データ前処理ナレッジ共有コミュニティ

BigGorilla

誰？

- 三宅 智久
- 株式会社リクルート
  R&D イノベーティブテクノロジー研究戦略室
  Megagonプロジェクトグループ 所属
- もともと**Webエンジニア**、後にPM
- Contact
  - miyaketomohisa[at]megagon.ai
  - facebook.com/miyaketomohisa

誰？

- 三宅 智久
- 株式会社リクルート
  R&D イノベーティブテクノロジー研究戦略室
  Megagonプロジェクトグループ 所属
- もともと**Webエンジニア**、後にPM
- Contact
  - miyaketomohisa[at]megagon.ai
  - facebook.com/miyaketomohisa

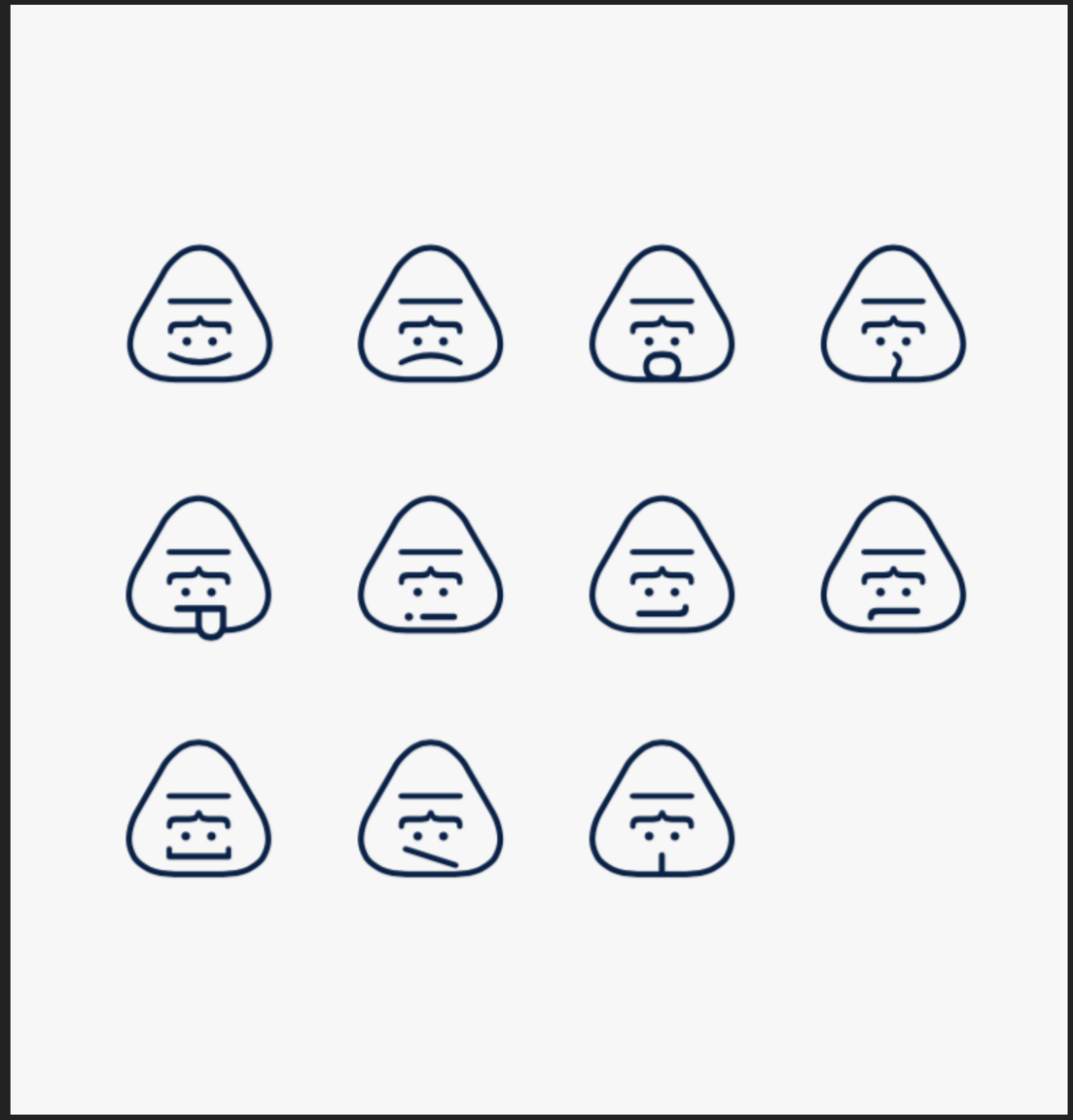
Megagon?

Megagon Labs

- 正式名称 Megagon Labs
  http://www.megagon.ai/
- 旧Recruit Institute of Technology
  アメリカ、シリコンバレーに拠点を構えるリクルートグループの先端技術研究所
- 研究所所長は元GoogleのAlon Halevy
- 自然言語処理、データベースなどのスペシャリストが所属

Agenda

1. **データ前処理の課題**
2. BigGorillaとは？
3. 主要なライブラリ
4. 包括的なチュートリアル



多岐にわたったタスクが存在し

閉じられていて

不可避で

工数が大きく

決定的な

データ前処理

Agenda

1. データ前処理の課題
2. **BigGorillaとは？**
3. 主要なライブラリ
4. 包括的なチュートリアル

BigGorillaとは？

BigGorilla.org

- Megagon Lab.とウィスコンシン大学によりスタート
- データ前処理にまつわる問題を解決するためのWEBコミュニティ＝**集合知を作り上げる**
- 4つのカテゴリからなるオープンソースPythonライブラリデータベース
- チュートリアル
- 相談できるフォーラム
- **誰でも投稿可能**

データ収集
データ抽出
データ洗浄

エンティティ
マッチング

スキーマ
マッチング＆
マージング

その他
ワークフロー管理
など

なぜ「BigGorilla」なのか…

データキュレーションにおける問題は

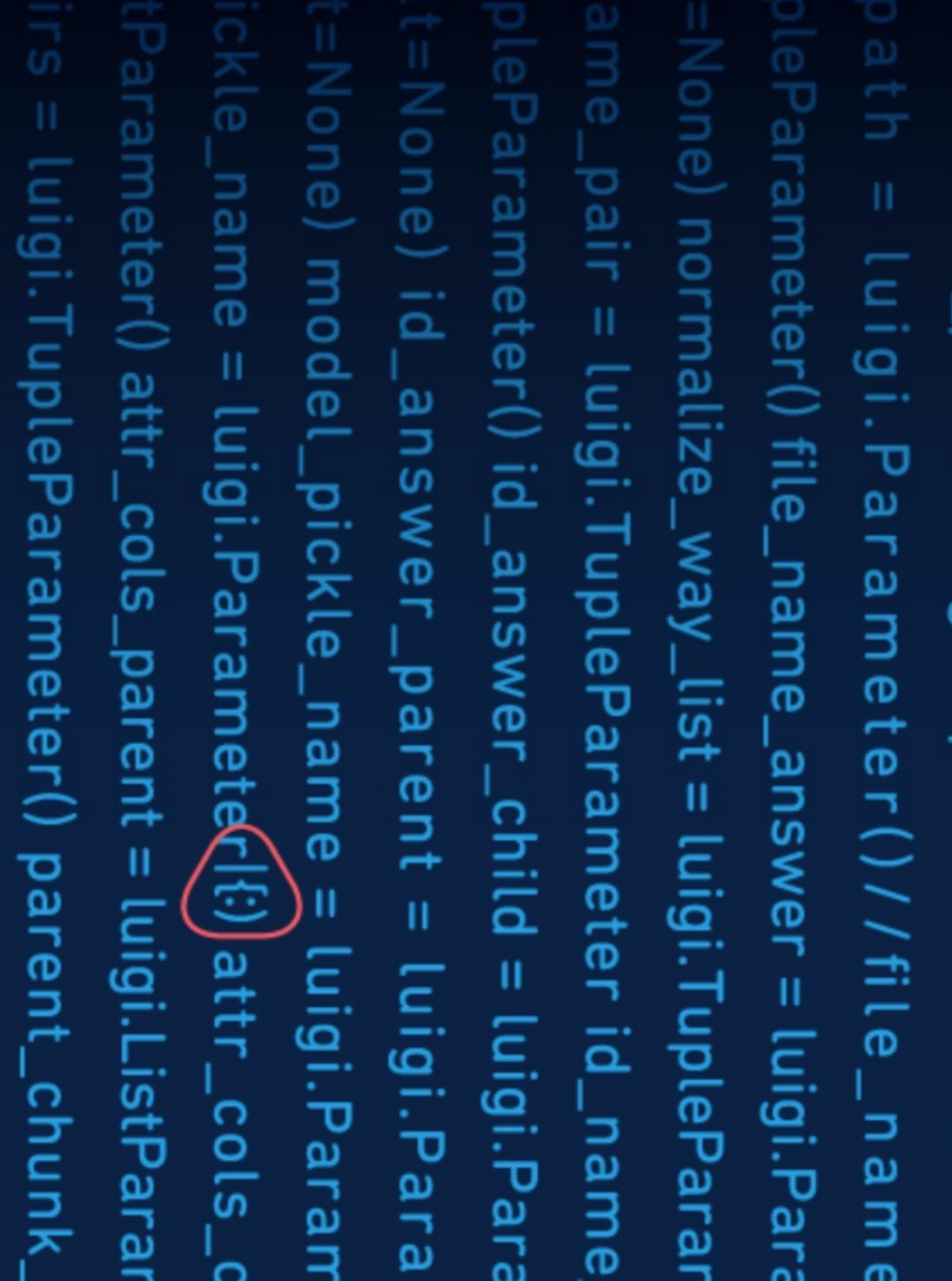
"800ポンド(362kg)のゴリラ"

https://www.youtube.com/watch?v=7I2uQei_j_4

Michael Stonebraker
(Postgres, MIT, Cambrigde)

Agenda

1. データ前処理の課題
2. BigGorillaとは？
3. **主要なライブラリ**
4. 包括的なチュートリアル



INNOVATIVE TOOLS

主要ライブラリ

Magellan

- https://sites.google.com/site/anhaidgroup/projects/magellan
- ウィスコンシン大学 **Prof.AnHai**と**WalmartLabs**、**Megagon Lab.** らによって開発。
- 以下3つによって構成
  - **py_entitymatching: 複数フィーチャ（カラム）類似度を計算しテーブルをジョイン**
  - py_stringsimjoin: 2つデータセットを類似度に基づきペアリング
  - py_stringmatching: 文字列トークナイザと類似度計算

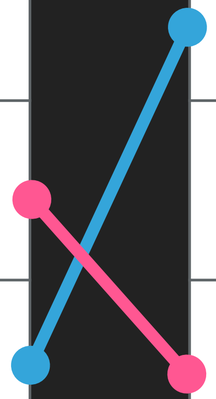
Magellan

テーブル A

year	movie	leading_actor
2001.8.2	Lord of the Rings	E.Wood
2008.6.27	Wall E	B.Burtt
1989.3.1	Back to the Future Part 2	M.J.Fox

テーブル B

year	movie	leading_actor
1989	Back to the Future II	Micheal J Fox
2010	Inception	Leonardo DiCaprio
2008	Wall-E	Ben Burtt



FlexMatcher

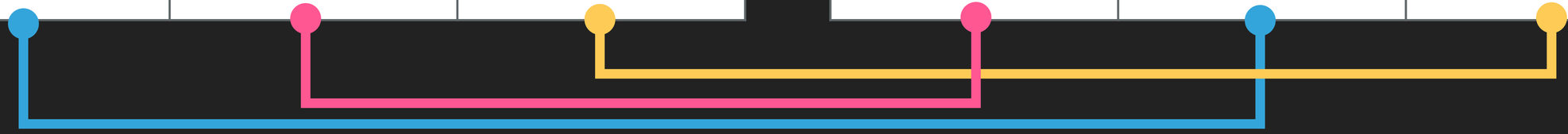
- https://pypi.org/project/flexmatcher/
- 複数のテーブルスキーマ結合を機械学習を用いて行なう

テーブル A

year	movie	imdb_rating
2001	Lord of the Rings	8.8
2010	Inception	8.7

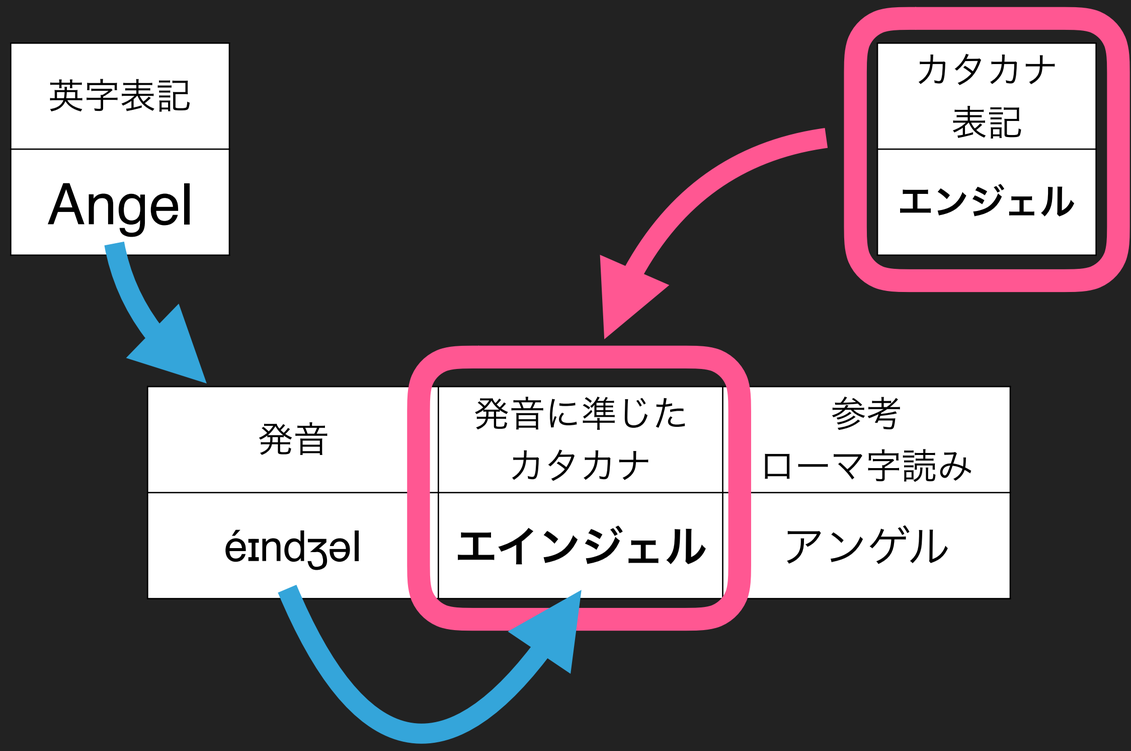
テーブル B

Title	Produced	movie_rating
The Godfather	1972	9.2
The Matrix	1999	8.7



En2Kana

- 英単語とカタカナをマッチングさせるための辞書と変換ライブラリ。
- 英単語を発音記号に変換（翻字）することで、カタカナと精度高くマッチングさせることが可能。
- 元言語と発音記号の辞書があれば、英語以外の言語も対応化。
- 活用例：
  - データ名寄せ、ゆらぎ解消
  - 検索エンジン「ゼロ件ヒット」問題
- オープンソース化に向けて準備中。



KOKO

- https://github.com/biggorilla-gh/koko
- 平叙文から、ルーズな条件文で文字列を取り出すことができるエクストラクションツール

- Strict left context:

```
extract "Ents" x from "doc.txt" if
      ("introduce" x)
```

- Loose left context:

```
extract "Ents" x from "doc.txt" if
      ("introduce" near x)
```

- Semantic right context:

```
extract "Ents" x from "doc.txt" if
      (x ~ "serves coffee")
```

- Strict right context:

```
extract "Ents" x from "doc.txt" if
      (x ", a cafe")
```

- Loose right context:

```
extract "Ents" x from "doc.txt" if
      (x near "cafe")
```

- Semantic left context:

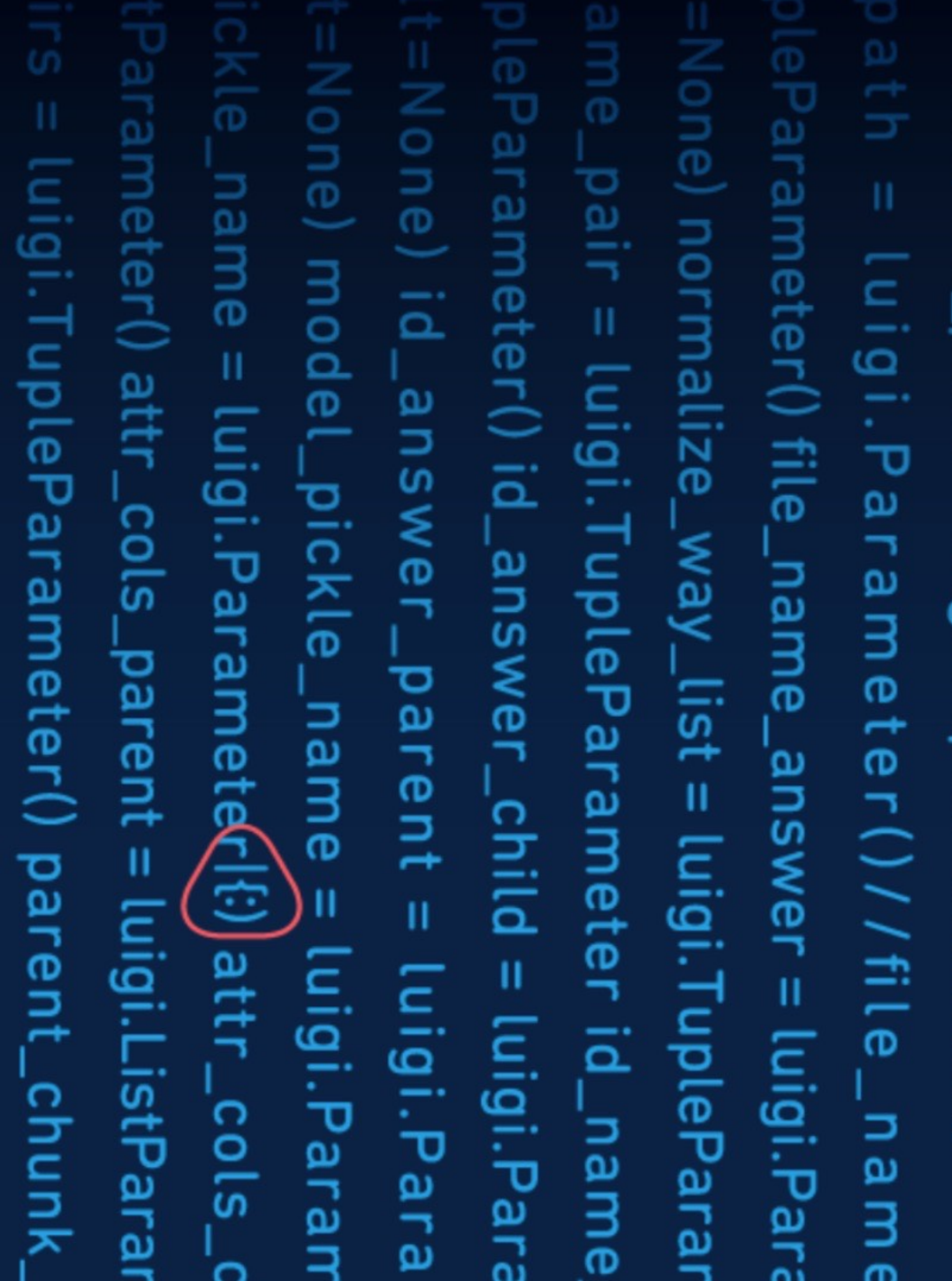
```
extract "Ents" x from "doc.txt" if
      ("introducing cafe" ~ x)
```

Usagi

- https://github.com/biggorilla-gh/usagi
- 複数データベースのメタデータを横断検索できるエンタープライズ向けデータディスカバリーシステム。リクルート内では「Meta-Looking（メタルキング）」として稼働。

Agenda

1. データ前処理の課題
2. BigGorillaとは？
3. 主要なライブラリ
4. **包括的なチュートリアル**

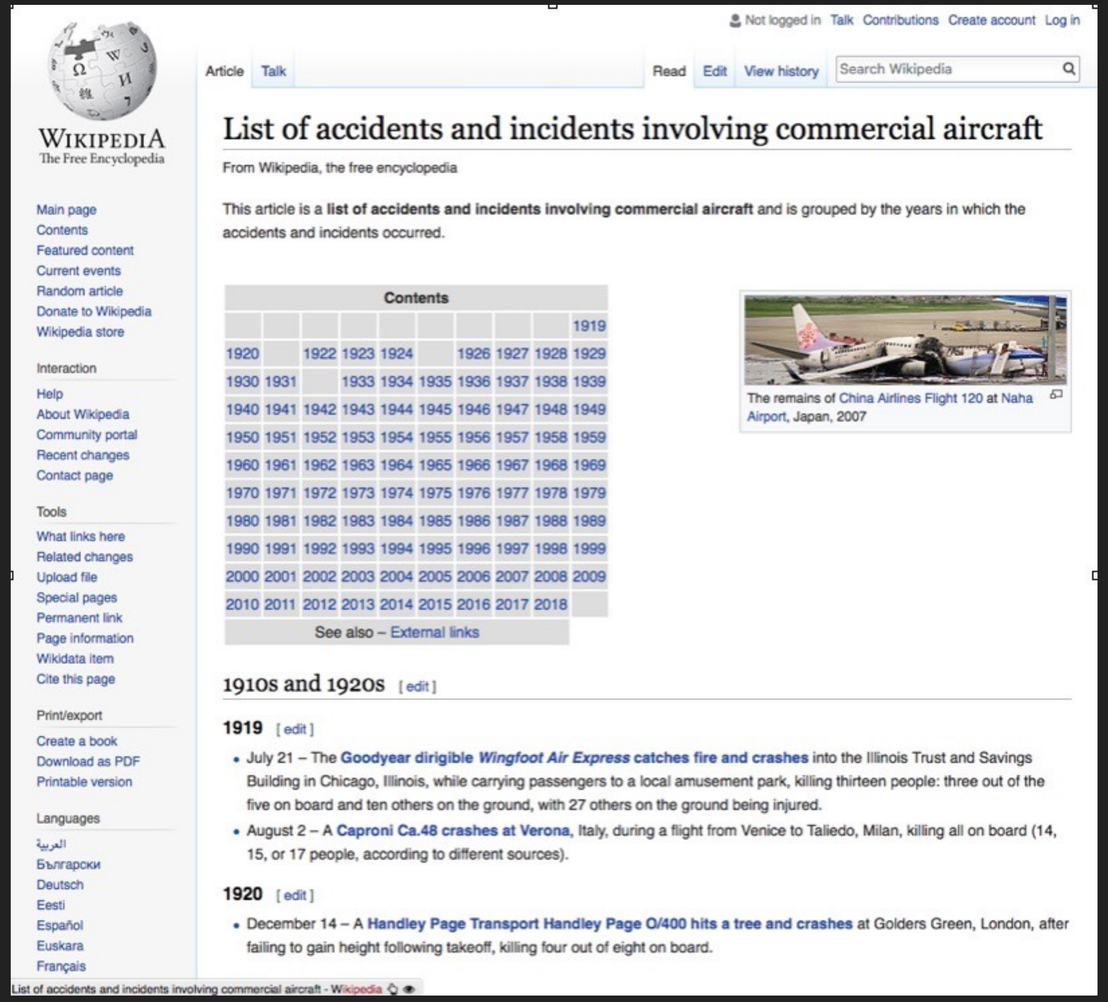


INTRODUCTORY TUTORIAL

チュートリアル

包括的チュートリアル

- https://www.biggorilla.org/data-preparation-using-koko
- Wikipediaの「航空機事故一覧」から htmlをスクレーピング＆データ整形
- 前述のKOKOを使用し、航空会社名を抽出
- 航空会社, 日付, 便名, 事故現場の地名をデータフレーム(構造化データ)に格納



BigGorilla

1930s [edit]

1930 [edit]

- February 10 – An **Air Union Farman F.63 Goliath crashes** during an emergency landing at Marden Airfield, Marden, Kent, England, following failure of the right tailplane, killing two of six on board.
- October 5 – On its maiden voyage from England to British India, the **British civil airship R101 crashes and burns** in Allonne, Oise, France, while flying at low altitude at night in a rainstorm, killing 48 out of 54 on board, the worst civil airship disaster in history.

1931 [edit]

- March 21 – An **Australian National Airways Avro 618 Ten, *Southern Cloud*, disappears** in severe weather on a flight from Sydney to Melbourne, killing all eight on board in Australia's first significant airline disaster; the crash site in the Snowy Mountains remains undiscovered until 1958.
- March 31 – A **Transcontinental & Western Air Fokker F-10 Trimotor crashes** near Bazaar, Kansas, after a wing breaks off in flight, killing all eight aboard, including University of Notre Dame football coach Knute Rockne.

1933 [edit]

- March 28 – The **1933 Imperial Airways Dixmude crash** in Belgium of an Armstrong Whitworth Argosy II is the first suspected case of air sabotage; all 15 on board are killed.
- October 10 – The **United Airlines crash near Chesterton**: a Boeing 247 is destroyed by a bomb over Chesterton, Indiana, United States, in the first proven case of air sabotage on a commercial aircraft; all seven on board are killed.
- December 30 – In the **1933 Imperial Airways Ruysselede crash** in Belgium, an Avro Ten collides with a radio mast, killing all 10 on board.

1934 [edit]

- February 23 - A **United Air Lines Boeing 247** crashes into a Utah canyon in bad weather, killing all eight on board.
- May 9 – An **Air France Wibault 282T crashes** into the English Channel off Dungeness, Kent, killing all six on board.
- July 27 – A **Swissair Curtiss T-32 Condor II crashes** near Tuttlingen, Germany, after a wing separates in a thunderstorm, killing all 12 passengers and crew on board.
- October 2 – A **Hillman's Airways de Havilland Dragon Rapide crashes** into the English Channel off Folkestone, Kent, due to pilot error, killing all seven on board.

1935 [edit]

- May 6 - **TWA Flight 6**, a Douglas DC-2 flying from Albuquerque, New Mexico, to Kansas City, Missouri, flying at low altitude through poor visibility at night while desperately low on fuel, crashes near Atlanta, Missouri, killing five out of thirteen on board, including a U.S. Senator. The political aftermath transforms U.S. civil air regulation.
- October 7 – **United Airlines Trip 4**, a Boeing 247D, crashes near Silver Crown, Wyoming, United States, due to pilot error; all 12 on board die.

https://en.wikipedia.org/wiki/List_of_accidents_and_incidents_involving_commercial_aircraft

	Airline	Date	Flight	Location
0	the Illinois Trust and Savings Building	July 21, 1919,		ChicagoIllinois\n
1	Caproni Ca.48	August 2, 1919,		VeronaItalyVeniceTaliedoMilan\n
2	Golders Green	December 14, 1920		London\n
3	Thieulloy	April 7, 1922,		France\n
4	An Air Union	May 14, 1923		MonsuresSommeFrance\n

• • •

	Airline	Date	Flight	Location
1127	Boeing	January 16, 2017	Flight 6491	BishkekKyrgyzstan\n
1128	South Supreme Airlines	March 20, 2017		South Sudan\n
1129	Boeing	March 28, 2017	Flight 112	Jauja\n
1130	SFO	July 7, 2017		\n
1131	Airbus	September 30, 2017		Canada\n

包括的チュートリアル：続き

- https://www.biggorilla.org/schema-matching-entity-linking
- 前半で作成したテーブルと、
  出処の違う飛行機事故データベース(下記)を前述のFlexmatcherを用いてスキーママッチング。

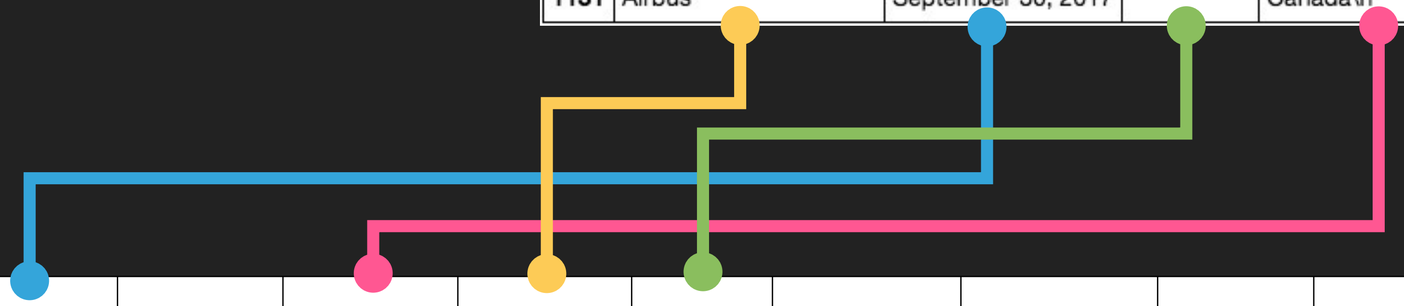
	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/In	Aboard	Fatalities	Ground	Summary
2584	08/28/1972	14:30	Papua, New Guinea	Military - Royal Australian Air Force	NaN	Lae - Port Moresby	de Havilland Canada DHC-4 Caribou	A4-233	NaN	29	25	0	While traveling through a valley, the pilot re...
182	06/24/1929	NaN	St. Paul, Minnesota	Northwest Orient Airlines	NaN	St. Paul - Minneapolis	Ford 5-AT-B Tri-Motor	NC7416	5-AT-002	8	1	0	Crashed near Indian Mounds park shortly after ...
980	12/23/1948	NaN	Near Madrid, Spain	Iberia Airlines	NaN	Madrid - Barcelona	Douglas DC-3 (C-47-DL)	EC-ABK	4256	27	27	0	Crashed into Pandols Mountain while en route.

かなりリッチなデータベース

BigGorilla

スキーママッチング

	Airline	Date	Flight	Location
1127	Boeing	January 16, 2017	Flight 6491	BishkekKyrgyzstan\n
1128	South Supreme Airlines	March 20, 2017		South Sudan\n
1129	Boeing	March 28, 2017	Flight 112	Jauja\n
1130	SFO	July 7, 2017		\n
1131	Airbus	September 30, 2017		Canada\n



	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/ln	Aboard	Fatalities	Ground	Summary
2584	08/28/1972	14:30	Papua, New Guinea	Military - Royal Australian Air Force	NaN	Lae - Port Moresby	de Havilland Canada DHC-4 Caribou	A4-233	NaN	29	25	0	While traveling through a valley, the pilot re...
182	06/24/1929	NaN	St. Paul, Minnesota	Northwest Orient Airlines	NaN	St. Paul - Minneapolis	Ford 5-AT-B Tri-Motor	NC7416	5-AT-002	8	1	0	Crashed near Indian Mounds park shortly after ...
980	12/23/1948	NaN	Near Madrid, Spain	Iberia Airlines	NaN	Madrid - Barcelona	Douglas DC-3 (C-47-DL)	EC-ABK	4256	27	27	0	Crashed into Pandols Mountain while en route.

エンティティマッチング

- スキーママッチングによって発見された「共通の項目」を比較対象に、
  py_entitymatchingを用いてエンティティ（行）を突き合わせる＝**名寄せ**

	_id	l_id	r_id	l_Airline	l_Date	l_Location	l_Flight	r_Operator	r_Flight #	r_Location	r_Fatalities	_sim_score
597	597	269	664	American Eagle	October 31, 1994,	Roselawn, Indiana, Chicago	4184	American Eagle	4184	Roselawn, Indiana	68	0.0
491	491	200	513	Aeroflot	October 11, 1984	Tupolev, Omsk, Russia	3352	Aeroflot	3352	Near Omsk, Russia	174	0.0
405	405	160	444	Aeroflot	January 13, 1977,	Tupolev, Tu-104	3843	Aeroflot	3843	Near Alma Ata, Kazakastan, USSR	96	0.0

おさらい

- データ前処理は、**不可避・膨大・決定的**、かつ**ノウハウが手に入りにくい**。
- BigGorillaは、その**手に入りくかった知識の共有・獲得の場**である。
- 一般的なツールから、**ニッチながら強力な道具**までリストアップされており、その模範的な**組み合わせ方、使い方が参照できる**。
- **誰でも投稿できる**。
- **役立ちそうなツールを「知ってる」「作った」という方はぜひ投稿してください。**

ご清聴
ありがとうございました。