

# LEARNING IN CYCLES

## IMPLEMENTING SUSTAINABLE MACHINE LEARNING MODELS IN PRODUCTION

*Andrew Therriault  
Chief Data Officer, City of Boston*

*Presented at the PyData Boston Meetup  
October 2017*



City of Boston  
Mayor Martin J. Walsh



Innovation & Technology

# GETTING STARTED

---

## MY BACKGROUND

Chief Data Officer for the City of Boston, leading our Analytics Team and steering the city's overall plans for using data.

Other resume highlights:

- *PhD in poli-sci from NYU*
- *Post-doc at Vanderbilt*
- *Data science consultant*
- *Director of Data Science at the Democratic National Committee (2014 and 2016 election cycles)*

## WHICH MEANS

I've made **lots** of models.

I've also made **LOTS** of mistakes.

Fortunately, at least some of what I've learned along the way should be useful to others.

## WHAT'S IN THIS TALK

This talk's about **sustainable machine learning models** - iterative models that use outputs of one iteration to generate data for future runs.

Example use cases:

- *Direct marketing & outreach*
- *Recommendation engines*
- *Email & ad customization*
- *QA and compliance audits*

We'll focus on practical approaches and considerations for data scientists, analysts, and engineers working on these kinds of models in the wild.

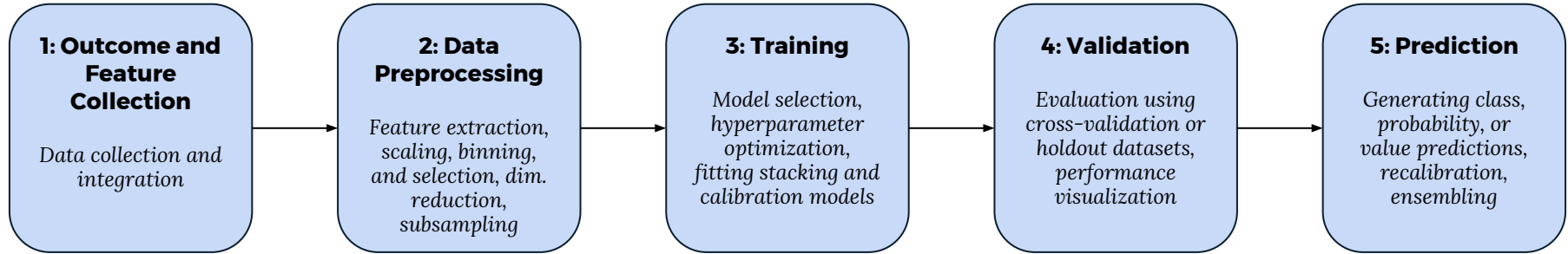
---

# WHAT'S DIFFERENT ABOUT SUSTAINABLE MACHINE LEARNING MODELS?



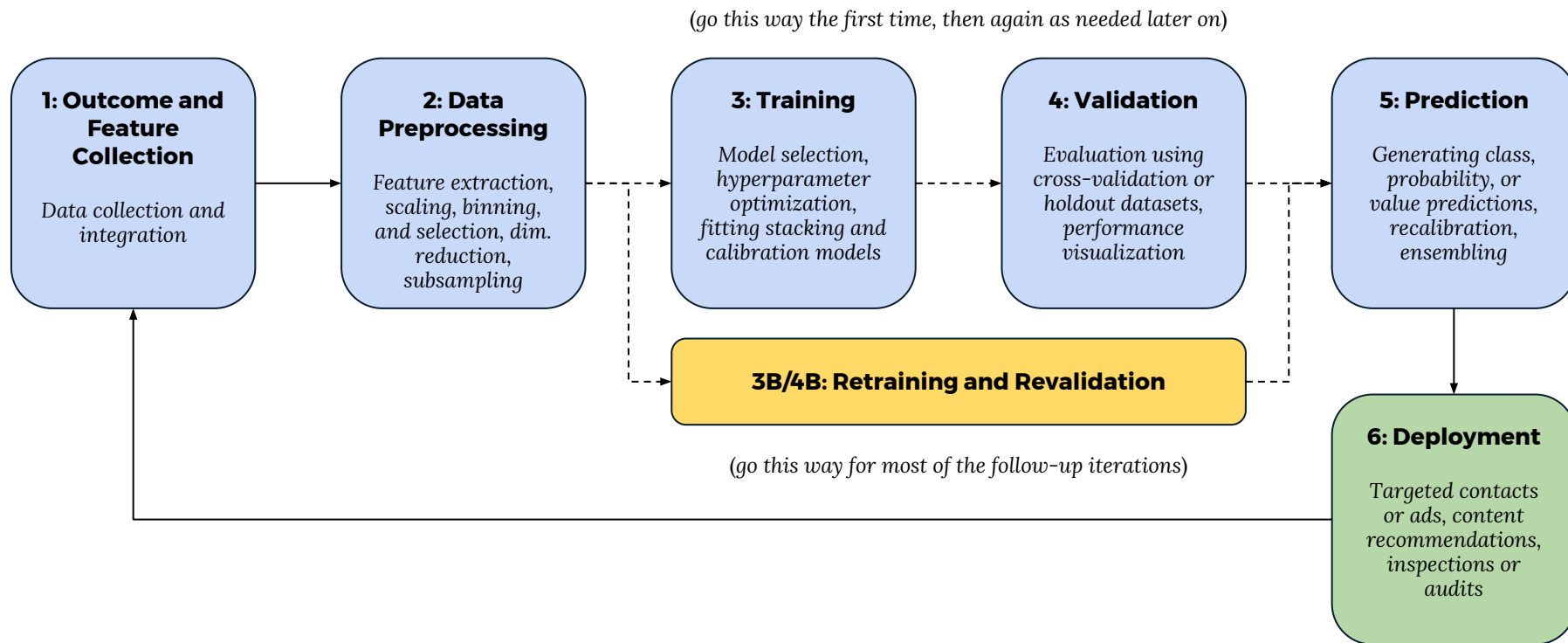
# A TRADITIONAL MACHINE LEARNING MODEL DESIGN

---



# A SUSTAINABLE MACHINE LEARNING MODEL DESIGN

---



# SO WHAT'S THE BIG DEAL?

---

## PICKING OPTIMIZATION METRICS

---

*A single-run model can have a clear technical goal, but a sustainable model has a systemic goal that's harder to translate to a simple metric*

## EXPLORATION OR EXPLOITATION?

---

*There are benefits to targeting sub-optimally in a given iteration, both to learn more and to preserve future value, but the cost is lower performance now*

## OFTEN NO FORMAL SOLUTION

---

*It's tempting to try to come up with a mathematical solution for the repeated optimization problem, but that's not feasible for most complex models*

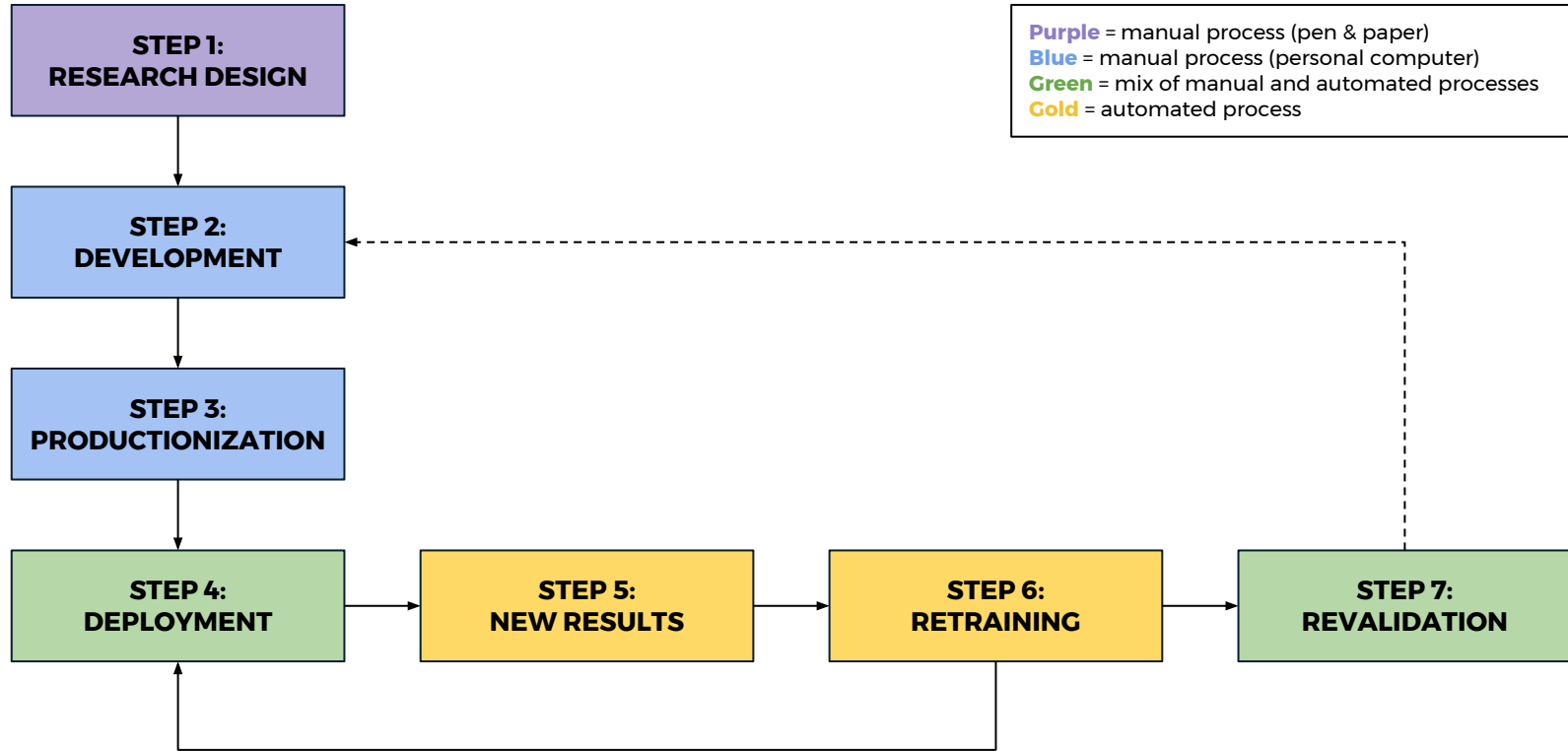
## WE'RE MAKING SOFTWARE NOW

---

*Running a model repeatedly is an obvious case for automation, so we need to think about things like version control, scheduling, monitoring, alerts, etc.*

# MY RECOMMENDED APPROACH TO SUSTAINABLE ML MODELING

---







# THE IMPORTANCE OF DIVERSE DATA



# WHY DOES DIVERSE DATA MATTER FOR SUSTAINABILITY?

---

## THE TRADEOFF

The key benefit of a sustainable ML model: using organic data for model updates, **new info is gathered at little or no cost.**

The downside of this approach: unlike data from a distinct research process (surveys, experimental tests, etc.), **this information will be biased.**

What makes a model “sustainable”: when the data collected is sufficiently diverse, **biases can often be overcome.**

## DATA COVERAGE

The most obvious problem with organic data is that it will only cover a narrow subset of the total possibilities (potential customers, ad variations, etc.).

When that happens, you don’t learn anything new about the rest of the universe of options.

That matters because **without new information about alternatives, it’s hard to learn anything that will suggest changing course.**

## DATA RELIABILITY

Even worse, having an unrepresentative sample of data can lead you to bad inferences about the rest of the universe.

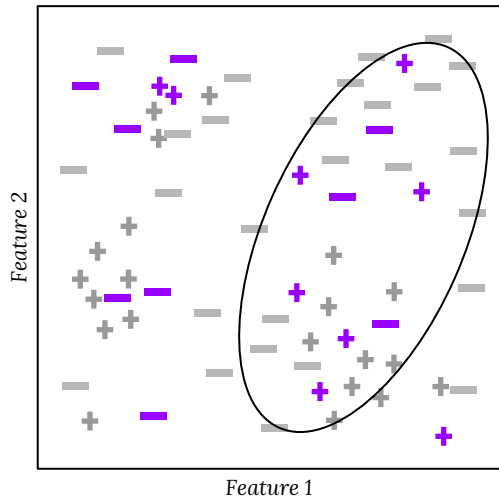
This is the same selection bias problem that shows up in other fields, most notably survey research.

Even when we know the data generating process, **models often can’t recognize outliers without some examples of “normal” observations to compare to.**

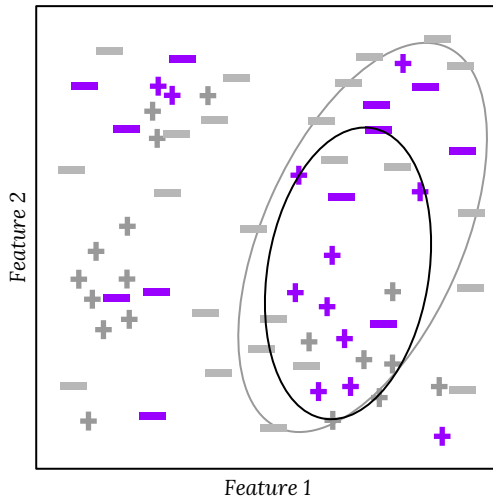
# WHAT NOT TO DO #1: TUNNELING DOWN TOO NARROWLY

Initial data and target universe  
selected by 1st model run

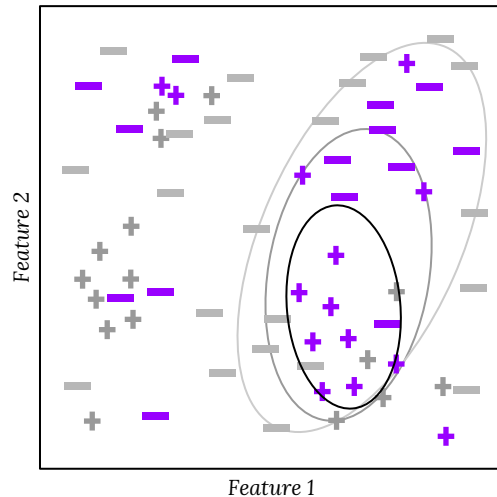
(purple = observed, gray = unobserved)



Data after some new results,  
and target universe selected by  
2nd model run



Data after more new results,  
and target universe selected by  
3rd model run



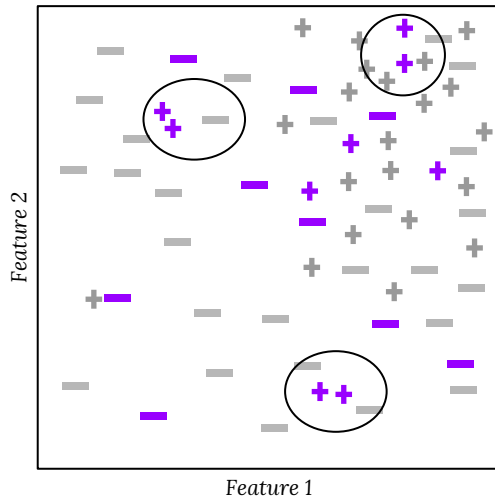
While this gives us a very high-propensity target universe by the time you get to the third run, it's a **very** narrow part of the overall universe and misses a lot of potential good targets. This will only get worse over time.

*Real-world example: Age filtering in direct mail fundraising*

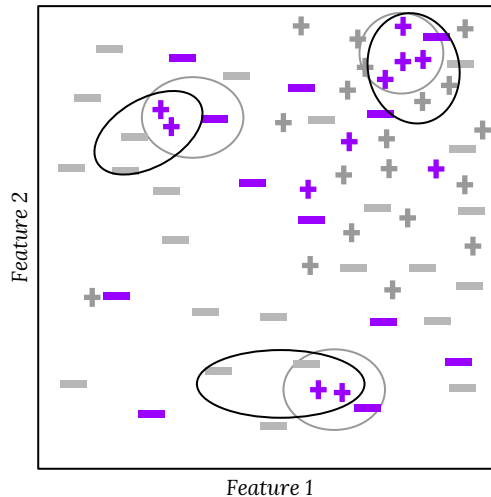
# WHAT NOT TO DO #2: PUTTING TOO MUCH TRUST IN BIASED DATA

**Initial data and target universe  
selected by 1st model run**

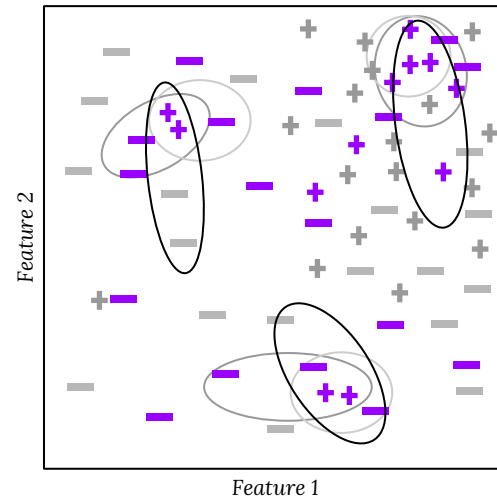
(purple = observed, gray = unobserved)



**Data after some new results,  
and target universe selected by  
2nd model run**



**Data after more new results,  
and target universe selected by  
3rd model run**



Knowing that bias is built-in to your data generating process, it's important to approach your data with a bit of skepticism and be ready to change course when the model's not working the way you expect.

***Real-world example: Republicans targeted by a Democratic volunteer recruitment model***

# FIVE STRATEGIES FOR KEEPING YOUR DATA DIVERSE

---

<b>BROAD-BASED SAMPLING</b>	<p>Deliberately collect new data from outside your optimal universe, in order to ensure wide coverage.</p> <p><i>Real-world example: Annual restaurant inspections</i></p>
<b>SUPPLEMENTATION</b>	<p>Bring in comparable data from other sources to expand coverage to a broader universe.</p> <p><i>Real-world example: Shared fundraising lists</i></p>
<b>WEIGHTING TO LEARN</b>	<p>Adjust your optimization calculations so that higher priority is given to options which are uncertain.</p> <p><i>Real-world example: Direct mail prospecting ensembles</i></p>
<b>ROUGH TARGETING</b>	<p>Add noise to your final predictions so that a wider range of options is tried.</p> <p><i>Real-world example: Rounding volunteer propensity scores</i></p>
<b>HIGH-CHURN MODELING</b>	<p>Build your model in such a way that it will quickly update its recommendations and try out a wider range of possibilities.</p> <p><i>Real-world example: Phone quality scores</i></p>



---

# RECOMMENDATIONS FOR MODEL DESIGN

# BUILDING FEATURES FOR SUSTAINABLE MODELS

---

## CYCLICAL PATTERNS

---

*Add hour of day, day of week, time until/since important events, flags for specific periods of time, etc., as features so that the model's predictions will account for when they're being generated*

## DYNAMIC FEATURES

---

*Bring in external sources of data that change over time, such phone / address / email validations or histories with other organizations or programs, so that you're not in an entirely-closed system*

## INDIVIDUAL HISTORIES

---

*Include results from prior contacts or assignments, exposure to previous treatments, behavior after treatments, and other history data so that the model can adapt over time on an individual level*

## COMPONENT MODEL OUTPUTS

---

*Predictions from other models (supervised or unsupervised) of fundamental characteristics that are regularly updated can keep your model fresh without drowning it in individually-weak features*



# MODEL SELECTION AND TUNING

---

## BUILD IN HETEROGENEITY

---

*Models that allow complex interactions (random forests, boosted trees) will give more variety within and across iterations than simpler models (linear regression, naive bayes)*

## RETRAIN FREQUENTLY

---

*Not every time you predict necessarily, but often enough for new results to be incorporated quickly*

## EMBRACE ENSEMBLES

---

*Tree ensembles, models with varied hyperparameter values, bagging and stacking, novel ensemble combination methods*

## ROTATE YOUR DATA

---

*As you add new data, start rotating out older data once you have a sufficient amount to keep the model from getting increasingly stale*

# EVALUATING AND MONITORING PERFORMANCE

---

## OUT-OF-CONTEXT VALIDATION

---

*Don't just look at random holdout set, hold out batches of results by time period or group - for example, use most recent results as test set to proxy for next batch's performance*

## HOLD YOURSELF ACCOUNTABLE

---

*Keep archives of predictions and go back later on and see how they did, to make sure that your validation strategy is giving you realistic expectations and to find opportunities for improvement*

## TRACK CHANGE OVER TIME

---

*Be on the lookout for sudden changes in performance, feature importances, or recommendation patterns, as these can be signs of problems with the model or data*

## STAY ON TOP OF THINGS

---

*Set up automated monitoring and alerts for both the models and incoming results - for example, keep a Jupyter notebook of all the things you check frequently and automate it with nbconvert*

# FINAL ADVICE

---

## KNOW YOUR DATA

Your model is only as good as its data, and when you're relying on data you know is biased, it's critical to know specifically how so that you can adjust for it.

Spend some time understanding the data generating process, from a human perspective, away from your IDE or notebooks.

Talk to the subject matter experts, log on to the app or website, and **make sure you understand what the data you're relying on really means.**

## DO SANITY CHECKS

The main benefit of these models is not that they deliver mysterious insights, but that they can synthesize a lot of obvious things at scale.

Make a habit of inspecting feature importances, looking at top/bottom tiers' summary stats, and spot-checking individual predictions to see if they make sense .

Even if the calculations aren't directly interpretable, **the overall patterns should still pass the smell test.**

## BE OPEN TO CHANGE

Don't expect to get your model right the first time: more likely than not, things will go wrong somehow, and even if not there's always room for improvement.

Though you can automate the periodic updates, plan to routinely go back to the model and make changes by hand.

**Just because a model is “sustainable” doesn't mean it does all the work on its own.**

For more information about the City of Boston's Analytics Team, go to  
[boston.gov/analytics](https://boston.gov/analytics).

And if you want to talk more:  
[andrew.therriault@boston.gov](mailto:andrew.therriault@boston.gov) (work)  
[andrew.therriault@gmail.com](mailto:andrew.therriault@gmail.com) (personal)  
[@therriaultphd](https://twitter.com/therriaultphd) (twitter)

**THANK YOU!**



City of Boston  
Mayor Martin J. Walsh



Innovation & Technology