
# Automated System for Cognitive Presence Coding

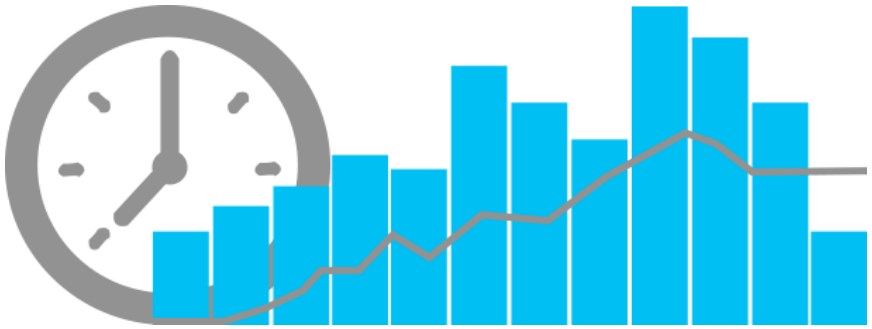
## Presentation of the outcomes

Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević
THE UNIVERSITY *of* EDINBURGH
s.joksimovic@ed.ac.uk
@s_joksimovic


UK Learning Analytics Network
London,
3rd February 2016

# The Main Idea

## Learning Analytics for Supporting Traditional DE/OL & MOOCs



# Overall Goal and Study Approach

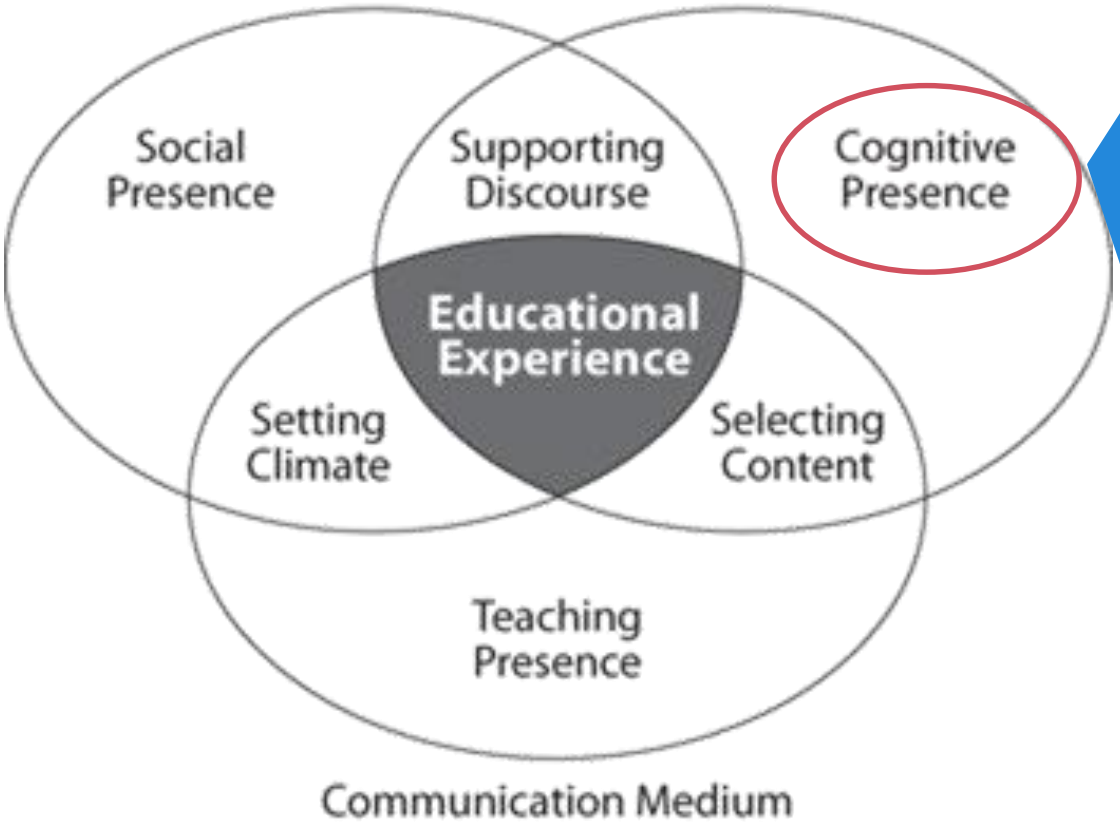
Through analytics:

- Drive pedagogical interventions
- Build models/frameworks of MOOC learning
- Empowering instructors and learners

Approach:

- Build Learning Analytics research on ***existing knowledge and models of Distance/Online education***
- Use Learning Analytics to **validate and extend** existing educational theories
- **Automate** as much as possible

# Community of Inquiry




1. **Triggering event**: Problem identification, sense of puzzlement
1. **Exploration**: Brainstorming, Idea exploration, divergence
2. **Integration**: Synthesis of relevant information
1. **Resolution**: Problem resolution, testing application

https://coi.athabascau.ca/coi-model/

Garrison, Anderson, and Archer (1999)

# Proposal:
# Automated CoI Content Analysis

**Build a system for automated coding of discussion messages for the levels of cognitive presence**

Advantages:

- Enable for broader adoption of CoI model
- Faster and cheaper adoption in research
- Provide detailed operationalization of CoI coding scheme
- Real-time feedback of learning in discussions
- Enable for development of various analytics dashboards

Future work:

- Eventually support other presences/models

# Proposed tasks and current implementation

| Task | Impl. | Note |
|---|---|---|
| Implementation of lag features |  | |
| Implementation of Coh-Metrix & LIWC features |  | |
| Implementation of Conditional Random Field classifier |  | Random Forest in the final version |
| Coding MOOC data |  | Extended the initial task |
| Evaluation on MOOC data |  | Extended the initial task |
| Dissemination of results & reporting |  | **Papers:** 2 published, 1 in preparation |

 Completed  In progress

# Data

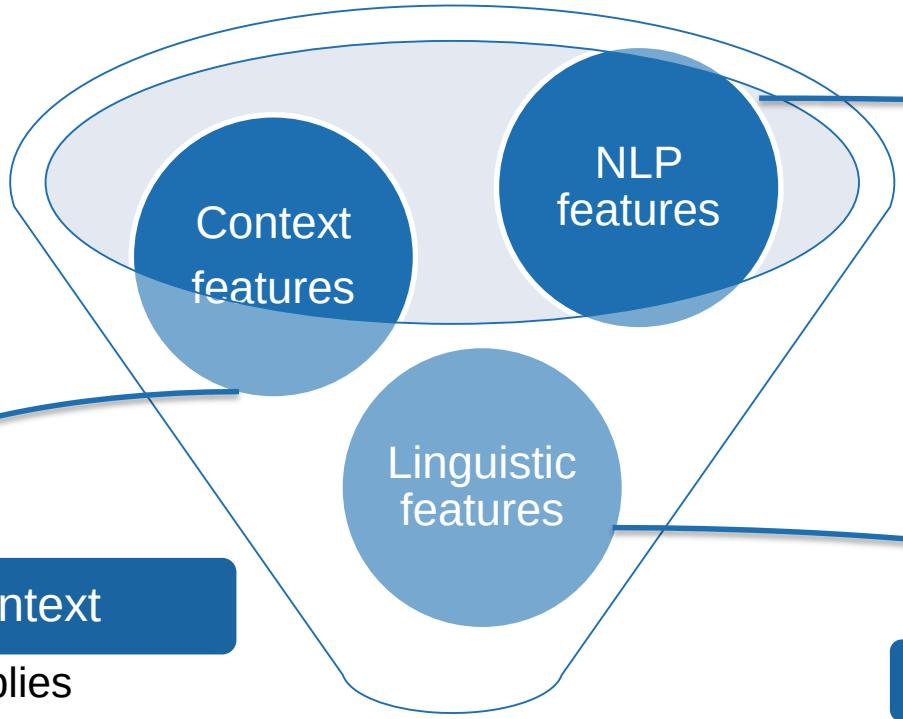
**Course offerings statistics**

| Course offer | # Students | # Msg |
|---|---|---|
| Winter 2008 | 15 | 212 |
| Fall 2008 | 22 | 633 |
| Summer 2009 | 10 | 243 |
| Fall 2009 | 7 | 63 |
| Winter 2010 | 14 | 359 |
| Winter 2011 | 13 | 237 |
| **Average (SD)** | **13.5 (5.1)** | **291.2 (192.4)** |
| **Total** | **81** | **1,747** |

**Distribution of cognitive presence phases**

| Phase | # Msg | % |
|---|---|---|
| Other | 140 | 8.0% |
| Triggering | 308 | 17.6% |
| Exploration | 684 | 39.2% |
| Integration | 508 | 29.1% |
| Resolution | 107 | 6.1% |
| **Average (SD)** | **349.4 (245.7)** | **20% (10%)** |
| **Total** | **1,747** | **100%** |

# Feature extraction




**Discussion context**

- Number of replies
- Message depth
- Cosine similarity to previous/next message
- Start/end indicators

**LSA**

- Text similarity

**Named entities**

- Text annotation (DBPedia Spotlight)

**Coh-Metrix**

- 103 measures of text coherence

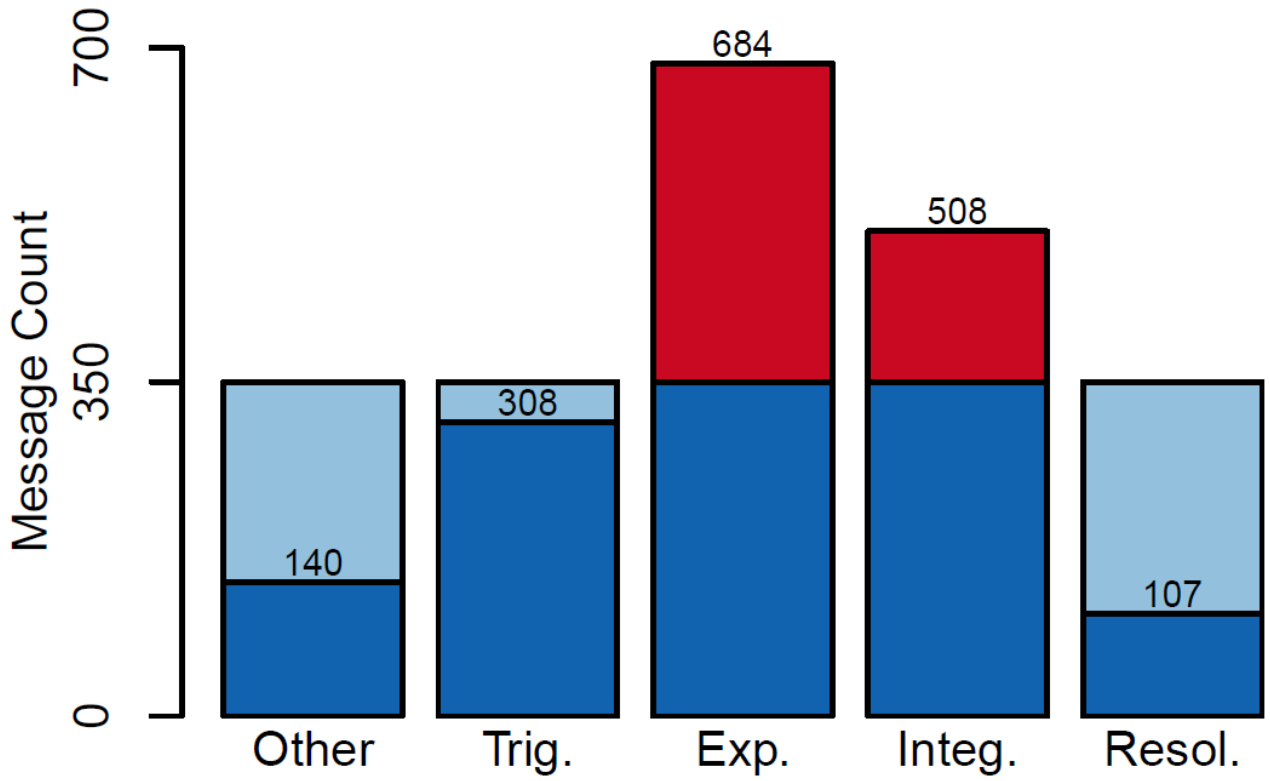
**LIWC**

- 93 measures representing psychological processes

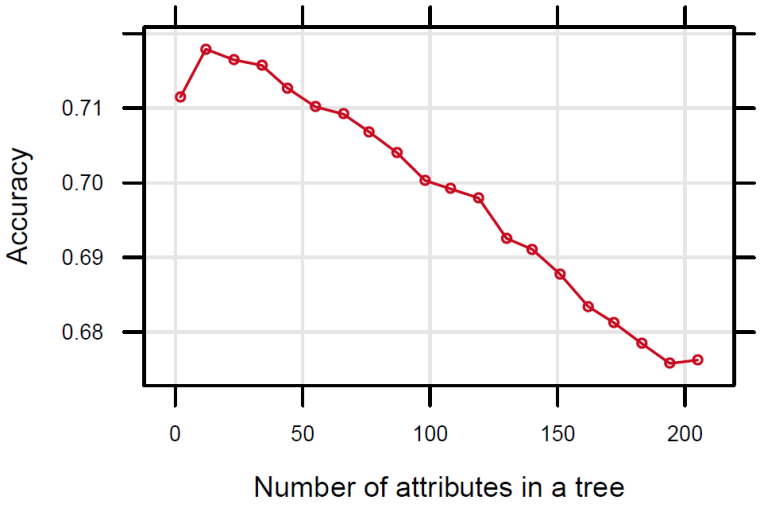
Feature space

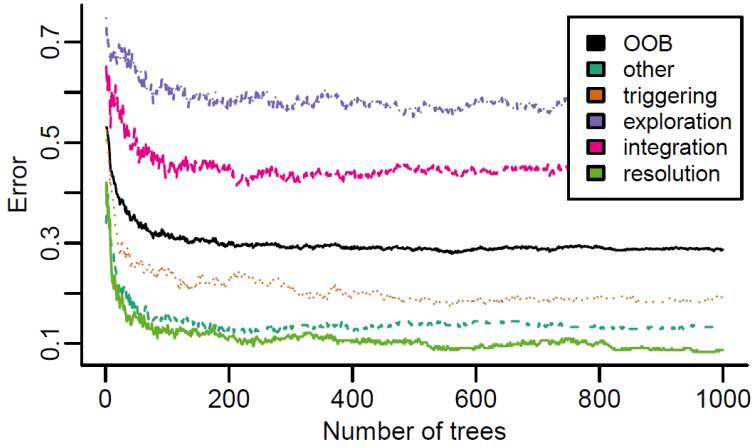
# Data pre-processing

**SMOTE** - **S**ynthetic **M**inority **O**ver-sampling **Te**chnique



# Model selection and results




| Study | Classifier | Accuracy | Cohen's Kappa |
|---|---|---|---|
| Kovanović et al. (2014) | Support Vector Machines | 53.38% | 0.41 |
| Waters et al. (2015) | Conditional Random Fields | 64.20% | 0.48 |
| **Kovanović et al. (2016)** | **Random Forest** | **70.30%** | **0.63** |

# Variable importance

| Variable | Other | Triggering | Exploration | Integration | Resolution |
|---|---|---|---|---|---|
| Number of words | 55.41 (61.06) | 80.91 (41.56) | 117.71 (67.23) | 183.30 (102.94) | 280.68 (189.62) |
| Number of named entities | 13.44 (15.36) | 21.67 (10.55) | 28.84 (16.93) | 44.75 (24.85) | 64.18 (32.54) |
| Lexical diversity, all words | 0.85 (0.12) | 0.77 (0.09) | 0.71 (0.10) | 0.65 (0.09) | 0.58 (0.09) |
| Position within discussion | 2.39 (1.13) | 1.00 (0.90) | 1.84 (0.97) | 1.87 (0.94) | 2.00 (0.68) |
| Lexical diversity, content words | 0.95 (0.06) | 0.90 (0.06) | 0.86 (0.08) | 0.82 (0.07) | 0.78 (0.07) |
| Avg. givenness of each sentence | 0.10 (0.07) | 0.14 (0.06) | 0.18 (0.07) | 0.21 (0.06) | 0.24 (0.06) |
| Number of question marks | 0.27 (0.85) | 1.84 (1.63) | 0.92 (1.26) | 0.58 (0.82) | 0.38 (0.55) |
| Similarity with previous message | 0.20 (0.17) | 0.06 (0.13) | 0.22 (0.21) | 0.30 (0.24) | 0.39 (0.19) |
| Lexical diversity, VOCD | 12.92 (33.93) | 28.99 (50.61) | 53.57 (54.68) | 83.47 (43.00) | 97.16 (28.95) |
| Avg. number of paragraphs sent. | 4.26 (2.98) | 6.37 (2.76) | 7.49 (4.11) | 10.17 (5.64) | 14.05 (8.88) |

# Project Insights

- Classification performance
  - Substantial level of agreement
- Feature Modeling
  - ~~surface features (e.g., unigrams, bigrams, POS-bigrams)~~
  - contextual features
  - domain-independent features
- “Class Balancing”
  - **S**ynthetic **M**inority **O**ver-sampling **Te**chnique

# Future Work

- Classification improvement
  - “Quoting” problem
  - Class balancing – different approaches
  - Extending feature set
- MOOC data – work in progress
- Learning Analytics platform
  - Integrate classification module
  - Provide associated probabilities

# Team

## Vitomir Kovanović

1. Third year PhD student (School of Informatics, UoE)
2. Msc Software Engineering
3. Four years industry experience as software developer
4. Best paper award at LAK 2015

## Srećko Joksimović

1. Third year PhD student (School of Education, UoE)
2. Msc Information systems
3. Six years industry experience as software developer

## Dragan Gašević (supervisor)

1. Professor at the University of Edinburgh (Informatics & Education)
2. President of SOLAR - Society for Learning Analytics research
3. PhD Information Systems

# References

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education*, *2*(2–3), 87–105.

Kovanović, V., Joksimović, S., Gašević, D., & Hatala, M. (2014). Automated Content Analysis of Online Discussion Transcripts. In *Proceedings of the Workshops at the LAK 2014 Conference*. Indianapolis, IN. Retrieved from http://ceur-ws.org/Vol-1137/

Z. Waters, V. Kovanovic, K. Kitto, and D. Gaševic. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Proceedings of the 11th Asia Information Retrieval Societies Conference, AIRS 2015*, 2015.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., & Siemens, G. (2016). Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In Proceedings of the Sixth International Conference on Learning Analytics And Knowledge, LAK 16 (accepted).

LAK 16
The 6th International
Learning Analytics & Knowledge Conference
University of Edinburgh, Edinburgh, UK, April 25-29, 2016
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
acm In-Cooperation
SIGCHI
sigweb

# Automated System for Cognitive Presence Coding

## Q&A

Vitomir Kovanović, Srećko Joksimović, and Dragan Gašević
THE UNIVERSITY *of* EDINBURGH
s.joksimovic@ed.ac.uk
@s_joksimovic

Jisc
UK Learning Analytics Network
London,
3rd February 2016