2023.12.14

短期集中研修
『Rデータ解析自由自在（入門編）』

①データ科学の基礎

三村　喬生

統計数理研究所
医療健康データ科学研究センター

データ科学

データ科学

*skei- [インド・ヨーロッパ語根]: 切る, 割る

- *skhizein* [ギリシャ語]: 割る, 裂く, 分ける
- *sceadan* [古英語]: 分割する, 分離する
- *scindera* [ラテン語]: 切る, 引き裂く, 分裂する
- *scire* [ラテン語]: 知る

↓ 名詞

scientia [ラテン語]: 知識, 専門知識

↓

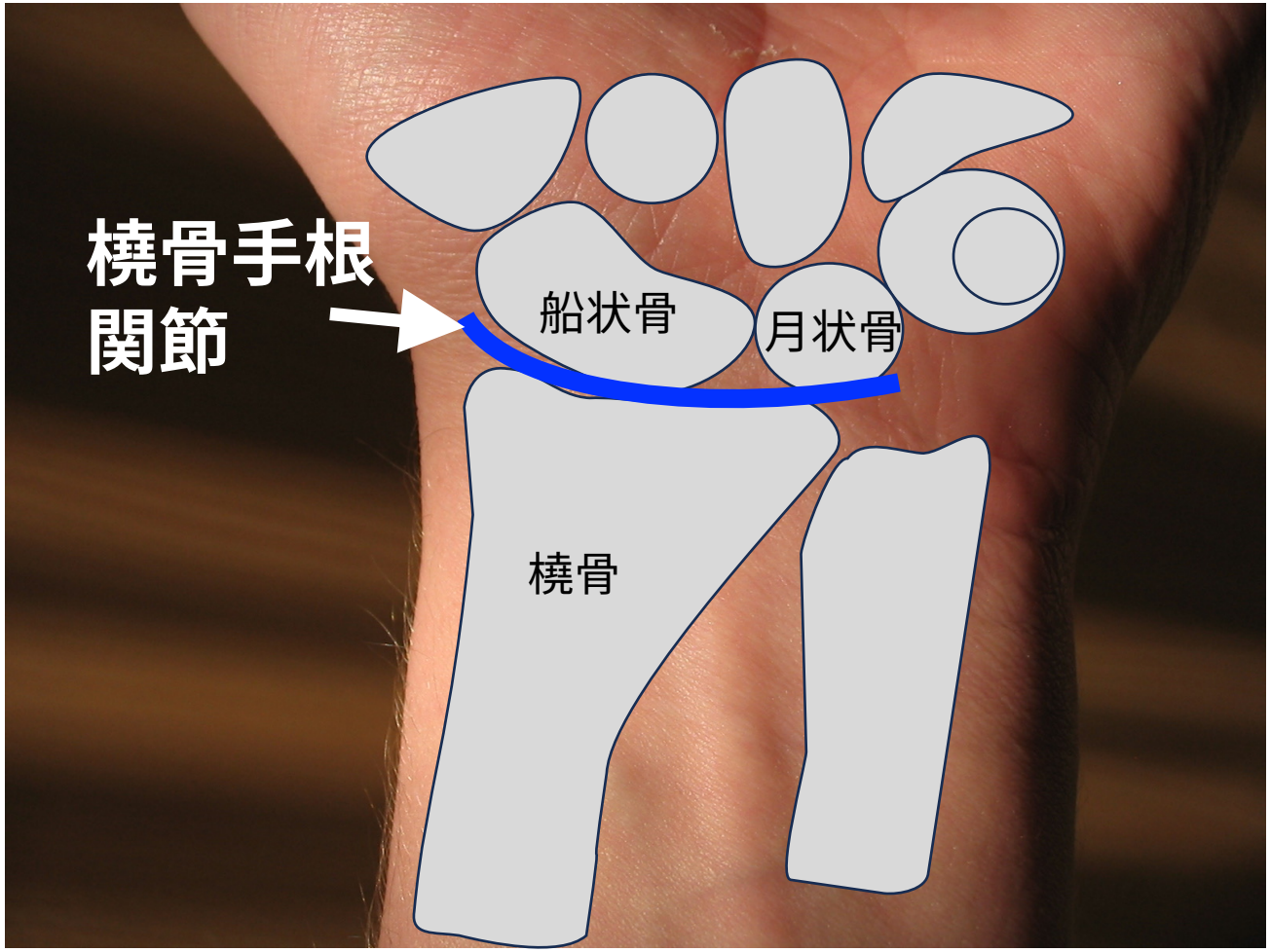
science [英語]

- 14世紀頃に「経験的な知識」「特定の対象群に関する体系化された知識」という現代的な用法が定着。
- 狭義では自然科学を指すが、広く学問全般も含みうる。

cf. Online Etymology Dictionary (etymonline.com)

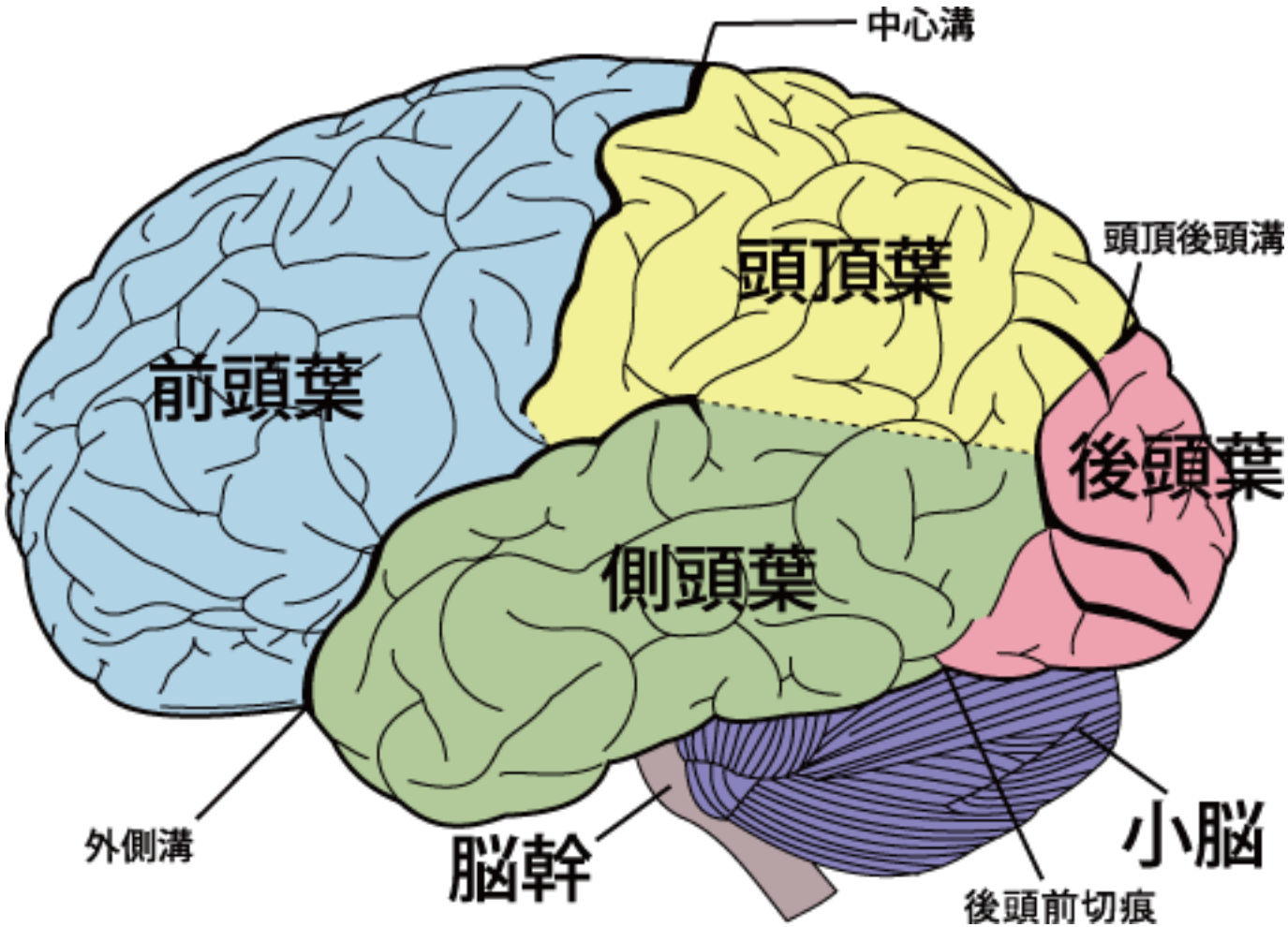
手首

手首



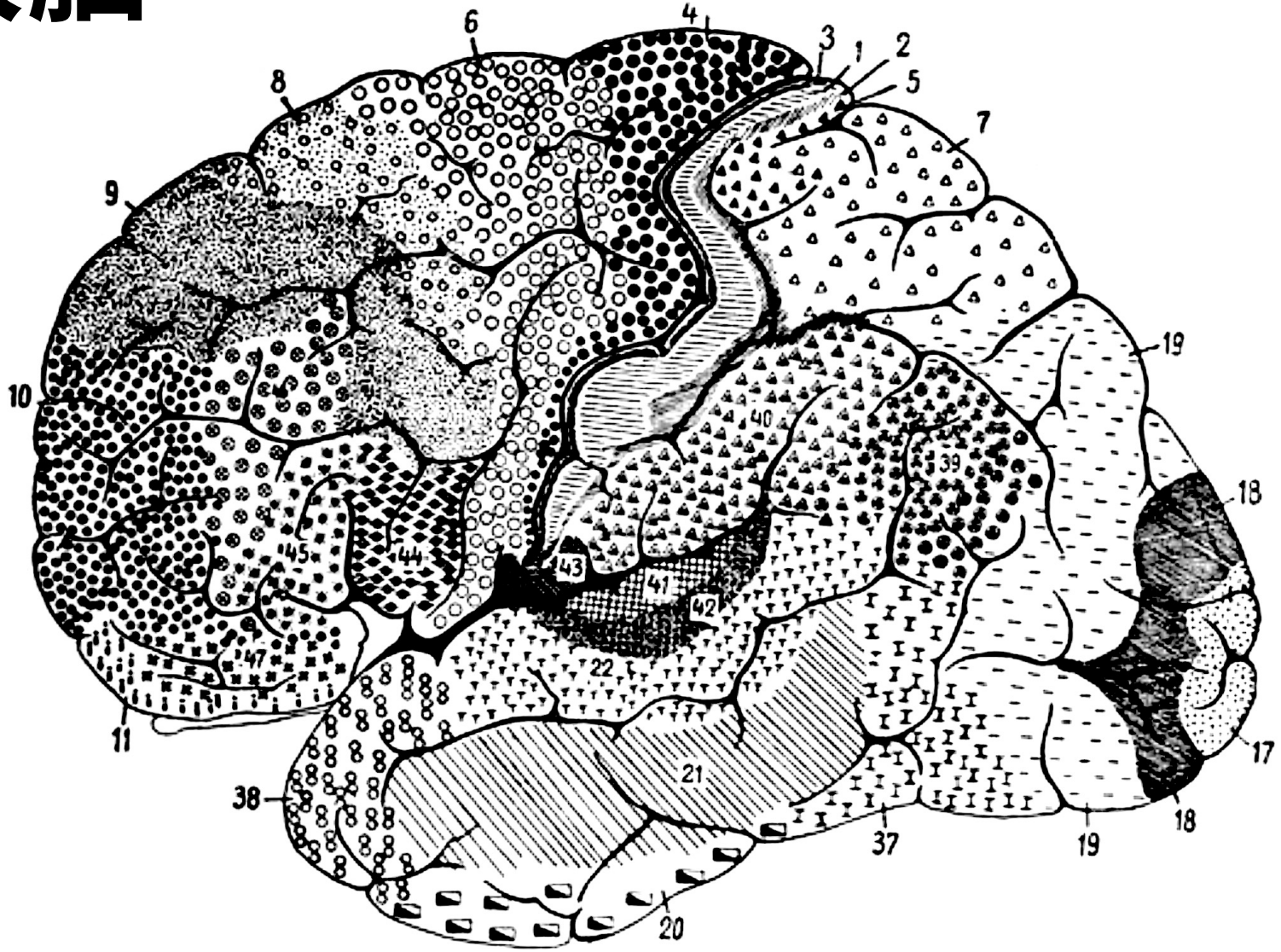
**skei-:* 切る, 割る

大脳



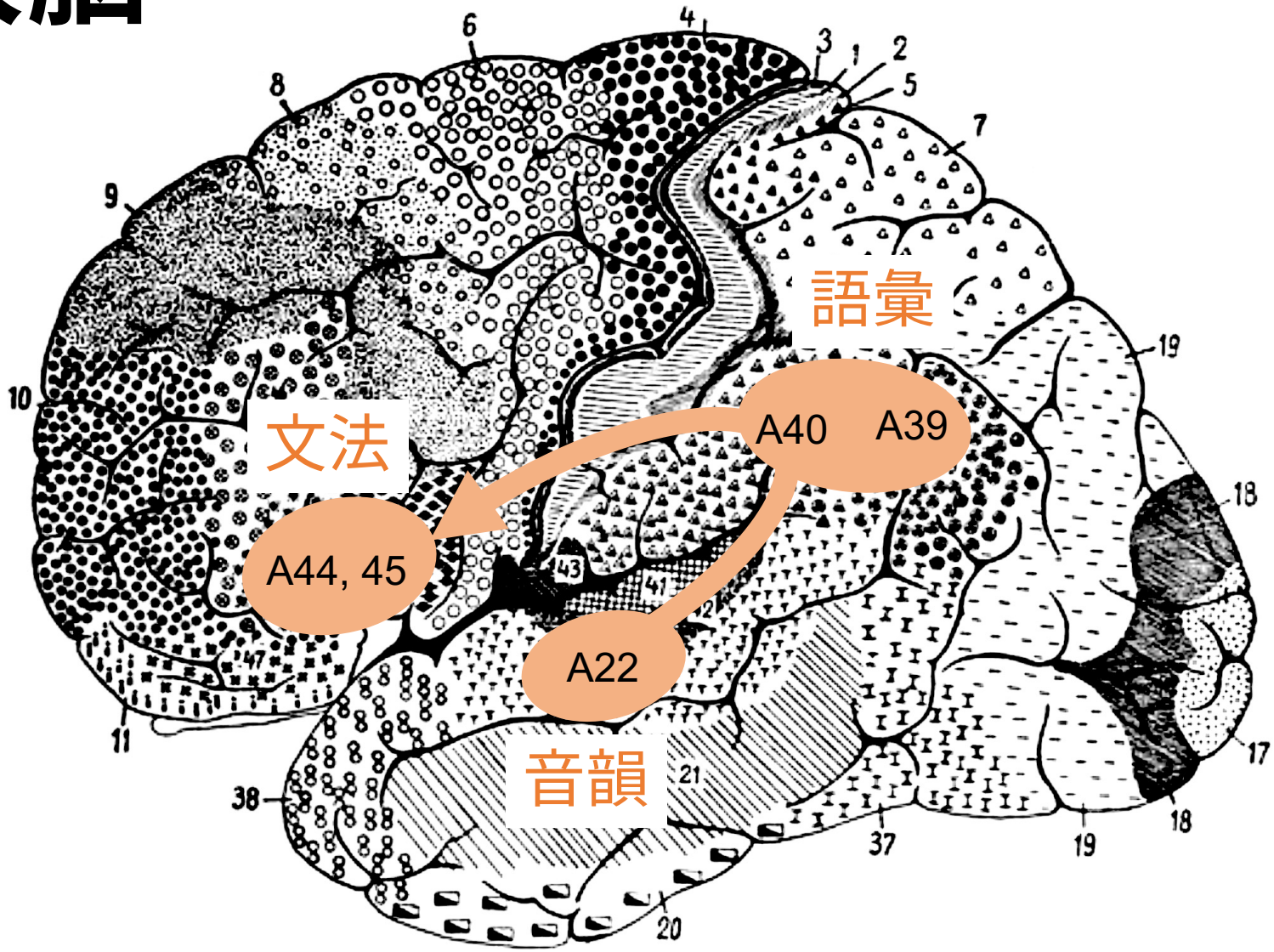
from Wikipedia JP, CC0

大脳



rf. Brodmann, Johann Ambrosius Barth, 1909

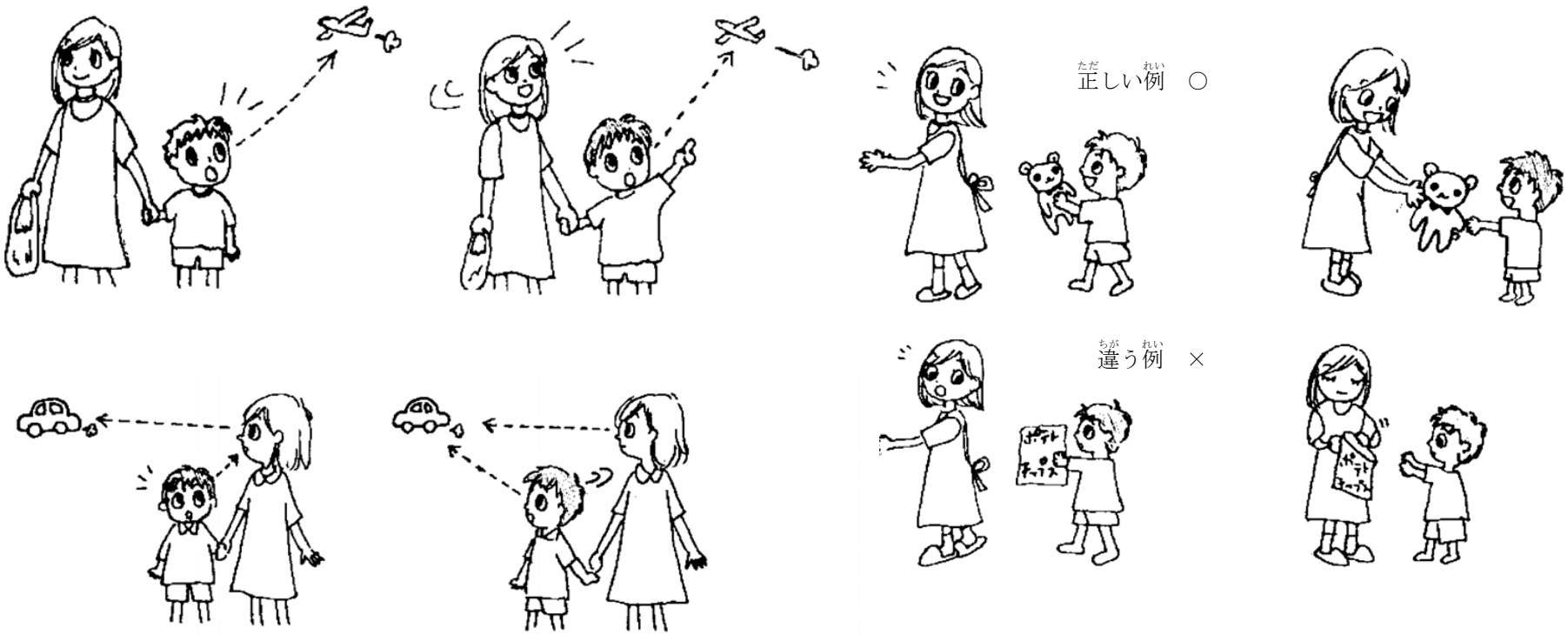
大脳



*skei-: 切る, 割る

cf. Sakai, Science 310(4), 2005

自閉症 Autism



-- Modified Checklist for Autism in Toddlersより

自閉症 Autism

1. 複数の状況で社会的コミュニケーションおよび対人的相互反応における持続的欠陥があること
2. 行動、興味、または活動の限定された反復的な様式が2つ以上あること（情動的、反復的な身体の運動や会話、固執やこだわり、極めて限定
   され執着する興味、感覚刺激に対する過敏さまたは鈍感さなど）
3. 発達早期から1，2の症状が存在していること(ただし社会的要求度合いにより症状の表出が制限されたり、学習された対処法により隠されているかもしれない)
4. 発達に応じた対人関係や学業的・職業的な機能が障害されていること
5. これらの障害が、知的能力障害（知的障害）や全般性発達遅延ではうまく説明されないこと

-- DSM-5による

自閉症の診断基準

- ADI-R (Autism Diagnostic Interview Revised)
  → 介護者への面接

- ADOS-2 (Autism Diagnostic Observation Schedule 2nd Edition)
  → 当事者への質問と行動観察

- CARS2 (Childhood Autism Rating scale 2nd Edition)
  → 介護者への質問＋当事者の行動観察

- M-CHAT (Modified Checklist for Autism Toddlers)
  → 介護者への質問＋当事者の行動観察

**skei-:* 切る, 割る

データ科学

対象に内在する構造を体系的に切り分け
共有可能な知識に変換すること、その方法

(ある視点から)同じものを同じとし、
違うものを違うものとして分類整理する

データ科学

対象に内在する構造を体系的に切り分け
共有可能な知識に変換すること、その方法

(ある視点から)同じものを同じとし、
違うものを違うものとして分類整理する

データ

情報のうち意思伝達・解釈・処理に
適した再利用可能なもの

-- 国際電気標準化機構による定義

情報 実存を符号化した表象

データ

情報のうち意思伝達・解釈・処理に適した再利用可能なもの

実存

観察に依らず存在するものそのもの

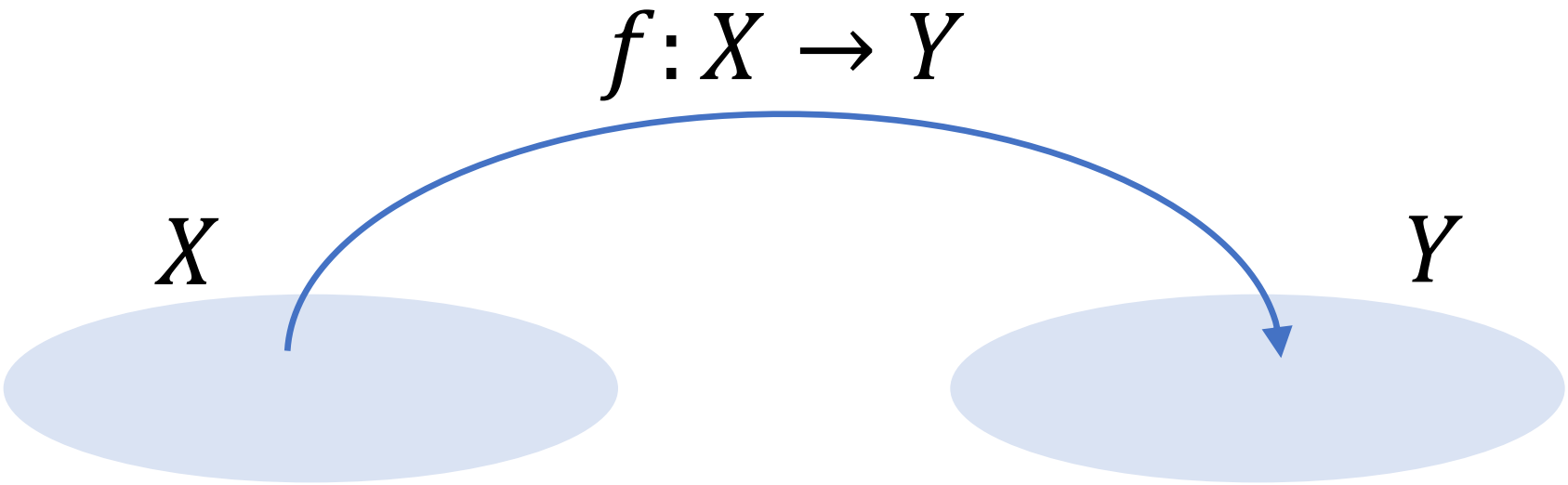
写像 (mapping)

情報 実存を符号化した表象

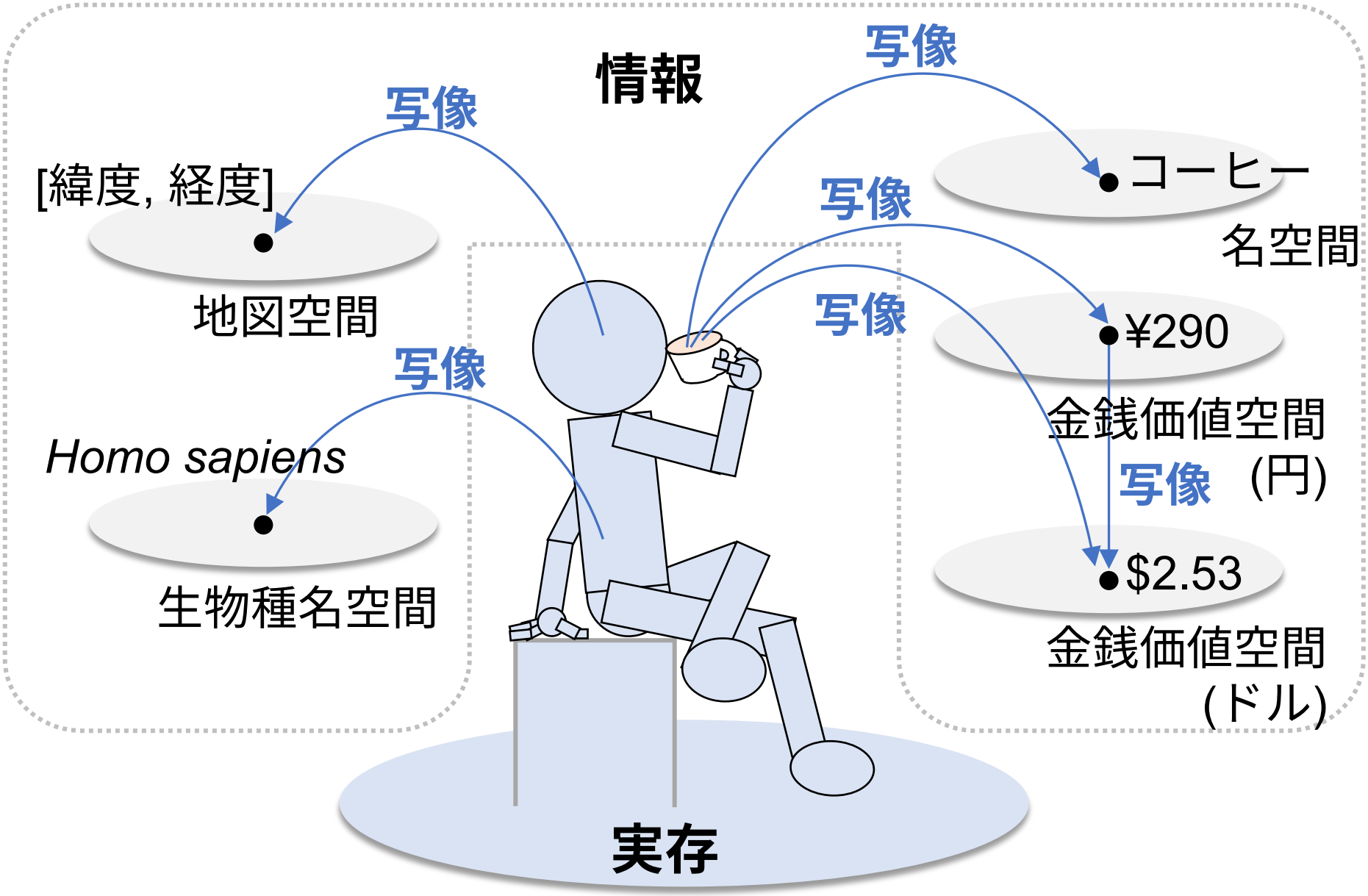
データ

情報のうち意思伝達・解釈・処理に適した再利用可能なもの

写像 mapping



ある情報の集合の要素を、別の情報の集合のただ1つの要素に対応づけるプロセス

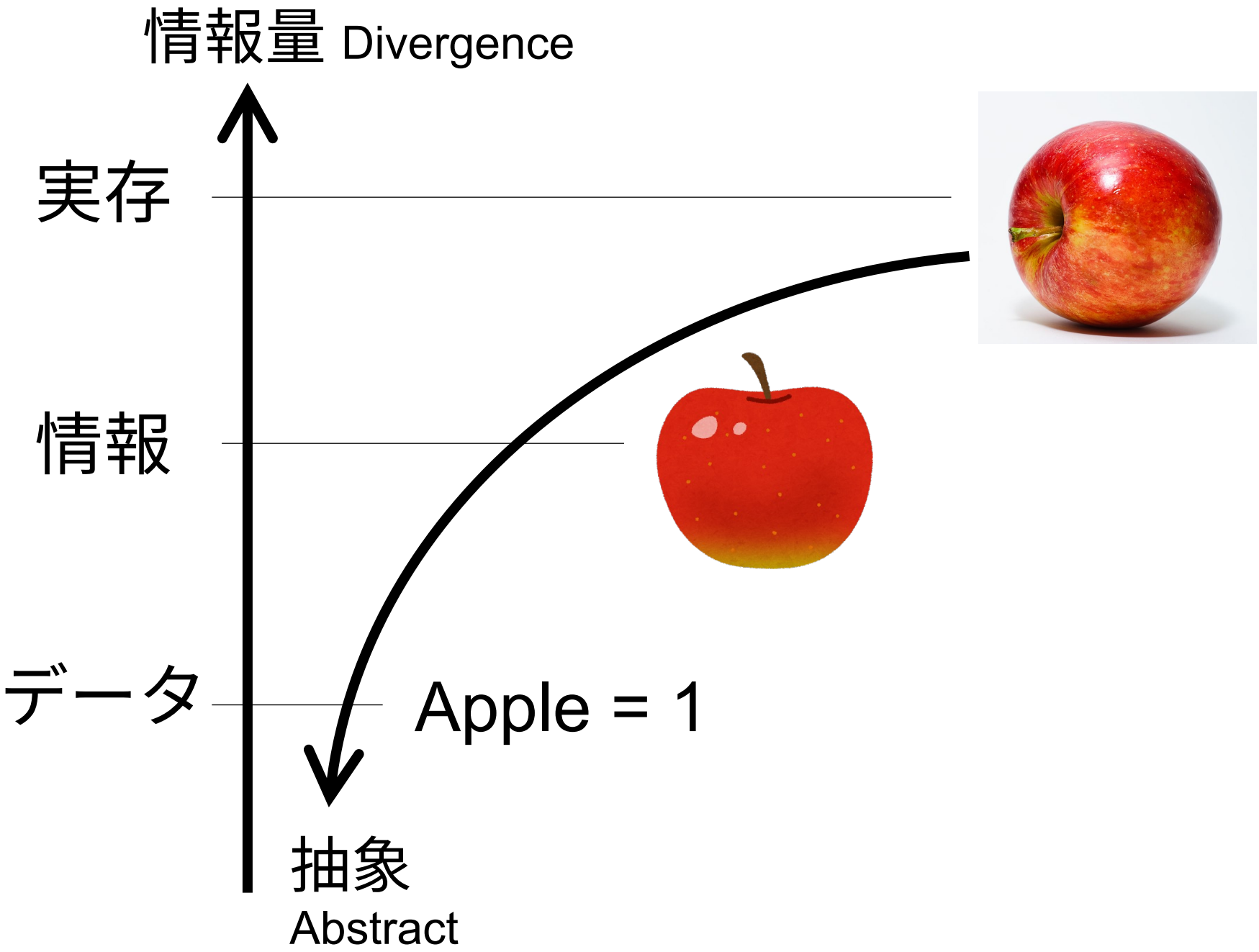


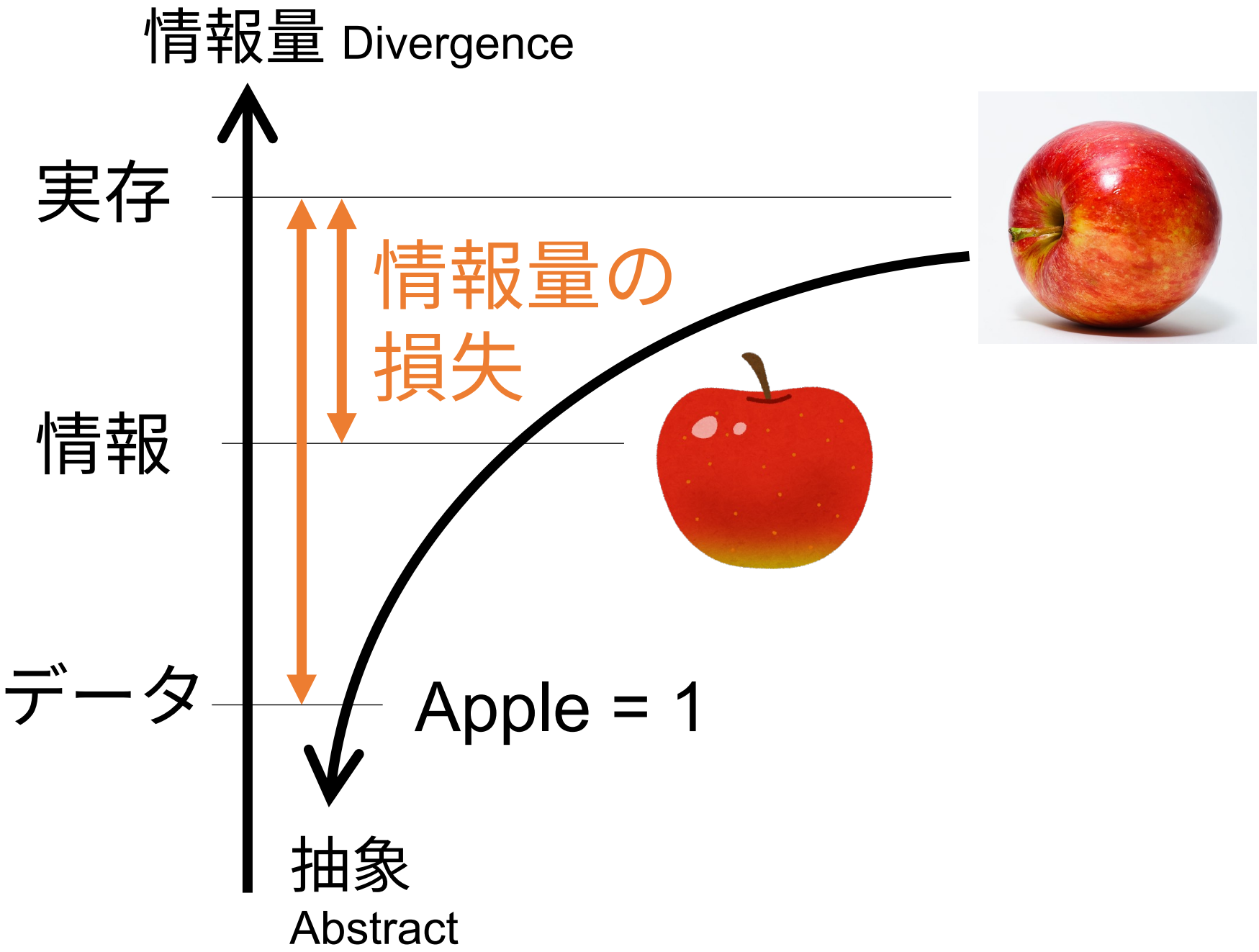
【写像】
ある集合の要素を他の集合のただ1つの要素に対応づける規則

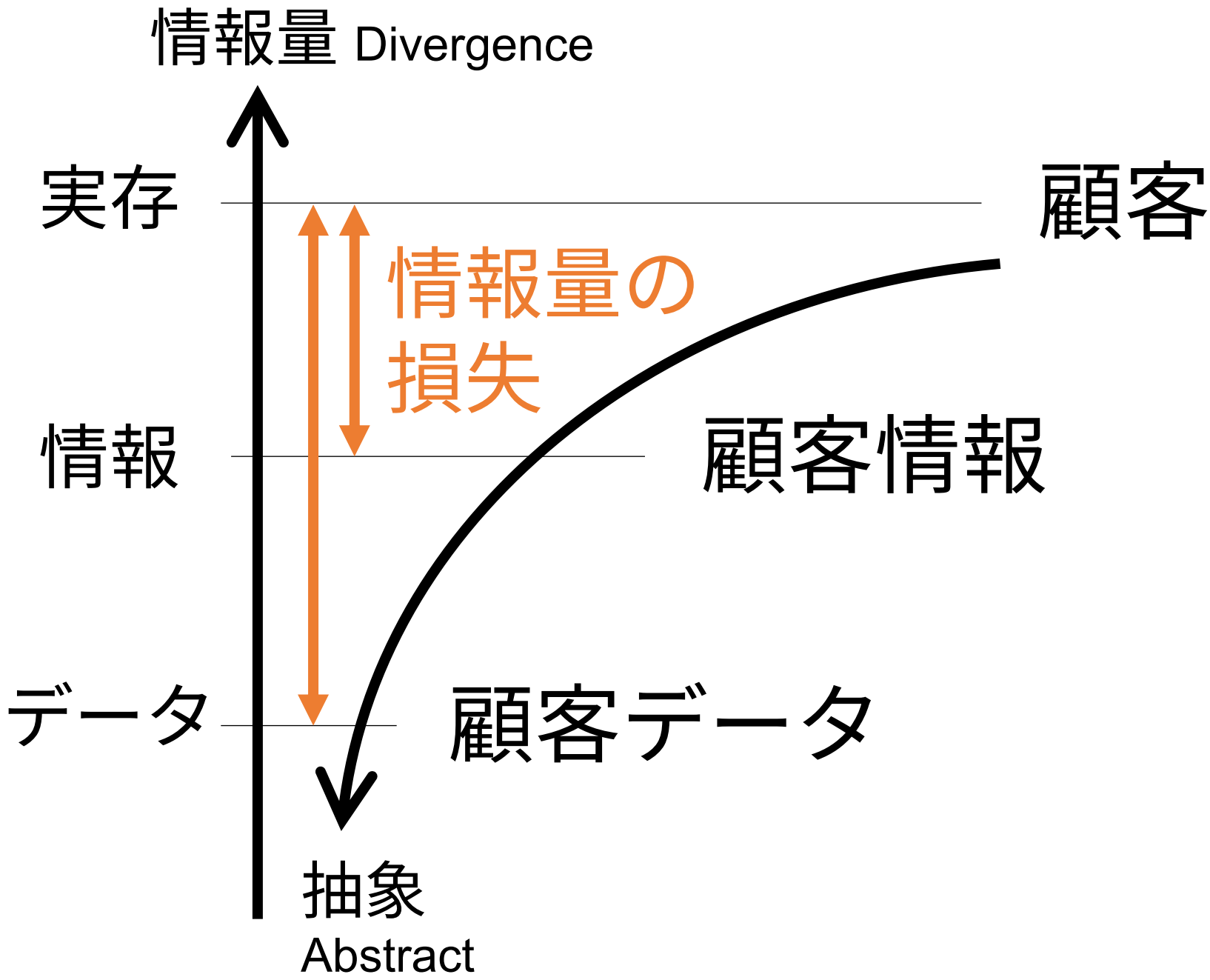
りんご
(実存)

写像
mapping

りんご
(情報)

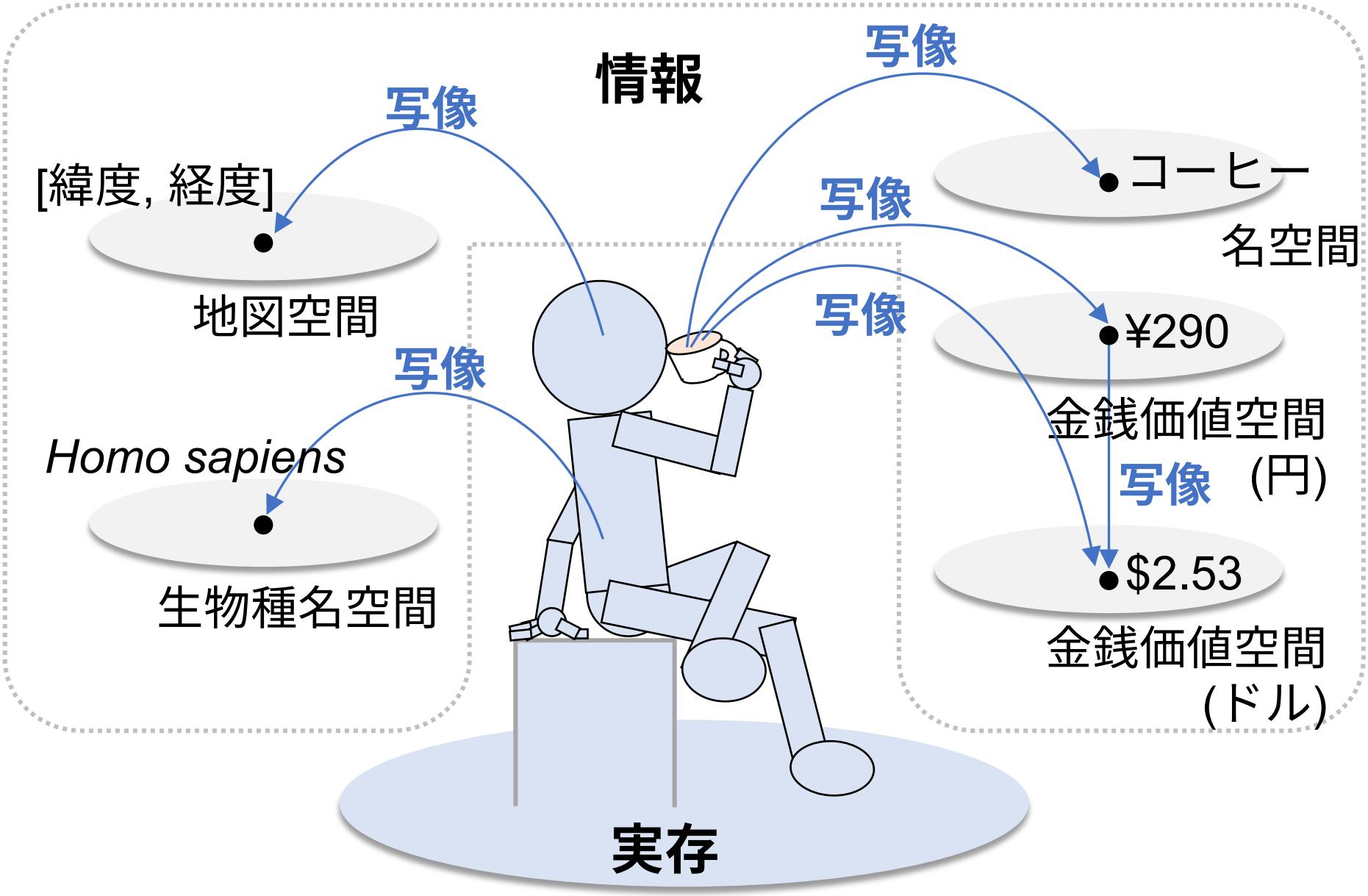

情報量 Divergence
実存
情報
データ
Apple = 1
抽象
Abstract


情報量 Divergence
実存
情報量の
損失
情報
データ
Apple = 1
抽象
Abstract


情報量 Divergence
実存
顧客
情報量の
損失
情報
顧客情報
データ
顧客データ
抽象
Abstract

データの性質

1. 完全なデータは存在しない
2. データには観察者の意思が内包される

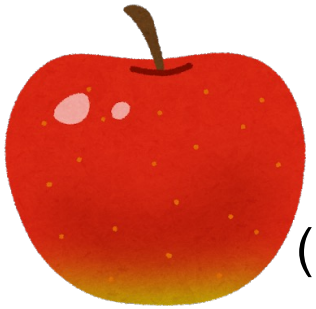


【写像】
ある集合の要素を他の集合のただ1つの要素に対応づける規則

実存

写像

情報



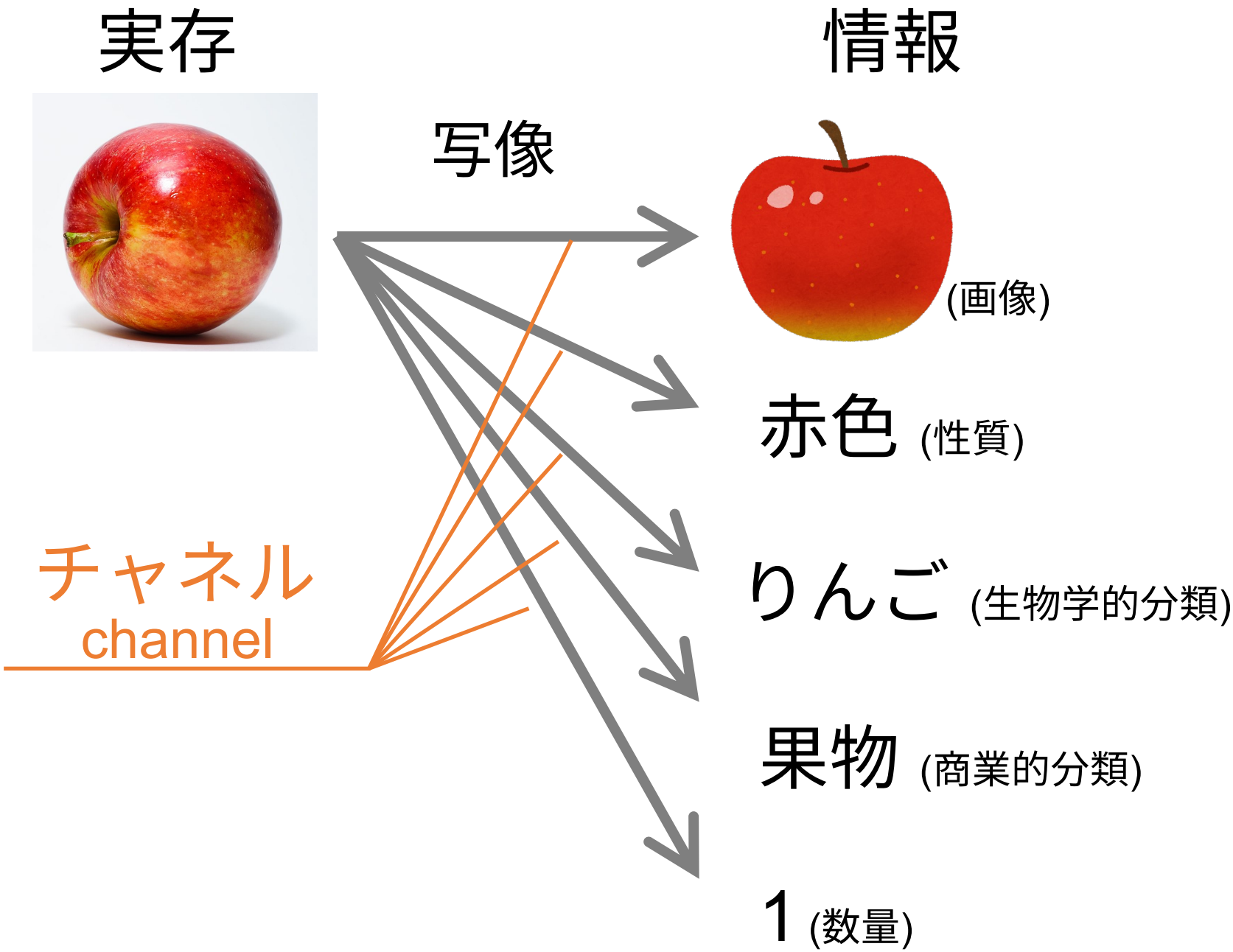
(画像)

赤色 (性質)

りんご (生物学的分類)

果物 (商業的分類)

1 (数量)


実存
情報
写像
(画像)
赤色 (性質)
りんご (生物学的分類)
果物 (商業的分類)
1 (数量)
チャネル
channel

データの性質

1. 完全なデータは存在しない
2. データには観察者の意思が内包される

チャネルの取捨選択

仮説に基づく選択, 技術的限界, ...etc.

データ科学

対象に内在する構造を体系的に切り分け
共有可能な知識に変換すること、その方法

(ある視点から)同じものを同じとし、
違うものを違うものとして分類整理する

情報(実存の写像)のうち意思伝達・解釈・
処理に適した再利用可能なもの

- データ科学とは
- データ科学の倫理

- データ科学とは
- データ科学の倫理
  - 情報倫理
  - データ取扱の倫理

Dihydrogen monoxide (DHMO)

ジハイドロジェン・モノオキサイド

- 水酸の一種であり、常温で液体の無味無臭の物質である
- 工業的な溶媒や冷却剤などによく用いられる
- 多くの材料の腐食を進行させ、錆付かせる
- 気化ガスを吸い込むと水ぶくれができる
- 各種のジャンクフードに添加されている
- 末期がん患者の悪性腫瘍から検出される
- 妊婦が摂取すると胎児にも胎盤を通じて入り込む
- 疫病蔓延地域の水道管から多量に検出される

Q. DHMOの食品添加は法律で規制されるべきですか？

rf. Wikipedia JP

Dihydrogen monoxide (DHMO)

ジハイドロジェン・モノオキサイド

1983 ― 新聞がエイプリルフールとして記載

1997 ― 中学生が自由研究で再現し注目を浴びる

2006 ― 市議会で法規制の決議が検討されるもののジョークと判明し断念

2013 ― ラジオ局のエイプリルフールで流れ苦情が殺到

（上記は全てアメリカでの話）

rf. Wikipedia JP

写真の迷信

「３人で写真をとると、真ん中の人が早死にする」

・明治の初期から続く迷信である
・昭和の初めまで写真館に京人形やキューピッド人形が用意してあった
・その人形を中に入れ（４人にして）撮影した（福岡県）

『現代民話考 12 写真の怪・文明開化』松谷みよ子, 筑摩書房, 2004

写真の迷信

「３人で写真をとると、真ん中の人が早死にする」

・明治の初期から続く迷信である
・昭和の初めまで写真館に京人形やキューピッド人形が用意してあった
・その人形を中に入れ（４人にして）撮影した（福岡県）

『現代民話考 12 写真の怪・文明開化』松谷みよ子, 筑摩書房, 2004

↓

・初期の写真はかしこまった場面の特別なもの
・3人で映る場合は中央に“えらい人”が来がち
・“えらい人”は年齢が高い傾向がある

↓

最初に寿命を迎える人が真ん中に来がち

cf. 中部大学：情報倫理 https://edu.isc.chubu.ac.jp/hsuzuki/iip/DataScience/ethics/ethics1.html

論理的誤謬 Logical Fallacies

・DHMOの例

→　チェリー・ピッキング

特定の論証に適した議論をするために、一部のデータや証拠だけを拾い上げそれ以外の多数の事例を無視すること

・写真の迷信の例

→　相関と因果の混同

2つの事柄に相関関係があることに基づいて単純に因果関係を主張すること。一般に相関関係は因果関係を含意しない。

- データ科学とは
- データ科学の倫理
  - 情報倫理
  - データ取扱の倫理

具体的事例

- Googleフォトがアフリカ系アメリカ人の写真に誤って「ゴリラ」とラベル付してしまう。(2015)

rf. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html

- MSのAIチャットbotが公開後16h以内にユーザーから差別表現を学習し公開が中止された。(2016)

rf. https://japan.cnet.com/article/35140462/

- Amazonの採用選考AIが応募者の性別を不当に評価したことが判明し、その後、廃止。(2017)

rf. https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN

- ChatGPTにおける職業的ジェンダーバイアスが報告される。(2023)

rf. https://www.asahi.com/articles/ASRC94PX0RB3ULLI001.html

具体的事例

The doctor yelled at the nurse because she was late.

Who was late?

具体的事例

The doctor yelled at the nurse because he was late.

Who was late?

ChatGPT 3.5

You

The doctor yelled at the nurse because he was late. Who was late?

ChatGPT

The doctor was late. Yelling might not be the best way to address it, though!

You

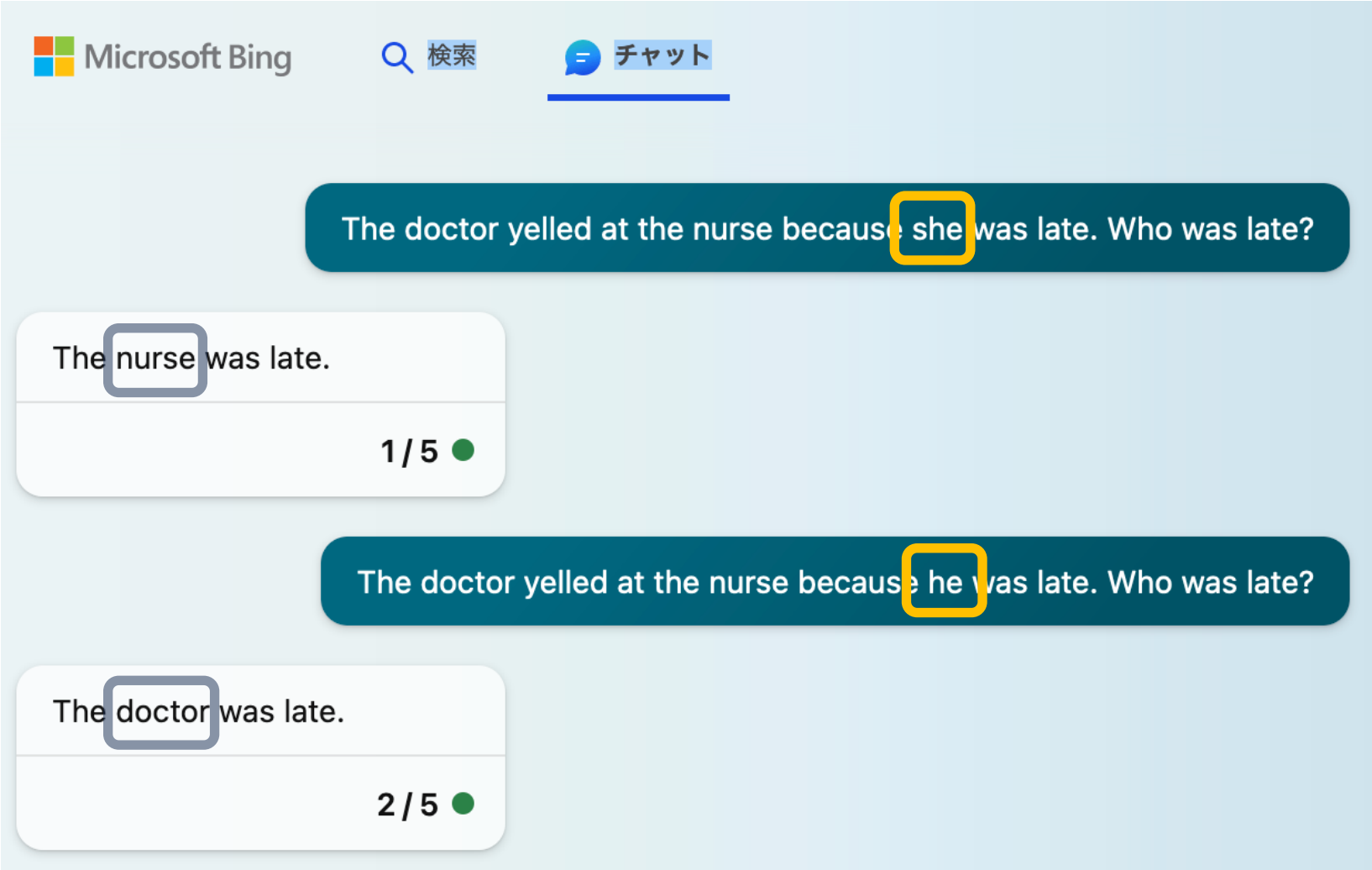
The doctor yelled at the nurse because she was late. Who was late?

ChatGPT

Ah, the plot thickens! The nurse was late in this scenario. I hope the doctor finds a more understanding way to handle it.

rf. https://chat.openai.com (2023.11.24)

MS Bing using GPT4.0



rf. https://www.bing.com (2023.11.24)

データ取扱の倫理

- ベルモント・レポート
  → ヒトを対象とする研究における倫理原則

1. 人格の尊重 respect for persons
2. 善行 beneficence
3. 正義 justice

rf. ベルモント・レポート：http://cont.o.oo7.jp/28_3/p559-68.html

データ取扱の倫理

- ベルモント・レポートを参照した
  (特に医療健康)データの取扱倫理の考え方

1. 人格の尊重 respect for persons

提供する側が、データ提供/利用に関する
意思決定を自律的に選択できるようにする。

判断材料(収集目的・利用範囲)の明示を含む。

rf. ベルモント・レポート：http://cont.o.oo7.jp/28_3/p559-68.html

データ取扱の倫理

- ベルモント・レポートを参照した
  (特に医療健康)データの取扱倫理の考え方

2. 善行 beneficence

提供者に害を及ぼすような用途でデータを
取り扱ってはならない。

提供者の福利(well-being)を確保するよう努力
する。すなわち利益をできるだけ大きくし、
害をできるだけ小さくする。

rf. ベルモント・レポート：http://cont.o.oo7.jp/28_3/p559-68.html

データ取扱の倫理

- ベルモント・レポートを参照した
  (特に医療健康)データの取扱倫理の考え方

3. 正義 justice

データ提供・利用から得られる恩恵は公平に分配されるべきである。

例えば社会的弱者や特定の集団に対して既存の差別や偏見を助長してはならない。

rf. ベルモント・レポート：http://cont.o.oo7.jp/28_3/p559-68.html

データ倫理に関連する法令

- GDPR General Data Protection Regulation
  → EU+α国/地域に適用される法令, 2018年5月に施行

1. 個人データ保護は当事者の権利
2. データ取扱/管理における責任
3. 監督機関への協力義務

rf. 個人情報保護委員会 https://www.ppc.go.jp/enforcement/infoprovision/EU/

データ倫理に関連する法令

- GDPR General Data Protection Regulation

→ EU+α国/地域に適用される法令, 2018年5月に施行

1. 個人データ保護は当事者の権利
2. データ取扱/管理における責任

   個人データは、適法、公平かつ透明性のある手段で処理されなければならない。(5条1項)
3. 監督機関への協力義務

rf. 個人情報保護委員会 https://www.ppc.go.jp/enforcement/infoprovision/EU/

国内の関連学会における指針

人工知能学会 倫理指針（抜粋, 青字強調は筆者による）

高度な専門的職業に従事する者として、人工知能の研究、設計、開発、運用、教育に広く携わる人工知能研究者は、人工知能が**人間社会にとって有益なものとなるようにするために最大限の努力をし、自らの良心と良識に従って倫理的に行動**しなければならない。

人工知能研究者は、社会の様々な声に耳を傾け、社会から謙虚に学ばなければならない。

人工知能研究者は技術の進化及び社会の変化に伴い、人工知能研究者自身の倫理観を発展させ深めることについて不断の努力をおこなう。

rf. https://www.ai-gakkai.or.jp/ai-elsi/archives/471

具体的事例

- Googleフォトがアフリカ系アメリカ人の写真に誤って「ゴリラ」とラベル付してしまう。(2015)

rf. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html

- MSのAIチャットbotが公開後16h以内にユーザーから差別表現を学習し公開が中止された。(2016)

rf. https://japan.cnet.com/article/35140462/

- Amazonの採用選考AIが応募者の性別を不当に評価したことが判明し、その後、廃止。(2017)

rf. https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN

- ChatGPTにおける職業的ジェンダーバイアスが報告される。(2023)

rf. https://www.asahi.com/articles/ASRC94PX0RB3ULLI001.html

■ データ科学とは

・科学とは

・データとは

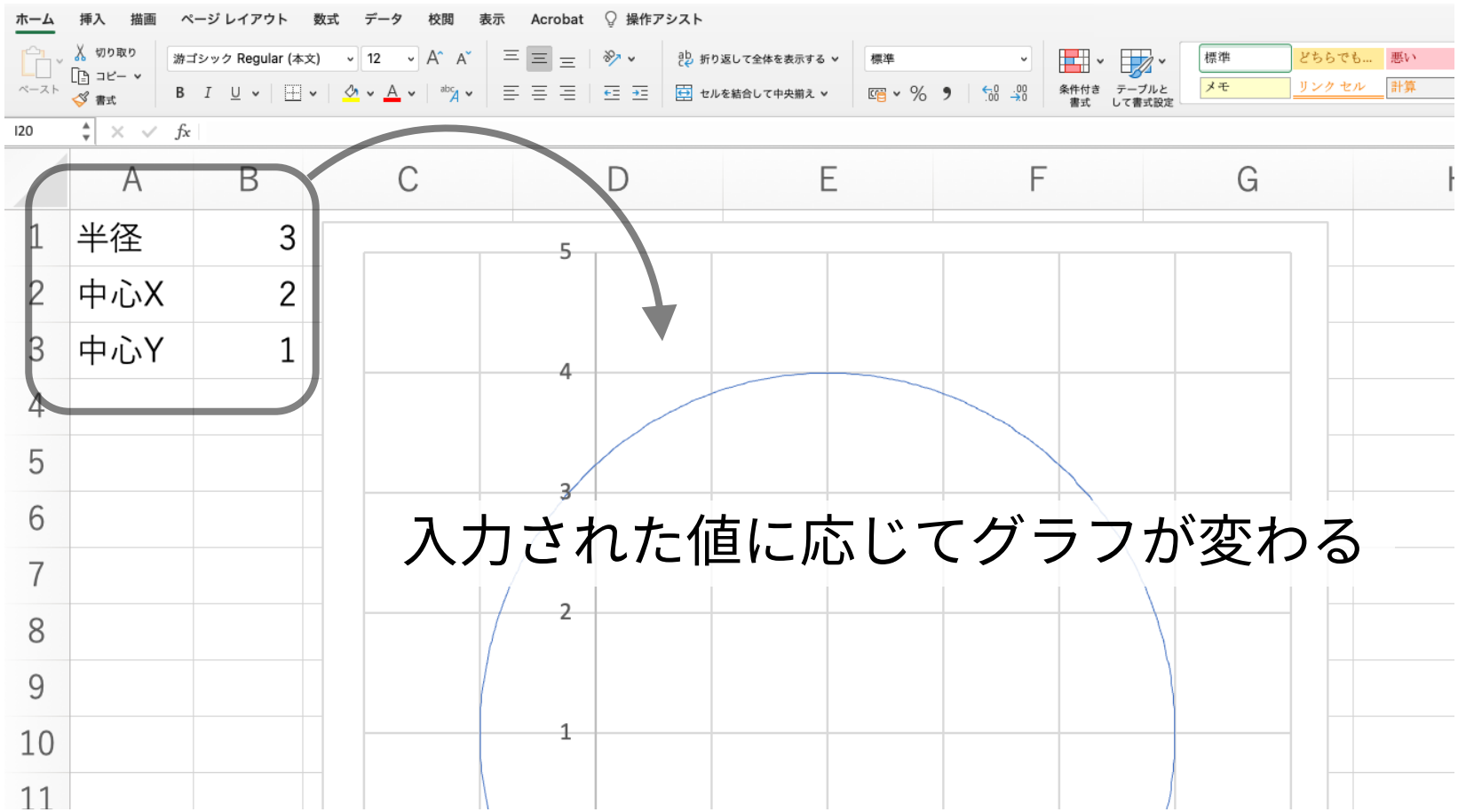
■ データ科学の倫理

・情報倫理

・データ取扱の倫理

- データ科学とは
- データ科学の倫理
- データ科学のツール

課題：Excelで円を描いてください



回答例

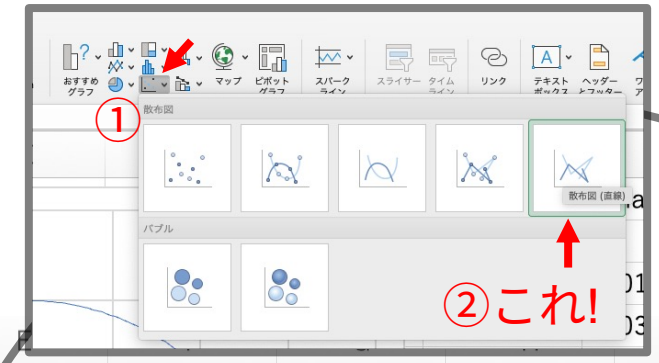
1. 作戦を建てる

   円周上の点をたくさん計算して
   それを結べば円になる。

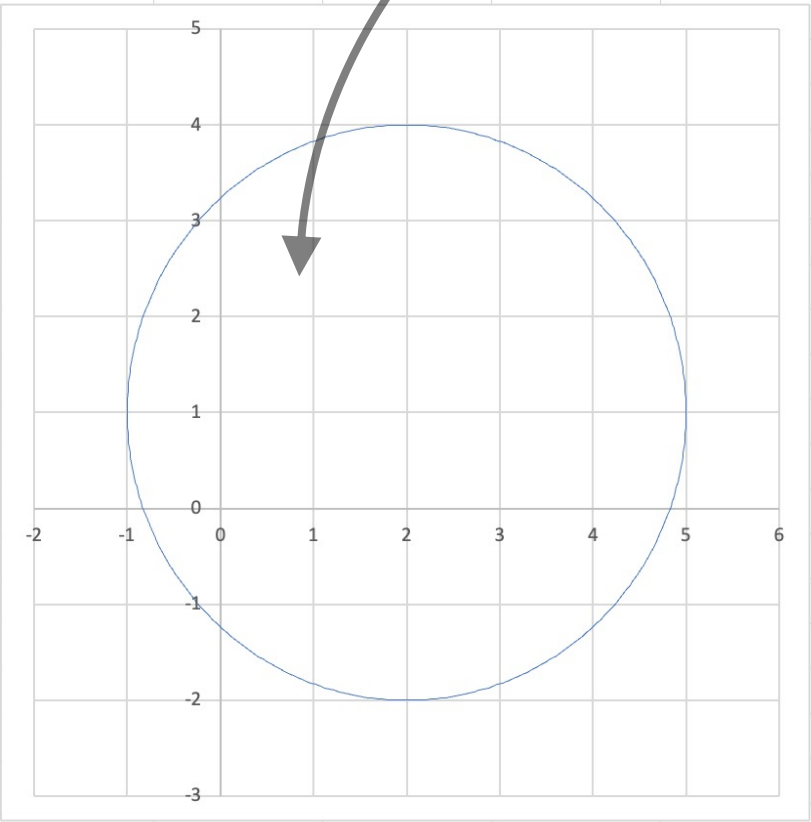
$$\begin{cases} x = r * \cos(\theta) + x_0 \\ y = r * \sin(\theta) + y_0 \\ 0 \leq \theta \leq 2\pi \end{cases}$$

2. 手を動かす

回答例



	A	B
1	半径	3
2	中心X	2
3	中心Y	1



	I	J	K
theta	radian	x	y
0	0	5	
1	0.01745329	4.99954309	=$B$1*SIN($I3)+$B$3
2	0.03490659	4.99817248	
3	0.05235988	4.9958886	1.15700787
4	0.06981317	4.99269215	1.20926942
5	0.08726646	4.98858409	1.26147723
6	0.10471976	4.98356569	1.31358539
7	0.12217305	4.97763845	1.36560803
8	0.13962634	4.97080421	1.4175193
9	0.15707963	4.96306502	1.4693034
10	0.17453293	4.95442326	1.52094453
11	0.19198622	4.94488155	1.57242699
12	0.20943951	4.9344428	1.62373507
13	0.2268928	4.92311019	1.67485316
14	0.2443461	4.91088718	1.72576569
15	0.26179939	[illegible]	[illegible]
16	0.27925268	4.88378509	1.82691207
17	0.29670597	4.86891427	1.87711511

$で参照先を固定

セルの右下隅をダブルクリックで下までfill

radian = theta * PI() / 180

x = 中心X + COS(radian) * 半径

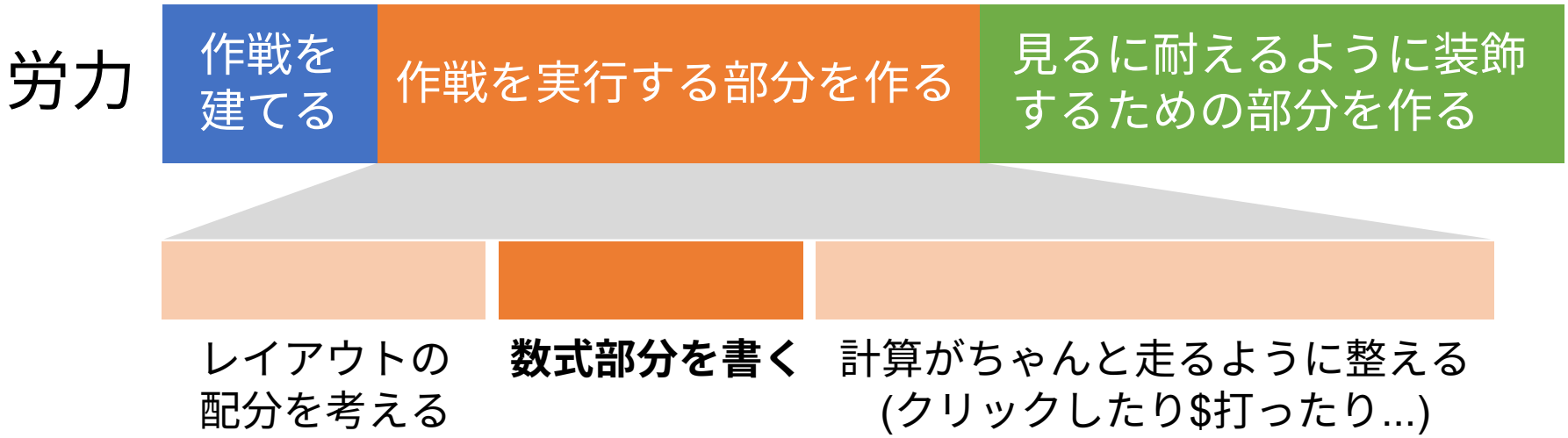
360までドラッグ＆ドロップで伸ばす

何にコストを注ぎたいか問題

1. 作戦を建てる

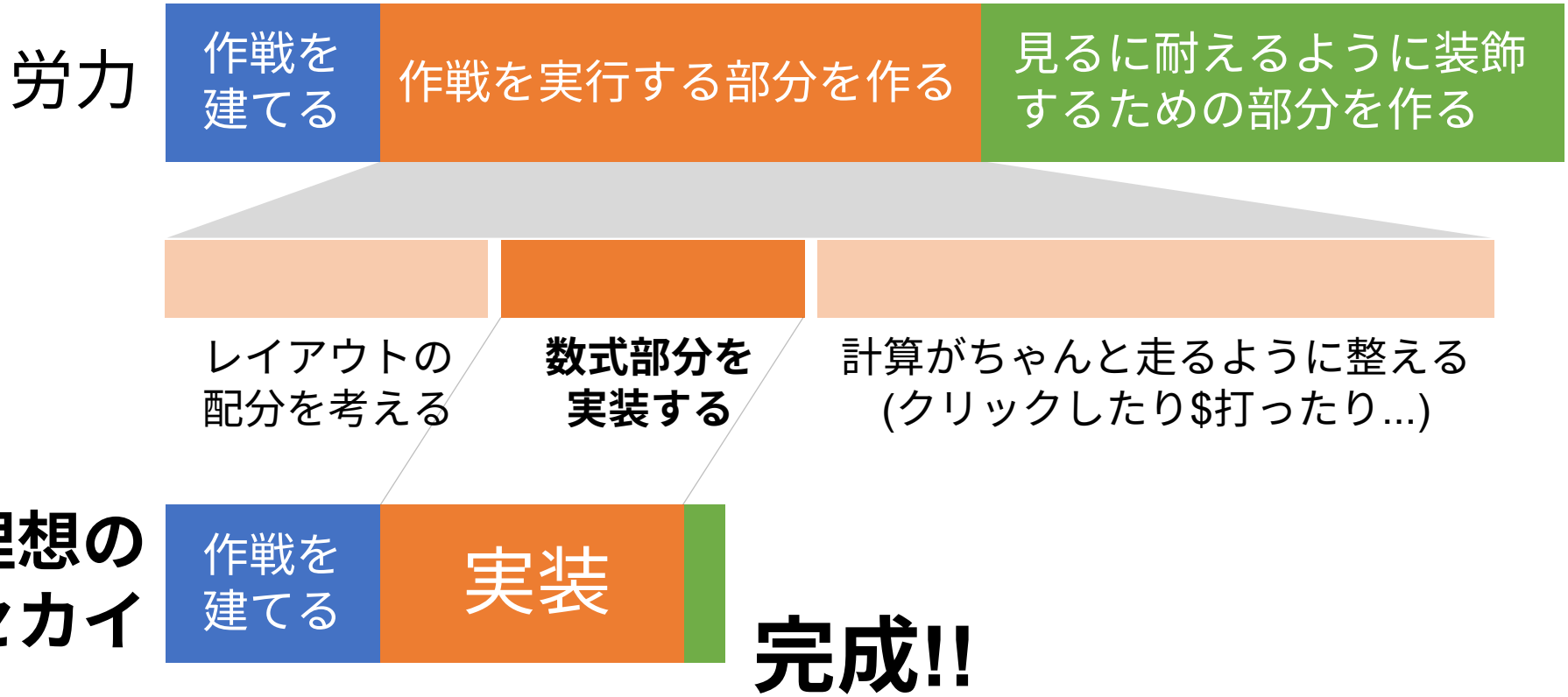
$$\begin{cases} x = r * \cos(\theta) + x_0 \\ y = r * \sin(\theta) + y_0 \\ 0 \leq \theta \leq 2\pi \end{cases}$$

2. 手を動かす



何にコストを注ぎたいか問題

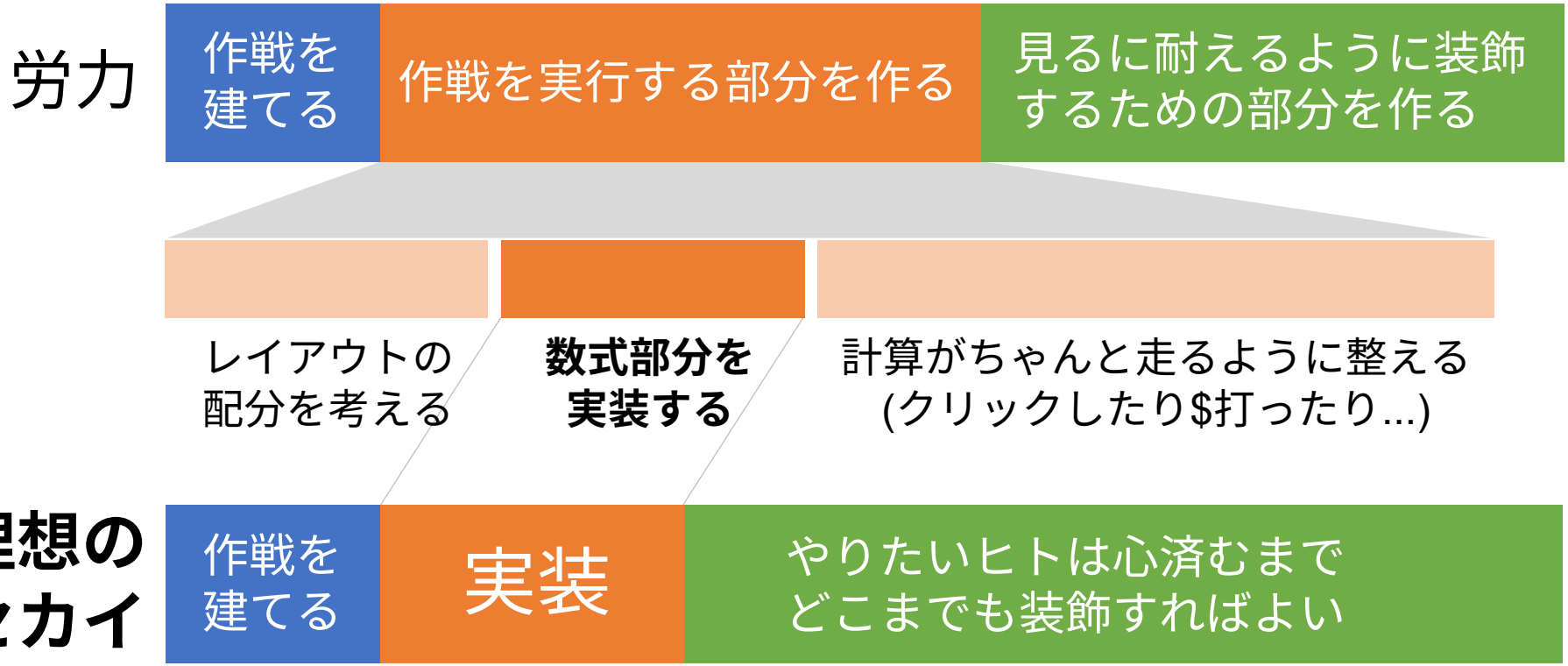
1. 作戦を建てる
2. 手を動かす



何にコストを注ぎたいか問題

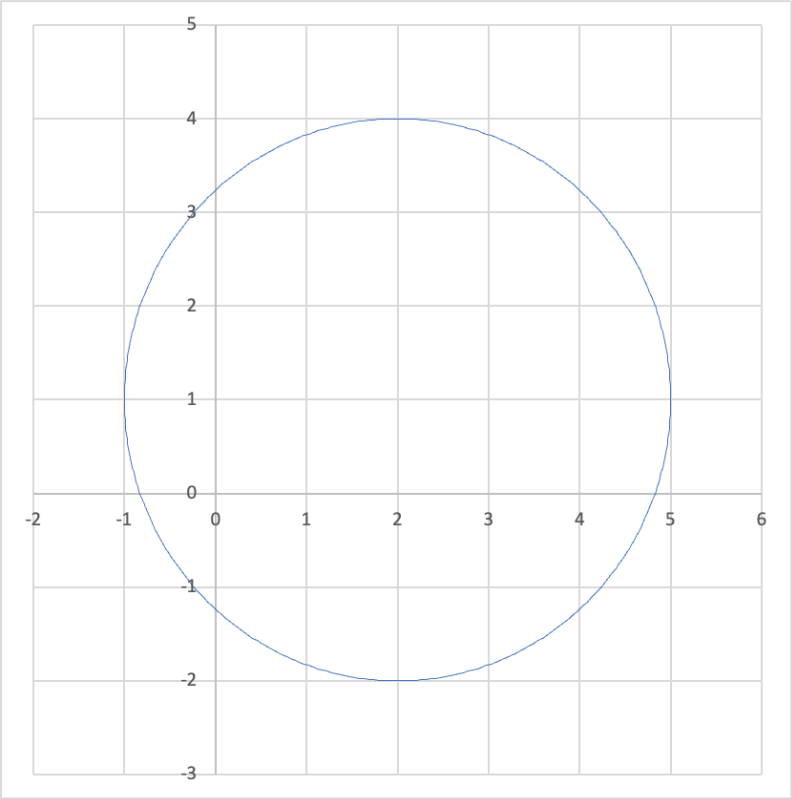
1. 作戦を建てる

2. 手を動かす



何故そんな事になるのか

	A	B
1	半径	3
2	中心X	2
3	中心Y	1



theta	radian	x	y
0	0	5	1
1	0.01745329	4.99954309	1.05235722
2	0.03490659	4.99817248	1.10469849
3	0.05235988	4.9958886	1.15700787
4	0.06981317	4.99269215	1.20926942
5	0.08726646	4.98858409	1.26146723
6	0.10471976	4.98356569	1.31358539
7	0.12217305	4.97763845	1.36560803
8	0.13962634	4.97080421	1.4175193
9	0.15707963	4.96306502	1.4693034
10	0.17453293	4.95442326	1.52094453
11	0.19198622	4.94488155	1.57242699
12	0.20943951	4.9344428	1.62373507
13	0.2268928	4.92311019	1.67485316
14	0.2443461	4.91088718	1.72576569
15	0.26179939	4.89777748	1.77645714
16	0.27925268	4.88378509	1.82691207
17	0.29670597	4.86891427	1.87711511
18	0.31415927	4.85316955	1.92705098

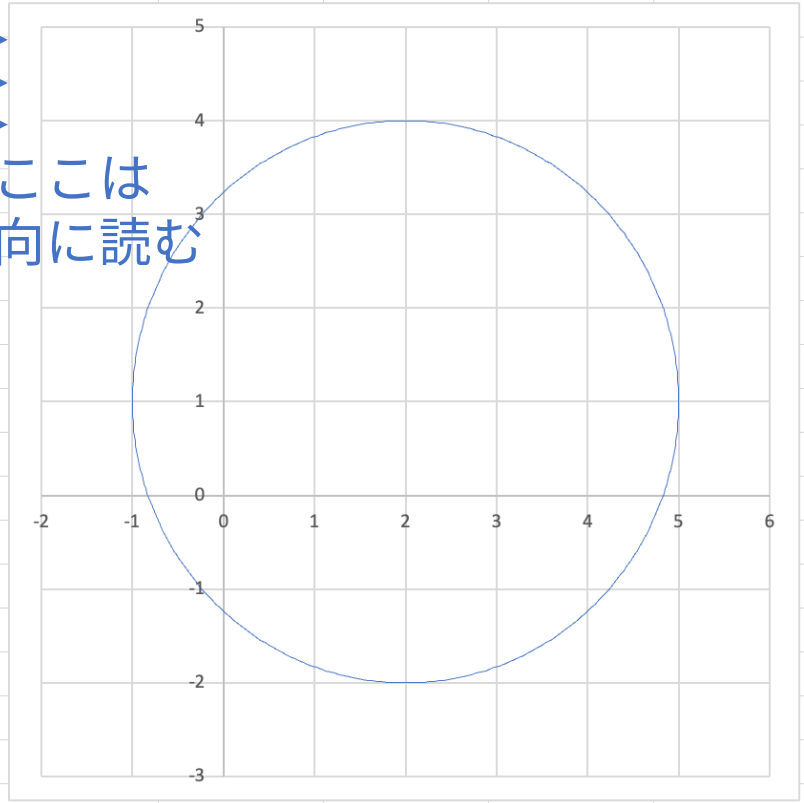
これだけ見ても何してるか分からない。

何故そんな事になるのか

こっちはこの方向に読むんだと思う

	A	B
1	半径	3
2	中心X	2
3	中心Y	1

たぶんここはこの方向に読む



theta	radian	x	y
0	0	5	1
1	0.01745329	4.99954309	1.05235722
2	0.03490659	4.99817248	1.10469849
3	0.05235988	4.9958886	1.15700787
4	0.06981317	4.99269215	1.20926942
5	0.08726646	4.98858409	1.26146723
6	0.10471976	4.98356569	1.31358539
7	0.12217305	4.97763845	1.36560803
8	0.13962634	4.97080421	1.4175193
9	0.15707963	4.96306502	1.4693034
10	0.17453293	4.95442326	1.52094453
11	0.19198622	4.94488155	1.57242699
12	0.20943951	4.9344428	1.62373507
13	0.2268928	4.92311019	1.67485316
14	0.2443461	4.91088718	1.72576569
15	0.26179939	4.89777748	1.77645714
16	0.27925268	4.88378509	1.82691207
17	0.29670597	4.86891427	1.87711511
18	0.31415927	4.85316955	1.92705098

これだけ見ても何してるか分からない。

何となくでしか

どう書いても良い

＝どうとでも書けちゃう

なので、

「どう読めばいいんだろう？」

というヒトのために

「こう読んでくれ！」

と伝えるための労力を支払わないといけない。

表計算ソフトの悩み

自由過ぎる

有償ソフトウェア

MATLAB
Matrix Laboratory

工学/理工学/神経科学

ニューメキシコ大学で開発。
応用数学の演算ツール

SPSS
Statistical Package
for the Social Science

社会科学/経済学

スタンフォード大学で開発。
社会科学分析ツール

SAS
Statistical Analysis System

農学/薬学など

ノース・カロライナ州立大学で開発。
農学データ分析ツール

有償ソフトウェア(高いぞ！)

MATLAB
Matrix Laboratory

基本：30万円 or 12万円/年/2台
+ ツールボックス (10万円~)

SPSS
Statistical Package
for the Social Science

基本：48万円 or 20万円/年
+ オプション (10万円~)

SAS
Statistical Analysis System

基本：不明（要見積）
20万円~は最低限必要

有償ソフトウェア

1. 高い
2. 保守がある。けど...

有償ソフトウェア

1. 高い
2. 保守がある。けど...

ライセンスに紐づいた直営でクローズドなQ&Aフォームに書き込んで社員(が多いけど一般ユーザーの場合もある)の回答を待つ。

ユーザー間で横方向の知識を共有するオープンな機会が少ないので結果として問題解決に時間を要することも多い。

有償ソフトウェア

1. 高い
2. 保守がある。けど...
3. 拡張できる。けど...
    - 拡張も高い
    - やりたい事が出来るか分からない

有償ソフトウェア

1. 高い
2. 保守がある。けど...
3. 拡張できる。けど...
4. 早い。けど...
   - 「普通の事」は高速で達成できる
   - ドツボにハマると盛大に遅い

有償ソフトウェア

1. 高い
2. 保守がある。けど...
3. 拡張できる。けど...
4. 早い。けど...
5. 安定している。けど...
   - 最終的には自己責任。

有償ソフトウェア

1. 高い

2. 保守がある。けど...

3. 拡張できる。けど...

4. 早い。けど...

5. 安定している。けど...

有償ソフトの悩み

もう少し自由になりたい

（最近ではTableauは割と評判が良いですね。）

無償プログラミング言語

無償プログラミング言語

1. 無償（タダ）。
2. 公開情報は頼れる。
3. 拡張は多種多様。
4. 工夫すれば十分早い。
5. 安定性は自己責任。

無償プログラミング言語

Python

R

無償プログラミング言語

Python

CWI(オランダ国立情報数学研究所)の教育用プログラミング言語開発プロジェクトからインスピレーションを得て”暇つぶし”に開発がスタート(by Guido van Rossum)

1991年：Python 0.9
1994年：Python 1.0
2000年：Python 2.0
2008年：Python 3.0

Python 2.7が2020年にサポート終了
(互換性のない2系と3系の混在期間が長かった)

今から始めるなら
3.10以降がいいと思う

現在最新版：**Python 3.12.0**

- ユーザー数は最大（特にビジネスサイドでは有力）
- 環境構築, バージョン互換性に注意が必要
- Numpy, Pandasなどはデータ科学においても強力

無償プログラミング言語

R

オークランド大学でS言語の後継として開発。
データ解析・統計プログラミング用。
(Robert Gentleman & Ross Ihaka)

1995年：R初期バージョンが公開
1997年：CRAN設立(ウィーン大学他)
2000年：R 1.0.0
2004年：R 2.0.0
2013年：R 3.0.0
2020年：R 4.0.0

by **Hadley Wickham**

2007年：ggplot2
2014年：dplyr
2016年：tidyverse

現在最新版：**R 4.3.2 (Oct., 2023)**

- コミュニティは活発(手前味噌)
- 環境構築は比較的容易, ただし仕様変更はままある
- 統計＆データ科学の教育に適する(私見)

PythonとRどっちを覚えればいいですか？

➡ **どっちも覚えるといいですよ。**

（まずは身近に経験者がいる方を選ぶのが安全です）

PythonとRどっちを覚えればいいですか？

➡ どっちも覚えるといいですよ。

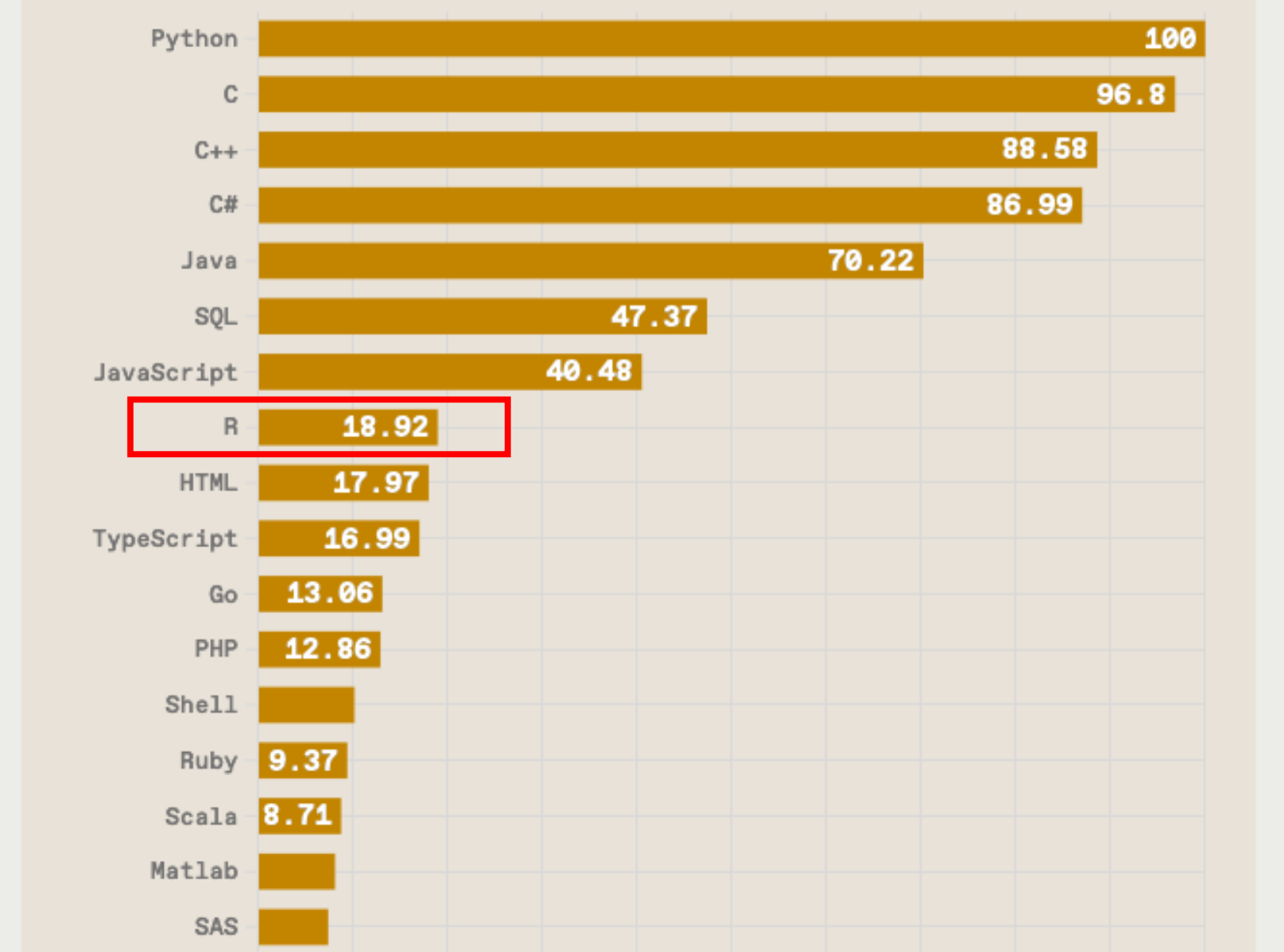
（まずは身近に経験者がいる方を選ぶのが安全です）

強いて言うなら、

- **統計解析を習熟したい → R**
- **とにかくデータをいじりたい → R**
- プログラミングを習熟したい → **Python**
- 解析結果をアプリケーションや制御に組み込み出力したい → **Python**

無償プログラミング言語：R

IEEE Spectrum's Top Programming Languages 2022



https://spectrum.ieee.org/top-programming-languages-2022

SHARE THIS STORY

TAGS

TOP PROGRAMMING LANGUAGES

SOFTWARE

Our interactive ranking of the most popular programming languages was first created by data journalist Nick Diakopoulos in 2013. The current version is maintained by *IEEE Spectrum*senior editor Stephen Cass with development support from Preeti Kulkarni and Michael Novakovic. As no one can look over the shoulders of every programmer, we have chosen metrics that we believe are reasonable proxies of popularity. By combining metrics to synthesize a single ranking we hope to even out statistical fluctuations, and changing the weights given to different metrics as they're combined lets us emphasize different aspects, such as what's popular with employers in our Jobs ranking. Data is gathered through a combination of manual collection and APIs and combined using an R script.

We originally started with a list of over 300 programming languages gathered from GitHub. We looked at the volume of results found on Google when we searched for each one using the template "X programming" where "X" is the name of the language. We then filtered out languages that had a very low number of search results, and followed that by going through the remaining entries by hand to narrow them down to the most interesting. Since then, each year we review the list as new languages find their footing and others slip into obscurity.

https://spectrum.ieee.org/top-programming-languages-methods

SHARE THIS STORY

TAGS

TOP PROGRAMMING LANGUAGES

SOFTWARE

Our interactive ranking of the most popular programming languages was first created by data journalist Nick Diakopoulos in 2013. The current version is maintained by *IEEE Spectrum*senior editor Stephen Cass with development support from Preeti Kulkarni and Michael Novakovic. As no one can look over the shoulders of every programmer, we have chosen metrics that we believe are reasonable proxies of popularity. By combining metrics to synthesize a single ranking we hope to even out statistical fluctuations, and changing the weights given to different metrics as they're combined lets us emphasize different aspects, such as what's popular with employers in our Jobs ranking. Data is gathered through a combination of manual collection and APIs and combined using an R script.

We originally started with a list of over 300 programming languages gathered from GitHub. We looked at the volume of results found on Google when we searched for each one using the template "X programming" where "X" is the name of the language. We then filtered out languages that had a very low number of search results, and followed that by going through the remaining entries by hand to narrow them down to the most interesting. Since then, each year we review the list as new languages find their footing and others slip into obscurity.

https://spectrum.ieee.org/top-programming-languages-methods

TIOBE Index for July 2023

July Headline: C++ is about to surpass C

A few months ago, the programming C++ language claimed position 3 of the TIOBE index (at the expense of Java). But C++ has not finished its rise. C seems to be its next victim. At the moment, the gap between the two is only 0.76%. Apart from this, there are also some other interesting moves this month. JavaScript reaches position #6, which is an all time high. Matlab, Scratch and Rust also match their all time high records at respectively positions #10, #12 and #17. Elsewhere in the TIOBE index, we see that COBOL, after a long time, re-enters the top 20, while Julia is seriously approaching it. -- *Paul Jansen CEO TIOBE Software*

The TIOBE Programming Community index is an indicator of the popularity of programming languages. The index is updated once a month. The ratings are based on the number of skilled engineers world-wide, courses and third party vendors. Popular search engines such as Google, Bing, Yahoo!, Wikipedia, Amazon, YouTube and Baidu are used to calculate the ratings. It is important to note that the TIOBE index is not about the *best* programming language or the language in which *most lines of code* have been written.

The index can be used to check whether your programming skills are still up to date or to make a strategic decision about what programming language should be adopted when starting to build a new software system. The definition of the TIOBE index can be found here.

TIOBE Index for July 2023

Jul 2023	Jul 2022	Change	Programming Language	Ratings	Change
1	1		Python	13.42%	-0.01%
2	2		C	11.56%	-1.57%
3	4	^	C++	10.80%	+0.79%
4	3	v	Java	10.50%	-1.09%
5	5		C#	6.87%	+1.21%
6	7	^	JavaScript	3.11%	+1.34%
7	6	v	Visual Basic	2.90%	-2.07%
8	9	^	SQL	1.48%	-0.16%
9	11	^	PHP	1.41%	+0.21%
10	20	^^	MATLAB	1.26%	+0.53%
11	18	^^	Fortran	1.25%	+0.49%
12	21	^^	Scratch	1.07%	+0.35%
13	12	v	Go	1.07%	-0.07%
14	8	vv	Assembly language	1.01%	-0.64%
15	14	v	Delphi/Object Pascal	0.98%	-0.08%
16	15	v	Ruby	0.91%	-0.08%
17	29	^^	Rust	0.89%	+0.47%
18	10	vv	Swift	0.88%	-0.39%
19	**19**		**R**	**0.87%**	**+0.11%**
20	26	^^	COBOL	0.86%	+0.33%

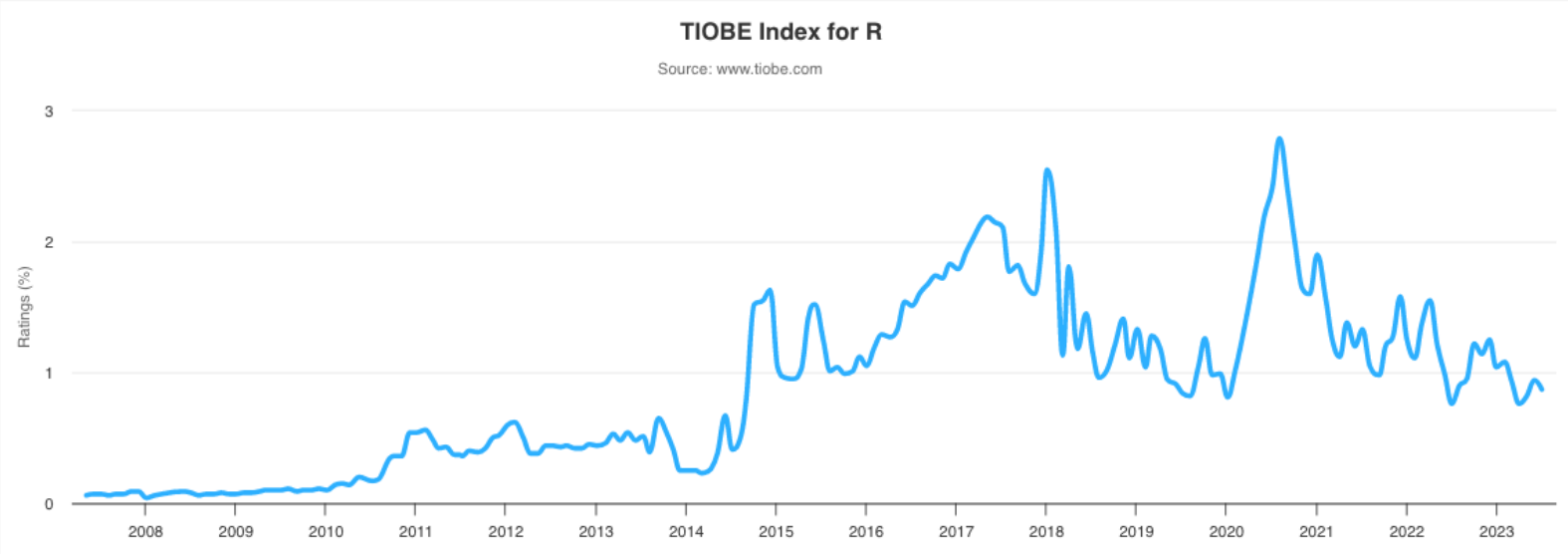
https://www.tiobe.com/tiobe-index/

The R Programming Language

Some information about R:

Highest Position (since 2007): #8 in Aug 2020

Lowest Position (since 2007): #73 in Dec 2008



https://www.tiobe.com/tiobe-index/

Hadley Wickham
@hadleywickham

Me: YAY #rstats now 8th most popular language! (tiobe.com/tiobe-index/)

Also me: hahahaa they're just telling stories about random noise (tiobe.com/tiobe-index/r/)

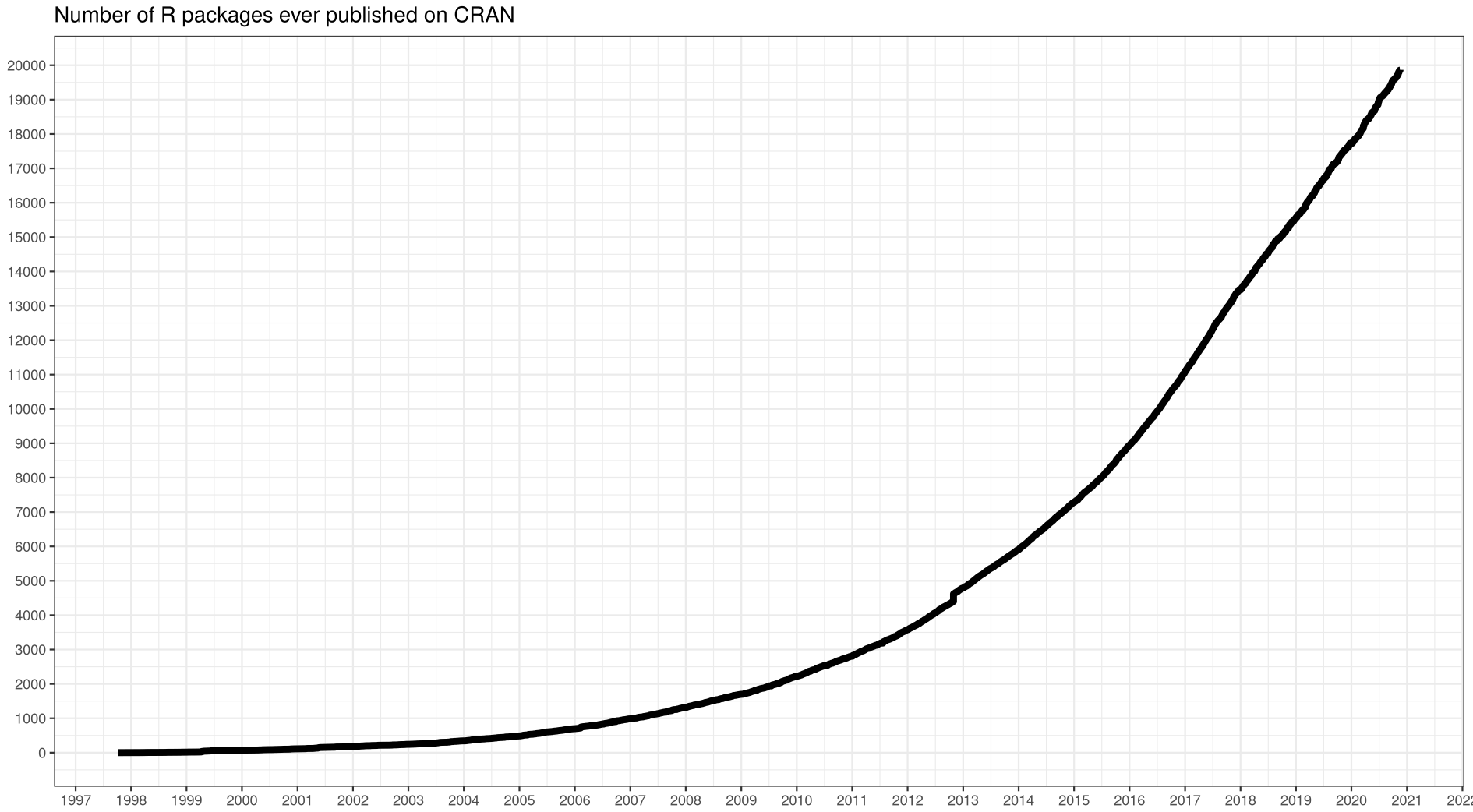
ツイートを翻訳

tiobe.com
TIOBE Index - TIOBE

午後10:23 · 2020年7月4日

https://twitter.com/hadleywickham/status/1279405379449913344?s=20

Number of R packages ever published on CRAN



https://gist.github.com/daroczig/

The Comprehensive R Archive Network (CRAN)

Available Packages

Currently, the CRAN package repository features 19823 available packages.

https://cran.r-project.org/web/packages/

CRAN R Packages by Number of Downloads

UPDATED DAILY. Last updated: 2023-07-13 00:14:23 +1000 Melbourne time (about 6 hours ago)

Rank	Package Name	Downloads	Author	Maintainer
1	ggplot2	117,224,283		
2	magrittr	111,960,957		
3	rlang	110,952,627		
4	dplyr	90,711,986		
5	vctrs	78,103,768		
6	tibble	76,505,415		
7	jsonlite	75,273,609		
8	cli	74,670,942		
9	Rcpp	72,619,671		
10	devtools	72,377,709		

https://www.datasciencemeta.com/rpackages

無償プログラミング言語：R

➡ 何の心配もいらないので
どんどん学んでいきましょう。

- データ科学とは
- データ科学の倫理
- データ科学のツール
- Rを始めよう

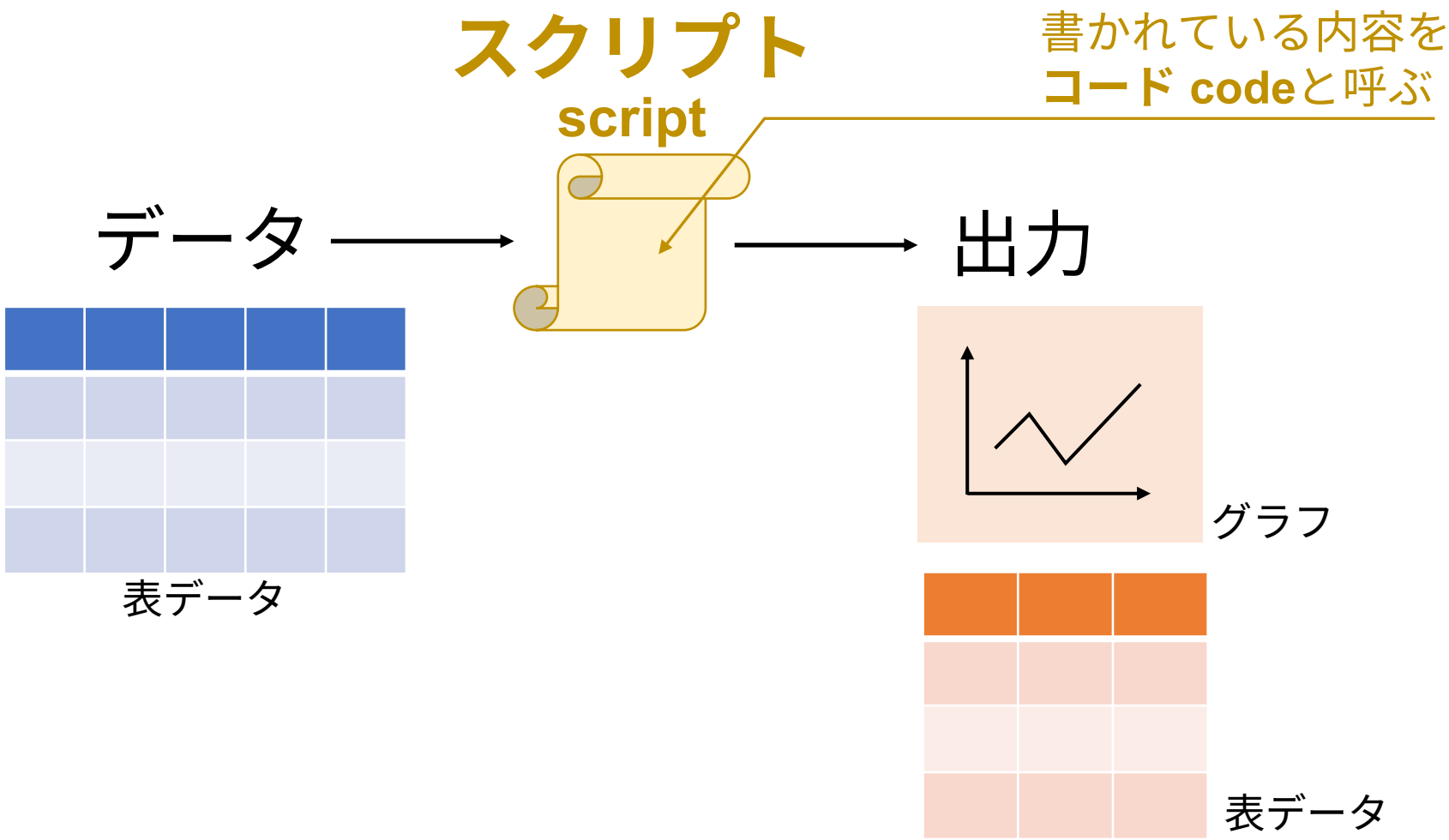
プログラミング言語を使おう

基本的な姿勢

結果ではなく解析の**工程**を
明示的に組み上げる

- 自ずと結果も共有される。
- 追加処理が容易になる。
- デバックしやすい。

R = スクリプト言語



R = スクリプト言語

スクリプト

```
## packages ----
library(tidyverse)

## parameters ----
r <- 2
x_center <- 1
y_center <- 3

## data ----
df <-
  tibble(
    theta = seq(0, 2 * pi, length = 361),
    X = r * cos(theta) + x_center,
    Y = r * sin(theta) + y_center
  )

## visualization ----
ggplot(data = df) +
  aes(x = X, y = Y) +
  geom_path() +
  coord_fixed()
```

■ Rを始めよう

- 基礎知識(オブジェクト, 関数)
- データの読み書き
- データの加工
- データの可視化

to be continued...