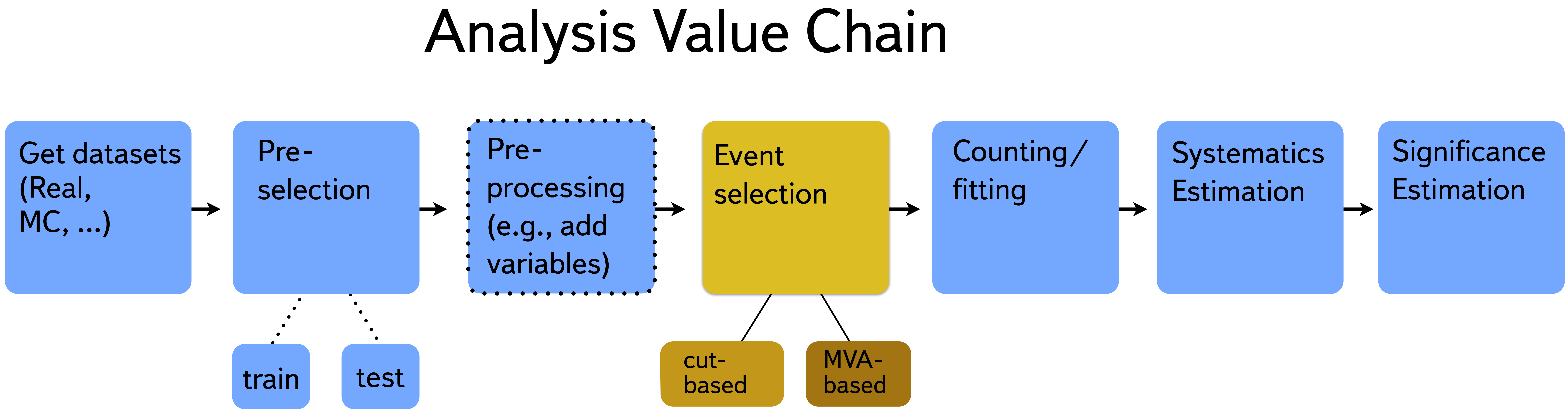
Research on Event Search

Andrey Ustyuzhanin
Yandex, NRC «Kurchatov Institute» Moscow,
Imperial College, London

Quest for analysis sensitivity



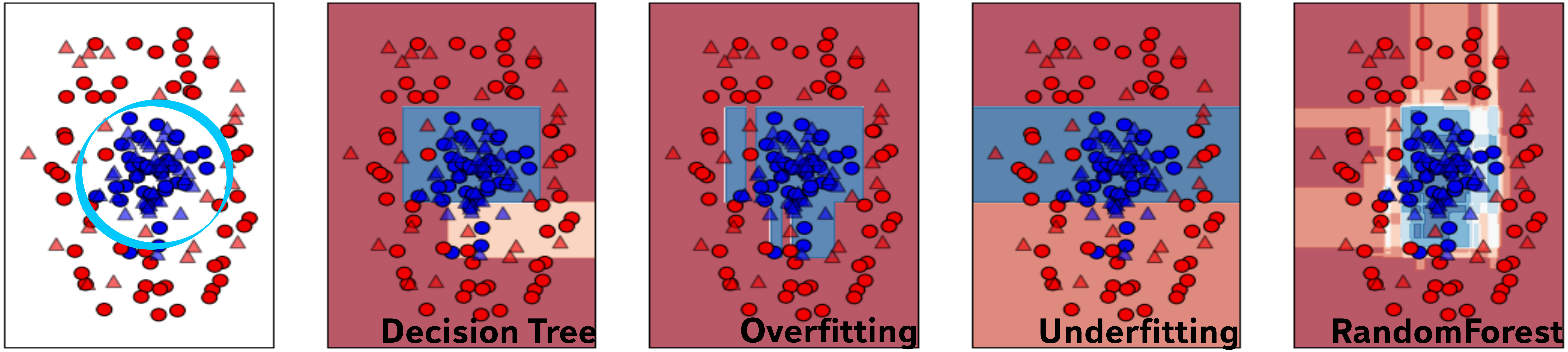
Sources of better sensitivity

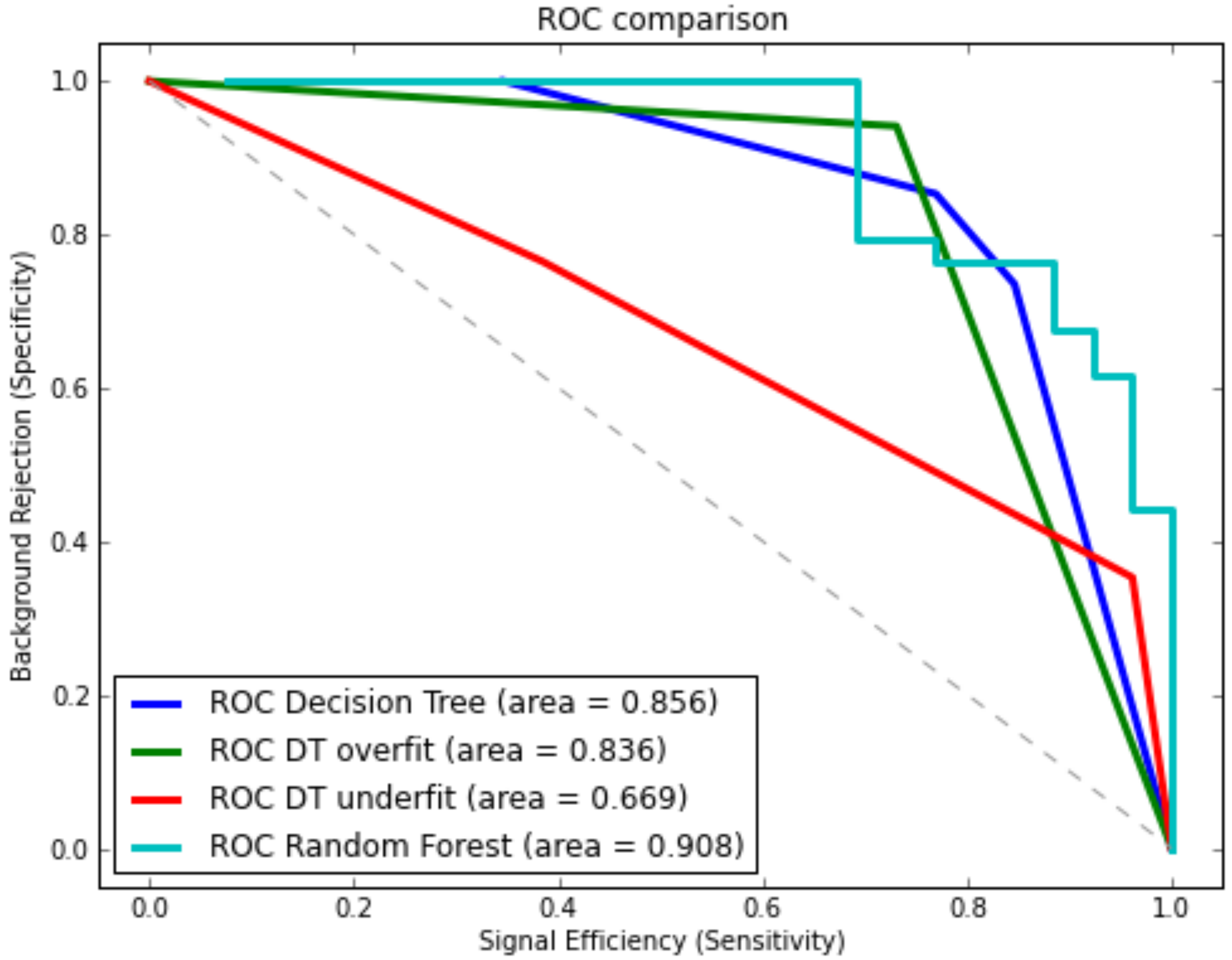
1. more powerful algorithms (e.g. BDT, Deep Neural Networks)
2. improved features (e.g. «isolation» variables or particle identification)
3. complex training scenarios (e.g. n-folding, ensembling, blending, cascading)

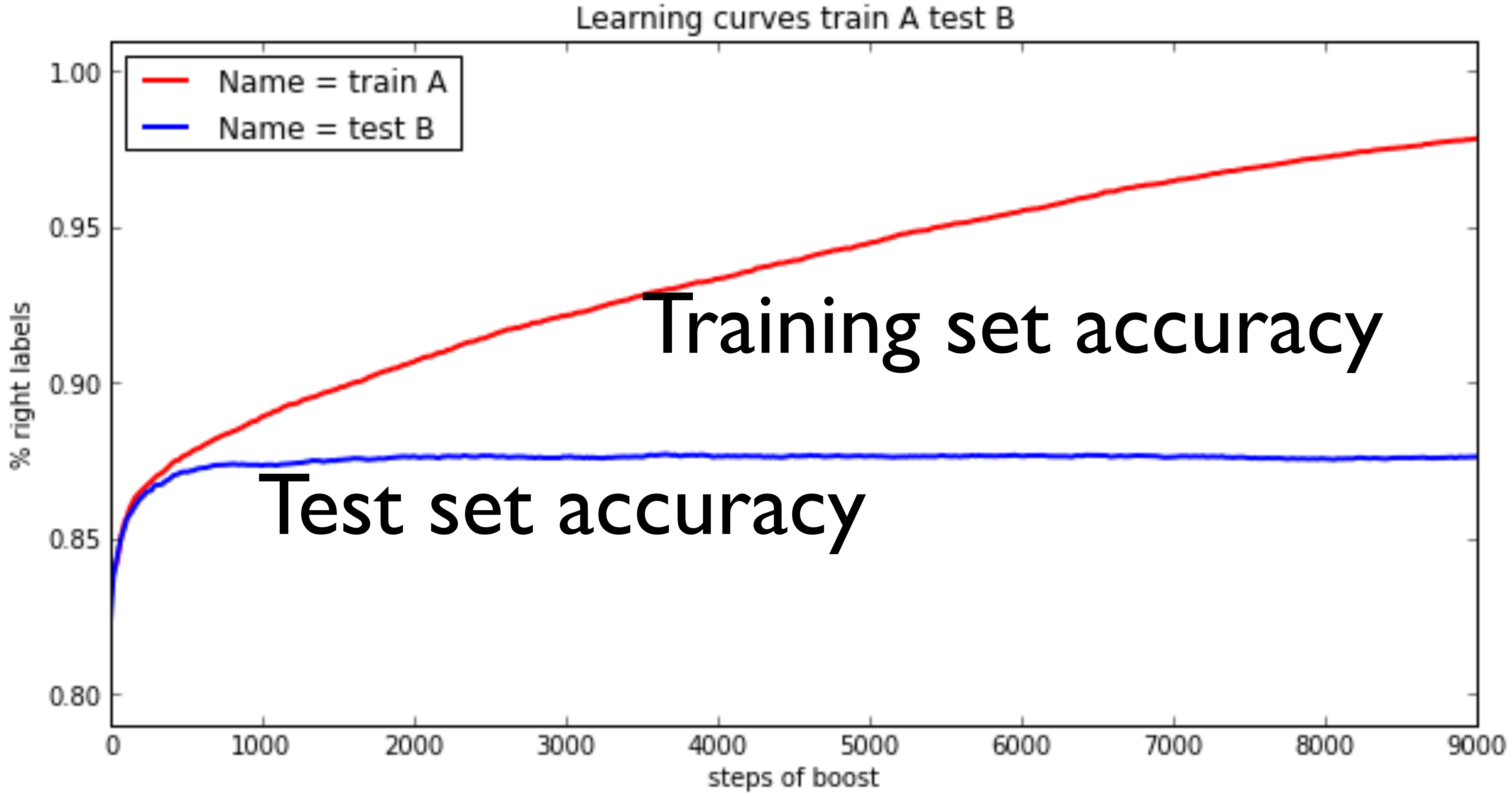
Price for sensitivity

- **How do I check quality of event discriminating function?**
  - Overfitting?
  - Correlations?
  - Relevance of figure of merit to analysis significance?
- **How do I deal with complexity?**
  - Estimate influence of model parameters
  - Extra computation
  - Organization (cross-checks, collaboration)

MVA Performance (ROC, Learning curve)



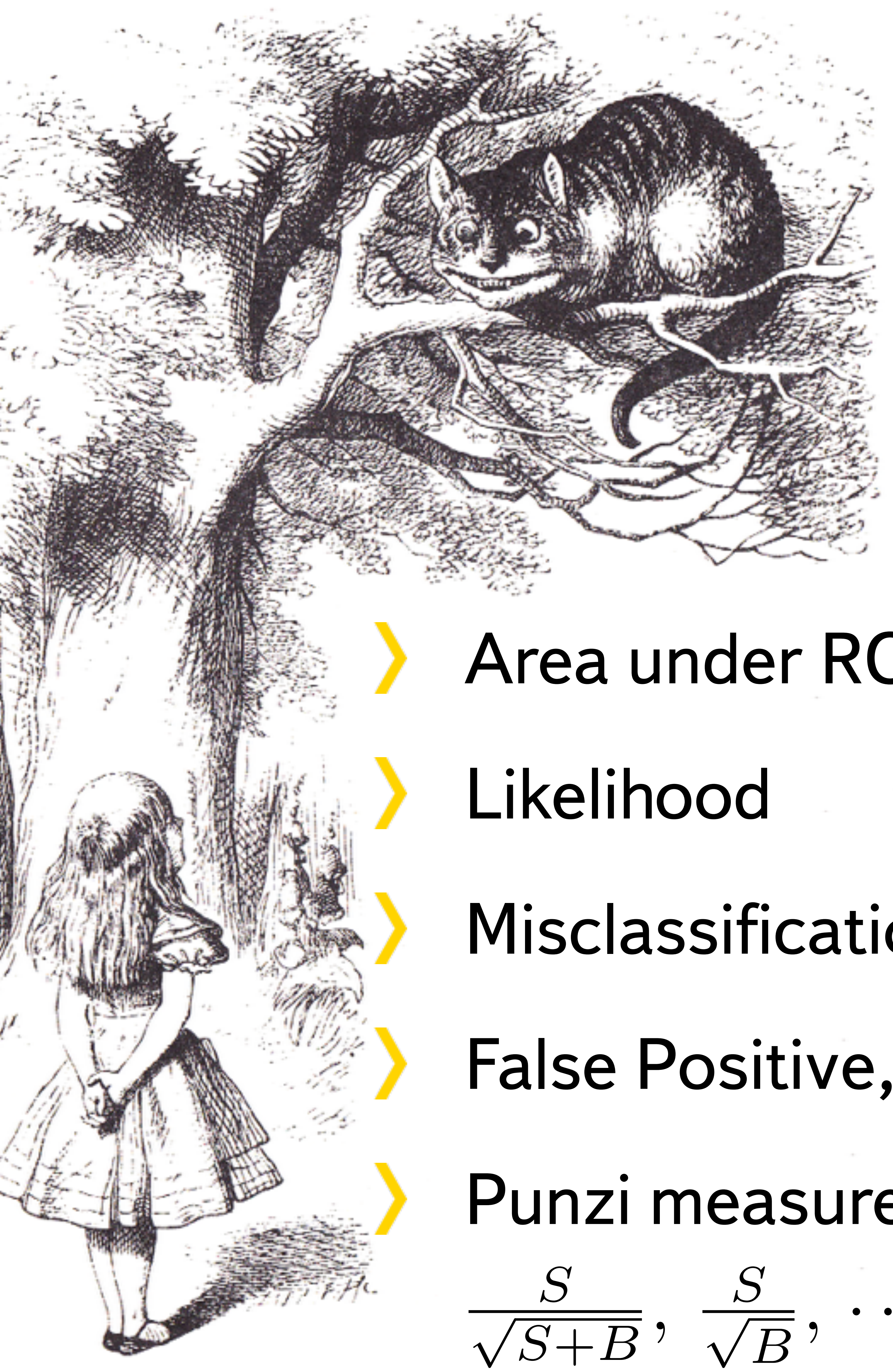




MVA algorithms: easy to find, hard to choose

- **Families:**
  - Boosted Decision Trees (BDT)
  - Artificial Neural Network (ANN)
  - Support Vector Machine (SVM)
  - Clustering, Bayesian Networks, …
- **Implementations**
  - TMVA (60+ algorithms)
  - NeuroBayes
  - python scikit-learn
  - R packages
  - Private (Matrixnet, predict.io)
  - XGBoost, …

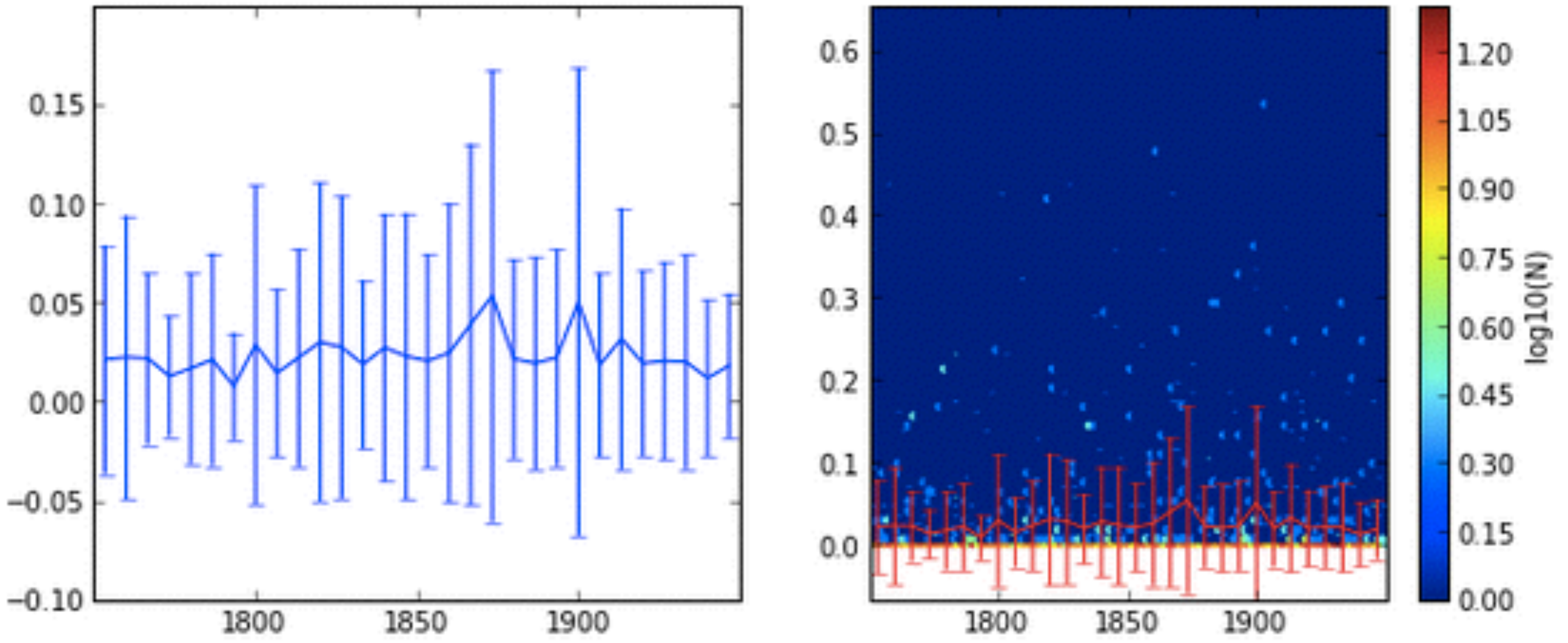
Figure-of-Merits Land



- Area under ROC
- Likelihood
- Misclassification
- False Positive, False Negative
- Punzi measure

$$\frac{S}{\sqrt{S+B}},\ \frac{S}{\sqrt{B}},\ \ldots$$

Efficiency flatness?



Complexity indicators

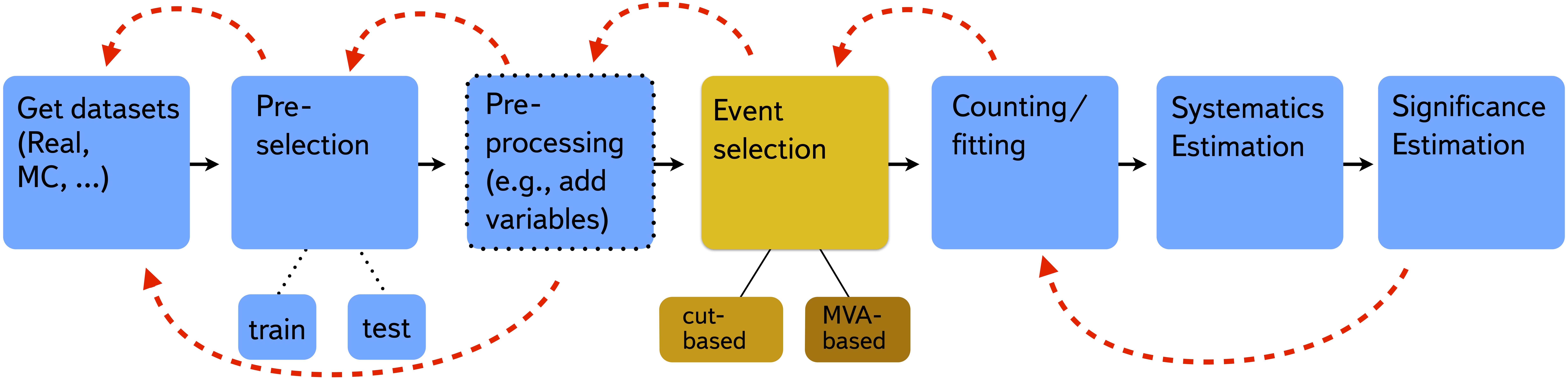
- ‘I can’t remember which version of the code I used to generate figure 13’
- ‘The new student wants to reuse that model I published three years ago but he can’t reproduce the figures’
- ‘I thought I used the same parameters but I’m getting different results!?’
- ‘It worked yesterday!’
- ‘Why did I do that?’
- ‘Where are events selected with previous version of reconstruction software?’

Complexity sources

- Domain (Physics)
- Datasources & formats
- Analysis strategy (http://bit.ly/SqDDE4)
- Analysis step details (algorithms)
- (Distributed) team communication

Analysis complexity

Case: $\tau \to 3\mu$ (LHCb)



Repeat count: 10^2 10^2 10^3 10^2 10^2 10^2

Trained models: ~1500

Requires dedicated framework!

Reproducible Experiment Platform (REP)

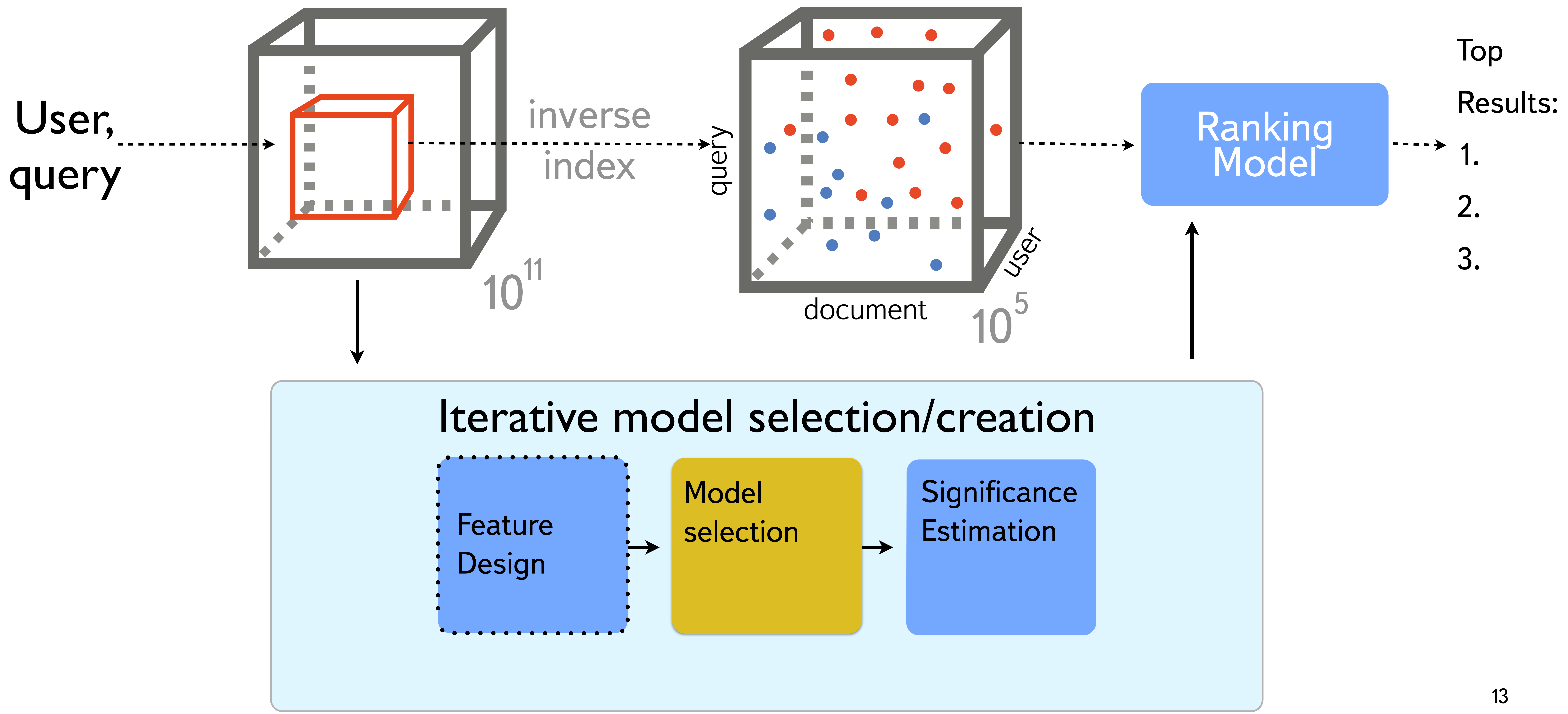
software infrastructure to support a collaborative ecosystem for computational science. It is a solution for team of researchers that allows

- running computational experiments on big shared datasets,
- obtaining reproducible and repeatable results,
- comparing measurable result consistently.

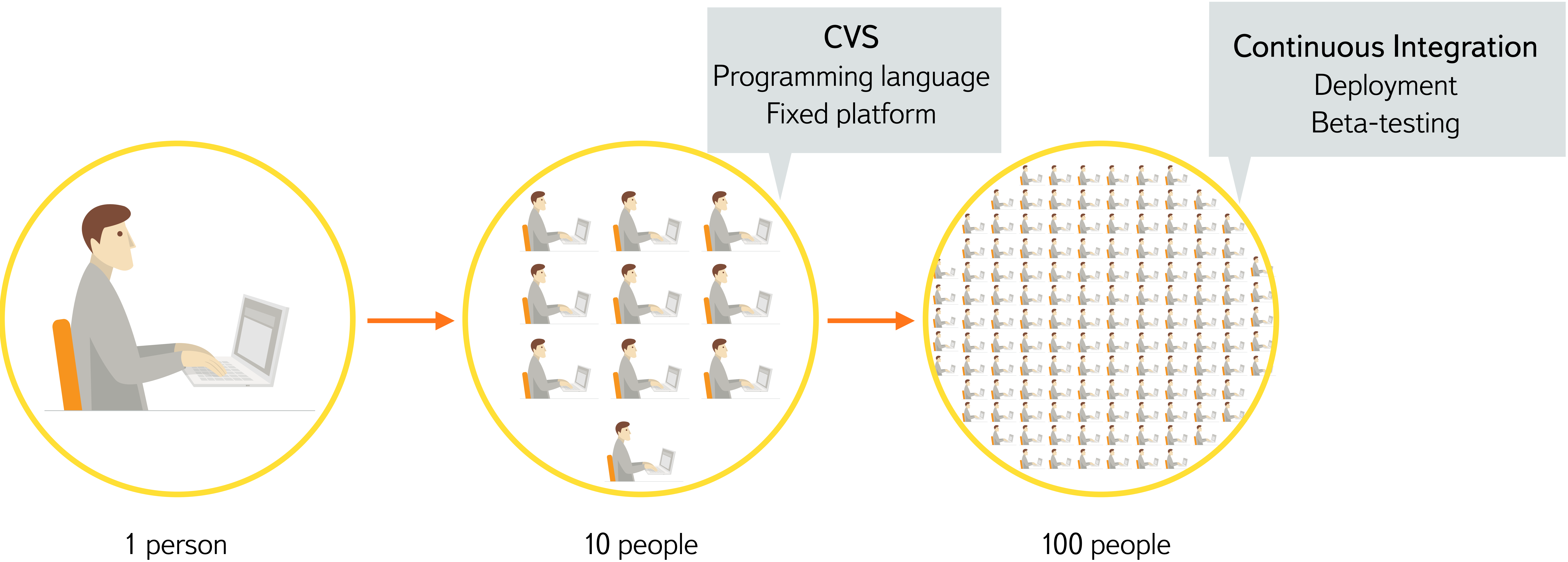
REP features / requirements

1. research automation, i.e. defining modules that can be reused later on,
2. consistent automatic cross-check,
3. online visually enhanced shared interactive environment,
4. result reproducibility (code / data provenance),
5. support for existing standard modules,
6. scalability (performance increase as additional [hardware] resources are available),
7. [flat learning curve]

Web Search Workflow



Collaborative work redux

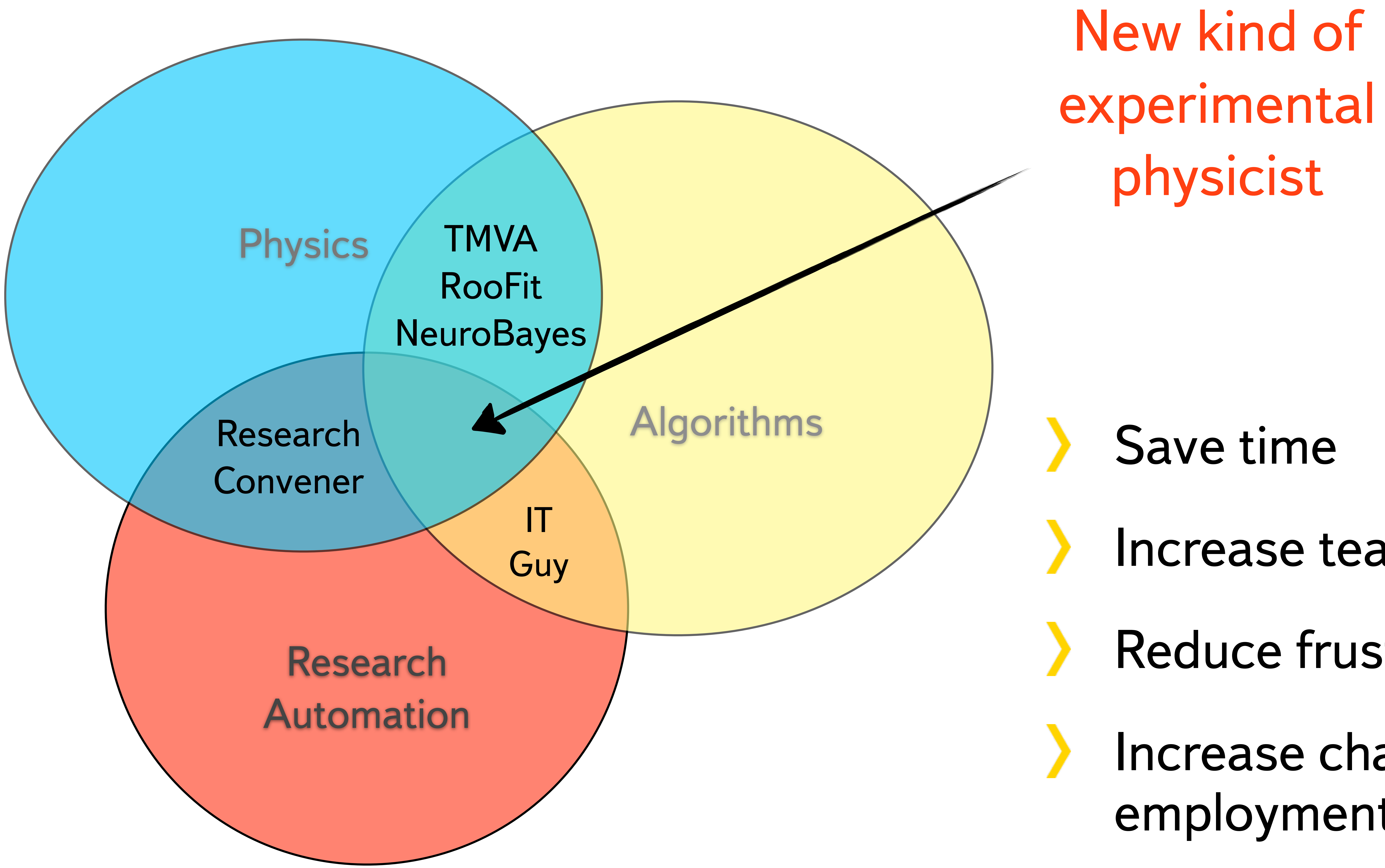


- Total «freedom»

- Formal agreements
- Experiments repository
  - share of experience, source code reuse
  - data specification, parameters, version

- Regulative infrastructure
- Automated hypotheses testing
  - **10s per week ⟹ 1000s per week**

Skills for a physicist



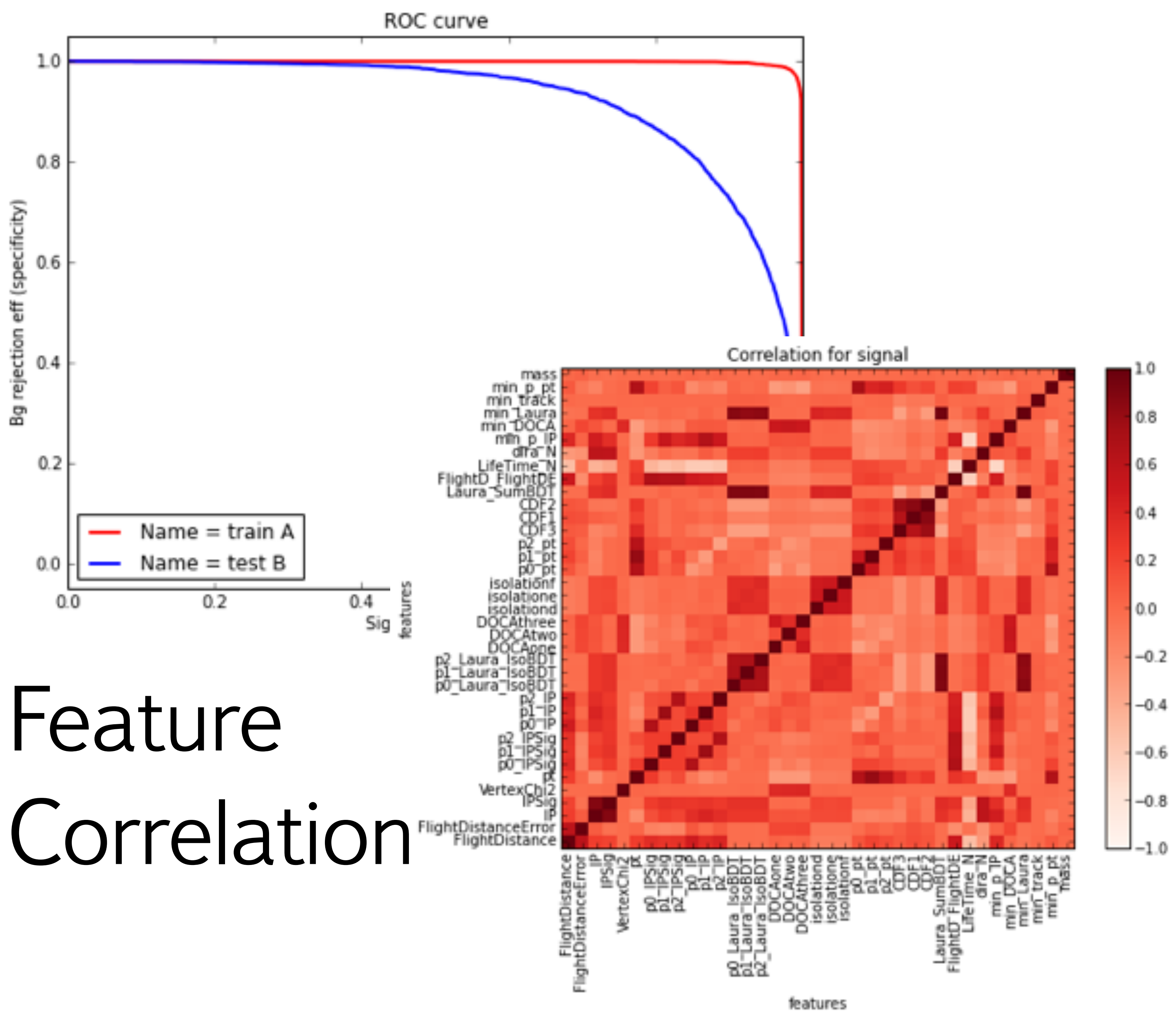
- Save time
- Increase team productivity
- Reduce frustration
- Increase chances of employment

REP for HEP

- Online & Interactive
- Support for ROOT & Python & TMVA
- Support for 3rd party classifier (e.g Matrixnet and SKLearn)
- Run heavy jobs on cluster

ROC

Feature Correlation



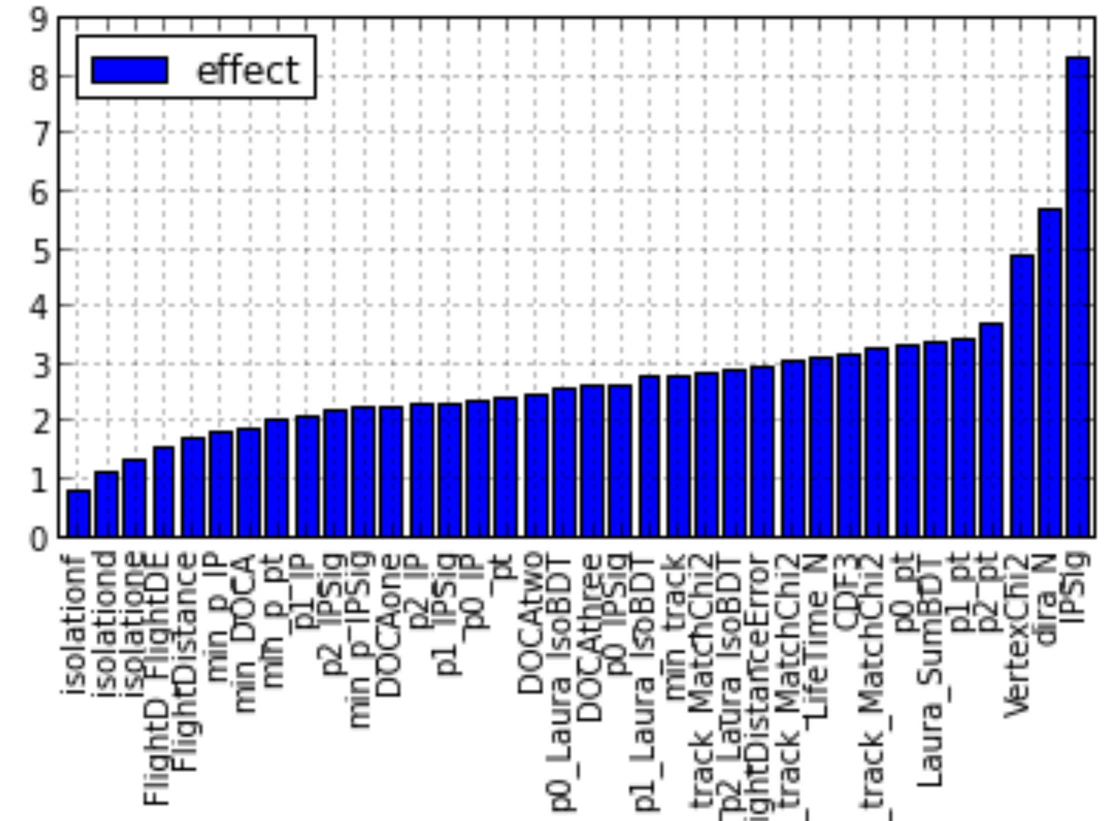
Code Example

```
[*]: import train_strategy

folding_scheme = train_strategy.TrainStrategy(directory=work_dir + 'folding/', classifier_type='TMVA')
folding_scheme.set_params(nfolds=10, features=variables, spectators=['mass'])
folding_scheme.fit(train_data_description)
folding_scheme.predict(test_file)

report = folding_scheme.get_model_report()
```

Feature Strength



More details: http://bit.ly/1fCjEqg (tomorrow)

Cases

$$B_s \to \mu^+ \mu^-$$
$$B_s \to 4\mu$$
$$\tau \to 3\mu$$
$$B \to K^* \mu^+ \mu^-$$
$$\ldots$$

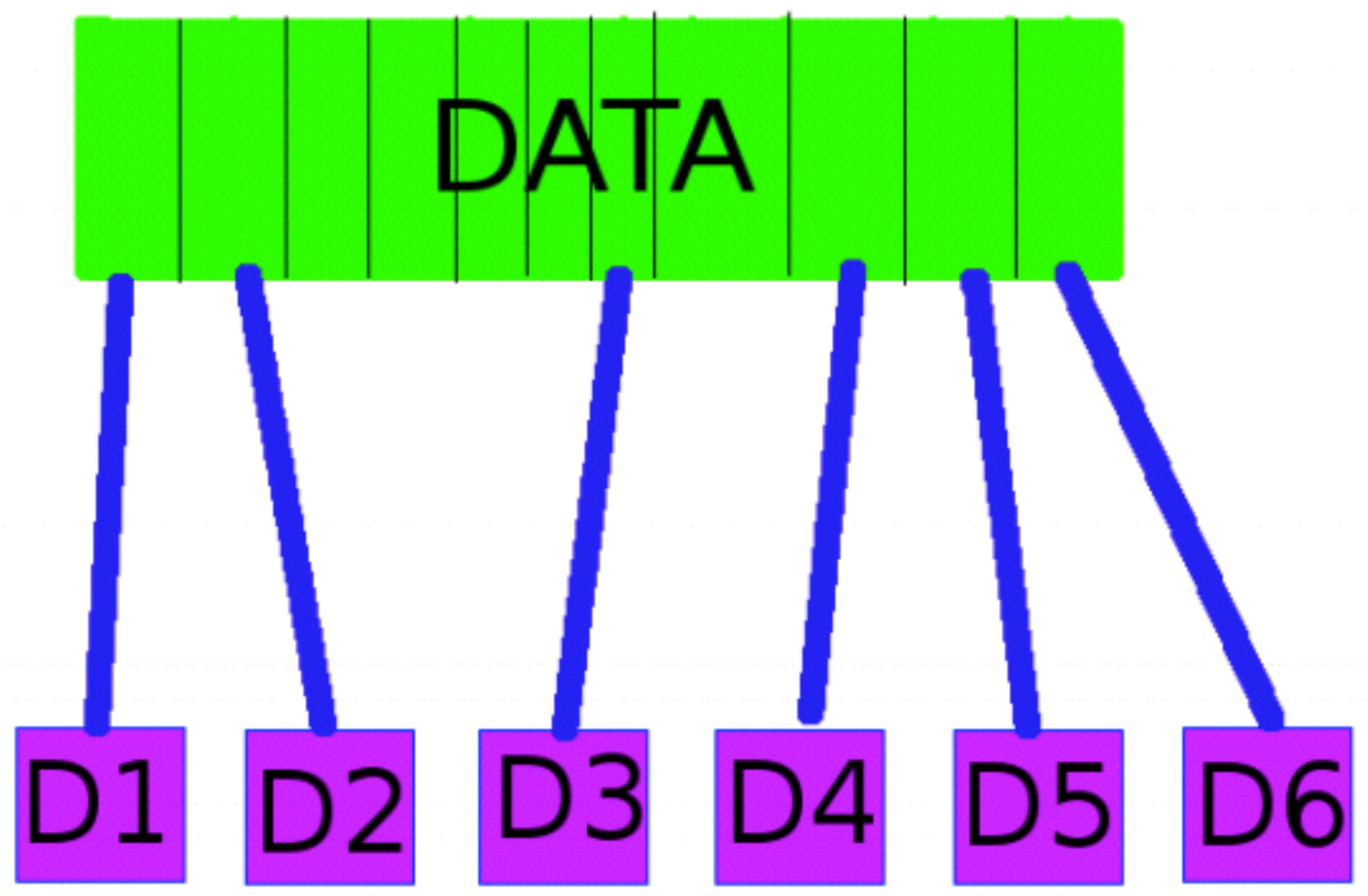
Instead of Conclusion

- New source of tools & metrics: **data science**
  - ...as well as source of complexity
- Research reproducibility = defeat of complexity
  - Environment (http://bit.ly/1fCjEqg)
  - Status: looking for new cases, adopters
- How to try?
  - Hands-on introduction **tomorrow at 16:15**
  - andrey.ustyuzhanin@cern.ch

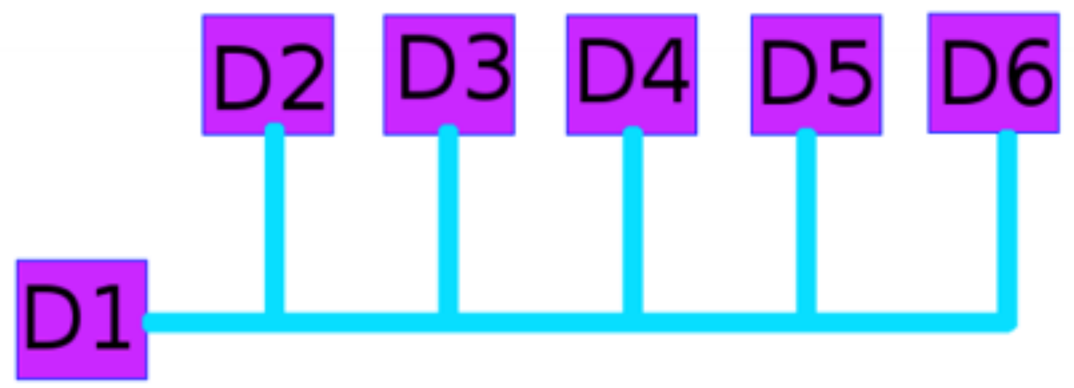
Backup

N-folding, training scheme example

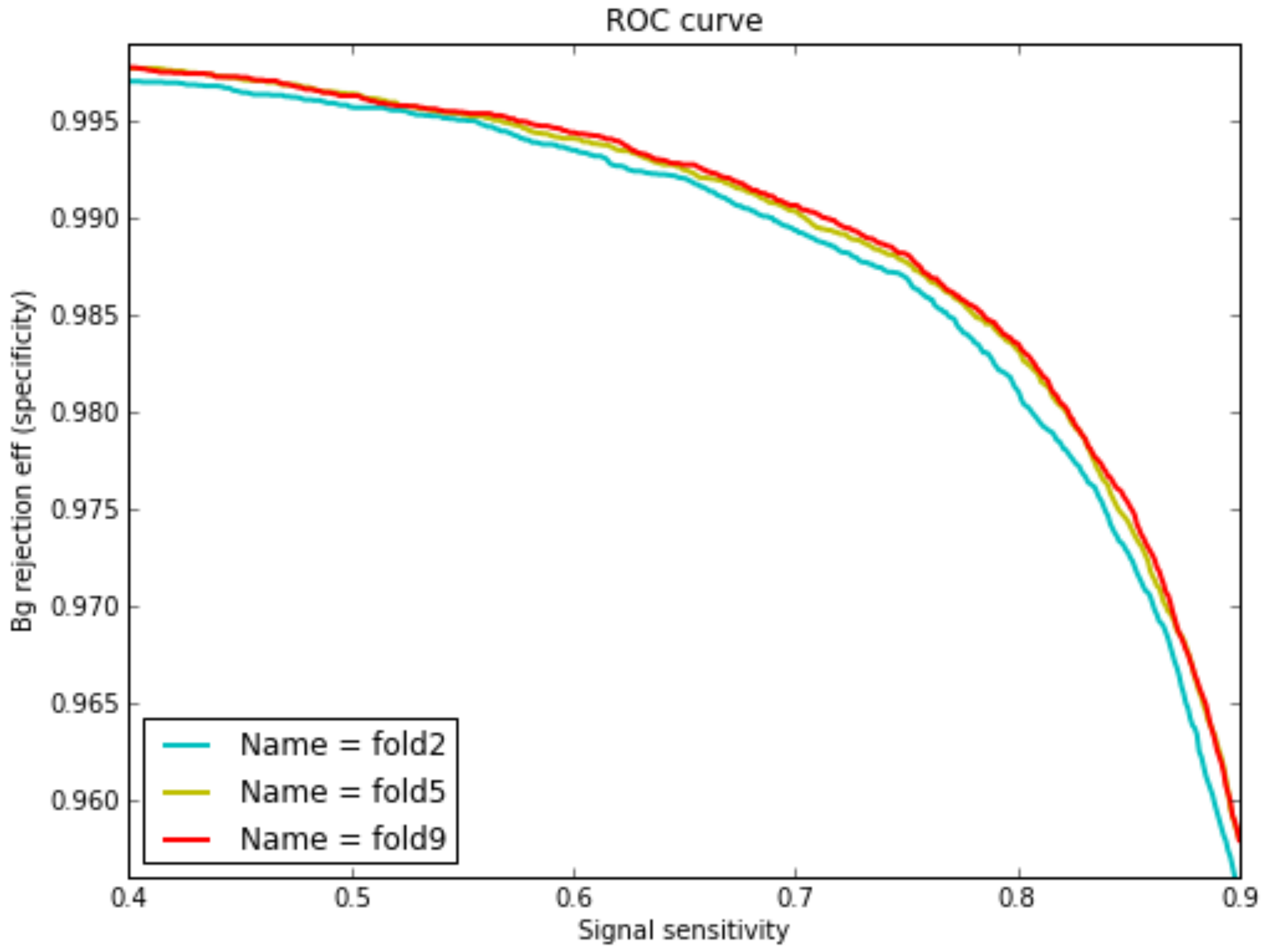
(works well for limited statistics)



Split data in N folds randomly



Take i-th fold,
train formula on remaining folds,
apply to selected one



See the difference