BigSurv18 – A Conference and a Monograph

Lars Lyberg, Inizio

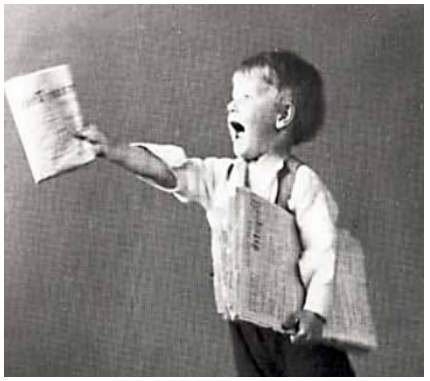
Presentation at Frimis, November 28, 2018

INIZIO


BigSurv18
BIG DATA MEETS SURVEY SCIENCE

Top stories - Read all about them

1. Anyone can take pictures of you from a satellite and there is nothing you can do about it
2. Better data can drive travel and conference savings
3. The Health and Human Services Department will mine data from its internal social network
4. Vietnam taps Big Data to avoid China's traffic chaos
5. Tweets can foretell votes



A black swan

A black swan is an undirected and unpredicted event.

It is rare, has an extreme impact but in retrospect we saw it coming

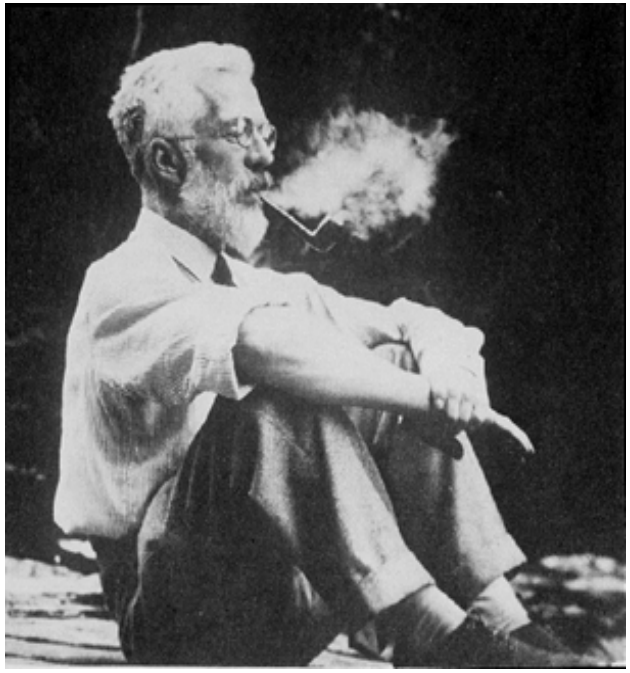
- Internet - yes
- 9/11 - yes
- The Lehman Brothers crash – yes
- Decreasing response rates-yes
- The advent of Big Data –not really
- New data sources other than BD-yes
- Nonprobability sampling-yes and no

Monograph Contents

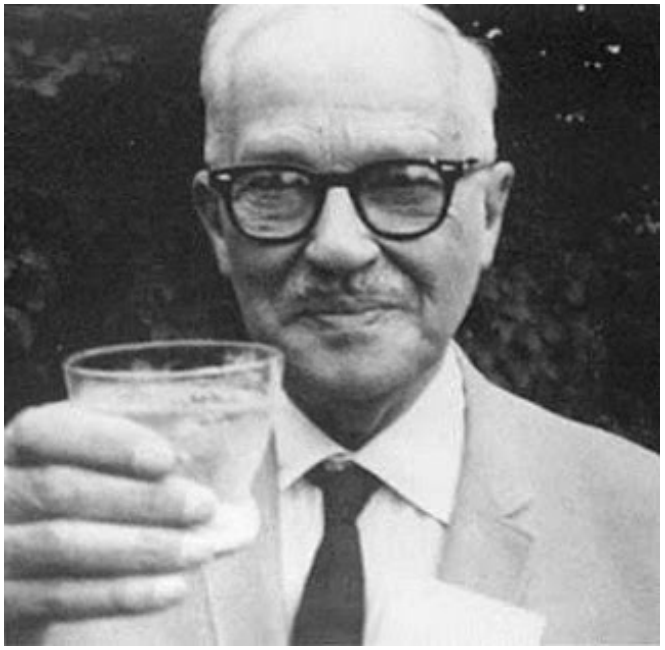
- The new survey landscape
- Total error and data quality
- Big data in official statistics
- Combining big data with survey statistics: methods and applications
- Combining big data with survey statistics: tools
- Regulations, ethics, privacy

A Couple of Giants

Sir Ronald Fisher



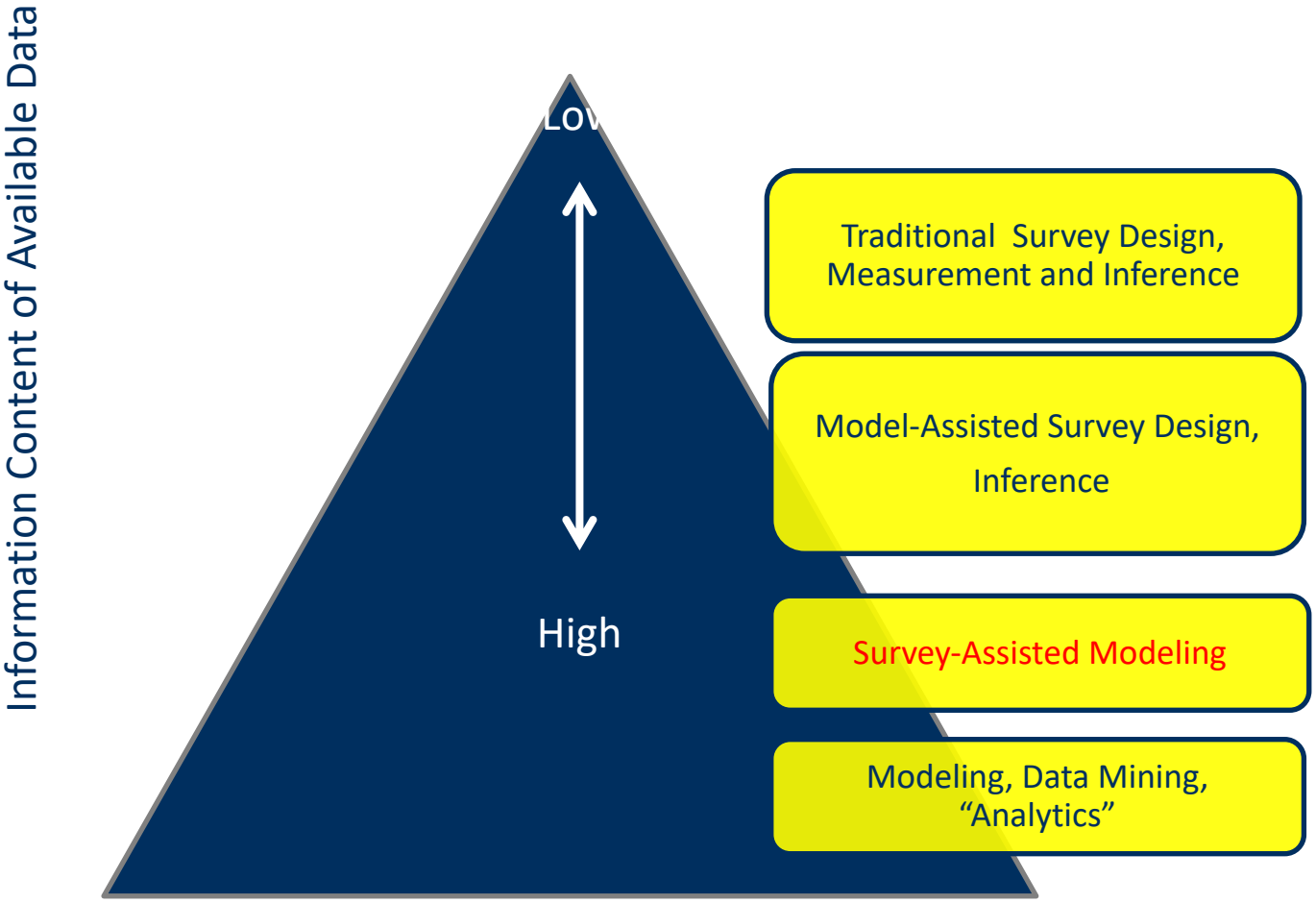
Jerzy Neyman



A Stunning Statement Made by One of the Keynote Speakers

Surveys are the last resort

Design, Measurement, Inference Adaptation to Available Data Resource



Source: Heeringa 2018

Why Did Data Become "BIG"

- Technological advances associated with data science and computational tools and methods.

- Information-based Decision Making
  - "Evidence-based", "Data-Driven", "Analytics", "Machine Learning"

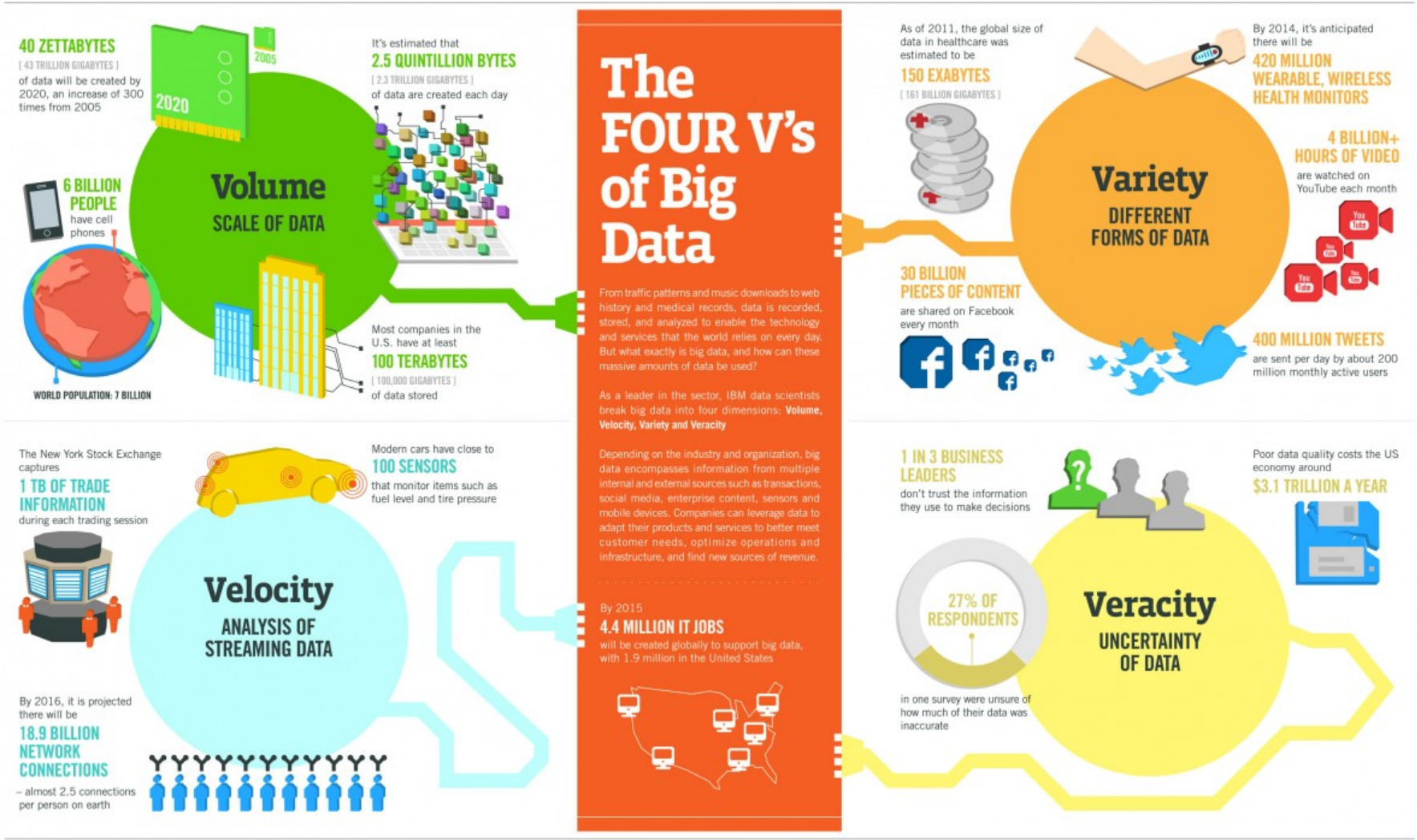
- Focus on short-run prediction
  - Business decision making
  - Health risks (e.g. Google Flu)
  - Financial markets
  - Political processes

- Style points: "Tail Fins"

Source: Heeringa 2018

INIZIO

The “V” Taxonomy for Big Data: 3→4→5(?, Variability, Value)



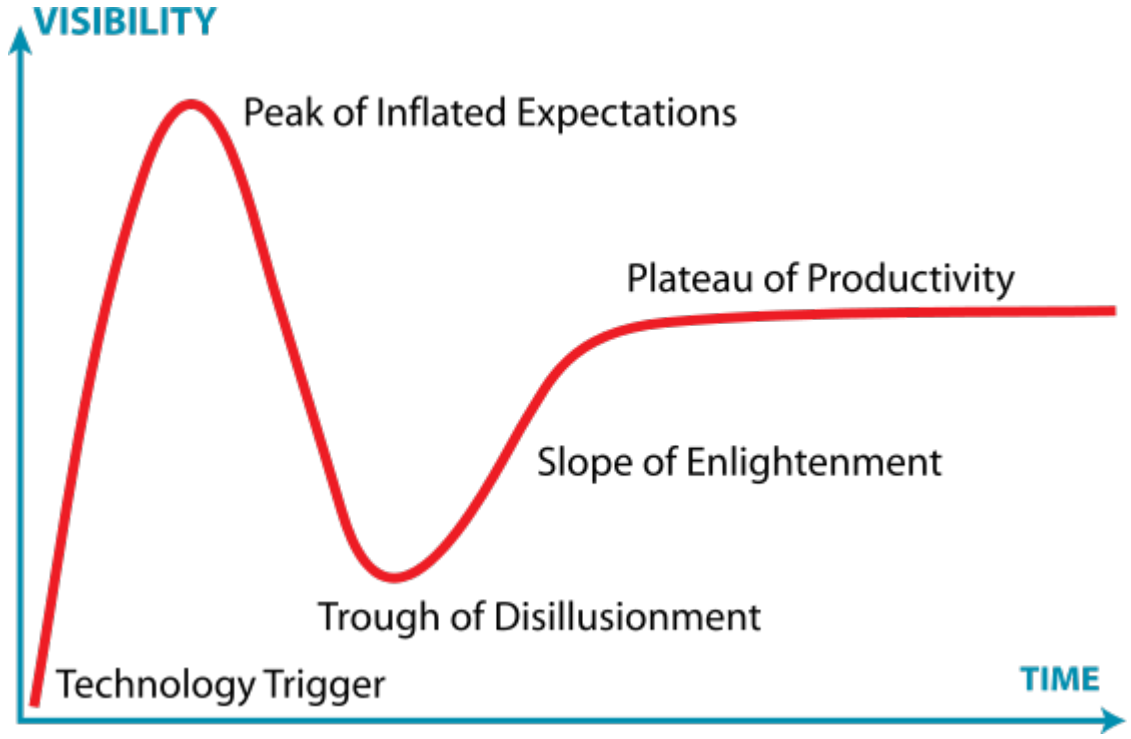
INIZIO

Some Concepts

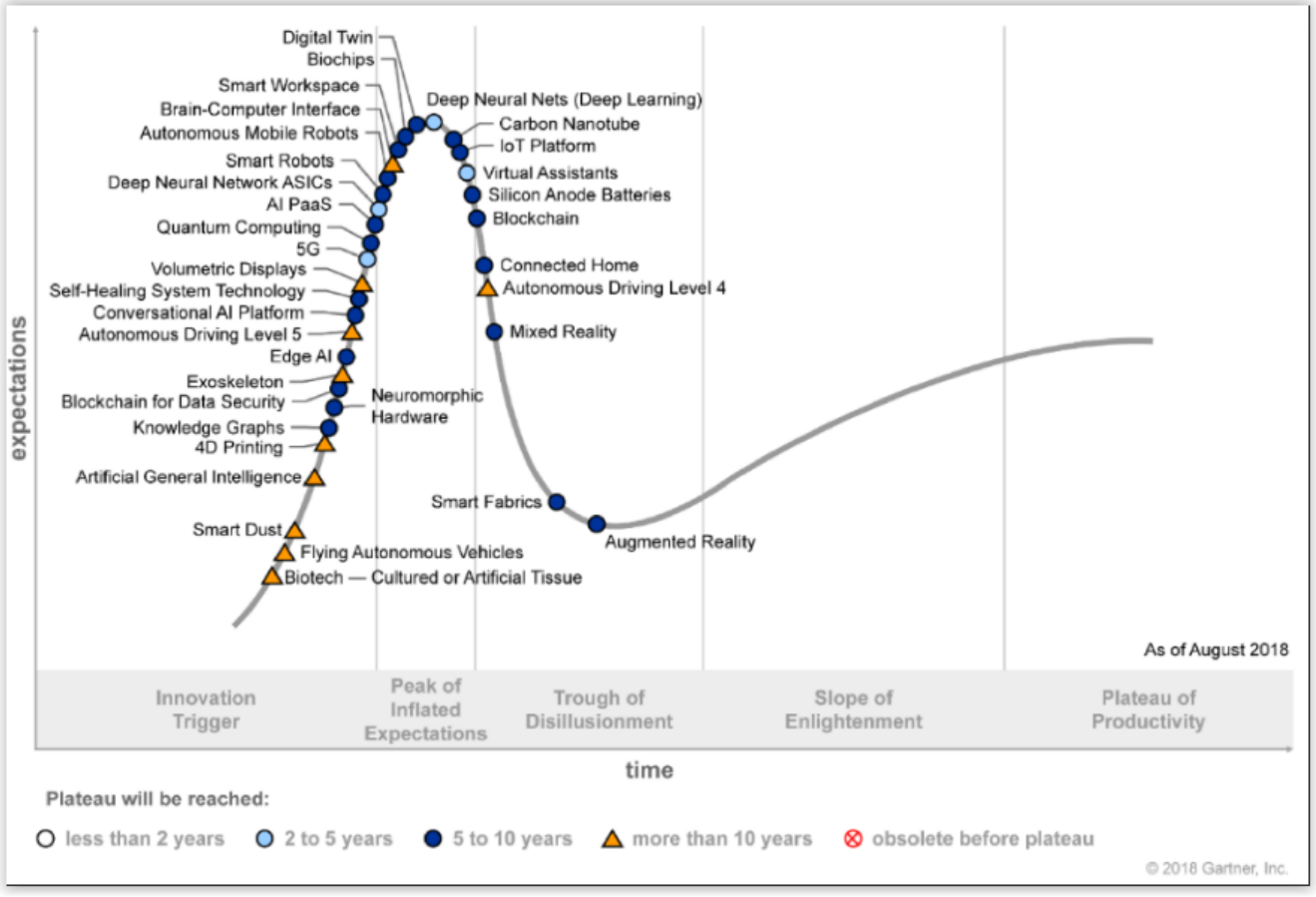
- Artificial Intelligence-machines being able to carry out tasks in a smart way
- Machine Learning-application of AI where we give machines access to data and let them learn for themselves via neural networks and natural language processing
- Data Mining-builds intuition about what is really happening in some data
- Data Science-combines the application of computer science, statistics, programming and business management

Hype of Big Data

Gartner's hype curve



Source: Wikipedia


Digital Twin
Biochips
Smart Workspace
Brain-Computer Interface
Autonomous Mobile Robots
Smart Robots
Deep Neural Network ASICs
AI PaaS
Quantum Computing
5G
Volumetric Displays
Self-Healing System Technology
Conversational AI Platform
Autonomous Driving Level 5
Edge AI
Exoskeleton
Blockchain for Data Security
Knowledge Graphs
4D Printing
Artificial General Intelligence
Smart Dust
Flying Autonomous Vehicles
Biotech — Cultured or Artificial Tissue
Neuromorphic Hardware
Deep Neural Nets (Deep Learning)
Carbon Nanotube
IoT Platform
Virtual Assistants
Silicon Anode Batteries
Blockchain
Connected Home
Autonomous Driving Level 4
Mixed Reality
Smart Fabrics
Augmented Reality
expectations
As of August 2018
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
© 2018 Gartner, Inc.

An example of Big Data analytics

- Evolv has a database containing information on 984 000 hourly workers in 20 companies
- Data sources: online employee background checks, time and attendance tracking software, and performance ranking programs
- Selected results:
  - Employees with a criminal record performed slightly better than others
  - Experienced employees did no better than the inexperienced

An example of Big Data analytics (cont'd)

- Potential problems:
  - Weak conceptualization
  - Not clear if data sources are compatible across clients
  - No background variables
  - No questions asked to real persons
  - Inference is based on big rather than theory
  - The analysis might be business-driven

Having said that:

How can data on 984 000 workers be used effectively?

Happiness and Well-being

The common survey question: How satisfied are you with your life?

BD alternative

- 10 million tweets that are coded for happiness (rainbow, love, beauty, hope, wonderful, wine…) and non-happiness (damn, boo, ugly, smoke, hate, lied,…)
- Happiest states: Hawaii, Utah, Idaho, Maine, Washington
- Saddest states: Louisiana, Mississippi, Maryland, Michigan, Delaware

The Potential Use of Big Data in Statistics Production

- Produce statistics based on BD that can replace surveys
- Combine BD with admin data, sample surveys, and nonprobability sources in order to improve statistics
- Explore new topics and concepts
- Data mining to identify new patterns and models

INIZIO

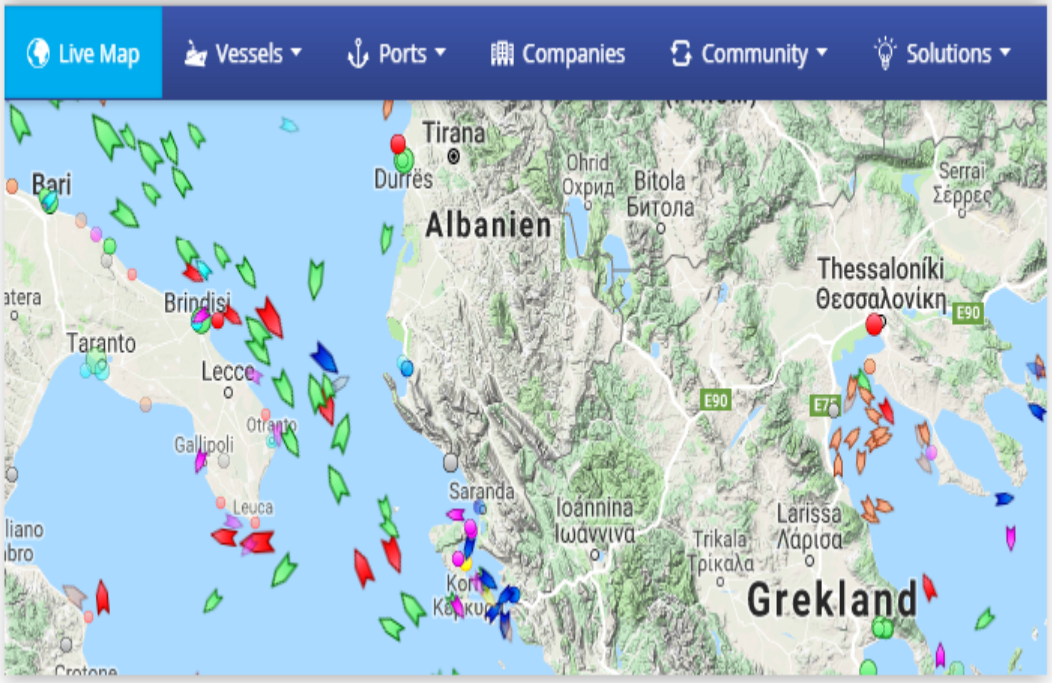
Examples of Sources of Data

- Censuses
- Other survey programs
- Administrative data systems
- Medical records systems
- Commercially compiled data
- Financial data
- Satellite imagery
- GPS and GIS
- Social media
- Mobile devices
- Wearable measurement devices
- Sensors (Internet of Things)
- Visual data: pictures and video
- Genetic profile data
- Transactional data systems

Adapted from Heeringa 2018

AIS data

- AIS - Automatic Identification System
- Data can be used to follow all vessels
- Messages from vessels are transmitted with high speed
- Monitor marine traffic in real time



Source: www.marinetraffic.com

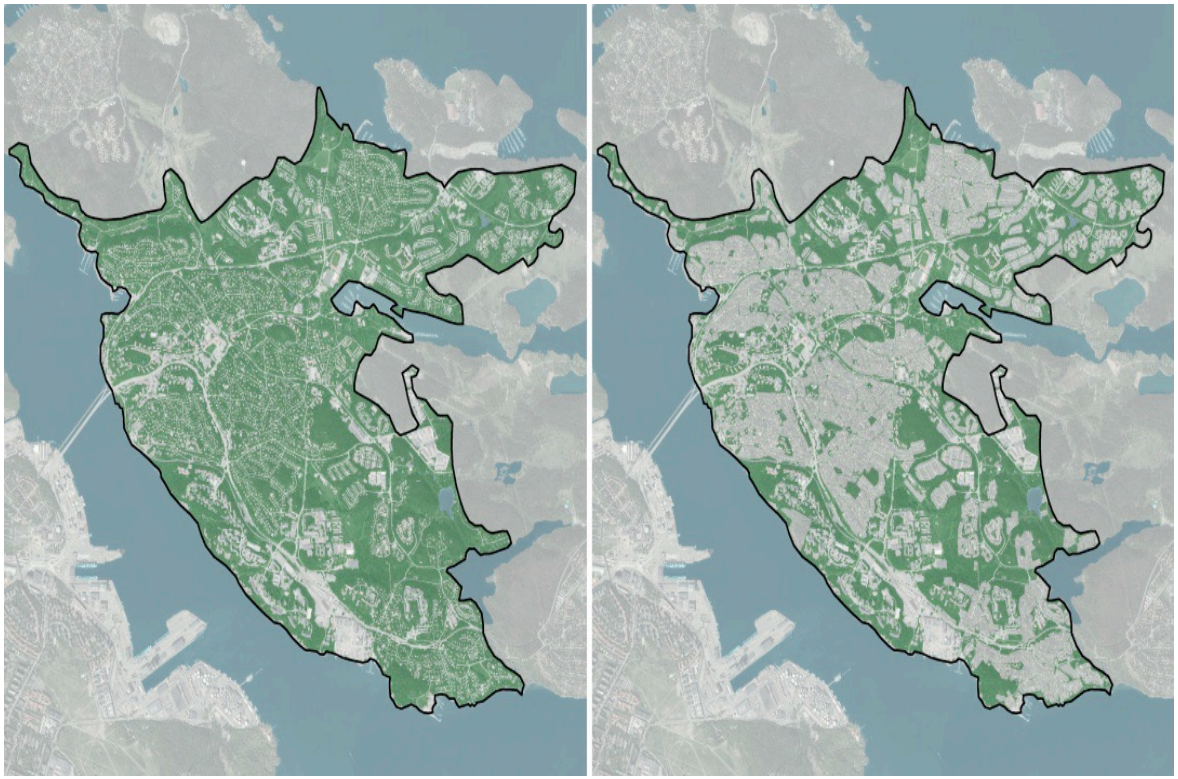
INIZIO

Improve current statistics and produce new statistics (de Wit et al 2017)

- Statistics on marine traffic to estimate emission and identify areas with heavy traffic
- Port statistics to monitor how vessels move between different ports

INIZIO

Combining different data sources- example from Statistics Sweden



The left map shows green areas in Lidingö, Stockholm Sweden. Combining the data with data from the real estate register we get the green areas that are accessible to the public in general (the right map).

***Source*: SCB. Bakgrundskarta © Lantmäteriet**

INIZIO

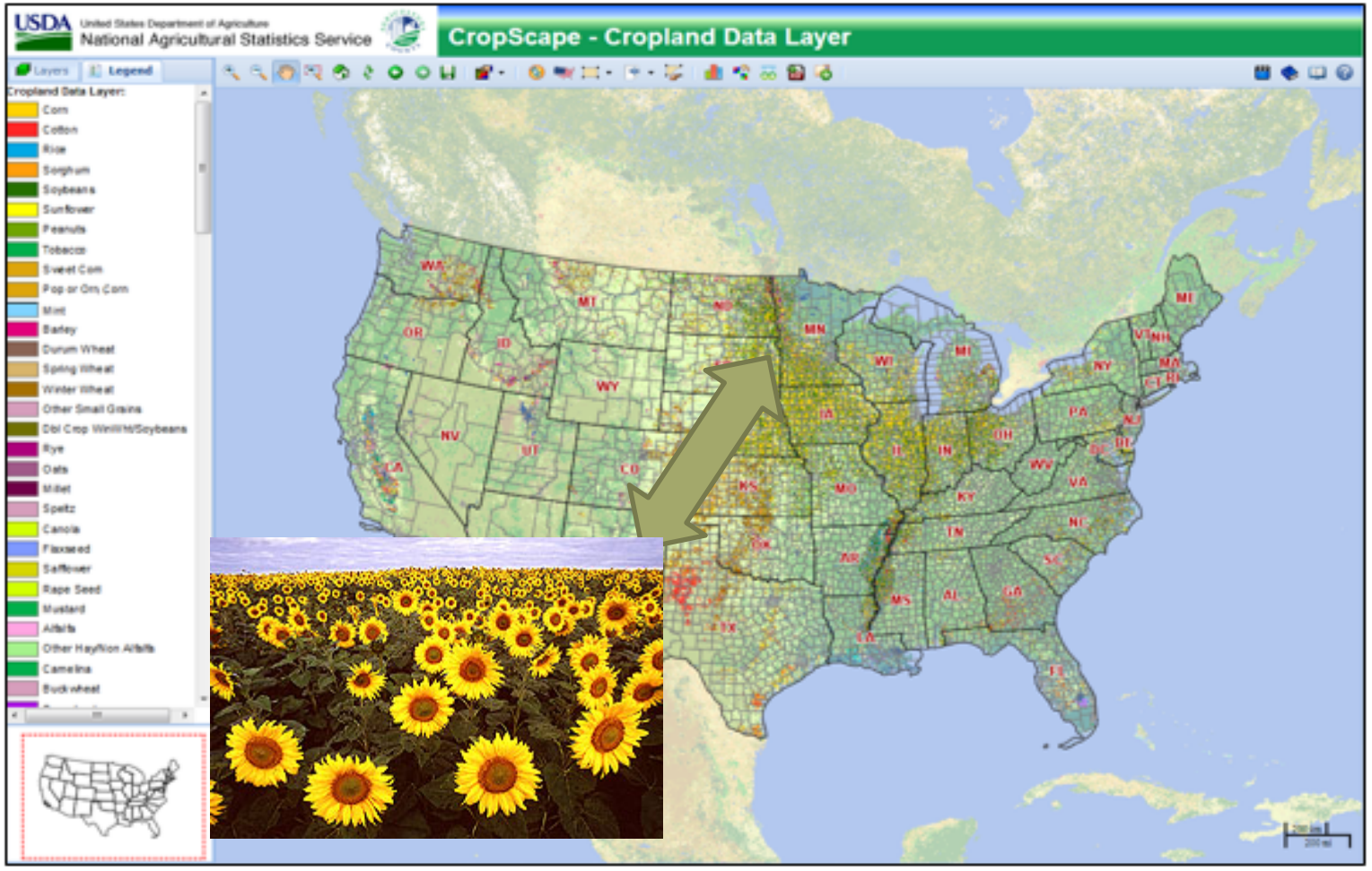
Combining Different Data Sources- Example from Statistics Netherlands

- Solar energy - power use estimates:
  - transmission grid load,
  - metrological data,
  - areal images,
  - electricity meter readings, and
  - energy efficient home improvements.

- How to combine these data sources? That's the question.

INIZIO

Satellite imagery: LANDSAT Crop Layer



80 acre sun flower field. Fargo, North Dakota.

Source: USDA

New set of legal and ethical issues in the big data era

- Data are often collected for one purpose but combined with other data sources and used for another purpose
- Risk of privacy and confidentiality breaches
- The old way – to get access to survey and administrative data by using statistical disclosure control techniques and provision of controlled access through research data centers.
- The new way – unclear legal situation – who owns the new type of data?
- Consent statements that foresee all potential future use of data - too complex for anyone to grasp
- There are no data stewards controlling access to individual data
- GDPR does not explicitly mention BD

INIZIO

Take-away points

- New survey developments are taking place
- Our industry needs innovations, less fighting and more collaboration
- We need to merge with other research cultures
- We need to know more about combining data sources
- We need to account for all major sources of uncertainty that are associated with data collection and analysis of data
- We need to develop new theories for handling error structures and combining data sources

Over and Out