Glossary
グロッサリ

■ コールドスタート問題
スタートアップ問題ともいう．データが不足しているために，学習処理が不十分となり，結果としてシステムが適切な判断ができない問題を示す．　（壁谷佳典）

■ 集合間類似度（ダイス係数, ジャッカード係数, コサイン類似度, シンプソン係数）
2つの集合の類似度合いであり，基本的には共通の要素を持つ集合間ほど高くなるが，

ダイス係数$=\frac{2|X\cap Y|}{|X|+|Y|}$，　ジャッカード係数$=\frac{|X\cap Y|}{|X\cup Y|}$，

コサイン類似度$=\frac{|X\cap Y|}{\sqrt{|X||Y|}}$，シンプソン係数$=\frac{|X\cap Y|}{\min(|X|,|Y|)}$，

などさまざまな定義がある．たとえば，X={a, b, c, d}, Y={a, c, e}の場合，$|X|$=4, $|Y|$=3, $|X\cap Y|$=2, $|X\cup Y|$=5となる．　（土田正明）

■ テキスト含意認識
テキストTと仮説Hがそれぞれ自然文で与えられたとき「TならばH」と論理的にいえるか否かを自動判定するタスク．たとえば，「T：東京にいる」と「H：日本にいる」の場合，東京にいるならば日本に必然的にいるため「TならばH」といえる．本例のTとHの文を逆にすると「TならばH」はいえない．　（土田正明）

■ テキストマイニング
自然言語処理技術と統計的な分析技術を組み合わせ，大量の自然言語のテキストデータに内在する特徴を見つけ出す方法やシステム．たとえば，特定の商品に対する意見に偏って出現する単語を調査してその商品の改善案を検討したり，マーケティング活動として，最近急増した単語から世の中の流行を把握したりできる．　（土田正明）

■ ビジネスルール
ビジネス活動を導き，ビジネス運用上の判断や決定を行うために使用される基準．ビジネスルールはシステムに実装されるシステムロジックだけではなく，経営戦略レベルのものなども含まれる．　（壁谷佳典）

■ 分散表現
ベクトル表現の一種．Bag-of-wordsのような，1要素を1次元に割り当てる表現方法を局所表現と呼ぶのに対して，1要素をすべての次元に分散させた分布として表現する方法全般を分散表現と呼ぶ．一般的に局所表現と比べて，低次元かつ密なベクトルとなる．　（大倉俊平）

■ ランダムフォレスト
ランダムにサンプリングされたデータ・特徴量から学習した複数の決定木を統合することで精度を向上させる集団学習アルゴリズム．分類，回帰，クラスタリングに用いられる．決定木の学習は並列処理が可能なため，大規模データに対してもスケールしやすい．　（菊池匡晃）

■ ルールベース
「もし～ならば～」（IF-THEN）のような規則で記述された推論規則の集合．　（壁谷佳典）

■ A/Bテスト
Webサービスにおいて，2つの手法の良し悪しを判定する一般的なテスト手法．アクセスをランダムにグループAとBの2つに振り分け，それぞれ対応する手法で生成されたページを表示する．それぞれのユーザの反応を比較して，2つの手法の良し悪しを判定する．　（大倉俊平）

■ RNN（Recurrent Neural Network）
ニューラルネットワークの構造の一種．自身の出力を入力の一部として受け取り，繰り返し同じユニットを呼び出す再帰構造を持つことを特徴とする．テキスト処理や音声処理など系列長が不定の入力を扱うためによく採用される構造である．　（大倉俊平）