

Practical Machine Ethics

And implications for AI,
ML and DS

Jesus Ramos
[@xuxoramos](#)
[fb.com/xuxoramos](#)
[linkedin/xuxoramos](#)
[aixsw.mx](#)

Who's this bloke?

- Born as Software Engineer
- MSc Computational Finance @ Unottingham
- Led the development of 2 software solutions on a national scale
- Formed Data Science groups and teams for companies since 2013
- Founded The Data Pub in 2015, the largest Data Science community in México [2000 members]
- Cofounder of the Mexican Society for Data Science in 2015
- Founded Datank.ai in 2016 to empower AI organizations
- Founded AI x Social Wealth in 2019 to address social problems with ML
- Speaker at several national and international forums on ML done right





And now, a story...



Sep 19th, 2017



Sep 19th, 2017



#REVISAMIGRIETA

ND Mantarraya



SISMOMX

Centro de respuesta rápida para la ciudadanía.

The Project

- 600 photos reliably tagged by experts
- To train a convolutional neural network model
- To obtain 2 types of classifications: “structural damage detected”, or “no structural damage detected”
- Model then exposed as API
- That could be invoked via Twitter and Telegram

The caveat

- 600 examples to learn from are too few for any neural network
- Any NN usually need thousands of examples
- Such a small training set will cause the NN to be wrong very frequently
- Many false positives
- Many false negatives

False
positives



False
negatives



There are social costs associated with algorithms not doing their work!

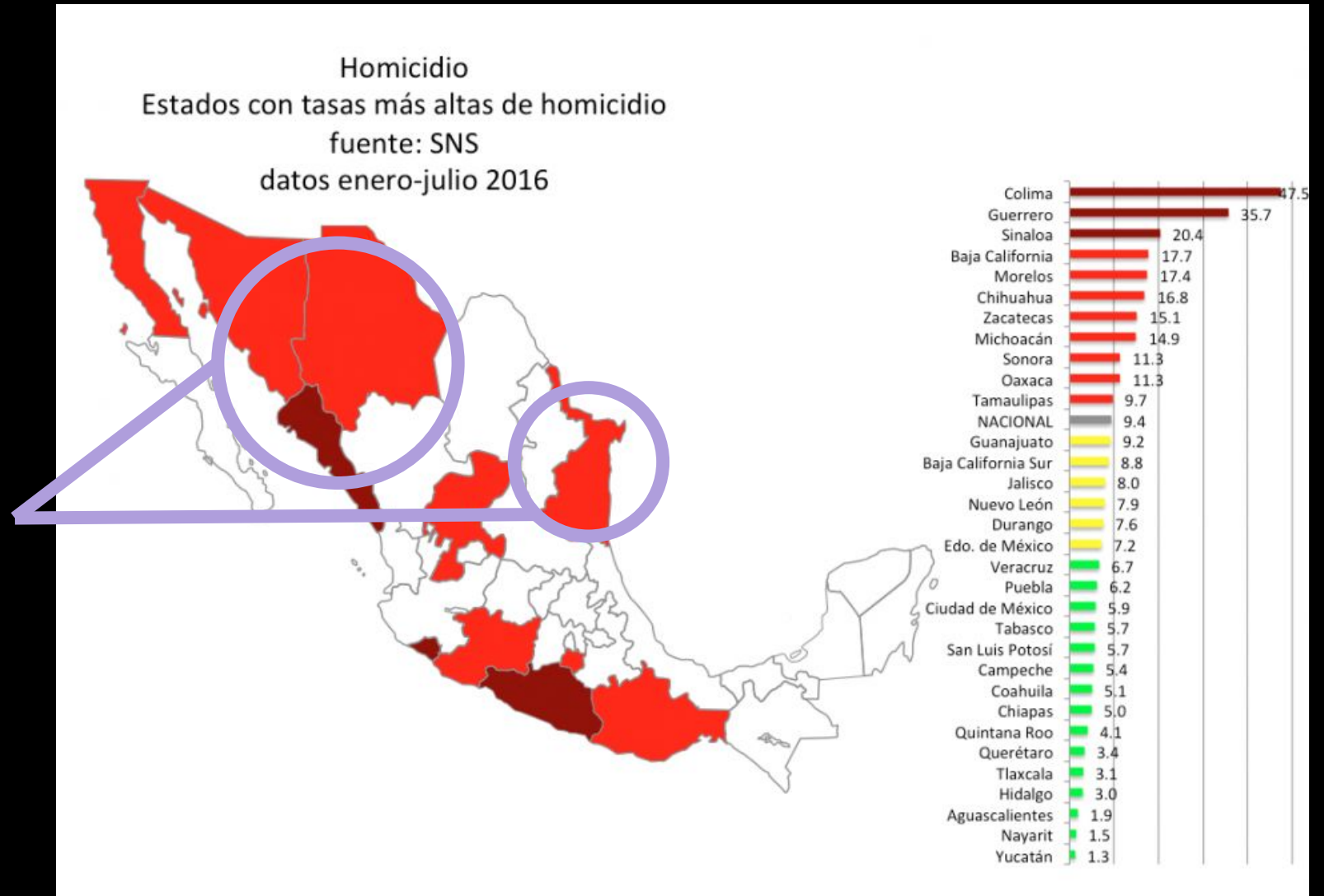
Another story, this one from the private sector..

Phone company from LATAM detects users with prepaid phones who spend over 400 USD a week in airtime.

Company creates an airtime credit product where a small line is extended when you use up all your phone minutes.

This product becomes financially successful. On a closer look, however...

Georeferencing of the users that spent +400 USD a week in phone credit.



What kind of users do you think can spend +400 USD a week on a nearly-anonymous cell phone line? 😞

What went wrong?

Lack of context

team left out the subject matter expert.

Correlation == Causality

team left out mathematicians and scientists

Biased data



FACT:

Machine Learning classification algorithms
are created to discriminate*.

* *Discriminate*

From latin *discrimināre*. tr. To select by excluding.

"A big step towards countering discriminatory algorithms is the ability to understand them..."

-Ethan Chiel, writer @ Fusion

European Union, Apr 2016*

European Union regulations on algorithmic decision-making and a “right to explanation”

Bryce Goodman,^{1*} Seth Flaxman,²

¹Oxford Internet Institute, Oxford

1 St Giles', Oxford OX1 3LB, United Kingdom

²Department of Statistics, University of Oxford,
24-29 St Giles', Oxford OX1 3LB, United Kingdom

- You didn't get accepted in your university of choice?
- You were given a medical treatment instead of another?
- You were not granted a mortgage?
- You were not selected as part of a tax break program?

You have the right to an explanation!

Yes, yes, but what are the REAL, PRACTICAL implications of INTERPRETABILITY for the Data Science, Machine Learning and AI disciplines?

Change of core paradigm:

Model interpretability >> Good Predictions

Implications for Supervised Learning

- Privilege simplicity over complexity
- Be thorough in your feature selection
- Models with fewer, but more significant variables
- So we can train them all in less time
- And thus avoid the curse of dimensionality
- And reduce overfitting

Implications for Unsupervised Learning

- Privilege reproducibility*
- Choose highly parameterized algorithms (DBSCAN, Gower metric)
- Define experiment design protocols

Implications for Data Engineering

- Version control! Of data, of experiments, of results!
- Provision infrastructure for training models in parallel
- Also for RETRAINING!
- Observe the “only orcs and goblins overwrite data” rule
- In general, imbue the ML discipline with the noble art of Engineering :D

And what about Deep Learning?

“Explaining the model’s decision in terms of weights and layers is unacceptable to the affected party”

So, we don't use it anymore?

The problem is time!

Self-driving cars

Action:
Aggressive steer



Consequence:
Avoid hitting
pedestrian

Credit scoring

Action:
Give credit



Consequence:
Client
default/reward

DL interpretability in the making

DARPA¹, MIT², Cambridge, Oxford, *et al* are working on it.

1.<http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>

2.<http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028>

Is there an ethical framework to help
us think about this (while deep learning interpretability
arrives)?

ACM, January 2017

- Awareness of bias
- Access & redress
- Accountability
- Explanation
- Data provenance
- Auditability
- Validation & testing

Institute for Ethical AI & ML, UK, 2018

- Human augmentation
- Bias evaluation
- Explainability
- Reproducible operations
- Displacement strategy
- Practical accuracy
- Trust by privacy
- Security risks

The Modeler's Oath (Wilmott & Derman, 2009)

- I will remember that I didn't make the world, and it doesn't satisfy my equations.
- Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.
- I will never sacrifice reality for elegance without explaining why I have done so.
- Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.
- I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.

Thanks!

[@xuxoramos](#)

[fb.com/xuxoramos](#)

[Inkdin/xuxoramos](#)

[www.aixsw.mx](#)