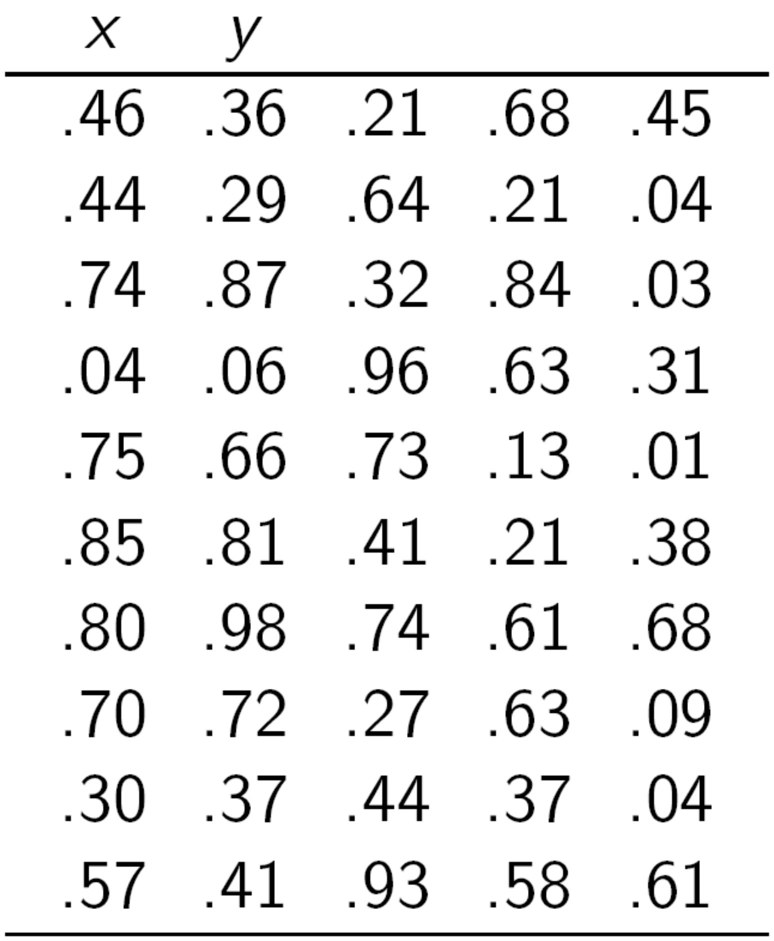
Randomization of Real-Valued Matrices for Assessing the Significance of Data Mining Results

- Problem:
  - Original *m x n* real-valued matrix *A*
  - Data mining result *S(A)*
  - How to assess the significance of *S(A)*?
- Our solution:
  - Randomization-based significance testing
  - Empirical *p*-value
  - Preserve the **row and column means and variances**

Markus Ojala, Niko Vuokko, Aleksi Kallio, Niina Haiminen and Heikki Mannila

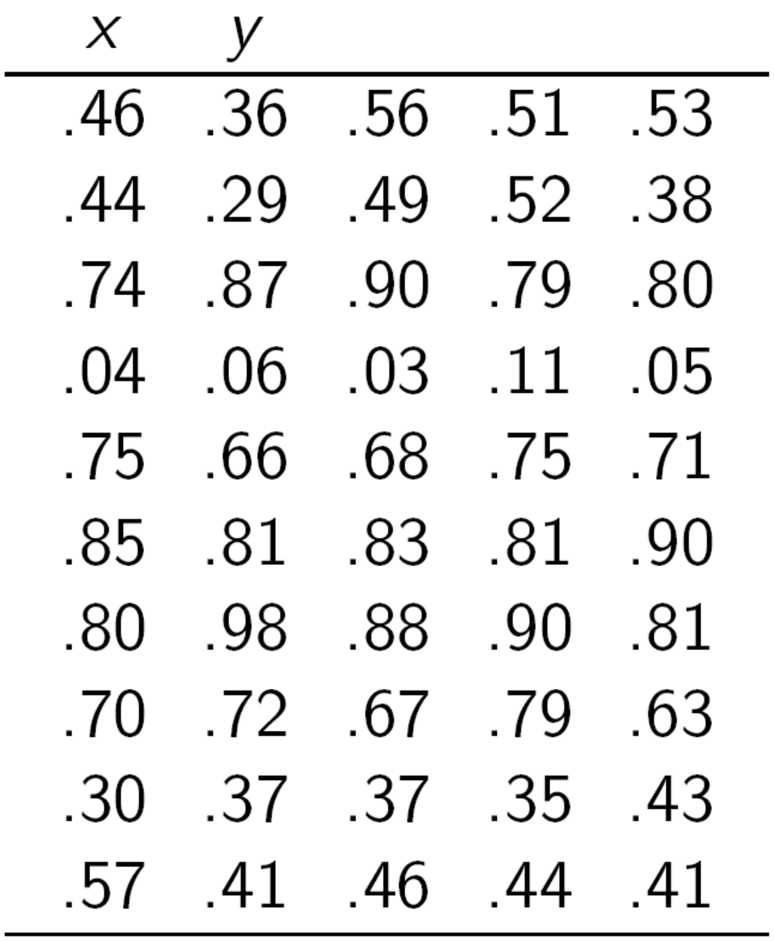
HIIT, Helsinki University of Technology, CSC, University of Helsinki

Example of using the approach



x	y			
.46	.36	.21	.68	.45
.44	.29	.64	.21	.04
.74	.87	.32	.84	.03
.04	.06	.96	.63	.31
.75	.66	.73	.13	.01
.85	.81	.41	.21	.38
.80	.98	.74	.61	.68
.70	.72	.27	.63	.09
.30	.37	.44	.37	.04
.57	.41	.93	.58	.61

Matrix A



x	y			
.46	.36	.56	.51	.53
.44	.29	.49	.52	.38
.74	.87	.90	.79	.80
.04	.06	.03	.11	.05
.75	.66	.68	.75	.71
.85	.81	.83	.81	.90
.80	.98	.88	.90	.81
.70	.72	.67	.79	.63
.30	.37	.37	.35	.43
.57	.41	.46	.44	.41

Matrix B

- Data mining: correlation between columns *x* and *y* (= *0.92*)
- Significance testing (1000 samples): $p_A = 0.001$, $p_B = 0.4156$