

Demography, meet Big Data; Big Data, meet Demography: Reflections on the Data-Rich Future of Population Science

Emmanuel Letouzé

Director & Co-Founder, Data-Pop Alliance


UNITED NATIONS EXPERT GROUP MEETING ON STRENGTHENING THE
DEMOGRAPHIC EVIDENCE BASE FOR THE POST-2015 DEVELOPMENT AGENDA
*Session on 'Complementing traditional data sources with alternative
acquisition, analytic and visualization approaches to ensure better utilization
of data for sustainable development'*

United Nations HQ, New York | October 6, 2015




HARVARD
HUMANITARIAN
INITIATIVE






1—What are we talking about?
2—What has been done?
3—What could / should be done?



1—What are we talking about?
2—What has been done?
3—What could / should be done?



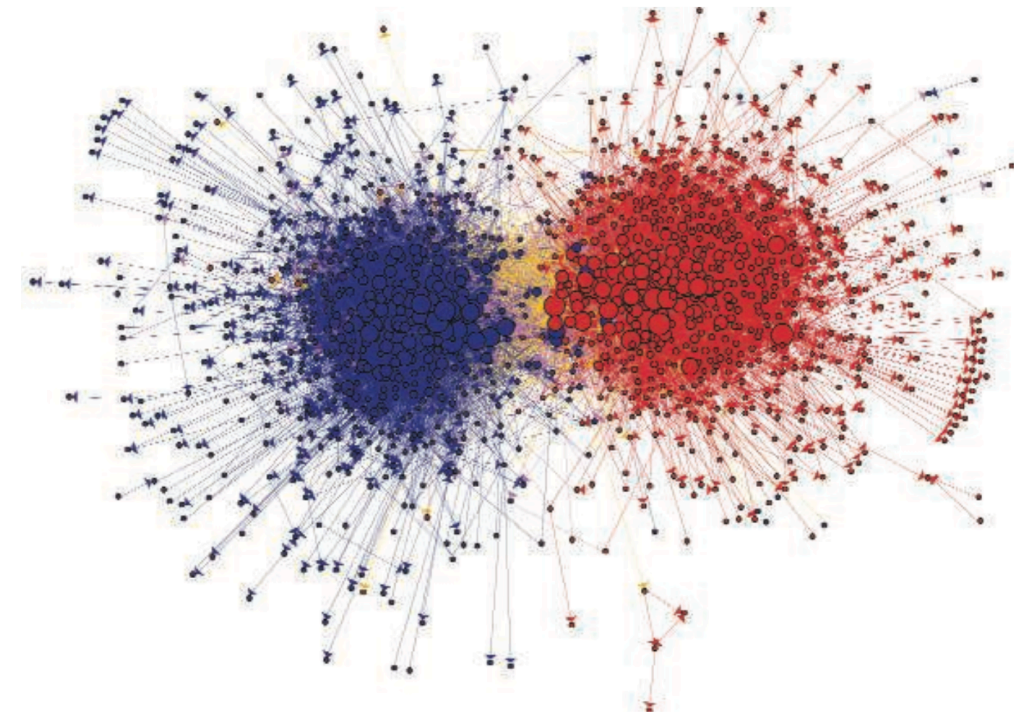
1—What are we talking about?
2—What has been done?
3—What could / should be done?

PERSPECTIVE

Ensuring the Data-Rich Future of the Social Sciences

Gary King

Massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress. I address these changes and challenges and suggest what can be done.



Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (8) with permission from the Association for Computing Machinery]

SOCIAL SCIENCE

Computational Social Science

David Lazer,¹ Alex Pentland,² Lada Adamic,³ Sinan Aral,^{2,4} Albert-László Barabási,⁵
Devon Brewer,⁶ Nicholas Christakis,¹ Noshir Contractor,⁷ James Fowler,⁸ Myron Gutmann,³
Tony Jebara,⁹ Gary King,¹ Michael Macy,¹⁰ Deb Roy,² Marshall Van Alstyne^{2,11}

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

2009

Demography is ^[finally] entering the ‘Data Revolution’ conversation

Demographers Coming Into the Big Tent of the Data Revolution

May 22, 2015 — By Kristen Stelljes



Plenty of room. (Photo Credit: [Steve & Michelle Gerdes](#), licensed under [CC BY 2.0](#))

Monitoring demographic indicators for the post 2015 Sustainable Development Goals (SDGs)

A review of proposed approaches and opportunities

Prepared by **Stephane HELLERINGER**, Johns Hopkins University

4/1/2015

This report reviews the indicators proposed by the Sustainable Development Solutions Network (SDSN) for the post-2015 SDG monitoring period that require access to population data or refer to demographic processes. We make recommendations to strengthen the proposed monitoring framework. The report was conducted as part of the IUSSP's activities related to the post-2015 data revolution with funding from UNFPA.



Annex: Uses of Big Data for SDG monitoring

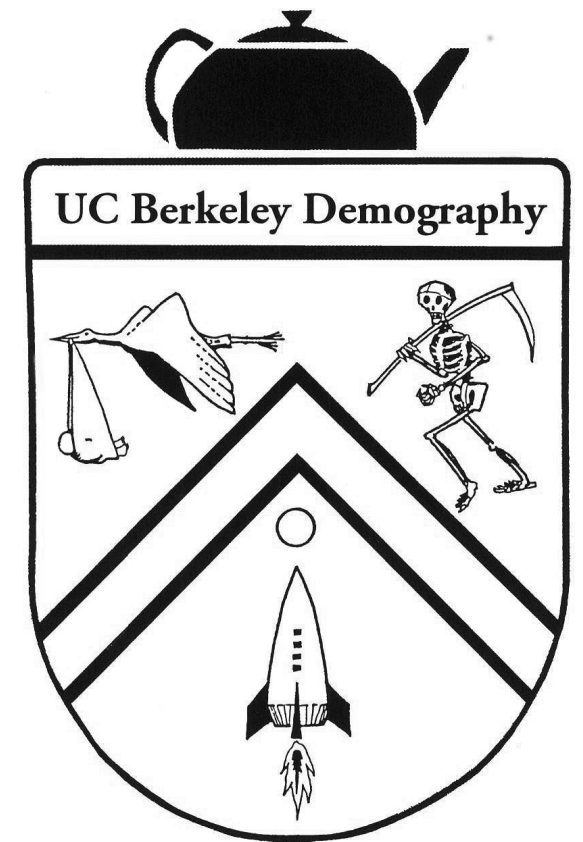
SDGs adopted by the OWG	Big data examples	What is monitored	How is monitored	Country(ies)	Year	Advantages of using big data
1. Poverty eradication	Satellite data to estimate poverty ^{viii}	Poverty	Satellite images, night-lights	Global map	2009	International comparable data, which can be updated more frequently
	Estimating poverty maps with cell-phone records ^{ix}	Poverty	Cell phone records	Cote d'Ivoire	2013-4	
	Internet-based data to estimate consumer price index and poverty rates ^x	Price indexes	Online prices at retailers websites	Argentina	2013	Cheaper data available at higher frequencies
	Cell-phone records to predict socio-economic levels ^{xi}	Socio-economic levels	Cell phone records	“Major city in Latin America” (Actually Mexico-City)	2011	Data available more regularly and cheaper than official data; informal economy better reflected
2. End hunger, achieve food security and improved nutrition, and promote sustainable agriculture	Mining Indonesian Tweets to understand food price crises ^{xii}	Food price crises	Tweets	Indonesia	2014	
	Uses indicators derived from mobile phone data as a proxy for food security indicators ^{xiii}	Food security	Cell phone data and airtime credit purchases	A country in Central Africa	2014	
	Use of remote-sensing data for drought assessment and monitoring	Drought	Remote sensing	Afghanistan, India, Pakistan ^{xiv}	2004	
				China ^{xv}	2008	
3. Health	Internet-based data to identify influenza breakouts ^{xvi}	Influenza	Google search queries	US	2009	Real-time data; captures disease cases not officially recorded; data available earlier than official data
	Data from online searches to monitor influenza epidemics ^{xvii}	Influenza	Online searches data	China	2013	
	Detecting influenza epidemics using twitter ^{xviii}	Influenza	Twitter	Japan	2011	
	Monitoring influenza outbreaks using twitter ^{xix}	Influenza	Twitter	US	2013	
	Systems to monitor the activity of influenza-like-illness with the aid of volunteers via the internet ^{xx,xxi}	Influenza	Voluntary reporting through the internet	Belgium, Italy, Netherlands, Portugal, United Kingdom, United States	ongoing	
	Cell-phone data to model malaria spread ^{xxii}	Malaria	Cell-phone data	Kenya	2012	

Source: Martinho and Letouzé (2015)

First things first...what is demography?



© mathias the dread, Koosinger, suze / photocase.com



Demography—from the Greek **demo**, for people, and **graphy**, for writing, or analysis, or field of study, is ***“the study of changes (such as the number of births, deaths, marriages, and illnesses) that occur over a period of time in human populations”***, or just ***“science of population”***

1. <http://www.merriam-webster.com/dictionary/demography>
2. [p://www.demogr.mpg.de/En/education_career/what_is_demography_1908/default.htm](http://www.demogr.mpg.de/En/education_career/what_is_demography_1908/default.htm)

« DEMOGRAPHY? WHAT'S THAT? »



This is demography

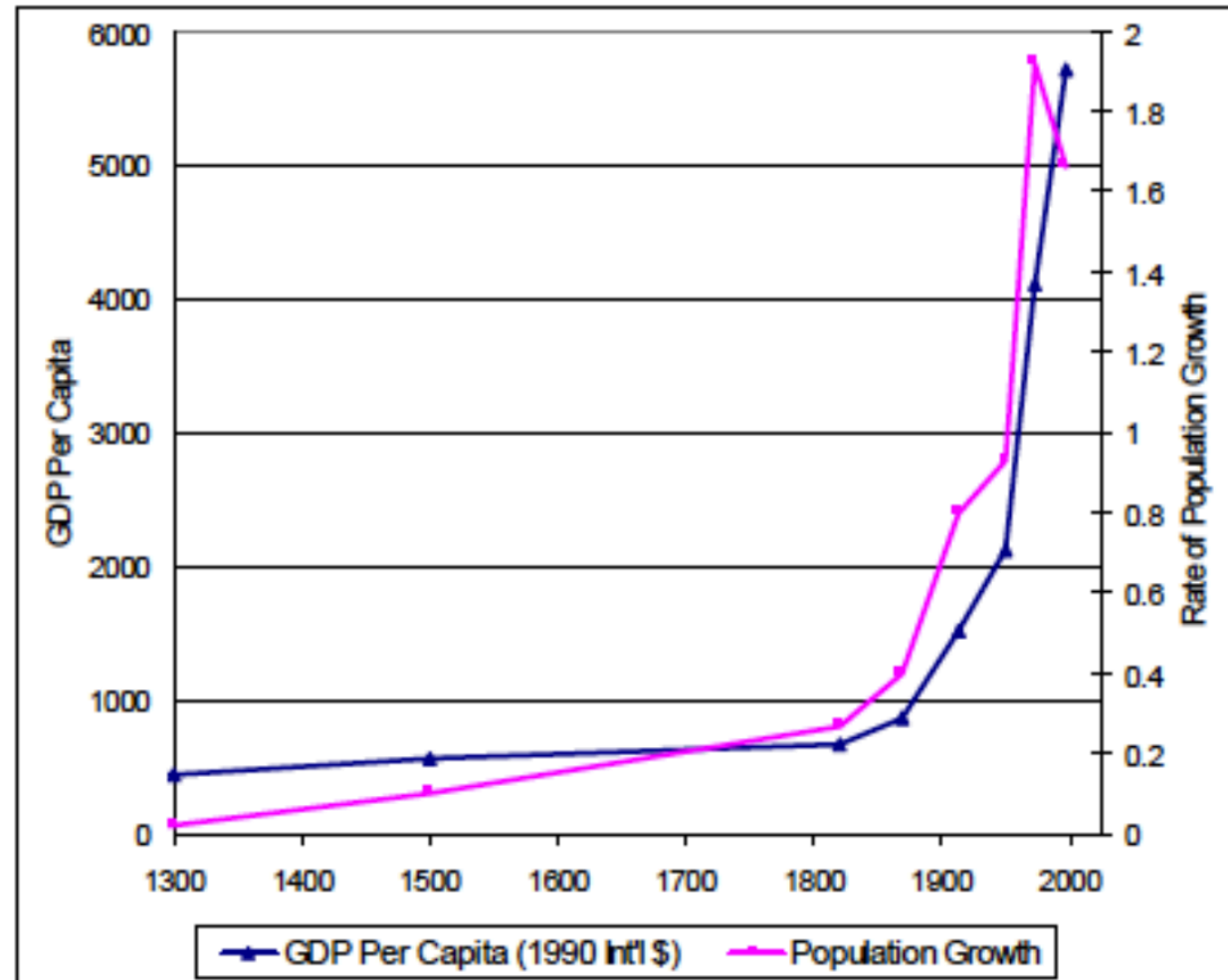
Table 1

Global Population Trends Over the Transition: Estimates, Guesstimates and Forecasts, 1700–2100

	<i>Life Expectancy</i> (Years at Birth)	<i>Total Fertility Rate</i> (Births per Woman)	<i>Pop Size</i> (Billions)	<i>Pop Growth Rate</i> (%/Year)	<i>Pop < 15</i> (% of Total Pop)	<i>Pop > 65</i> (% of Total Pop)
1700	27	6.0	.68	0.50	36	4
1800	27	6.0	.98	0.51	36	4
1900	30	5.2	1.65	0.56	35	4
1950	47	5.0	2.52	1.80	34	5
2000	65	2.7	6.07	1.22	30	7
2050	74	2.0	8.92	0.33	20	16
2100	81	2.0	9.46	0.04	18	21

Source: Lee, 1993

This is demography



Source: Galore PpT: http://www.econ.brown.edu/fac/Oded_Galor/UGT-Dec18-2012-handout.pdf

This is demography

The Slave Trade: Formal Demography 265

262 SOCIAL SCIENCE HISTORY

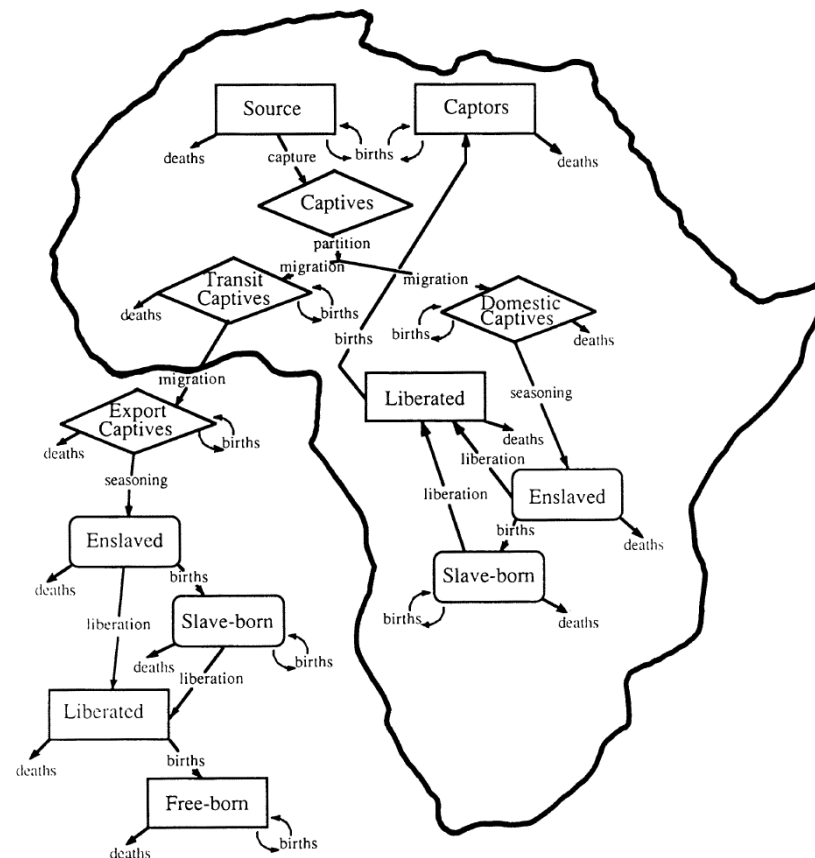


Figure 2 Reproduction and migration in the African and occidental slave trades

Mortality in each age group above birth is calculated as the age-specific survivorship rate multiplied by the previous population. For the Captors:

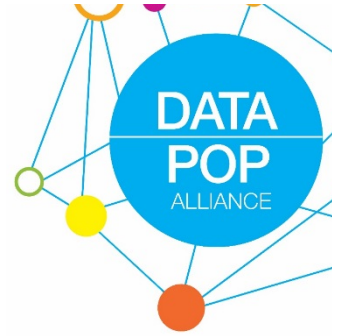
$${}_5N_x^t = {}_5N_{x-5}^{t-5} \times \frac{{}_5L_x}{{}_5L_{x-5}} = {}_5N_{x-5}^{t-5} \times {}_5P_{x-5}$$

For survivors to age 80+, the survival rate is taken as $T80/T75$.⁸ The population in each age group, as estimated for the middle of each five-year period, for Captors, is then

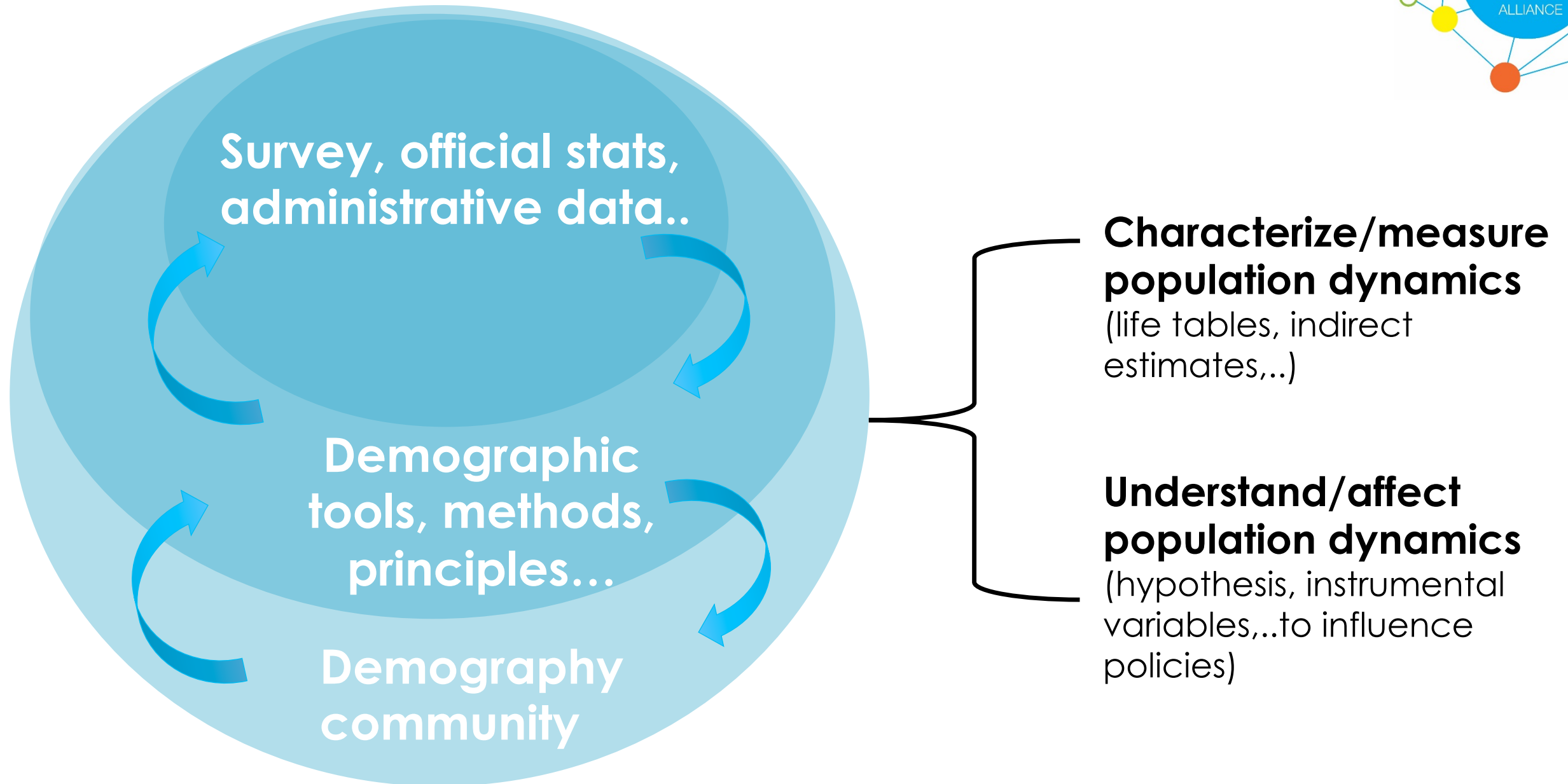
$$\frac{{}_5N_x^{t-5} + {}_5N_{x-5}^{t-5} \times {}_5P_{x-5}}{2} = \frac{{}_5N_x^{t-5} + {}_5N_x^t}{2}$$

Female births in each population are calculated as the age-specific annual fertility rate multiplied by the midperiod population, then by five years of exposure, and summed over all childbearing years.⁹

$${}_5B^{f,t} = \sum_{x=\alpha}^{\beta} \left[\frac{({}_5N_x^{f,t-5} + {}_5N_{x-5}^{f,t-5} \times {}_5P_{x-5})}{2} \times {}_5F_x^f \times 5 \right]$$

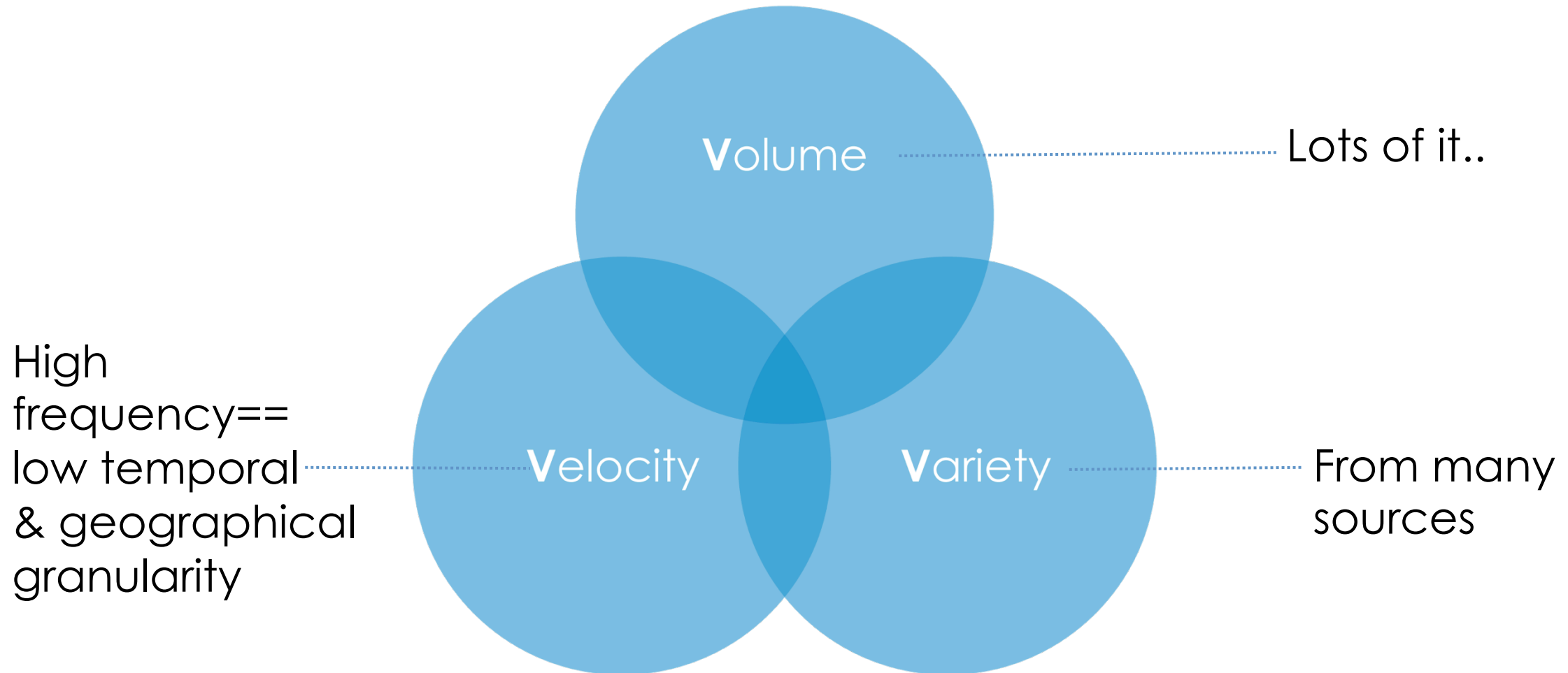
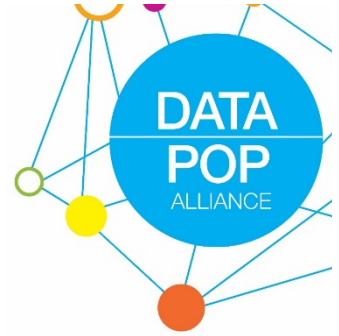


Demography as / is a (vibrant, complex) discipline

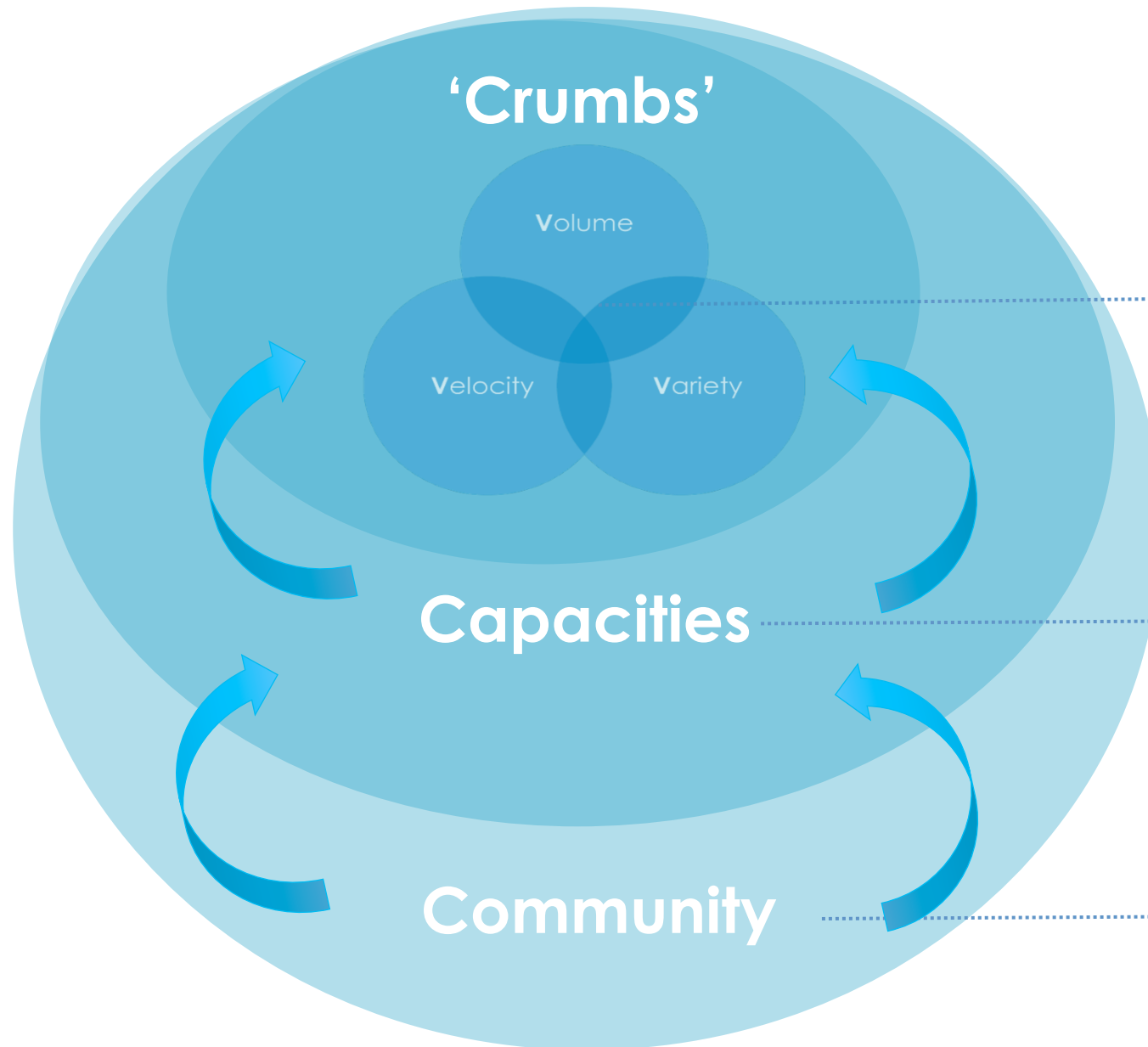
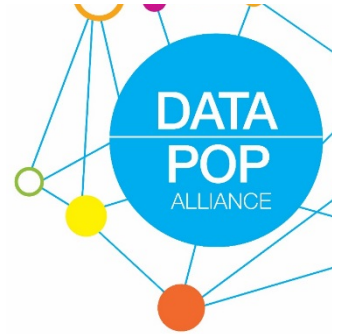


What is Big Data?

From the 3 Vs of big data...



...to the 3 Cs of Big Data *as an ecosystem*



Crumbs

1. Exhaust data
2. Web content
3. Sensing data

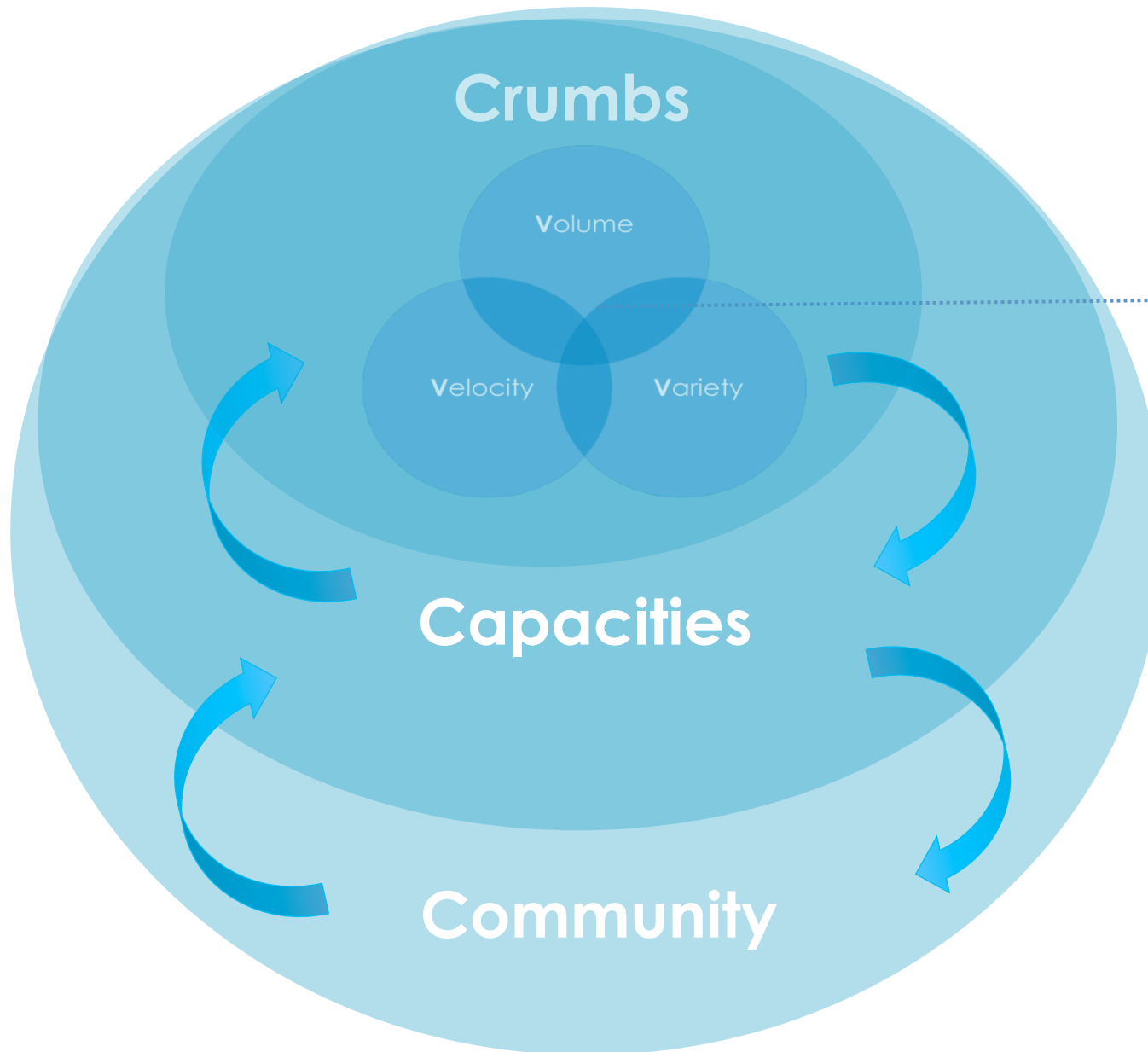
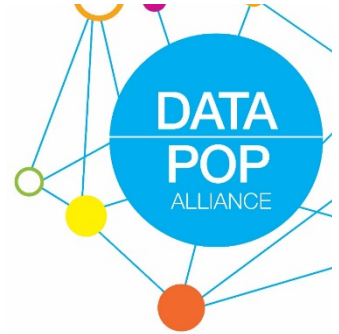
Capacities

1. Soft & hardware
2. Methods & tools
3. Institutional & human

Communities

1. Organizations
2. Objectives
3. Outputs/venues..

...to the 3 Cs of Big Data as an ecosystem



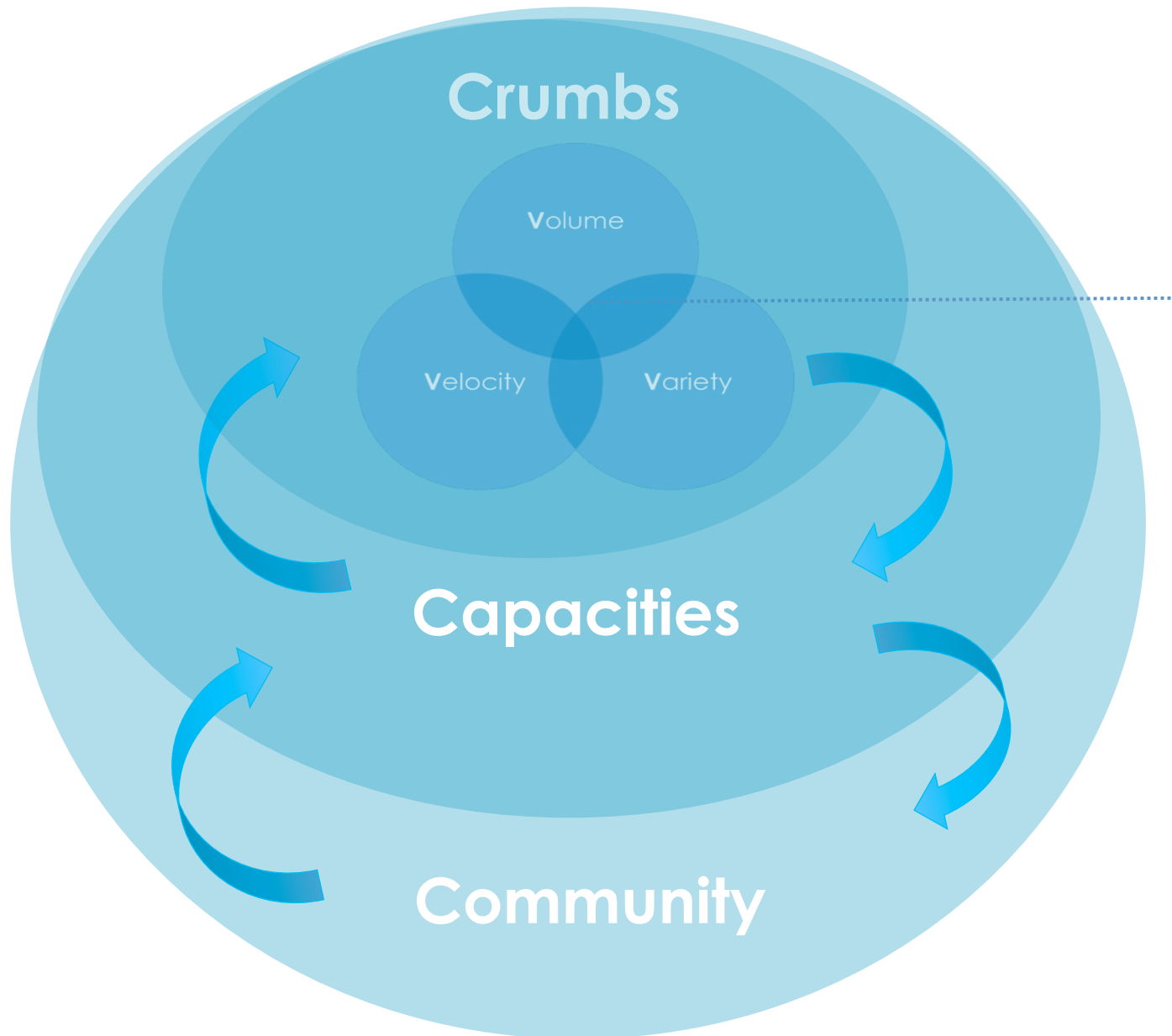
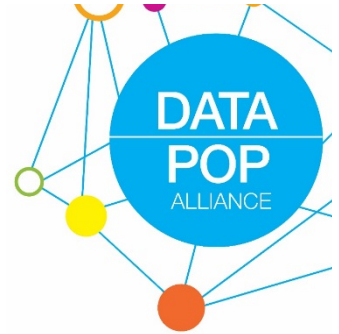
Crumbs

1. Exhaust data
2. Web content
3. Sensing data

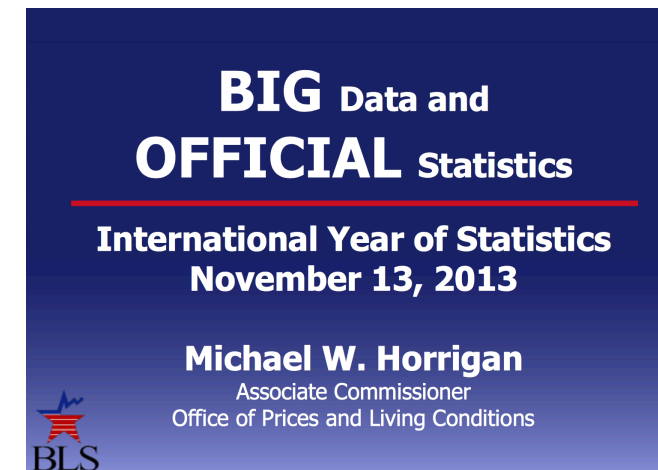
As data, big data is not primarily about size; it is:

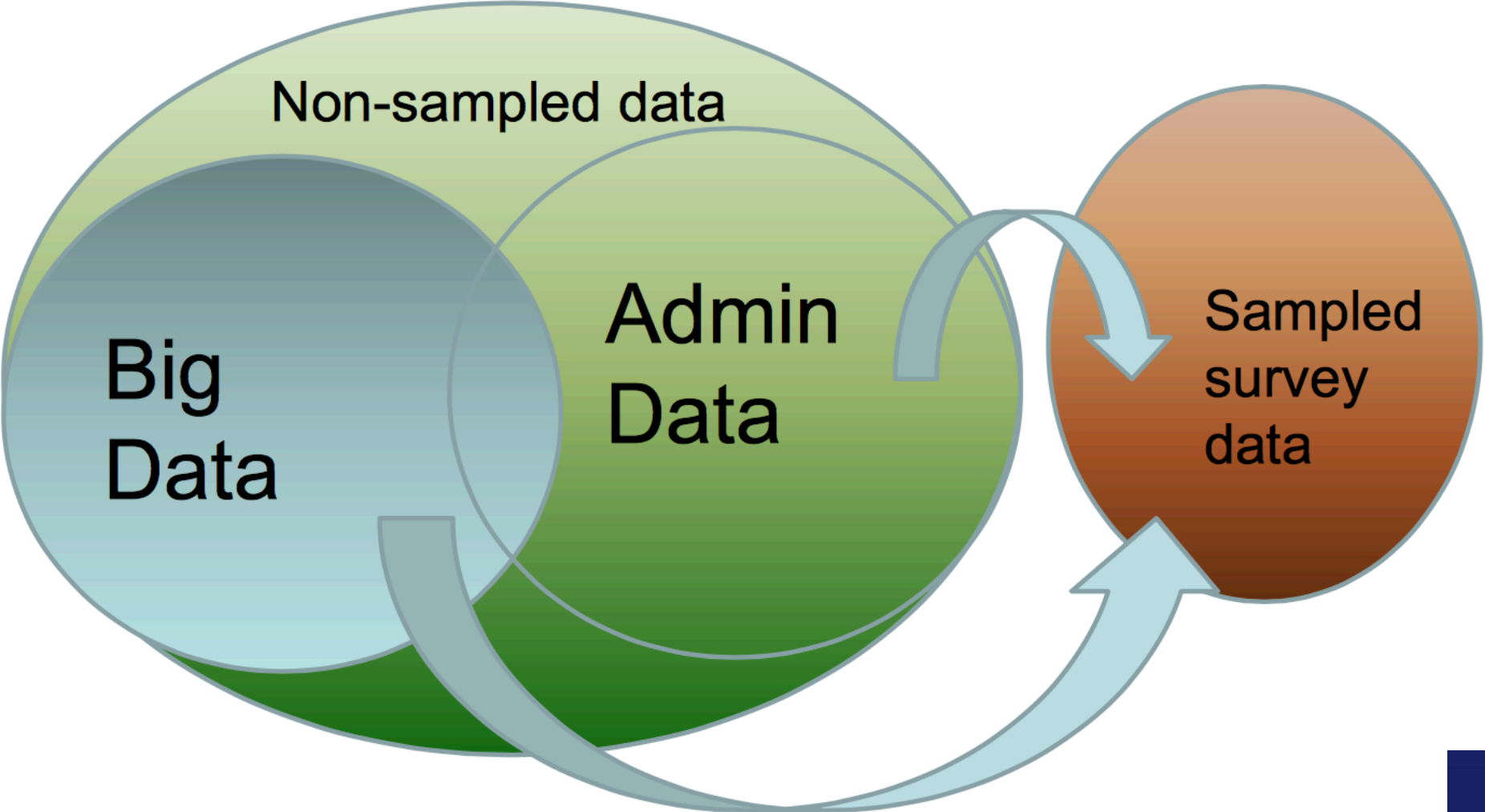
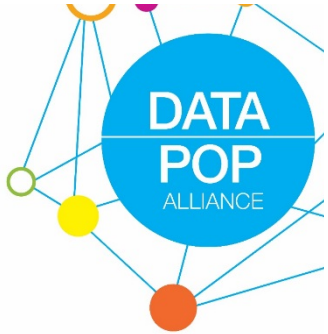
“the digital translation of human behaviors and beliefs passively emitted and/or picked up by digital devices”

...to the 3 Cs of Big Data as an ecosystem



Big data are “**non-sampled** data, characterized by the creation of databases from electronic sources whose **primary purpose is something other than statistical inference.**”
Michael Horrigan, USBLS





**BIG Data and
OFFICIAL Statistics**

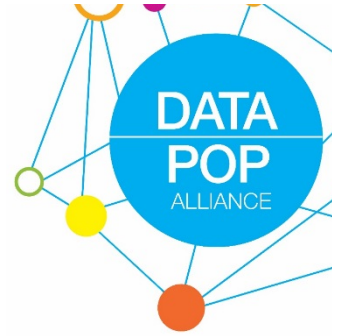
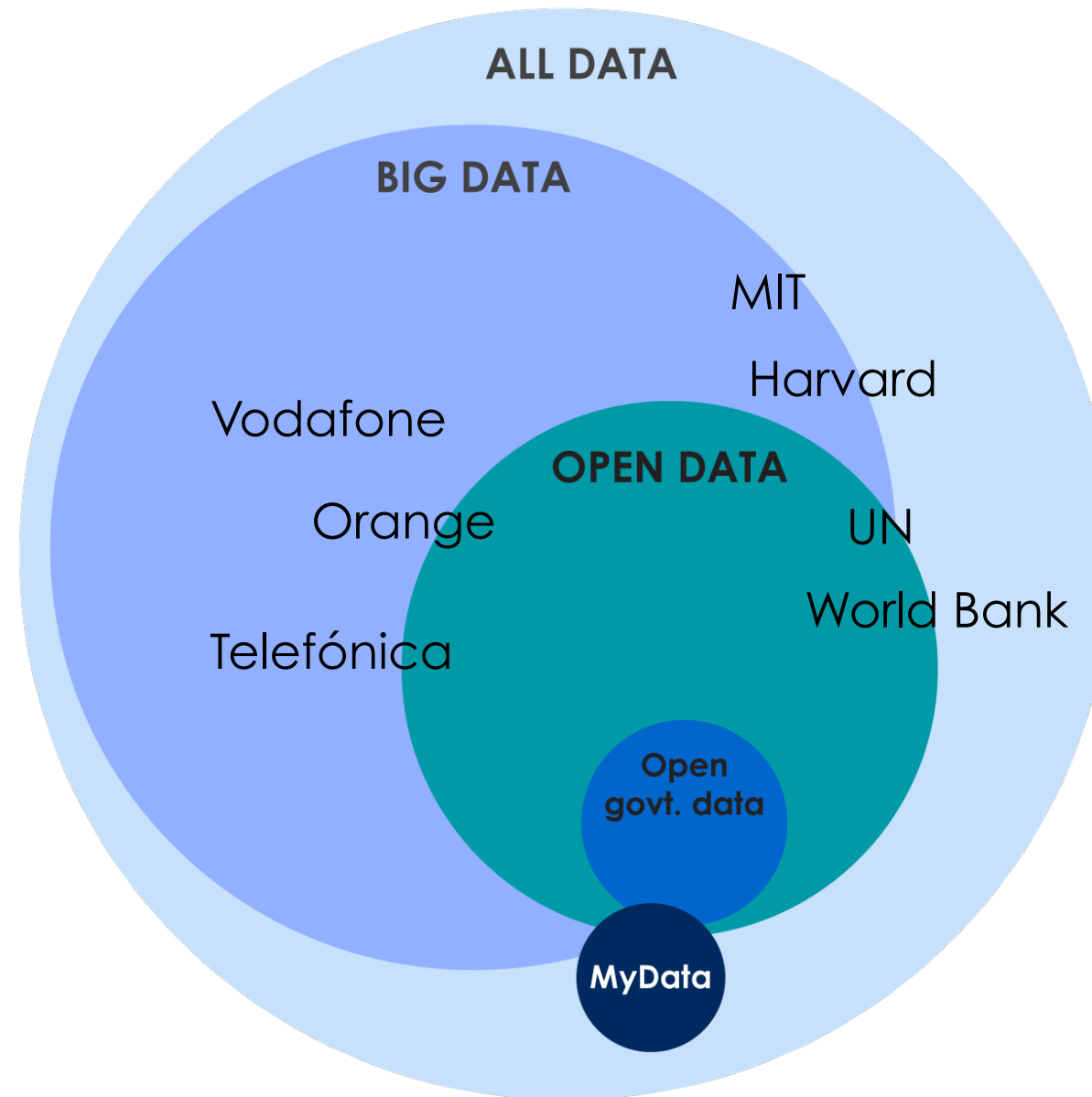
International Year of Statistics
November 13, 2013

Michael W. Horrigan
Associate Commissioner
Office of Prices and Living Conditions

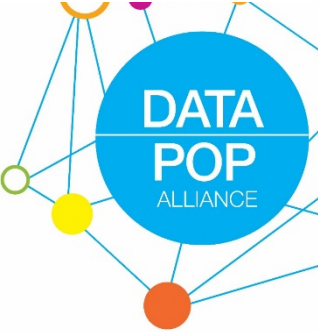
BLS

Source: McKinsey Global Institute

The new data ecosystem

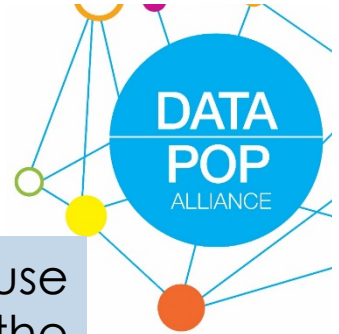


Source: McKinsey Global Institute



1. Exhaust data

Types	Examples	Opportunities
<i>Category 1: Exhaust data</i>		
Mobile-based	Call Details Records (CDRs) GPS (Fleet tracking, Bus AVL)	Estimate population distribution and socioeconomic status in places as diverse as the U.K. and Rwanda
Financial transactions	Electronic ID E-licenses (e.g. insurance) Transportation cards (including airplane fidelity cards) Credit/debit cards	Provide critical information on population movements and behavioural response after a disaster
Transportation	GPS (Fleet tracking, Bus AVL) EZ passes	Provide early assessment of damage caused by hurricanes and earthquakes
Online traces	Cookies IP addresses	Mitigate impacts of infectious diseases through more timely monitoring using access logs from the online encyclopedia Wikipedia



1. Exhaust data—the example of CDRs

Called Detail records (CDRs) are metadata (data about data) that capture subscribers' use of their cell-phones — including an identification code and, at a minimum, the location of the phone tower that routed the call for both caller and receiver — and the time and duration of call. Large operators collect over six billion CDRs per day.

CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24' 22.14", 35°49' 56.54"	A81UTC93KK52	3°26' 30.47", 31°12' 18.01"	2013-11- 07T15:15:00	01:12:02

Source: http://www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev

Note: these are **structured** data==**answers** actively sought by the collector...but the emitter emits them passively; i.e. as a by-product, typically without full knowledge/ informed consent / choice....

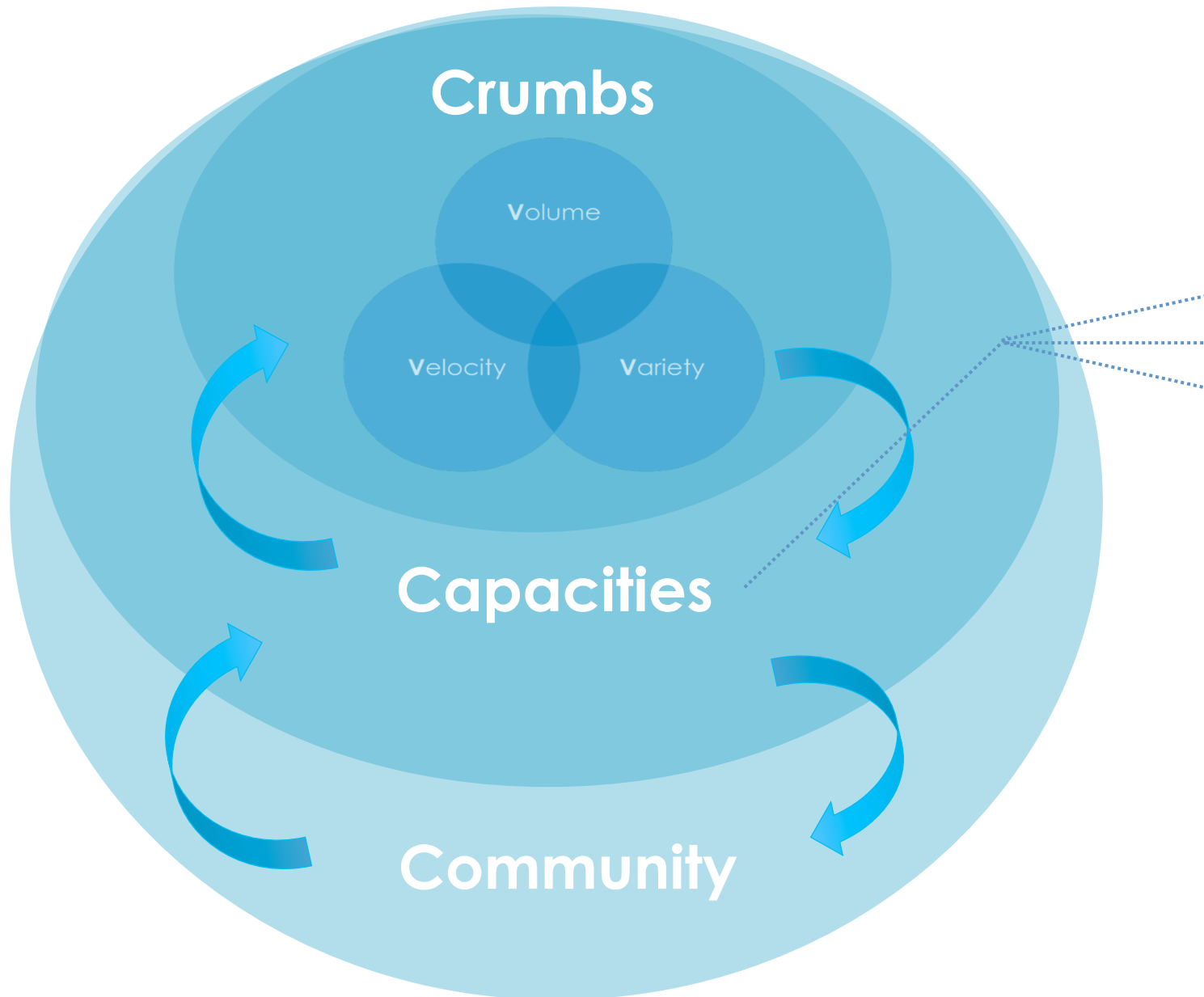
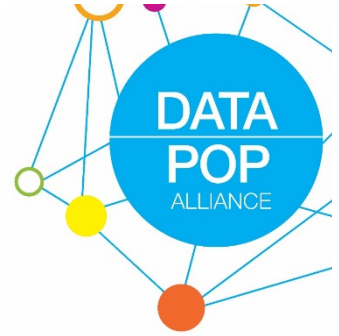


2. Web content

3. Sensing data

Types	Examples	Opportunities
Category 2: Digital Content		
Social media	Tweets (Twitter API) Check-ins (Foursquare) Facebook content YouTube videos	Provide early warning on threats ranging from disease outbreaks to food insecurity
Crowd-sourced/ online content	Mapping (Open Street Map, Google Maps, Yelp) Monitoring/ Reporting (uReport)	Empower volunteers to add ground-level data that are useful notably for verification purpose
Category 3: Sensing data		
Physical	Smart meters Speed/weight trackers USGS seismometers	Sensors have been used to assess the demand for using sensors to estimate demand for high efficiency cook-stoves at different price points in Uganda or willingness to pay for chlorine dispensers in Kenya
Remote	Satellite imagery (NASA TRMM, LandSat) Unmanned Aerial Vehicles (UAVs)	Satellite images revealing changes in, for example, soil quality or water availability have been used to inform agricultural interventions in developing countries

...to the 3 Cs of Big Data as an ecosystem



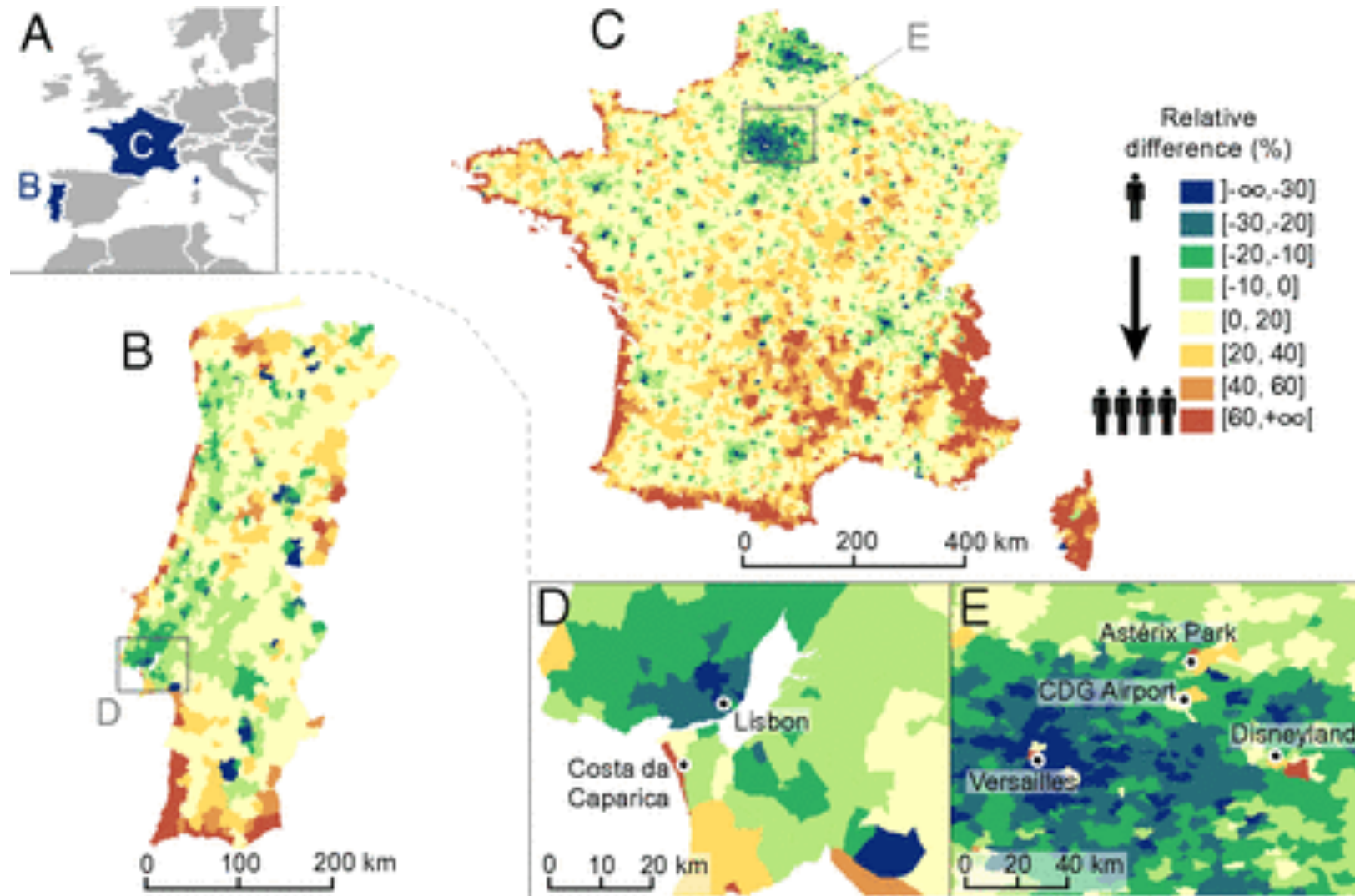
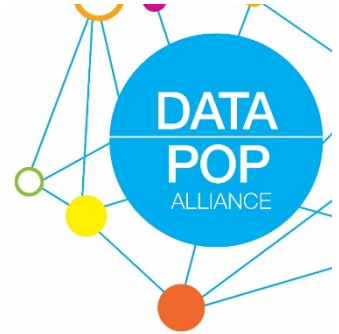
Capacities

1. Soft & hardware
2. Methods & tools
3. Institutional & human

- **Machine-learning**
- **Statistical machine learning**
- **New measures/concepts**
e.g. radius of gyration, entropy....
- **Visualizations...**

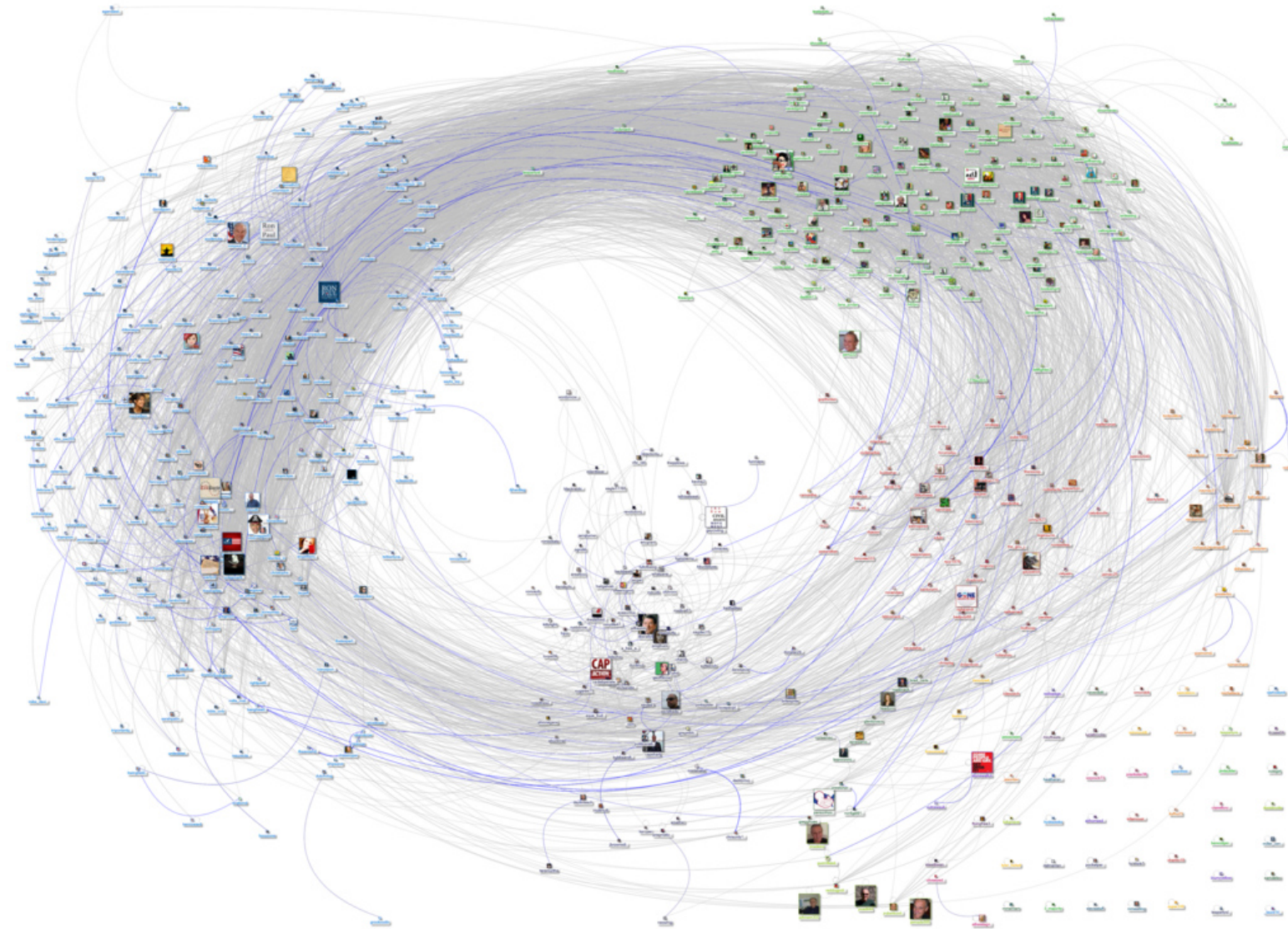
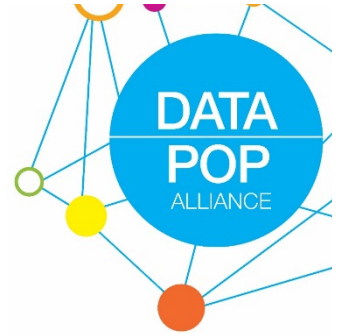
Example: Dynamic Population Mapping Using Mobile Phone Data

France and Portugal (2014)



Source: Deville, Linard et al (2014), PNAS, vol. 111 no. 45. <http://www.pnas.org/content/111/45/15888.abstract>

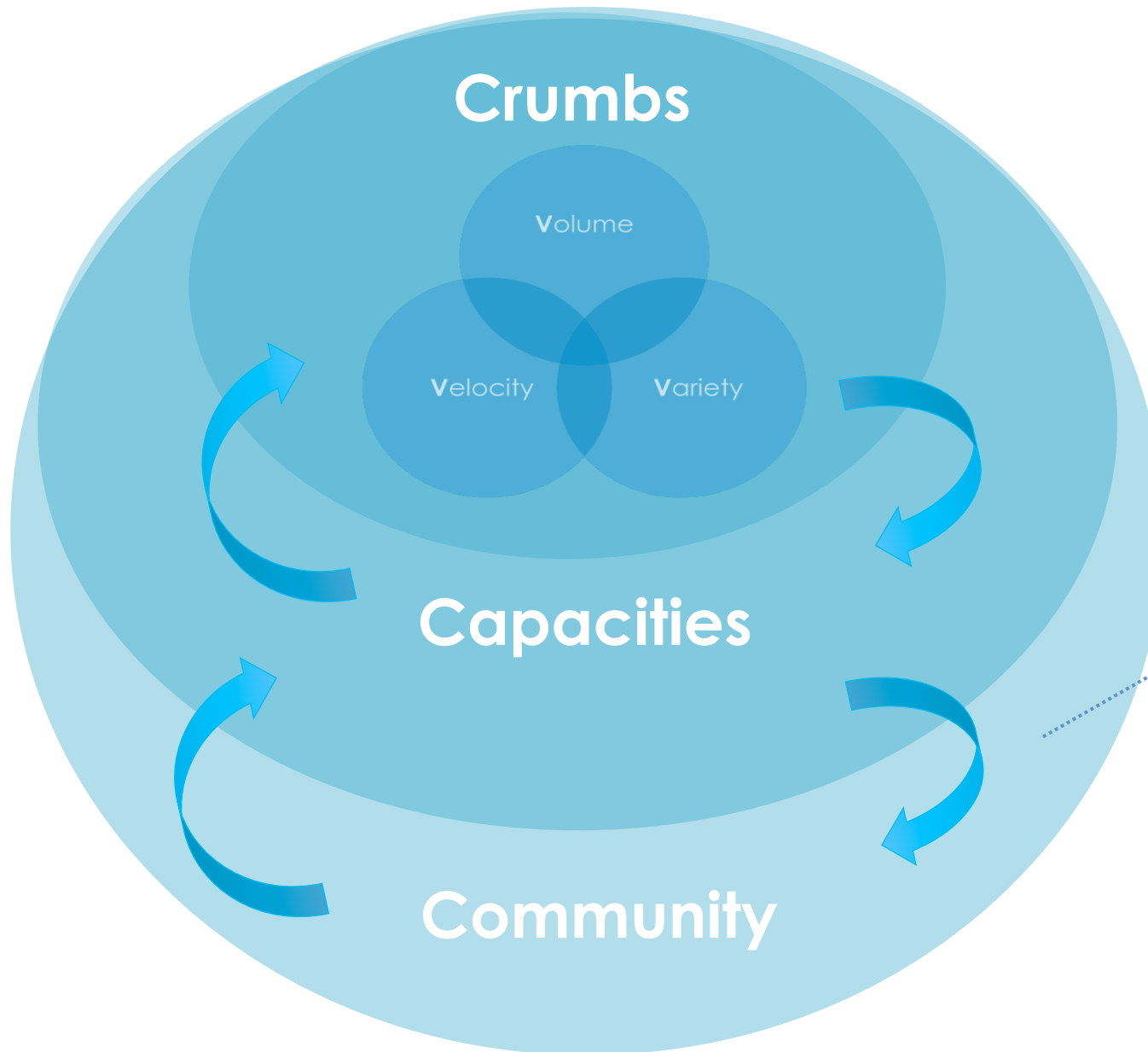
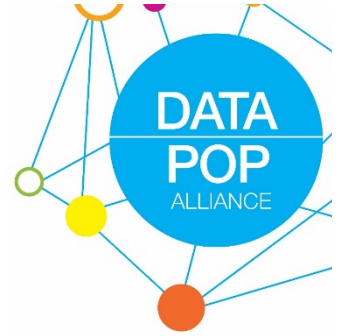
Example: Tea Party vs. Occupy Wall Street on twitter USA (2011)



Created with NodeXL (<http://nodexl.codeplex.com>) from the Social Media Research Foundation (<http://www.smrfoundation.org>)

Not demography? What if pro-life vs. pro-choice?

...to the 3 Cs of Big Data as an ecosystem

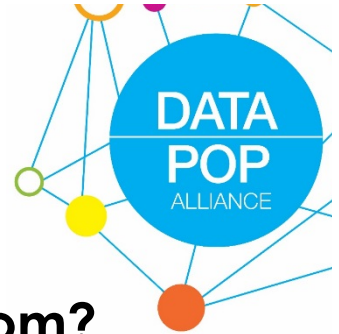
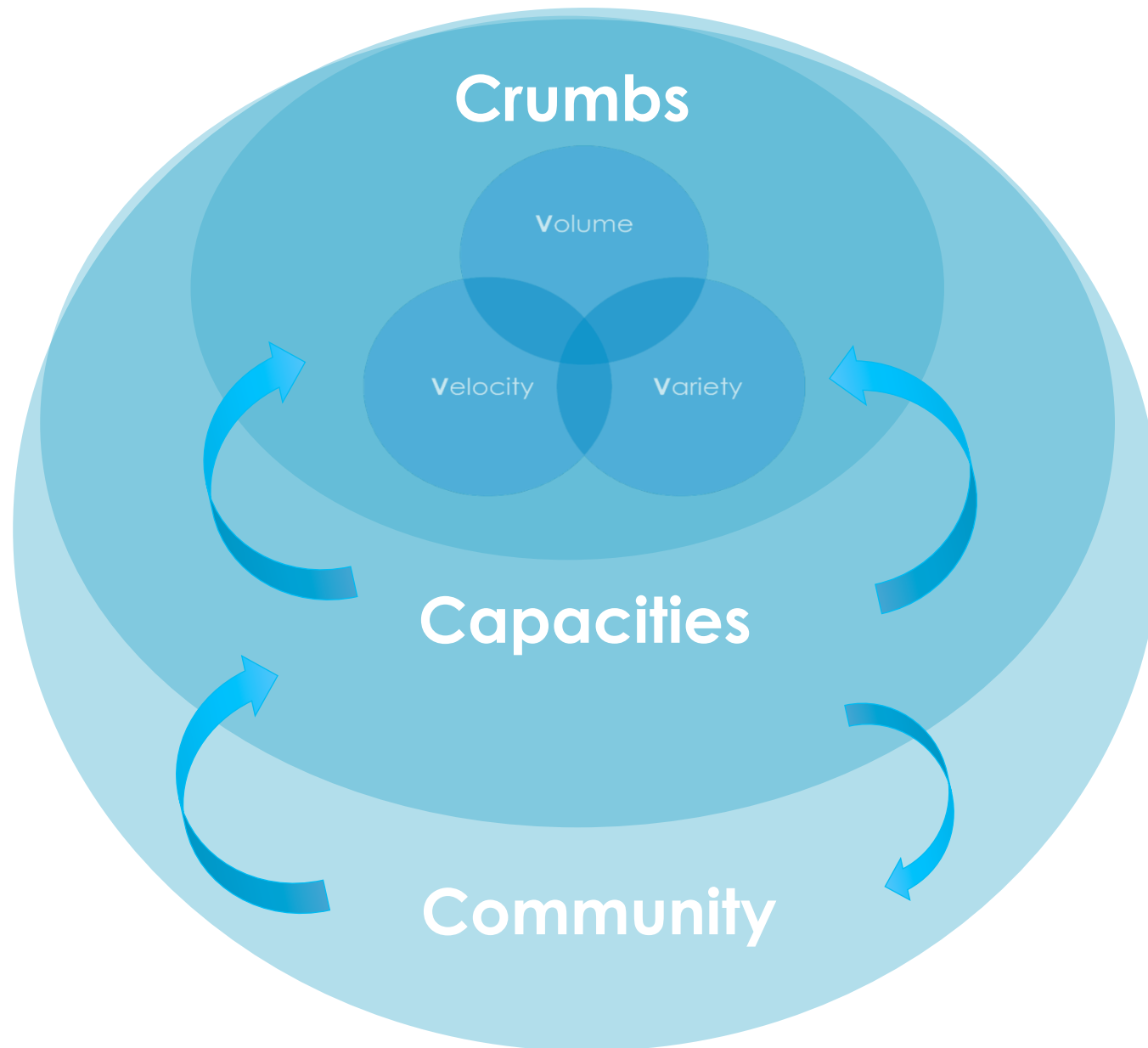


Big Data *communities*:

➔ What for, with and by whom?

- **Machine-learning**
- **Statistical machine learning**
- **New measures/ concepts==radius of gyration, entropy....**
- **Visualizations...**

...to the 3 Cs of Big Data as an ecosystem



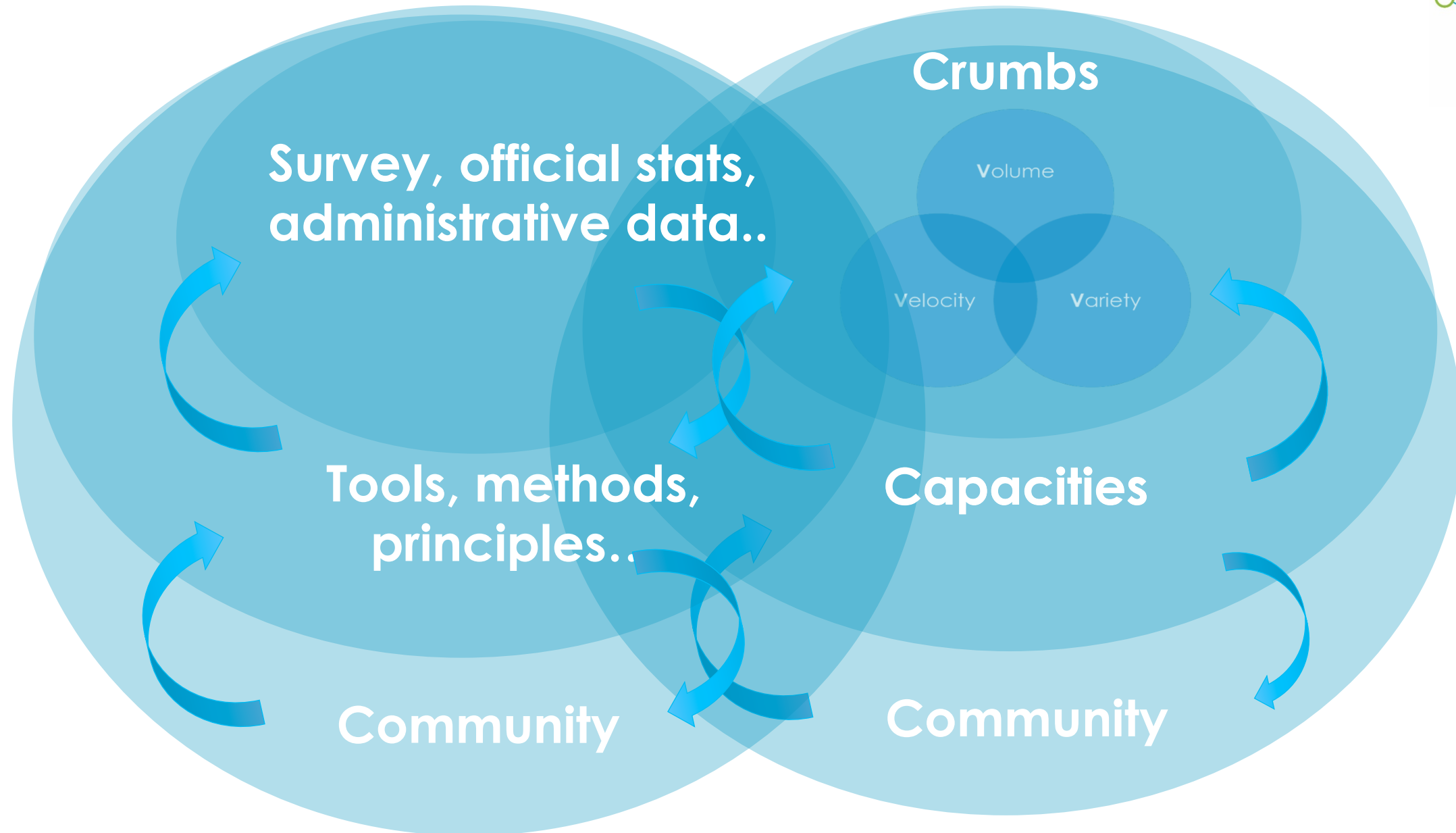
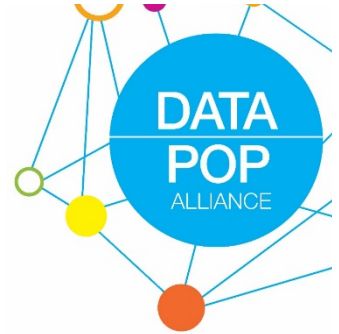
Big Data *community*:

➔ What for, with and by whom?

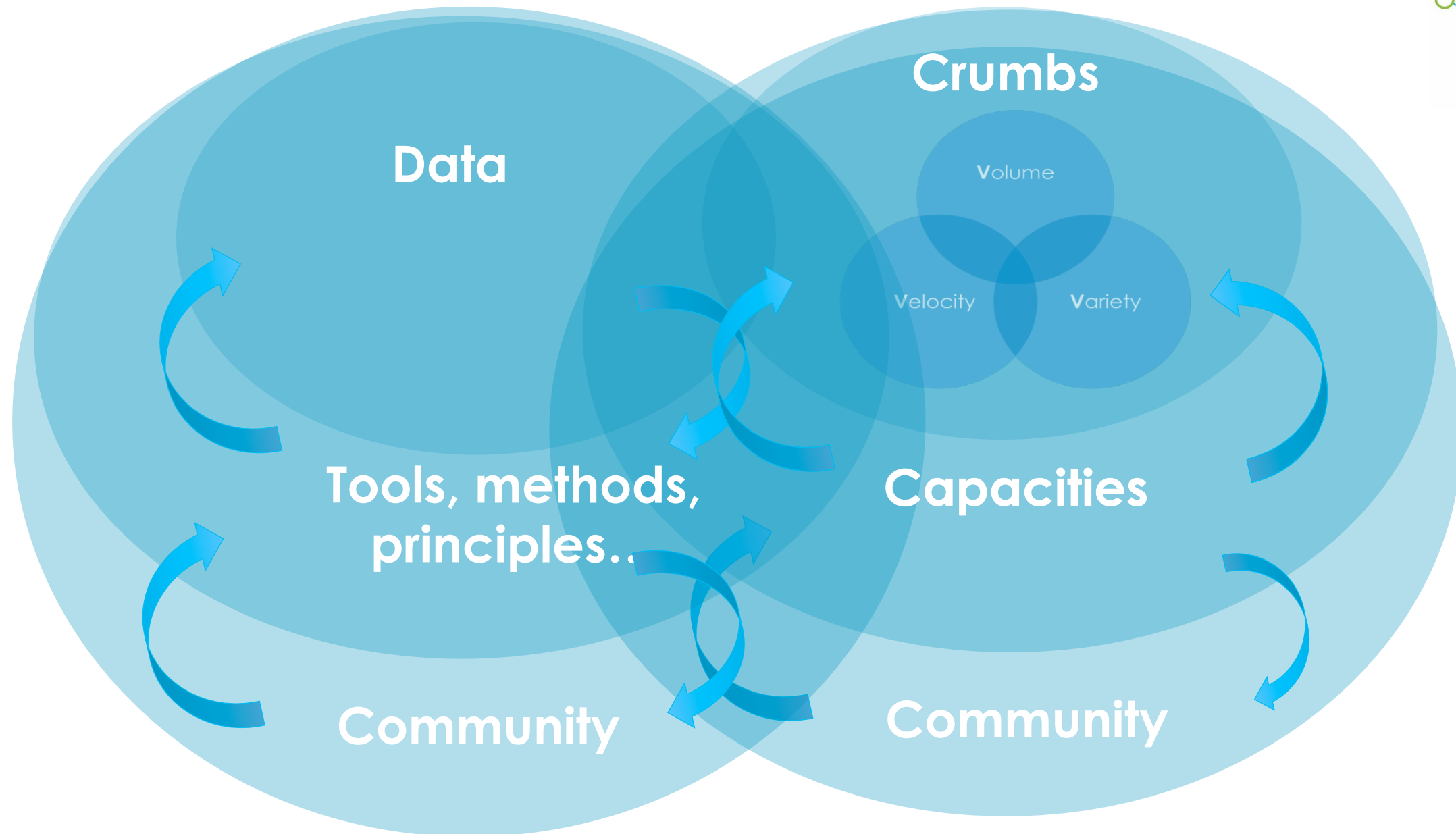
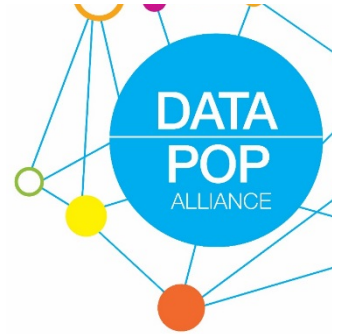
Big Data can have four main roles or functions:


1. **Descriptive** (e.g. maps);
2. **Predictive**, includes what has been called ‘now-casting’ or inference as well as forecasting
3. **Prescriptive** (or *diagnostic*), by establishing causal relations (=
4. **Discursive** (or *engagement*), concerns spurring and shaping dialogue within and between communities ==
“datadvocacy”—like DHS?

Demography meets Big Data, Big Data meets demography



Demography meets Big Data, Big Data meets demography



- 
- 1—What are we talking about?
- 2—What has been done? (a few cases)
- 3—What could / should be done?



Predicting Population Density from Cell-Phone Activity

Senegal (2015)



Scientific Prize and Ethics Mention: Construction of socio-demographic indicators with digital breadcrumbs

F. Bruckschen ⁽¹⁾, T. Schmid ⁽²⁾, T. Zbiranski ⁽¹⁾

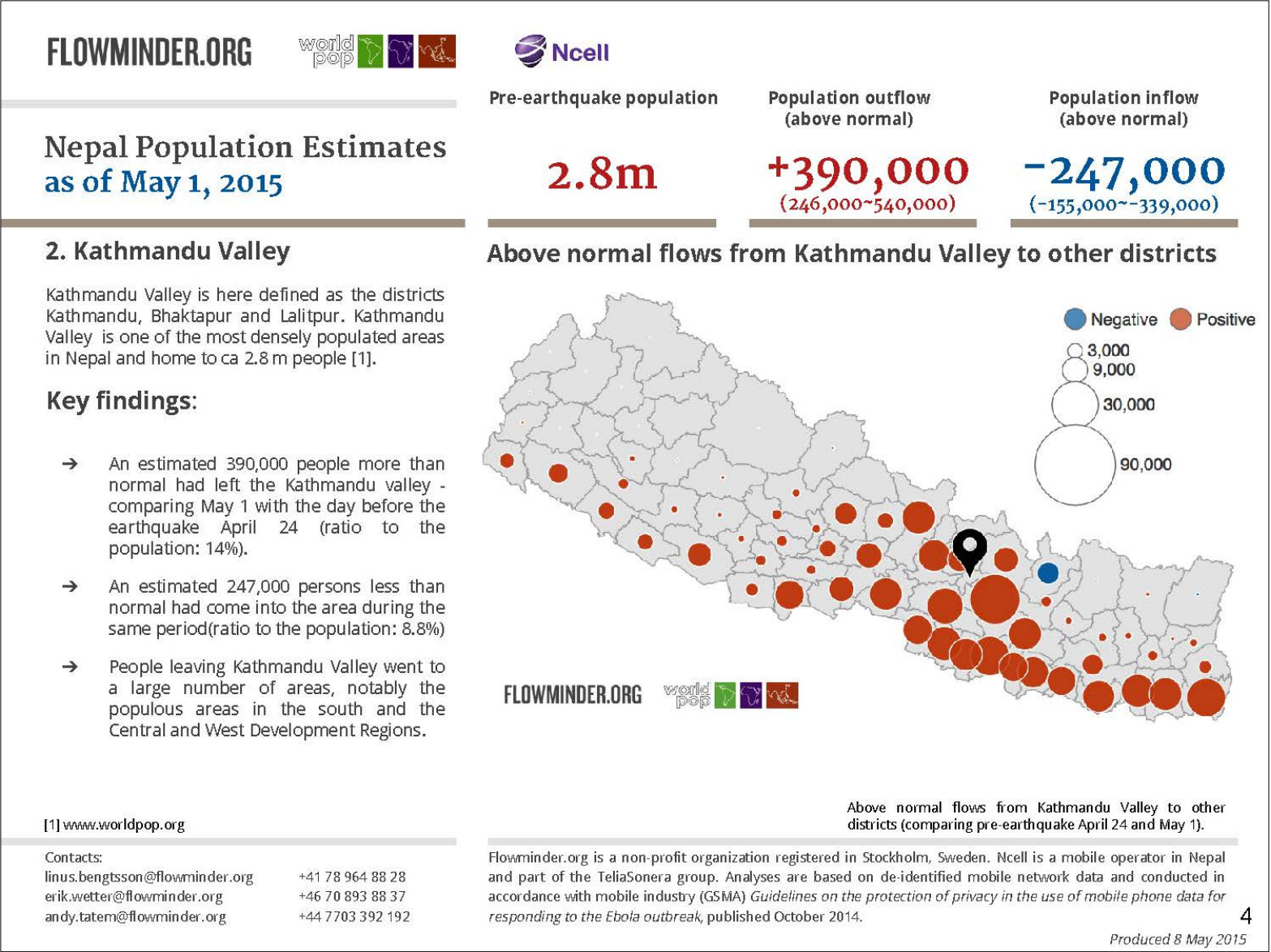
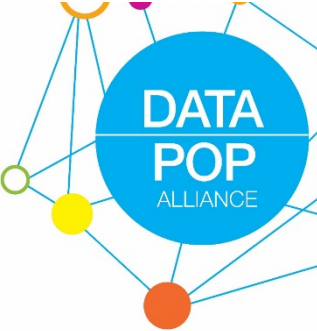
We show that socio-demographic indicators such as population, age, literacy, poverty, religion, ethnicity, electricity supply and others can be estimated in unprecedented detail and virtually ad-hoc using antenna-to antenna traffic data only. We offer a uniform approach that can be easily extended to other variables. Results are tested for spatio-temporal robustness and visualized as heat maps.

(1) Humboldt Universität Berlin, Germany - (2) Freie Universität Berlin, Germany



Post-Earthquake Population Movement

Nepal (2015)



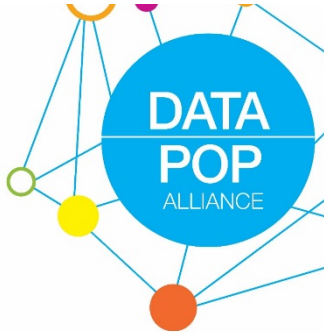
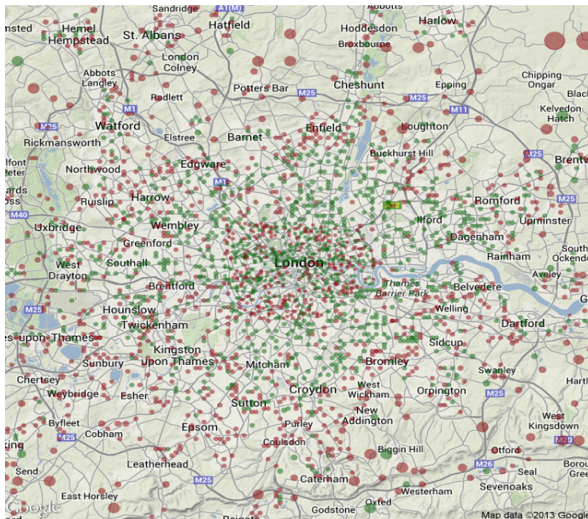
Predicting Crime Hotspots from Cell Phone Data

London (2013-14)

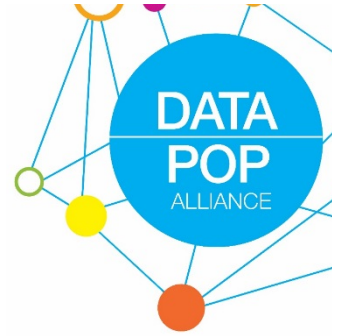
- Binary classification task using Random Forest

Table 3: Metrics Comparison

Model	Acc.,%	Acc. CI, 95%	F1,%	AUC
Baseline Majority Classifier	53.15	(0.53, 0.53)	0	0.50
Borough Profiles Model (BPM)	62.18	(0.61, 0.64)	57.52	0.58
Smartsteps	68.37	(0.67, 0.70)	65.43	0.63
Smartsteps + BPM	69.54	(0.68, 0.71)	67.23	0.64



Source: Moves on the Streets: Predicting Crime Hotspots Using Aggregated Anonymized Data on People Dynamics
Bogomolov A., Lepri B., Staiano J., Letouze, Oliver N., Pentland A., Pianesi F.



Risk Sharing in Natural Disasters through “Mobile Money”

Rwanda (2010)

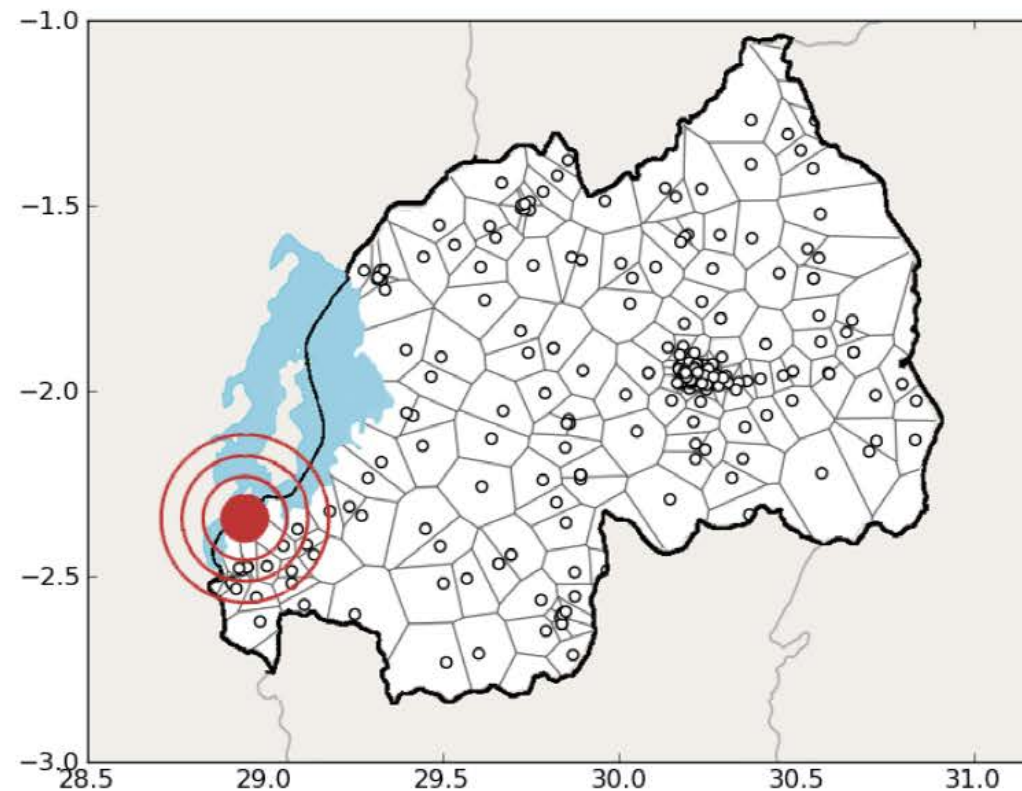
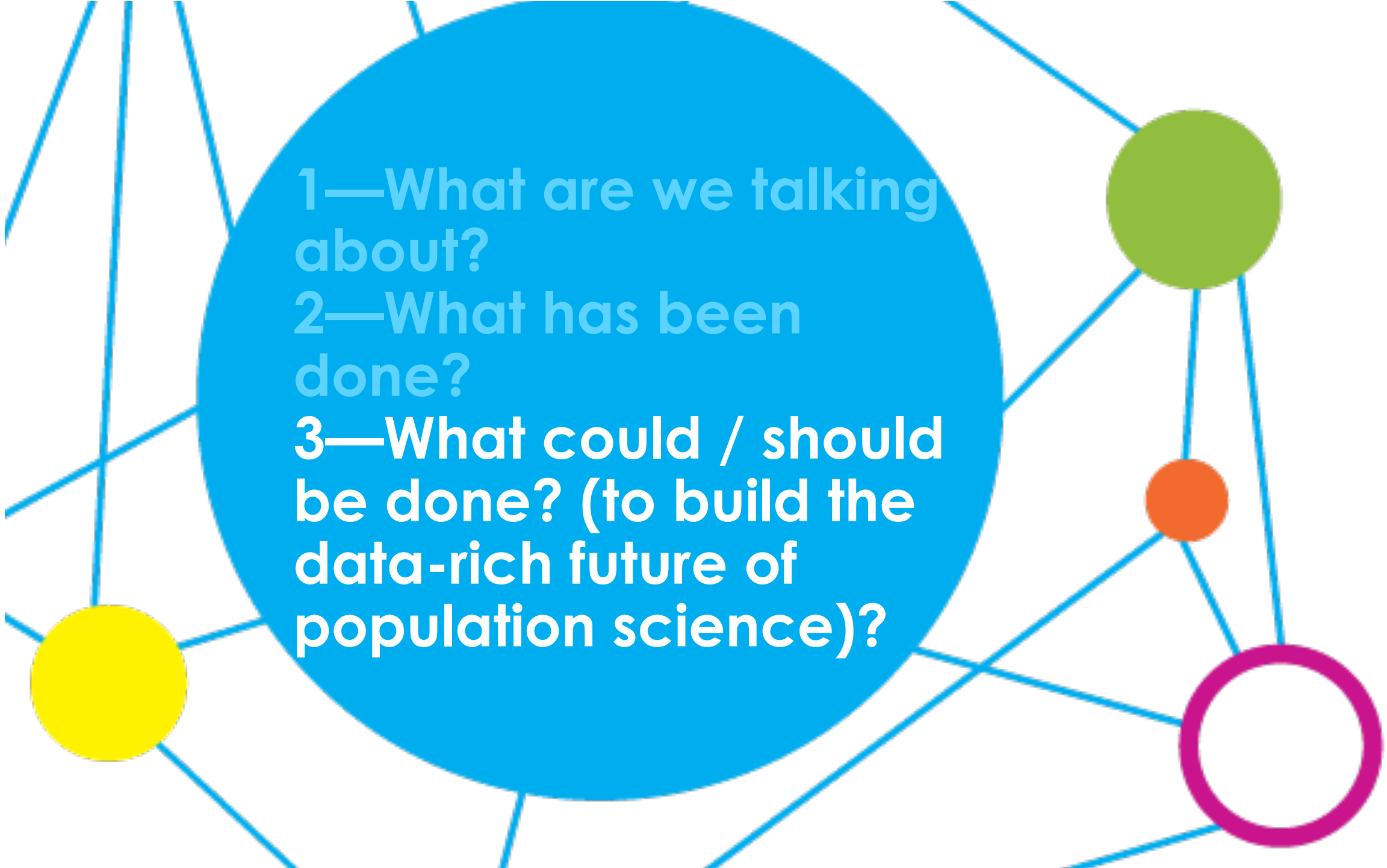
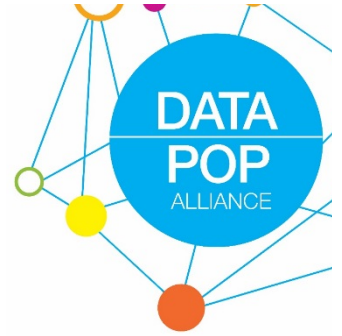


Figure 1: Map of Rwanda showing the location of mobile phone towers (as of February 2008) and the location of the Lake Kivu earthquake of 2008. Each black dot represents a cell tower, with the approximate area covered by the tower demarcated by adjacent Voronoi cells. The epicenter of the earthquake is shown with red concentric circles.

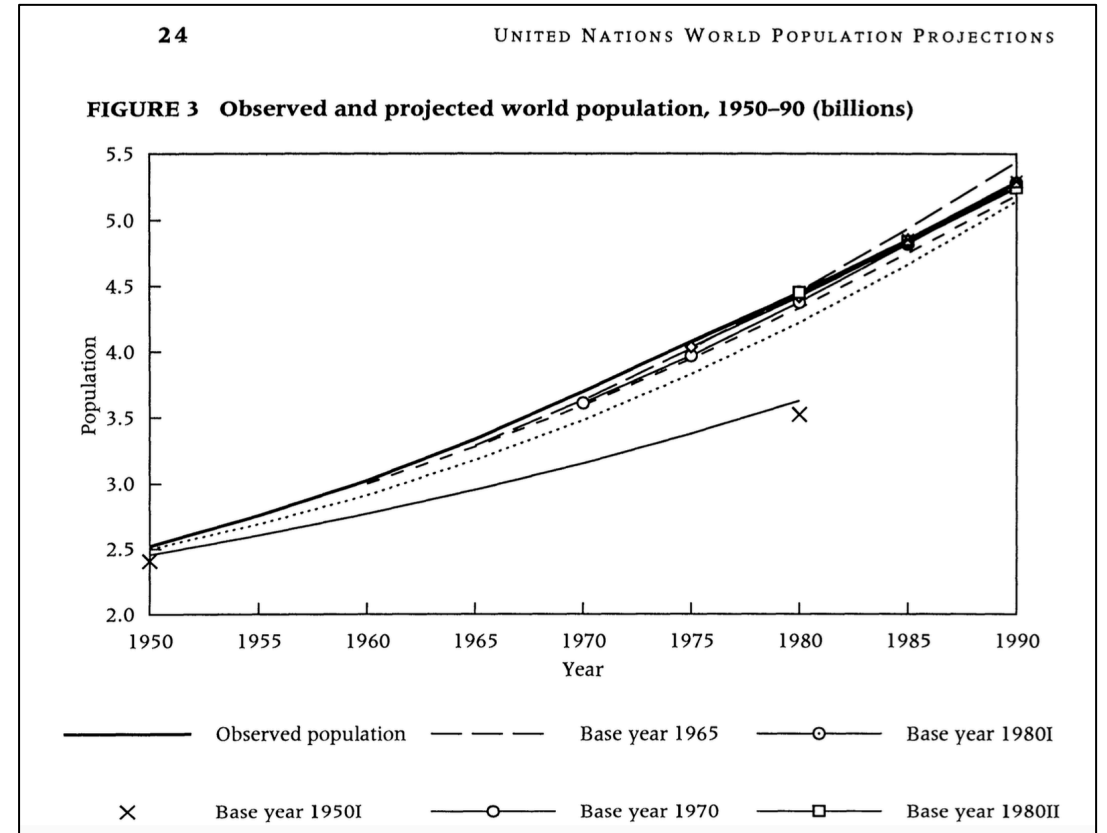
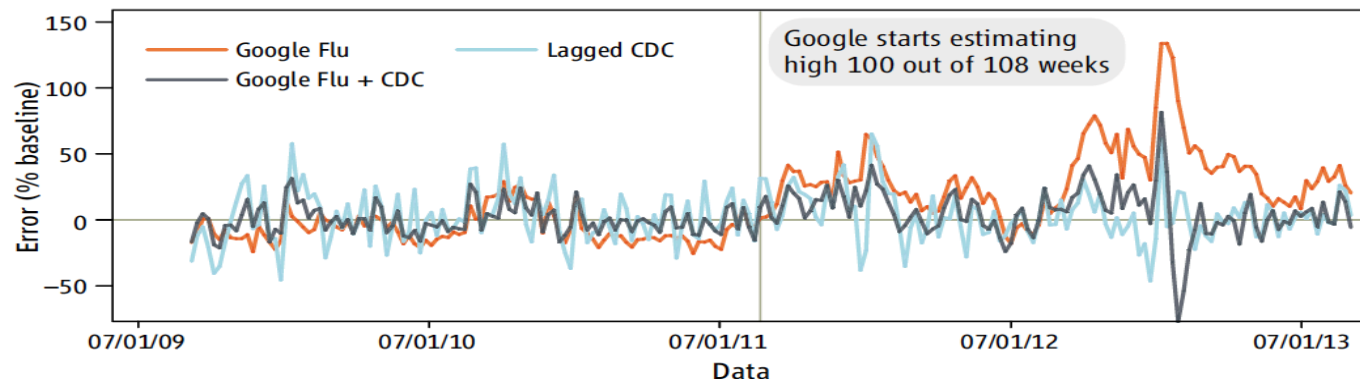
Source: Risk Sharing and Mobile Phones: Evidence in the Aftermath of Natural Disasters, 2014
Blumenstock J., Eagle N., Fafchamps M.,

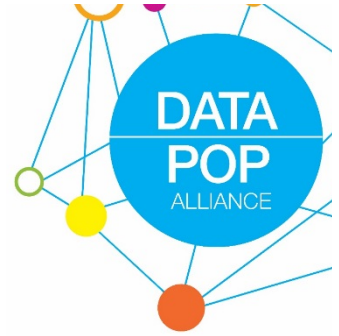
- 
- 1—What are we talking about?
- 2—What has been done?
- 3—What could / should be done? (to build the data-rich future of population science)?



Learning.

1—Learn from past mistakes (evolution)





2—From old recipes (using new ingredients)

Modeling and correcting sample bias in non-sampled data

Selection bias correction

The profiles by age and gender are adjusted in two ways. First, they are rescaled: the entire profile is shifted upward or downward by a factor that makes estimates for European countries match the order of magnitude of annual age-specific rates provided by Eurostat for the year 2009. Second, the estimated number of migrants, by age group and gender, is multiplied by a correction factor to adjust for over-representation of more educated and mobile people in groups for which the Internet penetration is low. The correction factor, CF is:

$$CF = \frac{p_{gac}(e^{-k} - 1)}{(e^{-kp_{gac}} - 1)}$$

where p_{gac} is the Internet penetration rate for gender g , age group a , and country c . k is a parameter that measures the intensity of the impact of lower Internet penetration rates on the selection of the more mobile people in the sample of users. For a large number of countries, data on Internet penetration rates, by age and gender, are available from the UNECE Statistical Database on Internet use. Therefore, the correction factor is in practice a function of k . Figure 1

Zagheni & Weber, 2012

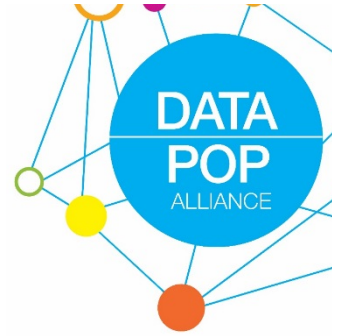
$$\log(P_i) = k \log(U_i) + bias_i + \varepsilon_i$$

$$bias_i = f(\text{mobile phone penetration}_i)$$

$$Y_i = \beta_0 + \beta_1 \text{mobile phone penetration}_i + \varepsilon_i$$

Letouzé, Zagheni & al, Weber, 2015

Then: **blending** of hypothesis based vs. supervised machine learning methods to model bias



3—From and for others

IUSSP PANEL PROPOSAL

Big Data and Population Processes

Justification:

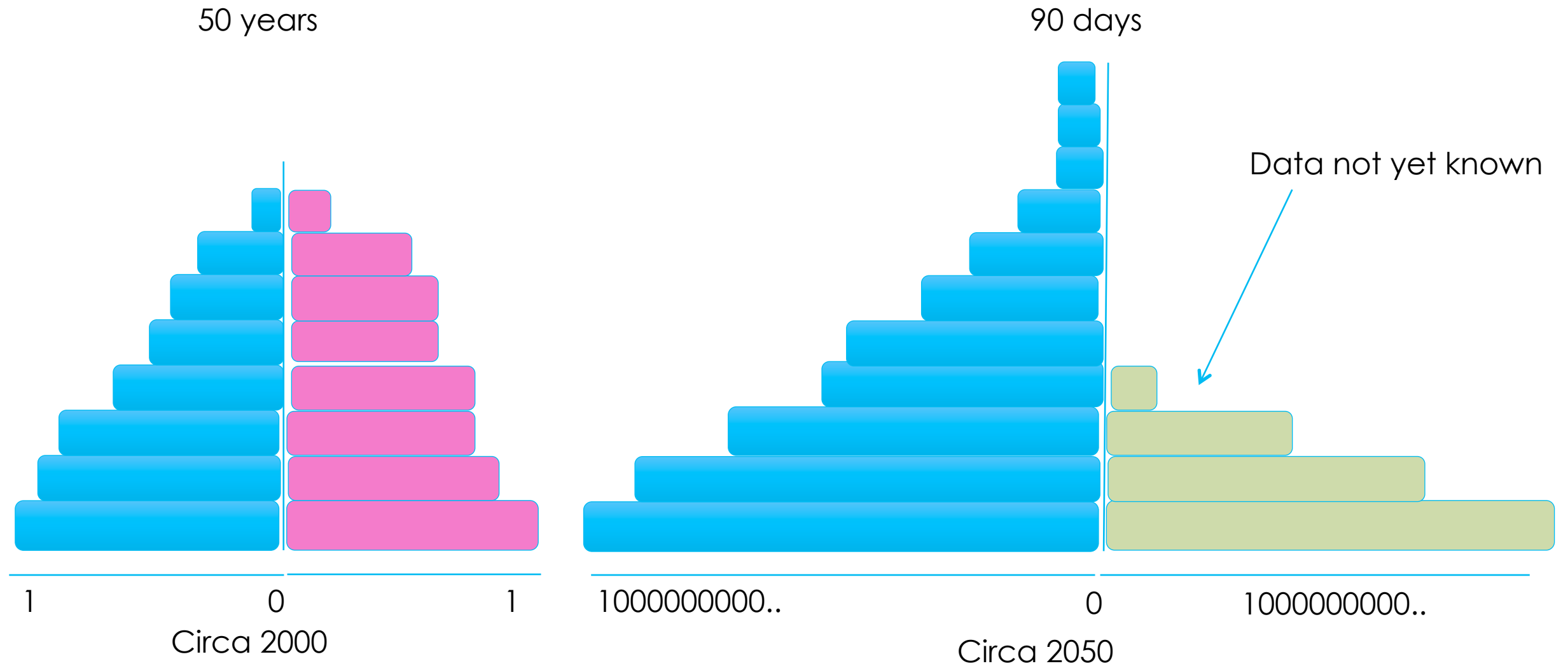
Demography has been a data-driven discipline since its birth, about 350 years ago. In 1662, John Graunt published the very first life table after compiling data that had been collected partially for public health and 'marketing' purposes: data on number of deaths were originally collected at the request of the merchants of London, who wanted to evaluate the number of potential customers (i.e., live people by age) in London at times of epidemics. Since then, data collection and development of formal methods have sustained most of the major advances in our understanding of population processes.

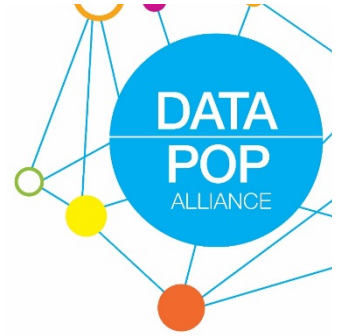
Today's increasing availability of digital records (e.g., social media data, digitized historical data, call detail records, smartphone apps, and blogs) holds the promise of an exceptional development of new demographic knowledge, which will lead to theoretical advances in population studies. 'Big data' is widely seen as a sea change. The global spread of Internet and digital technologies has radically transformed the way in which we communicate with each other. As a consequence of the digital revolution, individuals leave an increasing quantity of digital breadcrumbs. These records can be aggregated and mined to advance our understanding of social processes. Web-based research has been rapidly increasing its prominence in the social sciences.



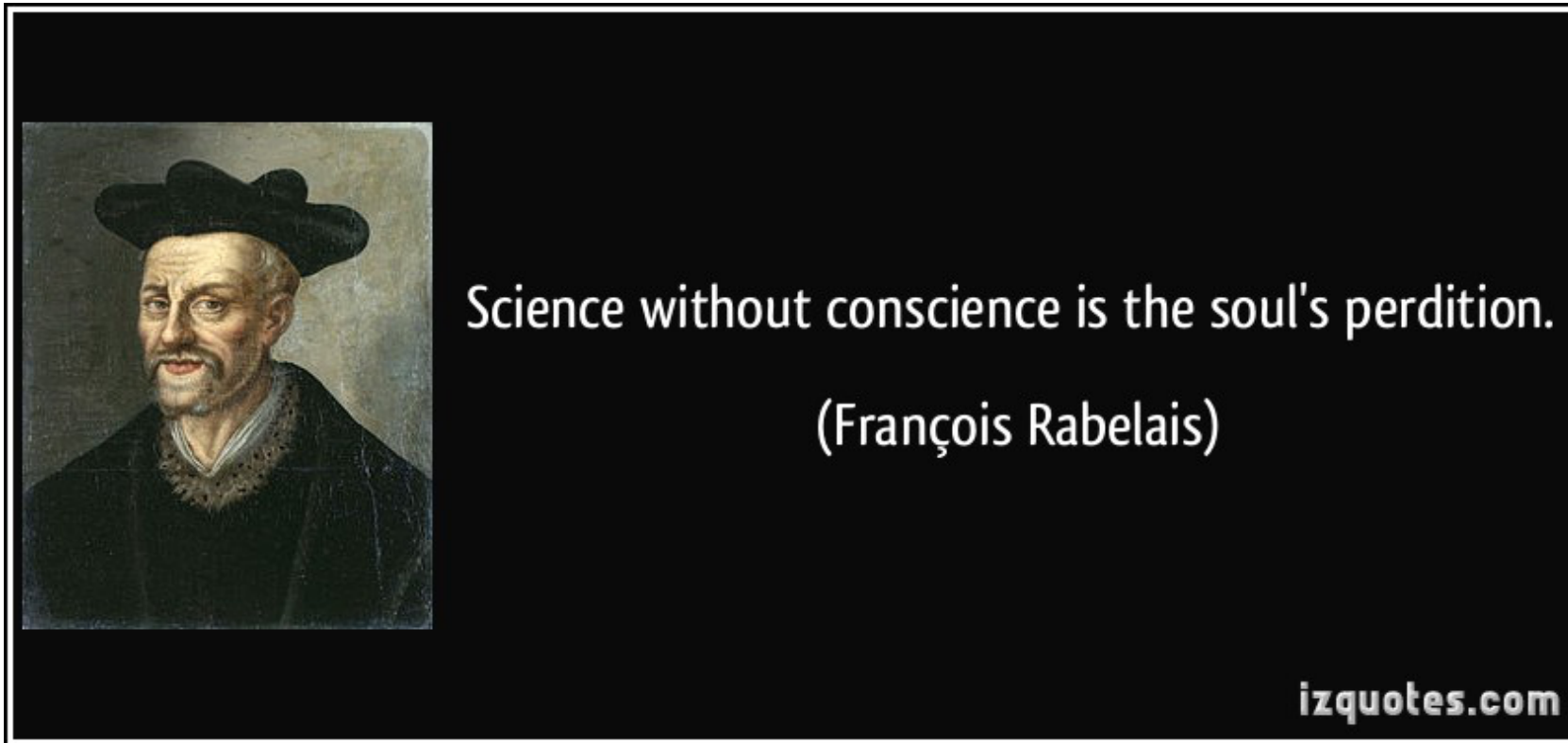
https://c1.staticflickr.com/9/8570/16070333273_5139661ba4.jpg

Modeling and projecting world population of *data*?



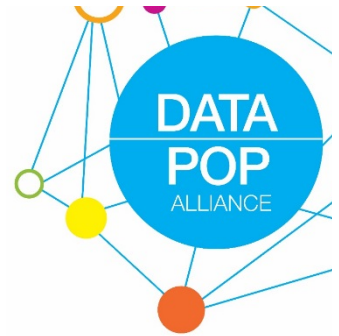


4—From good quotes

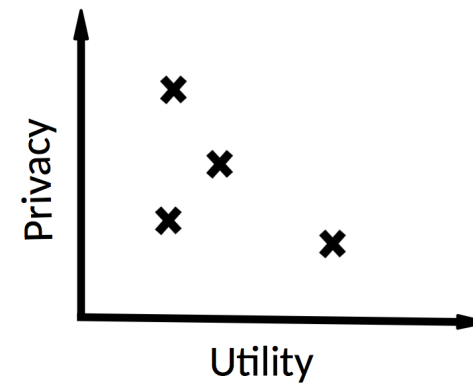
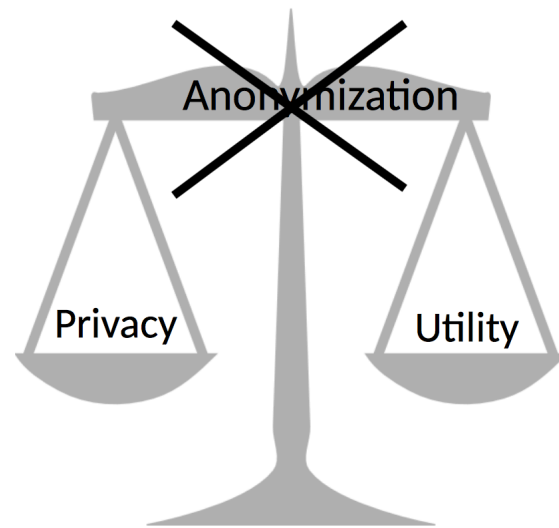


**“Change the instruments, and you will change
the **entire social theory** that goes with them”**

Bruno Latour 2009



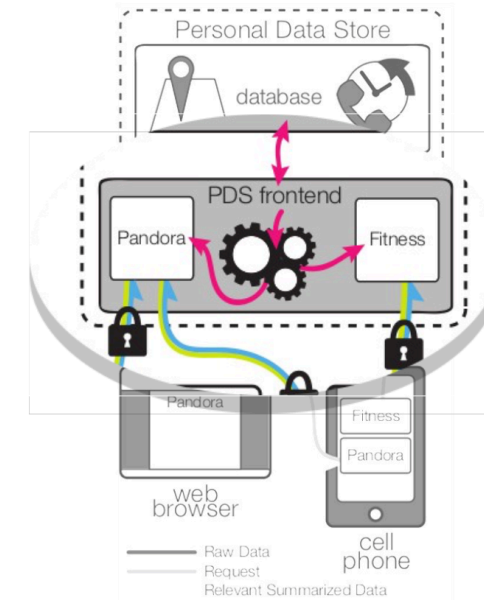
5—From innovation and imagination



Computational Privacy

Yves-Alexandre de Montjoye, MIT

Dynamic systems



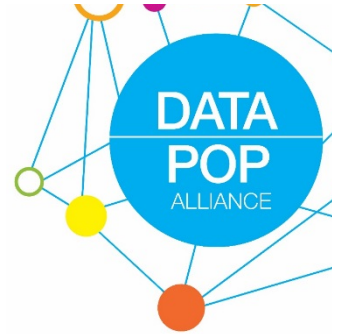
Inferring friendship network structure by using mobile phone data

Nathan Eagle^{a,b,1}, Alex (Sandy) Pentland^b and David Lazer^c

The Future of Cities

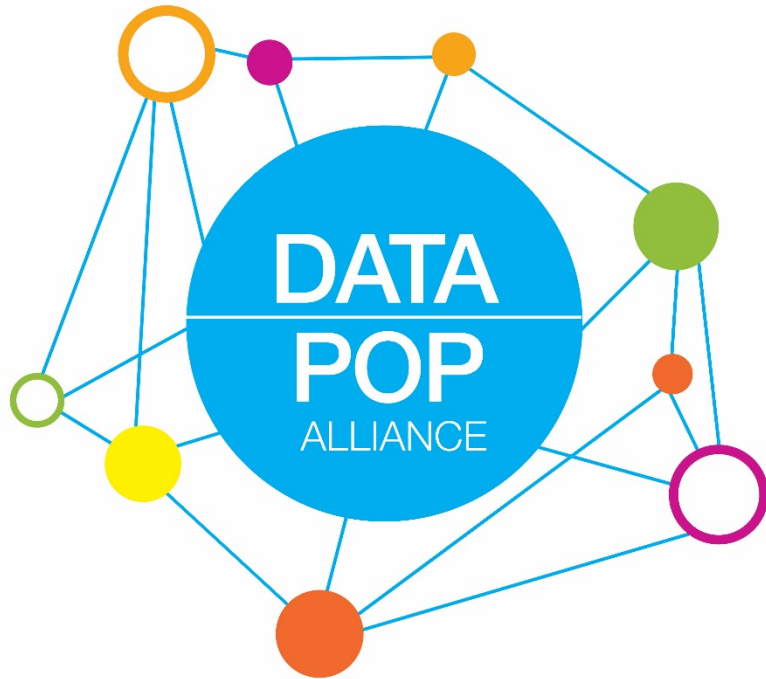
The Internet of Everything will Change How We Live

By John Chambers and Wim Elfrink



Conclusions:

1. Demography has been **slow to enter the data revolution** but it is well positioned to **catch up fast**—and has a lot to contribute
2. What is at play & stakes is not **just about new kinds of data; it's an entirely new ecosystem of data, tools and actors emerging**
3. Demography should and probably will **reinvent itself by becoming population science**, a science that should strive to both **measure and understand to positively affect old and new population processes** based on old, sound, principles
4. This should happen **gradually in the next 15 years** but it will require and significant efforts and investments to build **new mindsets, systems and capacities**



Thank you

eletouze@datapopalliance.org