JUAS 一般社団法人 日本情報システム・ユーザー協会
Japan Users Association of Information Systems

2018年度JUAS活動成果報告会

-ビジネスデータ研究会-

2019年4月18日

ビジネスデータ研究会
下田　朋彦（日本航空株式会社）

JUAS

本日のアジェンダ

1.研究概要

① ビジネスデータ研究会の紹介

② 研究テーマの設定

2. 研究成果

① データ連携

② データドリブン

③ データサイエンス

1.研究概要

① ビジネスデータ研究会の紹介

～研究会のミッション～

私たちは、データを事業価値につなげるために必要な
“考え方”、“技法・実行プロセス”、“組織体制・人材スキル”を
探求し、あらゆる事業従事者へ提案することによって、
データに携わる全ての人々が豊かな未来を描き
実現できる世界を目指します。

① ビジネスデータ研究会の紹介

～研究会活動内容～

《実施期間》

2018年5月22日 ～ 2019年3月23日

《主な活動実績》

・定例研究会 ： 11回
・講演会 ： 3回
・合宿 ： 2回
（沼津合宿＋JUAS日帰り合宿）

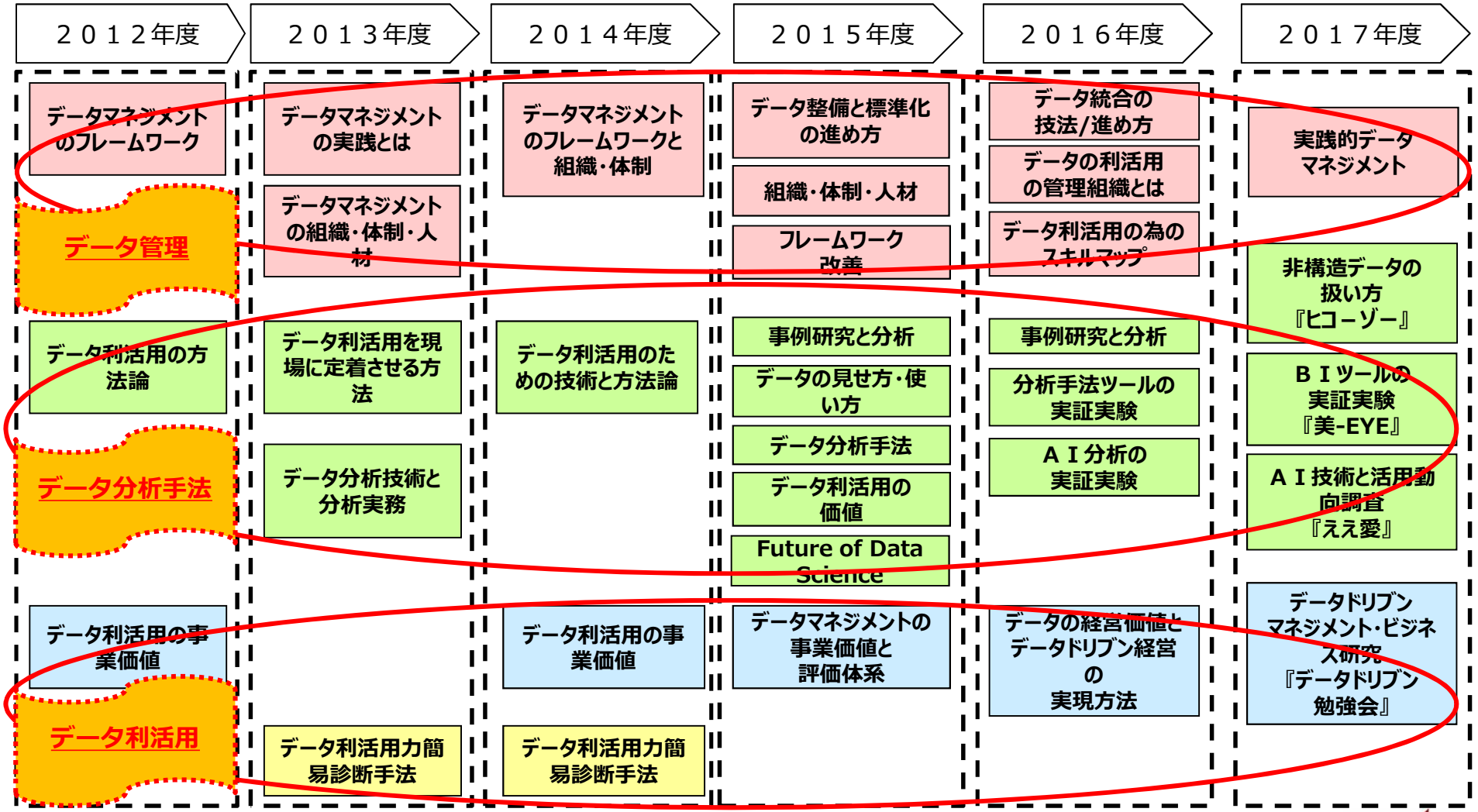
＊その他 各分科会を実施

《参加人数》

47名

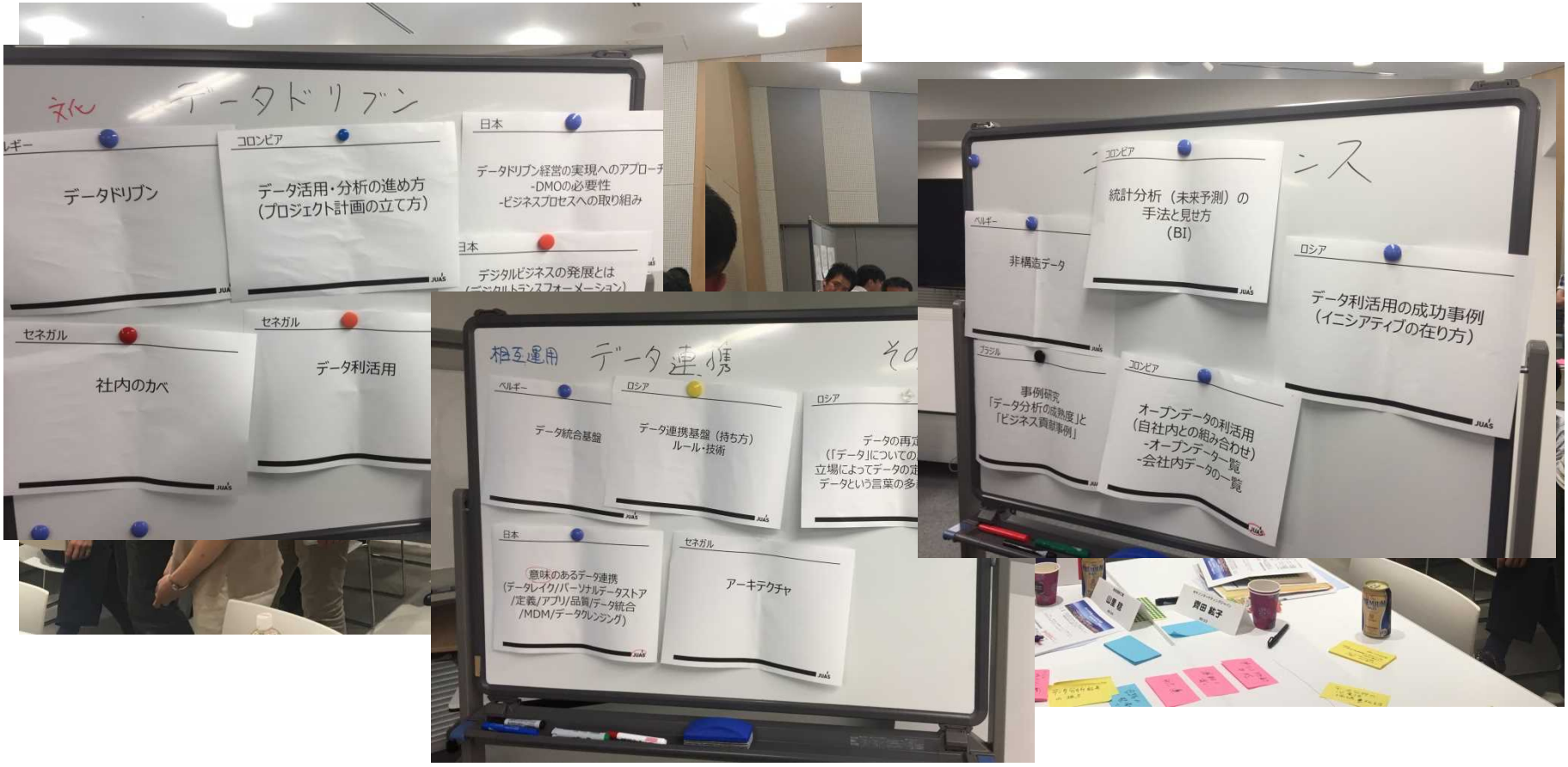
② 研究テーマの設定

・2012年度にデータマネジメント研究会として発足。これまで、6年間データを中心に幅広く研究。



② 研究テーマの設定

・7月27日　沼津プラザヴェルデで行われた合宿で、**全員参加でディスカッション**し、メンバーの課題認識を共有して一気に議論を深めました。
・本来は、合宿２日間で研究テーマを決定ですが、台風接近のため2日目は中止。第4回の定例会で研究会テーマを決定。



② 研究テーマの設定

・３つの分科会に分かれて研究

【第１分科会】；データ連携　10名
・データを社内外間で連携する場合の課題、社会動向について研究

【第２分科会】；データドリブン　16名
・データ利活用組織の成功要因について研究

【第３分科会】；データサイエンス　20名
・データ分析手法の研究及び実践

2.研究成果

全ての分科会で193ページ分となりますので、
一部のご紹介になります。

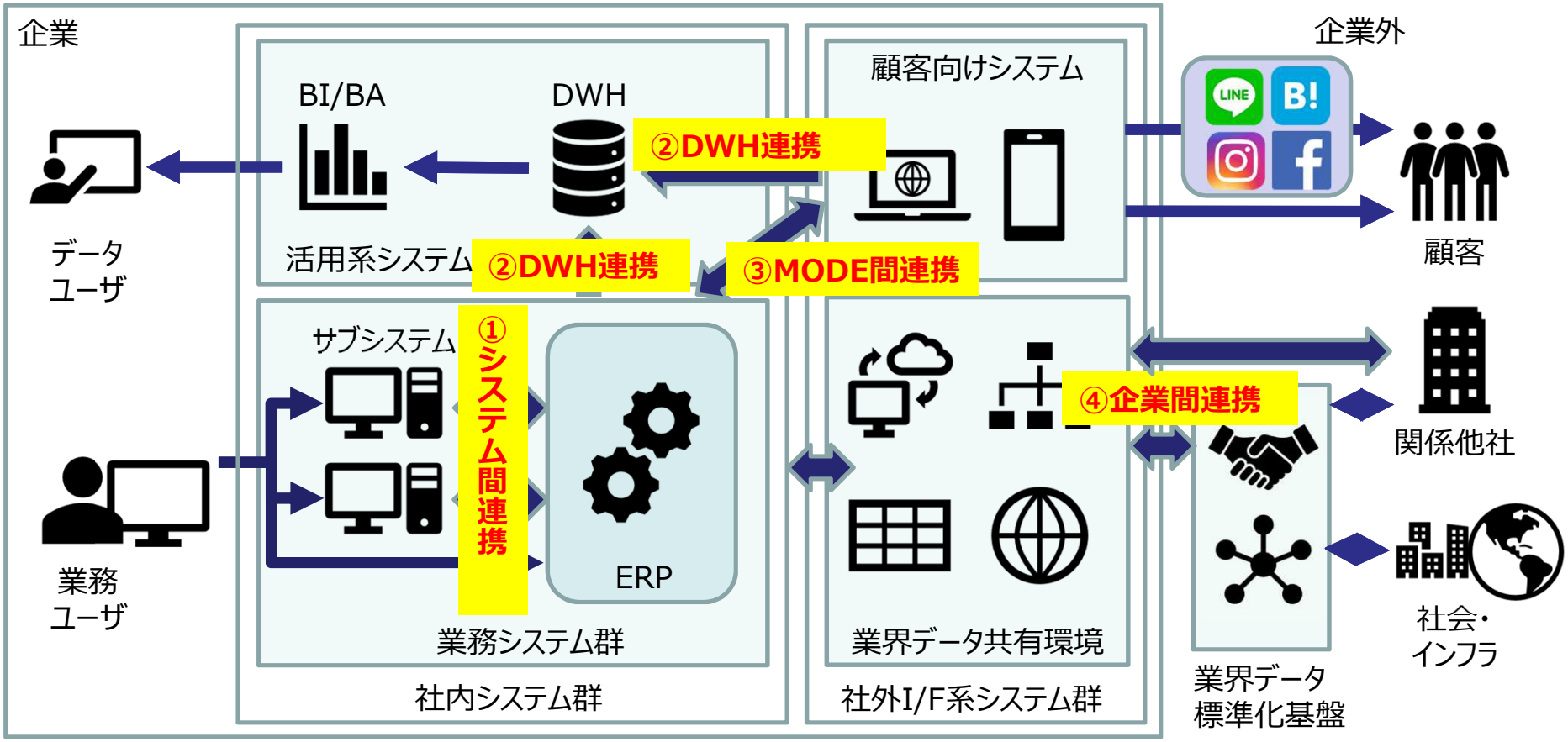
2. 研究成果「①データ連携」

2. 研究成果「①データ連携」

- 本テーマは今年生まれた新しいテーマ
- データ連携について、メンバー各社の置かれている状況を把握する為、全員でのブレストを繰り返した
- 業種、業界が異なる為、悩みごとは各社各様であるが、類型化は可能と判断し、アンケート形式で各社の状況を集約
- 集約した課題を類型化し、解決策や指針を取りまとめることを分科会の目標とした

2. 研究成果「①データ連携」

・まず、メンバー間で課題の共通認識を持つため、データ連携として考えられるアーキテクチャ（モデル）を設定（ブレストと整理）



2. 研究成果「①データ連携」

・６つの課題を洗い出し、今年度は3つの課題について研究。（課題の類型化）

データ連携基盤に関する課題

課題①
データHUBの構築

・マスタデータハブとトランザクションデータハブの構築
・マスタ登録機能、メタデータ管理機能の装備
・協業・関係会社向けAPIの提供　等

課題②
多様な連携方式の提供

・多様なデータ連携手段・タイミング・形式への対応
・それによる要件変化への柔軟な対応や利便性向上
・利用部門の所有するデータの活用　等

課題③
稼働環境設計手法の確立

・複数クラウド・自社DC間連携や分離ネットワーク対応
・大量データ処理や様々なアプリ連携・疎結合の実現
・データ活用と個人情報保護の両立を可能とする環境
等

データ構造や運用に関する課題

課題④
データ仕様の標準化

・コード体系の標準化・共通化、汎用コード体系の採用
・データ項目・オリジナル/コピーの明確な定義
・社外も含めた標準的なデータ仕様への準拠　等

課題⑤
データオーナーの明確化

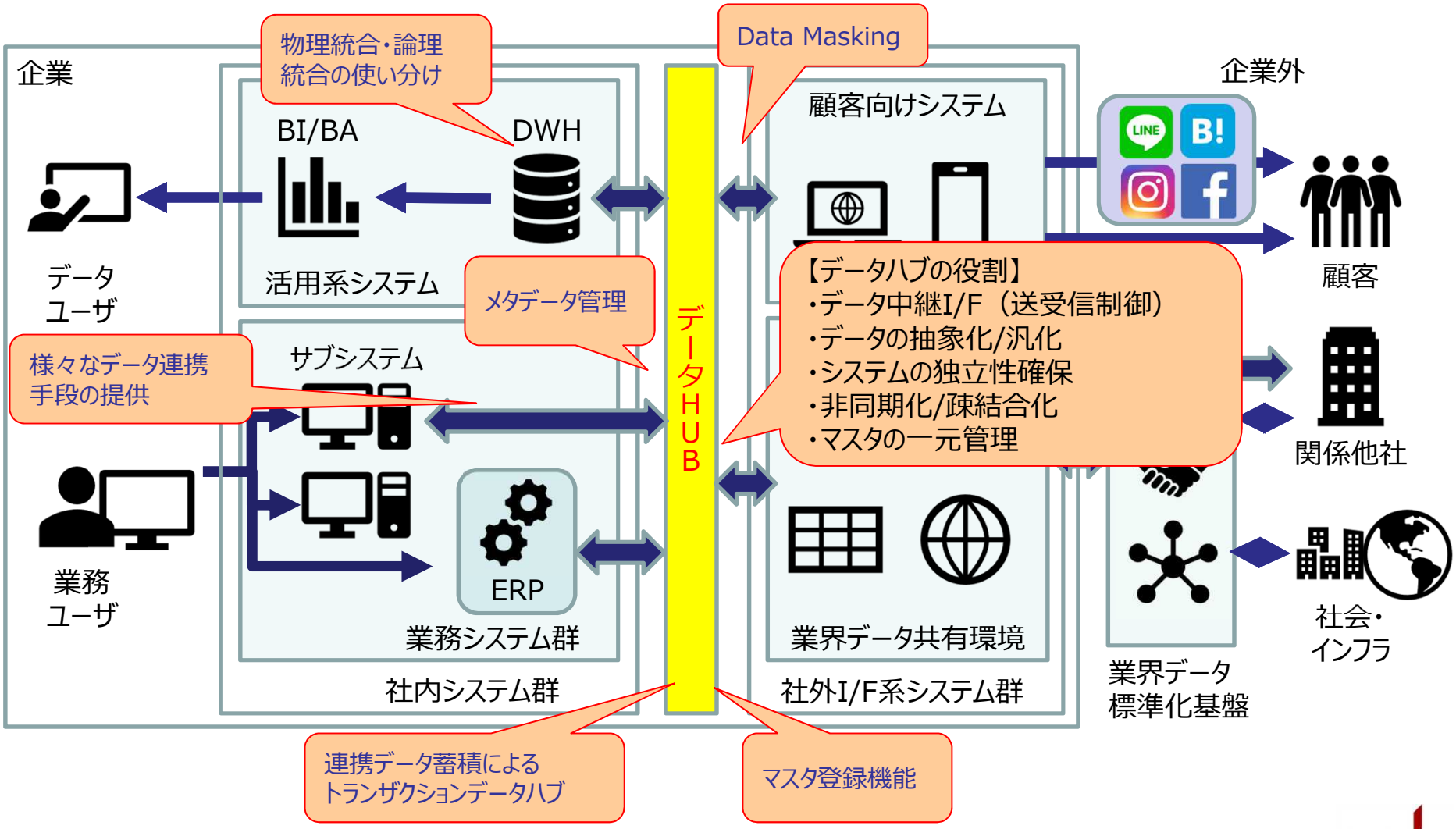
・データオーナーによるコードの管理
・データオーナーによるデータ利用ニーズの取捨選択、
データ定義とその周知　等

課題⑥
データ流通の実現

・データ流通のための市場価格の設定
・標準的なデータ交換ルールの確立　等

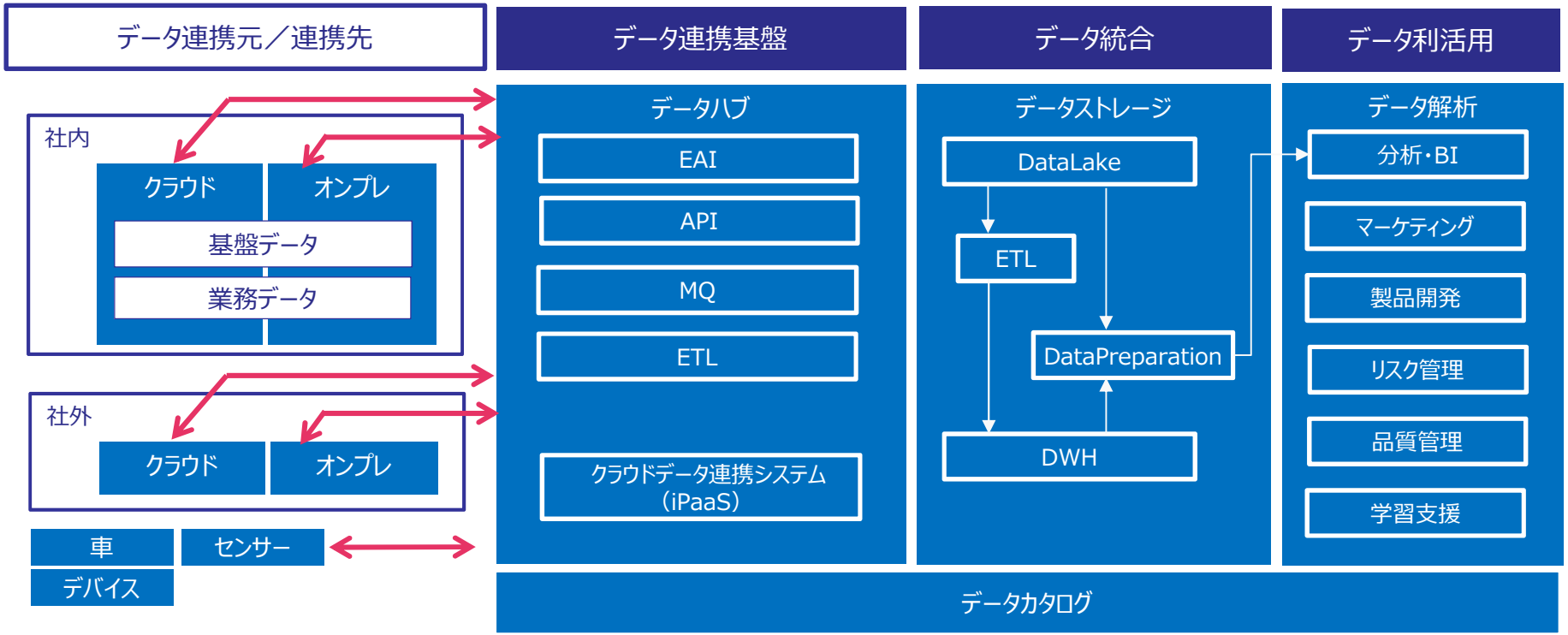
2. 研究成果「①データ連携」~データHUB~

・データHUBの役割を構築する上での課題（留意事項）を整理



2. 研究成果「①データ連携」~多様な連携方式の提供~

・連携方式の一覧化、特性を整理。



- 様々なシステム、端末との連携
- あらゆる機器やサービスとつながる
- データ連携基盤はデータカタログでデータオブジェクトの情報を持つ
- メタデータ管理で必要なデータがどこにあるかを持ち、必要なタイミングで必要なデータを連携
- オーケストレーションに外部クラウドとの柔軟な連携を実現する
- プロトコル変換を行う

- 大量データの蓄積
- リアルタイムでのデータ配信、蓄積
- 高セキュリティでのデータ保護
- データ変換、加工

- リアルタイムでの集計・可視化・分析
- 機械学習等の高度な分析
- ビジネスの提供
- ユーザーへの価値提供

2. 研究成果「①データ連携」~データ流通の実現~

・社外間のデータ連携について社会動向について整理。

5-3-1. 日本におけるデータ流通の課題
5-3-2. データの連携技術動向
5-3-3. データ項目の明確化
5-3-4. データの取引契約（商取引）
5-3-5. データの価格政策（概要）/（プロセス）/（実態）
5-3-6. データ連携に係わる政策（概要）/（事例）

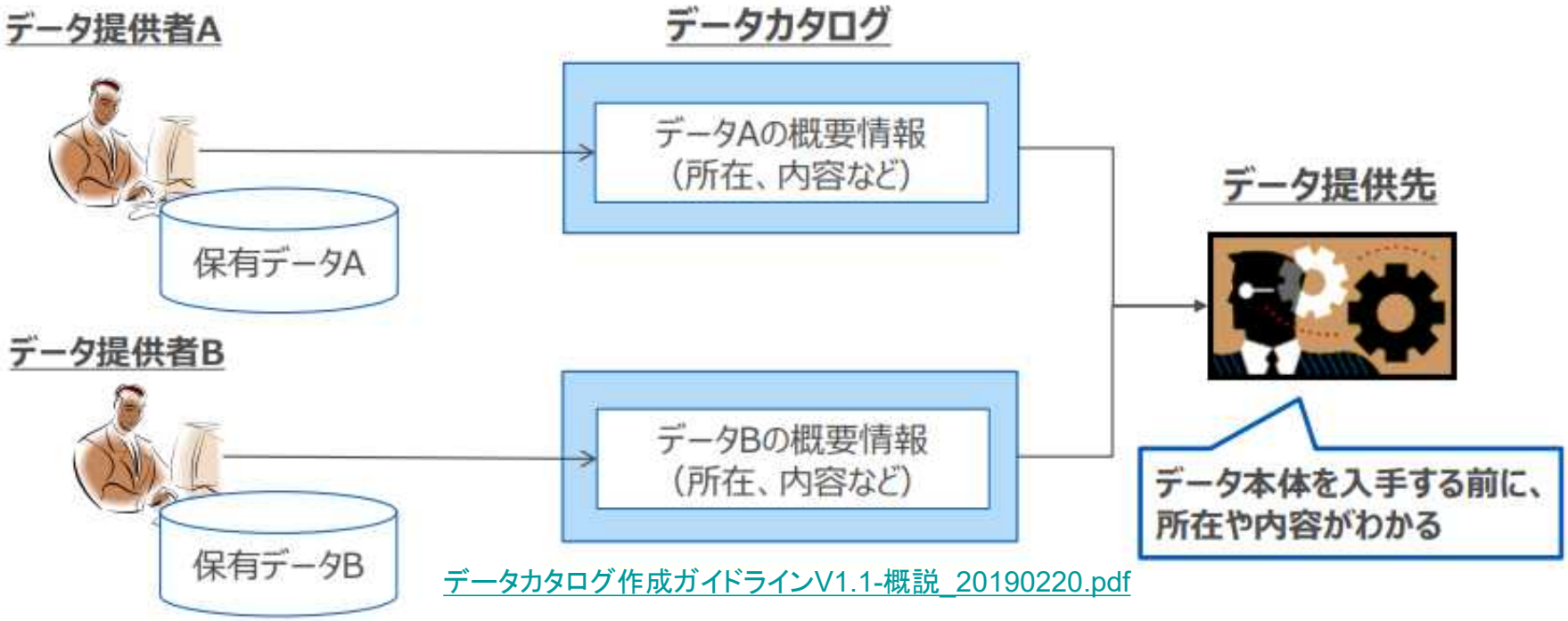
2. 研究成果「①データ連携」~データ流通の実現~

・メタデータはカタログを活用することで、社外利用が促進される。

メタデータ　事例①『データカタログ作成ガイドラインV1.1』

DTA（一般社団法人データ流通推進協議会）が策定

- データ本体を入手する前に、データの概要がわかる
- データの検索性を向上でき、データのやり取りを促進できる



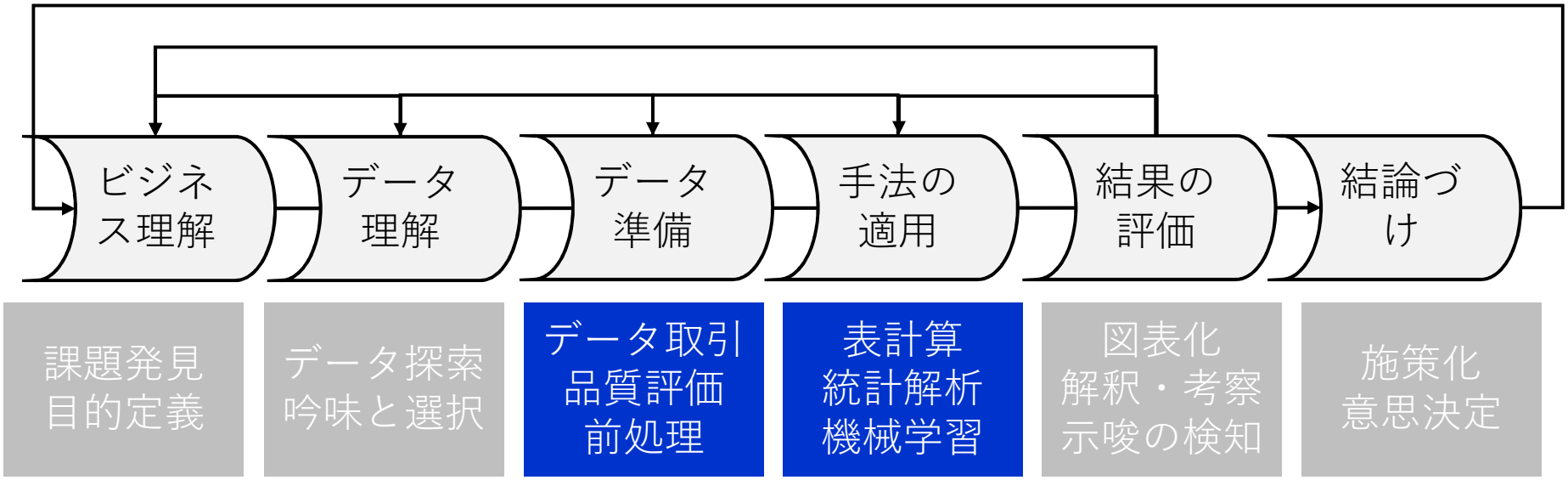
データカタログ作成ガイドラインV1.1-概説_20190220.pdf

2. 研究成果「①データ連携」~データ流通の実現~

・データの値決めはプロセス全体を考慮する。

データ活用プロセス全体をCostとして考える

CRISP-DM（CRoss-Industry Standard Process for Data Mining



- 注目されがちな「データ準備」「手法の適用」だけでなく、工程全体に注目を。
- 自組織が何に習熟し、どこまで内製/外部連携するかを見極めることが重要。
- 各工程を高スキル/低コストで実現できれば、投資対効果も得やすい。

2. 研究成果「①データ連携」~データ流通の実現~

～まとめ～

- 業種、業界、経験が異なるメンバーが集まり、データ連携についての課題の抽出を実施
  ⇒各社データ連携について多くの課題を抱えていることが判明

- 数多く出た課題を類型化
  ⇒分類が難しかったが、研究しやすい6つの課題へ分類

- 6つの課題内、「課題①データハブの構築」「課題②多様な連携方式の提供」「課題⑥データ流通の実現」の研究を実施

②データドリブン

2. 研究成果「②データドリブン」

2014 ▶ **VS（バリューストラクチャ）モデル**
経営/業務/業務・IT/ITの4層で効果や価値のつながりを整理

2015 ▶ **VSSC（Value Structure Score Card）**
企業におけるミッションとVSを紐付けてアプローチ

2016 ▶ **データドリブン経営と云うデータマネージメントが提供する価値**
定義と状態の整理

2017 ▶ **データドリブン経営の成熟度モデル**
経営/業務/組織・人材/IT・情報システムの4視点での評価に挑戦

2018 ▶ **そして　2018年度は！！**

2. 研究成果「②データドリブン」

・これまでも研究されてきたが、事例を基に再度「成功の鍵」を研究。

事例を通して自社の課題や確認したいポイントをクリアにする。

◆ 目的、ゴール
事業戦略、経営課題
新しいビジネスの創出、既存ビジネスの効率化

◆ 取組み、体制
プロセス、ルール、IT、組織、人材、予算

◆ 実現に向けた課題、どう解決したか
経営層のコミットメント、体制、位置づけ
業務側、IT側の連携、問題意識
データの精度、粒度、意味、有無
対象スコープ、MDM、アーキテクチャ
人材、CDOの存在

検討結果を
「成功の鍵」
としてまとめる

2. 研究成果「②データドリブン」

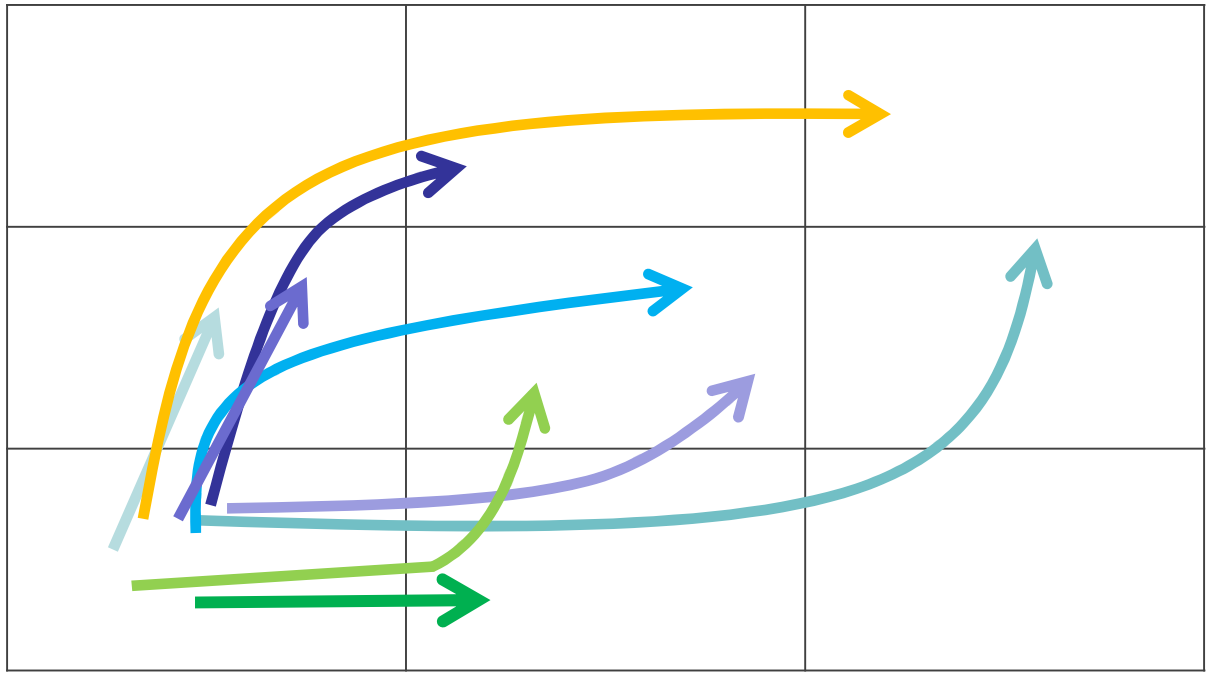
・2社への実際にヒアリング及びメンバー各社の状況を可視化。

変革
- 基準が全社でルール化されている
- 重要経営指標数値を経営陣がダッシュボードなどで確認でき、迅速な経営判断に結び付いている

経験
- 事業部門と中央の部門が連携し、全社的なデータ管理の基準ルール作りの議論がはじまっている

入門
- 基準がほとんど設定されていない
- プロダクト毎、部門毎にサイロ化された構造化データを特定業務で活用している

データ管理の熟練度



データ分析の習熟度

入門
- 主にスプレッドシートを利用
- 必要に応じて分析は個々に実施
- アナリティクス担当者の採用は困難

経験
- アナリティクスツールのポートフォリオの拡大中
- アナリティクス担当は事業部門に所属
- 部門研修と外部人材採用に注目が集まっている

変革
- 包括的なツールのポートフォリオを活用し、高度なアナリティクスモデリングをサポート
- 事業部門が高度なアナリティクススキルとガバナンス機能を備えた中央の部門とが結びついている企業が多い

参考資料；IBM社
「データドリブン経営」現実の裏側ーデータサイエンスを武器とした企業に変革できない理由
https://www.ibm.com/blogs/think/jp-ja/data-driven-management/

2. 研究成果「②データドリブン」

～データ管理の熟練度～

■ 成功要因

- 経営TOPの強い意志
- 事業部からの要望対応
- 顧客からの要望対応
- 社員の分析スキル平準化への危機感
- 個人情報保護法の施行によるデータ管理方針の明確化されたため
- 労働基準法の改正などにより労務管理の強化が必要となったため

■ 階段を上れていない要因

- 社員の分析スキル不足
- 縦割り組織の弊害
- いまはその時期ではない

2. 研究成果「②データドリブン」

～データ分析の習熟度～

■ 成功要因

＜環境変化＞

- 経営TOPの強い意志とそれを実行できる環境つくり
- 世間トレンドを利用して危機感醸成
- 本業がシュリンクすることの危機感
- ツールの進化とそれを活用できる人材の登用
- 事業の環境変化（上場のタイミングでのコンプライアンス強化、労務管理の強化等）
- 外部から専門家を採用

＜データ活用部門の努力＞

- 小さい案件を選び、成功事例を早い段階で作った
- 短いサイクルでのPDCA
- ユーザに分析結果を可視化して説明、積極的なコミュニケーション

2．研究成果「②データドリブン」

～まとめ～

- データドリブン経営実現に向けた参加各社の問題意識に対し、確認したいポイントを事例の検討や先進企業へのヒアリングを通して具体化し、まとめる事で、課題をブレークスルーするヒントを得る事ができた。

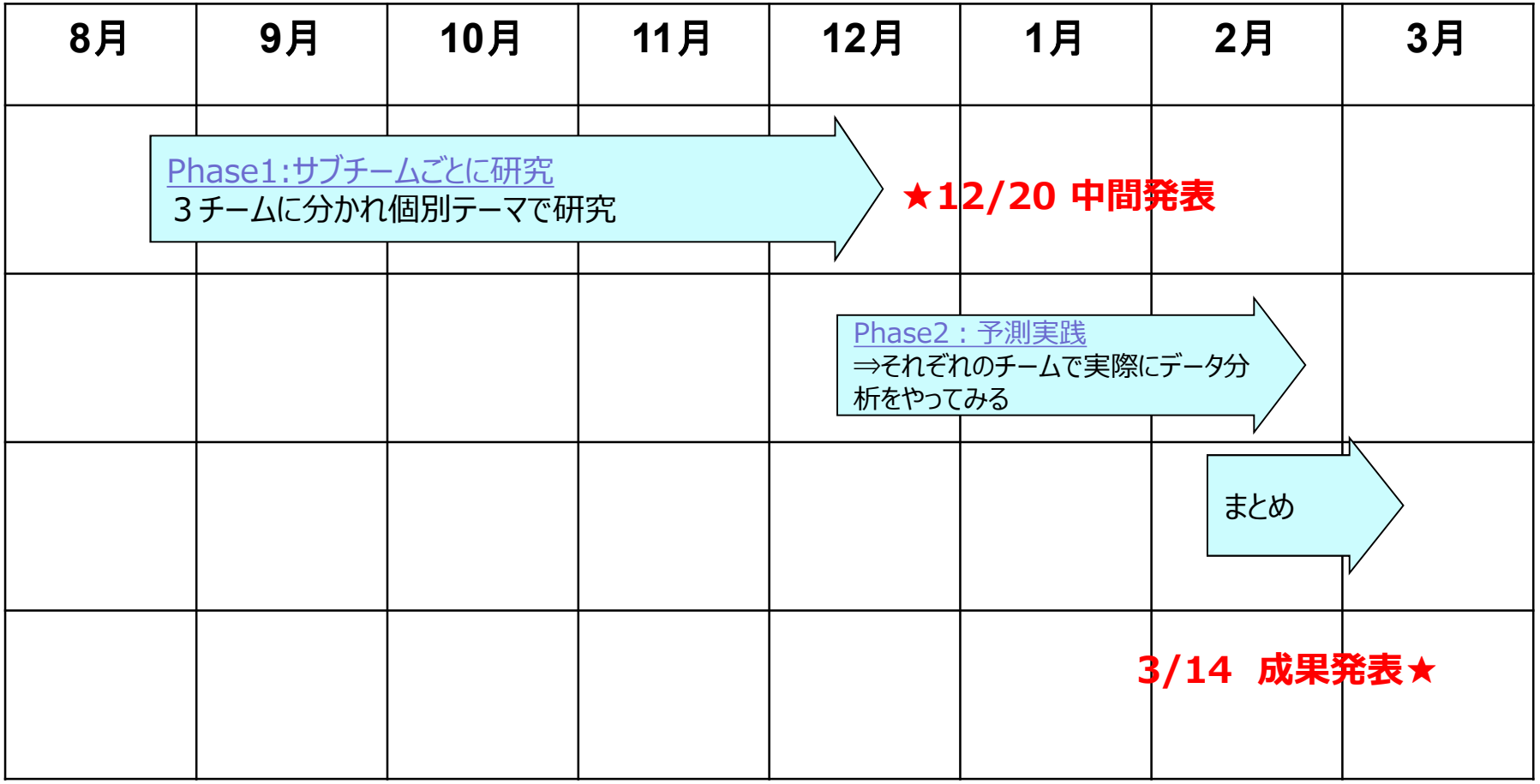
- 「データ分析」「データ管理」という２軸の熟練度で自社の立ち位置やアプローチをマッピングする事で「データドリブンな経営の実現」に向けて、大きく２つの方向がある事と、今後取るべきアプローチの示唆や共通の理解を得やすい事が分かった。

「預言者の秘密道具」
(データサイエンスを用いた未来予測)

～データサイエンスを用いた～
未来予測

2. 研究成果「③データサイエンス」

～スケジュール～



2. 研究成果「③データサイエンス」

～Phase1 サブチーム毎の研究～

・3つのサブチームに分かれて研究実施

分析・予測対象データ（オープンデータなど）の収集・加工

分析・予測手法（統計手法）の調査・研究

分析・予測手法（ツール・PythonやRなどプログラミング言語）の調査・研究

2．研究成果「③データサイエンス」

～オープンデータの収集～

・オープンデータの定義、目的、歴史、分類、加工技術を整理

【利活用目的の視点】

分類	利用データ	目的
スタンドアロン型	オープンデータ単体	単一データにて時系列分析・属性別比較
レイヤー型	複数のオープンデータの組合せ	複数の粒度を合わせて多目的分析
価値創造型	オープンデータ（複数）*独自の分析モデル	推定などによって新たな指標を導出
自社ビジネス高度化型	（オープンデータ+自社データ）*独自の分析モデル	自社データの不足分を補完

【ビジネスモデルの視点】

分類	特徴	主な企業
付加価値型	・既存ビジネスの価値を高めるためにオープンデータを利用 ・ データの加工は可視化などが主であり複雑な処理はしない ・競合相手もオープンデータを自由に利用できるため、既存ビジネスの優劣を極端に変えることはない	市場のリーダー
プラットフォーム型	・特定の領域のデータを大量に集め、プラットフォーム化する ・集めたデータを利用しやすく提供することで最初の価値を生み出す ・データの利用状況や利用者の状況を分析することで、さらに新しい価値を生み出していく	スタートアップ
新価値創造型	・オープンデータを含む多様なデータをかけ合わせ、高度な分析によって未来を予測する ・価値を生み出す源泉は新しく開発したアルゴリズムや分析モデル ・オープンデータはアルゴリズムや分析モデルを開発する際にも利用される	スタートアップ

2. 研究成果「③データサイエンス」

～分析・予測手法（統計手法）の調査・研究～

- 「タイタニック号の生存者予測」を実施し、その結果をkaggleに投稿する

実施したこと①

環境構築

Anaconda

kaggle

・Anaconda3.7のインストール
・kaggleよりタイタニック分析用データのダウンロード

実施したこと②

分析

①サイトを参考にして、Anacondaにサンプルソースをコピー＆ペースト

②分析結果ファイルをkaggleにアップロード

③結果を受けて他サイトの内容を試行
以降、①～③を繰り返す

kaggle

2．研究成果「③データサイエンス」

【実施結果】

試行回	分析手法	評価項目	分析手順	正答率
１回目	ランダムフォレスト	性別、等級	Web上の参考ソースをそのまま コピペ	76%
２回目	ランダムフォレスト	性別、等級、年齢	１回目の項目に年齢を追加 欠損値は中央値を補完	72% （※１）
３回目	回帰分析 （※２）	性別	生存０、死亡：１ 男性：０、女性：１ として回帰分析 ⇒解答項目を四捨五入	-

※１：項目値を増やしたが欠損値の補完がうまくいっていないため、
　　却って正答率をさげた
※２：分析する内容（生死判定）に適当な手法でなかったため、
　　正しく分析できず。無理やり四捨五入して解答を作成も、投稿せず

2．研究成果「③データサイエンス」

～分析・予測手法（プログラミング言語）の調査・研究～

◆テーマ詳細

データ分析、未来予測をやる上で使用するツール、プログラミング言語が多数あることが分かっている。これらの内容を把握し、可能なものについて実際に試用する

◆取り組み内容

1．多数あるツール、プログラミング言語の一覧をまとめる
   - Web、各社での事例を元に収集

2．ツールを実際に使ってみてデータ分析を試してみる
   - 無償のツールを実際にインストールし試す
   - 実データを元にツールで分析

2. 研究成果「③データサイエンス」

～分析・予測手法（プログラミング言語）の調査・研究～

No	ツール名	種類１	種類2	有償or無償
1	R	プログラム言語	分析	無償
2	Python	プログラム言語	分析・可視化	無償
3	Julia	プログラム言語	分析	無償
4	Matlab	ツール	分析	有償
5	SPSS	ツール	分析・可視化	有償
6	SAS	ツール	分析・可視化	有償
7	Rapidminer	ツール	分析・可視化	有償/無償版あり
8	DataRobot	ツール	分析	有償
9	Tableau	ツール	可視化	有償/無償デスクトップ版あり
10	Power BI	ツール	可視化	有償/無償デスクトップ版あり
11	QlikView/QlikSense	ツール	可視化	有償/無償デスクトップ版あり
12	MotionBoard	ツール	可視化	有償
13	Alteryx	ツール	前処理、加工	
14	Excel	ツール	分析・可視化？	無償版/有償

2. 研究成果「③データサイエンス」

～分析・予測手法（プログラミング言語）の調査・研究～

No	ツール名	種類１	種類2	有償or無償
15	Azure ML			有償
16	KNIME	ツール	加工・分析	無償版/有償
17	Glue	ツール	加工	有償
18	DataStudio	ツール	可視化	?
19	Selenium	言語	収集	?
20	CasperJS/phantomJS	言語	収集	?
21	KH Coder	ツール（テキストマイニング）	可視化	?
22	見える化エンジン	ツール（テキストマイニング）	可視化	?
23	BigQuery	ツール	蓄積	?
24	CloudDataFlow	ツール	加工	?
25	Data Diver	ツール	分析	?
26	Ｄａｔａ Ｐｒｅｐ	ツール	加工	?
27	D3.js	ツール	可視化	?

2．研究成果「③データサイエンス」

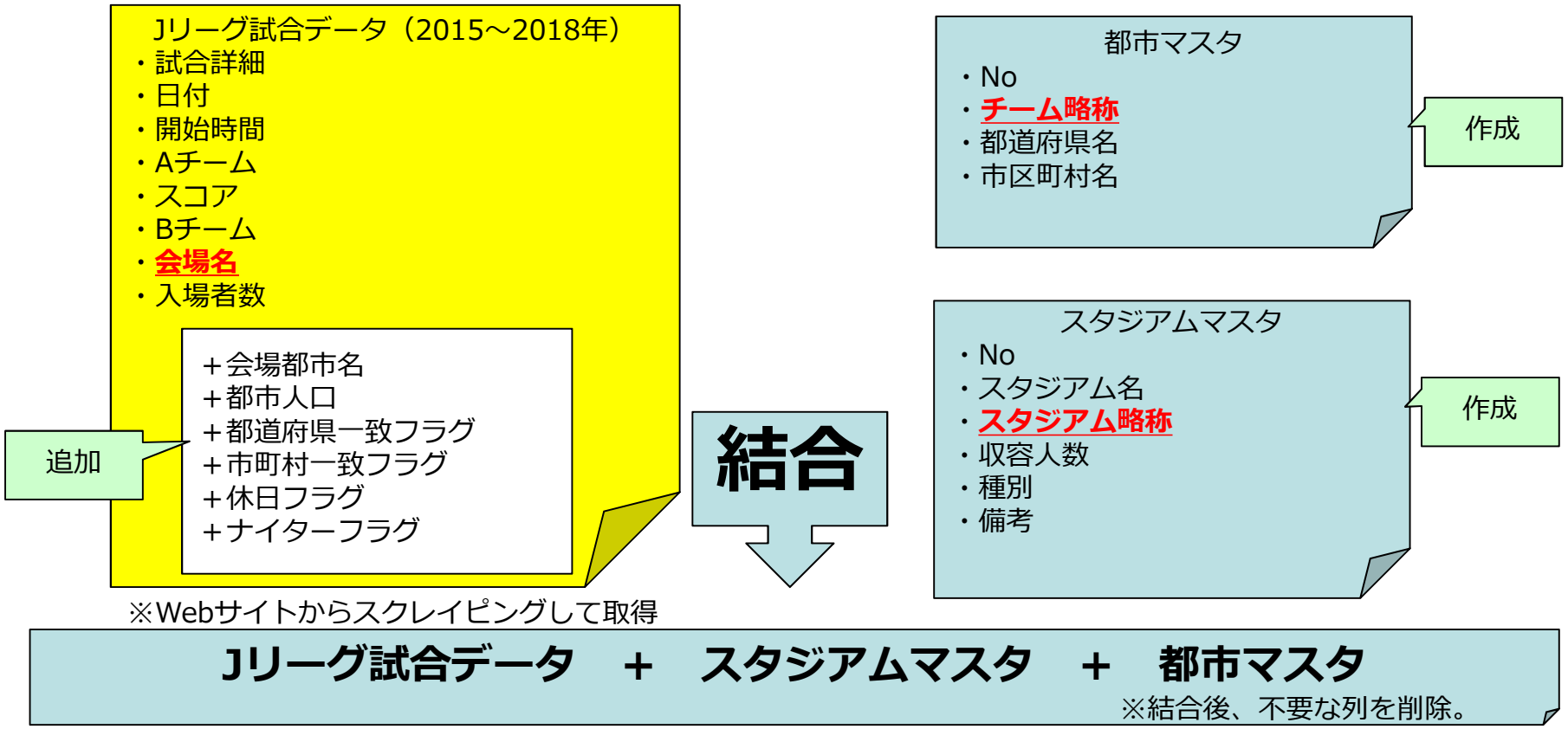
～Phase2 予測実践～

Jリーグの2019年開幕戦の入場者数予測

- 学習データはオープンデータを加工して利用
- 利用ツールは「Knime」
- 分析手法は「重回帰分析」

３つのチームに分かれて各チームで利用データを決定し、予測実践

2. 研究成果「③データサイエンス」



2015～2017年分をモデル作成用データ、
2018年分をテストデータとして活用。
2019年開幕戦の入場者数を予測。

2. 研究成果「③データサイエンス」

～Aチーム～

目的変数

目的変数	加工内容
入場者数	入場者数（例：19,032人⇒1932人となってしまう）の不整値を修正

説明変数に採用した項目

説明変数	初期項目	追加項目	追加理由
試合名			
試合詳細			
Aチーム	○		
Bチーム	○		
休日フラグ		○	休日の方が入場者が多い
両チームの合計総年棒		○	有名選手が多い方が観客が増える
開始時間フラグ		○	日中のほうが観客が増える
ホームの平均入場者数		○	ホームの人気に左右される
両チームの日本代表メンバーの合計		○	実力ある選手は集客効果がある

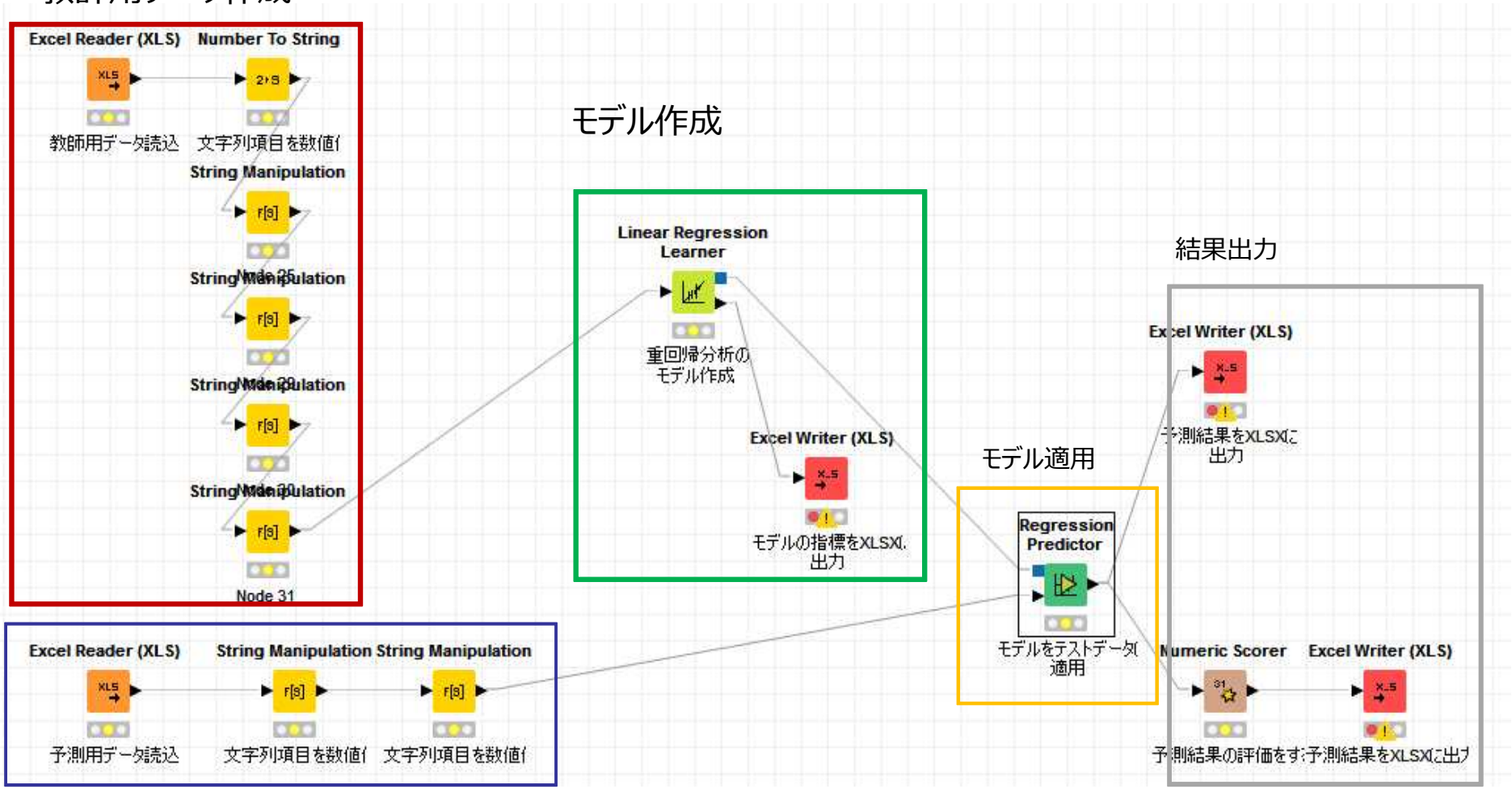
説明変数から除外した項目

説明変数	除外理由
日付	休日フラグに編集
会場名	毎年変わる
年	意味を持たない
No.	意味を持たない

2. 研究成果「③データサイエンス」

～Aチーム～

教師用データ作成



テストデータ作成

2．研究成果「③データサイエンス」

～Bチーム～

目的変数

目的変数	加工内容
入場者数	入場者数の不整値を修正（例：19,032人⇒1932人となってしまう）

説明変数に追加

説明変数	Bチームのみ	追加理由（仮説）
Homeチームの前年度の入場者数（最大、最小、平均、標準偏差）	○	・昨年度の実績値をもってくることでより説得力のある説明変数になると思慮。 ・経験的にフラグ値のみで数値を予測することは難しい。連続値を予測する際には連続値を上手く特徴量として加えてあげることで予測精度が上がる。
雨フラグ	○	天気が悪いと入場者数がすくなくなるはず。
HomeチームとAwayチームの前年度の順位	○	昨年度の順位が高いほど翌年の入場者数は多いはず。Ｊ１～J2の順位を連番にすることで、カテゴリの差を表現。（Ｊ１の最下位は１８、J2の１位は１９）
同郷フラグ	○	人口の多い都市開催した方が入場者数が多い傾向がある
収容人数		会場の規模が大きいと入場者数の最大値が上がる想定
休日フラグ		土日祝日は入場者数が多い傾向にある

2. 研究成果「③データサイエンス」

～Bチーム～

- チューニングは実施せず、デフォルトパラメータで実施。
- いくつか試した結果、勾配ブースティングをベースとした「Catboost（機械学習）」の精度が一番良い。
- 連続値の予測といっても、重回帰分析が良いとは限らず。データの性質や量によってアルゴリズム毎に分析結果に差が出るため、まずは主要なアルゴリズム（デフォルトパラメータ）にて分析モデルを構築し、モデルを選別していくことが必要。

分析モデル	結果			
	訓練データ		検証データ	
	決定係数	RMSE	決定係数	RMSE
ランダムフォレスト	0.97	1624.5	0.78	3960.186
XGBoost	0.91	2641.84	0.81	3660.91
LightGBM	0.91	2665.46	0.82	3551.5
CatBoost	0.89	2903.69	0.83	3472.04
重回帰	0.82	3793.844	0.77	4038.1577
ラッソ回帰	0.82	4033.489	0.77	3793.873
リッジ回帰	0.82	3797.814	0.77	4025.85
エラスティックネット	0.8	4860.93	0.78	4018.98

2. 研究成果「③データサイエンス」

～Cチーム～

目的変数

目的変数	加工内容
入場者数	入場者数の不整値を修正（例：19,032人⇒1932人となってしまう）

説明変数に追加した項目

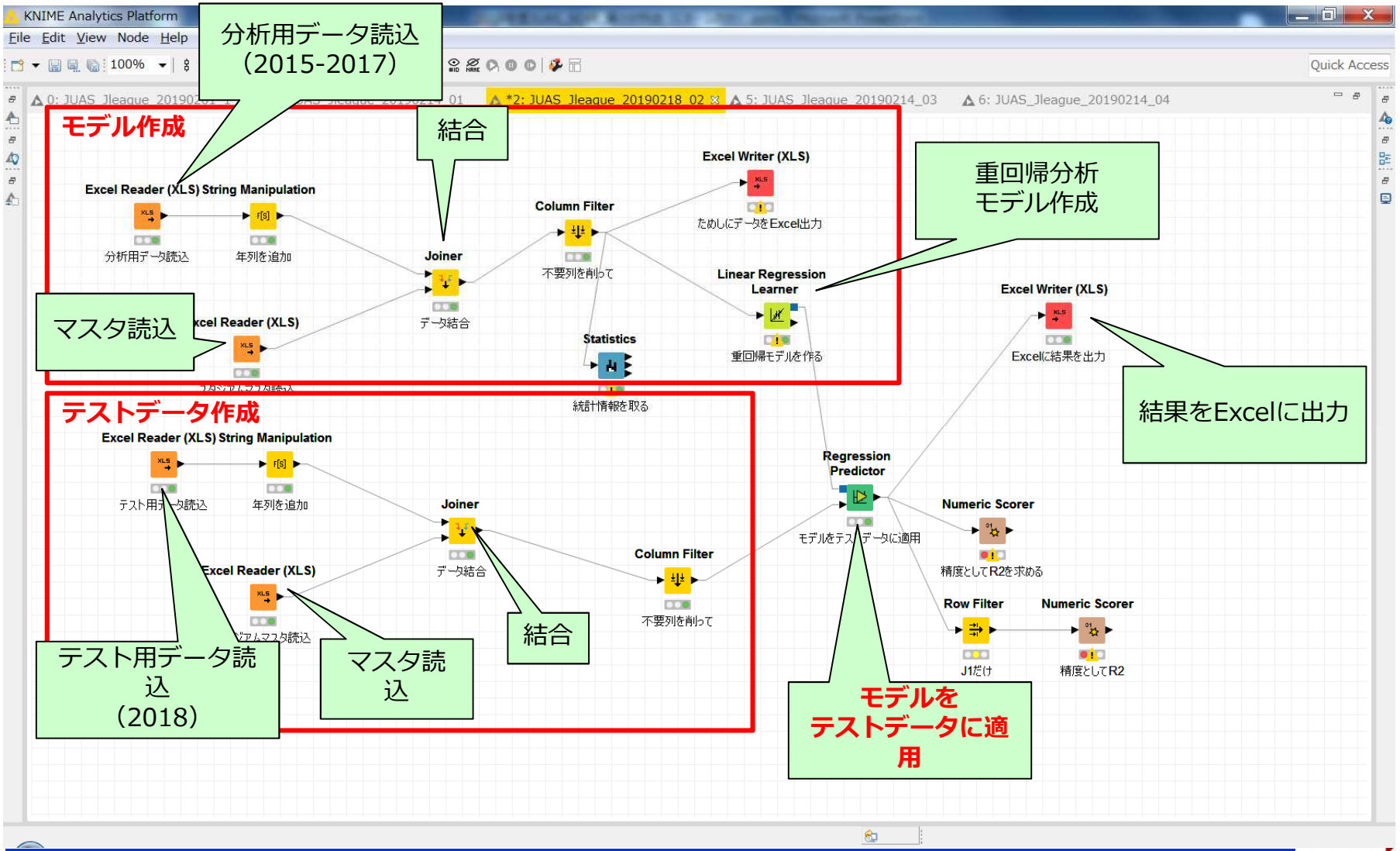
説明変数	追加理由（仮説）
休日フラグ	土日祝日は入場者数が多い傾向がある
ナイターフラグ	デーゲームより、ナイターの方が入場者数が多い傾向がある
会場都市	政令都市開催は入場者数が多い傾向がある
都市人口	人口の多い都市開催した方が入場者数が多い傾向がある
市町村一致フラグ	ホームタウン以外での主催試合は入場者数が下がる傾向がある ※ホームタウンと会場の都市が一致していたら1を設定
収容人数	会場の規模が大きいと入場者数の最大値が上がる想定

説明変数から除外した項目

説明変数	除外理由
日付	休日フラグに編集
会場名	毎年変わる
年	意味を持たない
No.	意味を持たない

2. 研究成果「③データサイエンス」

～Cチーム～



2．研究成果「③データサイエンス」

～まとめ～

【結果】

		Cチーム	Aチーム		メンバー		Bチーム
			平均入場者数無	平均入場者数有	単純なツリーモデル	若干複雑なモデル	
ホーム	アウェイ	差	差	差	差	差	差
C大阪	神戸	26080	-9170	17527	27398	18349	17801
仙台	浦和	-1013	-5768	439	-1282	-2095	-237
鹿島	大分	2564	1238	-183	3452	-1322	1261
川崎Ｆ	FC東京	624	-3179	-1060	3057	-1242	1245
湘南	札幌	2050	-3837	3146	2733	1785	-333
磐田	松本	-910	-4188	-2899	-5380	-3197	-4630
G大阪	横浜FM	5200	6538	3432	7215	228	3703
広島	清水	-1149	-5704	-2048	1001	-3703	-4137
鳥栖	名古屋	5212	-1534	3541	5299	595	3020
	合計	38658	-25604	21895	43493	9398	17693

差=実入場者数-予測入場者数

- トライアンドエラーを繰り返し、段階的に精度を上げていく必要があった
- オープンデータで取れるデータが限られる、精度、サイズなどハンドリングも大変
- 話題の選手(神戸：ビジャ、ポドルスキ、イニエスタ、鳥栖：トーレス)が多いと予測よりも実数が多めに出る。（ただし、Ａチームは人気も加味できた？？）

2019年度
「ビジネスデータ研究会」
引き続き開催します！

～2019年度　研究会ミッション～

ビジネスにおけるデータ利活用の重要性と可能性を追究し、データに携わる多く方々に提案することによって、事業活動の未来に希望を持つことができる研究会を目指します。

～研究テーマ案～

①データ利活用、データドリブン経営に必要な組織・体制・人材を研究

②データに価値を付加するプロセス・分析手法を研究

③業界を超えたデータ連携実現に向けた、データ形式・連携方法・セキュリティ・制度について研究

皆さまのご参加をお待ちしています！

ご清聴ありがとうございました