Big Data for Development:

What May Determine Success or failure?

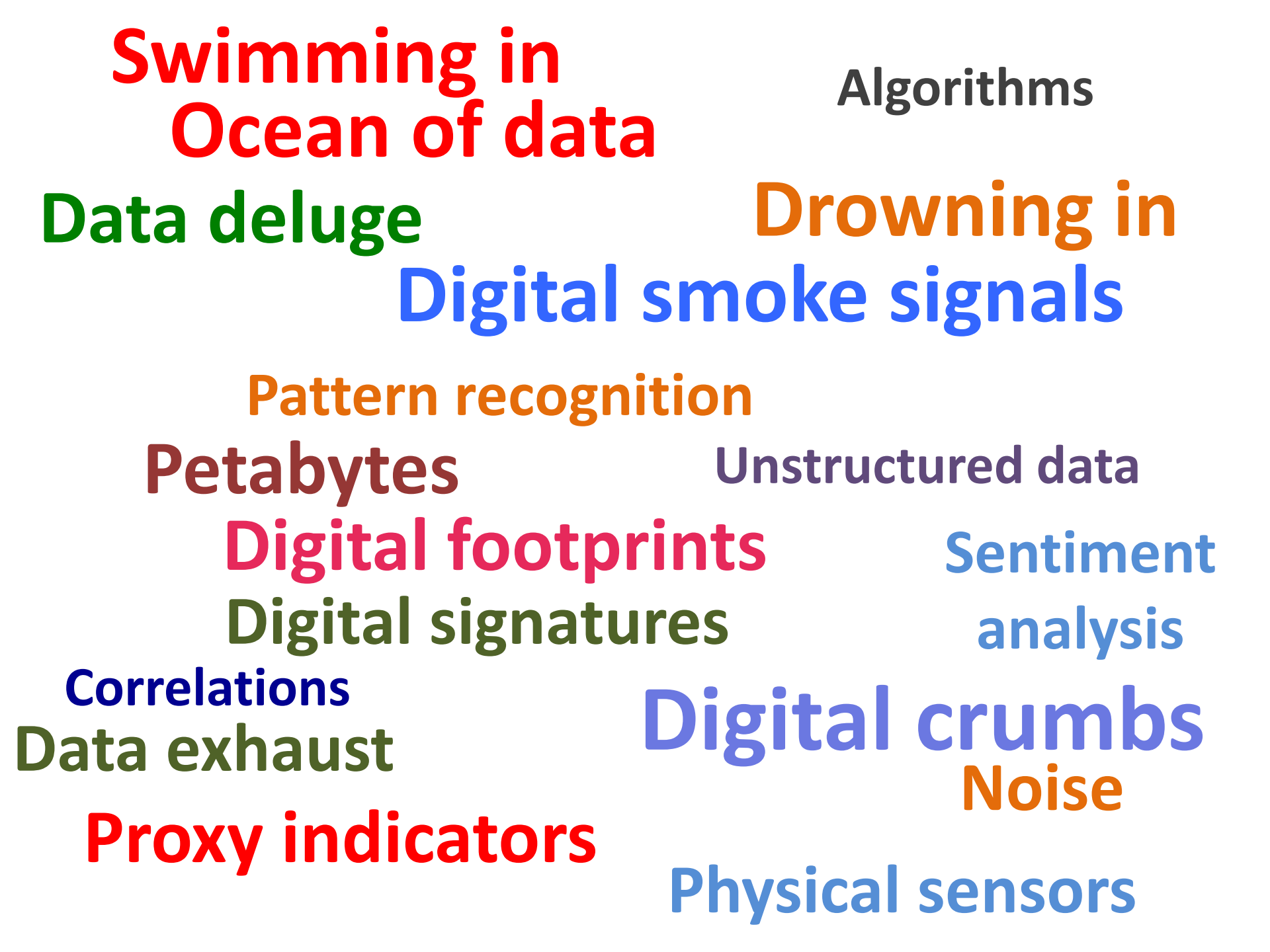
Emmanuel Letouzé

letouze@unglobalpulse.org

OECD Technology Foresight 2012
Paris, October 22

Big Data for Development: Challenges & Opportunities

May 2012


Swimming in Ocean of data
Algorithms
Data deluge
Drowning in
Digital smoke signals
Pattern recognition
Petabytes
Unstructured data
Digital footprints
Sentiment analysis
Digital signatures
Correlations
Digital crumbs
Data exhaust
Noise
Proxy indicators
Physical sensors

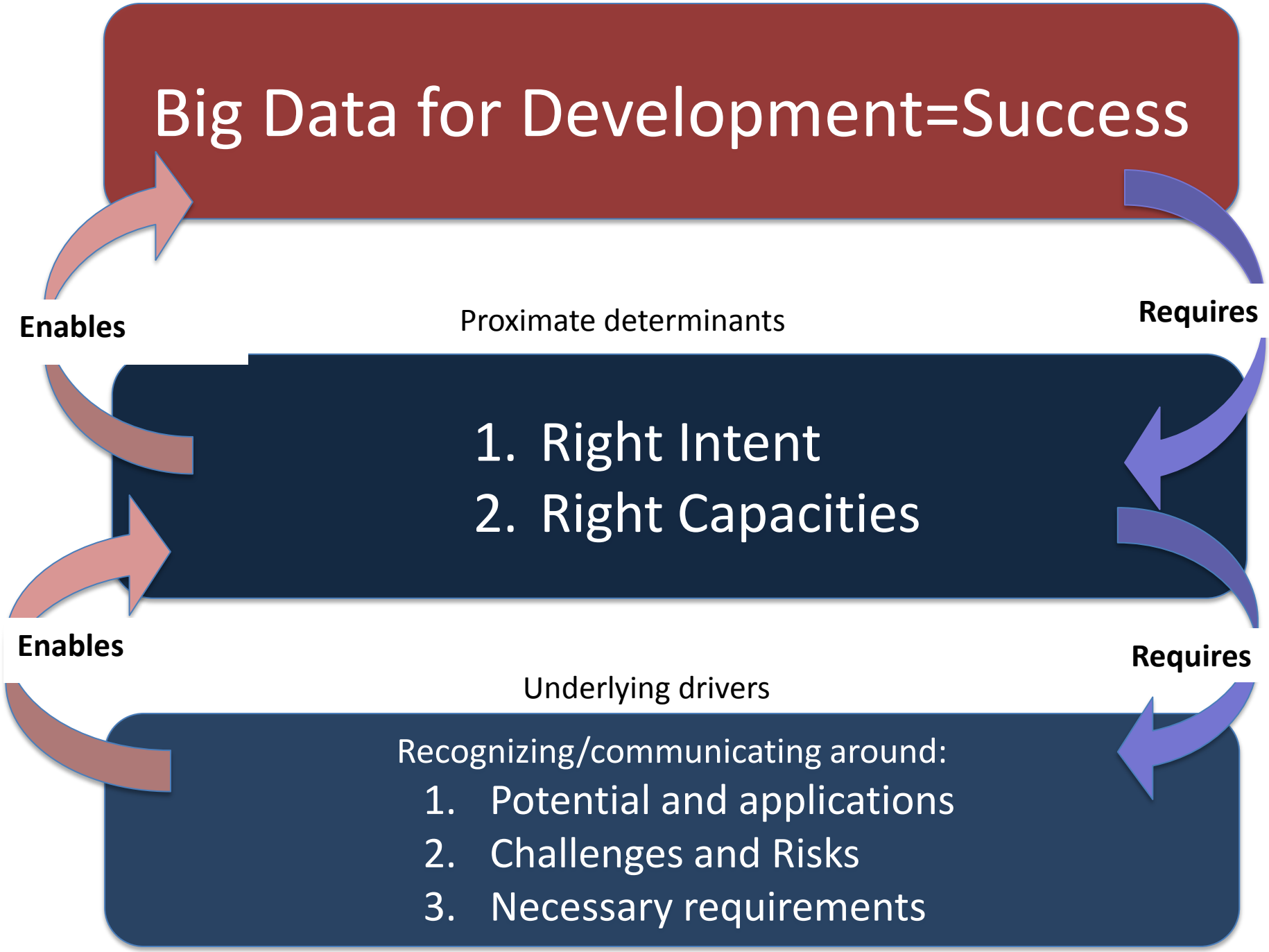
Background and Question(s)

1. **Complex, hyperconnected, volatile world**. Shocks, risks, vulnerability, resilience, agility, have become mainstream. Policies need to be more agile.
2. **Big Data. Big excitement. Big skepticism**. "New Industrial Revolution"? "New oil" that "needs to be refined"? "Big enough data speak for themselves"? "Big Data, Big Deal"? Worse: a dangerous path? (think conflict settings)? Can it help / change development—how?

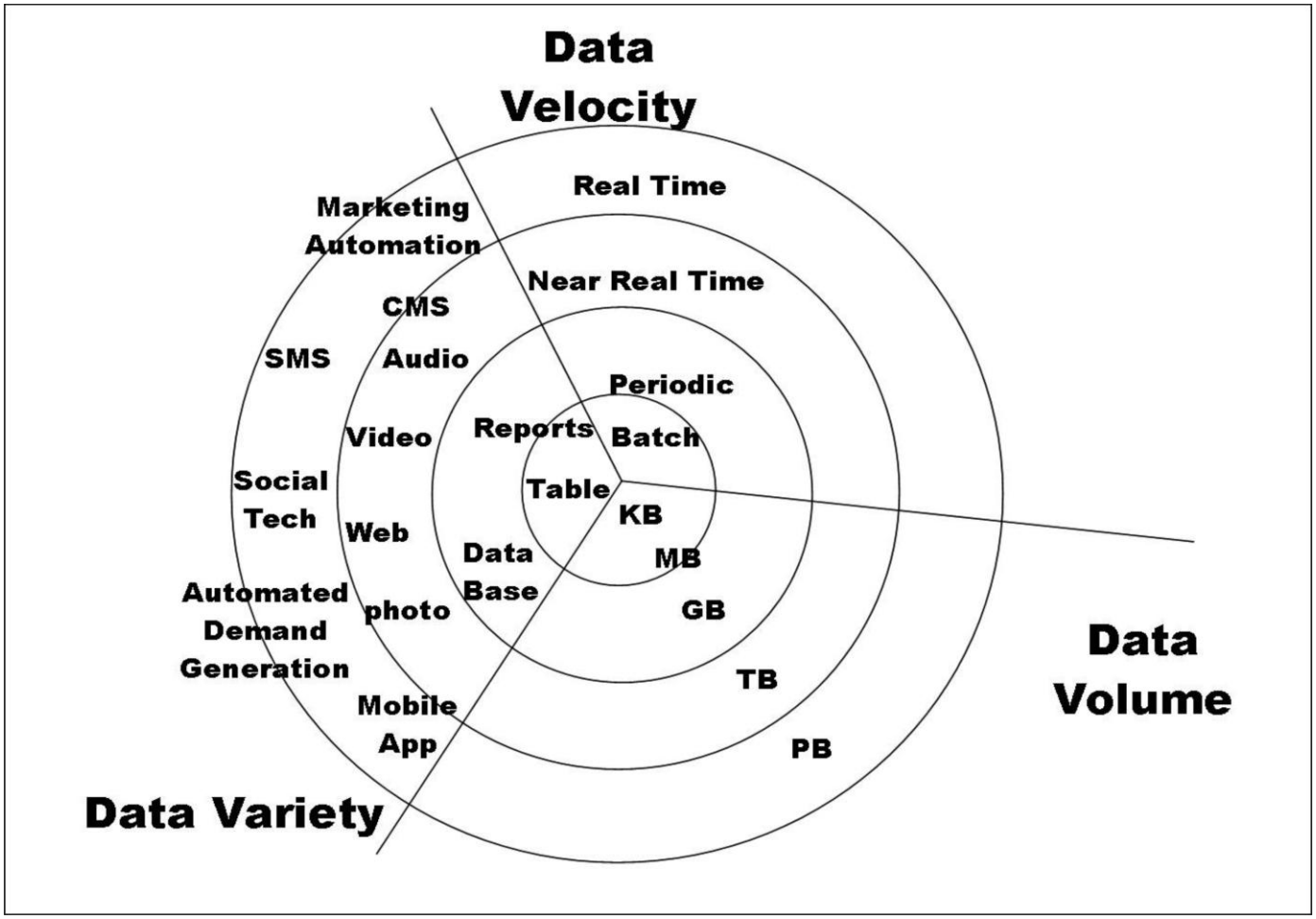
➔ What is the opportunity, what are the challenges?

➔ A better question is:

What will determine whether Big Data for Development succeeds or fails?


Big Data for Development=Success
Enables
Proximate determinants
Requires
1. Right Intent
2. Right Capacities
Enables
Requires
Underlying drivers
Recognizing/communicating around:
1. Potential and applications
2. Challenges and Risks
3. Necessary requirements

From Big data ➔ Big Data for development



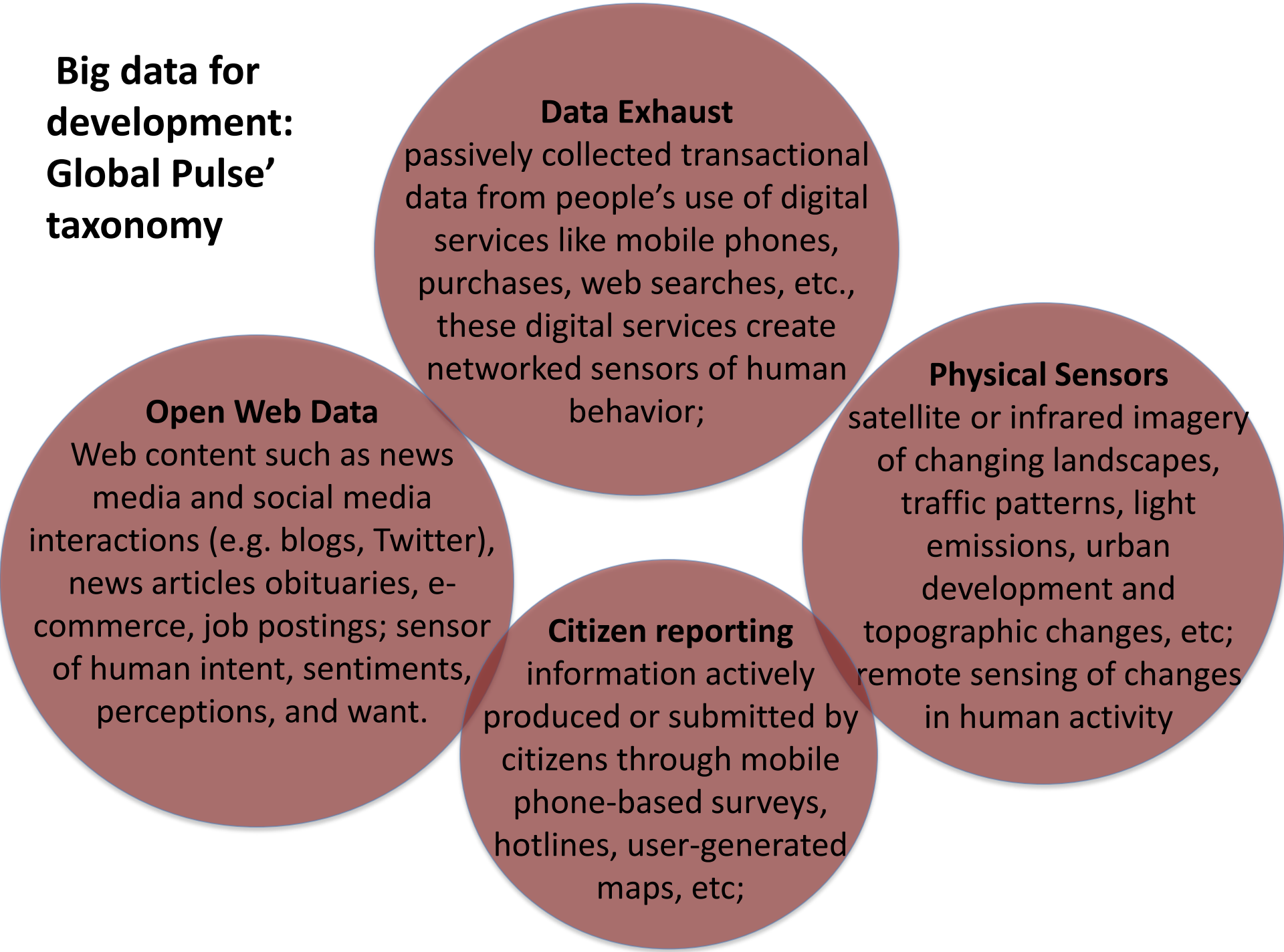
+ purpose (*intent*)

+ computing tools, systems (*capacity*)

It's *relative*

It's *contextual*

Big data for development: Global Pulse' taxonomy



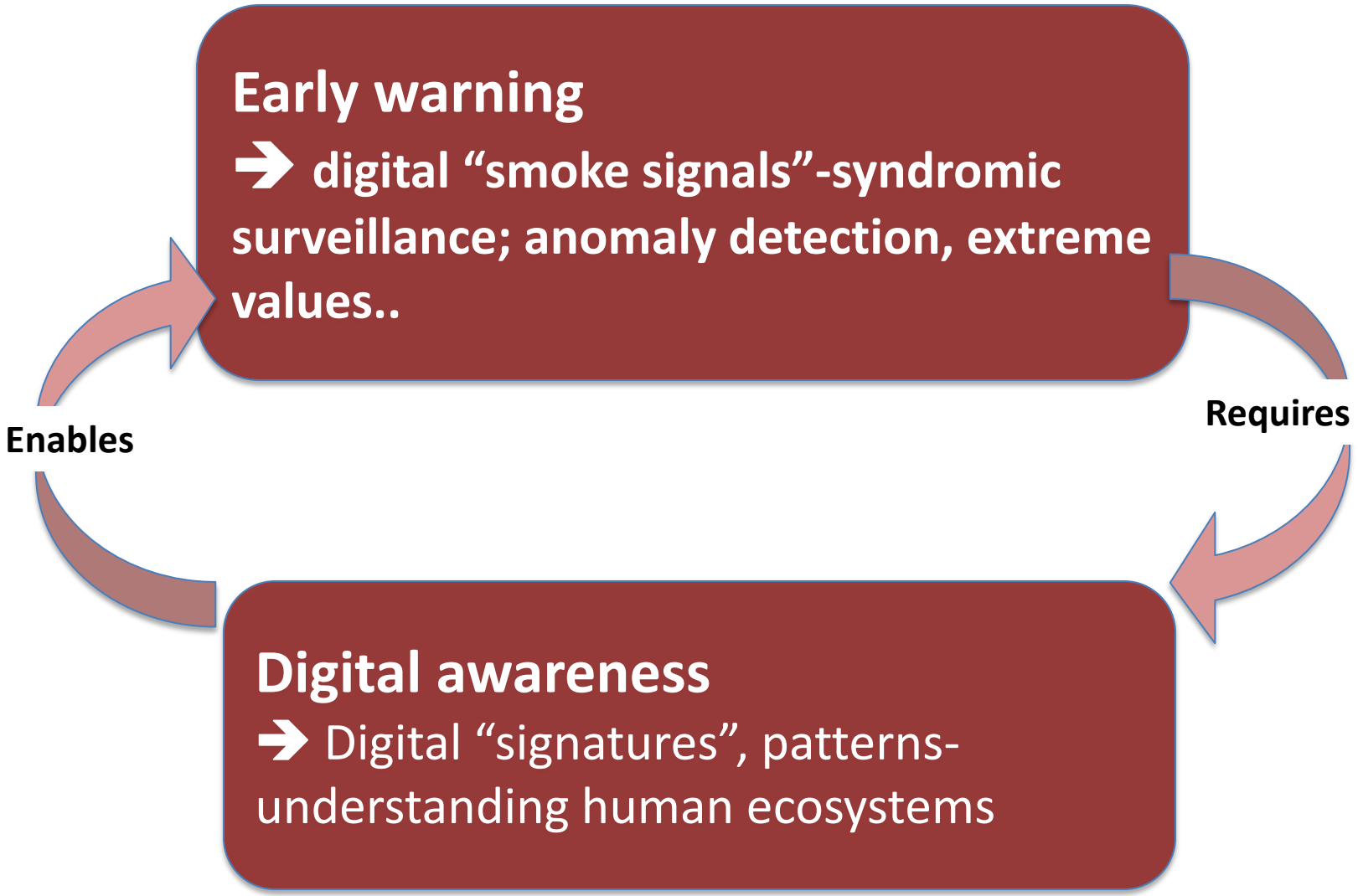
Potential applications according to Global Pulse

Early warning: "early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis";

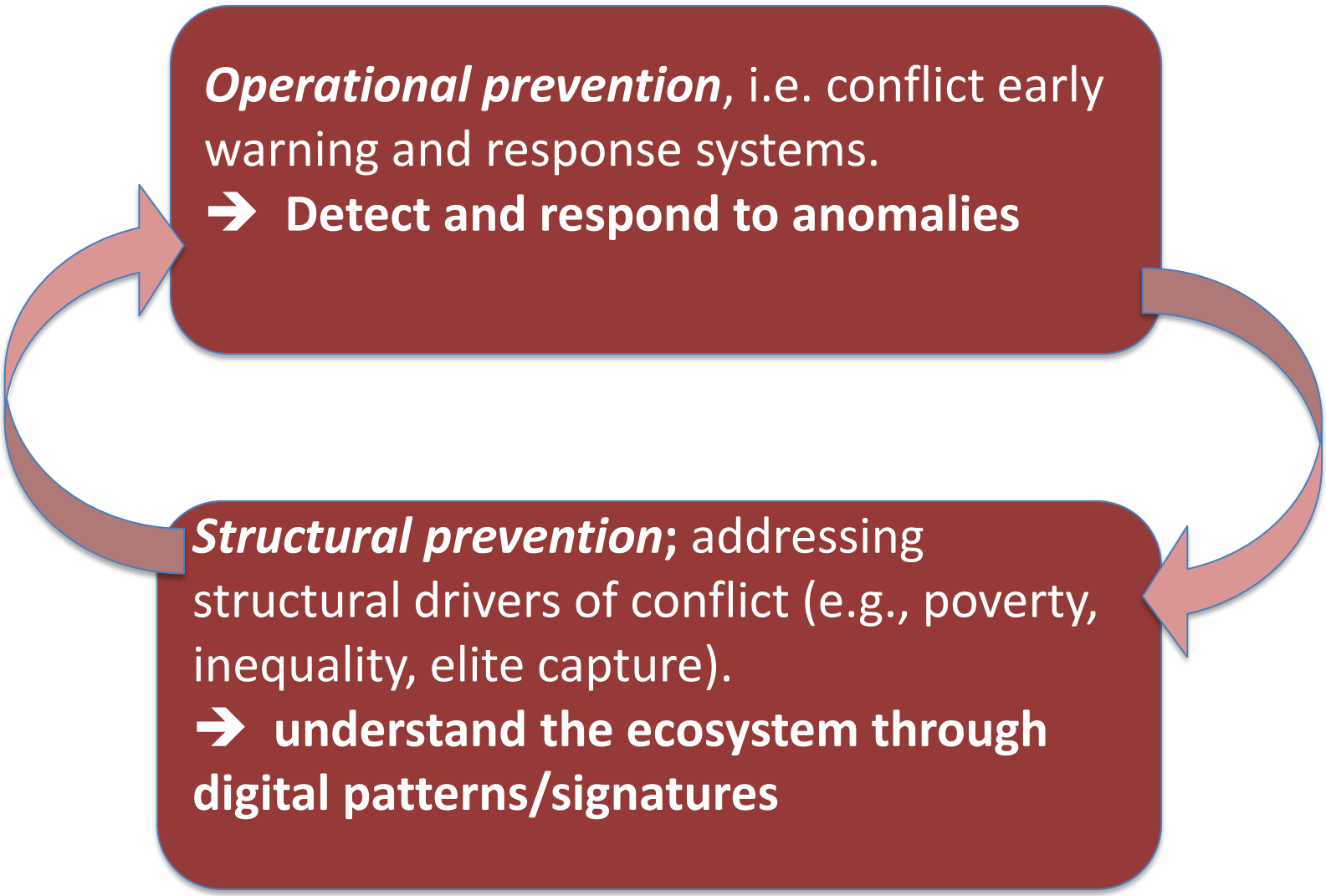
Digital awareness: "Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies;"

Real-time feedback: "monitor a population in real time makes it possible to understand where policies and programs are failing and make adjustments."

Intent ←→ Capacity

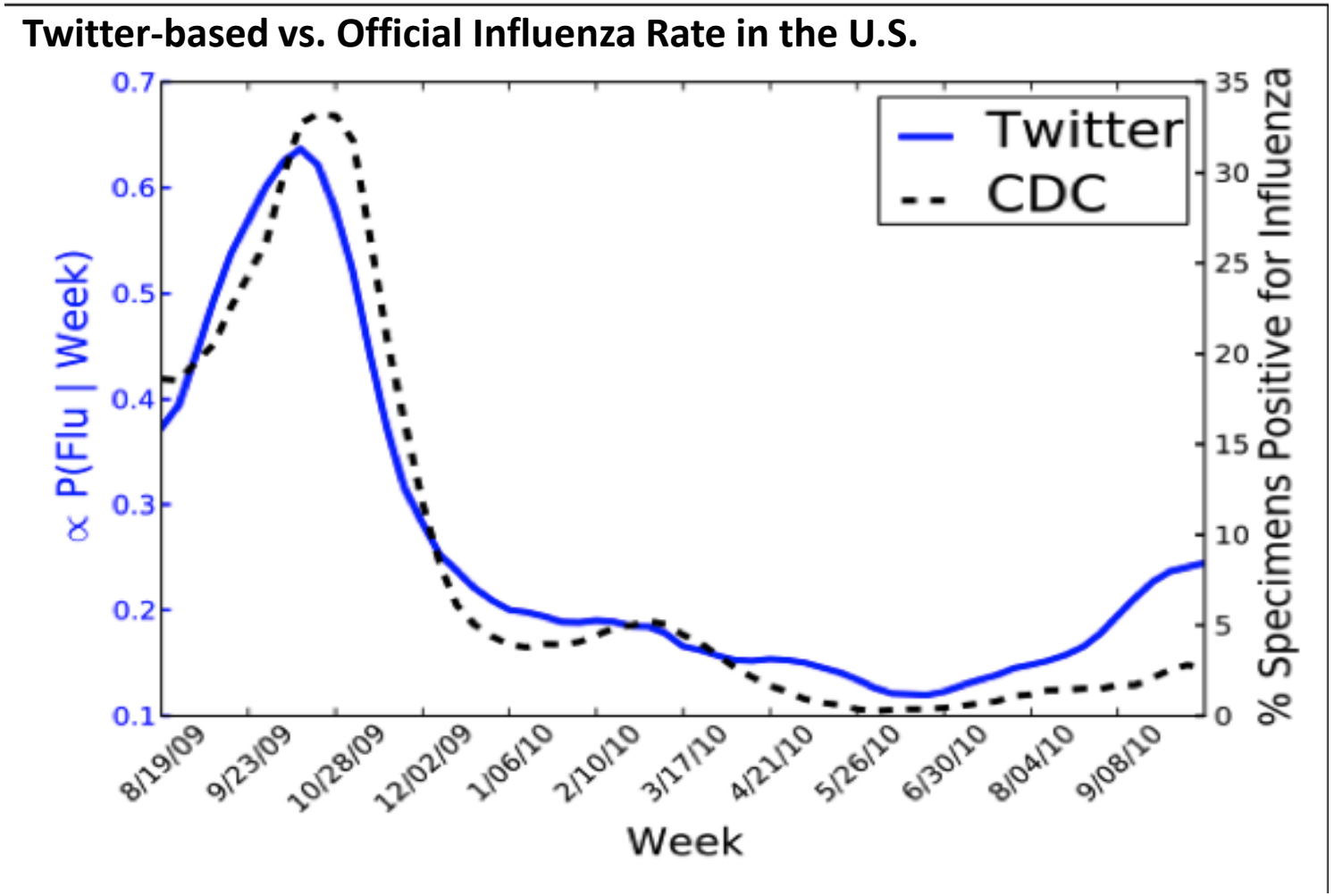


Ex: Conflict prevention (highly simplified)



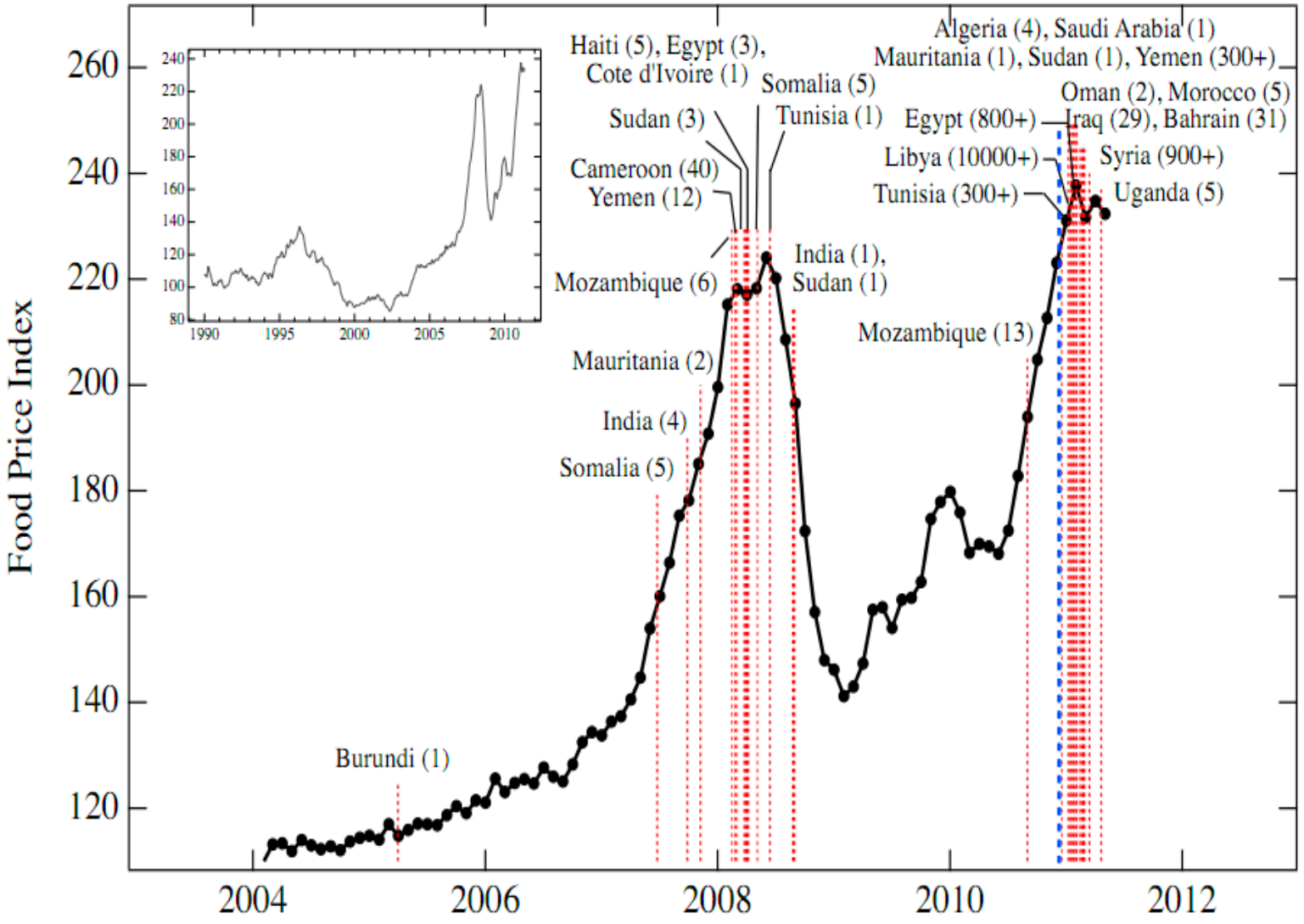
Examples

Early warning, syndromic surveillance and proxies= pick up / model anomalies



« You Are What You Tweet: Analyzing Twitter for Public Health. M. J. Paul and M. Dredze, 2011.
http://www.cs.jhu.edu/%7Empaul/files/2011.icwsm.twitter_health.pdf

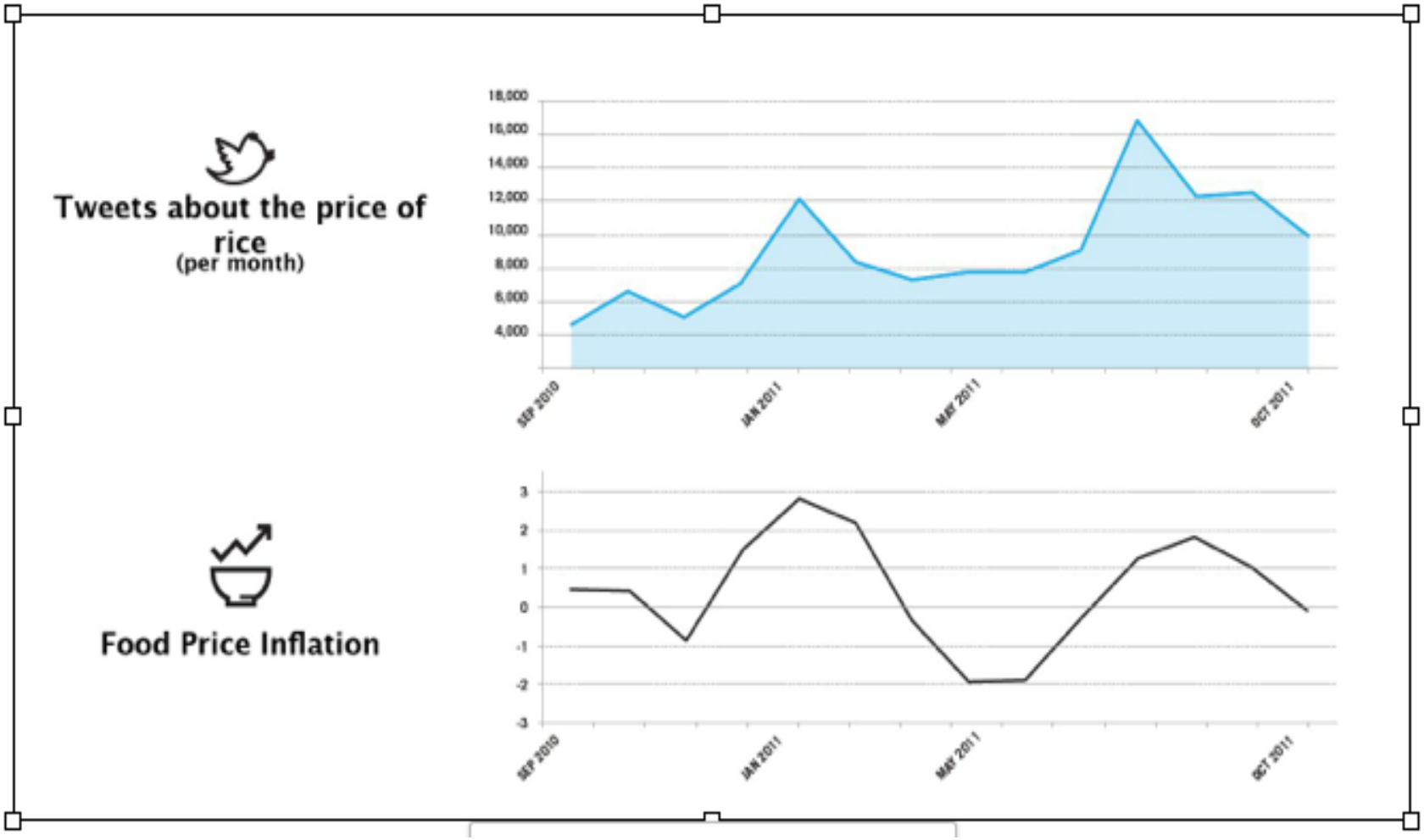
Modeling & predicting food riots



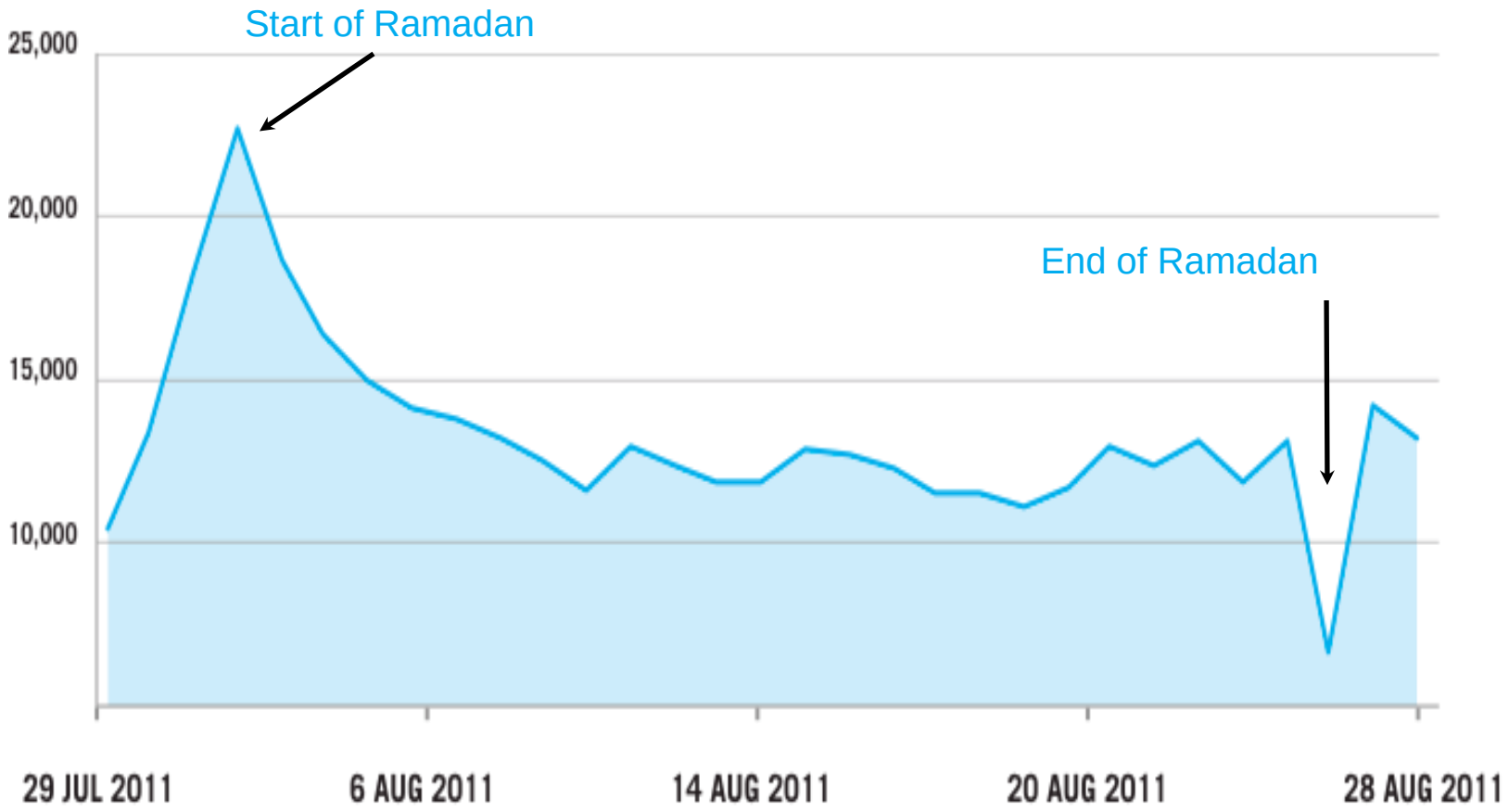
http://necsi.edu/research/social/food_crises.pdf

Finding proxy indicators (Global Pulse and SAS project)

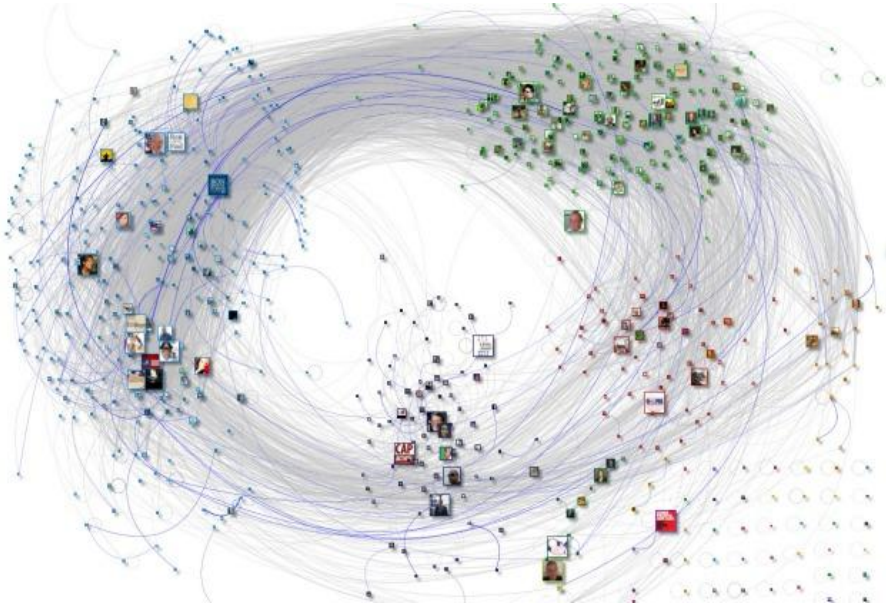
Tweets about the price of rice vs. actual price of rice in Indonesia



Tweets per day about meals during Ramadan

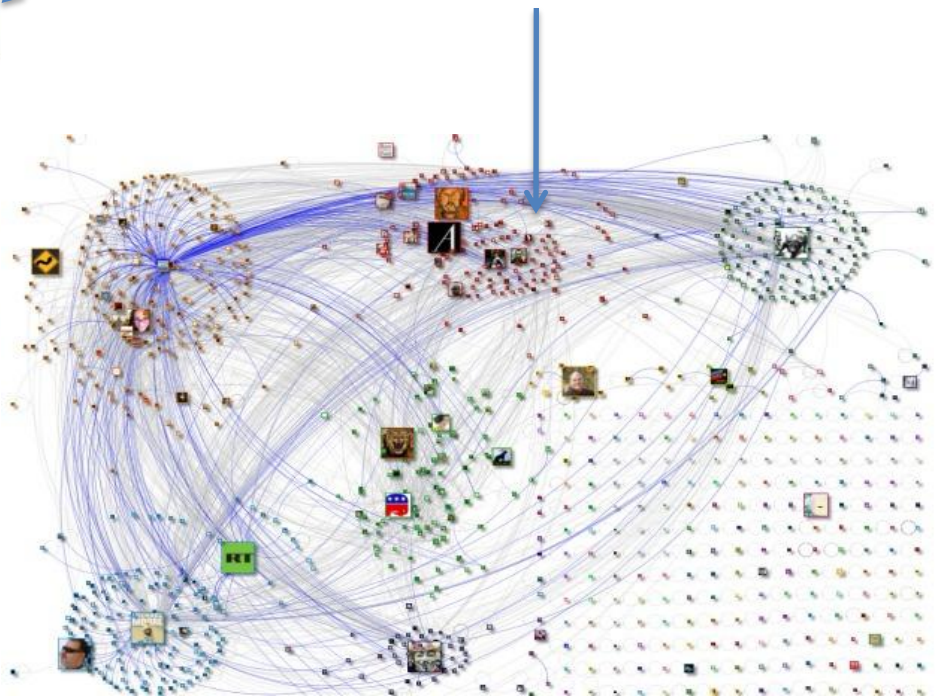


Digital awareness = understand ecosystem



Tea Party on Twitter

Occupy Wall Street on Twitter

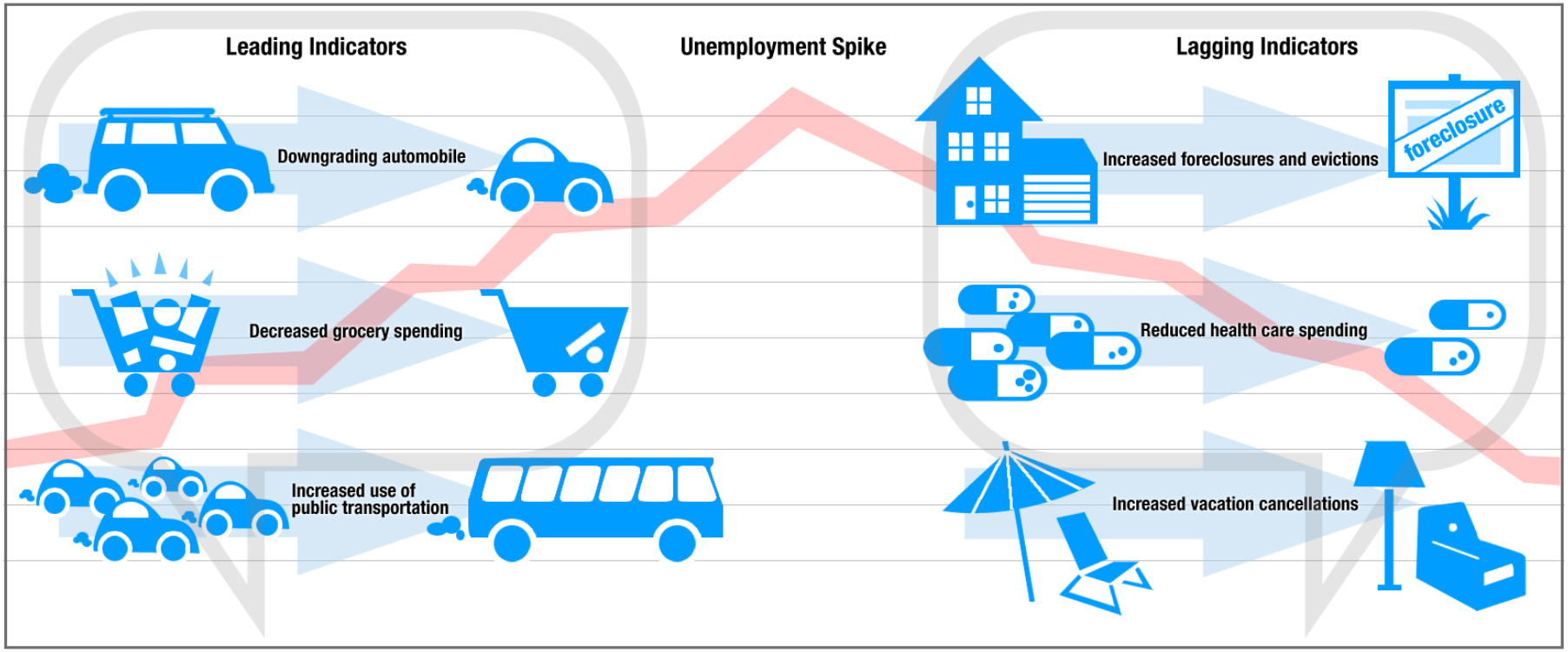


"The **Tea Party appears as a far more tight-knit** group (..) whereas **Occupy is made up of looser clusters with a few high-profile** accounts. In the lower right-hand corner = matrix of "isolates:" People who are talking about the ideas of Occupy or the Tea Party, but (…) don't have connections to others on the graph. **For Occupy, the number of isolates is greater**, which (..) could indicate a **larger potential for growth and stronger brand cachet**."

Source: M. Smith cited in FP

➔ What if we could do the same thing for militias, rebel groups—not today, but in N years, where N=5, 10, 15..?

Mood analysis = understand ecosystem + detect anomalies



Analysis of social media using SAS shows increases in chatter about certain topics that are leading and lagging indicators of a spike in unemployment.

Summing up potential / applications

i. ***digital 'signatures'*** that help better understand human ecosystems
ii. ***digital 'smoke signals'*** for anomaly detection

may be *especially* relevant in risky/poor places high rates of growth of tech penetration with typically poor / little more traditional data

But also gives a sense of some of the <u>risks and challenges</u> posed by Big Data for Development

Recognizing the difficulties and requirements

1. Data challenges

<u>Data</u>: How do we access data and protect their producers?

Challenge #1: *Availability:* Depends on technological penetration and use. But mobile phone use is increasing fast and supports/will support Internet access.

Challenge #2: *Reliability.* Attempts at playing the system(s), plus suppressing signals? Especially prevalent in conflict settings.

1. Data challenges

Challenges #3: Access. Not all data produced in easily accessible, storable (e.g. Twitter). This is compounded in poor countries. How to access data from a mobile phone carrier?

Challenge #4: *Privacy.* Because it's accessible does not make it ethical or safe. True in all settings. In **conflict/crisis settings, the privacy challenge may soon become a security risk**.

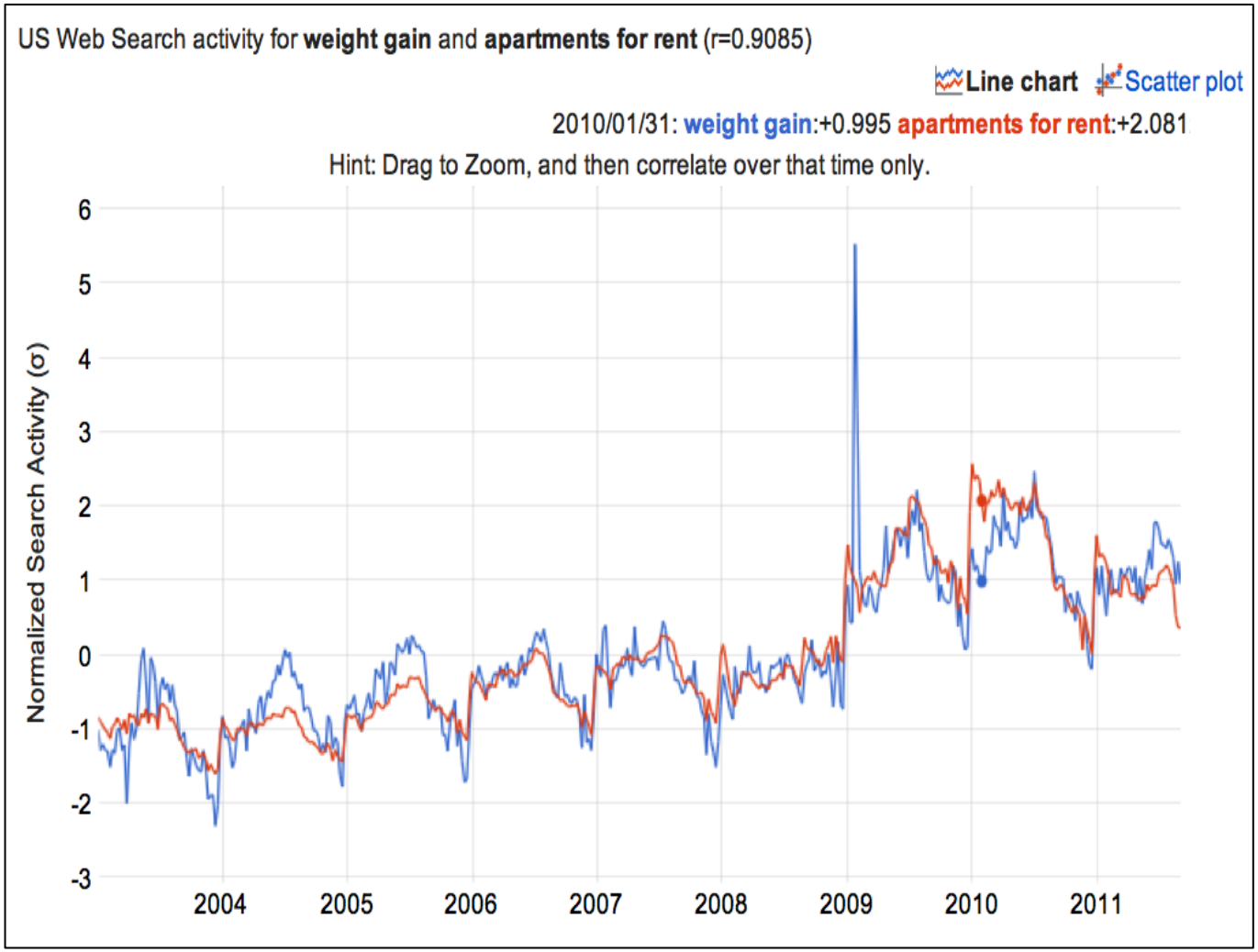
2. Analytical challenges

Challenge #1: Arrogance/naivety. Believing that the truth is in the (big) data, that the (big enough) data 'speak for themselves', that all it takes is to beep digging and mining to unveil the truth, that more is always better etc. **Apophenia**—seeing patterns where none exists. If you torture the data long enough, it will eventually start talking. S&P index and butter production in Bangladesh? (Leinweber, 2007)

2. Analytical challenges

Challenge #2: “Making sense of the data”..

Not having the right computing capacities. Unstructured data. Sentiment analysis. Translation. Sample bias. No understanding of context.



3. Operational/systemic challenges

Challenge #1: Legal. Data privacy? Intellectual property?

Challenge #2: changing systems, policies, frameworks, decision making processes

2. Risks

Risk #1: *Security and privacy* of individual and communities may be jeopardize—dark side of Big Data. Conflict zones!

Risk #2: *Creating new digital divides*, between the Small and the Big world, as well as within them**, between communities.** Who has access, who has the knowledge? Those who can answer the questions are often those who ask the questions.

Risk #3: *Respond foolishly*. i.e. in a way that may exacerbate tensions, poverty..

What to avoid

- Vertical down→top→down
- a-theoritical—no questions asked
- a-contextual
- automated, autarchic, unresponsive to the recommendations of practitioners and social scientists

Desirable principles and policies

Right intent, sound principles:

- *Big Data is a tool that can help answer QUESTIONS—**start from questions***
- *Contextualization is key—especially when lives are in the line. Rely on **local insights** to make sense of the data.*
- *Responsibility: **do not harm. Privacy and security: privacy preserving data analytics***
- *Build on existing systems and knowledge, e.g. thinking around strategic peacebuilding*
- *Think **incremental, complementary, iterative** and long term*

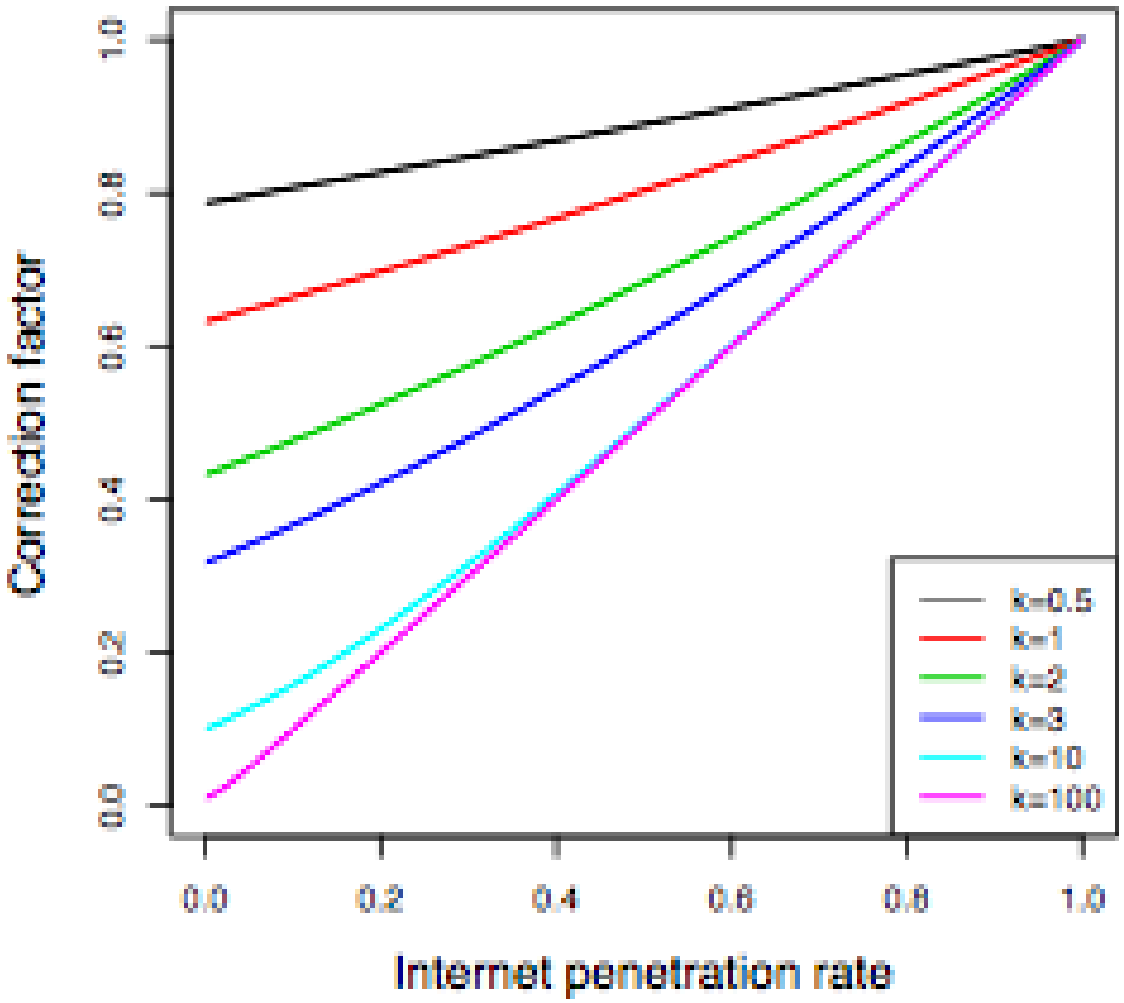
Right Capacities → Policies and systems:

- ***Bring together*** *data and social scientists, private sector and public sector, traditional statisticians and Big Data people*
- *Set up Small World-Big World interdisciplinary facilities and agreements to ask questions, test and exchange, draw lessons*
- *Create work flows as feedback loops*
- *Engage in debates on privacy, security, responsibility, design legal frameworks*

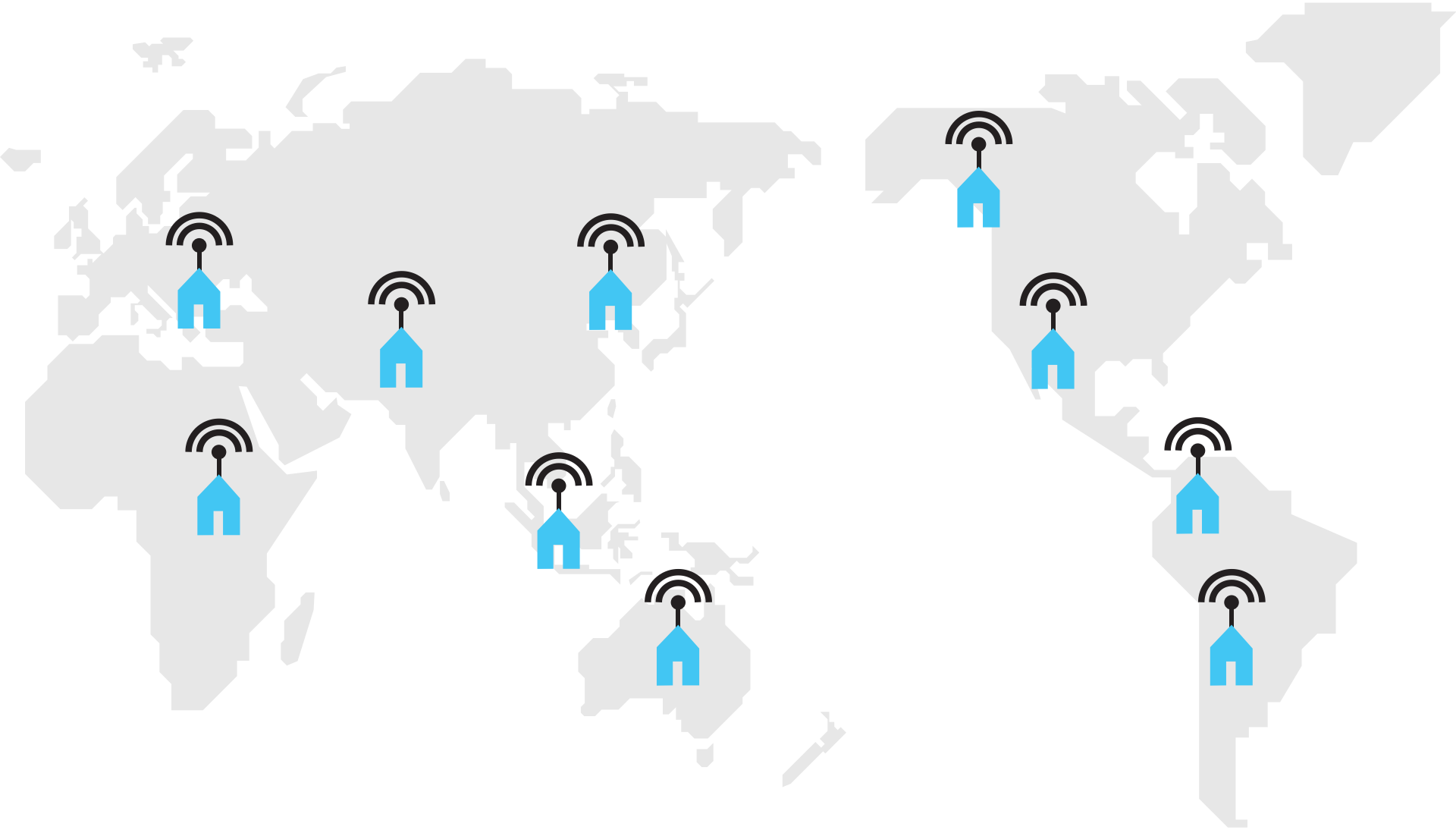
Research on big data sample bias correction exists

"You are where you E-mail: Using E-mail Data to Estimate International Migration Rates"

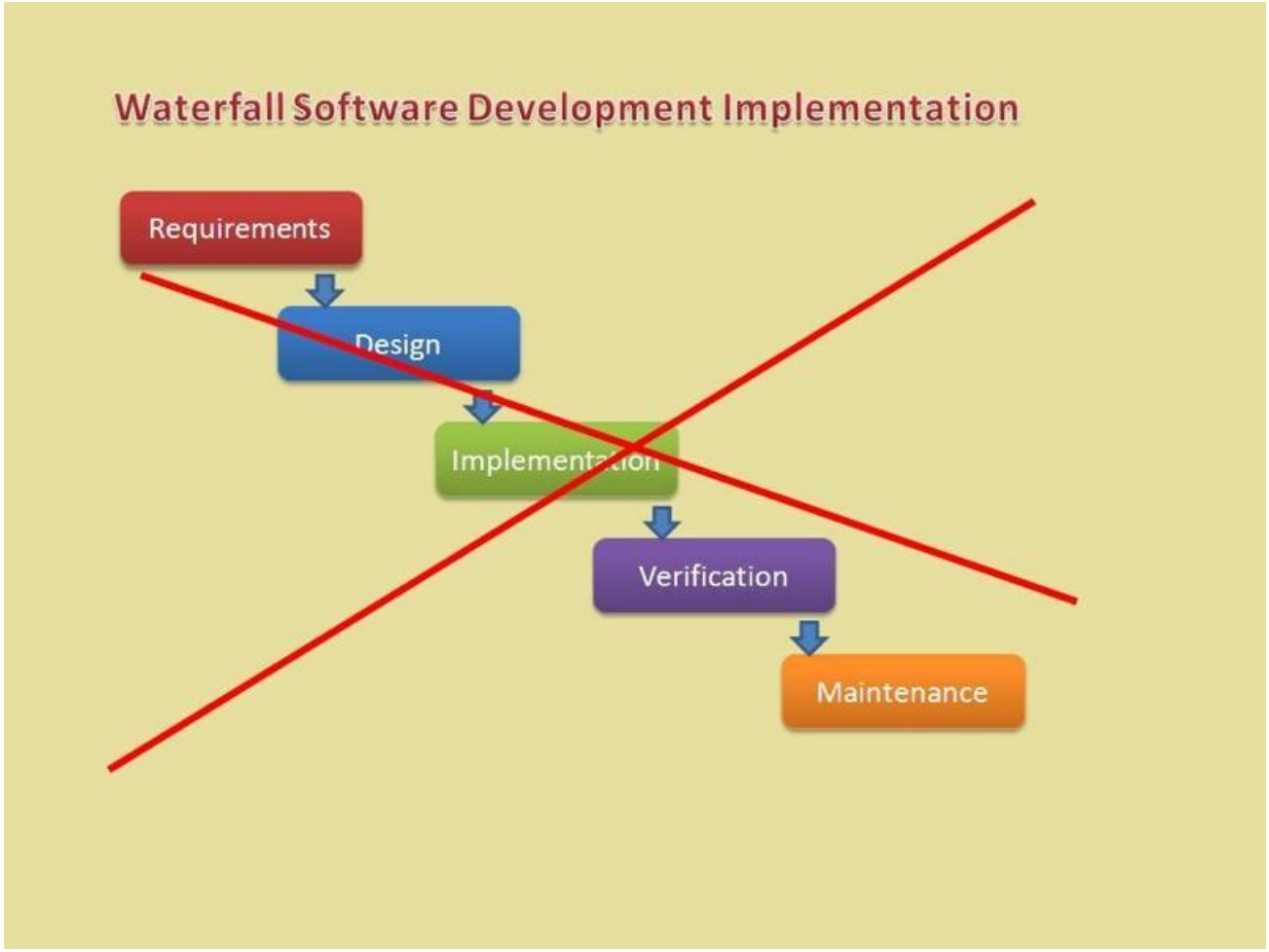
Illustrative relationship between the correction factor for selection bias and Internet penetration rates for different choices of the parameter k.



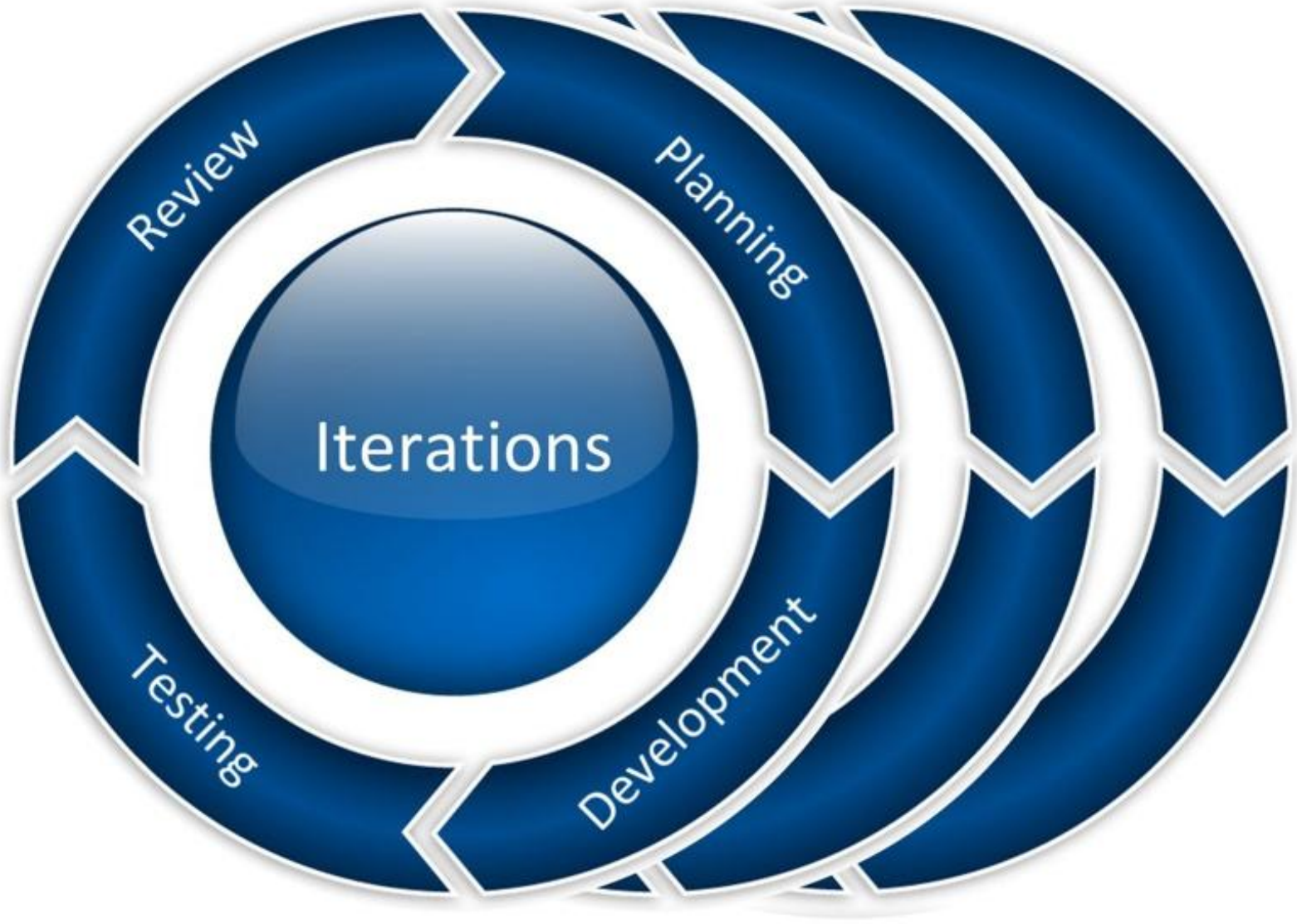

GLOBAL PULSE
GOVERNMENT
UNITED NATIONS
PRIVATE SECTOR AND CORPORATIONS
FOUNDATIONS AND NGOs
OPEN SOURCE TECHNOLOGISTS
LOCAL ENTREPRENEURS
COMMUNITY GROUPS
ACADEMIA



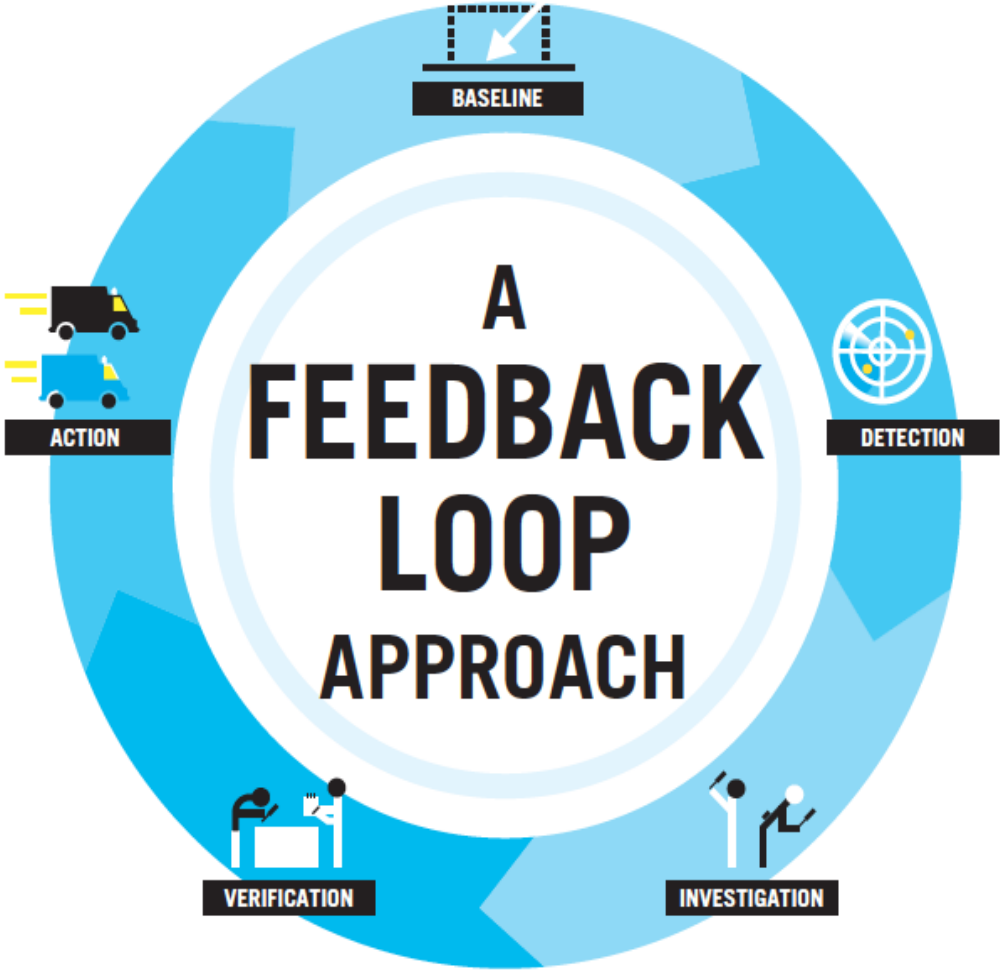
Where may this take us?
Lessons from the Technology Sector



Agile Software Development Model

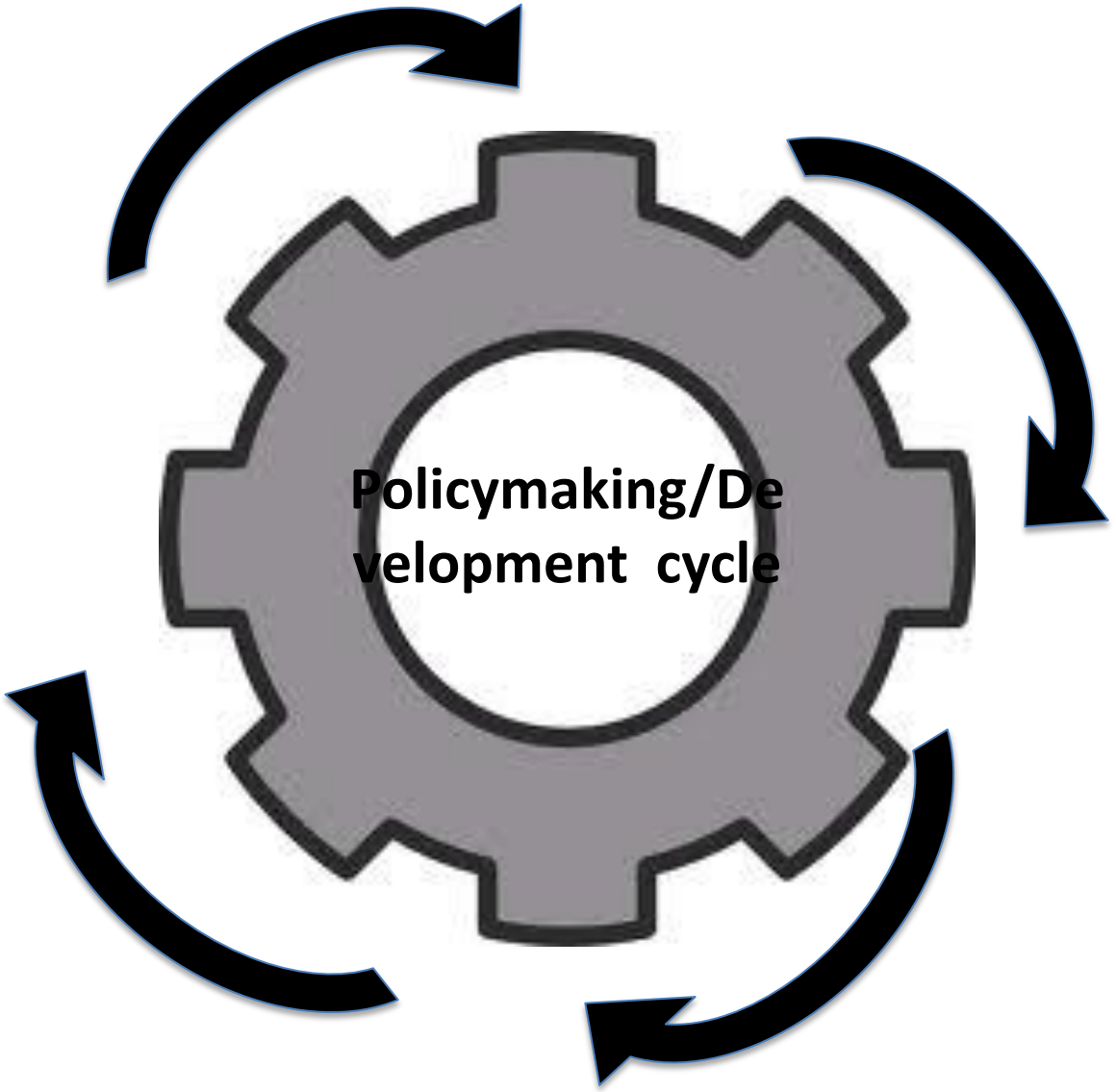


Agile Global Development?



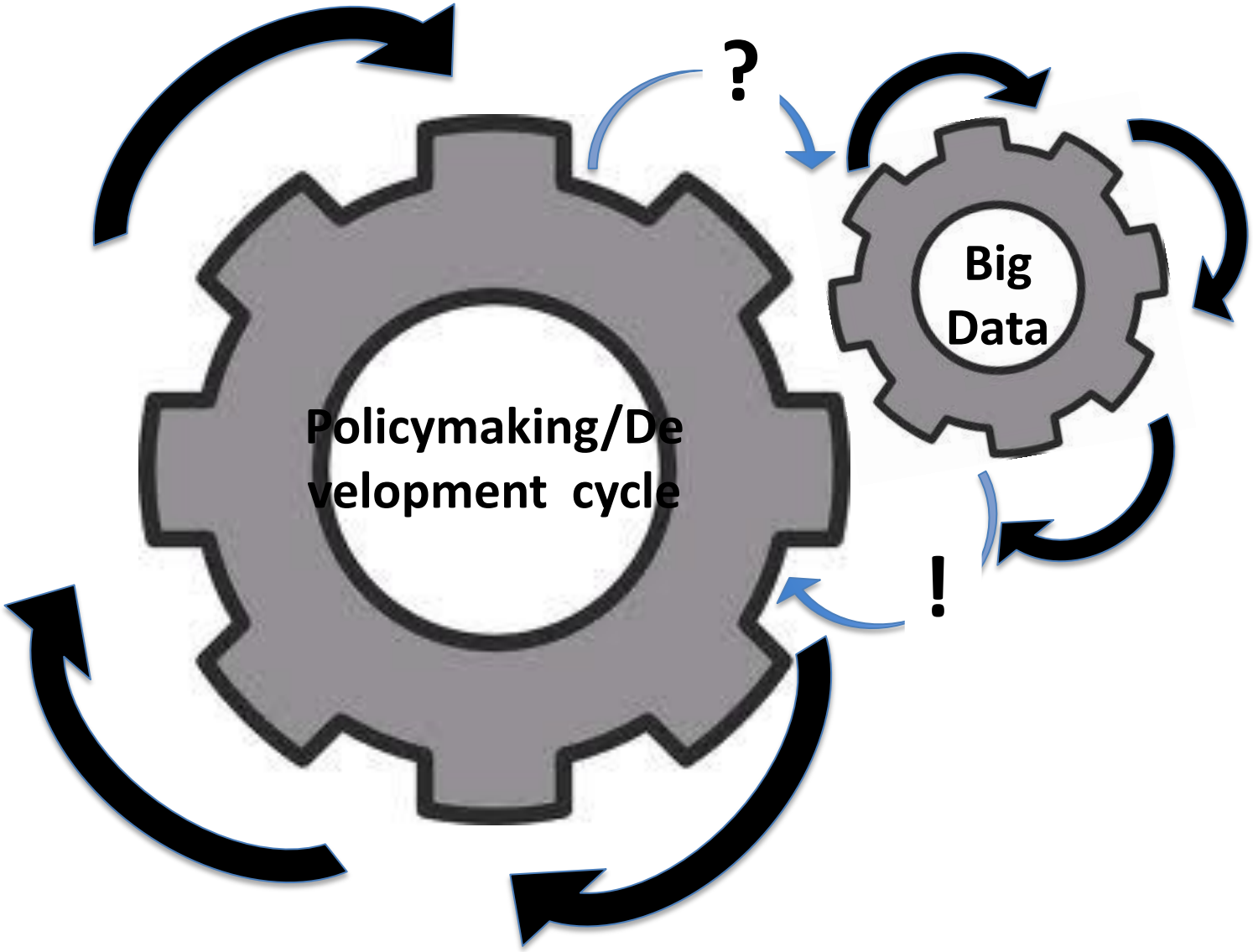
→ 2000

Traditional planning-monitoring policymaking cycle



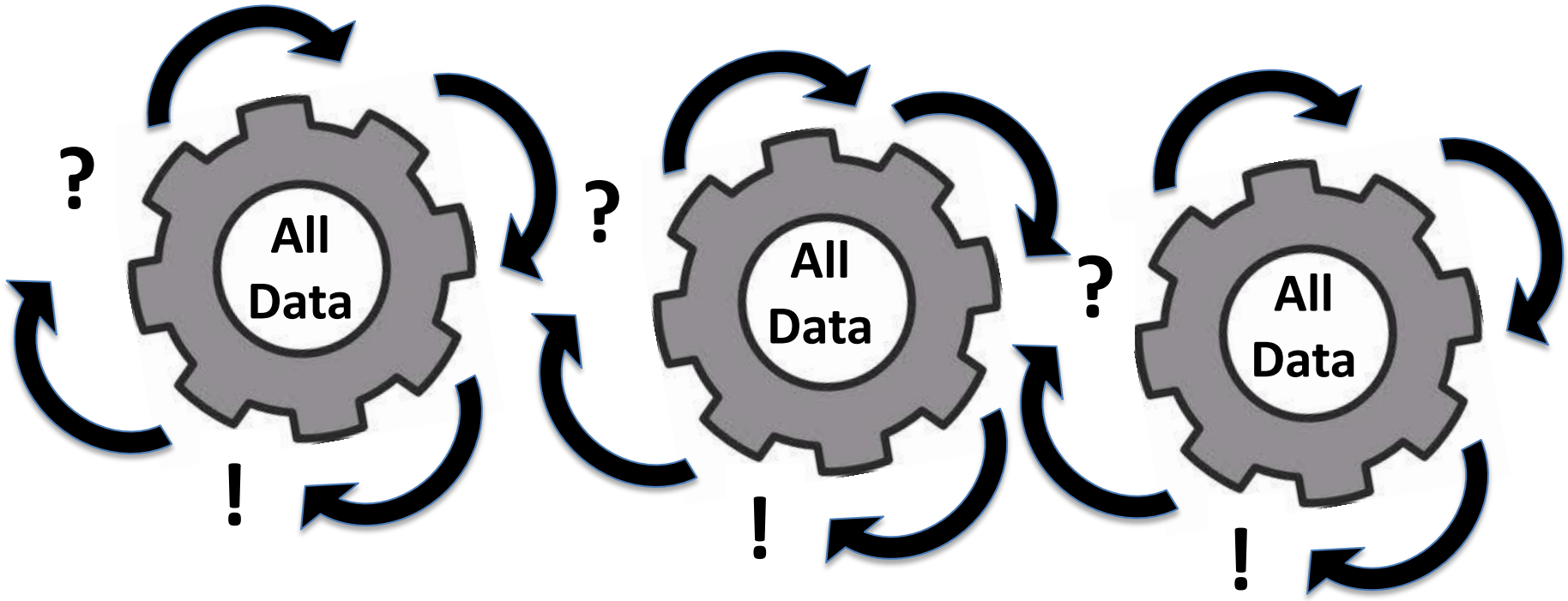
→ 2020

Responsive policymaking cycle?



2020→

Agile policymaking cycle?



10 years ahead?

Data streams will continue to have grown in volume, velocity and variety. It is very unlikely that Big Data won't be playing a significant role in Development. Question is: what do we want to achieve and how?

Big Data for Development should start small, and grow.

Thanks