Big Data:
Status quo + quo vadis

Stanisław Matwin

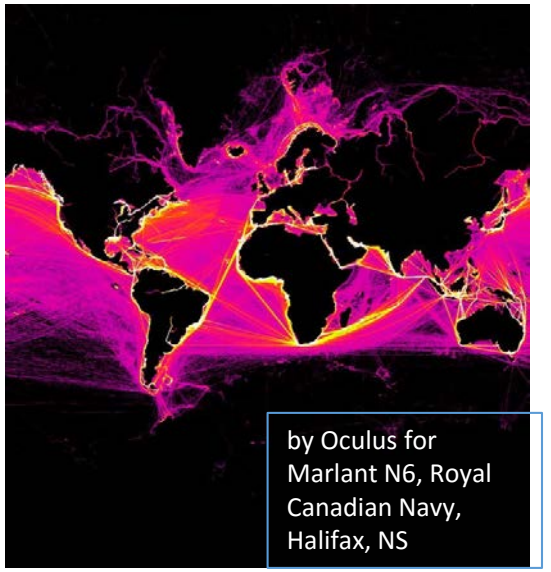
stan@cs.dal.ca

Plan

- Próba definicji + uwagi
- Przykład zastosowania
- Nieco historii
- Big Data meets Big Water
- Wyzwania
  - korelacja-przyczynowość/interpretowalność
  - do kogo naprawdę należą dane?
  - prywatność

Automatic Identification System (AIS)

IMO/ITU standard



Courtesy of ExactEarth, Inc.

Big Data – 5 Vs

- Volume
- Velocity
- Variety
- Veracity
- Value


TYGODNIK POWSZECHNY

MOJE STRONY | (STANMATWIN)

AUTORZY | BLOGI | ARCHIWUM | REDAKCJA | WYDAWCA | FUNDACJA | PRENUM

MENU

STRONA GŁÓWNA » NAUKA » ZŁUDZENIE BIG DATA

ZŁUDZENIE BIG DATA

MIROSŁAW SZREDER | 28.02.2016

Czy wielka ilość informacji, do których mamy dziś dostęp, a także możliwości przetwarzania ich na skalę dotąd niespotykaną, przekładają się na naszą bardziej wartościową wiedzę?

Nieco inne spojrzenie:

- Dane jako wartość
- Dane często z sensorów
- Integracja różnych rodzajów danych
- Użycie metod ML do budowy modeli

„Memory law”

- Usama Fayad: objętość pamięci które wypełniamy danymi podwaja się co 9 miesięcy (wobec 18 miesięcy dla szybkości procesorów [Moore])
- Oznacza to że

Nasza zdolność do przetworzenia wszystkich spływających danych maleje wykładniczo

Technologie umożliwiające big data:

- Uzyskiwanie danych
- Analiza danych
  - Machine Learning
  - Data mining
  - Statystyka

- HPC
  - Hadoop
- Bazy danych
  - NOSQL
  - Streaming
- Wizualizacja

http://hint.fm/wind/ [Martin Wattenberg]

Wartość danych AIS

- Monitoring statków
- Rybołówstwo
- Wycieki ropy naftowej
- Analiza ryzyka
- ….

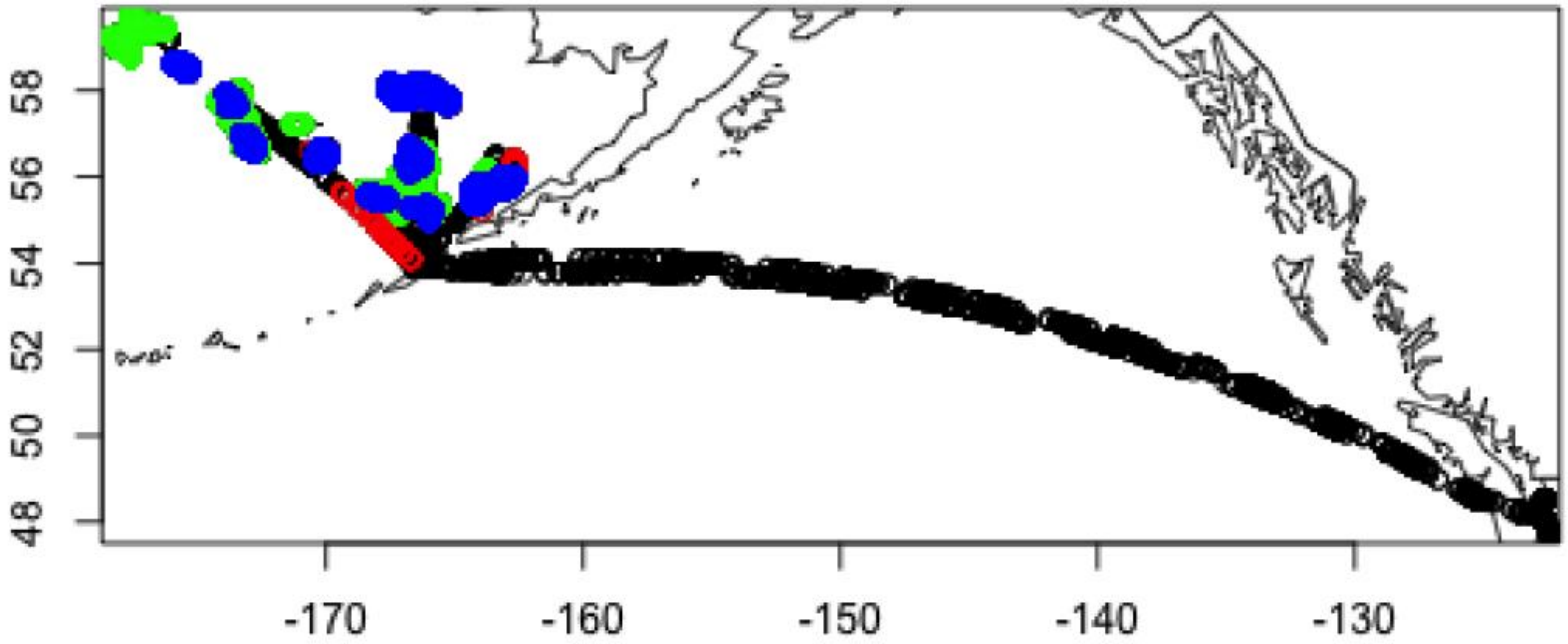
AIS - ilościowo

- 400,000 statków
- 100M rekordów dziennie
- Luki w danych:
  - Technologiczne
  - Błąd ludzki
  - Manipulacja

Zastosowanie AIS – ekologia morza

- Cel: globalny wgląd w ekploatację ryb
- Marine Protected Areas
- Biologowie chcą móc ustalić, na podstawie trajektorii statku:
  - typ statku/rodzaj sieci
  - czy w danym punkcie prowadzi odłów

Long liner results visualization

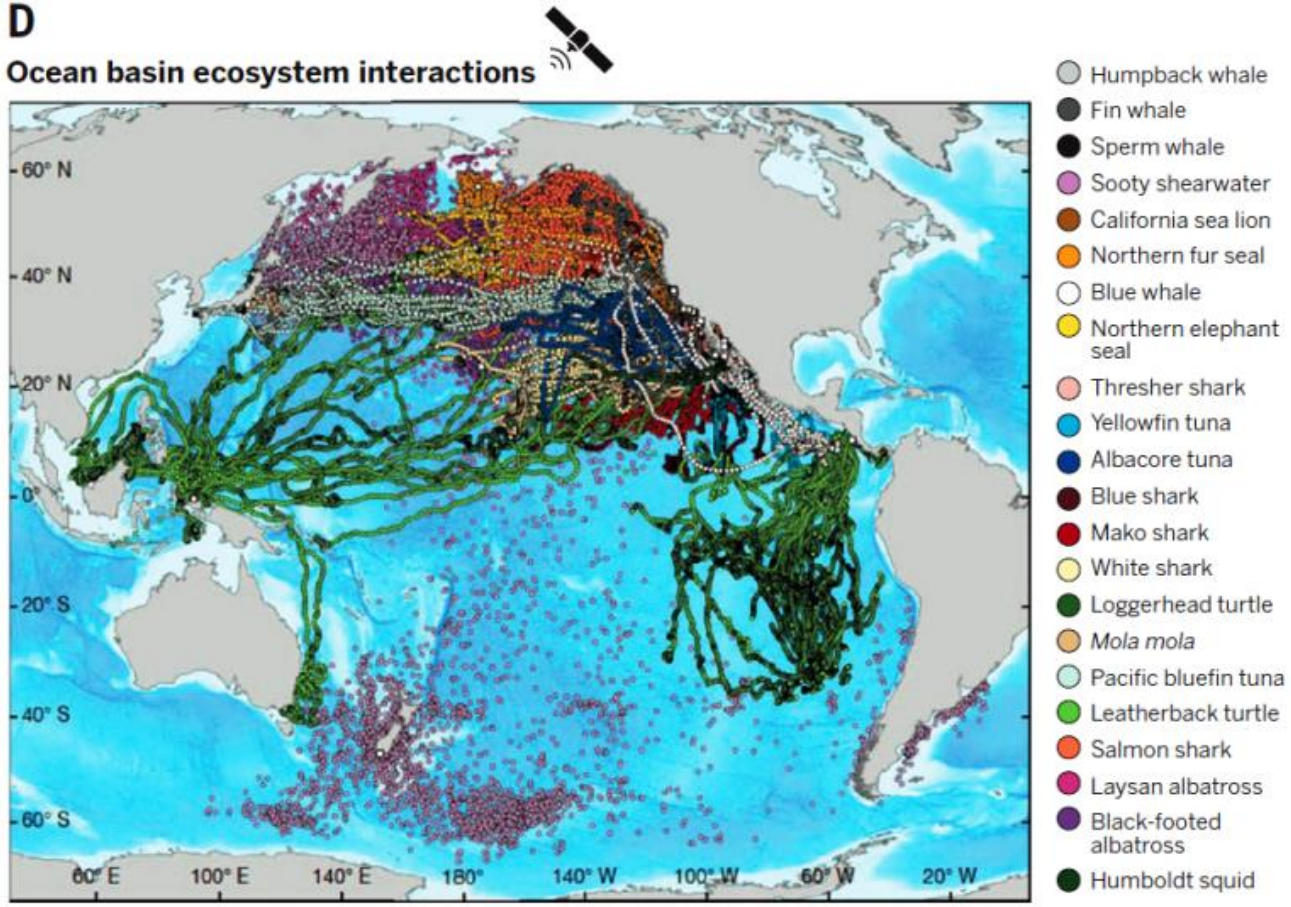


(d) Resutls long liner vessel number 2 (Accuracy: 54%).

Fishing – expert and algorithm
Fishing – expert label
Fishing – algorithm label
Non-fishing - expert and algorithm

de Souza, Boerder, Matwin, Worm Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning
PLoS One, to appear

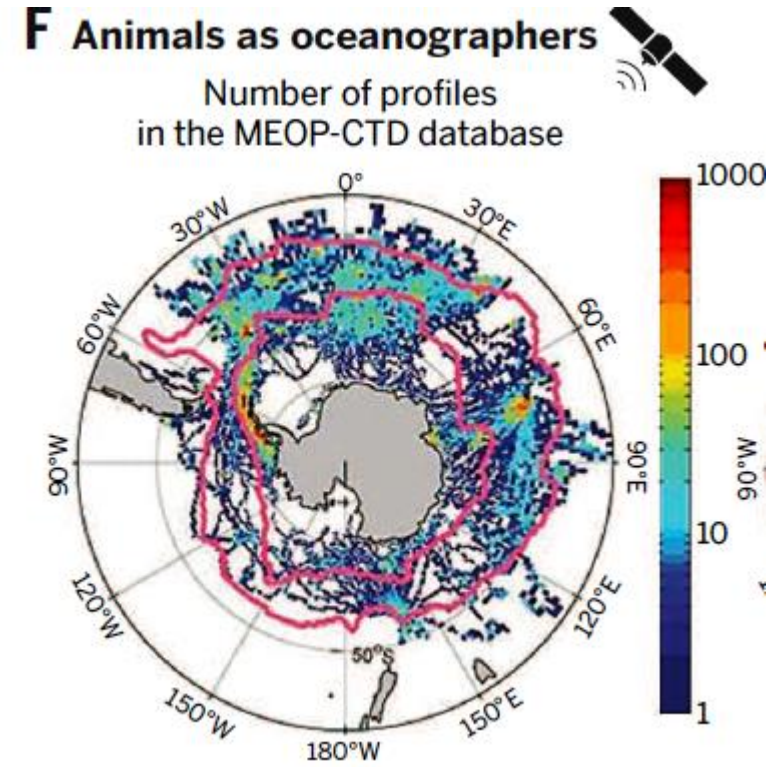
Interakcja ekosystemów na podstawie danych akustycznych



Telemetryczne trajektorie mobilności 23. gatunków drapieżników morskich na Oceanie Spokojnym

Hussey et al.
Aquatic animal telemetry: A panoramic window into the underwater world
Science, 12 Jun 2015:Vol. 348

Zwierzęta jako oceanografowie



Foki z wszczepionymi instrumentami pomiaru przewodności/temperatury/głębokości

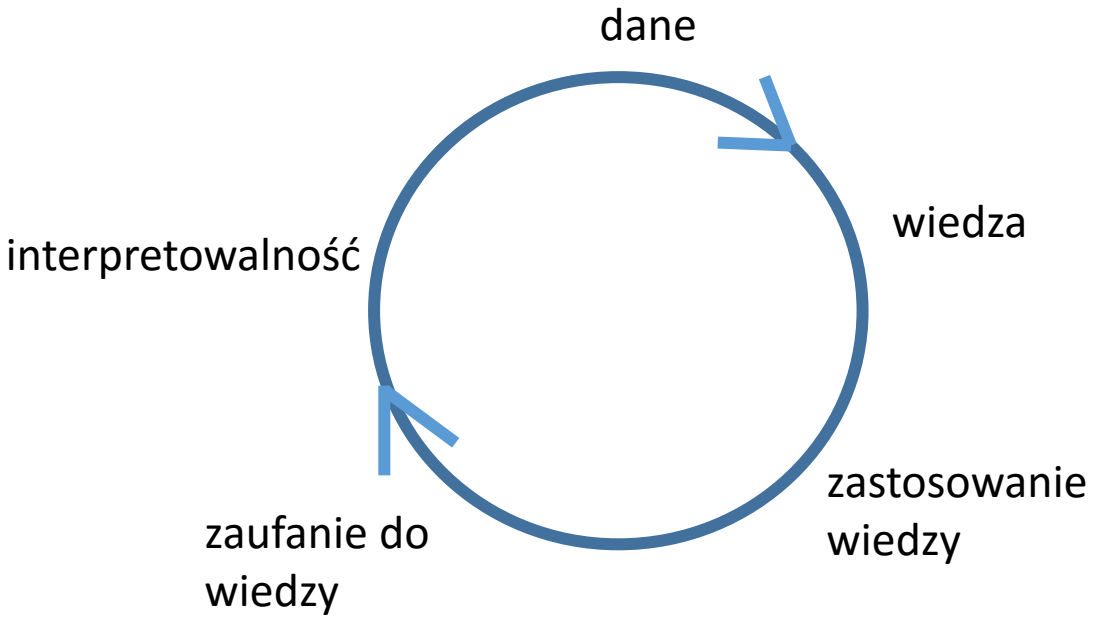
Hussey et. Al.
Aquatic animal telemetry: A panoramic window into the underwater world
Science, 12 Jun 2015: Vol. 348

MERIDIAN

- Marine Environmental Research Infrastructure for Data Integration and Application Network
- Budowa „centrum danych” dotyczących morza (dane akustyczne)
- Cztery ośrodki oceanograficzne, trzy informatyczne („lead” IBDA Dalhousie, SM)
- Szukamy osób zainteresowanych udziałem

Korelacje a przyczynowość

- Zasadnicze wyzwanie dla Big Data:
- Przyczynowość wymaga wiedzy i interpretacji wyników (np. klasyfikatora) w kategoriach tej wiedzy
- Zob. [Ribeiro et *al.* 2016] – ciekawe nowe podejście



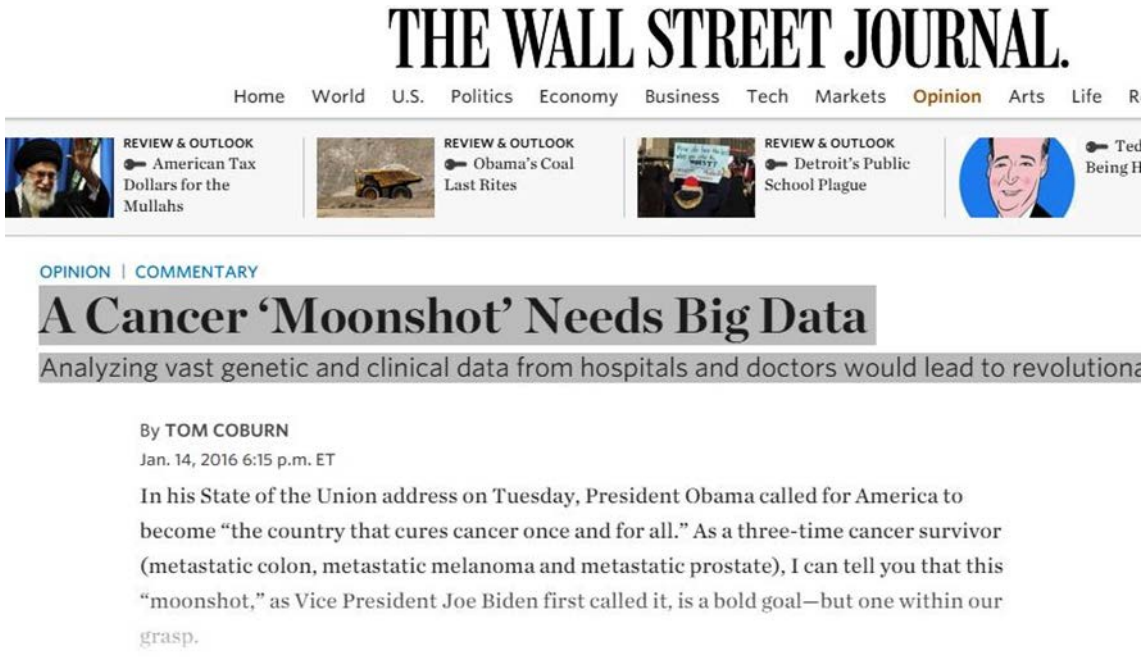
Ribeiro, Sing, Guestrin
„Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Własność danych

- Większość ciekawych danych jest „zamknięta”
  - Mimo iż są one zbudowane z danych stanowiących własność indywidualnych konsumentów
  - W szczególności dane dot. mobilności ludzi
- Olbrzymia niekomercyjna wartość tych danych
- Potrzeba rewizji tego modelu
  - Na podobieństwo własności intelektualnej?
- Ekonomiczny model wartości danych

Prywatność

- Realne niebezpieczeństwo
- Przeszkoda w dostępie do danych
- Reakcja – prewencja
- Potrzeba wielowymiarowgo modelu



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life R

REVIEW & OUTLOOK
American Tax Dollars for the Mullahs

REVIEW & OUTLOOK
Obama's Coal Last Rites

REVIEW & OUTLOOK
Detroit's Public School Plague

Ted Being H

OPINION | COMMENTARY

A Cancer 'Moonshot' Needs Big Data

Analyzing vast genetic and clinical data from hospitals and doctors would lead to revolutiona

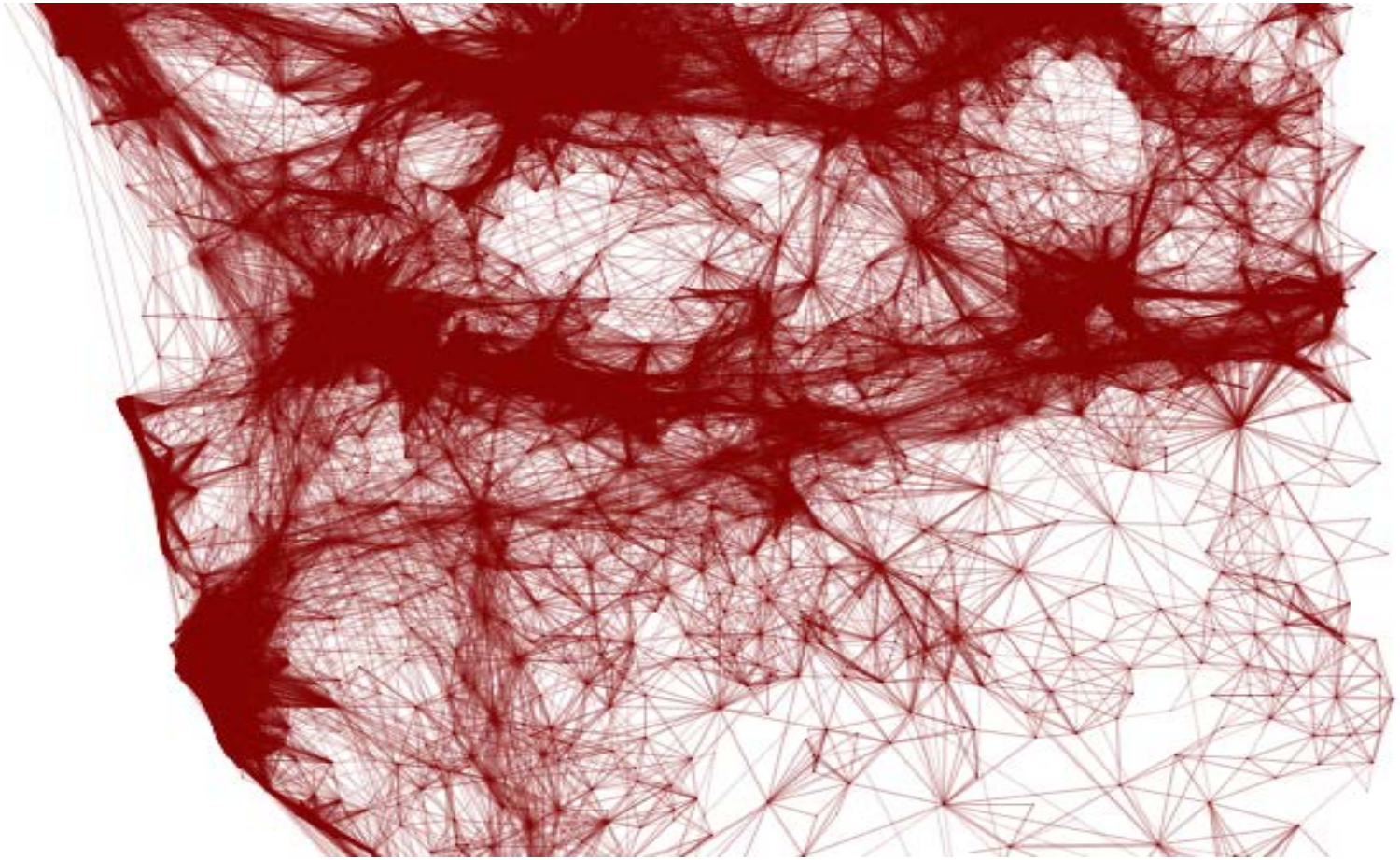
By TOM COBURN
Jan. 14, 2016 6:15 p.m. ET

In his State of the Union address on Tuesday, President Obama called for America to become "the country that cures cancer once and for all." As a three-time cancer survivor (metastatic colon, metastatic melanoma and metastatic prostate), I can tell you that this "moonshot," as Vice President Joe Biden first called it, is a bold goal—but one within our grasp.

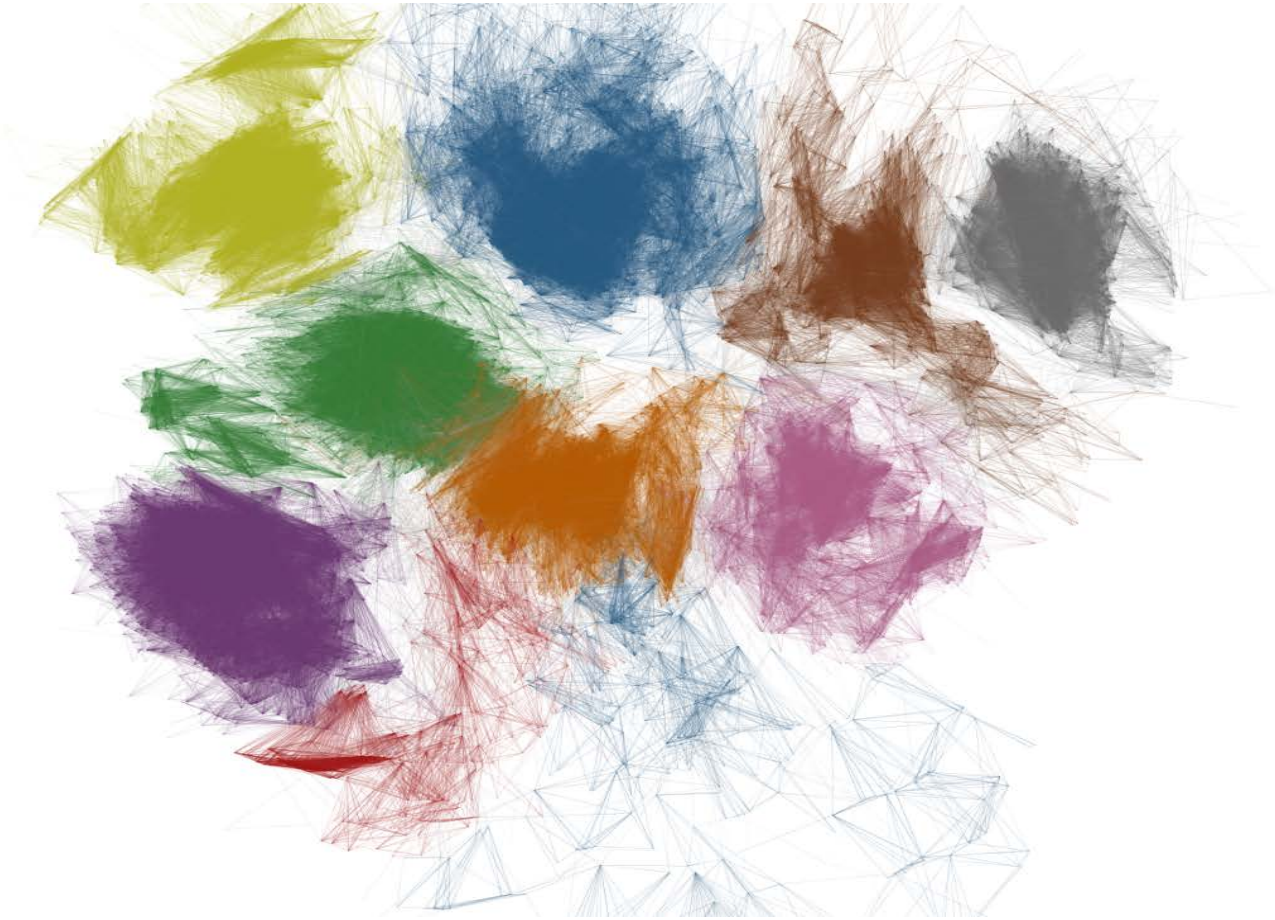
Podsumowanie

- Big Data jest czymś nowym
- Fascynujące zastosowania o olbrzymim potencjale
- Ambitne wyzwania do rozwiązania

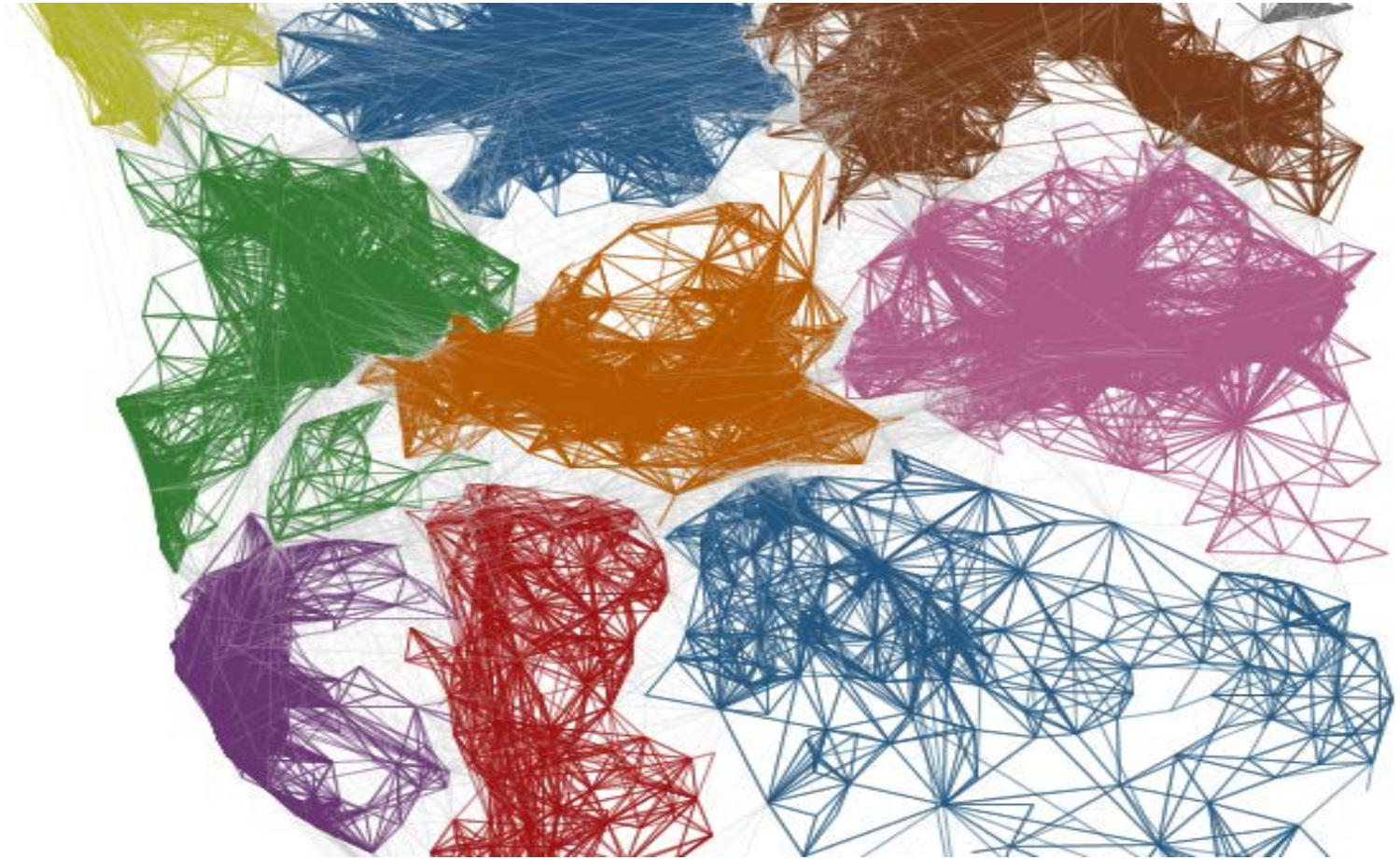
Mobilność kierowców w Toskanii



Community detection



Z powrotem do geografii



granice