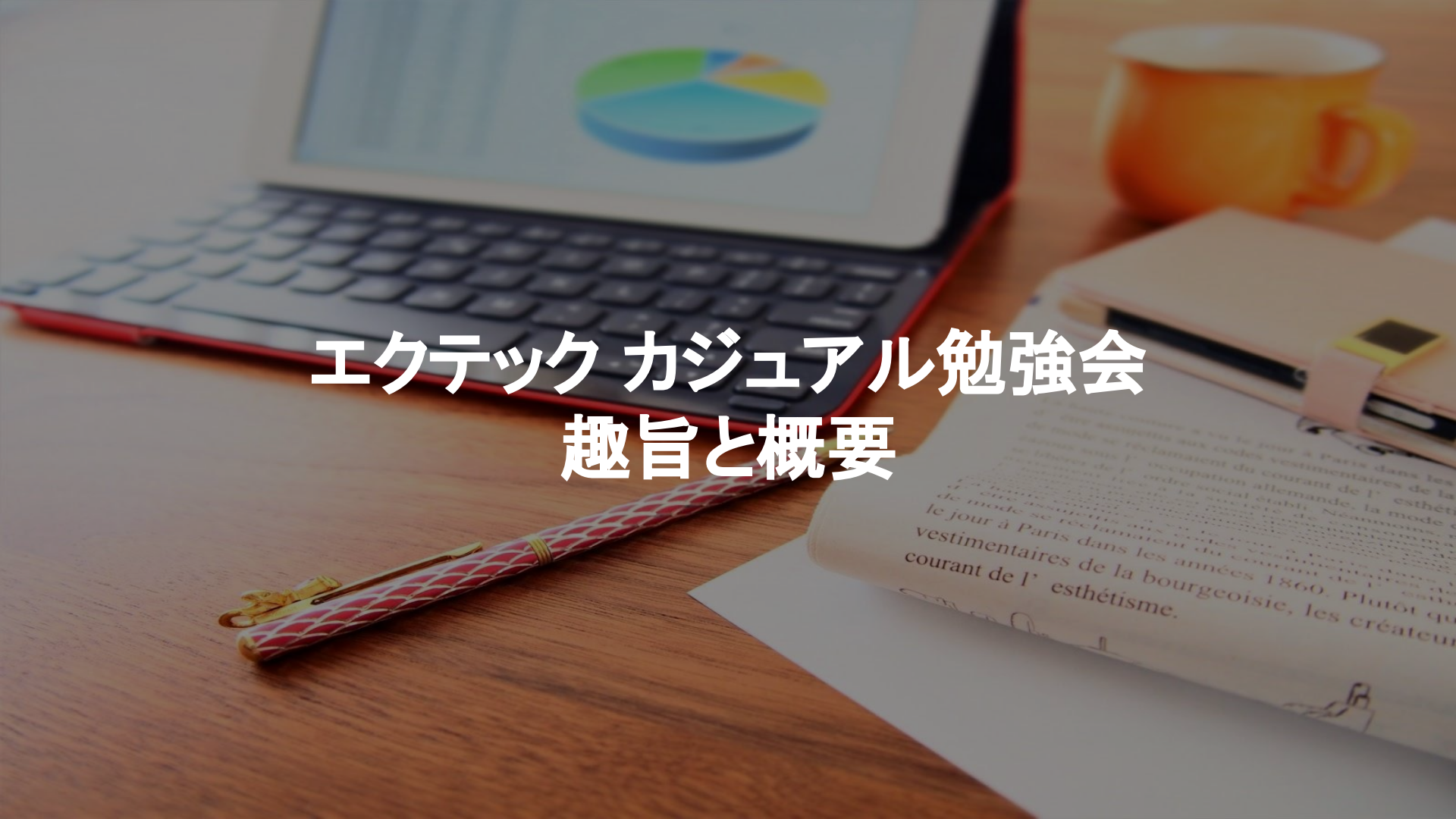
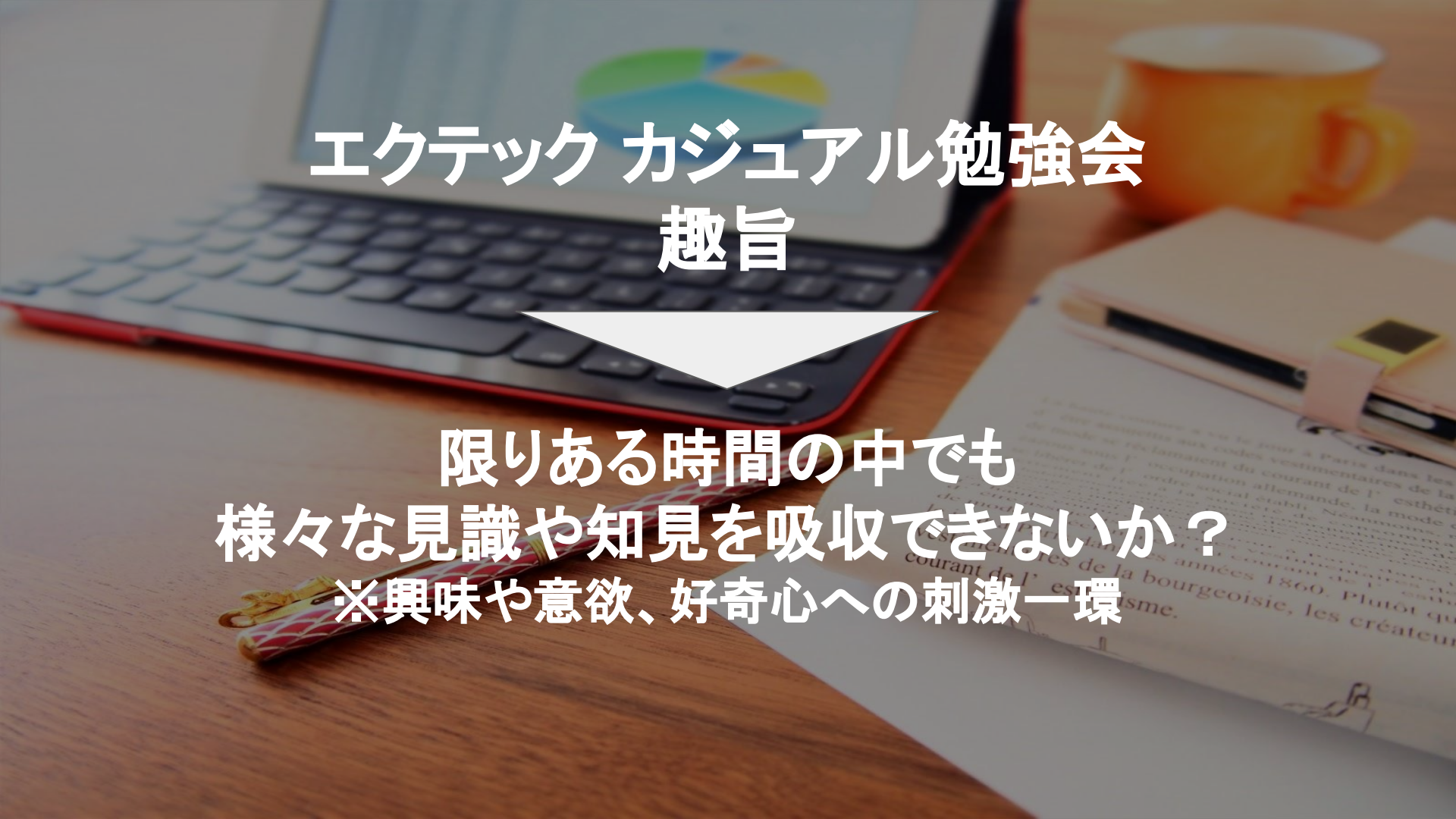
# エクテック
# カジュアル勉強会

Pythonで自然言語処理＆トピックモデルを学ぶ



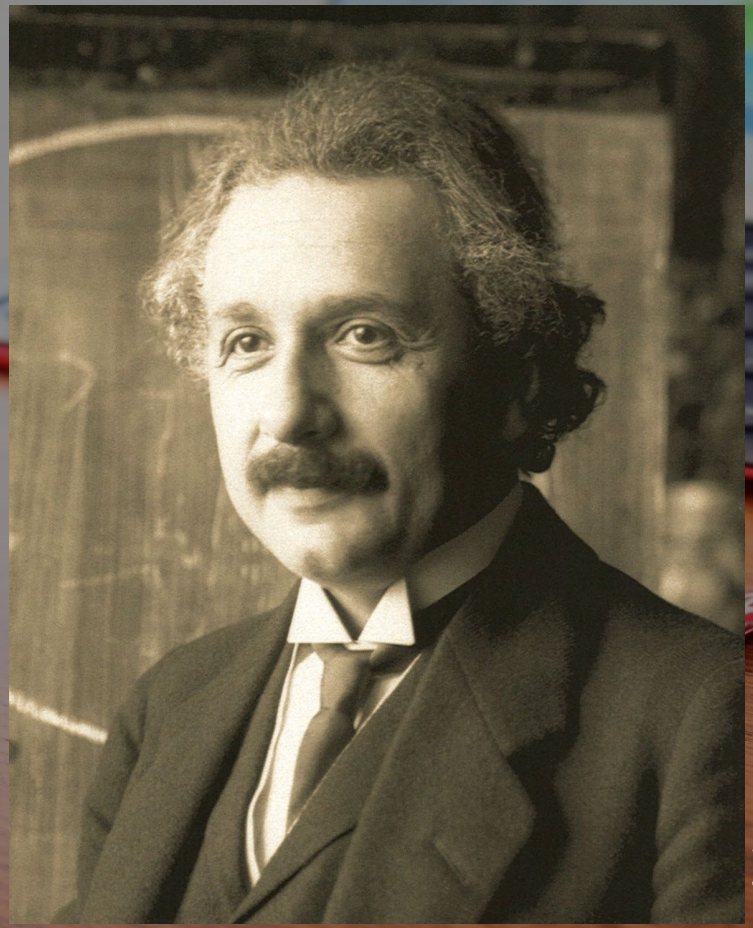

# エクテック カジュアル勉強会
# 趣旨と概要



# エクテック カジュアル勉強会
# 趣旨

## 限りある時間の中でも
## 様々な見識や知見を吸収できないか？

※興味や意欲、好奇心への刺激一環

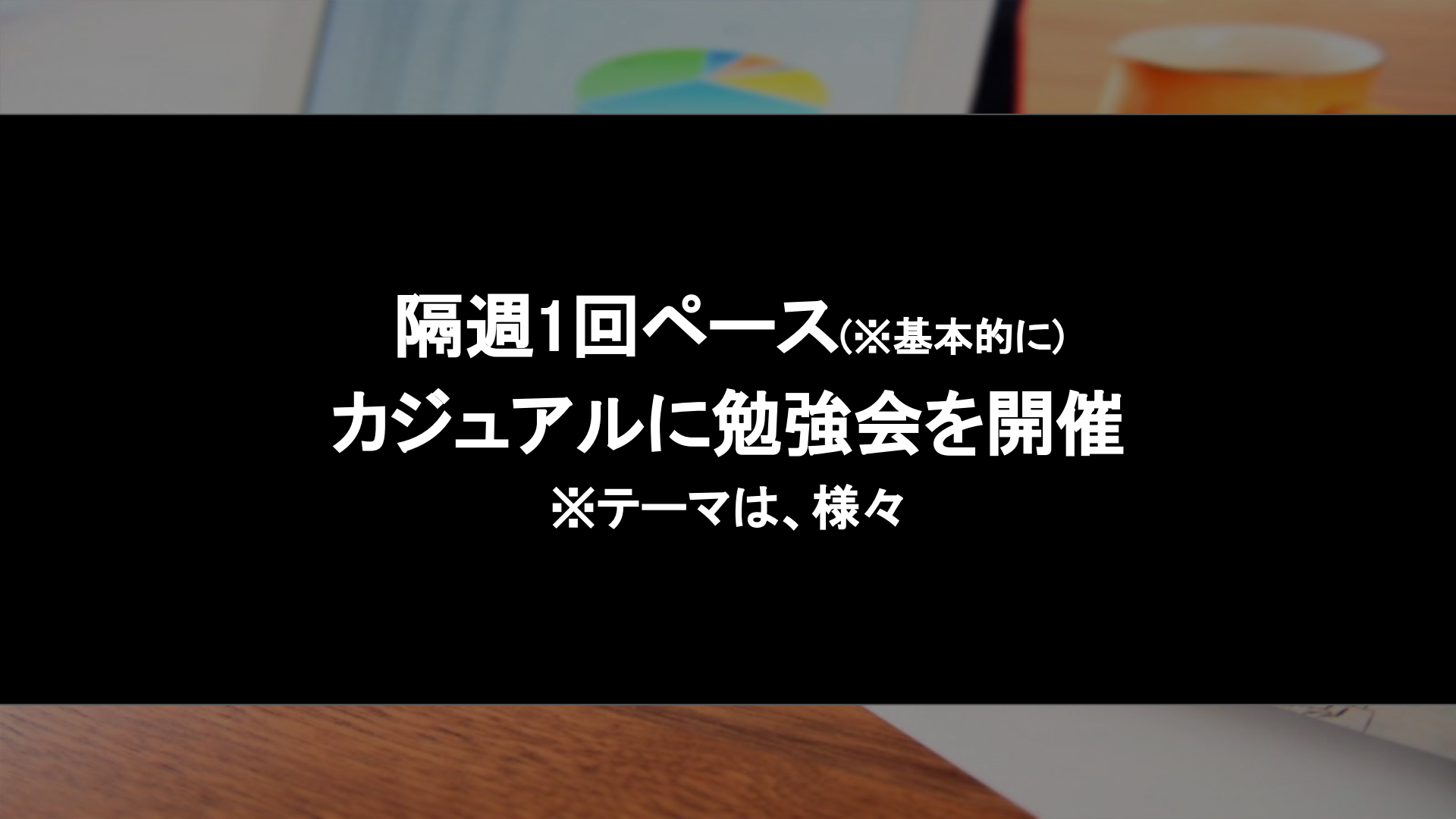


# 宇宙で最も力強いのは
# 幅広い興味である

理論物理学者
アルベルト・アインシュタイン

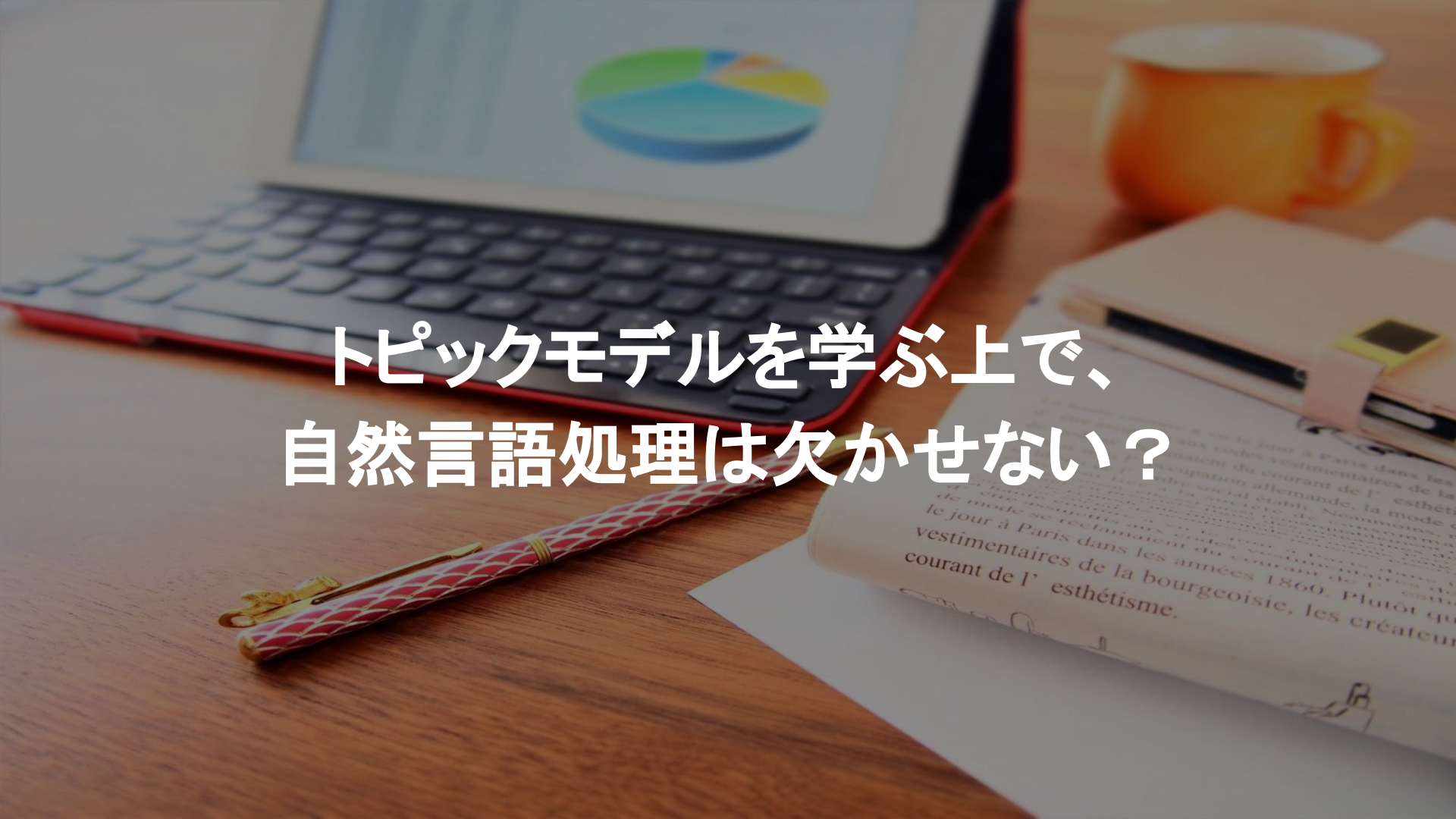
# 勉強したいと思う源泉は、新しいもの、珍しいもの、自分とは違うものに対する好奇心です

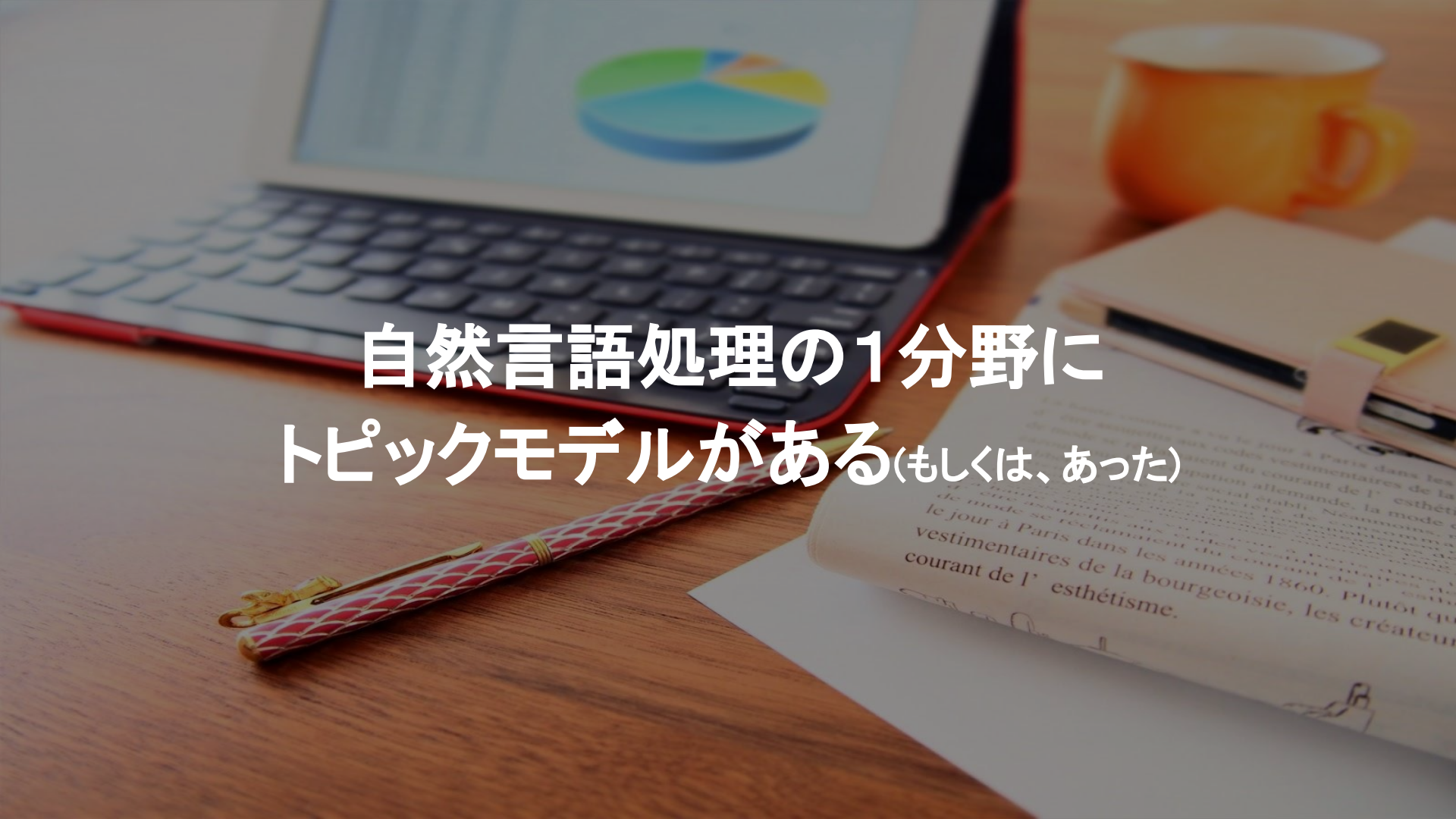
ファーストリテイリング 代表取締役会長兼社長 柳井 正

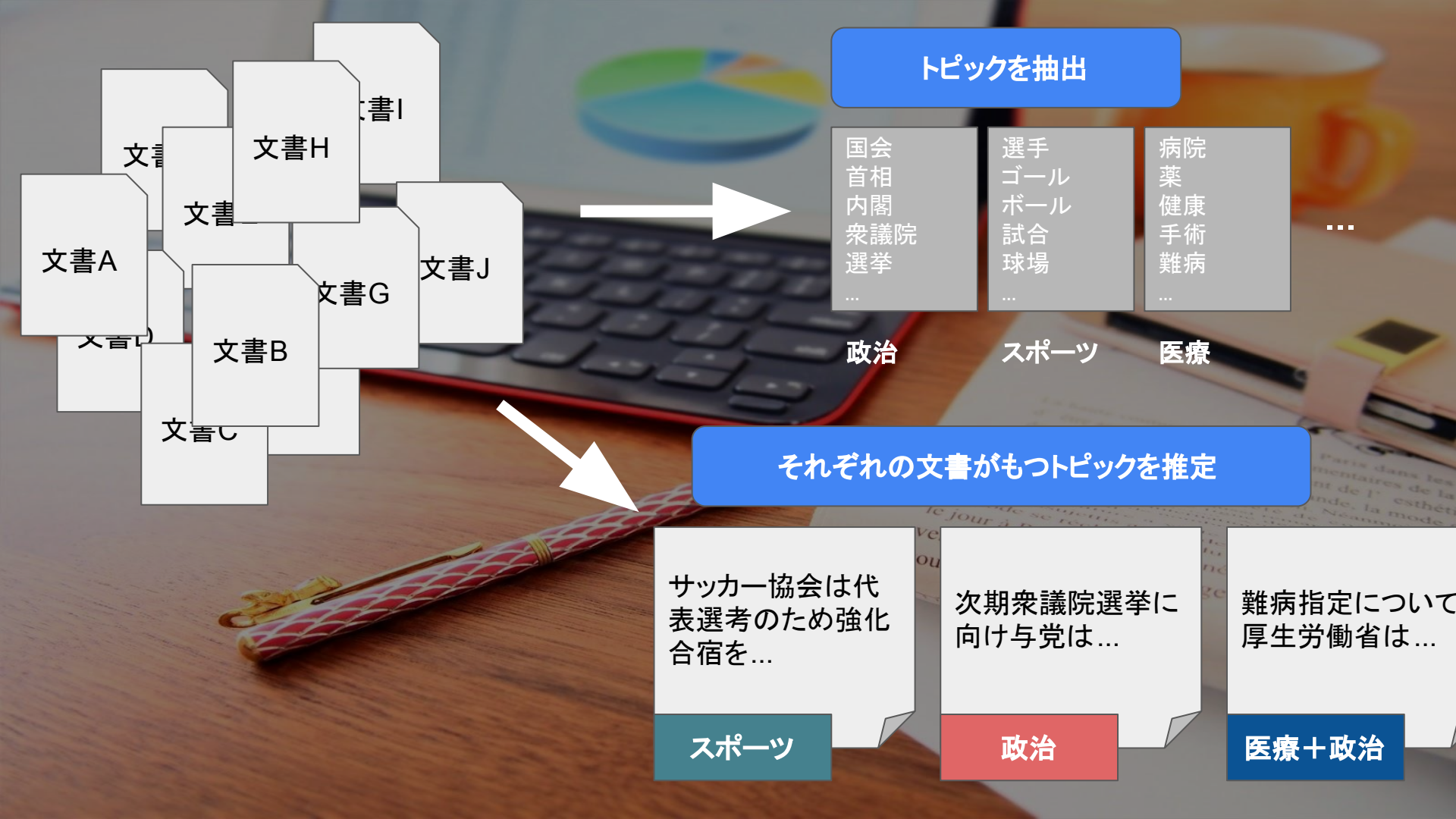


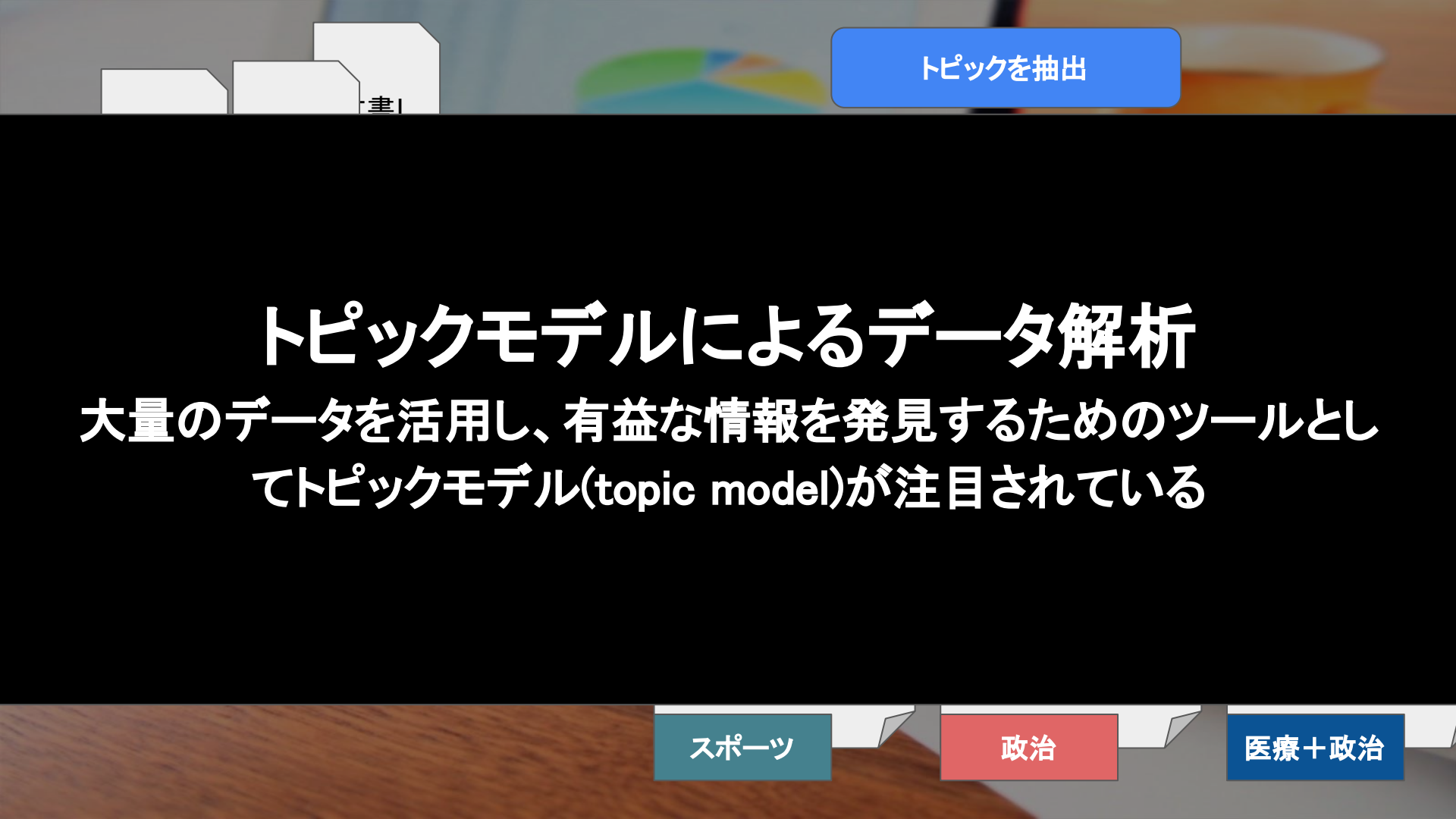
# 隔週1回ペース(※基本的に)
# カジュアルに勉強会を開催
## ※テーマは、様々


今回は...
自然言語処理とトピックモデル


トピックモデルを学ぶ上で、
自然言語処理は欠かせない？

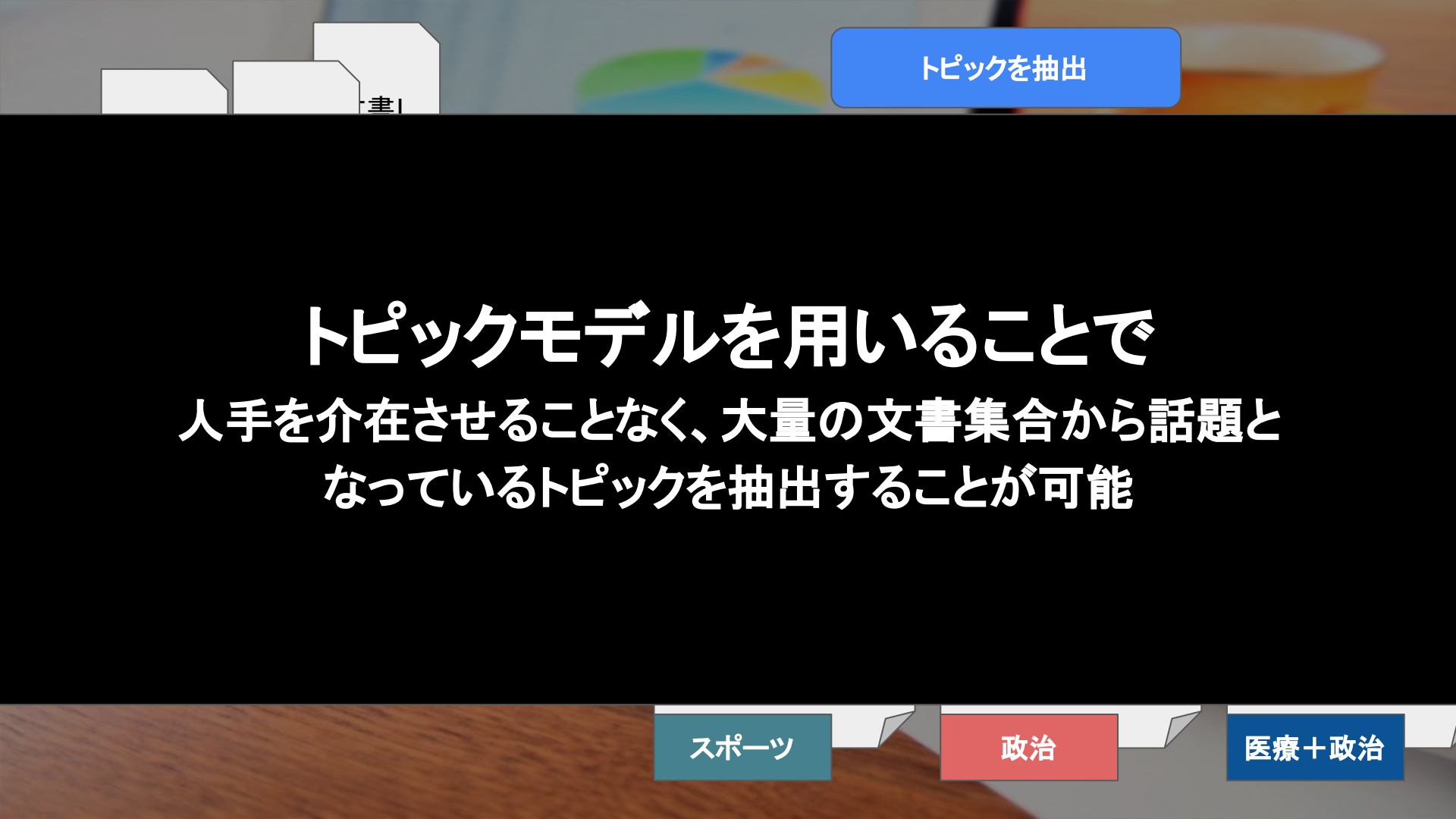

自然言語処理の1分野に
トピックモデルがある(もしくは、あった)


文書A
文書B
文書G
文書H
文書I
文書J
トピックを抽出
国会
首相
内閣
衆議院
選挙
…
選手
ゴール
ボール
試合
球場
…
病院
薬
健康
手術
難病
…
…
政治
スポーツ
医療
それぞれの文書がもつトピックを推定
サッカー協会は代表選考のため強化合宿を…
スポーツ
次期衆議院選挙に向け与党は…
政治
難病指定について厚生労働省は…
医療＋政治



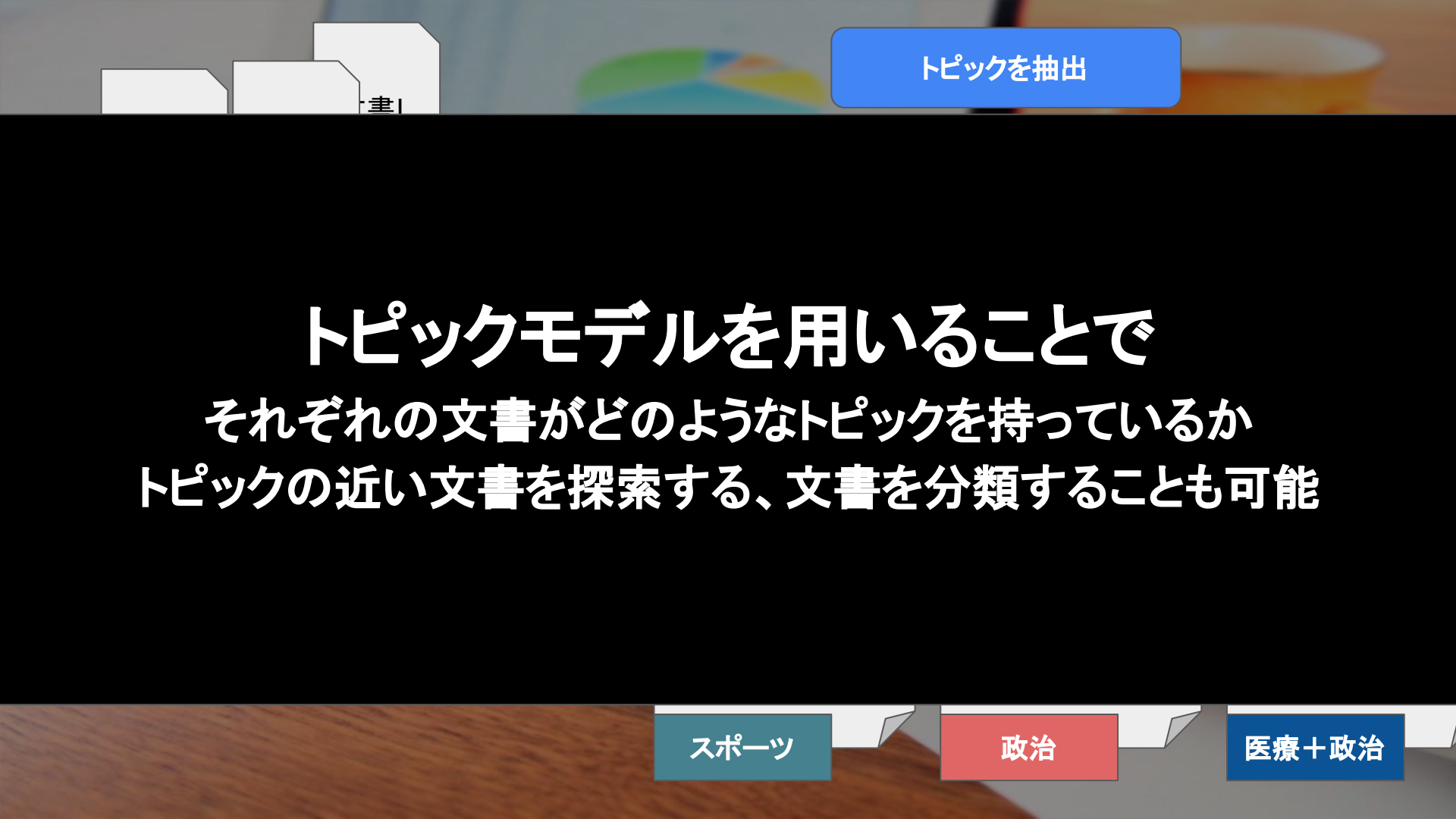

# トピックモデルによるデータ解析

## 大量のデータを活用し、有益な情報を発見するためのツールとしてトピックモデル(topic model)が注目されている




# トピックモデルを用いることで

## 人手を介在させることなく、大量の文書集合から話題となっているトピックを抽出することが可能



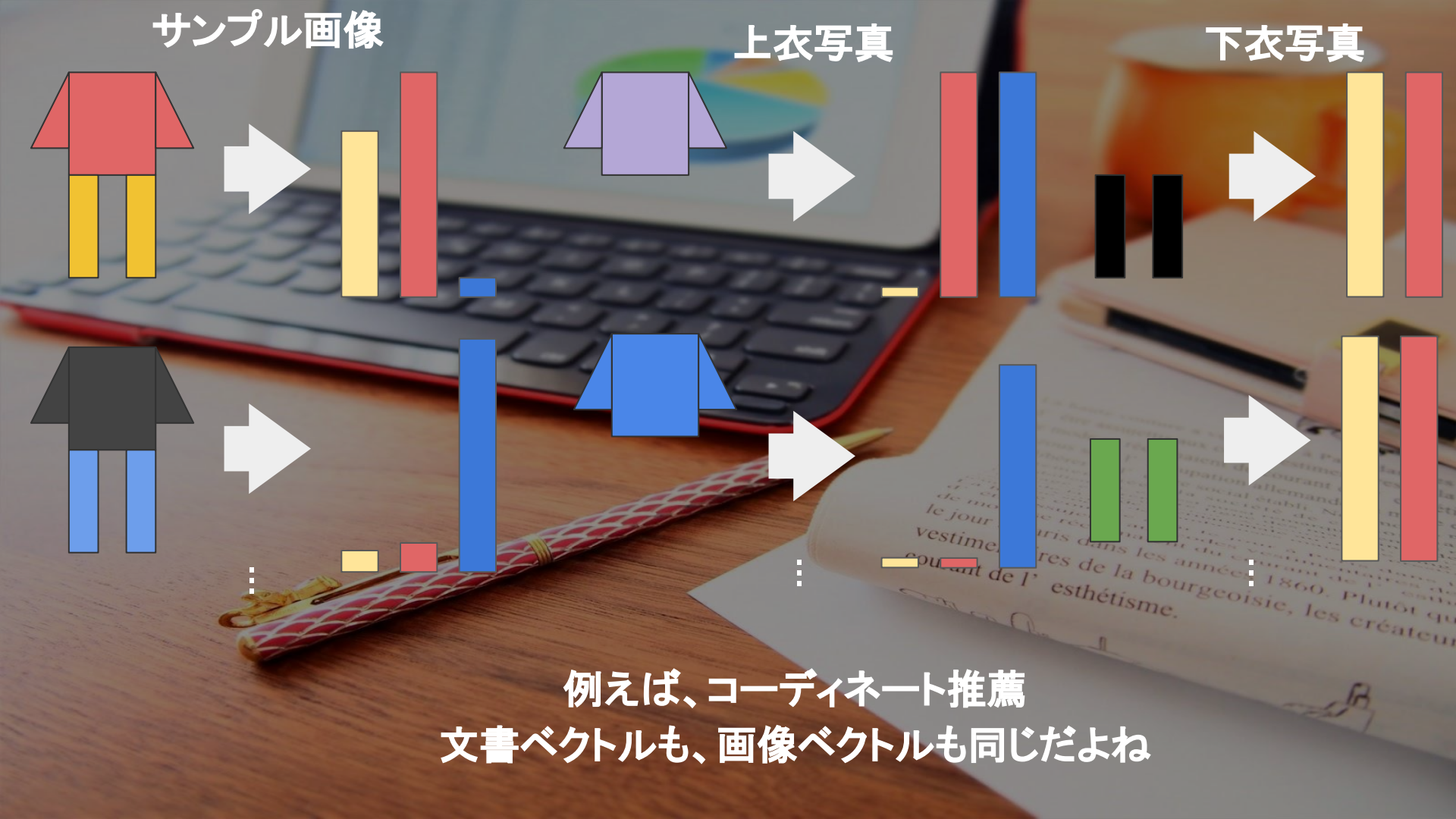

# トピックモデルを用いることで

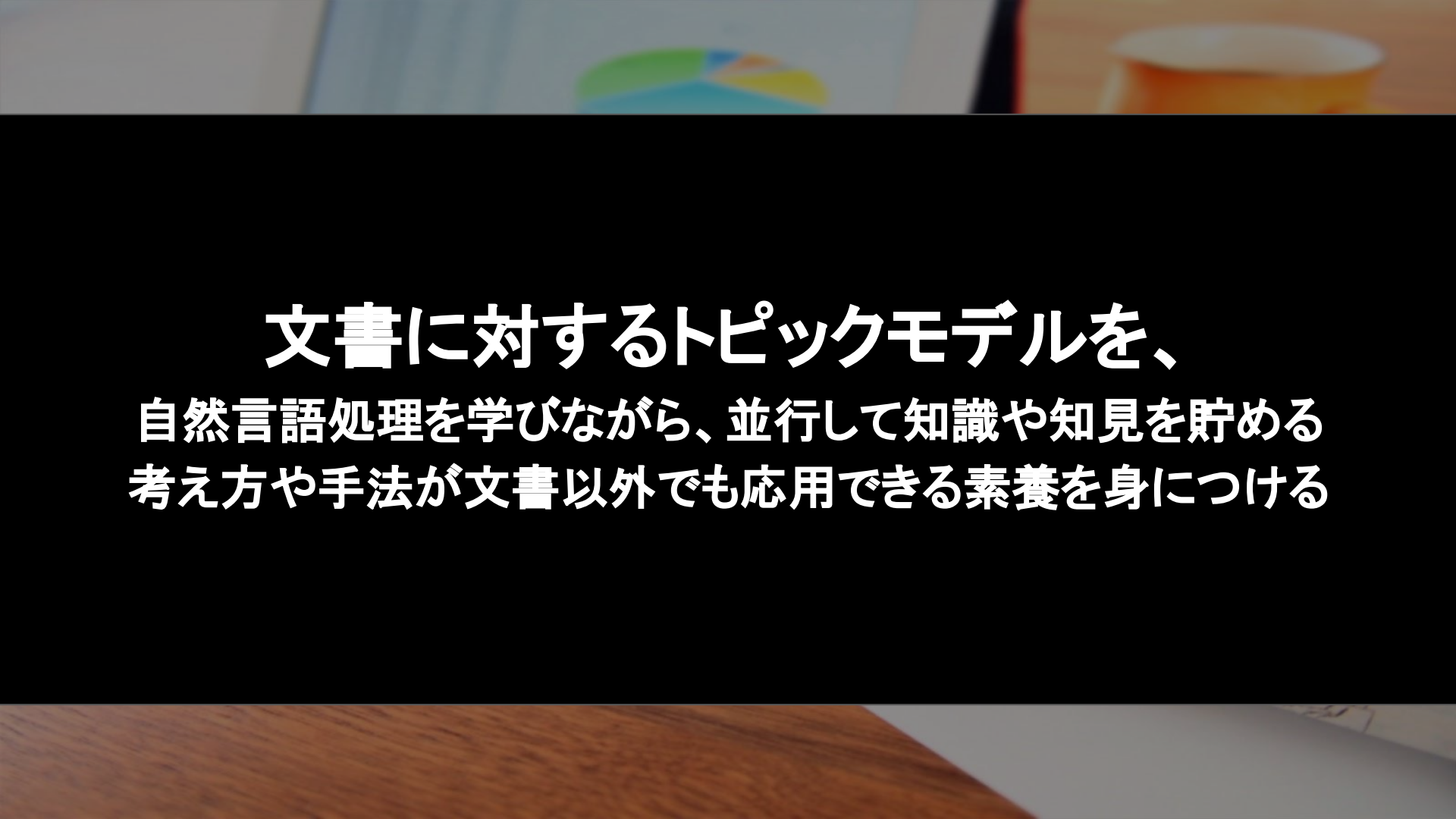
## それぞれの文書がどのようなトピックを持っているか

## トピックの近い文書を探索する、文書を分類することも可能

# トピックモデル応用分野

- 画像処理
- 推薦システム
- ソーシャルネットワーク
- バイオインフォマティクス
- 音楽情報処理
- 予測・異常検知

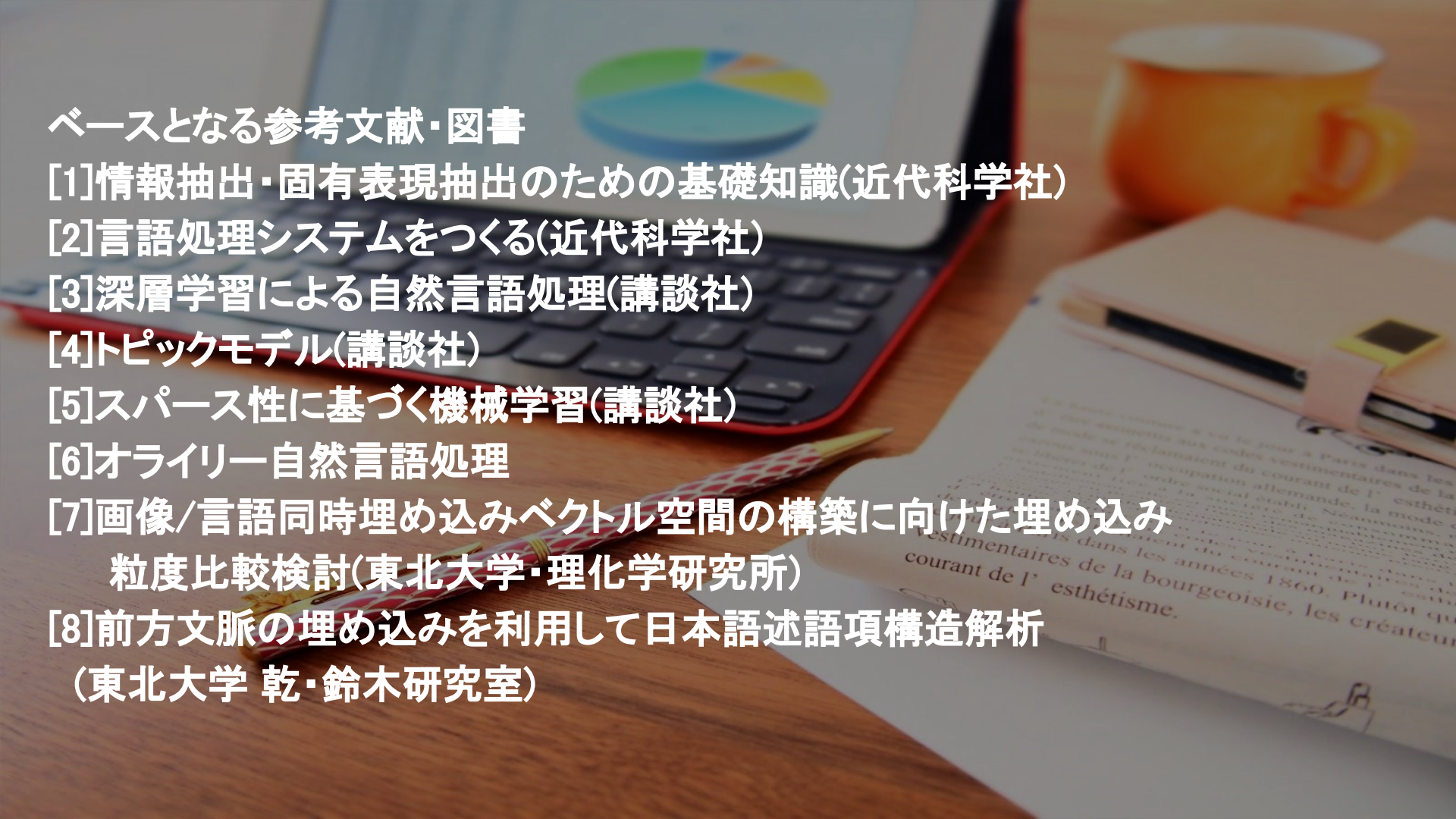

サンプル画像
上衣写真
下衣写真
例えば、コーディネート推薦
文書ベクトルも、画像ベクトルも同じだよね



# 文書に対するトピックモデルを、

自然言語処理を学びながら、並行して知識や知見を貯める

考え方や手法が文書以外でも応用できる素養を身につける



ベースとなる参考文献・図書

[1]情報抽出・固有表現抽出のための基礎知識(近代科学社)
[2]言語処理システムをつくる(近代科学社)
[3]深層学習による自然言語処理(講談社)
[4]トピックモデル(講談社)
[5]スパース性に基づく機械学習(講談社)
[6]オライリー自然言語処理
[7]画像/言語同時埋め込みベクトル空間の構築に向けた埋め込み
粒度比較検討(東北大学・理化学研究所)
[8]前方文脈の埋め込みを利用して日本語述語項構造解析
(東北大学 乾・鈴木研究室)

# Google Colaboratory


第2回に入る前に...

# データの1例: 一般的な予測(異常検知を含む)について

カテゴリー変数としての順序尺度・名義尺度、数値変数としての連続変数・離散変数によって属性と観測値を分析して、未来予測にあたる目的変数を、特徴をよく表す説明変数を用いて予測モデル定義を行います。

| | timestamp | location_id | device_id | variable1 | variable2 | variable3 | variable4 | variable5 |
|---|---|---|---|---|---|---|---|---|
| 0 | 2020-07-01 00:09:57 | 10 | 4 | 2.250 | 0.083 | 0.228 | 0.147 | 54.50 |
| 1 | 2020-07-01 00:09:58 | 10 | 5 | 1.602 | 0.098 | 0.187 | 0.173 | 58.30 |
| 2 | 2020-07-01 00:10:01 | 10 | 1 | 1.501 | 0.039 | 0.123 | 0.127 | 52.49 |
| 3 | 2020-07-01 00:10:01 | 10 | 3 | 1.487 | 0.106 | 0.117 | 0.093 | 61.58 |
| 4 | 2020-07-01 00:10:06 | 10 | 2 | 1.532 | 0.079 | 0.171 | 0.143 | 61.41 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 21595 | 2020-07-30 23:55:57 | 10 | 5 | 1.945 | 0.097 | 0.165 | 0.107 | 60.79 |
| 21596 | 2020-07-30 23:57:44 | 10 | 2 | 2.167 | 0.113 | 0.128 | 0.153 | 61.99 |
| 21597 | 2020-07-30 23:59:13 | 10 | 1 | 2.005 | 0.059 | 0.143 | 0.133 | 52.75 |
| 21598 | 2020-07-31 00:00:16 | 10 | 4 | 1.395 | 0.096 | 0.104 | 0.083 | 51.53 |
| 21599 | 2020-07-31 00:05:19 | 10 | 3 | 1.857 | 0.107 | 0.124 | 0.177 | 54.81 |

属性

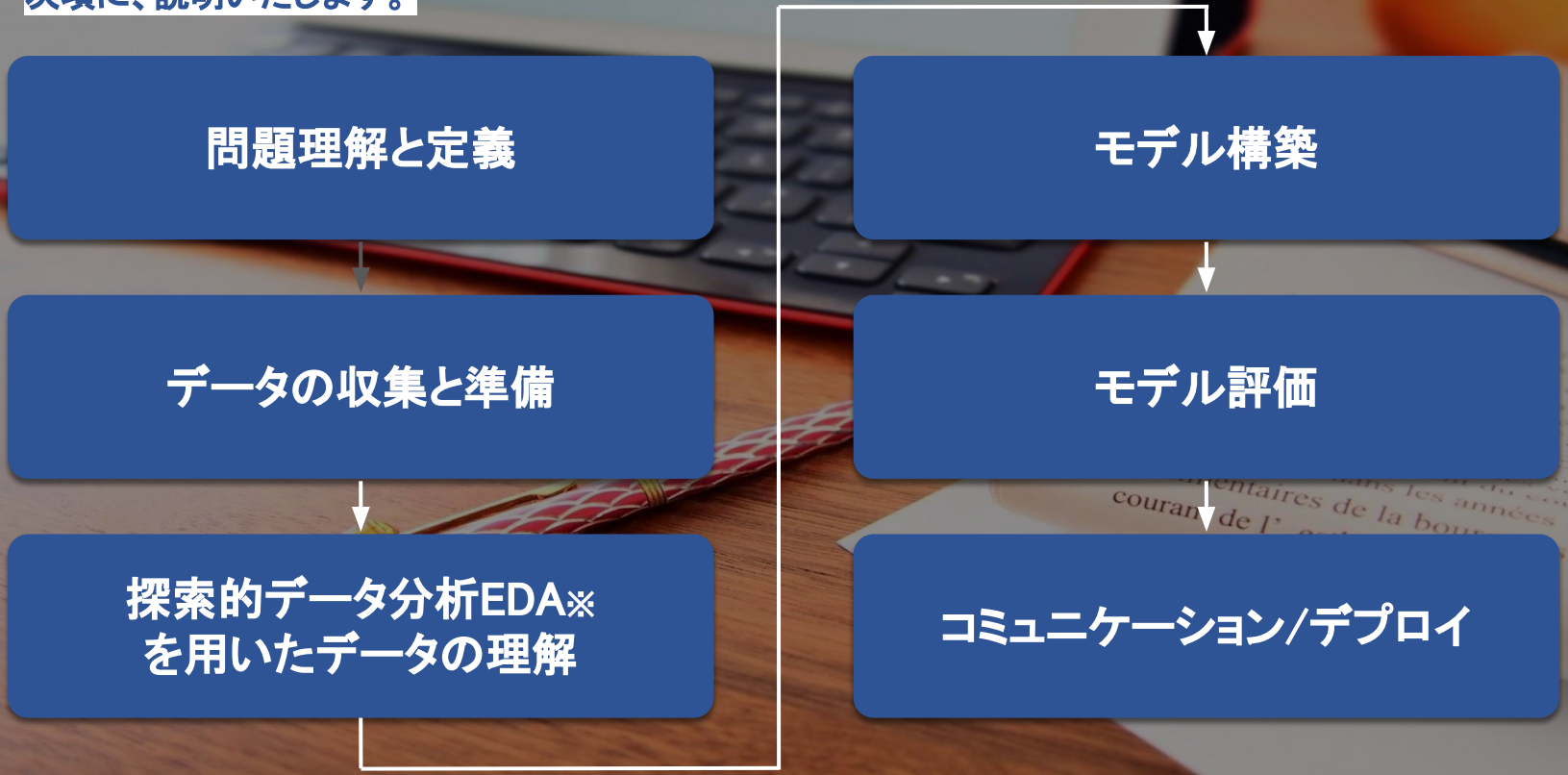
観測値

未来予測

?

# 分析の基本プロセス

AIやデータサイエンスなプロジェクトを進めるにあたっての大枠となるワークフローとなります。
次項に、説明いたします。




※Exploratory Data Analysis(EDA) データの特徴を探求し、構造を理解することを目的とした手法・概念

# 分析の基本プロセス: 問題理解と定義

**ゴール: 問題と潜在的な解決策がどのようなものか理解し、問題解決のための要件を定義する。**
**※プロセスの最初の段階であると同時に、キーとなる重要な段階となります。**

## 問題理解と定義

ステークホルダー(社内外利害関係者)とともに、予測モデルの目的が何であるかを定義するフェーズとなります。解決する必要のある問題を理解し、ビジネスの観点からのソリューションがどのようなものであるかを明確にします。同時に、プロジェクトの要件も明示的に定義します。
要件は、入力の観点から考慮する必要があり、例えば、ソリューションの作成に必要なデータは何か、必要な形式は何か、必要なデータの量、分析および予測モデルの出力がどのようなものか、議論されている問題の解決策をどのように提供するか、などです。

# 分析の基本プロセス: データの収集と準備

**ゴール: 分析の準備が整ったデータセットを用意する。**
**※利用可能なデータを、例えば DB管理者とやりとりを行い、データ提供を依頼する場面もあります。**

## データの収集と準備

必要なデータを取得するために、多くの異なるデータソースにアクセスする必要があるかもしれません。
場合によっては、データがまだ存在していない可能性があり、それらデータを取得する計画を練る必要があるかもしれません。この段階でのゴールは、予測モデルの構築に使用するデータセットを得ることです。
データを取得する過程において、データに関する潜在的な問題が特定される場合もあります。
データセットを準備するためのタスクを実行しながら、利用可能なデータがビジネス理解の段階で定義した問題を解決するために十分ではないことに気づいた場合、前段の『問題理解と定義』に戻り、再度ステークホルダーと議論して、問題と解決策を再定義する必要があります。
この段階では、データ内部の欠損値や外れ値は考慮しません。
次の『EDAを用いたデータの理解』にて、これらデータ内部の特性を考えていきます。

# 分析の基本プロセス: EDAを用いたデータの理解

**ゴール: データを理解する。**
**※データセットを準備したのち、データセット・変数・変数間の潜在的な関係を理解していきます。**

## EDAを用いたデータの理解

1変数に対して棒グラフ、円グラフなど用いてデータの特性を可視化・分析することを、1変量EDA、
2変数に対して散布図、散布図行列など用いてデータの特性を可視化・分析することを、2変量EDA、
多変数に対して複合的な可視化手法を用いて分析をすることを、多変量EDAと呼びます。
EDAそのものの作業は非常に面倒である一方、多くの役に立たない作業を試みることで、矛盾した情報を見つけ出し、アイデアを修正し、興味深い特性や事実を発見し、それらプロセスを経て、新たな画期的なソリューションを作り出せる可能性があります。
『データセットにはどのような種類の変数がありますか？』
『それらの分布はどのようになっていますか？』
『欠損値はありますか？冗長な変数はありますか？外れ値はありますか？』
『変数と、予測するべき目的変数の関係は何ですか？』、... etc

# 分析の基本プロセス: モデル構築とモデルの評価

**ゴール: 問題を解決するモデルを作成し、最適なモデルを選択し、
モデルがソリューションを提供する上で、どれほど優れているかを判断する。**

## モデル構築とモデルの評価

Pythonをはじめとした、機械学習、ディープラーニング、ベイズ統計など様々なアプローチを扱います。
モデルのトレーニングは機械学習に関連付けられ、推定は統計に関連付けられます。アプローチ、モデルの種類、およびトレーニング・推定プロセスは、解決しようとしている問題と探しているソリューションによってのみ決定されなければなりません。
いくつかのモデルを構築した上で、それら候補となり得るモデルやサブセットの優劣を評価します。
予測分析プロセスの評価は、解決する問題によって決まります。通常は1つ以上の指標(KPI)をもとに、モデルのパフォーマンスを評価します。プロジェクトによっては、計算量、解釈性、使いやすさ、方法論などの指標以外のことも考慮される場合があります。
最良のモデルは、最も派手なものでも、最も複雑なものでもありません。
可能な限り最良の方法で問題を解決できるモデルが、最良のモデルといえます。

# 分析の基本プロセス: コミュニケーション/デプロイ

**ゴール: モデルとその結果を利用する**
**※モデルをどのように利用するかはプロジェクトによって異なります**

## コミュニケーション/デプロイ

結果や予測は、主要なステークホルダーに共有するためのレポート対象となります。
コミュニケーションは、これらステークホルダーに結果や予測を効果的に共有するための手段となります。
モデルは、Web、デスクトップ、モバイルなどのソフトウェアアプリケーションの一部として組み込まれる場合があります。この場合、アプリケーションに実装するソフトウェア開発チームと密にやりとりする、メンバーとしてチームに参画するなど場面に応じてコミュニケーションの礎(いしずえ)が必要になります。
モデルの実装には、例えば、”Dash”や”tableau”を用いることもあります。

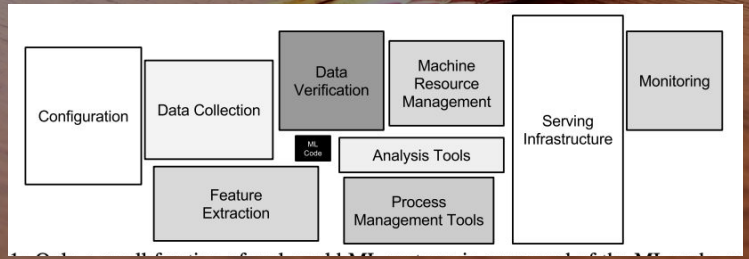
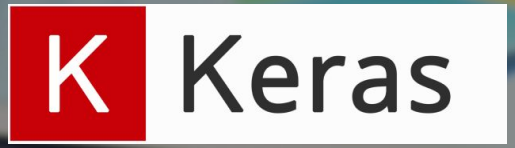



Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

# 補足1: オープンソースソフトウェアの活用







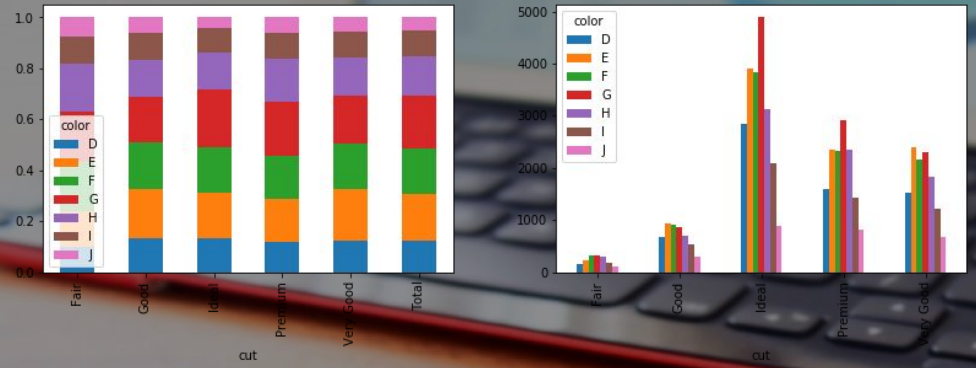

```
model.summary()

_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_2 (InputLayer)         (None, 30)                0         
_________________________________________________________________
embedding_2 (Embedding)      (None, 30, 26)            260000    
_________________________________________________________________
spatial_dropout1d_2 (SpatialDro (None, 30, 26)         0         
_________________________________________________________________
bidirectional_2 (Bidirectional) (None, 30, 256)        119040    
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 29, 64)            32832     
_________________________________________________________________
global_average_pooling1d_2 (Glo (None, 64)             0         
_________________________________________________________________
global_max_pooling1d_2 (GlobalM (None, 64)             0         
_________________________________________________________________
concatenate_2 (Concatenate)  (None, 128)               0         
                                                                 
_________________________________________________________________
dense_2 (Dense)              (None, 11)                1419      
=================================================================
Total params: 413,291
Trainable params: 413,291
Non-trainable params: 0
_________________________________________________________________
```

データ収集・加工から分析、モデル構築・検証までPythonをはじめとした各分野で必要となるオープンソースソフトウェアを適時、利用します。これらは、オープンソースであるため、商用利用が可能となっており、商用サービスとして展開することも可能です。

※ 自然言語処理や時系列予測、異常検知・外れ値検知の分野において、よく利用される深層学習モデル(CNN+RNN)の1例

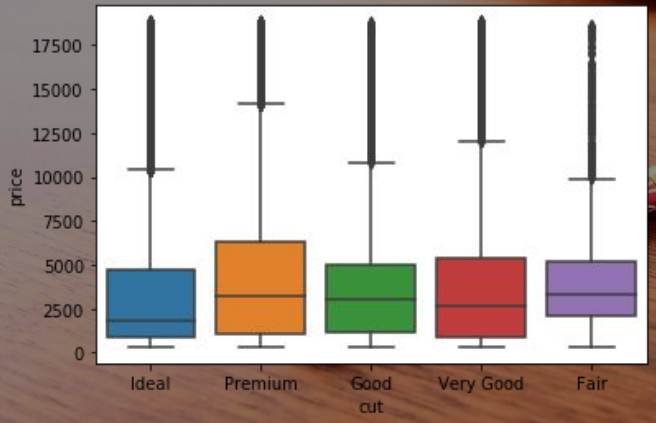
# 補足2: EDAによるデータへの理解1例

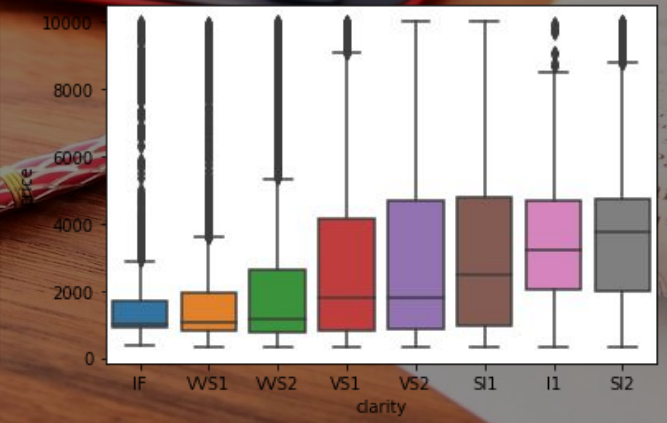


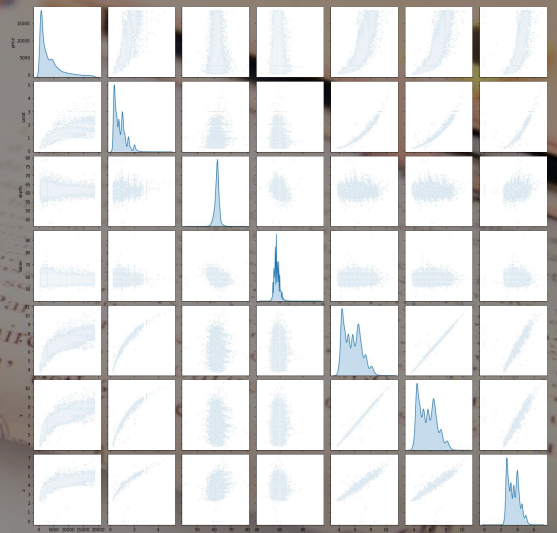




データの中身、1つ1つの変数やカラムを丁寧に分析していき、これまで意識していなかった考え方や捉え方、まったく新しい知見や発見を見つけ出します。

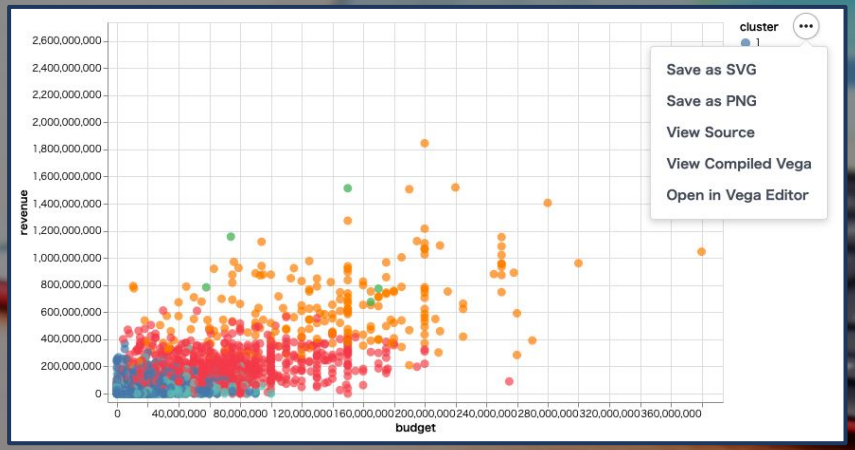


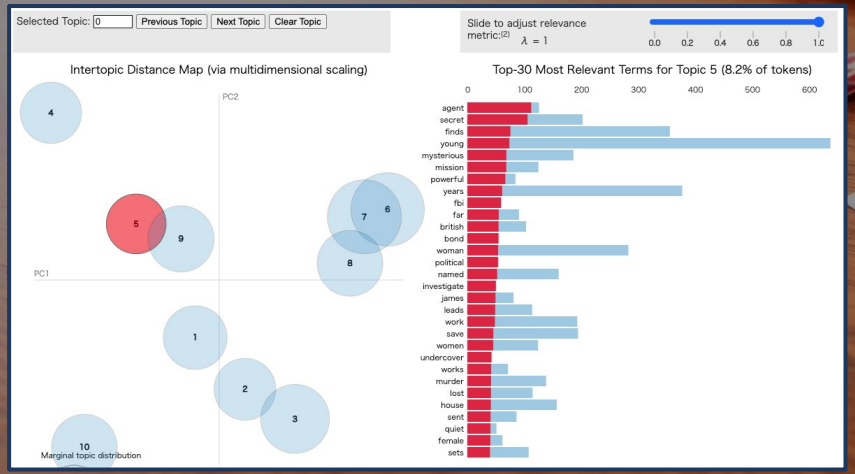



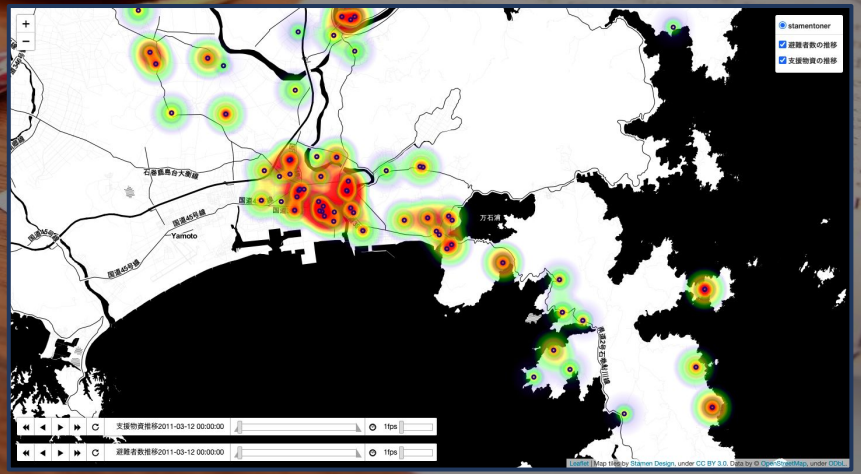
※ ダイヤモンドのオープンデータを用いた、”CUT”, “CLARITY”それぞれに対する1変量EDAの取り組み1例や、すべての変数のペアの相関を可視化する2変量EDAの取り組み1例

# 補足3: 実験的な取組みから実用システムへのシフト1例




『探索的なデータ分析 (EDA)をチーム内や組織内で共有できないか』、『構築済みモデルからアウトプットされる結果をチーム内や組織内で共有できないか』、『分析内容を外部で公開できないか』といった課題などを、様々な取組みを、社内外に公開できるツールとして開発し、システムとして展開することも可能です。

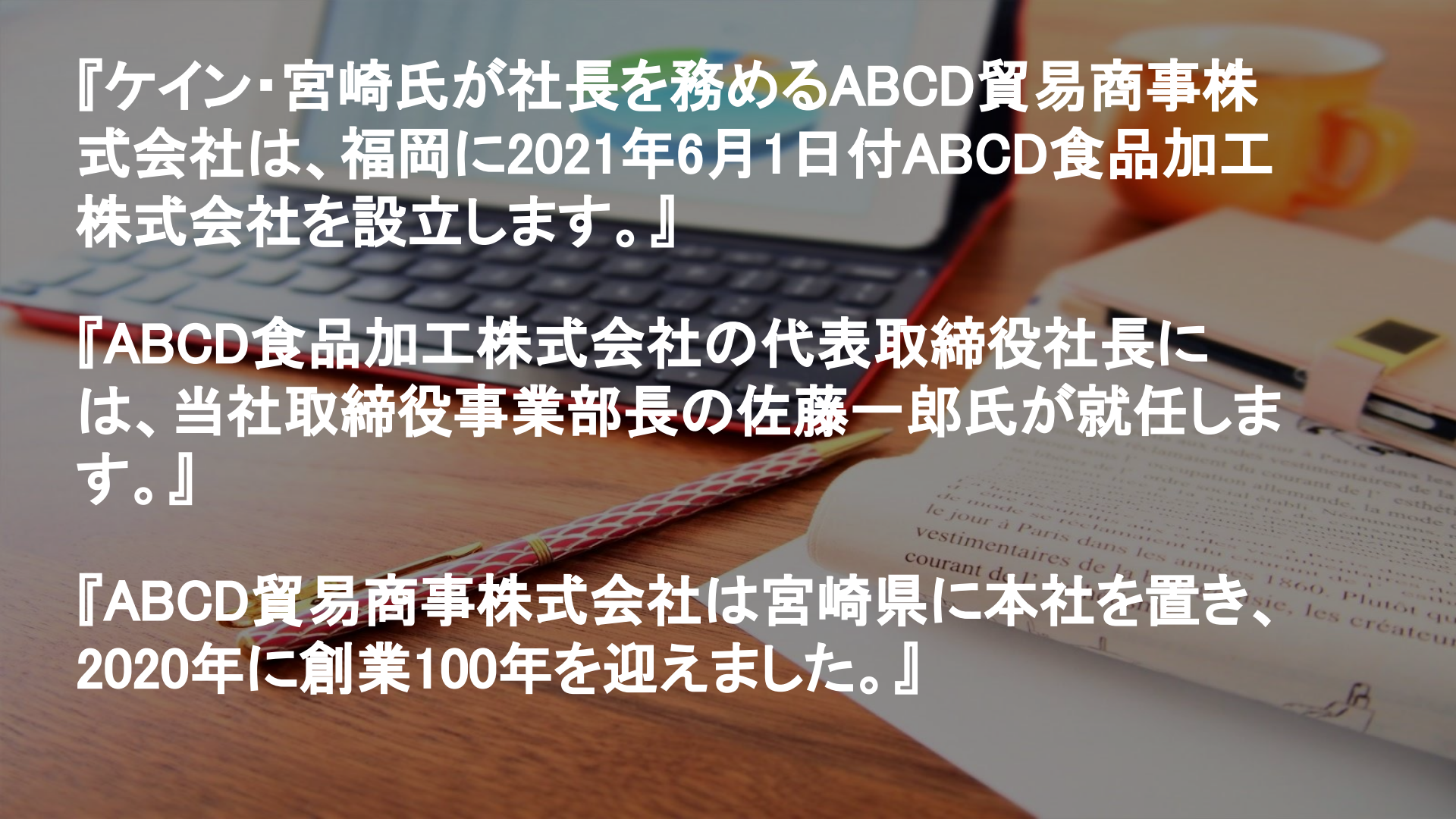


# Pythonで自然言語処理
# &トピックモデルを学ぶ

第2回目


自然言語処理パート
情報抽出

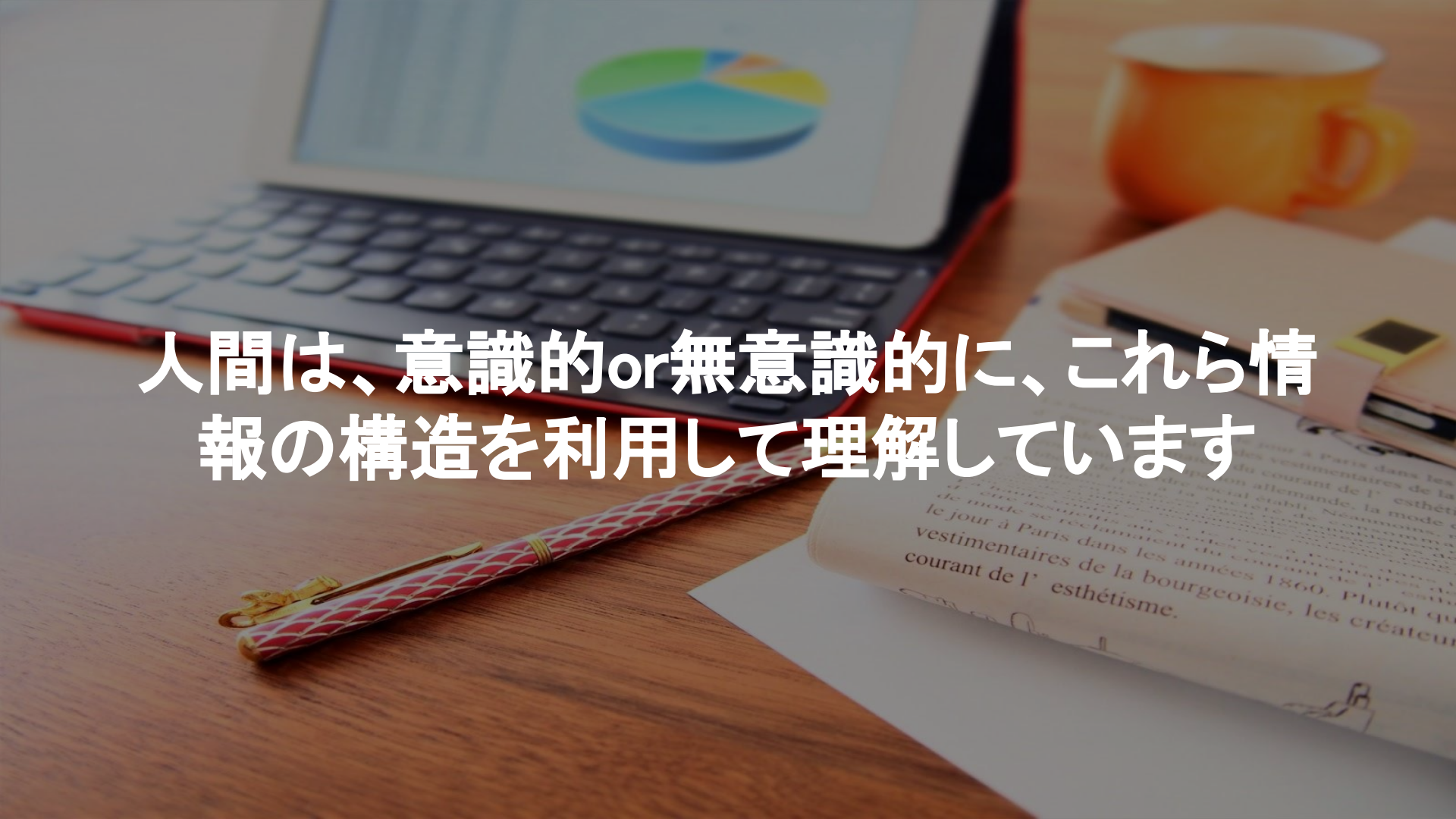


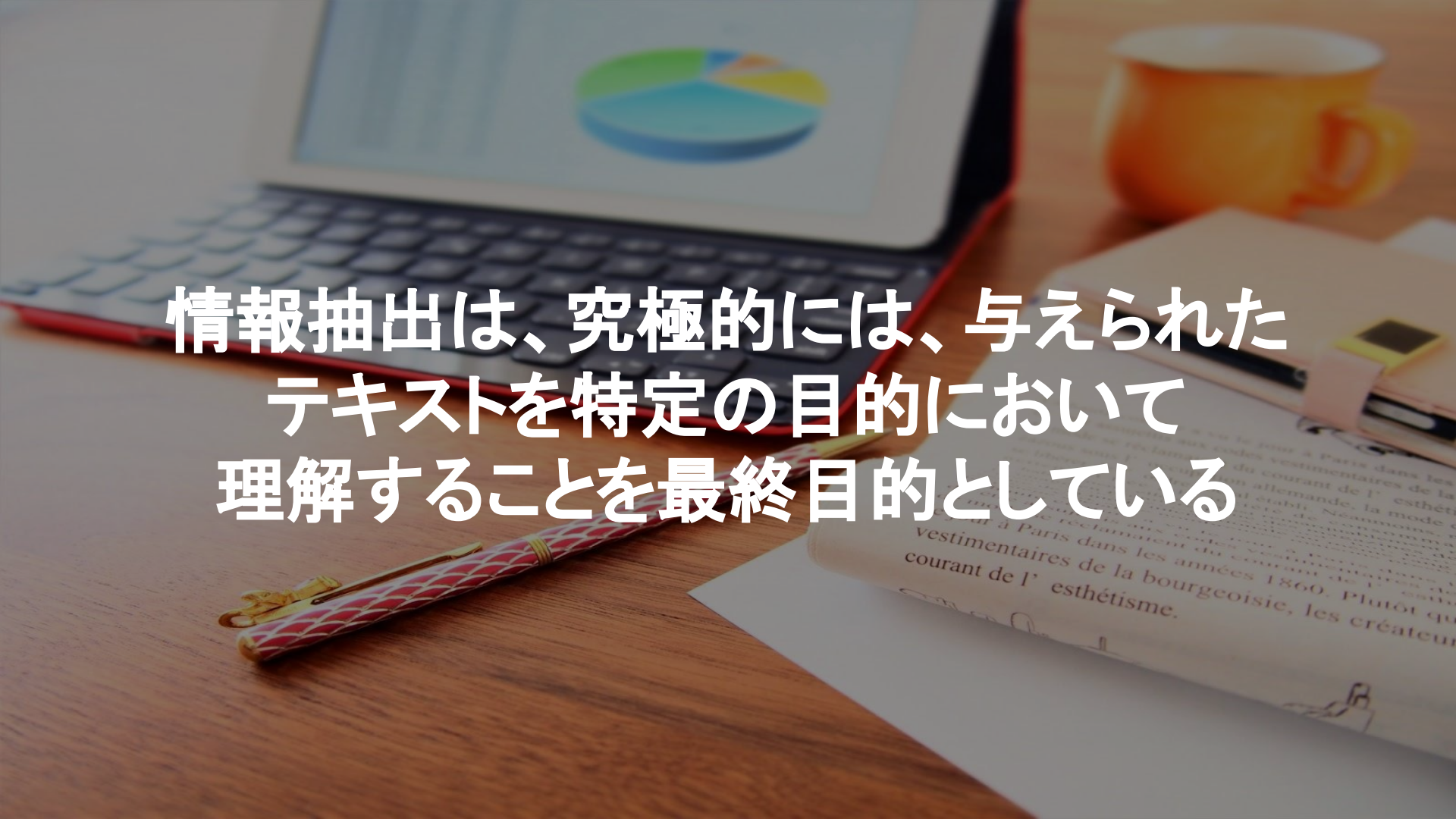
『ケイン・宮崎氏が社長を務めるABCD貿易商事株式会社は、福岡に2021年6月1日付ABCD食品加工株式会社を設立します。』

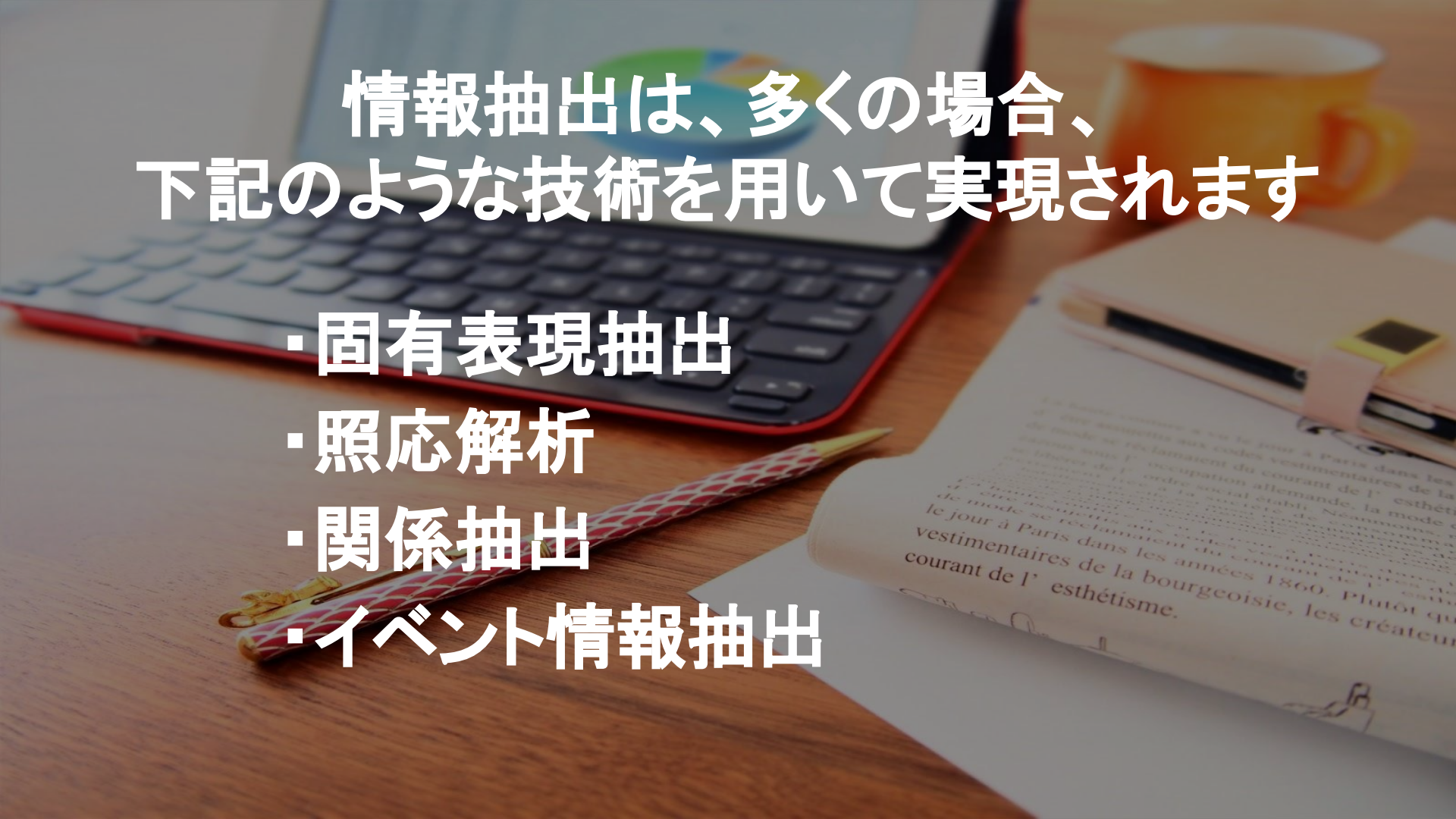
『ABCD食品加工株式会社の代表取締役社長には、当社取締役事業部長の佐藤一郎氏が就任します。』

『ABCD貿易商事株式会社は宮崎県に本社を置き、2020年に創業100年を迎えました。』

| | |
|---|---|
| 会社名 | ABCD食品加工株式会社 |
| 親会社 | ABCD貿易商事株式会社 |
| 代表者 | 佐藤一郎 |
| 設立日 | 2021年6月1日 |
| 所在地 | 福岡 |

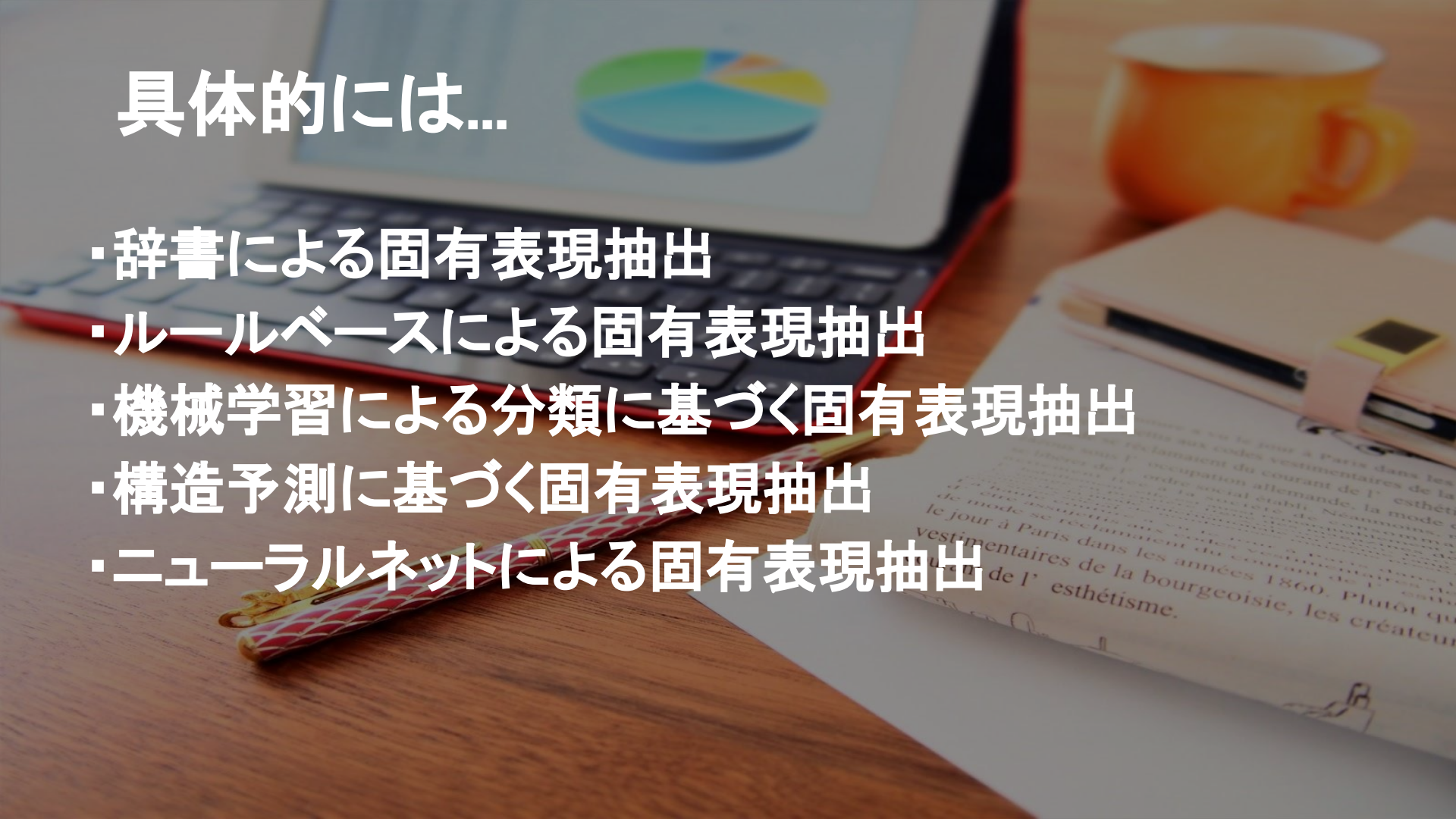

人間は、意識的or無意識的に、これら情
報の構造を利用して理解しています


情報抽出は、究極的には、与えられた
テキストを特定の目的において
理解することを最終目的としている



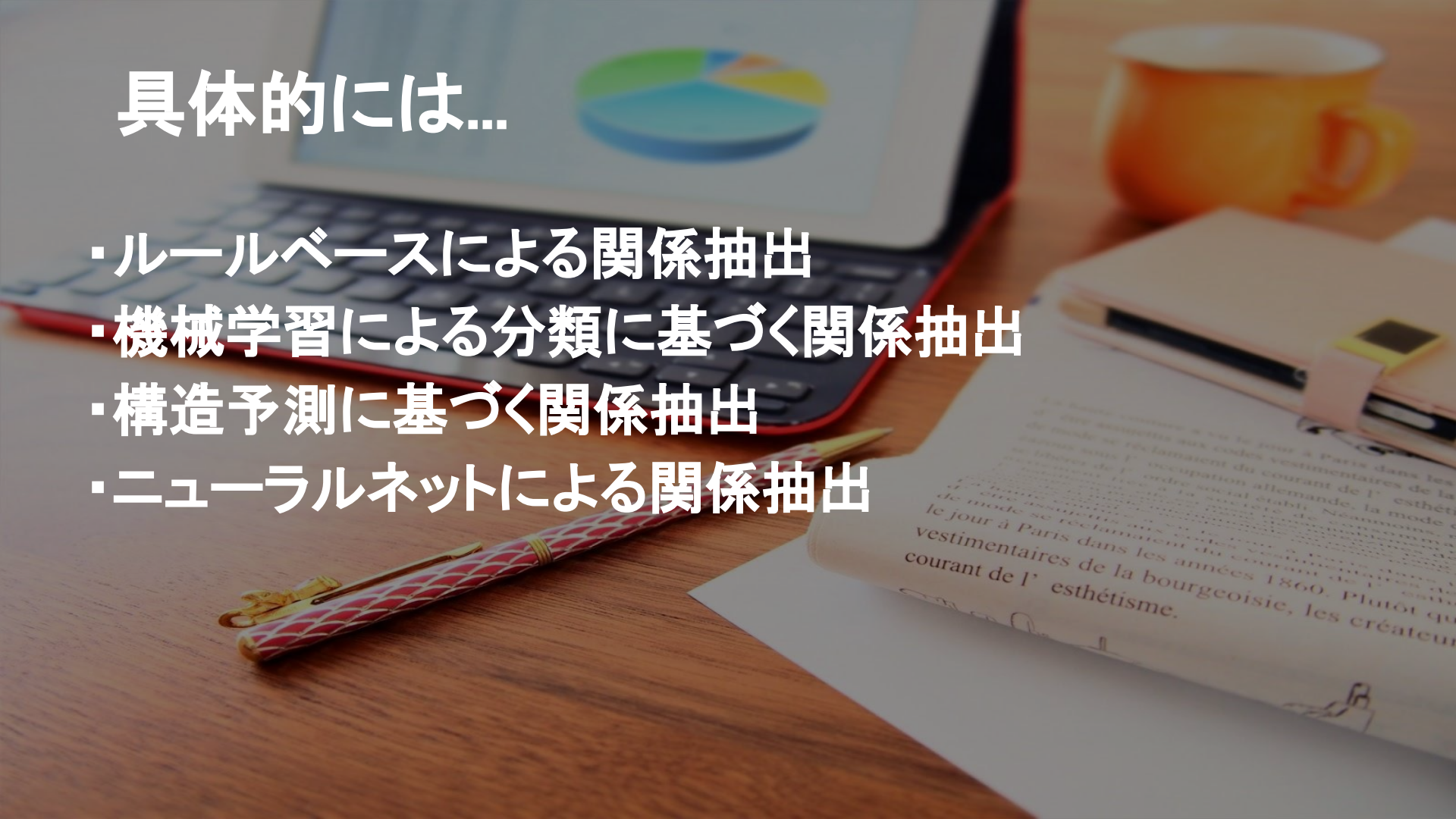
# 情報抽出は、多くの場合、
# 下記のような技術を用いて実現されます

- 固有表現抽出
- 照応解析
- 関係抽出
- イベント情報抽出



# 具体的には...

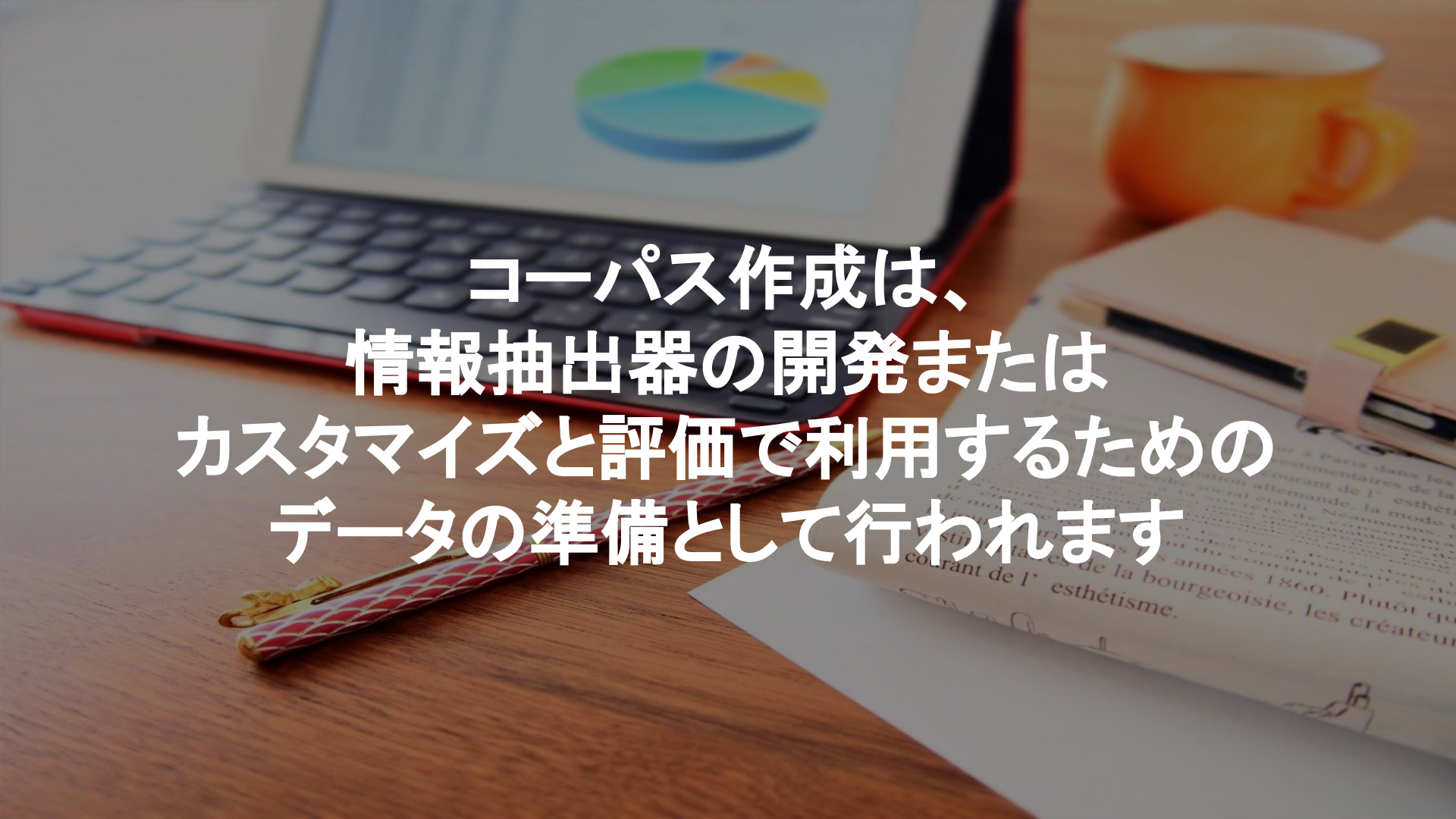
- 辞書による固有表現抽出
- ルールベースによる固有表現抽出
- 機械学習による分類に基づく固有表現抽出
- 構造予測に基づく固有表現抽出
- ニューラルネットによる固有表現抽出

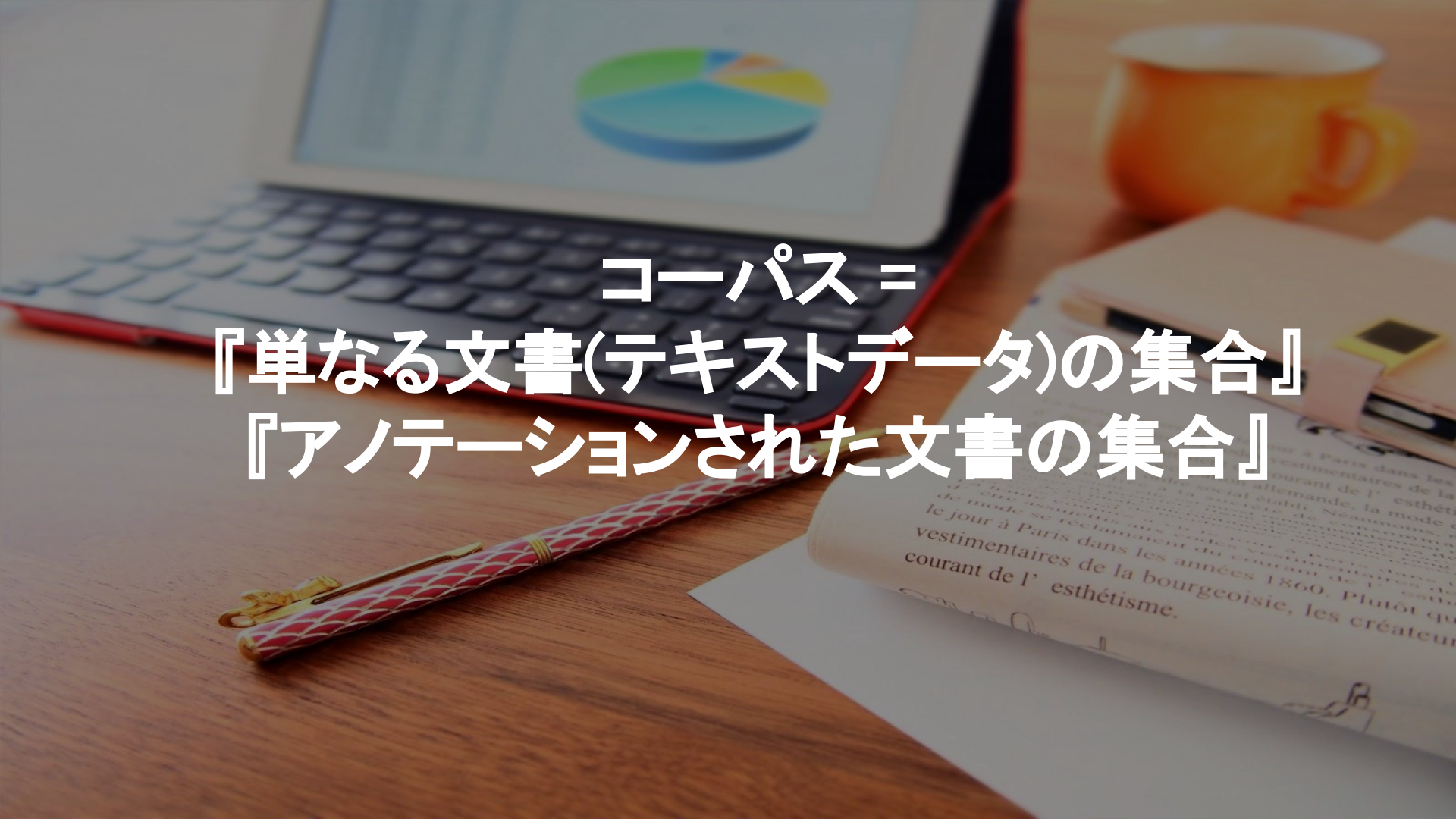


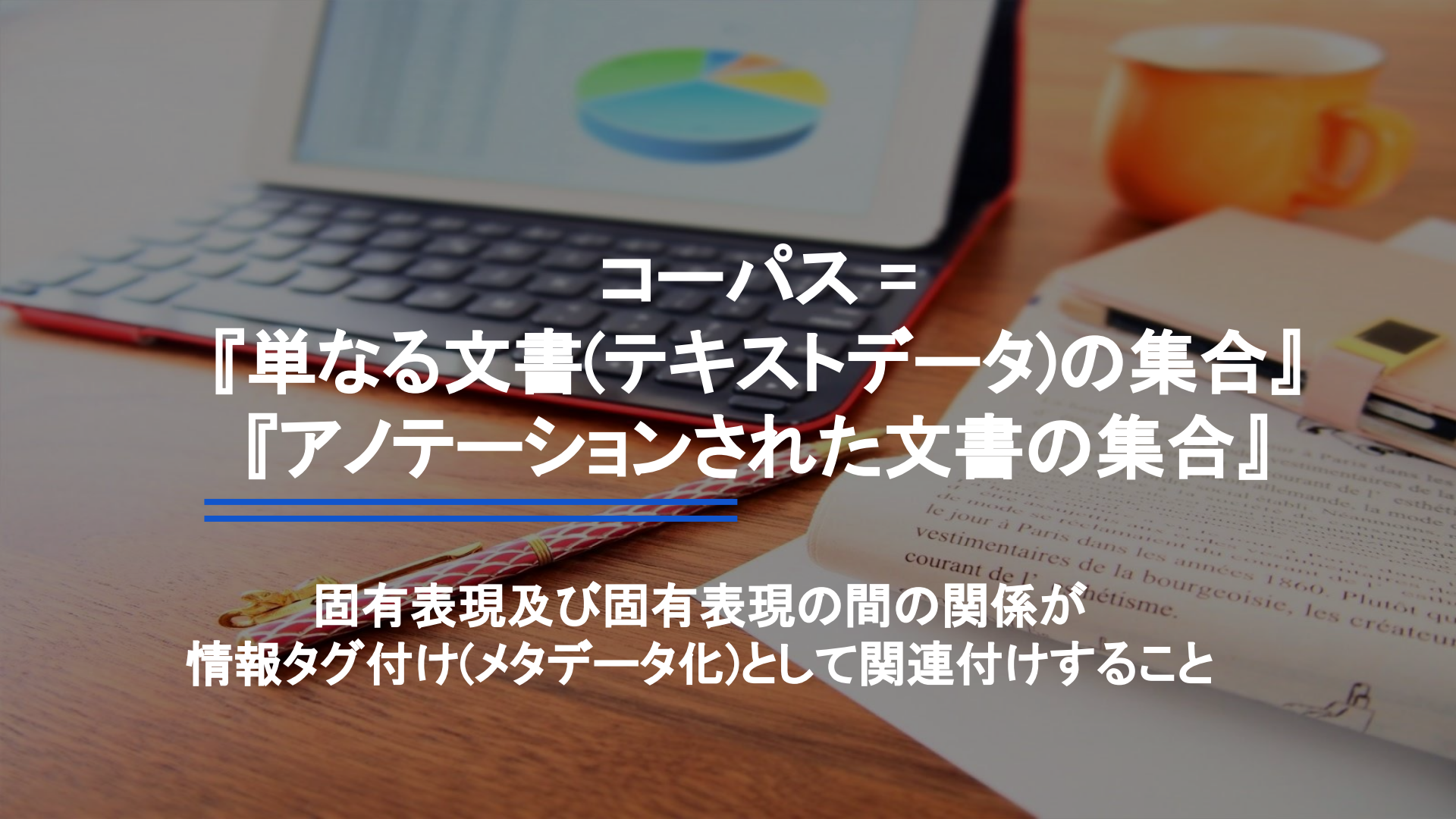
# 具体的には...

- ルールベースによる関係抽出
- 機械学習による分類に基づく関係抽出
- 構造予測に基づく関係抽出
- ニューラルネットによる関係抽出


情報抽出のための
コーパス作成

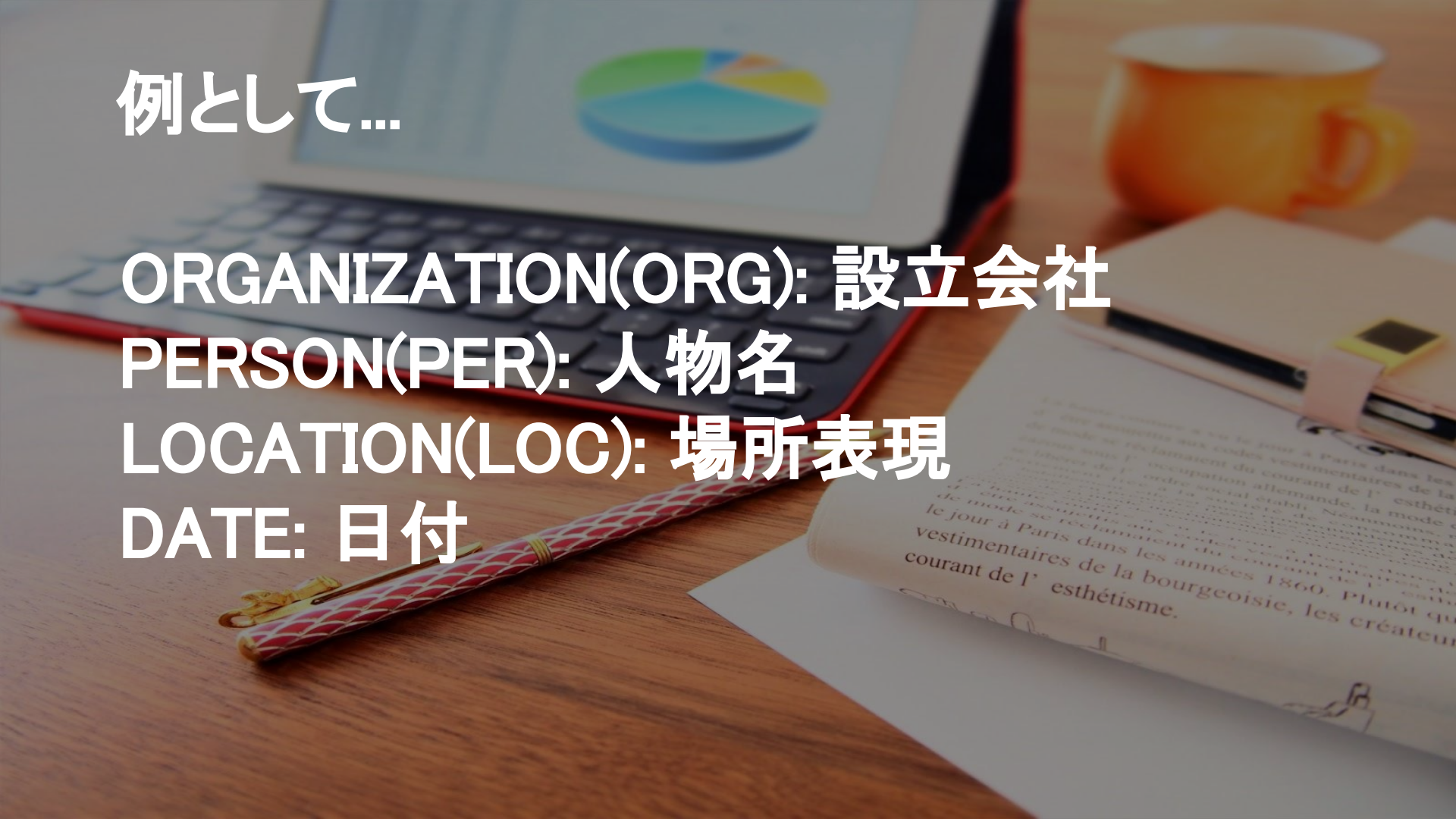

コーパス作成は、
情報抽出器の開発または
カスタマイズと評価で利用するための
データの準備として行われます


コーパス＝
『単なる文書(テキストデータ)の集合』
『アノテーションされた文書の集合』



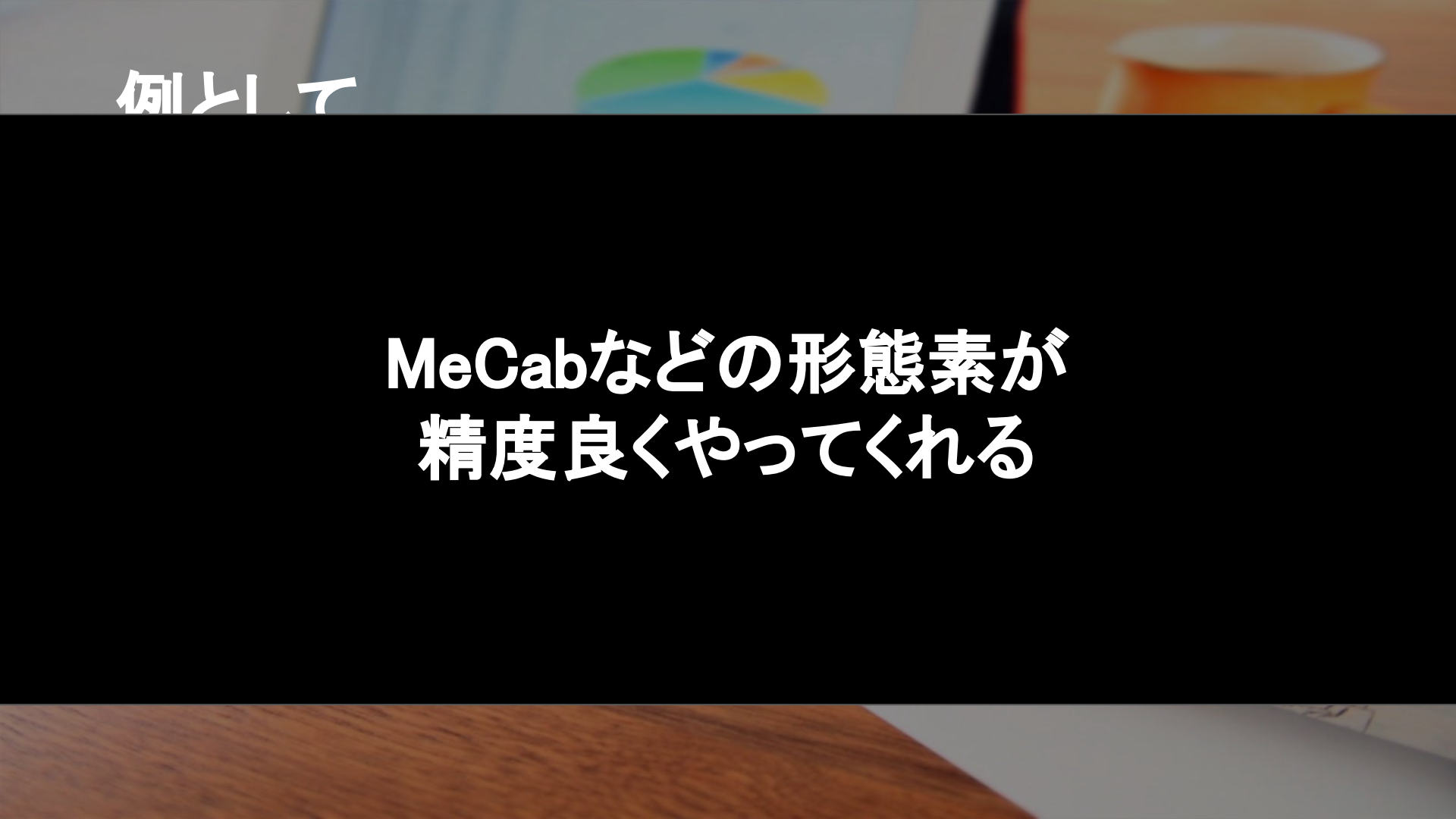
# コーパス＝
# 『単なる文書(テキストデータ)の集合』
# 『アノテーションされた文書の集合』

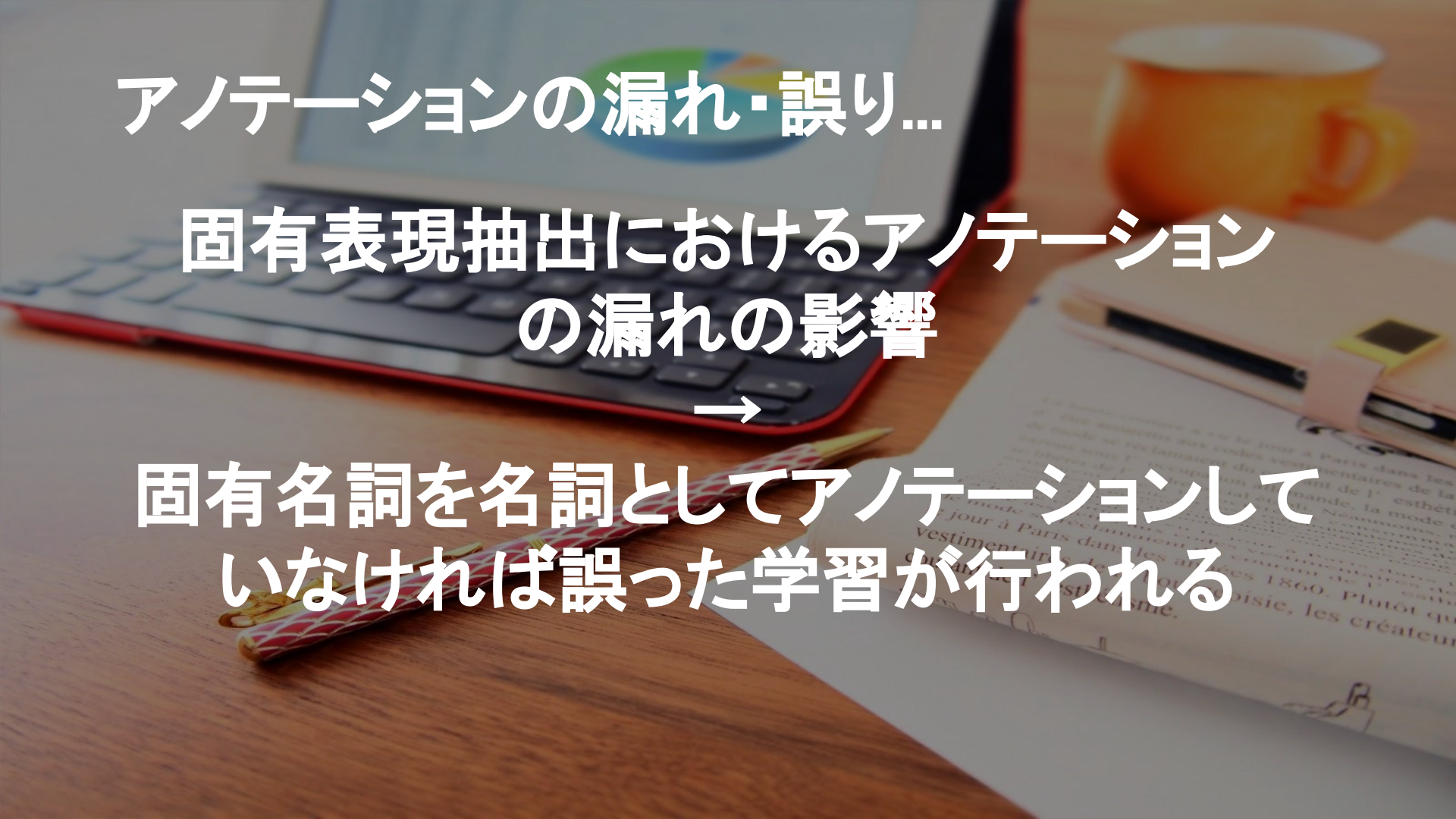
固有表現及び固有表現の間の関係が
情報タグ付け(メタデータ化)として関連付けすること



# 例として...

ORGANIZATION(ORG): 設立会社
PERSON(PER): 人物名
LOCATION(LOC): 場所表現
DATE: 日付

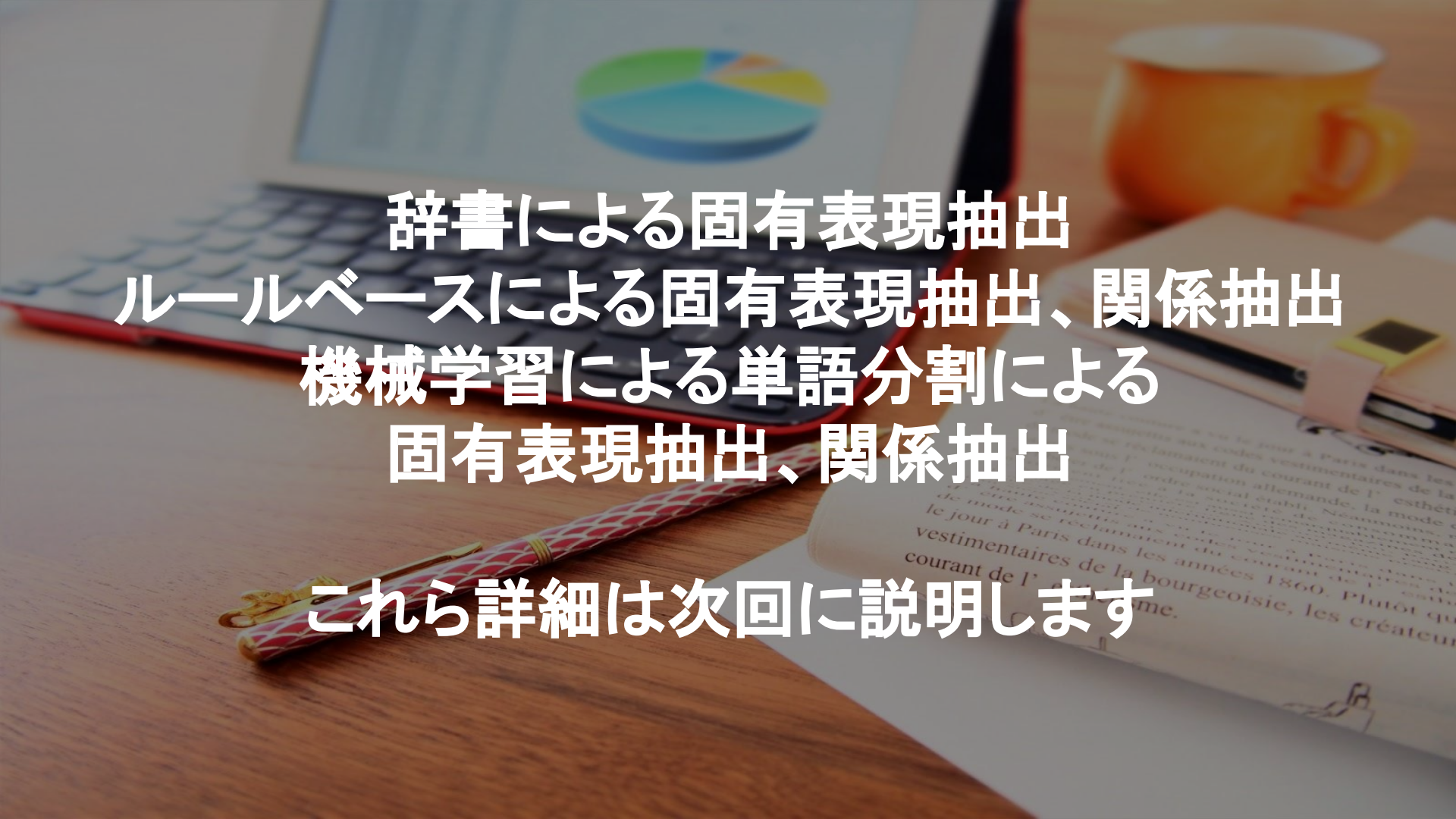

例として
MeCabなどの形態素が
精度良くやってくれる



# アノテーションの漏れ・誤り…

固有表現抽出におけるアノテーション
の漏れの影響
→
固有名詞を名詞としてアノテーションして
いなければ誤った学習が行われる

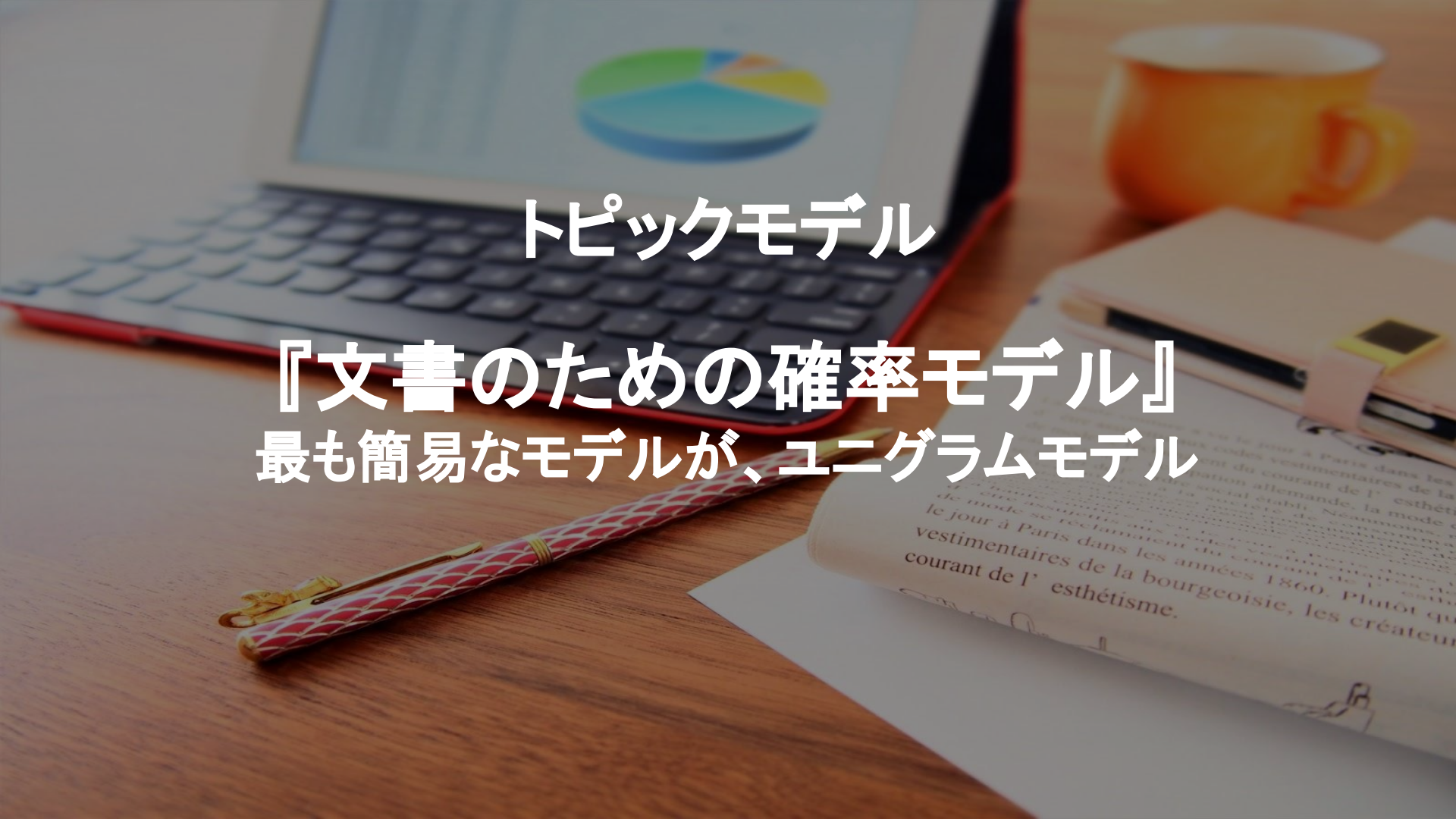
# 辞書による固有表現抽出



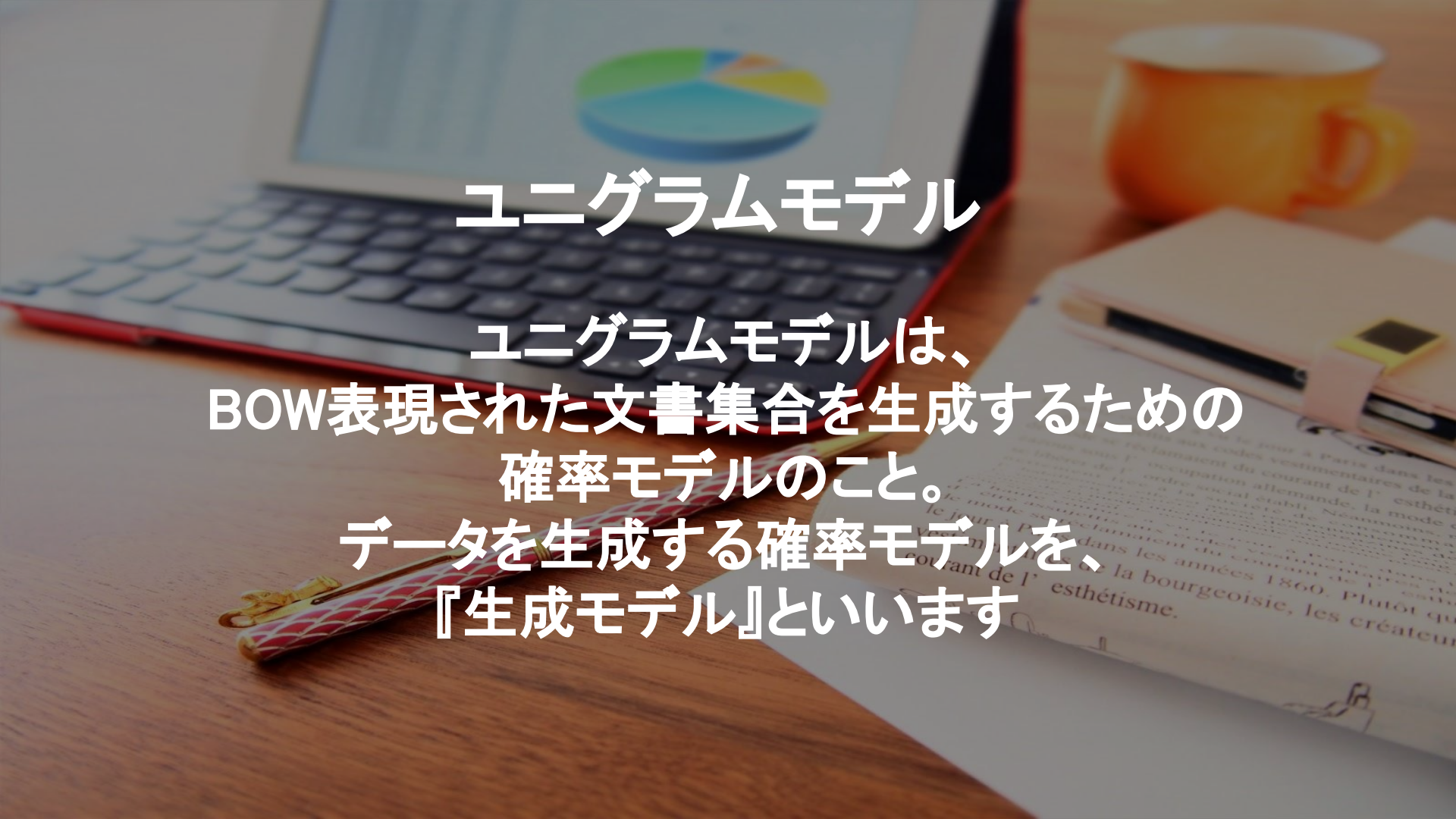
辞書による固有表現抽出
ルールベースによる固有表現抽出、関係抽出
機械学習による単語分割による
固有表現抽出、関係抽出

これら詳細は次回に説明します

# トピックモデル概論

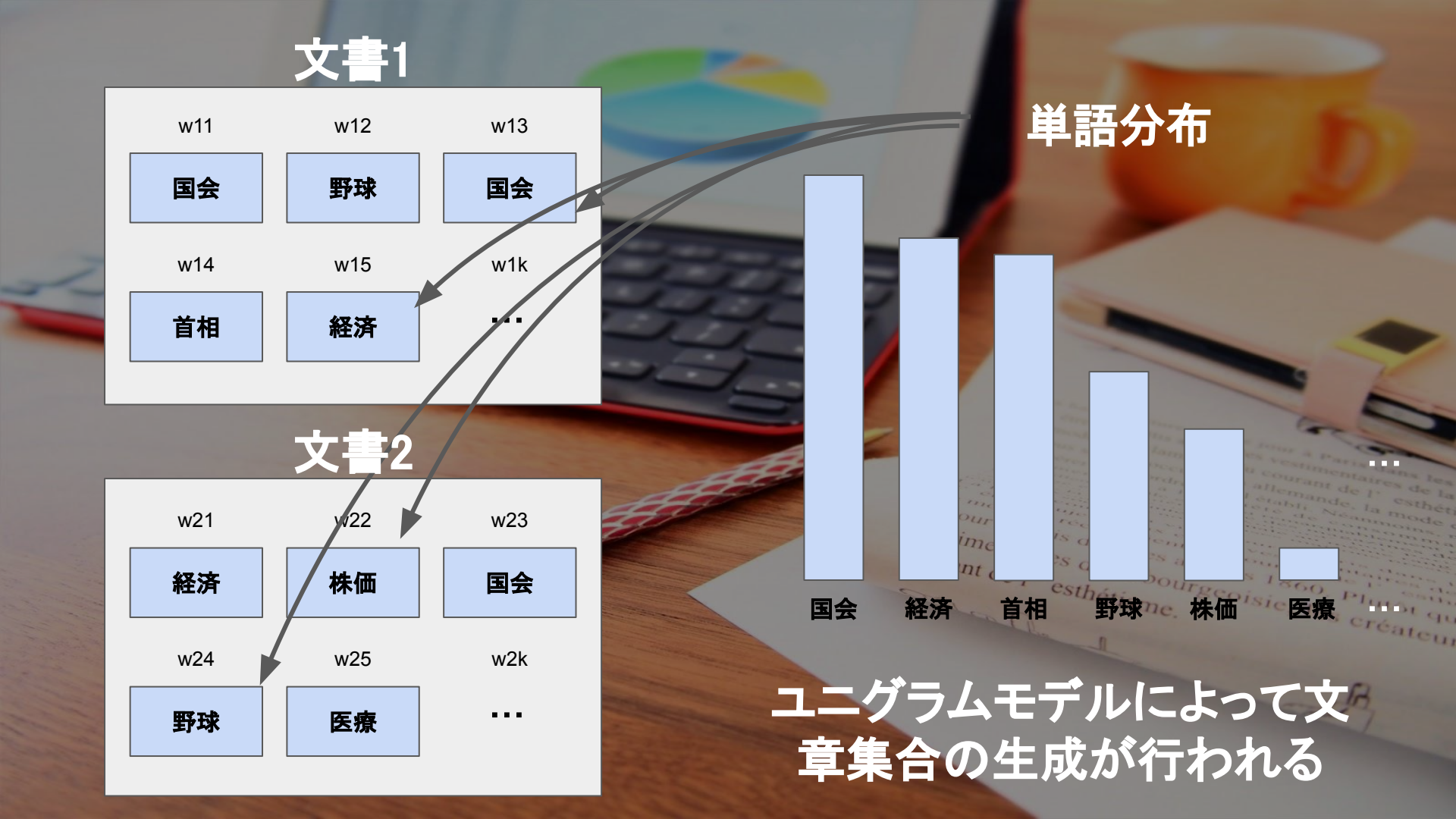

トピックモデル
『文書のための確率モデル』
最も簡易なモデルが、ユニグラムモデル

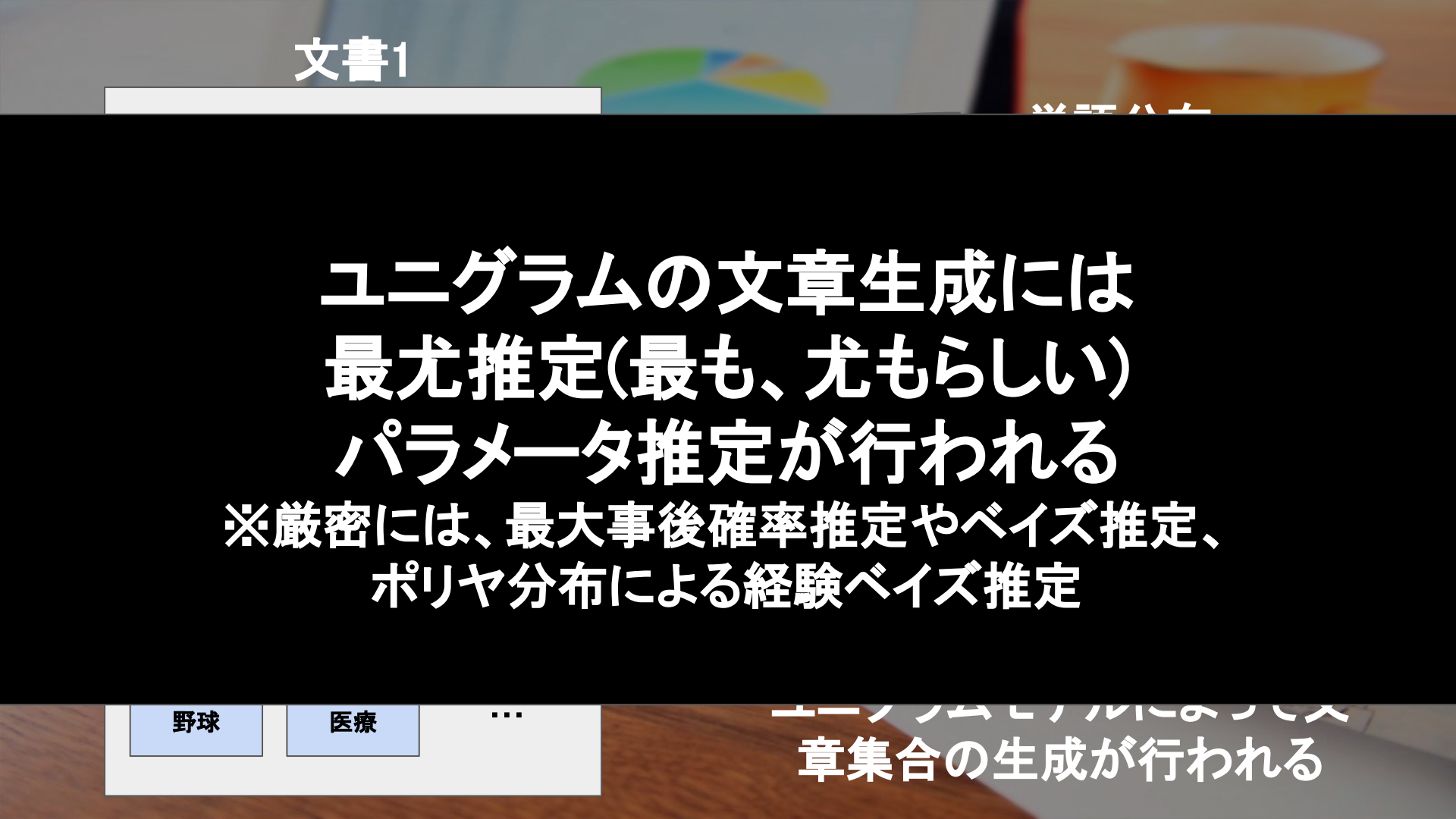

文章集合のBOW(Bag Of Words)表現
文章1は『トピックモデルは文章の確率モデルとして提案する…』
文章2は『英語と数学は科目であり、英語の場合、学習モデルに…』
このとき、ストップワードとして助詞などを除いている



# ユニグラムモデル

ユニグラムモデルは、
BOW表現された文書集合を生成するための
確率モデルのこと。
データを生成する確率モデルを、
『生成モデル』といいます

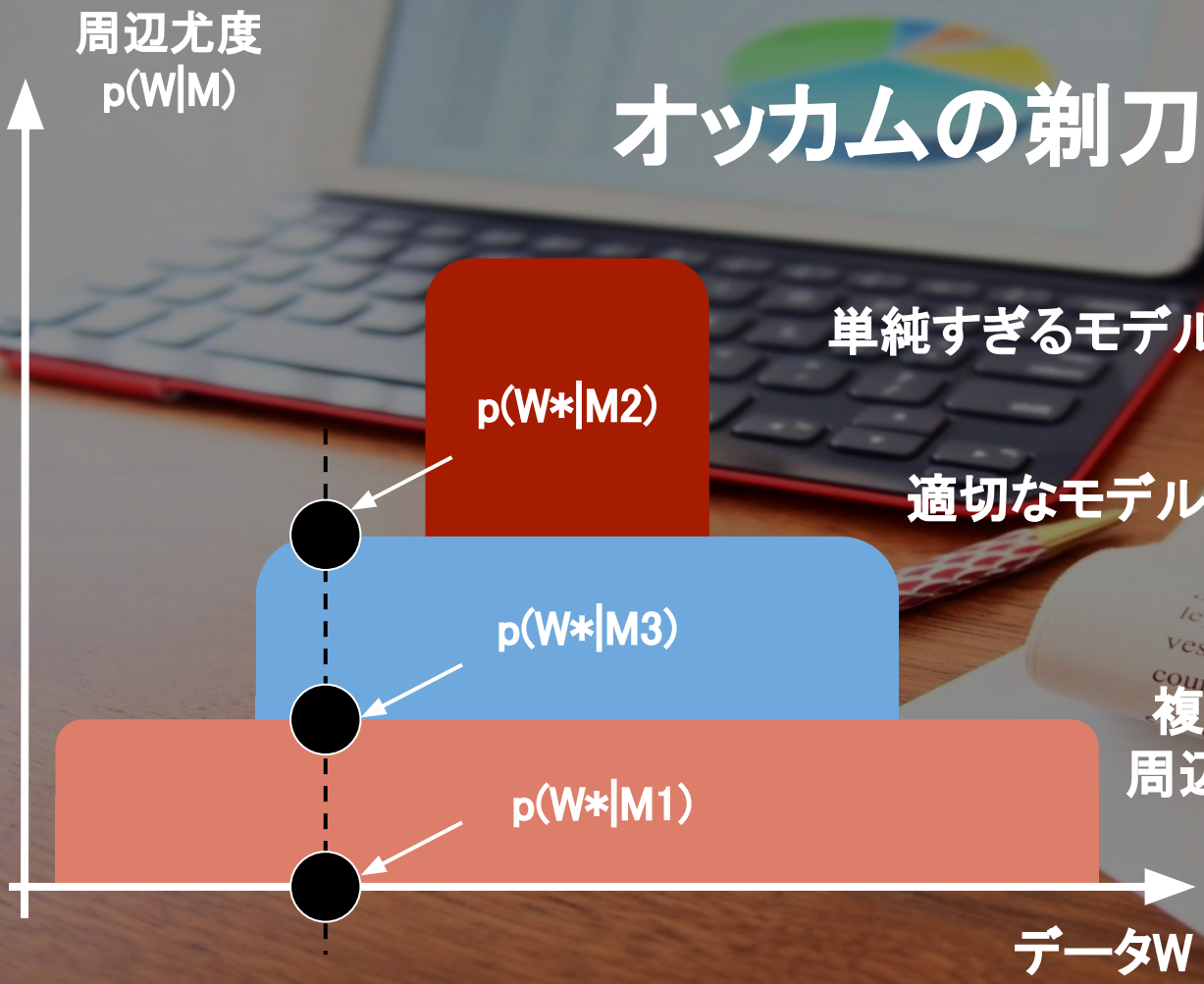

文書1
w11
国会
w12
野球
w13
国会
w14
首相
w15
経済
w1k
…
文書2
w21
経済
w22
株価
w23
国会
w24
野球
w25
医療
w2k
…
単語分布
国会
経済
首相
野球
株価
医療
…
ユニグラムモデルによって文章集合の生成が行われる




# ユニグラムの文章生成には最尤推定(最も、尤もらしい)パラメータ推定が行われる

※厳密には、最大事後確率推定やベイズ推定、ポリヤ分布による経験ベイズ推定

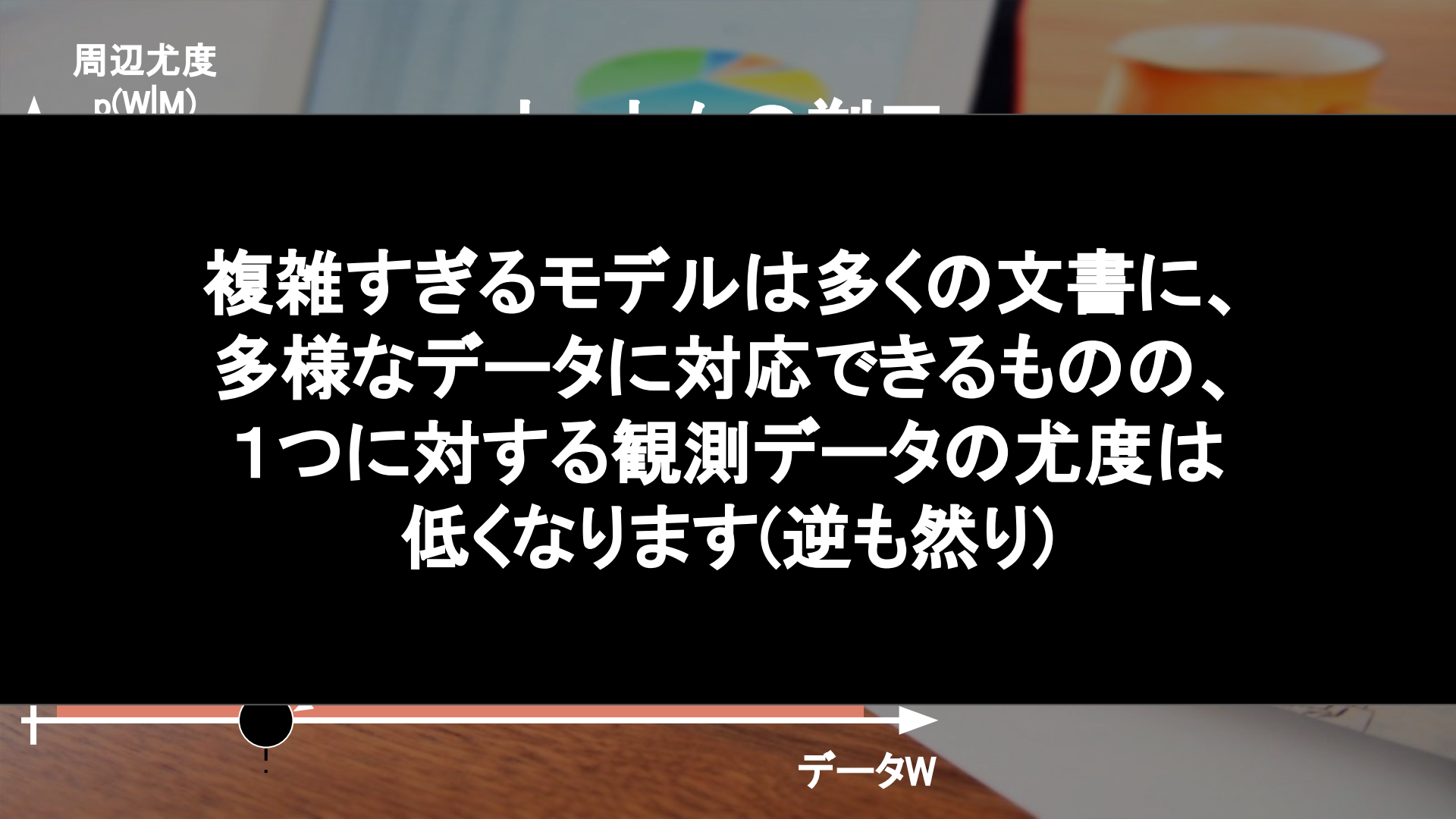
# オッカムの剃刀



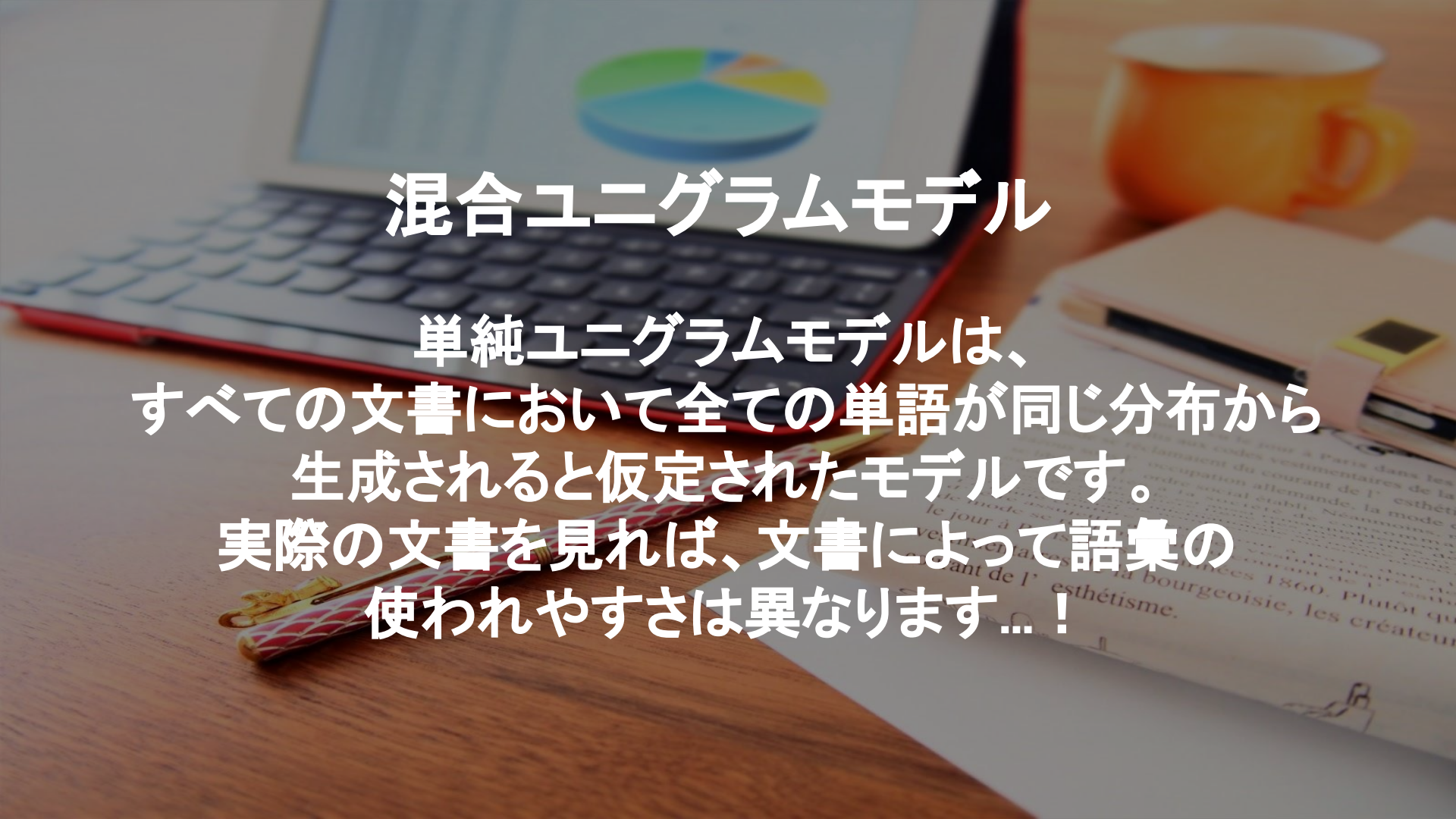

単純すぎるモデルの周辺尤度関数 p(W|M1)

適切なモデルの周辺尤度関数 p(W|M2)

複雑すぎるモデルの
周辺尤度関数 p(W|M3)

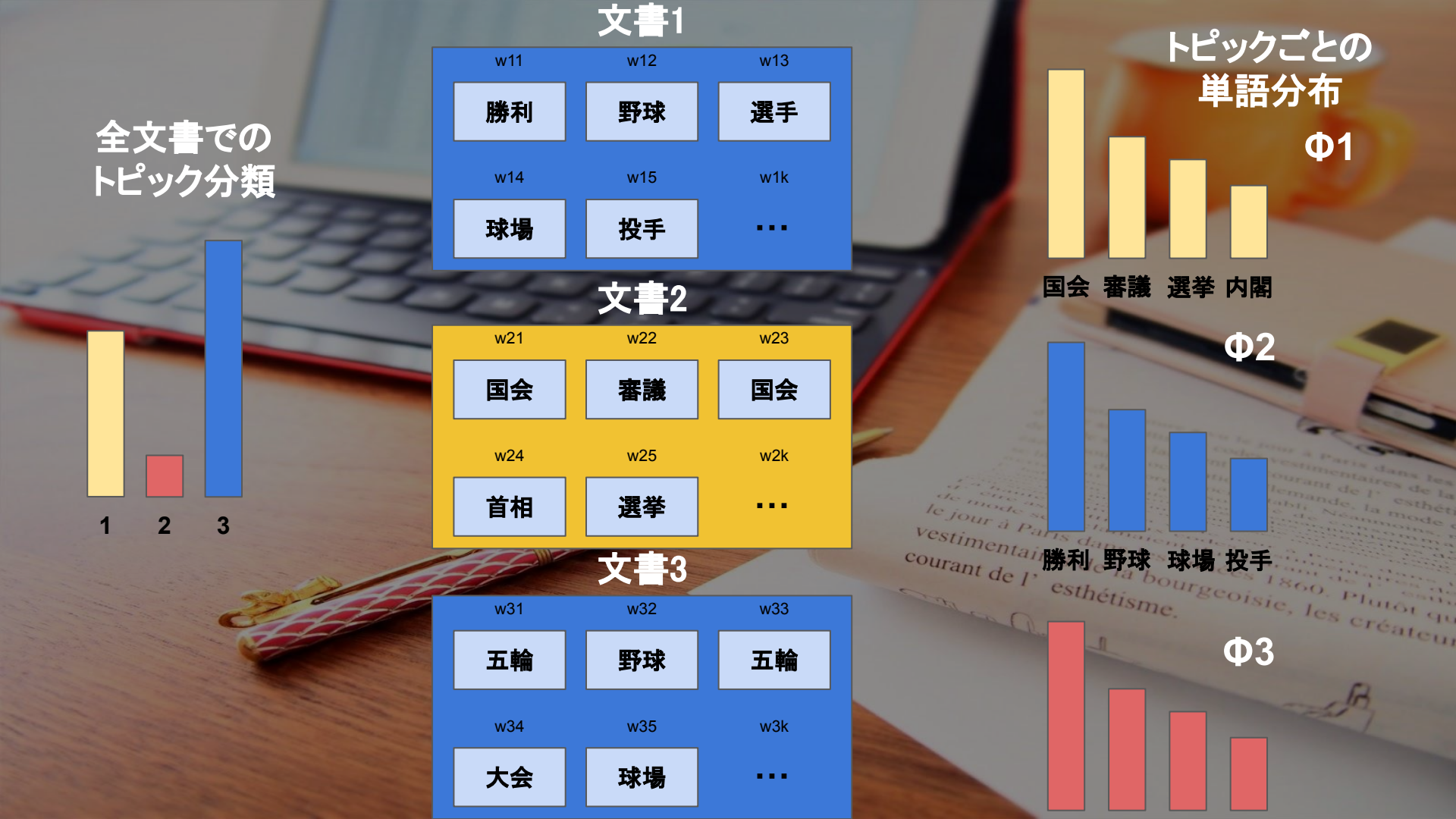


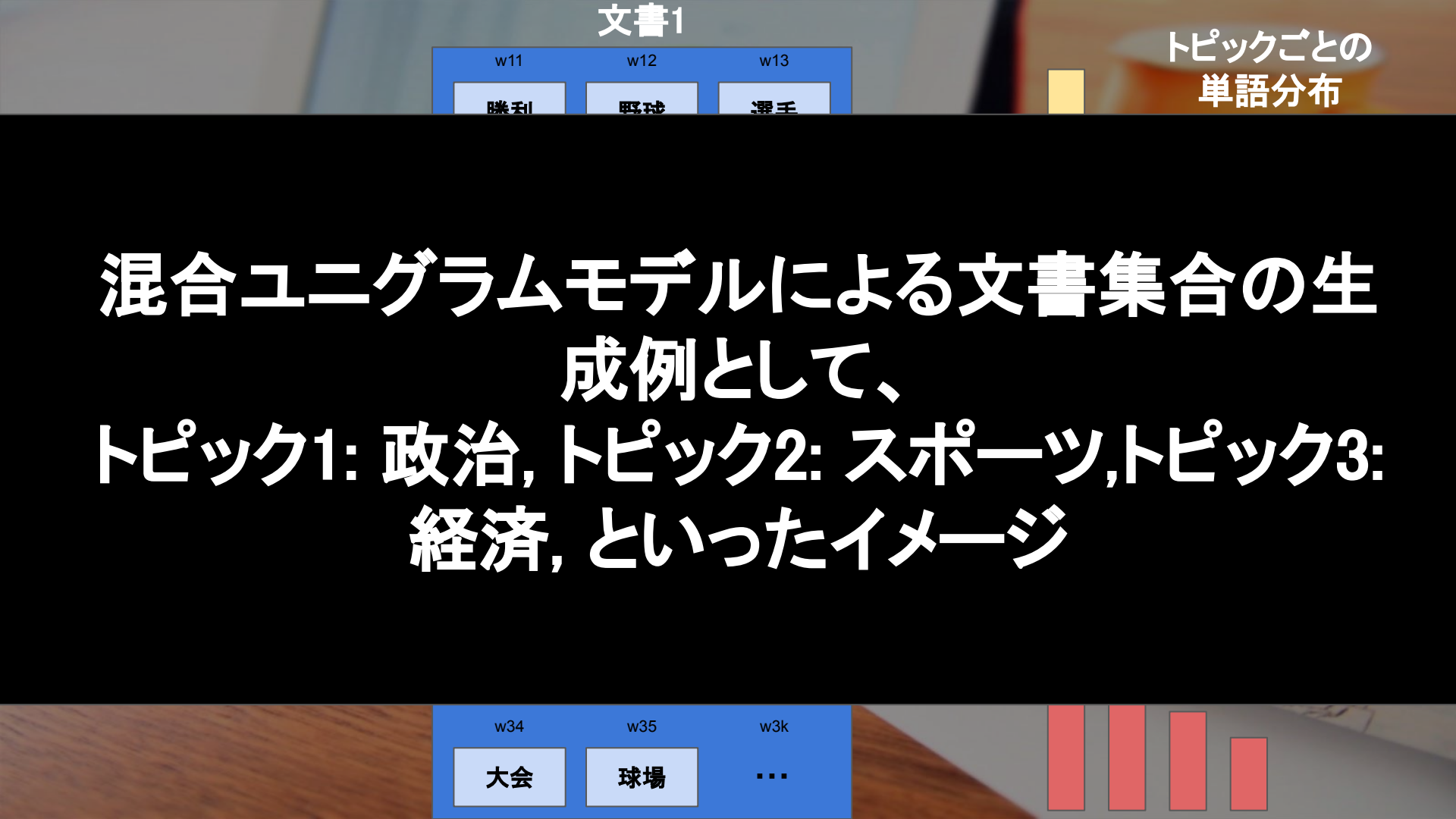

複雑すぎるモデルは多くの文書に、
多様なデータに対応できるものの、
1つに対する観測データの尤度は
低くなります(逆も然り)

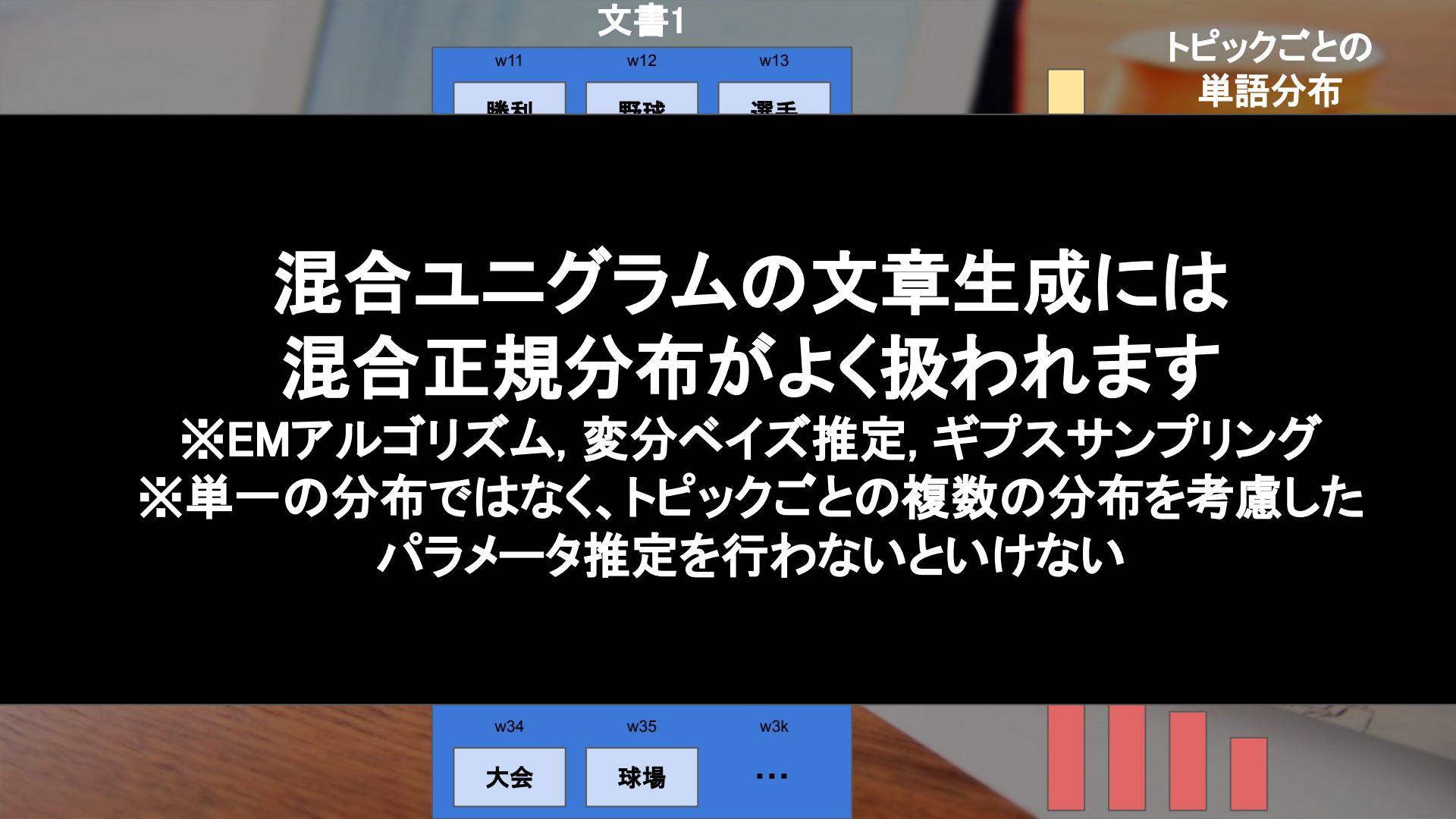


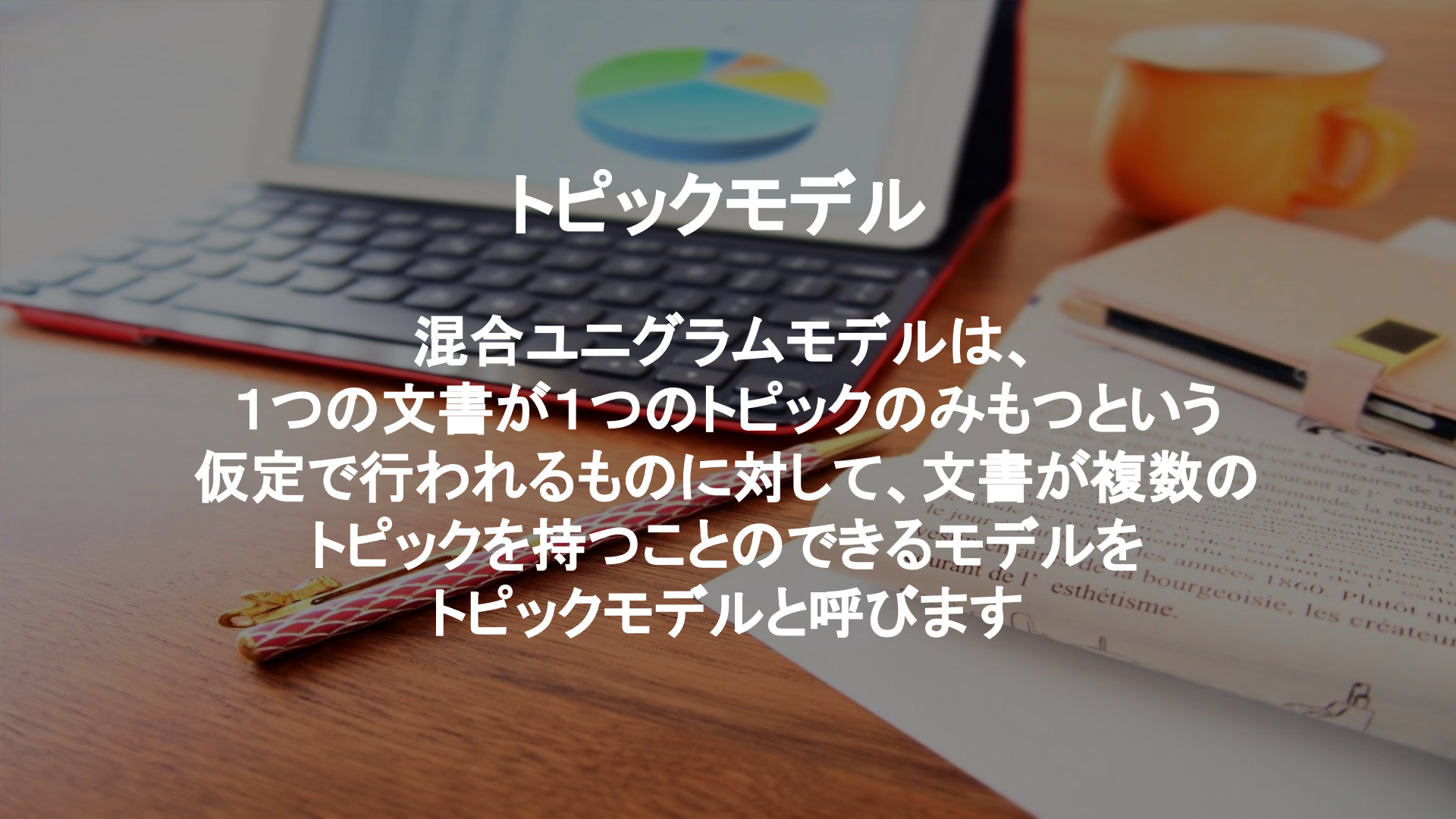
# 混合ユニグラムモデル

単純ユニグラムモデルは、
すべての文書において全ての単語が同じ分布から
生成されると仮定されたモデルです。
実際の文書を見れば、文書によって語彙の
使われやすさは異なります…！


全文書での
トピック分類
1
2
3
文書1
w11
勝利
w12
野球
w13
選手
w14
球場
w15
投手
w1k
・・・
文書2
w21
国会
w22
審議
w23
国会
w24
首相
w25
選挙
w2k
・・・
文書3
w31
五輪
w32
野球
w33
五輪
w34
大会
w35
球場
w3k
・・・
トピックごとの
単語分布
Φ1
国会
審議
選挙
内閣
Φ2
勝利
野球
球場
投手
Φ3

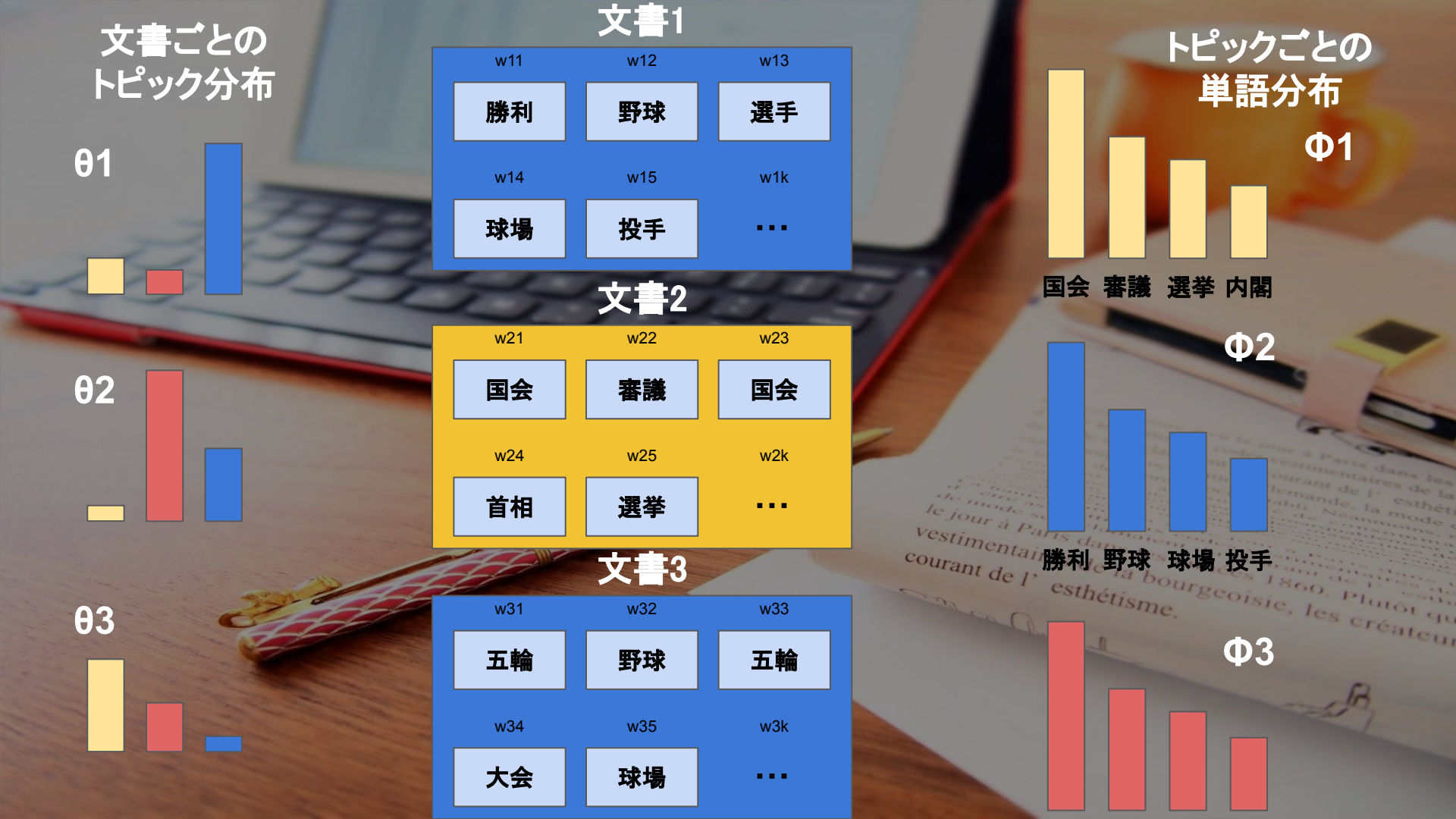

文書1
w11
w12
w13
トピックごとの
単語分布
混合ユニグラムモデルによる文書集合の生成例として、
トピック1: 政治, トピック2: スポーツ,トピック3: 経済, といったイメージ
w34
w35
w3k
大会
球場
・・・


文書1
w11
w12
w13
トピックごとの
単語分布
混合ユニグラムの文章生成には
混合正規分布がよく扱われます
※EMアルゴリズム，変分ベイズ推定，ギプスサンプリング
※単一の分布ではなく、トピックごとの複数の分布を考慮した
パラメータ推定を行わないといけない
w34
w35
w3k
大会
球場
・・・



# トピックモデル

混合ユニグラムモデルは、
1つの文書が1つのトピックのみもつという
仮定で行われるものに対して、文書が複数の
トピックを持つことのできるモデルを
トピックモデルと呼びます


文書ごとの
トピック分布
θ1
θ2
θ3
文書1
w11
w12
w13
勝利
野球
選手
w14
w15
w1k
球場
投手
...
文書2
w21
w22
w23
国会
審議
国会
w24
w25
w2k
首相
選挙
...
文書3
w31
w32
w33
五輪
野球
五輪
w34
w35
w3k
大会
球場
...
トピックごとの
単語分布
Φ1
国会
審議
選挙
内閣
Φ2
勝利
野球
球場
投手
Φ3

文書ごとの
トピック分布

文書1


トピックごとの
単語分布

# トピックモデルそのものは
# 次回以降、紹介いたします