データサイエンス入門

［概要］　膨大なデータを収集して傾向など分析し、社会やビジネス課題の解決につなげるデータサイエンスがデジタル時代を生きる社会人のリテラシーとして必要な時代になっています。

このコースでは、いま求められているデータサイエンスの基礎知識をコンパクトに身に着けることができます。データサイエンスシステムの概要・全体構成・構成要素、データ分析に必要な統計解析や機械学習の概要を理解することができます。

なお、「SE のためのデジタルシステム入門（ビッグデータ分析編）、（機械学習・ディープラーニング編）」両コースと重複する部分があるため、これらのコースとの重複受講は避けてください。

［学習目標］　・データサイエンスの概要、データサイエンスシステムの全体構成・構成要素を理解する
・データ分析手法としての機械学習の基礎を理解する

［受講対象者］・データサイエンスの基礎を理解したい方
・データサイエンスシステムをこれから企画・開発・運用する方

［前提条件］　特になし

［日数］　１日間

［形式］　講義＋演習（講義中心、90：10）

［研修に必要なもの］

講師用　：（オンラインの場合）オンライン研修環境
（オフラインの場合）プロジェクター、ホワイトボード

受講者用：（オンラインの場合）オンライン研修環境、
Word の使える PC（確認クイズをファイル配布する場合）

［タイムスケジュール］

日程	内容
	Ⅰ． データサイエンスの概要 １．データサイエンスシステムとは データサイエンスの定義、必要な知識領域、扱うデータ ２．データサイエンスシステムの構成 全体構成、構成要素の概要 ３．分散処理の基礎 分散処理の必要性、分散計算：Hadoop MapReduce, Hive 分散ストレージ：Hadoop HDFS、オブジェクトストレージ、 レプリケーション、リードレプリカ、データのローカリティ ４．データ収集 バッチデータ収集とストリームデータ収集 バッチデータ収集 ファイルデータ収集、SQL による収集、データ出力による収集、 更新ログ同期による収集、その他 ストリームデータ収集： 各プロセスの構築方法、データレイクの構築方法 ５．データ蓄積 一次加工：一次加工の構築方法 データウェアハウス:オペレーショナル DB とアナルティック DB 列指向フォーマット SQL on Hadoop その他のアナルティック DB ６．データ分析 データ分析手法としての統計解析と機械学習 統計解析の概要：統計解析の種類、記述統計、推測統計、 ベイズ統計 データ活用方法：アドホック分析、データ可視化、 データアプリケーション ７．メタデータ管理 静的メタデータ管理 動的メタデータ管理 データ構造管理 データリネージ管理 （参考）3 大クラウドにおけるビッグデータ分析システム構成 AWS、GCP、Microsoft Azure （確認クイズ）

日程	内容
	Ⅱ．機械学習 １．機械学習の概要 機械学習とは 機械学習の分類：教師あり学習、教師なし学習、強化学習 ２．機械学習システム 概要 設計・学習 モデルの設計、前処理、学習、モデルの評価、 ハイパーパラメータの調整 システムへの組み込み 機械学習システム構築環境 開発言語、機械学習用ライブラリ、 ディープラーニングのフレームワーク、機械学習サービス ３．機械学習の代表的なアルゴリズム 回帰分析 サポートベクターマシン カーネル法 決定木 ロジスティック回帰 k 平均法 ４．ディープラーニング ディープラーニングとは： ニューラルネットワーク、AI のニューラルネットワーク、 ディープニューラルネットワーク、ディープラーニング ディープニューラルネットワークのコア技術： パーセプトロン、誤差逆伝搬法、 ニューラルネットワークの最適化 ディープニューラルネットワークの代表的なアルゴリズム： 畳み込みニューラルネットワーク（CNN） 再帰型ニューラルネットワーク（RNN） オートエンコーダ （確認クイズ）