Przemysław Polak
Warsaw School of Economics

The phenomenon of big data: dangers and possible applications

Big data - definitions

- a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data processing application software
- data sets characterized by huge amounts (**volume**) of frequently updated data (**velocity**) in various formats, such as numeric, textual, or images/videos (**variety**) – 3V model
- the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data
- Big data is where parallel computing tools are needed to handle data. It represents a distinct and clearly defined change in the computer science used, via parallel programming theories, and losses of some of the guarantees and capabilities made by Codd's relational model.

Big data – key features

- huge volume of data
- unstructured and semi-structured data
- various sources of data (also Internet!)
- finding patterns
- extracting new knowledge
- instead of searching for causality, finding correlations: not knowing **why** but only **what**

Finding patterns Walmart case

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data — the equivalent of 167 times the information contained in all the books in the US Library of Congress
- Walmart finds patterns that can be used to provide product recommendations to users based on which products were bought together or which products were bought before the purchase of a particular product

Finding patterns Walmart case

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data — the equivalent of 167 times the information contained in all the books in the US Library of Congress
- Walmart finds patterns that can be used to provide product recommendations to users based on which products were bought together or which products were bought before the purchase of a particular product

It’s NOT Big data

But when added

- Tracking customers
- Opinions from social media
- Information from Video Surveillance Systems

Detecting influenza epidemics

- US doctors were requested to report new flu cases to the CDC (Centers for Disease Control and Prevention)
- But the data available for analysis were one-two weeks out of date
  - People may wait days before consulting a doctor
  - Data were reported once a week

Detecting influenza epidemics

- Google could detect the spread of flu in near real time by looking at what people were searching on Internet
- Google compared the 50 millions most common search terms with the CDC data on the spread of flu in previous years
- They found a combination of 45 search terms that had a strong correlation with the spread of flu

Cows and deers sense Earth's magnetism

- Farmers know that cows in herds tend to face the same way when grazing (suspected reasons: wind & sun)
- Source of information: satellite images from Google Earth
- Finding: cows and deers tend to face either magnetic north or south when grazing or resting
- A major influence from the sun was ruled out: the satellite images also recorded the position of the animals' shadows
- Influence of wind: Google Earth images aren't time-stamped with enough accuracy for the researchers to be able to compare the shots with weather data, but according to the maps of prevailing winds – it's very unlikly

Project:
OptiMAL (Optical methods for Marine Litter detection)

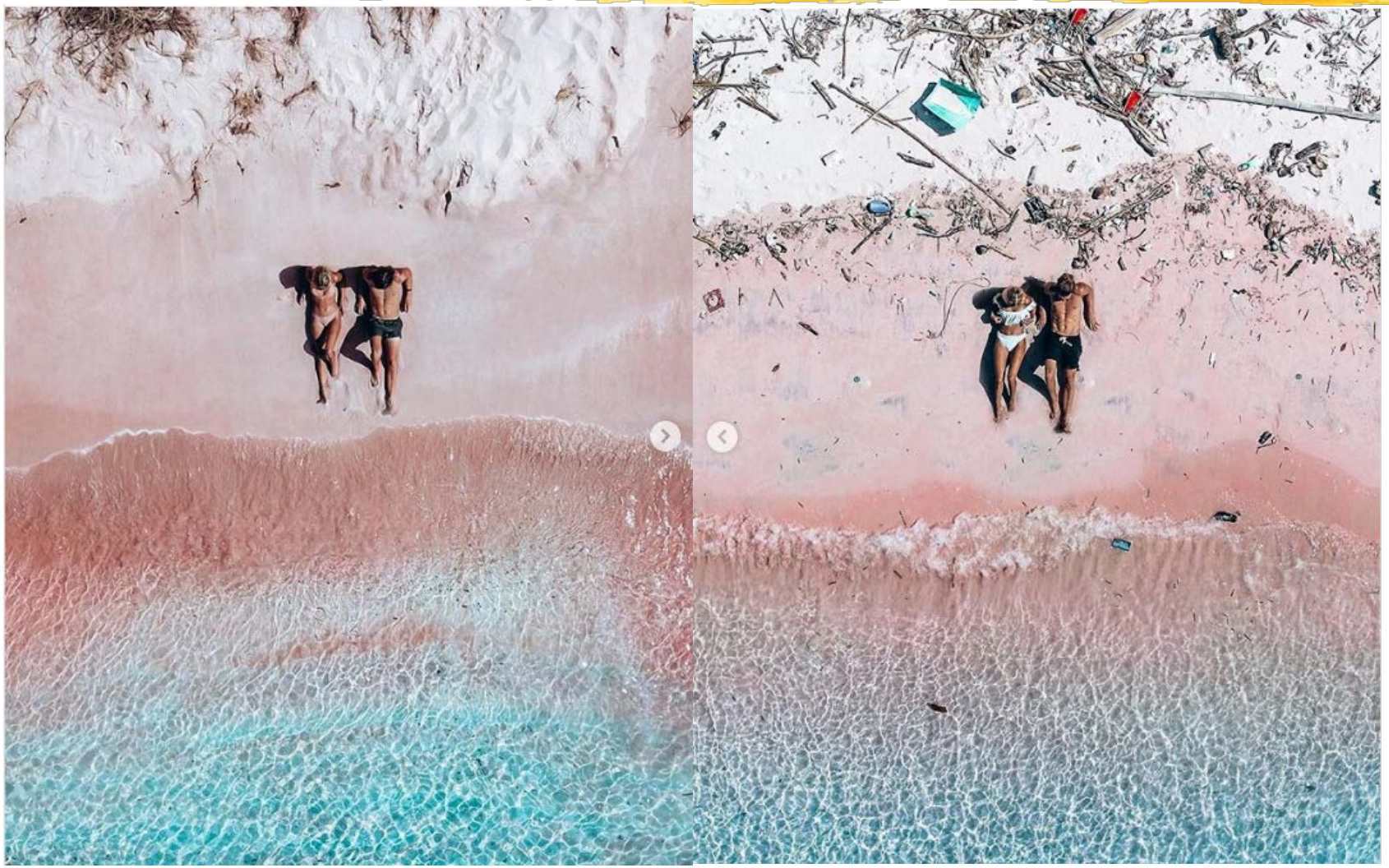
- The European Space Agency is developing technology to allow satellites to identify the concentration, movement and origin of plastic debris across the world's oceans
- deploying satellites to monitor marine litter on a global scale could give researchers working on plastic pollution data about its abundance, concentrations and movement

Project:
OptiMAL (Optical methods for Marine Litter detection)

- The European Space Agency is developing technology to allow satellites to identify the concentration, movement and origin of plastic debris across the world's oceans
- deploying satellites to monitor marine litter on a global scale could give researchers working on plastic pollution data about its abundance, concentrations and movement

How much can it cost?

When I saw that ...



Groupwork

- Propose a draft idea of a research related to sustainable development using big data
  - tracing sth.
  - searching for patterns and correlations
- Indicate potentials sources of information, types of information

Thank you for your attention!

Przemysław Polak, Ph.D

Warsaw School of Economics

Al. Niepodległości 162

02-554 Warszawa, Poland

e-mail: ppolak@sgh.waw.pl

www.sgh.waw.pl