

Gemini による動画検索革命と バンダイナムコ「ClipSearch」事 例紹介

Google
Cloud
Next

Tokyo

Proprietary



吉田 望

Google Cloud, Customer Engineer



玉置 絢

株式会社バンダイナムコエンターテインメント
CE事業部 第2プロダクション
スペシャリスト／プロデューサー



アジェンダ

01

Gemini による動画検索革命

Gemini は動画が得意

シーン検索とは？(デモ)

シーン検索の仕組みと実現技術

02

「ClipSearch」事例紹介

ClipSearch(デモ)

なぜ動画検索なのか？

AI 化モチベーションの見つけ方

Gemini / TAP で得た知見

Gemini は動画が得意

Gemini Model

Google の基盤モデル

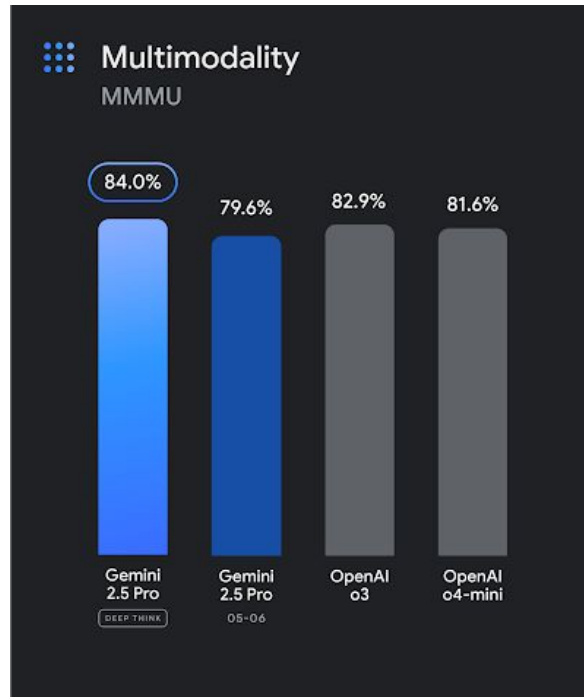
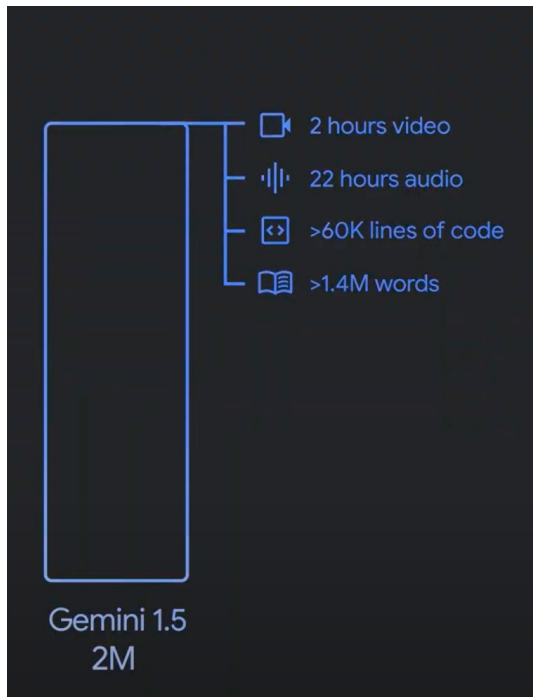


Gemini

長いプロンプトに対する
マルチモーダルな推論

Gemini Model

Google の基盤モデル



動画に含まれる情報の種類

動画全体の
ストーリー

音声

テロップ以外の
テキスト

文字以外の
映像

テロップ
(字幕)

タイムコード



前半のセッション

「動画中の特定のシーンを検索する」
という具体的なユースケースを見ながら、
Gemini ✕ 動画 の活用例をご紹介します。



シーン検索とは(デモ)

こんなシーンを見つけない！

「アクション映画のカーチェイスで赤い車が爆発するシーン」

「先生がホワイトボードに重要な数式を書いている場面」

「先週の会議で山田さんが来季の予算について話している場面」



目的のシーンはすぐには見つからず...

長い動画ではシークバーとの果てしない戦い

大量の動画があるので埋もれる(サムネイル識別では限界)

該当シーンが複数あり、探しきれない or 見逃す



出発点

「アクション映画のカーチェイスで赤い車が爆発するシーン」

「先生がホワイトボードに重要な数式を書いている場面」

「先週の会議で山田さんが来季の予算について話している場面」



上記をそのまま検索したら
その該当シーンがヒットする
ような仕組みを作れるか？

(デモ) Gemini で実現するシーン検索システム

- Gemini を活用した動画シーン検索アプリ

シーン検索の例→

The screenshot shows a web browser window displaying a web application titled "Scene Search". At the top, there is a "Search Option" dropdown menu currently set to "Scene Search". Below this, the main heading "Scene Search" is displayed. Under the heading, there is a "Search Query" input field with a red border and a cursor. To the right of the input field is a "Top N" control with a question mark icon and the value "1". Below the input field is a "Search" button. A tooltip is visible over the "Search" button, containing the text "AI/ML/LLM/シーン検索" and "ニューラルネットワークを用いたシーン検索". The browser's address bar shows the URL "frontend-608635780090.asia-northeast1.run.app".

(デモ)シーン検索対象の動画サンプル素材

ストーリーが明確な
短い動画



千葉県印西市に
データセンターを開設

複数のシーンに別れた
会話が多めの動画



Google Cloud Next Tokyo '24
- Innovators Hive の Day 0 に潜入！ -

スライドが常時表示
された講演動画



大規模言語モデルを支える
分散学習インフラ Pathways

(デモ) Gemini で実現するシーン検索システム



シーン検索の仕組みと実現技術

シーン検索の仕組みを理解するポイント

1. シーン検索の処理の流れをアーキテクチャを通して理解する

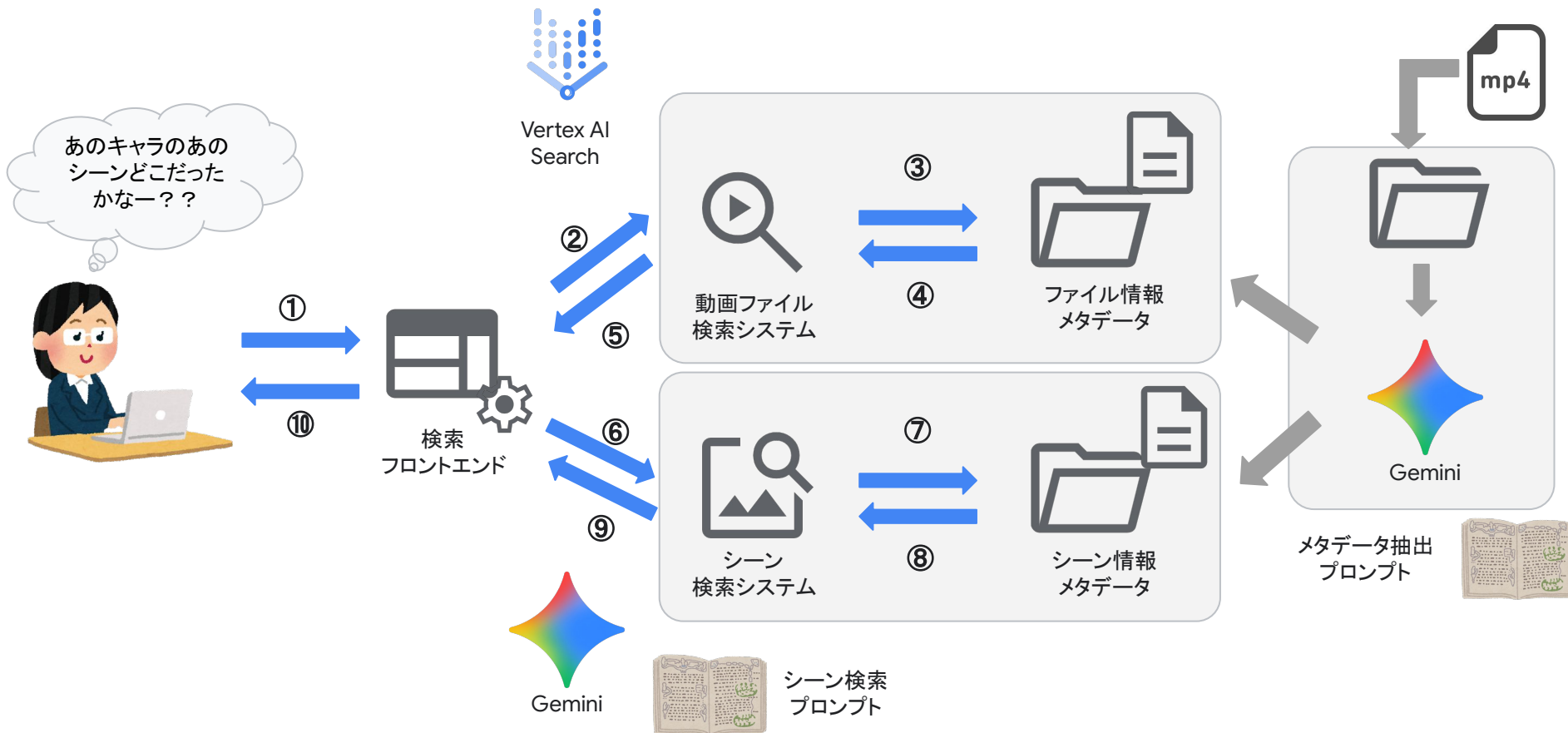
- ✦ a. Gemini で動画から事前抽出したメタデータ テキスト
(=動画に関連する情報を言語化したテキスト ファイル)を使用し、
- b. ファイル単位の検索を (Google Cloud のベクトル検索サービスを用いて) 行った後に、
- ✦ c. Gemini が「タイムスタンプ情報を含むメタデータ テキスト」を解釈することで、ユーザーが知りたいシーン特定する

2. シーン検索に使用する メタデータ テキストの必要性 を理解する

- a. Gemini は直接動画を処理できるにも関わらず、なぜメタデータ テキストの抽出が必要なのか

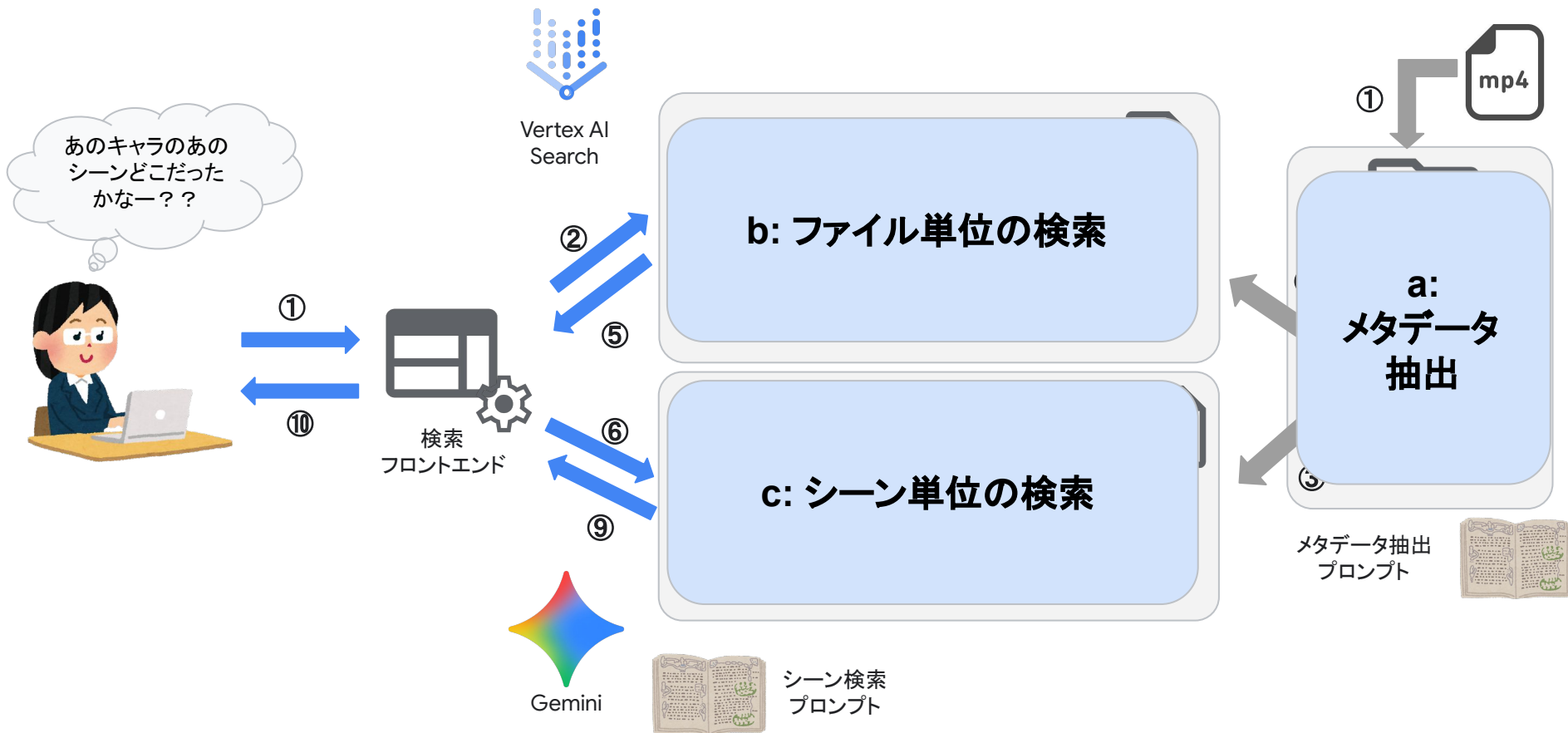
アーキテクチャ

Proprietary



アーキテクチャ(ざっくり)

Proprietary



a: メタデータテキスト抽出

Proprietary



実際のメタデータ テキストの例

[summary]

```
{'Summary': '2023年、Google は千葉県印西市に日本最大規模のハイパースケール
データセンターを開設しました。\\n'
  'データセンター建設の責任者である金井匡彦氏によると、このデータセン
ターは Google '
  'データセンターのグローバル設計基準と日本の技術・文化を融合させて設
計されており、環境にも配慮しています。\\n'
  'このデータセンターにより、日本の企業は Google Cloud を含む
Google '
  '製品をより快適かつ安全に利用できるようになります。\\n'
...}
```

[important scenes]

```
Timestamp, Description
00:00-00:07, Google データセンター建設の責任者である金井匡彦氏が、 2023年に千
葉県印西市にデータセンターを開設したことを発表する重要なシーンです。
00:07-00:25, 金井匡彦氏が、データセンターの特徴として、 Google データセンターのグ
ローバル設計基準と日本の技術・文化を融合させて設計されていること、環境にも配慮して
いること、日本最大規模のハイパー スケールデータセンターであることを説明する重要な
シーンです。
...
```

[visual info]

```
Timestamp, Description
00:00-00:01, 金井 匡彦さんがGoogleの看板の前で話しています。金井 匡彦さんは白
いTシャツを着ており、背景には Googleのデータセンターが見えます。晴天で、空は青く澄
み渡っています。
00:01-00:03, データセンターの内部の映像です。サーバーがずらりと並んでおり、通路
の奥には光が見えます。通路の床は白く、天井には蛍光灯が設置されています。
...
```

[transcription]

```
[{'Name': '金井 匡彦',
  'Transcription': '2023年 Googleとして データセンターを ここ 千葉県印西市に
開設しました',
  'timestamp': '00:00-00:07'},
{'Name': '金井 匡彦',
  'Transcription': 'このデータセンターの特徴は 何と言ってもGoogleデータセンター
のグローバル設計基準と 日本技術 '
  '文化の融合にて設計されていること そして日本で最大規模のハ
イパースケールデータセンターであることです '
  '当然環境にも配慮した設計になっています このデータセンター
によって日本の企業の皆様が Google '
  'Cloudを含むGoogle製品を さらに快適に そして安全にお使い
いただけるようになります',
  'timestamp': '00:07-00:25'},
...]
```

[texts]

```
[{'Layout': '画面左側に地図が表示され、地図の中にテキストが表示されている ',
  'Text': ['千葉県 印西市'],
  'Timestamp': '00:06-00:08'},
{'Layout': '画面左側に縦書きで二行のテキストが表示されている ',
  'Text': ['金井 匡彦', 'データセンター建設', 'ジャパンリード'],
  'Timestamp': '00:08-00:10'},
...]
```

実際のメタデータ テキストの例

[summary]

{'Summary': '2023年、Google は千葉県印西市に日本最大規模のハイパースケールデータセンターを開設しました。\\n' 'データセンター建設は、金井 匡彦氏によると、このデータセンターは Google ' 'データセンターの設計は、最新の技術・文化を融合させて設計されており、環境にも配慮している。\\n' 'このデータセンターは Google Cloud を含む Google ' '製品をより快適かつ安全に利用できるようになります。 \\n' ...

要約

[important scenes]

Timestamp, Description
00:00-00:07, Google が、2023年に千葉県印西市にデータセンターを開設しました。\\n
00:07-00:25, 金井 匡彦氏が、データセンターのグローバル設計基準と日本最大規模のデータセンターであることを説明する重要なシーンです。\\n
...

重要シーン (タイムスタンプ付き)

[visual info]

Timestamp, Description
00:00-00:01, 金井 匡彦さんは白いTシャツを着ており、晴れた空の下で歩いています。\\n
00:01-00:03, データセンターの奥には光が見えます。\\n
...

視覚情報 (タイムスタンプ付き)

[transcription]

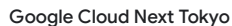
[{'Name': '金井 匡彦', 'Transcription': '2023年、Google は千葉県印西市にデータセンターを開設しました。', 'timestamp': '00:00-00:07'}, {'Name': '金井 匡彦', 'Transcription': 'データセンターの設計は、最新の技術・文化を融合させて設計されており、環境にも配慮している。', 'timestamp': '00:07-00:25'}] \\n
このデータセンターは Google Cloud を含む Google 製品をより快適かつ安全に利用できるようになります。 \\n
...

文字起こし (タイムスタンプ付き)

[texts]

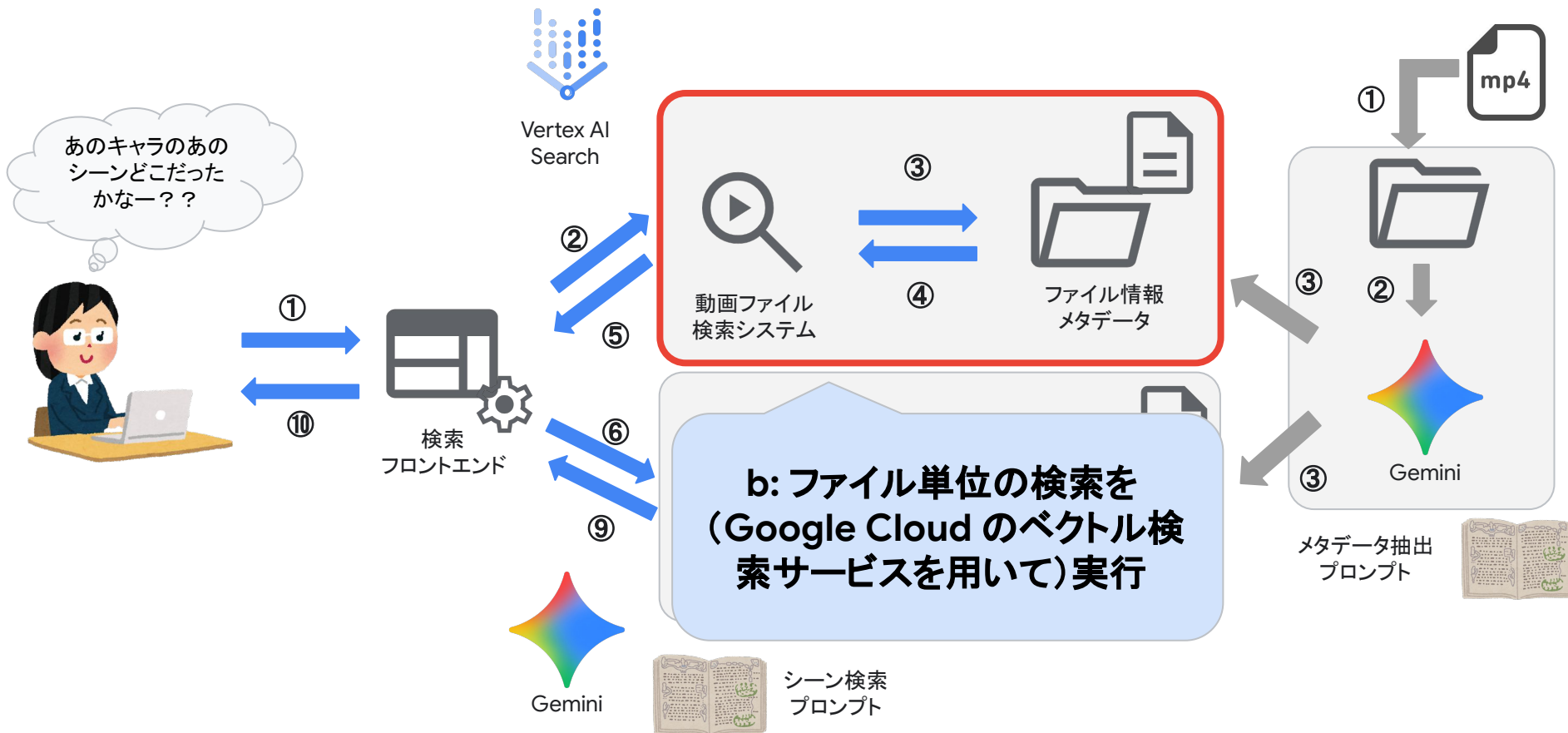
[{'Layout': '画面左側', 'Text': '千葉県のデータセンター建設は、金井 匡彦氏によると、このデータセンターは Google Cloud を含む Google 製品をより快適かつ安全に利用できるようになります。', 'Timestamp': '00:00-00:07'}, {'Layout': '画面左側', 'Text': '金井 匡彦氏は白いTシャツを着ており、晴れた空の下で歩いています。', 'Timestamp': '00:01-00:03'}] \\n
...

テキスト情報 (タイムスタンプ付き)



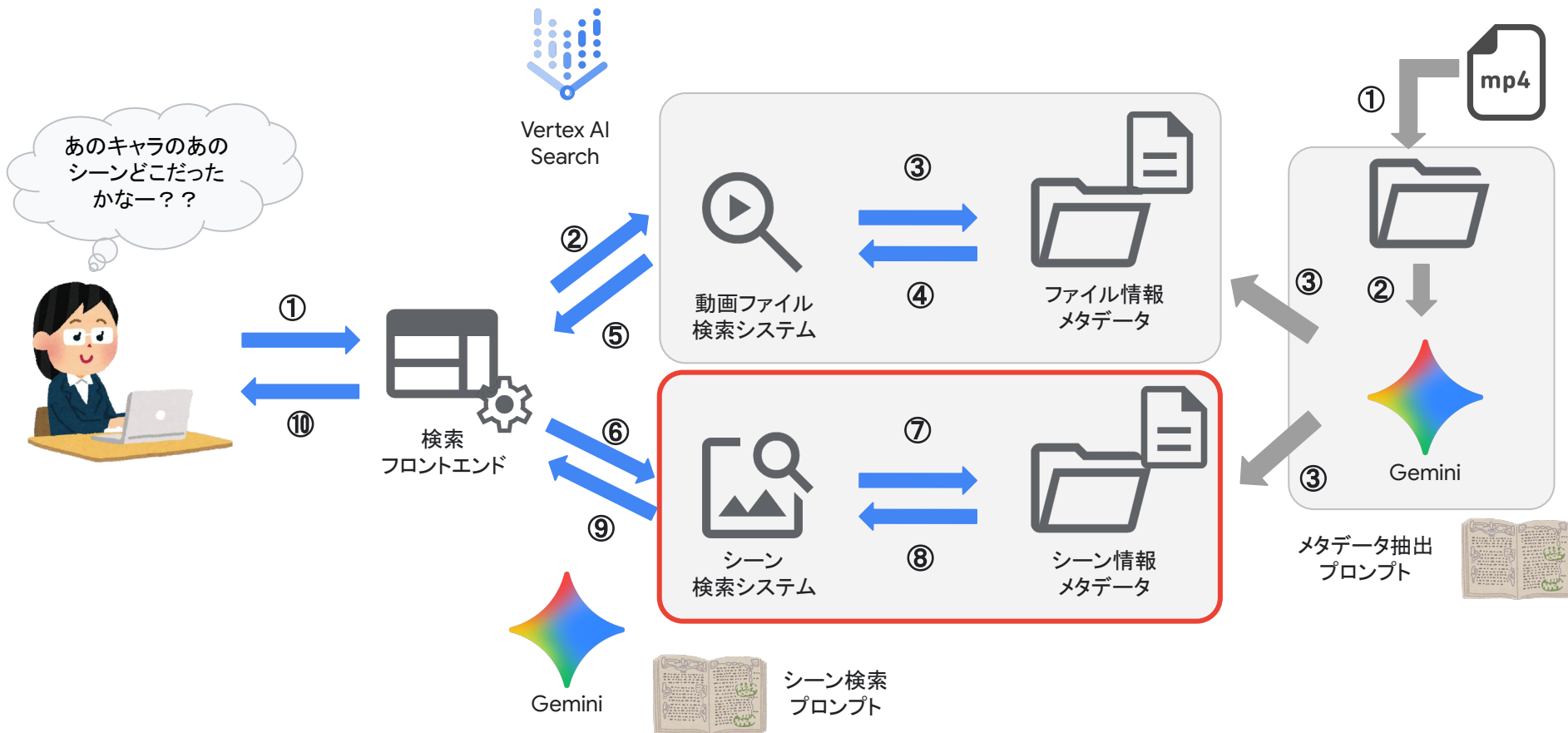
b: ファイル単位の検索

Proprietary



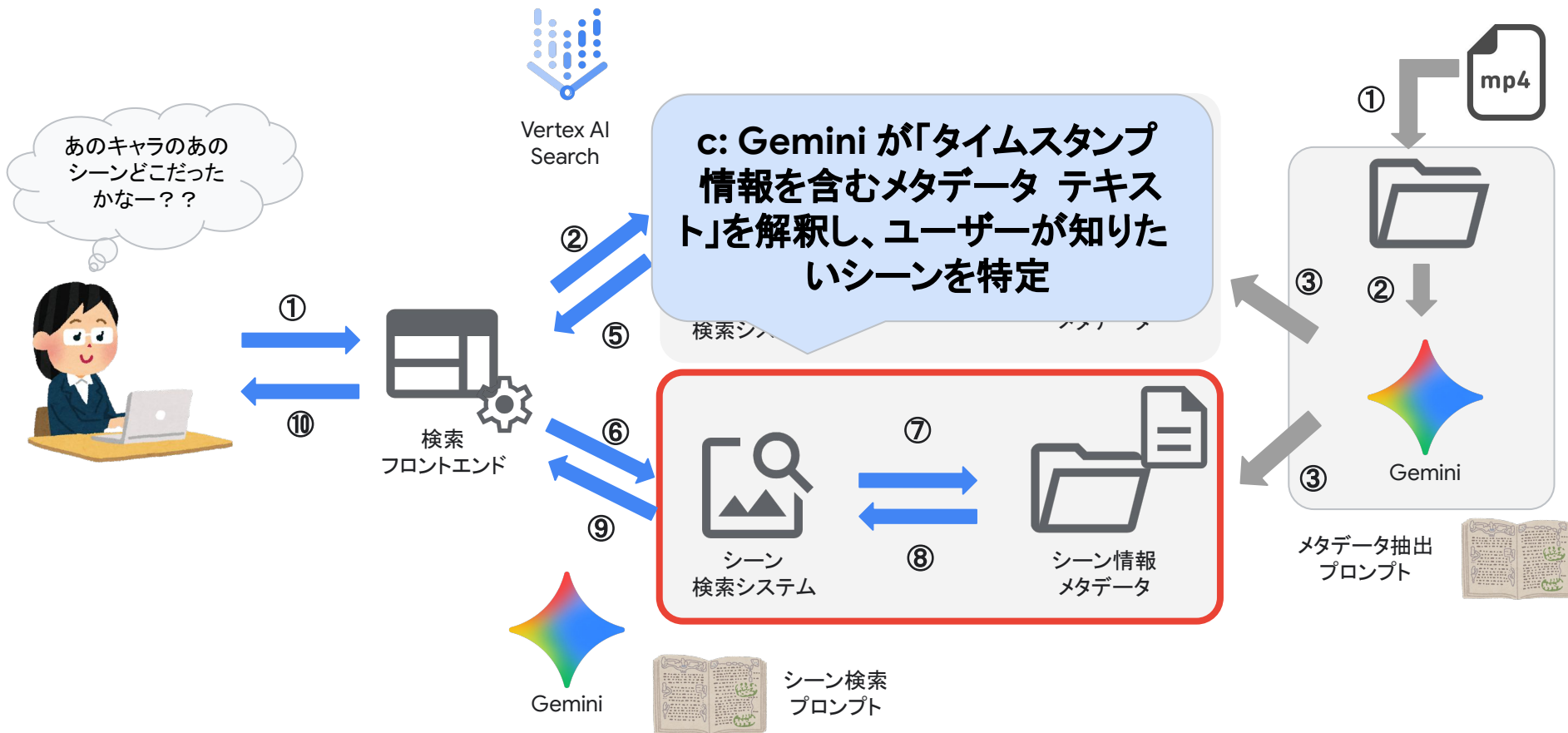
c: シーン単位の検索

Proprietary



c: シーン単位の検索

Proprietary



Gemini は何を検索しているのか？

- メタデータ テキストをプロンプトに埋め込んで、質問にマッチするシーンのタイムスタンプを Gemini で出力します。

```
prompt_content_search = '''  
You are a video content editor.
```

Given the following information of a movie:

- The [summary] section contains the summary of the movie.
- The [important scenes] section contains the important scenes of the movie with timestamps.
- The [visual info] section contains the visual information on what's happening in each scene with timestamps.
- The [transcription] section contains speech transcription with timestamps.
- The [text] section contains text information with timestamps.

Find one to three scenes that match the user query with timestamps.

[format instruction]

Output in Japanese. Output is a JSON list with "scene dict".

Each "scene dict" is a JSON dict with the following format:

```
{{  
  "Timestamp": "<timestamp mm:ss-mm:ss>",  
  "Description": "<Explain how this scene matches the query.>",  
  "Evidence": [<List of data snippets that support your result>]  
}}
```

この部分に質問と
メタデータテキストを
まとめて埋め込む

```
[user query]  
{}  
  
[summary]  
{}  
  
[important scenes]  
{}  
  
[visual info]  
{}  
  
[transcription]  
{}  
  
[texts]  
{  
  ...  
}
```

Gemini は何を検索しているのか？

- メタデータ テキストをプロンプトに埋め込んで、質問にマッチするシーンのタイムスタンプを Gemini で出力します。

```
prompt_content_search = '''
You are a video content editor.
```

Given the following information of a movie:

- The [summary] section contains the summary of the
- The [important scenes] section contains the important
- The [visual info] section contains the visual information
- The [transcription] section contains speech transcripts
- The [text] section contains text information with

Find one to three scenes that match the user query with timestamps.

[format instruction]

Output in Japanese. Output is a JSON list with "scene dict".

Each "scene dict" is a JSON dict with the following format:

```
{{
  "Timestamp": "<timestamp mm:ss-mm:ss>",
  "Description": "<Explain how this scene matches the query>",
  "Evidence": [<List of data snippets that support your answer>]
}}
```

動画編集者として、ユーザーの探しているシーンを、添付する動画関連情報(メタデータ テキスト)を参考に探してね

シーンを見つけたら、メタデータを参考にタイムスタンプ情報を出力してね

ユーザーの探したいシーン

```
[user query]
{}

[summary]
{}

[important scenes]
{}

[visual info]
{}

[transcription]
{}

[texts]
{}
...
```

動画関連情報
(メタデータ テキスト)

実際のメタデータ テキストの例

[summary]

```
{'Summary': '2023年、Google は千葉県印西市に日本最大規模のハイパースケール
データセンターを開設しました。\\n'
  'データセンター建設の責任者である金井匡彦氏によると、このデータセン
ターは Google '
  'データセンターのグローバル設計基準と日本の技術・文化を融合させて設
計されており、環境にも配慮しています。\\n'
  'このデータセンターにより、日本の企業は Google Cloud を含む
Google '
  '製品をより快適かつ安全に利用できるようになります。 \\n'
  ...}
```

[important scenes]

Timestamp, Description

```
00:00-00:07:00 金井 匡彦氏が、2023年に千
葉県印西市に Google データセンターのグ
00:07-00:10:00 ローバル設計基準と、環境にも配慮して
いること、日本企業と Google Cloud を含む Google
シーンを説明する重要なシーンです
...
```

Gemini は、タイムスタンプ付の
メタデータから該当シーンの
タイムスタンプを発見している

[visual info]

Timestamp, Description

```
00:00-00:00:01 金井 匡彦さんが Google の看板の前で話しています。金井 匡彦さんは白
いTシャツを着ており、背景には Google のデータセンターが見えます。晴天で、空は青く澄
み渡っています。
00:01-00:00:03 データセンターの内部の映像です。サーバーがずらりと並んでおり、通路
の奥には光が見えます。通路の床は白く、天井には蛍光灯が設置されています。
...
```

[transcription]

```
[{'Name': '金井 匡彦',
  'Transcription': '2023年 Googleとして データセンターを ここ 千葉県印西市に
開設しました',
  'timestamp': '00:00-00:07'},
{'Name': '金井 匡彦',
  'Transcription': 'このデータセンターの特徴は 何と言ってもGoogleデータセンター
のグローバル設計基準と 日本技術 '
  '文化の融合にて設計されていること そして日本で最大規模のハイ
パースケールデータセンターであることです '
  '当然環境にも配慮した設計になっています このデータセンター
によって日本の企業の皆様が Google '
  'Cloudを含むGoogle製品を さらに快適に そして安全にお使い
いただけるようになります',
  'timestamp': '00:07-00:25'},
...]
```

[texts]

```
[{'Layout': '画面左側に地図が表示され、地図の中にテキストが表示されている ',
  'Text': ['千葉県 印西市'],
  'Timestamp': '00:06-00:08'},
{'Layout': '画面左側に縦書きで二行のテキストが表示されている ',
  'Text': ['金井 匡彦', 'データセンター建設', 'ジャパンリード'],
  'Timestamp': '00:08-00:10'},
...]
```

メタデータ テキストの必要性

- Gemini は直接動画処理できるにも関わらず、なぜメタデータテキストの抽出が必要？



- Gemini に直接動画を読み込ませてもシーン検索は実現可能ですが、多くの実世界ユースケースでは、拡張性や速度の観点でテキスト化(=メタデータテキスト作成)を挟むことをおすすめします。

Tech Acceleration Program (TAP)



生成 AI 関連技術を活用 して具体的な **ビジネス成果の獲得**を目標としたお客様専用の活用シナリオの策定 を Google Cloud のエンジニア がワークショップを通じて **アイディエーション** から **ソリューション** アーキテクチャの考案 、**取り組みのマイルストーン設定** などを支援。

ご支援内容

アイディエーション ワークショップ

- 生成 AI ソリューションを具体的なビジネス成果に紐づける活用シナリオの具体化をご支援

プロトタイピングサポート

- 開発予定アプリケーションのアーキテクチャ設計からプロトタイプ開発のご支援



メリット

Google Cloud の専門チームが伴走して活用シナリオの策定からプロトタイピングまで支援をすることで、生成 AI の効果的な導入ポイントを選定し、早期にプロトタイプ プロジェクトを立ち上げ生成 AI 活用のナレッジ蓄積と早期の成果獲得をご支援



- テーマ選定、内容のディスカッションをさせていただき、その上で、内容、スケジュールを別途提案
- 弊社プリセールスエンジニアによる無償でのご支援を想定





バンダイナムコホールディングス

デジタル事業
例：ゲーム

バンダイナムコエンターテインメント

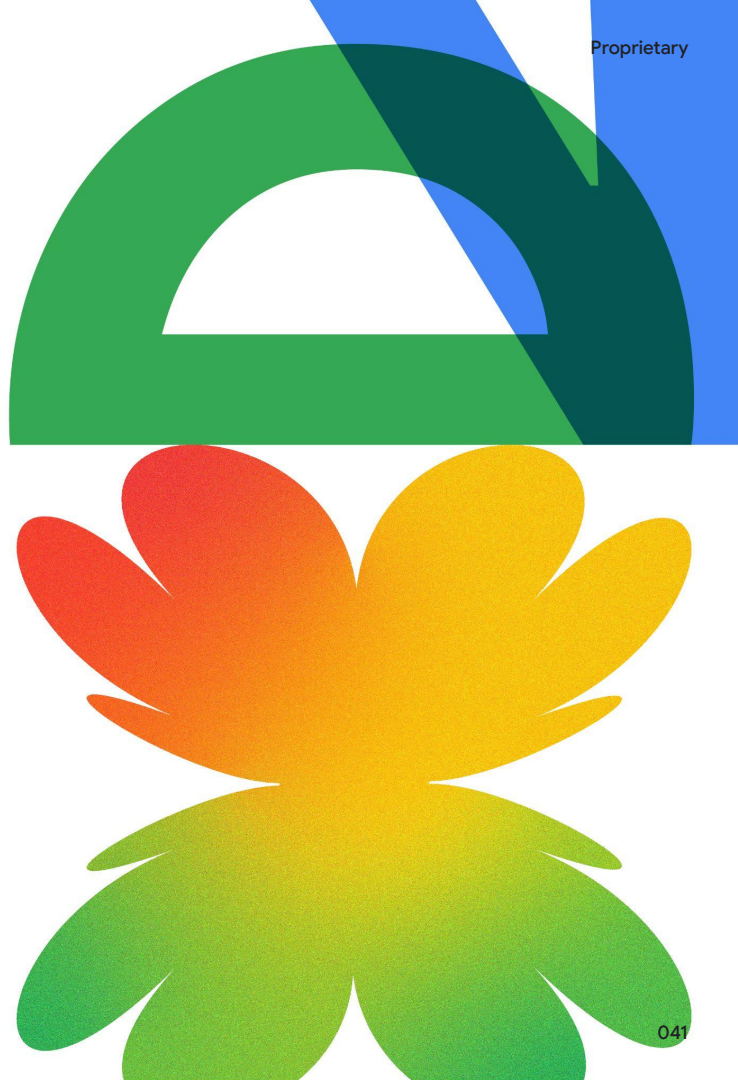
ClipSearch って こんな感じ

動画あり
アーカイブ動画をご視聴ください

Q.
なぜ動画検索なのか？

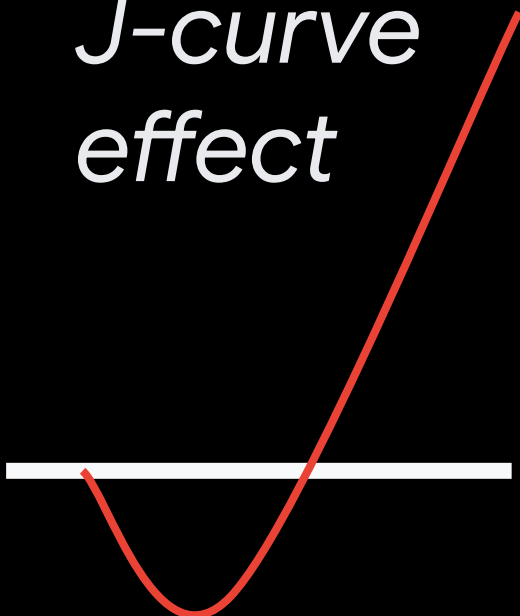


分かりやすく
「生成 AI って本当に
便利なんだ！」
を伝えるフェーズだから



AI 利用が 進まないケース

*J-curve
effect*



不定形の業務

多岐にわたる業務を一気にカバーするAIは開発も習熟もコストが大きすぎる。まずは LLM チャットボットで十分

ニッチすぎる

多部署・多職種を横断していない、特定の部署で年に数回しか必要としない業務は「喉元過ぎれば熱さ忘れる」

苦痛度が薄い

AI で代替したいというモチベーションが生まれない分野は実務利用が進まない。苦痛が大きい作業が優先

*J-curve
effect*

コールドスタートを
モチベーションで
乗り越えられる
ことが大事

“

ユーザーのモチベーション

∨

理論上のメリット見積もり

意味のない KPI を立てない

- “社内アカウント登録率 ○○ % !”
“利用したことがある社員が ○ 割 !”

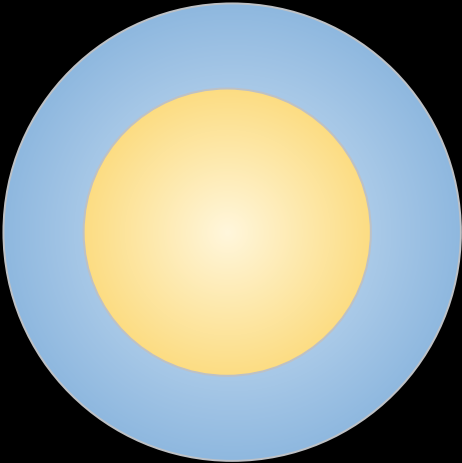
↑ 1 ~ 2 回触ってモチベを失って離脱

- 計算上は ○○○ 時間の圧縮だが、
そうならないとしたら
使わない社員の意識が課題だ !

↑ 使われないものを作ったほうが課題

モチベーションの発見

実務者の モチベ探索



1

取捨選択

「自分自身で行いたい、そこをAIに任せるとモチベが下がる」と周延的に仕方なく行っている作業を分離する

実務者の モチベ探索

1

取捨選択

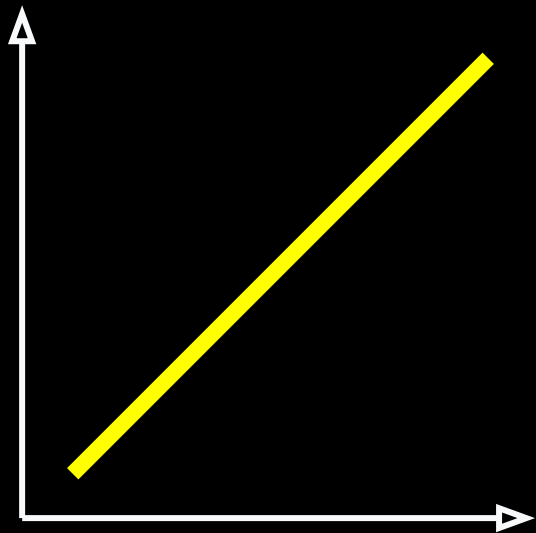
「自分自身で行いたい、そこをAIに任せるとモチベが下がる」と周道的に仕方なく行っている作業を分離する

2

スキルと パフォーマンス

自分の努力や訓練したスキルを発揮できる場所では、自分でやるモチベもパフォーマンスも高い(AIは不要)

発揮できる
パフォーマンス



努力・訓練の蓄積
キャリア・職階

実務者のモチベ探索

1

取捨選択

「自分自身で行いたい、そこをAIに任せるとモチベが下がる」と周回的に仕方なく行っている作業を分離する

2

スキルとパフォーマンス

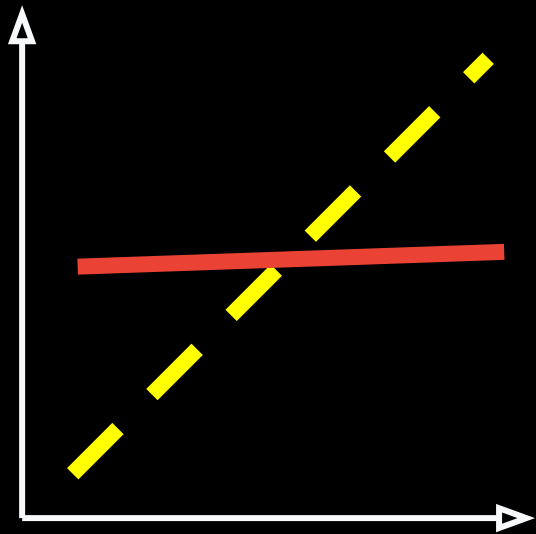
自分の努力や訓練したスキルを発揮できる場所では、自分でやるモチベもパフォーマンスも高い(AIは不要)

3

AI活用点が見えてくる

その人の努力や訓練の度合い、ジュニア-シニアの差と効率が関係しないタスクは苦痛・苦行 = **AI化モチベ高**

発揮できる
パフォーマンス

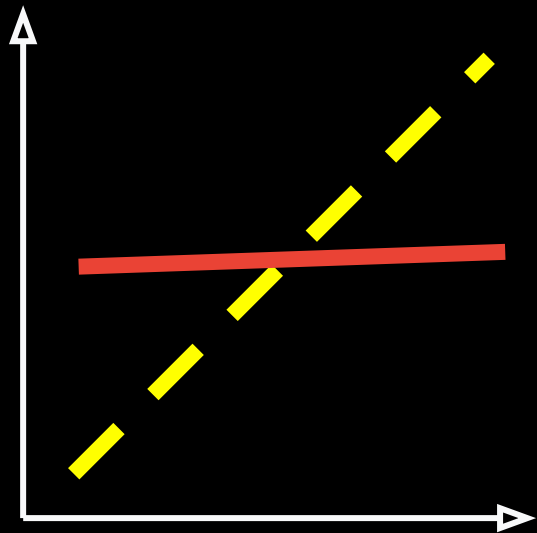


努力・訓練の蓄積
キャリア・職階

“

成長の実感が持てない仕事は
苦痛な作業である
＝ AI 導入チャンス！

発揮できる
パフォーマンス

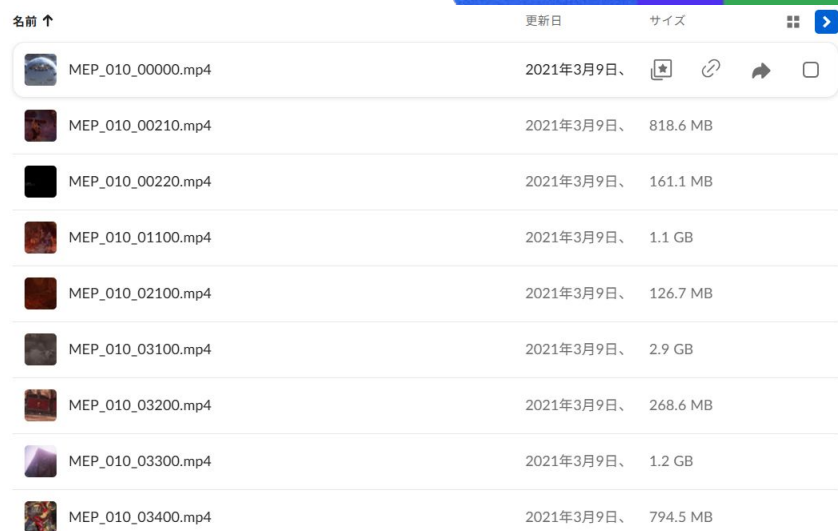


努力・訓練の蓄積
キャリア・職階

この典型例
としての
「欲しい場면을
大量の動画群から探す」
タスクの苦痛さ

ゲームの裏には 大量の録画ファイルがある

- 家庭用ゲームのプレイ時間は
数十時間あるものが多い
- ゲームを実際に走らせる環境以外
でもゲーム内容が確認できるよう
録画ファイルが大量に作られるよう
になってきている
- + 大量の宣伝映像 × 言語差分！

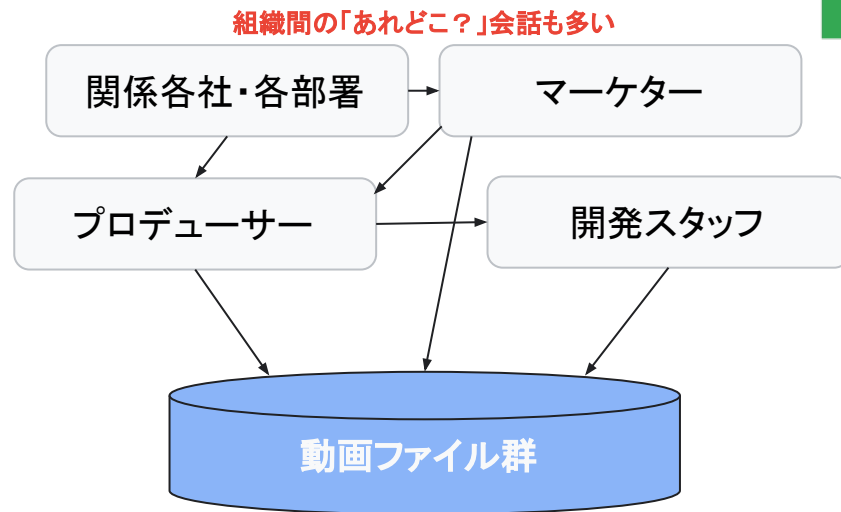


名前 ↑	更新日	サイズ
 MEP_010_00000.mp4	2021年3月9日、	   
 MEP_010_00210.mp4	2021年3月9日、	818.6 MB
 MEP_010_00220.mp4	2021年3月9日、	161.1 MB
 MEP_010_01100.mp4	2021年3月9日、	1.1 GB
 MEP_010_02100.mp4	2021年3月9日、	126.7 MB
 MEP_010_03100.mp4	2021年3月9日、	2.9 GB
 MEP_010_03200.mp4	2021年3月9日、	268.6 MB
 MEP_010_03300.mp4	2021年3月9日、	1.2 GB
 MEP_010_03400.mp4	2021年3月9日、	794.5 MB

重い = DL 長い、シーク遅い

ユースケース例

1. 新作 PV にゲーム中のあの名シーンを使いたい！でもどのファイルの何時何分だったけ？
2. 今日はあのキャラの誕生日。イケてるシーンを SNS に流したい。何かない？
3. このセリフを翻訳したいんだけど実際のシーンの映像が欲しい。
4. 過去に作ったゲームのあの画面みたいな UI にしてほしいんだけど、その UI 画面って画像残ってたっけ？



Q.
なぜ動画検索なのか？

A.
人間はどうやっても倍速視聴が限度
という肉体的限界が着目点 &
ニーズを持つ組織が多いから

検索ならやっぱり...
Google さんでしょ！



VLM (Vision Language Model)
としての Gemini の動画把握能力

Vertex AI Search で高速 PoC

TAP が決め手に

半日合宿的スタイルの TAP を
Google Cloud 様の TAP チーム主導で
実施していただき、全 3 社 4 部署にまたがる
チームビルディングの初動がしっかり完了。

→ TAP を起点にして知見を貯めていったことが
ClipSearch PoC の直接的成功要因

ClipSearch の知見

検索体験のデザイン

「見落としがない」
と「誤検知がない」
どっちが大事？

「検索速度」と「信頼性」
どっちが大事？

といった問いを通して
実務のリアルに迫るのが大事

隠れた前提

宣伝のユースケースであれば、
「あるキャラが映っている画」

とは、そのキャラの顔が一定量以上画面を占めていることを意味する。

といった「当たり前過ぎて言語化されない前提」を掘っていく

薄く広い一致は無意味

ClipSearch でユーザー（社員）が求めている欲しいシーンは

一瞬で、動画全体の要素がまんべんなくクエリと近いといった要素は無意味。

一瞬だけでも強いクエリと合致しているほうがよほど大事。

結び: AI を面白がろう！

安易な「導入ごっこ」に終わらせず、利用モチベを重視しよう。
GenAI-ops の長い旅の序盤フェーズとして何が適切かを、
自社理解の深化とセットで選択することが重要

AI を取り巻く利害や感情を超えて、
「面白がる、夢をみる」 姿勢こそが、
真に価値ある AI 活用へと繋がる