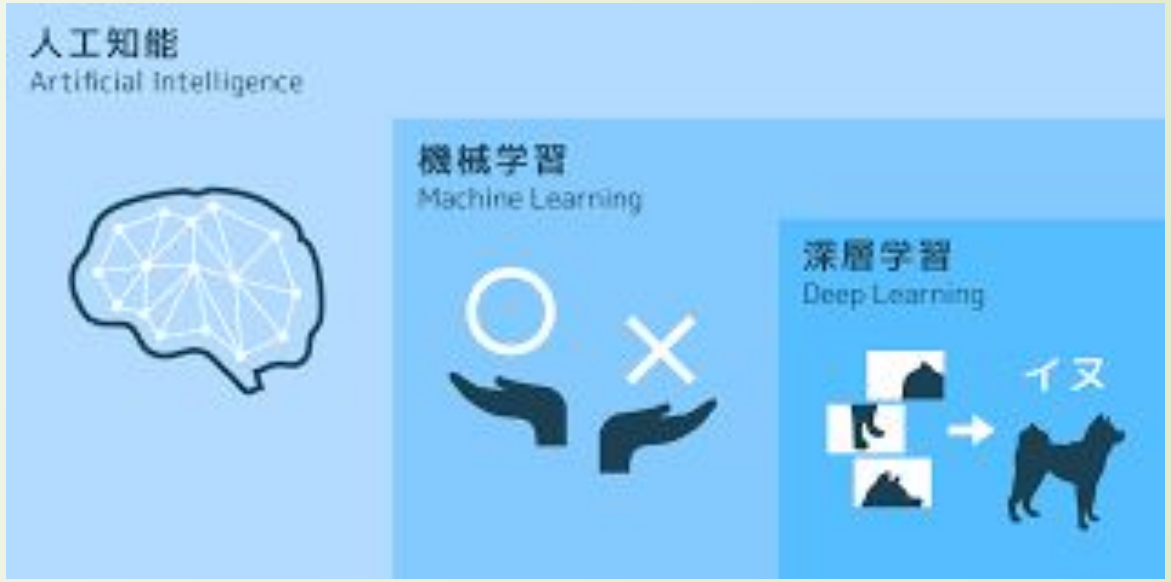
機械学習とCNNの触り

b2 安部政俊　山拓さんの画像をお借りしました。

AIって何？

大量のデータから“特徴”を抽出し
予測判断を行う

ex)犬の“犬らしさ”
　　ワインの“美味しさ”の根源

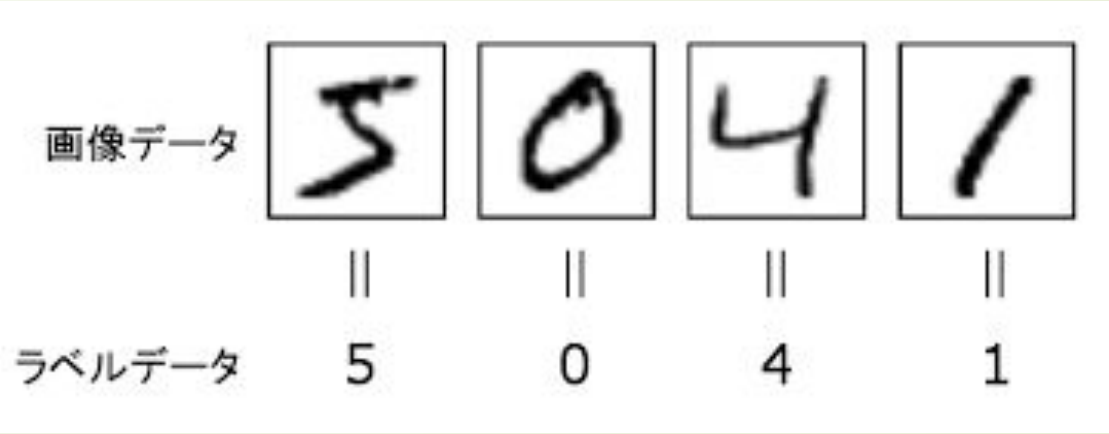


教師あり　と　教師なし

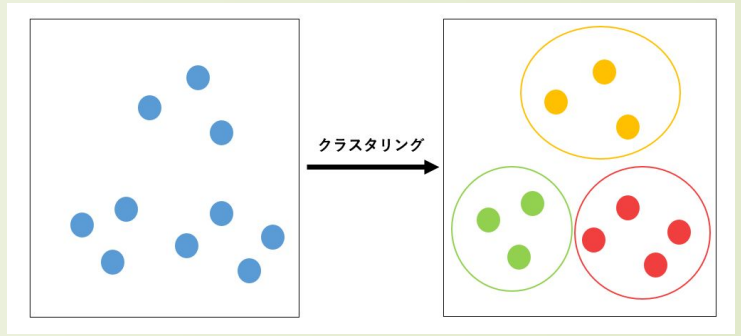
教師あり学習：**問題と解答のセット**を与えて正しい解答を導くようコンピュータが学習。問題（データ）に対する解答を**正解ラベル**と呼ぶ

人間の思い通りに動いて欲しいときなどに使う

e.g...値の予測、手書き文字から数字0～9を分類



教師あり　と　教師なし

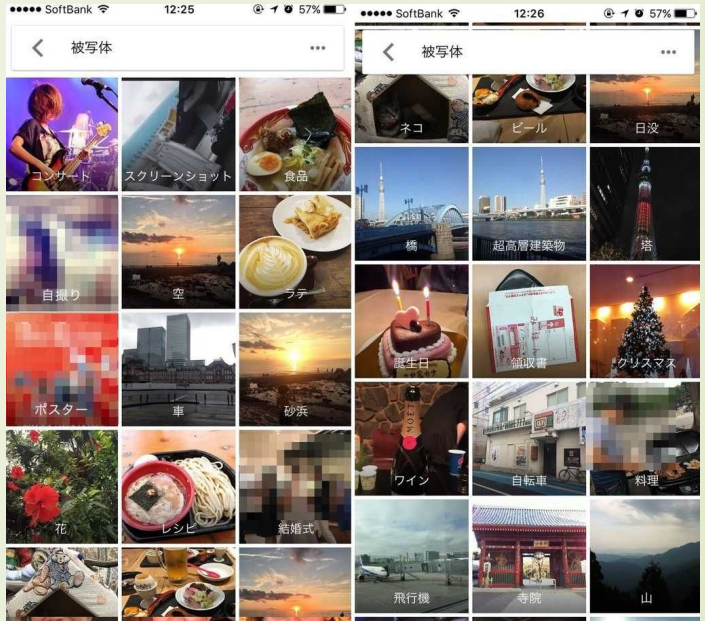


教師なし学習:問題しか与えない。問題（データ）の**分類や特徴の抽出**が得意

人間にはよくわからないことを見つける、正解を与えるのが難しいときに使う

e.g....クラスタリング（たくさんあるデータをコンピュータが何種類かに分類）

...google photoで写真が人物別に分類されているのかクラスタリングを用いている



機械学習で扱うタスクについて

- ①回帰タスク…物の値段来客数などの値を予想
- ②分類タスク
- -1…二値分類…作物が病気にかかっているか否かなどを予想。0or1の2種類どちらかを予測する、あるいは0～1の間の確率として予測（確率0.3なら30%で1に分類される）、の2通り
- -2…多クラス分類…複数のクラスのうちどれかに属すマルチクラス分類と同時に複数のクラスに属するマルチラベル分類に分かれる。マルチラベルでは二値分類を複数回するイメージ

機械学習で扱うタスク

③生成…画像を入力して似た画像を出力（GAN）

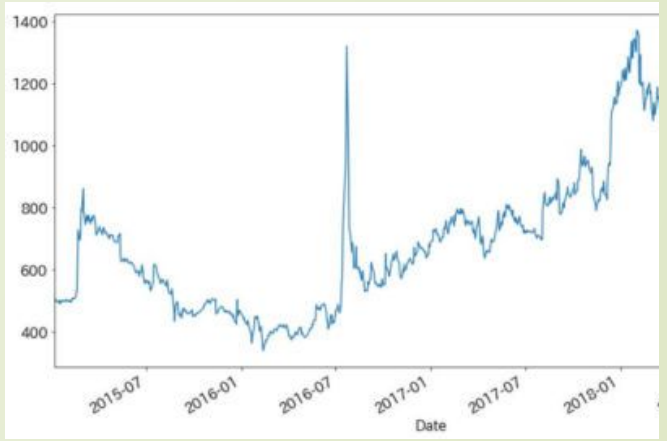
文章を入力してそれに即した画像を、逆に画像を入力して画像に対する適切な説明を出力

画像キャプション生成例 [Ushiku+, ICCV 2015]

Tourists are standing on the middle of a flat desert.

機械学習で扱うデータの種類

- ①テーブルデータ…表形式のデータ。複数与えられることもありその場合はテーブルを結合したりそのあとに処理をしたりしないといけない。
- ②時系列データ…時間の推移とともに観測されたデータのこと。心電図、音声データ、ユーザーの閲覧、購入履歴など
- ③画像や自然言語データ…自然言語とは人間が喋る普通の言葉のこと、機械翻訳で扱うデータはこれ



EMP_NAME	EMP_KANA	LOGIN_PASSWORD	TEL	FAX	DEPT_CODE	START_DATE
小杉優	こすぎゆう	MYLFPG	999-999-9999	999-999-9998	MBLH	2016/09/06 1
本村海翔	もとむらかいと	MIE	999-999-9999	999-999-9998	LMGPF	2016/09/06 1
藤岡優奈	ふじおかゆうな	LXBDTKL	999-999-9999	999-999-9998	OR	2016/09/06 1
福島康平	ふくしまこうへい	WDV	999-999-9999	999-999-9998	CNLQ	2016/09/06 1
吉井若菜	よしいわかな	XGFCEBGC	999-999-9999	999-999-9998	ERYUJ	2016/09/06 1
安藤奈央	あんどうなお	G	999-999-9999	999-999-9998	DZUN	2016/09/06 1
藤村暢子	ふじむらのぶこ	BQKZWQ	999-999-9999	999-999-9998	UQDLO	2016/09/06 1
中沢雅也	なかざわまさや	GVH	999-999-9999	999-999-9998	FMG	2016/09/06 1
森下要一	もりしたよういち	JMFEIYWM	999-999-9999	999-999-9998	WFD	2016/09/06 1
阿久津健司	あくつけんじ	XCBBLCH	999-999-9999	999-999-9998	JCAD	2016/09/06 1
西浦綾	にしうらあや	WZET	999-999-9999	999-999-9998	GE	2016/09/06 1
天野美結	あまのみゆ	KHHME	999-999-9999	999-999-9998	YQDX	2016/09/06 1
及川和馬	おいかわかずま	PXJTLRA	999-999-9999	999-999-9998	TW	2016/09/06 1

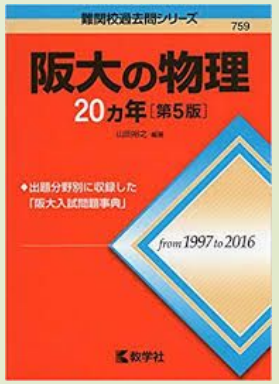
学習のためにデータを分ける？

全データを学習用（訓練用と検証用）データ、とテストデータに分けて学習を進めていく。今は無きセンター試験でイメージを掴むと、

学習用データ：昔の過去問を家でやって問題の傾向と対策を掴む。検証用を模試のように活用し、どう学習すれば良い点が取れるかを考えながら学習用の問題を解く

テストデータ：直近数年分の問題をしっかり時間を計って本番のつもりで

問題集



模試

データの分け方大丈夫？

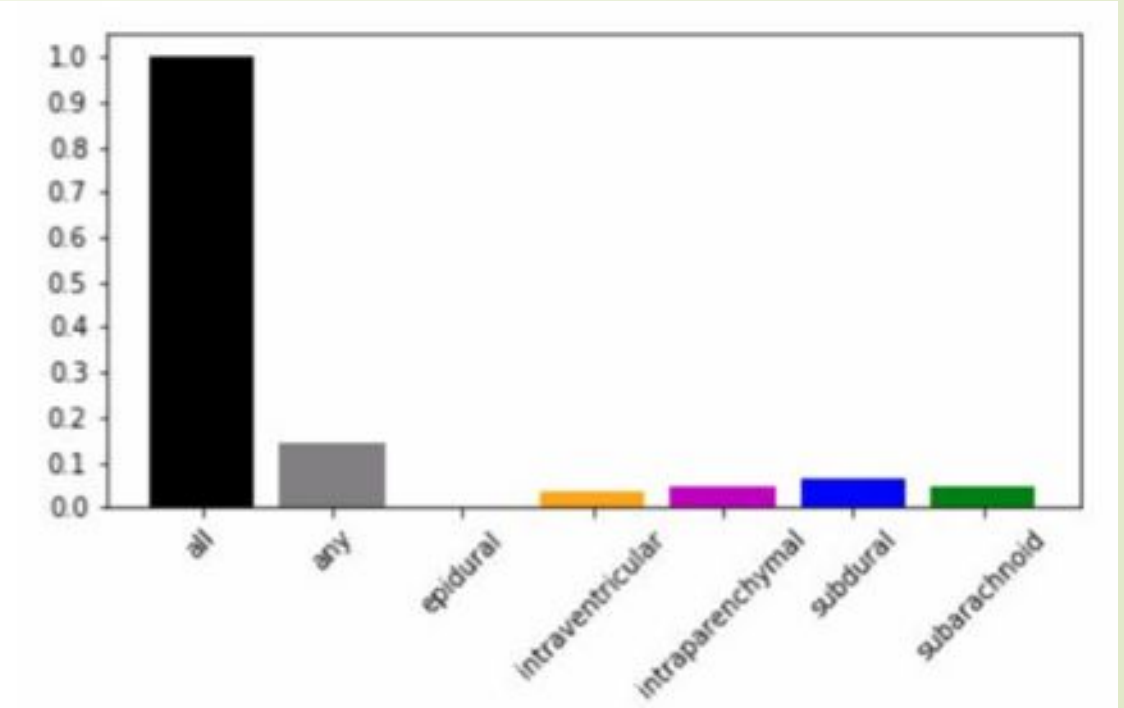
センター試験の例でいうと、データは時系列性がある（〇〇年度の問題というように）ので検証用を古い問題にして古い過去問がよくできるように学習してしまうと直近のテスト用データには対応できない！

あるいは、検証用の分け方によってたまたま良い精度になってるだけかもしれない

データの分け方大丈夫？

データが偏っている場合（1%で病気など）検証用に病気の例が入ってなかったら全部健康！と予想すれば良い精度にはなる（100%）が

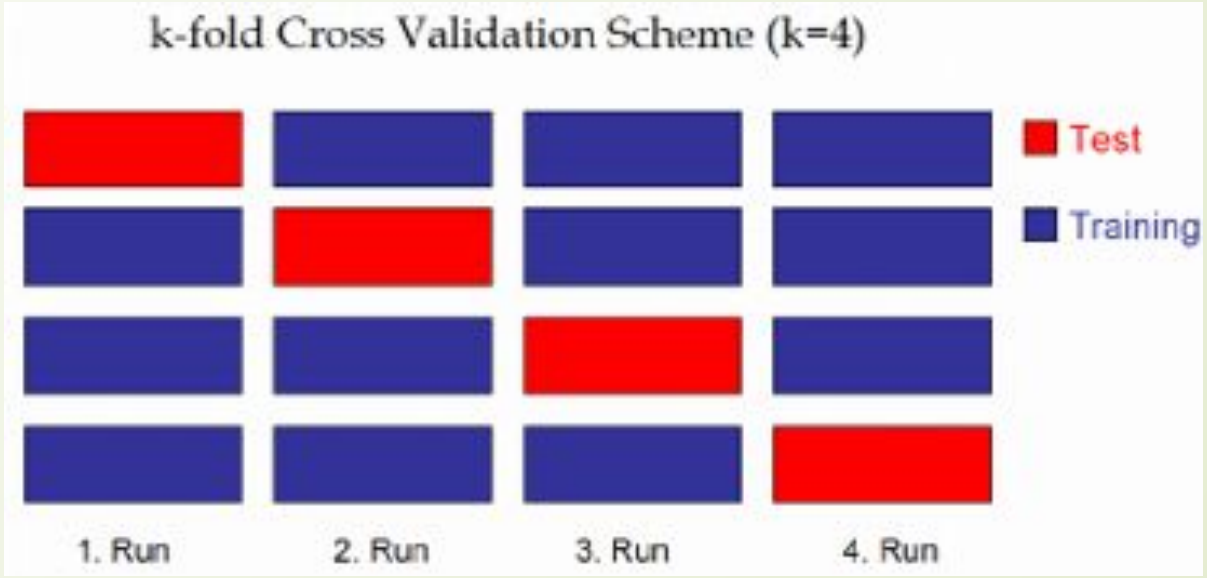
良いモデルとは言える？



陽性のデータ自体は結構少なかった。

しっかりとした分け方を！

検証用と訓練用の分け方による"ガチャ"のような精度にならないように学習用データを1回ずつ数回に分けて検証用に回す　**Kfold交差検証**、というものを用いる。



データに偏りがある場合は、偏りを考慮して層化抽出（検証用と元のデータの分布が同じになるように）をする。

学習しすぎも良くない？

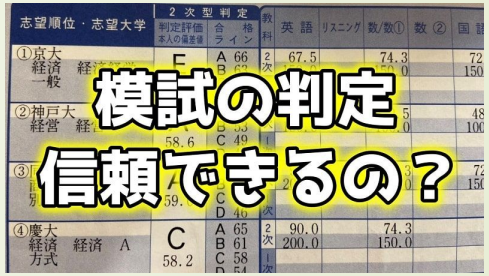
適切にデータを訓練用、検証用、テストデータに分けたとして

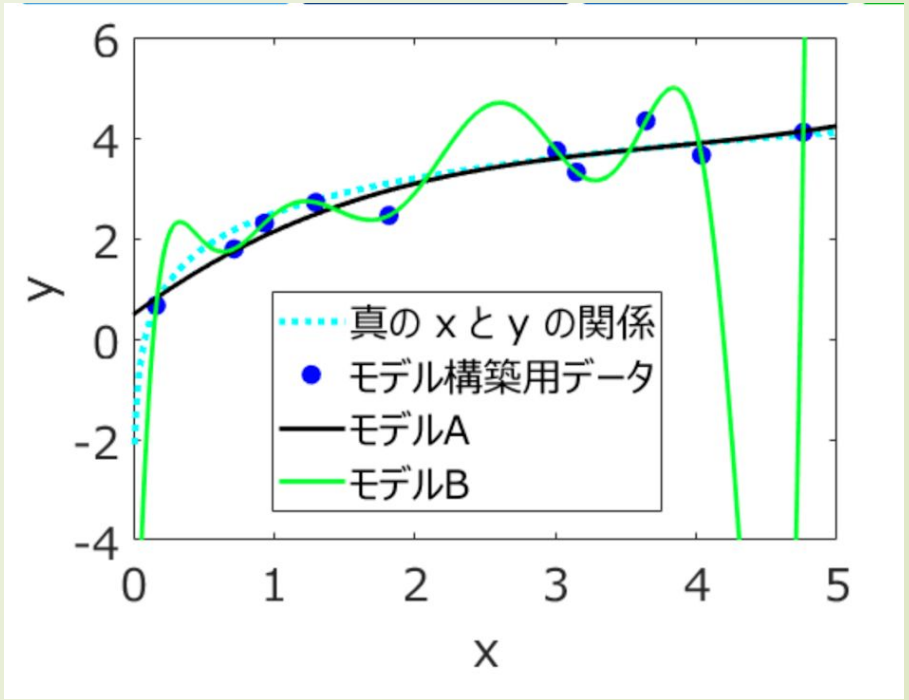
学習中：訓練用に対していくら高精度でも

モデルの評価について：検証用に対していくら高精度でも

未知データに対しての精度が良くないと使えない！

→このように未知のテストデータに対する識別能力を**汎化性能**といい、汎化性能が乏しく、既知のデータについての精度ばかりが上がってしまっている状態を**過学習**をしている、という。





汎化性能を担保するためには？

過学習を防ぐために一般的に行われるのは**正則化**というもの

パラメータを色々変えてうまい出力を得ようとするのが機械学習の"学習"であったが

パラメータの値が大きくなりすぎてモデルが複雑になりすぎないように下のような関数を加える

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \|h_\theta(x^{(i)}) - y^{(i)}\|^2$$

これは，単純に各学習データの計算結果と期待値の誤差の二乗の平均値である．

この式に，正則化を加えると，

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} \|h_\theta(x^{(i)}) - y^{(i)}\|^2 + \lambda \sum_{j=1}^{n} (\theta_j)^2$$

というようにパラメータΘの二乗和を追加するのである．この項を正則化項と呼ぶ．

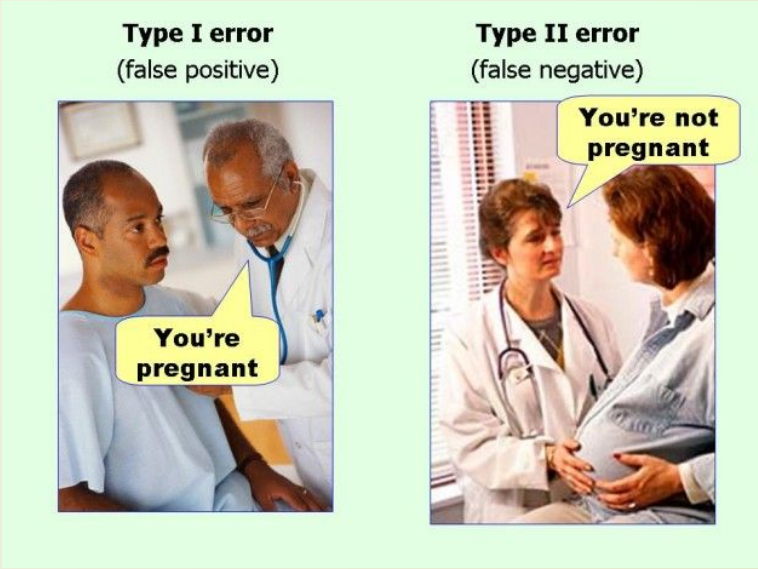
評価指標とは？

予測値の精度を測るモデルの良し悪しのモノサシ

分類タスクでもaccuracy（全予測に対する正解の割合）だけが精度を測るものではない

行うタスクや、データの分布、求める性能によって選択する。

	正解で正	正解で負
予測で正	真陽性 TP: True Positive	偽陽性 FP: False Positive
予測で負	偽陰性 FN: False Negative	真陰性 TN: True Negative



選ぶべき評価指標？

例1：回帰タスクにおいて、予測値と正解の値のは大きく外れて欲しくない！

→MAEでは外れ値に対する罰則がRMSEより大きいのでこちらを選択

Root Mean Squared Error (RMSE)

RMSE は、後述する MAE と共に平均化された誤差の値を表します。

$$RMSE = \sqrt{\frac{\sum_i (y_{\mathrm{obs},i} - y_{\mathrm{pred},i})^2}{n}}$$

ここで、n はサンプル数です。

Mean Absolute Error (MAE)

MAE は以下の式で計算され、RMSE と共に平均化された誤差の大きさを表します。

$$MAE = \frac{\sum_i |y_{\mathrm{obs},i} - y_{\mathrm{pred},i}|}{n}$$

対数平均二乗誤差（RMSLE)

- ある値は10、ある値は10億といった、値のレンジが大きな（対数正規分布に近い）データの学習に利用します。
- 先程のRMSEと見比べると下記のような感じ。
  - 対数（log）を計算してから、予測と実績を引き算。
  - 予測と実績の誤差を幅ではなく比率として表現
  - 対数を取る前に、予測、実績共に＋１
    - 予測または、実績が0の場合、log(0)となり計算できなくなるので、＋１する。

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(Pred_i + 1) - log(Act_i + 1))^2}$$

選ぶべき評価指標？

2値分類において

例2：偽陰性（FN）は出て欲しくない→陽性と判断したものの中で真の陽性
の割合（**再現率**）を最大にすればOK!

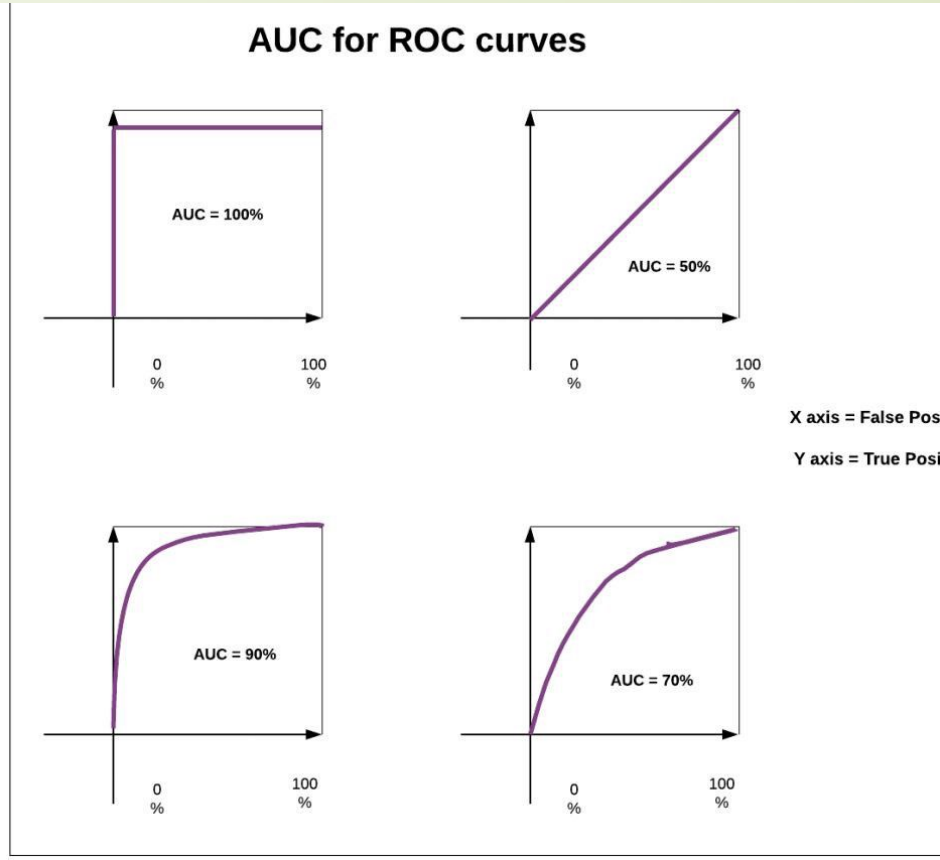
例3：偽陽性（FP）は出て欲しくない→陰性と判断したものの中で真の陰性
の割合（**適合率**、感度）を最大にすればOK!

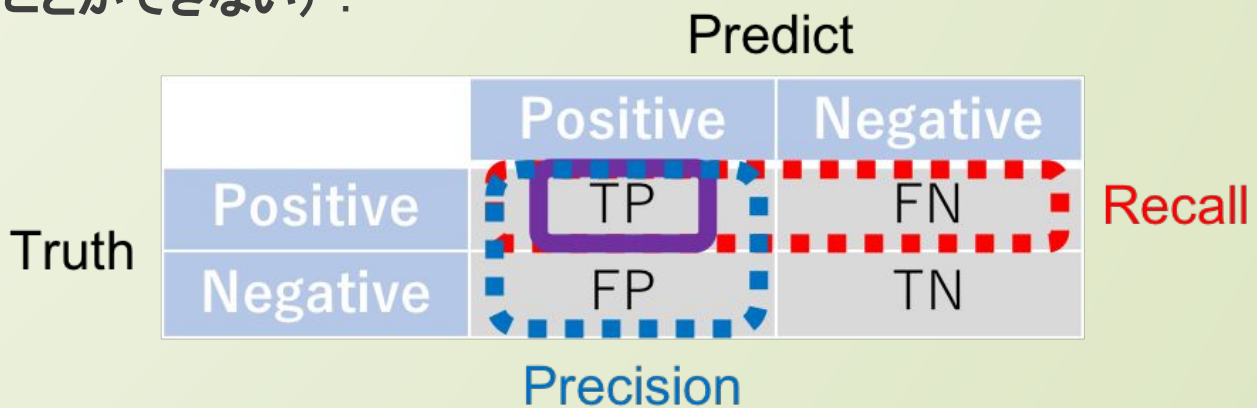
例4：間違い方は構わないので、とりあえず正解数を増やしたい→全体の
の正解の割合を最大にすれば良い！

例5：予測は確信度（確率）であり、閾値によらない精度が欲しい→**AUC**

～～注意！～～

再現率と適合率は**トレードオフ（どちらも追い求めることができない）**！

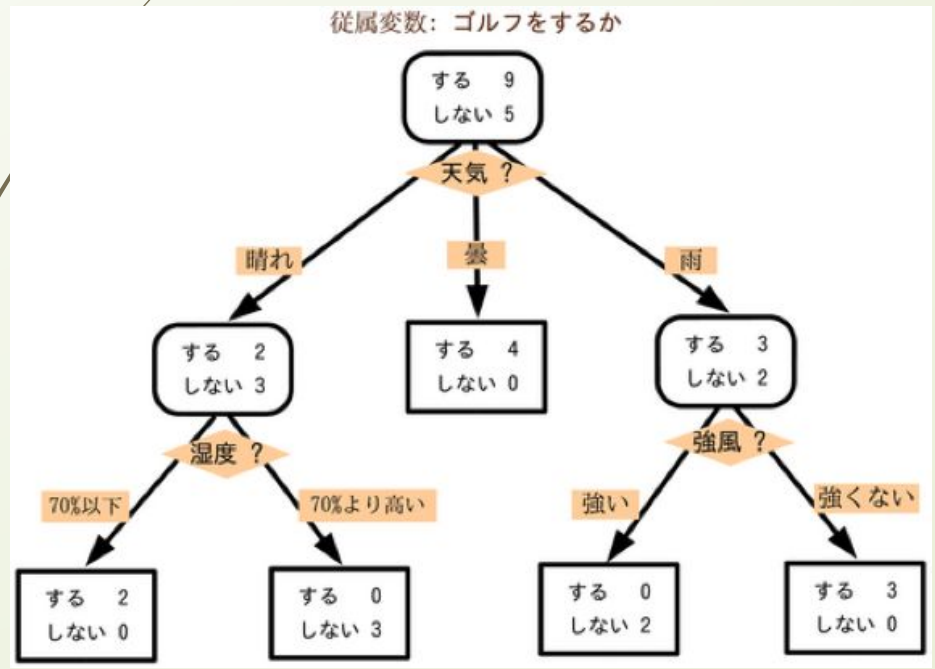


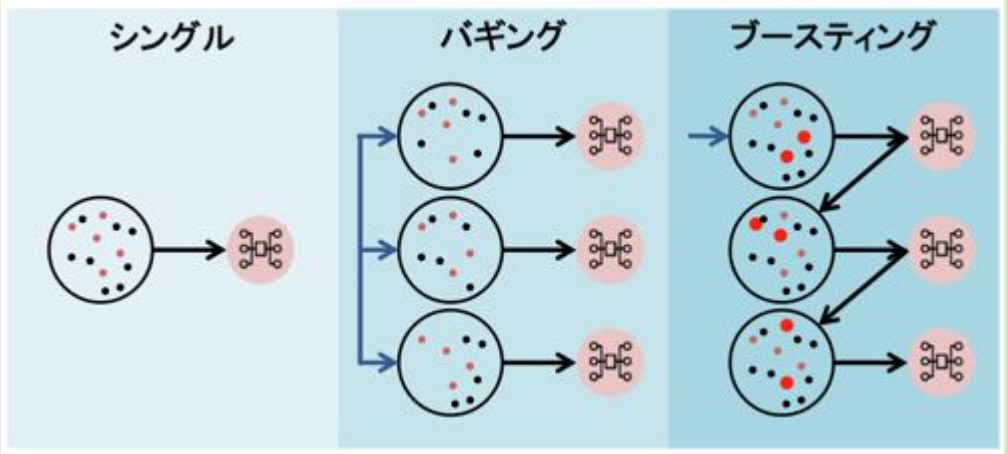


機械学習のモデルについて

使っておけば無難なのはGBDT（勾配ブースティング木）テーブルデータ（excelの表など）の扱いに長ける

（LightGBMなど）

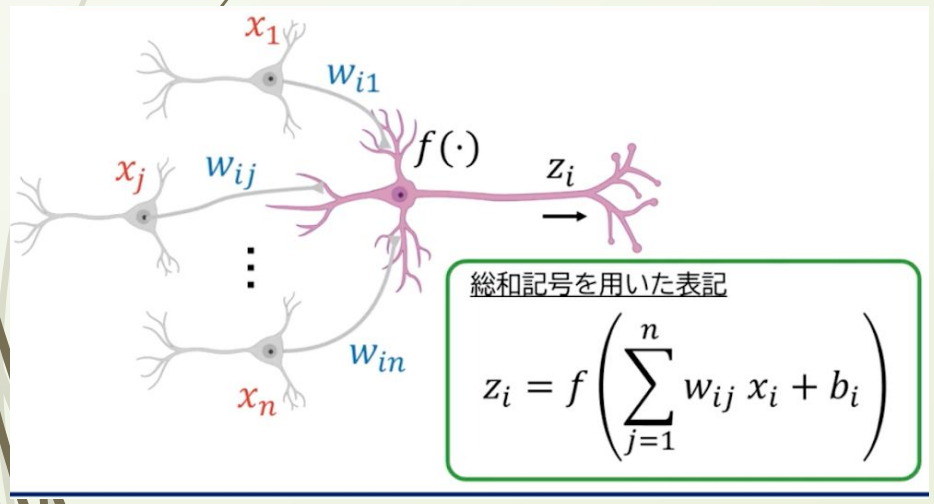


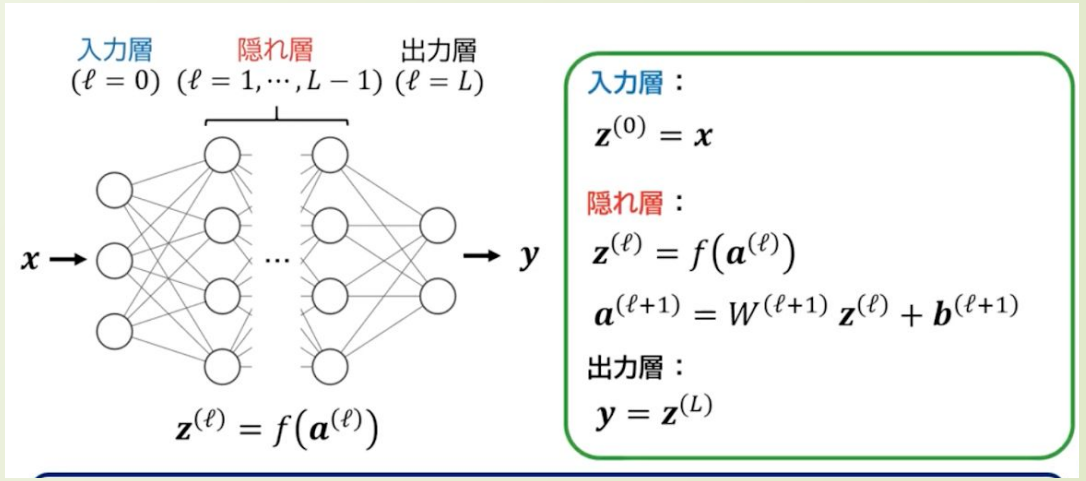



```
#lightGBM
import lightgbm as lgb
tree_lgb_model=lgb.LGBMRegressor()#Regressor()を変えれば分類にも使える
tree_lgb_model.fit(train_x,train_y)
forest_predict=tree_lgb_model.predict(test_x)
```

深層学習での”ニューラルネットワーク”

シナプスを模した以下のようなものを何層にも積み重ねたものを**ニューラルネットワーク**、隠れ層（中間層）が2つ以上のものをディープニューラルネットワーク(**DNN**)、これを学習させる手法のことを**ディープラーニング(深層学習)**、という



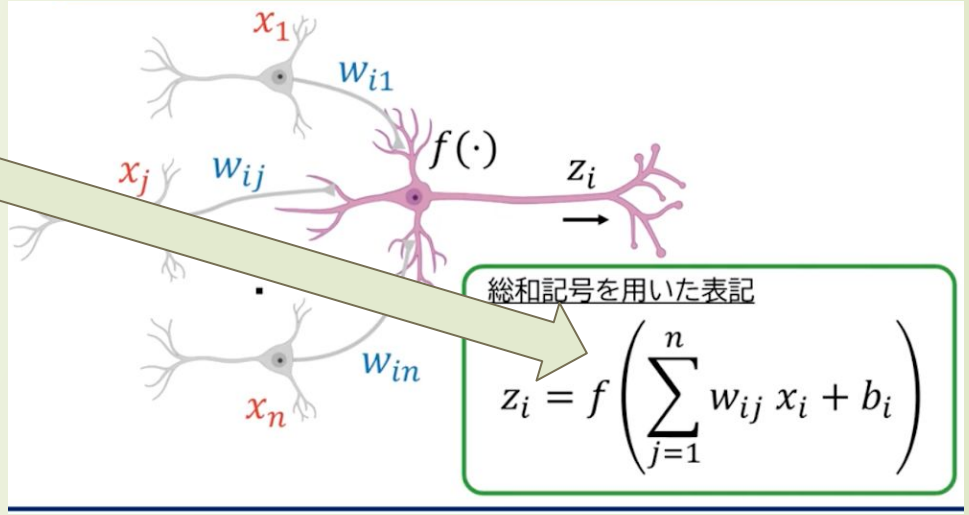


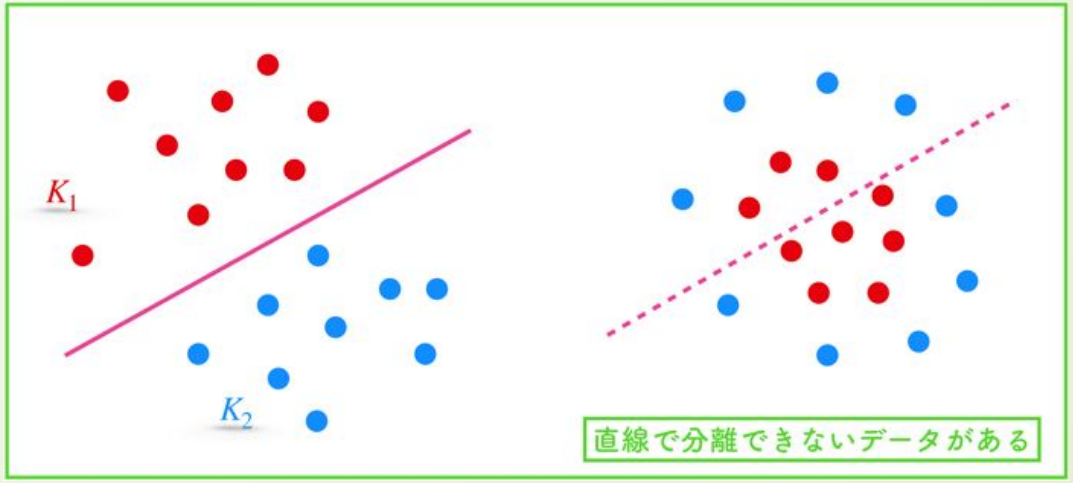
活性化関数について

重みw×入力xの和を入れる関数

線形より非線形の方が複雑なことができる（下図より）

→”重みw×入力xの和”の部分は線形だが活性化関数を非線形にすることで層の出力を非線形にしている



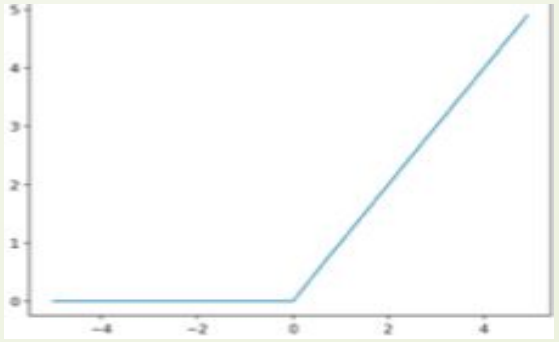


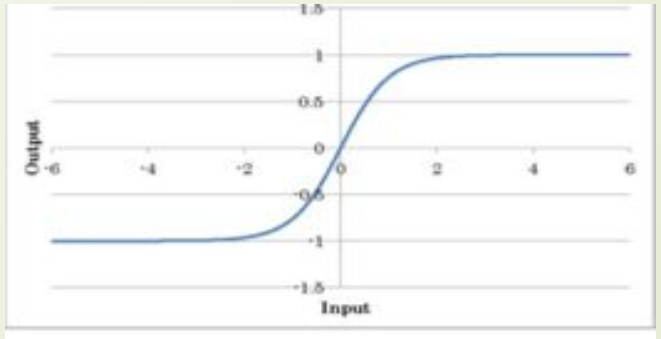
活性化関数の種類

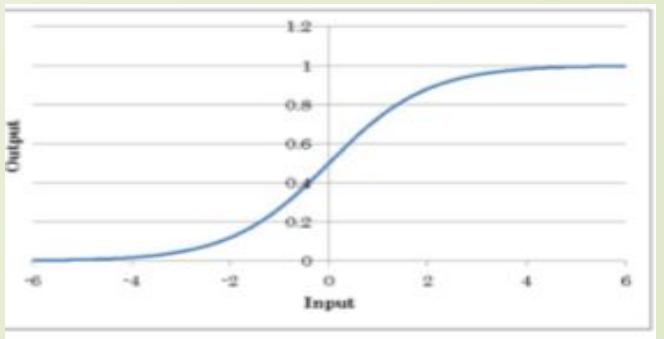
ニューラルネットワークの層の出力をそのまま次の層の入力とするのが恒等関数(f(x)=x)

ほかにも下のような関数がよく使われている（左から順にReLU,tanh,sigmoid）

最近のものはSwich,Mish,Leaky ReLU



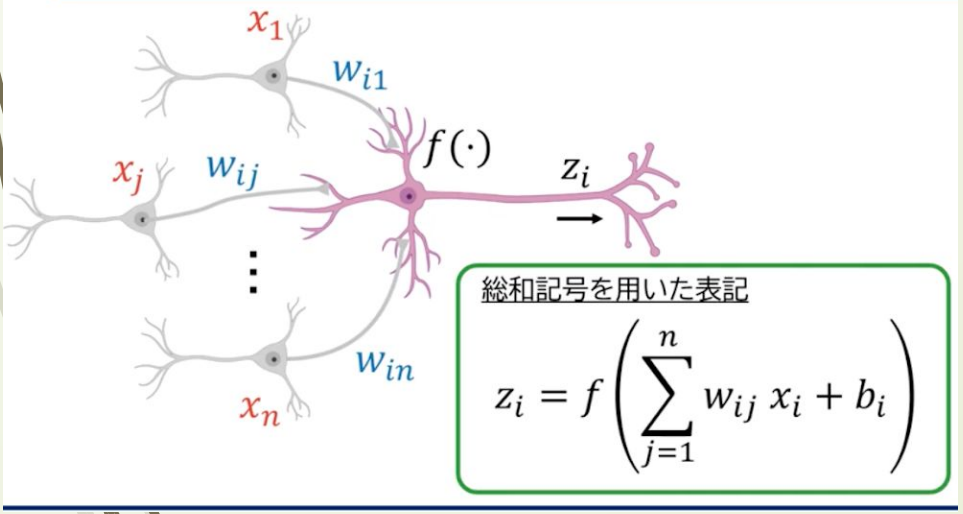


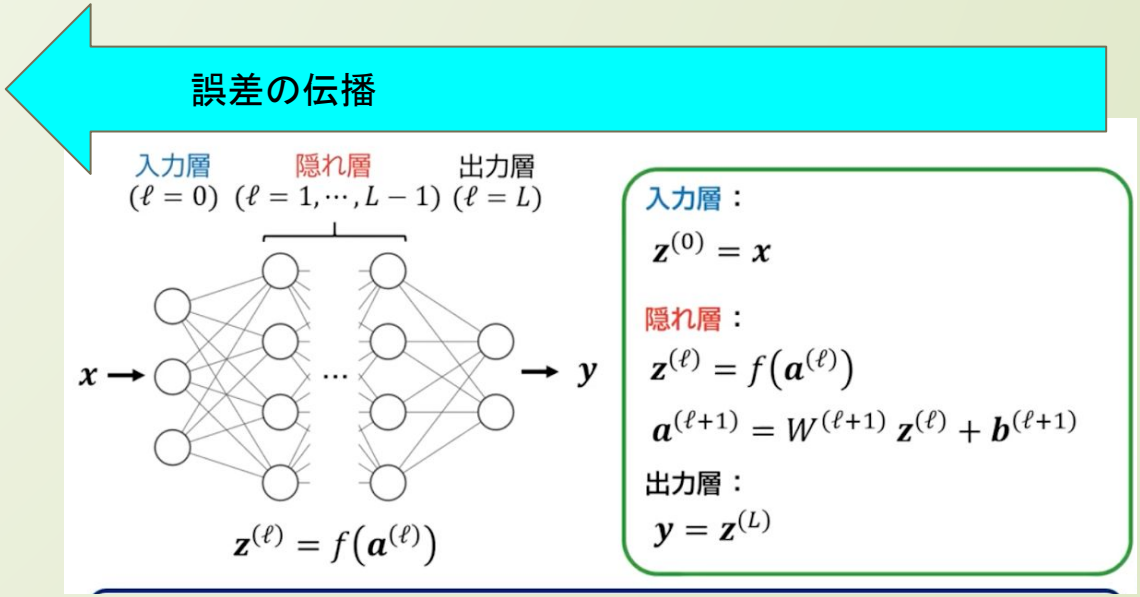


深層学習における"学習"について

ニューラルネットワークが"学習"しているのは適切な出力をするための**重みW**であり
出力が間違いだったとき、間違いに応じて重みWの値を更新していく。

予測と正解の誤差を入出力の向きと逆方向に計算していくのが**誤差逆伝播法**で、どのくらい
重みwの値を変更すれば出力がうまくいくのかを計算するの手法が**最適化手法**
（AdamやSGDなどがある）



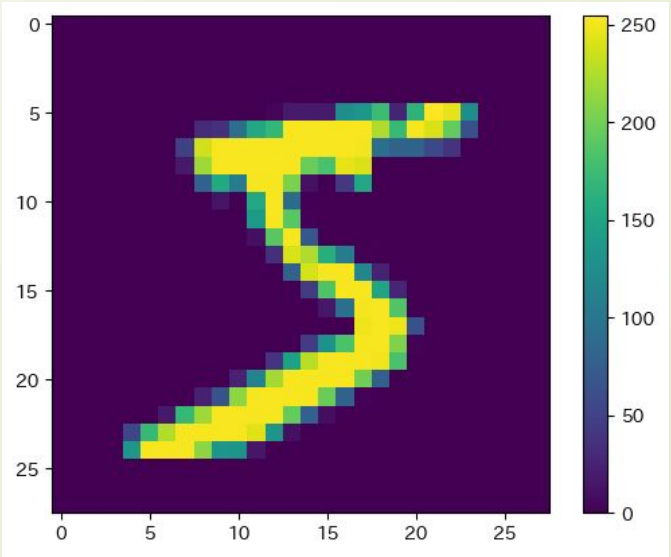


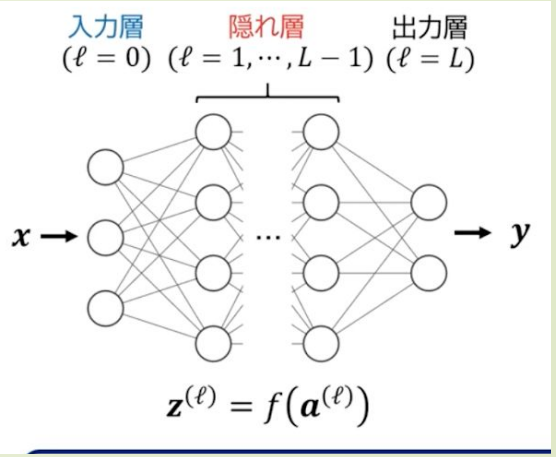
画像について

画像は縦＊横＊3(RGB)に0～255の値が入った行列(マス目に値が入ったもの)と見ることができる

1ピクセルについて周辺(縦横)のピクセルと関連があるがDNNの入力は1次元なので2(3)次元の画像データは1次元にしないといけない

これでは、縦横の関連をしっかり参照することはできない

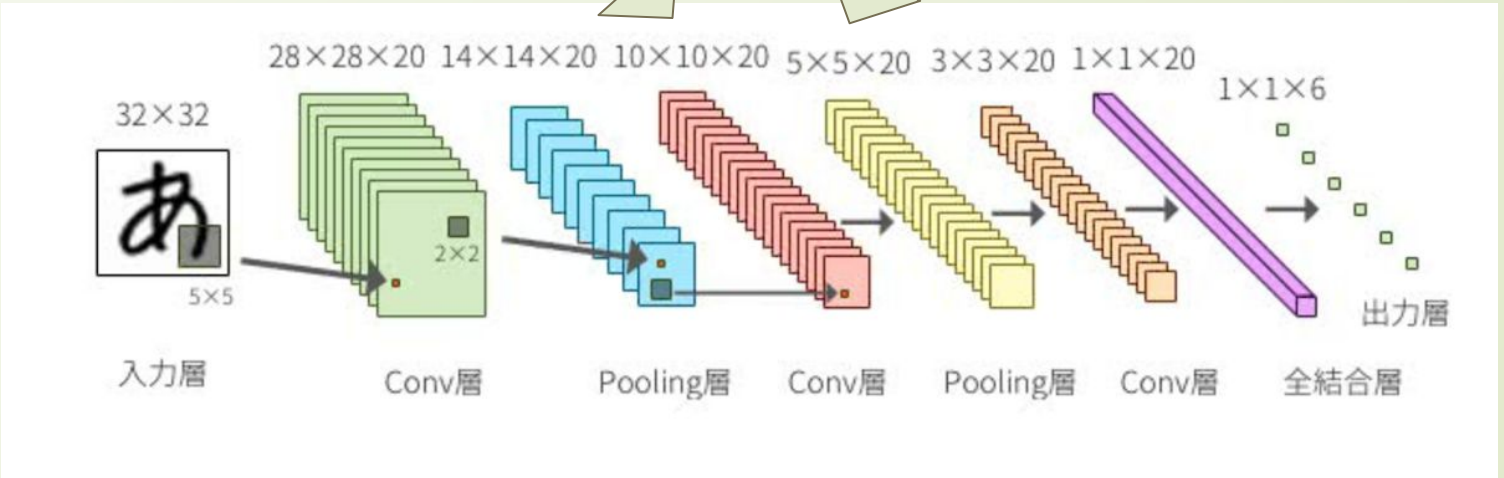




畳み込みニューラルネットワーク(CNN)

DNNでは全ての層は全結合層(前の層と後の層が全て繋がってる)だったが、CNNにはこれに**畳み込み層**(Convolution)と**プーリング層**を加えたもの

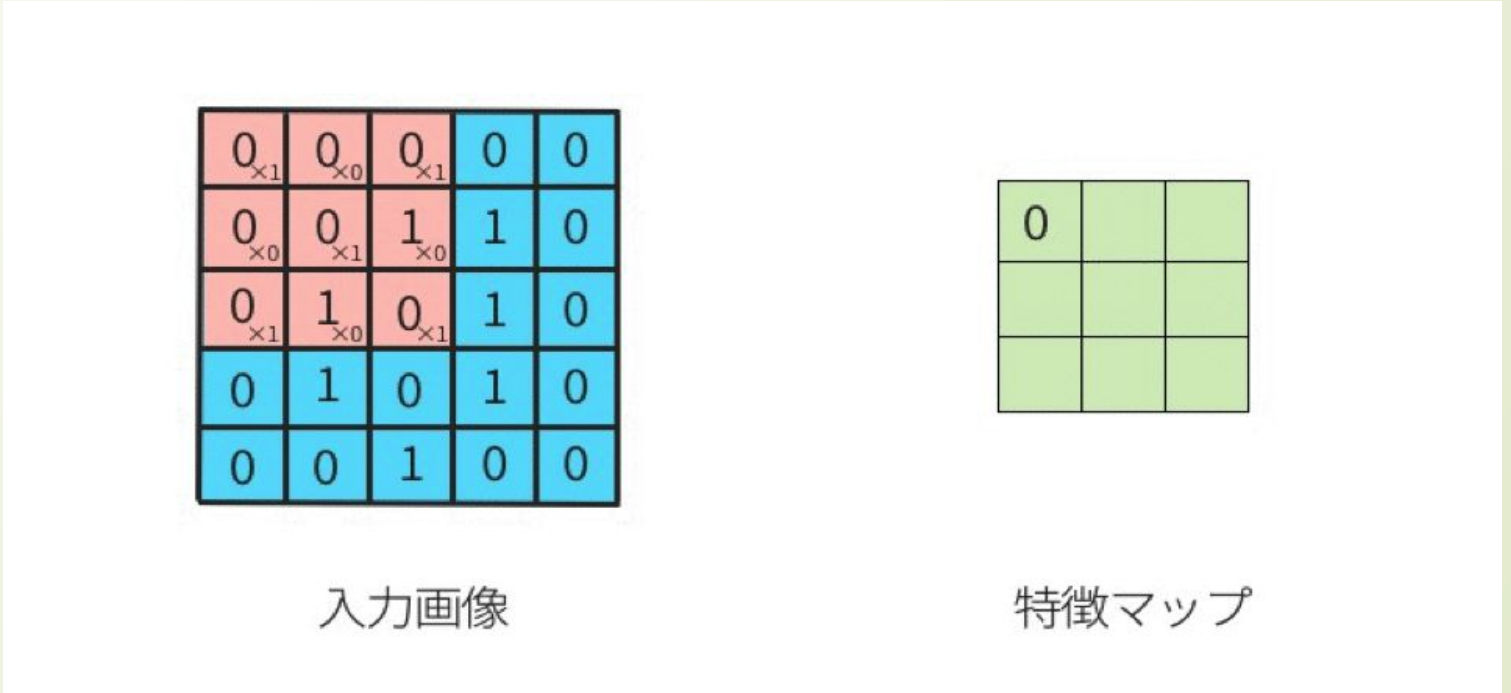
”画像”の扱いに長けている



畳み込みについて

赤い3＊3の行列がカーネル。カーネルの要素(I,j)と入力画像の要素を各々かけてその総和を一回り小さいマス目に書き込んでいく。これを**特徴マップ**といい畳み込み層の出力とする

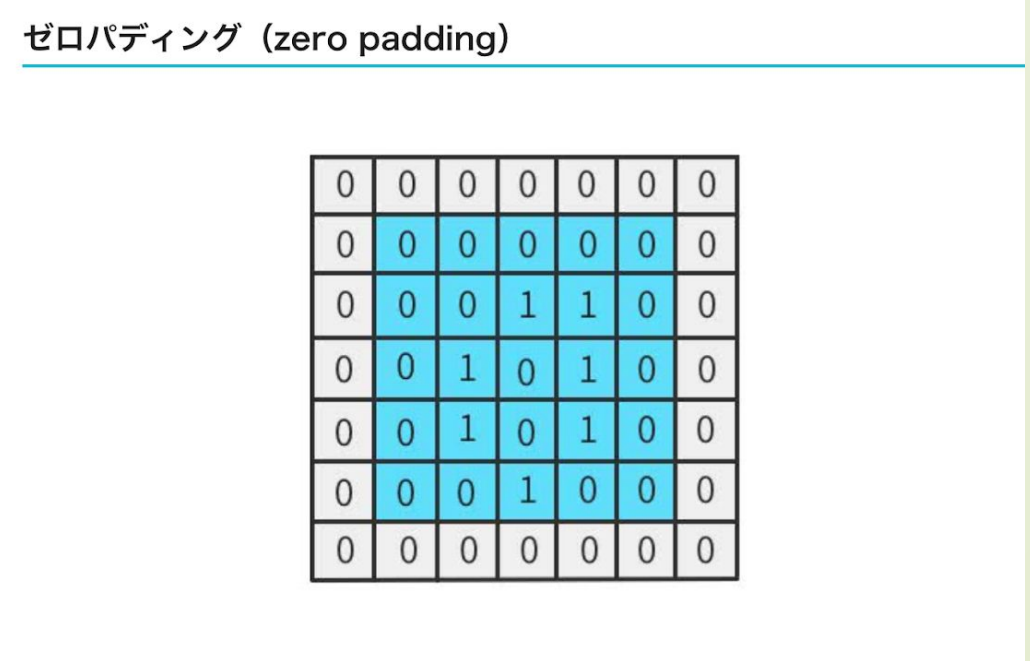
様々なカーネルを使用することで色々な角度から画像を"見て"、"特徴"をまとめている



パディング

データを一定の大きさにするために無意味な値を挿入すること。

これによって畳み込み層を通す前と後の大きさを同じにしたりできる。

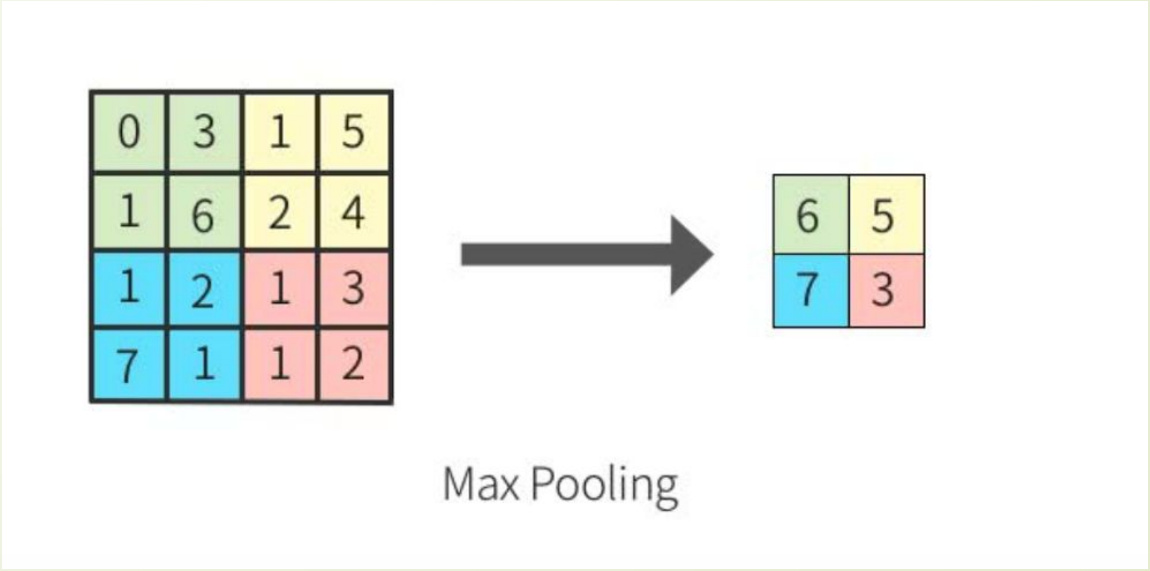


プーリング

通常畳み込み層の後に通すもので次元を圧縮する（特徴をキュッとまとめる）手法。

入力において（h＊w）のなかでの代表値をとる

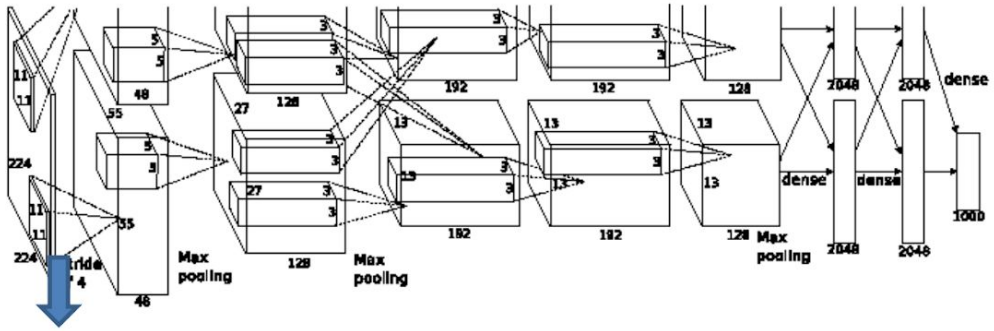
代表値は主に平均値か最大値

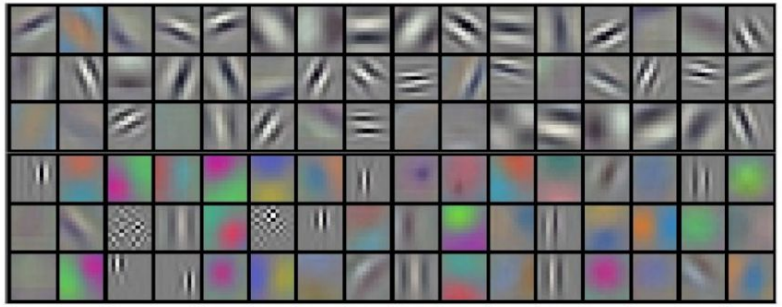


中間層の可視化…実際に機能しているのか？

人間の脳と同じ機能をCNNが果たしていることがわかった！→使えてる！

方位選択性が「初期視覚野」に現れたこと





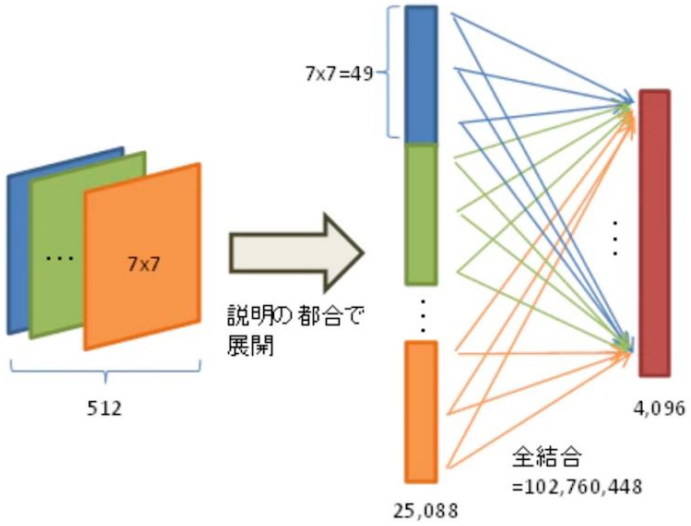
Hubel&Wiesel(1959)が発見した初期視覚野の方位選択性の再現

中間層には顔を見せなくても顔ニューロンができる

全結合の代わり？GAPについて

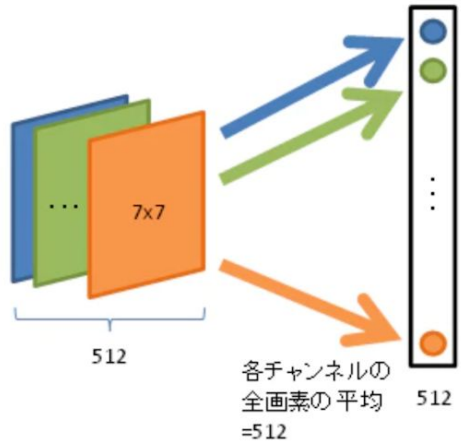
パラメータが少ない分、計算が早くなり、容量も節約できる

現状は、max poolingにより、7x7x512のデータができています。
これを1x1x4,096に全結合してますので、25,088×4,096=102,760,448の重みパラメータが存在しています。



Global Average Poolingとは

各チャンネル（面）の画素平均を求め、それをまとめます。
そうすると、重みパラメータは512で済みます。

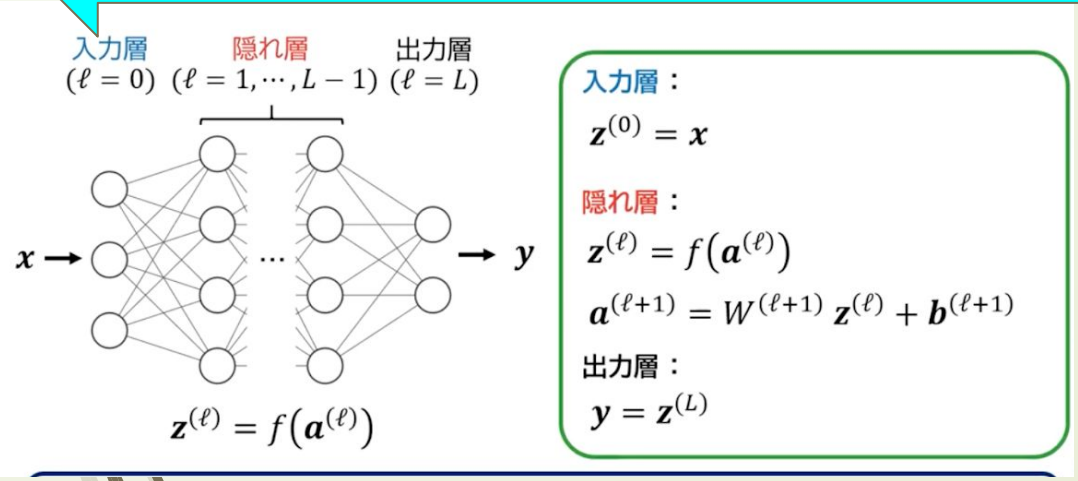


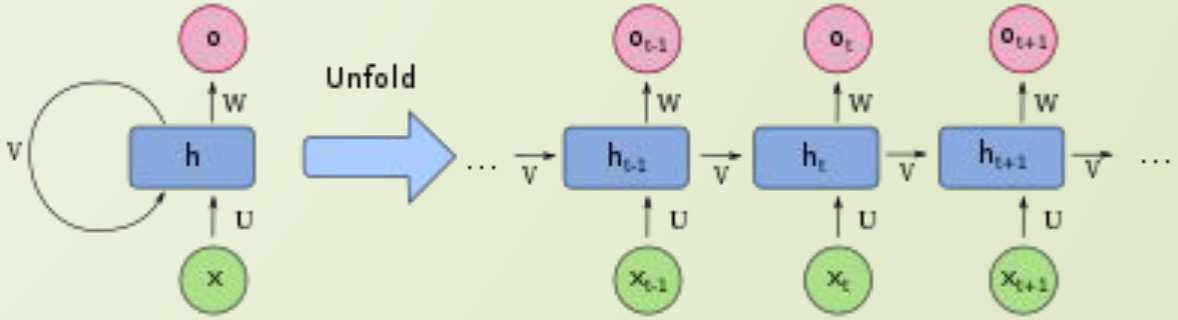
勾配消失問題とは？

重みを適切に更新していくのが深層学習における"学習"だが、層が深すぎると重みの更新が出来なってしまう

これを**勾配消失問題**といい、初期の重みを適切にすることによって従来は対策されていたが、現在では**BatchNormalization**などの正規化によって初期の重みを選択せずとも勾配消失問題は解決できているが、自然言語などの時系列データを扱うRNNでは勾配消失問題は解決していない（後続のLSTMで対応）

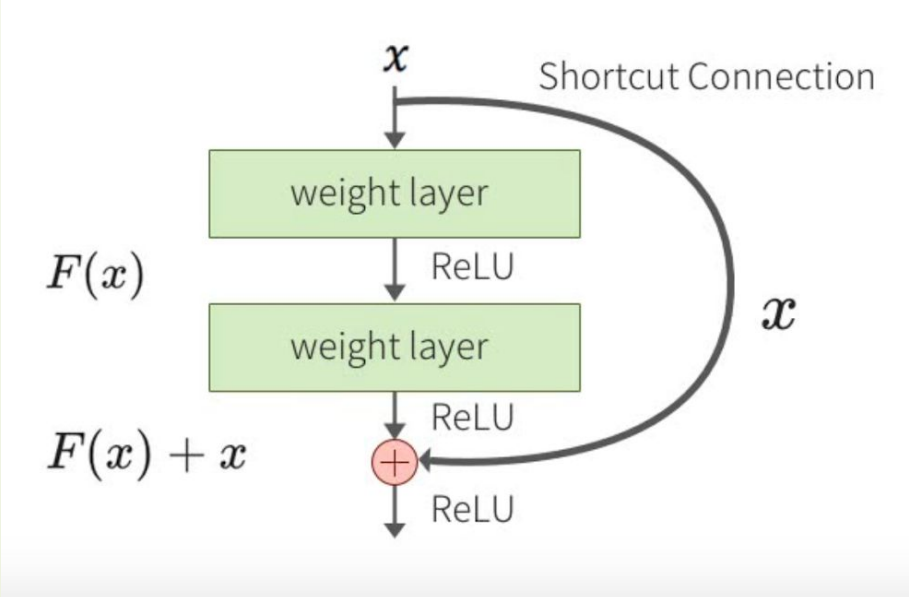
誤差が伝播しない





更に層を深くするために

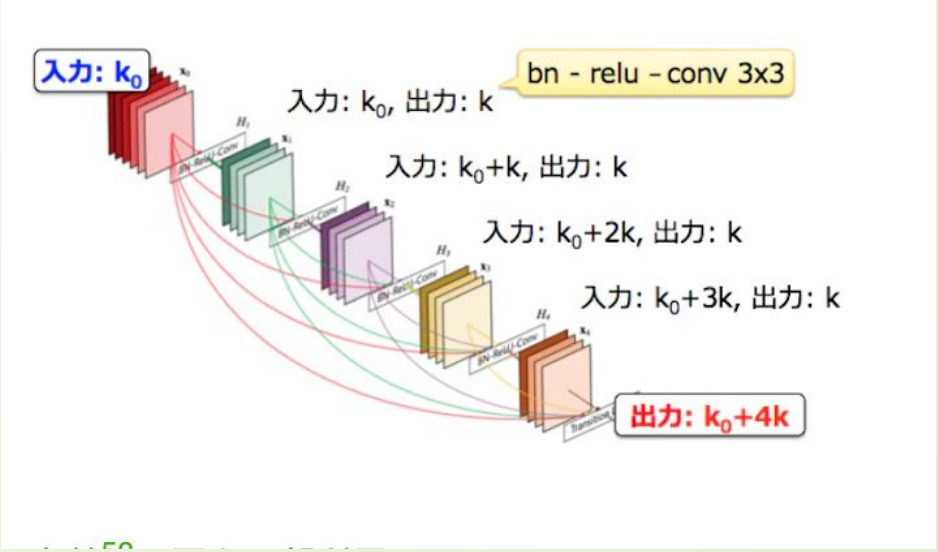
2012年から始まったCNNの精度競争は層を深くすることで加速していったが、層を深くすると前述の通り勾配消失問題が発生する。これを解決するアイデアが**スキップコネクション**で最先端モデルの一つであるResNeXtやEfficientNetなどにも組み込まれている



全部繋げてしまう：DenseNet

ResNetではn層の入力はn-1層からの残渣接続を用いたが、

DenseNetではn層の入力に1～n-1層からの残渣接続を用いる。

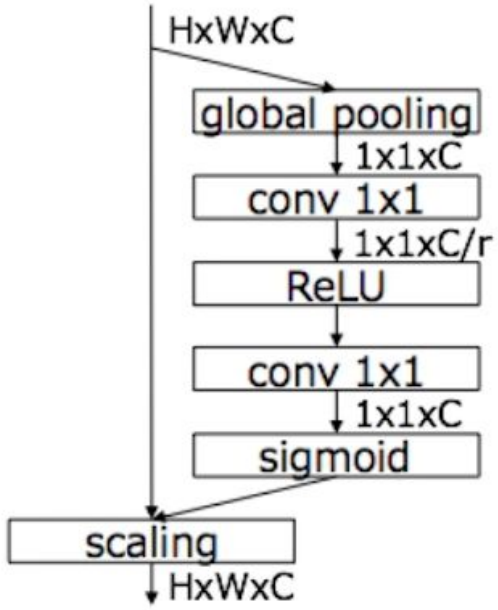


skip connectionのその先へ

流行りのAttention機構を画像分野に導入したもの

どのモデルにもつけることができるのが売り

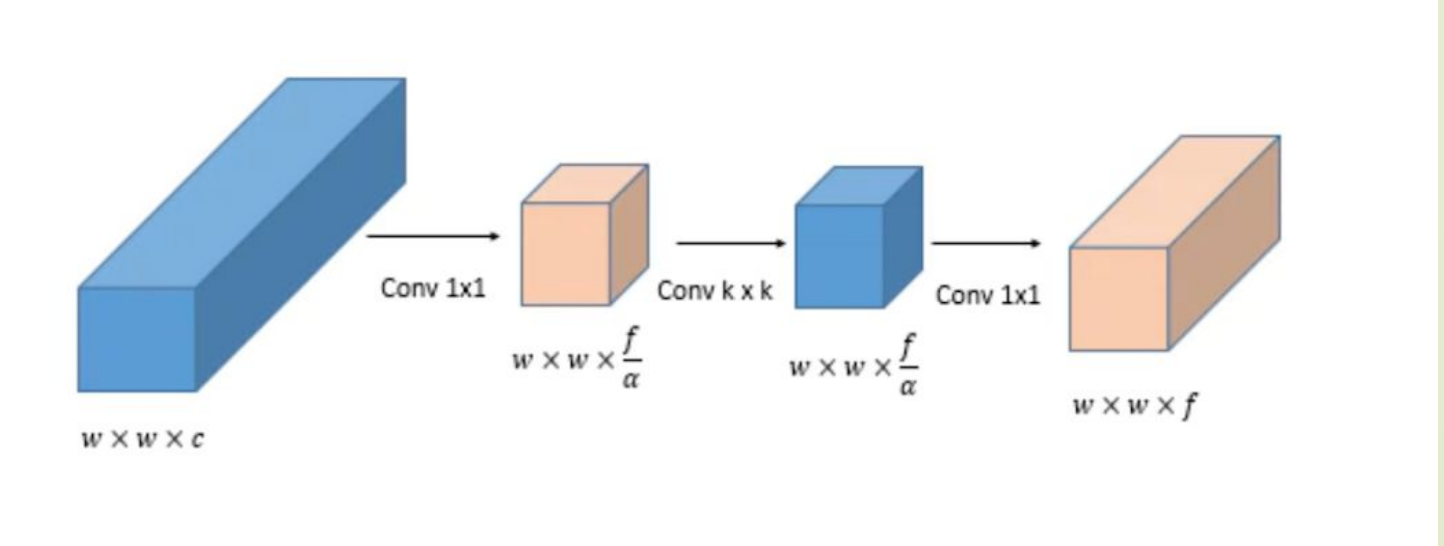
attentionについては自然言語分野がメインとなるのでここでは割愛



次元圧縮の畳み込み bottle neck

1*1畳み込みのことをbottle neck, pointwise 畳み込みという。

狙いとしてはattention機構の導入か次元圧縮による計算量の削減



augmentationについて

現在の最強モデルは3億枚の画像を用いて学習しているが、一般人はそんなに画像を用意できない

確率的に元データに変化を加えたものを学習に用いることであたかもN倍のデータで学習してる雰囲気を出そうというもの

強化学習を用いた効果的なaugmentationの探索方法が検討されている

おまけとしての”説明可能性”

深層学習では何に注目して学習しているのかは”神のみぞ知る”のだが、重みから何に注目しているかわかるものがあり、**Grad_CAM**などがある

重みが大きいほどそれを重視している、と解釈するもの

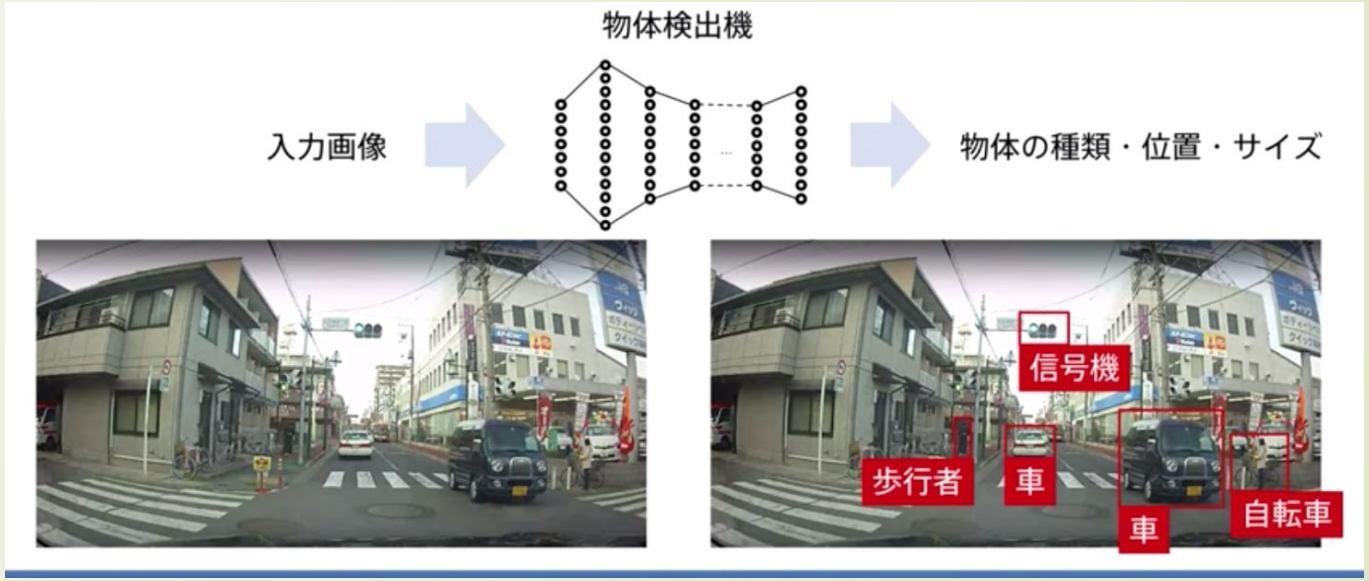
今春に高速化が試みられ、リアルタイムでのGrad_CAMのが可能になった？

入力画像	Grad-CAM	Guided Grad-CAM
		

CNNを使った分類や値の回帰以外のタスク

入力画像に対して**どこに何が写っているか**を出力するのが**物体検出**

どのクラスに属しているか、だけではなく、どこに写っているかの情報が正解データとして必要



CNNを使った分類や値の回帰以外のタスク

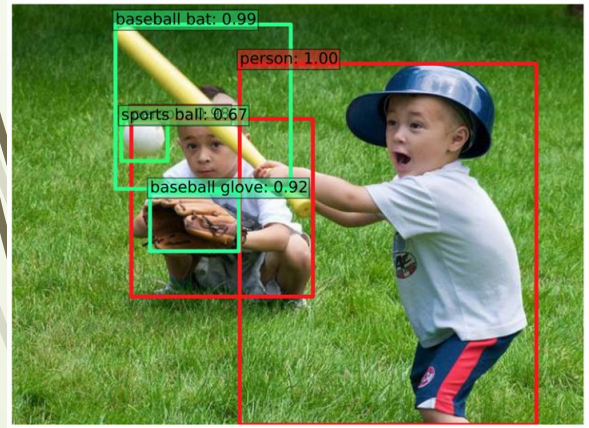
画像のピクセルごとに何が写っているかを予測するのが**セマンティックセグメンテーション**

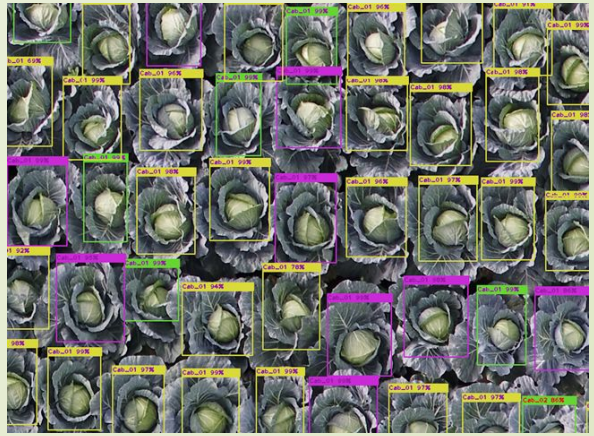
ピクセルごとに何が写っているかの正解データが必要

AIでできること（画像に絞ると）

一般的に…

- 犬か猫か、など画像を分類できる
- 画像の中のどこに人が写っているのか、わかる
- 画像について1ピクセルごとに何が写っているか判断する





AIが“学習”するためには？

- 大量のデータが必要
  →1から学習するなら1万枚の画像、excel1万行分のデータ　程度
- データにノイズがあまり入っていないことが望ましい(空白はない方が良い)
- データが偏りすぎていたら汎化性能がなくなってしまう
- 学習用のデータと実際に使うデータに乖離があると使えないことがある

```
Installing collected packages: efficientnet
Successfully installed efficientnet-1.1.0
```

[3]:

```
train_df = pd.read_csv('../input/prostate-cancer-grade-assessment/train.csv')
print(train_df.shape)
train_df.head()
```

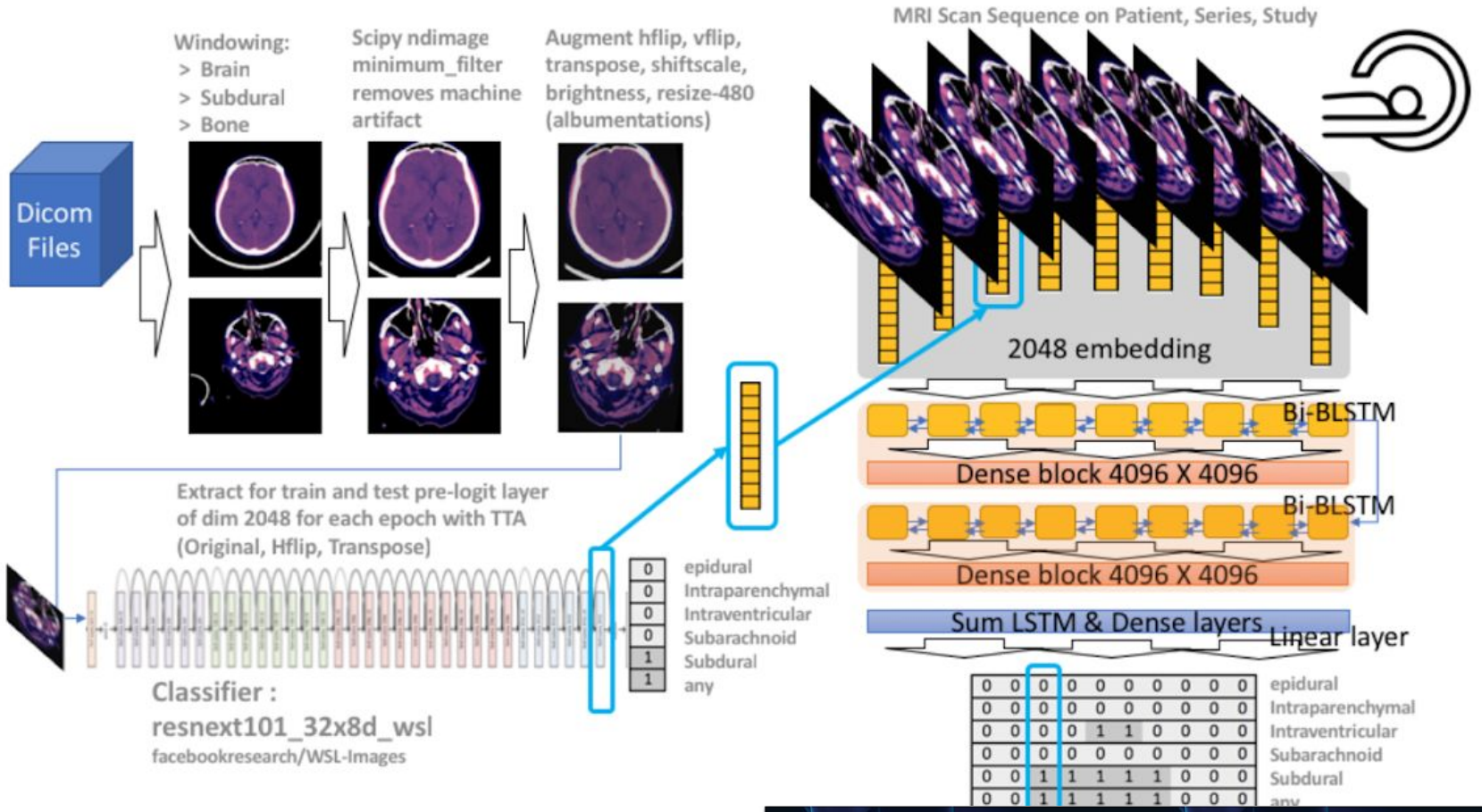
```
(10616, 4)
```

[3]:

	image_id	data_provider	isup_grade	gleason_score
0	0005f7aaab2800f6170c399693a96917	karolinska	0	0+0
1	000920ad0b612851f8e01bcc880d9b3d	karolinska	0	0+0
2	0018ae58b01bdadc8e347995b69f99aa	radboud	4	4+4
3	001c62abd11fa4b57bf7a6c603a11bb9	karolinska	4	4+4
4	001d865e65ef5d2579c190a0e0350d8f	karolinska	0	0+0

	A	B	C	D
1	製品	第 1 四半期	第 2 四半期	総計
2	Chocolade	$744.60	$162.56	$907.16
3	Gummibarchen	$5,079.60	$1,249.20	$6,328.80
4	Scottish Longbreads	$1,267.50	$1,062.50	$2,330.00
5	Sir Rodney's Scones	$1,418.00	$756.00	$2,174.00
6	Tarte au sucre	$4,728.00	$4,547.92	$9,275.92
7	Chocolate Biscuits	$943.89	$349.60	$1,293.49
8	合計	$14,181.59	$8,127.78	$22,309.37

We were very sad not to get a top3 in Recursion competition, now we are very happy 😄



工業分野

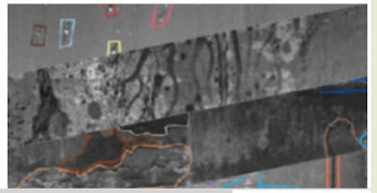
Overview Data Notebooks Discussion Leaderboard Rules Late Submission

Overview

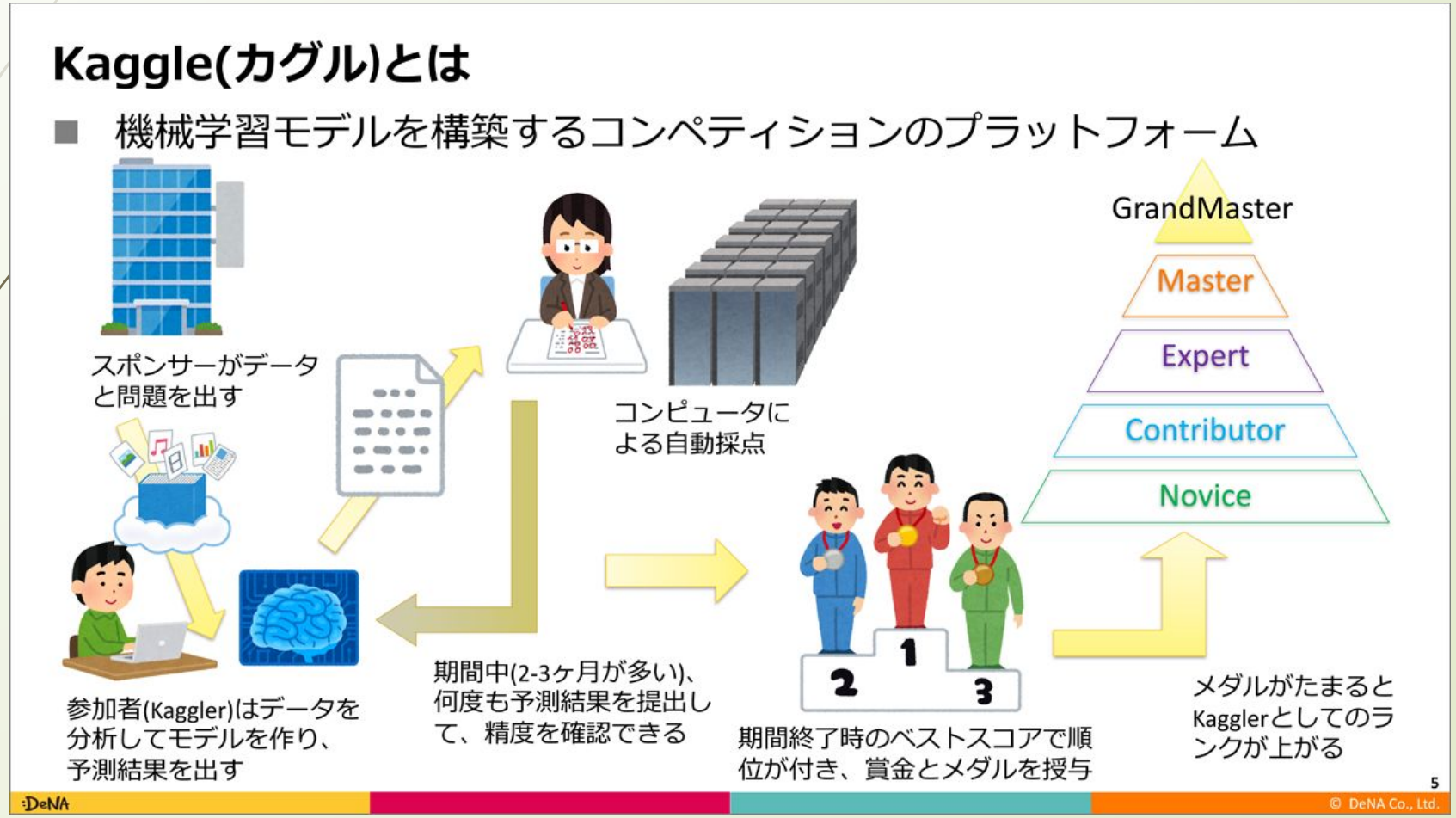
Description
Evaluation
Timeline
Prizes
Kernels Requirements

Steel is one of the most important building materials of modern times. Steel buildings are resistant to natural and man-made wear which has made the material ubiquitous around the world. To help make production of steel more efficient, this competition will help identify defects.

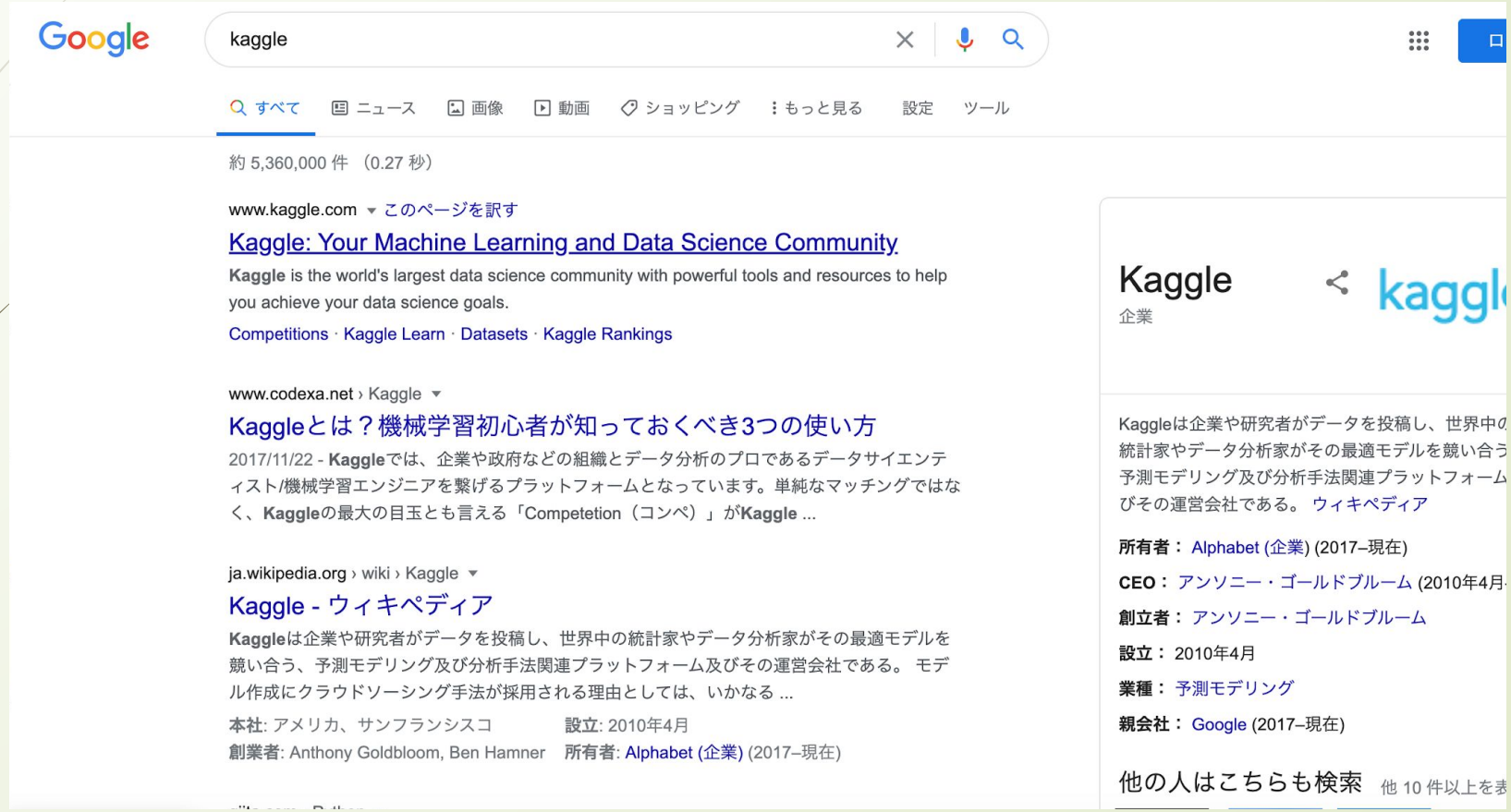
Severstal is leading the charge in efficient steel mining and production. They believe the future of metallurgy requires development across the economic, ecological, and social aspects of the industry—and they take corporate responsibility seriously. The company recently created the country's largest industrial data lake, with petabytes of data that were previously discarded.



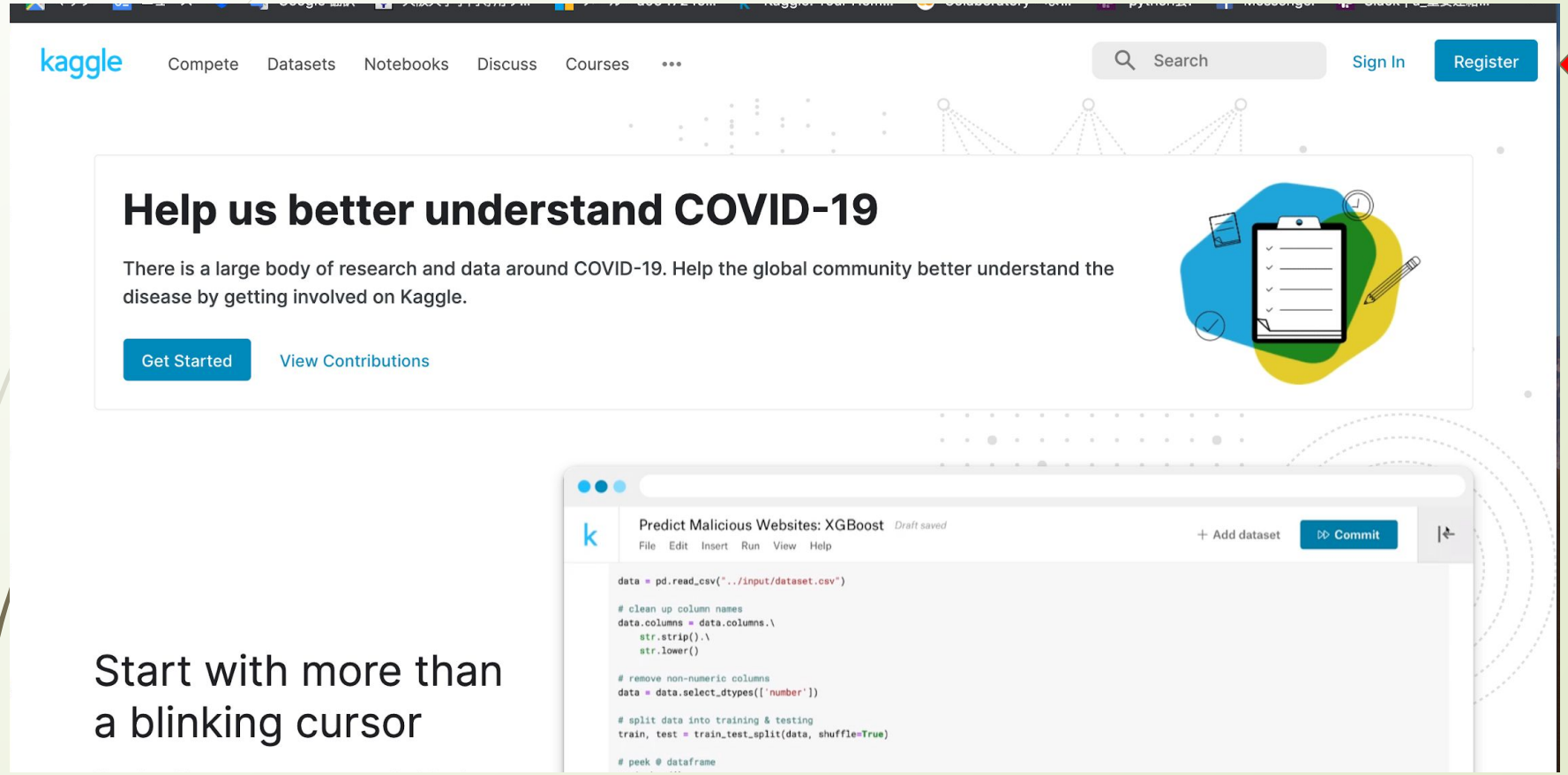
データ分析のホーム kaggle



kaggleに登録...

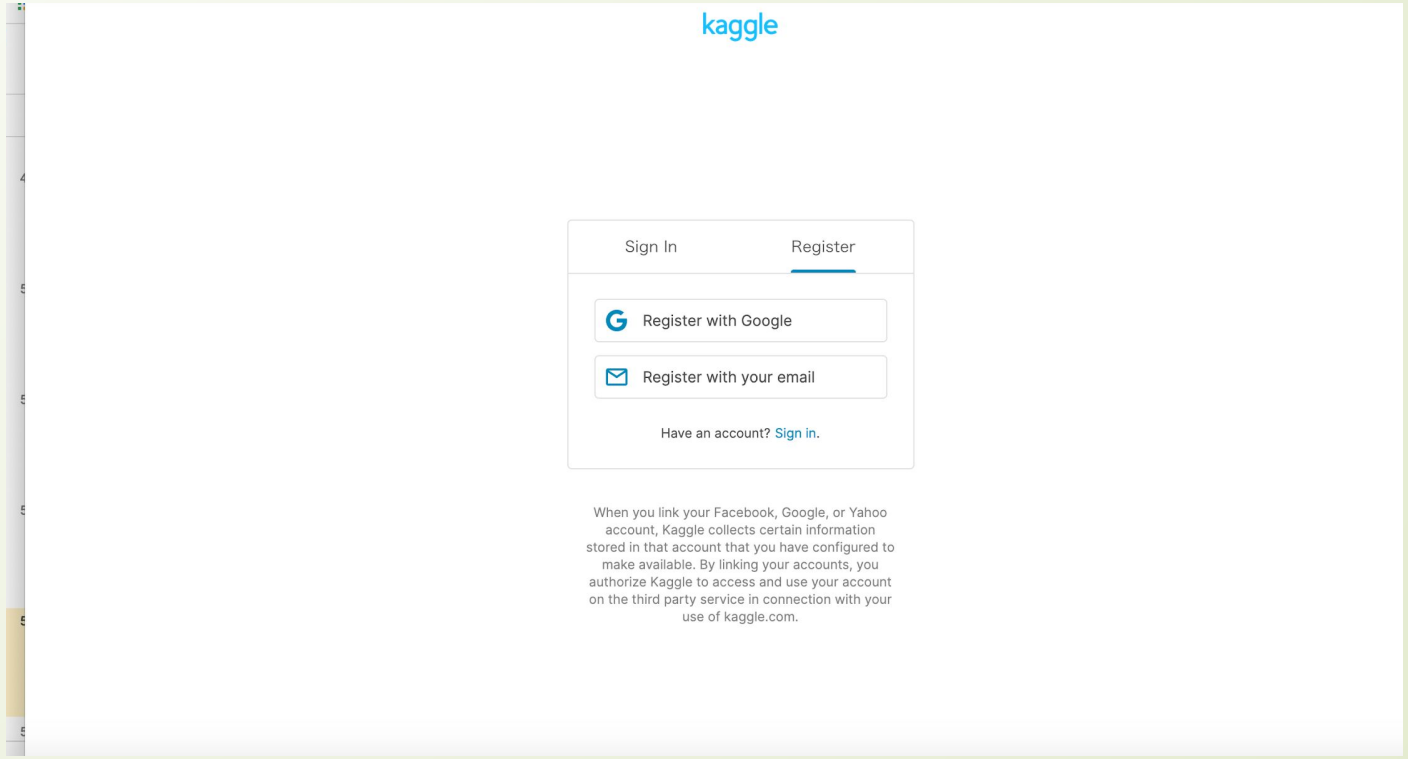

Google
kaggle
すべて
ニュース
画像
動画
ショッピング
もっと見る
設定
ツール
約 5,360,000 件 （0.27 秒）
www.kaggle.com このページを訳す
Kaggle: Your Machine Learning and Data Science Community
Kaggle is the world's largest data science community with powerful tools and resources to help you achieve your data science goals.
Competitions · Kaggle Learn · Datasets · Kaggle Rankings
www.codexa.net › Kaggle
Kaggleとは？機械学習初心者が知っておくべき3つの使い方
2017/11/22 - Kaggleでは、企業や政府などの組織とデータ分析のプロであるデータサイエンティスト/機械学習エンジニアを繋げるプラットフォームとなっています。単純なマッチングではなく、Kaggleの最大の目玉とも言える「Competetion（コンペ）」がKaggle ...
ja.wikipedia.org › wiki › Kaggle
Kaggle - ウィキペディア
Kaggleは企業や研究者がデータを投稿し、世界中の統計家やデータ分析家がその最適モデルを競い合う、予測モデリング及び分析手法関連プラットフォーム及びその運営会社である。 モデル作成にクラウドソーシング手法が採用される理由としては、いかなる ...
本社: アメリカ、サンフランシスコ
設立: 2010年4月
創業者: Anthony Goldbloom, Ben Hamner
所有者: Alphabet (企業) (2017–現在)
Kaggle
企業
Kaggleは企業や研究者がデータを投稿し、世界中の
統計家やデータ分析家がその最適モデルを競い合う
予測モデリング及び分析手法関連プラットフォーム
びその運営会社である。 ウィキペディア
所有者： Alphabet (企業) (2017–現在)
CEO： アンソニー・ゴールドブルーム (2010年4月
創立者： アンソニー・ゴールドブルーム
設立： 2010年4月
業種： 予測モデリング
親会社： Google (2017–現在)
他の人はこちらも検索
他 10 件以上を表

kaggle.comから　registerをクリック



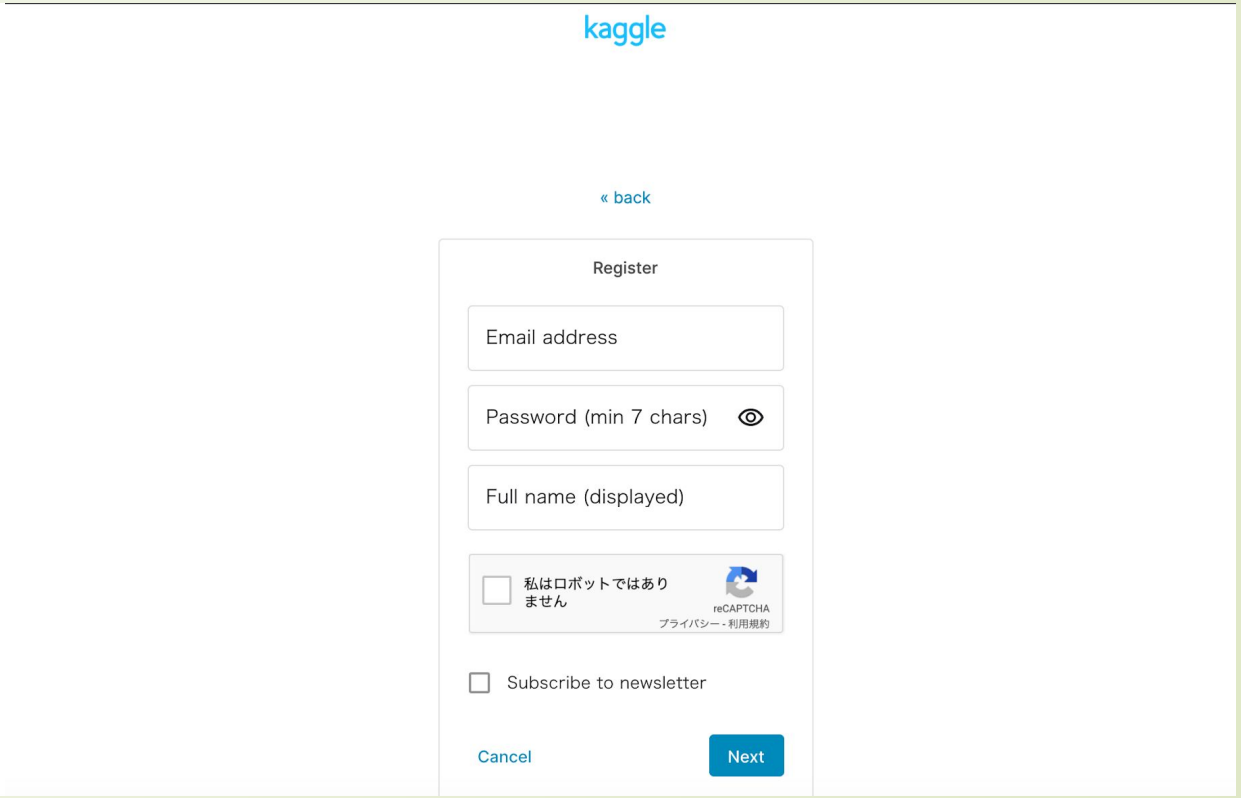
Start with more than a blinking cursor

どちらからでも良いです



どちらからか登録


Google にログイン
ログイン
「kaggle.com」に移動
メールアドレスまたは電話番号
メールアドレスを忘れた場合
続行するにあたり、Google はあなたの名前、メールアドレス、言語設定、プロフィール写真を kaggle.com と共有します。
アカウントを作成
次へ
日本語
ヘルプ
プライバシー
規約


kaggle
« back
Register
Email address
Password (min 7 chars)
Full name (displayed)
私はロボットではありません
reCAPTCHA
プライバシー - 利用規約
Subscribe to newsletter
Cancel
Next

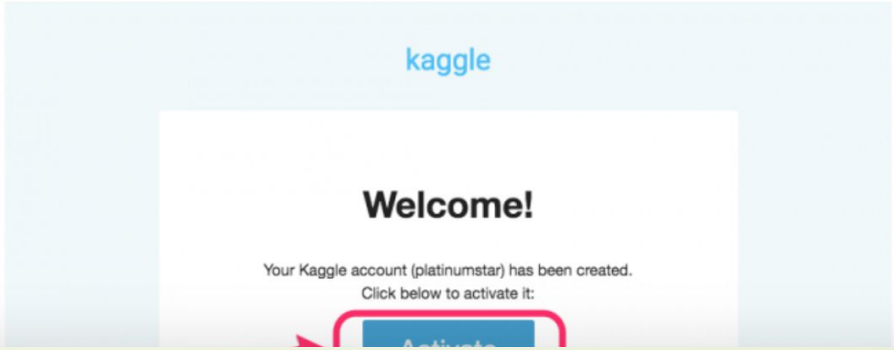
入力したらチェックボックスにチェックを入れ、**「Get Started（始める）」**を選択しましょう。入力完了すると、Kaggleのプライバシーポリシーの確認画面になります。

一番上と、一番下にある**「I agree.（同意する）」**に2つチェックを入れ、**「Create Account（アカウントを作成する）」**を選択しましょう。

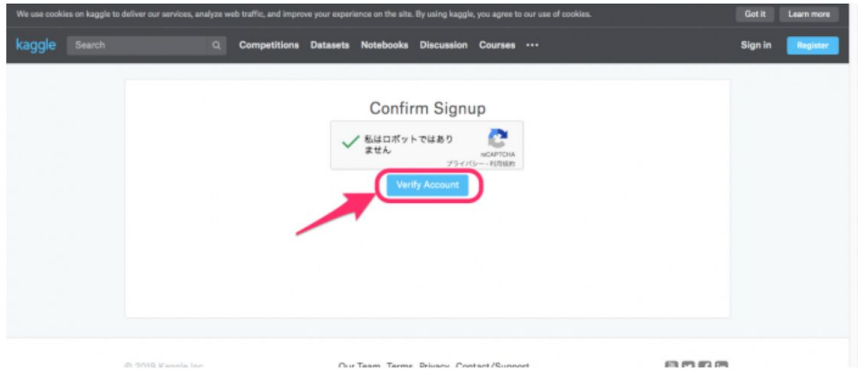
読み込みに時間がかかることがありますので、しばらく待ちます(^^) _旦~~
遅い場合は、メールボックスを見てみましょう。

Kaggle signup confirmationから「Welcome！」とメールが届いていればOKです。メールを開くと「Your Kaggle account (あなたのアカウント名) has been created.Click below to activate it:（Kaggleのアカウントが作成されました。有効化するには下をクリックしてください）」とあります。

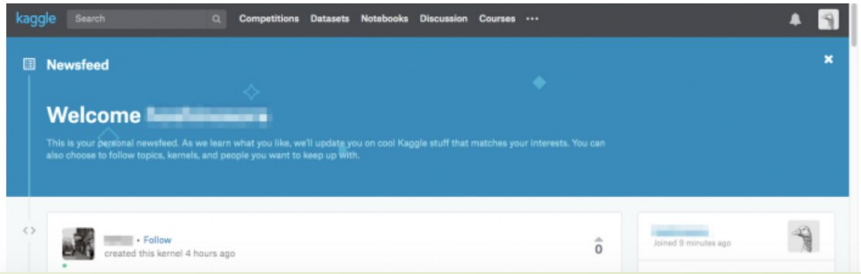
水色の**「Active」**ボタンを選択しましょう。



すると、Kaggleのサインアップページに飛びます。「私はロボットではありません」にチェックを入れ、**「Verify Account（アカウントを確認する）」**を選択します。



この画面になれば、アカウント作成完了です！ヽ(=´▽`=)ﾉ



pytorchのチュートリアルをやろう！

https://pytorch.org/tutorials/