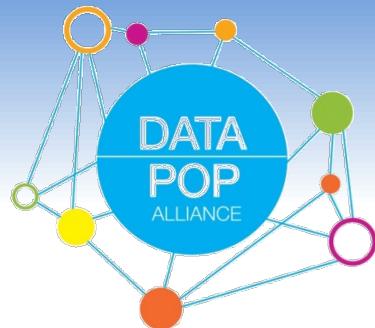


Big Data for Social Good

Using Kaggle for Business and Social Impact

Emmanuel Letouzé
@datapopalliance



Tobias Pfaff
@datalook



Peter Prettenhofer
@datarobot



Agenda

Big Data for Social Good

- > The Macro Perspective
- > The Micro Perspective

...and Kaggle

- > Short Intro
- > Interactive Crash Course

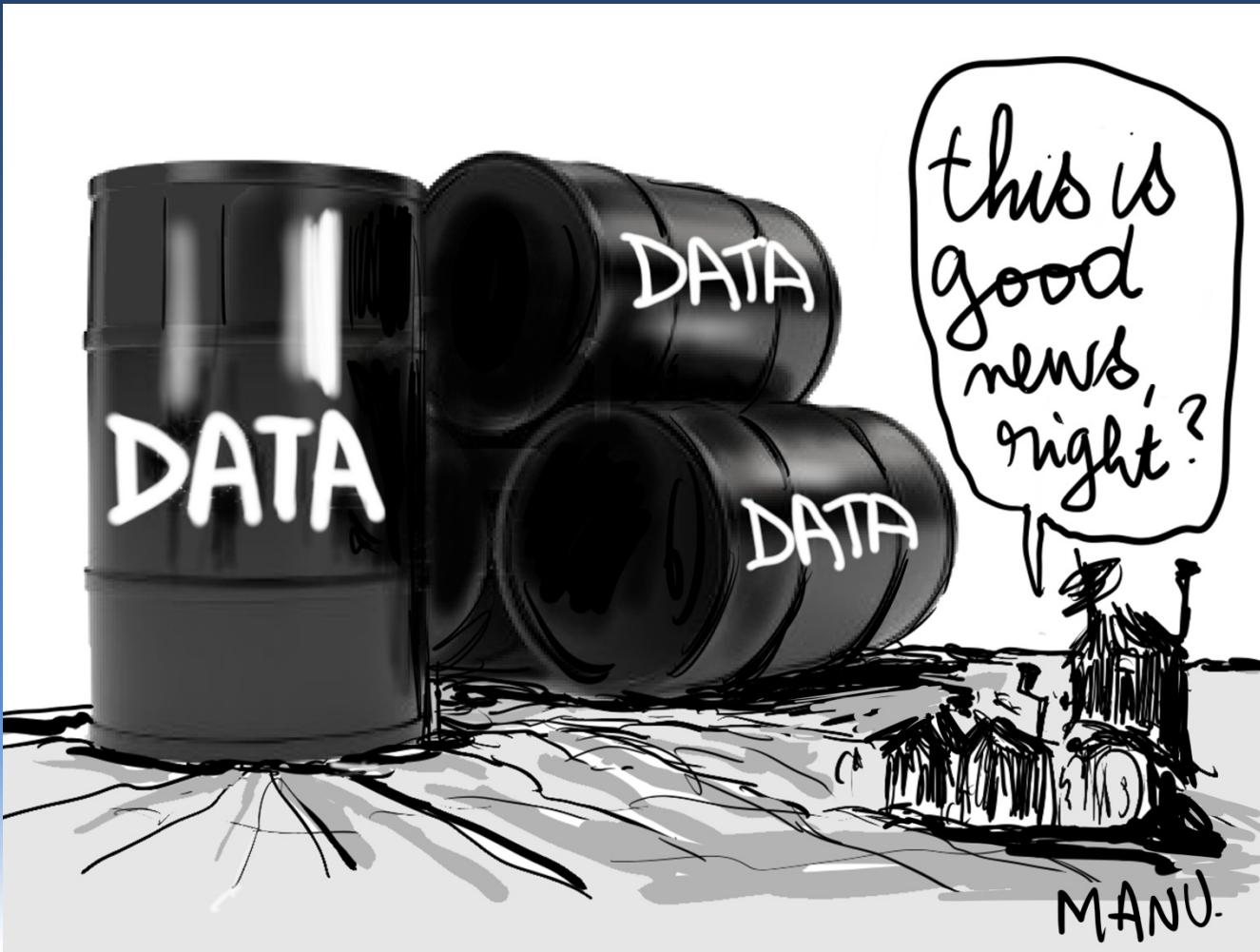
Interactive part: 1. Load bit.ly/bitkom15 and create account
2. Create account on kaggle.com

> The Macro Perspective

Big Data and Development

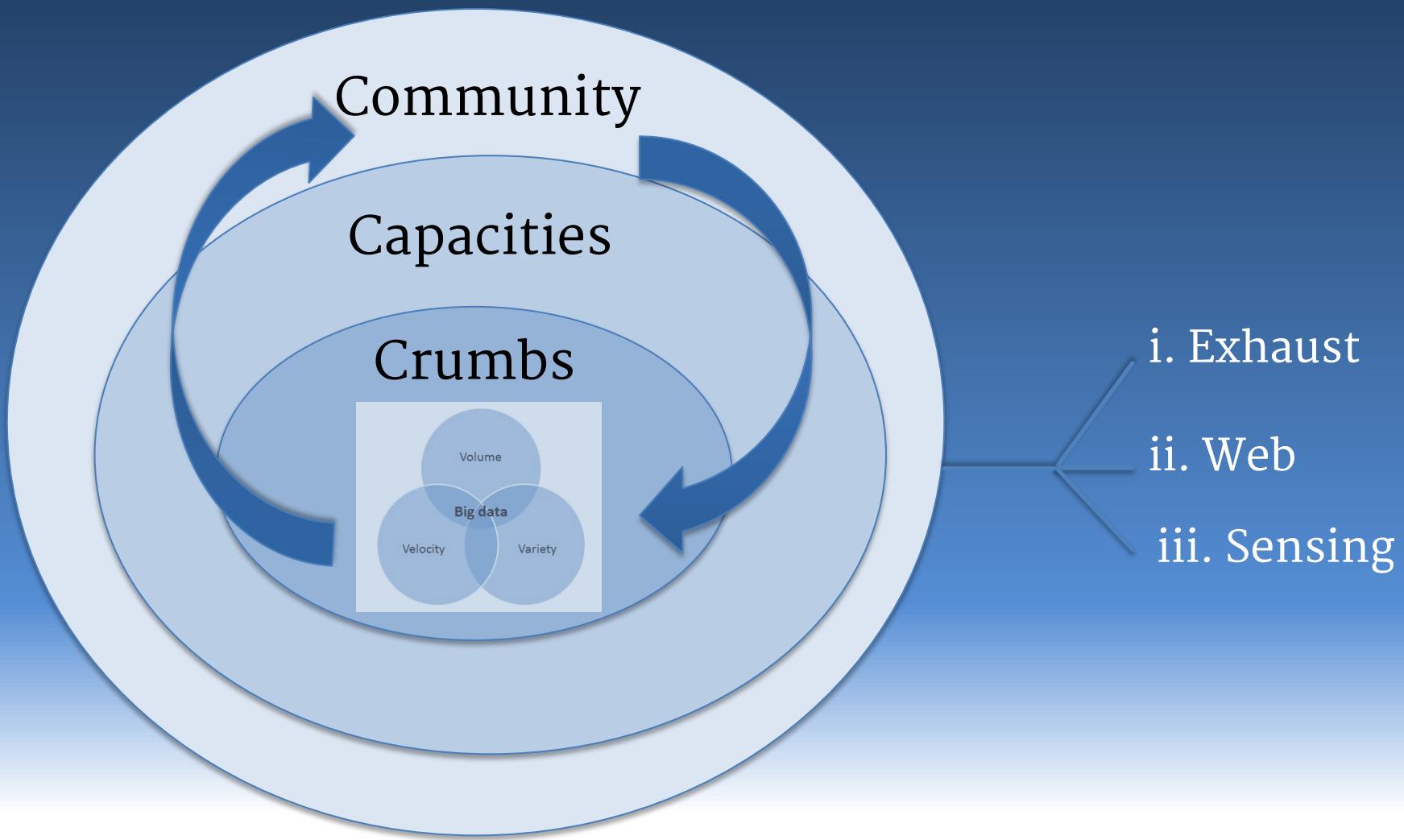


“The bottom half of the world’s population owns the same as the richest 85 people in the world”*



* Source: Oxfam International, citing Credit Suisse, Jan. 2014

Big Data as / is a new ecosystem: from the 3 Vs to the 3Cs of Big Data



Applications: Taxonomy

1. Descriptive
 - e.g. maps, clouds..
2. Predictive:
 - forecasting
 - inference
3. Prescriptive
 - causal inference

an illustrated introduction to

Predicting socioeconomic levels through cell-phone data

Question:



So, how is it possible to predict an area's socioeconomic - or poverty - level from the cell-phone data it emits?

Step ①

first, find or collect actual survey data..

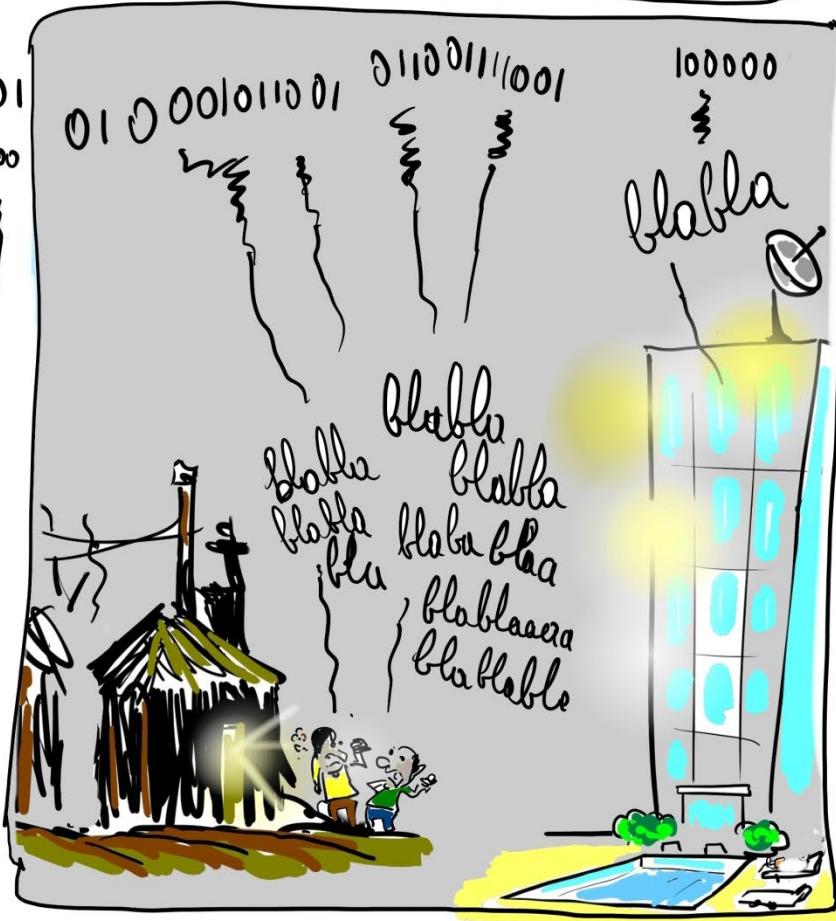
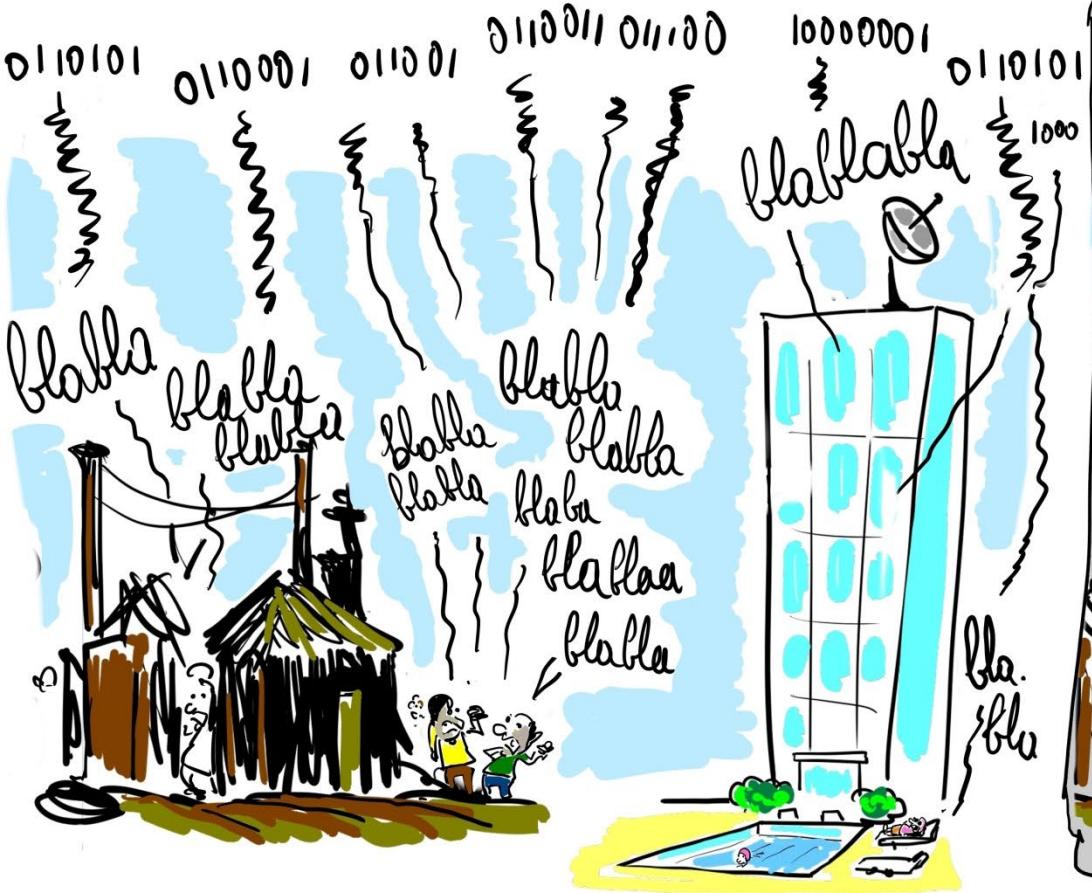


MANU.

Step ②



then notice how cell phone users leave digital traces, day & night..

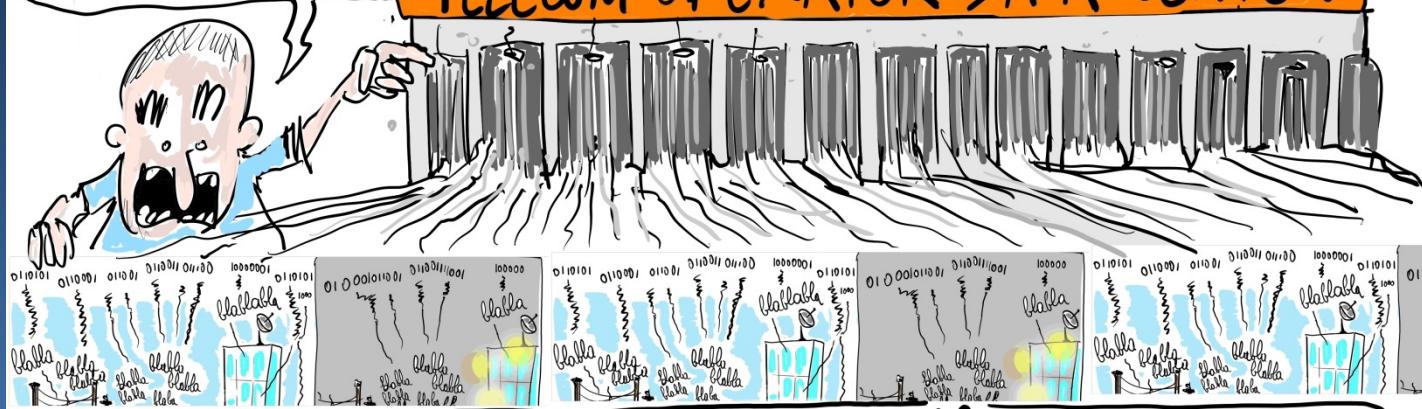


MANU.

"these 'digital traces' recorded by every telecom operator, are Call Detail Records, or CDRs, metadata that look like that"

CALLER ID	CALLER LOCATION	RECIPIENT ID	RECIPIENT LOCATION	CALL TIME	CALL DURATION
X36872 9748Y	2°24'22," 35°49'58"	A8C492 TC7364G	3°38'49" 31°12'22"	2014.04.01 ET 17 22	01.12.27

TELECOM OPERATOR DATA CENTER



"and these CDRs will show differences in calling patterns between different areas ..."



And using computers and algorithms, researchers try to find the strongest correlations

between CDRS and survey data to build a "predictive model"

...that can then take CDRs from a later time or different area

...and turn them into estimates of socioeconomic levels without a survey!

MANU.

Applications: ongoing



Moves on the Streets: Predicting Crime Hotspots Using Aggregated Anonymized Data on People Dynamics

Andrey Bogomolov, Bruno Lepri*, Jacopo Staiano, Emmanuel Letouzé,
Nuria Oliver, Alex 'Sandy' Pentland, Fabio Pianesi

Applications: Examples

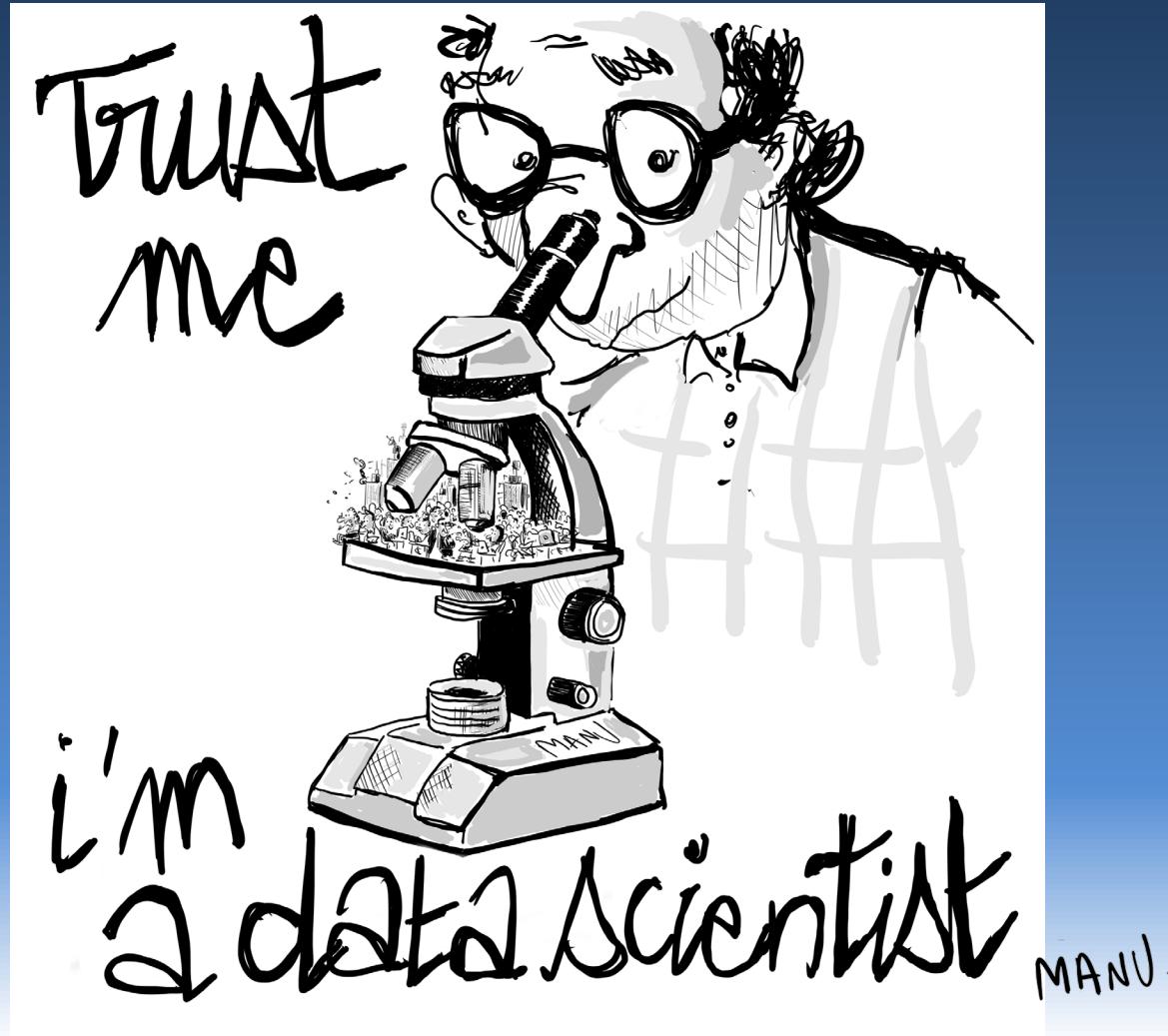
BLOG POST

Could Big Data provide alternative measures of poverty and welfare?

Cell-Phone Data Might Help Predict Ebola's Spread

Mobility data from an African mobile-phone carrier could help researchers recommend where to focus health-care efforts.

Implications: Ethics



Implications: Power

A development data revolution
needs to go beyond the geeks
and bean-counters

Will better data really lead to better development policies? Only if the right people have access and use it to make better policies

Jonathan Glemmie, The Guardian, Oct 3, 2013



Cartagena Data Festival

Better data for
a better tomorrow

20 - 22 APRIL 2015

CARTAGENA, COLOMBIA

Event Organisers



Empowered lives.
Resilient nations.

Event Partners



> The Micro Perspective



> The Micro Perspective

Bring the superpowers of data science to **nonprofit** organizations and to the **local administration**.

> The Micro Perspective

**Saving lives with predictive
analytics in New York City**



ROC curve pre-analysis

% of severe violations found



ROC curve post-analysis

% of severe violations found



Time that buildings are at risk of severe fire significantly reduced.

> The Micro Perspective

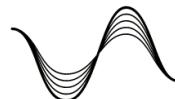
Which **social problems** can be solved with (Big) Data?

> The Micro Perspective

Who are the other players?

DataKind

 BAYES IMPACT



The Eric & Wendy Schmidt
**Data Science for
Social Good**
Summer Fellowship



THE UNIVERSITY OF
CHICAGO



> The Micro Perspective

...and how can my company get involved?



January 16 2015

0 comments

Pivotal For Good with Crisis Text Line: A First Look



Post by Noelle Sio, DataKind Data Ambassador and Principal Data Scientist at Pivotal

Note: This post originally appeared on [Pivotal's blog](#) on January 13 2015.

A couple months ago, Pivotal and DataKind [announced the launch](#) of the first Pivotal For Good (P4G) project, a collaboration between [Crisis Text Line](#) and yours truly. P4G

is a multi-year initiative to build a crisis text line system with Pivotal's

> Short Intro to Data Science Competitions with

kaggle™

How Data Science Competitions Work

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.72	12340	Audio	19.95	Mexico
0.41	31240	Computer	6.99	Taiwan
1.94	54323	Hardware	11.99	Taiwan
0.023	92356	Household	2.05	USA
0.08	78023	Computer	99.99	USA
2.09	12340	Computer	129.99	China
1.1	31240	Audio	18.99	China



Target

Features

How Data Science Competitions Work

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.64	78023	Household	229.9	Brazil
0.72	12340	Audio	19.95	Mexico
0.41	31240	Computer	6.99	Taiwan
1.94	54323	Hardware	11.99	Taiwan
0.023	92356	Household	2.05	USA
0.08	78023	Computer	99.99	USA
2.09	12340	Computer	129.99	China
1.1	31240	Audio	18.99	China

Ground Truth

Training

Testing

How Data Science Competitions Work

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
?	78023	Household	229.9	Brazil
?	12340	Audio	19.95	Mexico
?	31240	Computer	6.99	Taiwan
?	54323	Hardware	11.99	Taiwan
?	92356	Household	2.05	USA
?	78023	Computer	99.99	USA
?	12340	Computer	129.99	China
?	31240	Audio	18.99	China

Blanks

Training

Testing

How Data Science Competitions Work

Submissions

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.83	78023	Household	229.9	Brazil
0.65	12340	Audio	19.95	Mexico
0.52	31240	Computer	6.99	Taiwan
1.74	54323	Hardware	11.99	Taiwan
0.1	92356	Household	2.05	USA
0.02	78023	Computer	99.99	USA
2.9	12340	Computer	129.99	China
0.83	31240	Audio	18.99	China

Training

Testing

How Data Science Competitions Work

 MARINEXPLORE
WHALE
DETECTION

Completed • \$10,000 • 245 teams

The Marinexplore and Cornell University Whale Detection Challenge

Fri 8 Feb 2013 – Mon 8 Apr 2013 (22 months ago)

[Dashboard](#) ▾

Private Leaderboard - The Marinexplore and Cornell University Whale Detection Challenge

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts?
[Let us know.](#)

#	Rank	Team Name * In the money	Score ⓘ	Entries	Last Submission UTC (Best – Last Submission)
1	—	SluiceBox 🎉 *	0.98384	70	Sun, 07 Apr 2013 18:58:34
2	—	alfnie *	0.98379	27	Sun, 07 Apr 2013 22:47:36 (-1.2h)
3	↑1	RBM 🎉	0.98226	32	Sun, 07 Apr 2013 23:22:16 (-1.3h)
4	↓1	Free Willzyx 🎉	0.98210	38	Sun, 07 Apr 2013 23:52:09 (-1.8h)
5	—	Jure Zbontar	0.98080	24	Mon, 01 Apr 2013 15:52:11 (-5.1h)
6	—	Daniel Nouri	0.98070	16	Thu, 04 Apr 2013 00:45:00 (-6h)
7	↑1	Tree growers 🎉	0.97982	80	Sun, 07 Apr 2013 12:42:55 (-3d)
8	↓1	sedilem	0.97975	8	Sun, 07 Apr 2013 22:42:53 (-8.6h)
9	↑1	Paul Horsfall	0.97713	7	Sun, 07 Apr 2013 20:39:35

How Data Science Competitions Work

Public/Private Leaderboard

Return%	ProductID	Dept	Price	MFR
1.94	54323	Household	54.95	USA
0.023	92356	Household	9.95	USA
0.8	78023	Computer	4.5	China
0.01	12340	Audio	109.99	China
0.41	31240	Audio	29.99	Taiwan
0.97	12351	Hardware	54.95	Mexico
0.0115	90141	Hardware	4.99	USA
0.4	81240	Hardware	6.55	Taiwan
0.03	14896	Computer	211.99	Korea
0.205	62132	Computer	1100	USA
1.6878	54323	Audio	34.99	USA
0.0345	92356	Audio	7.99	USA
0.83	78023	Household	229.9	Brazil
0.65	12340	Audio	19.95	Mexico
0.52	31240	Computer	6.99	Taiwan
1.74	54323	Hardware	11.99	Taiwan
0.1	92356	Household	2.05	USA
0.02	78023	Computer	99.99	USA
2.9	12340	Computer	129.99	China
0.83	31240	Audio	18.99	China

Training

Testing

How Data Science Competitions Work



> Interactive Kaggle Crash Course

Interactive part: 1. Load bit.ly/bitkom15 and create account
2. Create account on kaggle.com

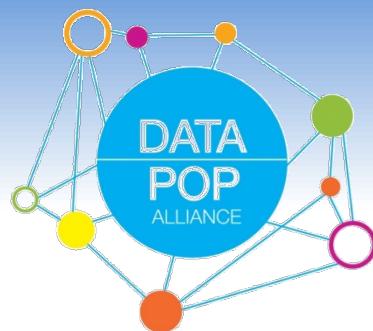
Slides and interactive material available at bit.ly/bitkom16

Thank you.

Questions?

Emmanuel Letouzé

eletouze@datapopalliance.org



Tobias Pfaff

tobias@datalook.io



Peter Prettenhofer

peter@datarobot.com

