Strategic Design and Analysis for Hosting Data Competitions

Christine Anderson-Cook, Los Alamos National Laboratory
Lu Lu, University of South Florida

https://sites.google.com/site/andersoncookluftctalk/

October 2018

National Nuclear Security Administration
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA

LA-UR-18-23796

Why Host a Data Competition?

- **Crowdsourcing**
  - New participants from other areas of expertise providing innovative solutions
  - Cheap labor (fixed prize money for winners, but can get many diverse solutions)

- **Competition spurs results**
  - Drive to improve solutions and win leads to accelerated improvements
  - Close to deadline for competition closing often sees frenzied activity

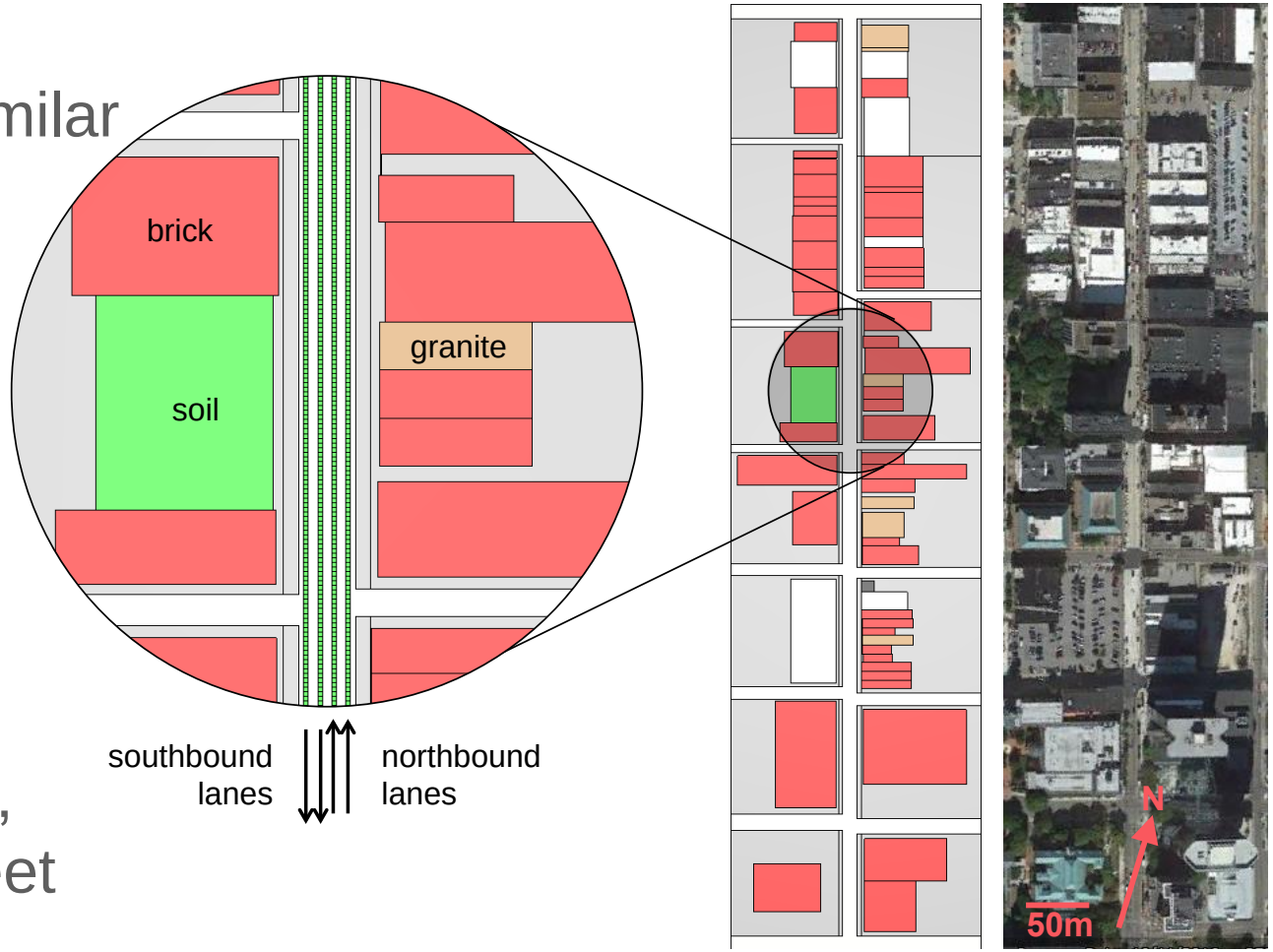
- **Enables objective comparisons of solutions**
  - Often many algorithms claim to have the best solution, but assumptions for the scenario where it is best not clearly specified
  - Competition allows host to intentionally specify the problem space for which solution is sought

Outline

- **Description of Data Competition Recently Hosted by LANL**
  - Urban Radiation Search
  - Basic Structure of Data
- **Opportunities in Design of Data**
  - Where can we be strategic?
- **Opportunities in Analysis of Data**
  - Limitations of the Leaderboard Scoring
  - Post-Competition Analysis
- **Conclusions: Design and Analysis combined**

The Urban Radiological Search Competition

- ORNL initially designed a 0.5 mile street model with characteristics similar to those of Knoxville's Gay Street.
- 56 buildings
  - 48 brick, 7 granite, 1 concrete
  - Hollow shells
- Side streets, sidewalks, 6 parking areas.
- Ability to vary levels of K, U, Th.
- In response to LANL's explorations, ORNL later developed multiple street layouts.



Basics of Data Competitions

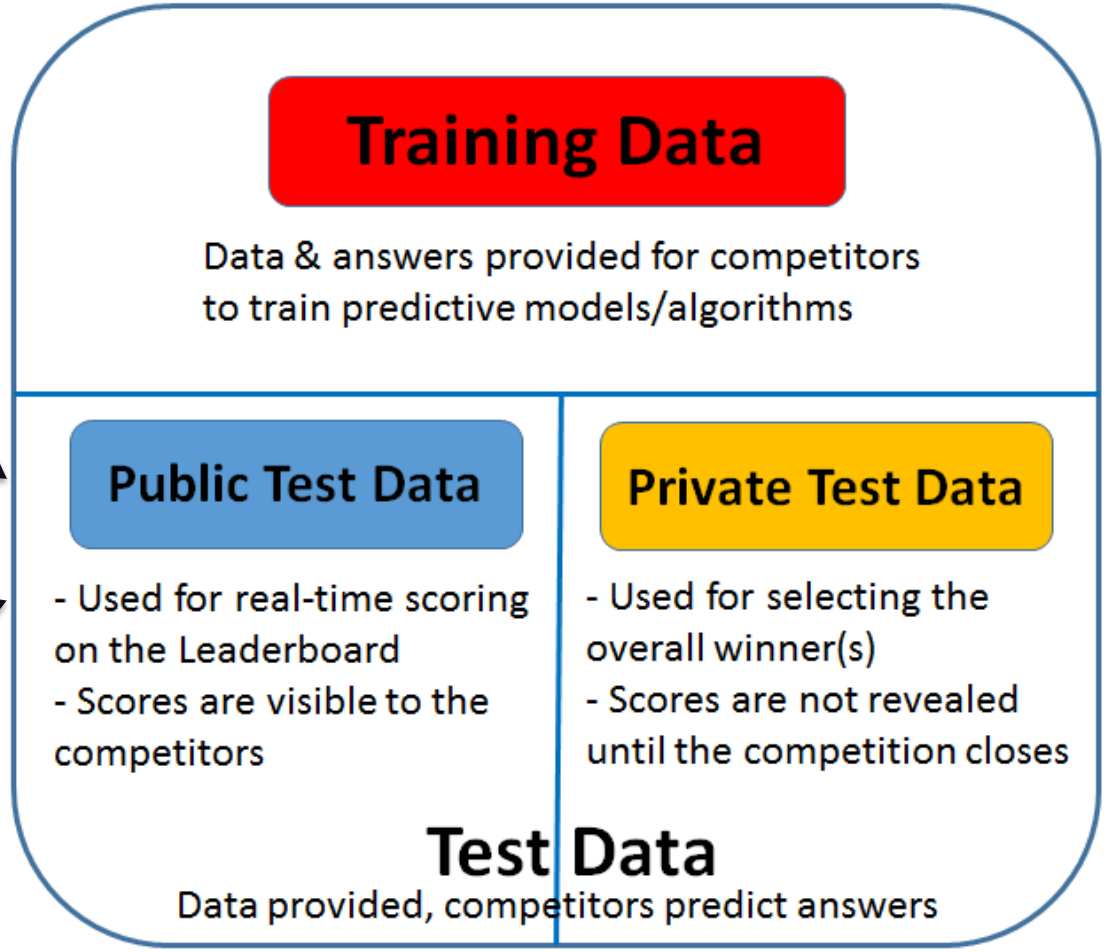
Training data: `<list mode data, source type and location>`

Test data: `<list mode data>`

Public and Private data combined – competitors don't know which runs will be used for each leaderboard

Leaderboards rank teams/competitors based on a score to assess how well they answered the problem of interest

- Same scoring function for public and private
- Multiple aspects of solution must be combined



Structure of the Competition Data Sets

Competition format: *datacompetitions.lbl.gov*

Competitors are provided with

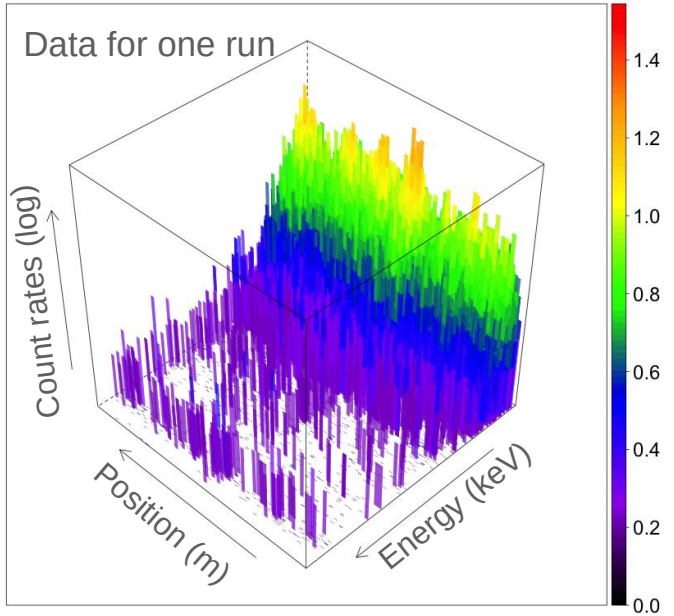
- A training set of **list mode data** for ~10k runs.
- A test set of ~16k runs: 43% **public**, 57% **private**.
- Energy spectra for each source type.

For each run in the test set, competitors must

- **Detect** whether there is an extraneous source.
- **Identify** the type of source.
- **Locate** when the detector is closest to it.

Competitors can submit up to 1000 entries for scoring. Final rankings are based on the best performance of all submissions on the private test set.

Time since last photon (μs)	Photon energy (keV)
1020	88.72
91	179.65
9453	446.41
820	942.51
4295	182.96
1313	262.20
2858	354.80
2687	1295.18

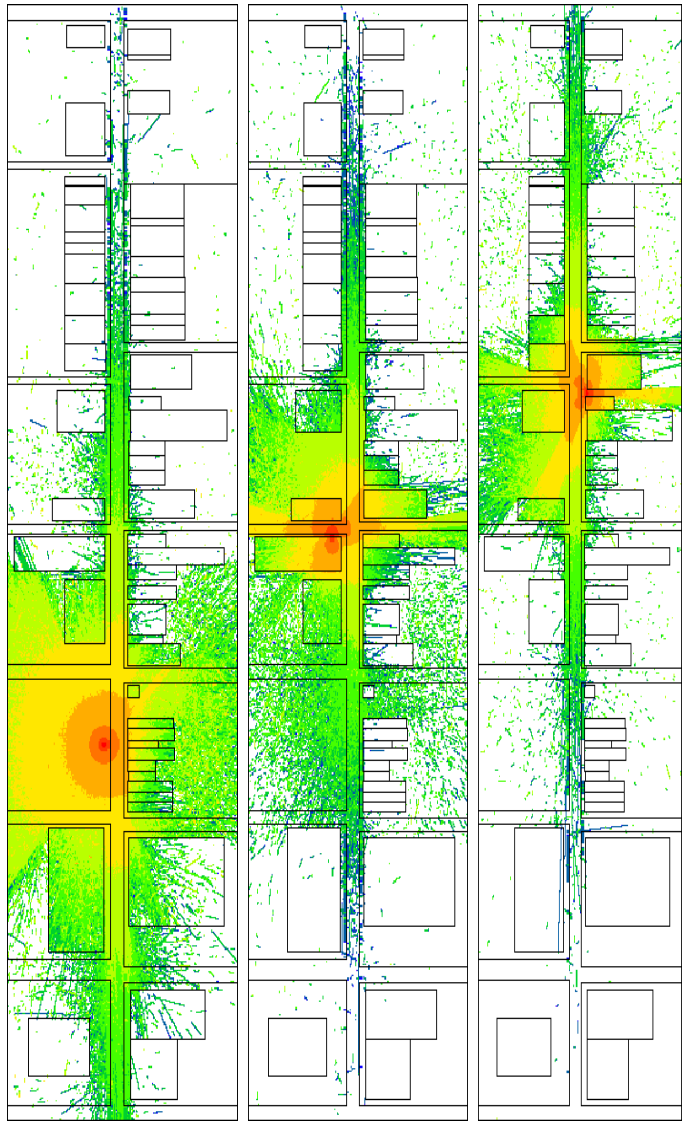


Factors considered for generating the competition data

A. **Background:** 8 background models, each with 82 parameters

B. Source
- **6 types**
- **2 shielding settings** (On/Off)
- **14 source locations**

C. Other
- **speed**
- **source strength** (individually tailored to source type and location)
- **proximity to source**
- length of path, starting location



Outline

- **Description of Data Competition Recently Hosted by LANL**
  - Urban Radiation Search
  - Basic Structure of Data
- **Opportunities in Design of Data**
  - Where can we be strategic?
- **Opportunities in Analysis of Data**
  - Limitations of the Leaderboard Scoring
  - Post-Competition Analysis
- **Conclusions: Design and Analysis combined**

Strategies for Improved Design for Data Competitions

1. **Select data that adequately cover the region of interest**
   - Match region explored to target study goals
2. **Encourage competitors from diverse technical backgrounds**
   - Provide sufficient background for those new to problem
   - Match current practices for established disciplines
3. **Emphasize data of maximum interest**
   - Avoid “too easy” and “too hard” regions to enhance ability to distinguish between competitor solutions
4. **Discourage overfitting by competitor algorithms**
   - Include regions of interpolation and extrapolation in test sets
5. **Balance standard design of experiment principles while avoiding unintentional clues for competitors**
   - Adapt replication, balance and randomization for competition data
6. **Create leaderboard scoring that ranks performance to match competition goals**
   - Find suitable balance between multiple criteria

Intentional choices in the design stage enable better analysis later

Select data that adequately cover the region of interest

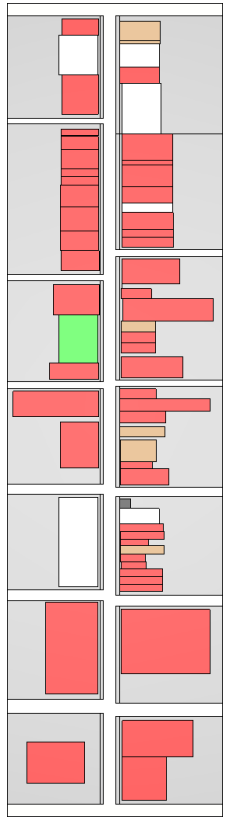
1. **Precisely define the goals of the competition**
   - What do we want to assess? Criteria for success?
   - Over what set of conditions do we need the solution?

2. **Assess the capability of data to match competition goals**
   - Constraints of data generator?
   - Sufficient heterogeneity possible?
   - What simplifying assumptions are made? Do they compromise the ability to address the goals?

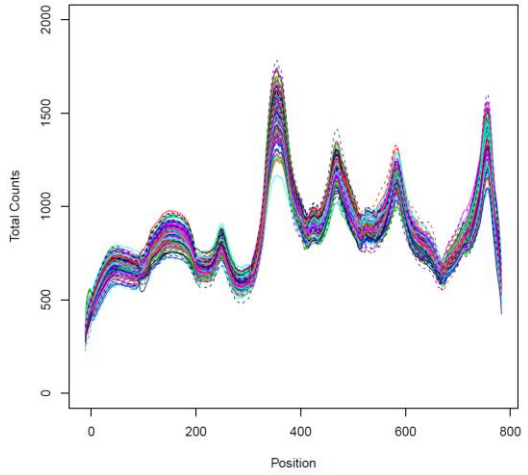
3. **Will we be able to adequately estimate within data size constraints?**
   - Does complexity of input space match the total data sets?

4. **Details of the inputs and their ranges?**
   - What is of interest? Will the problem be solvable here?

1. **Goal for Urban search competition:**
   - Detect, identify and locate multiple radioactive sources on a typical US street.



2. **Capability of simulator for background**
   - Initially, not very heterogeneous:



   - Solution:
     - Multiple street layouts
     - Circular street that allowed flexible start and finish
     - Adapted K, U, Th contributions to match regional differences

Encourage competitors from diverse technical backgrounds

1. **Remove obstacles to join for non subject matter experts.**
   – If something is a one-time discovery to help with solving the problem, provide it.

2. **Include some of fundamental assumptions currently used by subject matter experts (or think hard about why to not include them)**
   – Allow current best algorithm to provide a starting point for those in area

1. **For non-SMEs:**
   – Provided details about the source spectra.

2. **For current radiation detection experts:**
   – Generally, start with a period of calibration for algorithm
   – We announced that “no source would be located within the first 30 seconds of any run”

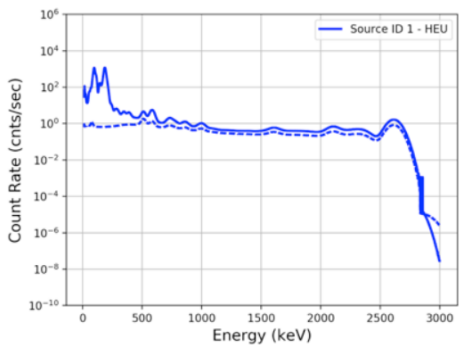
Source Types

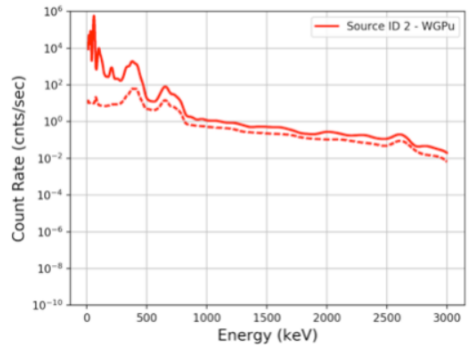
If a source is present in a run, it will be one of six types:

SourceID	Source Type
1	HEU: Highly enriched uranium
2	WGPu: Weapons grade plutonium
3	^{131}I: Iodine, a medical isotope
4	^{60}Co: Cobalt, an industrial isotope
5	^{99m}Tc: Technetium, a medical isotope
6	A combination of ^{99m}Tc and HEU

Enter your prediction of the SourceID in your answers file for each run. If no source is present in a run, the SourceID field for that run in your answers file should be entered as 0. For all runs, there will be no source within the first 30 seconds of measurements. That is, you're guaranteed at least 30 seconds of background-only measurements in each run.

Energy spectra for each source type are shown below for a significant quantity [1] of source types 1 and 2 and 1 microcurie (μCi) for source types 3-5. In each figure, the solid curve shows the unshielded spectrum and the dashed curve, the spectrum with 1 cm of lead shielding. The plots show sources 1 meter away from the center of the detector in a vacuum.





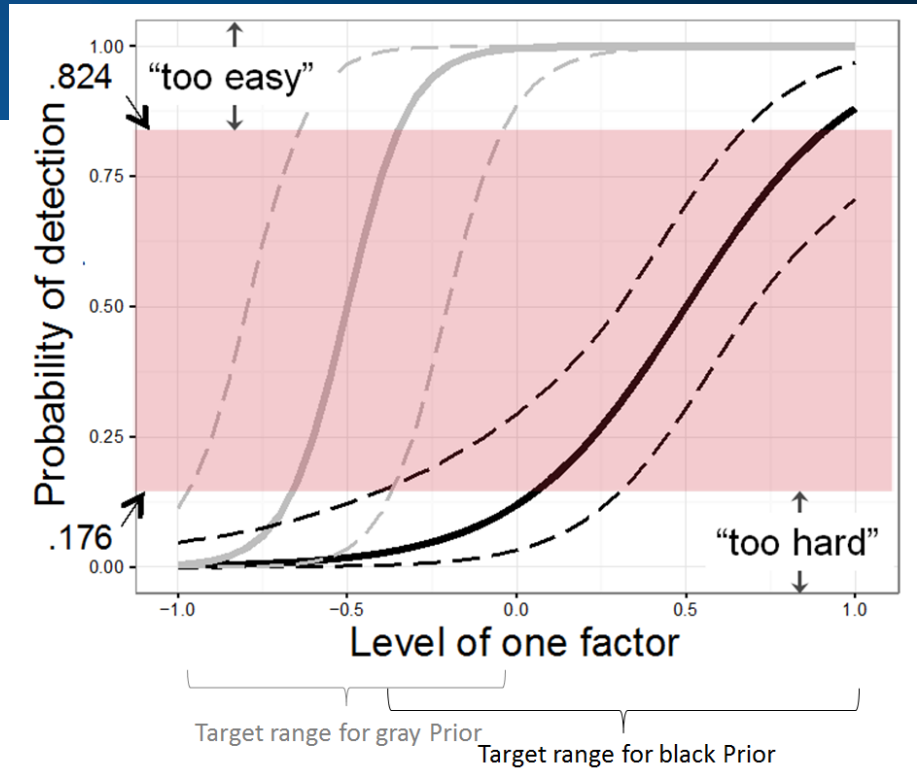
Emphasize data of maximum interest

1. For univariate case with known relationship:

- Best solution: ½ points each at P(success) = 0.176 & 0.824
- What is different for our scenario:
  - Multiple inputs
  - Multiple objectives – detect & identify (using logistic model)
  - Unknown a priori relationship between inputs and responses
  - Multiple competitors (with different performance anticipated)

2. Our solution:

- Bound our region with “current best algorithm” and “dream capability”
- Want robust estimation throughout span
- Generate a **“superset”** with 10x runs to select from
- Eliminate regions that are
  - “too easy” – all the good algorithms will get them right
  - “too hard” – no one will be able to get them right
- Emphasize region with good potential to distinguish between algorithms



- Used best algorithm available to predict answers for entire superset
- Consult radiation detection experts for upper bound
- Downweighted regions where current algorithm has P(detect) > 0.5

Formal methodology: Quinlan et al (2017), Quinlan & Anderson-Cook (2018)

Discourage overfitting by competitor algorithms (cont'd)

Multiple entries on a single test set tempt competitors to tune to the idiosyncrasies of the data. Overfitted solutions are unlikely to predict well for new scenarios

Private test set:

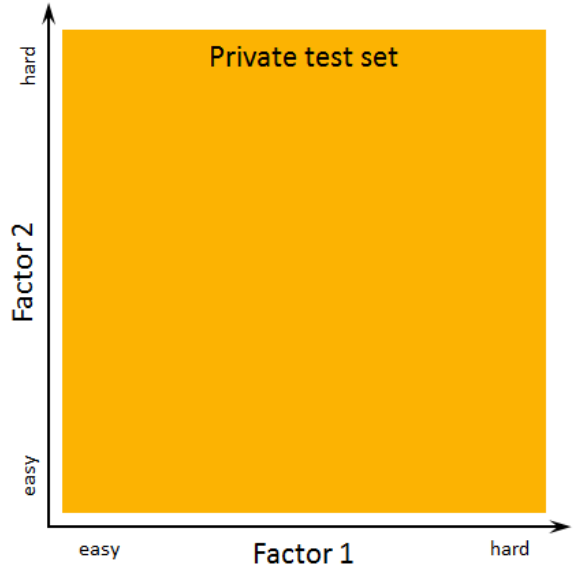
- Final choice of best algorithm based on this
- Set should span the entire space of interest

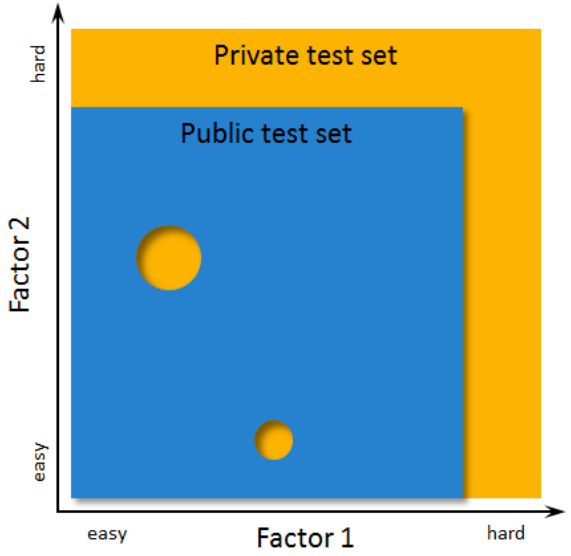
Public test set:

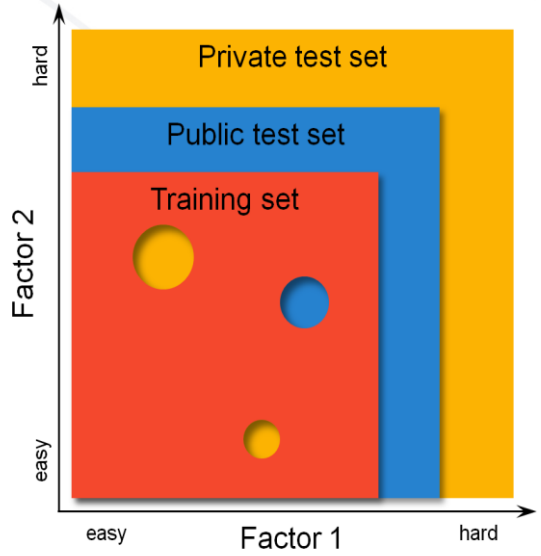
- Strategically exclude some difficult and internal regions
- Allows assessment on new situations

Training set:

- Labels provided to competitors
- Strategically further shrink region







Balance Design of Experiment principles & avoid unintentional clues

1. **Traditional design of experiments does not have an adversary to consider.**
2. **Seasoned data competition participants often gain advantage by exploiting unintended artifacts in the data**

Our goal: force competitors to solve the intended problem, while still collecting ideal data

Design of Experiment concepts to leverage:

Replication –luxury of a large data set, better information for binary response

Balance – improved estimation of model parameter estimates and prediction throughout space

Randomization – avoids systematic patterns

- **Randomization**
  - Sorted training data
  - Randomized order of public and private runs in overall test set
  - Randomized source / no source in test set

- **Replication**
  - Generated 3 replicates of each scenario in superset
  - Use starting location and length of detector path to make runs look different from each other
  - Vary number of replicates included in final data (1, 2 or 3)
- **Balance**
  - For the number of run for different sources:
    - Decided separately on fraction of “no source” needed estimate false positive rate throughout region
    - Varied number of runs for each of 6 sources in [MinSize, 1.2*MinSize]

Create leaderboard scoring to match competition goals

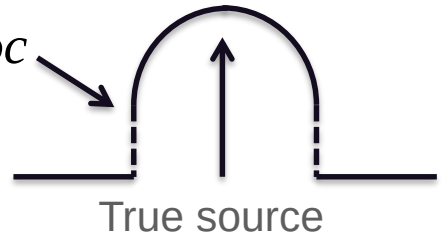
- **Recall goals of competition are to detect, identify and locate multiple sources**
- **Leaderboard scoring requires a single static formula to evaluate and compare algorithm performance – need to get this right for declaring the right winner!**

Our strategies:

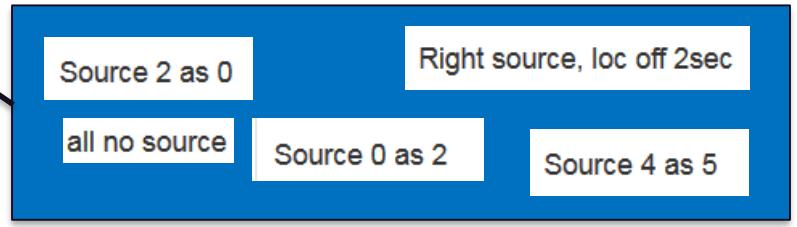
1. Not all runs contribute equally to the total ← "no source" runs contributed ½ as much as "with source" runs
2. Use a weighted average of different contributions for different goals

$$Score_{source} = w_{\det} Det + w_{iden} Iden + w_{loc} Loc$$

$$Score_{NoS} = 0 \text{ or } 1$$



3. Constructing different types of algorithm mistakes, experiment with different weights to get a desirable final ranking that matches subject matters priorities



4. Check, check, check for correct implementation! There are no second chances without embarrassment!

Outline

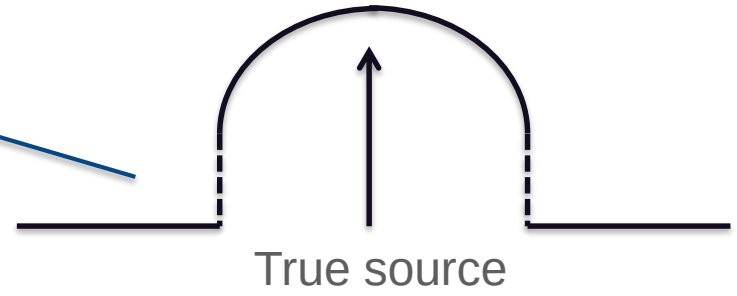
- **Description of Data Competition Recently Hosted by LANL**
  - Urban Radiation Search
- **Opportunities in Design of Data**
  - Basic Structure of Data
  - Where can we be strategic?
- **Opportunities in Analysis of Data**
  - Limitations of the Leaderboard Scoring
  - Post-Competition Analysis
- **Conclusions: Design and Analysis combined**

Why Post-Competition Analysis

- **We have three separate objectives for the competition: detect, identify and locate**
  - We are interested in the performance of each of these aspects separately – the best algorithms for each individual aspect might be different
  - We were forced to pick a fixed weight for the leaderboard scoring – what if we got it wrong? Or we consider a different scenario

$$Score_{source} = w_{\det} Det + w_{iden} Iden + w_{loc} Loc$$

$$Score_{NoS} = 0 \text{ or } 1$$



- **The goals of the competition include:**
  - Ranking the competitors' algorithms (hopefully, primarily handled by leaderboard)
  - Understanding the strengths and weaknesses of each algorithm
  - Understanding the relative impacts of the inputs on the difficulty of the problem
  - Understanding the regions of the input space where
    - All top algorithms can get the right answer
    - All do poorly
    - Different algorithms perform differently

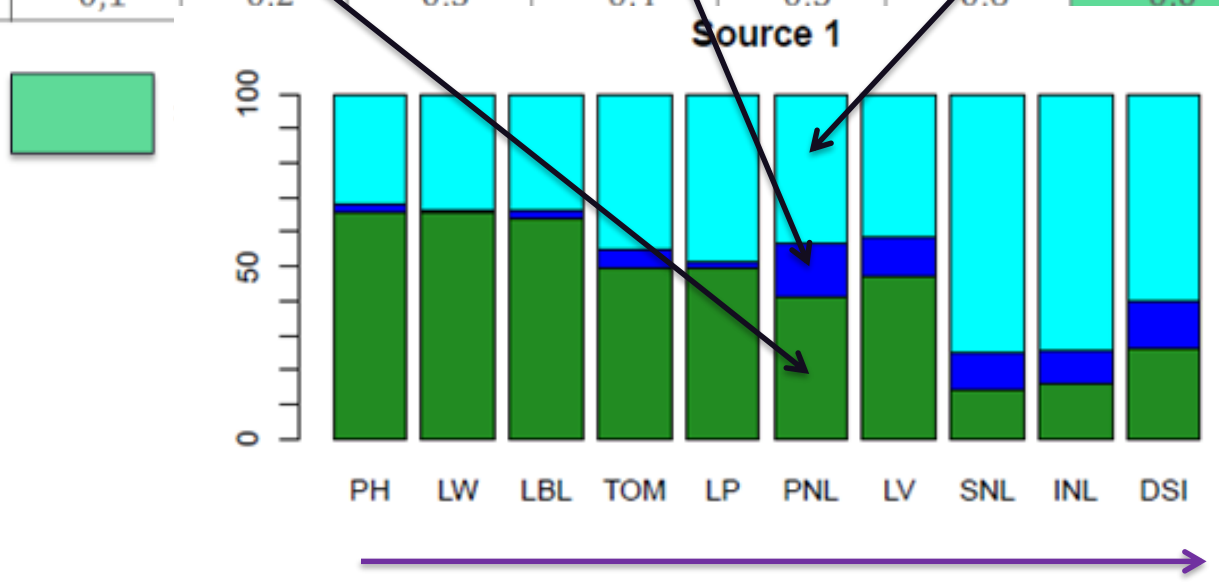
Elements for More Detailed Study

Public	Label							
Type	1	2	3	4	5	6	NoS	Total
Source 1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$n_{1,5}$	$n_{1,6}$	$n_{1,0}$	n_1
…								…
Source 6	$n_{6,1}$			…		$n_{6,6}$	$n_{6,0}$	n_6
No Source	$n_{0,1}$	$n_{0.2}$	$n_{0.3}$	$n_{0.4}$	$n_{0.5}$	$n_{0.6}$	$n_{0,0}$	n_0

To construct the traditional confusion matrix: $\frac{n_{i,j}}{n_i}$

For detection:

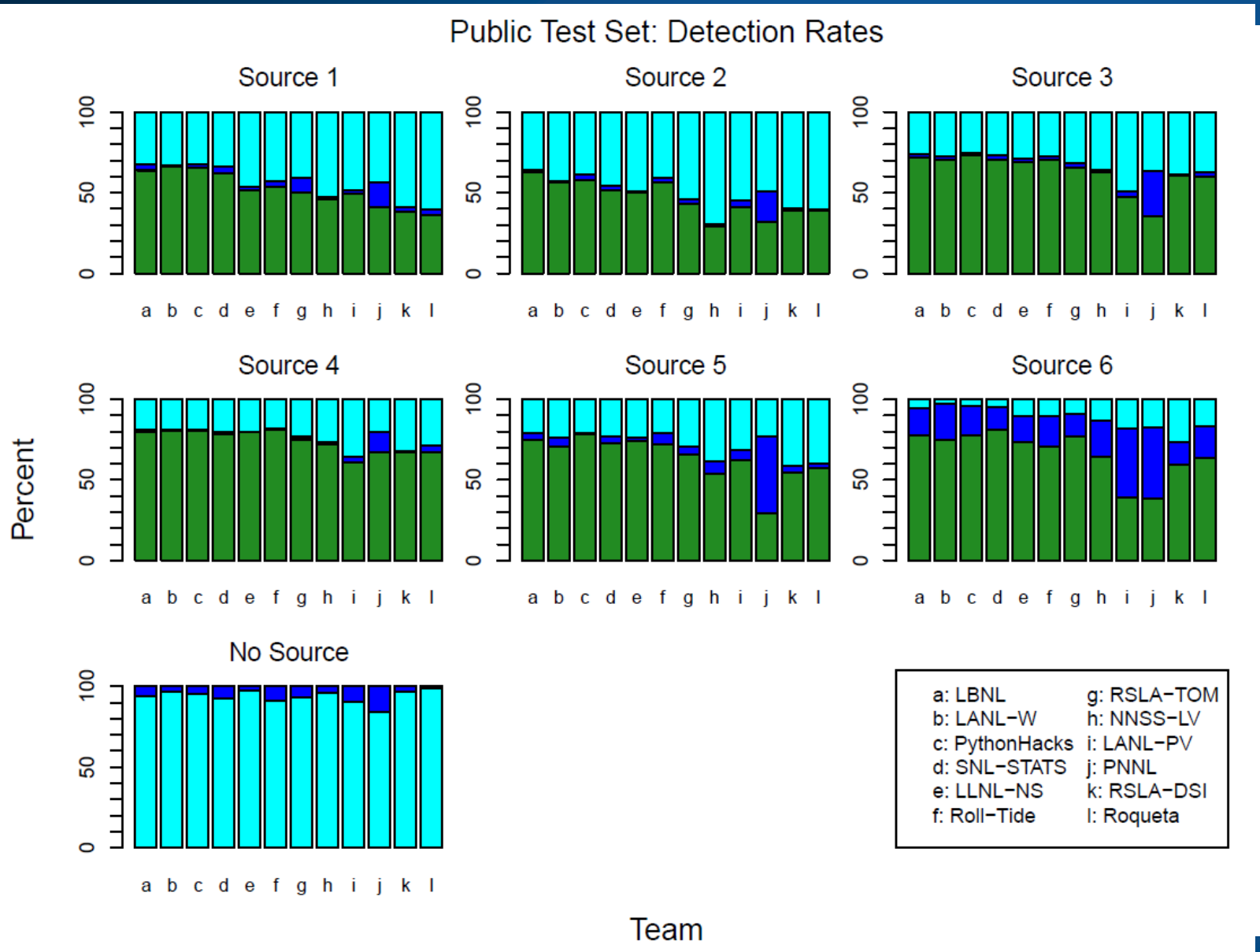
$n_{\text{all S},all\ S}$	$n_{\text{all S},0}$
$n_{0,all\ S}$	$n_{0,0}$



ction

Sorted based on current best leaderboard score

Comparison of Performance for Each Source By Team



1. HEU: Highly enriched uranium
2. WGPu: Weapons grade plutonium
3. ^{131}I: Iodine, a medical isotope
4. ^{60}Co: Cobalt, an industrial isotope
5. ^{99m}Tc: Technetium, a medical isotope
6. A combination of HEU and ^{99m}Tc

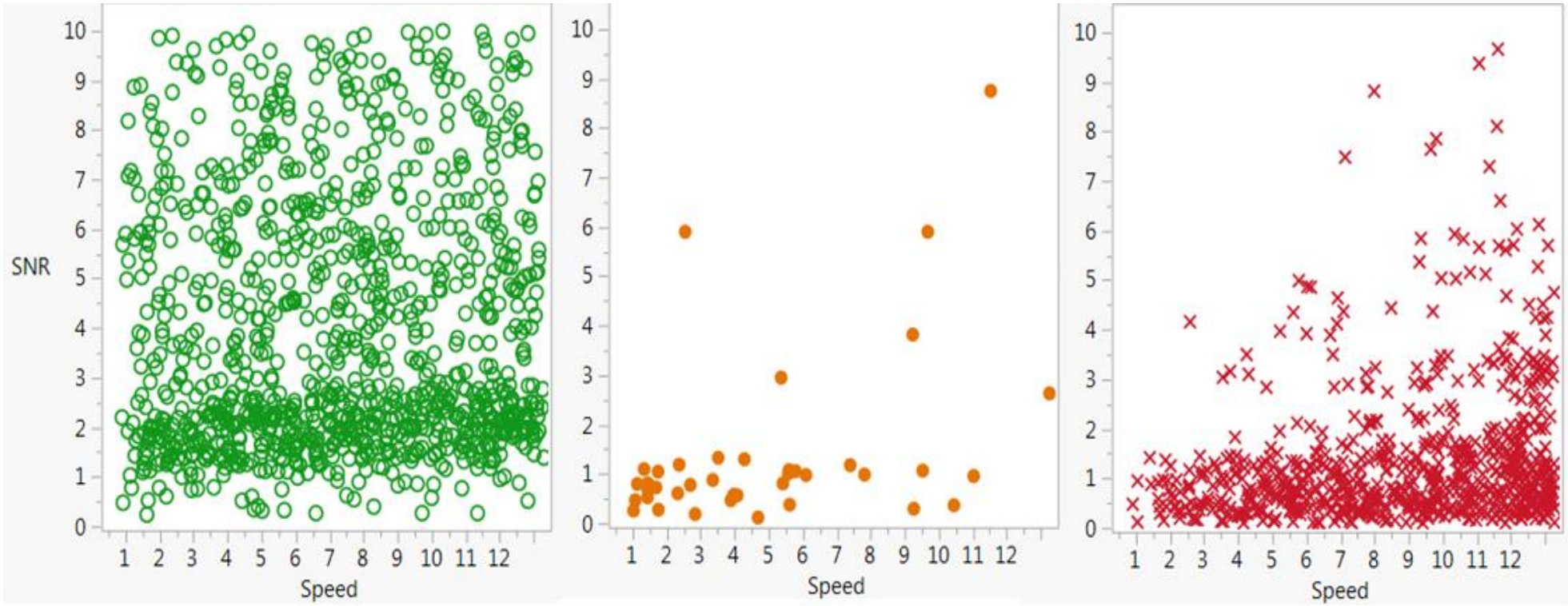
Patterns:

- Harder/easier sources
- False positive rates
- Non-monotonic performance

Run-by-Run Performance

Source 2

For each team, we can see where they
- Correctly identified (and detected)
- Correctly detected (incorrect identify)
- Missed source



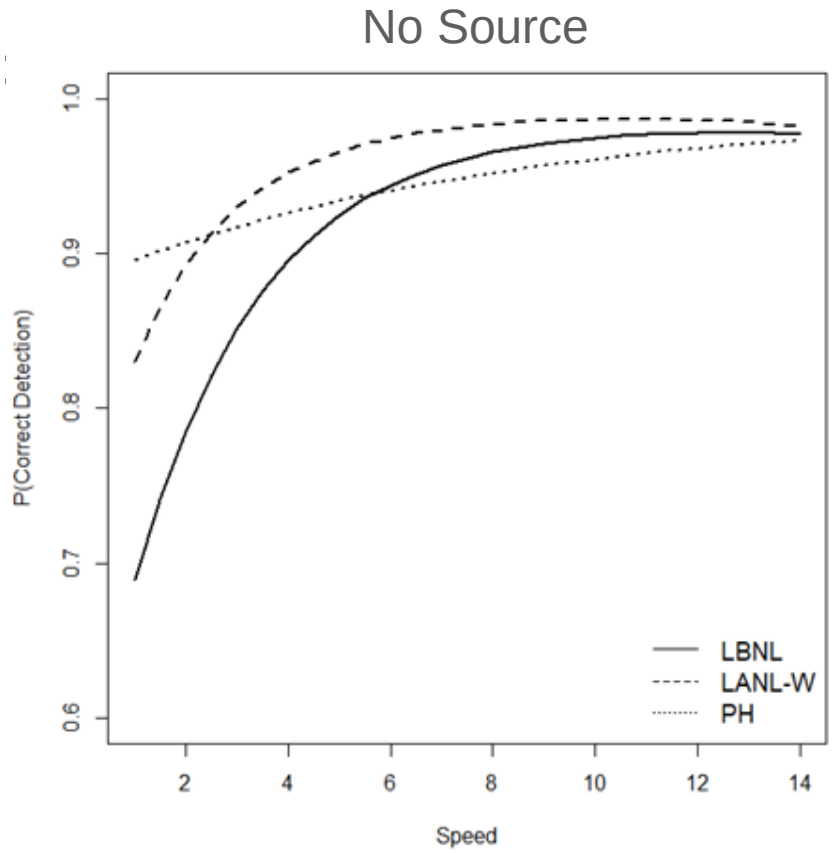
Model-based Analysis

To understand the impact of different inputs on algorithm performance as well as prediction throughout the region, we model each of detect, identify and locate separately

Logistic regression model for correctly identifying no source:

$$P(\text{detection}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$$

$$\boldsymbol{x}'\boldsymbol{\beta} = \underbrace{\beta_0 + \beta_1 X_{speed} + \beta_2 X_{speed}^2}_{\text{Significant}} + \underbrace{\beta_{back} I_{back}}_{\text{Not significant}}$$



Understanding Contributions From Different Input Factors

GLM for detection and identification for runs with source:

$$P(\text{Identify}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$$

This analysis will be performed for the best submission(s) from each team

$$x'b = b_0 + b_{SNR}X_{SNR} + b_{Shield}I_{Shield} + b_{Background}I_{Back} + b_{Lane}X_{Lane} + b_{Speed}X_{Speed}$$

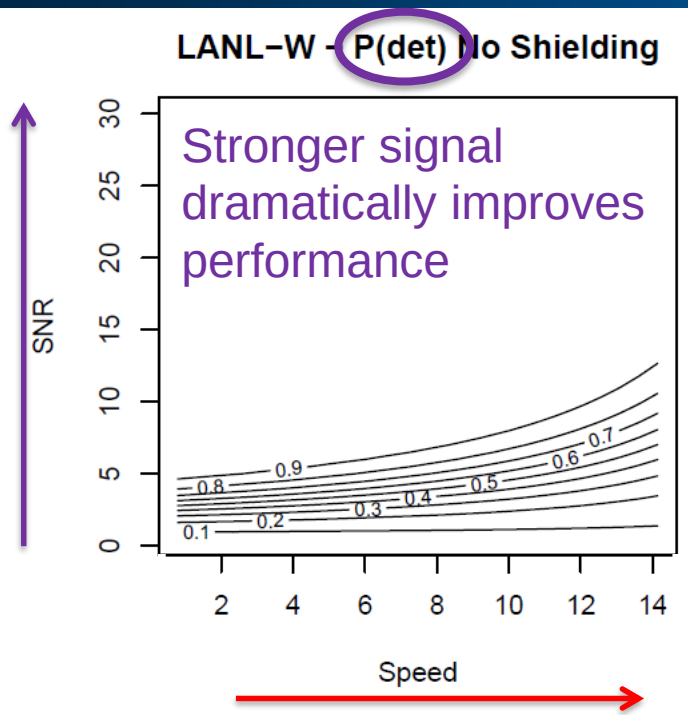
Significant

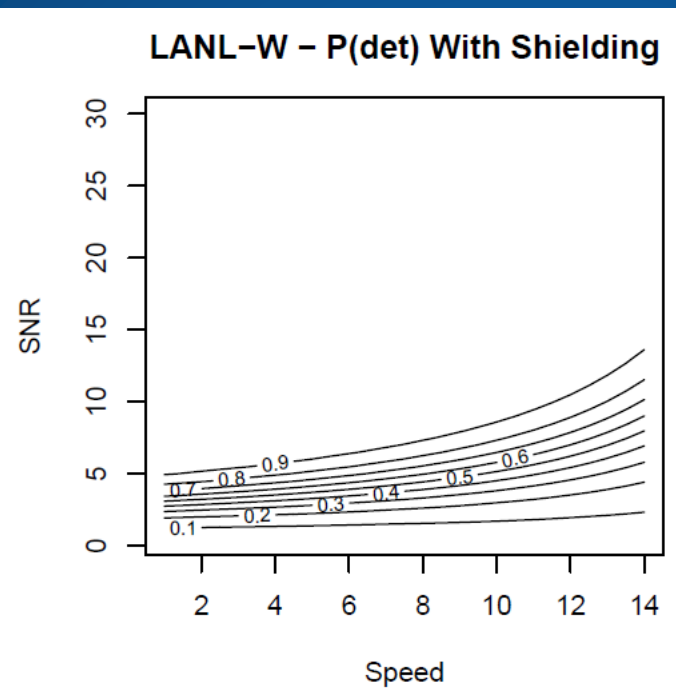
Linear regression model for location:

$$\text{Absolute miss} = \beta_0 + \beta_{SNR}X_{SNR} + \beta_{Shield}I_{Shield} + \beta_{Background}I_{Back} + \beta_{Lane}X_{Lane} + \beta_{Speed}X_{Speed} + \varepsilon$$

Significant

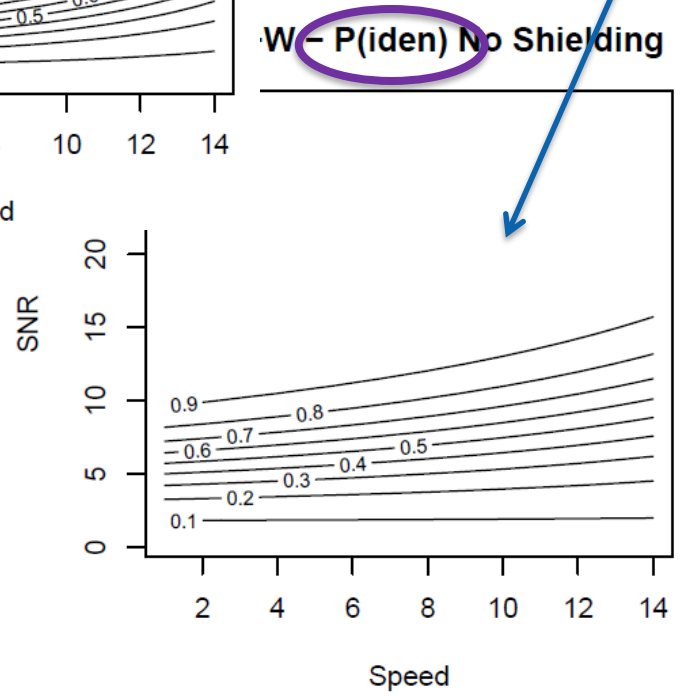
Performance Throughout the Input Region

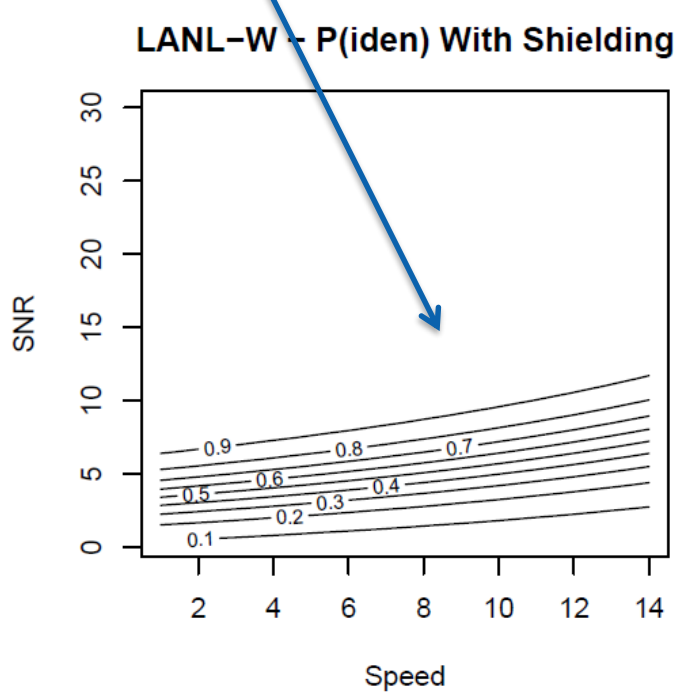




Increasing speed is somewhat harder

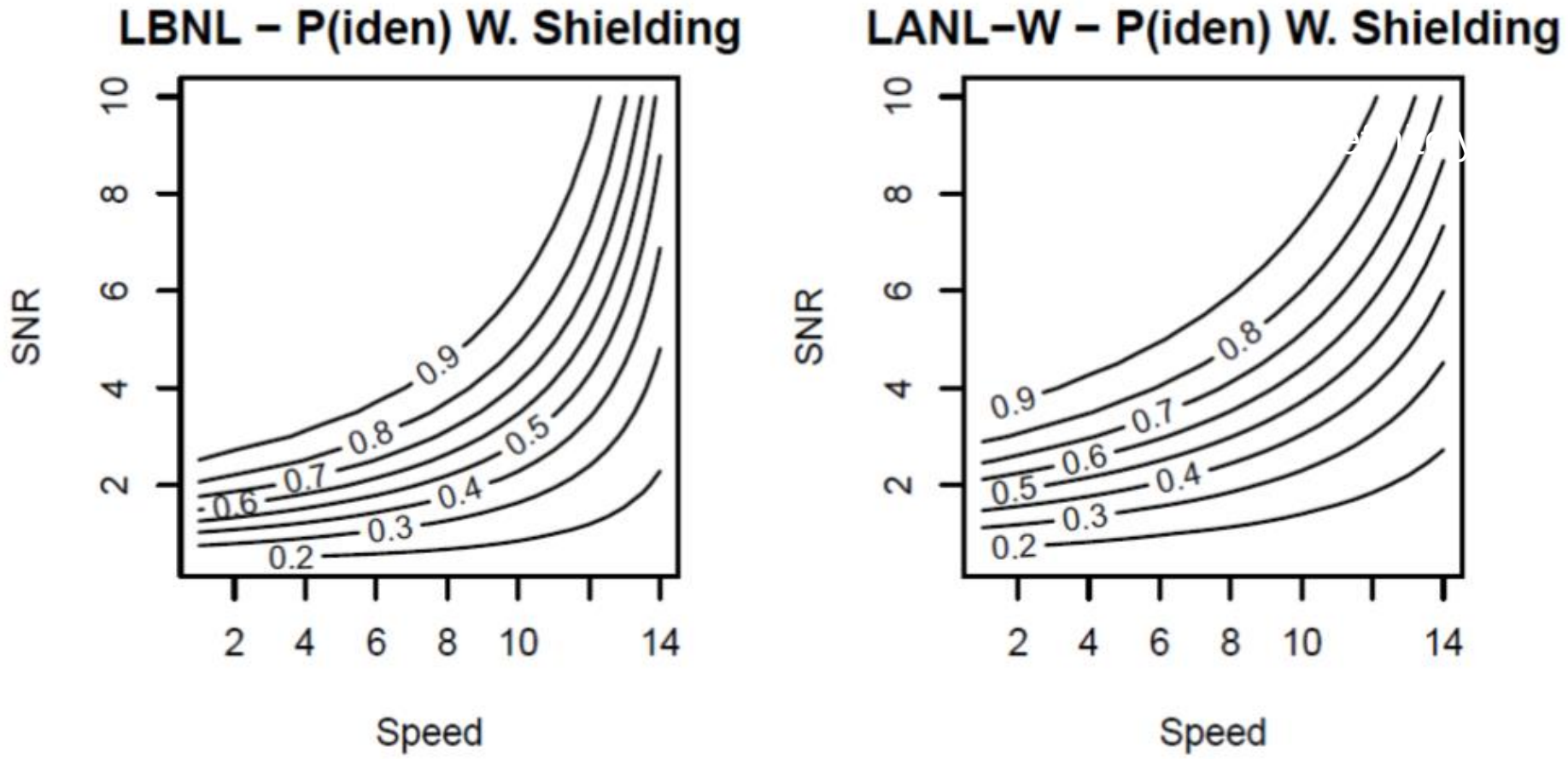
For this source, with shielding is a bit easier





Comparison of Performance Between Teams

– Do submissions from different teams have different performance?
– Are the difference similar across the input space?



Comparison of Performance Between Teams

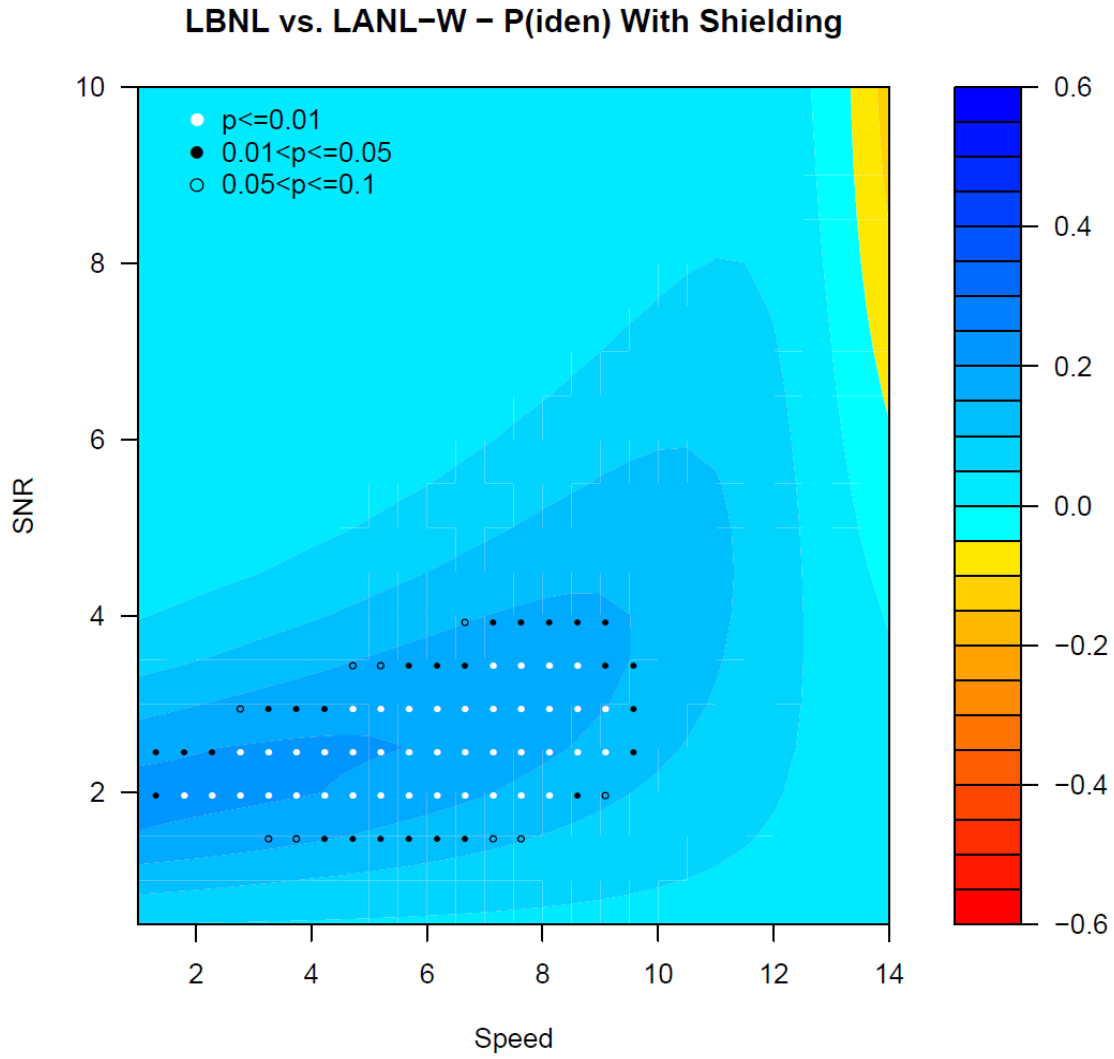
- $\widehat{Iden_i}(x)$ = estimated correct identification rate at location x for team i
- Difference in correct identification rate: $\widehat{Iden_1}(x) - \widehat{Iden_2}(x)$

- We used Bonferroni adjustment for multiple comparisons across the grid of locations:

  $$p_B = n \times p_Z,$$

  p_Z is the p-value for a regular z-test

- To improve the power of the test, users can use less conservative approach for multiple comparison adjustment or use coarser grid of locations.

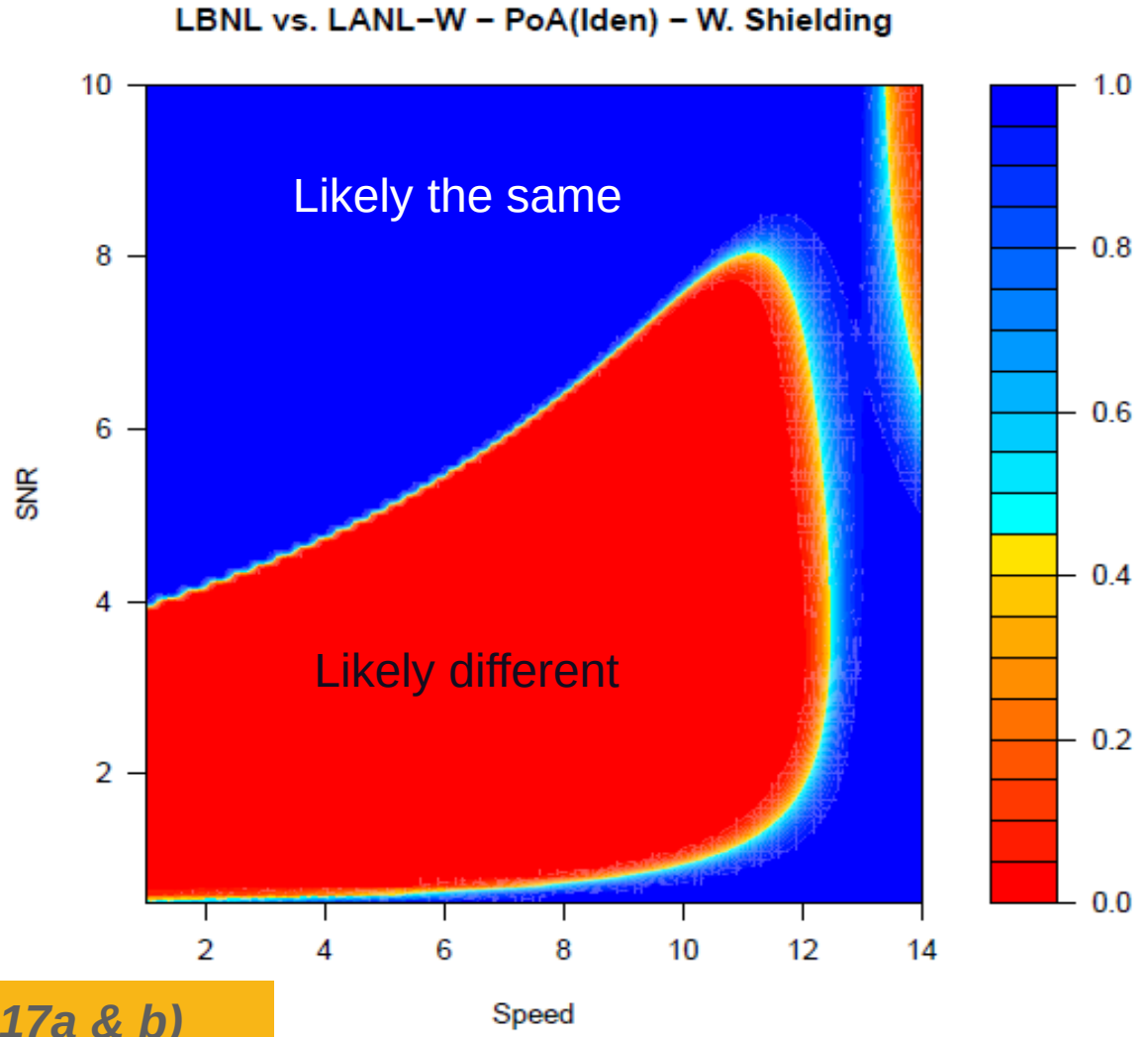


Probability of Agreement (PoA) for Comparing Teams

- Similar to equivalence testing with user specified threshold, δ
- Difference $> \delta$ is considered of practical importance

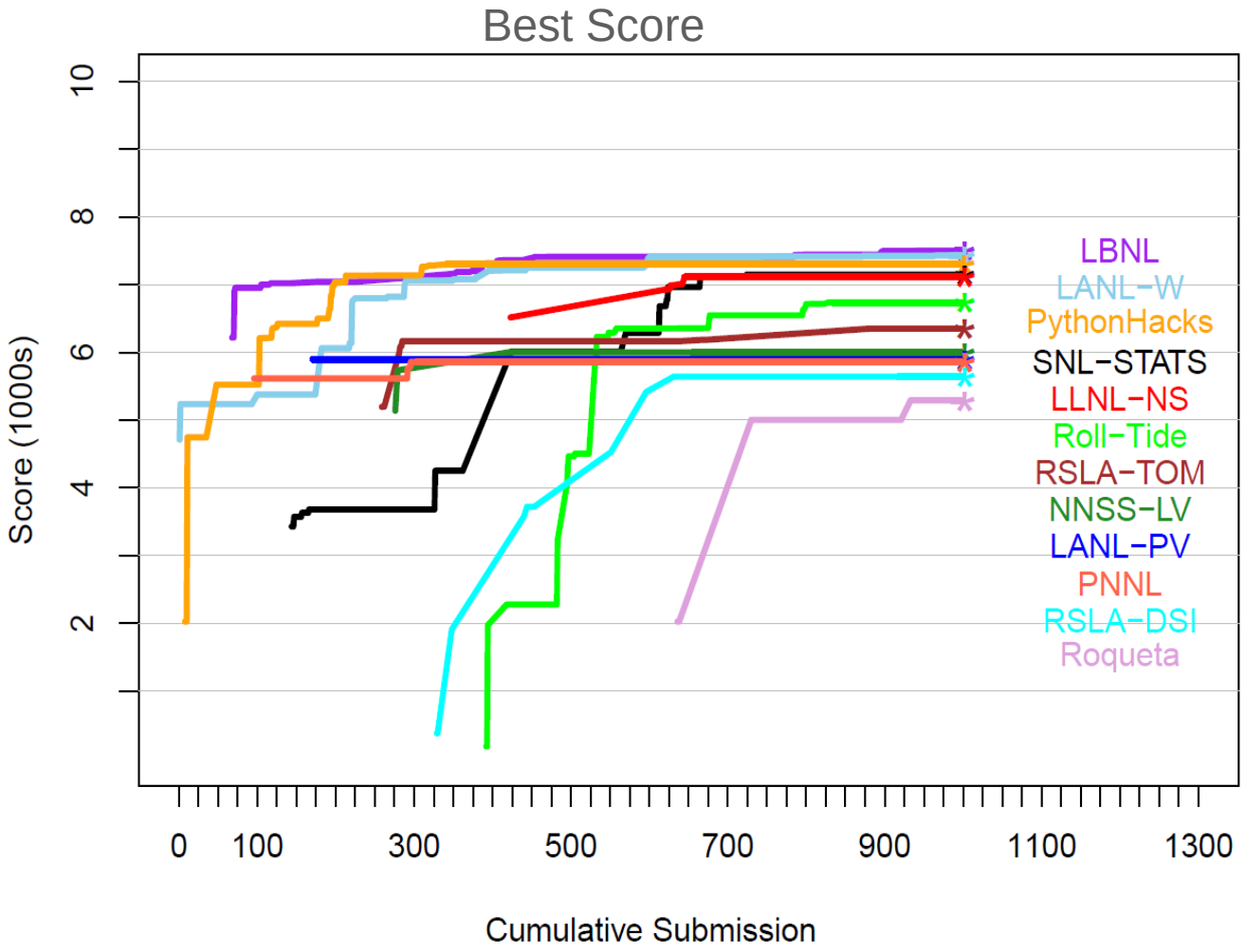
$$\delta = 0.05$$

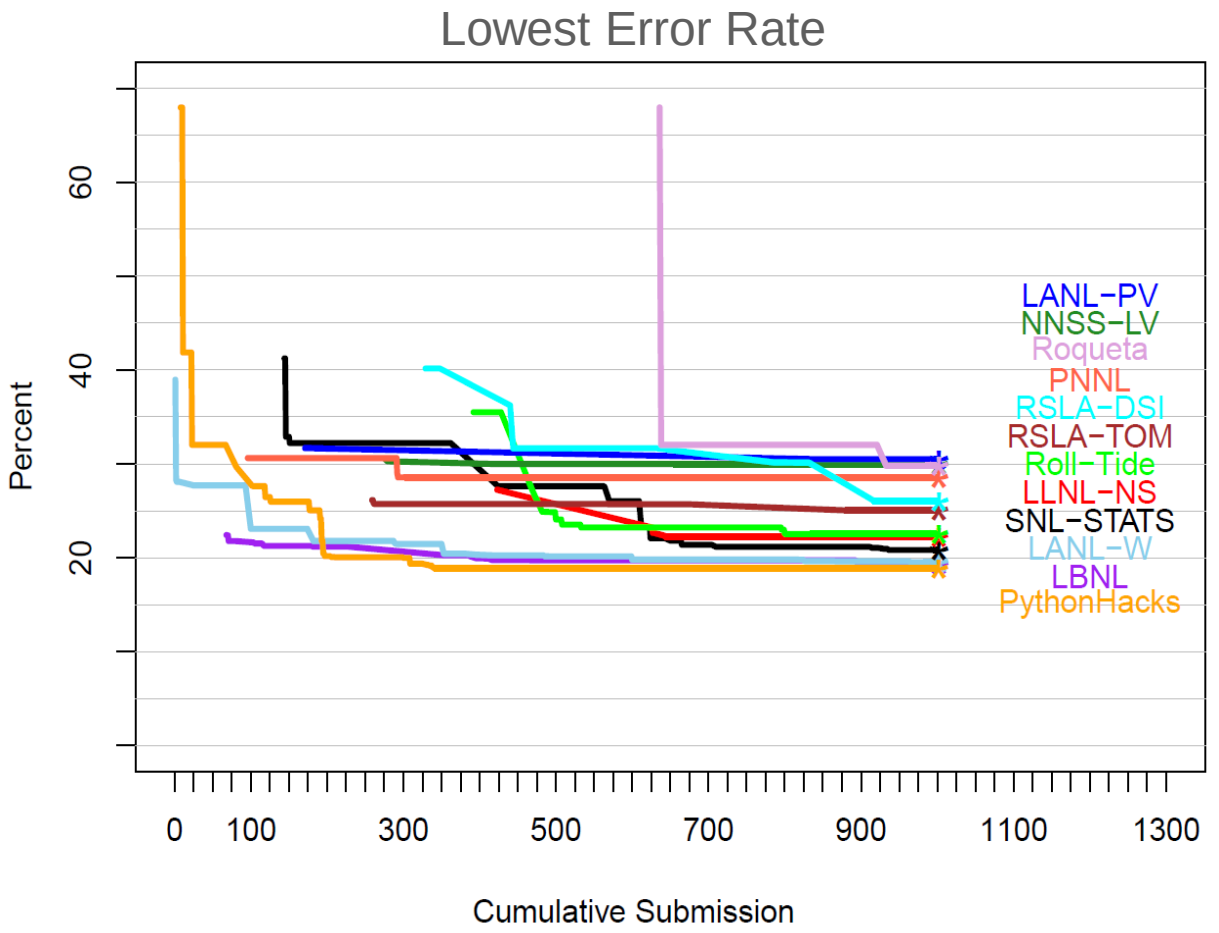
- $PoA(x) = \Pr\left(\left|\widehat{Iden}_1(x) - \widehat{Iden}_2(x)\right| \leq \delta\right)$
- Formal quantification of uncertainty
- Estimate based on asymptotic theory



Formal methodology: Stevens & Anderson-Cook (2017a & b)

Evolution of Competitor Scores During Competition





Conclusions: Design and Analysis – working in harmony

- **By intentionally designing the data to train and test the competitors, we**
  - Gain confidence in our ability to address the goals of the competition
  - Maximize the amount of usable information obtained
  - Hopefully, avoid artifacts that allow the competitors to solve a different problem
  - Set up the desired analysis to match the design for maximum learning outcome

- **By carefully constructing the leaderboard scoring, we**
  - Rank the teams to match our competition objectives

- **By using EDA and model-based post-competition analyses, we**
  - Gain additional insights about performance for individual components of competition (detect, identify, locate)
  - Take pressure off the assigned weighting of components on the leaderboard
  - Can understand and compare algorithm strengths and weaknesses
  - Can understand patterns in performance throughout the inputs space
  - Can calibrate the scoring system for future competitions

https://sites.google.com/site/andersoncookluftctalk/

References

- K.R. Quinlan, C.M. Anderson-Cook “Bayesian Design of Experiments for Logistic Regression to Evaluate Multiple Forensic Algorithms” ***Applied Stochastic Models in Business and Industry*** (in press), 2018.
- K.R. Quinlan, K., C.M. Anderson-Cook, K.L. Myers, “The Weighted Priors Approach for Combining Expert Opinion in Logistic Regression Experiments” ***Quality Engineering*** 29 (2017), 484-498.
- Stevens, N., Anderson-Cook, C.M. (2017) “Comparing the Reliability of Related Populations with the Probability of Agreement” ***Technometrics*** 59(3) 371-380.
- Stevens, N., Anderson-Cook, C.M. (2017) “Quantifying Similarity in Reliability Surfaces Using the Probability of Agreement” ***Quality Engineering*** 29(3) 395-408.