PR-SOCO

Personality Recognition in SOurce COde

PAN@FIRE 2016

Kolkata, 8-10 December

Francisco Rangel
Autoritas Consulting

Fabio A. González & Felipe Restrepo-Calle
MindLab - Universidad Nacional Colombia

Manuel Montes
INAOE - Mexico

Paolo Rosso
PRHLT - Universitat Politècnica de Valencia - Spain

Introduction

Author profiling aims at identifying **personal traits** such as age, gender, native language or **personality traits** from writings.

This is crucial for:

- Marketing
- Security
- Forensics

Task goal

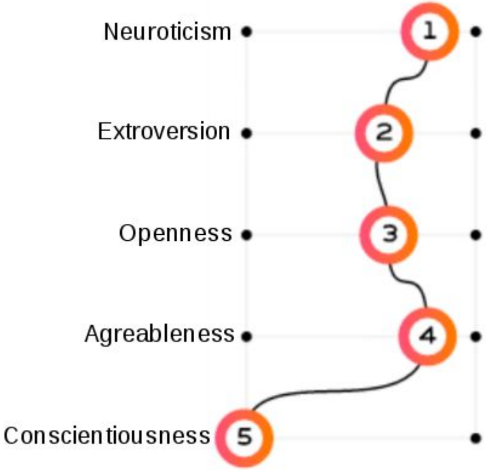
To predict **Personality Traits** from **Source Codes**.

This is crucial for:

- Human resources management for IT departments.

Corpus

- Java programs by computer science students at Universidad Nacional de Colombia
- Allowed:
  - Multipe uploads of the same code
  - Errors (compiler output, debug information, source codes in other languages such as Python…)



SOURCE CODES	**2,492**
AUTHORS	**70**
TRAINING	**TEST**
49	**21**

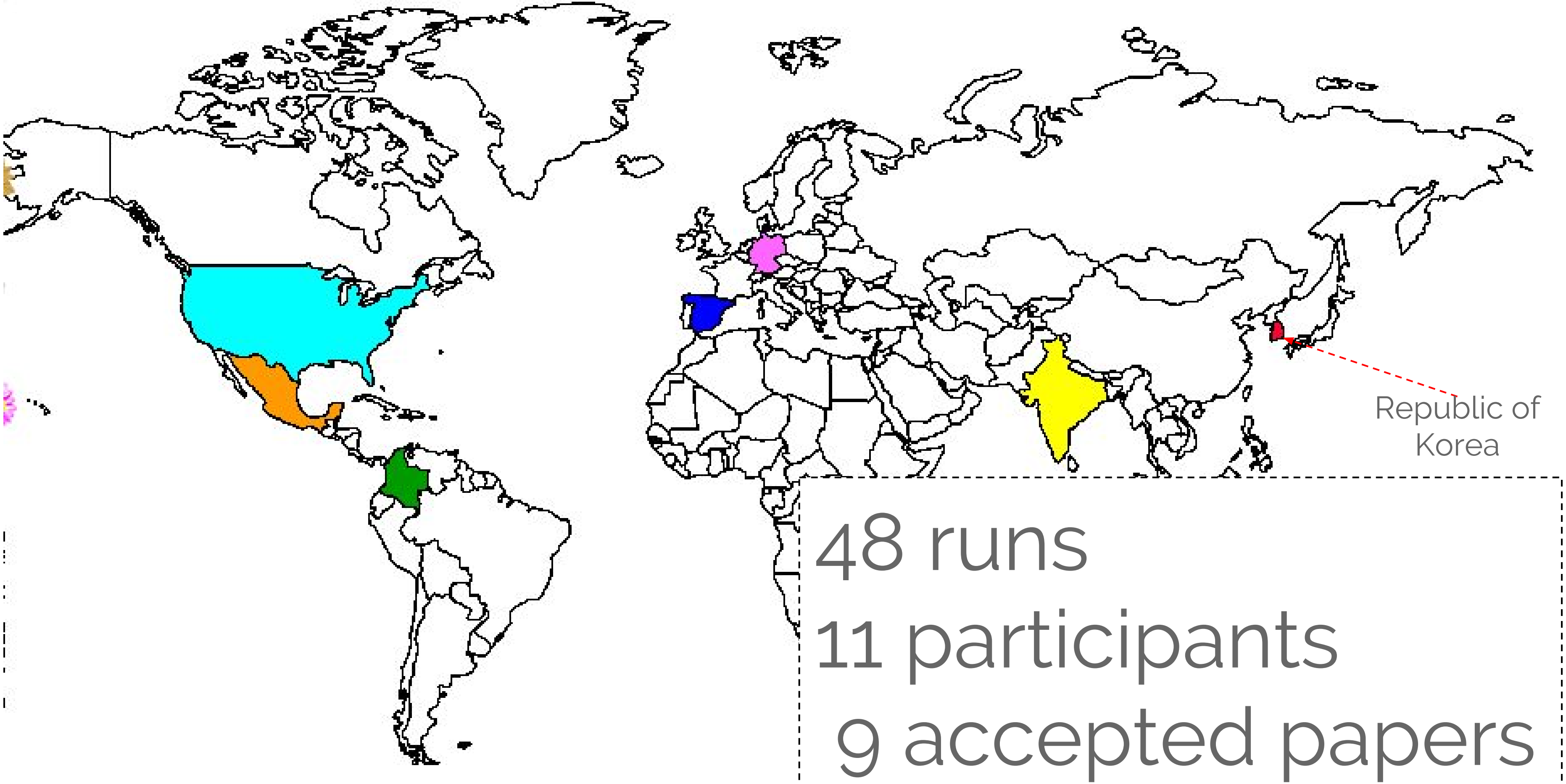
Evaluation measures

Two complementary measures per trait:

- Root Mean Squared Error to measure the goodness of the approaches.
- Pearson Product-Moment Correlation to measure the random chance effect.

$$RMSE_t = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$



48 runs

11 participants

9 accepted papers

7 countries

Approaches - Features

Bag of Words, word n-gams or char n-grams	Besumich, Gimenez, Besumich
Word vectors (skip-thought encoding)	Lee
Byte streams	Doval
ToneAnalyzed	Montejo
Code structure (ANTLR syntax)	Bilan, Castellanos
Specific features related to coding style - Length of the program, length of the classes... - Average length of variable names, class names… - Number of methods per class, ... - Frequency of comments and length - Identation, code layout, …	Bilan, Delair, Gimenez, HHU, Kumar, Uaemex
Halstead metrics (software engineering metrics)	Castellanos

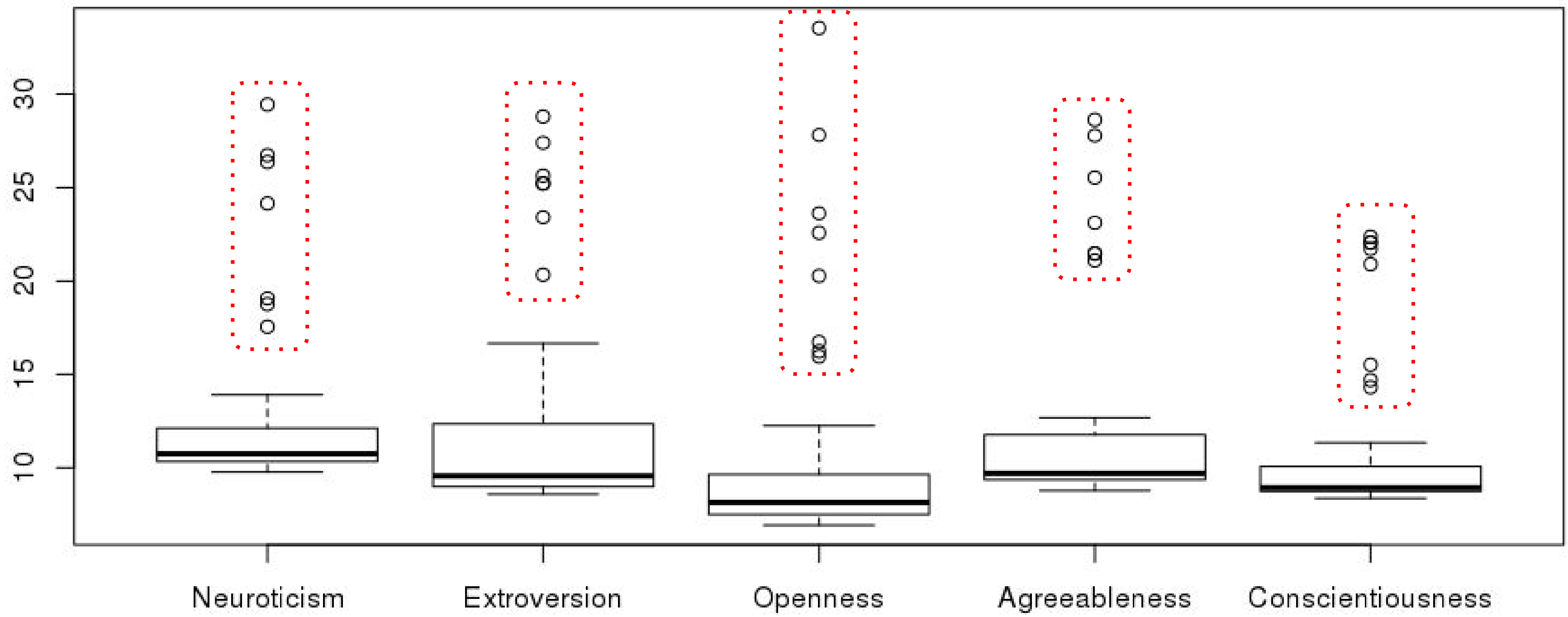
+ 2 baselines: char 3-grams and the observed mean.

Approaches - Methods

Logistic regression	Lee, Gimenez
Lasso regression	Besumich
Support vector regression	Castellanos, Delair, Uaemex
Extra trees regression	Castellanos
Gaussian processes	Delair
M5, M5 rules	Delair
Random trees	Delair
Neural networks	Doval, Uaemex
Linear regression	HHU, Kumar
Nearest neighbour	HHU, Uaemex
Symbolic regression	Uaemex

RMSE distribution

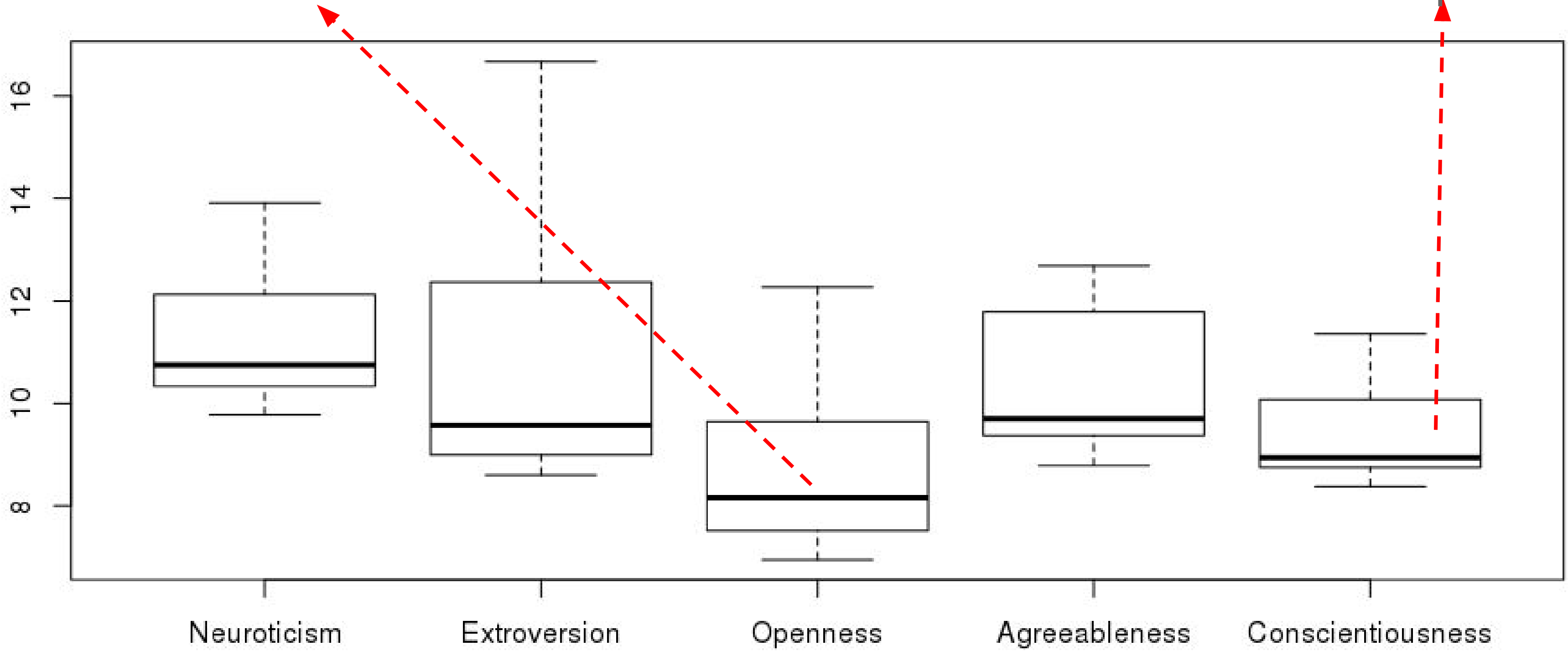
Too many outliers with poor performance...



RMSE distribution (without outliers)

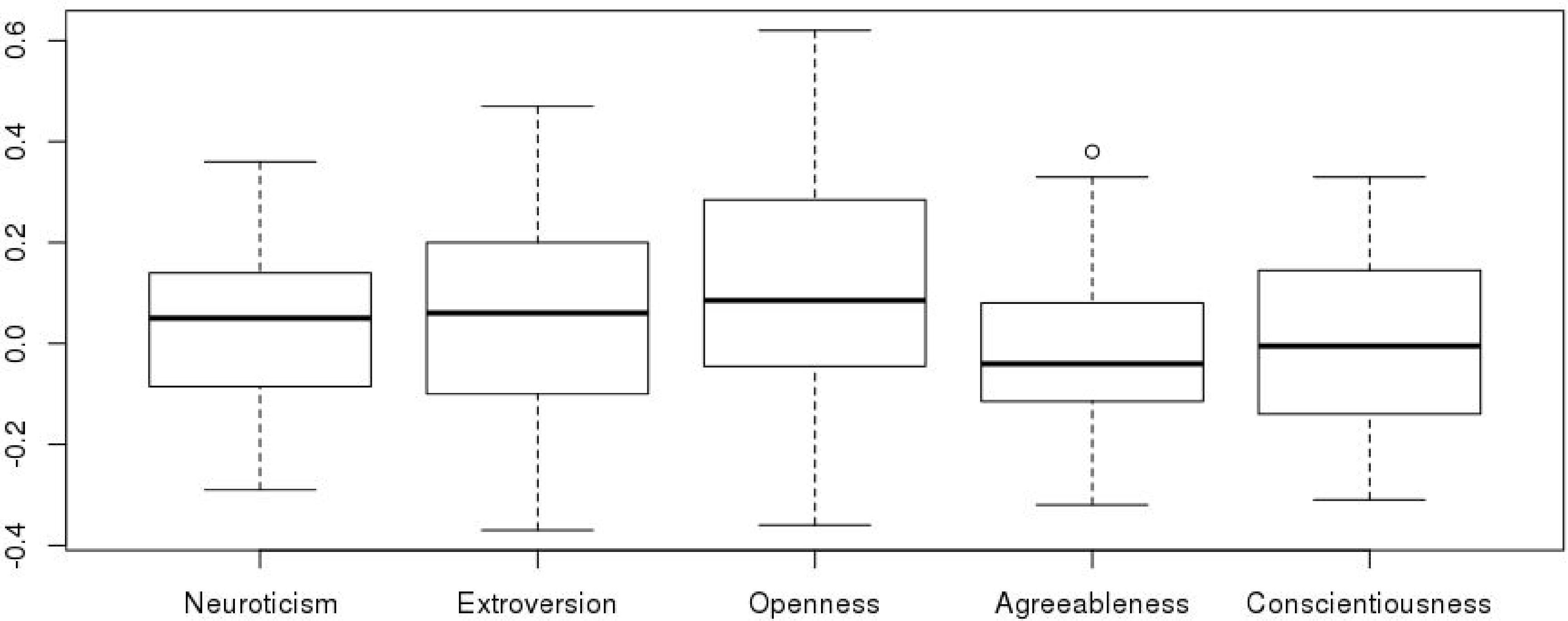
The best results (state of the art)

The lowest sparsity

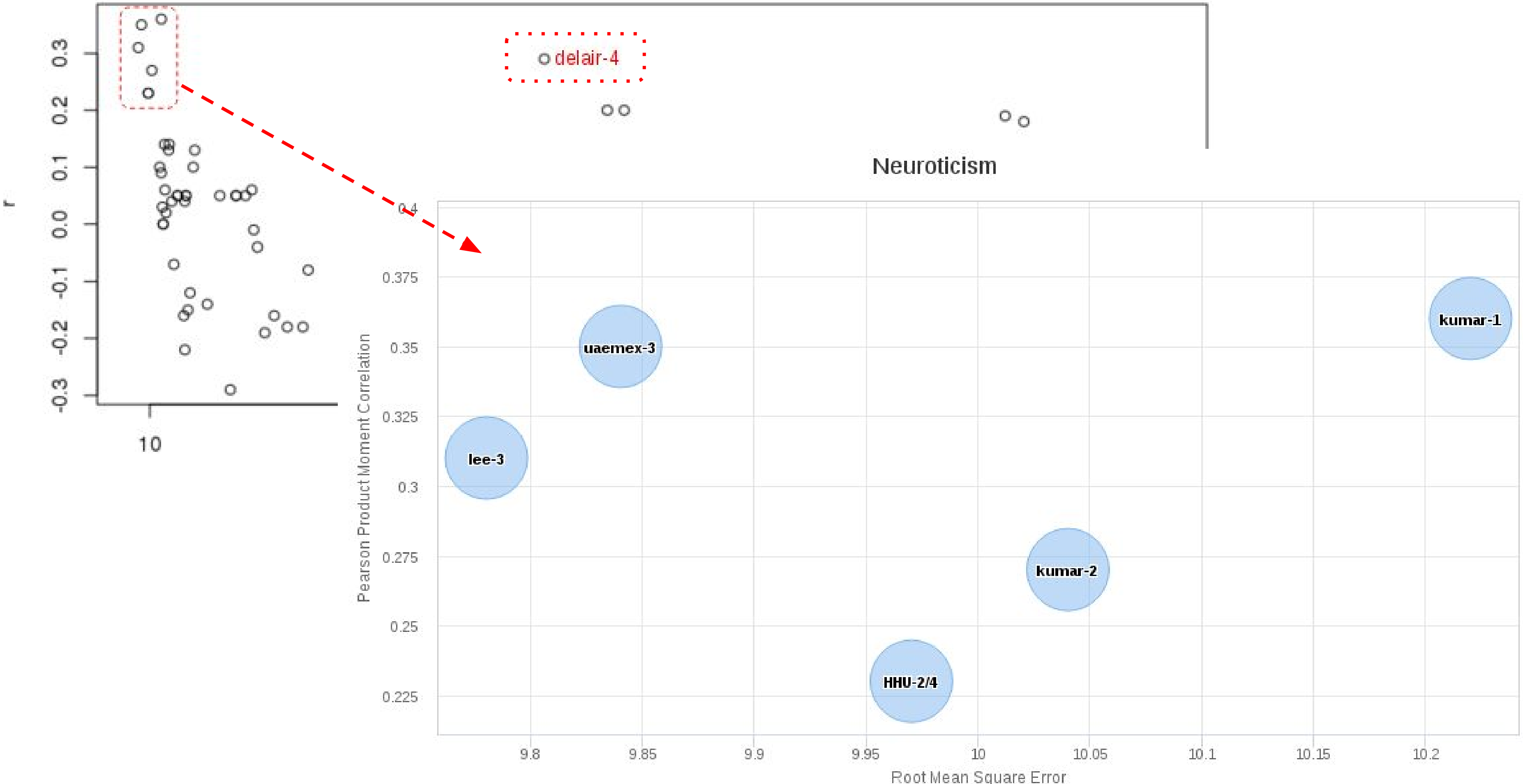


Pearson distribution

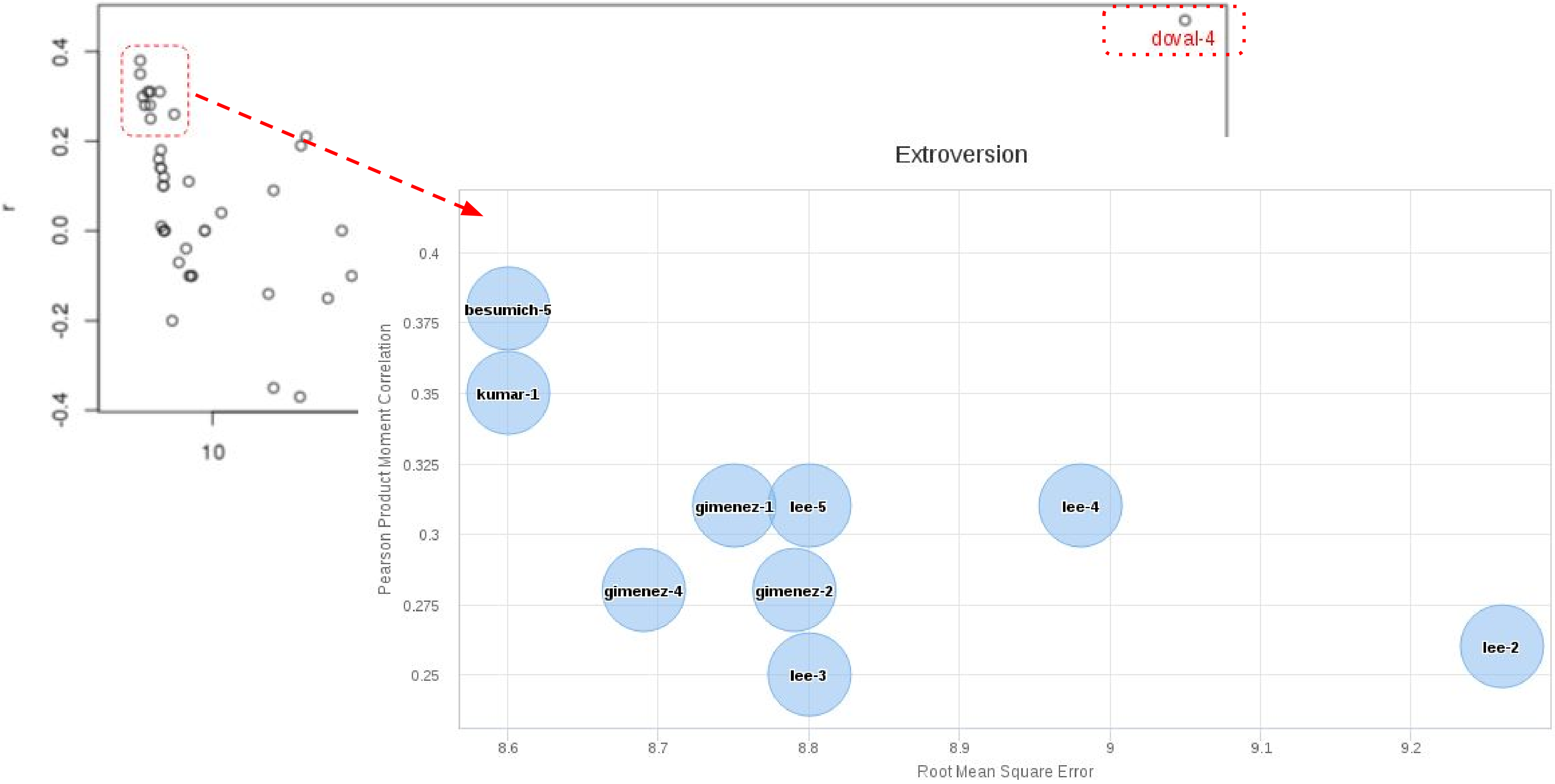
- Results much similar than for RMSE
- The average value is poor (lower than 0.3)



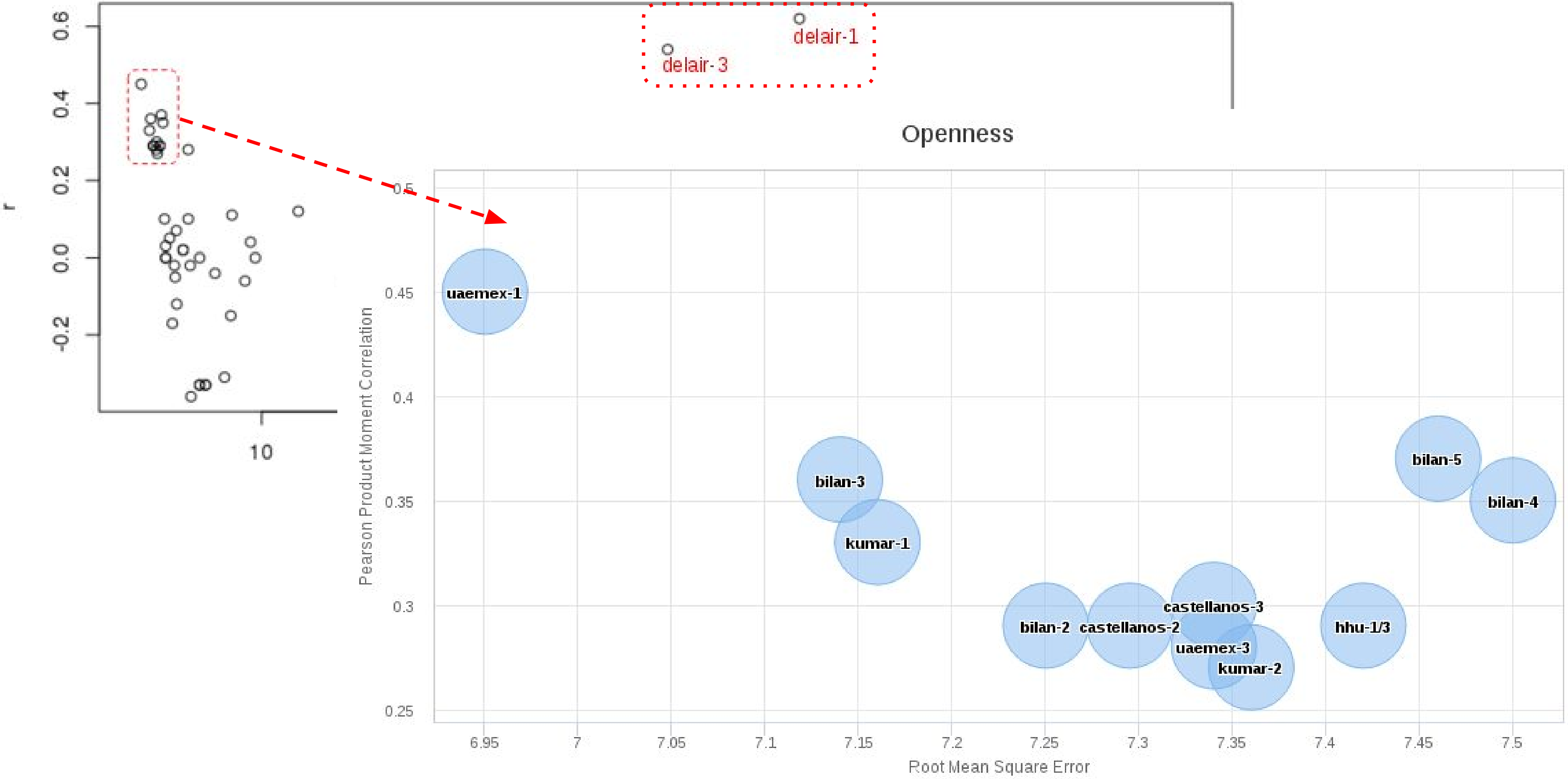
Neuroticism



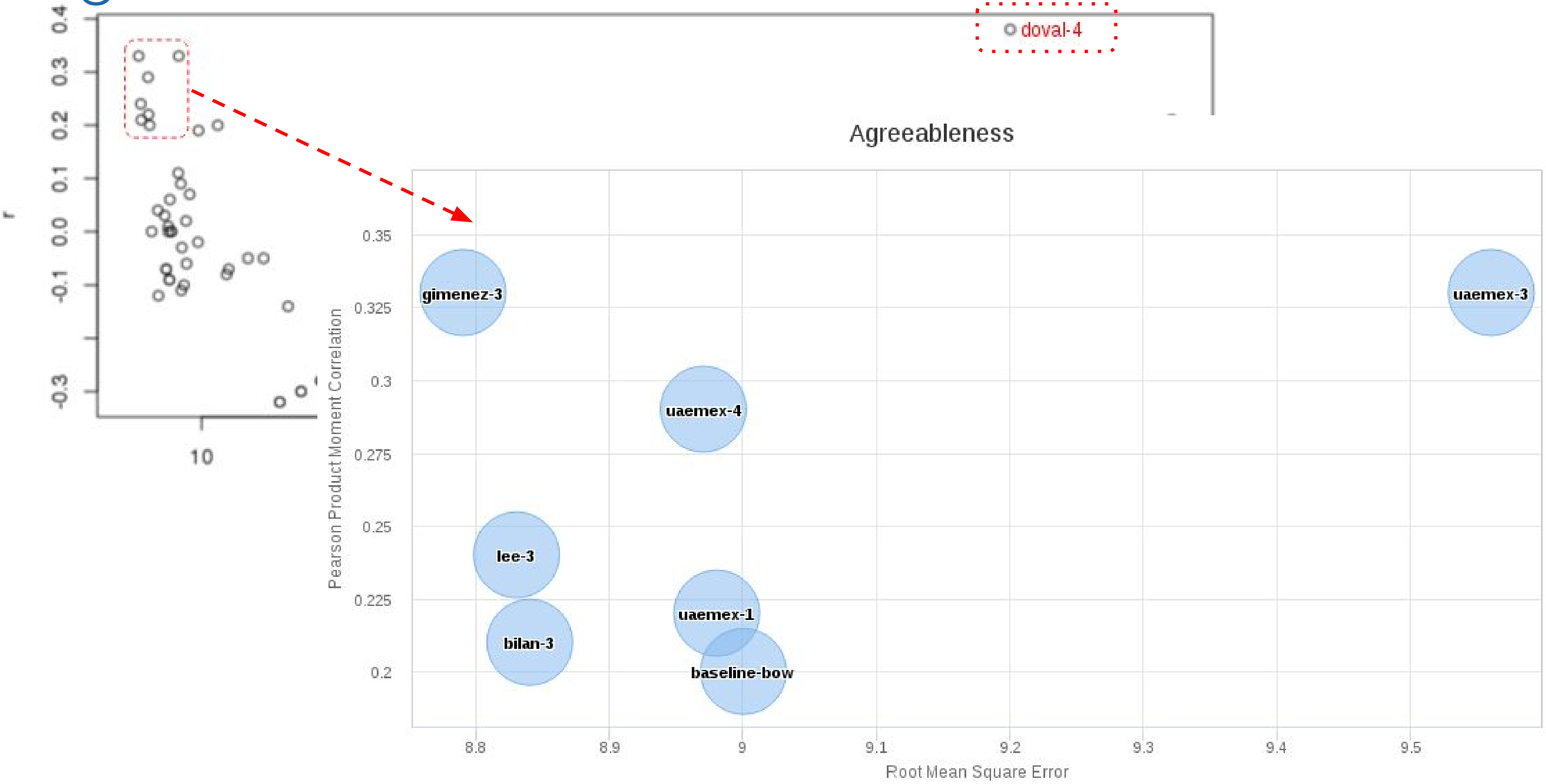
Extroversion



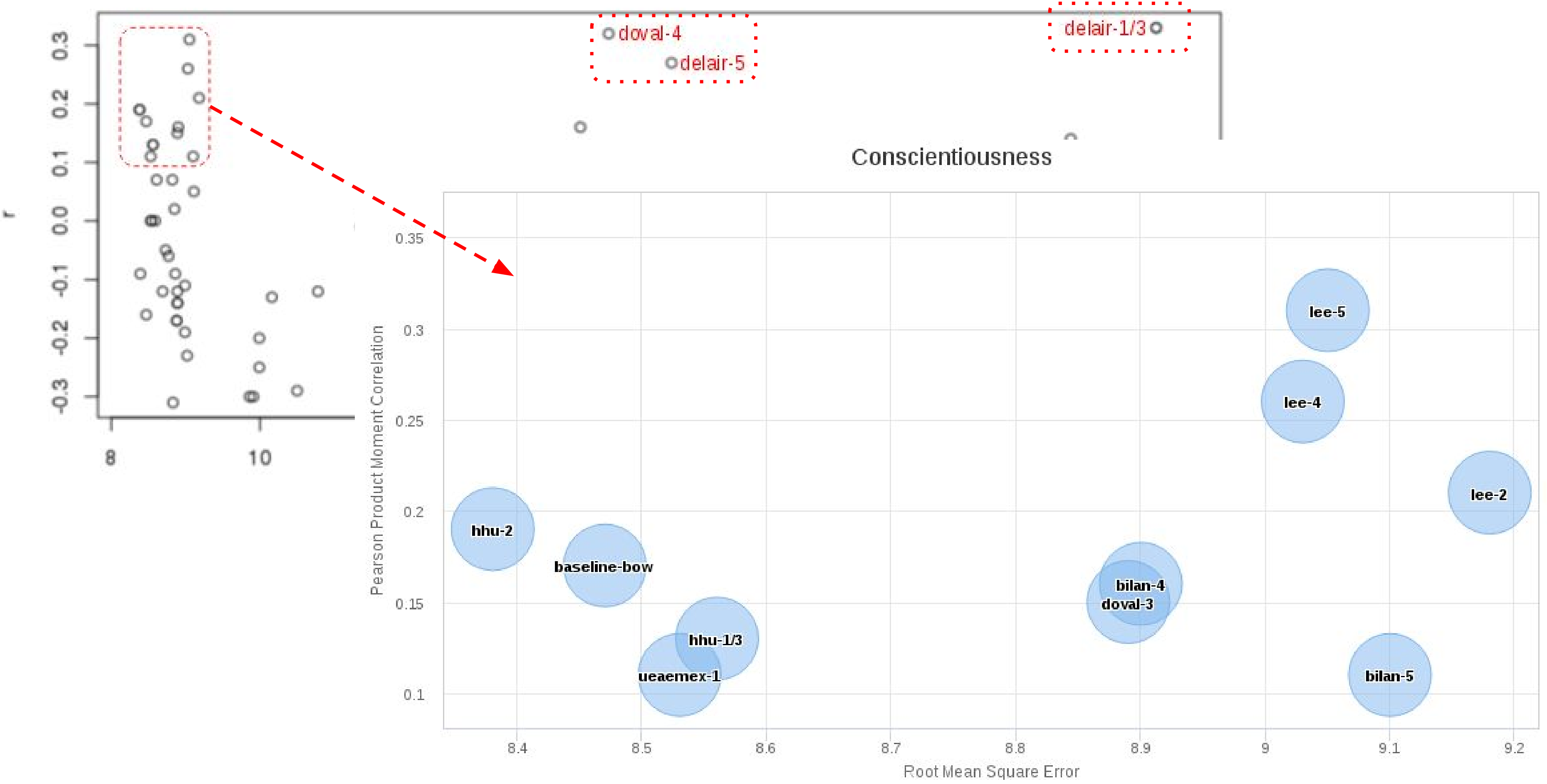
Openness


0.6
0.4
0.2
0.0
-0.2
r
10
delair-1
delair-3
Openness
0.5
0.45
0.4
0.35
0.3
0.25
Pearson Product Moment Correlation
uaemex-1
bilan-3
kumar-1
bilan-5
bilan-4
bilan-2
castellanos-2
castellanos-3
uaemex-3
kumar-2
hhu-1/3
6.95
7
7.05
7.1
7.15
7.2
7.25
7.3
7.35
7.4
7.45
7.5
Root Mean Square Error

Agreableness



Conscientiousness



Conclusions

- The task aimed at identifying big five **personality traits** from Java **source codes.**
- There have been **11 participants** sending **48 runs.**
- **Two complementary measures** were used:
  - **RMSE**: overall score of the **performance.**
  - **Pearson Product-Moment Correlation**: whether the performance is due to **random** chance.
- Wrt. **results**:
  - Quite **similar** in terms of **Pearson** for all traits.
  - Higher differences wrt. **RMSE**: **the best** results for **openness (6.95)**
- Several different **features**:
  - **Generic** (word and character n-grams) vs. **specific** (obtained by parsing the code, analysing its structure, style or comments)
  - **Generic** features obtained **competitive** results in terms of **RMSE**...
  - ... but with **lower Pearson** values.
  - They seemed to be **less robust.**
- **Baselines** obtained low RMSE with low Pearson -> this highlights the need of using both complementary measures.

On behalf of the PR-SOCO task organisers:

Thank you very much for participating
and hope to see you next year!!