Presented by:

ODSC

ACCELERATE AI

ACCELERATE AI EAST

BOSTON | April 30-May 4

2019

THE LEADING DATA SCIENCE CONFERENCE

BOSTON

APR 30 - MAY 3

What to Expect When You’
Putting AI in Production

David Talby, PhD

CTO,
Pacific AI

WHAT TO EXPECT WHEN YOU'RE PUTTING AI IN PRODUCTION

Dr. David Talby

PacificAI

MODEL DEVELOPMENT ≠ SOFTWARE DEVELOPMENT

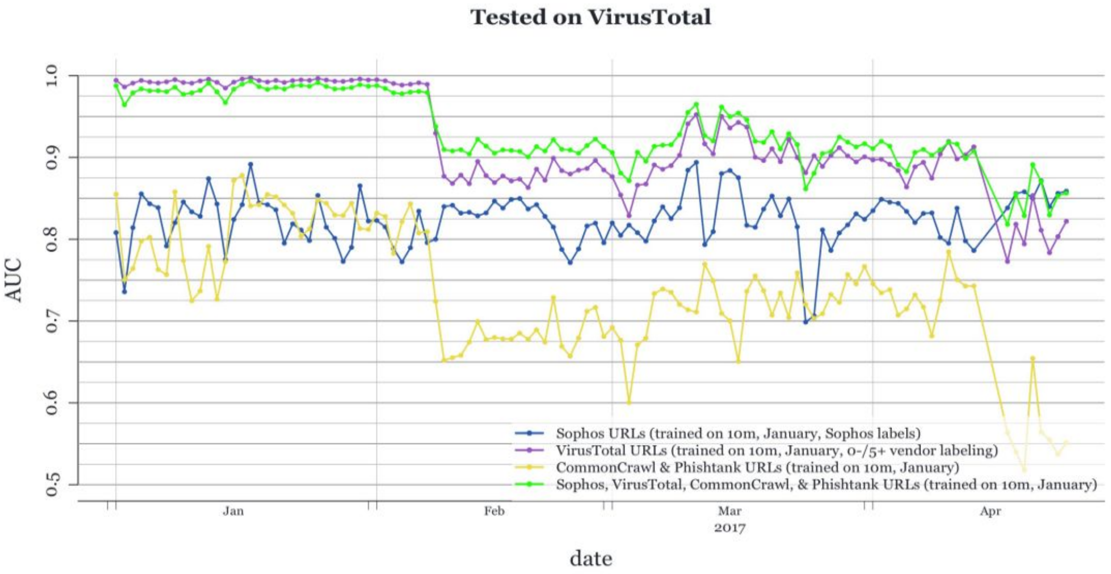
1.

The moment you put
a model in production,
it starts degrading

GARBABE IN, GARBAGE OUT

[[Sanders & Saxe, Sophos Group, Proceedings of Blackhat 2017]]

"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."



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

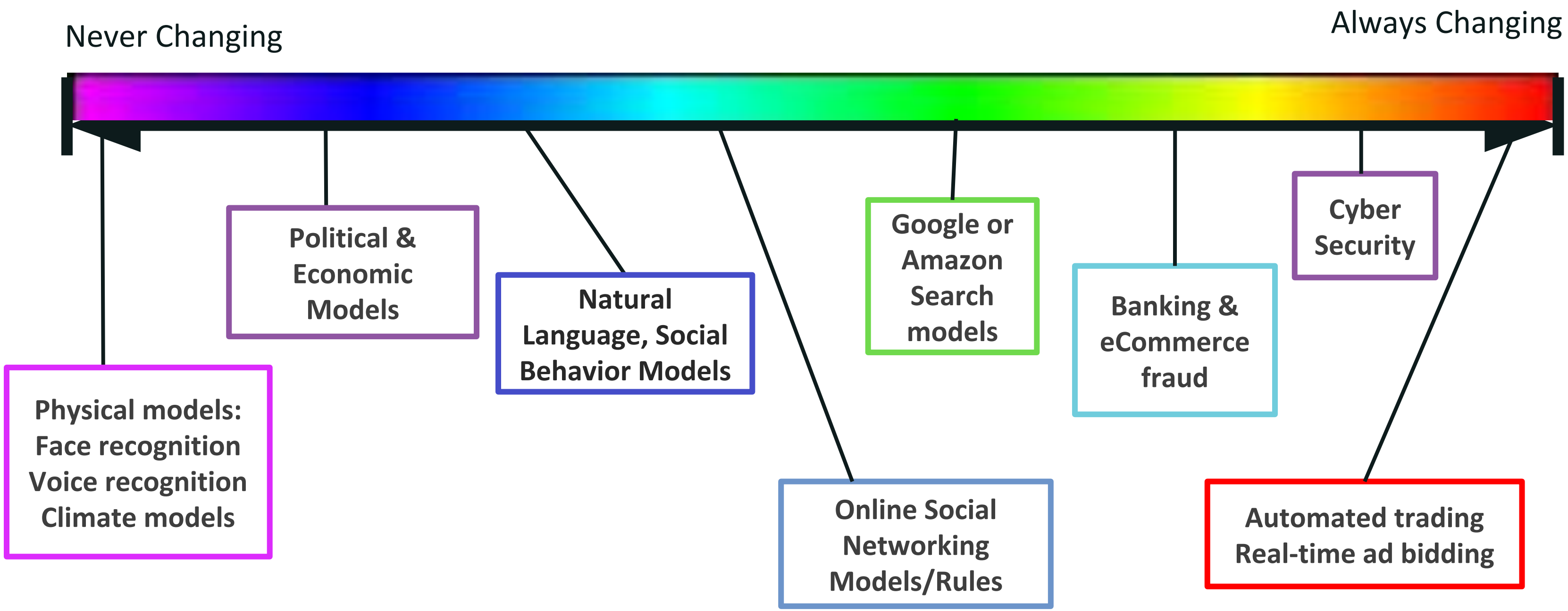
Hidden Technical Debt in Machine Learning Systems

[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.

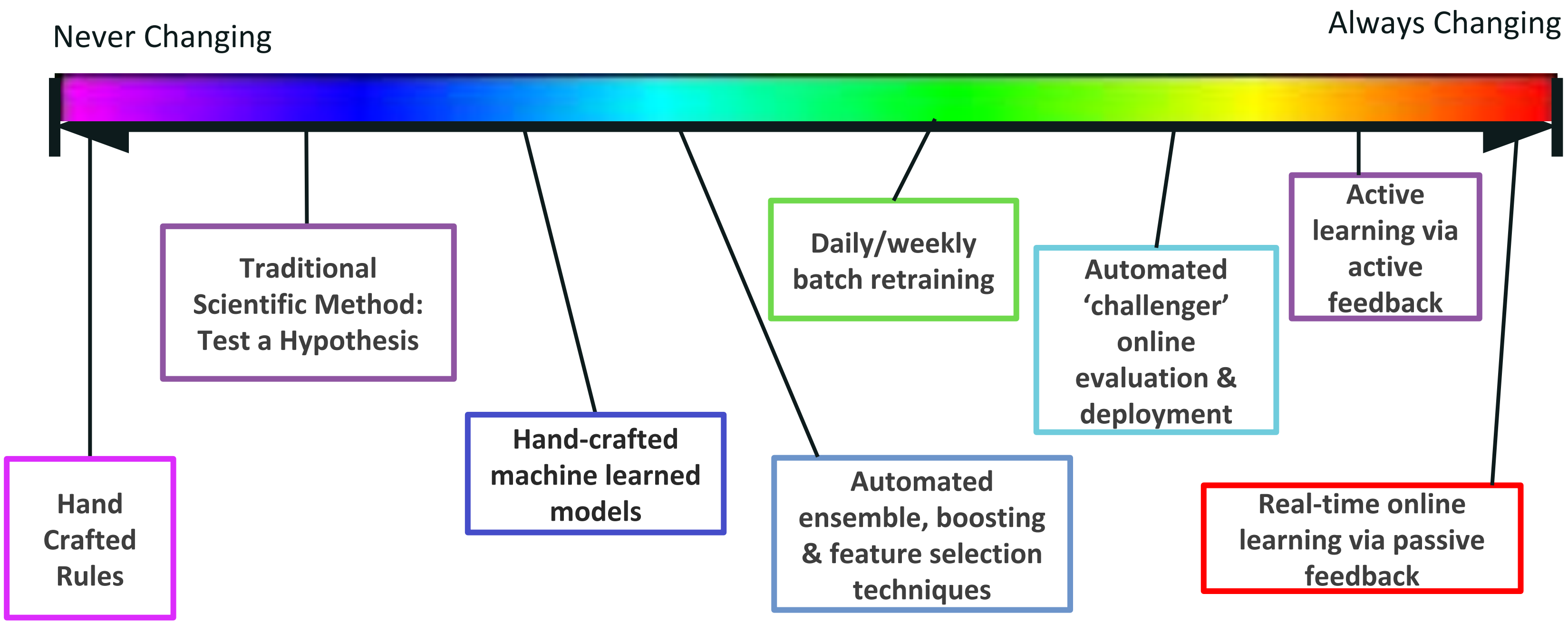
HOW FAST DEPENDS ON THE PROBLEM

(MUCH MORE THAN ON YOUR ALGORITHM)



SO PUT THE RIGHT PLATFORM IN PLACE

(MEASURE, RETRAIN, REDEPLOY)



2.

You rarely get to deploy the same model twice

REUSING MODELS IS A REPUTATION HAZARD

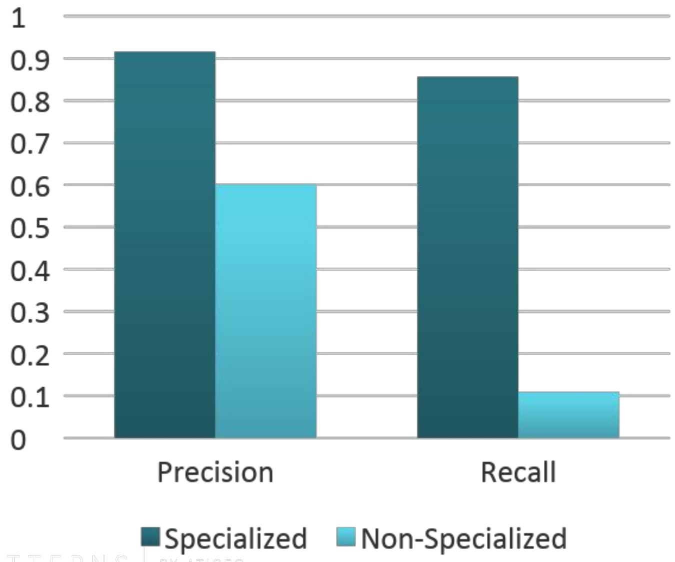
Model	Model's Goal	Sample size	Context
LACE index (2010)	30-day mortality or readmission	4,812	11 hospitals in Ontario, 2002-2006
Charlson morbidity index (1987)	1-year mortality	607	1 hospital in NYC, April 1984
Elixhauser morbidity index (1998)	Hospital charges, length of stay & in-hospital mortality	1,779,167	438 hospitals in CA, 1992

Cotter PE, Bhalla VK, Wallis SJ, Biram RW. Predicting readmissions: **Poor performance of the LACE index in an older UK population**. *Age Ageing*. 2012 Nov;41(6):784-9.

DON'T ASSUME YOU'RE READY FOR YOUR NEXT CUSTOMER

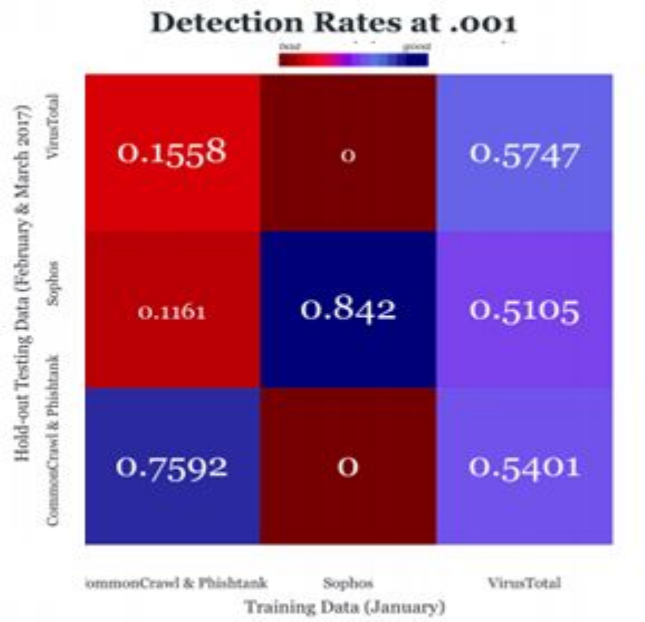
Healthcare / Natural Language

- 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)

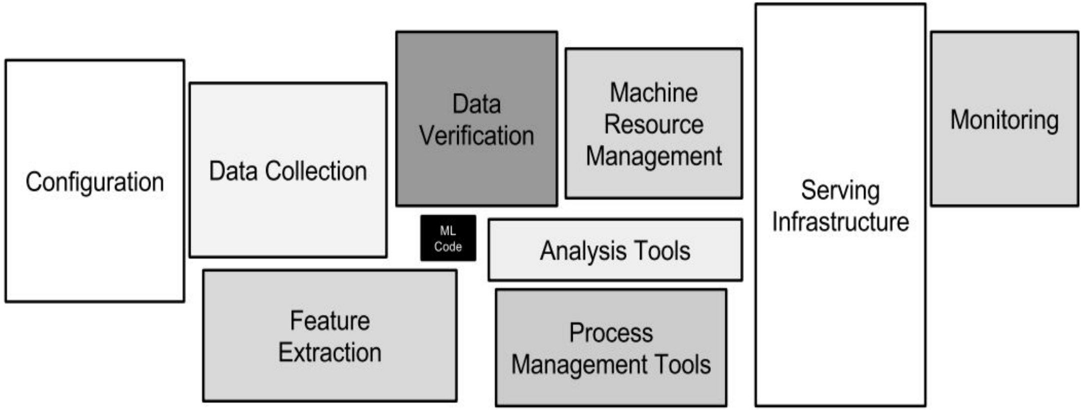


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

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

3.

It's <u>really</u> hard to know how well you're doing

HOW OPTIMIZELY (ALMOST) GOT ME FIRED

[Peter Borden, SumAll, June 2014]

“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.”

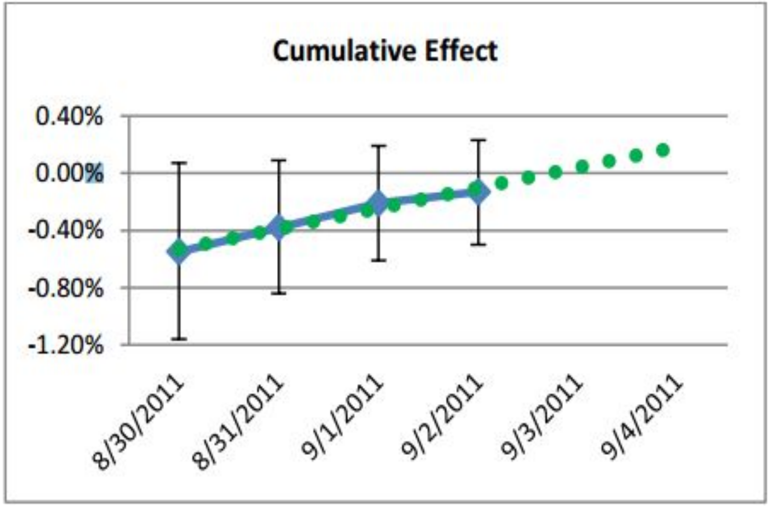
THE PITFALLS OF A/B TESTING

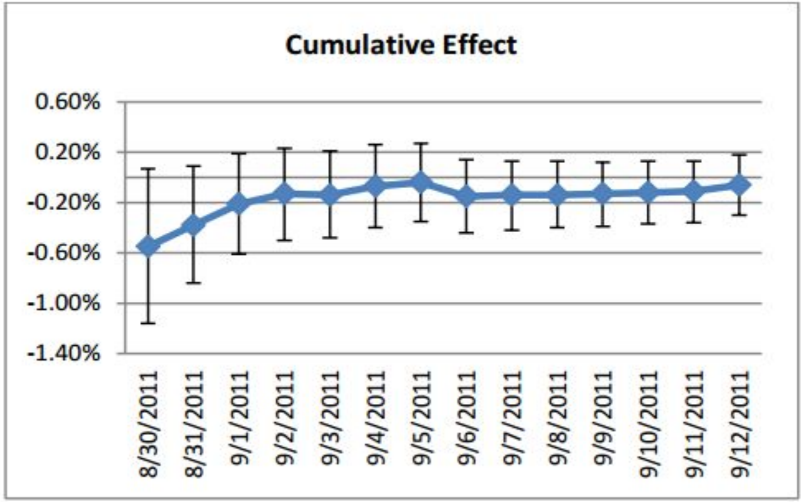
[Alice Zheng, Dato, June 2015]

separation of experiences	How many false positives can we tolerate?	What does the p-value mean?
Which metric?	How many observations do we need?	Multiple models, multiple hypotheses
How much change counts as real change?	Is the distribution of the metric Gaussian?	How long to run the test?
One- or two-sided test?	Are the variances equal?	Catching distribution drift

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

4.

Often, the real modeling work only starts in production

SEMI SUPERVISED LEARNING

BloombergBusiness

JPMorgan Algorithm Knows You're a Rogue Employee Before You Do

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

UBS fined £30m over rogue trader

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

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

Bank of America To Pay Record $16.65 Billion Fine

INVESTING 4/25/2015 @ 11:22AM | 6,590 views

Deutsche Bank's Record Fine Reveals Its Rotten Heart

IN NUMBERS

99.9999%

'Good' messages

6+

Months
per case

50+

Schemes
(and counting)

ADVERSARIAL LEARNING

Medicare Fraud Horror: Cancer Doctor Indicted for Billing Unnecessary Chemo

Michigan oncologist Farid Fata allegedly squeezed profits out of patients by prescribing unneeded treatments and inventing diagnoses

Fairfield Doctor Pleads Guilty to Illegally Prescribing Oxycodone, Health Care Fraud

Several pharmacists stopped filling the doctor's prescriptions to patients who showed obvious signs of addiction, according to prosecutors.

3 Identity Theft Horror Stories That Will Make Your Toes Curl

by David on October 25, 2013 in Law, Money, Organization, Regret, Security

Is This Fraud Too Big Even For 60 Minutes?

March 10, 2012 | 306,904 views

ADVERSARIAL LEARNING

Medicare And Medicaid Fraud Is Costing Taxpayers Billions

Forbes

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 <u>after</u> going to production

SOFTWARE DEVELOPMENT

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 operating in production

THANK YOU!

david@pacific.ai

@davidtalby

in/davidtalby