Data Science for International Development (and public policy)

Data Science in (Astro)particle Physics and the Bridge to Industry Symposium
March 16, Lisbon

Kiwako Sakamoto
Former Data Scientist at the World Bank

My background

(as one of the first batch of data scientist at the World Bank)

One minute CV

2011: Entered economics Ph.D. program at University of Wisconsin-Madison

2013-2014: Quit economics and joined IceCube project at UW-Madison, mainly worked with computer scientists to develop computational resource monitoring tools. Took AI course, self-taught programming and various data science methods.

2014-2017: Worked at the World Bank; became the first female person to be hired as a data scientist there and the first data scientist at its research division.

2018: Going independent to work between public sector and private sector to bridge data for public use.

Quick Introduction to the World Bank

Quick glance

Established in 1944, HQ in Washington DC, originally to rebuild WWII affected countries (originally focused on Europe/Japan)

Now owned by its 189 member countries, it is a global development cooperative. It is **the largest development bank in the world**, and it provides financing, knowledge, and convening services to help client countries address their most important development challenges.

The World Bank Group as a whole has more than 10,000 employees in more than 120 offices worldwide.

FY2016 commitment sums to USD 45.9 billion to partner countries, distributed in credits, loans, grants, and guarantees (USD 64.2 billion as the World Bank Group)

The World Bank’s famous “Twin Goals”

Adapted in 2013:

1) **ending extreme poverty** (reduce extreme poverty in the world to less than 3 percent by 2030)
2) **boosting shared prosperity** (foster income growth of the bottom 40 percent of the population in each country)

This became a mantra - every operations, every actions and decisions made by the World Bank is based on these fundamental Twin Goals.

Breadth of domains

GLOBAL PRACTICES

- Agriculture
- Education
- Energy
- Environment & Natural Resources
- Finance, Competitiveness & Innovation
- Governance
- Health, Nutrition & Population
- Jobs and Development
- Macroeconomics, Trade & Investment
- Poverty
- Social Protection
- Social, Urban, Rural, & Resilience
- Transport
- Digital Development
- Water

REGIONAL UNITS

- Africa
- East Asia & Pacific
- Europe & Central Asia
- Latin America & the Caribbean
- Middle East & North Africa
- South Asia

GLOBAL THEMES

- Climate Change
- Fragility, Conflict, and Violence
- Gender
- Infrastructure and Public-Private Partnerships
- Knowledge Management

Data Science at the World Bank

Data Science at the World Bank

Data science at the World Bank **is almost always tied to “big data” analytics**. Data scientists are needed to:

- use novel data sets to improve measurement and decision making; that is, the types of data traditionally not used by economists (i.e. remote-sensing, call detail records (CDR), various transaction metadata, sensor data, administrative records, social media/text data; most often involves geospatial, imagery, and/or language data)
- change how we use the old sources of data (i.e. household survey data) by applying machine learning instead of conventional econometrics methods
- propose potential solution to the existing/upcoming problems
- prepare the IT environment to support big data analytics

Example questions

- How can we measure poverty at higher resolution to better serve the poor? (Satellite, CDR)
- How can we monitor food security risks to better predict/prepare/respond? (satellite agricultural yield estimation)
- How can we create the complete road network map of the world? (satellite, mobile phone)
- How can we measure urban mobility? (CDR, mobile phone)
- How can we monitor electrification progress at village level? (nightlight from satellite)
- How can we detect/measure the extent of illegal lodging in forest resources? (high-resolution satellite/drone, sensor network)

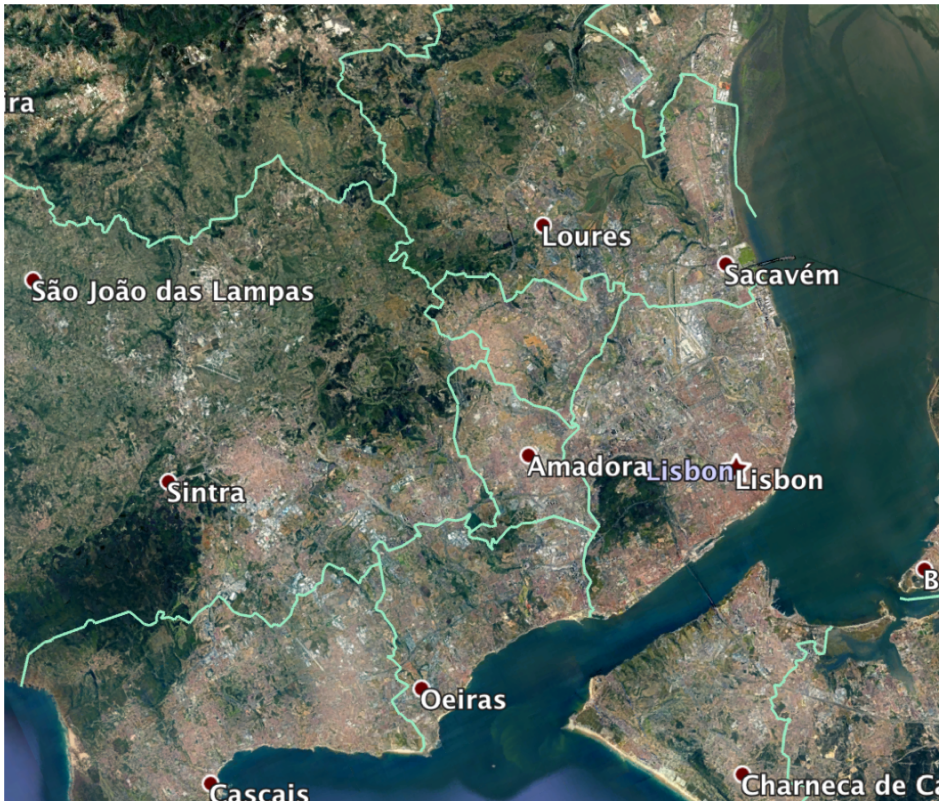
Poverty estimation using high-res satellite imagery: Stanford research (Jean et al, 2016*)

Main idea: Given that groundtruth survey data is limited, use night lights as a noisy but widely available proxy to poverty to improve poverty estimation.

1) use ImageNet-trained CNN to transfer edge and corner detection
2) teach computers which daytime features translate to night light (transfer learning part 2)
3) train a model to directly estimate local per capita outcomes from daytime image features (ridge regression)

* http://science.sciencemag.org/content/353/6301/790

Daytime satellite gives higher granularity of human activities



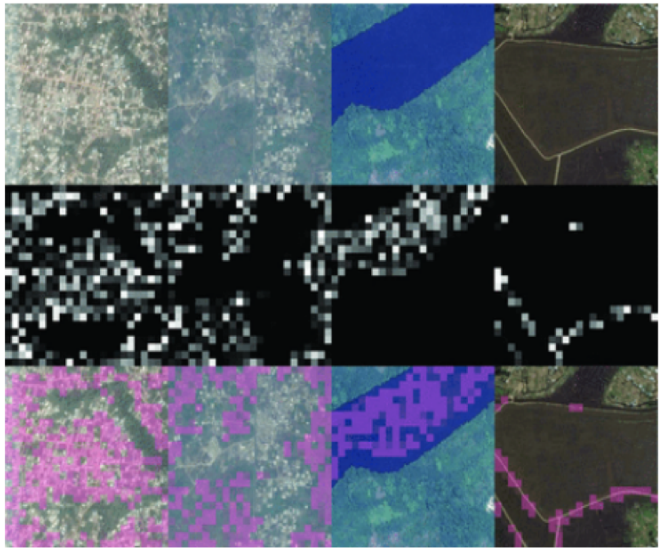
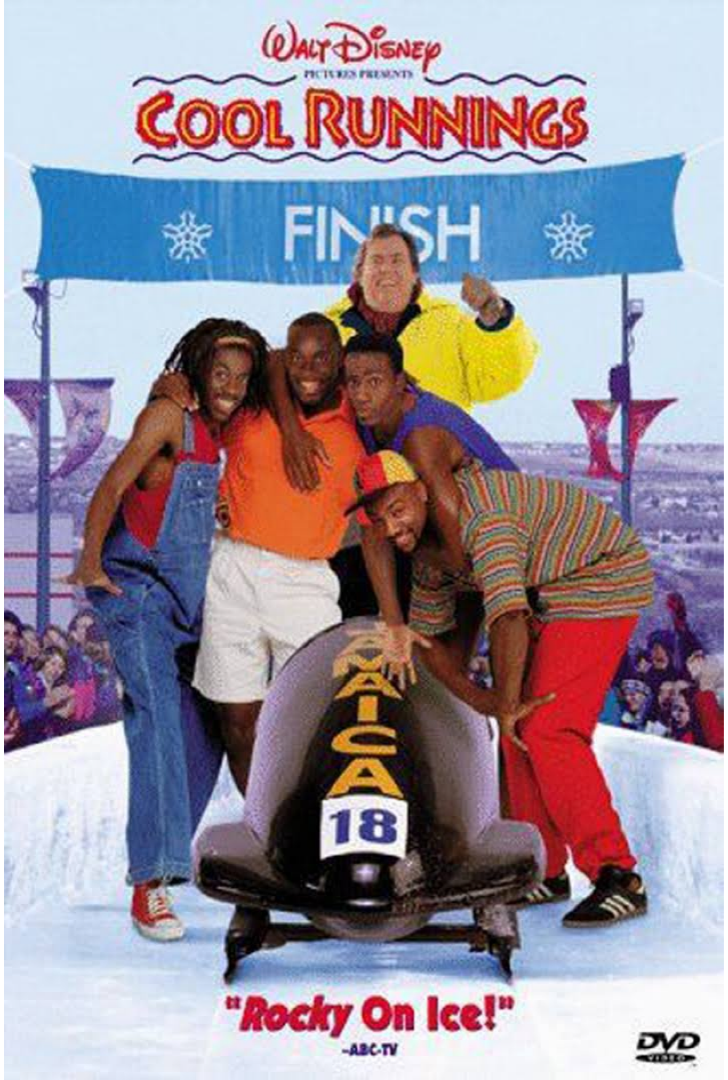


Fig. 2 Visualization of features.

By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Transfer learning



Poverty estimation using high-res satellite imagery: World Bank research (Engstrom et al, 2017*)

Instead of feeding satellite imagery directly to machine, they took the approach to:

1) extract features that are known/suspected to be correlated with poverty from high-resolution satellite imagery (objects, textual and spectral features)
2) feature selection
3) apply econometric analysis

*http://documents.worldbank.org/curated/en/610771513691888412/Poverty-from-space-using-high-resolution-satellite-imagery-for-estimating-economic-well-being

Figure 3: Example Developed Area (Buildings) Classification

Notes: above image shows raw (left) and classified (right) for developed area building classifier from raw satellite imagery. Areas in green show are true positive building classifications. Images in red are false positives: erroneously classified areas as buildings.

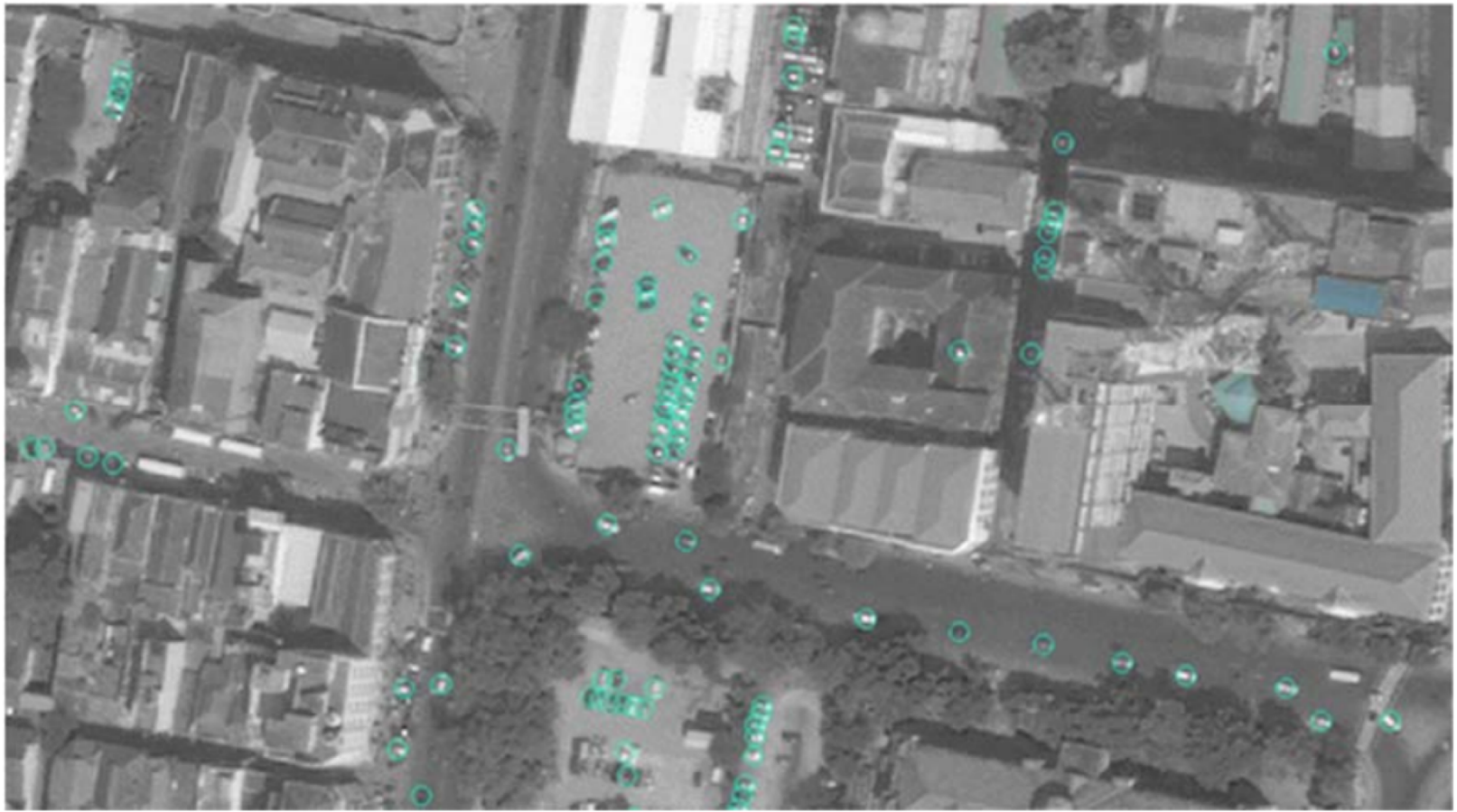


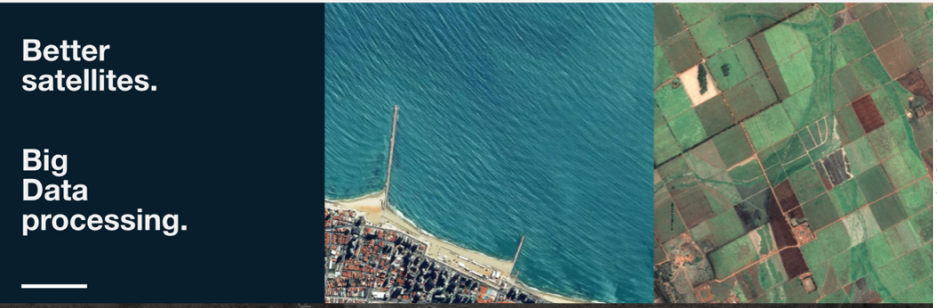
Figure 4: Example Car Classification

Notes: Cars identified by the convolutional neural network shown in blue.

Highlights

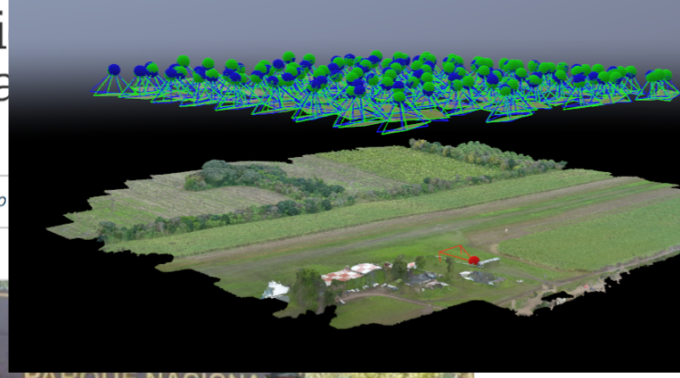
1. Why these approach? Very limited groundtruth data; household survey is very expensive!
2. Jean’s research uses not-so-fancy deep learning method but resulted in huge improvement over the existing method, gathered strong interest across academia, press, and large aid organizations.
3. Two examples highlight the difference of two fields: social scientist and computer scientist/engineers (causation vs. high accuracy rate in classification).

Satellites in Global Development

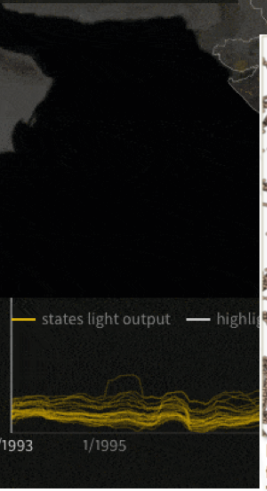


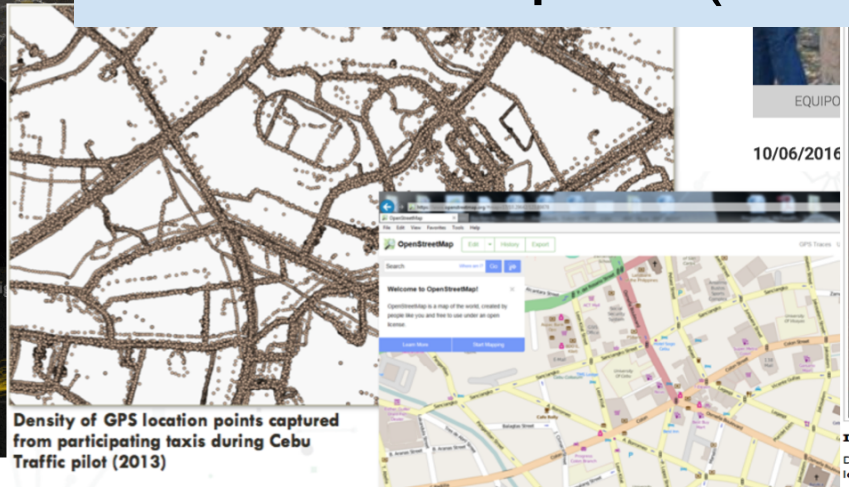
El Parque Naci
estudios con la
drones

Monitorean el estado de pastizales debido



Other examples (and many more!)





Density of GPS location points captured from participating taxis during Cebu Traffic pilot (2013)

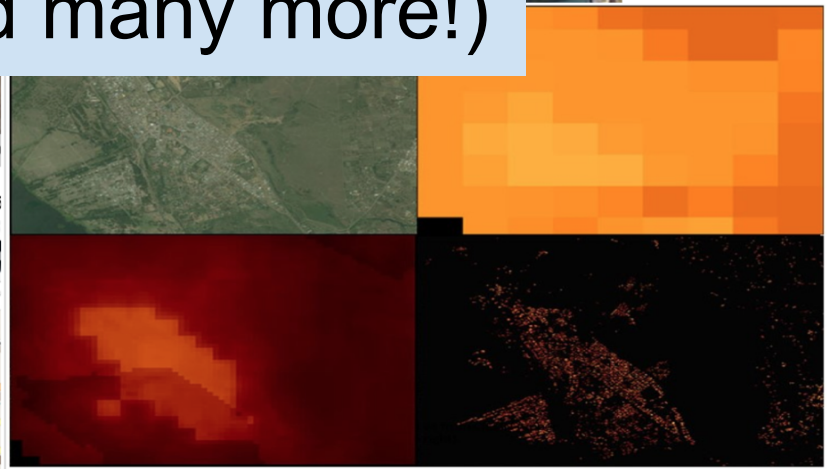


Image 1: Naivasha, Kenya.

DigitalGlobe satellite (upper left), gridded population of the world v4 from CIESIN (upper right), WorldPop (botton left), output from Facebook model (bottom right).

The “Bridge” Part

(why you might want to consider a career in international development/public policy?)

Photo by Ethan Tweedie

Hey, astroparticle physicists go to all the remote places in the world (adventurers!)

Courtesy of Mitsui kinzoku


TOYOTA

Modern (experimental) physicists are some or all of these:

- Software engineer
- Hardware engineer
- Statistician
- Mathematical modeler

(But maybe not:

- Science communicator
- Diplomat, etc

which are important qualities to work in public policy but can be learned on the job.)

Why it worked out for me & what I gained

- Stimulating work environment - domain experts in every fields, meeting with government officials and policy makers, affecting policy at large scale
- Every researcher wants to collaborate with you, with government contacts and access to rare data sets
- As an early facilitator/accelerator of big data projects, I was able to have a large voice in the organization + worked with innovative projects
- Learning balance between science and diplomacy
- Delegated most of the technical work to researchers -> became a master of none, but with valuable connections and experience in applications and problems

There are so many problems to be solved, with many low-hanging fruits!

The technical researchers in international development has traditionally been economists (and some domain expert engineers, etc), but there are many problems nowadays that can be solved/improved by data science methods due to newly available data.

Often times these are easily-solved problems with potential real-world impact, left unsolved due to small talent pool.

Additional Resources

Getting started with Satellite: http://landscape.satsummit.io/

World Bank data: https://data.worldbank.org/

(If you are interested in survey data: http://microdata.worldbank.org/)

If you want to replicate/extend Jean et al (2016): https://github.com/nealjean/predicting-poverty