すごいデータサイエンス入門書の話

みんなのPython勉強会 #76
辻真吾（@tsjshg）

お前、誰よ？（自己紹介）

- 大学の研究所に勤めています
  - エネルギーシステムとバイオインフォマティクス
- 2005年ごろJavaからPythonに乗り換えました
  - 2010年に「Pythonスタートブック」初版をだしてから技術書を何冊か執筆しています
- この1年で一番の思い出は「7月にRustに入門したこと」
- 最近読んで面白かった本「宇宙の終わりに何が起こるのか（講談社）」
- www.tsjshg.info

【実践Data Scienceシリーズ】

ゼロからはじめる

データサイエンス入門

―― R・Python一挙両得 ――

辻　真吾
矢吹太朗 著

RとPython両方学べる。コスパ最強の一冊!

コードが理解の試金石!

- 「データサイエンスの準備」にページを割いているから、プログラミング経験ゼロで大丈夫!
- 自分に合った言語を見つけたい、言語を乗り換えたいという方にもおすすめ!

講談社

おしながき

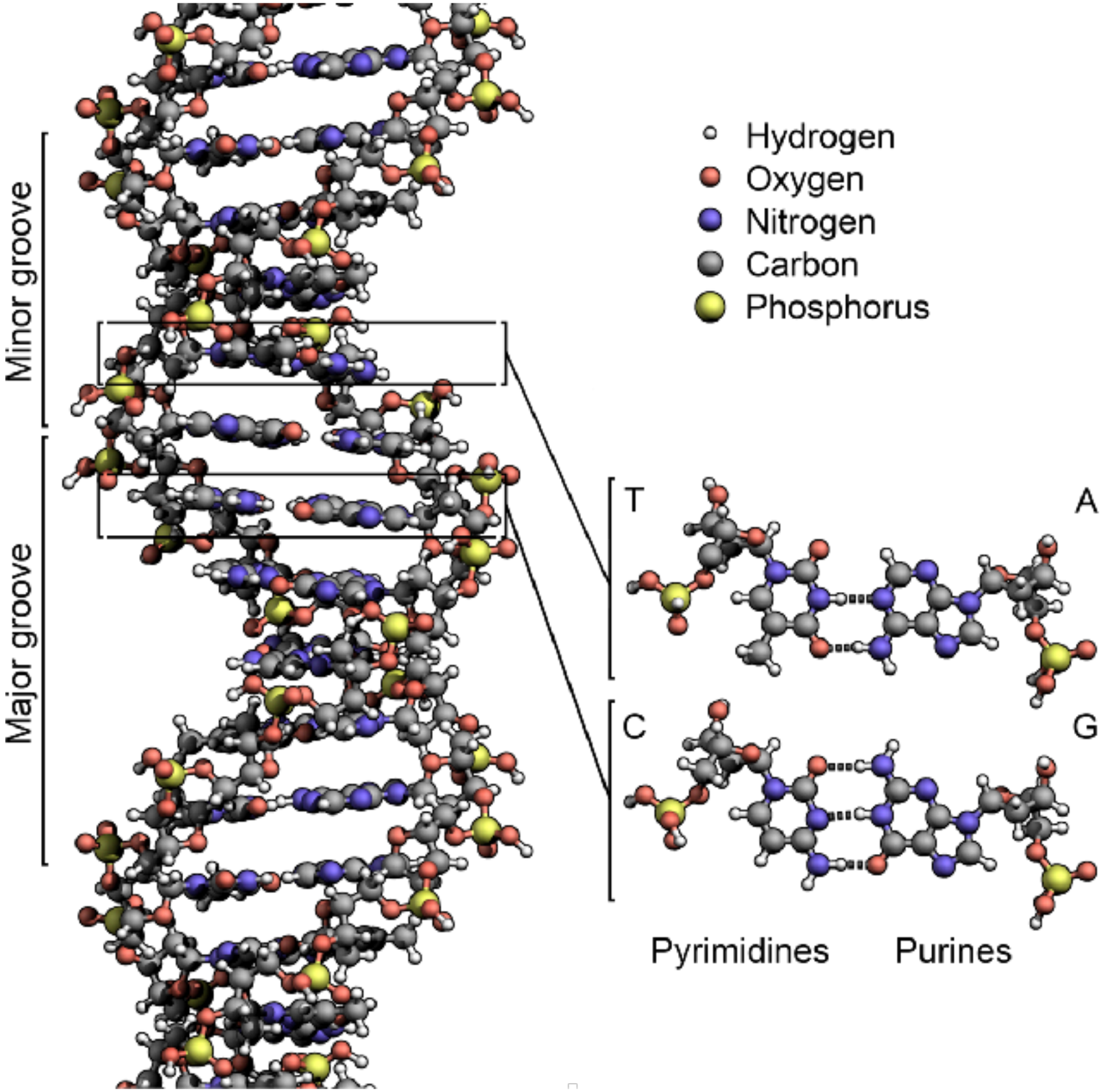
- データサイエンスとは？
- この本のすごいところ（まじめな話）
- この本のすごいところ（くだけた話）

データサイエンスってなに？

データサイエンスは「データ駆動型サイエンス」に由来

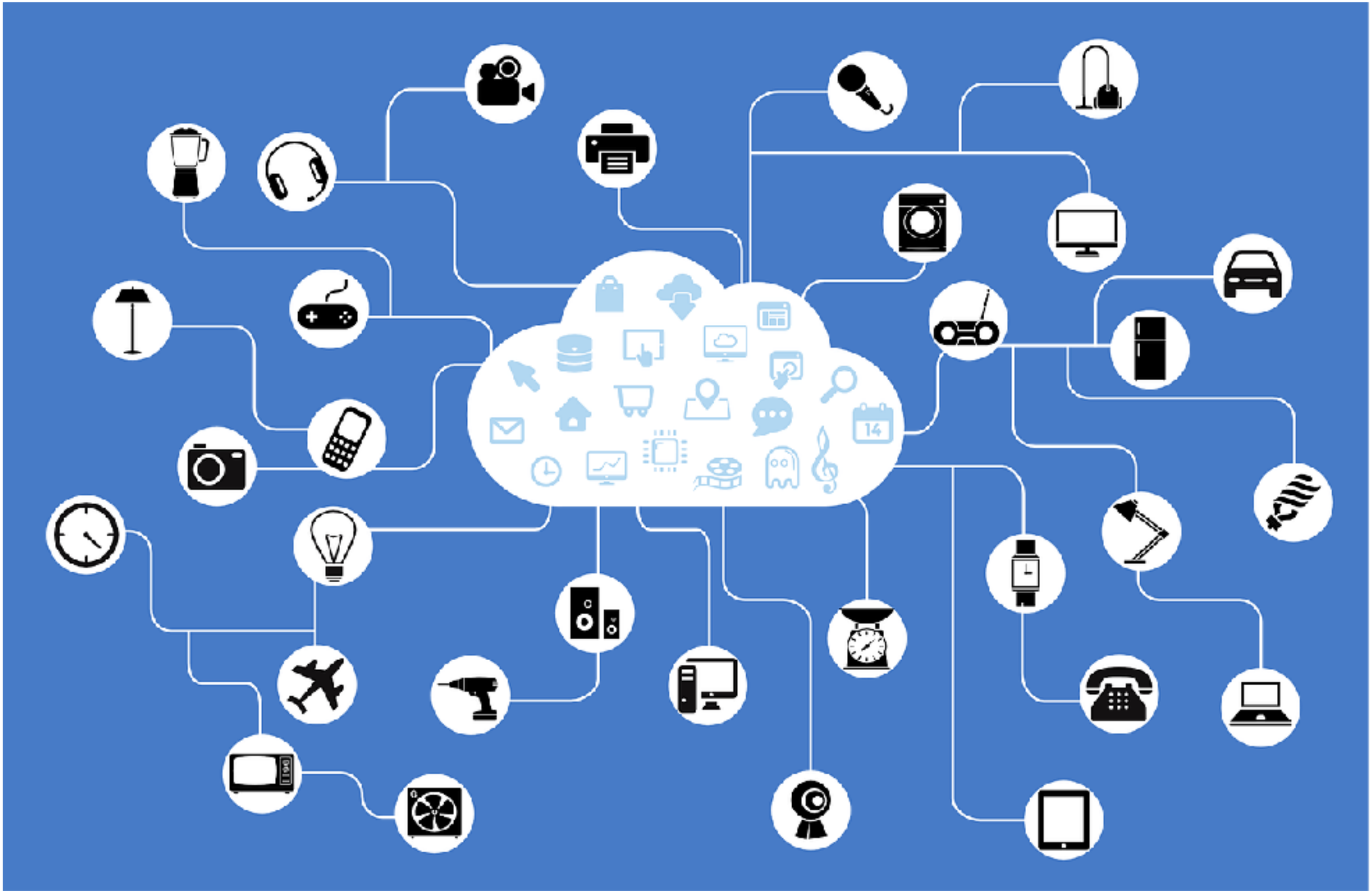
観測データを分析すると動きが全然違う惑星を発見できる

いま何がおきているか？



ヒト1人1つの細胞だけで32億文字

いま何がおきているか？



すべてのモノがインターネットに繋がって情報が蓄積される

なにが問題か？

- データが多すぎる・・・
- 夜空をぼーっと眺めていて惑星の存在に気が付くレベルの量では無い
- データ駆動型のサイエンスやデータ駆動型のビジネス（意志決定）を実践するにはスキルが必要
- それがデータサイエンス

データサイエンスを実践するには？

- コンピュータとネットワーク
- プログラミング
- データサイエンス固有の知識

本書で一通り網羅できる

- コンピュータとネットワークと環境構築（1, 2章）
- RとPythonの基本文法（3章）
- 統計入門と前処理（4, 5章）
- 機械学習アルゴリズム（6, 13章）
  - 深層学習と時系列データ解析を含みます

TCP/IPの基本もわかるようになる

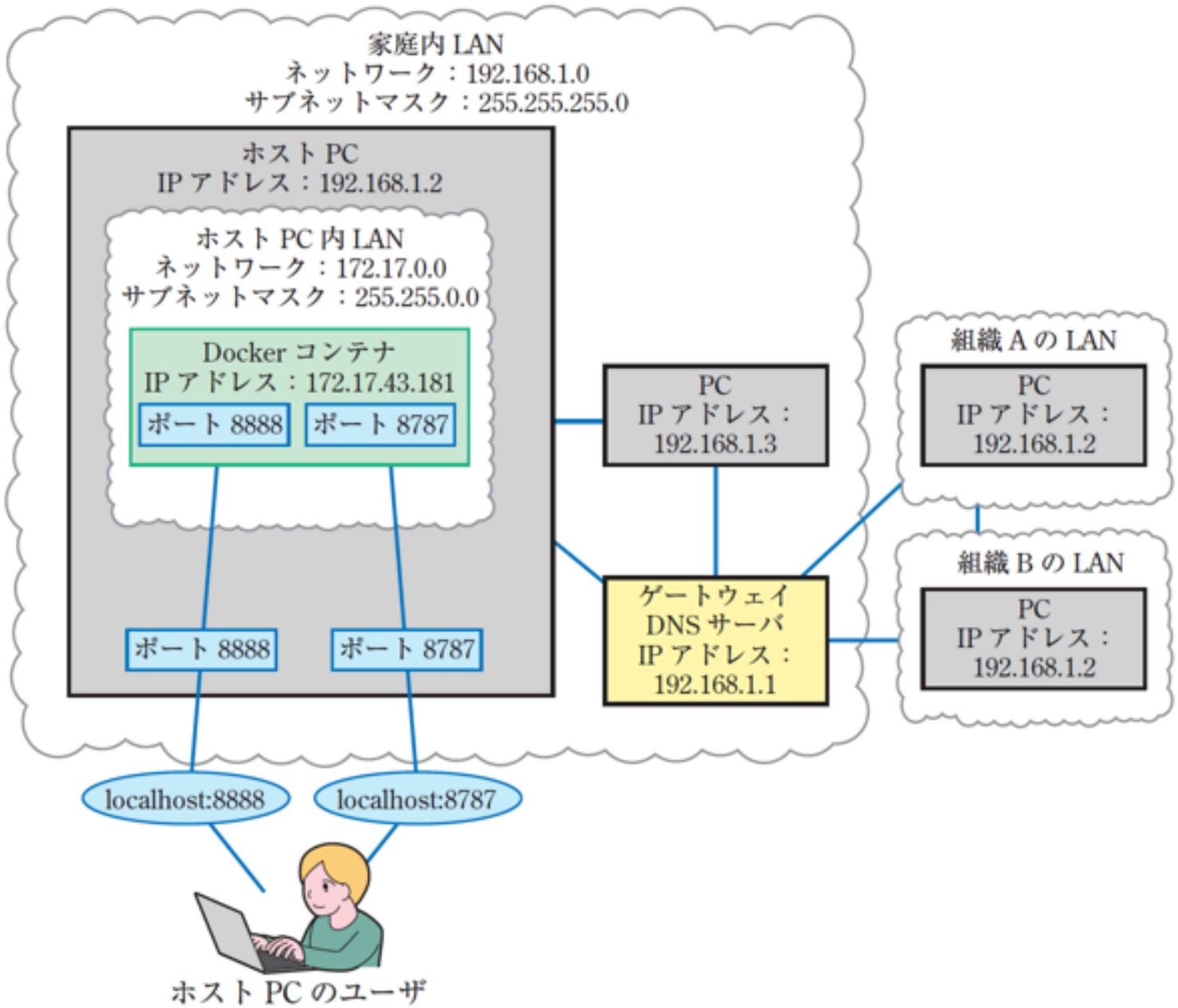


図 1.4　LAN について理解するための素材

P15から抜粋

Dockerについて詳しく解説

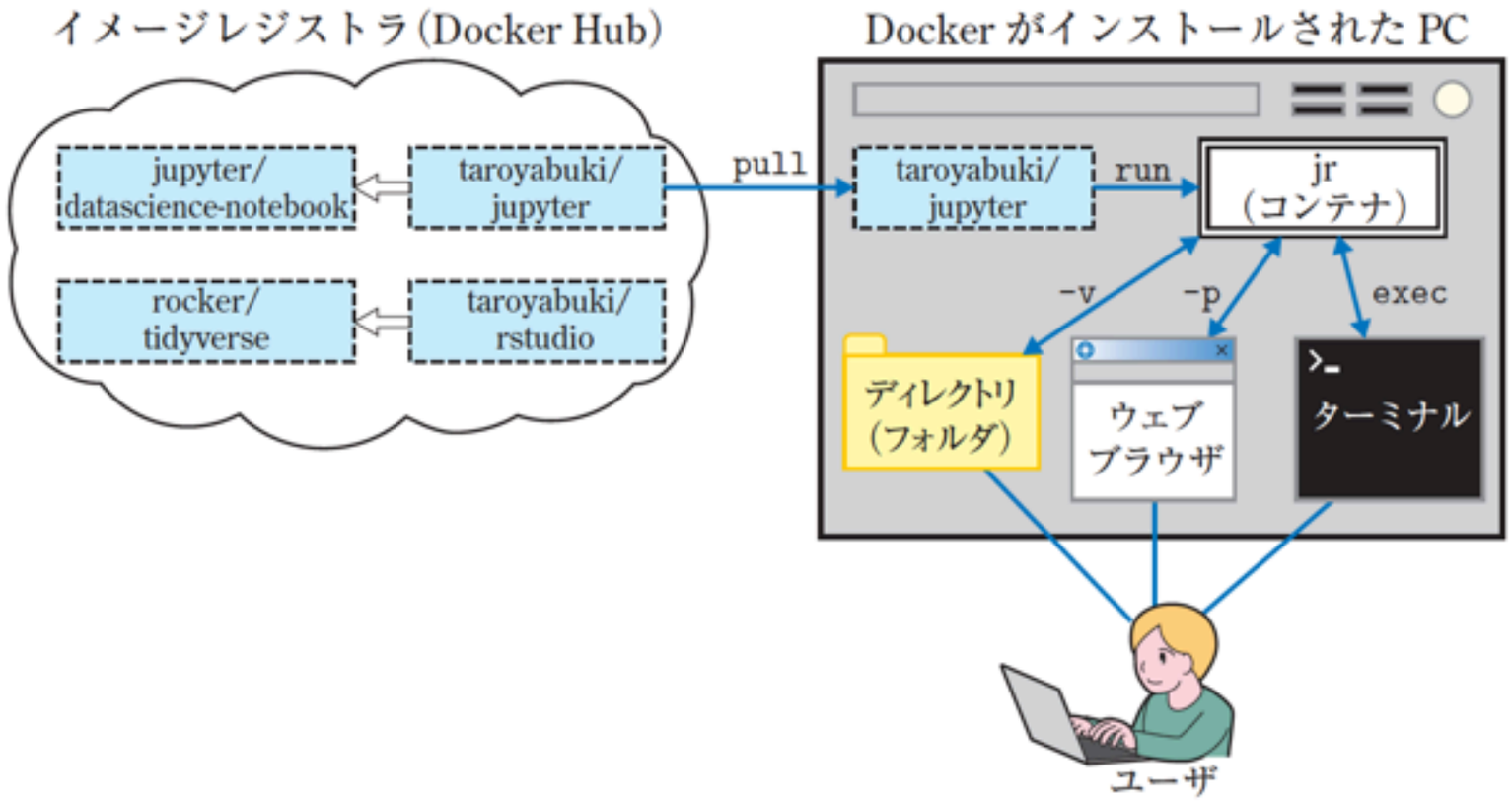


図 2.2　Docker の全体像（破線の長方形がイメージ，二重線の長方形がコンテナ）

P23から抜粋

ほぼすべてのコードをRとPythonで記述

R

```
1:5
#> [1] 1 2 3 4 5
```

Python

```
list(range(5))
#> [0, 1, 2, 3, 4]
```

0 以上 10 以下（11 未満）の偶数の 1 次元データを作ります．

P: わかりやすさのために range の結果をリストに変換しています．リストに変換しなくてもリストと同じように使える場面が多いです．

R

```
seq(from = 0, to = 10, by = 2)
#> [1]  0  2  4  6  8 10
```

Python

```
list(range(0, 11, 2))
#> [0, 2, 4, 6, 8, 10]
```

0 から 1 まで間隔 0.5 の 1 次元データを作ります．

P: range の 2 番目の引数は結果に属さないことに注意．

R

```
seq(from = 0, to = 1, by = 0.5)
#> [1] 0.0 0.5 1.0
```

Python

```
import numpy as np
np.arange(0, 1.1, 0.5)
#> array([0. , 0.5, 1. ])
```

0 から 100 までを分割し，5 個の数値からなる 1 次元データを作ります．

P: 引数が整数でない場合は np.arange を使います．

R

```
seq(from = 0, to = 100, length.out = 5)
#> [1]   0  25  50  75 100
```

Python

```
np.linspace(0, 100, 5)
#> array([  0.,  25.,  50.,  75., 100.])
```


P67から抜粋

RとPython（ロジスティック回帰の例）

R

```
library(caret)
library(PRROC)
library(tidyverse)

my_url <- str_c("https://raw.githubusercontent.com/taroyabuki",
                "/fromzero/master/data/titanic.csv")
my_data <- read_csv(my_url)

my_model <- train(form = Survived ~ ., data = my_data, method = "glm",
                  trControl = trainControl(method = "LOOCV"))
```

Python

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score, LeaveOneOut
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import OneHotEncoder

my_url = ('https://raw.githubusercontent.com/taroyabuki'
          '/fromzero/master/data/titanic.csv')
my_data = pd.read_csv(my_url)

X, y = my_data.iloc[:, 0:3], my_data.Survived

my_pipeline = Pipeline([('ohe', OneHotEncoder(drop='first')),
                        ('lr', LogisticRegression(penalty='none'))])
my_pipeline.fit(X, y)
```

P297から抜粋

RとPythonの本質的な違いにも言及

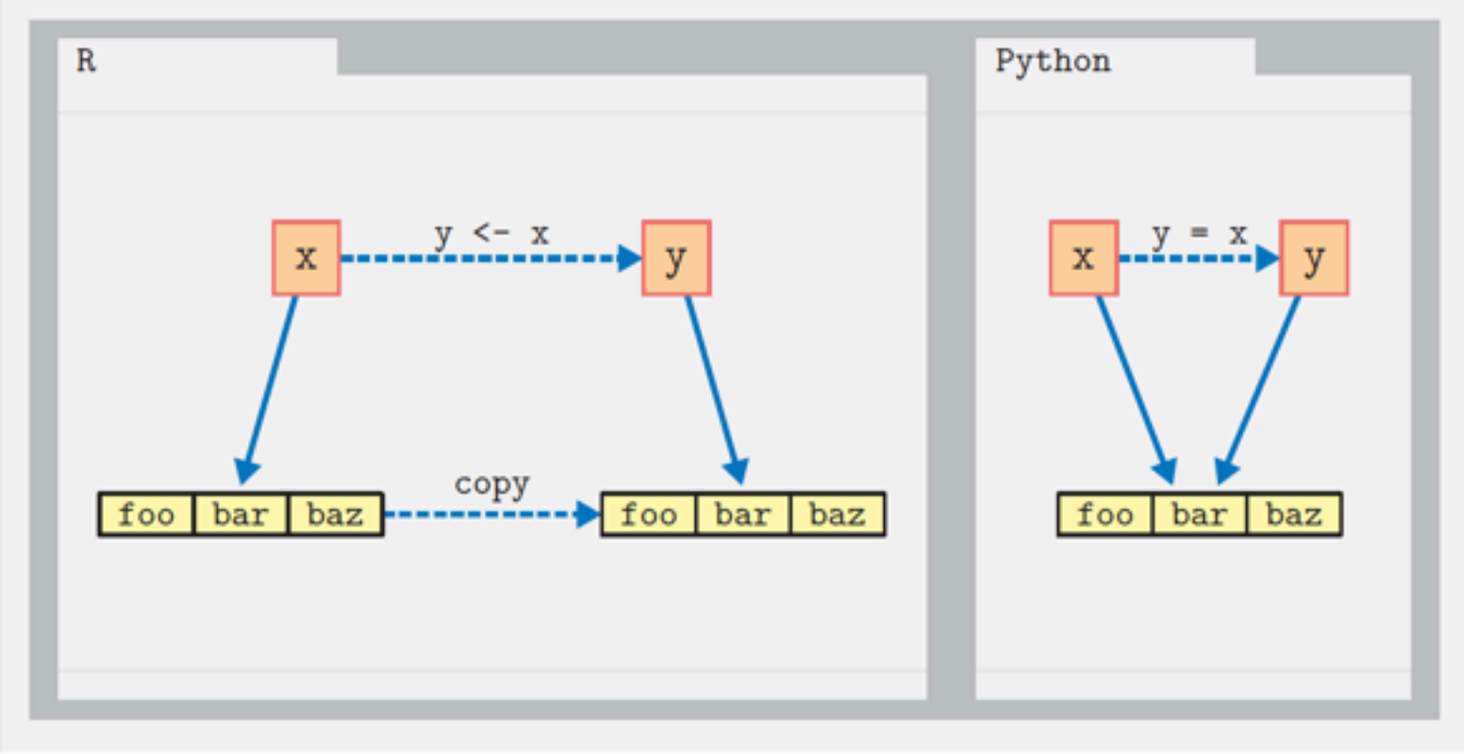


図 3.1　R と Python での「割り当て」の違い

P72から抜粋

この本のすごいところ（くだけた話）

著者の2人は小学校1年生からの友達

辻真吾（私）

担任の
酒井先生

矢吹太朗
（共著者）

1982年4月　東京都足立区立千寿第五小学校入学式の写真

執筆に5年を要した

- 普通1年〜2年（当初とだいぶ内容が変わった）
- それぞれ、5年の間に別の本を企画して出している
    - やる気あんのか？
- 完成を待ち望んでいた（かどうかはわからない）私の母は3年前に他界😅
- いつまでもあると思うな親と金

というわけで

- 総ページ数400ページで3,520円（ちょっとお安い）
- コンピュータとネットワーク→プログラミング→統計と前処理→機械学習アルゴリズム
  - ゼロからこの流れに乗れるようになっています
- （謝辞）『ディープラーニング　学習する機械　ヤン・ルカン、人工知能を語る』が大好評発売中の横山真吾さん（講談社サイエンティフィク）には大変お世話になりました

講談社サイエンティフィクの超敏腕編集者
横山真吾さんから3冊プレゼントがあります

応募方法

https://twitter.com/tsjshg/status/1470682849074622468

このツイートに**思い浮かべること#stapyを付けて**引用リツイートしく
ださい（関係者をフォローすると当選確率が上がるかも！？）

12月21日(火)23:59締め切り（の予定）

当選者は私のツイートでメンションします（クリスマス頃を予定）

辻へメール（shingo.tsuji@gmail.com）かツイッターDMで発送先お知ら
せください

今年も1年「みんなのPython勉強会」にご参加いただきありがとうございました。
来年もよろしくお願いいたします！