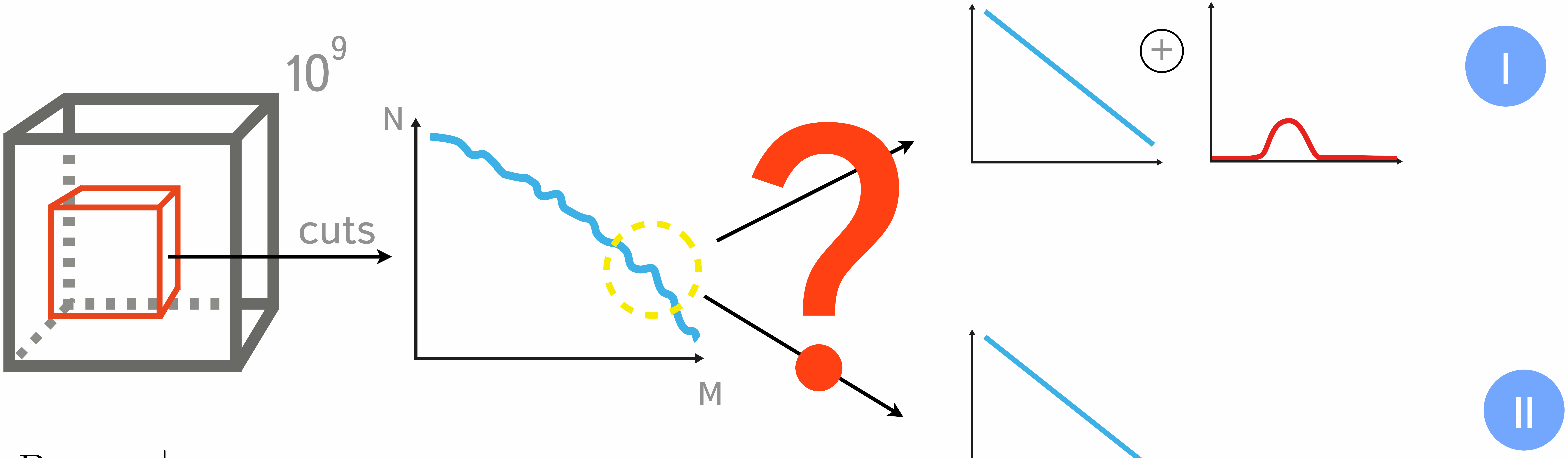
Research on Event Search

Andrey Ustyuzhanin
Yandex, Moscow

Search for rare decays



$B_s \to \mu^+\mu^-$

$B_s \to 4\mu$

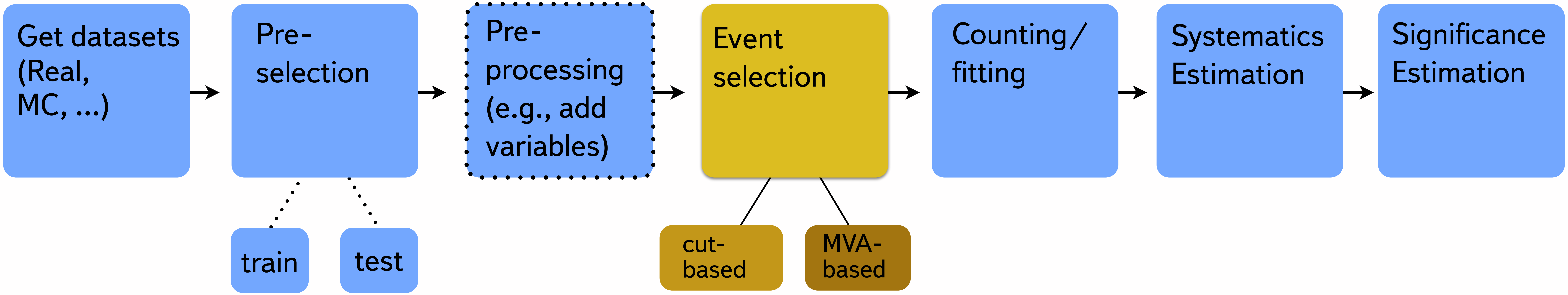
$\tau \to 3\mu$

$B \to K^*\mu^+\mu^-$

$\ldots$

Quest for analysis sensitivity

Analysis Value Chain



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 schemes (e.g. n-folding, ensembling, blending, cascading)

Data Science

«How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?»

Tom Mitchell, CMU

Price for sensitivity

- **How do I check quality of 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)

Growing a tree

M_t - invariant mass
P_t - jet transverse momentum
H_t - sum of P_t for all objects

Pros:

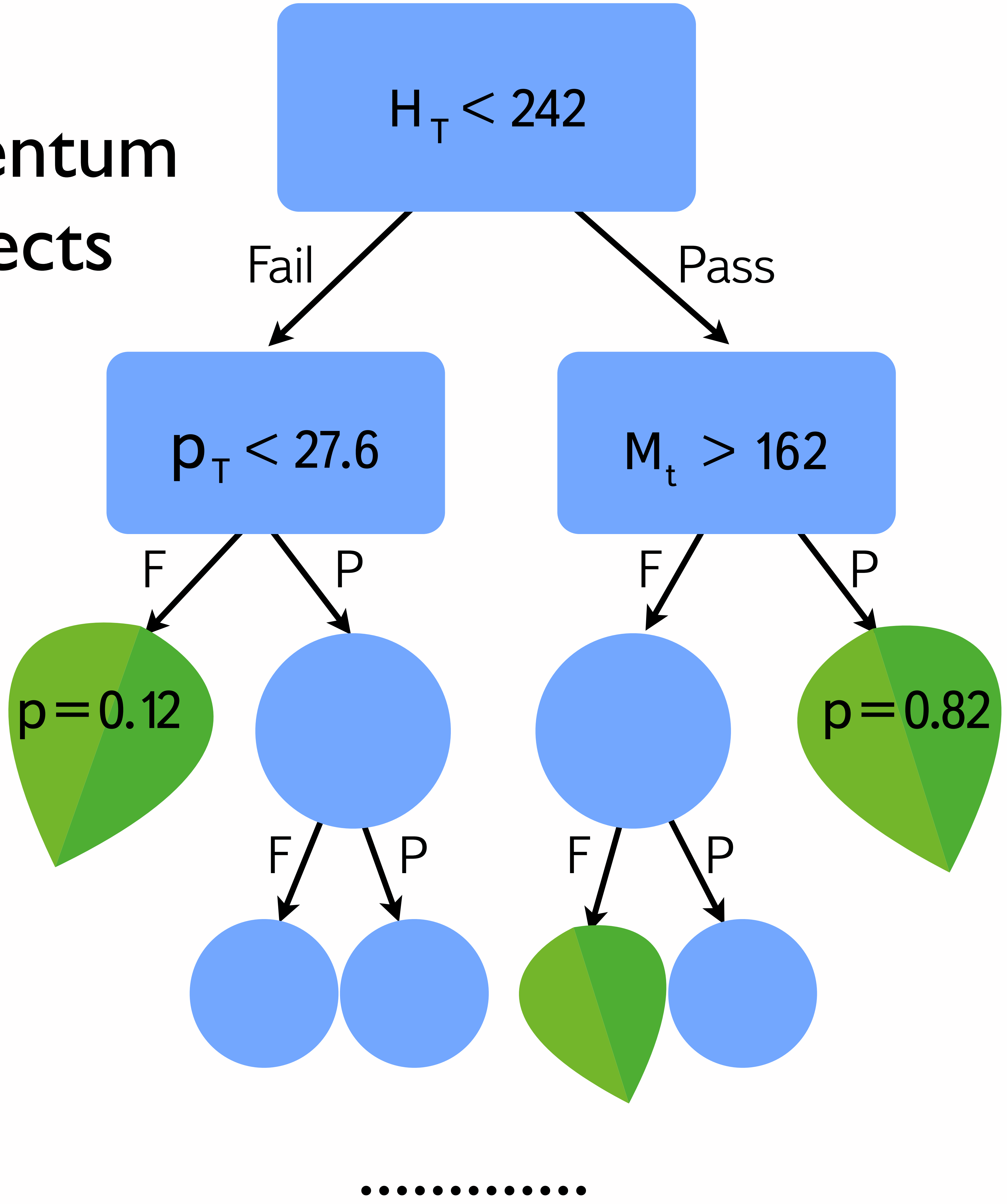
- easy to build

- interpretable

Parameters:

- max depth
- splitting criteria
- stopping criteria

...



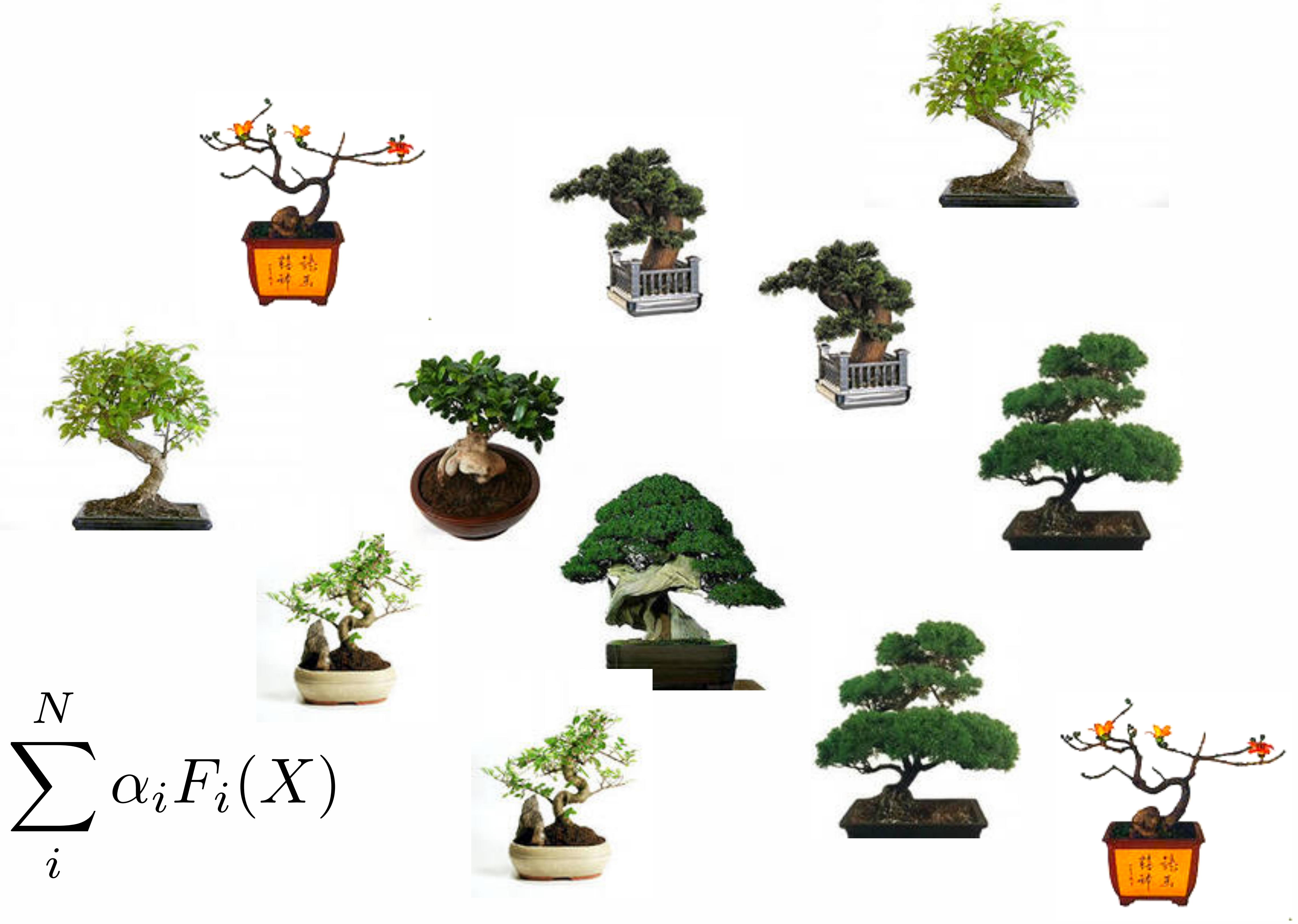
Cons:

- not very accurate (prone to overfitting)

- do not represent real probability distributions

Combining weak trees into strong forest

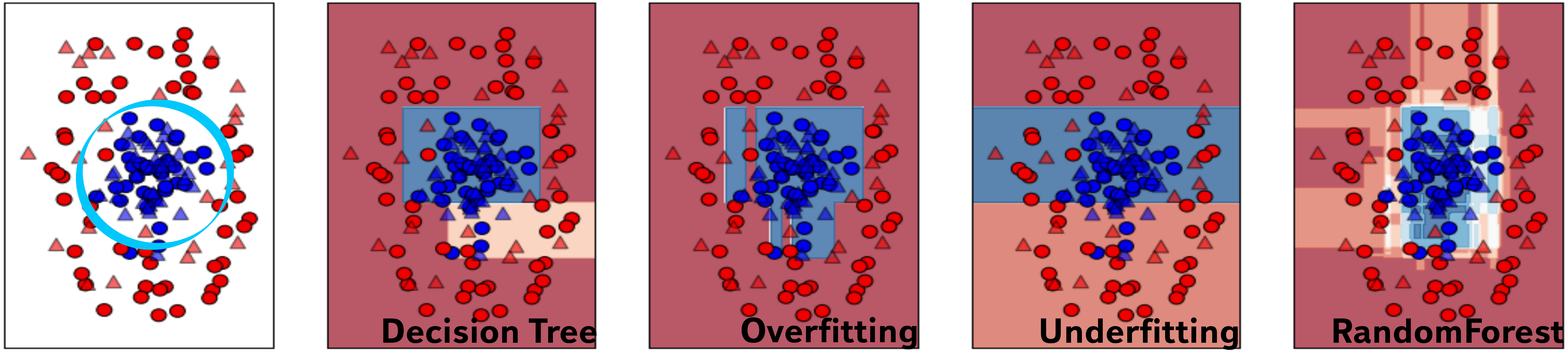
VS

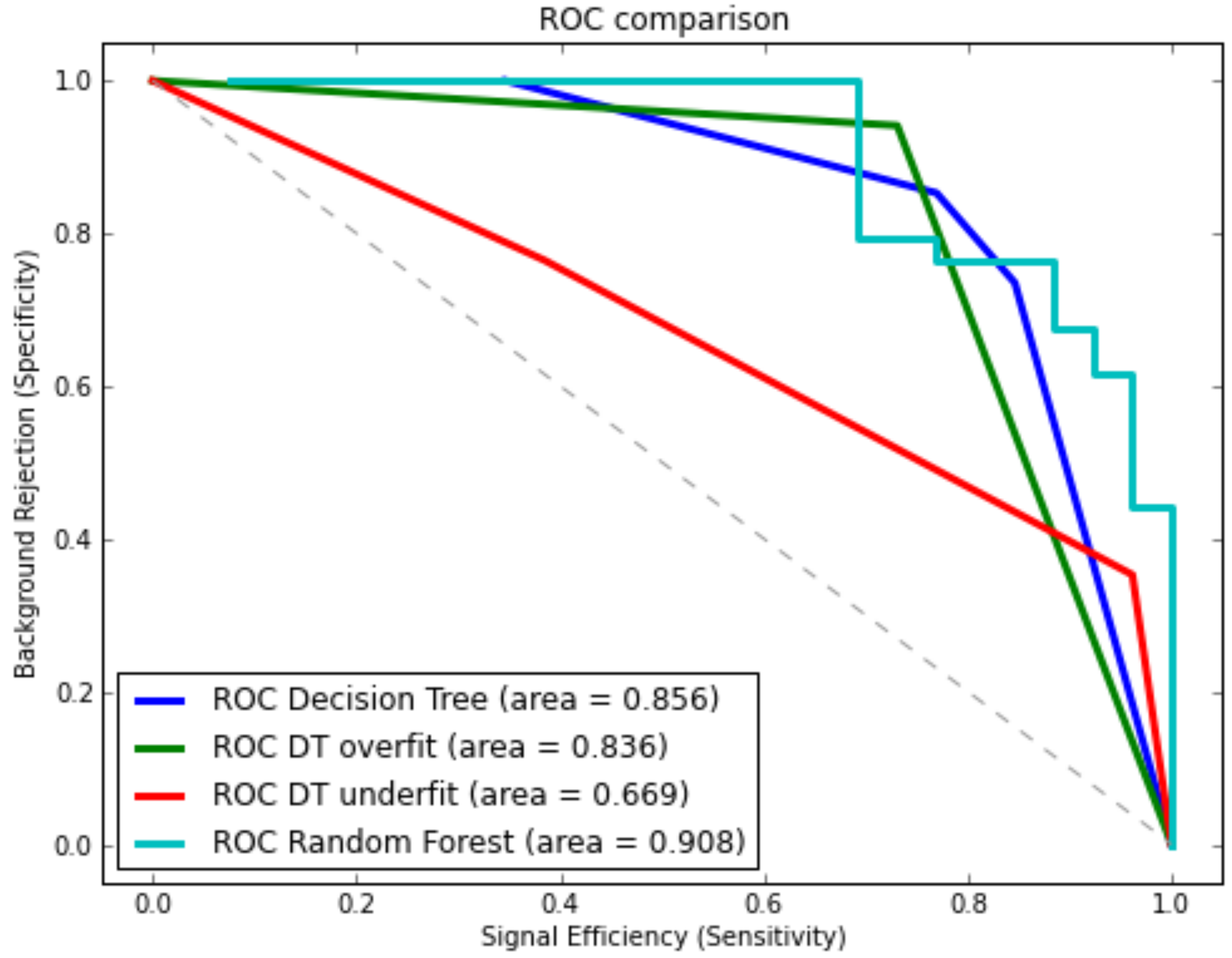


$$F(X) = \sum_{i}^{N} \alpha_i F_i(X)$$

$$\varepsilon \leq \prod_{i}^{N} 2\sqrt{\varepsilon_i(1-\varepsilon_i)}$$

MVA Performance (ROC, Learning curve)





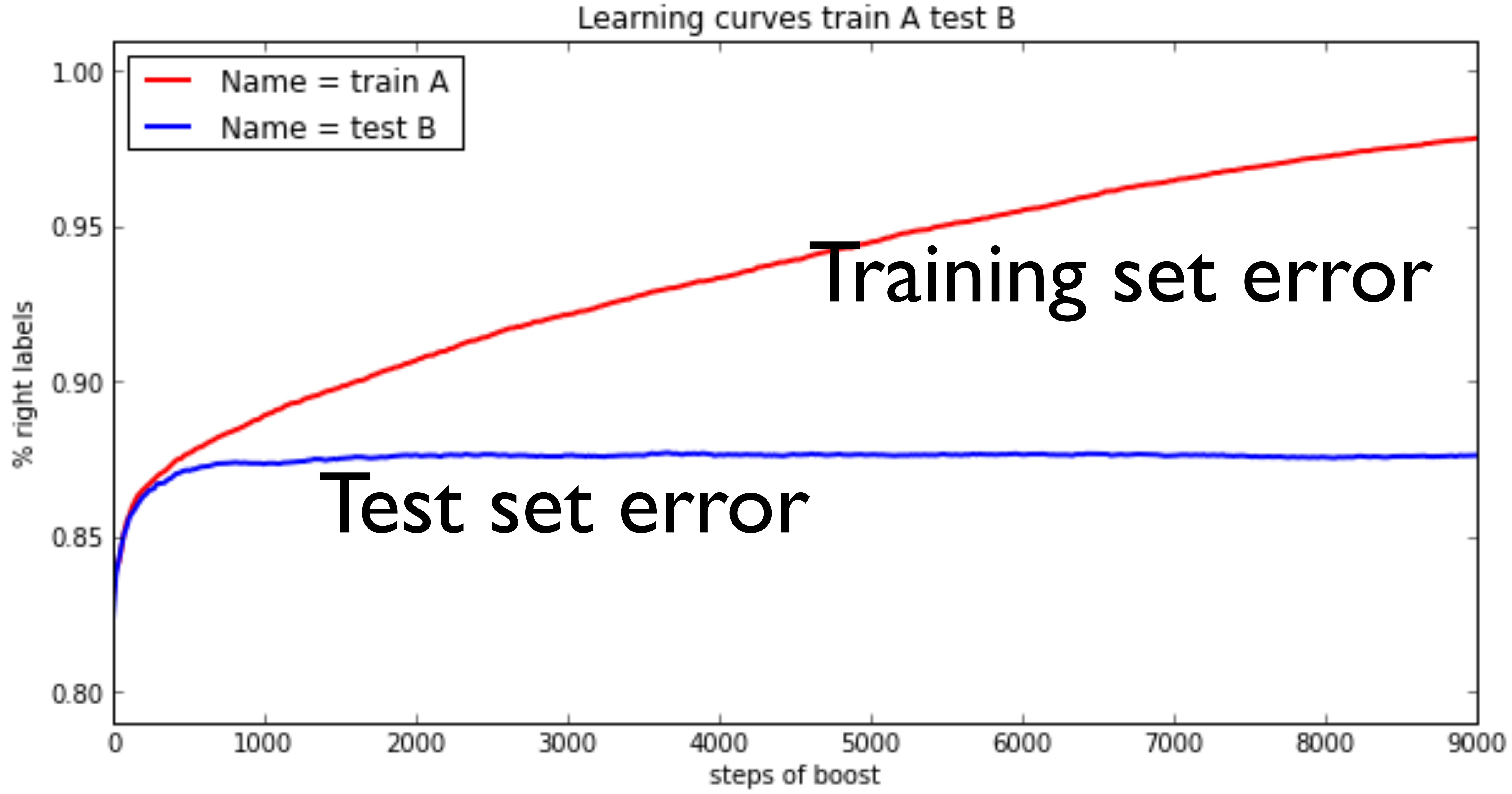
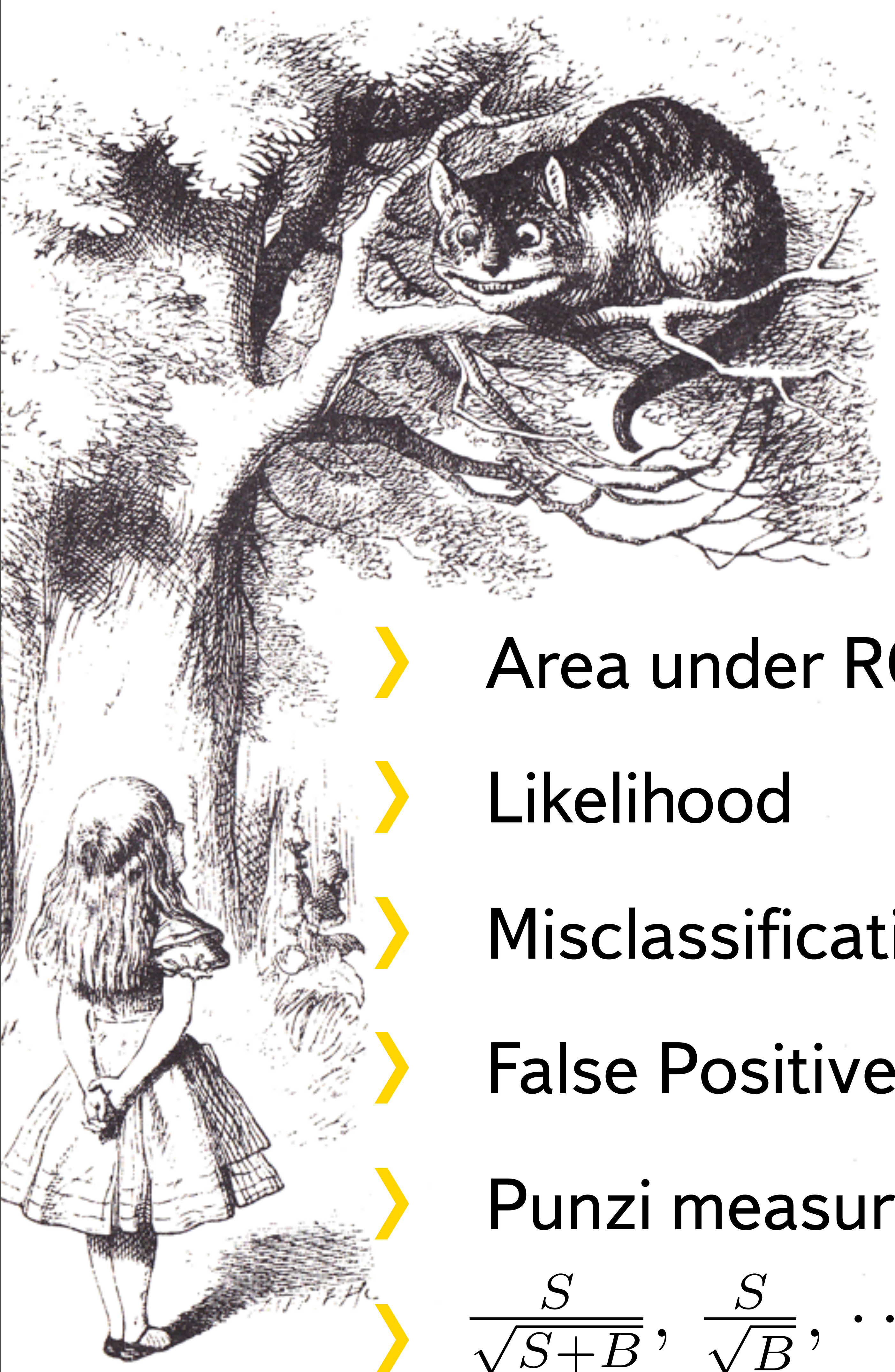
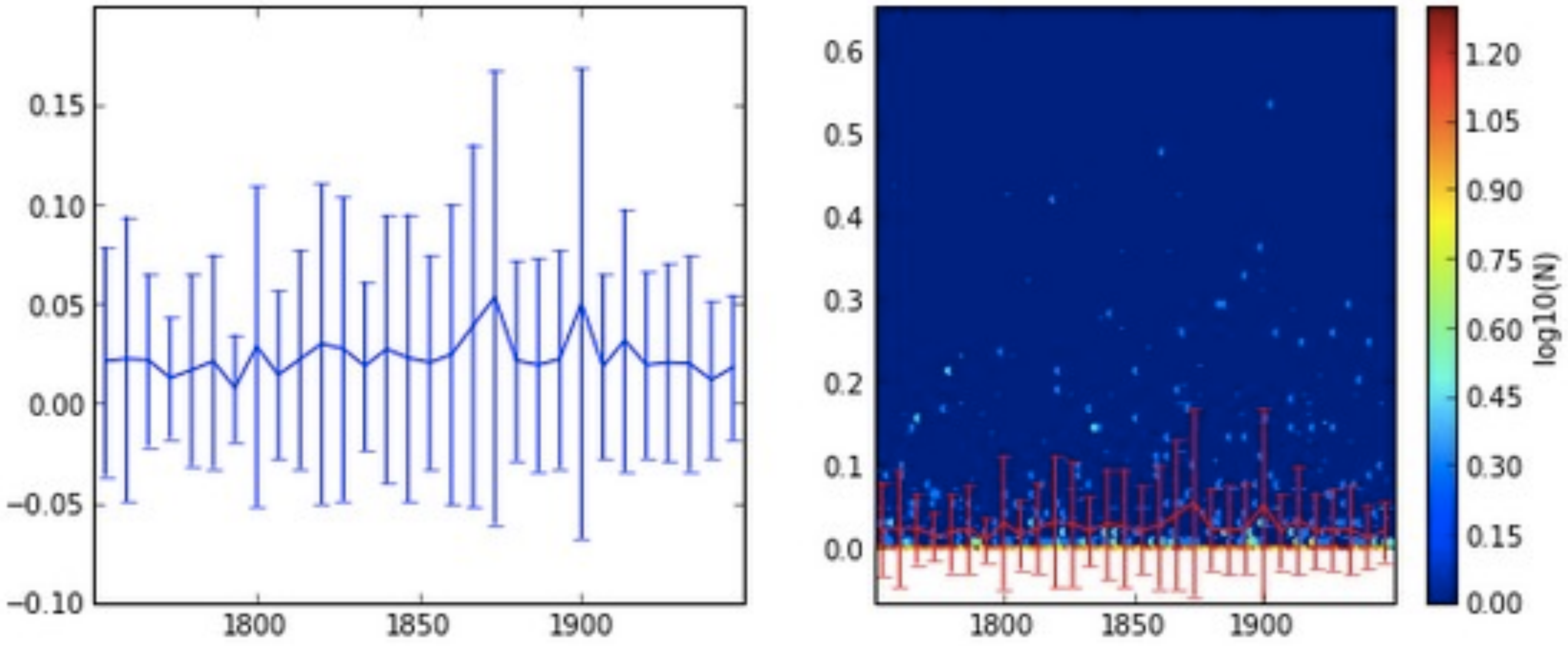


Figure-of-Merits Land



- Area under ROC
- Likelihood
- Misclassification
- False Positive, False Negative
- Punzi measure
- $\frac{S}{\sqrt{S+B}}, \frac{S}{\sqrt{B}}, \ldots$

Efficiency flatness?

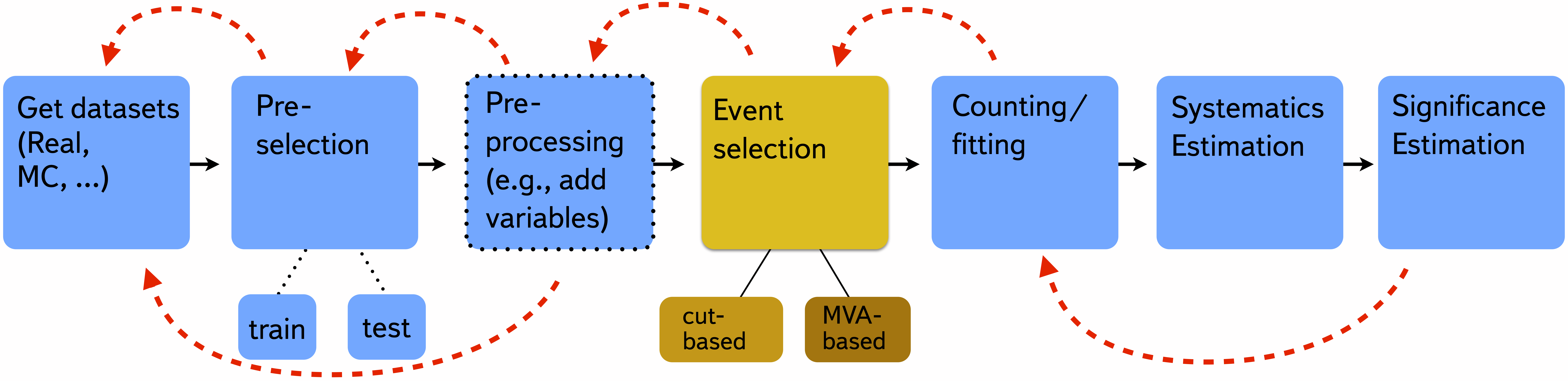


Complexity indicators

- ‘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?’

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

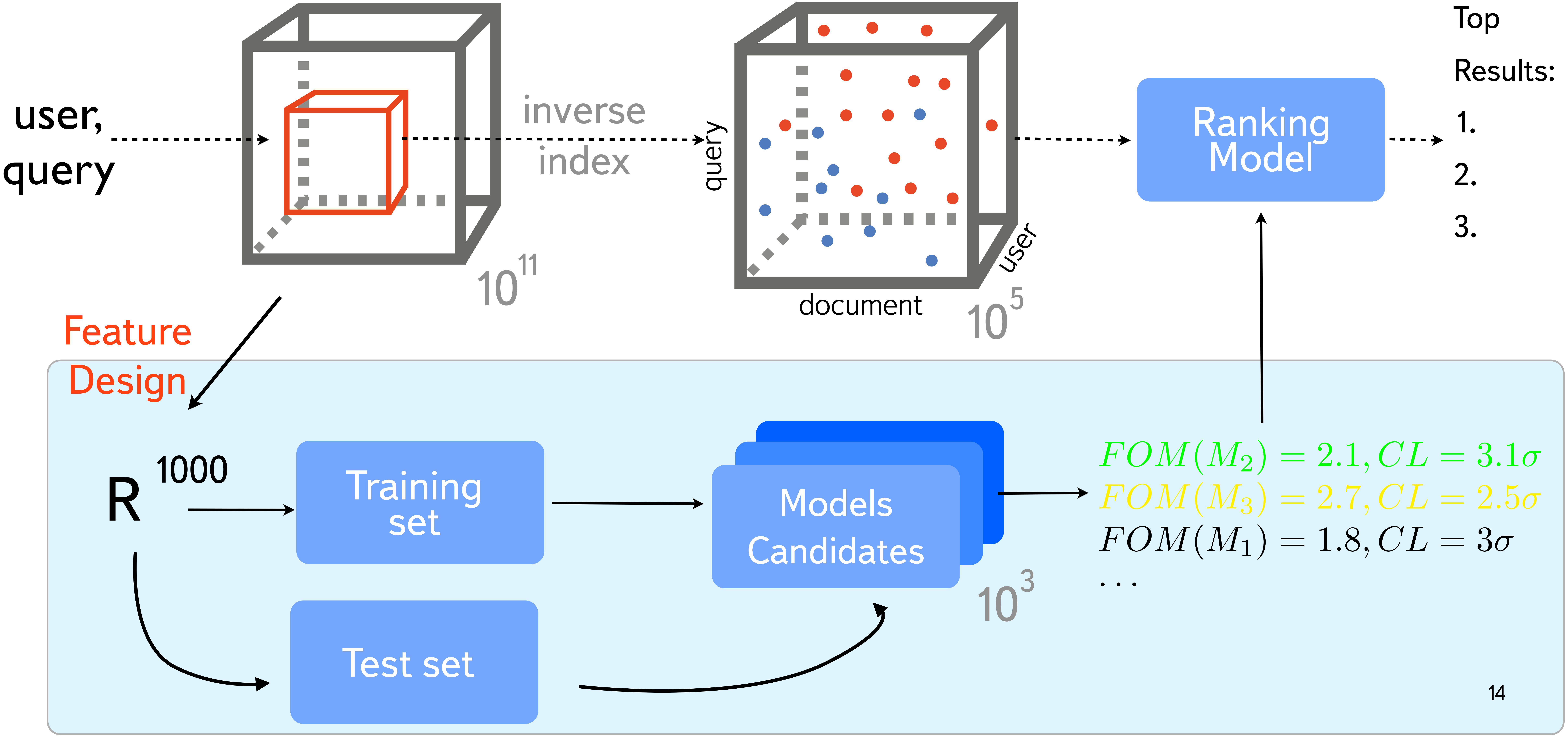
Requires dedicated framework!

Research reproducibility

- By yourself
- By your team members
- By member of another team in the same domain (HEP, Cosmology, …)
- By someone else

Requires dedicated framework!

Web Search Workflow



user, query

inverse index

10^{11}

query

user

document

10^5

Ranking Model

Top Results:

1.

2.

3.

Feature Design

R^{1000}

Training set

Test set

Models Candidates

10^3

$FOM(M_2) = 2.1, CL = 3.1\sigma$

$FOM(M_3) = 2.7, CL = 2.5\sigma$

$FOM(M_1) = 1.8, CL = 3\sigma$

$\ldots$

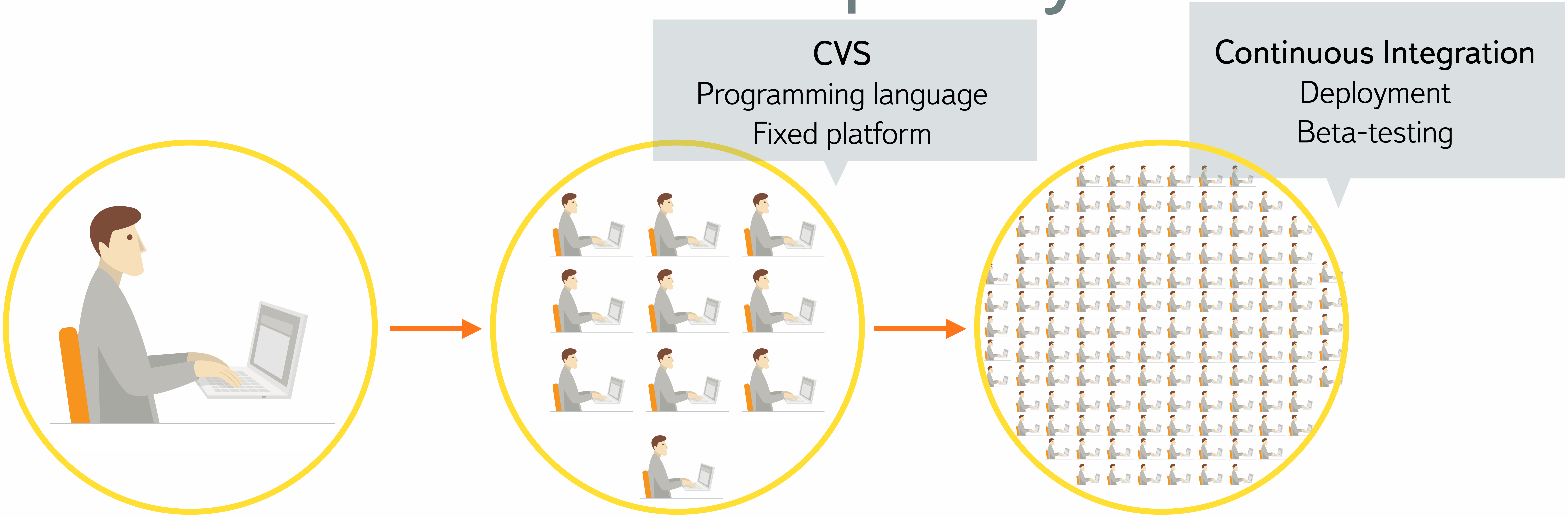
Old model

- Low level of shared knowledge
- No well-defined quality criteria
- Not scaleable
- Ineffective
- Slow
- Difficult to change

Collaborative Model

- **Consistent automatic cross-checks**
- **Ready-to-use tools & components**
- **Changes management**
- Online shared environment
- Reproducibility of results
- Easy to play

Collaborative work as complexity dimension



CVS
Programming language
Fixed platform

Continuous Integration
Deployment
Beta-testing

1 person

10 people

100 people

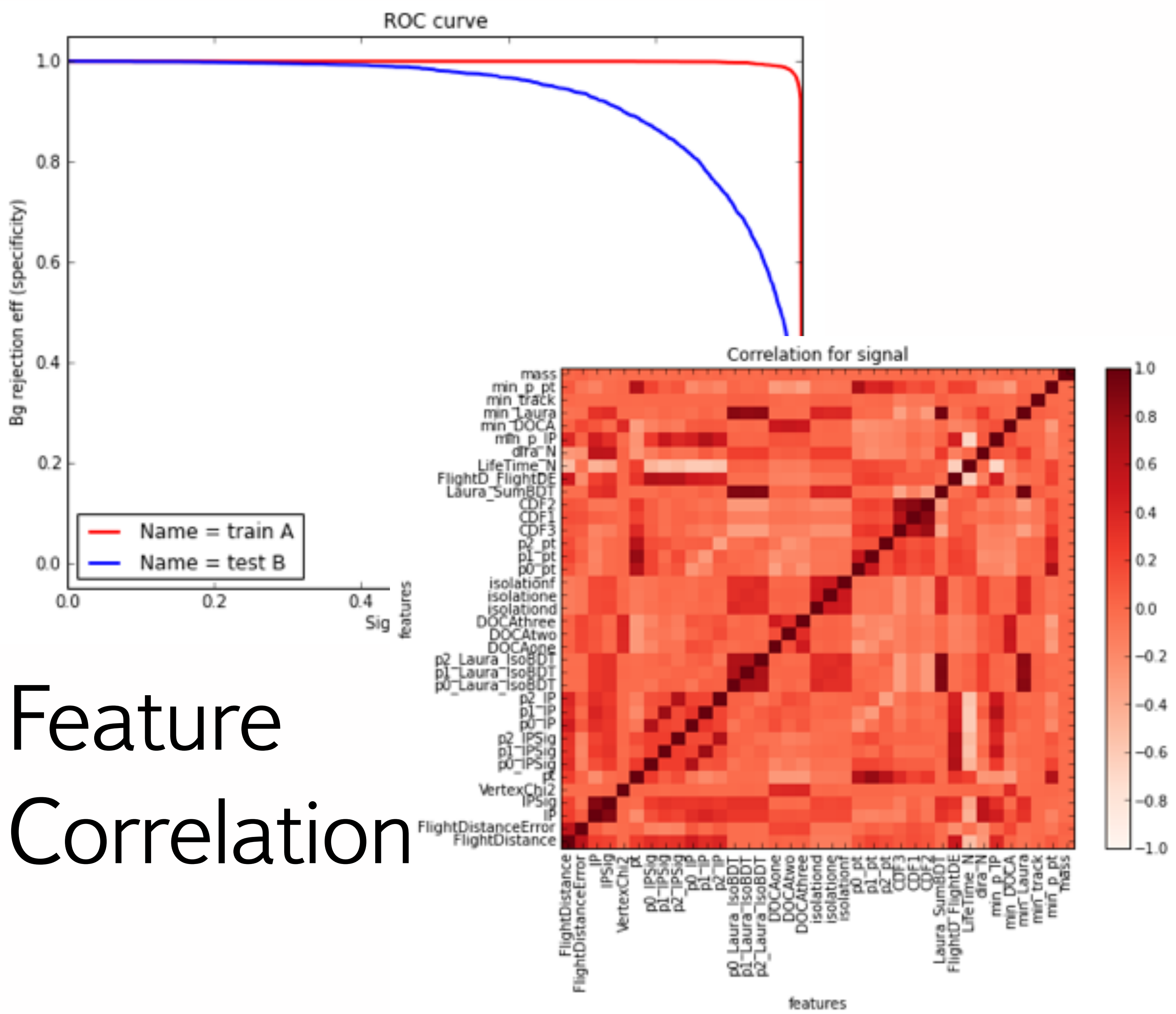
- Total «freedom»

- 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**

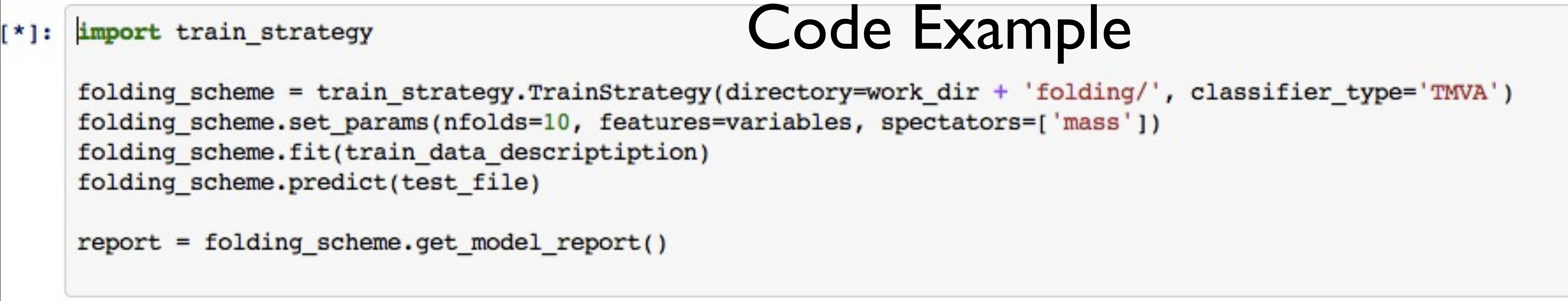
Prototype for HEP: Event Filter + IPython

- Online & Interactive
- Runs on lxplus.cern.ch
- support for ROOT & Python & Bender
- Train Matrixnet
- Run heavy jobs on cluster



ROC

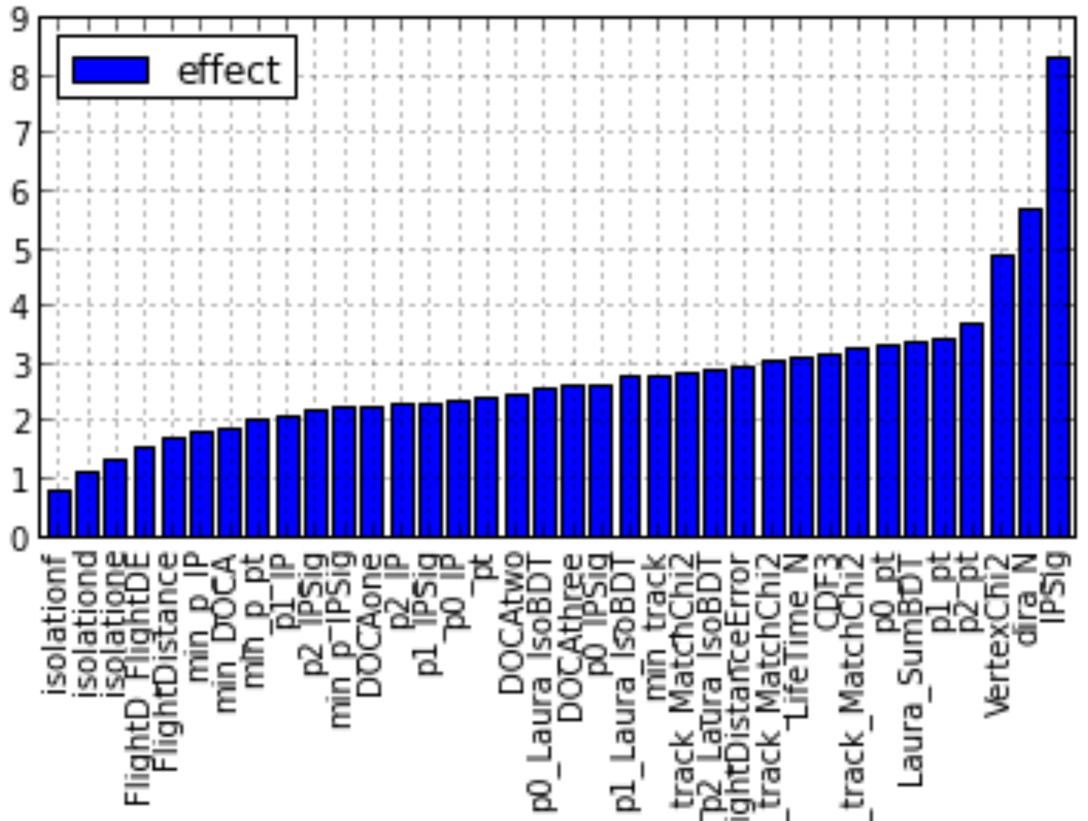
Feature Correlation



Code Example

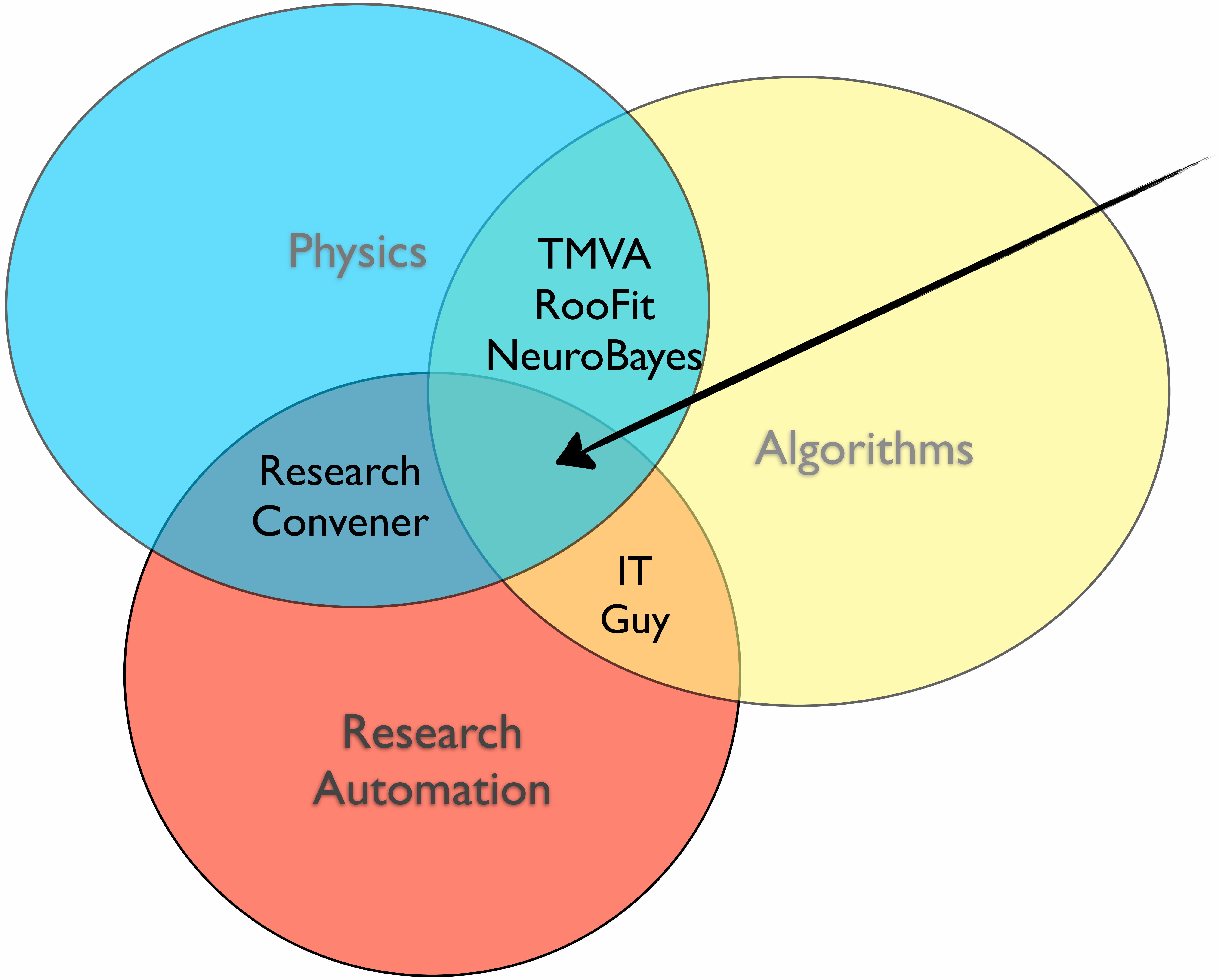
```
[*]: import train_strategy

     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 (~10th April, LHCb)

Skills for a physicist



new kind of experimental physicist

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

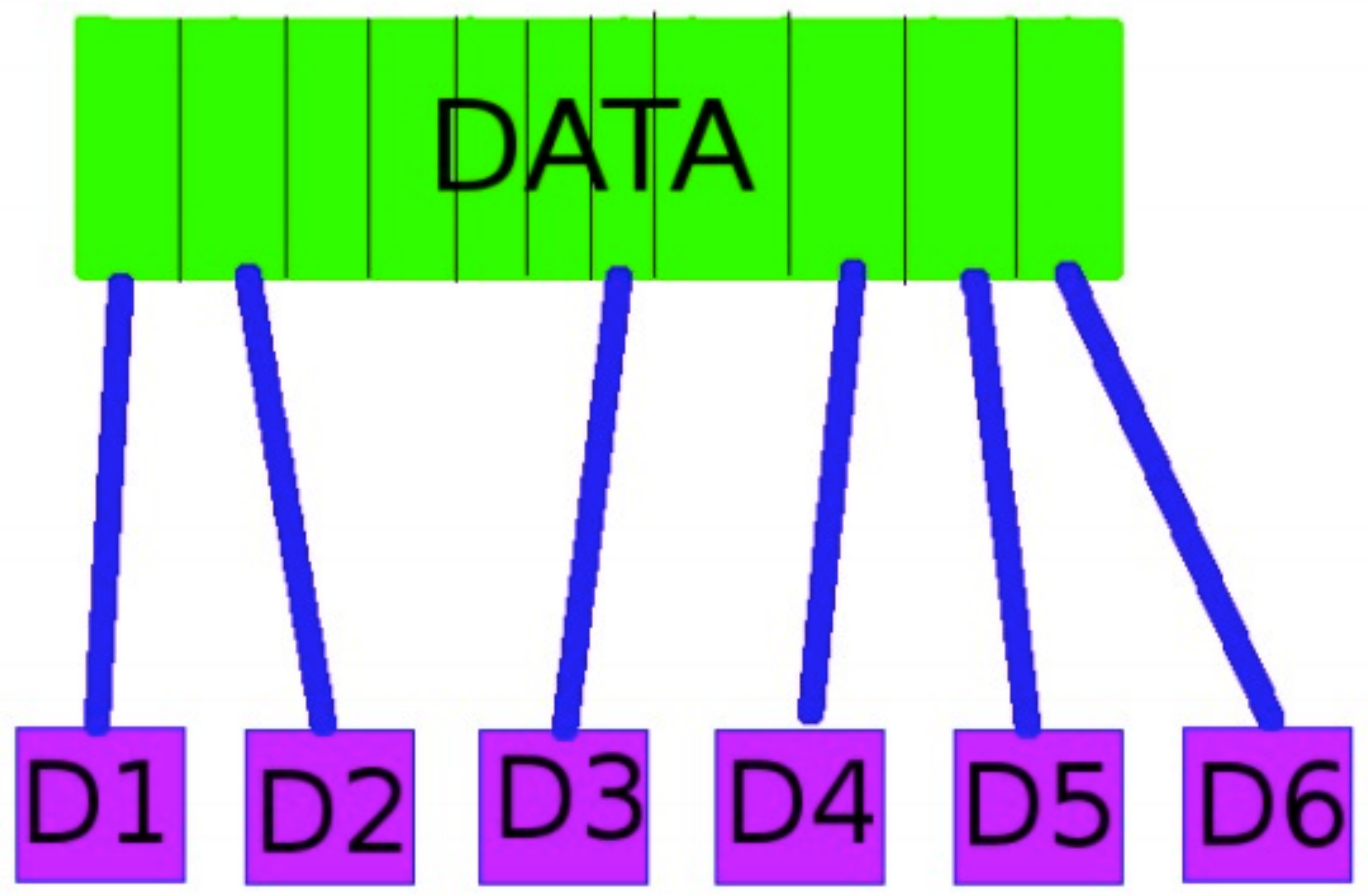
Conclusion

- New source of tools & metrics: **data science**
  - ...as well as source of complexity
- Reproducibility as indicator of mastering complexity
  - Environment (http://bit.ly/1fCjEqg, ~10th April, LHCb Analysis week@CERN)
  - New research methodology emerging

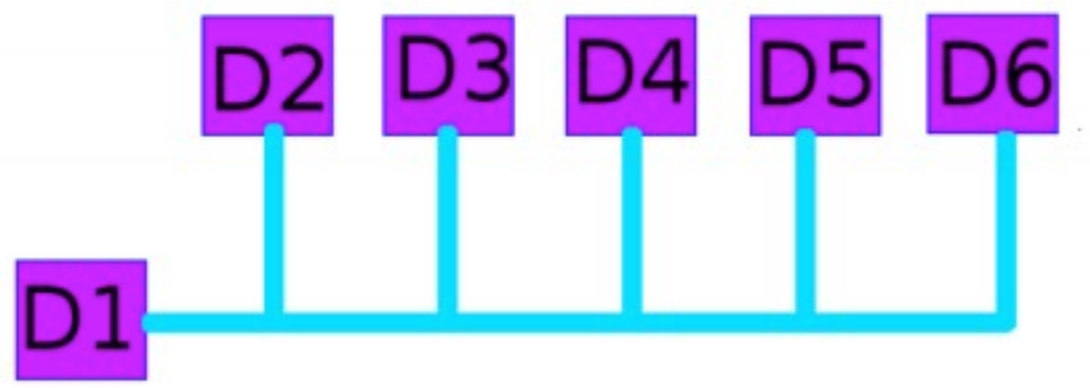
Backup

N-folding, training scheme example

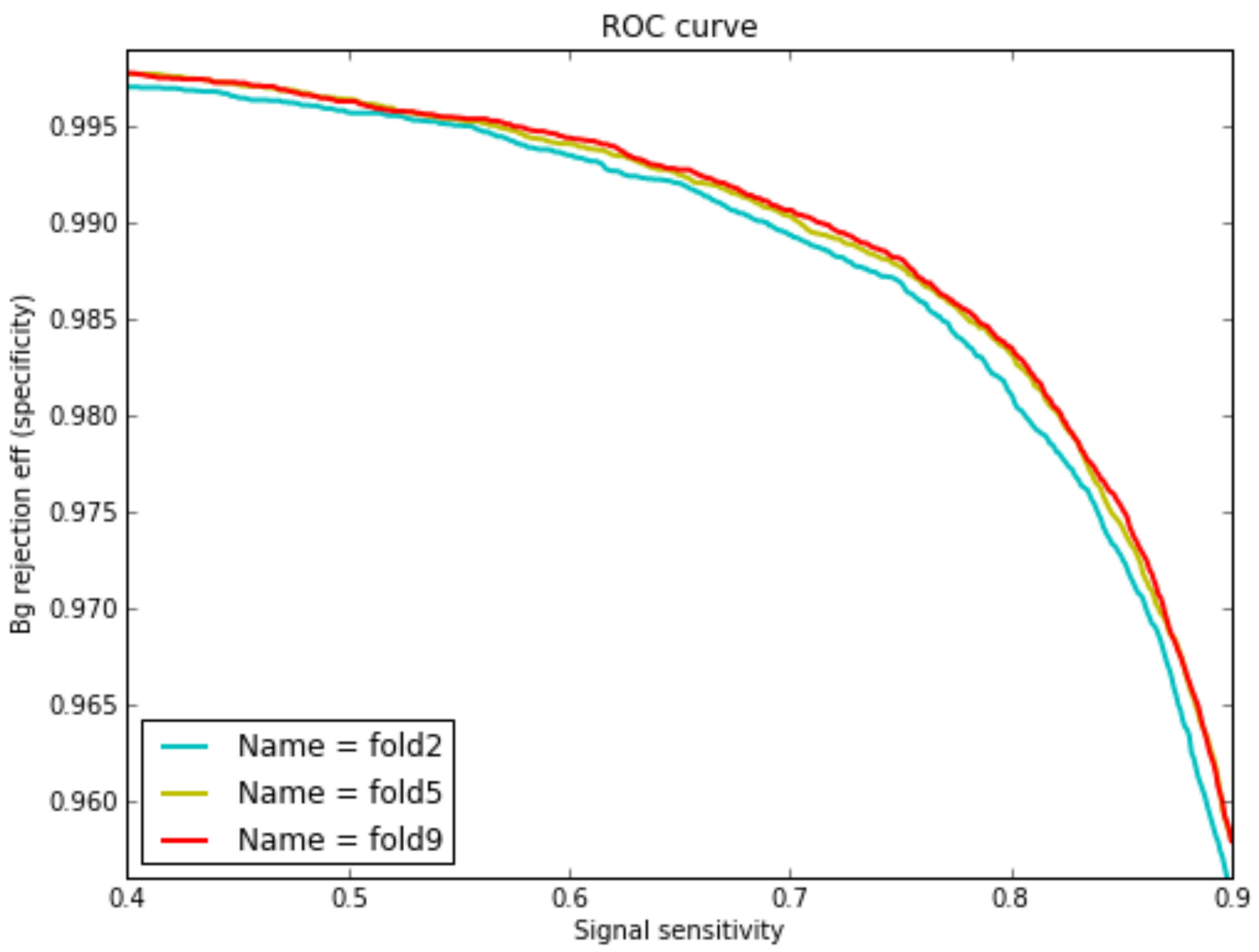
(works well for limited statistics)



Split data in N folds randomly



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



See the difference