Glossary
グロッサリ

■ データサイエンティスト
データ分析（アナリティクス）を通して価値を創造するプロフェッショナル．データサイエンティストと呼ばれる人の多くはサイエンティストというよりは，むしろプラクティショナ（実務家）である．　（丸山　宏）

■ クラウドソーシング
インターネットを通して個人が仕事の委託を受ける，労働市場の仕組み．ここでいうクラウド（crowd）は「群衆」の意味であり「雲」（cloud）ではない．Webデザインや翻訳など専門スキルを要するもの，文字認識の後処理など単純作業を行うもの，などがある．　（丸山　宏）

■ IoT（Internet of Things）
インターネットに，通常のPCやサーバに加えて，いわゆるコンピュータでない多くのデバイスが接続され，物理世界と密に連携するシステム，またそのトレンドを指す言葉．CPS（Cyber-Physical Systems），IoE（Internet of Everything）などと呼ばれることもある．（山田　敦）

■ ウェアラブルセンサ
ウェアラブルセンサとは，腕時計型やメガネ型など，直接身につけて持ち歩くことができるコンピュータを指す．センサやカメラなどを搭載し，人の振舞いを計測する．これによって，ユーザの嗜好や生活スタイルに合ったアドバイスを提供することができる．　（佐藤信夫）

■ CRISP-DM
Cross Industry Standard Process for Data Miningの略．データマイニングのための業界非依存の標準プロセス．ここでいうデータマイニングは，データ分析（アナリティクス）一般を指す．欧州ESPRITプロジェクトの一環として作られた．　（山田　敦）

■ 予測モデル
入力データから出力の反応を数理的に表現した関数で，線形，決定木・回帰木，ニューラルネットワークなどが代表例．数値の予測モデルは回帰モデル，カテゴリの予測モデルは判別モデルと呼ばれる．万能な予測モデルはなくデータとの相性や精度・分かりやすさに違いがある．　（藤巻遼平）

■ 説明変数
予測モデルの入力となる変数．線形予測モデル$y=ax+b$では，xを説明変数，yを被説明変数と呼ぶ．説明変数は，特徴量，属性とも呼ばれる．高精度な予測を行うためには，予測モデルの選定以上に，適切な説明変数を設計することが重要である．　（藤巻遼平）

■ 情報量基準
統計モデルのデータへの当てはまりとモデルの複雑度を考慮して，そのモデルの良さを測る指標．情報量基準に従ってモデルを選ぶことで，モデルがデータに過適合する（過学習）を防ぐことができる．正則モデルを対象とした赤池情報量基準，ベイズ情報量基準から，非正則モデルを扱えるWAICやFICといった基準がある．（藤巻遼平）

■ BI（Business Intelligence）
ビジネス上のデータを収集，分析，可視化することで，経営の意思決定に役立てる手法．歴史的にはビッグデータブームの前に現れた概念であり，レポーティングに主眼があるのに対して，現在のデータ分析（アナリティクス）は，予測や最適化も含めたより広い概念である．（山田　敦）

■ 的中率と捕捉率
予測をするシステムの精度を示す指標．たとえば故障予測の場合，的中率は故障と予測した事例の何％が実際の故障だったかを示し，捕捉率は，実際の故障の何％を正しく予測できたかを示す．的中率と捕捉率は，通常トレードオフの関係にある．　（河本　薫）

■ ETL，データクレンジング
ETLは外部の情報源からデータを抽出（Extract），それを必要に応じて変換（Transform）し，分析対象データとして書き出す（Load）という一連の処理．デー

タクレンジングはデータの重複・誤記や異常値・欠損値等を加工して分析対象データの品質を高める処理．分析の前処理として，このようなデータの準備・整備が必要になる．（福島俊一）