Research Opportunities in AutoML

Rich Caruana
Microsoft Research

Motivation

75% of Machine Learning is preparing to do machine learning

75% of Machine Learning is
preparing to do machine learning

and 15% is what you do afterwards…

75% of Machine Learning is
preparing to do machine learning

and 15% is what you do afterwards…

most ML research about the middle 10%

75% of Machine Learning is
preparing to do machine learning

and 15% is what you do afterwards…

~~most ML research about the middle 10%~~

Goals For This Talk

- Foster research on the complete ML pipeline
- Describe a few open problems in AutoML
- Suggest future challenge/competition problems

- Ultimate goal is to make the practice of ML more reliable so you don't need a Ph.D. in ML + 10 years experience to do ML well

- How/Where do we start?
  - Start by looking at difference between ML in Lab and ML in the field


MATT DAMON JODIE FOSTER
ELYSIUM
FROM THE DIRECTOR OF 'DISTRICT 9'
IN THEATERS AND IMAX
www.moviedeskback.com


MATT DAMON JODIE FOSTER
ELYSIUM
FROM THE DIRECTOR OF 'DISTRICT 9'
IN THEATERS AND IMAX
www.moviedeskback.com

ML Research vs. **Engineering**

UC-Irvine/CIFAR — Real-World

UC-Irvine/CIFAR vs. Real-World

UC-Irvine/CIFAR vs. Real-World

- Download data
  - No collection, cleaning, …
- Know how well others did
- Metric(s) pre-defined
- Change algs, params, and coding until doing well on metric
- Sometimes add data

UC-Irvine/CIFAR vs. Real-World

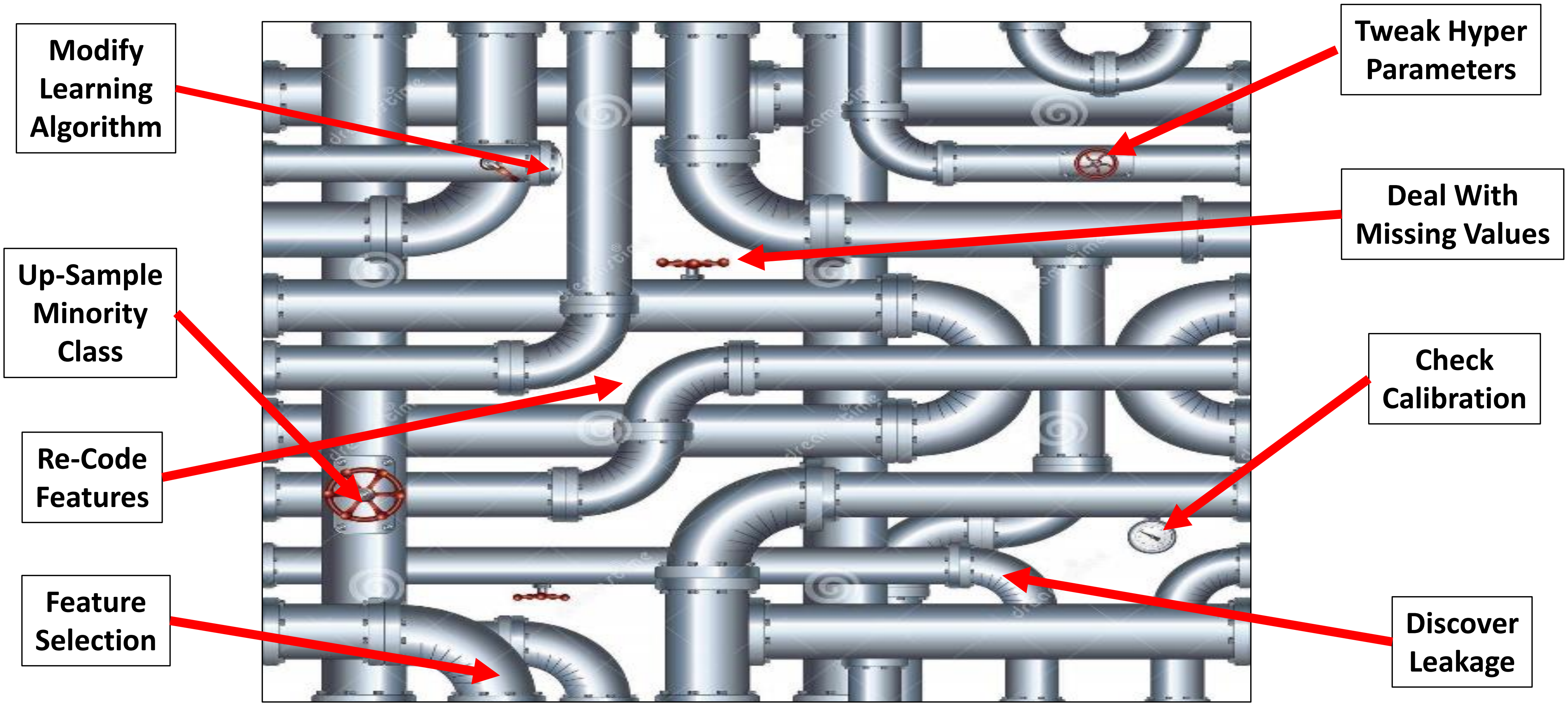
- Download data
  - No collection, cleaning, …
- Know how well others did
- Metric(s) pre-defined
- Change algs, params, and coding until doing well on metric
- Sometimes add data

- Problem undefined
- Don’t know how well you can do
- Add new features and data feeds
  - Clean, clean, clean
  - Most effort goes into the data!
- Coding of data is critical
- Choose practical algorithms
- Debug, debug, debug
- Wash, rinse, repeat
  - month after month after month!

Surprisingly, the research pipeline is complex because

we assume the researcher is an expert

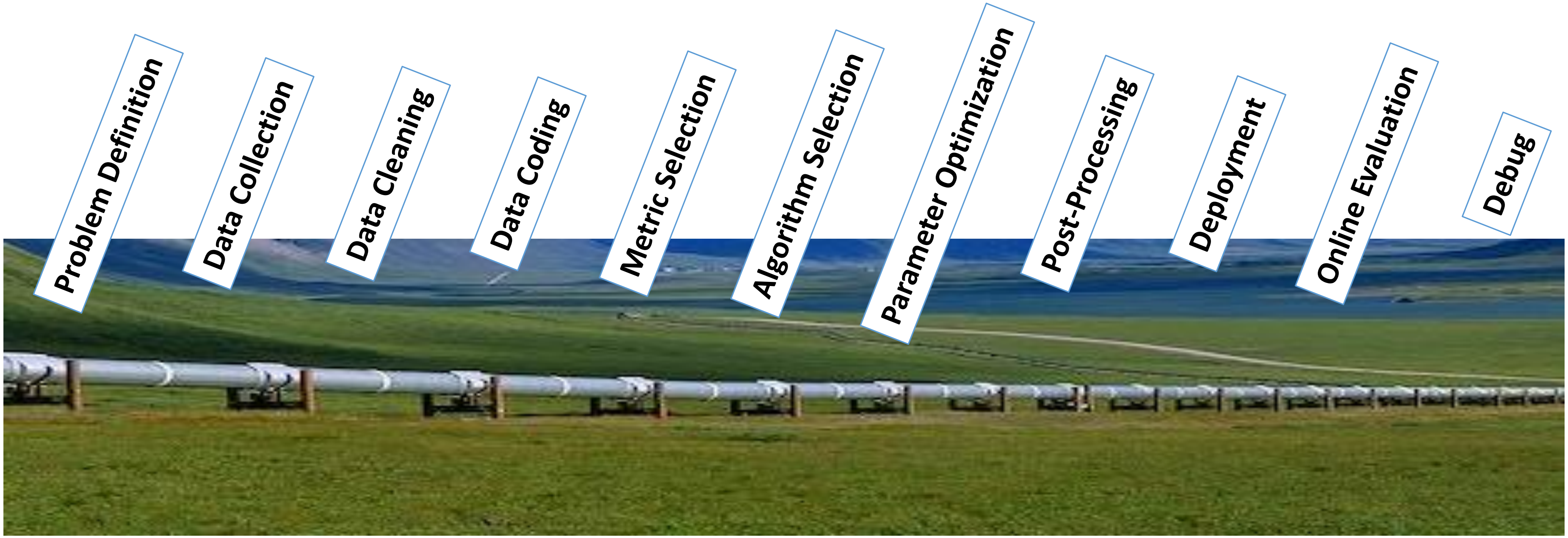
Machine Learning (Research) Pipeline



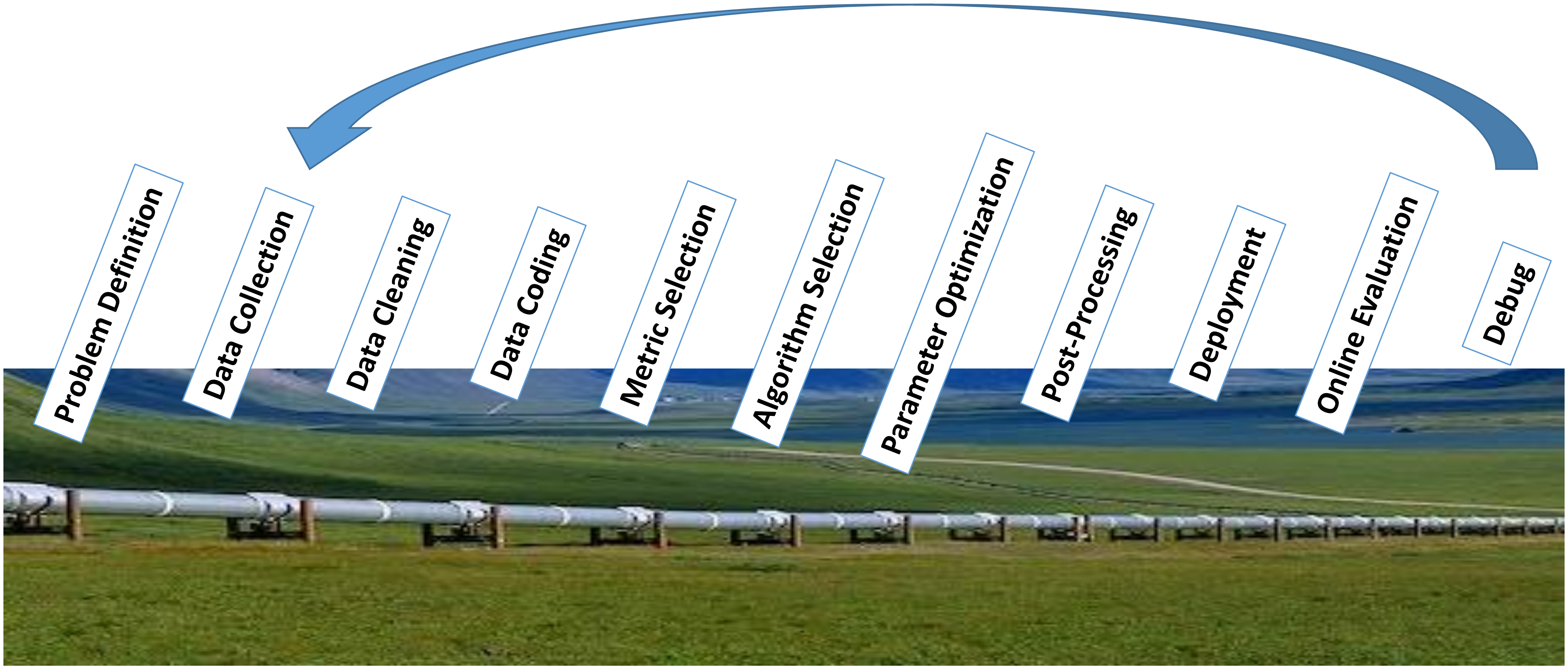
Machine Learning (Engineering) Pipeline

- By real engineers, teams of engineers, …
- On real data, to real metrics, …
- On schedule, on budget, …
- Must be maintainable, repeatable, documentable, …

Machine Learning (Engineering) Pipeline

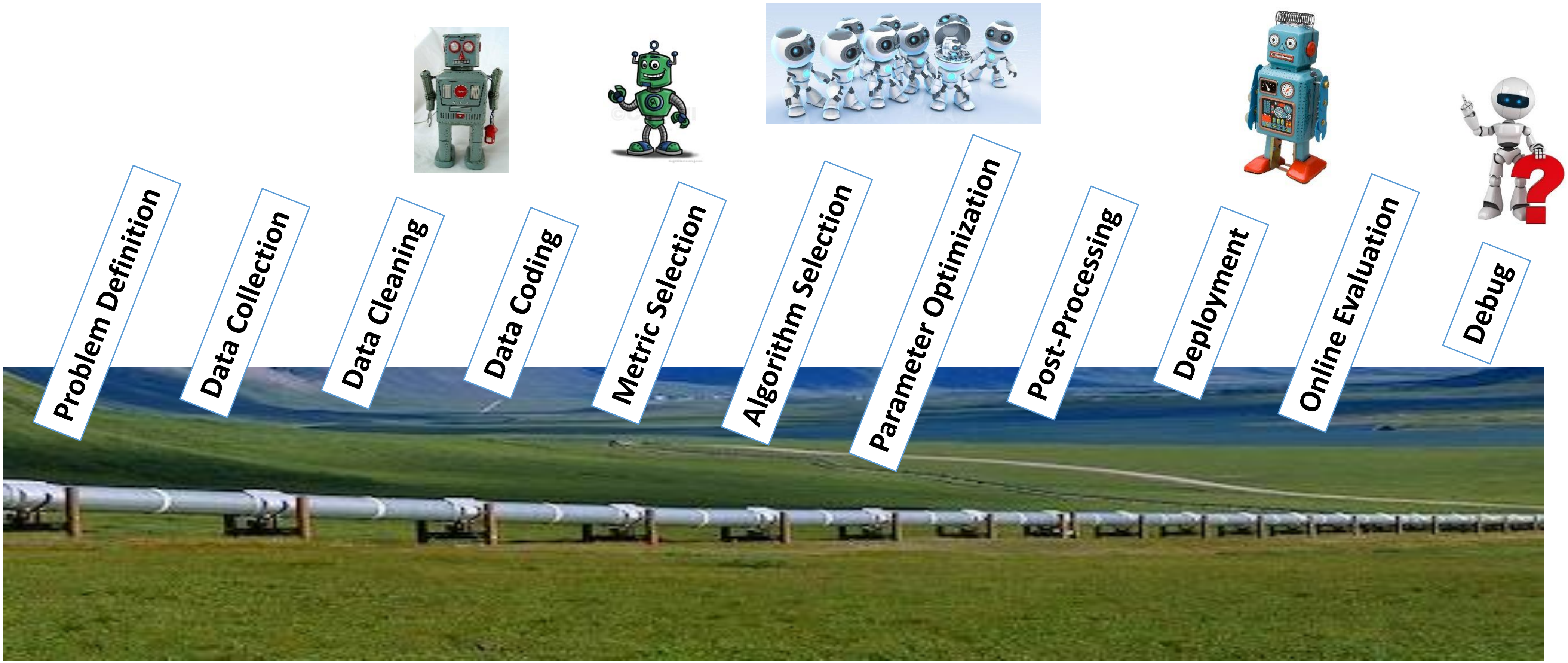


Machine Learning (Engineering) Pipeline



Each step in the pipeline is an opportunity to do AutoML research

Future AutoML (Engineering) Pipeline

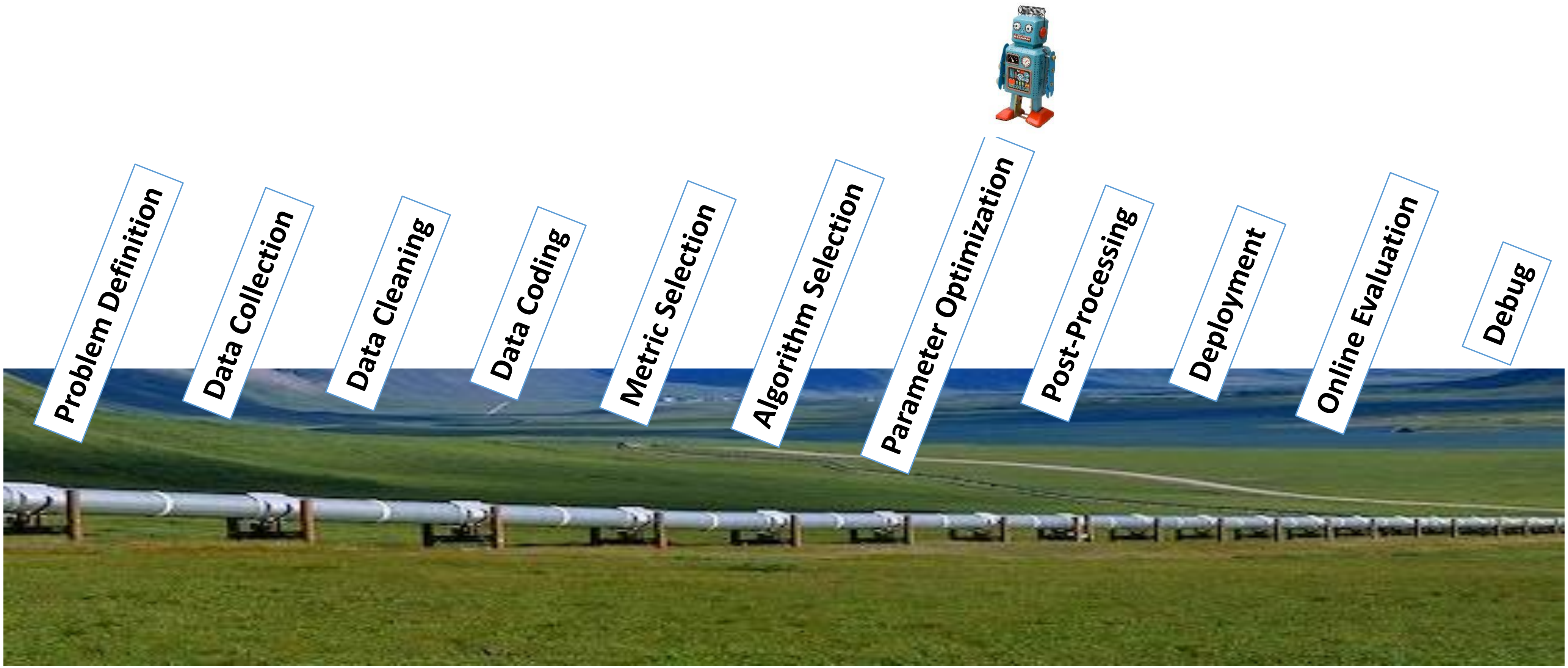


Goals For This Talk

- Foster research on the complete ML pipeline
- Describe a few open problems in AutoML
- Suggest future challenge/competition problems

- So let’s just jump in…

Future AutoML (Engineering) Pipeline



Importance of Hyper-Parameter Optimization

- Hyper-Parameter Optimization is most mature subarea in AutoML
  - Manual heuristic search: surprisingly sub-optimal
  - Grid search: effective with small number of parameters
  - Random search: better than grid with larger number of parameters
  - Bayesian Optimization: better than random with very large # parameters
  - …
- With modern algorithms (boosting, deep neural nets, …) parameter optimization is much more critical than you might think…
  - … because modern high-flying algorithms are all low-bias, high variance
- How many people here use automatic hyper-parameter optimization?

Importance of Hyper-Parameter Optimization

- Around 2000-2005, some thought supervised learning was done

- Quiz: they thought best algorithm was:

  - Neural Nets?
  - Boosting?
  - Random Forests?
  - SVMs?

Importance of Hyper-Parameter Optimization

- Around 2000-2005, some thought supervised learning was done
- Quiz: they thought best algorithm was:
    - Neural Nets?
    - Boosting?
    - Random Forests?
    - **SVMs?**

Importance of Hyper-Parameter Optimization

- ✓SVMs (circa 2000-2005)
- ✓Bing Ranker: FastRank vs. NeuralNet Ranker (circa 2010)
- ✓Best DNN on CIFAR-10 and -100 use massive parameter optimization
  - ✓Optimize usual hyper-parameters such as learning rate, initialization, drop-out
  - ✓Optimize hyper-parameters per layer(s)
  - ✓Optimize augmentations
  - ✓Optimize network architecture
- ✓Our results: +1-4% on DNNs by doing careful Bayesian Optimization
- ✓TIMIT benefits from careful hyper-parameter optimization
  - ✓Why didn’t deep nets get discovered in mid 90’s?
  - ✓Didn’t explore the space and hyper-parameters thoroughly enough?

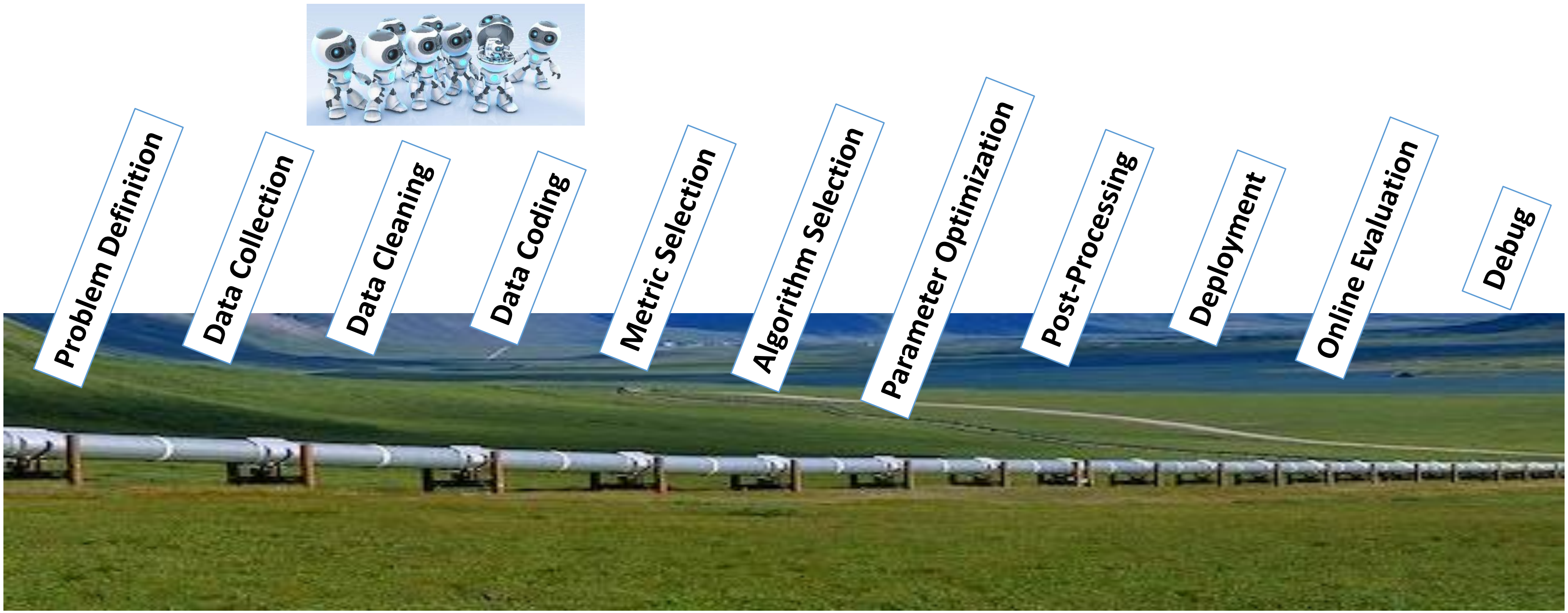
ML Algorithm is an Important Hyper-Parameter

	Threshold Metrics			Rank/Ordering Metrics			Probability Metrics			
Model	**Accuracy**	**F-Score**	**Lift**	**ROC Area**	**Average Precision**	**Break Even Point**	**Squared Error**	**Cross-Entropy**	**Calibration**	**Mean**
BEST	**0.928**	**0.918**	**0.975**	**0.987**	**0.958**	**0.958**	**0.919**	**0.944**	**0.989**	**0.953**
BST-DT	**0.860**	**0.854**	**0.956**	**0.977**	**0.958**	**0.952**	**0.929**	**0.932**	**0.808**	**0.914**
RND-FOR	**0.866**	0.871	**0.958**	**0.977**	**0.957**	**0.948**	0.892	0.898	0.702	0.897
ANN	**0.817**	**0.875**	**0.947**	0.963	0.926	0.929	0.872	0.878	**0.826**	0.892
SVM	0.823	0.851	0.928	0.961	0.931	0.929	0.882	0.880	0.769	0.884
BAG-DT	0.836	0.849	0.953	0.972	**0.950**	0.928	0.875	0.901	0.637	0.878
KNN	0.759	0.820	0.914	0.937	0.893	0.898	0.786	0.805	0.706	0.835
BST-STMP	0.698	0.760	0.898	0.926	0.871	0.854	0.740	0.783	0.678	0.801
DT	0.611	0.771	0.856	0.871	0.789	0.808	0.586	0.625	0.688	0.734
LOG-REG	0.602	0.623	0.829	0.849	0.732	0.714	0.614	0.620	0.678	0.696
NAÏVE-B	0.536	0.615	0.786	0.833	0.733	0.730	0.539	0.565	0.161	0.611

Importance of Hyper-Parameter Optimization

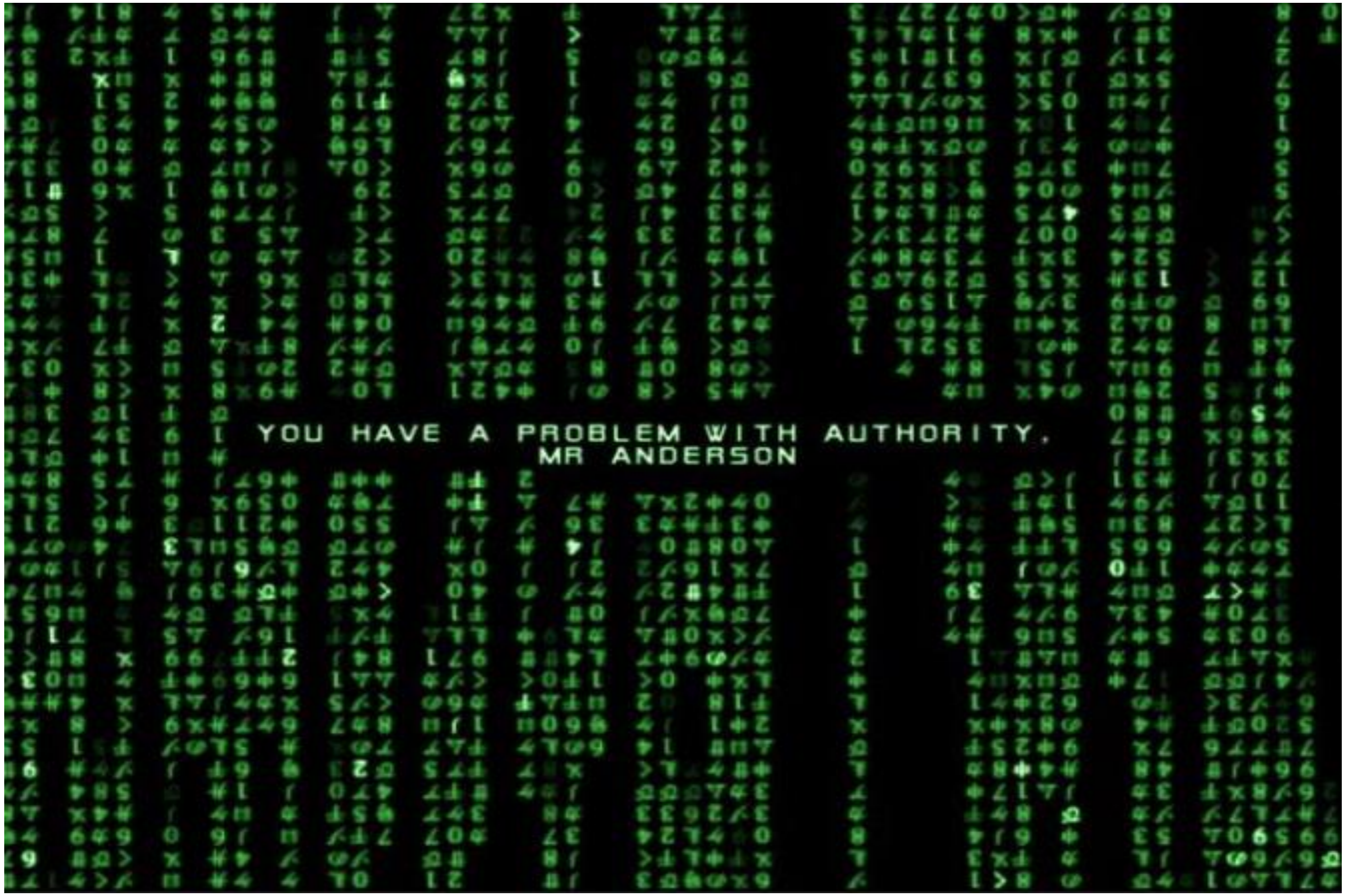
- Hyper-parameter optimization is example of what AutoML can achieve
- 20 years ago selecting hyper-parameters were part of the craft of ML
  - Neural nets: number of hidden units, learning rate, momentum, …
  - Knowing how to select hyper-parameters is part of what made you an expert
- Now, multiple papers and algorithms for hyper-parameter optimization
- Thriving research community with multiple workshops
- Makes a significant difference in accuracy of trained models
- Open source code
- Need to view other steps in ML pipeline as new research opportunities

Future AutoML (Engineering) Pipeline



Tools to Better Understand Data

NEVER Trust the DB/Data Spec!!!



AutoML Tools to Better Understand Data

- Auto variable type determination
  - 0, 1, 2, 3, 4, 5: nominal, ordinal, integer, continuous?
  - Are there dates in fields?
  - Is a field a unique identifier or sequence number?
- Auto coding
  - Different coding needed for NNs, SVMs, KNN vs. decision tree-based methods
- Auto missing value detector
  - 0, 1, 2
  - -1, 0, +1
  - Can’t just try everything --- missing variables often cause leakage!
- Auto anomaly detection
  - spurious strings, missing entries in table, …

First Real Data Set I Worked With (1995 ☹)

- Pneumonia data from 1992-1995
- 14,199 patients
- < 200 features
- mix of Booleans, categorical, and continuous variables
- missing values
  - MAR --- Missing At Random
  - missing correlated with target class (caused leakage!)

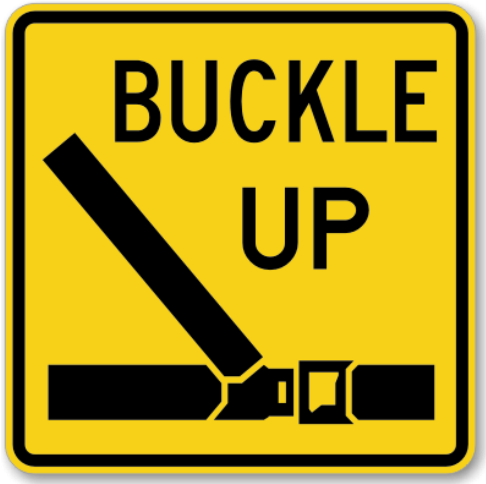
- Quickly wrote simple unix utility to help better understand the data colstats demo…

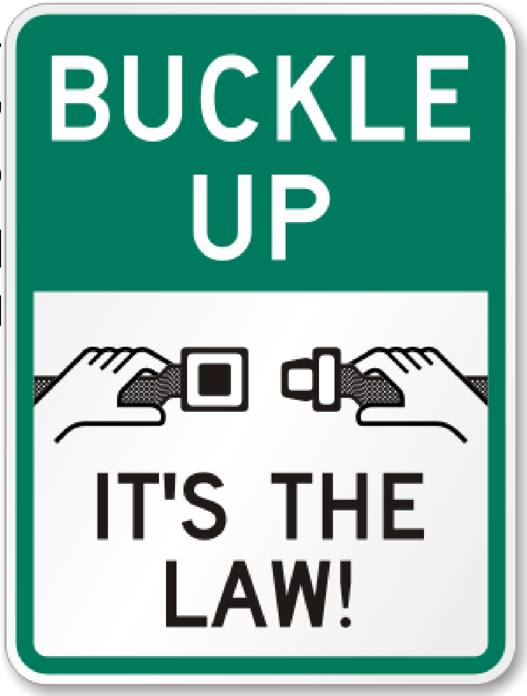
DataDiff

- Automatically recognize changes in data
  - changes in DB design, broken sensors, new semantics, new feeds, …
  - In real world, DBs and data sources are living, breathing, evolving entities
  - Humans make mistakes, forget what they did 1st time, retire, …
    - C-Section 1993-1995 vs. 1996-1998 data (missing values recoded, …)
    - 30-day Hospital Re-Admission 2011-2013 vs new 2014 sample
- dDiff is not as trivial as it might seem:
  - Density estimation is hard in high dimensions (but this is a special case)
  - Don’t care about simple drift if learned model can handle it
    - E.g., from 50-50 male-female to 40-60 male-female
  - Care most about changes that affect model accuracy or utility
  - Warning flags, default to more robust model, auto-retrain/adapt, …

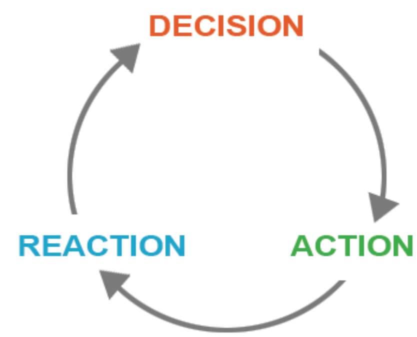
Model Protection Wrappers

- Model trained to predict 30-day re-admission was deployed at a children's hospital
- Real-
  - w
  - t ferent from train data
- Mod was traine tects run-t eaningful
- Can ard practi





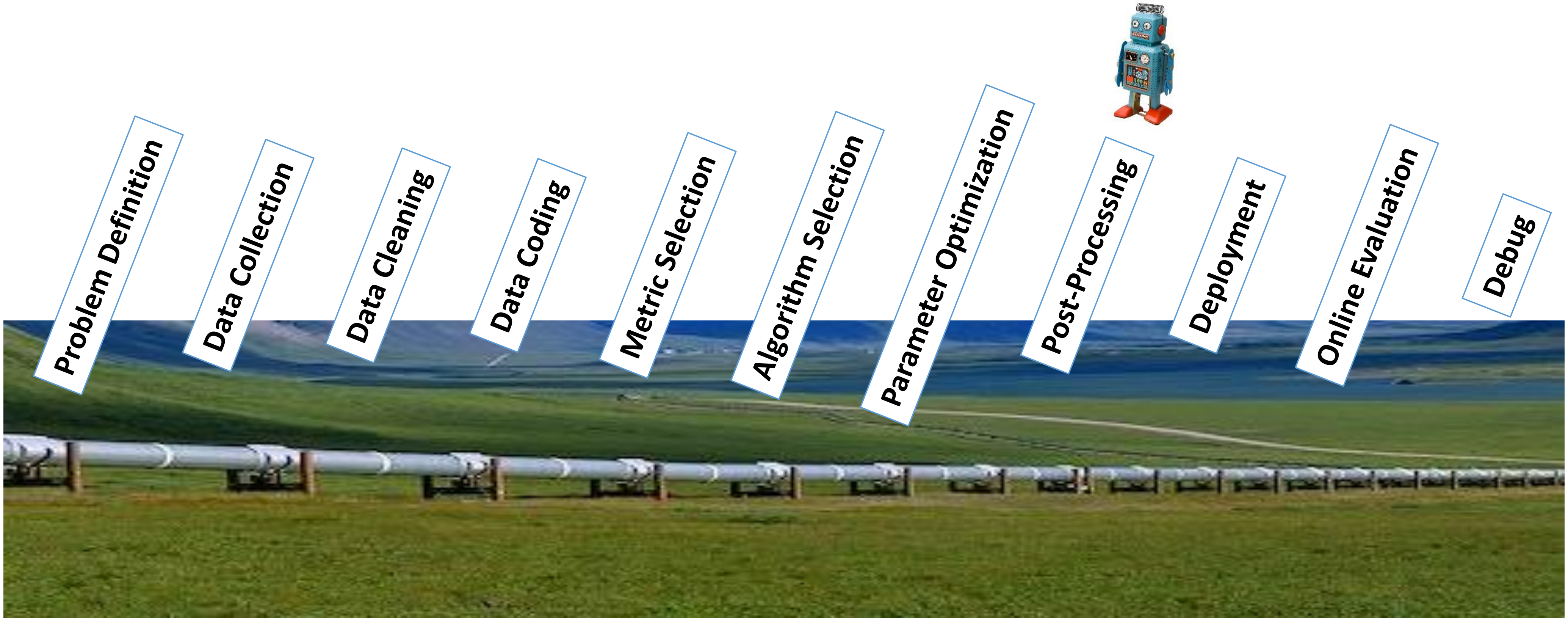
Feedback --- the Future Curse of ML!



- If you train a model on patient data
- And model is used to change practice of medicine (intervention)
- Next time you collect data it is affected by model…
- …so how do you collect unbiased data 2nd time around?

- This is a deep, fundamental problem that in some domains is not easy to solve (ethically, or efficiently) --- problem with non-causal learning
- Could approach this as a dDiff problem that looks not just at input features but at labels and relationship between inputs and outputs

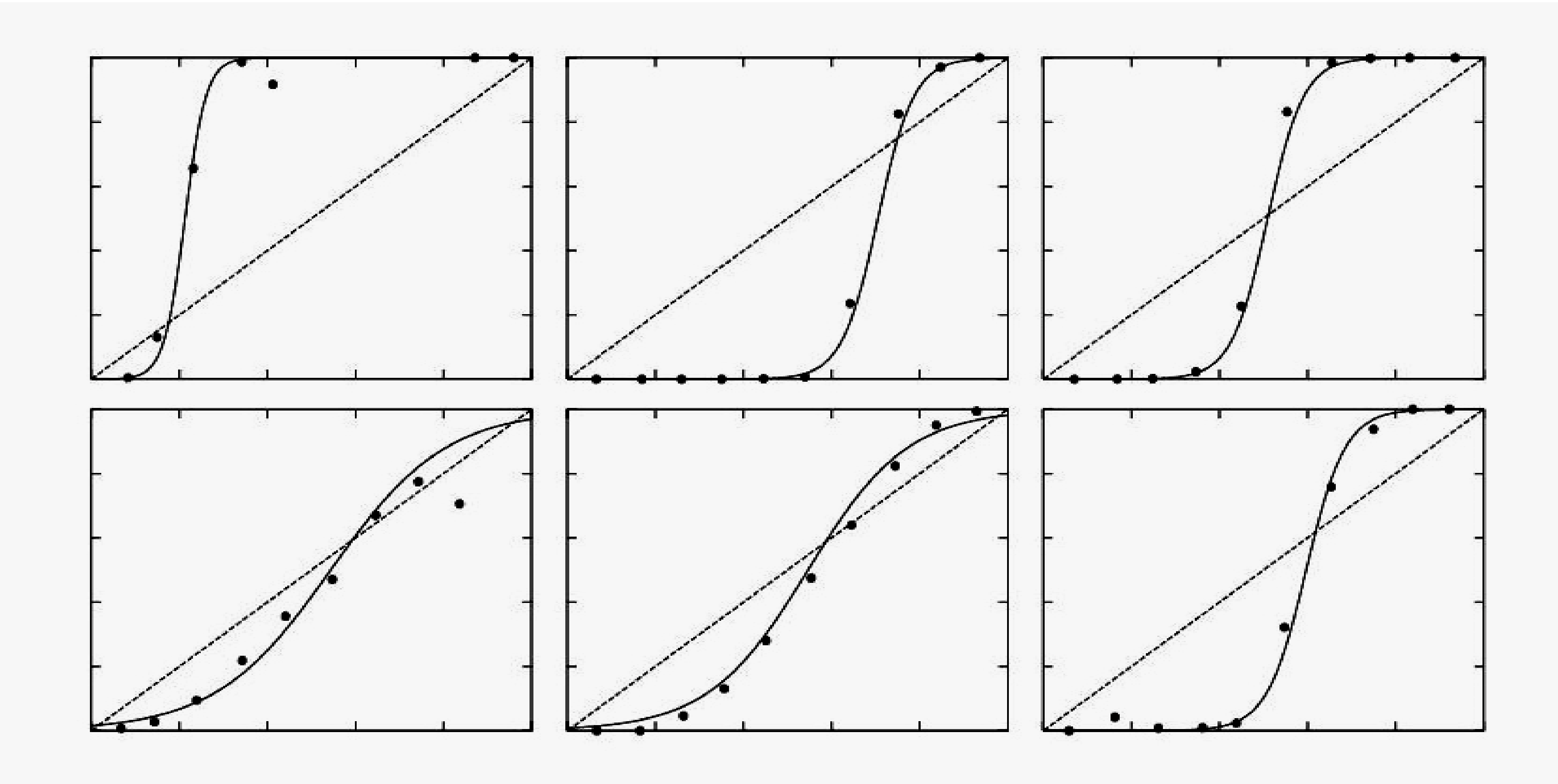
Future AutoML (Engineering) Pipeline



Post-Processing: Calibrated Probabilities

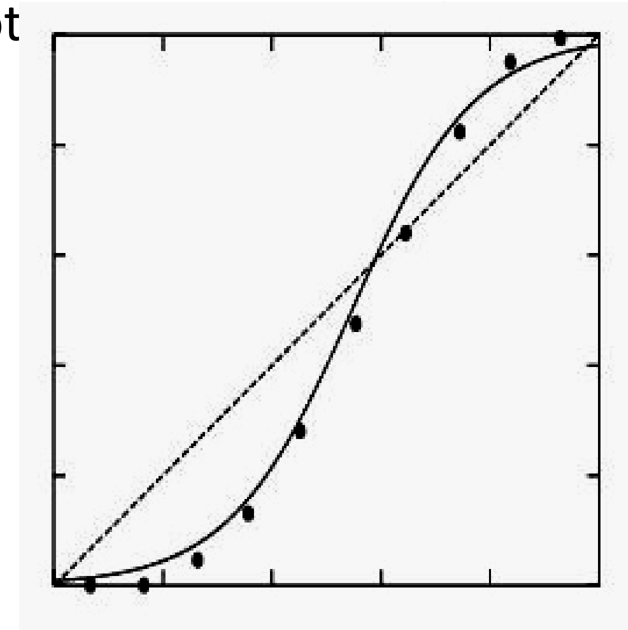
- Probabilities make complex systems easier to engineer
- Uniform language that is easy to explain/understand
- Consistent from rev to rev (eliminates threshold effects)
- Where do probabilities come from?
  - Careful choice of learning algorithm?
    - Most learning algorithms do NOT generate good probabilities
    - Even the best can often be improved
  - Post-Calibration?

SVM Reliability Plots



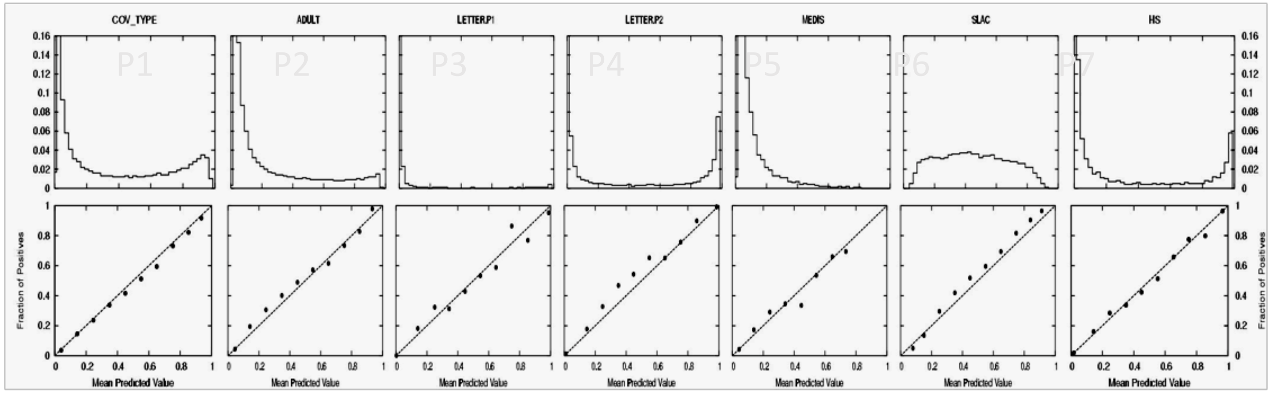
Platt Scaling by Fitting a Sigmoid

- Linear scaling of SVM [-∞,+∞] predictions to [0,1] is bad
- Platt’s Method [Platt 1999]:
  - scale predictions by fitting sigmoid on a *validation set* using 3-fold CV and Bayes-motivated smoot

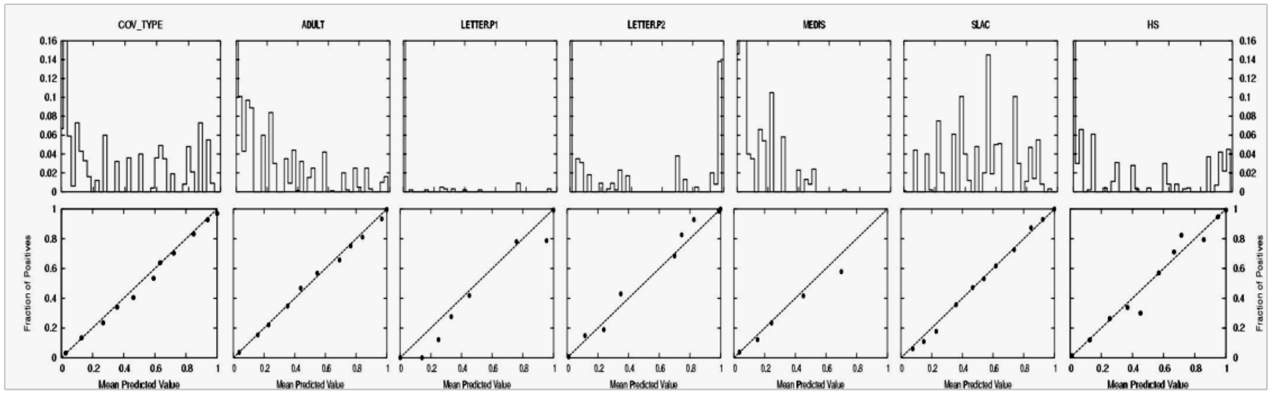


Platt Scaling vs. Isotonic Regression

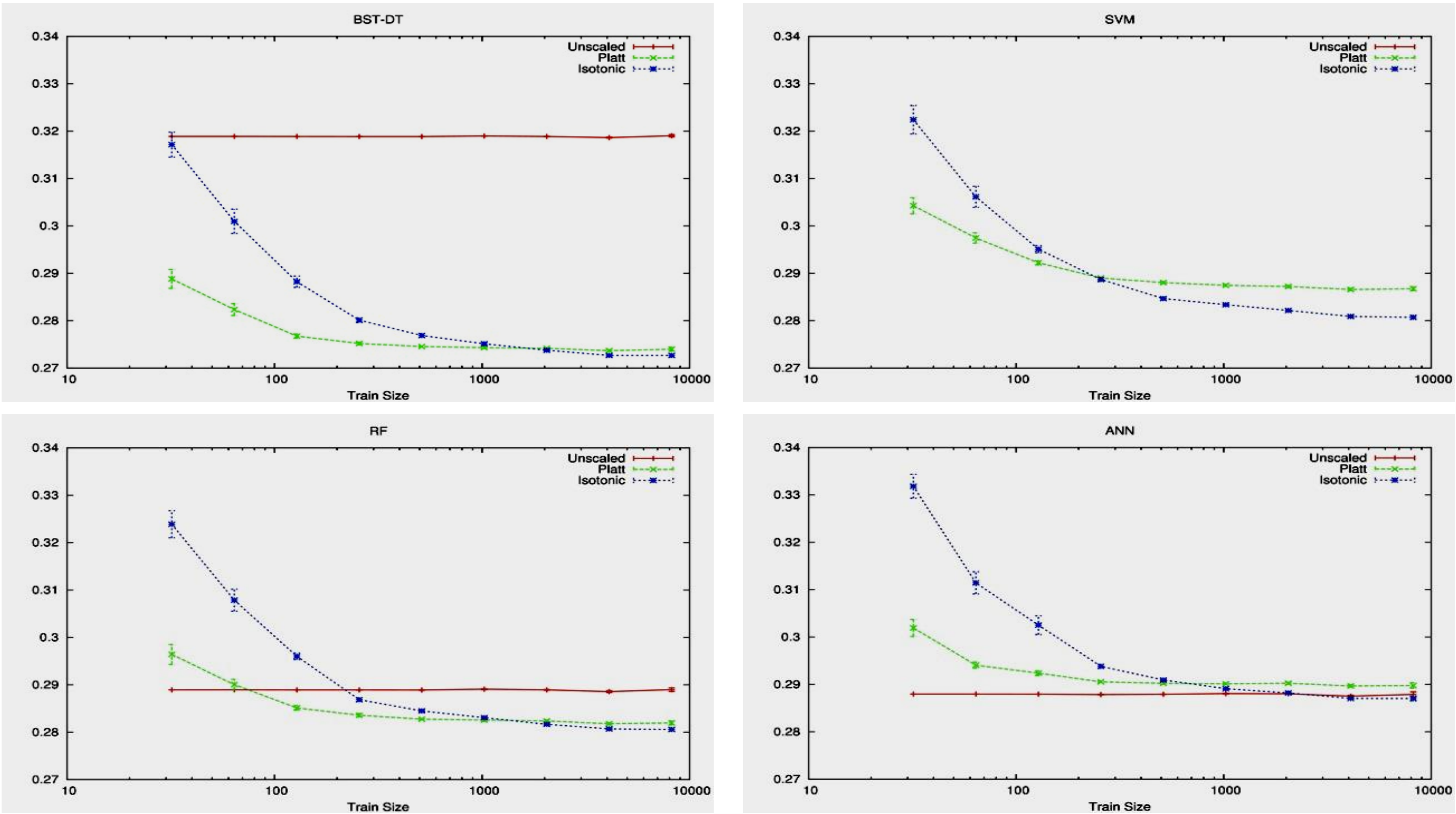
- Platt Scaling:



- Isotonic Regression:



Platt Scaling vs. Isotonic Regression



Auto-Calibrate

- Not as easy to make bulletproof as you might think
  - Depends on sample size
  - Depends on data skew
  - Depends on ROC
  - Probably depends on source model that generated scores in 1st place
  - Try multiple methods and *reliably* pick best…
- Automatically select sample to be used for post-training calibration?
  - Use cross-validation for calibration samples when small data?
- Easy to use tool for automatic calibration would see widespread use
  - Current tools require expertise and careful use
- Data mining challenge problem on calibration
  - Foster new research on new calibration methods

AutoML Open Problems

- **robust attribute typing and coding** --- the spec is never right
- **dDiff** --- because the world never stops changing
- **runtime wrappers** --- a model has to know its limitations
- **feedback cycle detection** --- and we never stop changing the world
- **auto calibration** --- probabilities are good, but not easy to automate
- **auto leakage detection** --- because data is never good enough

Leakage and other "accidents"

- 50% of data mining competitions have leakage!!!
  - win data mining competitions
- KDD2011 best paper award:

"**Leakage in Data Mining: Formulation, Detection, and Avoidance"**

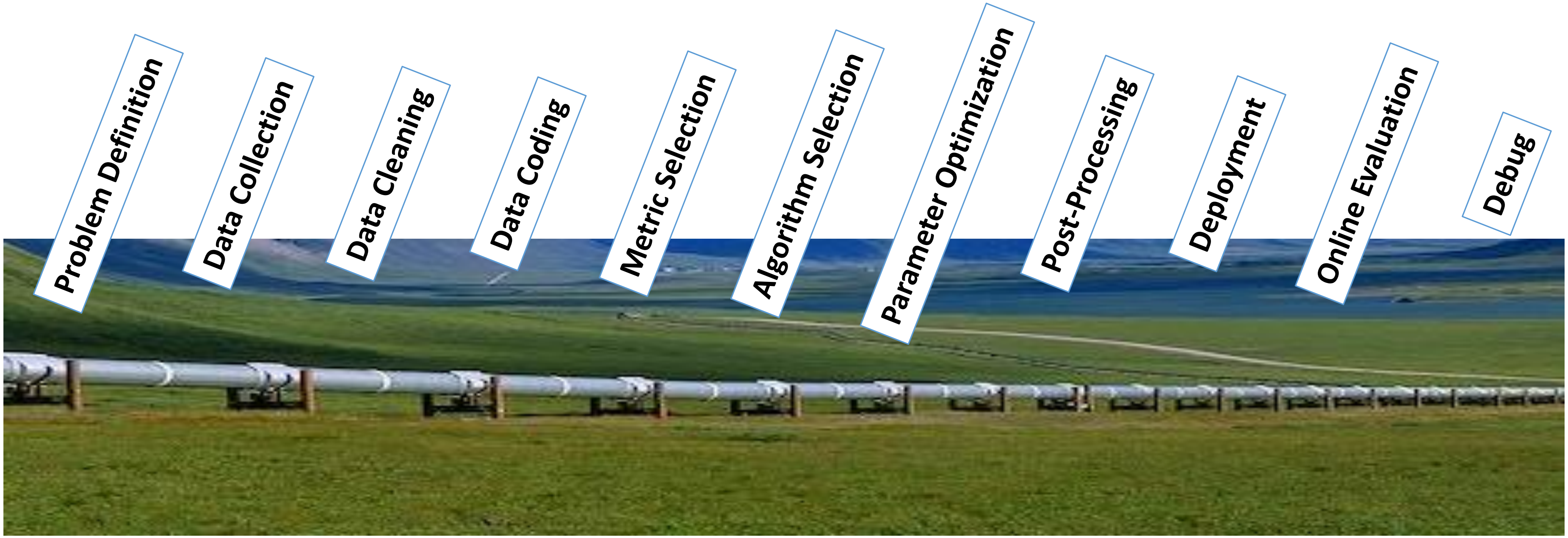
Shachar Kaufman, Saharon Rosset, Claudia Perlich, Ori Stitelman

- Pneumonia leakage 1: missing values
- Pneumonia leakage 2: 4k features (AUC = 0.99)
  - important to have expectations and know when they are violated
- Automatic leakage detection:
  - sequential analysis, missing value analysis, feature analysis, dDiff train to real test, …

AutoML Open Problems

- **robust attribute typing and coding** --- the spec is never right
- **dDiff** --- because the world never stops changing
- **runtime wrappers** --- a model has to know its limitations
- **feedback cycle detection** --- and we never stop changing the world
- **auto calibration** --- probabilities are good, but not easy to automate
- **auto leakage detection** --- because data is never good enough
- **skewed data expert** --- because rare classes are very common
- **auto cross-validate** --- because cross-validation isn’t really as simple as you think
- **auto metric selection** --- which metrics are sensitive to changes
- **auto compression** --- make small model as small and fast as possible
- **auto transfer** --- sometimes transfer helps, sometimes transfer hurts

Want to do New Research that Gets Cited?



Want to do New Research that Gets Cited?

- Pick a part of the ML pipeline that’s still largely manual
- Define what it would mean to make it (more) automatic
- Develop and publish methods
  - Fully-automatic “robot” that solves problem
  - Assistant that helps human recognize and solve problem
  - Tools that alert when problem (probably) exists
- Make data sets publicly available
- Make code available for use as a baseline (and possibly openSource)
- Organize a challenge competition on that part of the pipeline
- Good way to pick a thesis topic!

Summary

- AutoML is a growth research are
  - Community has neglected 85% of the challenges of doing real ML
  - Many independent sub-problems all worthy of attention
  - Every time you stub your toe on real problem => opportunity for new research
- Hyper-Parameter Optimization often critical --- start using it!
- Suggest we all do research and write paper on dDiff this year
  - dDiff worksop in 1-2 years?
  - dDiff challenge/competition in 1-2 years?
  - Make dDiff tools Open Source and available in R and Linux
  - Tools will immediately see widespread use

Thanks!