データサイエンティストのスキルとその育成

～事例を交えて～

2013年11月28日

EMCジャパン株式会社
コンサルティング部
内田　信也

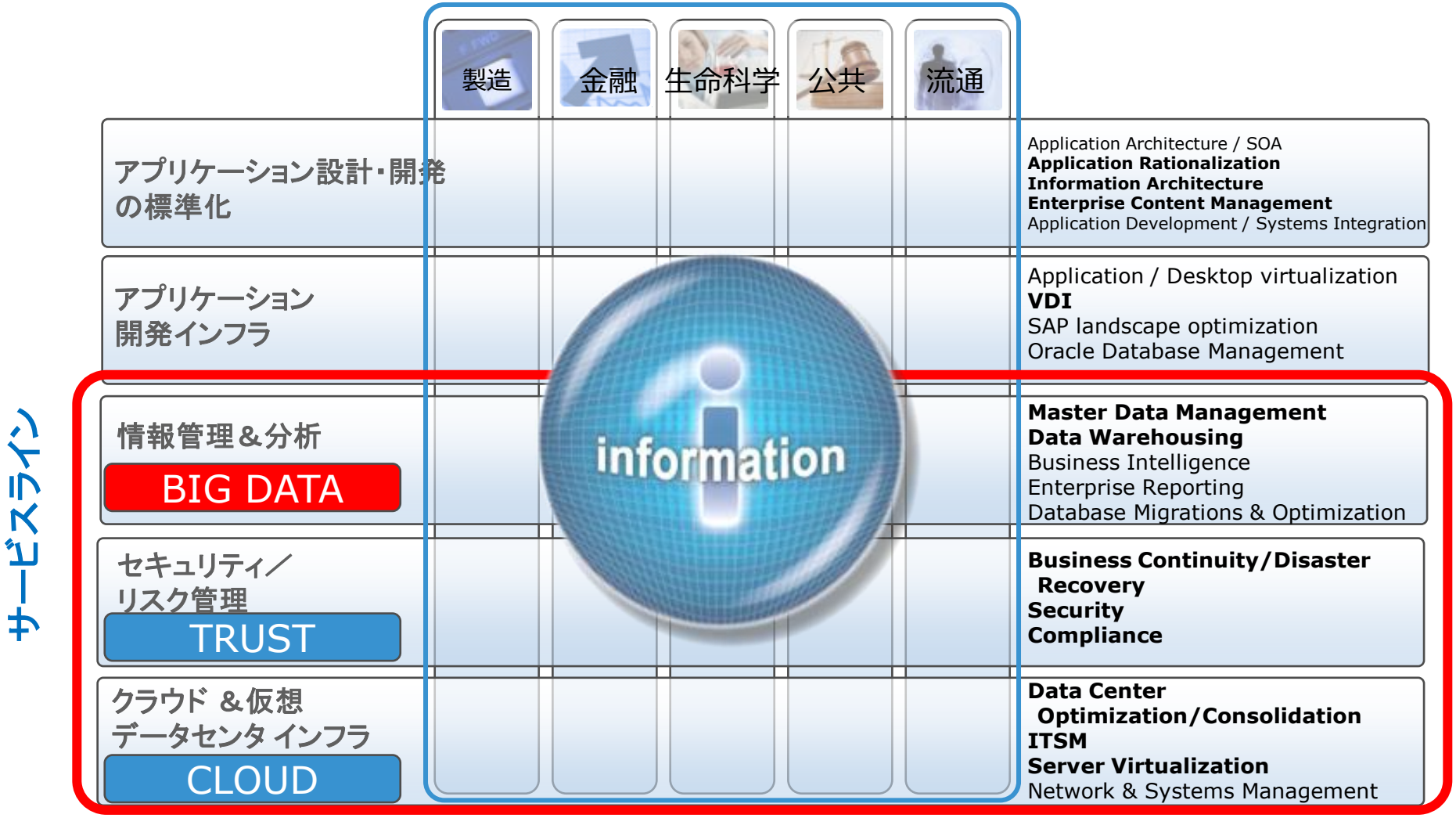
アジェンダ

- EMCコンサルティングのご紹介

- データサイエンティストとは？
- データサイエンティストをどう育成するか？
- データサイエンティストの仕事
  - ビジネスとITをどう橋渡しするか？

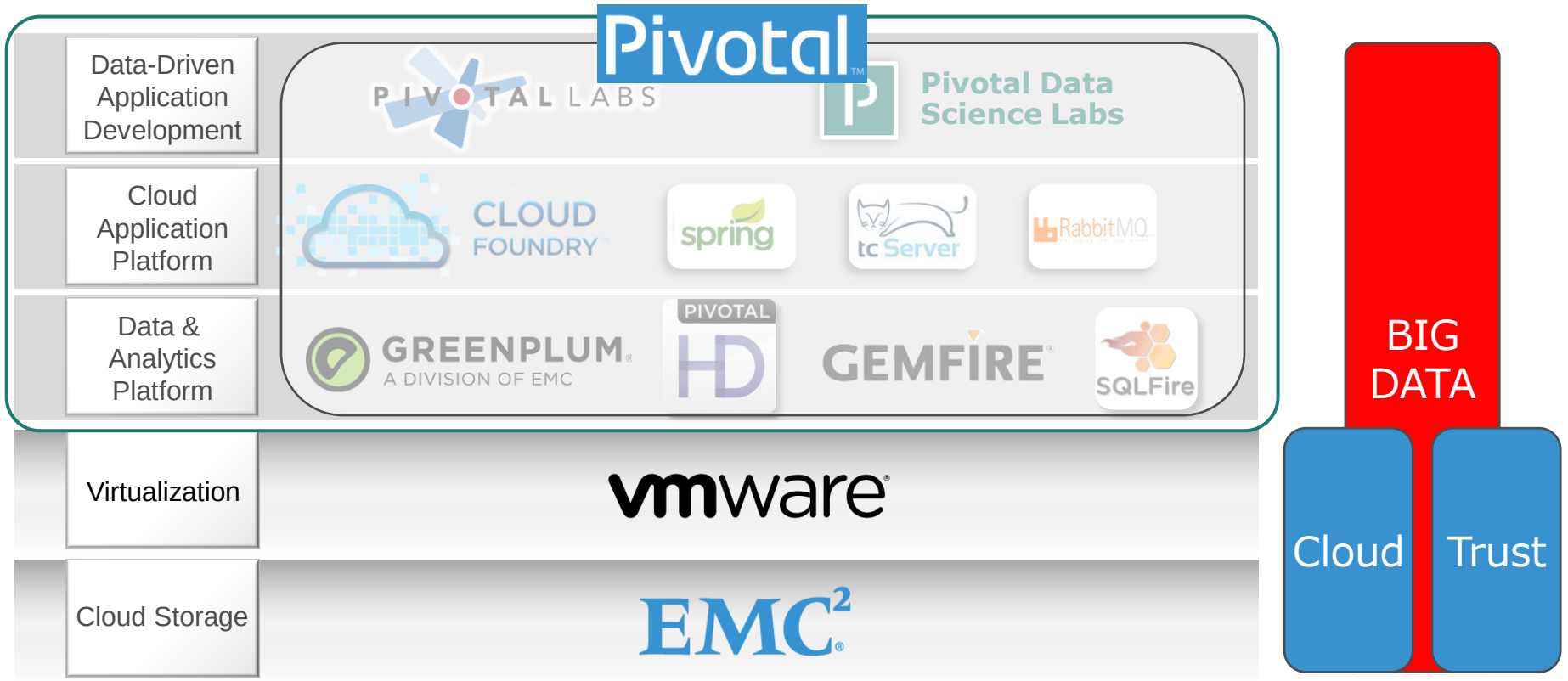
EMCコンサルティング・サービスラインのご紹介

EMCは「情報インフラの最適化サービス」を出発点として、ITコンサルティングのサービスラインを拡充して参りました。日本でご提供中の3つのサービスラインのうち、本日は「情報活用＆分析」を中心にご紹介します。



コンサルティング・サービスラインと製品ライン

EMCの製品戦略は、「お客様に選択肢をご提供する」ことで、相互にベンダーロックさせないところに特徴があります。
EMCコンサルティングは、製品ラインを跨って、お客様のあるべき姿を提案しています。



Pivotal Data Scientist Teamとのコラボレーション

EMCコンサルティングのBIG DATAチームは、Pivotal Data Science Teamと連携して
お客様の「分析支援」「データサイエンスチーム立ち上げ・サービス」を展開しています。

Pivotal Data Scientist Team (Senior Members)

Pivotal Data Science Labs

- **Annika Jimenez** – データサイエンスサービス・グローバル責任者 (Sr. Director, Audience and Advertising Analytics at Yahoo!, M.I.A. in International Management, UCSD)
- **Kaushik Das** – エネルギー、小売、通信業界における数学的モデリング (Director of Analytics at M-Factor, M.S. in Mineral Engineering, UC Berkeley)
- **Rahel Jhirad** –定量モデリングと貿易と金融におけるリスク管理 (Global Risk Management at Saloman Brothers, Morgan Stanley, Ph.D. in Economics, Princeton, M.S. in Mathematics, Courant Institute)
- **Michael Brand** –小売、金融、ゲーム業界におけるテキスト、音声、ビデオ分析 (Chief Scientist at Verint Systems, M.S. in Applied Mathematics, Weizmann Institute)
- **Woo Jung** – ベイズ推定と需要予測(Sr. Statistician at M-Factor, M.S. in Statistics, Stanford)
- **Noelle Sio** – デジタルメディア分析及び数学的モデリング(Sr. Analyst at eHarmony, Fox Interactive Media (Myspace), M.S. in Applied Mathematics, Cal Poly Pomona)
- **Rashmi Raghu** – 数値シミュレーションと分析 (Ph.D. in Mechanical Engineering, Stanford)
- **Jarrod Vawdrey** – マーケティング分析及びSAS (Analytics Consultant at Aspen Marketing, B.S. in Mathematics, Kennesaw State University)
- **Sarah Aerni** – 遺伝子工学 及び機械学習 (Ph.D. in Biomedical Informatics, Stanford)
- **Srivatsan Ramanujam** – 非線形計画法及びテキストマイニング(Natural Language Scientist at Sony, Salesforce.com, M.S. in Computer Sciences, UT Austin)
- **Emily Kawaler** –臨床情報学及び機械学習 (M.S. in Computer Sciences, University of Wisconsin-MClinical Informatics adison)
- **Niels Kasch** – テキスト分析及び非線形計画法 (Ph.D. in Computer Science, UMBC)
- **Regunathan Radhakrishnan** – 機械学習、信号処理、マルチメディアコンテンツ分析、指紋認証及び電子透かし (Research Staff at Dolby Laboratories, MERL, Ph.D. in Electrical Engineering, NYU-Poly, Brooklyn)
- **Michael Natusch** – データサイエンスサービス・欧州責任者 (Chief Analyst at Cumulus Analytics, Ph.D. in Theoretical Condensed Matter Physics, University of Cambridge)
- **Hulya Farinas** – ヘルスケア業界における最適化及びリソース配置(Modeler at M-Factor, IBM, Ph.D. in Operations Research, University of Florida)
- **Derek Lin** – ネットワークセキュリティ、不正検知、音声及び言語処理(Principal Scientist at RSA, M.S. in Signal Processing, USC)
- **Kee Siong Ng** – ヘルスケア業界におけるデータマイニング (Sr. Data Miner at Medicare Australia, Ph.D. in Computer Science, and Postdoctoral Fellow, Australian National University)
- **Jin Yu** – 確率的最適化、機械学習におけるロバスト統計及びコンピュータビジョン (Research Associate at U of Adelaide, Ph.D. in Machine Learning, Australian National University)
- Hong Ooi – 保険、金融業界のリスクモデリング(Statistician at ANZ, Ph.D. in Statistics, Australian National University)
- **Vivek Ramamurthy** – 適応学習、確率論的モデリング及び凸最適化(Ph.D. in Operations Research, UC Berkeley)
- **Noah Zimmerman** – 統計学及び免疫学 (Ph.D. in Biomedical Informatics, Stanford)
- **Mariann Micsinai** – 次世代塩基配列決定 (Market Risk Management Associate at Lehman Brothers, Ph.D. in Computational Biology, NYU / Yale)
- **Victor Fang** – 画像分析、グラフ分析及び機械学習(Sr. Scientist at Riverain Medical, Ph.D. in Computer Sciences, University of Cincinnati)
- **Anirudh Kondaveeti** – 軌跡データマイニング及び機械学習 (Ph.D. in Computing & Dec. Systems Eng, Arizona State University)
- **Joseph Zadeh** – IT/ネットワークトラフィック分析及び金融モデリング(Ph.D. in Mathematics, Purdue)
- **Alexander Kagoshima** – 時系列分析、統計学及び機械学習 (M.S. in Economics/Computer Science, TU Berlin)

データサイエンティストとは？

BIG DATAに対するお客様の動向

日本でも、BIG DATAへの意識変化

2011年末

・BIG DATAって何？
・BIG DATAはバズワード？

2013年初以降～

・BIG DATAの取り組みを始めねば・・
・BIG DATAへのスタンスは決めたい

・「IT部門がビジネス貢献」するには
BIG DATAが鍵になる

アクションに移る際の問題意識

1.総論賛成・各論反対
- 自社のどんな業務にどんな形で役に立つのか？

2.組織・体制の整備
- データサイエンティストとは？
- 人材を自社で育成すべきか？

3.情報インフラの整備
- 非構造化データを含めた情報系の姿は？
- SNS分析の基盤は自社で準備すべきか？

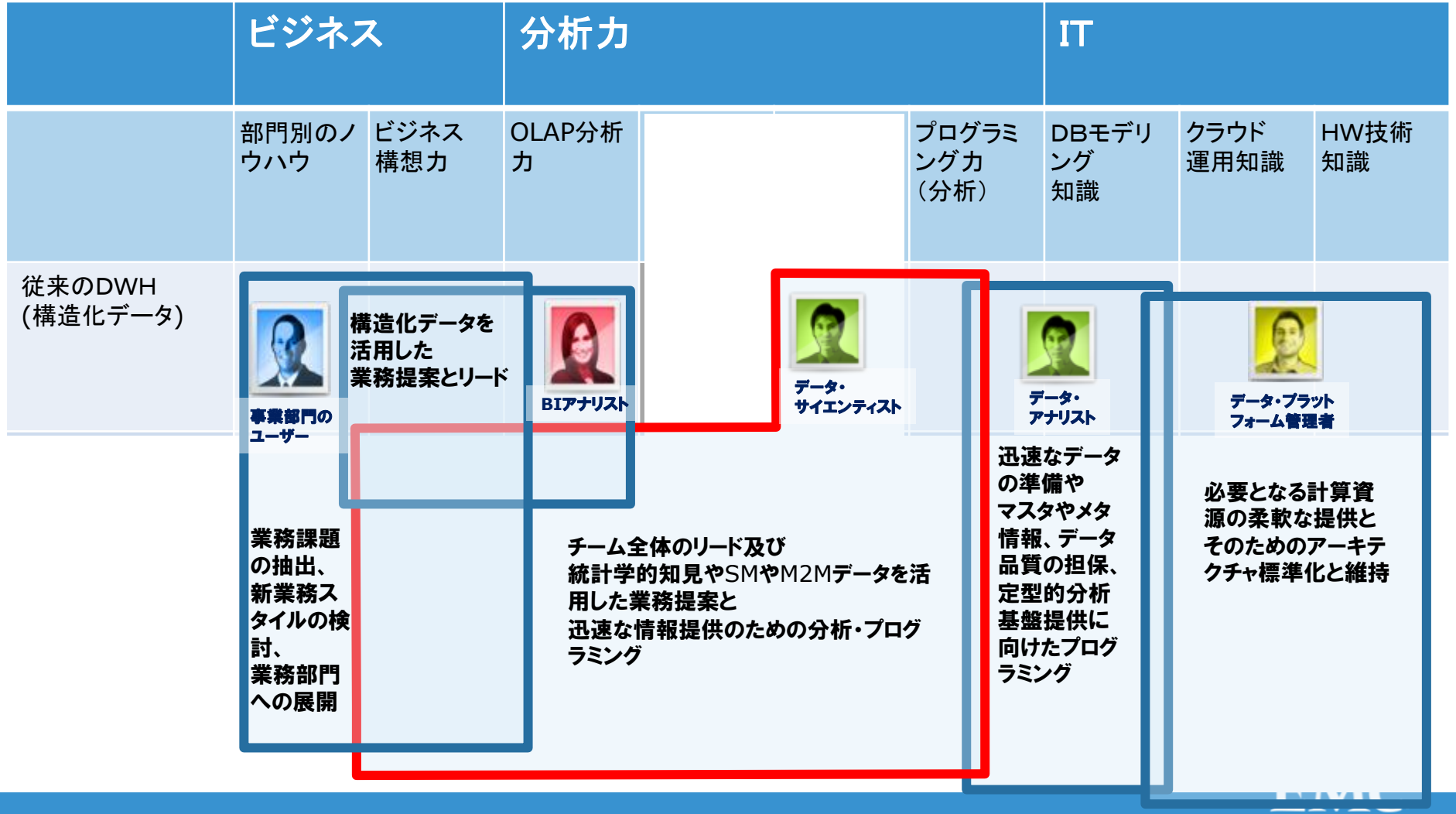
•活用可能性の検証

•人材像の見極め

•実際の動き方

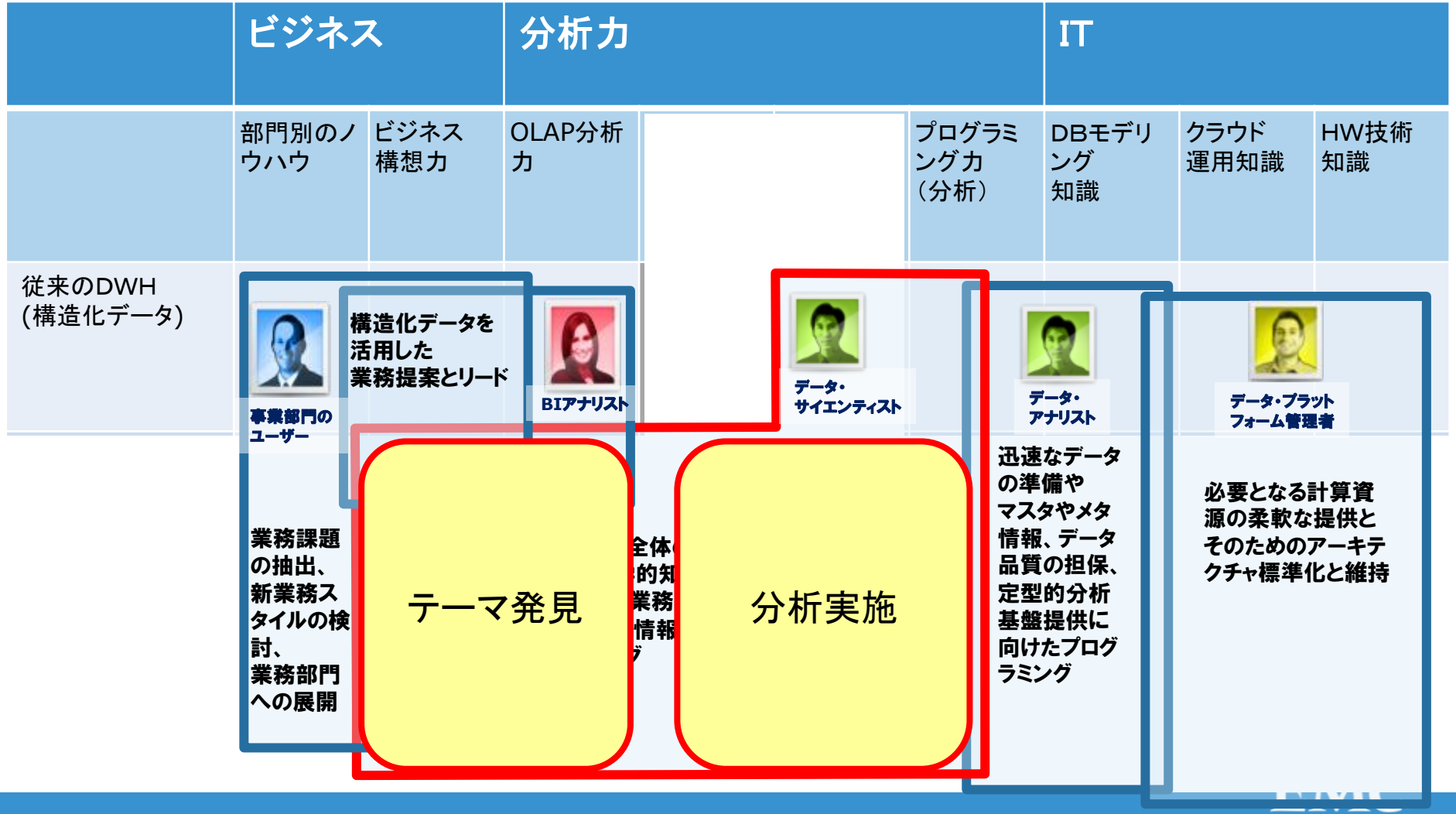
データサイエンティストの役割

ビジネス部門から施策テーマや分析テーマを引き出してくるのが、データサイエンティストです。データサイエンティストが社内にいることが重要と考えています。



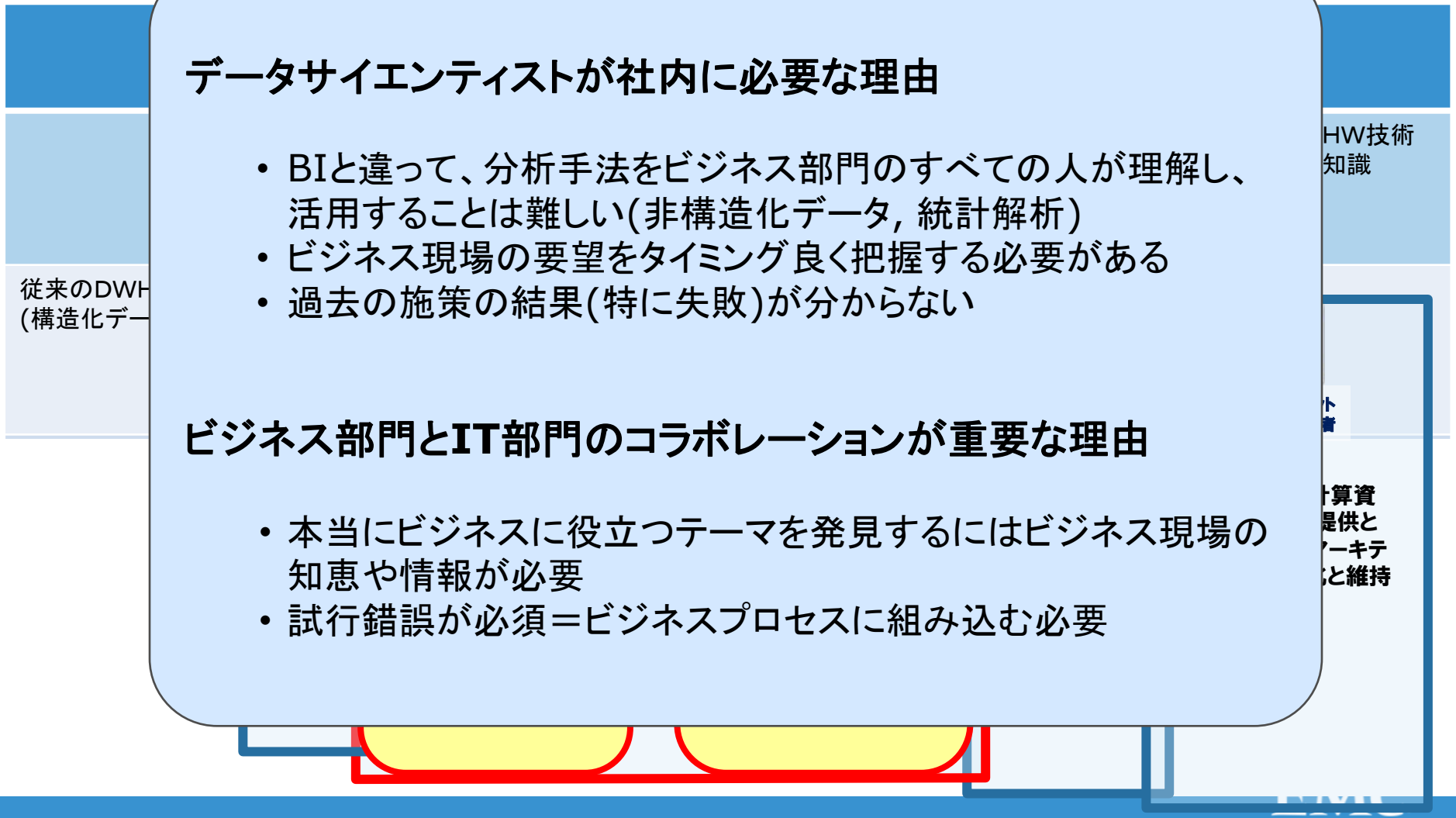
データサイエンティストの役割

ビジネス部門から施策テーマや分析テーマを引き出してくるのが、データサイエンティストです。データサイエンティストが社内にいることが重要と考えています。



データサイエンティストの役割

ビジネス部門から施策テーマや分析テーマを引き出してくるのが、データサイエンティストです。データサイエンティストが社内にいることが重要と考えています。



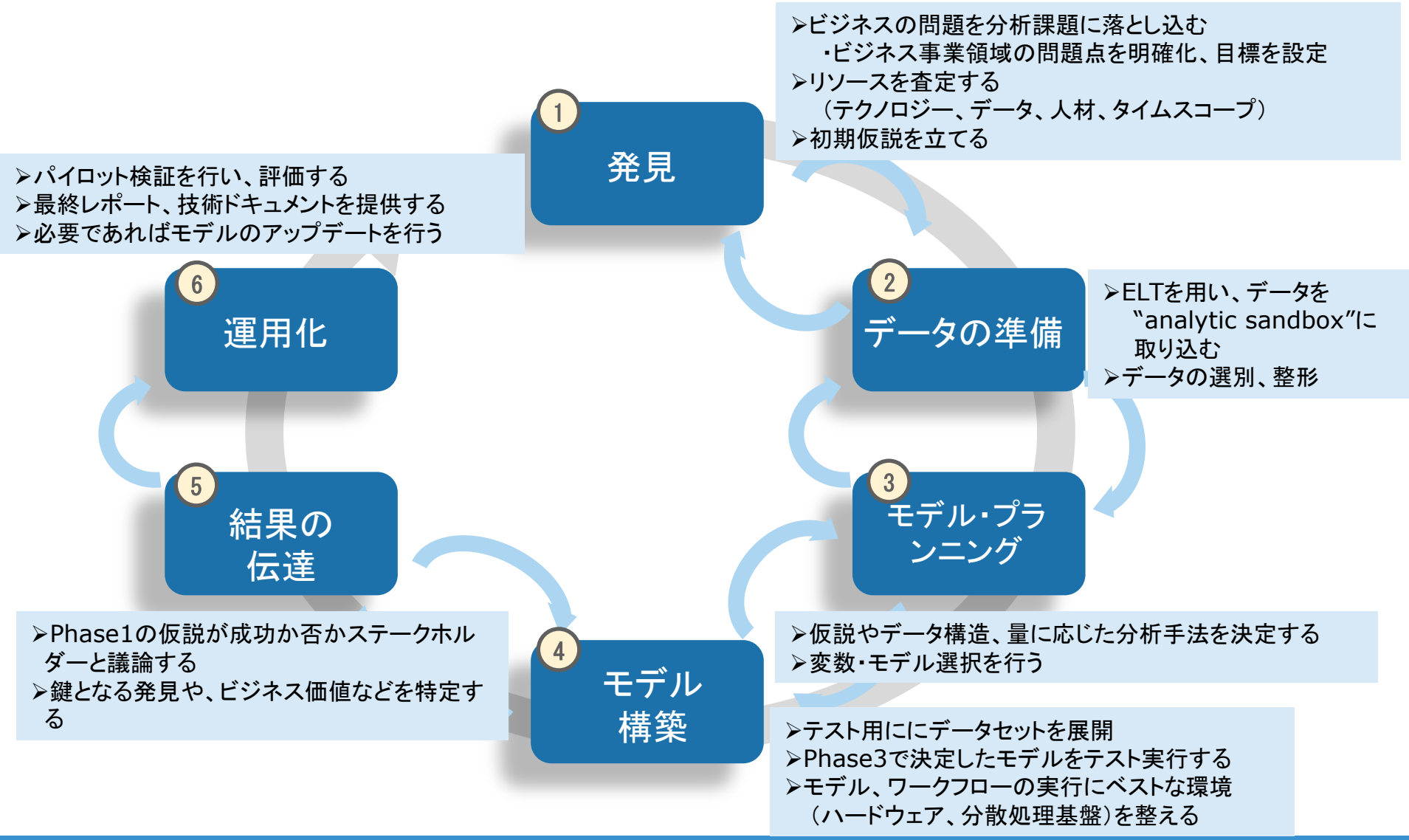
データサイエンティストが社内に必要な理由

- BIと違って、分析手法をビジネス部門のすべての人が理解し、活用することは難しい(非構造化データ, 統計解析)
- ビジネス現場の要望をタイミング良く把握する必要がある
- 過去の施策の結果(特に失敗)が分からない

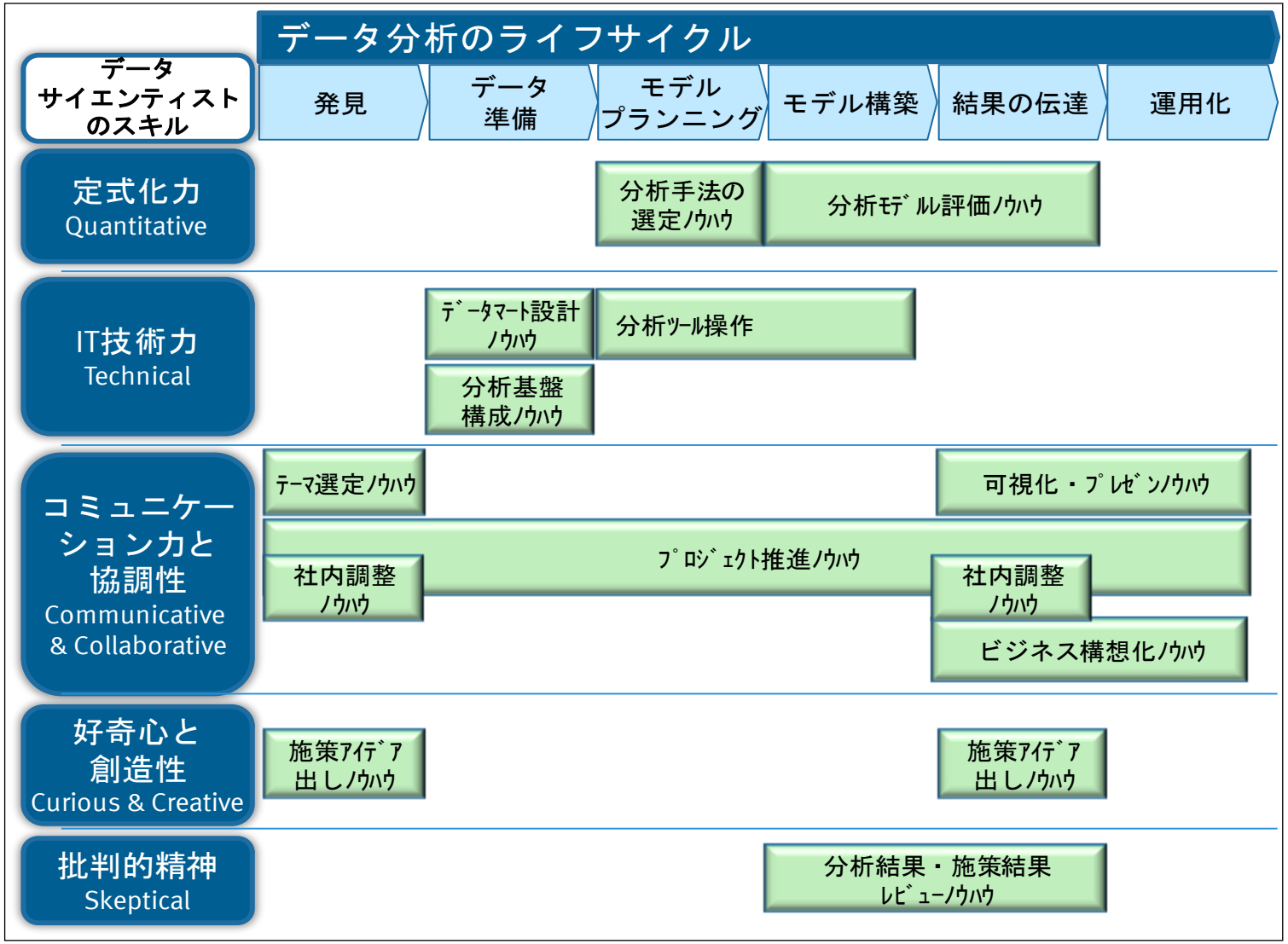
ビジネス部門とIT部門のコラボレーションが重要な理由

- 本当にビジネスに役立つテーマを発見するにはビジネス現場の知恵や情報が必要
- 試行錯誤が必須＝ビジネスプロセスに組み込む必要

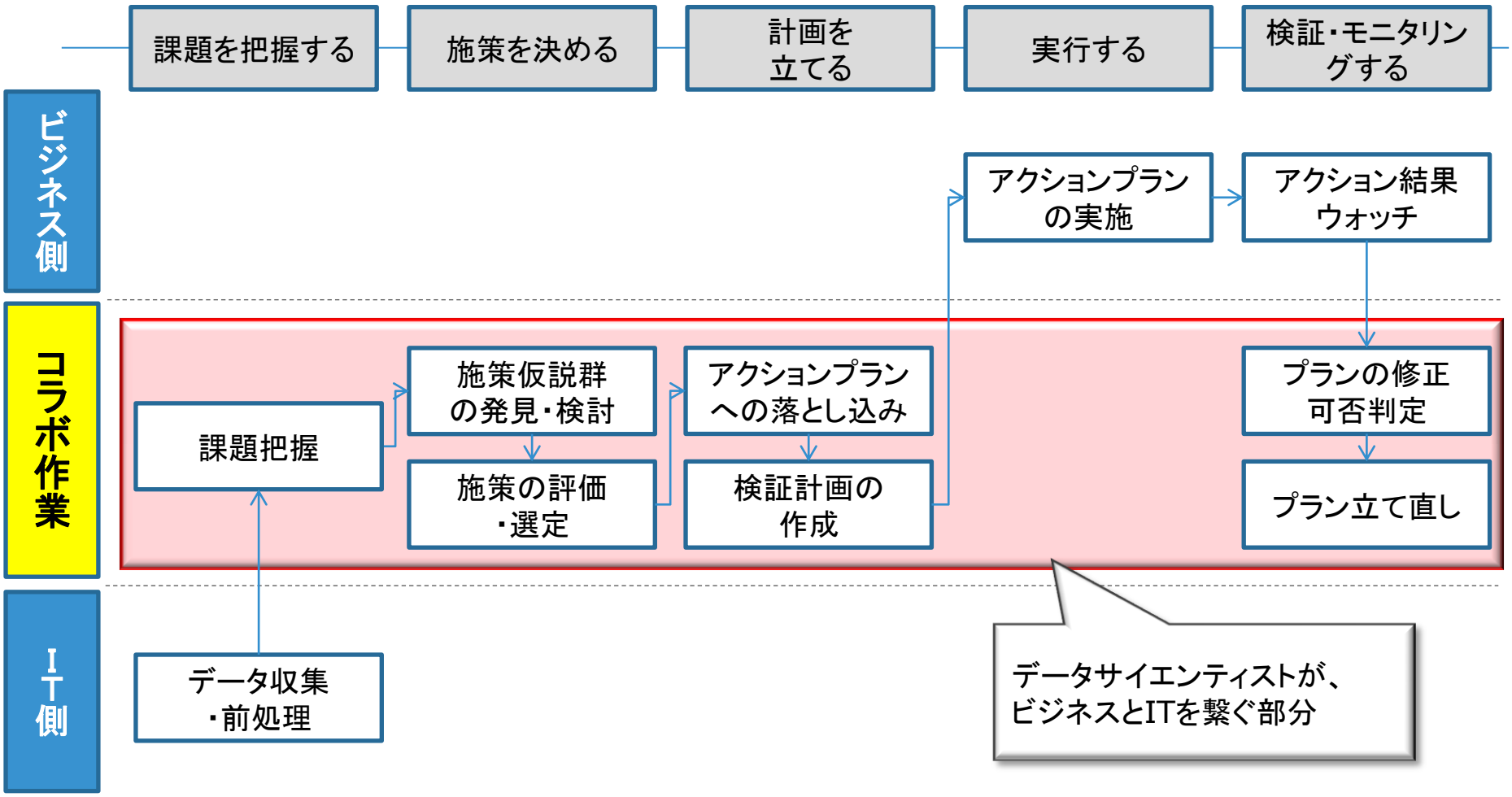
データ分析のライフサイクル



データサイエンティストのスキル



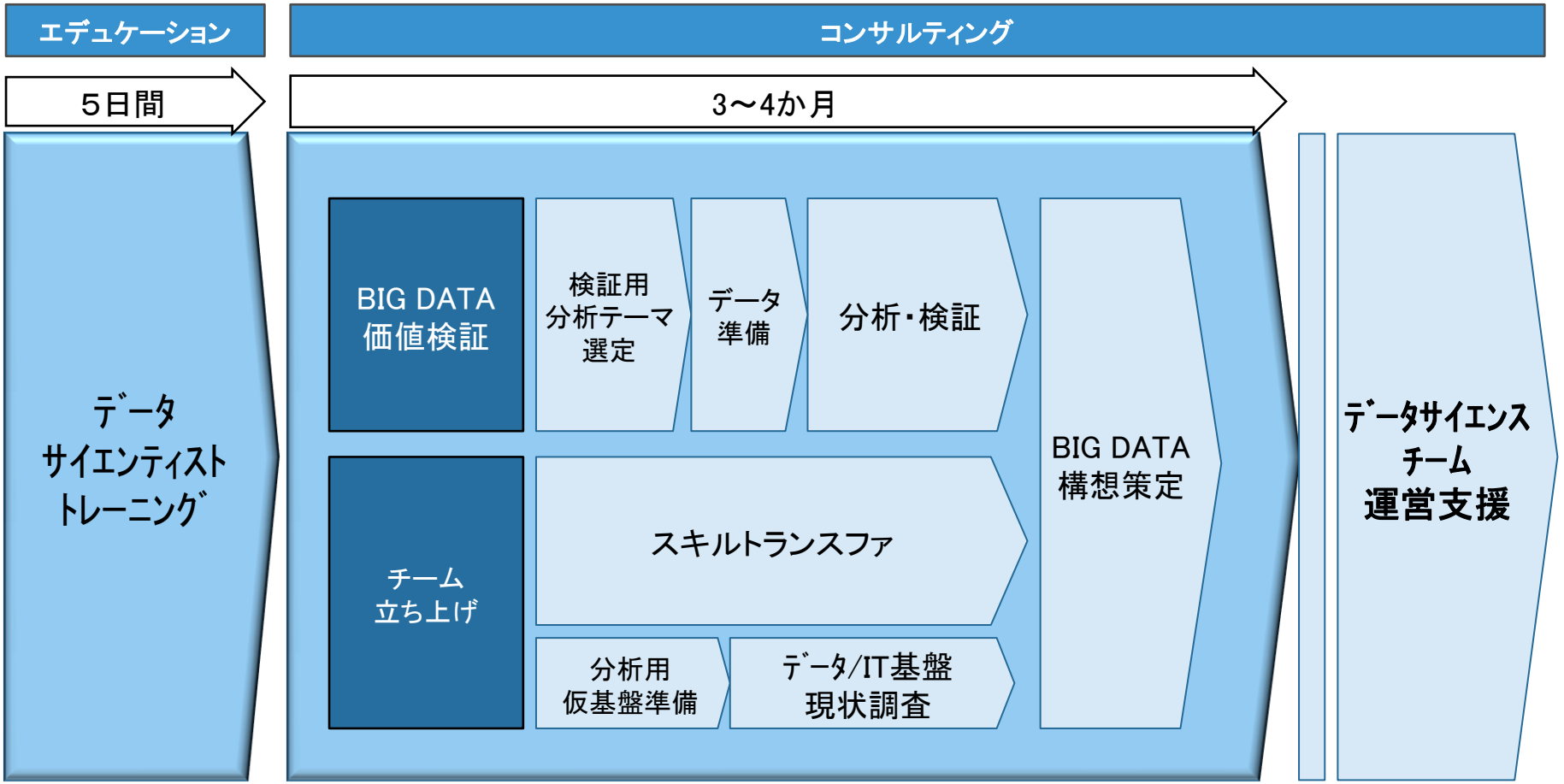
BIG DATA活用するにはビジネスとITが密に協働する必要がある



データサイエンティストをどう育成するか

データサイエンティスト育成支援の構成

- Pivotal Data Scientist Teamの方法論を日本向けにアレンジして、データサイエンティストの育成を支援しています。



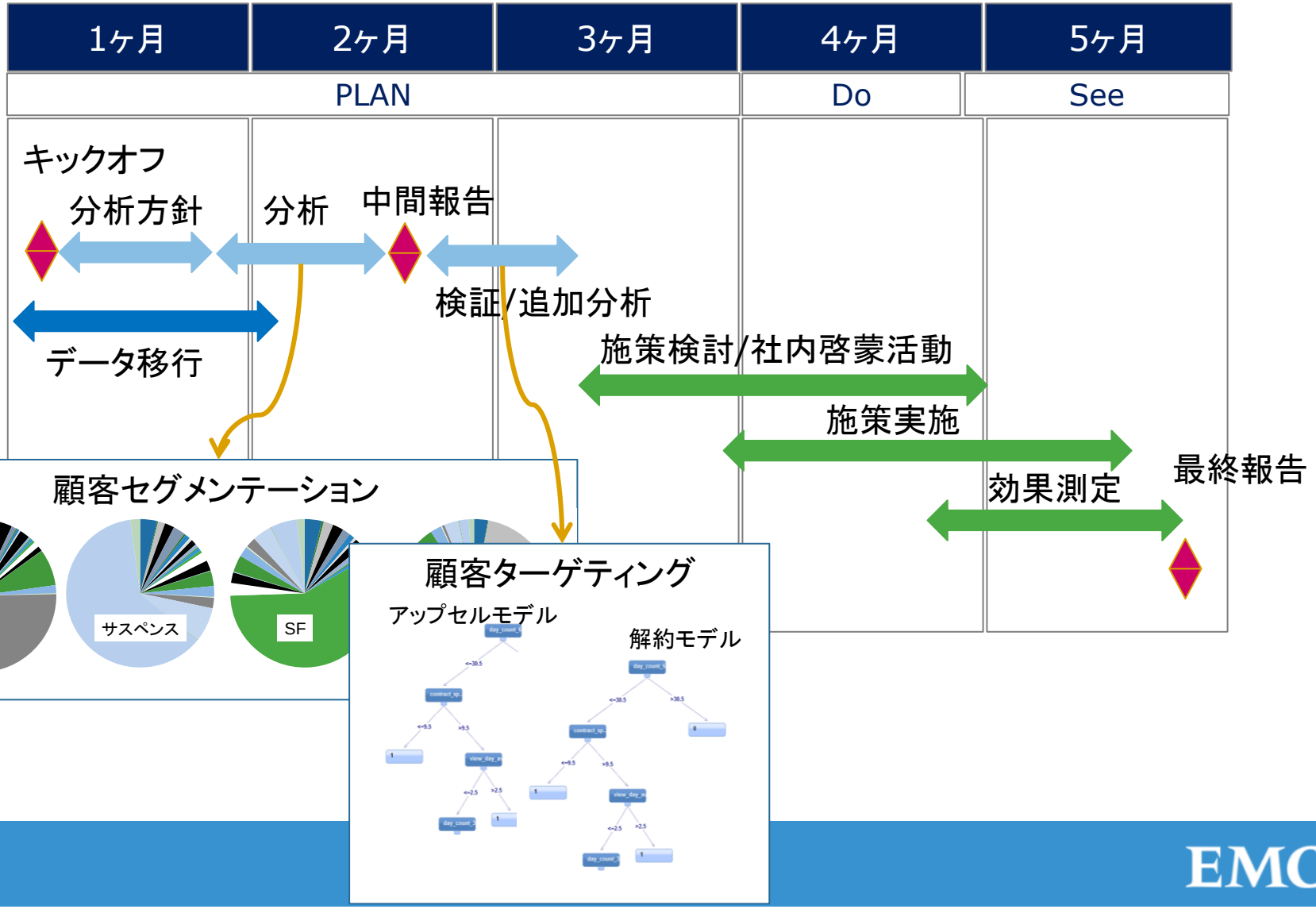
データ・サイエンティスト・トレーニング概要

「Data Science and Big Data Analytics」

1日目	2日目	3日目	4日目	5日目
ビッグデータ分析入門 ＋ データ分析のライフサイクル	**データ分析の基本～「R」を使って**	**先進的分析の理論と手法**	**先進的分析の技術とツール**	**分析結果の可視化とプレゼンテーション ＋ 課題チャレンジ**
・Big Dataの概要 ・分析実務の現状 ・データサイエンティストとは ・業界別のBig Data 分析 ・データ分析のライフサイクル	・R言語の基礎知識 ・データの調査と分析 ・モデル構築と評価	・K平均法クラスタリング ・アソシエーション・ルール ・線形回帰 ・ロジスティック回帰 ・単純ベイズ分類（Naïve Bayesian Classifier） ・決定木 ・時系列分析 ・テキスト分析	・非構造化データの分析（MapReduceとHadoop） ・Hadoopエコシステム ・In-database 分析 – SQLの要点 ・In-database分析で活用する先進SQLとMADlib	・分析プロジェクトの実施と運用 ・最終成果物の作り方 ・ビジュアル化のテクニック + 課題チャレンジ データ分析ライフサイクルの適用業務（ケーススタディ）

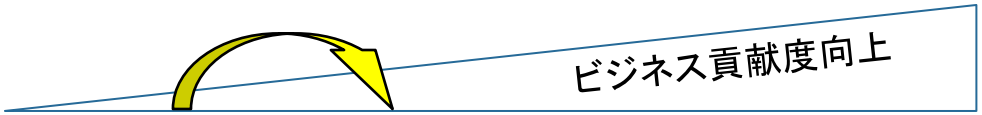
データサイエンスチーム育成プロジェクト事例

社内のデータサイエンスチームへのスキルトランスファーを実施しながら、
分析プロジェクトを実施。後半はお客様メンバーが分析を実際に担当。



データサイエンスチーム立ち上げサービスで目指すゴール

本サービスで分析への信頼を獲得（第１段階から第２段階への移行）することを目指します。



		第１段階	第２段階	第３段階
意思決定への定着	分析への信頼	半信半疑	腹落ち	ビジネスモデル転換
	分析活用組織	個人スキル依存	全社集中組織	ビジネスラインに分散
分析高度化	分析手法	集計 OLAP 多変量解析	データマイニング 機械学習 自然言語解析	複合分析 独自ロジック分析 自動化
	データ整備	既存データ	データ整備・蓄積（マスタ・実績） 施策結果蓄積	ビジネスラインによる自主的収集

BIG DATA活用成熟度

分析事例

主要分析対象	統計・機械学習系			テキスト系		数理計画系
	相関・分類	行動予測・事象予測	定量予測	トピック分析	ネットワーク分析	最適解探索
顧客・消費者	顧客分類（CATV） 顧客分類（金融ｸﾞﾙｰﾌﾟ）	ｱｯﾌﾟｾﾙ候補者予測（CATV） ﾀﾞｳﾝｾﾙ候補者予測（CATV） 解約候補者予測（CATV） ﾀｰｹﾞｯﾄ広告（金融ｸﾞﾙｰﾌﾟ） 犯罪・違反発見（金融） 営業訪問先ﾘｽﾄ（EMC社内） TCE（EMC社内）		話題分析（通信） 話題分析（小売） 話題検知（公共） リスク検知（公共） 解約ｲﾝﾌﾙｴﾝｻ抽出（通信）	ｲﾝﾌﾙｴﾝｻ抽出（公共）	ｱﾄﾘﾋﾞｭｰｼｮﾝ分析（Webサービス）
商品・サービス	商品ﾚｺﾒﾝﾃﾞｰｼｮﾝ（CATV） 商品ﾚｺﾒﾝﾃﾞｰｼｮﾝ（銀行）		需要予測（中食小売） 広告・販促効果分析（加工食品） 広告・販促効果分析（雑貨）	機器障害問合せ傾向分析（EMC社内）		最適価格設定（航空）
事業活動・その他		設備故障予測（石油精製） 停電予測（電力）	交通量予測（公共）	競合ﾎﾟｼﾞﾈｶﾞ分析（小売） リスク管理（公共）	研究開発ｲﾉﾍﾞｰﾀ検知（EMC社内）	最適航路設定（航空）

データサイエンティストの仕事

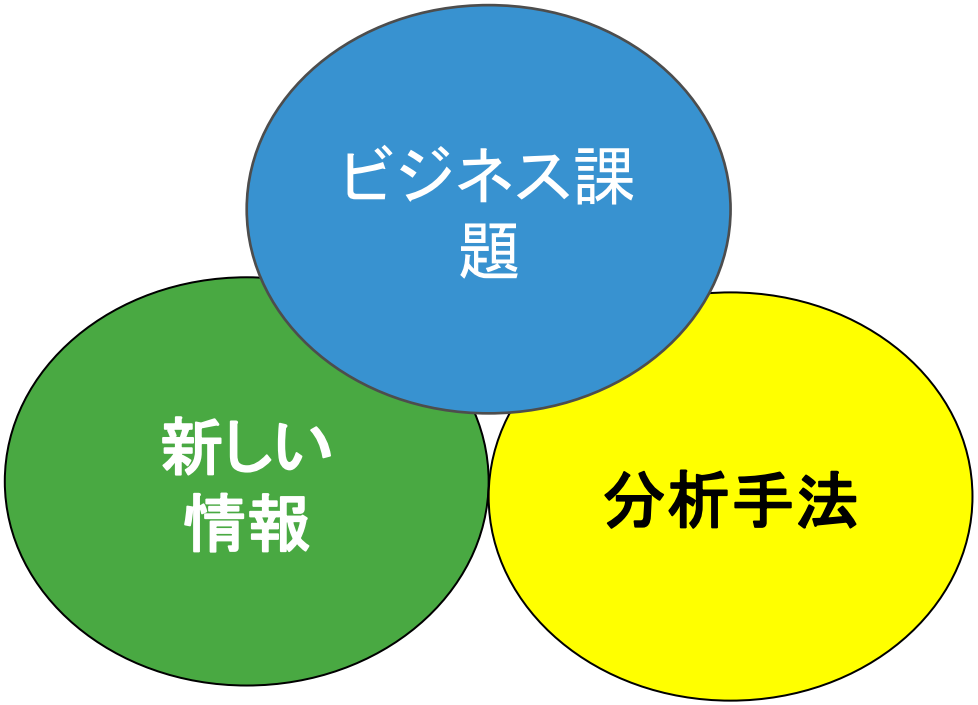
ビジネスとITをどう橋渡しするか？

ビジネスとITを橋渡しするポイント

	ビジネス部門と検討の合意	分析テーマ（施策）の発見	実行
トップダウン	•方針出し •体制（人）		•支援策（お金） •トライアンドエラー方針（時間）
ボトムアップ	•事例提供 •既存データの整理 •BIとの違い整理	•分析テーマ発見の3要素 •ビジネスでの活用可能性の検証	•スモールスタート •インフォーマルコミュニケーション

分析テーマの発見の3要素

3要素に対する理解度が高まるとともに、分析テーマ発見の効率は良くなる。



分析手法

統計・機械学習系			テキスト系		数理計画系
相関・分類	行動・事象予測	定量予測	トピック分析	ネットワーク分析	最適解探索
類似点の発見や似たもの同士をグループ化する手法	2つのグループを分けている要因を発見する手法	ある連続的な変数の値を決定する要因を発見する手法	発言の特徴を要約する手法	複数要素の間のつながり関係を分析する手法	ある命題の最適解(値)を探す手法

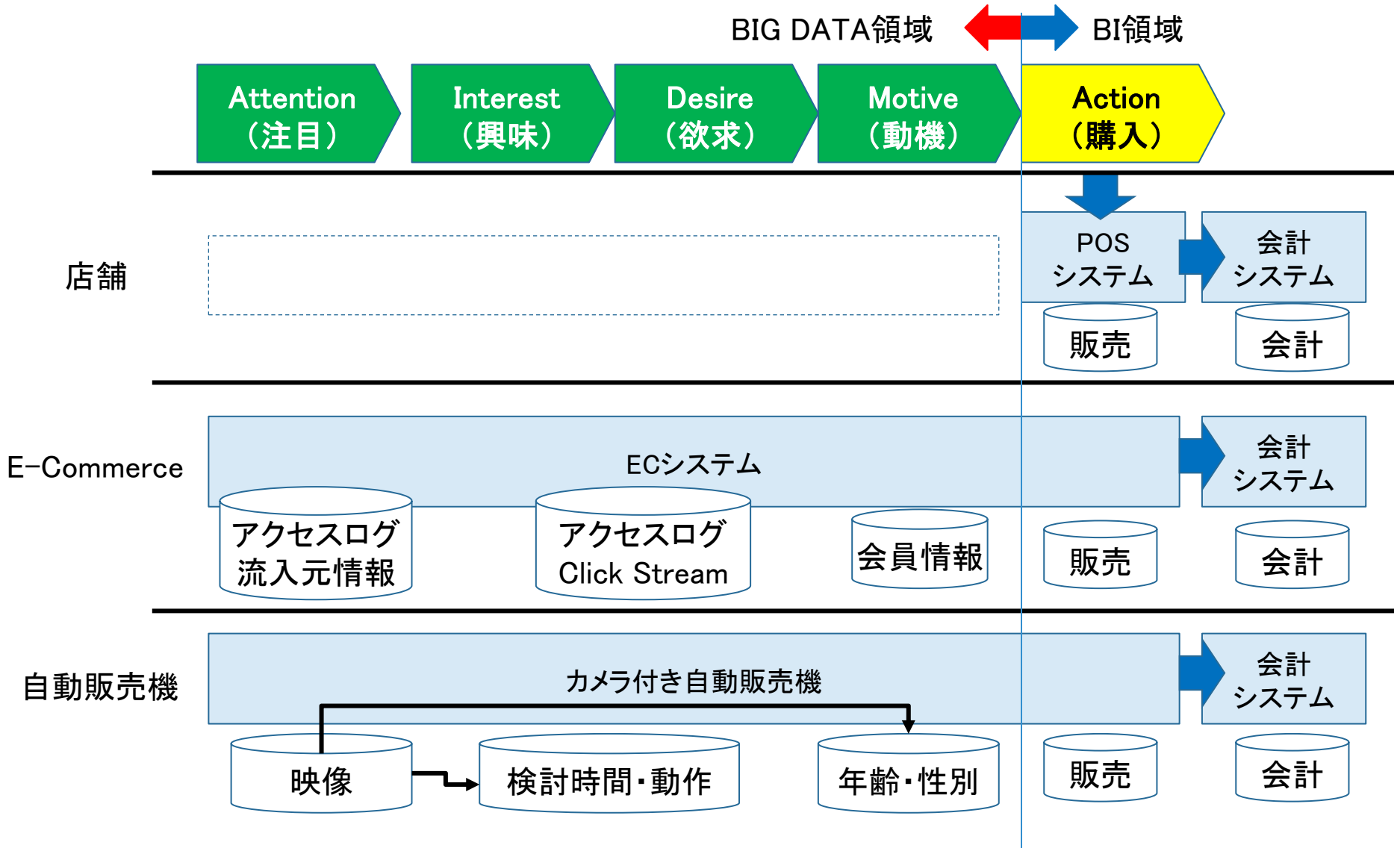
主要分析目的

顧客分類	顧客特性把握	主成分分析 因子分析
		K平均法
商品レコメンデーション		アソシエーション分析
アップセル ターゲティング広告 解約防止 故障予測 異常値検知	原因分析 要因分析 影響度分析	ロジスティクス回帰 決定木 数量化Ⅱ類
売上/来客予測		回帰/重回帰分析
話題把握		ポジネガ分析
インフルエンサー特定		インフルエンサ分析
最適価格設定	最適ルート設定	回帰/重回帰分析 線形計画法 動的計画法

時間がかかるのは
・応用力
・リカバリー力

新しい情報の例

顧客が購買に至る心理的変化や嗜好を理解・推定できる



ビジネス課題（分析テーマ）の検討

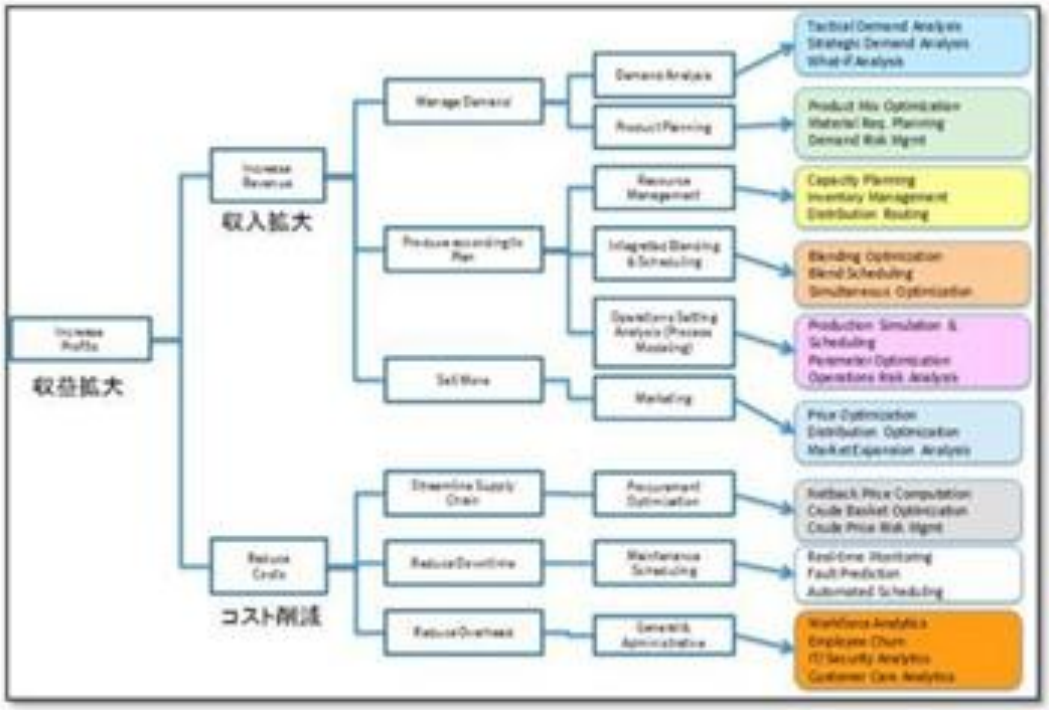
ビジネス課題の整理と新しい情報（今まで見えなかった情報）を組み合わせて
分析テーマを抽出する。

ビジネス課題の整理			課題	今まで見えなかった情報				
				ユーザーの生活・消費行動	ユーザーの購入決定プロセス	ユーザーの実際の利用状況	製造工程の詳細状況	・・
	成功率の悪い業務	売上への貢献が大きな業務	成約率向上 製品競争力向上	•販売ﾁｬﾈﾙ選定 •ﾚｺﾒﾝﾄﾞ商品選定	•販促施策検討	•広告ｺﾋﾟｰ検討	•納期保証	
		コストが大きい業務	リコール対応			•故障要因把握 •離反防止	•不良原因把握	
	リスクの高い業務	資産の大きな業務	在庫削減 設備稼働率向上	•製品需要予測	•競合製品把握	•部品需要予測	•生産設備故障予知	
		結果がすぐ出ない業務	ブランド力向上	•製品評判分析			•鮮度保証	

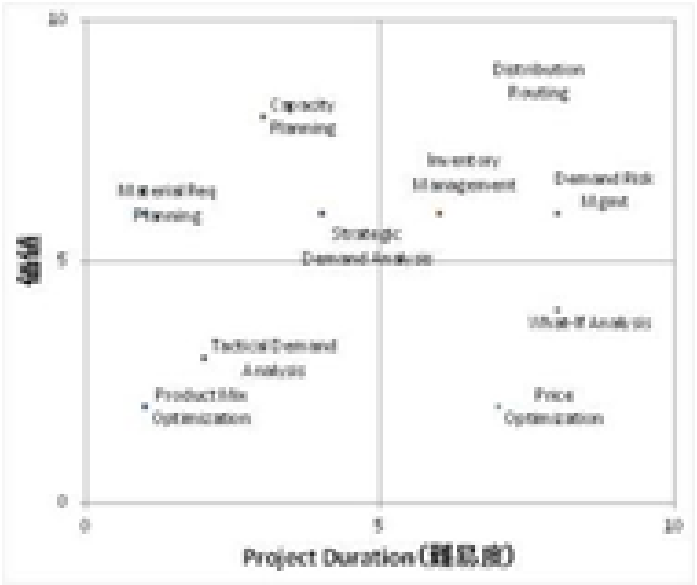
分析テーマ選定（例）

分析テーマをリストアップしたのち、価値や実施難易度で評価して選定します。

分析テーマのリストアップ



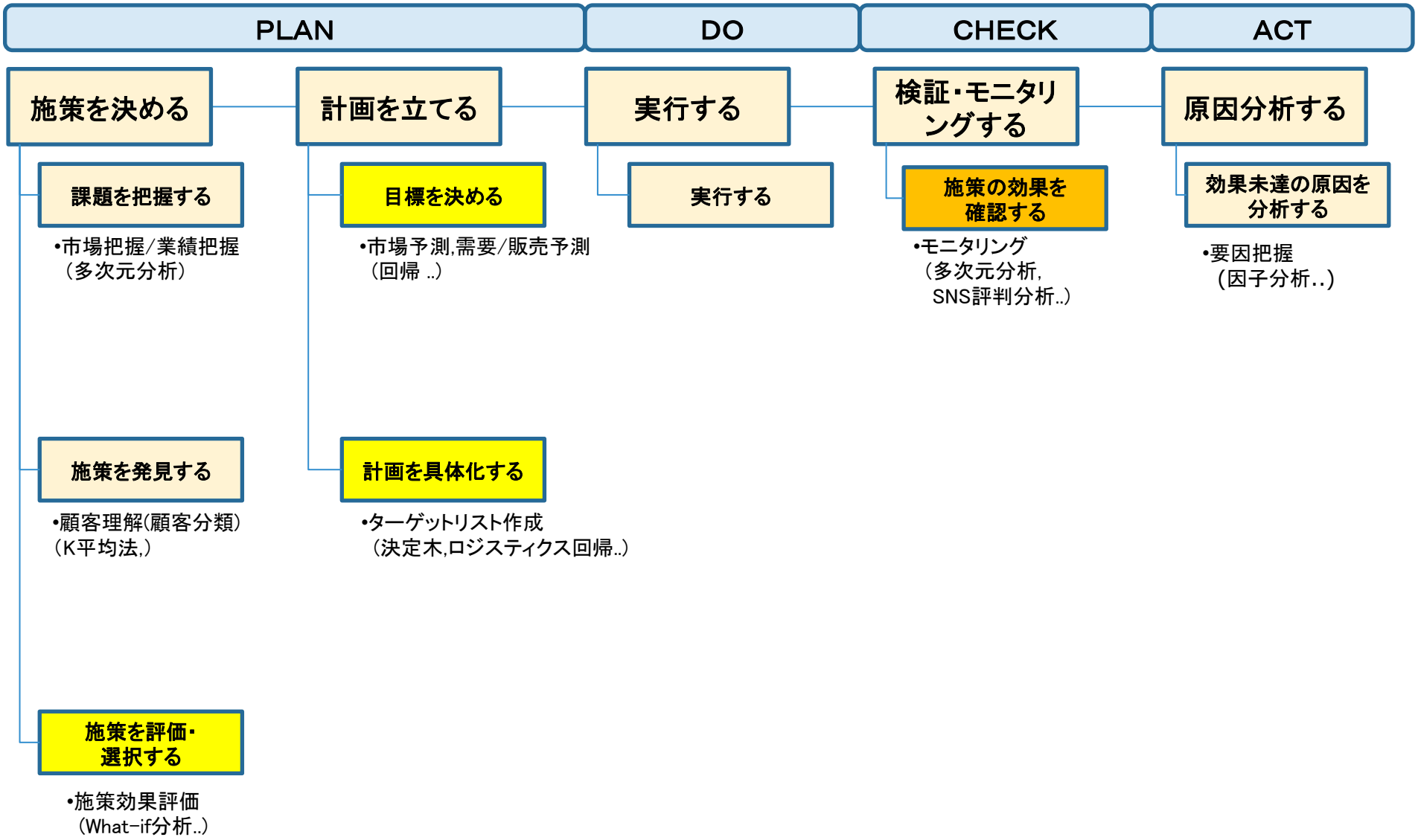
分析テーマの選定



ビジネスでの活用可能性のチェック

ビジネスにおける活用可能性の検証

効果の出しやすいテーマか？



分析事例抜粋

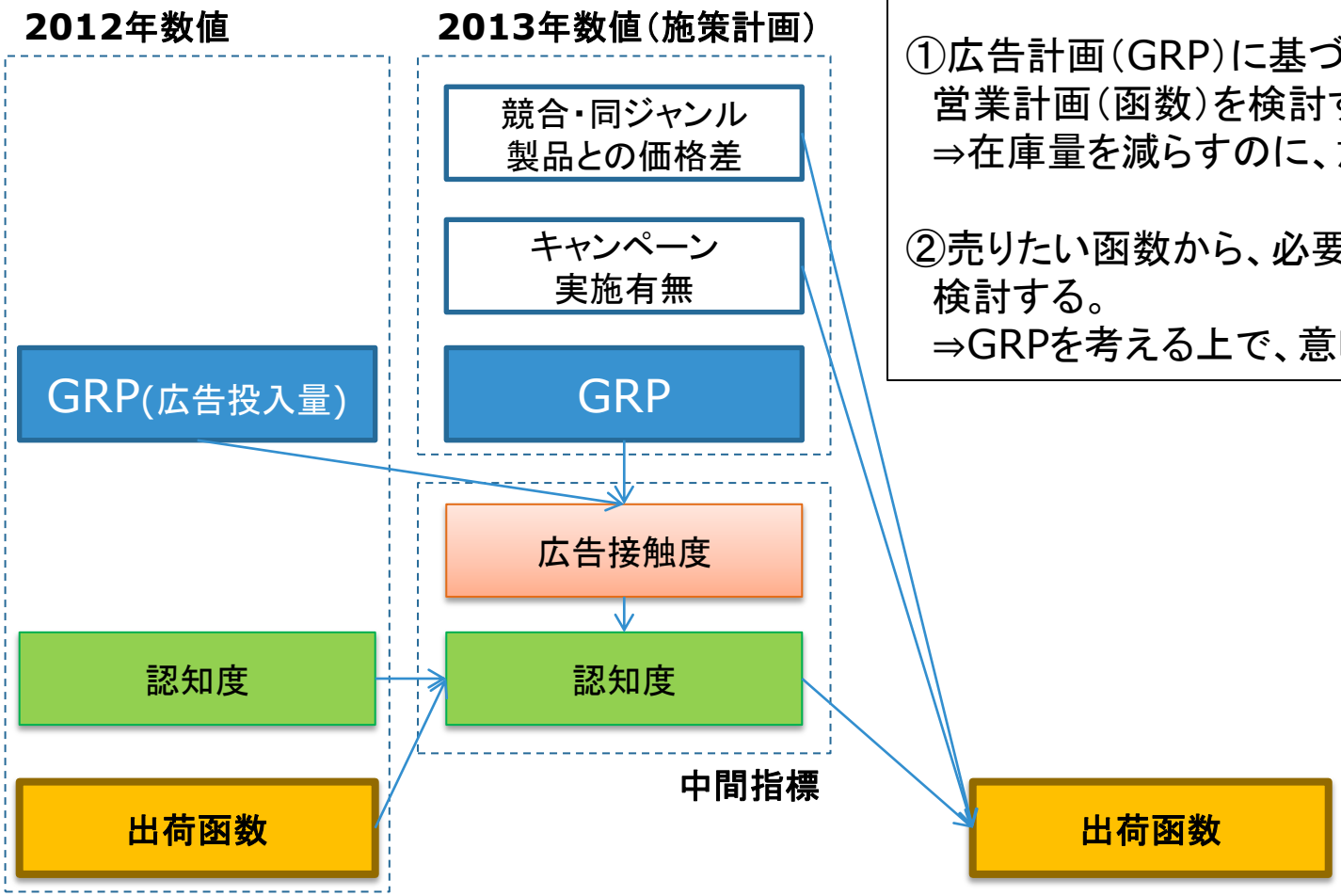
効果の出しやすいテーマか？

主要分析対象	統計・機械学習系			テキスト系		数理計画系
	相関・分類	行動予測・事象予測	定量予測	トピック分析	ネットワーク分析	最適解探索
顧客・消費者	顧客分類（CATV） 顧客分類（金融ｸﾞﾙｰﾌﾟ）	ｱｯﾌﾟｾﾙ候補者予測（CATV） 解約候補者予測（CATV） ﾀｰｹﾞｯﾄ広告（金融ｸﾞﾙｰﾌﾟ） 犯罪・違反発見（金融） 営業訪問先ﾘｽﾄ（EMC社内） TCE（EMC社内）		話題分析（通信） 話題分析（小売） 解約ｲﾝﾌﾙｴﾝｻ抽出(通信）		
商品・サービス	商品ﾚｺﾒﾝﾃﾞｰｼｮﾝ（CATV） 商品ﾚｺﾒﾝﾃﾞｰｼｮﾝ（銀行）		需要予測（中食小売） 広告・販促効果分析（加工食品） 広告・販促効果分析（雑貨）	機器障害問合せ傾向分析（EMC社内）		最適価格設定（航空）
事業活動・その他		設備故障予測（石油精製）	交通量予測（公共）	競合ﾎﾟｼﾞﾈｶﾞ分析（小売）	研究開発ｲﾉﾍﾞｰﾀ検知（EMC社内）	最適航路設定（航空）

ビジネスでの活用可能性検証

業務の精度に合うか？

食品メーカーの需要予測（出荷函数）



業務によってアローワンスが異なる

①広告計画（GRP）に基づいて
営業計画（函数）を検討する。
⇒在庫量を減らすのに、意味のある精度は？

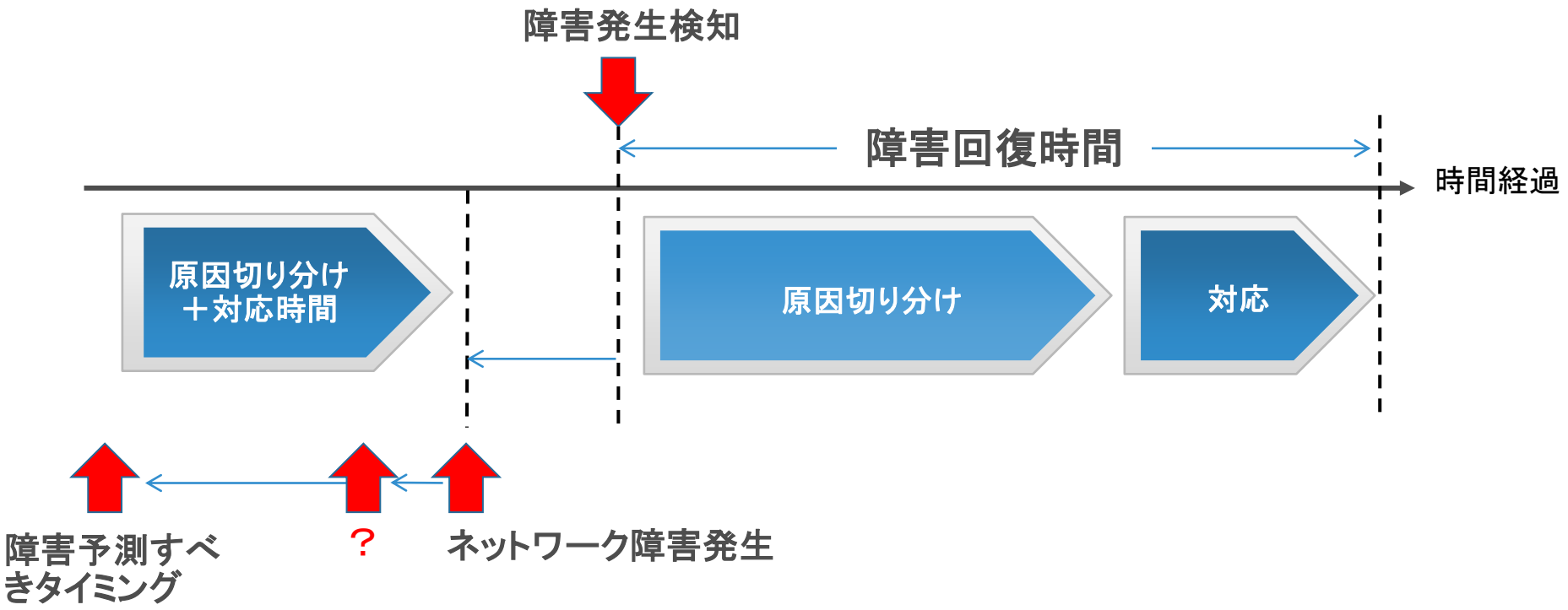
②売りたい函数から、必要GRPを
検討する。
⇒GRPを考える上で、意味ある精度は？

ビジネスでの活用可能性検証

業務のタイミングに合うか？

ネットワーク障害の予測

どのタイミングに予測ができればよいか？

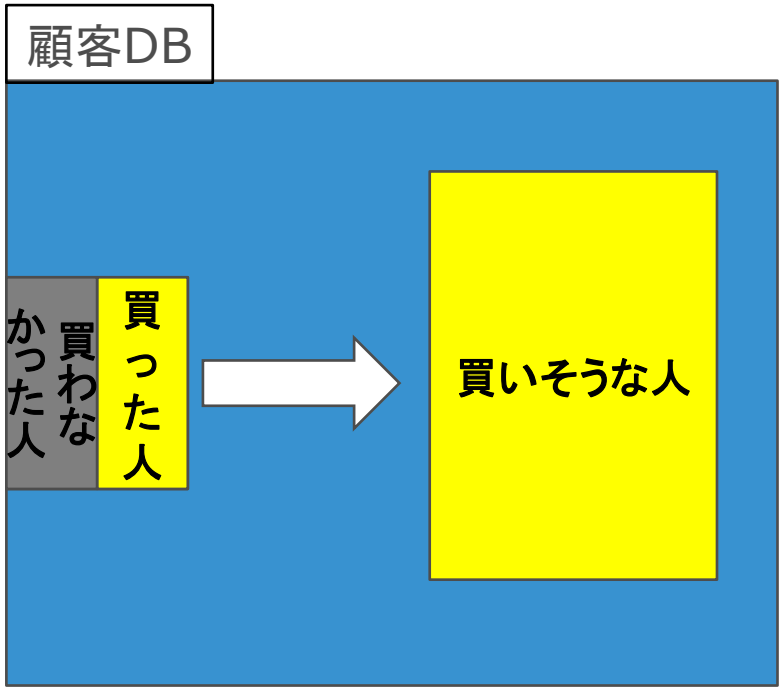


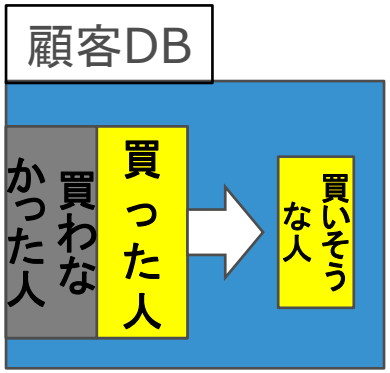
予測ができても、対応できなければ意味がない
対応時間は、環境によって異なる

ビジネスでの活用可能性検証

データ量は十分か？

アップセルの予測は、基本的に成功の横展開





まとめ

- データサイエンティストは手が届くものになってきている
  - データサイエンティストの育成方法が整備されてきている
- ビジネスで実行に移すところに辿り着くところが勝負
  - 何よりトップダウンの方針が重要
  - 効果が出やすいテーマから始める
  - ビジネスで活用できるところまで詰める
  - スモールスタートで半信半疑を克服

- **何より大切なのは**
  「新しい施策」「新しいビジネスモデル」を思いついて、
  失敗を恐れずトライアンドエラーを繰り返す

EMC²