


ChatGPTを使った分析例

宮崎大学
田中秀典

 ChatGPT (GPT-4) が出力したスクリーンショットを使用しています (2023年8月28日実行)



解説編の資料

1. 「Google」等で「OIRE」をキーワード検索
2. HP上部の「活動内容」をクリック



3. URL欄にある

<https://ir.yamagata-u.ac.jp/activity/>

の後ろに「20230829pm」と入力



ChatGPT



OpenAIが2022年11月に公開した人工知能チャットボットで生成AIの一種

2022年11月30日にプロトタイプとして公開され、幅広い分野の質問に詳細な回答を生成できることから注目を集めた。しかし、人間が自然と感じる回答の生成を特徴としていることから、**一見自然に見えるが事実とは異なる回答を生成**することもあり、大きな欠点とされた。

参照：Wikipedia



ChatGPTにおけるデータの取り扱い



[All Collections](#) > [OpenAI API](#) > [Policy](#) > [How your data is used to improve model performance](#)

How your data is used to improve model performance

Learn more about OpenAI's data usage policies for our API, ChatGPT and DALL-E



Written by Michael Schade
Updated over a week ago

<https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance>



ChatGPTにおけるデータの取り扱い

ChatGPT

When you use our non-API consumer services ChatGPT or DALL-E, we may use the data you provide us to improve our models. You can switch off training in ChatGPT settings (under Data Controls) to turn off training for any conversations created while training is disabled or you can submit [this form](#). Once you opt out, new conversations will not be used to train our models.

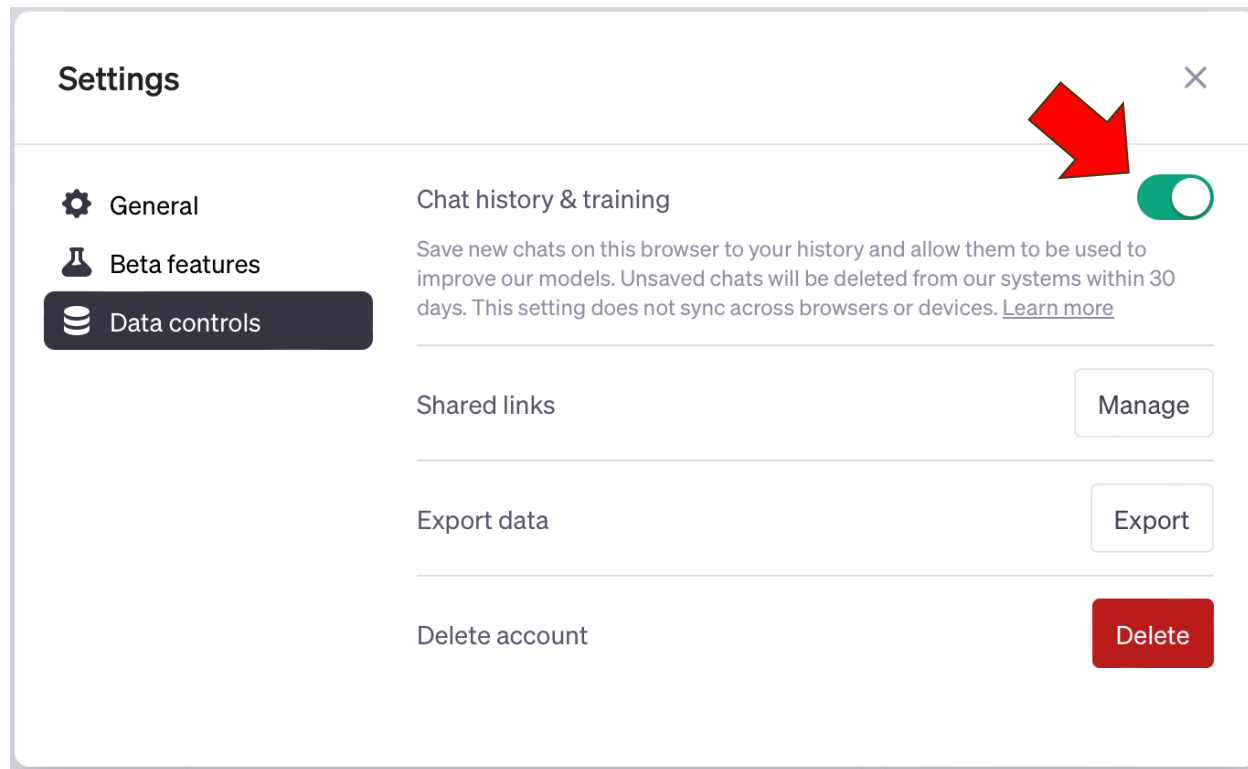
非APIコンシューマサービスChatGPTやDALL-Eを使用するとき、**私たちのモデルを改善するために、あなたが私たちに提供したデータを使用することがあります。**

ChatGPTの**設定でトレーニングをオフに切り替えるかフォームを送信すること**で、オフになっている間に作成されたいかなる会話からの**トレーニングをオフにできます。**

一度オプトアウトすると、新しい会話は私たちのモデルを訓練するために使用されません。



オプトアウト



しかし機能が制限されてしまう。



ChatGPTにおけるデータの取り扱い

とはなっていますが、学内のデータを学外のサービスへ提供（送信）すること自体が（学内規定等で）問題となる可能性が考えられる。

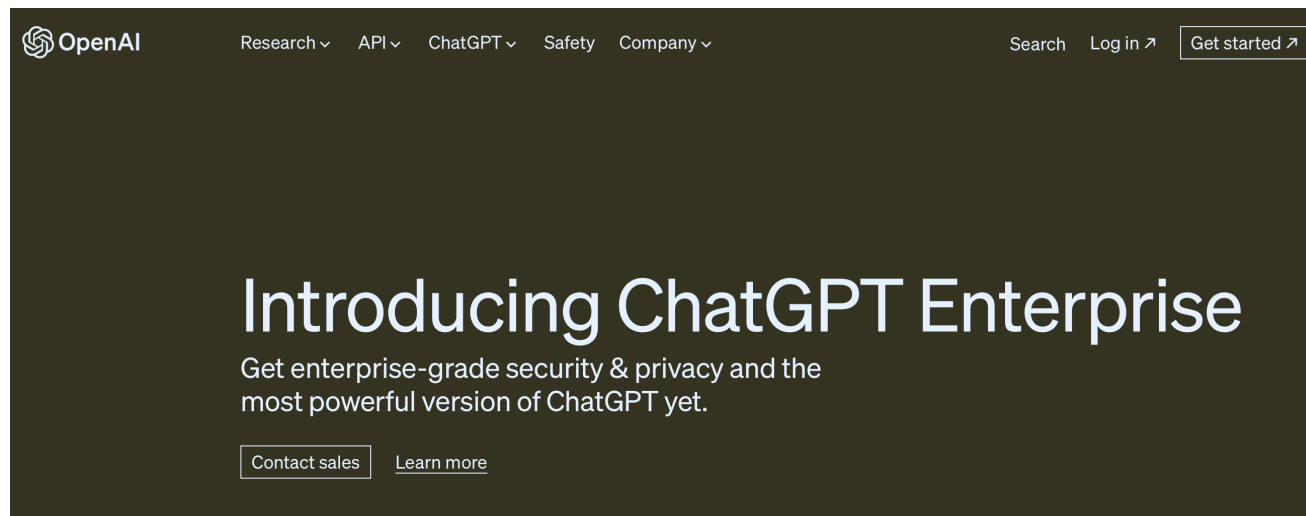


安全側に振っておくことが好ましい



ChatGPT Enterprise

2023年8月28日リリース



高速なGPT-4への無制限アクセス（Plusは50メッセージ/3時間）

4倍長い入力が可能

データがトレーニングに使われない

保存、転送時のデータ暗号化

組織で利用の管理を行う管理コンソールの提供やSSOをサポート



その前提で使ってみる

今回のデータはダミーなので、遊んでみましょう。
ぜひ、後ほど各自で試してみてください。
(もちろん自己責任ですが)

今回はエクセルデータをアップロードして分析するため、
データのアップロードを受け付ける、Advanced Data
Analysis (旧Code Interpreter)機能が必要になります。
そのため、ChatGPT Plus (有料版) を使います。
(月額20USD)



問1：欠席が3回以上連続する学生

このファイルは学生のある授業における出席データです。授業は全部で15回あります。

このデータから、欠席が3回以上連続する学生を抽出して、その男女別の人数と平均点数を求めてください。

ただし、出席と欠席を別の値に置き換えた上で一気に検索する方法でカウントせずに、初回の授業から順番に走査する方法でカウントしてください。



問 1 : 出力例



Data250_20230831.xlsx

このファイルは学生のある授業における出席データです。授業は全部で15回あります。このデータから、欠席が3回以上連続する学生を抽出して、その男女別の人数と平均点数を求めてください。ただし、出席と欠席を別の値に置き換えた上で一気に検索する方法でカウントせずに、初回の授業から順番に走査する方法でカウントしてください。



わかりました。まず、提供されたファイルを読み込んで内容を確認しましょう。

Finished working

Show work ▼

データの内容を確認しました。次に、以下の手順でデータを分析します。

1. 各学生について、出席データを順番に走査して、欠席が3回以上連続するかどうかを判断します。
2. 3回以上連続で欠席した学生を抽出します。
3. 男女別の人数と平均点数を求めます。

それでは、手順1から進めていきます。

Finished working

Show work ▼

3回以上連続で欠席した学生を抽出しました。次に、この抽出した学生の中で、男女別の人数と平均点数を求めます。

Finished working

Show work ▼

3回以上連続で欠席した学生の男女別の統計は以下の通りです：

- 男性：29人、平均点数は約40.03点
- 女性：36人、平均点数は約40.14点
- その他：3人、平均点数は約39.33点



問2：連続欠席が2回以上ある学生

元のデータから、欠席が2回以上連続することが2回以上ある学生を抽出してください。その時に、連続した欠席の間に出席が入る場合は、その出席を境にして新たな連続した欠席としてカウントできるように、工夫してください。ただし、初回の授業を欠席した時は、その前に欠席が入ることはないので注意してください。その男女別の学生の人数と平均点数を求めてください。



言いたいことはわかるが回りくどい



問2：出力例



元のデータから、欠席が2回以上連続することが2回以上ある学生を抽出してください。その時に、連続した欠席の間に出席が入る場合は、その出席を境にして新たな連続した欠席としてカウントできるように、工夫してください。ただし、初回の授業を欠席した時は、その前に欠席が入ることはないので注意してください。その男女別の学生の人数と平均点数を求めてください。



指示に従ってデータを分析します。

1. 各学生について、出席データを順番に走査して、欠席が2回以上連続する回数をカウントします。
2. 2回以上連続で欠席したことが2回以上ある学生を抽出します。
3. 男女別の人数と平均点数を求めます。

それでは、手順1から進めていきます。

Finished working Show work ▼

2回以上連続で欠席したことが2回以上ある学生を抽出しました。次に、この抽出した学生の中で、男女別の人数と平均点数を求めます。

Finished working Show work ▼

2回以上連続で欠席したことが2回以上ある学生の男女別の統計は以下の通りです：

- 男性：27人、平均点数は約44.22点
- 女性：37人、平均点数は約45.03点
- その他：2人、平均点数は約41.50点



実は

「**連続欠席が2回以上ある学生は何人ですか？**」と書きたいところですが、実はそれでは全く違う答えを返してきます。何度も試行錯誤して正しい答えを導き出す必要がありました。

その上で、できるだけ1回の質問で正しい答えを導き出す質問の仕方を検討したところ、このような質問になりました。

(実は問1のグレー部分は、問2のための仕込みでした)



ちょっと寄り道して

実際に試してみると…

1. このファイルは学生のある授業における出席データです。授業は全部で15回あります。このデータから、3回以上連続欠席のある学生の人数と平均点数を男女別で求めてください。
2. 連続欠席が複数回ある学生の人数と平均点数を男女別で求めてください。





問 1 : 出力例



了解しました。まず、提供していただいたファイルを読み込み、データの中身を確認します。



Finished working

Show work ▼

データを読み込むことができました。

次に、3回以上連続で欠席している学生を特定するための処理を行います。そのために、各学生の出席データを文字列として連結し、3回以上連続で「欠席」という文字列が含まれているかを確認します。その後、男女別に、該当する学生の人数と平均点数を計算します。

Finished working

Show work ▼

3回以上連続で欠席した学生の統計は以下のとおりです：

- 男性：29人、平均点数は約40.03点
- 女性：36人、平均点数は約40.14点
- その他：3人、平均点数は約39.33点

問題なし



問2：出力例



前の文章を引きずっている



了解しました。連続欠席が複数回ある学生を特定するために、3回以上連続で「欠席」という文字列が2回以上出現するかを確認します。その後、男女別に該当する学生の人数と平均点数を計算します。

Finished working

Show work ▼

連続欠席が複数回ある学生の統計は以下のとおりです：

- 男性：4人、平均点数は約31.75点
- 女性：3人、平均点数は約30.67点

おかしい

「その他」の 카테고리には該当する学生はいませんでした。

0

これらの学生のIDを教えてください。



Finished working

Show work ▼

連続欠席が複数回ある学生のIDは以下のとおりです：

- 749
- 1235
- 957
- 457
- 214
- 146
- 1173



試行錯誤が始まる



分解して説明する

連続欠席が複数回ある学生



元のデータから、欠席が2回以上連続することが2回以上ある学生を抽出してください。その時に、**連続した欠席の間に出席が入る場合は、その出席を境にして新たな連続した欠席としてカウント**できるように、工夫してください。ただし、**初回の授業を欠席した時は、その前に欠席が入ることはない**ので注意してください。その男女別の学生の人数と平均点数を求めてください。

→状況を分解して論理的に説明する必要



実際に試してみると…

やはり人間の理解とは異なる
出てきた答えを鵜呑みにしてはいけない



融通が利かない、ロジカルな、思い込みの激しい、面倒くさいヤツ
でも、逆ギレせずについてくれる忍耐強いヤツ

性格（理解の癖）を理解した上で使うととても便利！



元に戻って



問3：成績との関係は？

欠席と成績の間に関係性があるか調べてください。
その上で、結果を可視化してください。



欠席と成績の関係性を調べるために、以下の手順を実行します：



1. 各学生の総欠席回数を計算します。
2. 総欠席回数と成績の相関を計算します。
3. 総欠席回数と成績の散布図を描画します。

まず、各学生の総欠席回数を計算し、その後相関を計算します。

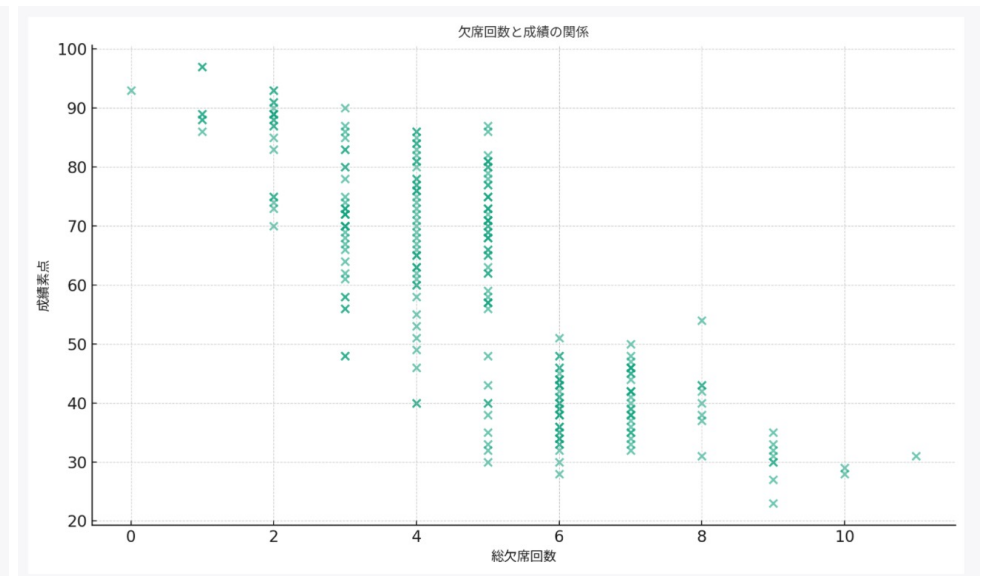
Finished working Show work ▾

欠席回数と成績素点の相関係数は約 -0.800 です。これは、欠席回数が増えると成績が低下することを示しています。この値は非常に強い負の相関を示しています。

次に、この関係性を散布図で可視化します。

Finished working Show work ▾




欠席回数と成績の関係を示す散布図を描画しました。このグラフから、欠席回数が増えるにつれて成績が低下する傾向が明確に見られます。相関係数の値もこれを裏付けています。



問4：性別との関係は？




さらに性別との関係性についても調べて、その結果を可視化してください。

了解しました。欠席、成績、および性別の関係性を調査するために、以下の手順を実行します：   

- 性別ごとの欠席回数の平均を計算します。
- 性別ごとの成績の平均を計算します。
- 欠席回数と成績の関係を性別ごとに散布図で可視化します。
- 棒グラフを用いて、性別ごとの欠席回数と成績の平均を比較します。

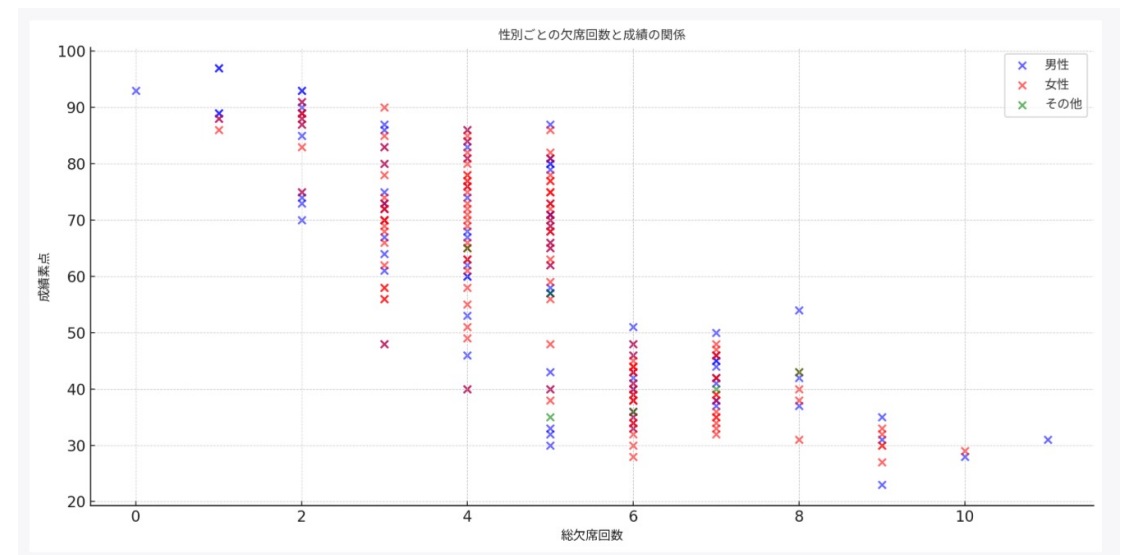
まず、性別ごとの欠席回数と成績の平均を計算します。

Finished working Show work 

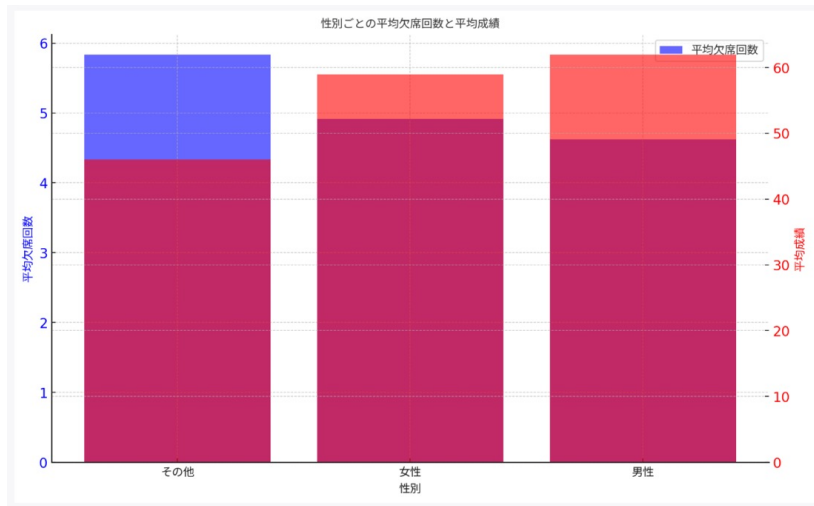
性別ごとの欠席回数と成績の平均は以下の通りです：

- その他
 - 平均欠席回数: 約5.83回
 - 平均成績: 約46.00点
- 女性
 - 平均欠席回数: 約4.92回
 - 平均成績: 約58.95点
- 男性
 - 平均欠席回数: 約4.62回
 - 平均成績: 約61.94点

次に、欠席回数と成績の関係を性別ごとに散布図で可視化し、その後性別ごとの欠席回数と成績の平均を棒グラフで比較します。



問4：性別との関係は？



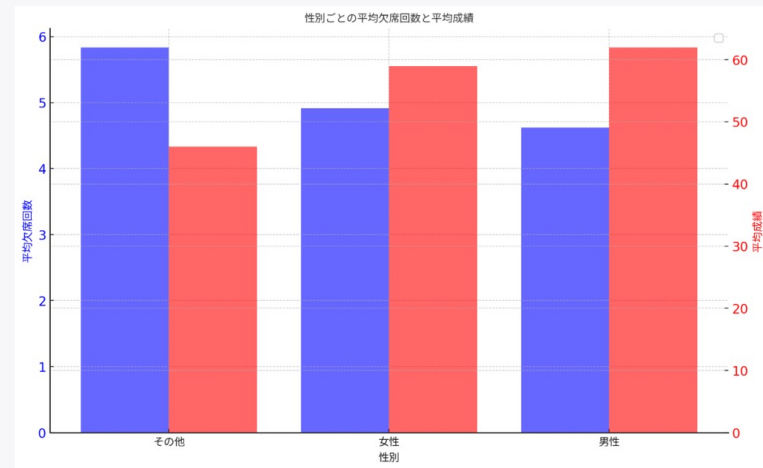
0 このグラフは適切ですか？



申し訳ございません。再度確認いたしましたところ、性別ごとの平均欠席回数と平均成績を示す棒グラフは、X軸のラベル（性別）が正しく配置されておらず、データの表示も重複しています。

このグラフは適切ではありません。正確なグラフを描画しなおします。再度お手順をおかけて申し訳ございません。

Finished working Show work



性別ごとの平均欠席回数と平均成績を示す棒グラフを修正して描画しました。



問5：まとめてもらう



これらの結果を上司に報告する必要があります。
誤解を与えない適切な文章を作成してください。

報告書：学生の出席状況と成績の関連性について

1. 背景:

- 分析の対象となったのは、ある授業の学生の出席データおよび成績データです。
- 授業は計15回行われました。

2. 欠席と成績の関連性:

- 全体の学生を対象に、欠席回数と成績の相関係数を計算した結果、約-0.800という非常に強い負の相関が確認されました。これは、欠席回数が増えるにつれて成績が低下することを示唆しています。
- 散布図でも、欠席回数が増加するにつれて成績が低くなる傾向が確認されました。

3. 性別ごとの欠席と成績の関連性:

- 各性別の平均欠席回数と平均成績は以下の通りです：
 - 男性：平均欠席約4.62回、平均成績約61.94点
 - 女性：平均欠席約4.92回、平均成績約58.95点
 - その他：平均欠席約5.83回、平均成績約46.00点
- 性別ごとの散布図を確認した結果、すべての性別で欠席回数が増加するにつれて成績が低下する傾向が見られました。

4. 結論:

- 学生の欠席回数は成績に強く影響していることが確認されました。特に、欠席が増加すると成績が低下する傾向が明確です。
- 性別による違いも確認されましたが、どの性別においても欠席が成績に負の影響を及ぼすことが明らかとなりました。



生成系AIの特性とまとめ

- モデルが変われば回答が変わる。
- 回答を都度“生成”するので、**回答が毎回異なる**可能性がある。
 - 再現するためにはコードを残しておくことが必要かも
 - 実データを使うときはコードを残してローカル環境で実行する必要もある
- 指示を出す側の理解と**異なる理解をする**可能性がある。
 - こちらの意図を正確に理解しているか確認が必要
 - 理解させるために分解して論理的に説明する
- **もっともらしく誤った回答を作成**する。
 - 鵜呑みにしない


あくまでも現時点でのまとめ

将来のバージョンでは、この多くが改善されるかも？



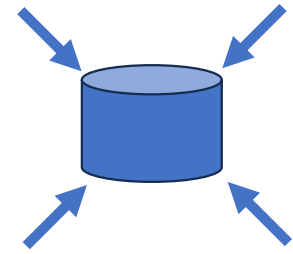
IR担当者の分析に必要な能力とは？

探索的データ分析 (EDA)

学内外の動向  データ



(現状は)
分析するための一定の技術力は必要



欠席率？
連続欠席？
前半or後半の欠席？



データを多面的に見る力

Research Question

