Data-driven Evaluation of Visual Quality Measures

Michael Sedlmair & Michaël Aupetit

universität wien

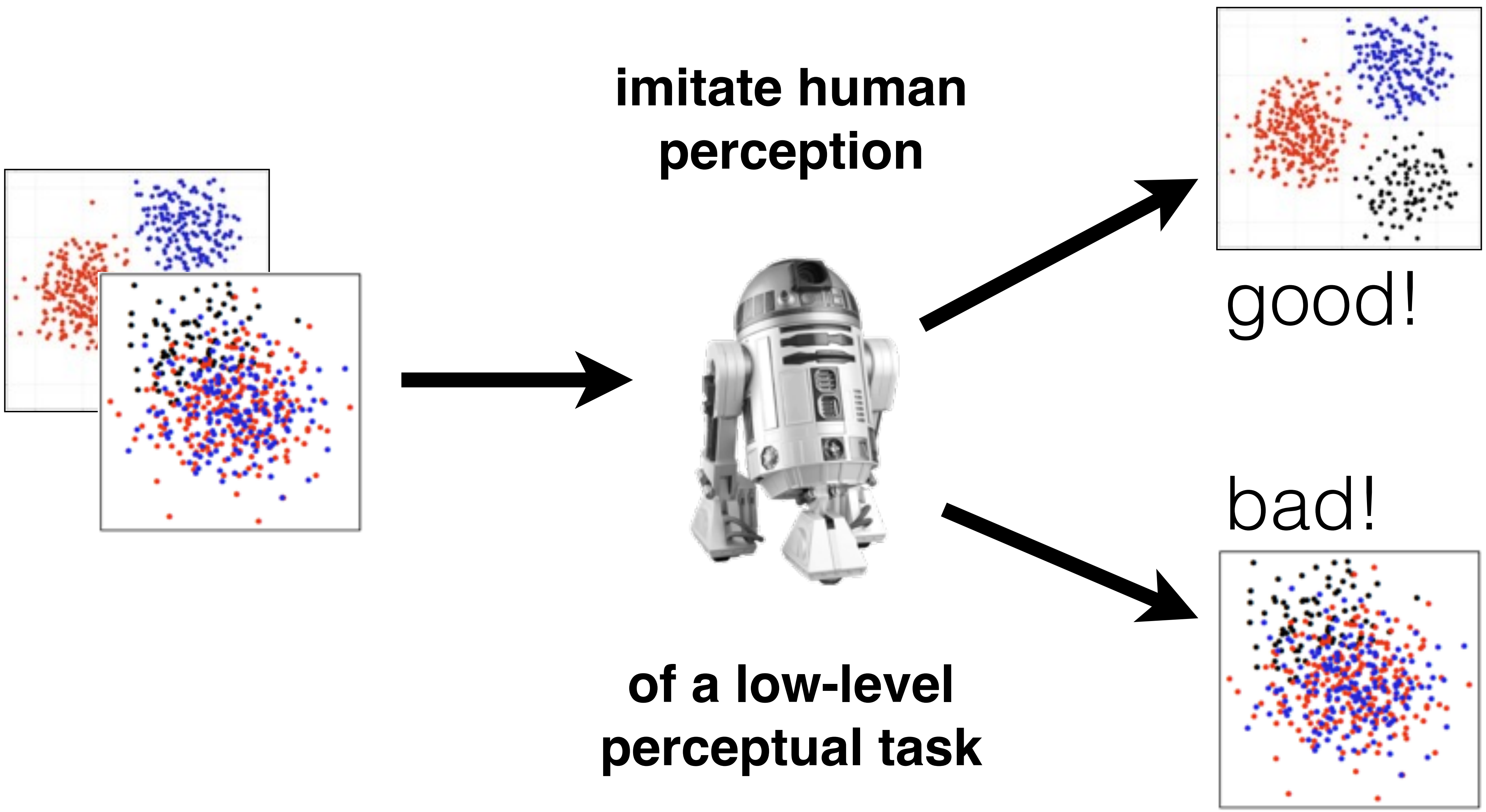
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

Visual Quality Measures

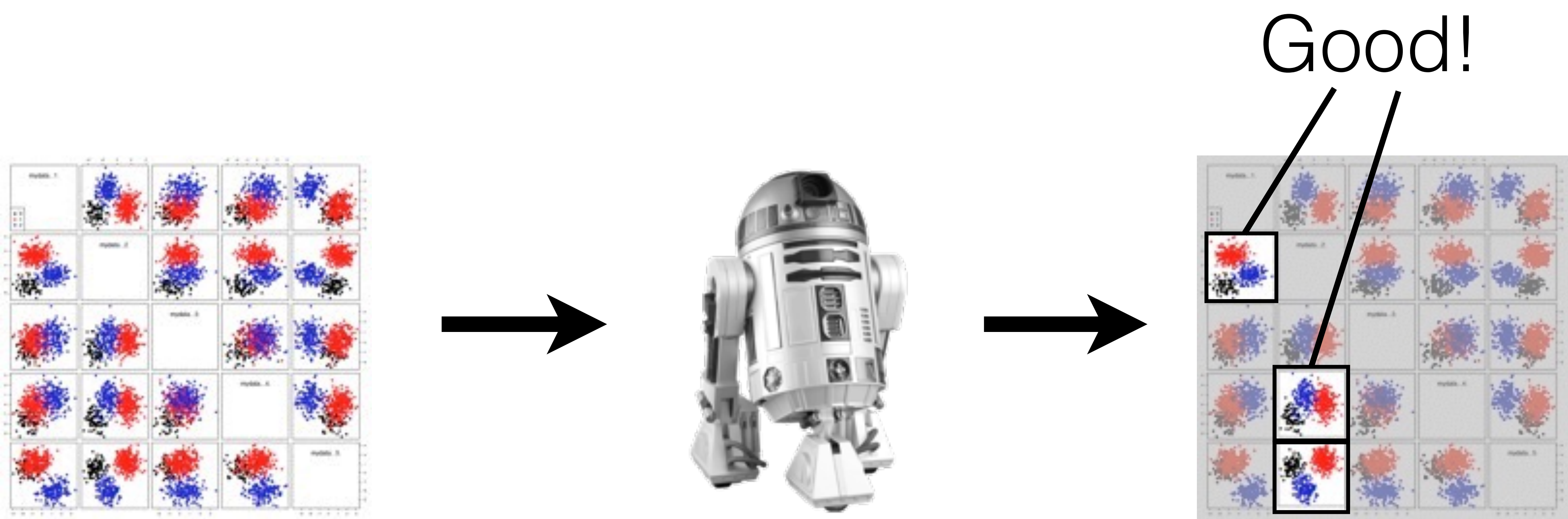
- Scagnostics
  [Wilkinson & Anand, 2005]
- Paragnostics
  [Dasgupta & Kosara 2010]
- Visual class separation measures
  [e.g., Sips et. al., 2009]
- Visual correlation measures
  [e.g., Tatu et. al., 2010]
- etc…

Visual Quality Measures

e.g., visual class separation measures



What can you do with them?



Measure evaluation

timeline

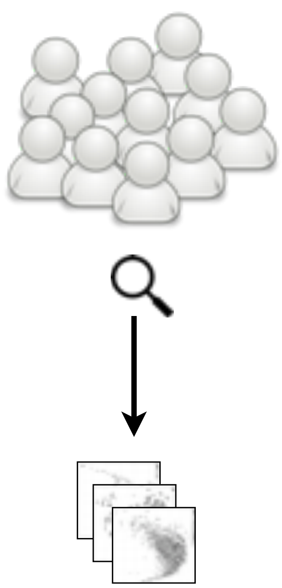
Sips et al. [EuroVis **2009**]

Tatu et al. [AVI **2010**]

Sedlmair et al. [EuroVis **2012**]

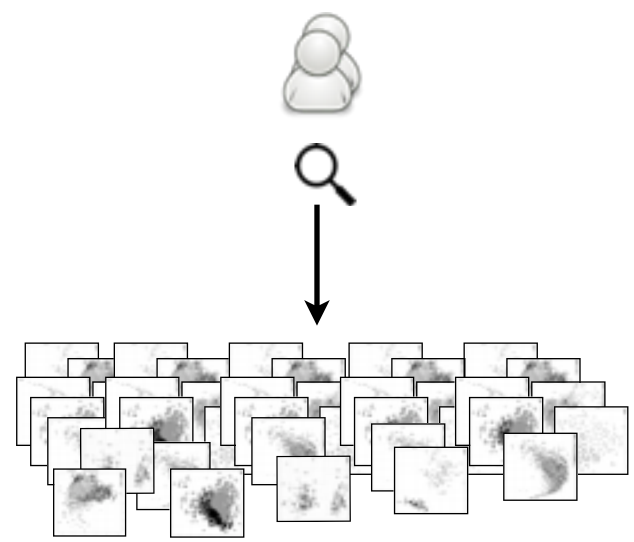
Today

User Study



1. Poor generalizability over dataset characteristics

(Manual) Data Study



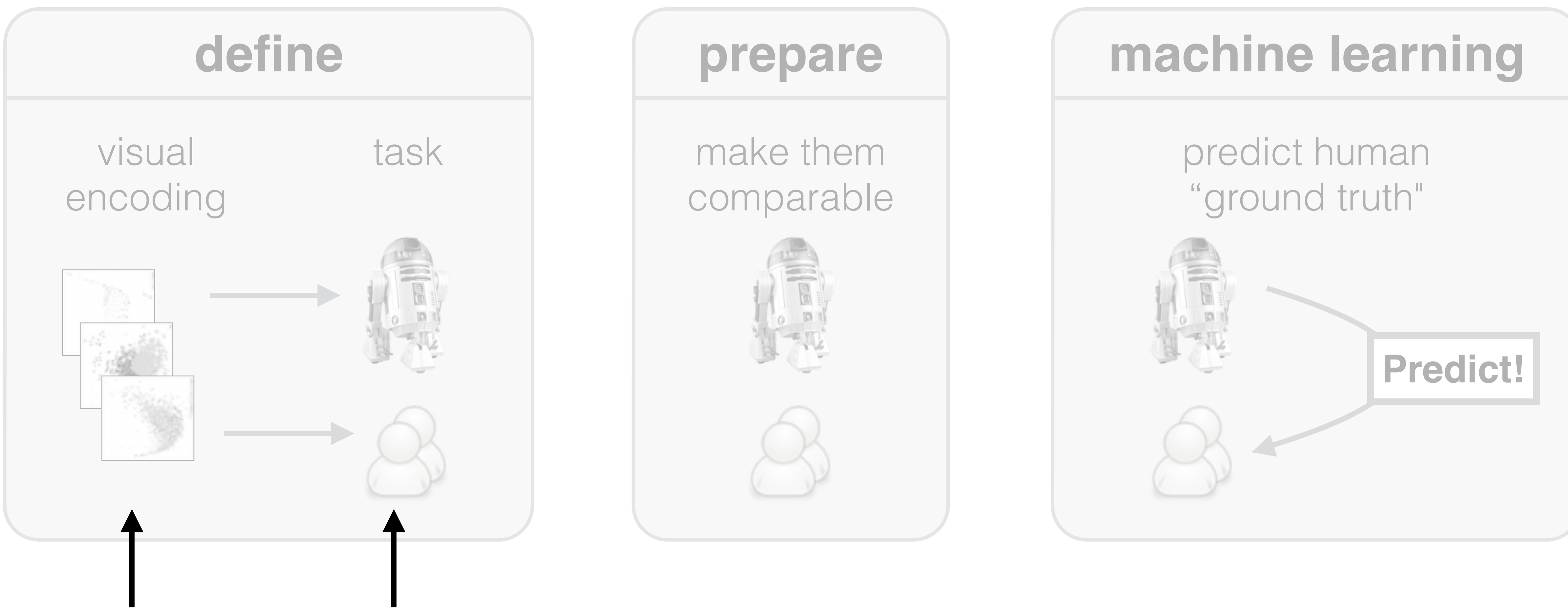
2. A lot of manual inspection work

Data-driven Evaluation

Contributions

- Framework for data-driven evaluation
- Instantiation of framework for class separation measures
- Evaluation of 15 class separation measures
- Guidelines for visual quality measure evaluation

Framework for Data-driven Evaluation (overview)



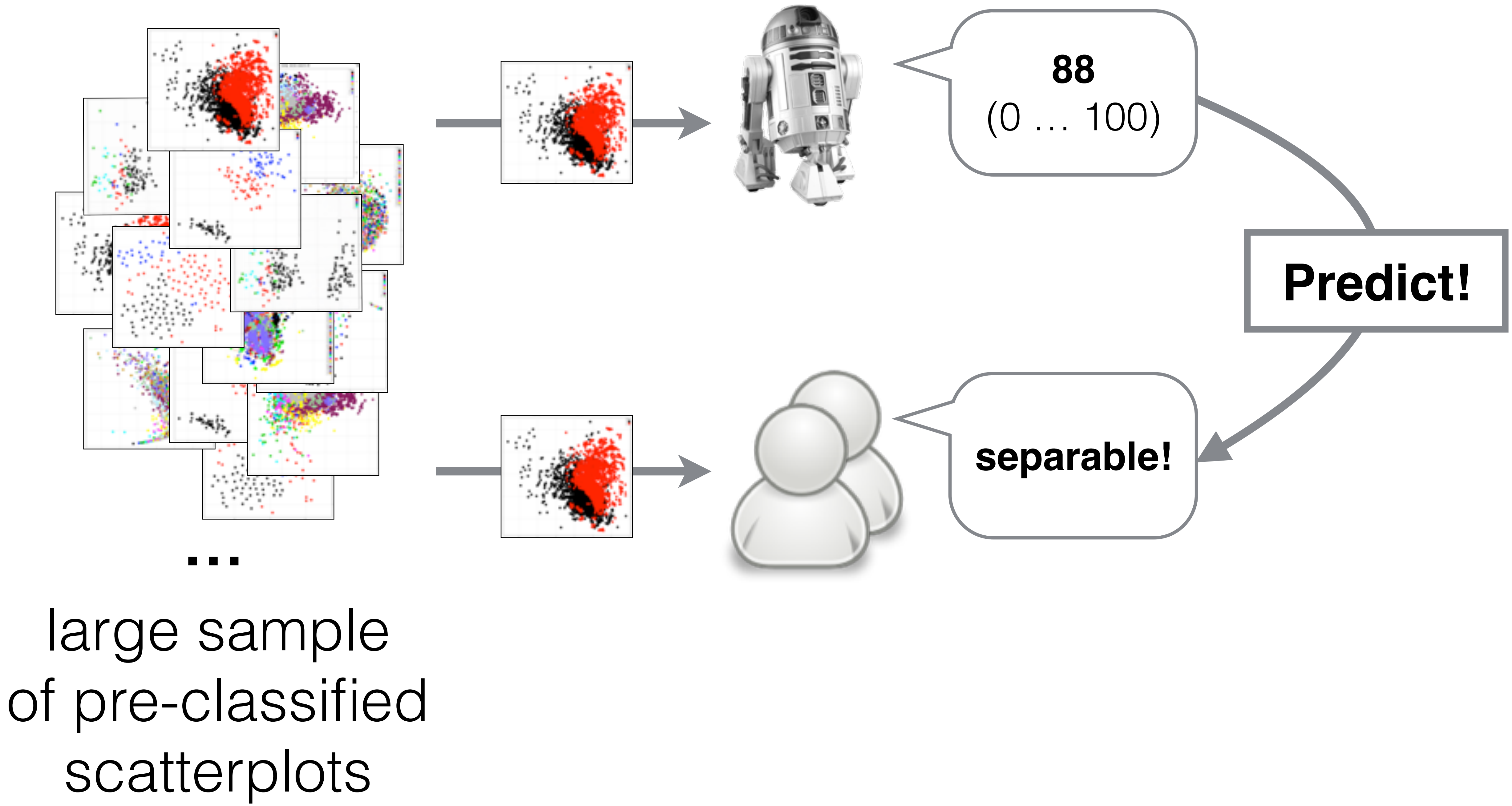
representative sample

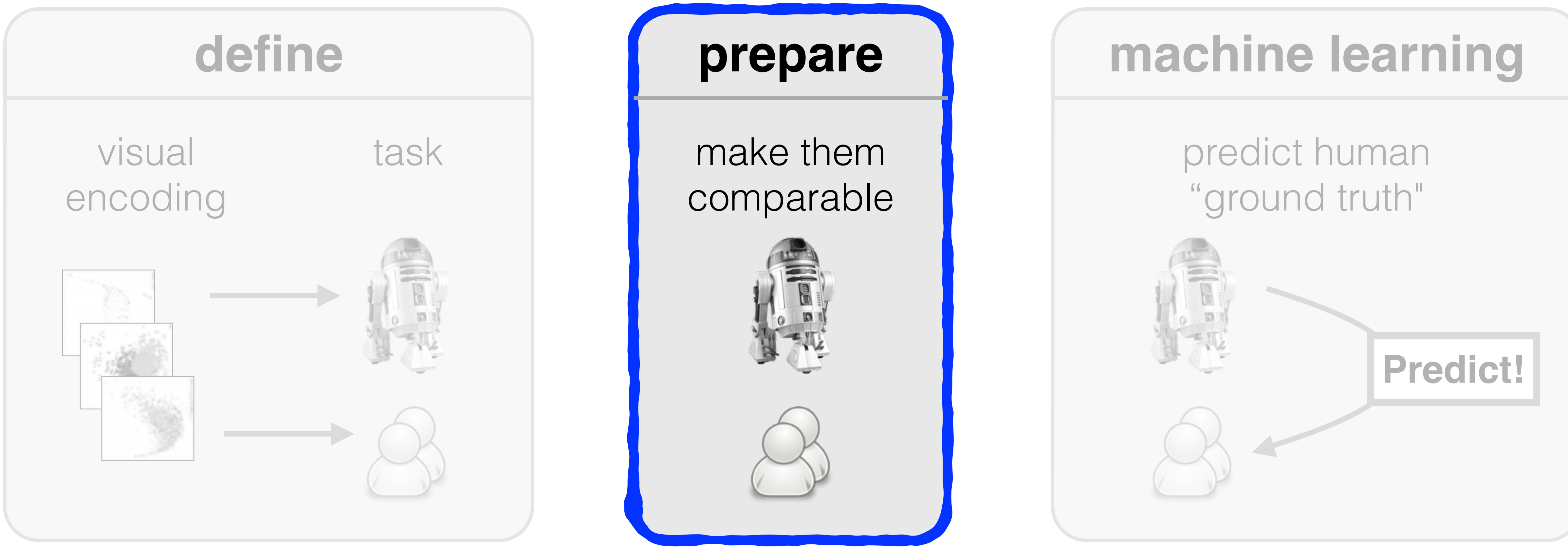
reliable judgments

1. **Evaluation based on how measures would perform on previously unseen data**
2. **Automatic evaluation**

Illustration
with
visual class separation

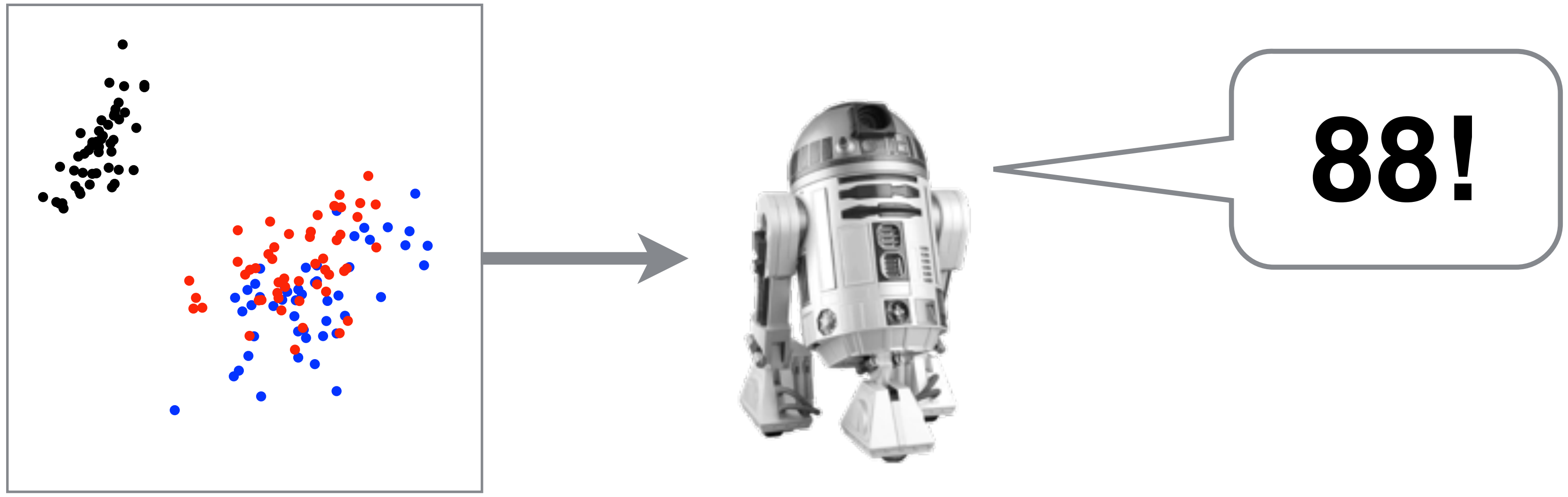
define the basic setting


88
(0 … 100)
Predict!
separable!
...
large sample
of pre-classified
scatterplots

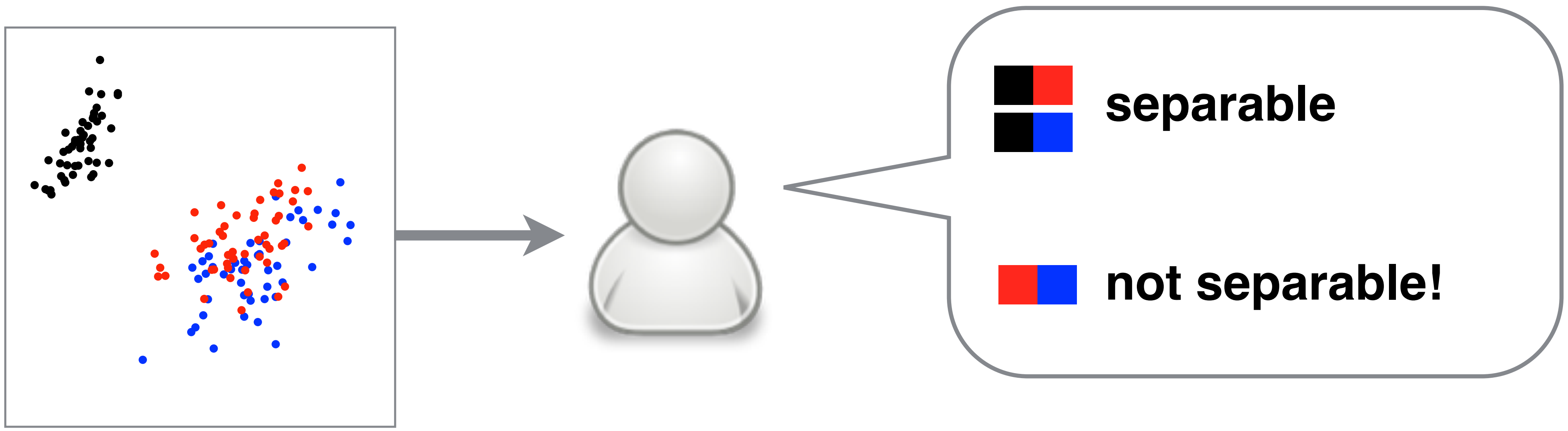


make them comparable

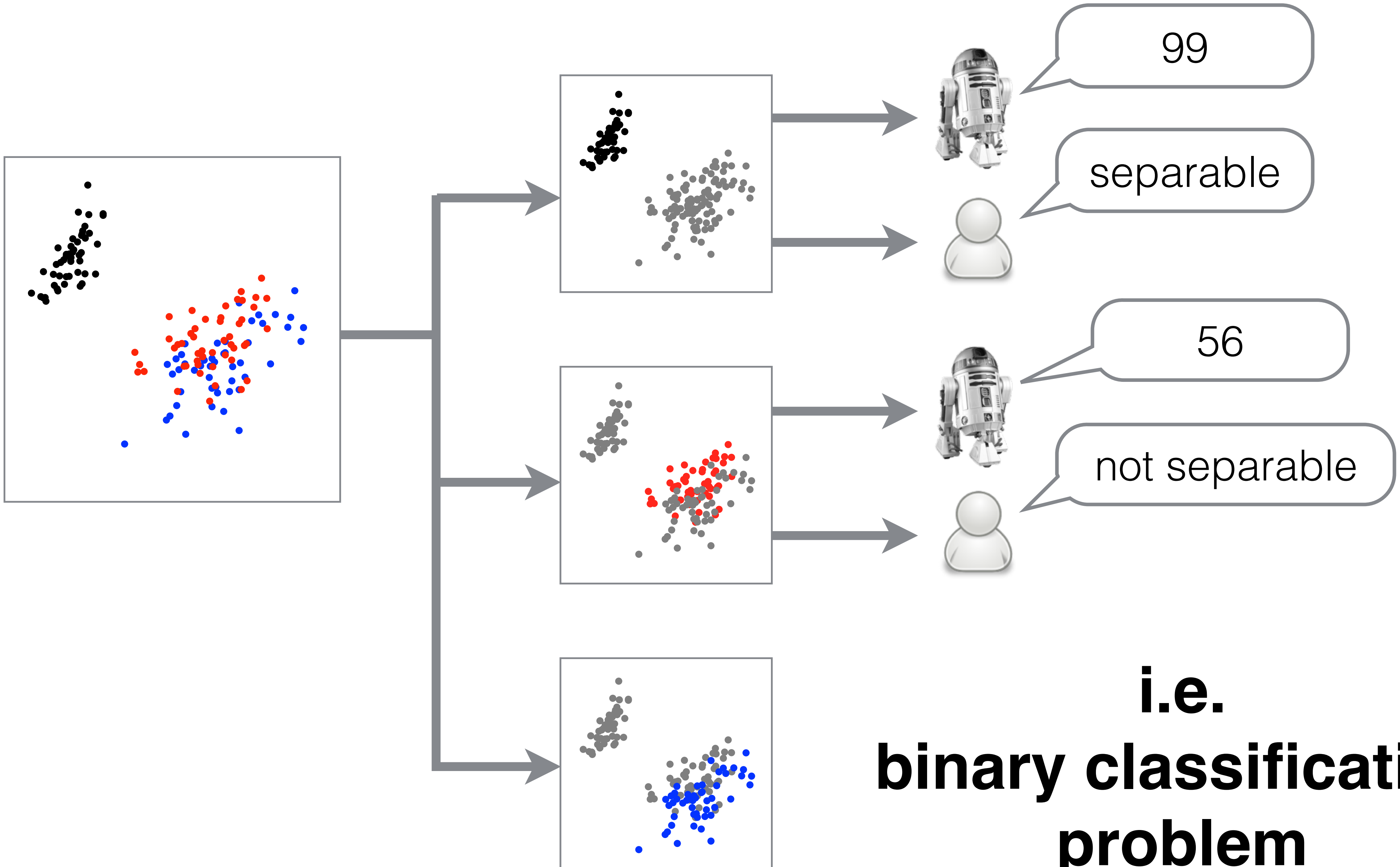
integrated judgments



class-wise judgments



1-vs-all scatterplots



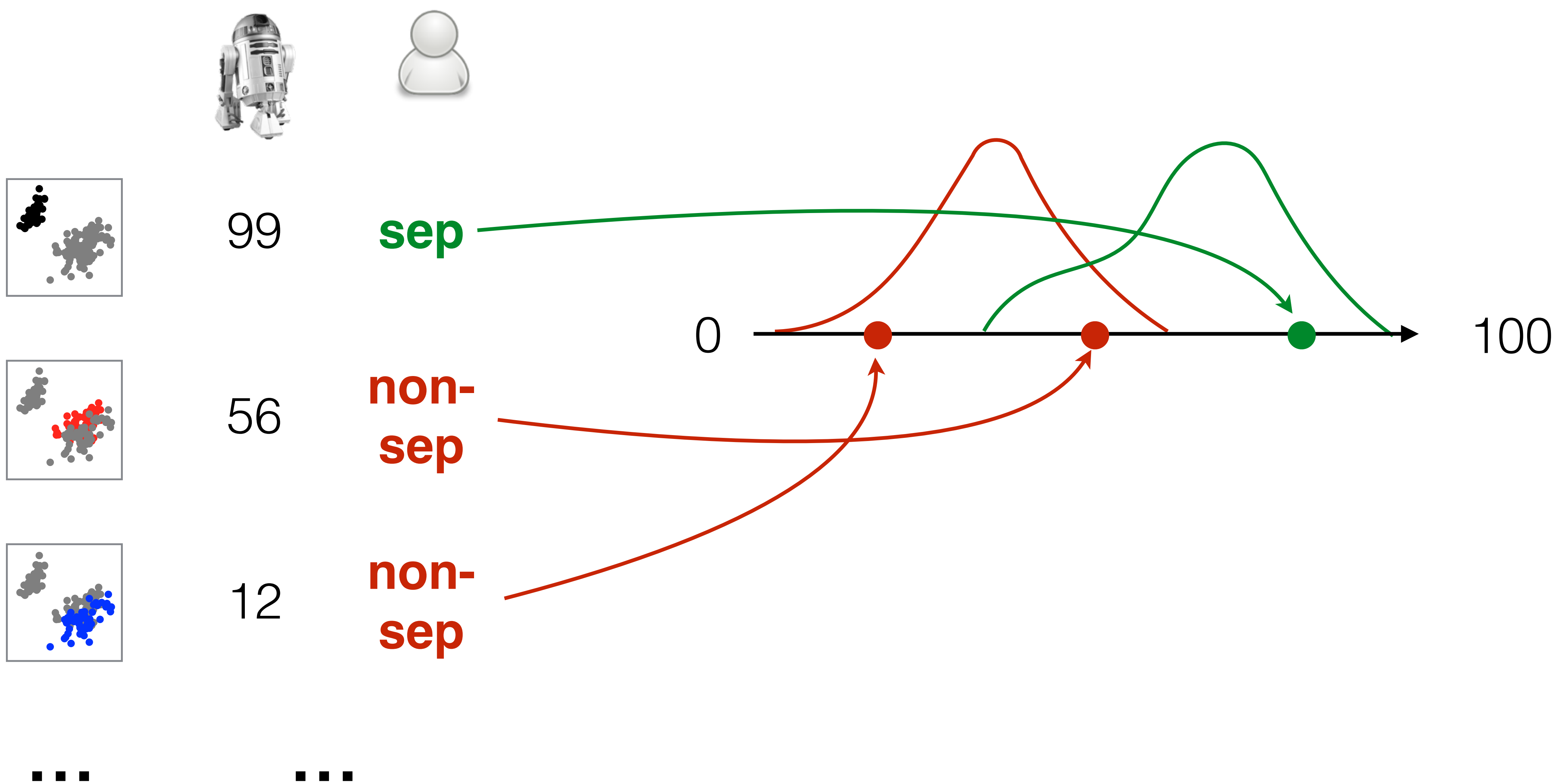
i.e.
binary classification problem

use ML pipeline

e.g., ROC & bootstrapping

ROC / AUC

(Receiver Operating Characteristic / Area Under the Curve)



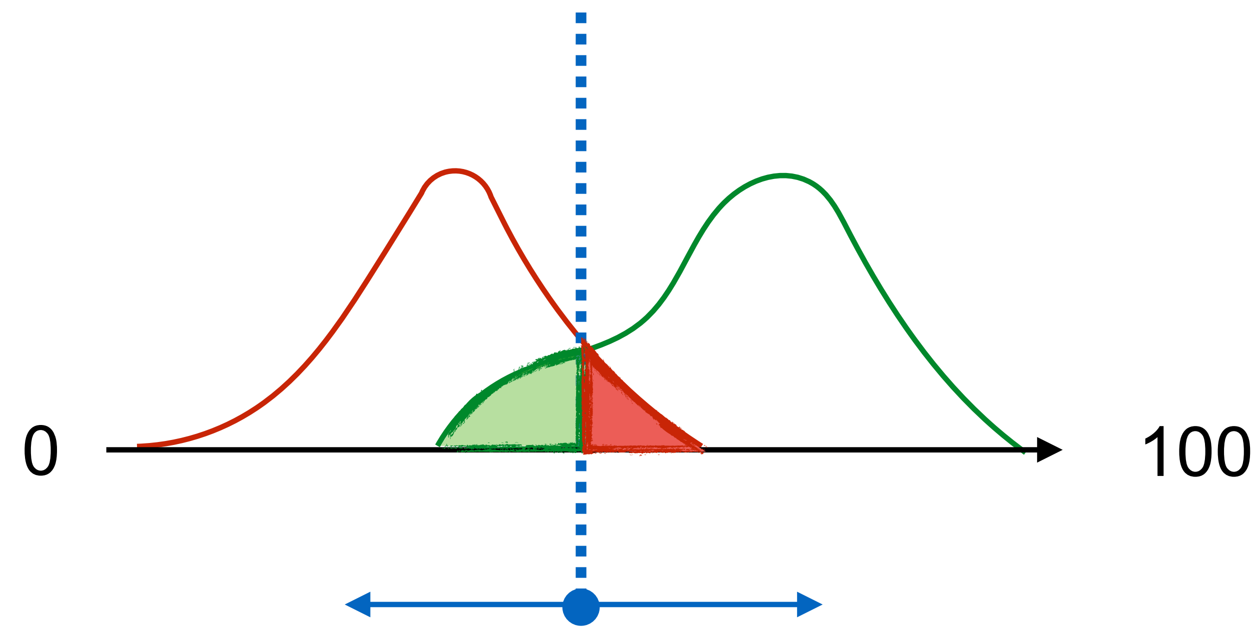
ROC / AUC

(Receiver Operating Characteristic / Area Under the Curve)

Decision Threshold?

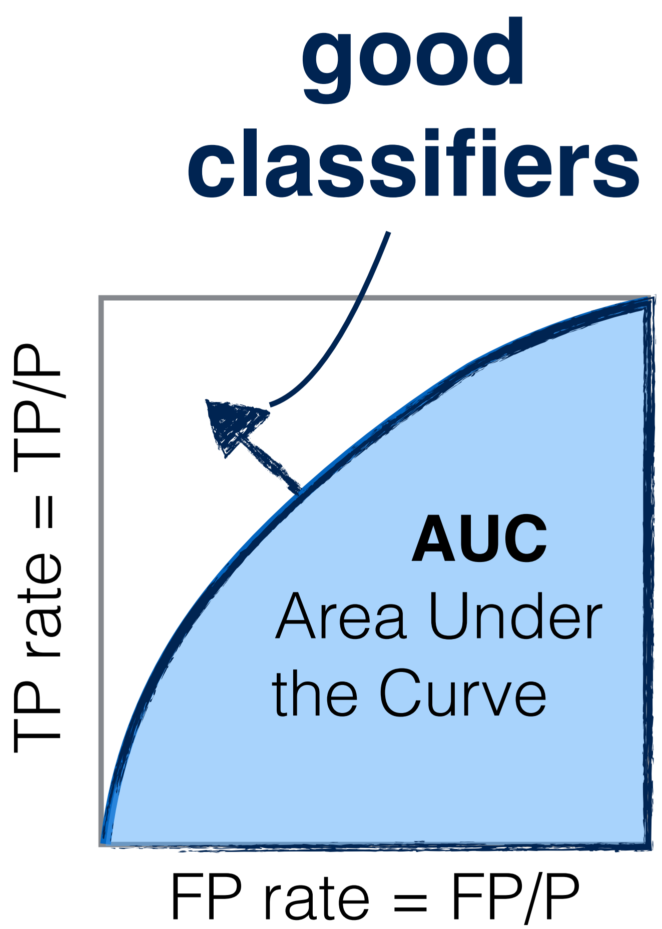
False Positives

False Negatives

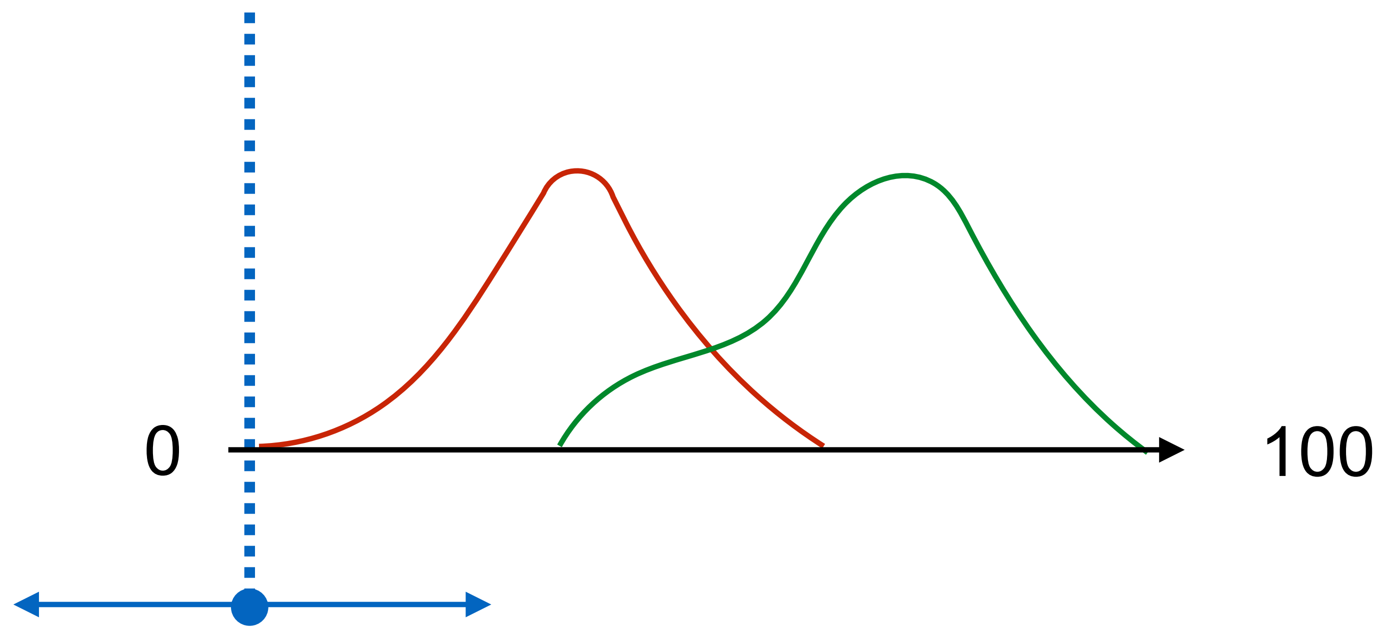


ROC / AUC

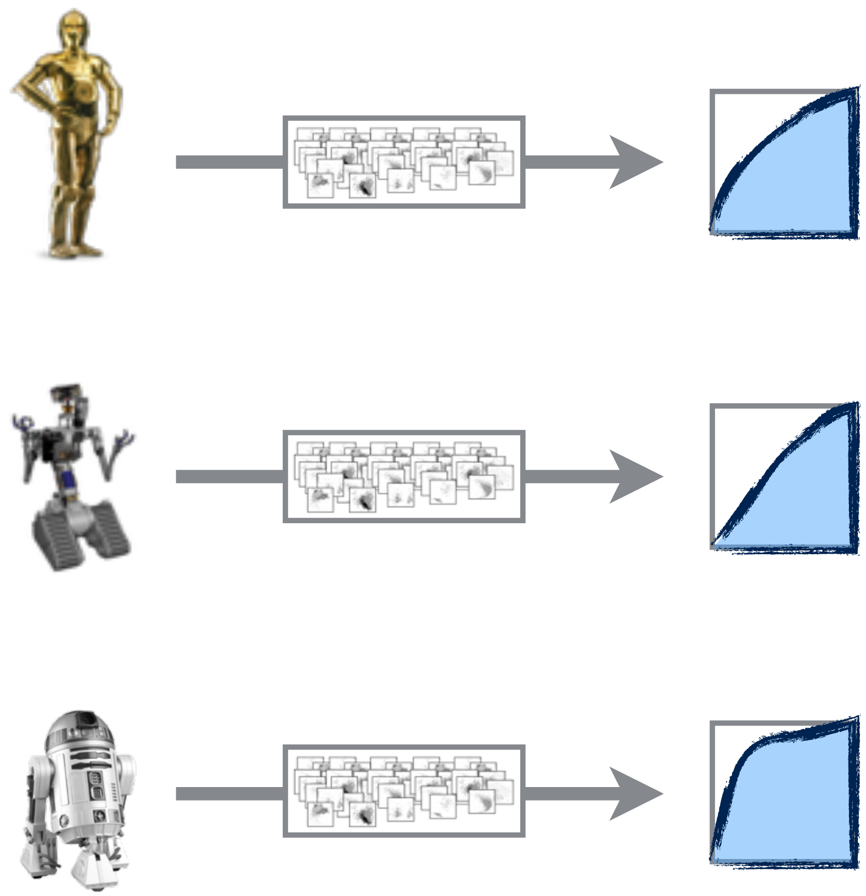
(Receiver Operating Characteristic / Area Under the Curve)



ROC Curve



Evaluate different measures …



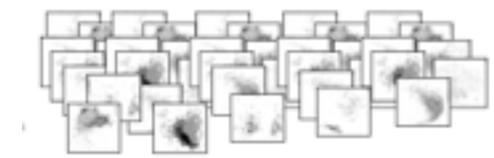
70%

55%

85%

…

Note: 50% means random guess

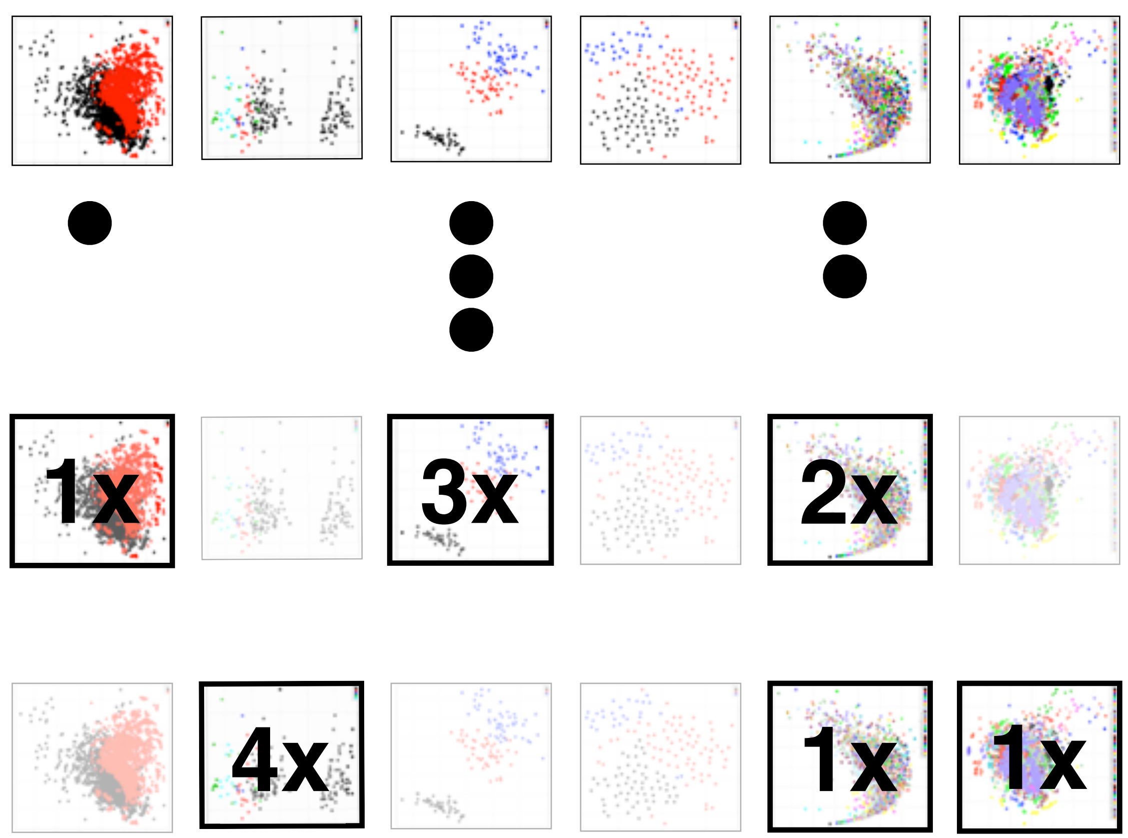


so far descriptive

initial goal:
predict performance on unseen data (inferential)

… cross validation
… bootstrapping

Bootstrapping



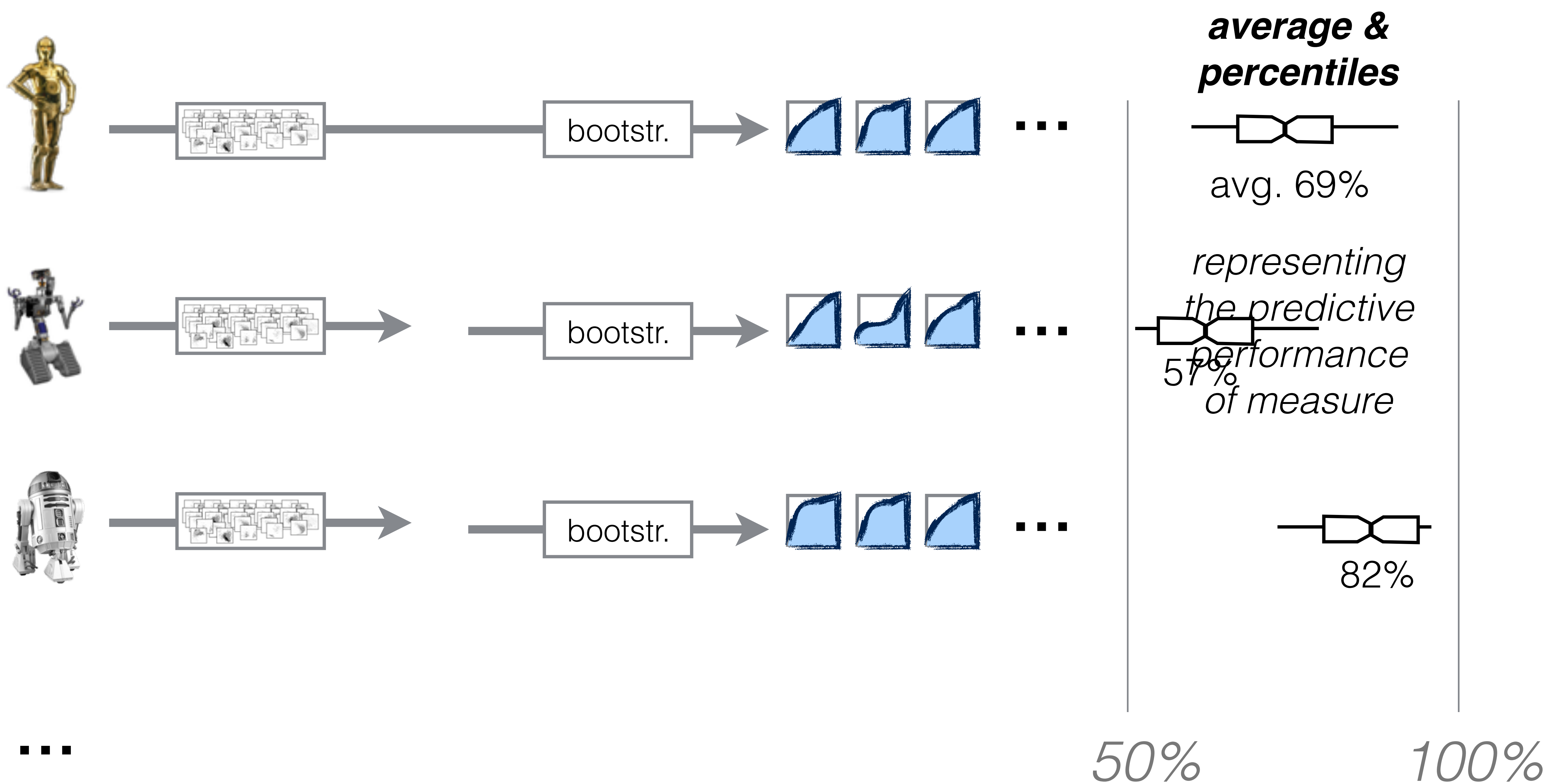
...

1. bootstrap sample

2. bootstrap sample

...

Evaluate different measures …

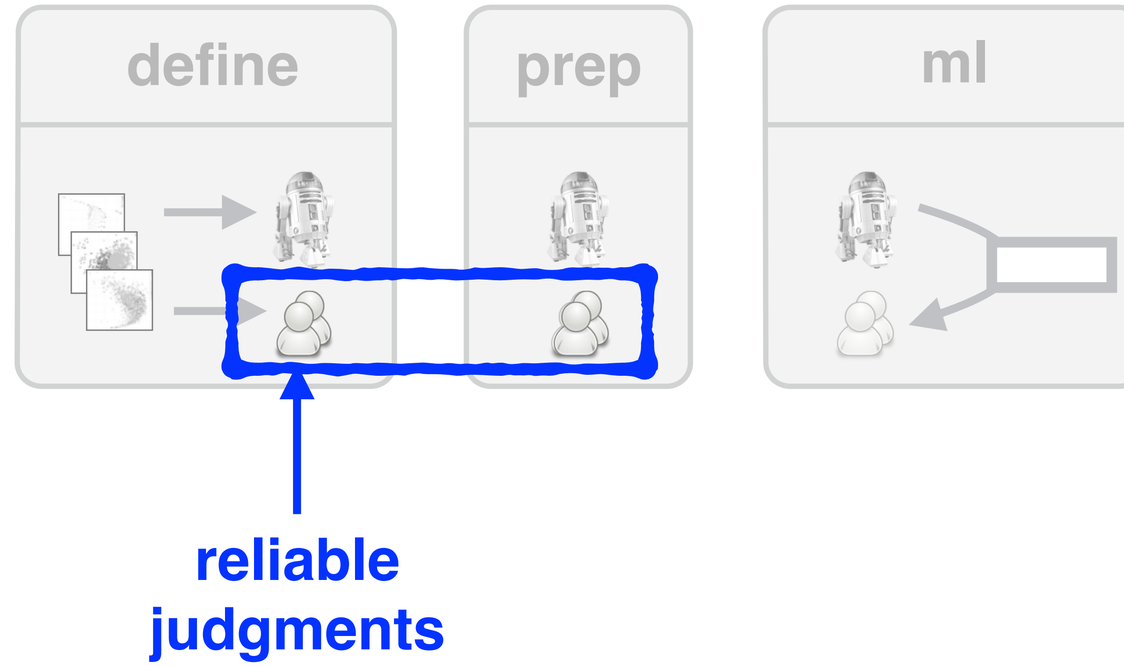


Evaluation of 15 measures

Setup:
Data

- from a previous study *[Sedlmair et al., InfoVis 2013]*
  - 272 pre-classified 2D scatterplots
  - from 75 real and synthetic datasets
- **828 1-vs-all scatterplots (420 real / 408 synthetic)**

Human Judgements



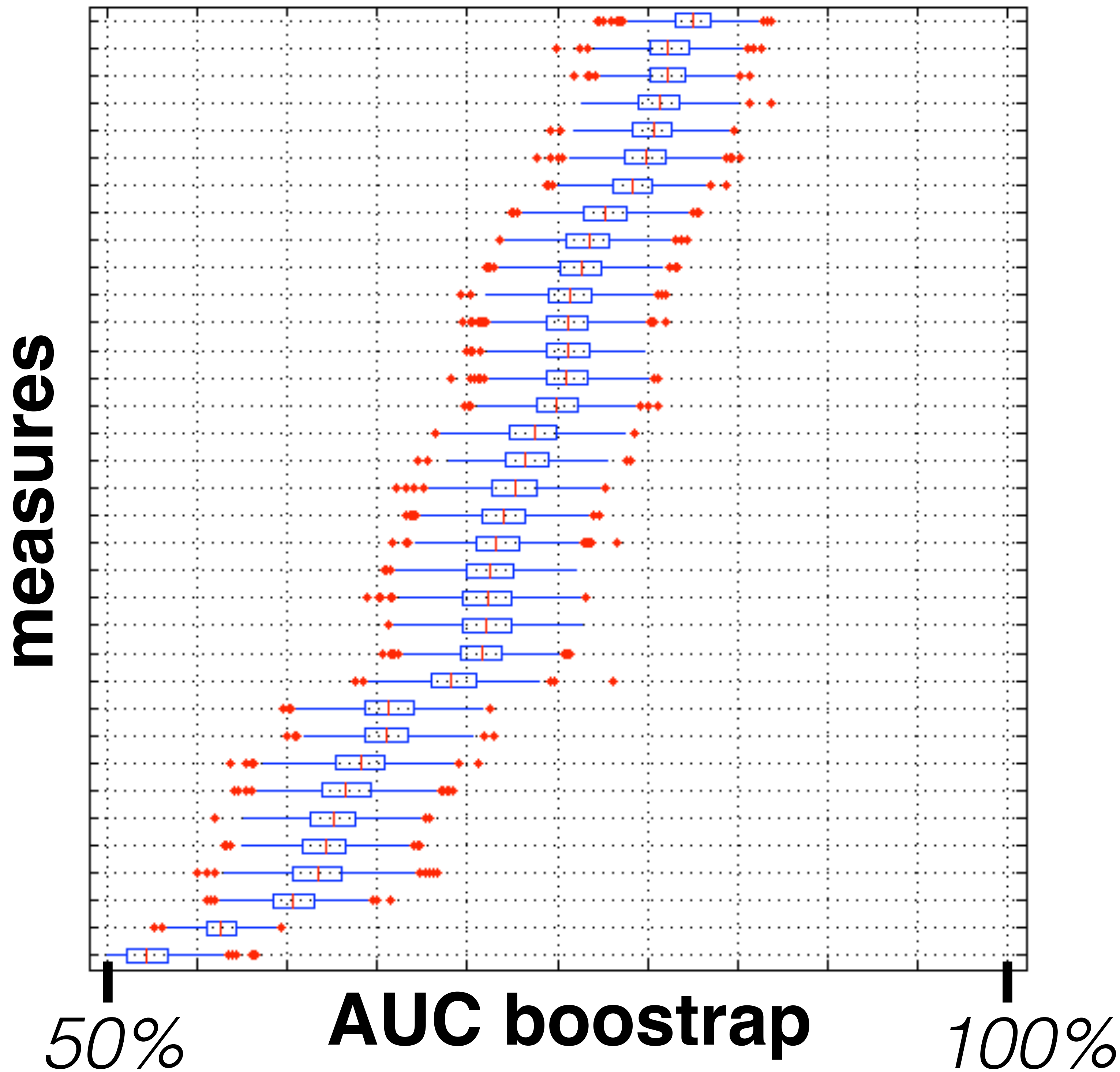
- from a previous study *[Sedlmair et al., InfoVis 2013]*
  - judged by two expert coders
  - 5-point scale: 1 - not separable … 5 - fully separable
  - Krippendorff's alpha = 0.85
- **Aggregation into binary judgments**
  - **(1,1) (1,2) (2,1) —> not separable**
  - **(5,5) (4,5) (5,4) —> separable**

Separation Measures

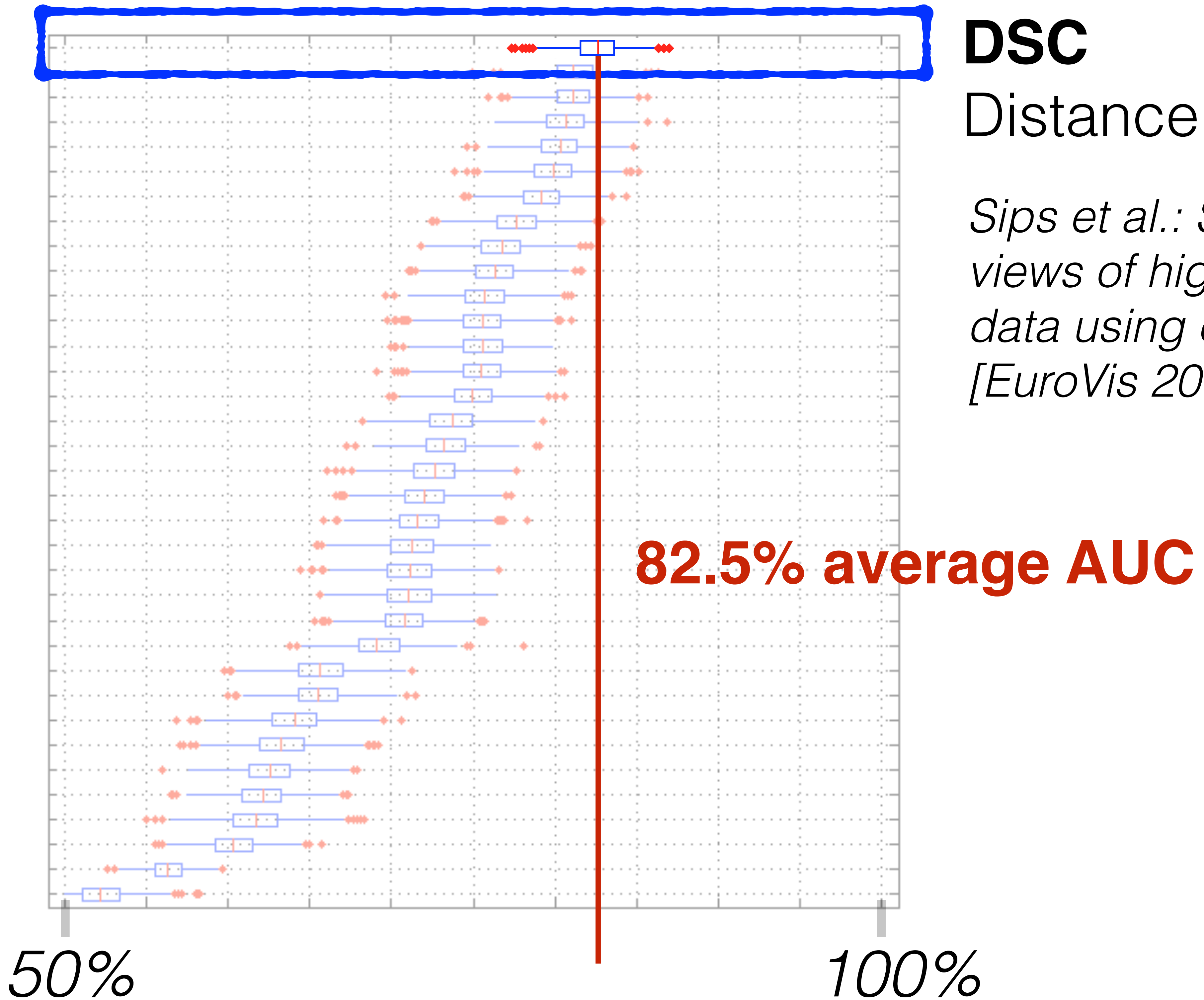
- 15 measures
  - from Visualization & ML community
  - 12 non-parametric
  - 3 parametric (different parameterization)
- **35 measure instances**

→ **10.000 bootstrap samples**

Results



The Winner

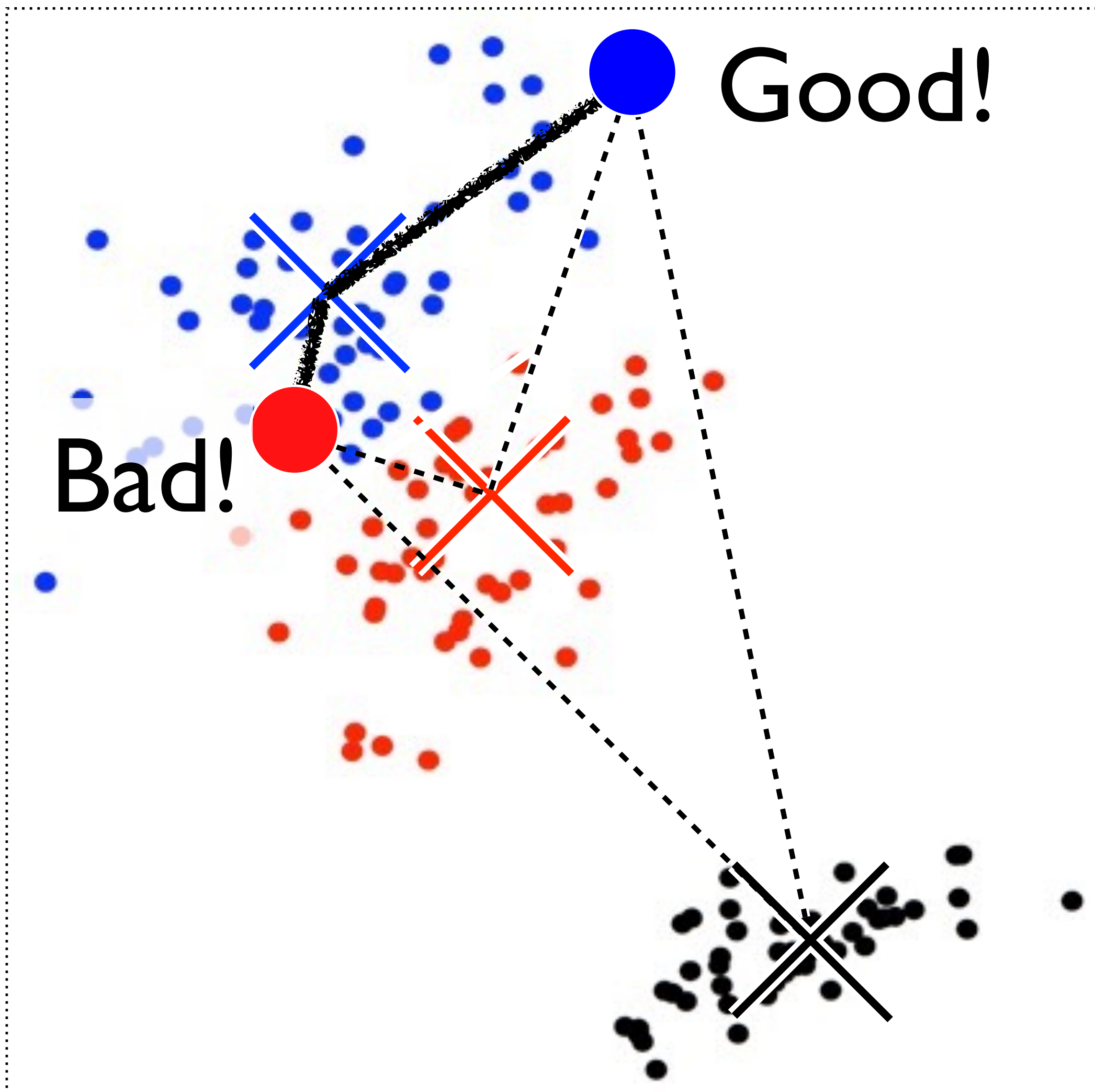


DSC

Distance Consistency

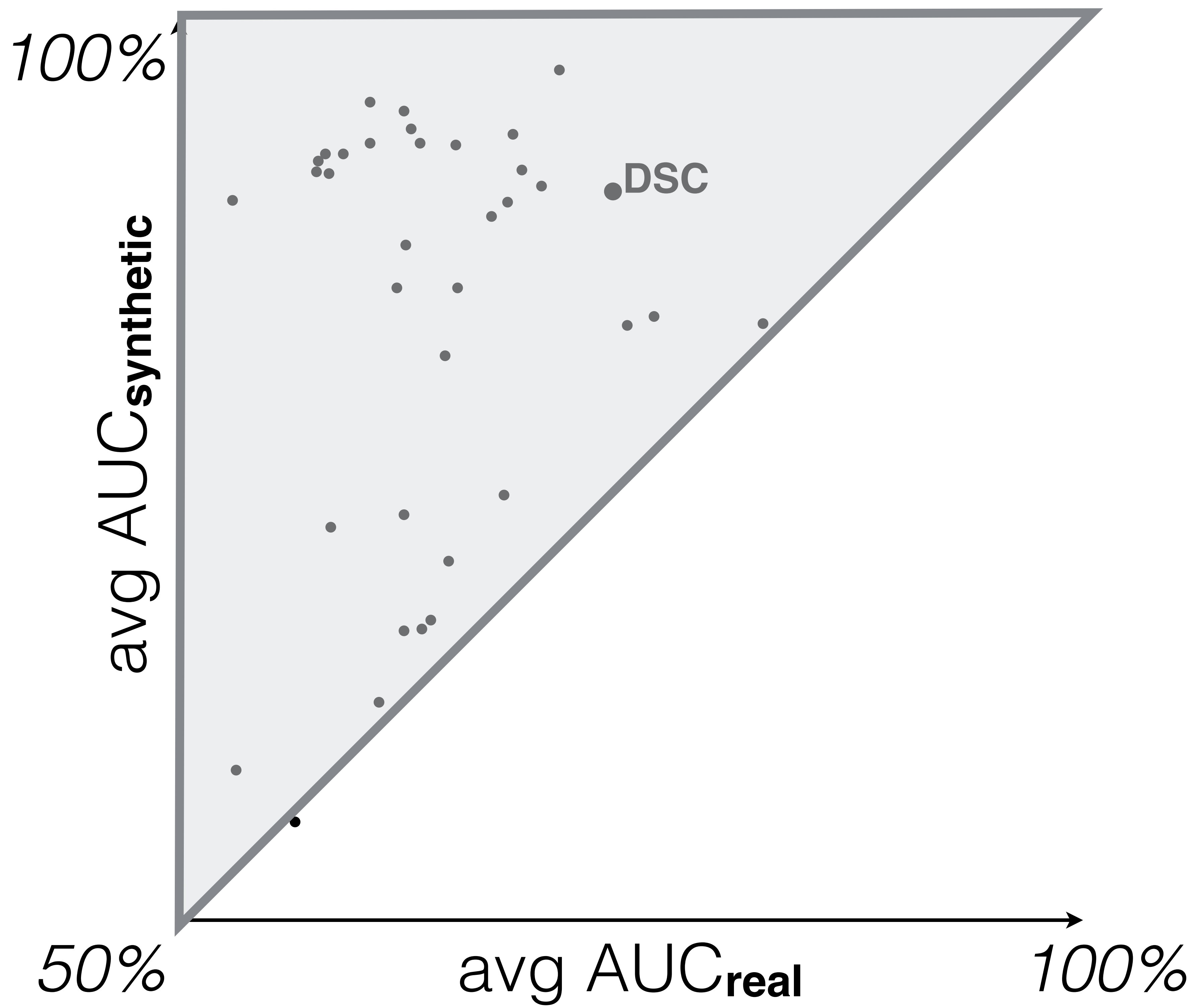
Sips et al.: Selecting good views of high-dimensional data using class consistency [EuroVis 2009]

DSC (Distance Consistency)

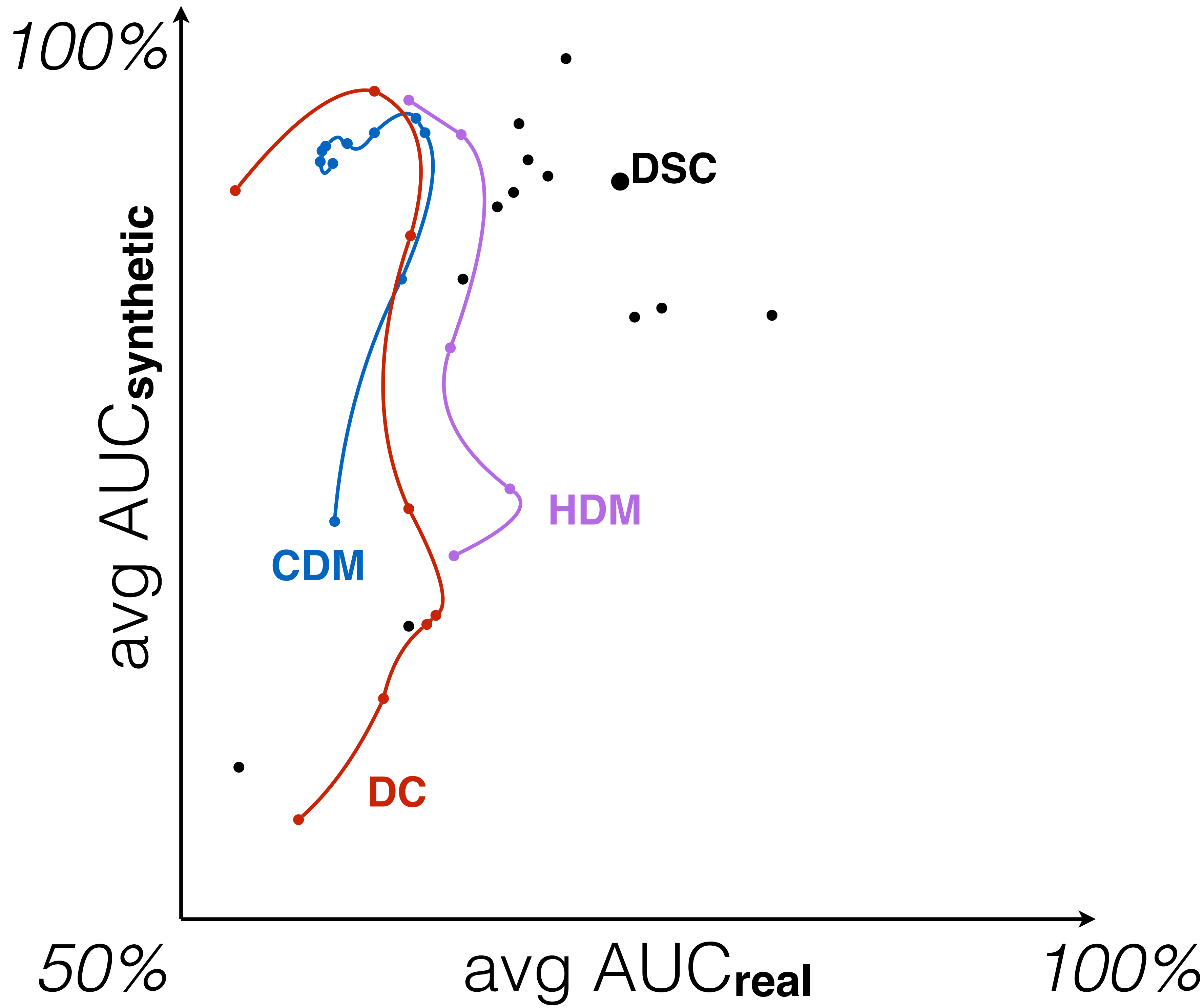


82.5% average AUC
… but still ~20%
room for improvement

Synthetic vs. real



Parameterizations



Discussion

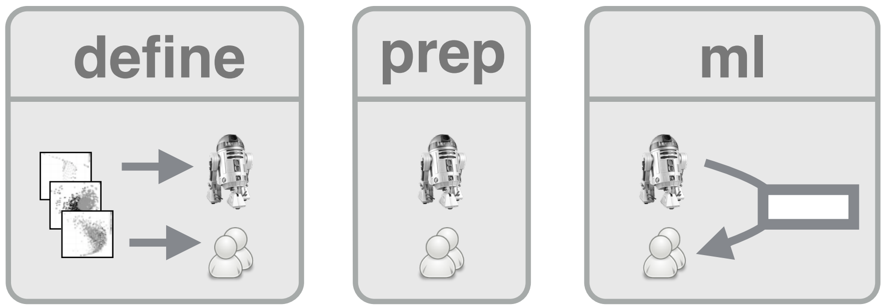
Guidelines

- Generalize over datasets
- Separate human judgment studies from measure studies (reuse instead of redo!)

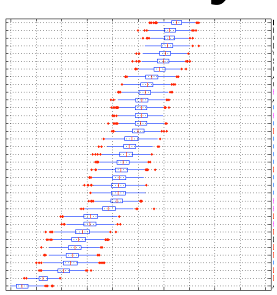
Data-driven evaluation framework,
not just for visual separation measures

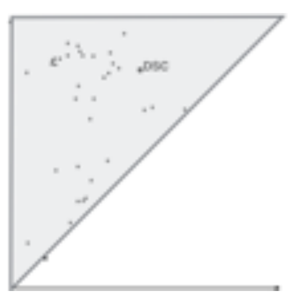
Summary

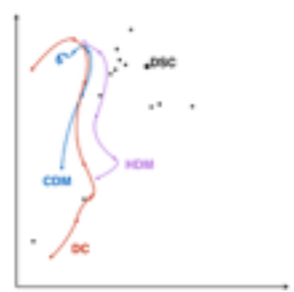
- Framework for data-driven evaluation



- Study of 15 measures: DSC —> 82,5% AUC avg.







- Guidelines: Generalization & Separation

Data-driven Evaluation of Visual Quality Measures

Michael Sedlmair & Michaël Aupetit

Thanks!