

東京大学グローバル消費者インテリジェンス寄付講座

第1回目

# データサイエンス入門講義

## 本日のお話

- 01 データサイエンスとは？
- 02 この講義の概略
- 03 データ分析を学習する上で大事なこと
- 04 講義事前学習
- 05 次回予告と参考文献

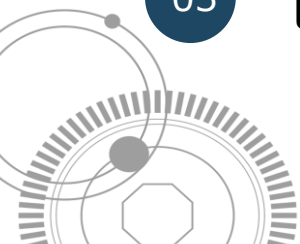


## 本講義のゴール



ビジネスでデータ活用をするための  
基本的なスキルを身に付けること

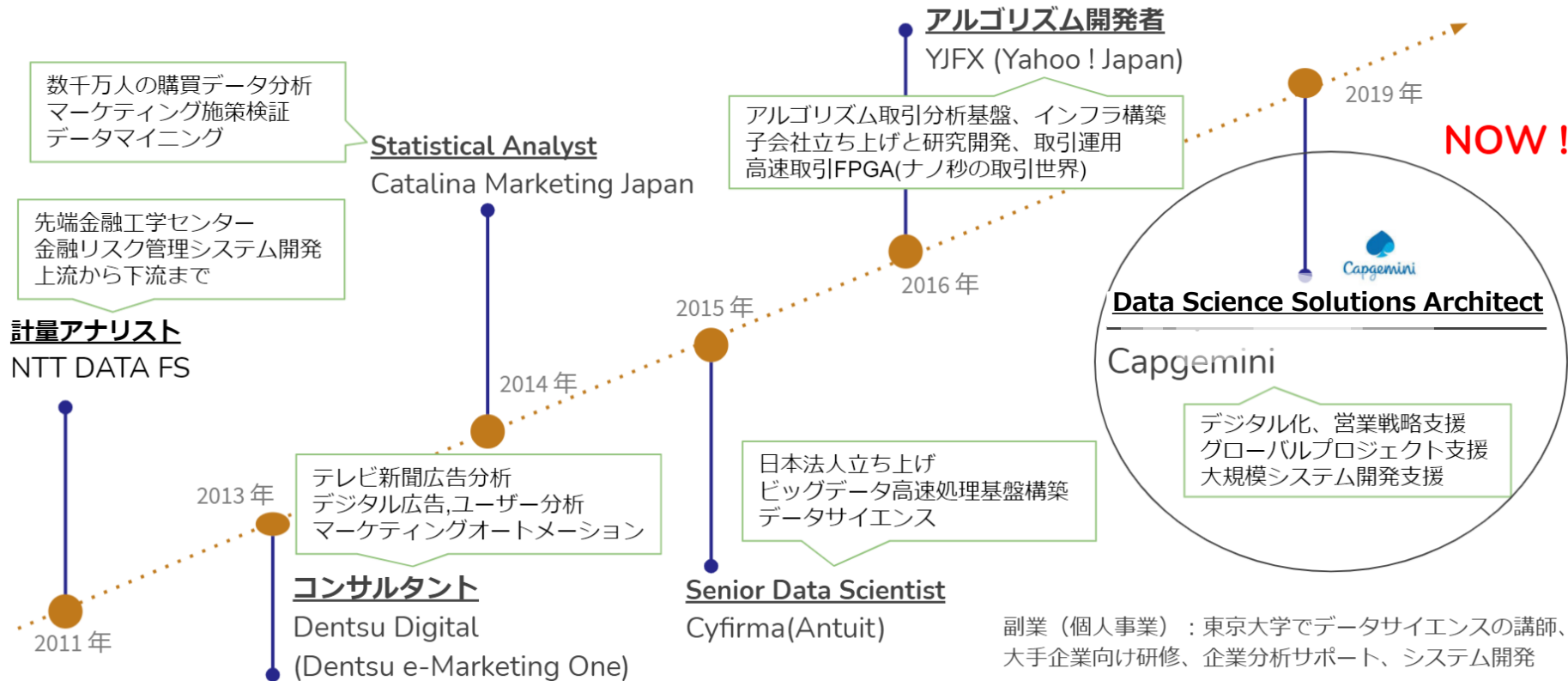
- 01 データサイエンスに必要な実装スキル（PythonやSQL・機械学習）
- 02 データサイエンスに必要なインフラスキル（DB）
- 03 ビジネスに必要なマーケティング思考（マーケティング・外部講師）



# 自己紹介

## 塚本 邦尊

職歴：データ戦略から環境構築、  
分析、システム・アルゴリズム開発、  
運用まで経験



# Capgeminiグループの概要

## Capgeminiグループ:

- 世界の大企業のおよそ2/3が弊社のクライアント
- 世界でトップ5のコンサルティングファーム
- 50年以上の歴史と業界固有の深い専門知識を基盤にサービスを提供
- Capgemini SE（本社）はパリに所在あり、ユーロネクスト・パリCAC40※の一つ  
ISIN code: FR0000125338

※ユーロネクスト・パリ（2000年以前はパリ証券取引所）における株価指数。  
同取引所に 上場されている株式銘柄のうち、時価総額上位40銘柄で構成される



## 従業員数/拠点数/国籍

従業員数

**269,470**

人以上

拠点数

**50**カ国以上

従業員の国籍数

**120**以上

(2019年12月末時点)



## 2019 full-year results

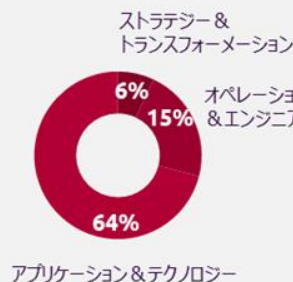
売上高

**€170億** (ユーロ)

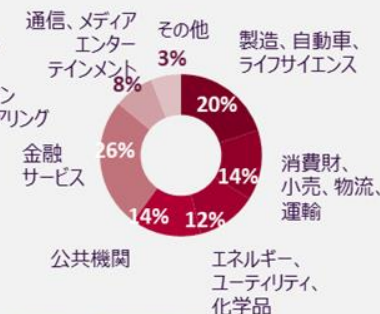
営業利益率※

**12.3%**

### 2019年度売上高内訳



### 業種別売上



※2019年買収のAltran社を除いた割合





データサイエンスに関するお話

# 01 データサイエンスとは

# Q:世の中にはどんなデータがあるの？

世の中には様々なデータがあり、日々あなたのデータが取られています！

## オンライン (ウェブ上での行動など)

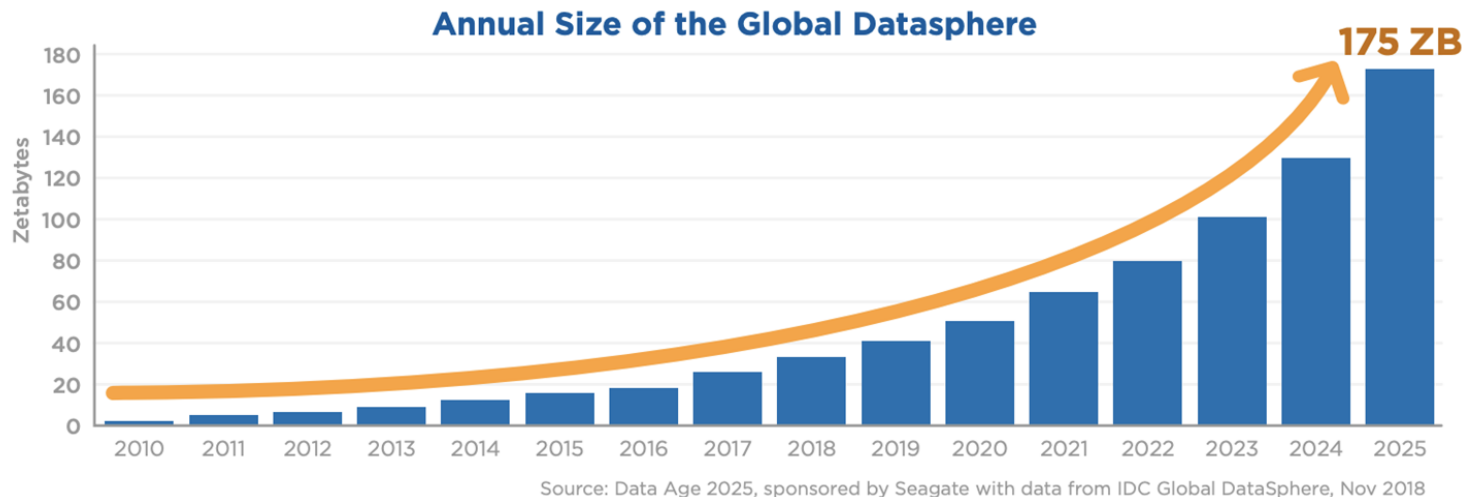
- ・検索エンジン
- ・ECサイトでの買い物
- ・SNS
- ・メール
- ・LINE、ゲーム（ポケモンGOなど）
- ・ウェブサイトサービス  
(ウェブの閲覧、広告閲覧) など

## オフライン (リアルでの行動など)

- ・コンビニやスーパーでの買い物データ
- ・公共交通機関
- ・気象データ
- ・スポーツ（野球、サッカーなど）
- ・医療・ヘルスケア
- ・HRデータなど

# データは日々、膨大に増えている

Figure 1 - Annual Size of the Global Datasphere



\*参照 『The Digitization of the World From Edge to Core (David Reinsel – John Gantz – John Rydning) 』 2018 IDC



## Q: データを活用している業界は？

### 金融

トレーディング  
リスク管理  
ニュースのテキスト解析  
不正取引

### ECサイト

POSデータ  
ウェブアクセス

### マーケティング

ウェブ広告  
TV効果  
施策の検証

### ゲーム

ユーザーの行動分析

### その他

ネット・医療  
ヘルスケア・放送  
人材・教育・スポーツ  
法曹等

### メーカー

自動車  
精密機器

### セキュリティ

異常検知  
不正アクセス

# データ分析ができれば何がうれしい？

サービスを利用しているユーザーを増やすことができる

▼  
売り上げを上げる

効率を上げることができる

▼  
コストを下げる

新しい価値を提供する（新しい示唆を得る）

PDCAを回し、サービスを改善していく

etc...



# なぜデータ分析ができてないのか？

そもそもデータがない！



今あるデータで何ができるかわかっていない

データはあるが、整理できてない

データ分析できそうだけど、  
リソース（人材）がない

元のデータがごちゃごちゃしすぎて大変！

データ分析の目的が曖昧  
（手段が目的化している）



Q:データってどんなもの？



データ サンプル

などのキーワードで調べてみましょう！



# 本講義で扱うデータの例

所属、性別、年齢、成績、etc

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
MS	M	20	U	LE3	A	2	2	services	services	...	5	5	4	4	5	4	11	9	9	9
MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	3	14	16	16
MS	M	21	R	GT3	T	1	1	other	other	...	5	5	3	3	3	3	3	10	8	7
MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	0	11	12	10
MS	M	19	U	LE3	T	1	1	other	at_home	...	3	2	3	3	3	5	5	8	9	9



# データ分析の8～9割はモデリングの前の作業に・・・

## ☹ データとして厄介な例

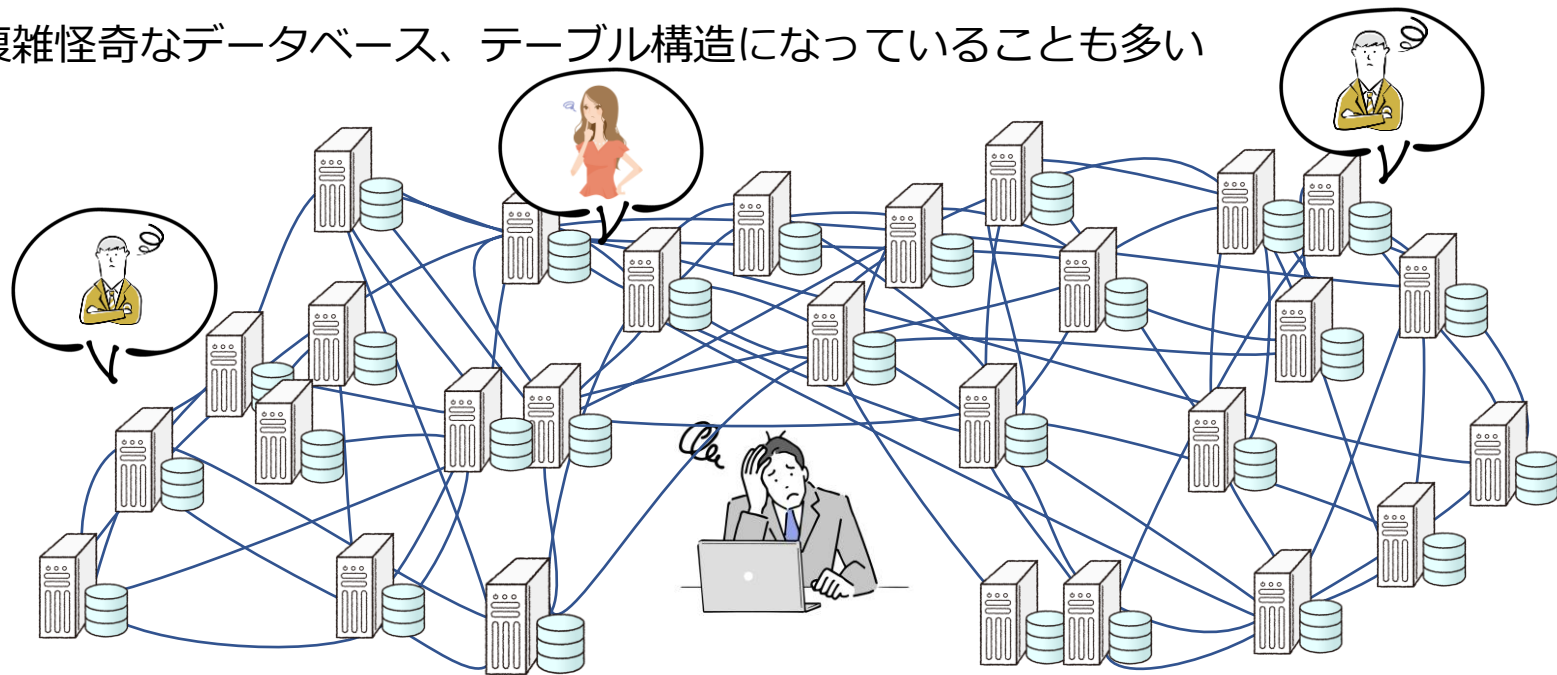
(※以下は金融の半構造化データ、テラバイト級/monthで約20億行以上)

2016-01-01 10:10:10.000 8=FIX.4.4/x019=122/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B/x019=122  
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072/x01  
/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020 /x0110=072/x012016-01-  
0110:10:10.0008=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B  
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072  
/x01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020/x0110=072/x01 2016-01-01  
10:10:10.000 8=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x01 2016-01-01 10:10:10.000 8=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020/x0110=072/x01

# 89%もの企業がレガシーシステムに悩んでいる

✓途中で改修されたシステムとの連携がたかつさん

✓複雑怪奇なデータベース、テーブル構造になっていることも多い



# 「何のためにデータを使うのか？」という目的や戦略が重要

✓ 実は、ほとんどのデータはそのままでは使えない、データはすぐに金にならない

✓ データをどういう目的で使うかが重要

顧客満足	・顧客満足度 ・NPS (ネットプロモータースコア) ・クレーム数 等	・SNS上の情報 ・テキスト等の定性情報 ・顧客音声等の感情分析 等
顧客動向	・顧客属性 ・購買データ ・顧客動線 等	・NFCデータ (交通ICカード等) ・気象データ ・Webアクセスログ ・カメラデータ 等
スピード	・納期遵守率 ・顧客待ち時間 ・所要時間 等	
量・効率	・単位時間あたりの生産量/ 販売量 ・在庫回転率 ・設備稼働率 等	
品質	・欠陥数 (率) ・ミス数 (率) ・予測精度 等	・機械/センサー/ITログ ・通信ログ ・M2M連携 ・GPS等のNSSデータ ・従業員の移動データ (動線、速度) ・気象/自然データ 等

プロセス改善に役立つ多くのデータ

図4-9 プロセス改善に役立つ新たなデータたち



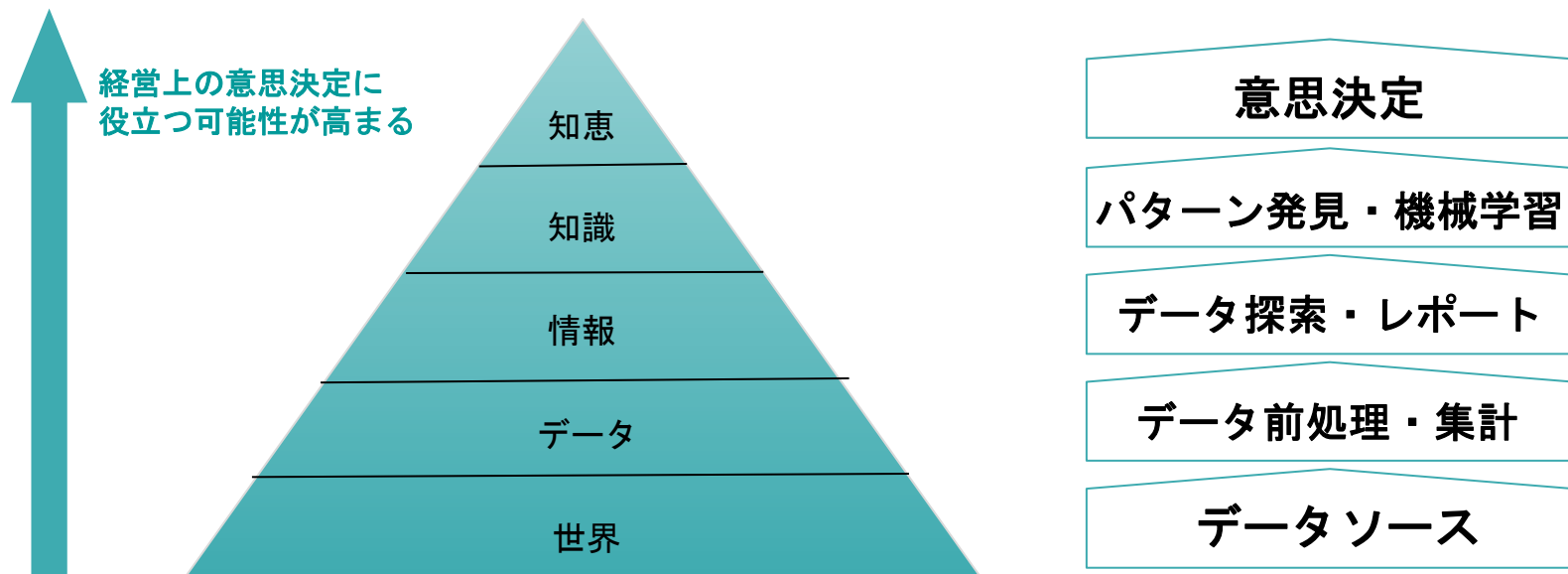
※参照『デジタルマーケティングの定石  
(垣内勇威)』日本実業出版社



※参照『ビジネスプロセスの教科書 (山本政樹)』  
東洋経済新聞社

# データ分析の結果を意思決定に

✓単にデータを蓄積するだけではなく、知恵にまで昇華し意思決定に役立てることが重要



＊参考『データサイエンス（ジョン・D・ケレハーら著）』NEWTON PRESS

＊出典元：キチン2014年a、ハン、カンバー、ペイ2011年

# データサイエンスとは

データサイエンス

=

データ

+

サイエンス(科学)

厄介な（？）データから、科学的なアプローチを使って示唆を得て、世の中の様々なビジネスの問題解決に挑戦し、データから価値に変換する（科学的なアプローチとは、数学や統計、ITの技術を使うこと）

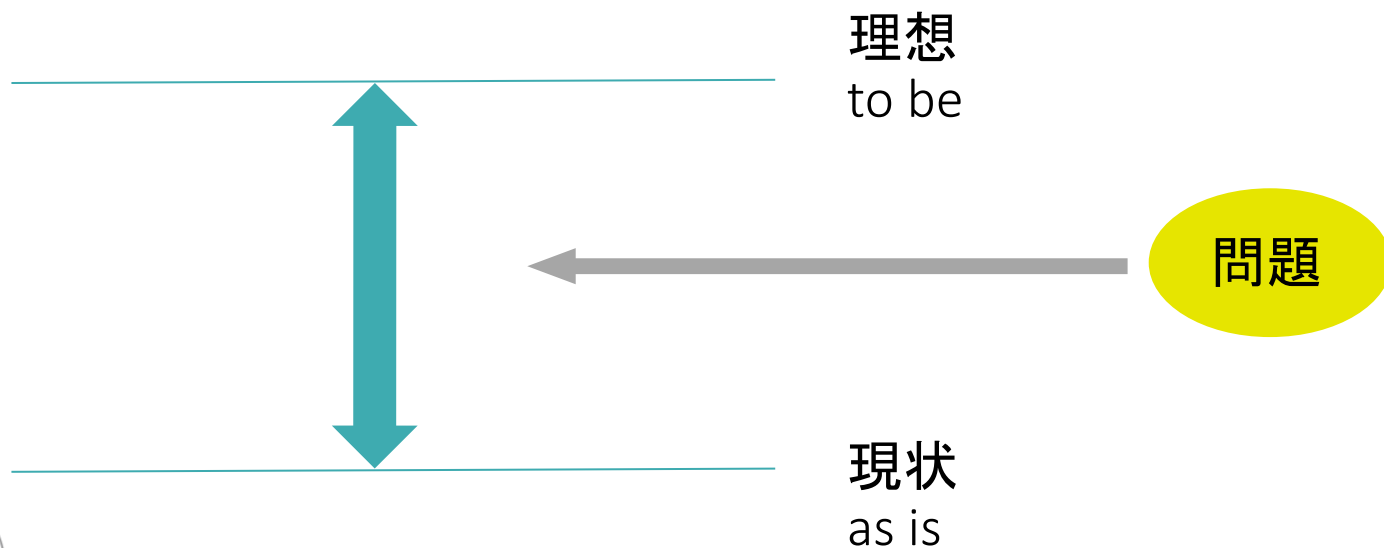
これらの仕事をする人が、**データサイエンティスト**



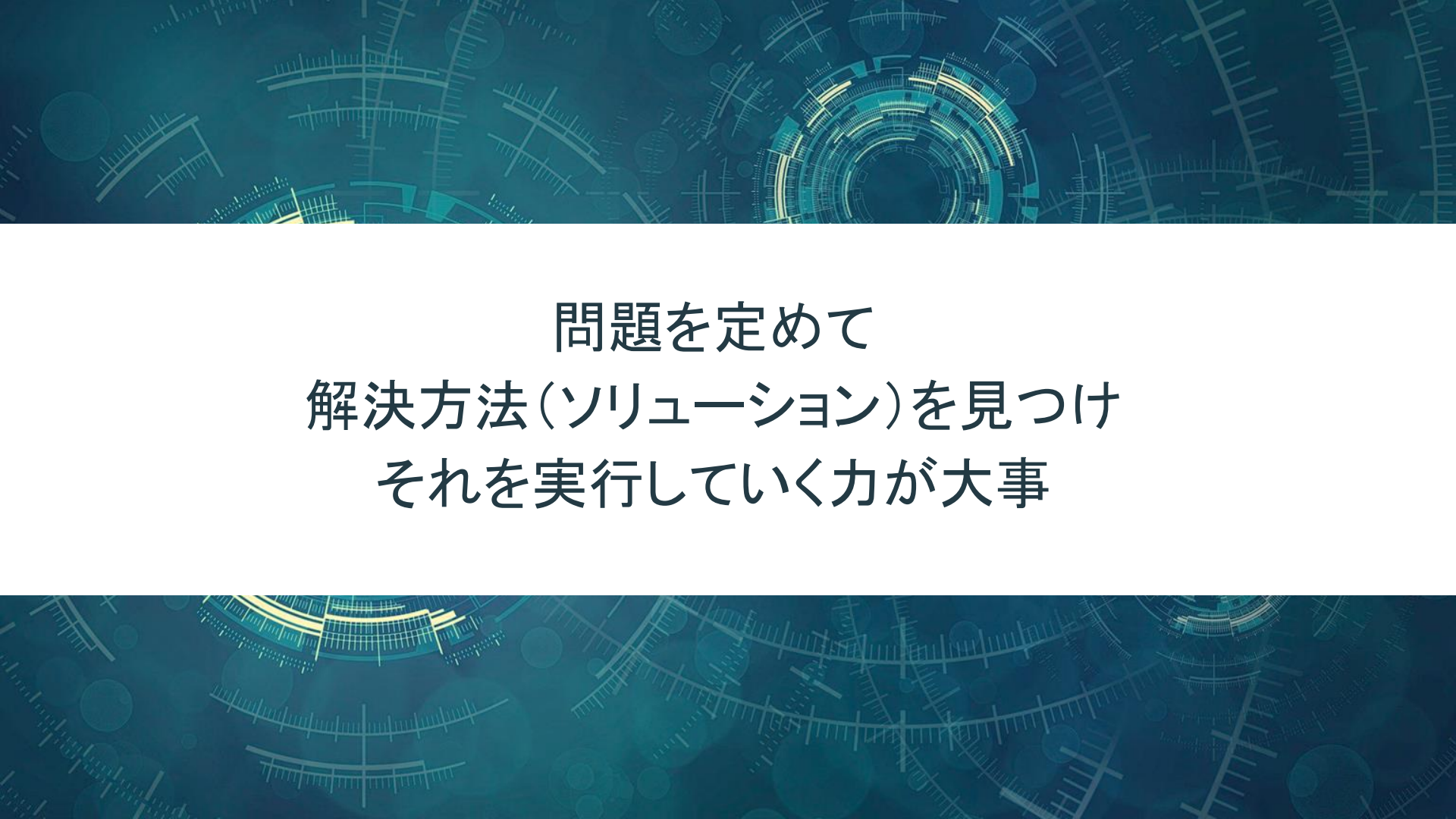
# 問題とは？

問題

理想と現状のギャップ



\*参照『問題解決プロフェッショナル「思考と技術」』ダイヤモンド社



問題を定めて  
解決方法(ソリューション)を見つけ  
それを実行していく力が大事

# データサイエンスの実態は？

✓データサイエンティストは注目されている職種だが、その実態は？

2009年2月、米グーグル チーフエコノミストのハル・バリアン氏は「今後10年で最もセクシーな職業は統計家だ」と発言。

さらに、米ハーバード・ビジネス・レビューの2012年10月号はデータサイエンティストを「21世紀で最もセクシーな職業」と表現した

2011年5月に米マッキンゼーが公表した

「McKinsey Global Institute「Big data: The next frontier for innovation, competition, and productivity」によると、米国では2018年までに、高度なアナリティクス・スキルを持つ人材が14万～19万人不足で、大規模なデータセットのアナリティクスを活用し意思決定のできるマネージャーやアナリストが150万人不足すると算出

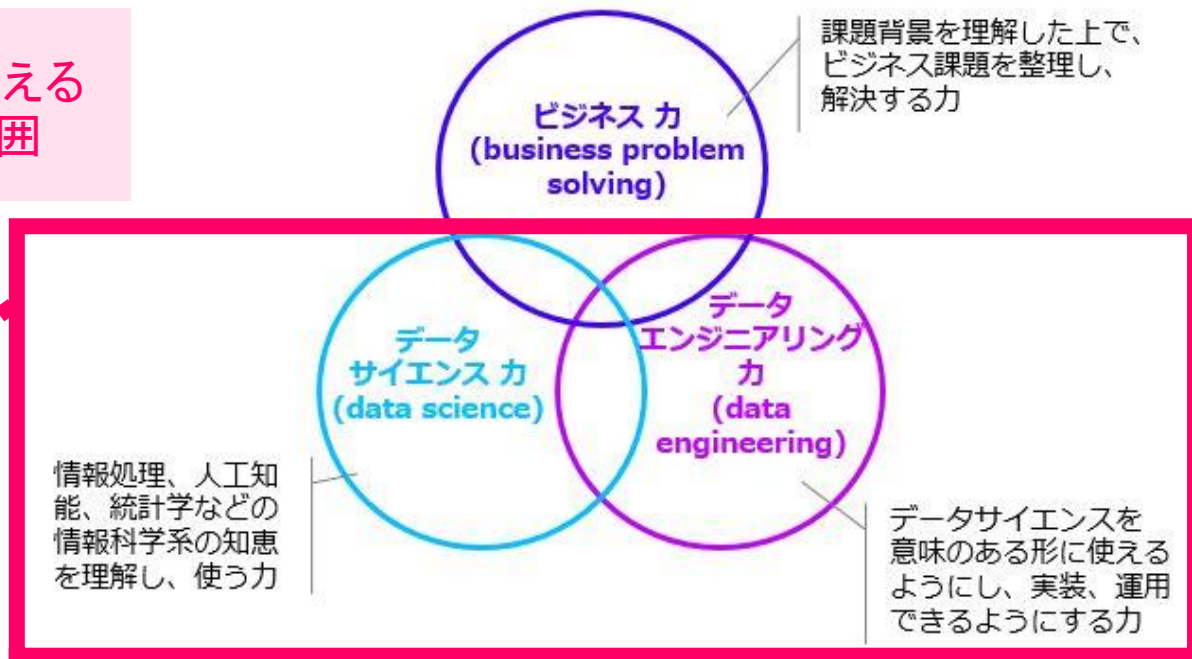
※でも最終的には  
「楽」したい  
モチベーションも  
あります

実態は泥臭くデータの確認、加工、検証をやっていく仕事

# データサイエンスの実態は？

## データサイエンティストに求められるスキルセット

本講義で鍛える  
スキル範囲

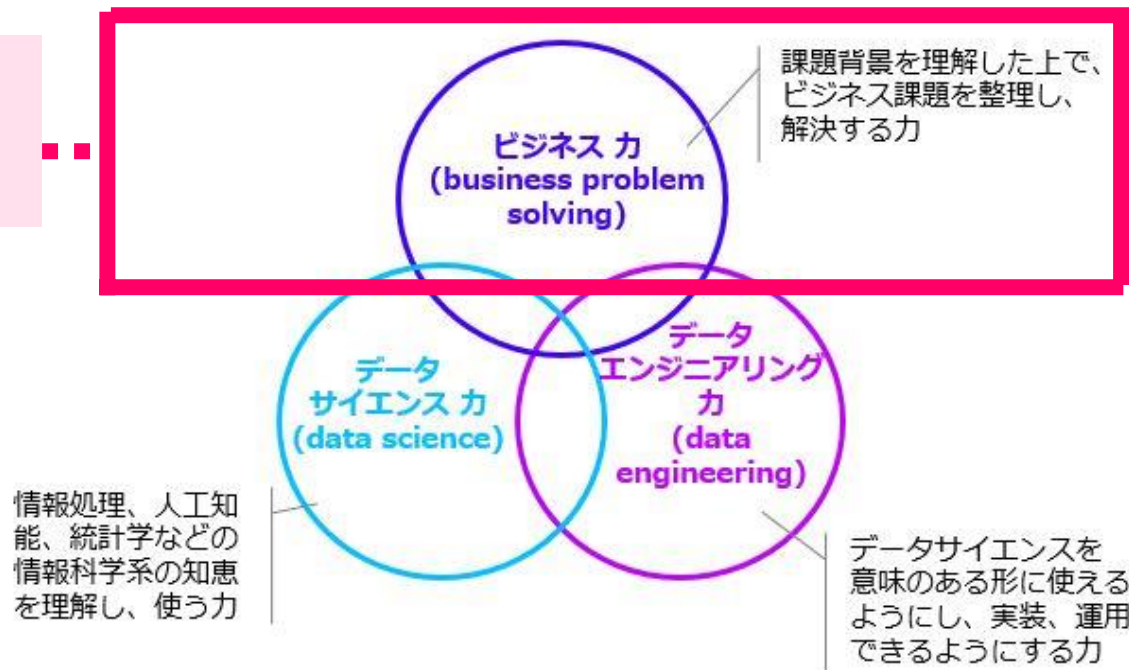


参照URL:<https://prt看es.jp/i/7312/8/resize/d7312-8-634846-2.jpg>

# データサイエンスの実態は？

## データサイエンティストに求められるスキルセット

実務視点が  
とても重要



参照URL:<https://prtimes.jp/i/7312/8/resize/d7312-8-634846-2.jpg>





# データ分析における**3つ**のポイント

01

仮説を持つ



ビジネス力・課題発見と解決方法

02

アクションにつながる



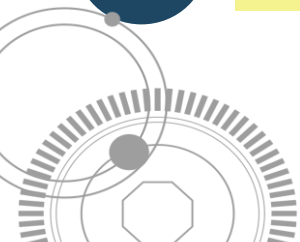
企画施策の実施・システム実装等

03

比較



統計力・データサイエンス





## データサイエンスに関する職

- 機械学習エンジニア
- クオオンツ
- アクチュアリー
- データアナリスト
- データエンジニア、など

## データサイエンスに関する選考・面接

- 企業と業界研究は様々
- やりたいことは何か？
- どんなスキルがあるか？
- 運(景気)

## その他必要そうなこと

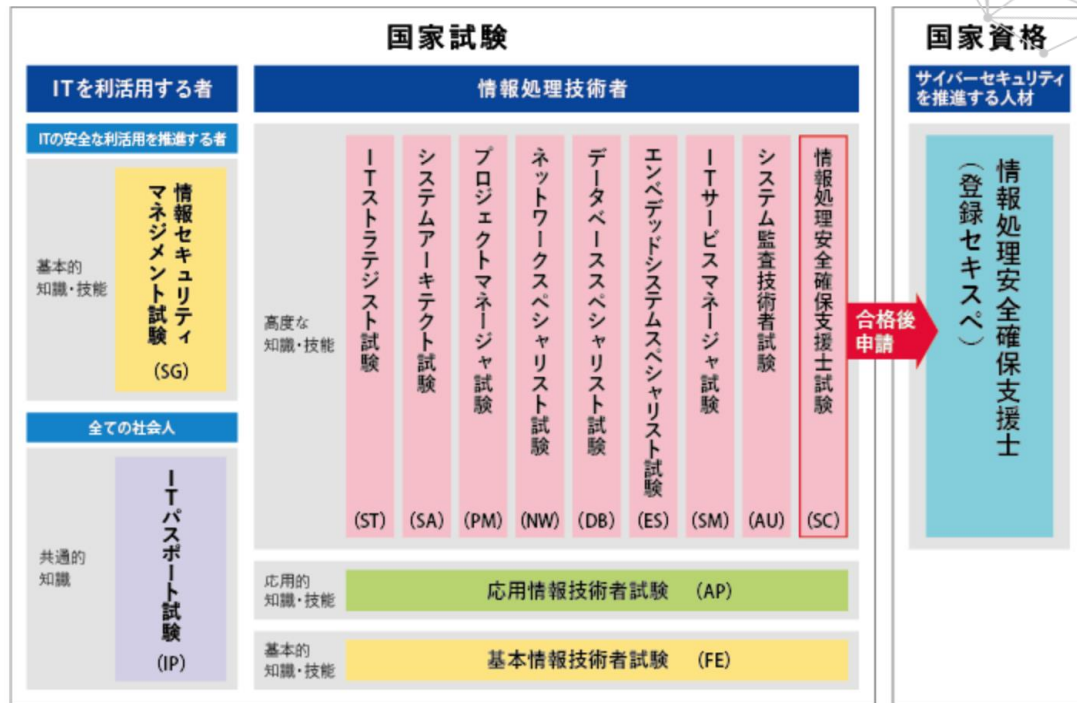
- **新しい技術を学ぶ意欲**
- 英語
- 教養



# 参考 資格試験(注意:必須ではありません)

## データエンジニア向け

- ✎ ～大学3年生：基本情報
- ✎ 大学4年生～大学院生：応用情報
- ✎ 余力がある人：  
データベーススペシャリスト  
ネットワークスペシャリスト



参考URL:[https://www.jitec.ipa.go.jp/1\\_11seido/seido\\_gaiyo.html](https://www.jitec.ipa.go.jp/1_11seido/seido_gaiyo.html)

# データサイエンスに関して、その他注目されているトピック



## データサイエンス視点

- ・ XAI (説明可能なAI)
- ・ 因果推論
- ・ 自然言語処理

- ・ AutoML
- ・ MLOps

## エンジニアリング視点

- ・ DevOps
- ・ アーキテクト
- ・ マイクロサービス
- ・ API

⇒ 昨今のトレンドは、「ローコード・ノーコード」

※〇〇カオスマップなどを参照したり、グーグルトレンドなどで調べてみてください



講義修了後は

なるべく早めに現場で働いてみる  
(バイトを募集している企業に自分から応募する)



## 参考:隠れたデータの存在

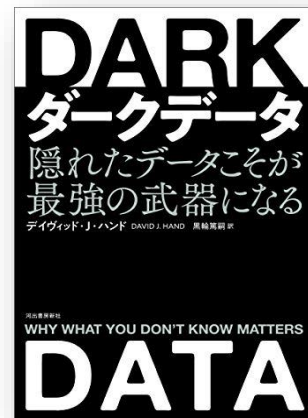


様々なデータを手に入れて分析できるようになったとはいえ  
保有しているデータだけに基づいた判断は誤っている可能性が高い！



見えているデータ

隠れたデータ



参考:『ダークデータ 隠れたデータこそが最強の武器になる』河出書房、デイビッド・J・ハンドより

# 参考：データ分析による失敗事例その 1

## 問題

シェア低下を食い止めるため、商品価格を変更するかどうか

## ソリューション

色々な分析を実施し、価格維持を提案

## 結果

シェアランキングのさらなる低下（1位からの転落）



＊参照『戦略コンサルタント仕事の  
本質と全技法（遠藤功）』  
東洋経済新聞社 p204



# 参考：データ分析による失敗事例その2



## 問題

不動産価格をどうやって決めるのか？

## ソリューション

AIを使って価格を予測

## 結果

ビジネスとして大失敗し、評価減額約570億円

※参照

「米不動産テック大手Zillowの大失敗に見るAI経営の教訓…「予測モデルの過信」「目標設定のミス」は他人事ではありません」

<https://www.businessinsider.jp/post-248629>

=> ブラックスワン（歴史の繰り返し）





## Q&Aセッション(1回目)



講義で何が必要で何をこの身につけることができるのか？

## 02

## 本講義の概要

# データサイエンティストに必要な3つの力



01

## ビジネス力

- ・マーケティングに関するデータを少し扱い、ビジネス的な観点も考えます


02

## データエンジニアリング力

- ・データを適切に処理、扱うためのプログラミング（PythonやSQLなど）スキル
- ・インフラ（DBなど）スキル

03

## データサイエンス力

- ・データから示唆等をえるための確率統計学
  - ・予測等をしたければ機械学習
- 

# データ分析を具体的に実行するために学ぶこと

受講対象者: Python、Jupyter経験、微分積分、線形代数、確率統計の基礎



第1回 導入(本日)

第2回 Pythonによる科学計算(Numpy)

第3回 Pythonによるデータ加工処理の基礎(Pandas)

第4回 Pythonによるデータ可視化の基礎(Matplotlib)

第5回 機械あり学習

第6回 機械なし学習





# データ分析を具体的に実行するために学ぶこと

受講対象者: Python、Jupyter経験、微分積分、線形代数、確率統計の基礎



第7回 SQL(データベース)

第8回 モデル検証とチューニング

第9回 特徴量エンジニアリング

第10回 マーケティング基礎、応用の一部

第11回 ゲスト講師

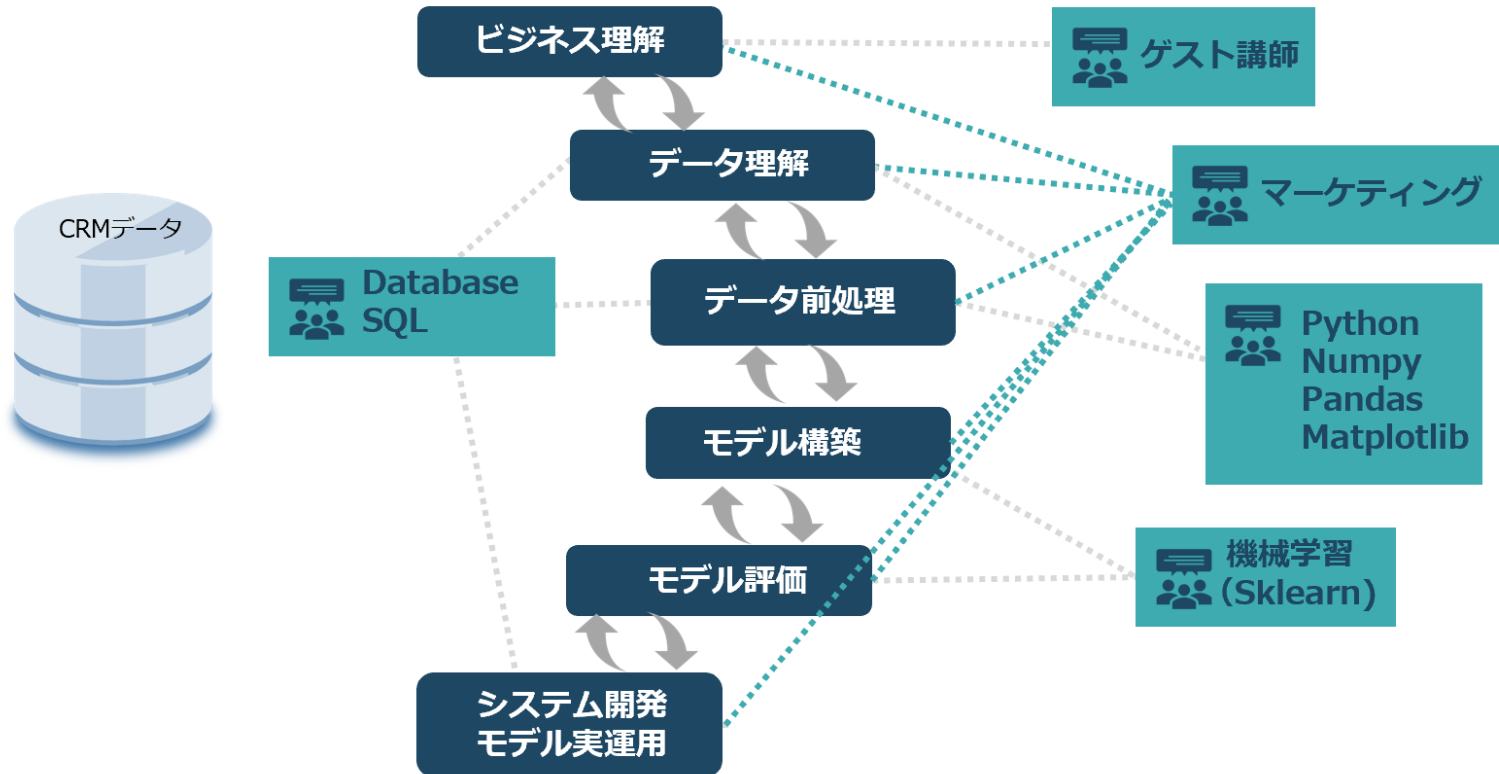
第12回 ゲスト講師

第13回 ゲスト講師



# 実データサイエンスプロジェクトと本講義の関係性

ビジネス理解からデータ確認と前処理、モデル構築と評価、実運用まで

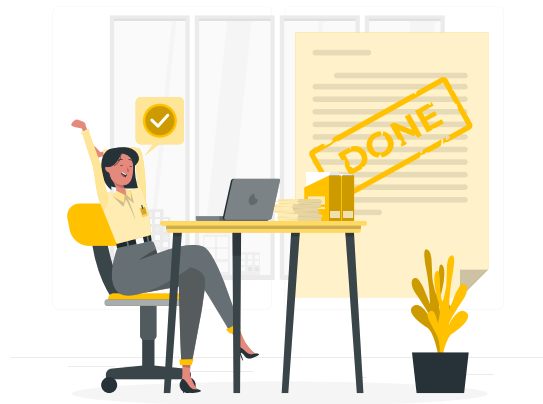




# 本講義の進め方・流れ

流れ 各回で学ぶ概略と理論 ▶ 実装の説明 ▶ 演習 ▶ 解説

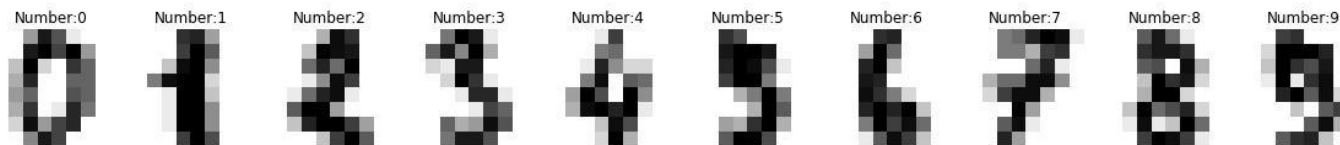
01. パワーポイントの資料とJupyterNote(Chapter~.jpynb)を中心に説明
02. 理論と実装（コード）の説明
03. 演習問題（JupyterNoteからピックアップ）
04. 演習の解説
05. 最後に宿題（基本的に毎回）



# 本講座で学ぶこと

Pythonを使って、色々なデータを読み込み、そのデータの加工処理、可視化、そして機械学習（AIの一部）のモデルの構築や検証ができるようになる

## 例 1 数字データの数値判定



## 例 2 数値データの予測モデルの構築（将来の価格予想）

過去10年の為替レートUSD/JPYデータを使って、  
明日のレートは111円になると予測する、など  
（ただし、必ずうまくいくとは限らない）

# 本講座で学ぶこと

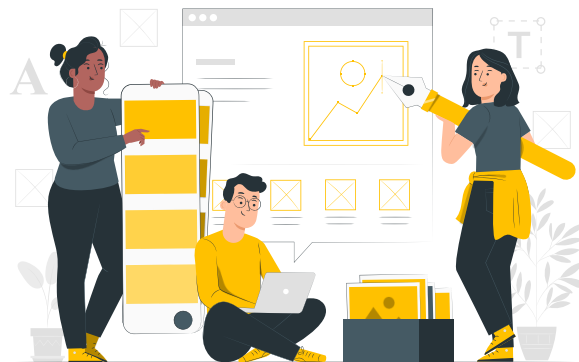
## 例 3 データの可視化

<https://dash-gallery.plotly.host/dash-uber-rides-demo/>

※ウーバーデータ

## 例 4 データの前処理

※次のページ参照



# 以下のデータを大量にZipファイルで渡されたらどうしますか？ (実装&目的)

金融の半構造化データテラバイト級/monthで約20億行以上

2016-01-01 10:10:10.000 8=FIX.4.4/x019=122/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B/x019=122  
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072/x01  
/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020 /x0110=072/x012016-01-  
0110:10:10.0008=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=2010022519:41:57.316/x0156=B  
/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=2010022519:39:52.020/x0110=072  
/x01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020/x0110=072/x01 2016-01-01  
10:10:10.000 8=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x01/x019=122/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0 /x0160=20100225-  
19:39:52.020/x0110=072/x01 2016-01-01 10:10:10.000 8=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-  
19:39:52.020/x0110=072/x018=FIX.4.4/x019=122/x0135=D/x0134=215/x0149=CLIENT12/x0152=20100225-  
19:41:57.316/x0156=B/x011=Marcel/x0111=13346/x0121=1/x0140=2/x0144=5/x0154=1/x0159=0/x0160=20100225-19:39:52.020/x0110=072/x01

# 本講座で学ぶこと

## ✓ データベースやSQLの基礎スキル

データベース、テーブル、RDBMSなど

## ✓ マーケティング思考(ユーザー理解)

マーケティング戦略、データ戦略、STP、因果推論  
リコmend、BIや自動化ツールなど

## ✓ ゲスト講師の方の実務的な視点



ただ講義を聞いている  
だけでは身につけません

## 03 データ分析を学習する上で大事なこと

この講義は手を動かすことがとても大事です

扱う範囲がとても広いため、必ず**自学自習**をしてください

聞いたことは忘れる

見たことは思い出す

体験したことは理解する

発見したことは身につく





# 少しでも進めていくことが大事！

**ショートカットを  
使いこなす！**

素振り練習みたいなもの



30分調べてもわからない場合  
Slack等で質問してください。

**調べる力**  
ググる力も重要



**周りは気にしない**  
自分のペースで



自身の説明力も上がる、  
周りとのつながりが持てる

**教えあう**

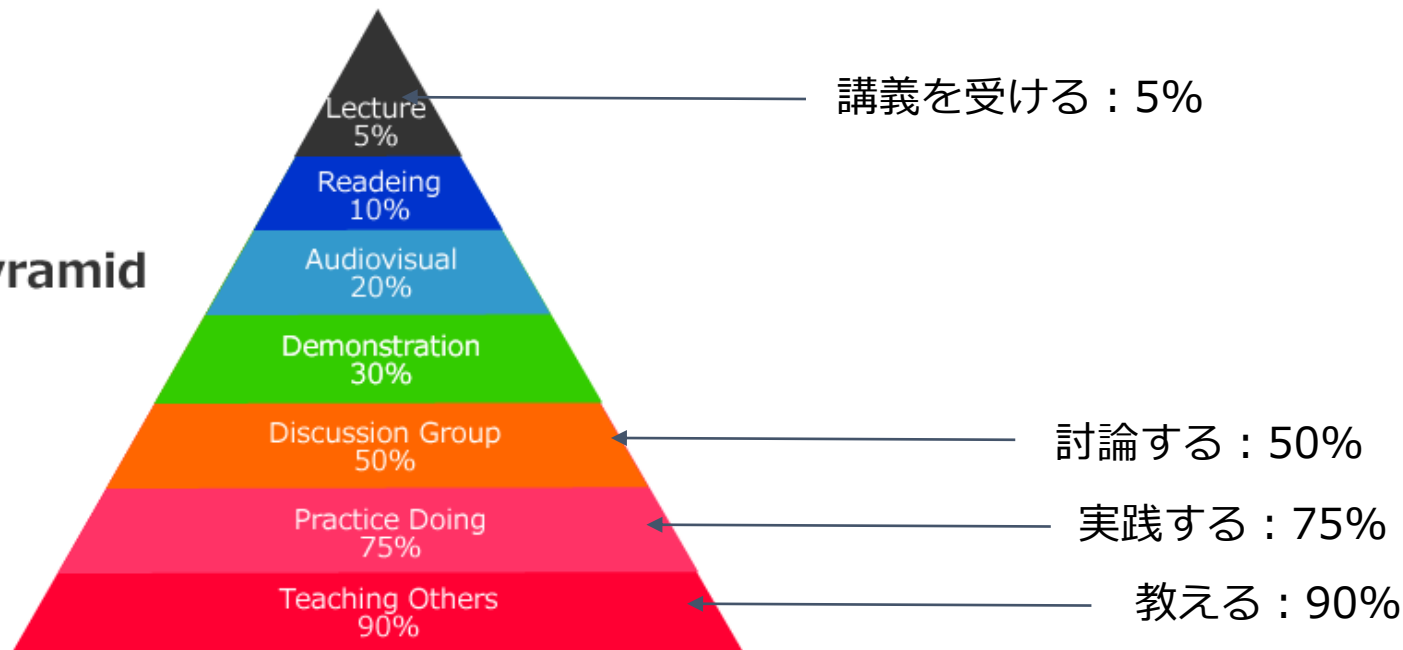
教えることでわかることもある  
足りない部分を補う



# アウトプットすることでより多くの学びを



## Learning Pyramid



Source: National Training Laboratories, Bethel, Maine



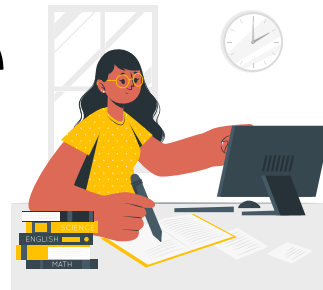
## ⚠ 注意事項



✎ **早いスピードで進めていくので予習をしてください**

✎ **コンテンツとして盛り沢山のため、ポイントだけ説明します**

✎ **ついていけない場合、復習でカバーしてください**



PythonとLinux、Jupyterは使えるようになってください

# 04

## 講義前の学習について

# 前提知識



✎ Python、Jupyter NoteBook、Linuxの基本的な使い方

✎ 微分積分学や線形代数の初歩、確率統計の基礎



## 他の環境準備方法：自分のPCで実行する方法



- ✎ 「Anaconda」をググる（Jupyterなどが使えるようになります）
- ✎ 自分の使うPCに合ったバージョンをインストールする
- ✎ 使い方もググる
- ✎ 他、GoogleColaboratoryも使用可能





次回はPythonの基礎

# 05

## 次回の予告と参考文献



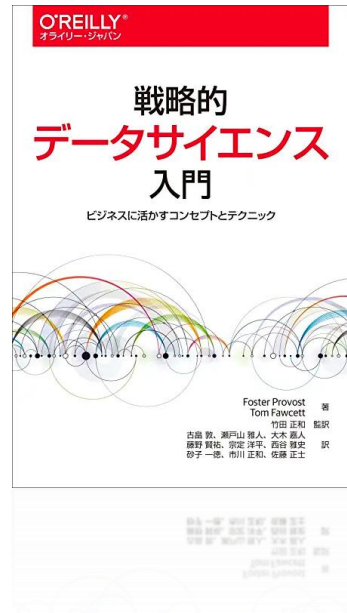
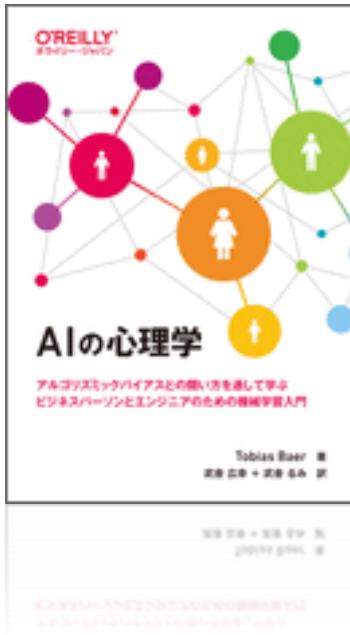


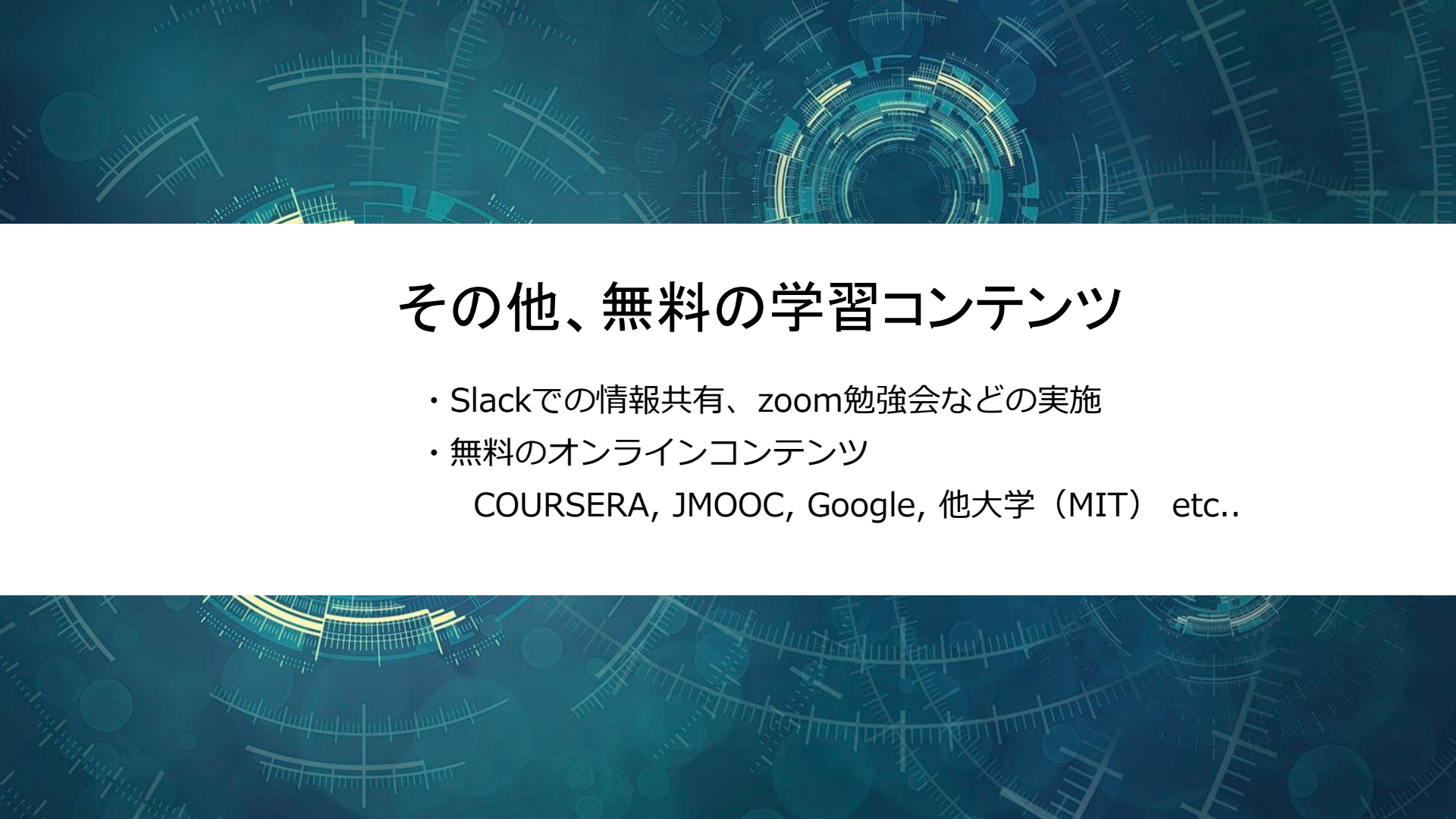
# 次回の予告

- 第2回 Numpyの基礎-
- Numpyの基本的な使い方



# 参考文献 データ分析に関連する読み物





## その他、無料の学習コンテンツ

- Slackでの情報共有、zoom勉強会などの実施
- 無料のオンラインコンテンツ  
COURSERA, JMOOC, Google, 他大学（MIT） etc..





## Q&Aセッション(2回目)

