EXECUTIVE BRIEFING: WHY MACHINE LEARNED MODELS CRASH AND BURN IN PRODUCTION (AND WHAT TO DO ABOUT IT)

Dr. David Talby

PacificAI
MODEL DEVELOPMENT ≠ SOFTWARE DEVELOPMENT
1. The moment you put a model in production, it starts degrading
"The greatest model, trained on data inconsistent with the data it actually faces in the real world, will at best perform unreliably, and at worst fail catastrophically."

[Sanders & Saxe, Sophos Group, Proceedings of Blackhat 2017]
CONCEPT DRIFT: AN EXAMPLE

Medicare Fines 2,610 Hospitals In Third Round Of Readmission Penalties

By Jordan Rau | October 2, 2014

| Medical claims                | > 4.7 Billion               |
|                              |                              |
| Pharmacy claims               | > 1.2 Billion                |
| Providers                     | > 500,000                    |
| Patients                      | > 120 million                |

- Locality (epidemics)
- Seasonality
- Changes in the hospital / population
- Impact of deploying the system
- Combination of all of the above
Hidden Technical Debt in Machine Learning Systems

[D. Sculley et al., Google, NIPS 2015]

Experience has shown that the external world is rarely stable. Indeed, the changing nature of the world is one of the sources of technical debt in machine learning.
HOW FAST DEPENDS ON THE PROBLEM
(MUCH MORE THAN ON YOUR ALGORITHM)

Never Changing

Political & Economic Models

Natural Language, Social Behavior Models

Google or Amazon Search models

Banking & eCommerce fraud

Online Social Networking Models/Rules

Automated trading
Real-time ad bidding

Always Changing

Physical models:
Face recognition
Voice recognition
Climate models

Cyber Security

(MUCH MORE THAN ON YOUR ALGORITHM)
SO PUT THE RIGHT PLATFORM IN PLACE

(MEASURE, RETRAIN, REDEPLOY)
2.

You rarely get to deploy the same model twice
**Table: Prediction Models**

<table>
<thead>
<tr>
<th>Model</th>
<th>Model’s Goal</th>
<th>Sample size</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>LACE index (2010)</td>
<td>30-day mortality or readmission</td>
<td>4,812</td>
<td>11 hospitals in Ontario, 2002-2006</td>
</tr>
<tr>
<td>Charlson morbidity index (1987)</td>
<td>1-year mortality</td>
<td>607</td>
<td>1 hospital in NYC, April 1984</td>
</tr>
</tbody>
</table>

Healthcare / Natural Language
• Clinical coding for outpatient radiology
• Infer procedure code (CPT), 90% overlap

Cyber Security / Deep Learning
• Detect malicious URL’s
• Train on one dataset, test on others
IT’S NOT ABOUT HOW ACCURATE YOUR MODEL IS

(IT’S ABOUT HOW FAST YOU CAN TUNE IT ON MY DATA)

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

[D. Sculley et al., Google, NIPS 2015]
3.

It’s really hard to know how well you’re doing
“it seemed we were only seeing about 10%-15% of the predicted lift, so we decided to run a little experiment. And that’s when the wheels totally flew off the bus.”
## THE PITFALLS OF A/B TESTING

**[Alice Zheng, Dato, June 2015]**

<table>
<thead>
<tr>
<th>separation of experiences</th>
<th>How many false positives can we tolerate?</th>
<th>What does the p-value mean?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which metric?</td>
<td>How many observations do we need?</td>
<td>Multiple models, multiple hypotheses</td>
</tr>
<tr>
<td>How much change counts as real change?</td>
<td>Is the distribution of the metric Gaussian?</td>
<td>How long to run the test?</td>
</tr>
<tr>
<td>One- or two-sided test?</td>
<td>Are the variances equal?</td>
<td>Catching distribution drift</td>
</tr>
</tbody>
</table>
FIVE PUZZLING OUTCOMES EXPLAINED

[Ron Kohavi et al., Microsoft, August 2012]

The Primacy and Novelty Effects
Regression to the Mean

Best Practice: A/A Testing
Often, the real modeling work only starts in production
SEMI SUPERVISED LEARNING

UK's big four banks face extra £19bn in fines, analysts predict

Ratings agency Standard & Poor’s estimates total costs for Barclays, HSBC, RBS and Lloyds on top of £42bn already paid in the five years to 2014

Bank of America To Pay Record $16.65 Billion Fine

UBS fined £30m over rogue trader

Terrorism, fines and money laundering: why banks say no to poor customers

The tightening of international banking standards is making it difficult for low-income people in the global south to get access to banking services

J.P. Morgan Adds $2.6 Billion to Its $25 Billion Plus Tally of Recent Settlements

Deutsche Bank's Record Fine Reveals Its Rotten Heart
IN NUMBERS

99.9999% 'Good' messages

6+ Months per case

50+ Schemes (and counting)
Medicare And Medicaid Fraud Is Costing Taxpayers Billions

Barely a day goes by without a major news story highlighting some new Medicare or Medicaid scam that has
5.

Your best people are needed on the project after going to production
DESIGN
Most important, hardest to change technical decisions are made here.

BUILD & TEST
Riskiest & most reused code components are built and tested first.

DEPLOY
First deployment is hands-on, then we automate it and iterate to build lower-priority features.

OPERATE
Ongoing, repetitive tasks are either automated away or handed off to support & operations.
MODEL DEVELOPMENT

MODEL
Feature engineering, model selection & optimization are done for the 1st model built.

DEPLOY & MEASURE
Online metrics is key in production, since results will often defer from off-line ones.

EXPERIMENT
Design & run as many experiments, as fast as possible, with new inputs, features & feedback.

AUTOMATE
Automate the retrain or active learning pipeline, including online metrics & labeled data collection.
To conclude...
MODEL DEVELOPMENT ≠ SOFTWARE DEVELOPMENT

- Rethink your development process
- Set the right expectations with your customers
- Deploy a platform & plan for the DataOps effort in production
THANK YOU!

Email: david@pacific.ai
Twitter: @davidtalby
LinkedIn: in/davidtalby