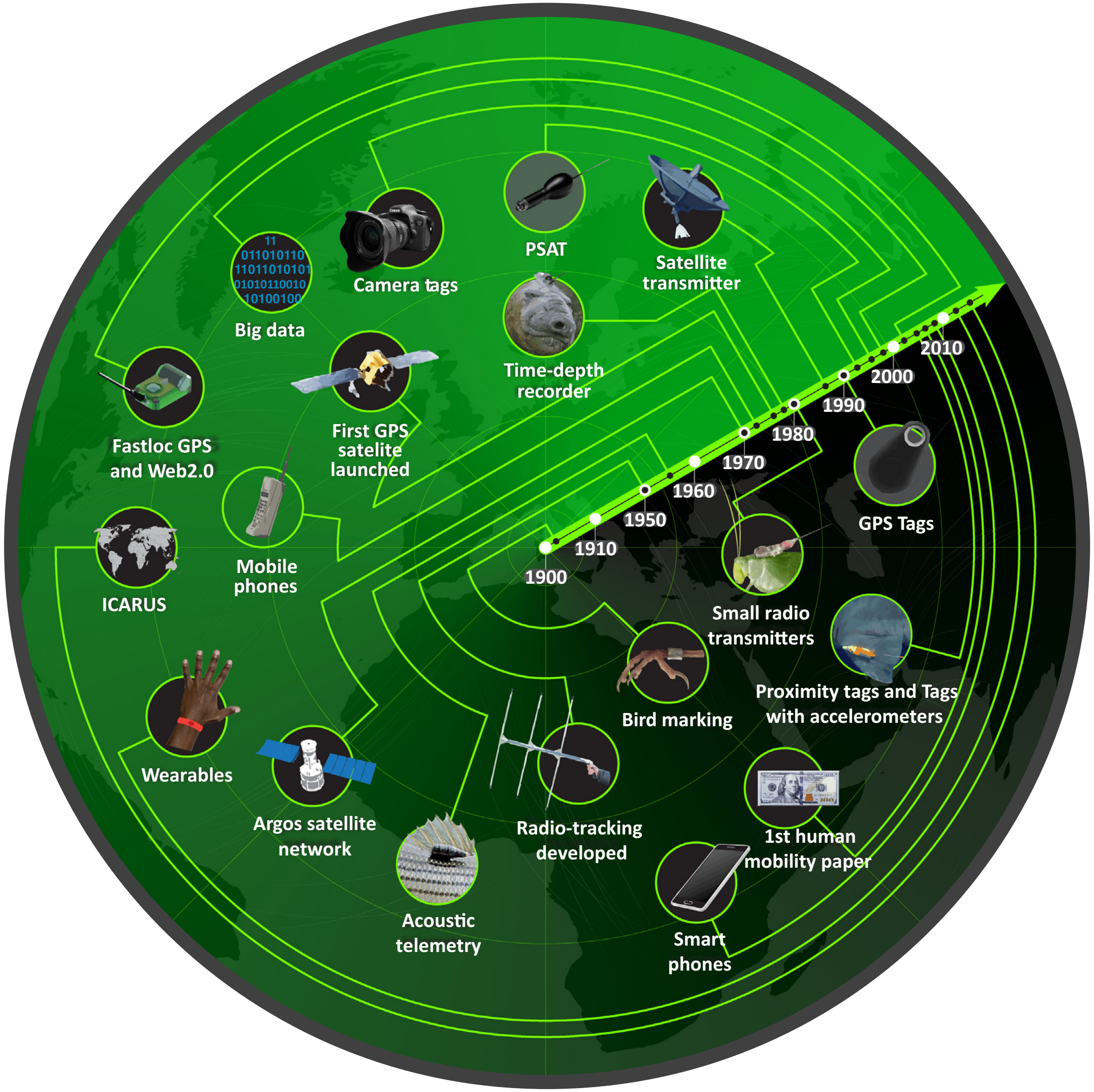


Quarantine – the first public health initiative? First mentioned in 12th century, the re-emergence of plague around Bari in the 1690s led to soldiers being used to enforce restrictions on movement ….

Other factors behind the rapid spread of some infectious diseases include land cover change, travel & trade, pollution etc



Just as modern technologies have facilitated the rapid spread of infectious diseases, we can also use modern technologies to monitor and anticipate their effects



Timeline of technological advances in animal movement and human mobility.

From Meekan et al. (2017) The ecology of human mobility. *Trends in Ecology & Evolution* 32: 198-210

No agreed definition of Big Data

Generally (e.g. Kitchin 2013) large datasets that are characterised by the *three vs ++*:

- Huge in *volume*
- High in *velocity*
- Diverse in *variety*

And a 4th "v": veracity (?)

And
- Exhaustive in **scope** – capture entire populations of data
- Fine-grained in **resolution**
- **Relational –** different datasets can be linked
- **Flexible** – easy to add to and to vary the scale

Also includes the technologies to analyse big data

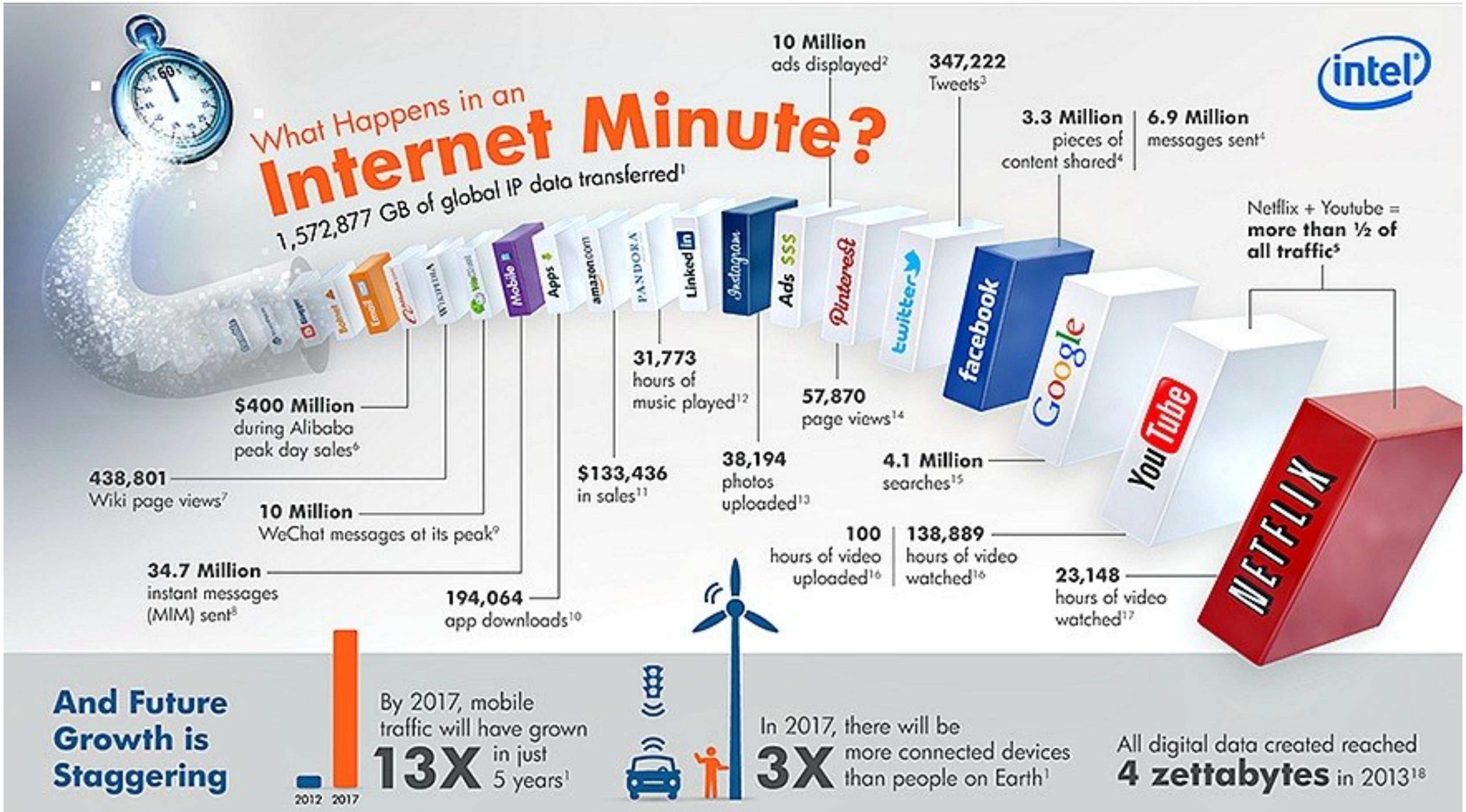
Different sources of big data:

1) **Directed** – generated by digital forms of surveillance

2) **Automated** – generated by an inherent, automatic function, e.g. credit card transactions, use of travel cards, tweets, interactions with the internet (clickstream data etc) etc

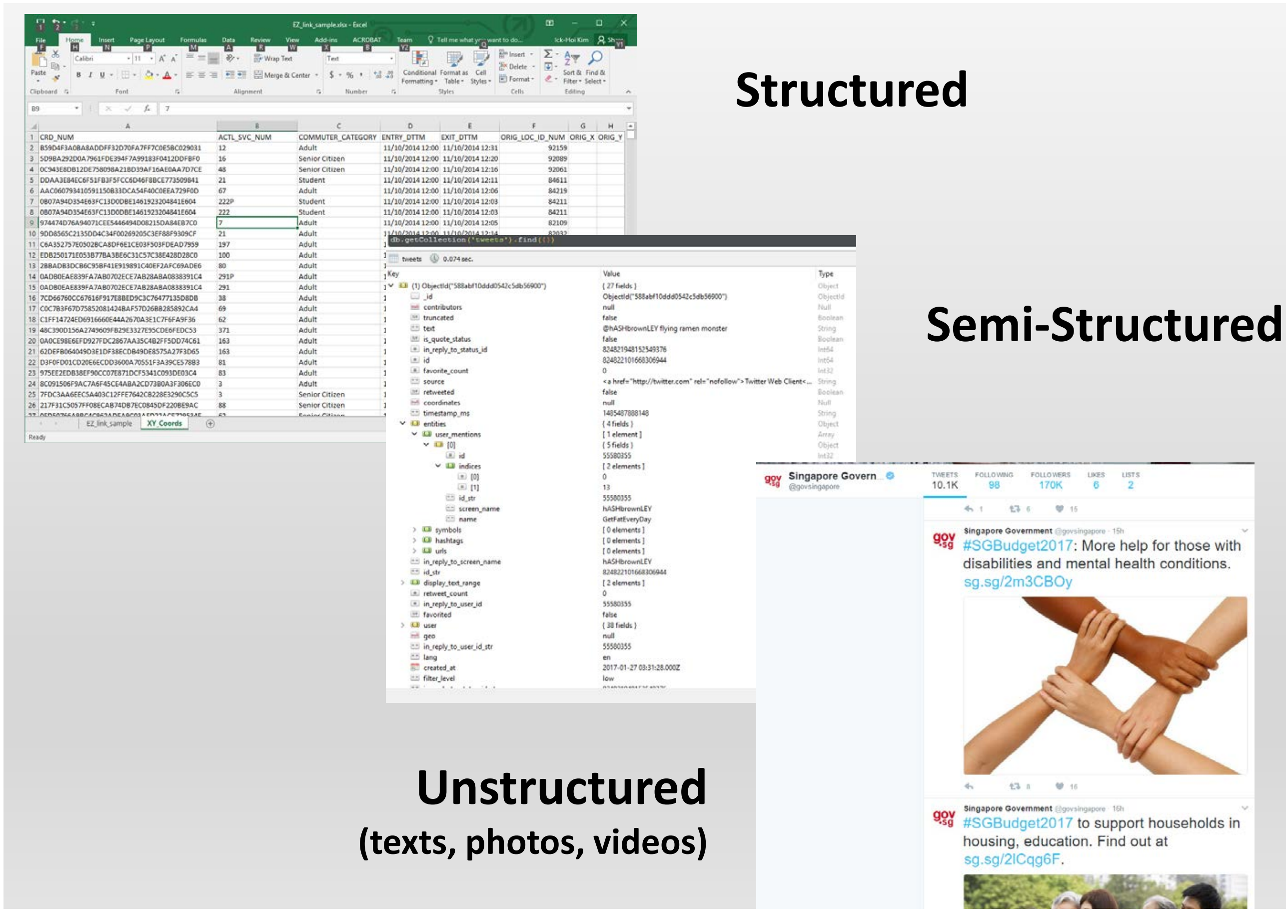
3) **Volunteered** – gifted by users, e.g. crowdsourcing of data, social media posts etc

The total amount of data produced and consumed is rapidly increasing



We all contribute to the production of data (create "**data shadows**"), much of it with spatial attributes ("**geotags**")

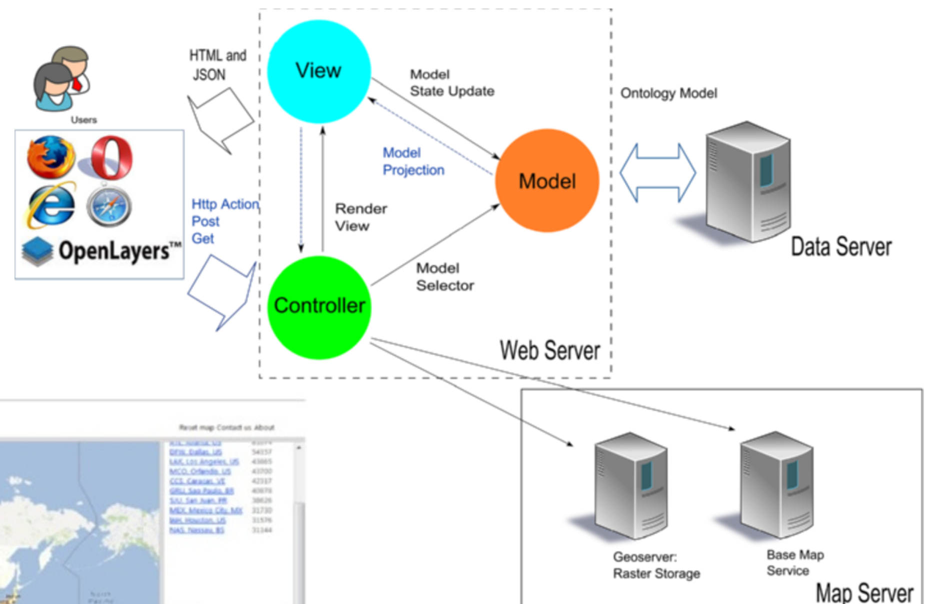
Big data comes in different forms, and generally needs “cleaning”



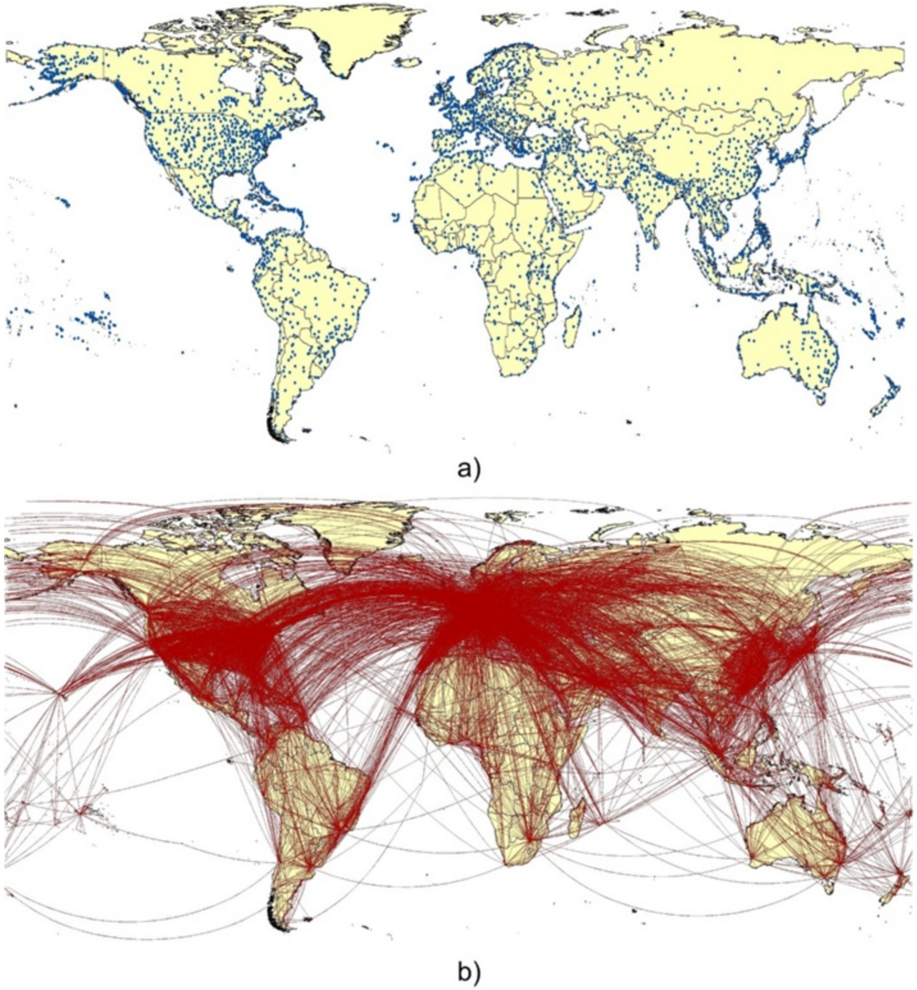
Need to separate signal from noise, identify misinformation (“fake news”) and determine how representative the information is

Big data creates new opportunities (e.g. new sub-disciplines of *infodemiology* and *macroecology*) and brings new insights (e.g. link between travel & spread of infectious disease)

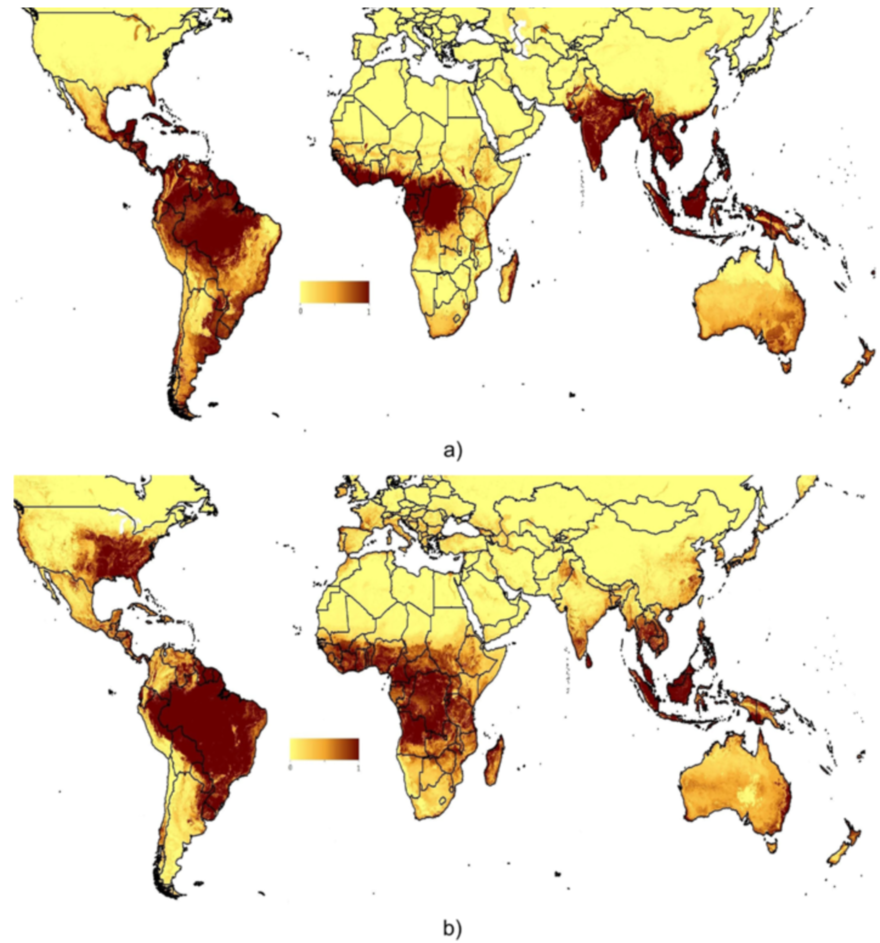
VBD-AIR tool is used to determine the disease importation risk at airports around the world – e.g. dengue for Miami airport, US



Huang, Z. et al. (2012) Web-based GIS: the vector-borne disease airline importation risk (VBD-AIR) tool. *International Journal of Health Geographies* 11: 33



Air network data (a) airports, (b) flight routes for 2011



Information on risk of chikungunya infection (a) and suitability for *Aedes alopictus* vector (b)

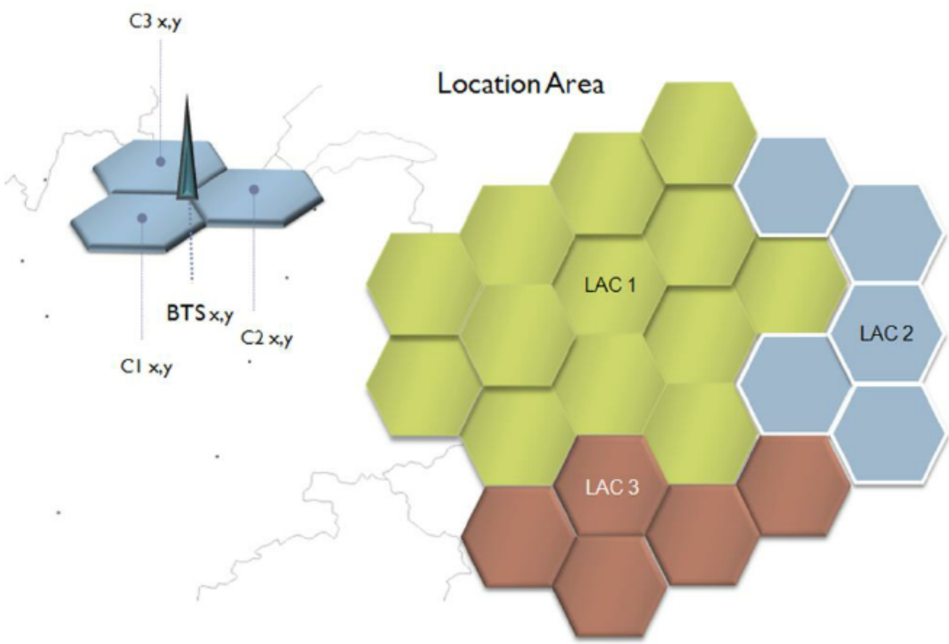
Big data have the potential to:

- Augment existing surveillance systems
- Provide an early warning of a disease outbreak
- Provide a basis for research on health and related matters

Big data have the potential to:

- **Augment existing surveillance systems**
- Provide an early warning of a disease outbreak
- Provide a basis for research on health and related matters

(Mobile) Cellular phone data provide up-to-date information on locations on individuals and their movements



Cellular network comprises a network of Base Transceiver Stations (BTS)

Each cell in the network has a Cell Identifier (ID)

During movement between cells, the network commands the mobile unit to switch to the next cell

The cellular handover records all of the cells through which the mobile unit passes

Cellular network is divided into larger, geo-administrative zones (Location Areas, LA)

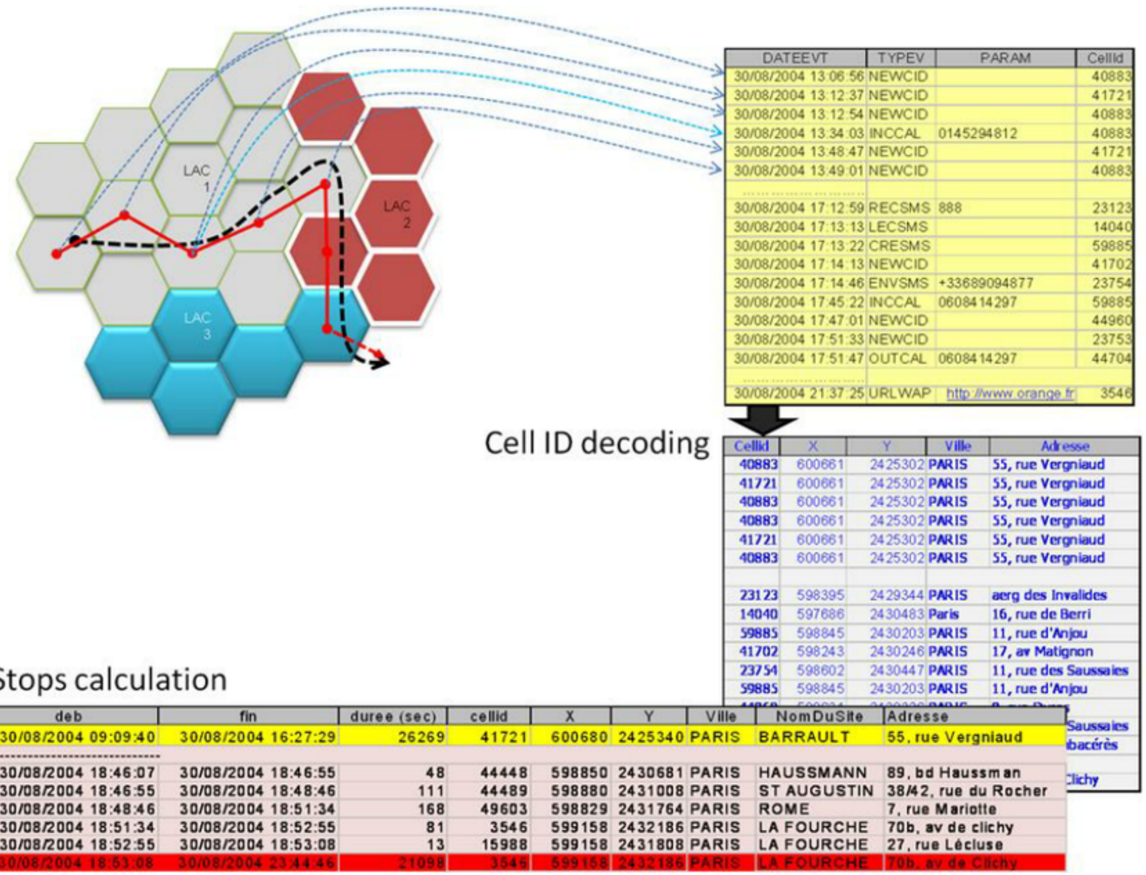
Cellular network must know the position of all mobile units at all times to facilitate high connection speeds

Mobile phone data collected either actively or passively

Active data collection involves provoking production of localisation information for a specific mobile device (e.g. through software or a GPS device included in the phone)

Can be complemented with semantic information

Generally requires agreement of mobile unit owner to participate



DATEEVT	TYPEV	PARAM	CellId
30/08/2004 13:06:56	NEWCID		40883
30/08/2004 13:12:37	NEWCID		41721
30/08/2004 13:12:54	NEWCID		40883
30/08/2004 13:34:03	INCCAL	0145294812	40883
30/08/2004 13:48:47	NEWCID		41721
30/08/2004 13:49:01	NEWCID		40883
...			
30/08/2004 17:12:59	RECSMS	888	23123
30/08/2004 17:13:13	LECSMS		14040
30/08/2004 17:13:22	CRESMS		59885
30/08/2004 17:14:13	NEWCID		41702
30/08/2004 17:14:46	ENVSMS	+33689094877	23754
30/08/2004 17:45:22	INCCAL	0608414297	59885
30/08/2004 17:47:01	NEWCID		44960
30/08/2004 17:51:33	NEWCID		23753
30/08/2004 17:51:47	OUTCAL	0608414297	44704
...			
30/08/2004 21:37:25	URLWAP	http://www.orange.fr	3546

Cellid	X	Y	Ville	Adresse
40883	600661	2425302	PARIS	55, rue Vergniaud
41721	600661	2425302	PARIS	55, rue Vergniaud
40883	600661	2425302	PARIS	55, rue Vergniaud
40883	600661	2425302	PARIS	55, rue Vergniaud
41721	600661	2425302	PARIS	55, rue Vergniaud
40883	600661	2425302	PARIS	55, rue Vergniaud
23123	598395	2429344	PARIS	aerg des Invalides
14040	597686	2430483	Paris	16, rue de Berri
59885	598845	2430203	PARIS	11, rue d'Anjou
41702	598243	2430246	PARIS	17, av Matignon
23754	598602	2430447	PARIS	11, rue des Saussaies
59885	598845	2430203	PARIS	11, rue d'Anjou

deb	fin	duree (sec)	cellid	X	Y	Ville	NomDuSite	Adresse
30/08/2004 09:09:40	30/08/2004 16:27:29	26269	41721	600680	2425340	PARIS	BARRAULT	55, rue Vergniaud
----------								
30/08/2004 18:46:07	30/08/2004 18:46:55	48	44448	598850	2430681	PARIS	HAUSSMANN	89, bd Haussman
30/08/2004 18:46:55	30/08/2004 18:48:46	111	44489	598880	2431008	PARIS	ST AUGUSTIN	38/42, rue du Rocher
30/08/2004 18:48:46	30/08/2004 18:51:34	168	49603	598829	2431764	PARIS	ROME	7, rue Mariotte
30/08/2004 18:51:34	30/08/2004 18:52:55	81	3546	599158	2432186	PARIS	LA FOURCHE	70b, av de clichy
30/08/2004 18:52:55	30/08/2004 18:53:08	13	15988	599158	2431808	PARIS	LA FOURCHE	27, rue Lécluse
30/08/2004 18:53:08	30/08/2004 23:44:46	21098	3546	599158	2432186	PARIS	LA FOURCHE	70b, av de Clichy

Data collection & decoding using cell tracing on a mobile phone (dotted line real trajectory, solid line = cell change based trajectory

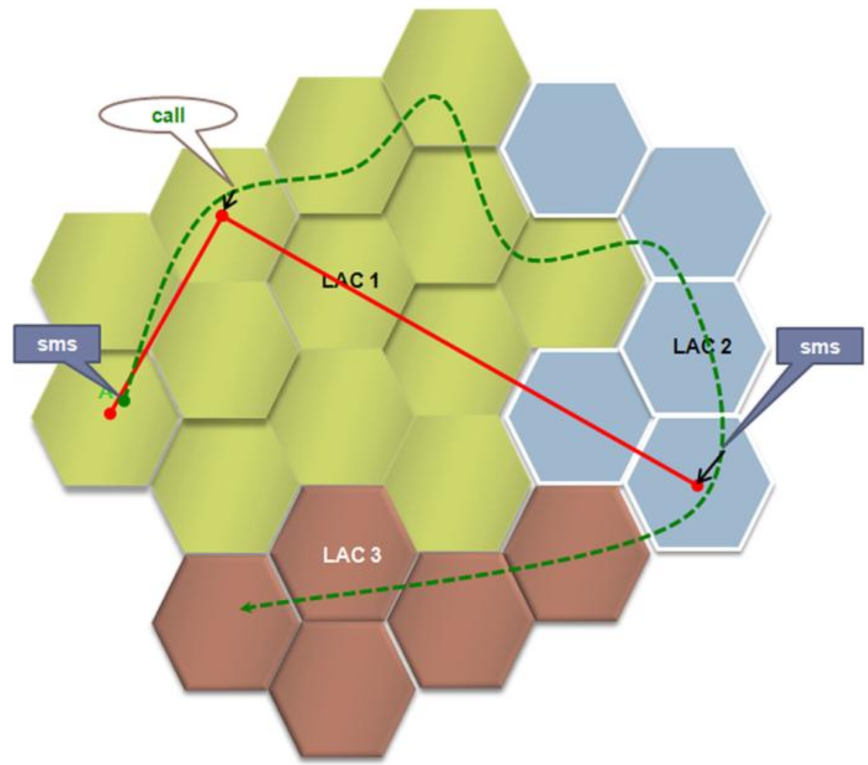
Passive data collection utilises billing data, or Call Detail Records

Mobile phone operators collect a large mass of data for billing purposes/system management

Mass of data available

Main draw-back is the lack of semantic information and the need for validation

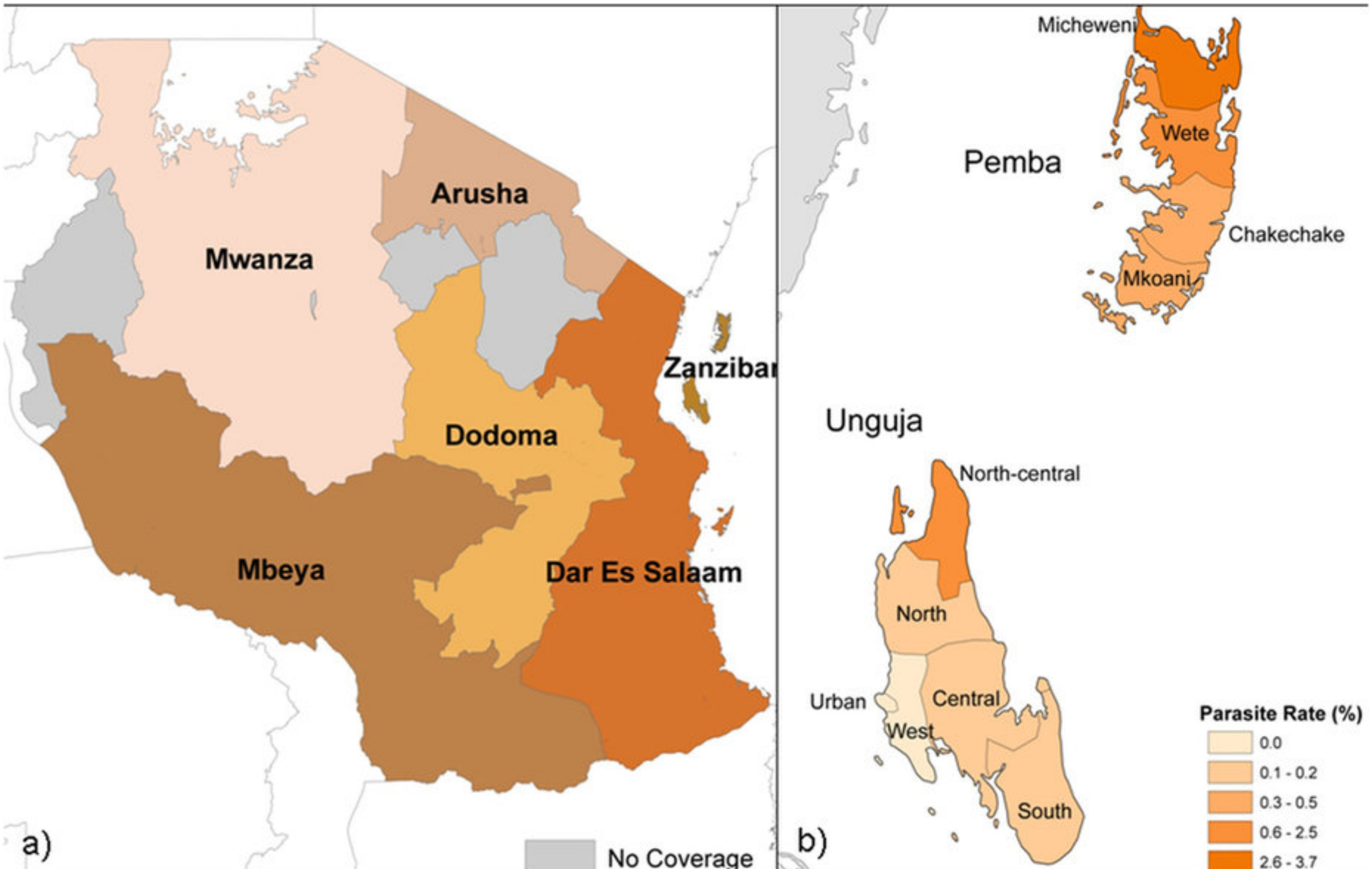
Also questions re privacy



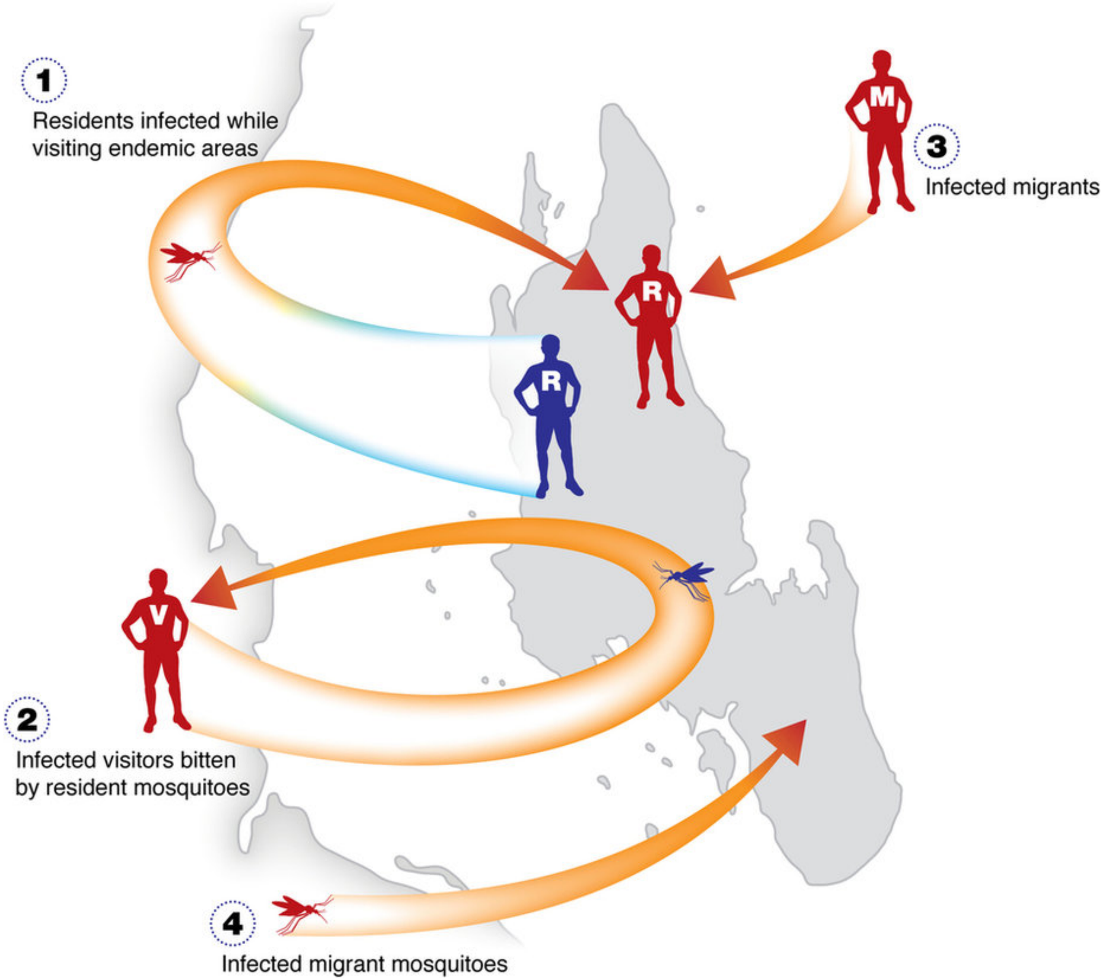
Location information using Call Detail Records (dotted line = real trajectory, solid line = route based on CDR (2 SMSs and one call)

e.g. malaria in Zanzibar

Malaria infections in Zanzibar largely result from malaria imported from parts of mainland Tanzania where malaria is endemic and subsequent transmission.



From Le Menach et al. (2011) Travel risk, malaria importation and malaria transmission in Zanzibar *Scientific Reports 1:93*



Mobile phone usage data used to refine models of importation of *Plasmodium falciparum* malaria to Unguja – help quantify risk from infectious visitors and returnees.

Most malaria on Unguja is from returning islanders, rather than visitors to the island.

Improved malaria control measures based as a result contribute towards elimination of malaria

Big data have the potential to:

- Augment existing surveillance systems
- **Provide an early warning of a disease outbreak**
- Provide a basis for research on health and related matters

Early warning of disease outbreaks through tracking consumer behaviour

Simple idea = human behaviour as reflected in our data shadows can be used as an early warning of disease outbreak and to track the spread of illness

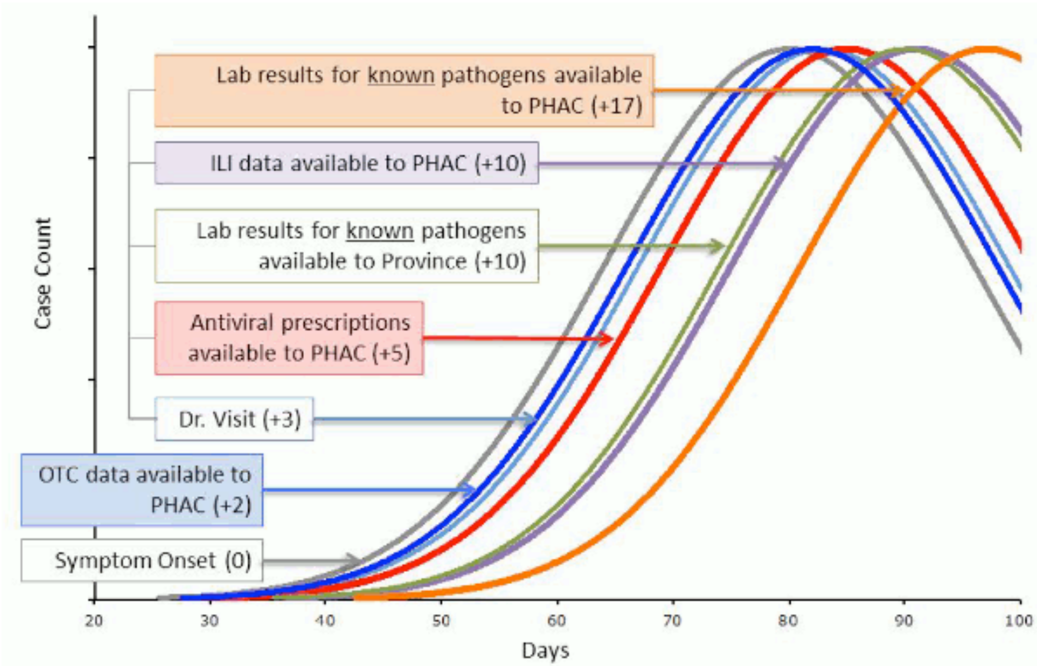


Figure depicts the timelines of pharmacy, clinical and laboratory data relative to the estimated onset of illness and the availability of the information for the purpose of respiratory surveillance.
From Muchaal et al. (2015) *Evaluation of a national pharmacy-based syndromic surveillance system* CCDR online 41 (9)

Data on the purchasing of drugs from pharmacies available several days/weeks before hospital data on positive disease tests available

Simple idea – but fraught with difficulties (e.g. agreement over definition of terms such as fever, malaria, dengue, uncertainties in self-diagnosing etc)

Big data have the potential to:

- Augment existing surveillance systems
- Provide an early warning of a disease outbreak
- **Provide a basis for research on health and related matters**

Infodemiology in an age of Twitter

Social media platforms (SMPs) mean that the public is no longer a passive recipient of information

Public now play a larger role in knowledge translation, including information generation, filtering and amplification

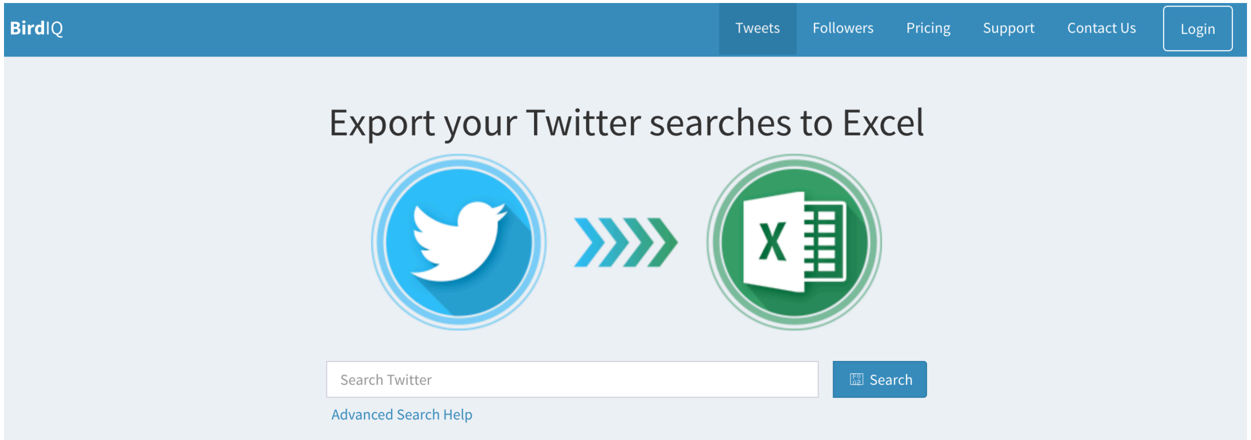
Public health authorities can use information from SMPs to monitor public perceptions of and responses to health risks, and the effectiveness/ penetration of health campaigns

E.g, public health researchers use Twitter – established in 2006 – to interact with the public and to mine the platform for data

Taxonomy of use of twitter generated data in health articles, 2010-2015*

Taxonomy	Description	Articles, No. (%)	Examples
Use of Twitter data			
Content analysis	Assessment of body of tweets for themes in relation to a specific subject	77 (56)	Smoking, diabetes, obesity, concussion
Sentiment analysis	Assessment of body of tweets for positive or negative discussion of a specific subject	21 (15)	Schizophrenia, vaccination, trans health
Image analysis	Assessment of images within body of tweets for themes in relation to a specific subject	1 (1)	#thinspo
Surveillance	Monitoring of Twitter traffic for mentions of a particular topic above the normal background level of discussion	36 (26)	Influenza, Ebola, adverse drug reactions
Prediction	Using Twitter to estimate prevalence of disease or behavior	7 (5)	Heart disease mortality, influenza infection, Affordable Care Act enrollment, asthma emergency department visits
Engagement	Assessing impact of discussion on Twitter by analyzing presence of an account, number of retweets, favorites, followers, etc.	19 (14)	Nutrition public health marketing campaign, social media impact of local health departments, social media adoption by pharmaceutical companies
Network analysis	Assessing the relationship and interactions between Twitter users about a certain topic	5 (4)	Communities of cancer patients, sharing of health information by health organizations

*See Sinneberg et al. (2017) Twitter as a tool for health research: A systematic review. *AJPH* 107, e1-8.



https://birdiq.net/twitter-search

Application Programme Interface (API)s such as BirdIQ provide a means of downloading information on tweets that can then be analysed (location, content etc).

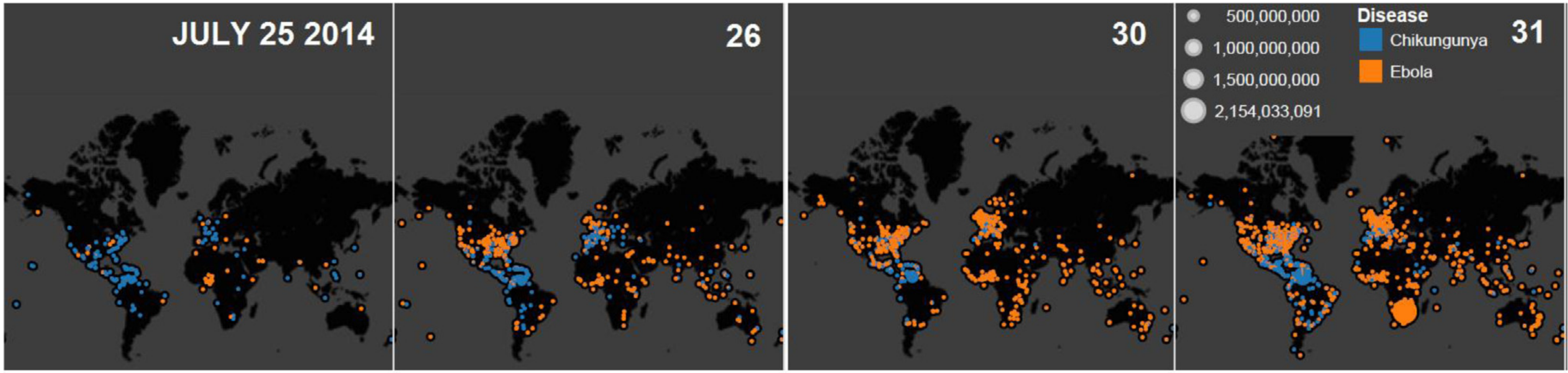
BirdIQ allows information to be exported as an EXCEL file

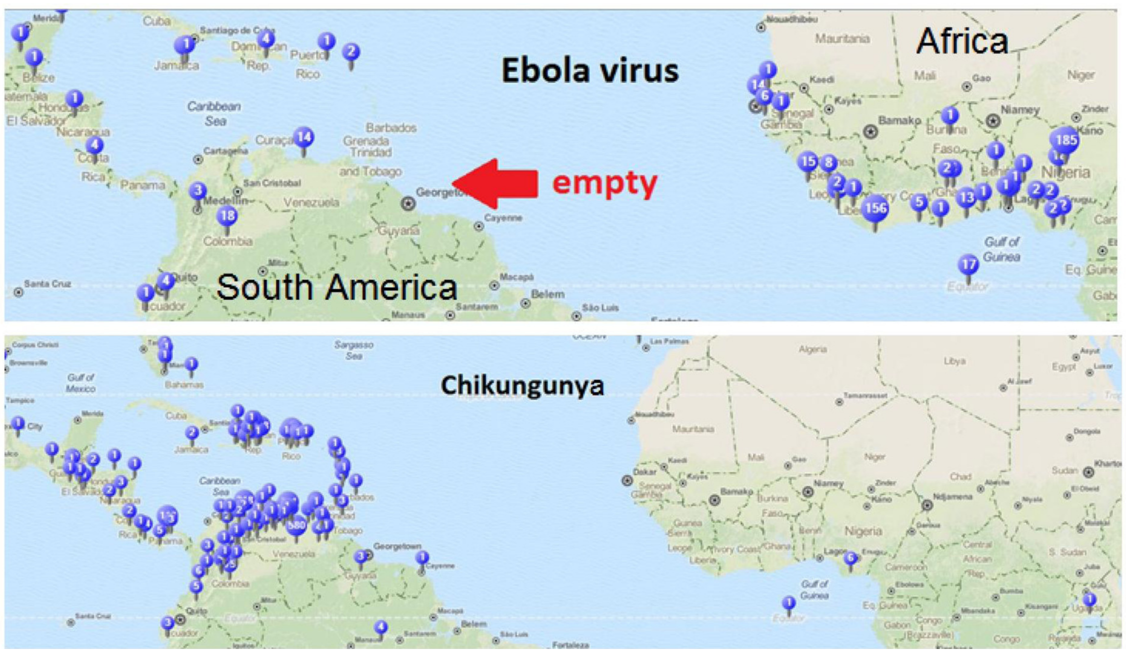
NCapture exports social media information to Nvivo
http://www.qsrinternational.com/support/faqs/what-is-ncapture

Other online tools – such as the One Million Tweet Map – maps the tweets for a particular hashtag or keyword

http://onemilliontweetmap.com/

e.g. Ebola Virus Disease outbreak, 2014



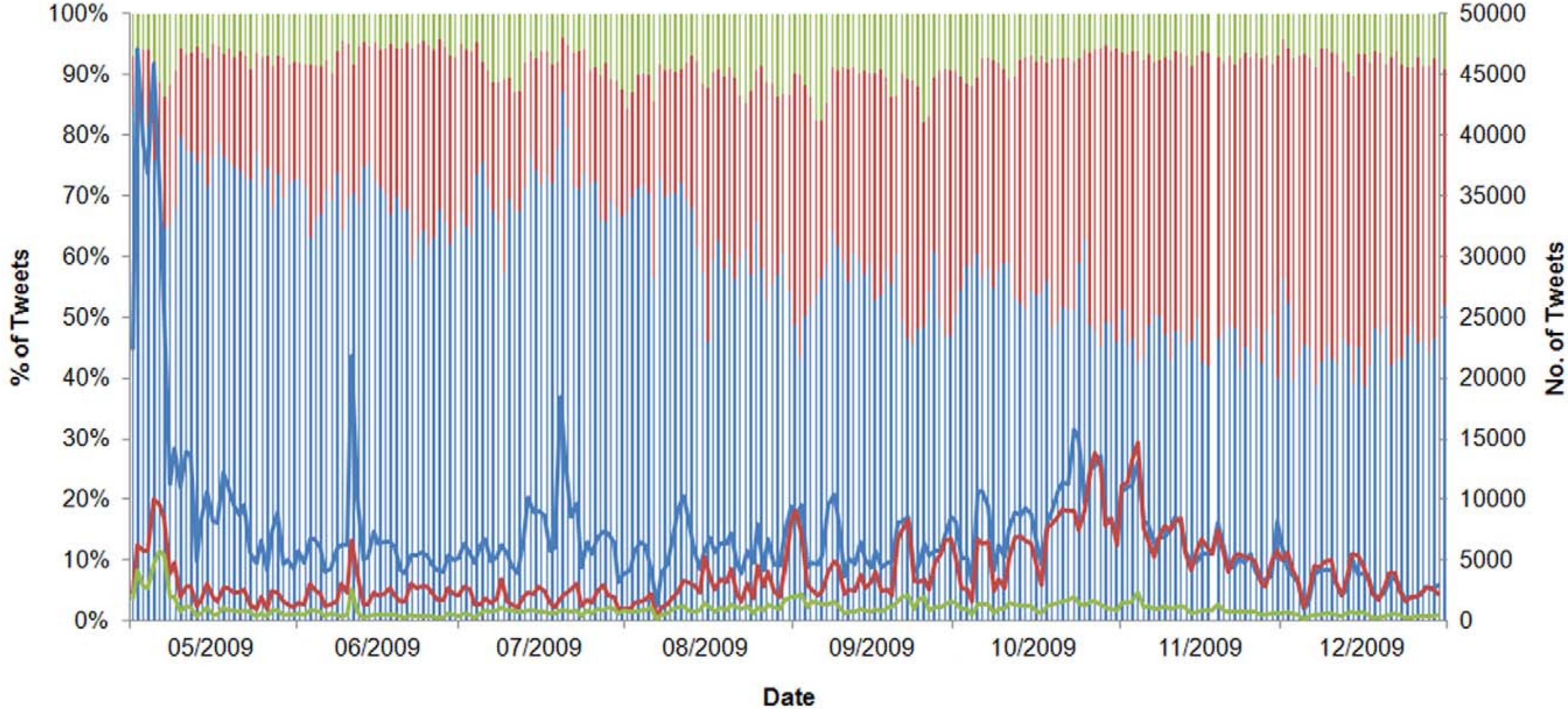


Daily geographic spread of tweets mentioning Ebola Virus Disease (EVD).

Following July 29 2014 health advisory announcement by CNC, US – awareness of EVD spread, but possibly at expense of attention to Chikungunya outbreak in Caribbean everywhere but in Caribbean

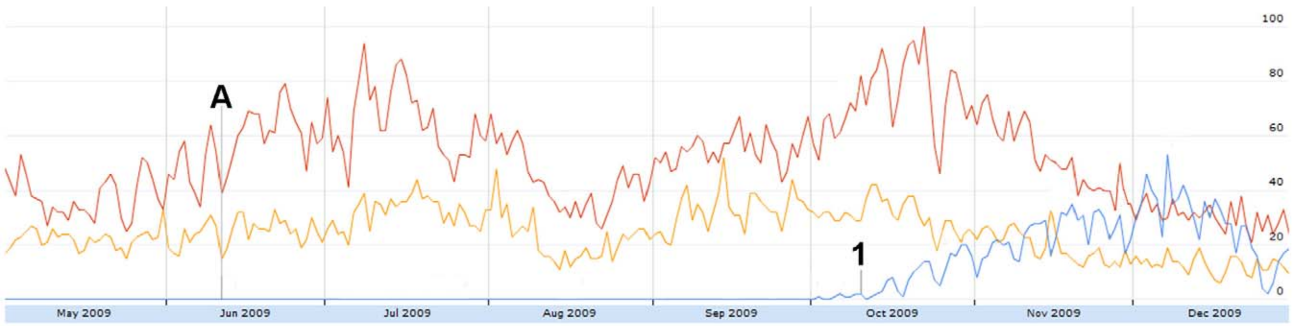
From Odium & Yoon (2015) What can we learn about the Ebola outbreak from tweets? *AJIC* 563-571

e.g. H1N1 pandemic, 2009

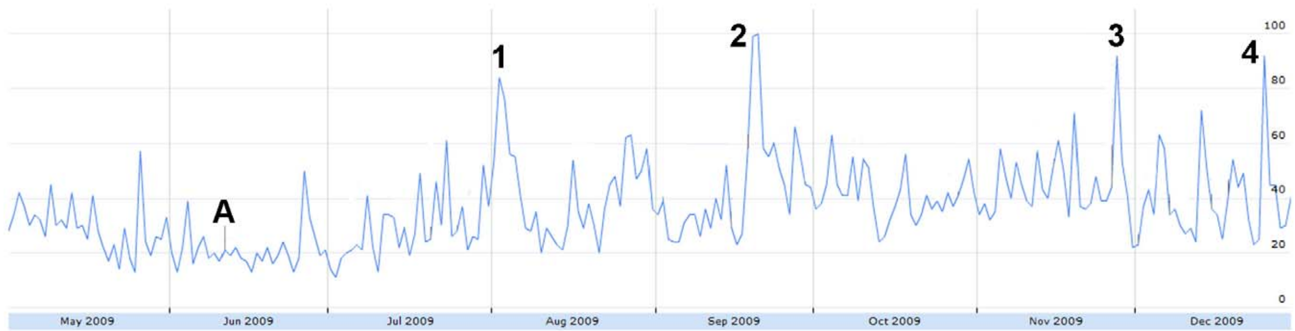


Tweets containing H1N1, swine flu, or both from May to December 2009. Lines = absolute number. Bars = relative percentage. Blue = ‘‘swine flu’’ or swineflu. Red = H1N1. Green = (‘‘swine flu’’ or swineflu) AND H1N1

From: Chew & Eysenbach (2010) Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak *PLoS ONE* 5, e14118



Relative proportion of tweets sharing personal experiences, 1 May-31 December, 2009. Figure was scaled to the highest peak on Oct. 20. Red = indirect (family/friend) experience. Yellow = personal/ direct experience. Blue = vaccination experience. **A** = June 11: WHO pandemic level 6 announcement. **1** = Oct 6: H1N1 vaccinations arrive in the US.



Relative proportion of tweets expressing misinformation, 1 May-31 December, 2009. Figure was scaled to the highest peak on Sept. 20. **A** = June 11: WHO pandemic level 6 announcement. **1** = Aug 2: CBS reports on parental concerns about H1N1. **2** = Sept 18–21: Ten swine flu lies told by the mainstream media. **3** = Nov 27: WHO and drug companies in collusion. **4** = Dec 25: Carbon monoxide poisoning can create same symptoms as H1N1

Quick summary

(Re)-emergence & spread of (existing and) new infectious diseases occur at a high rate

facilitated by human activity, including climate change impacts

Rate means that conventional surveillance methods – where they exist – are not fit for purpose

Big data offer opportunities to health researchers/practitioners

But not straightforward, as big data

are not the same as total data

may not be representative

require new forms of analysis – current statistical techniques not designed to handle the variety, volume and velocity of big data?