［依頼講演］

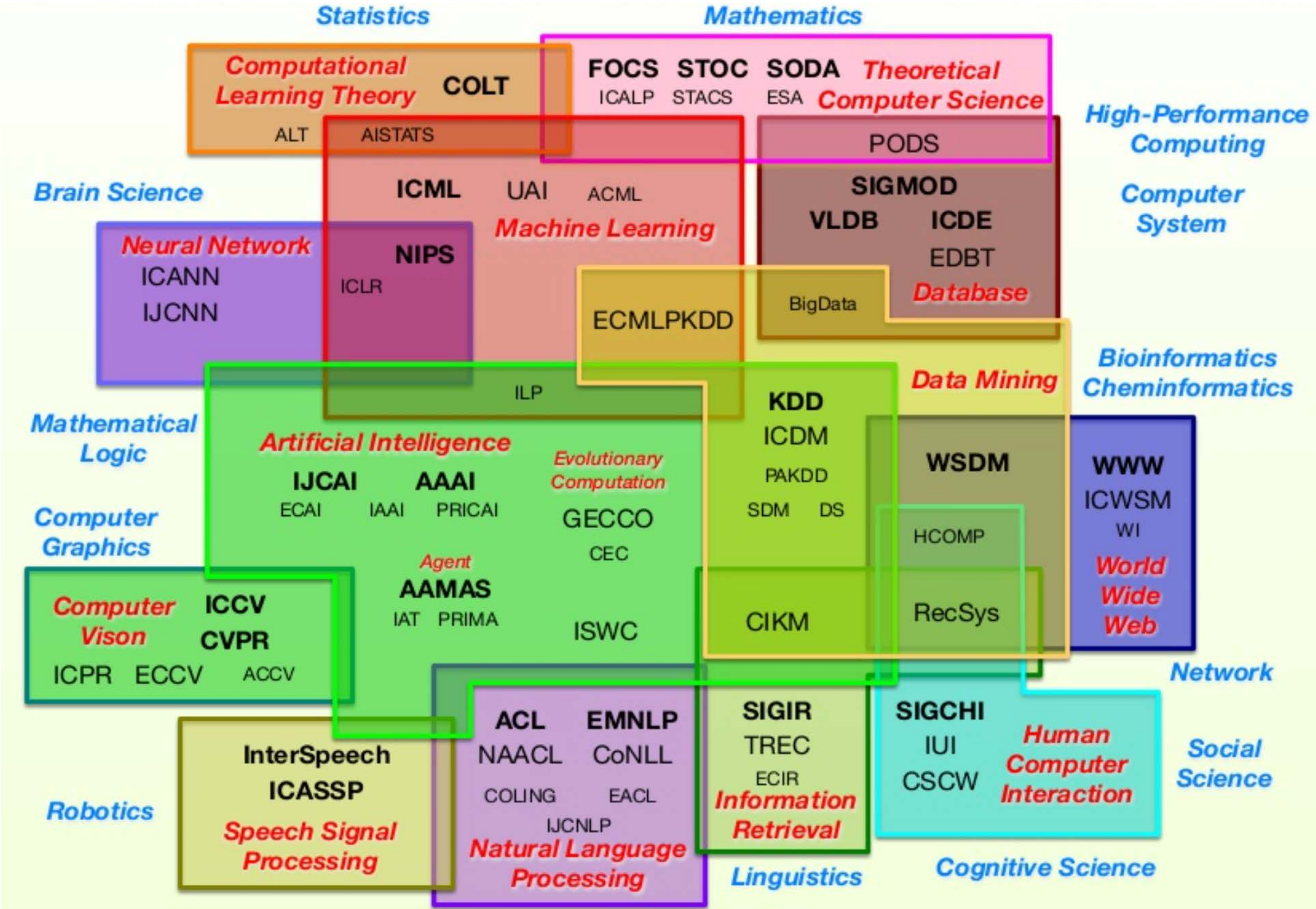
国際会議COLING2016 参加報告（2）

若宮 翔子
奈良先端科学技術大学院大学
研究推進機構
ソーシャル・コンピューティング研究室

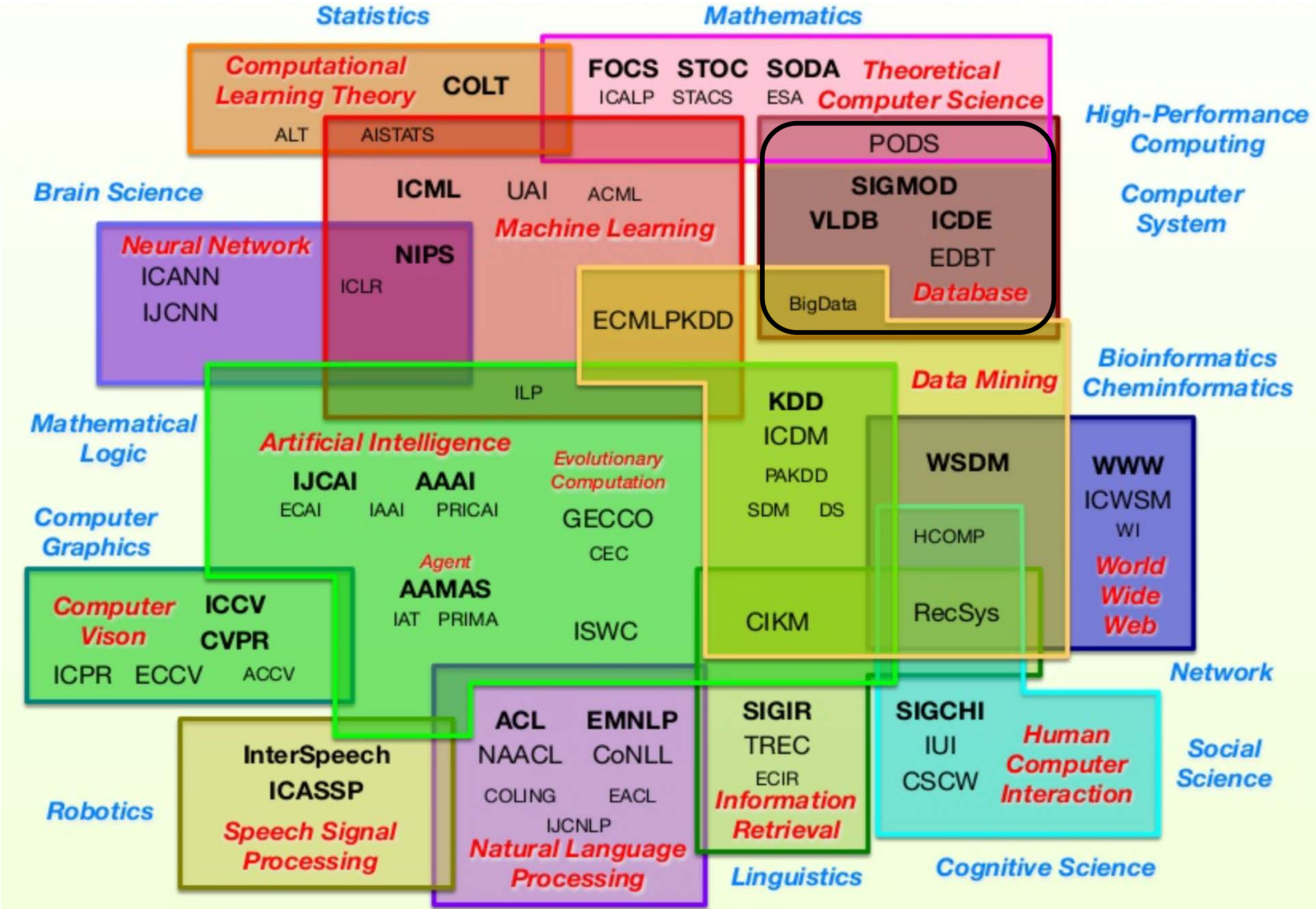
概要

1. DB系の人たちが狙う
   国際会議（の一部）の紹介

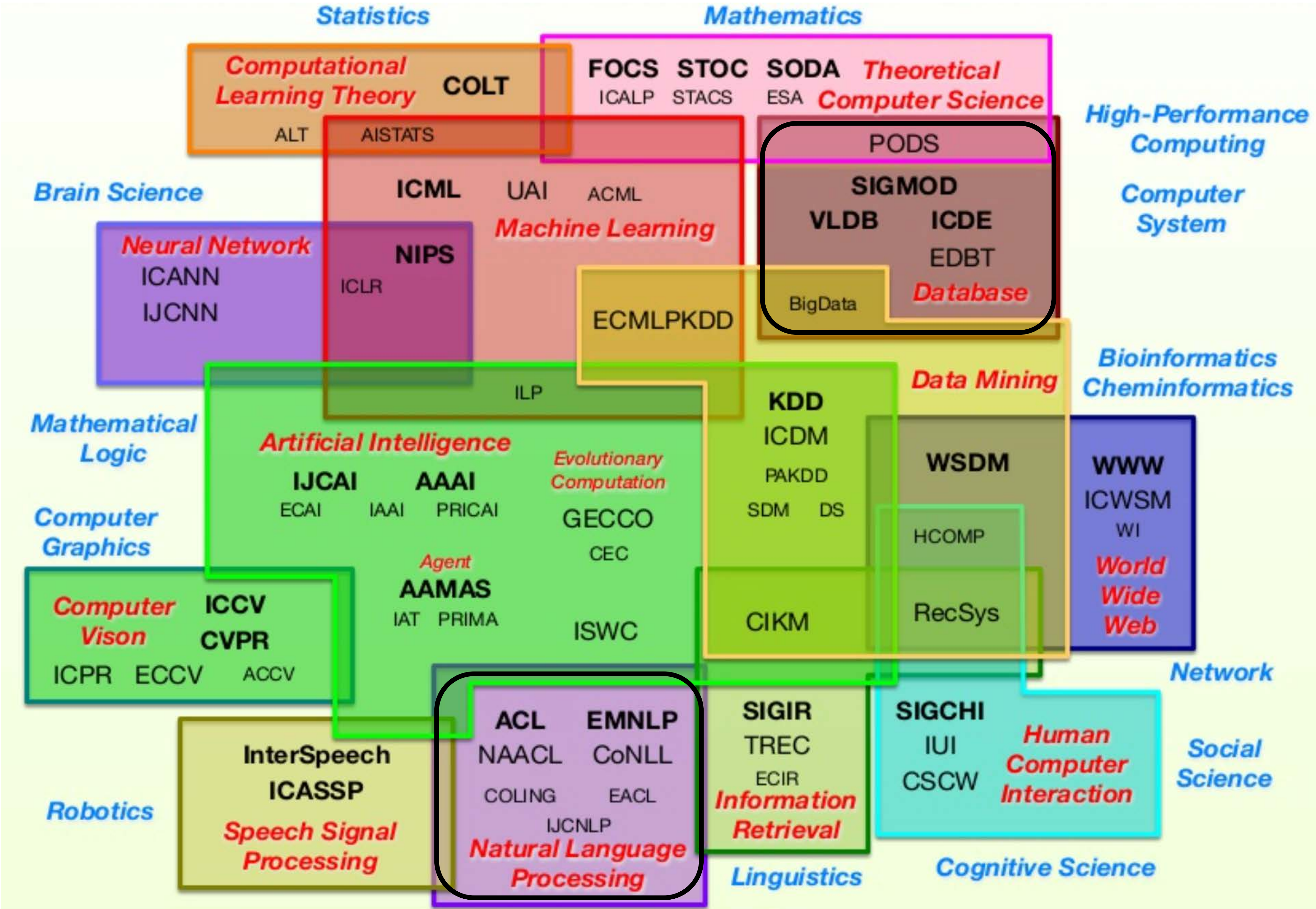
2. COLINGの論文の紹介



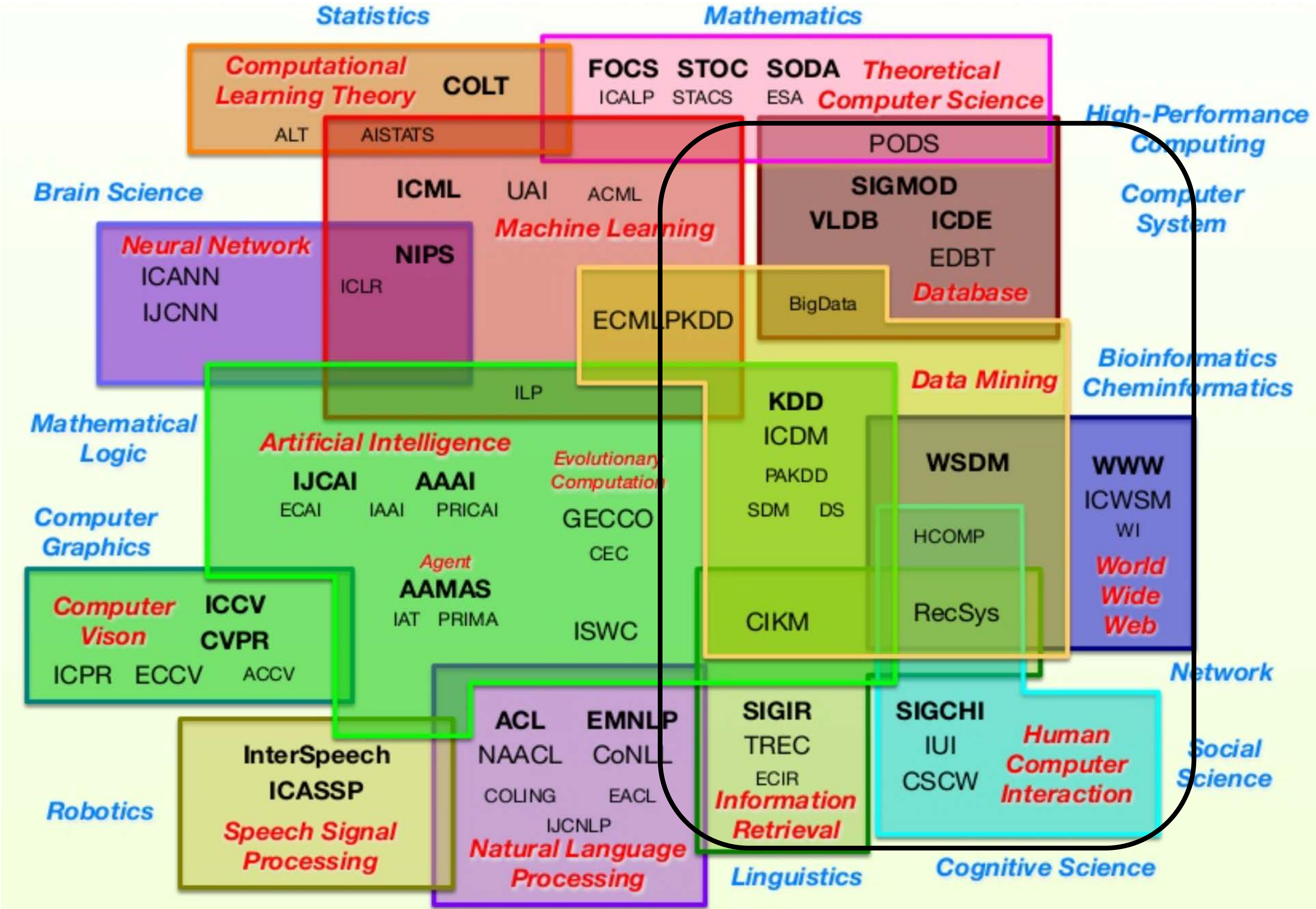
http://www.kamishima.net/soft/ (accessed 2016/12/20)



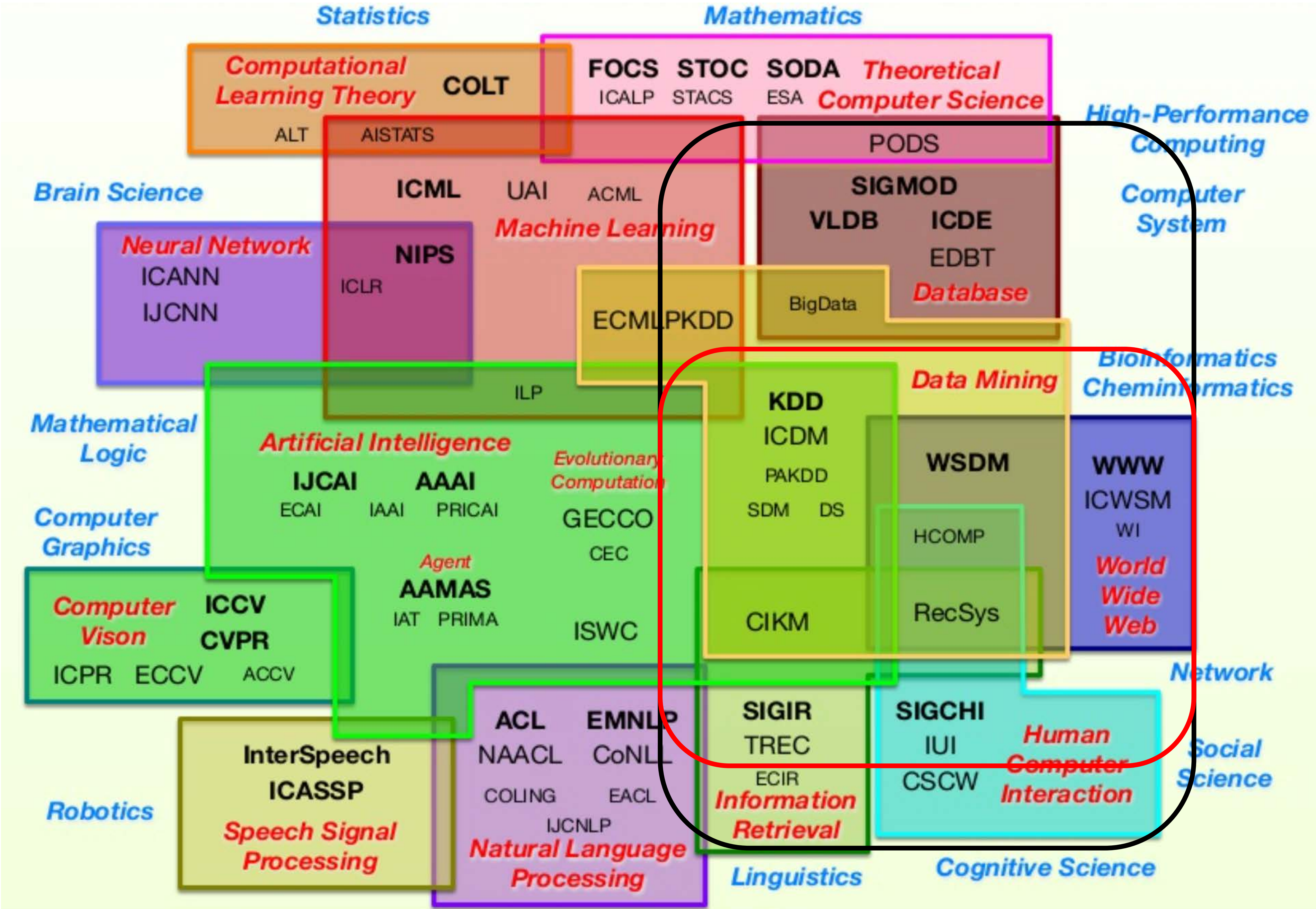
http://www.kamishima.net/soft/ (accessed 2016/12/20)



http://www.kamishima.net/soft/ (accessed 2016/12/20)

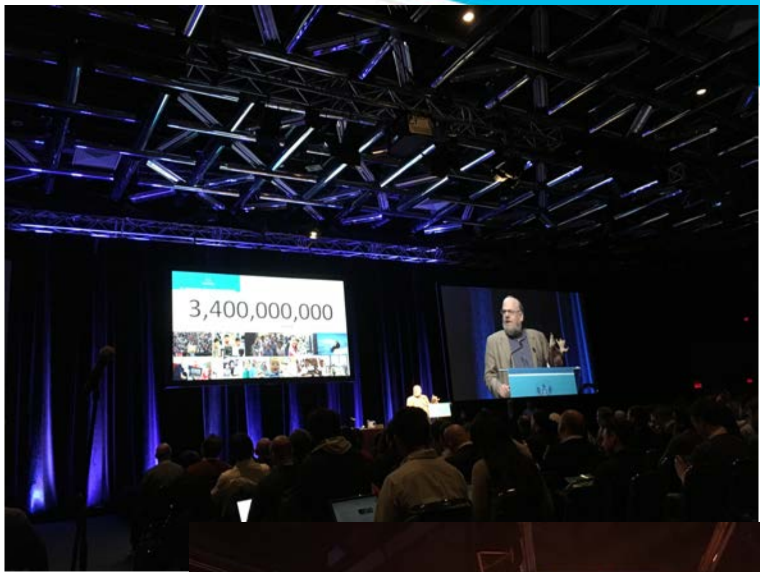


http://www.kamishima.net/soft/ (accessed 2016/12/20)



http://www.kamishima.net/soft/ (accessed 2016/12/20)

http://www2016.ca



←会場の様子

↓Banquet

- Web系
- Full Paper 採択率：15.8% (=115/727)
- 主なトピック：ソーシャルネットワーク，行動分析やパーソナライゼイション，セマンティックウェブやビッグデータ，セキュリティとプライバシーなど

- データマイニング系
- Full Paper 採択率
  - 2016：6% (=60/1115)
  - 2015：20% (=160/819)
- 主なトピック：機械学習，最適化，異常検知，推薦，クラスタリング，ソーシャルネットワーク
- 全員がポスター発表

Heidelberg, Germany September 12-16

UBICOMP 2016

http://ubicomp.org/ubicomp2016/

- ユビキタスコンピューティング分野
- 2013年からウェアラブルコンピューティングのトップ会議であるInternational Symposium on Wearable Computers と併設
- Full Paper/Notes 採択率： 23.7% (=114/481)
- 主なトピック：センサとパターン認識，人の位置情報や行動／ジェスチャ推定，セキュリティやプライバシ，健康管理

概要

1. DB系の人たちが狙う国際会議（の一部）の紹介

2. COLINGの論文の紹介

GAKE: Graph Aware Knowledge Embedding

Jun Feng*, Minlie Huang*, Yang Yang†, and Xiaoyan Zhu*
* State Key Lab. on Intelligent Technology and Systems,
Tsinghua National Lab. for Information Science and Technology,
Dept. of Computer Science and Technology, Tsinghua University
† College of Computer Science and Technology, Zhejiang University
feng-j13@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn
yangya@zju.edu.cn, zxy-dcs@tsinghua.edu.cn

Table 1

Method	Triple	Path	Edge
NTN(Socher et al., 2013)	✓	×	×
TransE(Bordes et al., 2013)	✓	×	×
TransH(Wang et al., 2014)	✓	×	×
TransR(Lin et al., 2015b)	✓	×	×
TransD(Ji et al., 2015)	✓	×	×
TranSparse(Ji et al., 2016)	✓	×	×
PTransE(Lin et al., 2015a)	✓	✓	×
Traversing(Gu et al., 2015)	✓	✓	×
GAKE(ours)	✓	✓	✓

- knowledge baseを有向グラフとして定式化
- Entity と Relation に対して3種類の context (Neighbor, Edge, Path) Embedding を学習
- Attention mechanism を導入し各 context の重みも学習

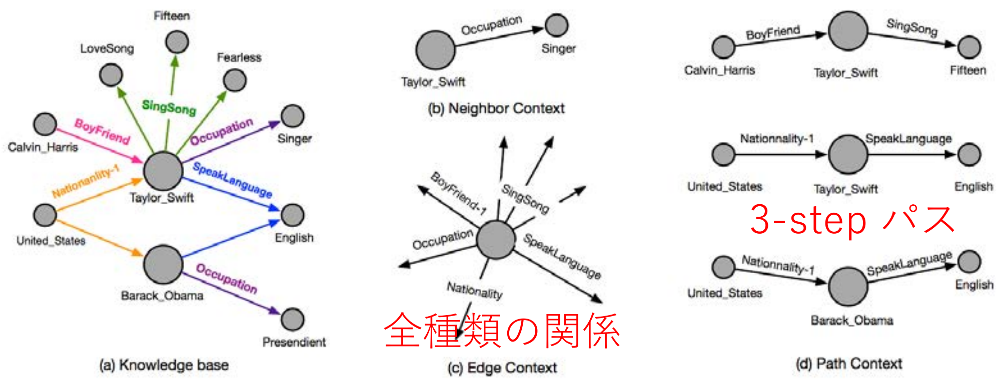
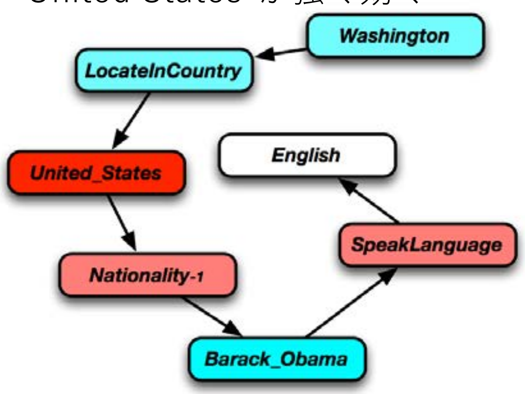
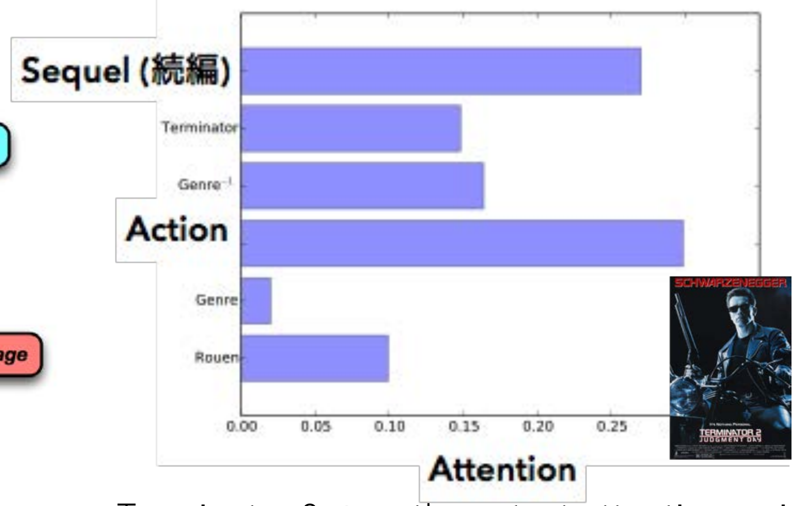


Figure 1: An illustration of three types of graph context, given by a knowledge base.

"English"を予想するとき
"United States"が強く効く





Terminator 2 の path context attention weights

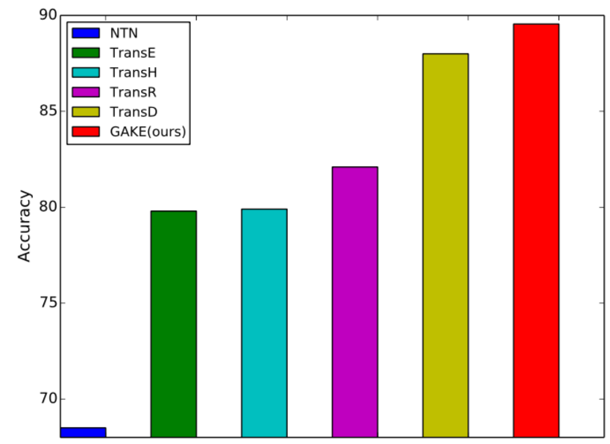


Figure 3: Evaluation results of triple classification.

Probabilistic Prototype Model for Serendipitous Property Mining

予期していなかったが関連する

Taesung Lee* IBM T. J. Watson Research Center taesung.lee@ibm.com

Seung-won Hwang Yonsei University seungwonh@yonsei.ac.kr

Zhongyuan Wang† Microsoft Research wzhy@outlook.com

- Serendipitous 情報の発見→楽しめるクイズ
- プロトタイプ理論に基づきknowledge baseを活用
  Probase：数百万のカテゴリとそれらの情報（entities, is-a, properties, typicalitiesなど）
- 従来の決定論的モデルDM（頻度ベース）に対し，確率的アプローチPMでプロトタイプを構築

Figure 1: Bing quiz scenario used by DM (Lee et al., 2016).

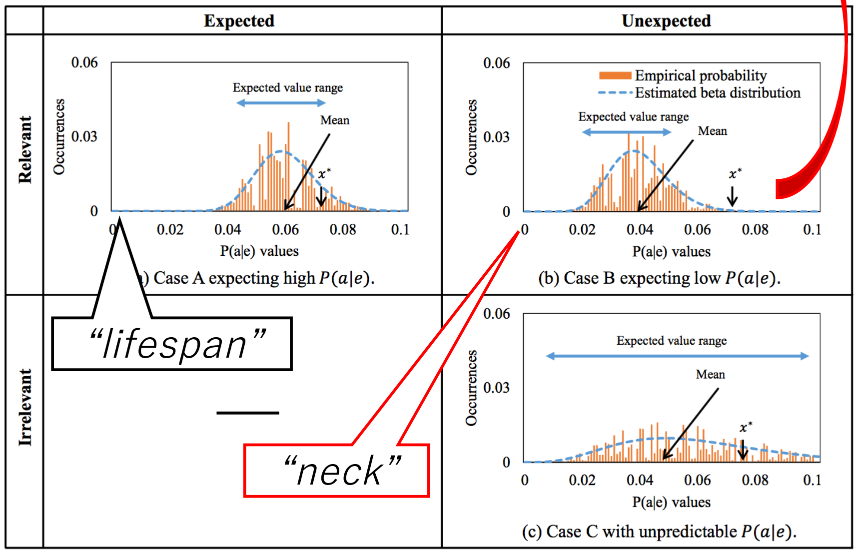


Figure 2: Distribution of $X_{a,c}$.

Table 2: Mined unexpected entity-property pairs for diverse categories.

Category	PM	DM
mammal	(marsupial, embryo), (giraffe, heart), (chimpanzee, brain), ...	(wolf, strength), (muskrat, best part), (otter, presence), ...
company	(microsoft, founder), (facebook, founder), (amazon.com, success), ...	(intel, fourth piece), (dell, model number), (coca-cola, original color), ...
country	(china, great wall), (china, population), (india, population), ...	(china, great wall), (china, choices), (india, reserve bank), ...
metal	(copper, resistivity), (gold, price), (gold, purity), (lead, density), ...	(copper, discovery), (lead, presence), (iron, presence), (aluminum, presence), ...
drug	(cocaine, price), (marijuana, legalization), (marijuana, odor), ...	(marijuana, 80 pounds), (cocaine, freebase form), (cocaine, last use), ...

Table 4: Average question relevance and serendipity at N.

Method	*Rel.*@5	*Rel.*@10	*SRDP*@5	*SRDP*@10
PM	**0.6689**	**0.6607**	**0.6662**	**0.6275**
DM	0.5672	0.5820	0.5562	0.5819
TF-IDF	0.6190	0.6048	-	-

Political News Sentiment Analysis for Under-resourced Languages

ノルウェー語

Patrik F. Bakken, Terje A. Bratlie, Cristina Marco, Jon Atle Gulla
Norwegian University of Science and Technology
Trondheim, Norway
bakken.patrik@gmail.com, terje.a.bratlie@gmail.com,
{cristina.marco, jag}@ntnu.no

Under-resourced languageに対する政治記事の感情の検出

手法: 2段階のバイナリ分類

データセット: 3961個の政治カテゴリの小記事 (Norwegian online)

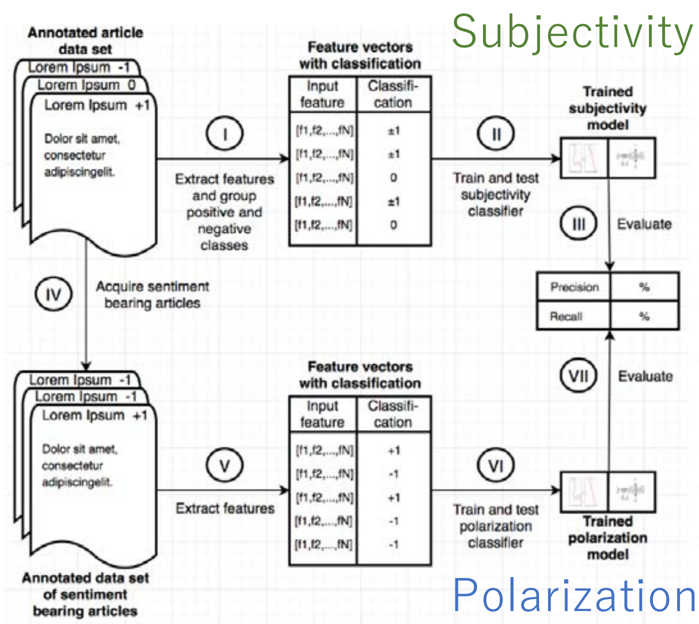


Figure 1: High-level system overview with two-step binary classification during testing and training.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		**67.1**
	✓	✓	✓	✓	66.6
				✓	59.9
RF	✓	✓	✓		62.5
	✓	✓	✓	✓	63.8
				✓	59.8
J48	✓	✓	✓		67.2
	✓	✓	✓	✓	67.0
				✓	61.8

Table 5: Subjectivity classification precision results with various feature combinations.

	PosCots	NeutCots	NegCots	Negations	Precision
NB	✓	✓	✓		72.8
	✓	✓	✓	✓	69.5
	✓		✓		**73.2**
	✓		✓	✓	71.4
				✓	57.4
RF	✓	✓	✓		64.2
	✓	✓	✓	✓	63.5
	✓		✓		66.1
	✓		✓	✓	64.2
				✓	55.6
J48	✓	✓	✓		72.2
	✓	✓	✓	✓	71.1
	✓		✓		72.3
	✓		✓	✓	71.6
				✓	—

Table 6: Polarity classification precision results with various feature combinations.

Class	Precision	Recall
Neutral	57.1%	**85%**
Subjective	**76%**	42.7%
Overall	67.1%	61.1%

Table 7: Precision and recall for subjectivity classification.

Class	Precision	Recall
Positive	**77.4%**	51.6%
Negative	70%	**88.2%**
Overall	73.2%	72.1%

Table 8: Precision and recall for polarity classification.

Automatically Processing Tweets from Gang-Involved Youth: Towards Detecting Loss and Aggression

Terra Blevins
Department of Computer Science
Columbia University
New York, NY, USA
tlb2145@columbia.edu

Robert Kwiatkowski
Department of Computer Science
Columbia University
New York, NY, USA
rjk2147@columbia.edu

Jamie Macbeth
Department of Electrical and Computer Systems Engineering
Fairfield University
Fairfield, CT, USA
jmacbeth@fairfield.edu

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY, USA
kathy@cs.columbia.edu

Desmond Patton
School of Social Work
Columbia University
New York, NY, USA
dp2787@columbia.edu

Owen Rambow
Center for Computational Learning Systems
Columbia University
New York, NY, USA
rambow@ccls.columbia.edu

- ギャングによる暴力を未然に検出，抑制
- 若くて力のある女性ギャングメンバーと彼女のコミュニケータからツイートを収集
- POS-tagger のドメイン適合
- 3 class classification
  - *loss* : 友人や家族を失った悲しみ
  - *aggression* : 脅し，報復
  - *others* : 他

💯	HUNDRED POINTS SYMBOL
🙏	PERSON WITH FOLDED HANDS
😂	FACE WITH TEARS OF JOY
😈	SMILING FACE WITH HORNS
💸	MONEY WITH WINGS

Figure 2: The five most common emojis in our dataset and their unabbreviated descriptions.

Experiment	Label	Precision	Recall	F-measure
TCF	**Aggression**	0.525	0.600	0.560
	Baseline (unigrams)	0.462	0.514	0.486
TCF	**Loss**	0.500	0.625	0.556
	Baseline (unigrams)	0.500	0.688	0.578
TCF	Average of **Aggression** and **Loss**	0.513	0.613	0.558
TCF	**Aggression** or **Loss**	0.588	0.800	0.678
CC	**Aggression**	0.471	0.923	0.623
CC	**Loss**	0.483	0.933	0.636
CC	Average of **Aggression** and **Loss**	0.477	0.928	0.630
BCS	**Aggression**	0.868	0.943	0.904
	Baseline (unigrams)	0.906	0.829	0.866
BCS	**Loss**	0.750	0.938	0.833
	Baseline (unigrams)	0.813	0.813	0.813

Table 3: Experimental Results on the test set. TCF is a Ternary Classification on the Full dataset (the three classes being **Aggression**, **Loss**, and neither). We provide separate results for our two classes of interest, as well as the macro-average for the two classes. We also give results for a binary task in which we collapse **Aggression** and **Loss** into one class (“**TCF Aggression** or **Loss**”). CC is the Cascading Classifier whose first step is an identification of **Aggression** or **Loss** (the system in line labeled “**TCF Aggression** or **Loss**”), and whose second step is a binary classification on the positively identified data points from the first step using the BCS system. We again provide separate results for our two classes of interest and the macro-average. BCS is Binary Classification on the aggression-loss Subset of the training data.

NTCIR13 organizerとして気になった研究 (ClinicalNLP WSの論文)

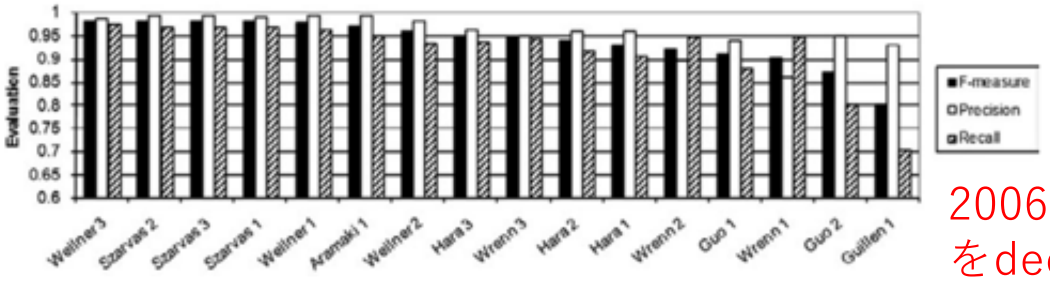
Feature-Augmented Neural Networks for Patient Note De-identification

I2b2 NLP (医療テキストコンペティション) Organizers

Ji Young Lee[1*], Franck Dernoncourt[1*], Özlem Uzuner[2], Peter Szolovits[1]
[1]MIT, [2]SUNY Albany
{jjylee,francky}@mit.edu, ouzuner@albany.edu, psz@mit.edu
* These authors contributed equally to this work.

電子カルテ文章

脱=特定
(匿名化)



2006年のShared task 課題（当時はCRFやSVM↓）
をdeepで再実験

Machine Learning method or Rules CRF Rules SVM SVM Rules, SVM Rules, SVM

典型的な単なるdeep でやっただけの研究かもしれない.
すでに精度は99.2%で人間の精度98%を超えている. この後, 何が起こるのだろうか？
タスク消滅？ 限りなく100%へ？ みまもりたい

	Binary HIPAA (optimized by F1-score)		
	Precision	Recall	F1-score
No feature	99.103	99.197	99.150
EHR features	99.100	99.304	99.202
All features	**99.213**	**99.306**	**99.259**

（言われているとおり）
素性のデザインの必要性
も僅少

←素朴な素性

←電子カルテ用に
デザインした素性

募集中

NTCIR13::MedWeb

The Fourth Medical NLP Shared Task

Home　Task　Dataset　Important Dates　Registration　Organizer　Link

The One and Only Medical Language Processing Contest

MedWeb NTCIR

Welcome to MedWeb (Medical Natural Language Processing for Web Document)

What's new: The task registration deadline is extended until March 31, 2017.

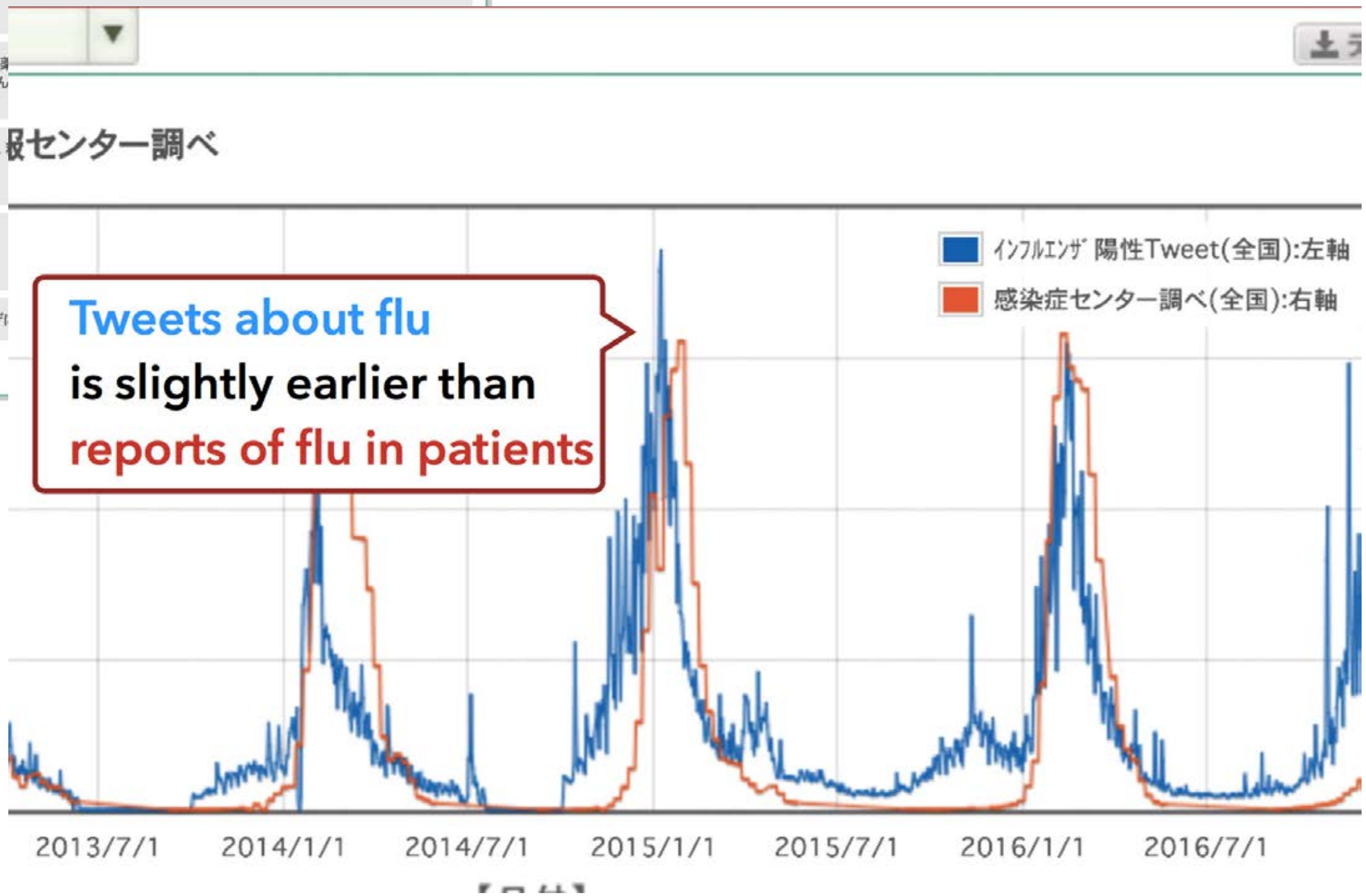
Recently, an increasing number of medical records is being stored in the form of electronic media instead of paper media -- making digital information processing in fields more and more necessary. Nowadays, this trend in information processing focuses not only on electronic health records but also on various data coming from patients. This data we call patient texts include social media texts, web blogs, and so on.

NTCIR-13 MedWeb (Medical Natural Language Processing for Web Document) task provides two different types of texts: Twitter message texts (in Japanese, English, and Chinese) and disease journal texts (in Japanese), and then requires to classify them or extract disease information from them. In detail, MedWeb consists of two subtasks: (1) Twitter subtask (in Japanese, English, and Chinese) and (2) Blog subtask (in Japanese). Since these subtask settings can be formalized as (1) binary-classification of disease/symptom-related texts and (2) medical codes labeling to disease or symptom names in patients' texts, the achievements of this task can almost be directly applied to a fundamental engine for actual applications.

http://mednlp.jp/medweb/NTCIR-13/

我々が発表した研究

インフルエンザサーベイランス（現状把握）



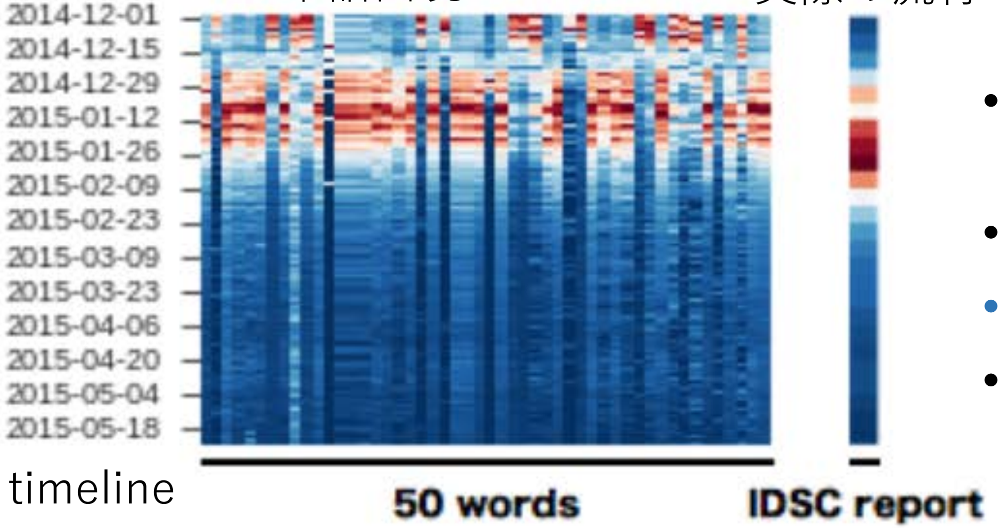
http://mednlp.jp/influ_map/

Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction

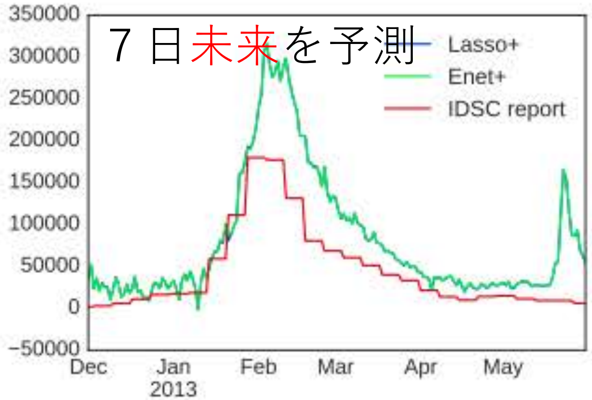
サーベイランス (now-cast) から予測 (fore-cast) へ

Hayate ISO, Shoko WAKAMIYA, Eiji ARAMAKI
Nara Institute of Science and Technology
{iso.hayate.id3,wakamiya,aramaki}@is.naist.jp

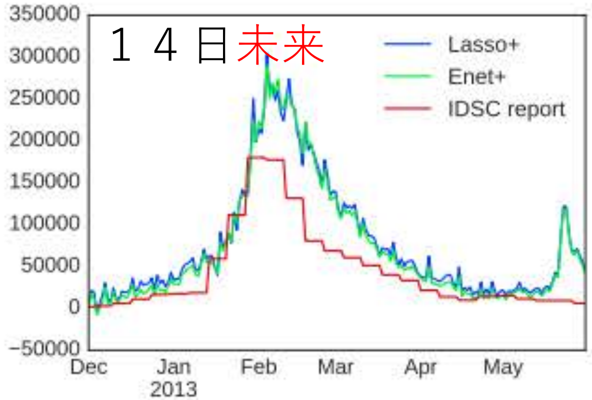
単語出現ヒートマップ 実際の流行



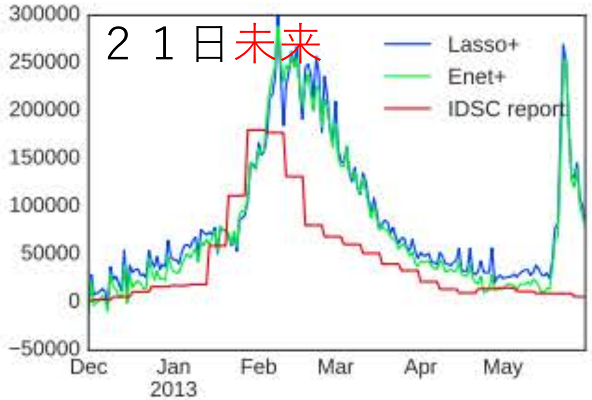
- 実際の流行の前に前兆となる単語がある
  - 例：「予防」「熱」など
- **現在**の単語で**現在**を予測　（これまで）
- 過去の単語で**現在**を予測　（本研究）
- **現在**の単語で未来を予測　（本研究）



(a) $\tau_{\min} = 7$ in Fig. 4a.



(b) $\tau_{\min} = 14$ in Fig. 4a.



(c) $\tau_{\min} = 21$ in Fig. 4a.