Data-Mining Go Games

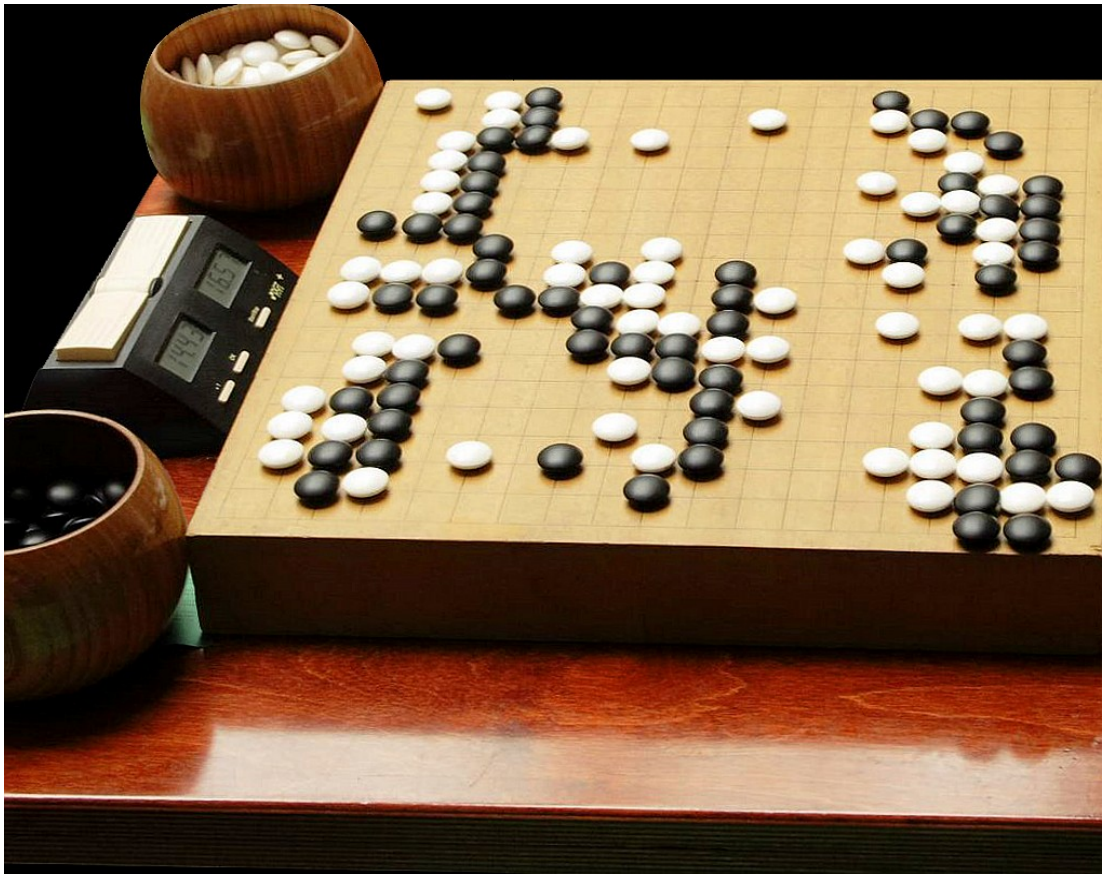
[Submitted to IEEE TCAIG]

Petr Baudis <pasky@ucw,cz>
Joe Moudrik <j.moudrik@gmail.com>
MFF, Prague, 2010

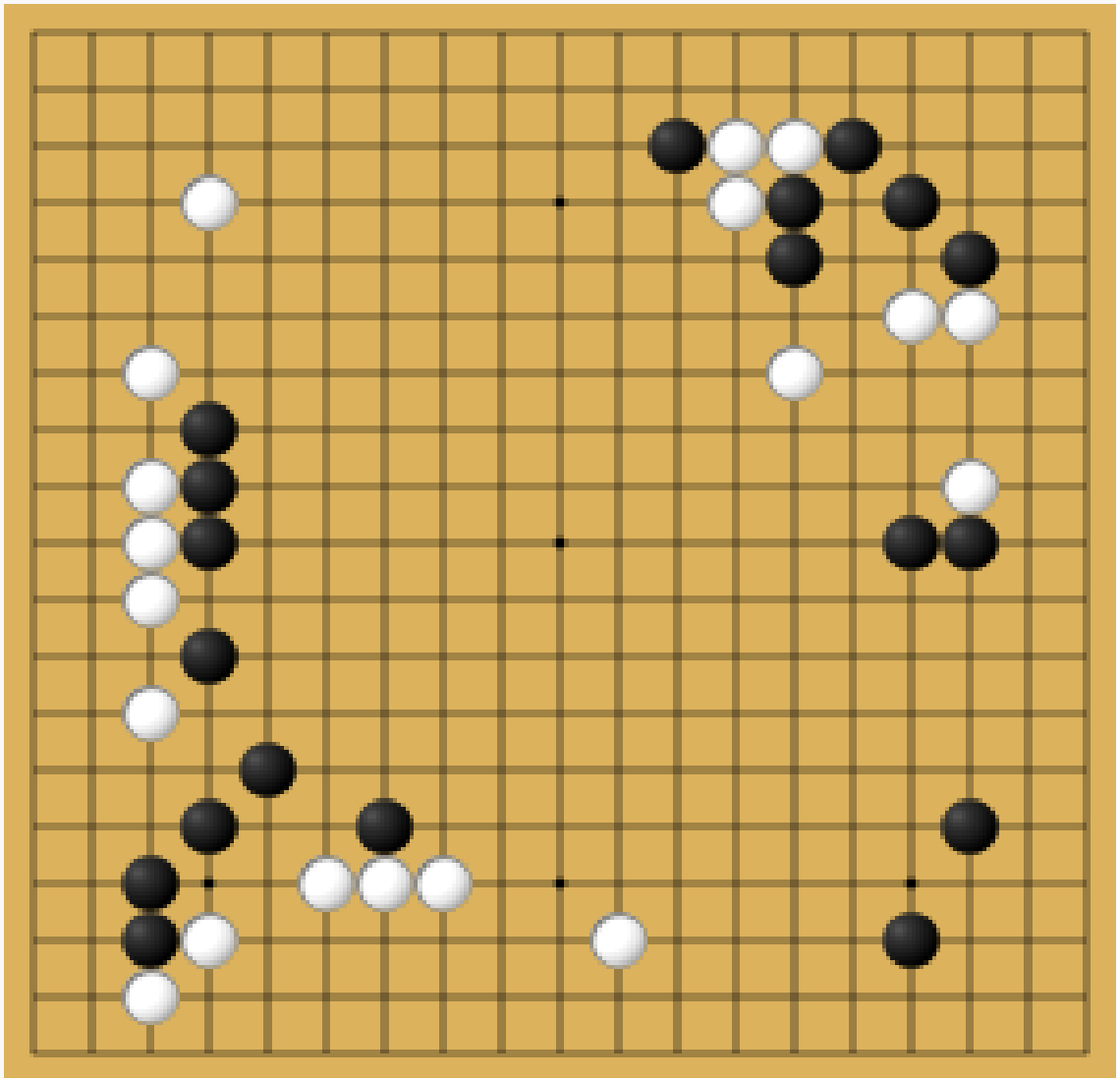
Table of Contents

Game of Go

- Ancient Asian game: square grid board, black&white stones
- Goal is surrounding most territory and capturing enemy stones
- Many professional players in Asia well known by amateurs



Example Game



White: territory

Black: aggressive

Side: Safe territory

Center: Hard to make territory, fight zone

Go Styles

- How the player steers the game, the kind of played moves (defensive/offensive, simple/complicated, …)
- Very hard to define, nuanced and ambiguous in the literature (“thick”, “orthodox”, “nimble”?!)
- Motivation:
  - Fun for players to find out style
  - Good studying guide
  - Learning other things than style

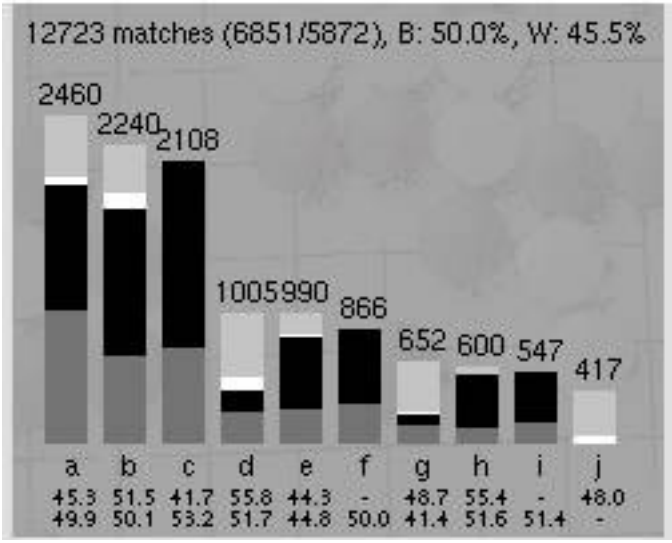
Go Strength

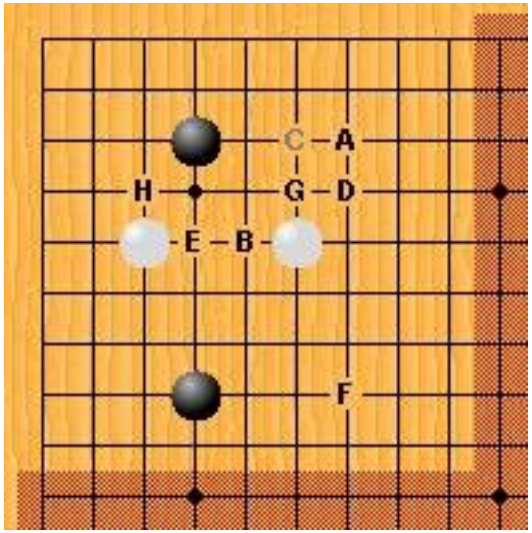
- Beginner: 30kyu ... 1kyu, 1dan ... 9dan :Expert
- Professionals: 1pro ... 9pro, equiv. to ~7d..9d
- Regular players distributed normally at around ~4kyu
- Correspondence to probability of win
  - d=0 means 1/2 chance of win
  - d=1 means 2/3 chance of stronger win
- (Elo rating scale)

Prior Data Mining

- Pattern-based program play
- Frequency of moves:

  In situation X, most played pro moves are A, B, C, D, then sometimes E;
  the followups are at ...
- Situation: Arbitrary board segment
- Fast matching: schemes based on Zobrist hashing





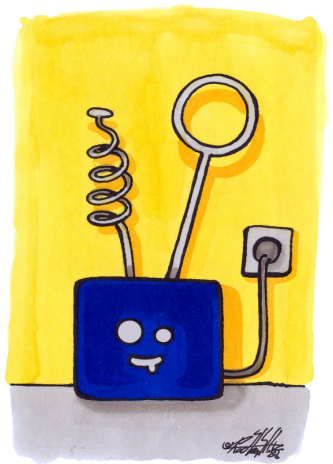
Methodology

Go Game Corpus

- **GoGoD 2009**
  - ~55000 games over 400 year time span
  - Professional-level games (simil. strength)
  - Styles research: Sample ~20 well-known current professionals, explore main differentiating factors
- **Go Teaching Ladder** review archive
  - ~7500 recent games 30kyu..4dan
  - Strength research: Differentiating factors

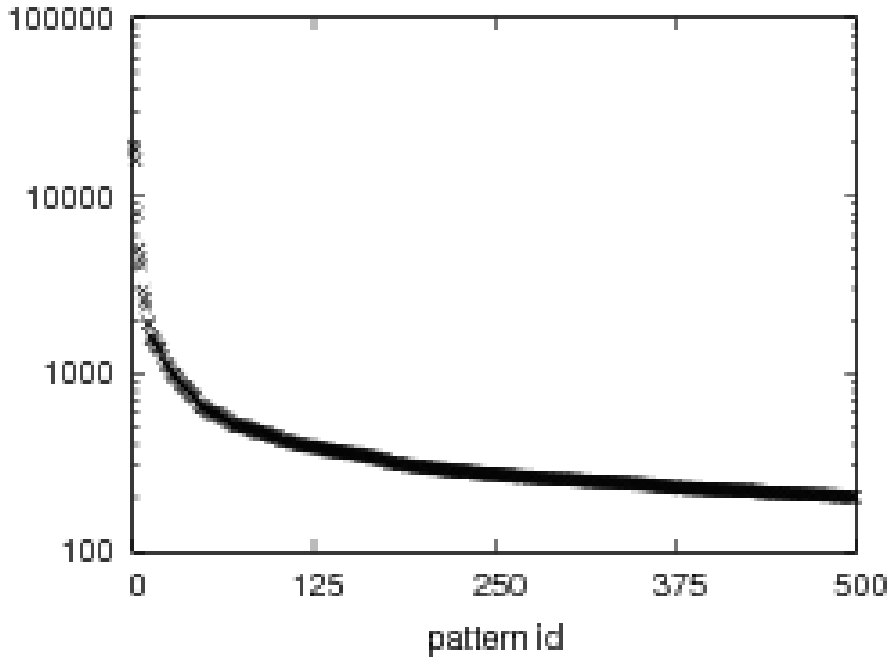
Input: Patterns

- Extracting **patterns** for each player:
  - Combination of **features** of chosen move
  - Edge distance, last move distance, …
  - Spatial feature: Hash of stone configuration in given gridcular radius
- Specificity tradeoff
- Implemented "patternscan" engine within the Pachi go framework



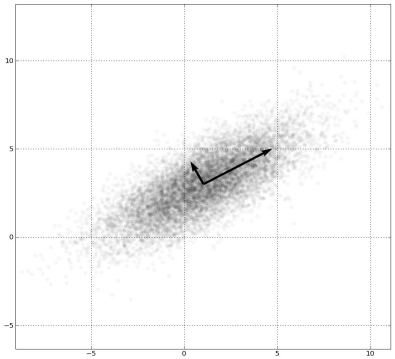
Input: Pattern Vectors

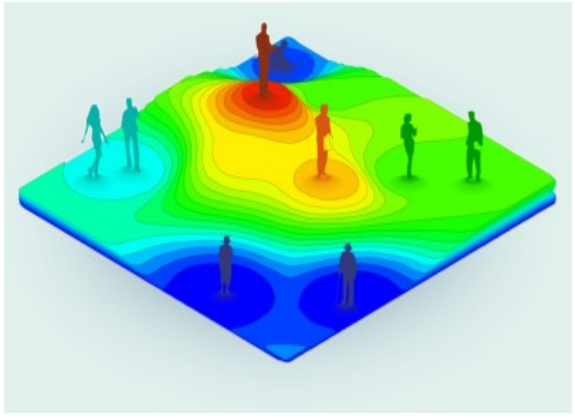
- Take n(=500) patterns most frequent in the corpus
- Build $[-1,1]^{500}$ vectors with per-subject frequencies
- Rescaling emphasizes different pattern types



Output: Statistical Exploration

- No prior knowledge, pure internal analysis
- **Principal Component Analysis**
  - Find eigenvectors of the data matrix base
  - Decomposes data to main linear dependencies





- **Sociomaps**
  - Spread players to a 2D plane, pretty pictures!
  - Layout keeps dataset ordering

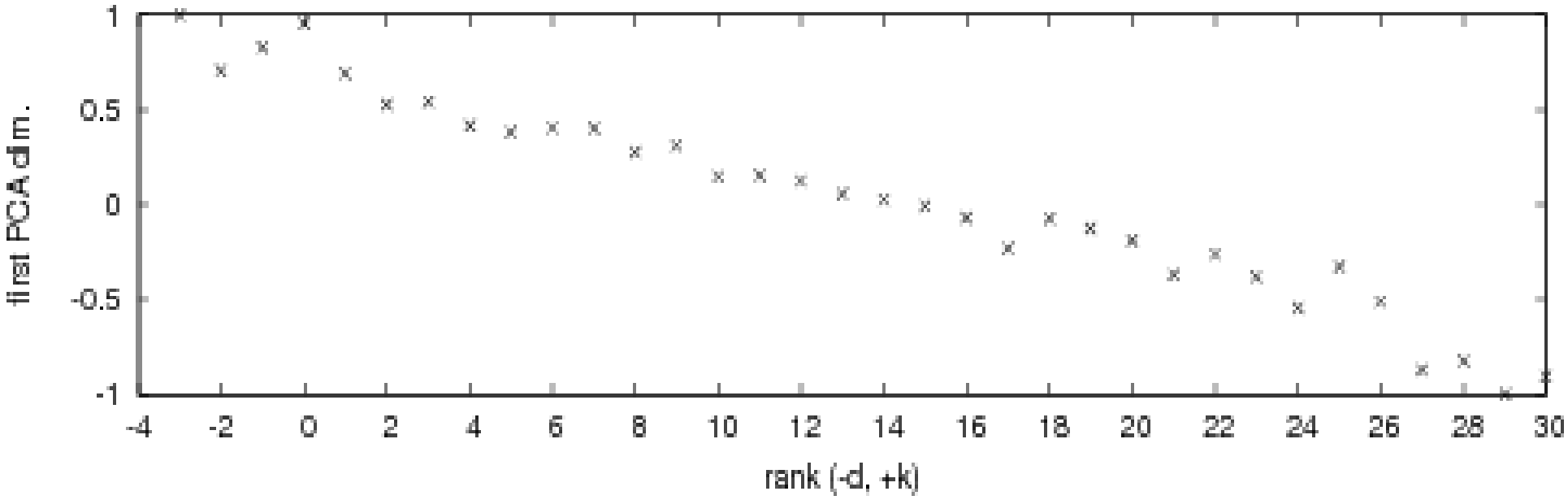
Output: Classifiers

- Assign input ***P*** classification ***O***
- Prior classifications by expert knowledge
- **k-Nearest Neighbors**
  - Weighted average of nearest patterns, weight exponentially decreasing with distance
- **Neural Networks**
  - Three-layer sigmoid RPROP
- **Naive Bayes Classifier**

Results

Go Strength Results

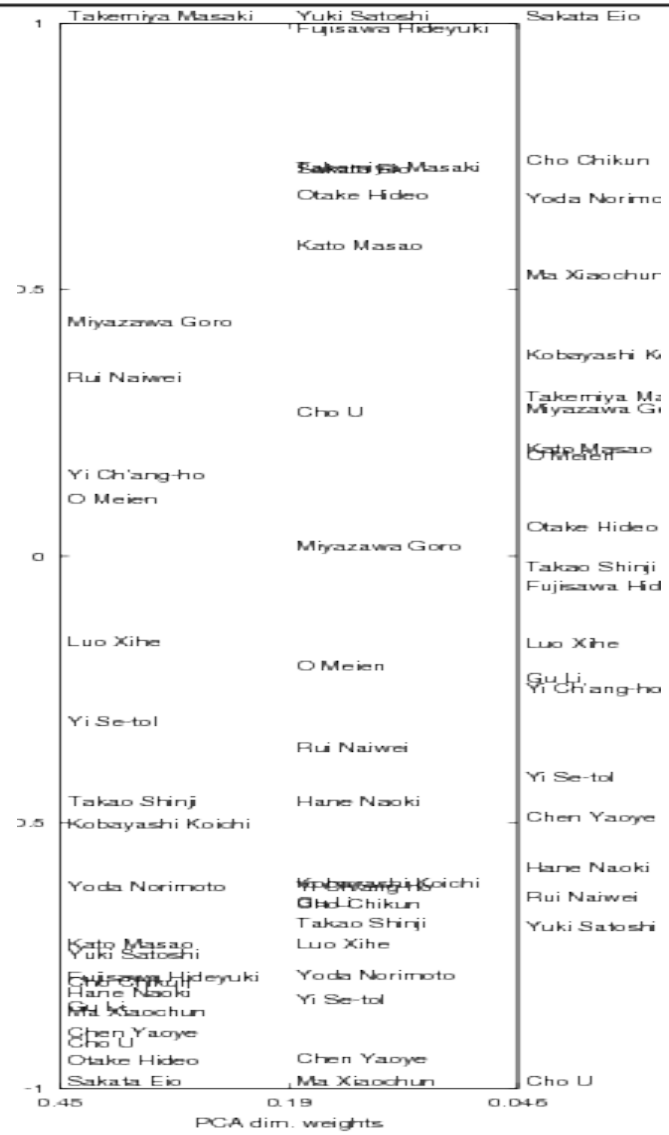
- GTL corpus, split by ranks
- One vector per rank, output is rank [-3,30]
- PCA shows very strong correlation r=0.979
- kNN over 15 games: +-6 ranks, 80: +- 4 ranks



Go Style Analysis

- GoGoD corpus, split by players
- One vector per player, output four style aspects plus "playing era" (median game year)
  - Territoriality, orthodoxity, aggressiveness, thickness – marks in [1,10]
- Expert information – style marks for 20 current pros by three strong players (3p, 7d, 4d)
  - Styles partially inter-correlated, standard error less than 1

Per-Player PCA



PCA – Style Correspondence

- Better correspondence: vector post-processing

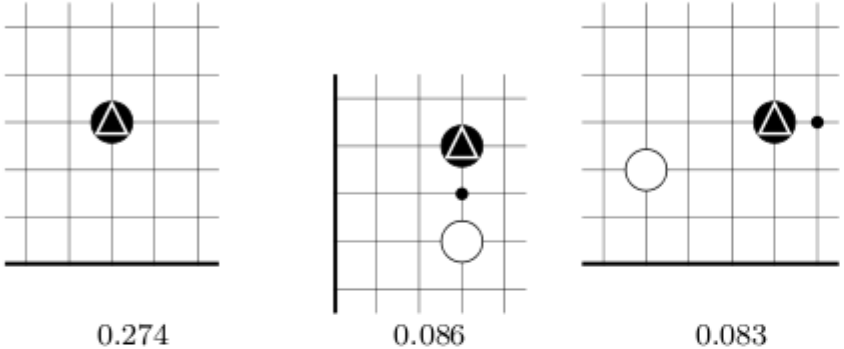
TABLE IV
COVARIANCE MEASURE OF PCA AND PRIOR INFORMATION

Eigenval.	τ	ω	α	θ	Year
0.4473697	**−0.530**	0.323	0.298	**−0.554**	0.090
0.1941057	**−0.547**	0.215	0.249	−0.293	**−0.630**
0.0463189	0.131	−0.002	−0.128	0.242	**−0.630**
0.0280301	−0.011	0.225	0.186	0.131	0.067
0.0243231	−0.181	0.174	−0.032	−0.216	0.352
0.0180875	−0.364	0.226	0.339	−0.136	0.113
0.0138478	−0.194	−0.048	−0.099	−0.333	0.055
0.0110575	−0.040	−0.254	−0.154	−0.054	−0.089
0.0093587	−0.199	−0.115	0.358	−0.234	−0.028
0.0084930	0.046	0.190	0.305	0.176	0.089

PCA Characteristic Patterns

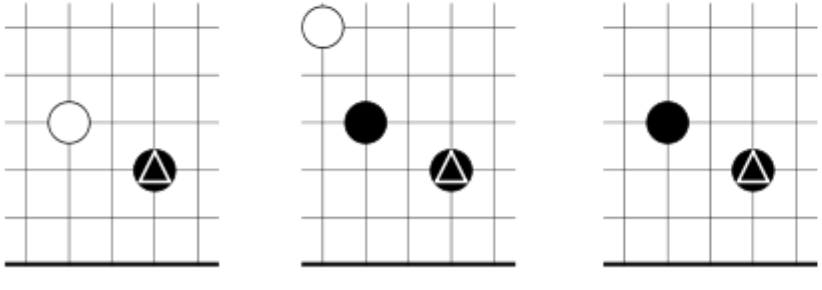
TABLE V
CHARACTERISTIC PATTERNS OF $PCA_{1,2}$ DIMENSIONS [1]

PCA_1 — Moyo-oriented, thin-playing player



0.274	0.086	0.083
high corner/side opening [2]	high corner approach	high distant pincer

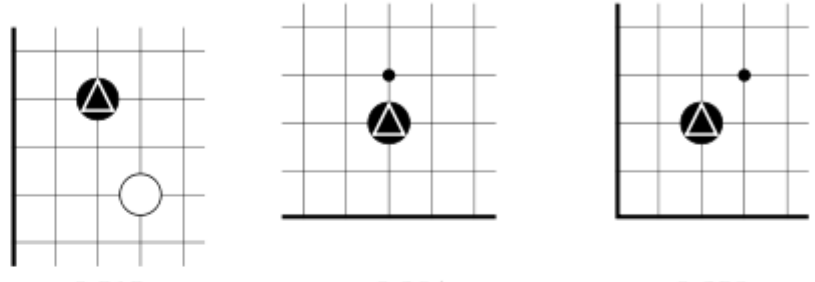
PCA_1 — Territorial, thick-playing player



−0.399	−0.399	−0.177
low corner approach	low corner reply	low corner enclosure

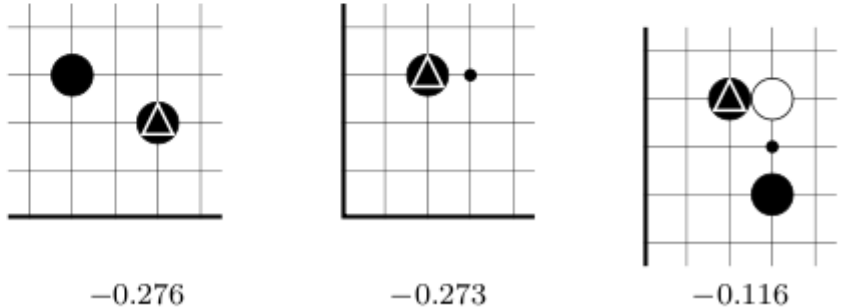
TABLE VI
CHARACTERISTIC PATTERNS OF PCA_3 DIMENSION [1]

PCA_3 — Old-time player



0.515	0.264	0.258
low corner approach	low side or mokuhazushi opening	san-san opening

PCA_3 — Current player

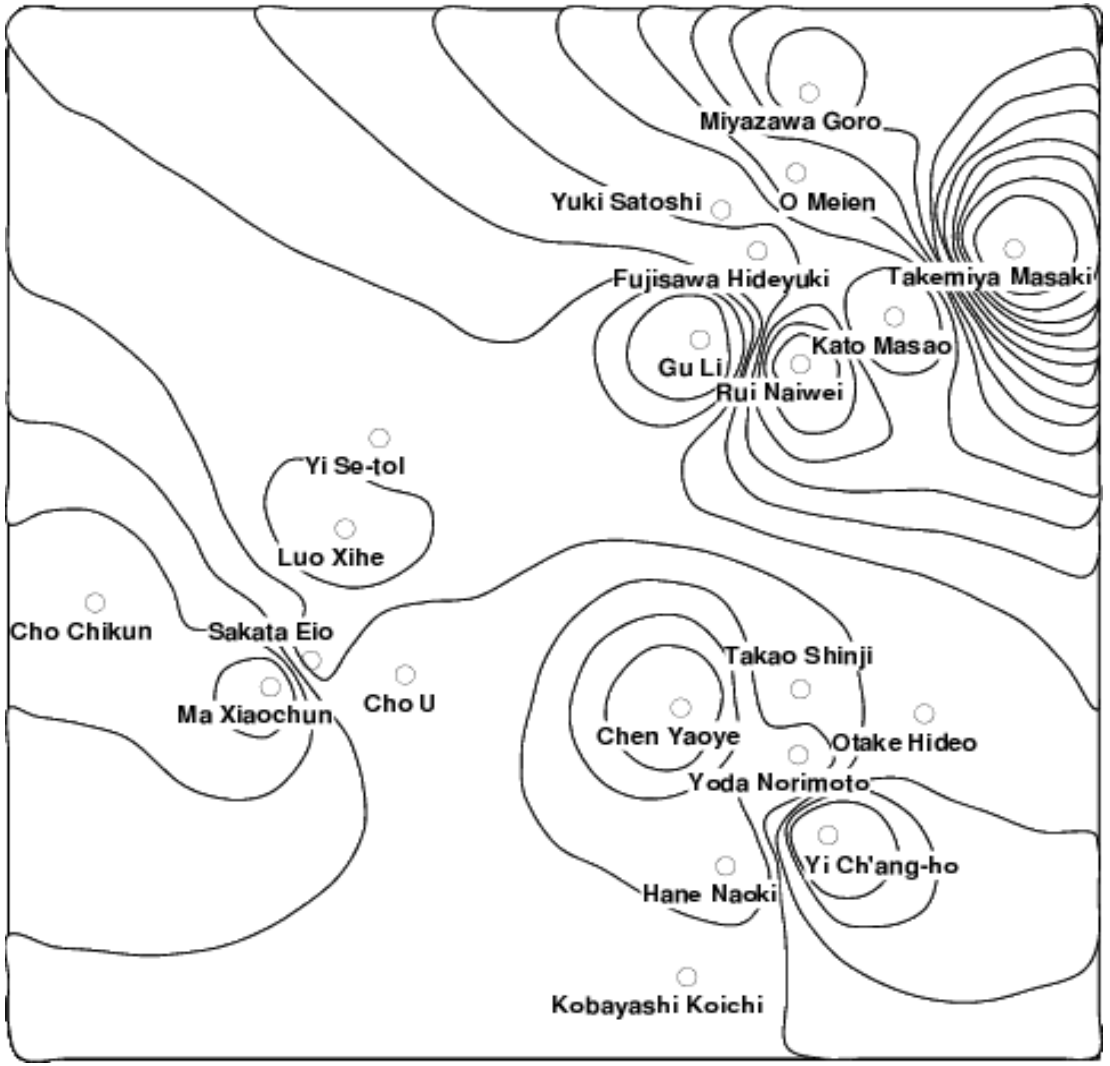


−0.276	−0.273	−0.116
low corner enclosure	3-4 corner opening [2]	high approach reply

[1] We cannot use terms “classic” and ”modern” in case of PCA_3

PCA – Style Sociomap

- Layout: Experts
- Countours: PCA1+PCA2
- Mostly smooth
- Gu Li **x** Rui Naiwei



Style Classification

- Guessing four style dimensions [1,10]
- Joint: Different dimensions kNN, NN

TABLE VIII
COMPARISON OF STYLE CLASSIFIERS

	MSE					
Classifier	τ	ω	α	θ	**Mean**	**Cmp**
Joint classifier[1]	4.04	**5.25**	**3.52**	**3.05**	**3.960**	2.97
Neural network	**4.03**	6.15	**3.58**	3.79	4.388	2.68
k-NN ($k = 2$)	4.08	5.40	4.77	3.37	4.405	2.67
k-NN ($k = 3$)	4.05	5.58	5.06	3.41	4.524	2.60
k-NN ($k = 1$)	4.52	**5.26**	5.36	**3.09**	4.553	2.59
k-NN ($k = 4$)	4.10	5.88	5.16	3.60	4.684	2.51
Naive Bayes	4.48	6.90	5.48	3.70	5.143	2.29
Random class.	12.26	12.33	12.40	10.11	11.776	1.00

Conclusions

- Clear correspondence in the dataset!
- More training data needed
- More factors can be investigated

- Classificators mildly successful, need more work
- Real-world classification application would be nice

Thank you!

Q&A

http://pasky.or.cz/~pasky/go/
http://www.goweb.cz/