Machine Learning with Python for Young Data Scientists

Kan Ouivirach

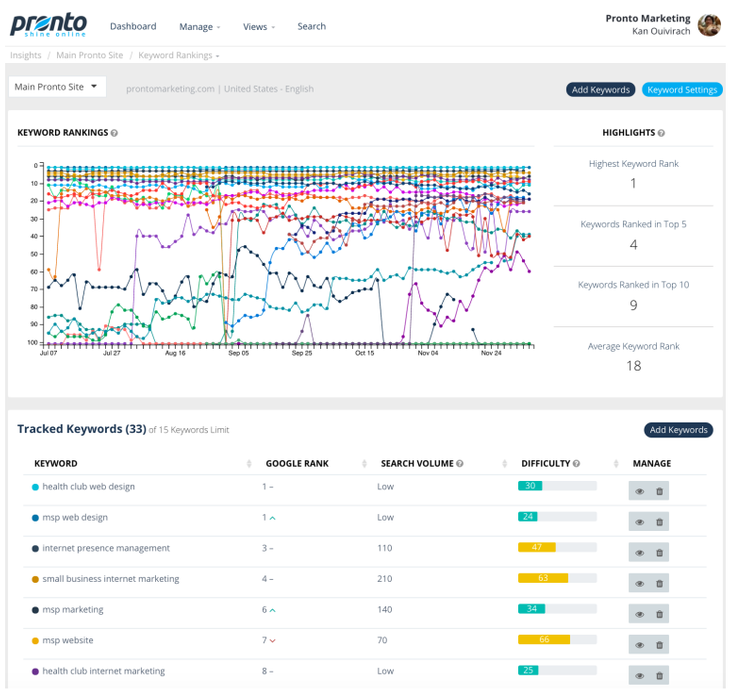
prontoi tools

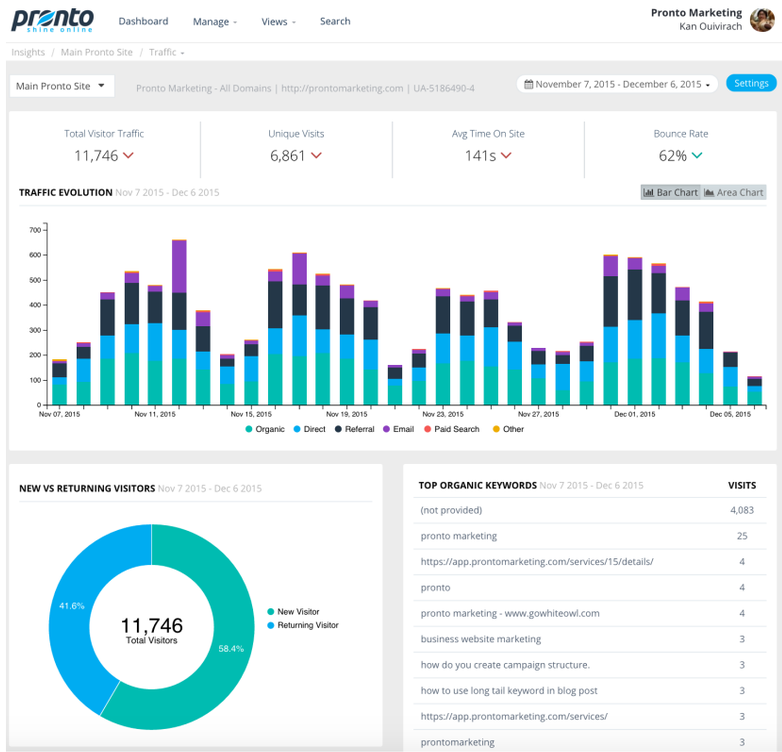
http://www.heardcountyrecreationdepartment.com/k/data-science-wallpaper

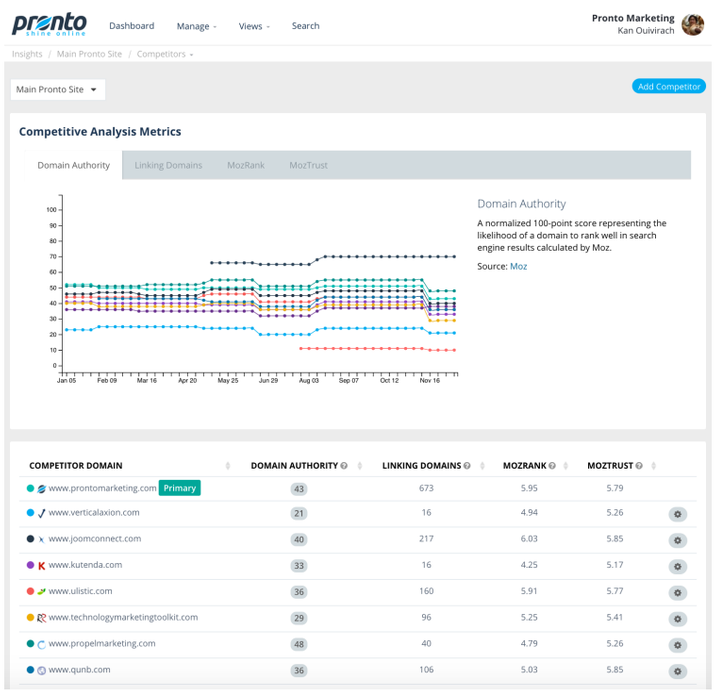
https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

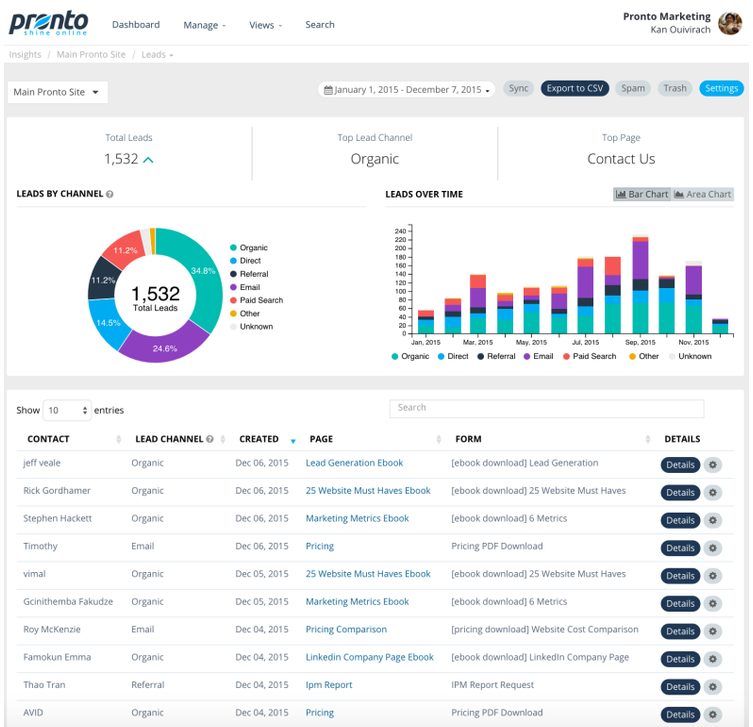
a bit about our work..

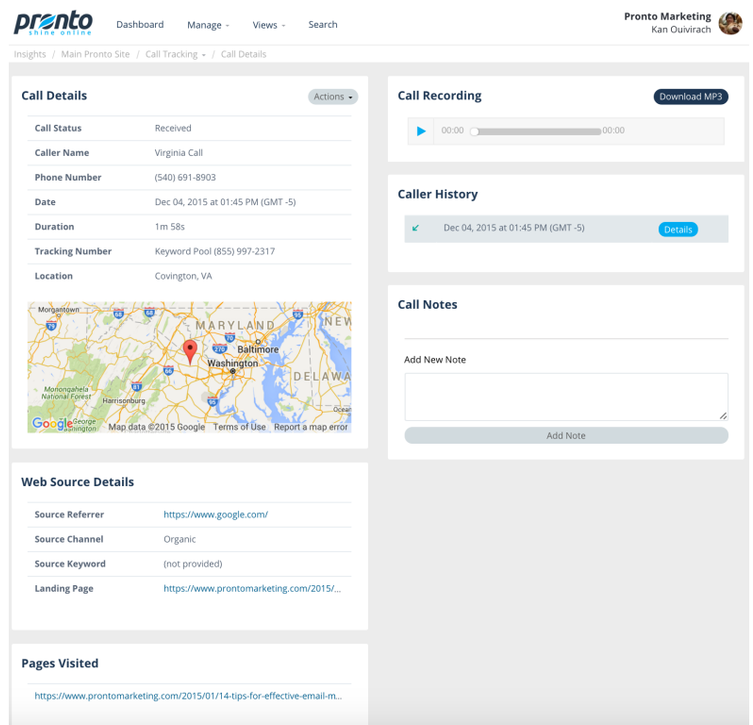
Better Marketing Visibility

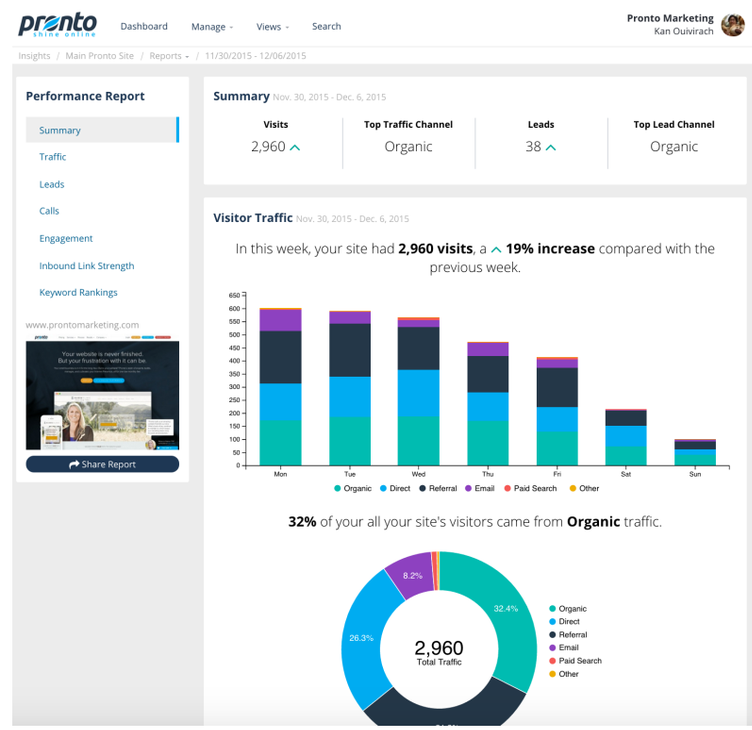




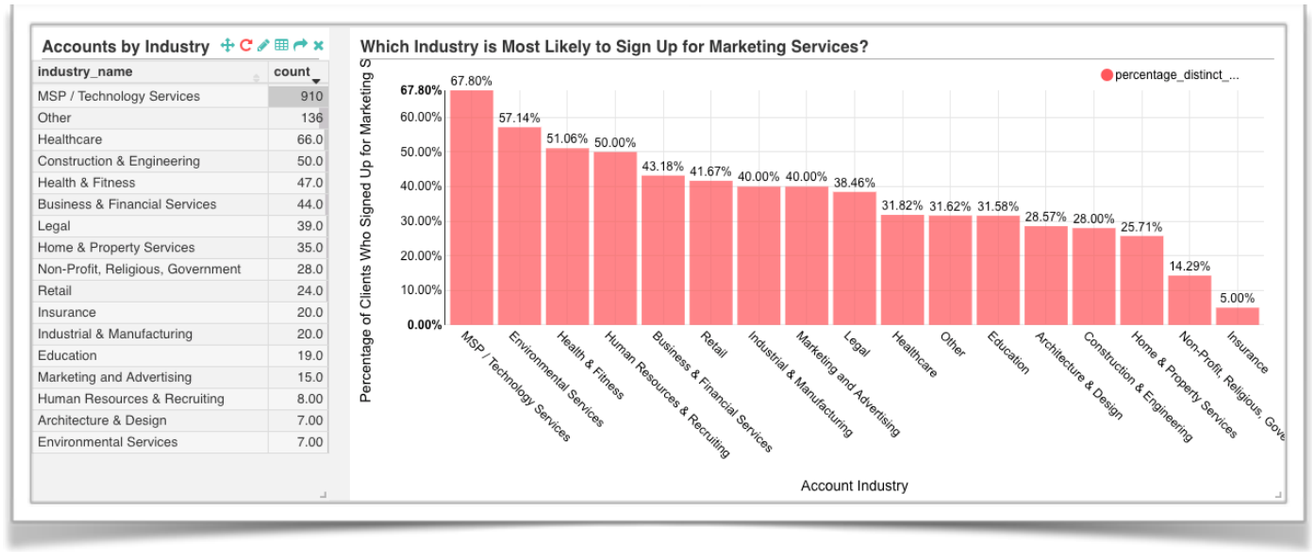




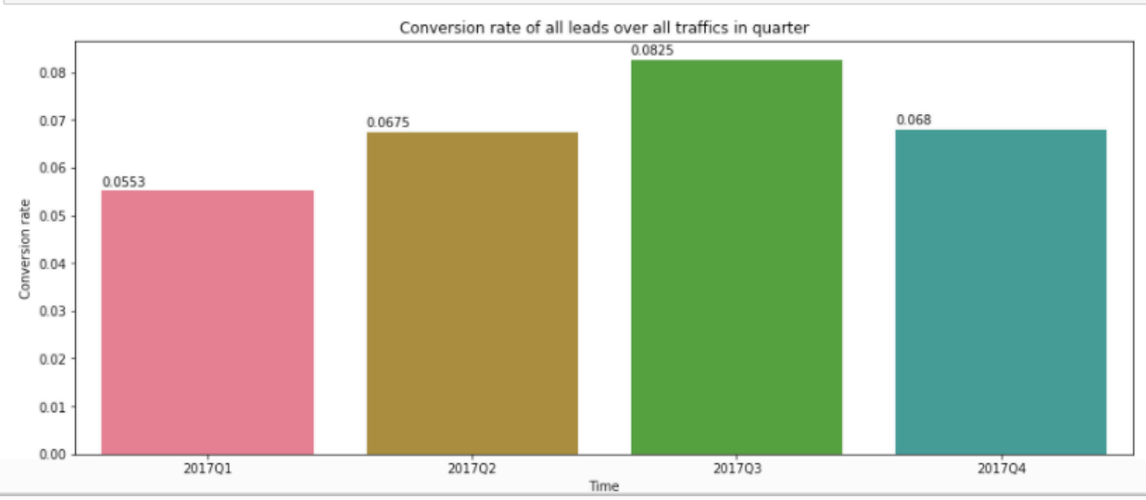


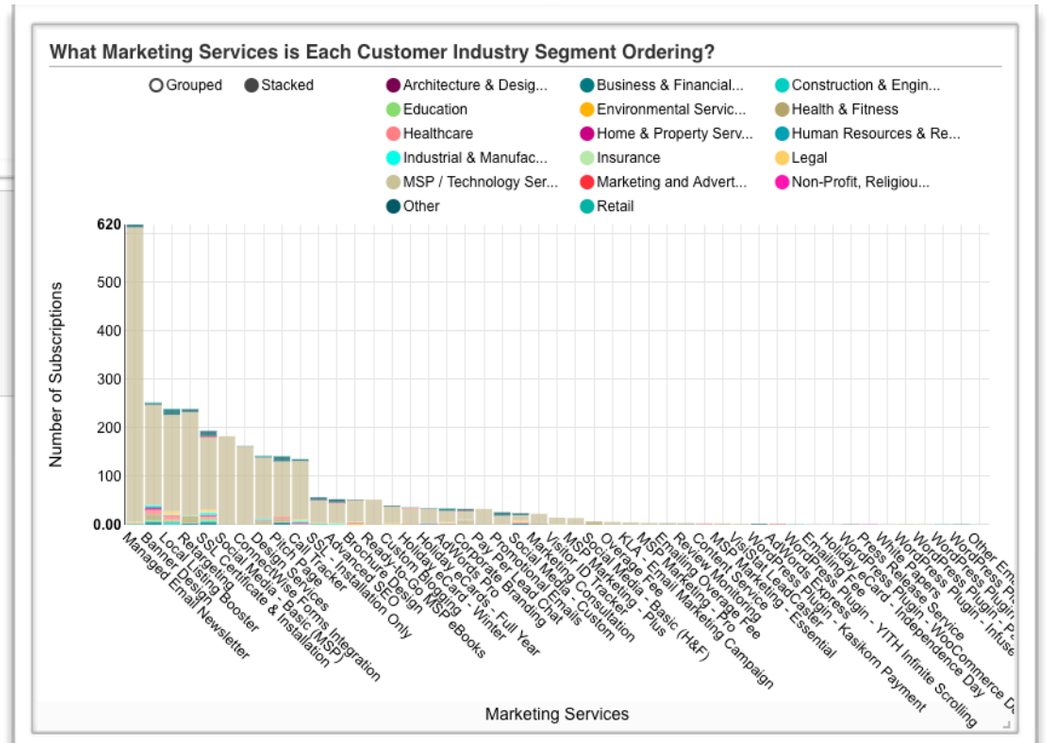


Data Science at Work

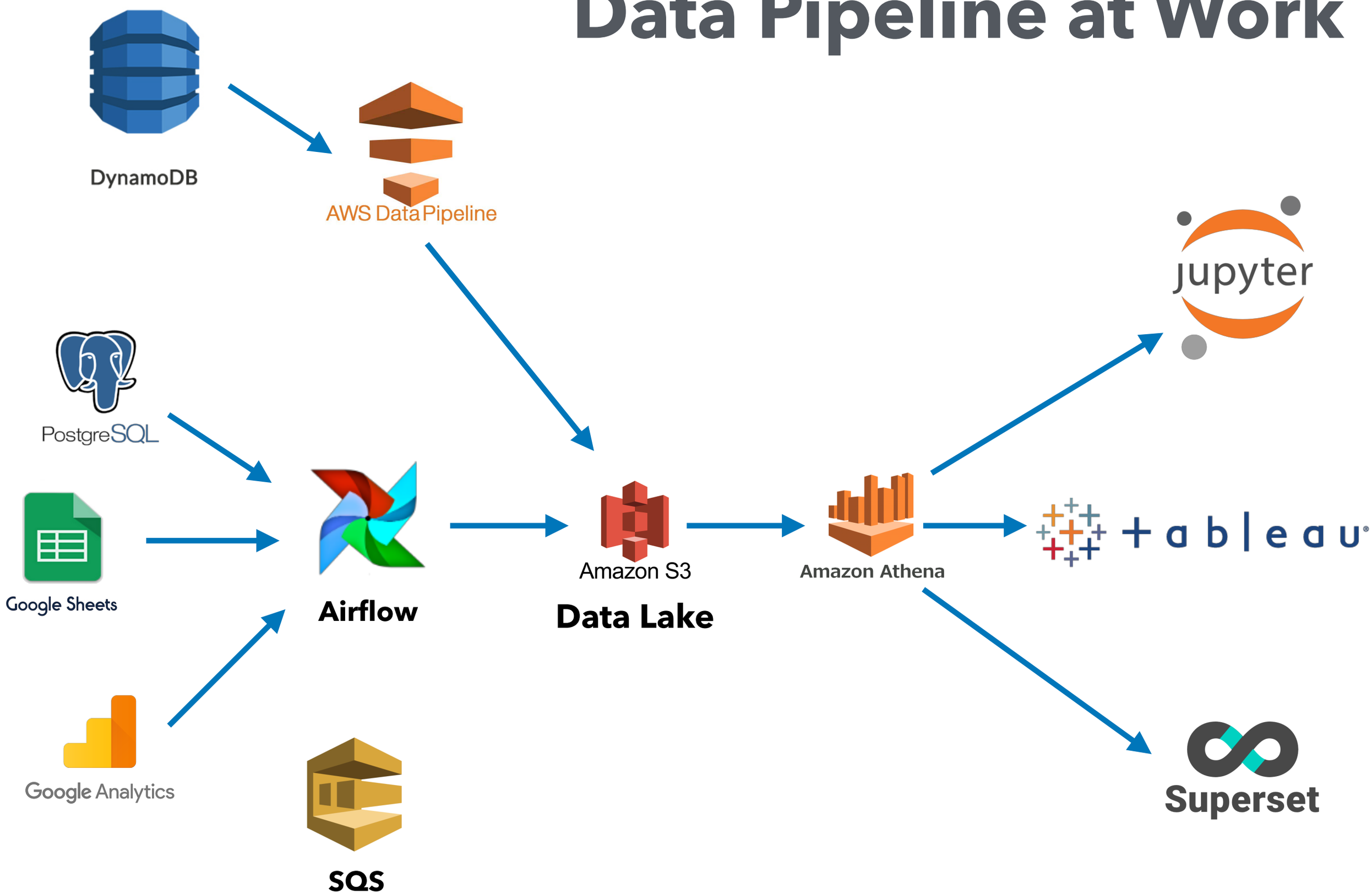



```
In [33]: plt.figure(figsize=(15, 6))
         palette = sns.set_palette('husl')
         df_quarter_count = df_quarter_count.round(4)
         g = sns.barplot(x='year_and_month', y='conversion_rate', data=df_quarter_count, palette=palette)
         g.set(xlabel='Time', ylabel='Conversion rate')
         g.set_title('Conversion rate of all leads over all traffics in quarter')
         for p in g.patches:
             g.annotate(str(p.get_height()), (p.get_x(), p.get_height() * 1.015));
         g.set_xticklabels(g.get_xticklabels(), rotation=0);
```





Data Pipeline at Work



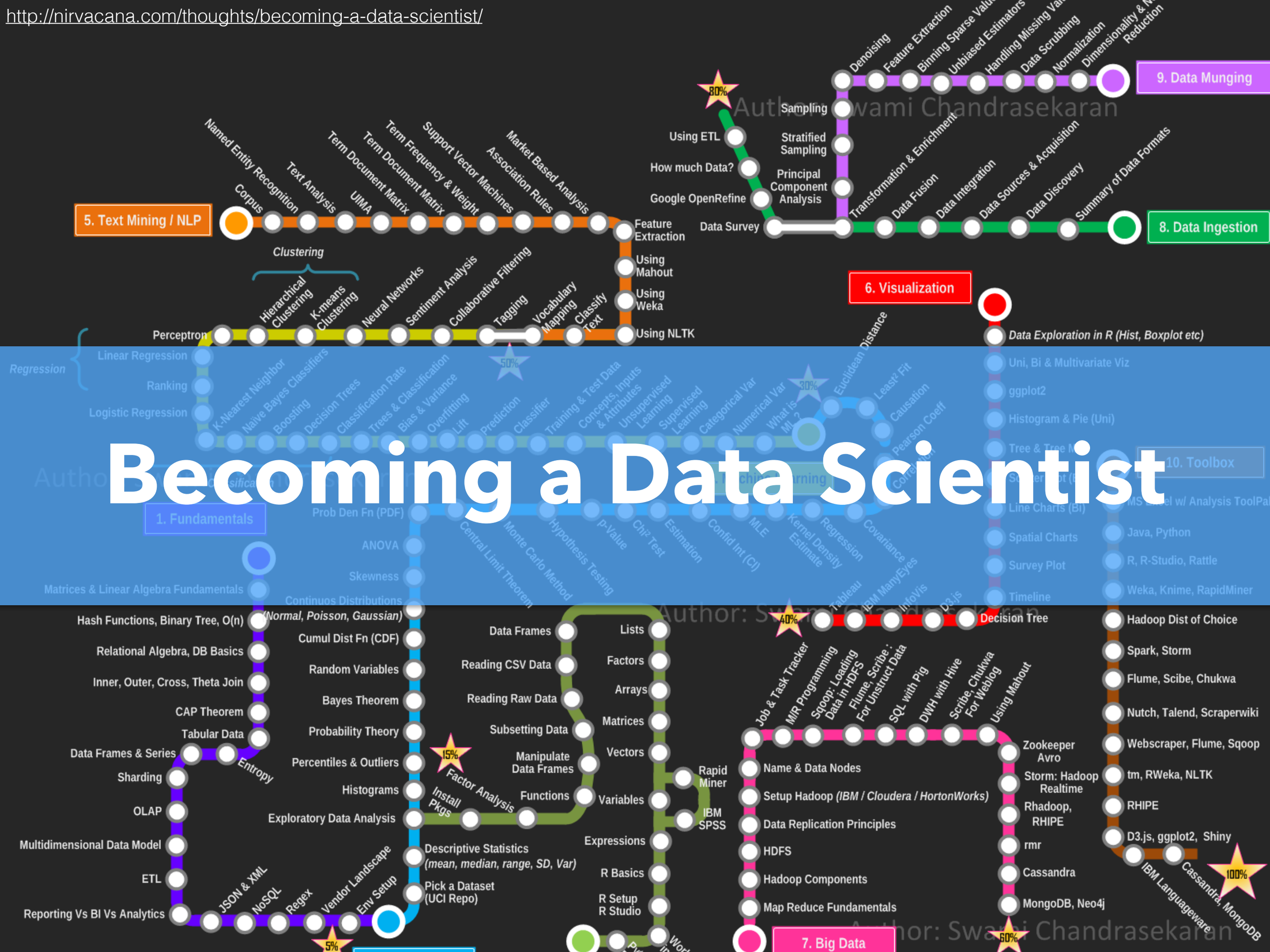
Course Outline

1. Data Science and Tools
2. Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

http://nirvacana.com/thoughts/becoming-a-data-scientist/

Becoming a Data Scientist



(((Josh Wills)))
@josh_wills

Following

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

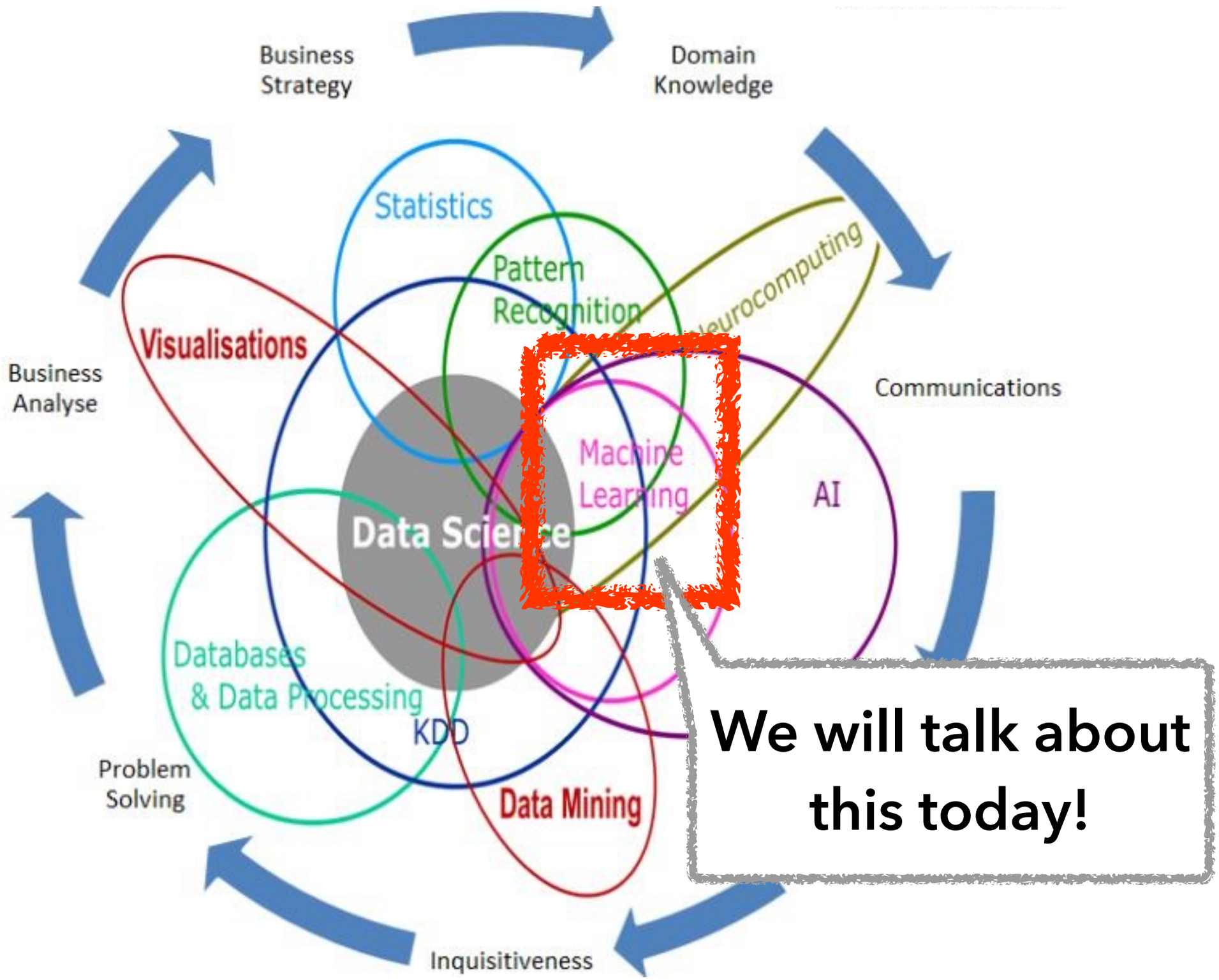
RETWEETS **1,474** LIKES **1,001**

11:55 PM - 3 May 2012

1.5K 1K

https://twitter.com/josh_wills/status/198093512149958656

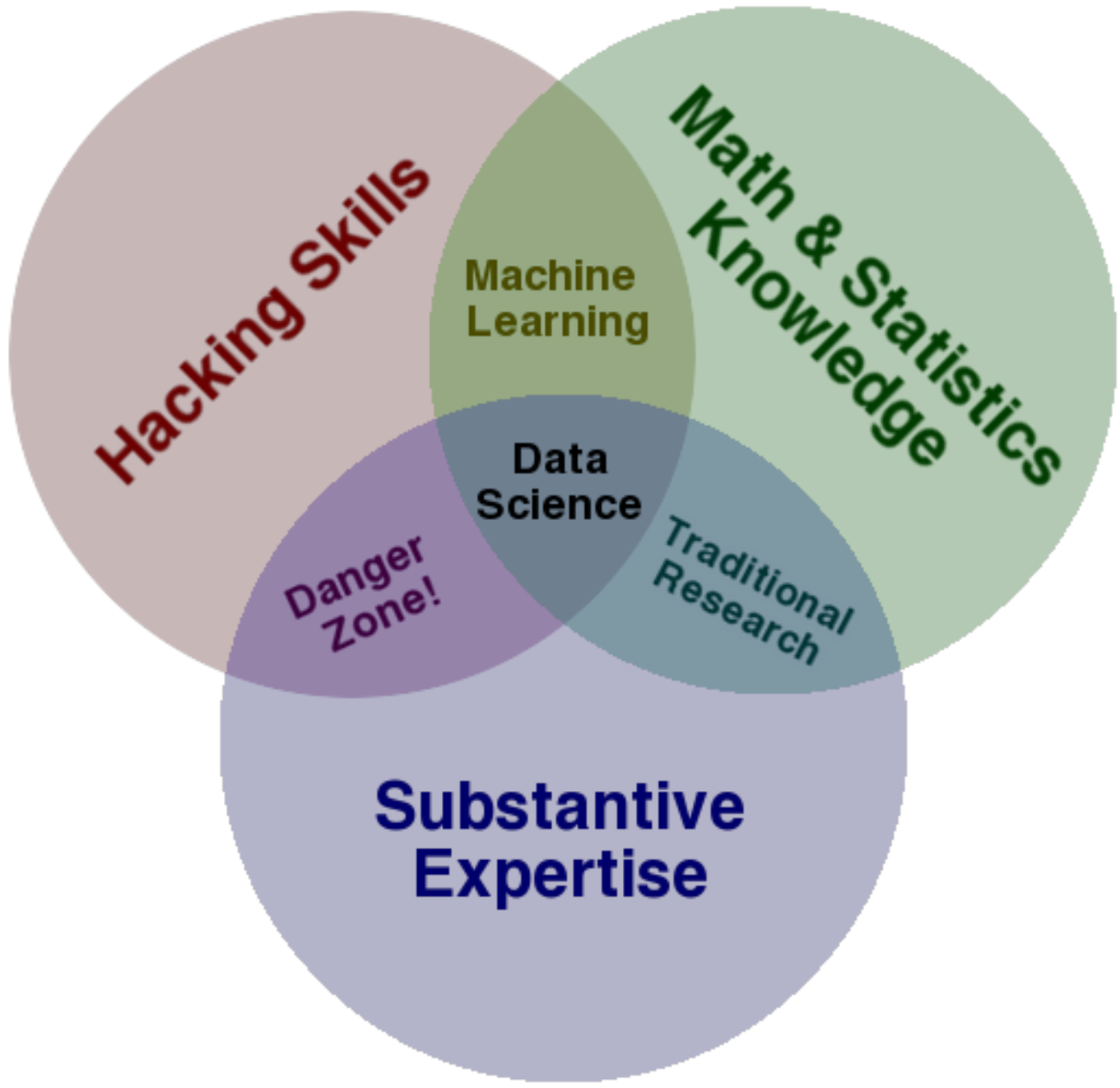
Data Science is Multidisciplinary



We will talk about this today!

http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html

Data Science Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Bad Data Scientists

1. Don’t like scientific enquiries
2. Make decisions on gut feeling
3. Like the details
4. Don’t like the details
5. Have closed mind
6. Hate failure
7. Have only one fraction of skills

https://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A348267

What do Data Scientists do?

Photo by Carlos Muza on Unsplash

Data Science Process in Organization

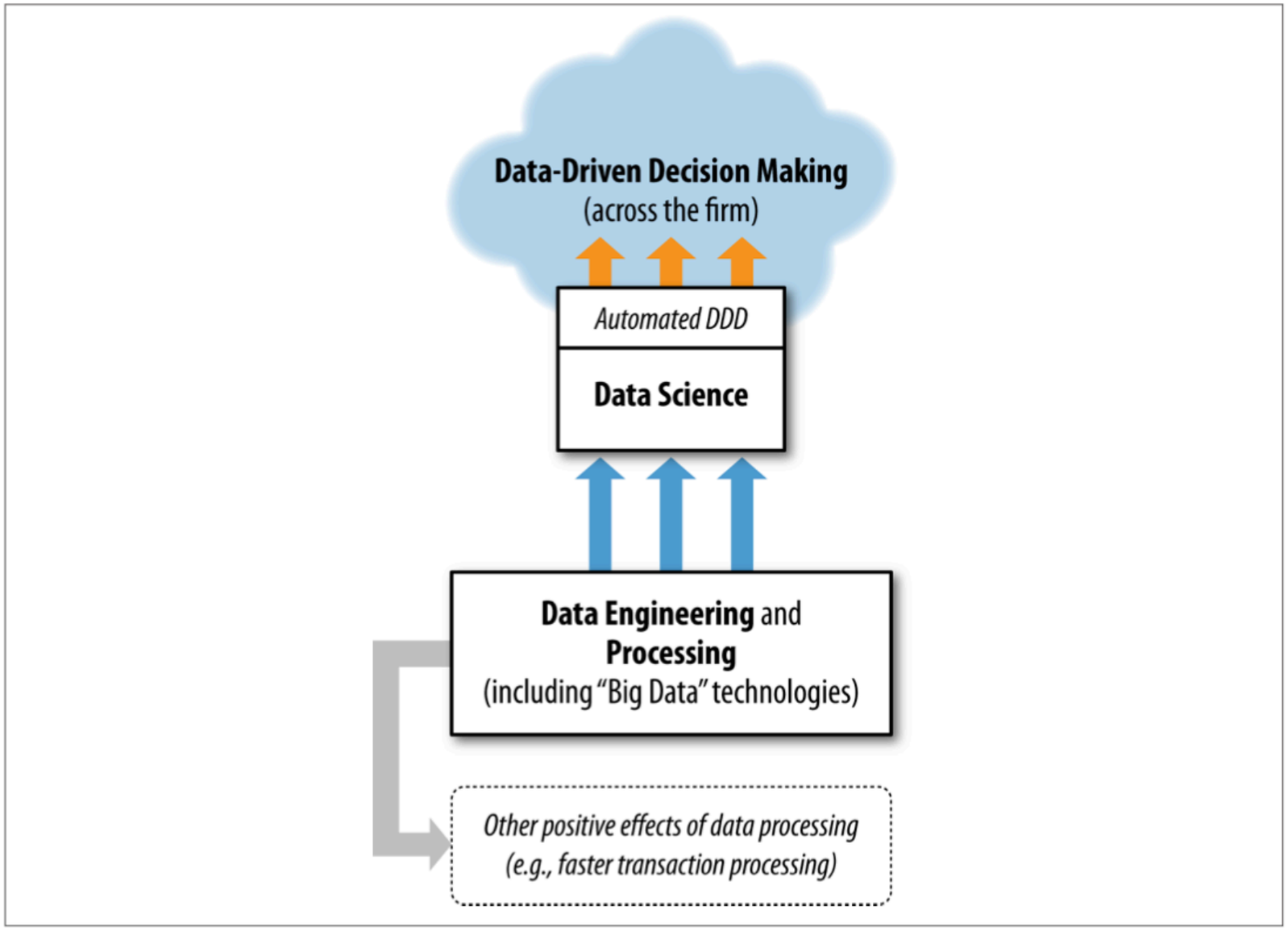
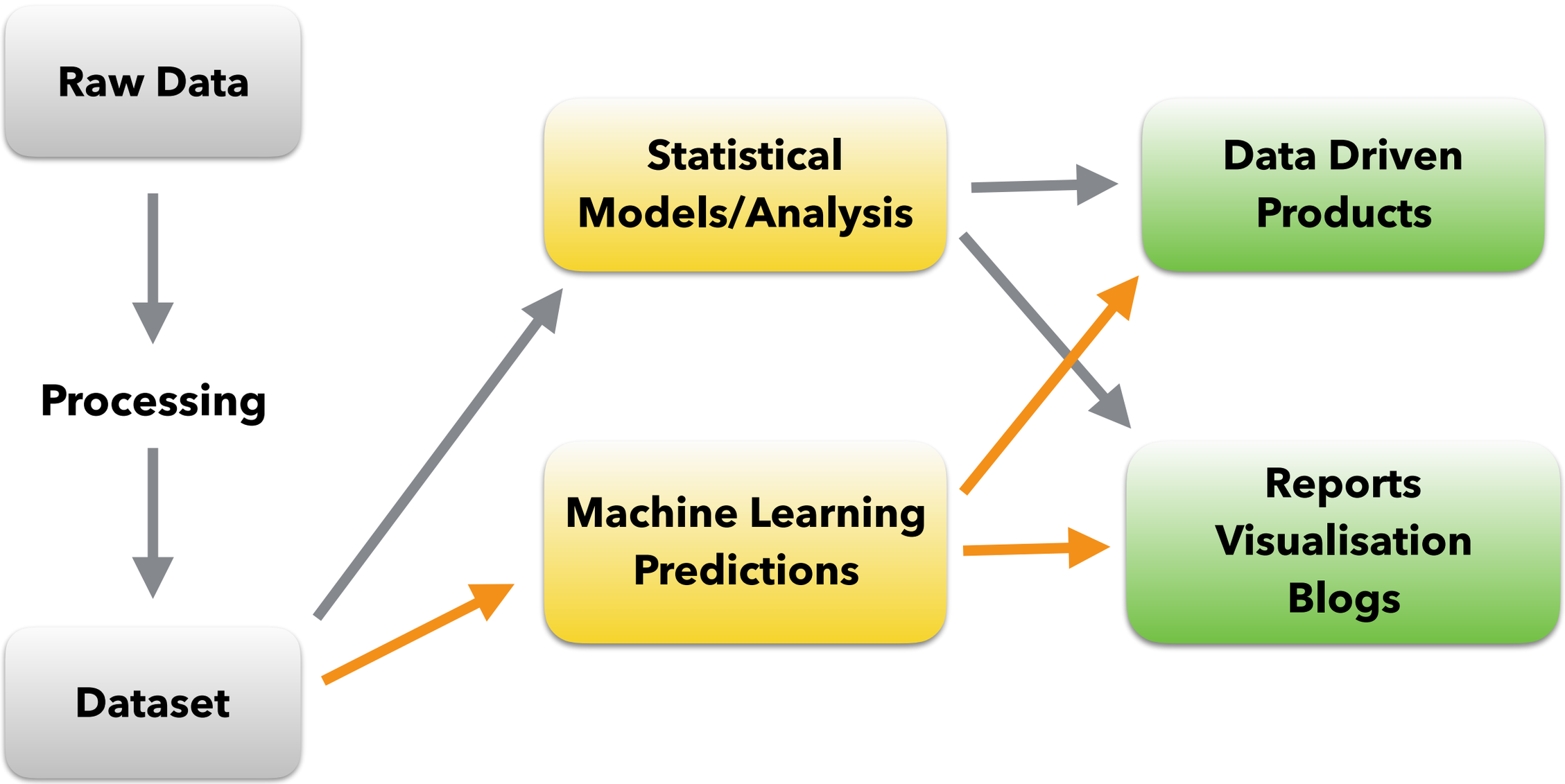


Figure 1-1. Data science in the context of various data-related processes in the organization.

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Provost & Fawcett

What do Data Scientists do?



As a Data Scientist, We Need

1. Good problem
2. Good data
3. Good **TOOLs**

Solving Data Science using Good TOOLS

Photo by Todd Quackenbush on Unsplash

WHY PYTHON?

- Simple!
- Powerful
- Flexible

WHY PYTHON?

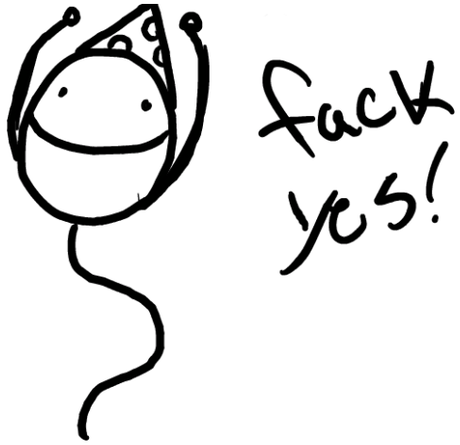
- **Simple!**
- Powerful
- Flexible

Ja..:

```
public class HelloWorld {
    public static void main(String[], args) {
        System.out.println("Hello, World!");
    }
}
```

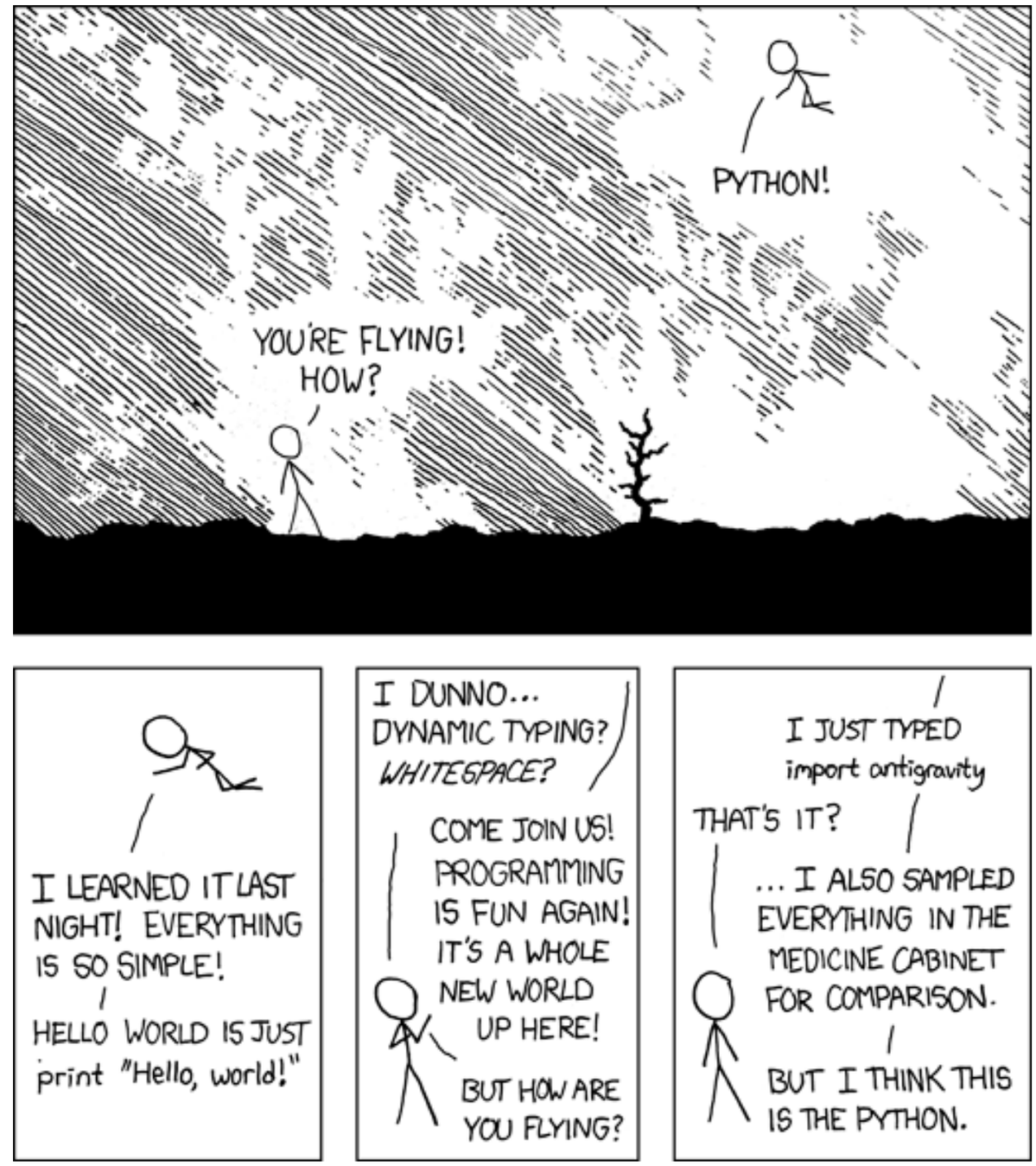
Python:

```
print(”Hello, World”)
```



WHY PYTHON?

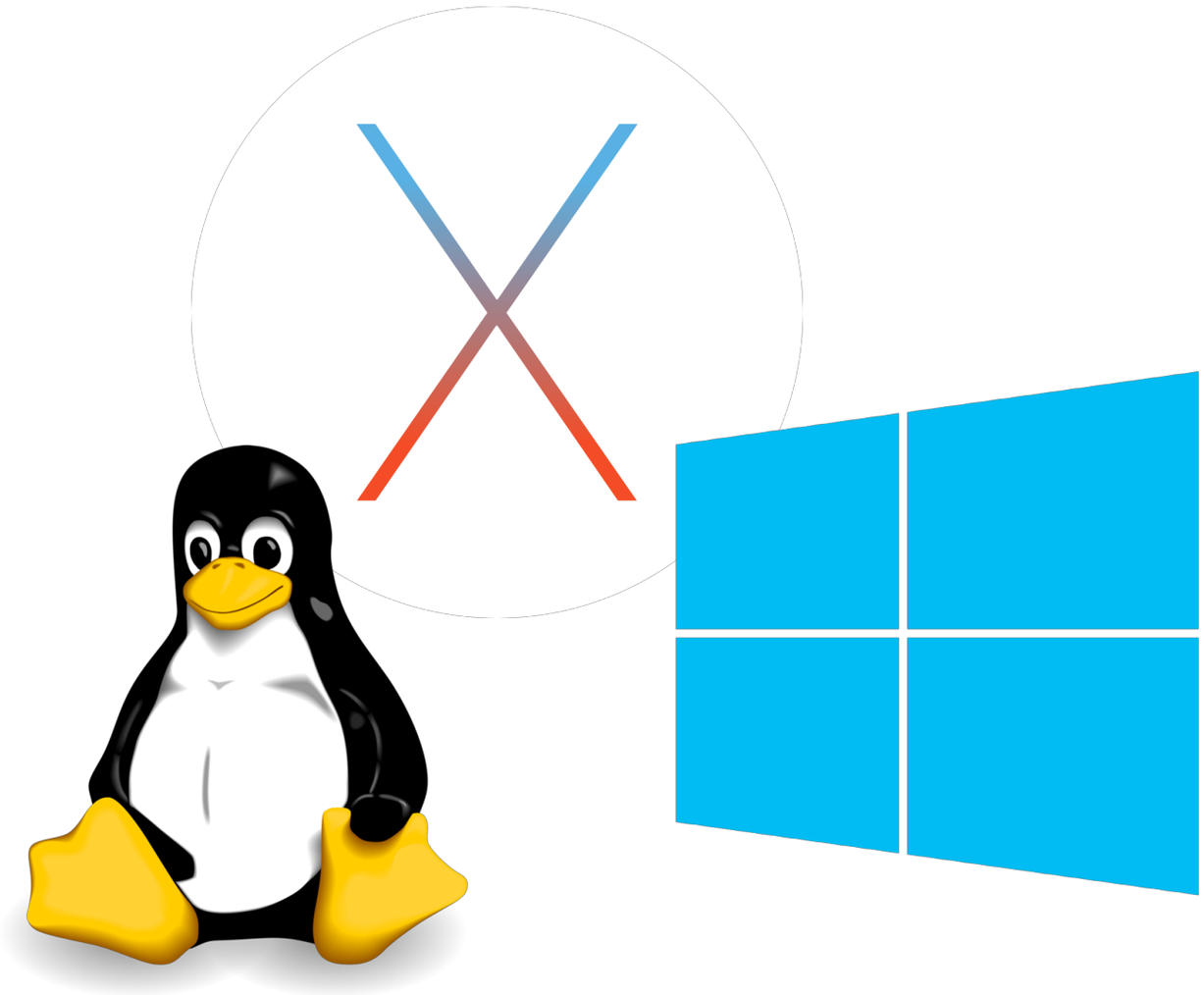
- Simple!
- **Powerful**
- Flexible



http://xkcd.com/353/

WHY PYTHON?

- Simple!
- Powerful
- **Flexible**



Popular

Quora

DISQUS

Google

TensorFlow

NASA

Pinterest

Insta

Coding Environments and Editors

and more..

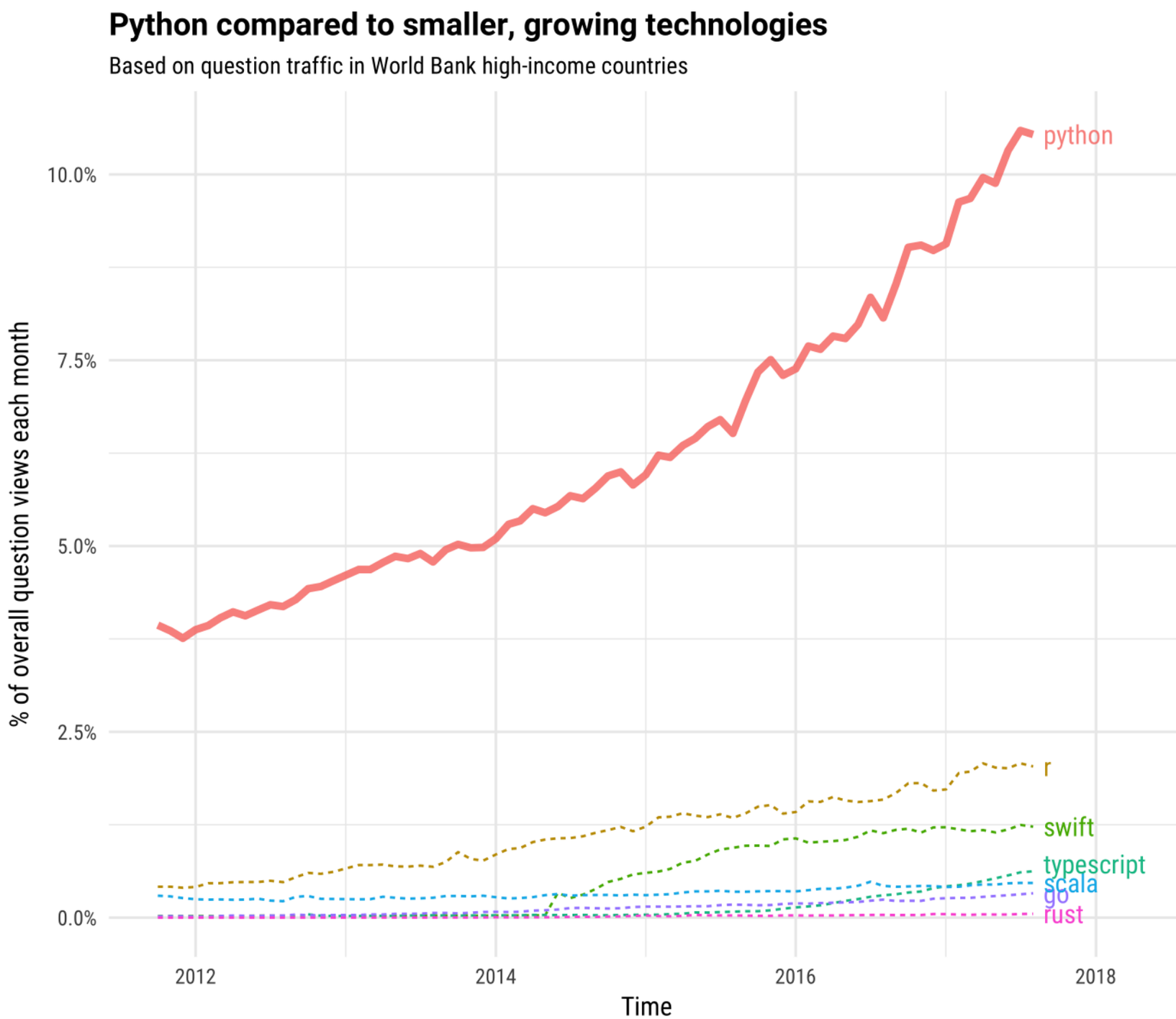
“Python = Potential”

-JEREMY ACHIN

https://www.youtube.com/watch?v=CoxjADZHUQA

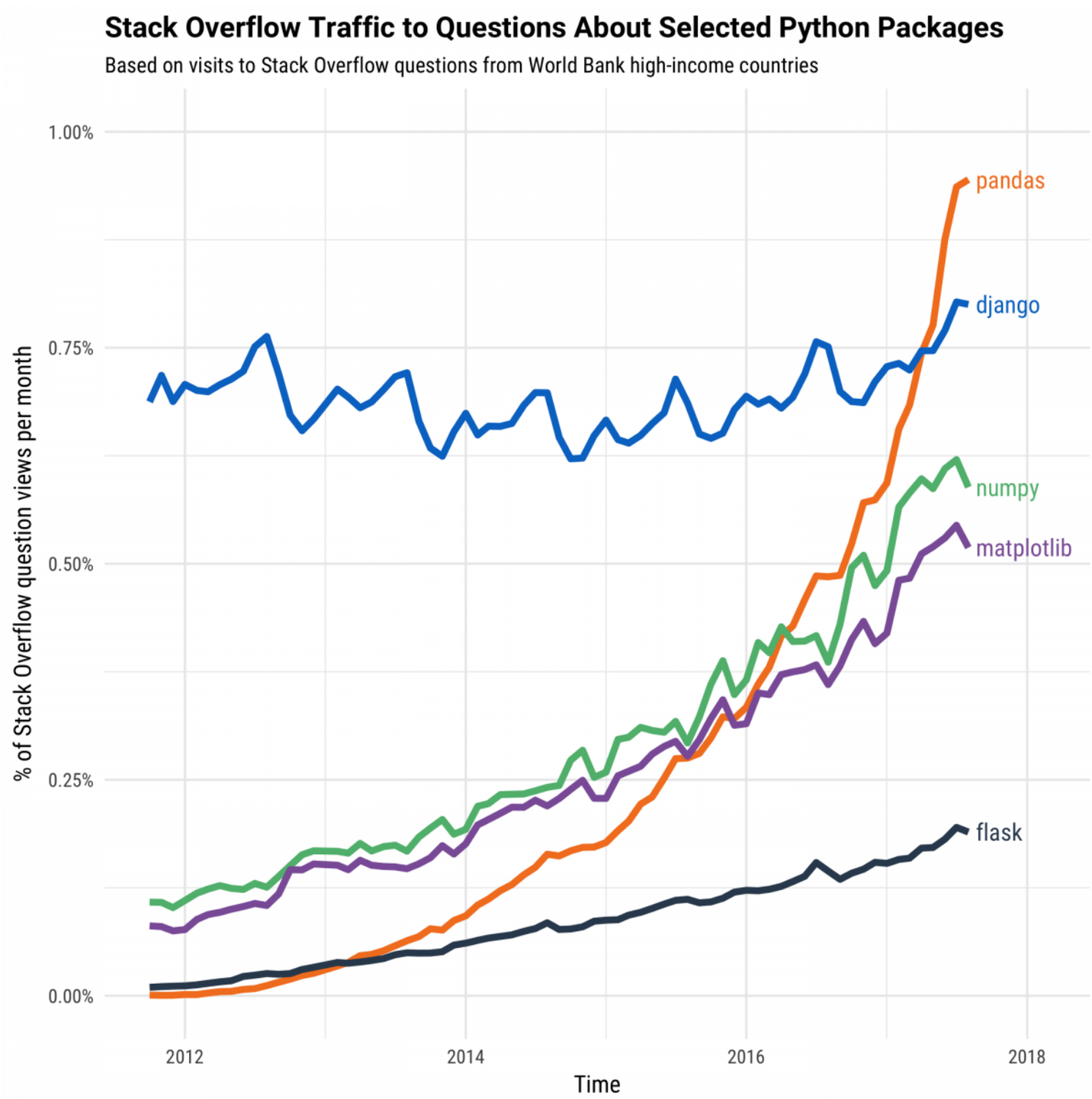
THE INCREDIBLE GROWTH OF PYTHON

https://stackoverflow.blog/2017/09/06/incredible-growth-python/



WHY IS PYTHON GROWING SO QUICKLY?

https://stackoverflow.blog/2017/09/14/python-growing-quickly/



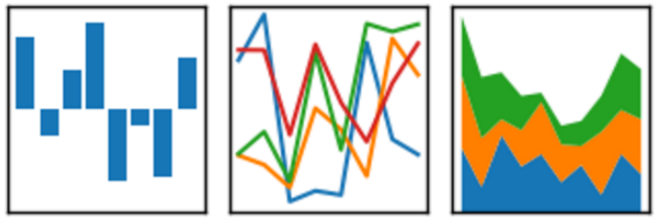
Pandas

Python Data Analysis Library

http://www.toonts.com/panda/

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NUMFocus sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

0.18.1 Final (May 3, 2016)

This is a minor release from 0.18.0 and includes a large number of bug fixes along with several new features, enhancements, and performance improvements. We recommend that all users upgrade to this version.

VERSIONS

Release
0.18.1 - May 2016
download // docs // pdf

Development
0.18.2 - July 2016
github // docs

Previous Releases
0.18.0 - download // docs // pdf
0.17.1 - download // docs // pdf
0.17.0 - download // docs // pdf
0.16.2 - download // docs // pdf
0.16.1 - download // docs // pdf
0.16.0 - download // docs // pdf
0.15.2 - download // docs // pdf
0.15.1 - download // docs // pdf
0.15.0 - download // docs // pdf
0.14.1 - download // docs // pdf

http://pandas.pydata.org/

panel **da**ta

(multi-dimensional data)

http://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf

scikit learn

http://scikit-learn.org/

High-quality, easy-to-use, implementations of popular Machine Learning algorithms

Well-written documentation & Consistent APIs

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search

Previou 9 1.9. Naive Bayes | Next 1.11. Ensemble m... | Up 1. Supervised le...

This documentation is for scikit-learn **version** 0.16.1 — Other versions

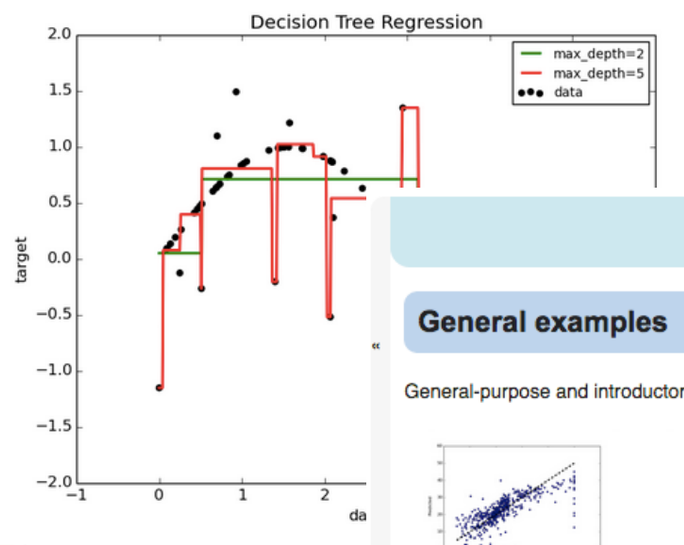
If you use the software, please consider citing scikit-learn.

1.10. Decision Trees
1.10.1. Classification
1.10.2. Regression
1.10.3. Multi-output problems
1.10.4. Complexity
1.10.5. Tips on practical use
1.10.6. Tree algorithms: ID3, C4.5, C5.0 and CART
1.10.7. Mathematical formulation
- 1.10.7.1. Classification criteria
- 1.10.7.2. Regression criteria

1.10. Decision Trees

Decision Trees (DTs) are a non-parametric supervised learning method used for *classification* and *regression*. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

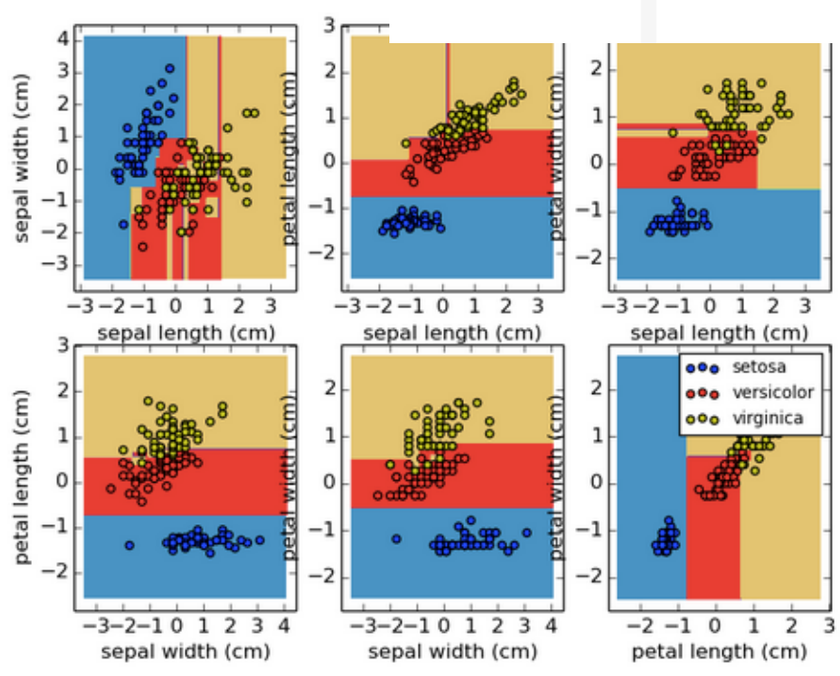
For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.



Alternatively, the probability of each class can be predicte leaf:

```
>>> clf.predict_proba(iris.data[:1, :])
array([[ 1.,  0.,  0.]])
```

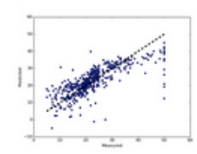
Decision surface of a d



Examples

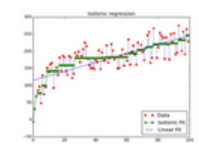
General examples

General-purpose and introductory examples for the scikit.



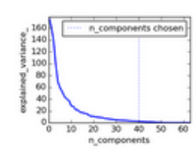
Plotting Cross-Validated Predictions

Concatenating multiple feature extraction methods

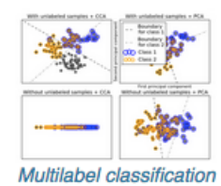


Isotonic Regression

Imputing missing values before building an estimator

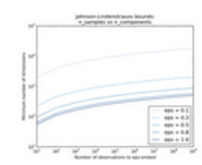


Pipelining: chaining a PCA and a logistic regression



Multilabel classification

Face completion with a multi-output estimators



The Johnson-Lindenstrauss bound for embedding with random projections

Models and algorithms chosen and implemented by experts

Cover most Machine Learning tasks

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *feature selection*, *non-negative matrix factorization*. — *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*. — *Examples*

Preprocessing

Feature extraction and normalization.

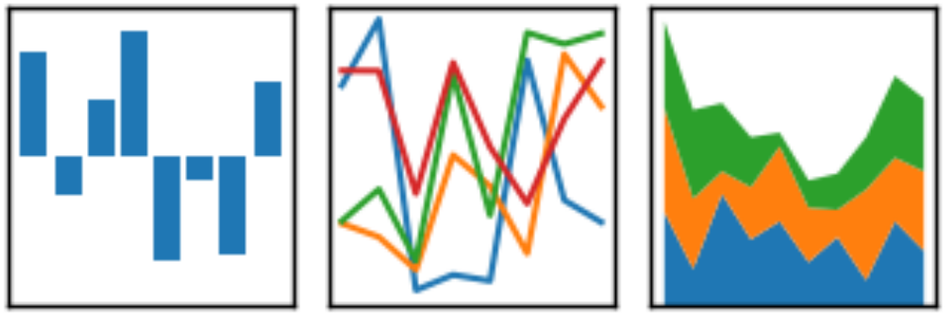
Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*. — *Examples*

Scale to most data problems

https://www.python.org/

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



http://pandas.pydata.org/

http://scikit-learn.org/

ANACONDA®

https://www.continuum.io/downloads

Our Goals in this Course!

- Take data and produce insights
- Build an exploratory data visualizations
- Model or predict something with data

Our Goals in this Course!

- Take data and produce insights
- Build an exploratory data visualizations
- **Model or predict something with data**

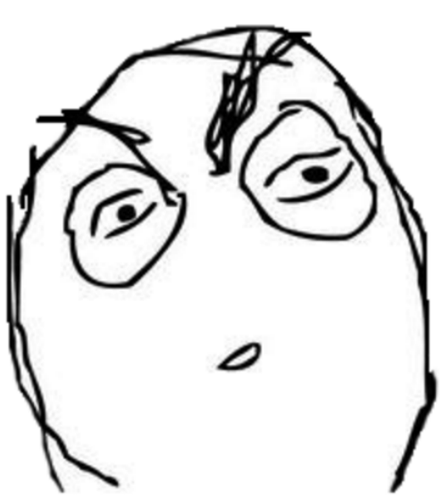
Learning Data Science

- Learn to love data
- Learn by doing
- Learn to communicate insights
- Learn from experts and peers
- Learn constantly the degree of difficulty

https://www.dataquest.io/blog/learn-data-science/

“การเรียนรู้ต้องสนุก”

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\ |\Sigma|^{1/2}}\ exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$



wat

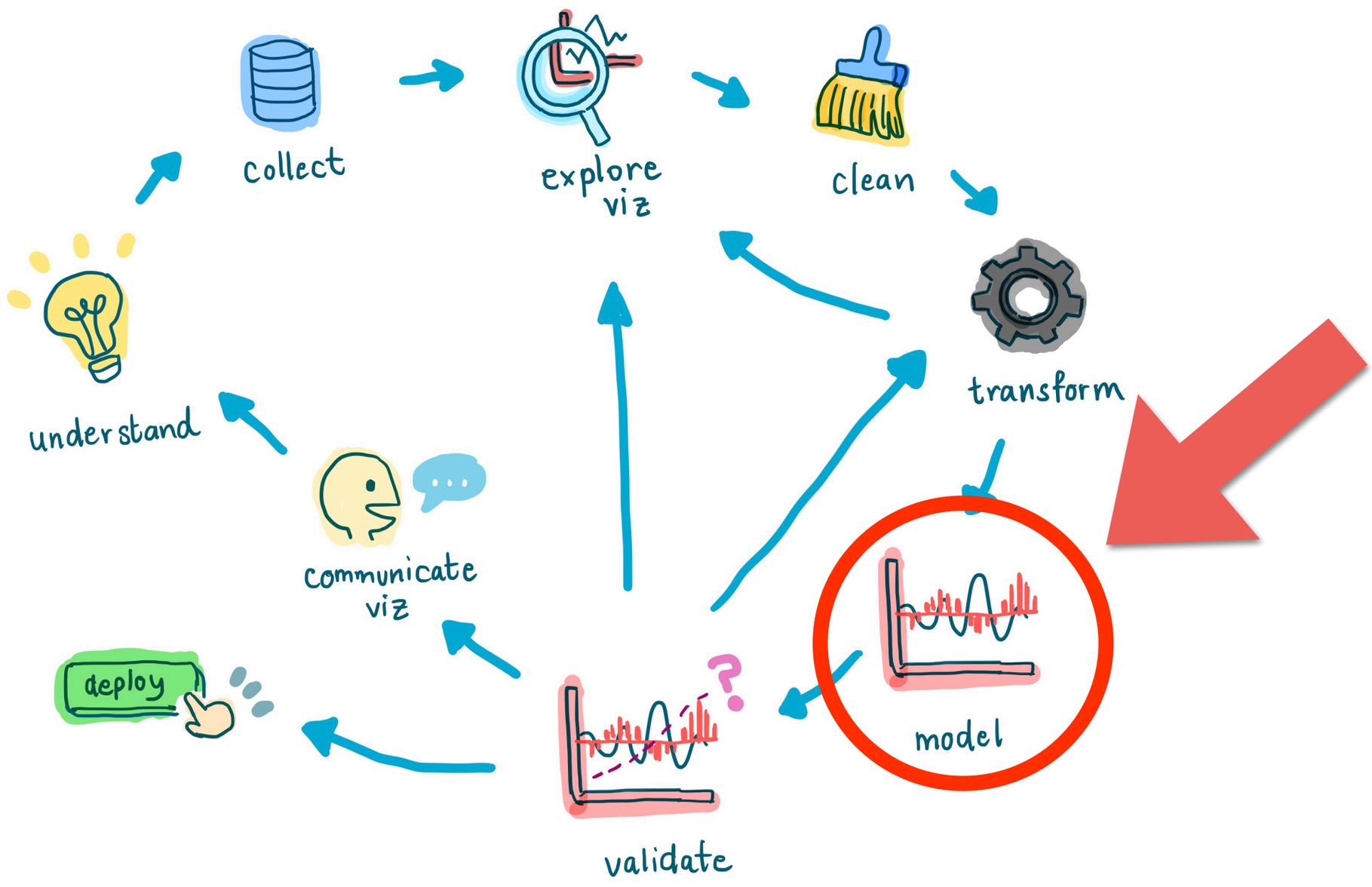
แนวทางการเรียนรู้
Machine Learning

“Top-Down” Approach

1. Process of **“applied”** machine learning
2. Tools
3. Practice, practice, and practice!
4. Theory..

Photo by Lweendo Hachifeka on Unsplash

Machine Learning Process



Refreshing Data Analysis with Python

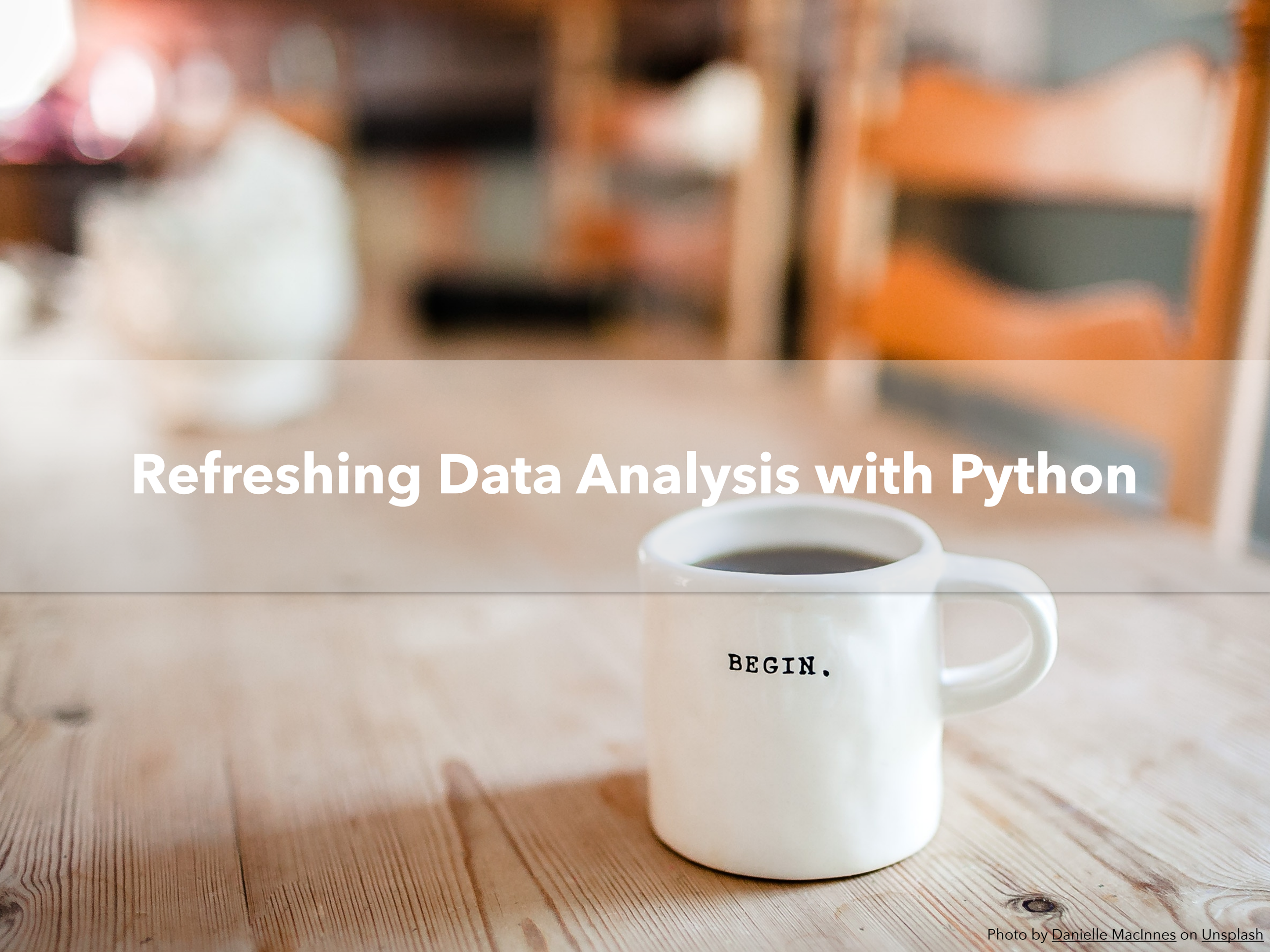


Photo by Danielle MacInnes on Unsplash

scikit learn

http://scikit-learn.org/

Course Outline

1. Data Science and Tools
2. Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Course Outline

1. ✅ Data Science and Tools
2. Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

What is this Deck?

- What is machine learning (ML) and what types of it exist?
- How does it work?
- How do we validate, select, and evaluate a model?

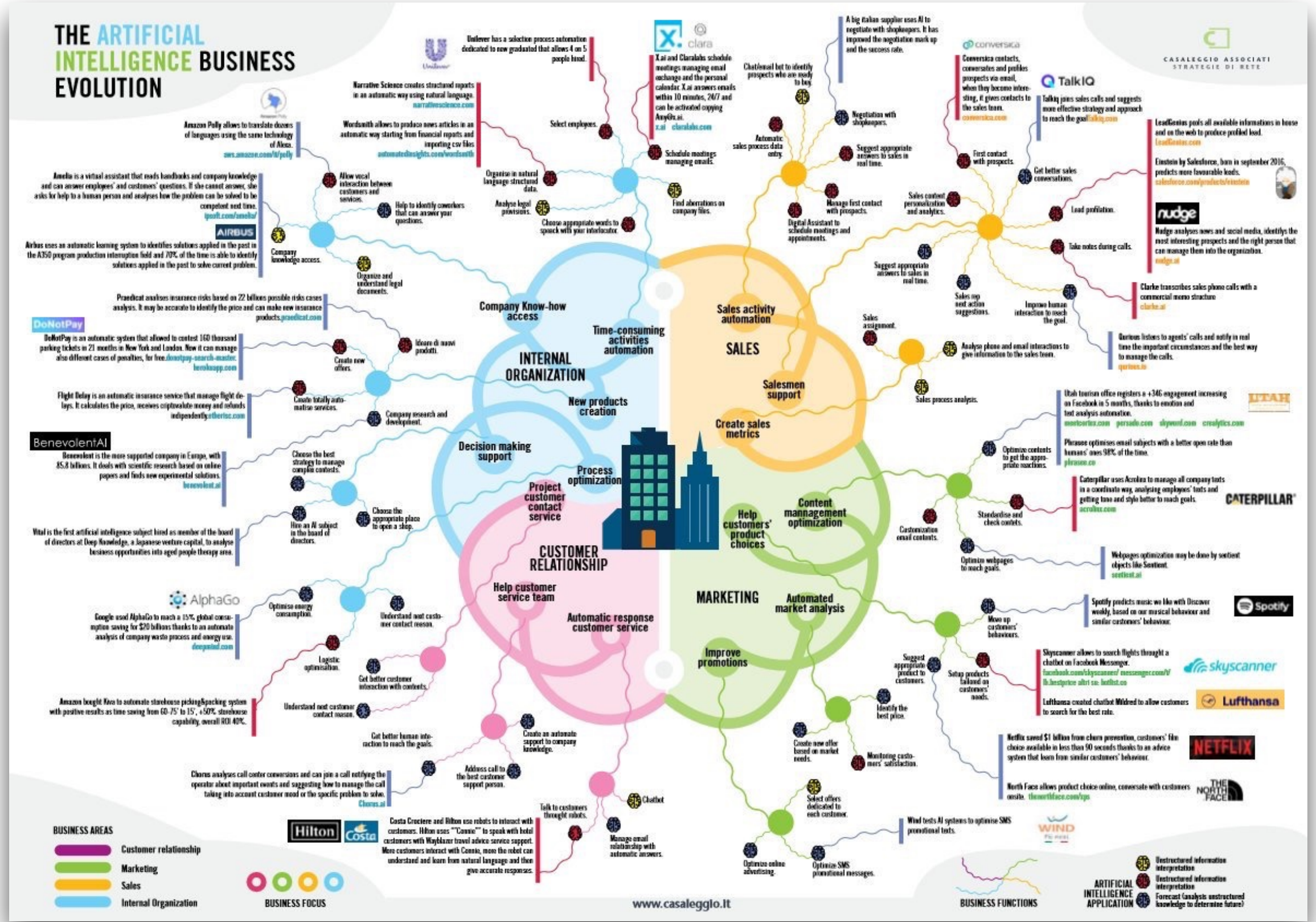
What's **Artificial Intelligence**, **Machine Learning**, and **Deep Learning**?

Pioneer in AI

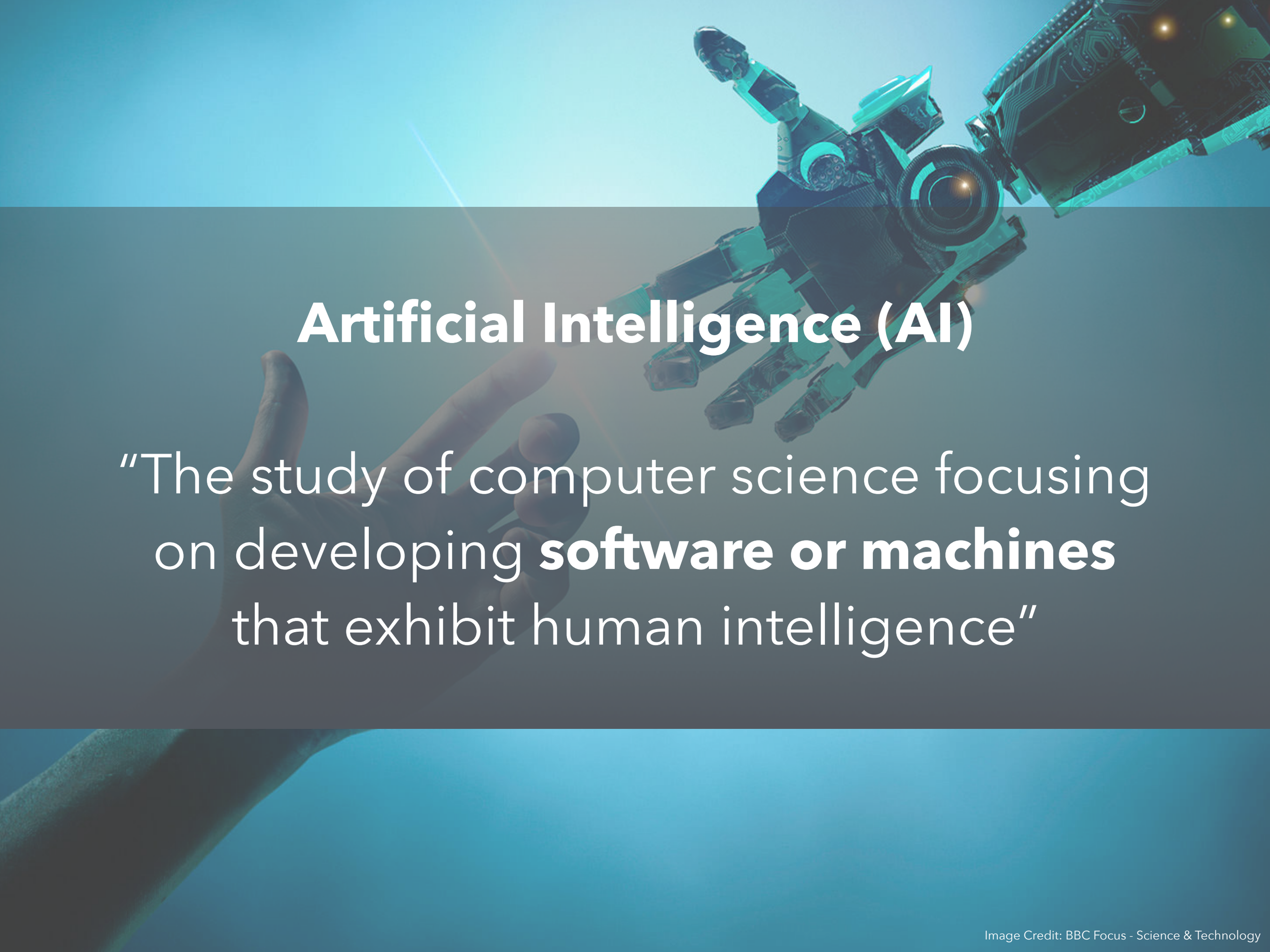
On February 24, 1956, Arthur Samuel's Checkers program

http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/ibm700series/impacts/

Impact of AI on Companies Today



https://twitter.com/inteliment/status/933876632112136193



Artificial Intelligence (AI)

"The study of computer science focusing on developing **software or machines** that exhibit human intelligence"

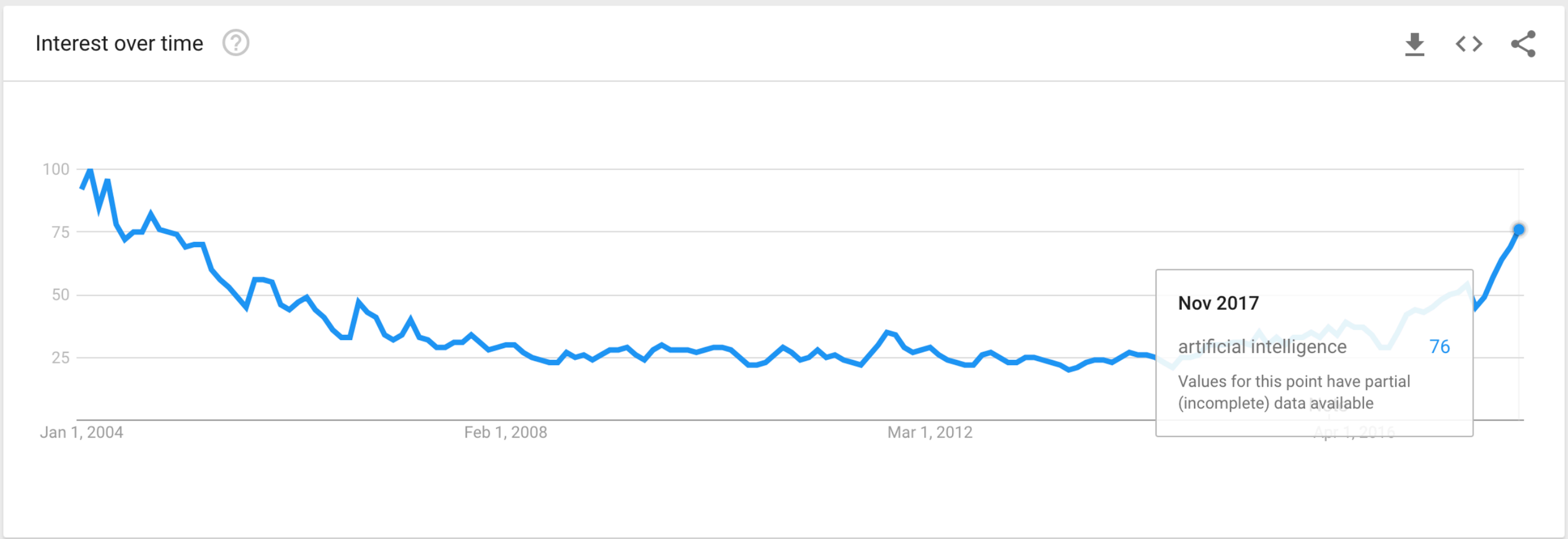
Image Credit: BBC Focus - Science & Technology

Artificial Intelligence

Artificial Intelligence (AI) can be defined as:

“Machine that see the world and think and understand the work and be intelligent the way that people are.”

The Rise of AI Adoption



https://trends.google.com/trends/explore?date=all&q=artificial%20intelligence

"AI is the new electricity."
–Andrew Ng
deeplearning.ai
https://www.wired.com/2013/05/neuro-artificial-intelligence/

There are 2 Terms of AI

Strong AI vs. Weak AI

Will Wilson
@WAWilsonIV

Follow

"What's the difference between AI and ML?"

"It's AI when you're raising money, it's ML when you're trying to hire people."

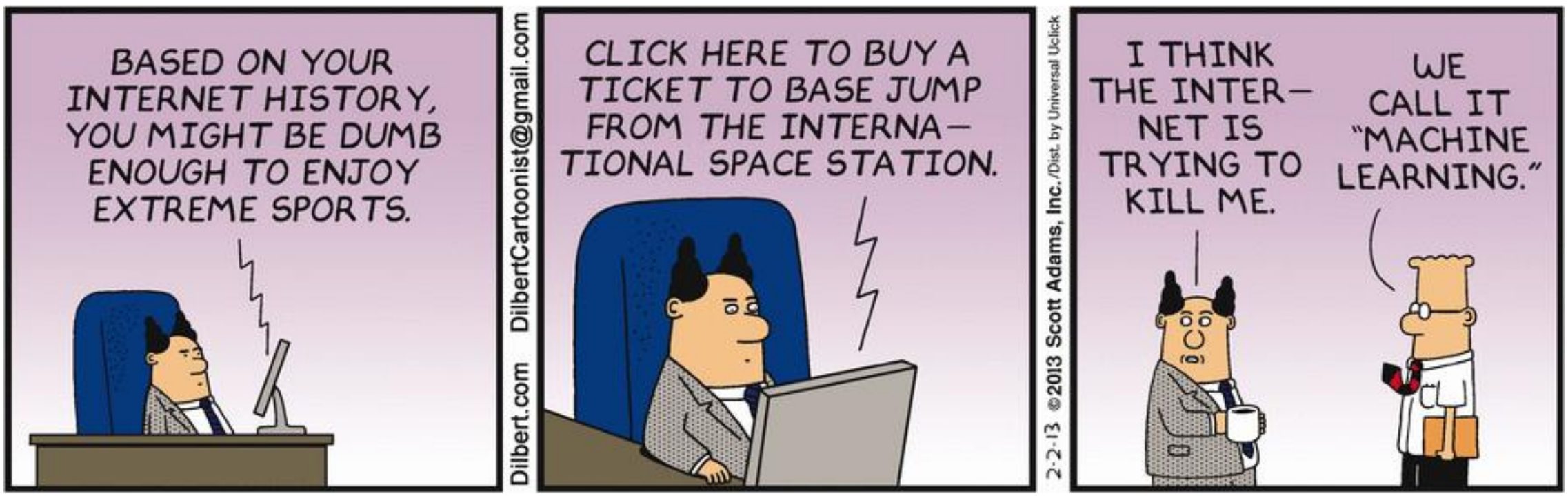
10:45 PM - 31 Oct 2017

4,687 Retweets **9,666** Likes

67 4.7K 9.7K

What is Machine Learning?

Machine Learning?



https://www.pinterest.com/pin/3237030963862645/?lp=true

Machine Learning

what society thinks I do

what my friends think I do

what my parents think I do

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

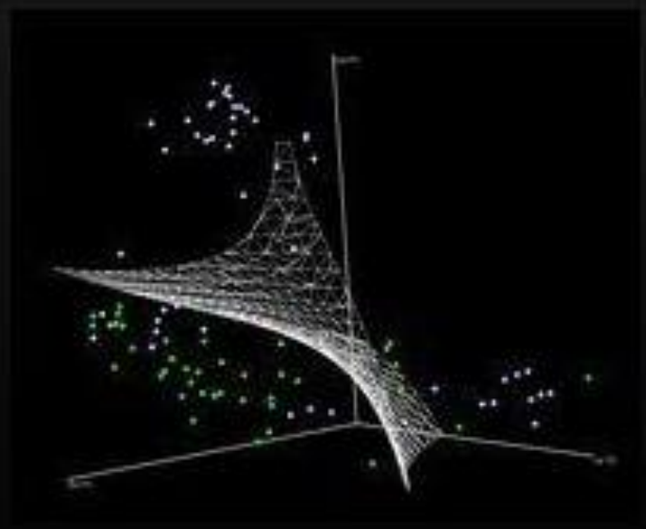
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do

```
>>> from scipy import SVM
```

what I really do

https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/

~~Rule-based programming~~

Machine Learning

Machine Learning (ML) can be defined generally as:

"Approach that can learn from experience to find patterns in data"

Machine Learning

- Summarizes data.
- Finds or discovers correlations in data.
- Performs many **narrow** tasks very well but still has no common sense!

ML: Definition by Tom Mitchell (1988)

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."

ML: Definition by Peter Norvig (2014)

- Give a specification of a function *f*.
- Implement *f* that meets the specification.

Programming

- Give example (x, y) pairs.
- Induce f such that $y = f(x)$ for given pairs and generalizes well for unseen x.

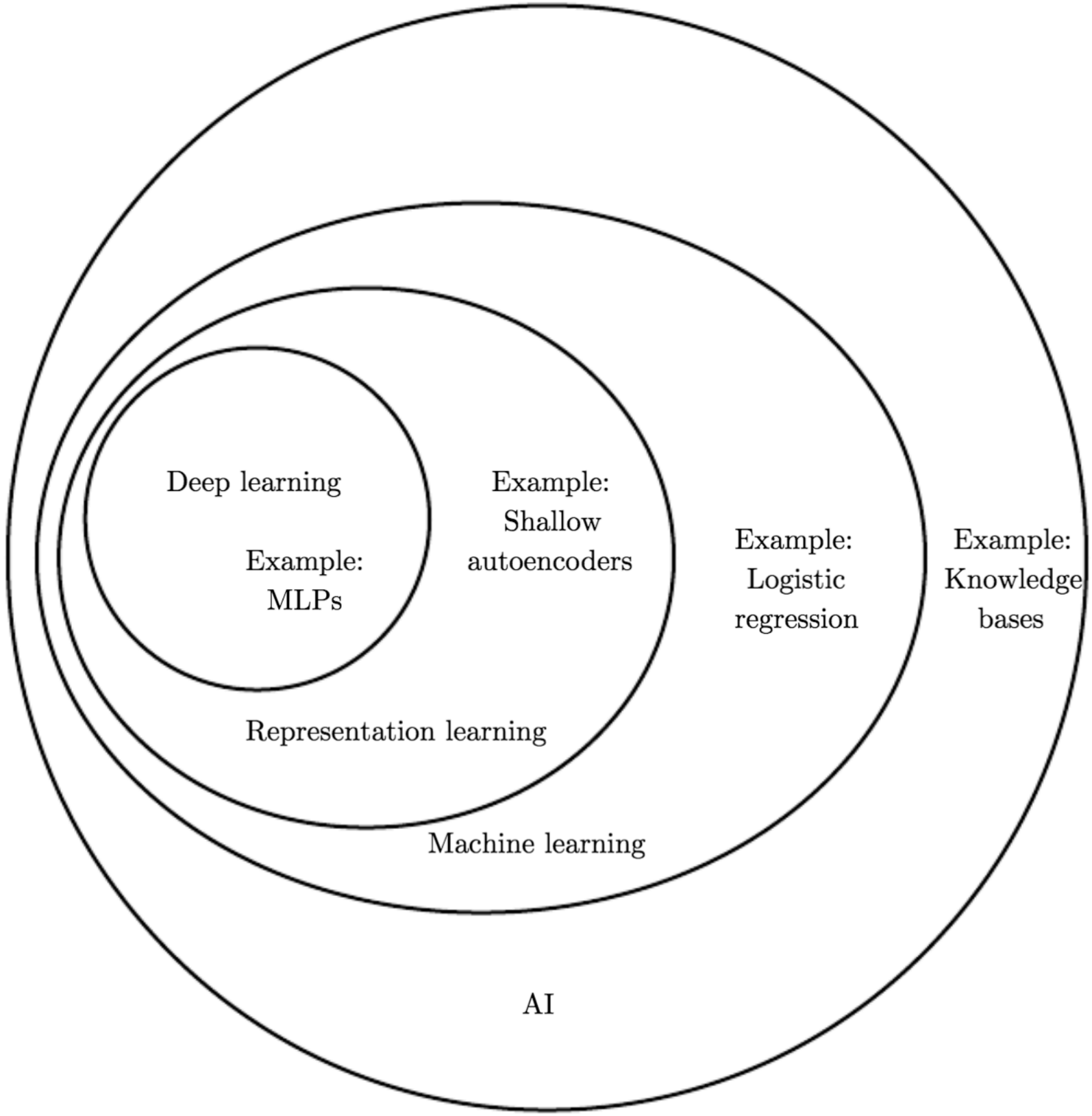
Machine Learning

Deep Learning

Deep Learning (DL) can be defined as:

"Technique in ML that is based on learning data representations"

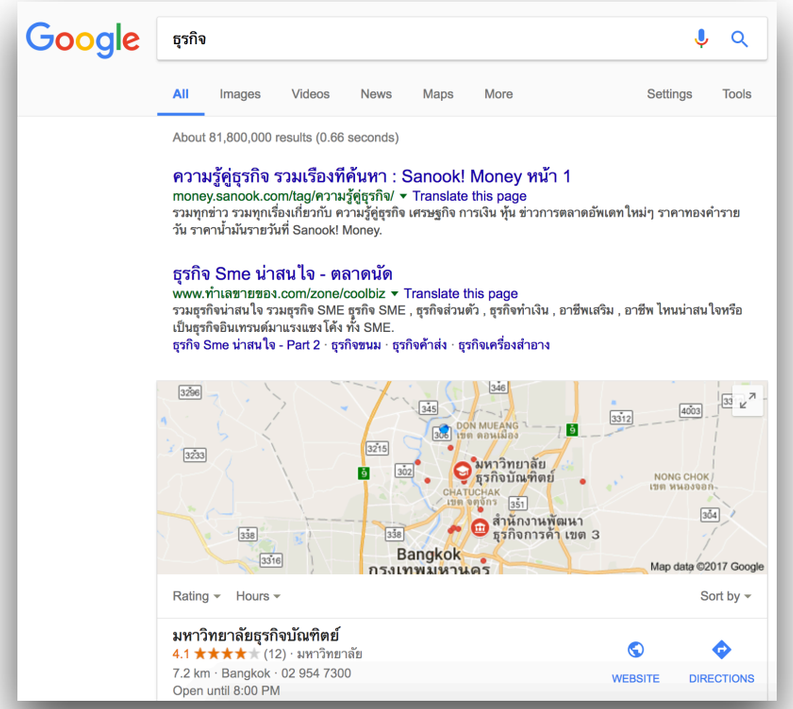
Summary



http://www.deeplearningbook.org/

Real-World Examples

- **Web Search and Recommendation Engines**
- Online Dating
- Crime Prediction
- Self-Driving Cars
- AI



NETFLIX

amazon

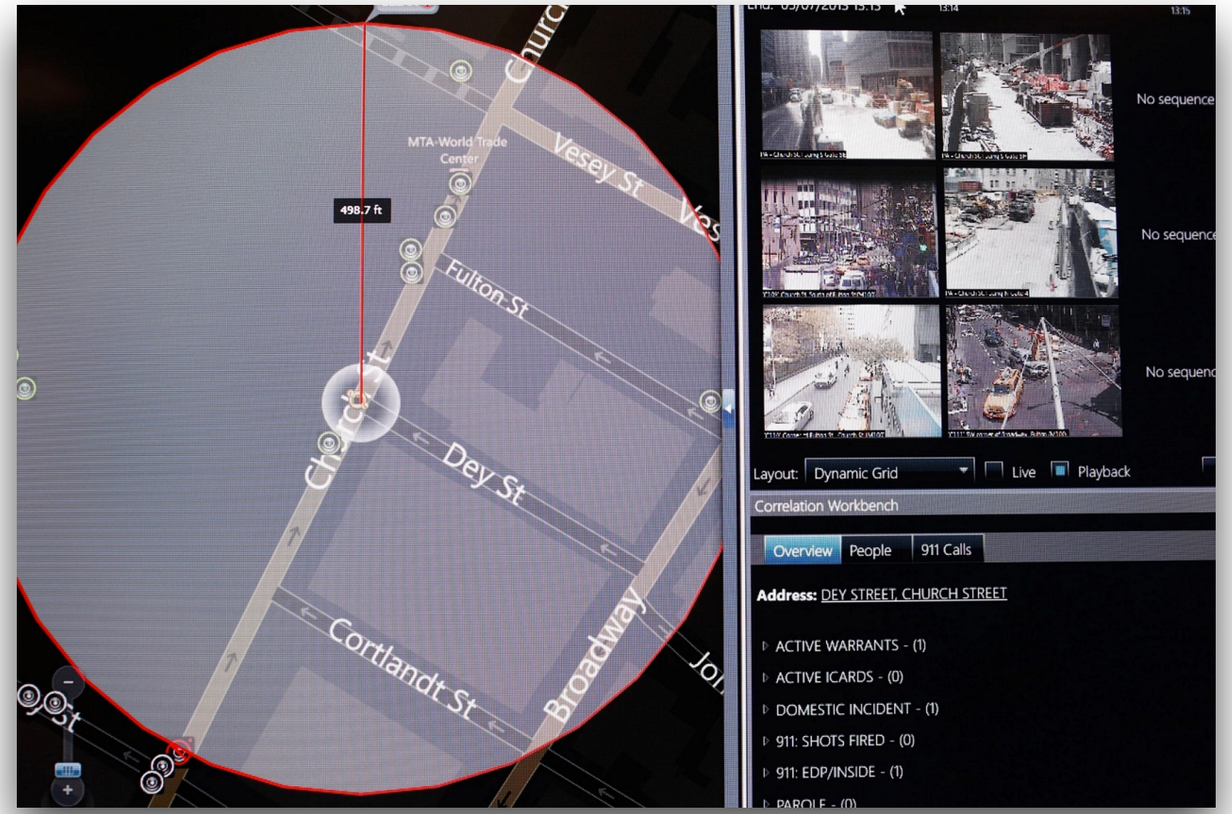
Real-World Examples

- Web Search and Recommendation Engines
- **Online Dating**
- Crime Prediction
- Self-Driving Cars
- AI

Real-World Examples

- Web Search and Recommendation Engines
- Online Dating
- **Crime Prediction**
- Self-Driving Cars
- AI

PredPol



http://www.predpol.com/

Real-World Examples

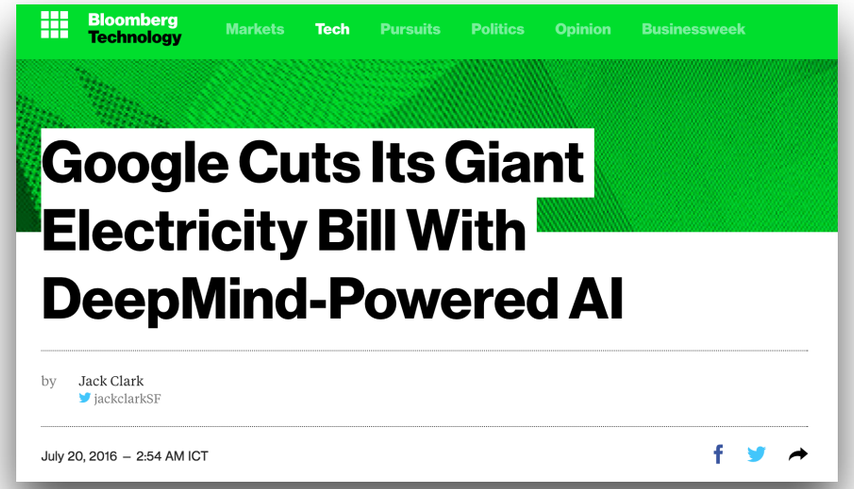
- Web Search and Recommendation Engines
- Online Dating
- Crime Prediction
- **Self-Driving Cars**
- AI

http://www.documentarytube.com/articles/self-driving-cars-when-we-will-have-them

https://www.recode.net/2017/11/7/16614780/alphabet-driverless-cars-phoenix-arizona

Real-World Examples

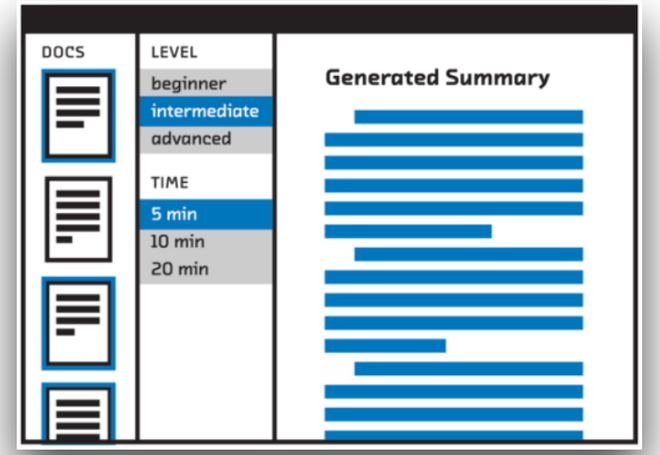
- Web Search and Recommendation Engines
- Online Dating
- Crime Prediction
- Self-Driving Cars
- **AI**



http://www.bloomberg.com/news/articles/2016-07-19/google-cuts-its-giant-electricity-bill-with-deepmind-powered-ai

AlphaGo

https://deepmind.com/alpha-go.html



http://blog.fastforwardlabs.com/2016/04/11/new-tools-to-summarize-text.html

More in Various Industries

- Artificial intelligence has learned to **spot suicidal tendencies** from brain scans (link)
- Microsoft's AI is learning to **write code by itself**, not steal it (link, arXiv)
- Algorithm that can **detect pneumonia from chest X-rays** at a level exceeding practicing radiologists (link)
- Neva **automates customer service** and support to deliver unprecedented precision and quality. (link)
- AI can **hunt down missile sites** in China (link)
- The titans of AI are **getting their work double-checked** by students (link)

Challenges in ML/AI Adoption

1. Accessing to data
2. No defined business case
3. Lack of people power
4. Lack of emotional intelligence
5. Better at specialized tasks
6. Difficult to collaborate between AIs

When to Apply ML/AI?

- Human expertise is absent
- Humans are unable to explain their expertise (speech recognition, vision, language)
- Solution changes with time (tracking, temperature control, preferences)
- Solution needs to be adapted to particular cases (personalization)
- Problem is too big for our limited reasoning capabilities (calculating webpage ranks, matching ads to facebook pages)

Why do we need Machine Learning?

- Hard to write programs that solve problems like object recognition.
- We don’t know what program to write since we don’t know how it’s done in our brain.
- If we know, the program might be super complicated!
- Rules change over time.

Mark Zuckerberg
5 hrs ·

Here's a good use of AI: helping prevent suicide.

Starting today we're upgrading our AI tools to identify when someone is expressing thoughts about suicide on Facebook so we can help get them the support they need quickly. In the last month alone, these AI tools have helped us connect with first responders quickly more than 100 times.

With all the fear about how AI may be harmful in the future, it's good to remind ourselves how AI is actually helping save people's lives today.

There's a lot more we can do to improve this further. Today, these AI tools mostly use pattern recognition to identify signals -- like comments asking if someone is okay -- and then quickly report them to our teams working 24/7 around the world to get people help within minutes. In the future, AI will be able to understand more of the subtle nuances of language, and will be able to identify different issues beyond suicide as well, including quickly spotting more kinds of bullying and hate.

Suicide is one of the leading causes of death for young people, and this is a new approach to prevention. We're going to keep working closely with our partners at Save.org, National Suicide Prevention Lifeline '1-800-273-TALK (8255)', Forefront Suicide Prevent, and with first responders to keep improving. If we can use AI to help people be there for their family and friends, that's an important and positive step forward.

71K

6.8K Comments 8.5K Shares
816 Saves

Like Comment Share

https://www.facebook.com/zuck/posts/10104242660091961

Are killer robots an imminent threat?

https://www.facebook.com/playgroundenglish/videos/566310110368990/

All approaches share common objective:

“Learn from Data.”

How are ML systems trained?

What do you see?

While the computer sees this..

11111110
11100101
00001010

ML and Feature Representation

Input

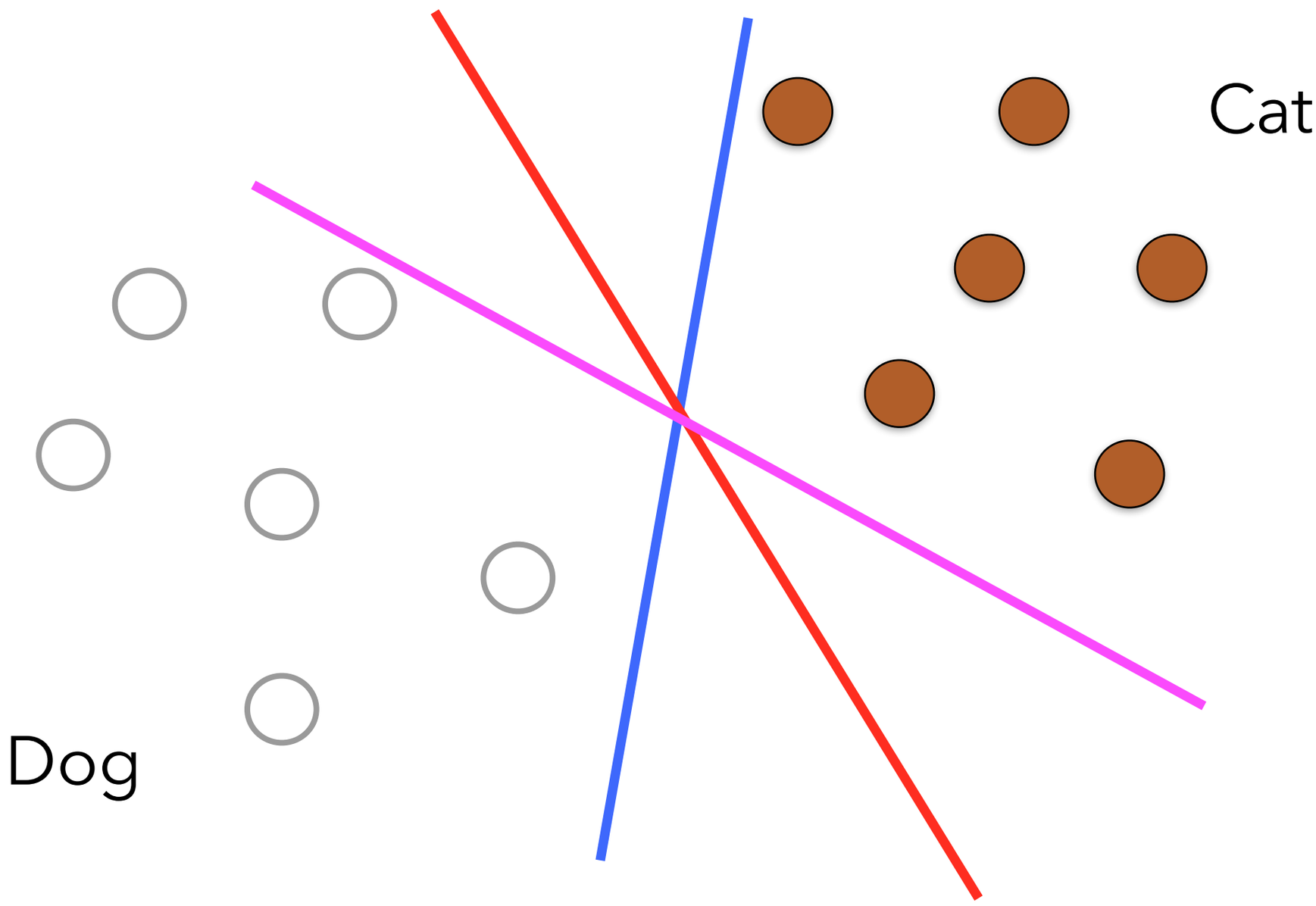
Feature Representation

Learning Algorithm

Learning Process in General

Representing Features

"Generalization"



Types of Learning

Incremental Learning

Reinforcement Learning

Supervised Learning

Active Learning

Unsupervised Learning

Semi-supervised Learning

Transduction

Types of Learning

Incremental Learning

Reinforcement Learning

Supervised Learning

Active Learning

Unsupervised Learning

Semi-supervised Learning

Transduction

Supervised Learning

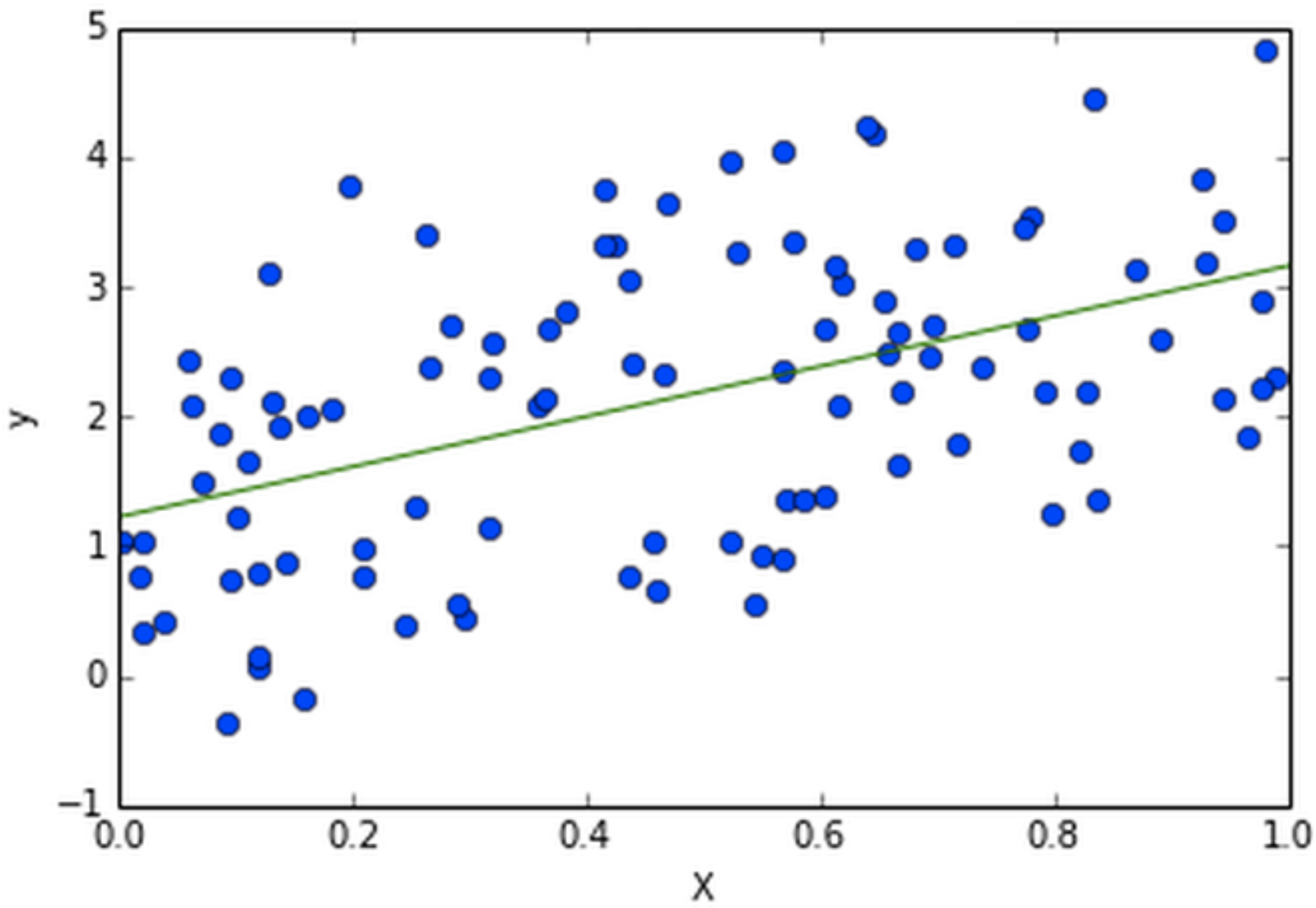
$$y = f(x)$$

Given *x*, *y* pairs, find a function f that will map new *x* to a proper *y*.

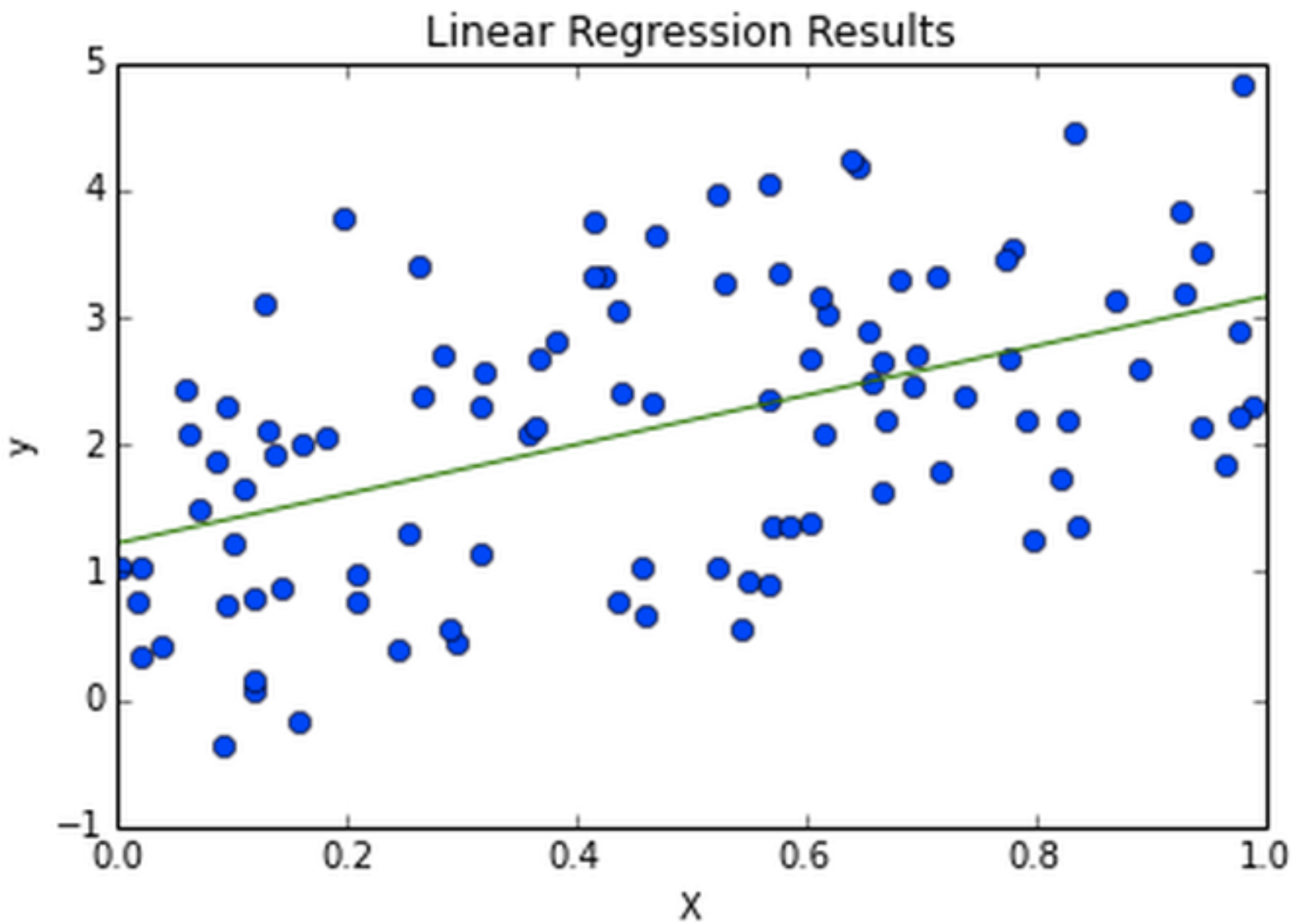
Supervised Learning: Problems

- Regression
- Classification

Supervised Learning: Regression



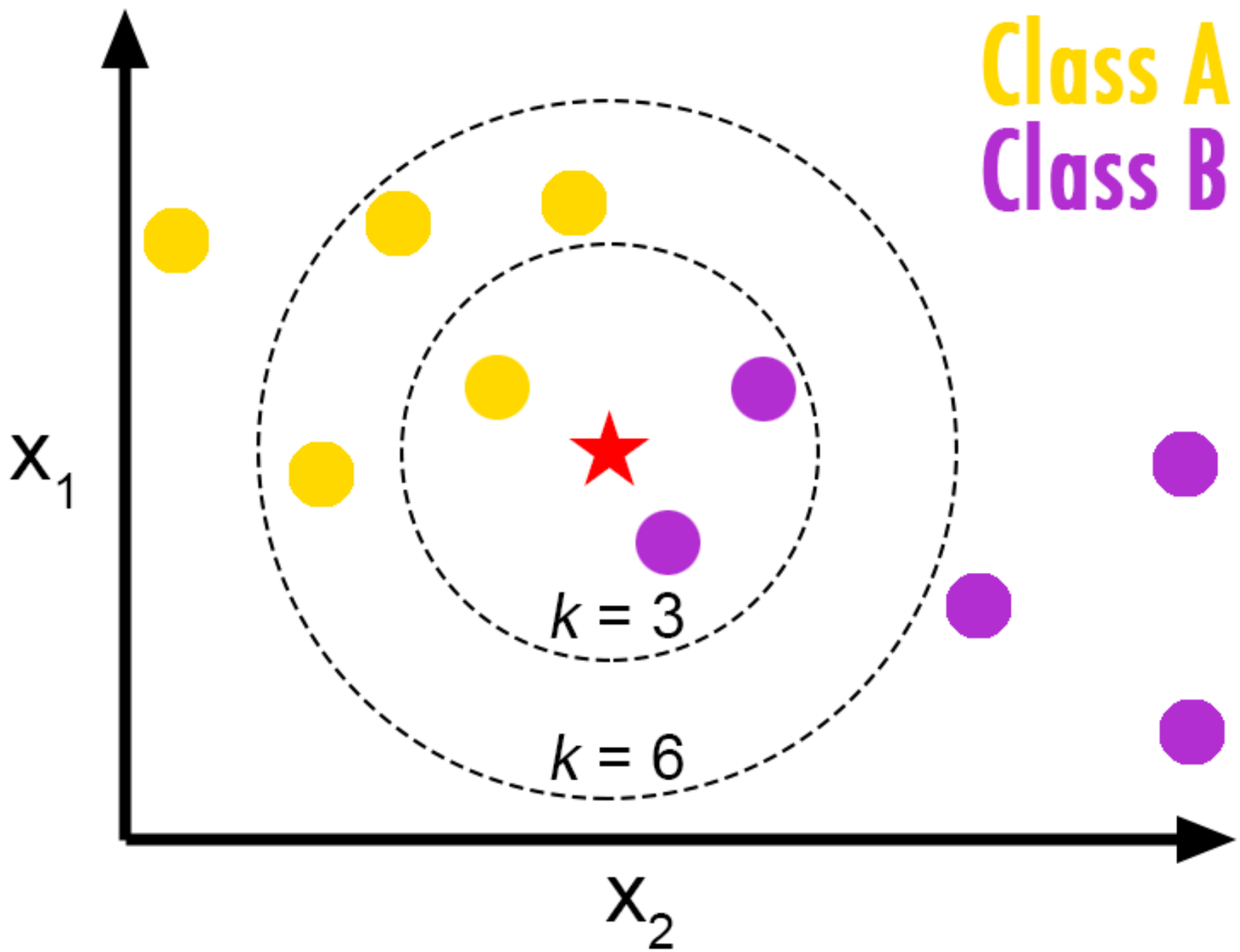
Supervised Learning: **Linear** Regression



$y = wx + b$

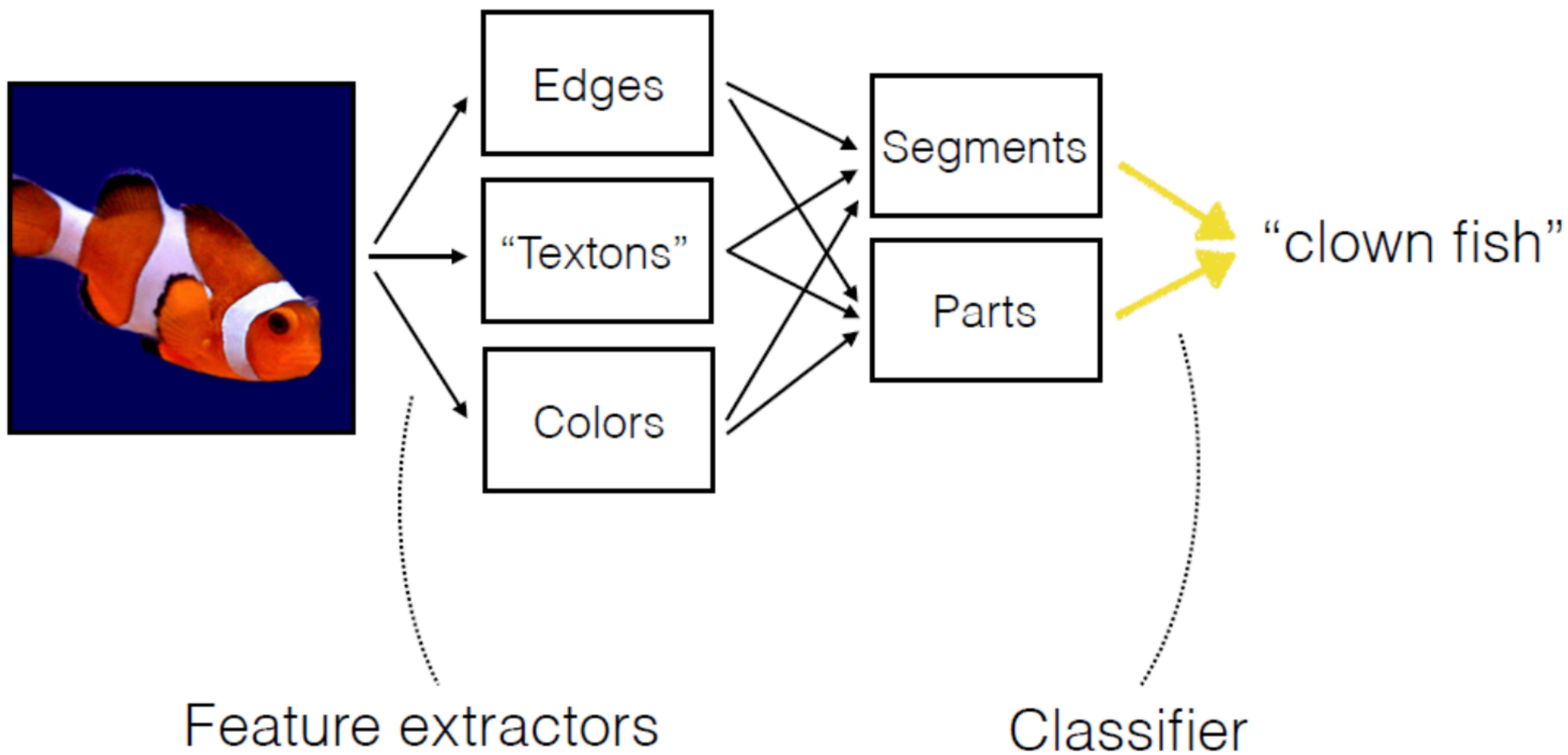
Supervised Learning: Classification

Supervised Learning: *k*-NN



http://bdewilde.github.io/blog/blogger/2012/10/26/classification-of-hand-written-digits-3/

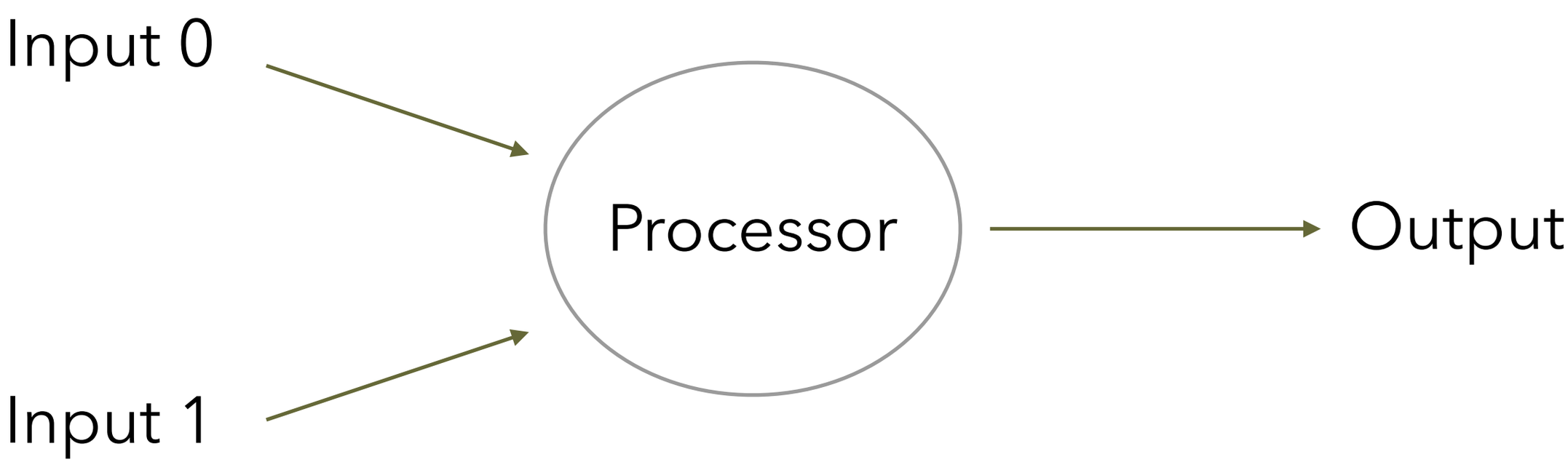
Supervised Learning: Neural Network



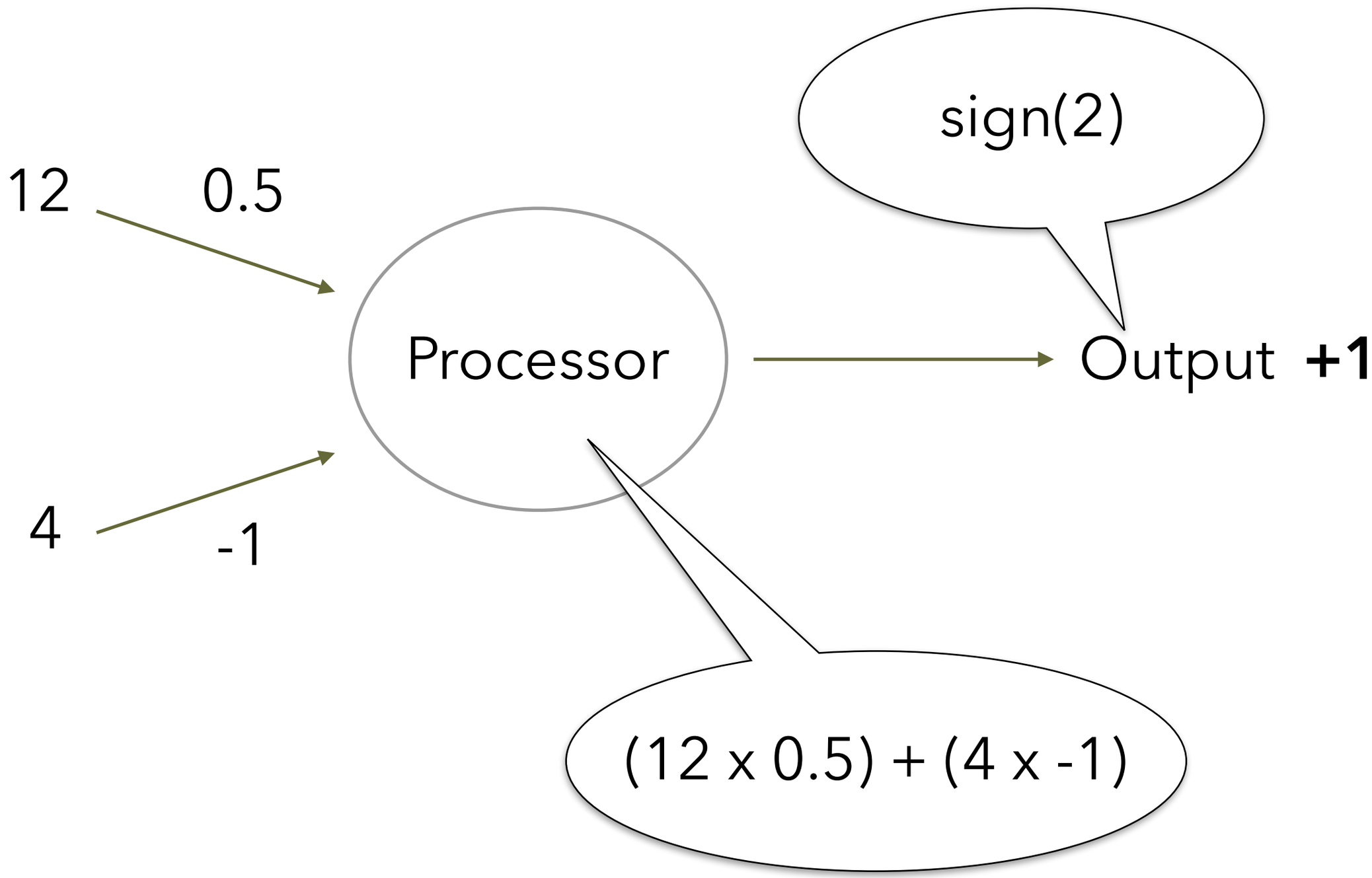
http://i-systems.github.io/HSE545/machine%20learning%20all/16%20Deep%20learning/CNN.html

Supervised Learning: Perceptron

One or more inputs, a processor, and a single output



Supervised Learning: Perceptron



Supervised Learning: Perceptron

Data:

Input 0	Input 1	Label
12	4	-1
5	9	+1

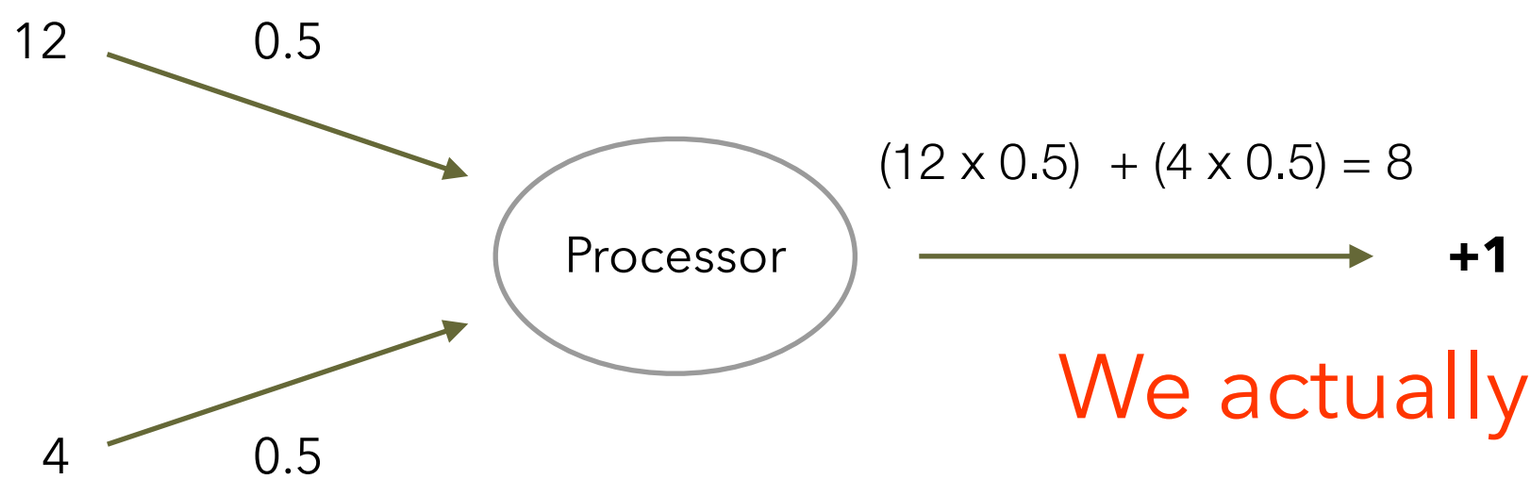
Initial Weight:

Weight 0	Weight 1
0.5	0.5

We define:

Desired	Guess	Error
-1	-1	0
-1	+1	-2
+1	-1	2
+1	+1	0

Supervised Learning: Perceptron (1st Iteration)



We actually need -1!

Input 0	Input 1	Label
12	4	-1
5	9	+1

ΔWeight = Error x Input

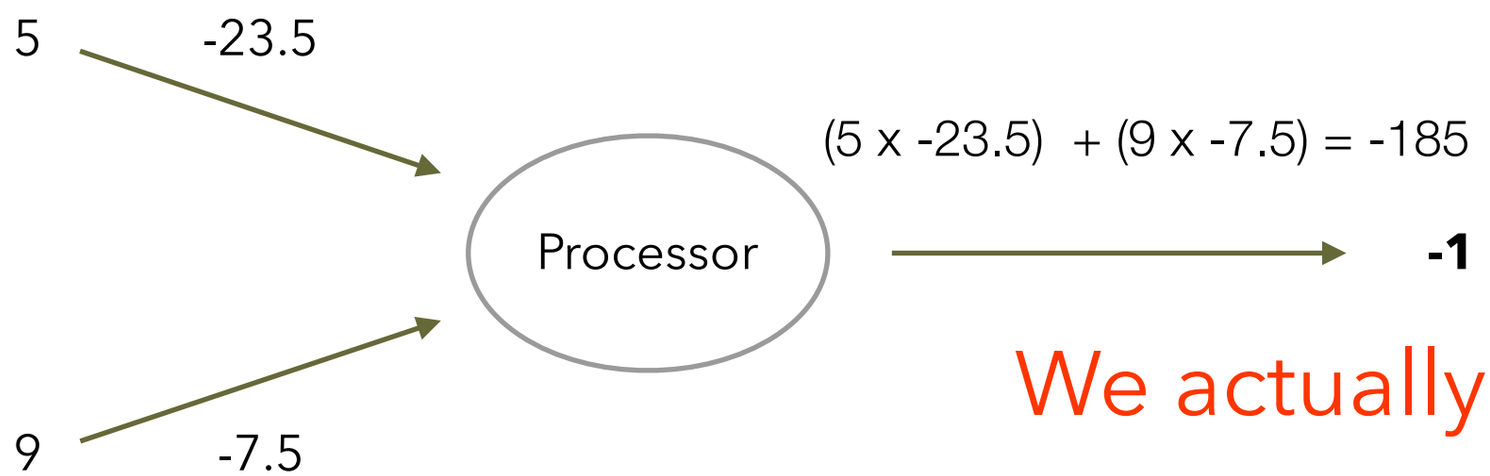
New Weight = Weight + ΔWeight

We then get:

New Weight 0 = 0.5 + (-2 x 12) = -23.5

New Weight 1 = 0.5 + (-2 x 4) = -7.5

Supervised Learning: Perceptron (2nd Iteration)



We actually need +1!

Input 0	Input 1	Label
12	4	-1
5	9	+1

Δ Weight = Error x Input

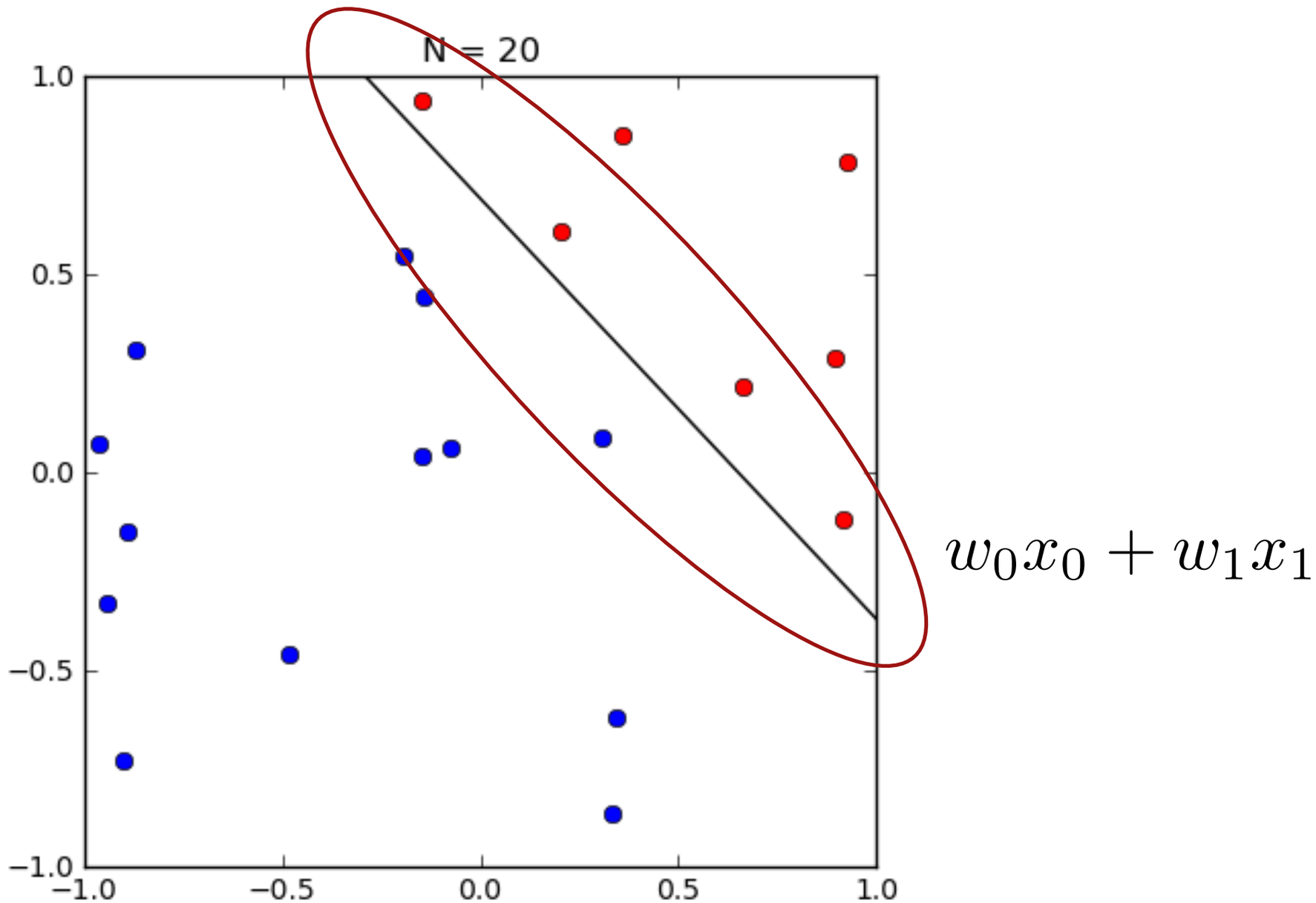
New Weight = Weight + ΔWeight

We then get:

New Weight 0 = -23.5 + (2 x 5) = -13.5

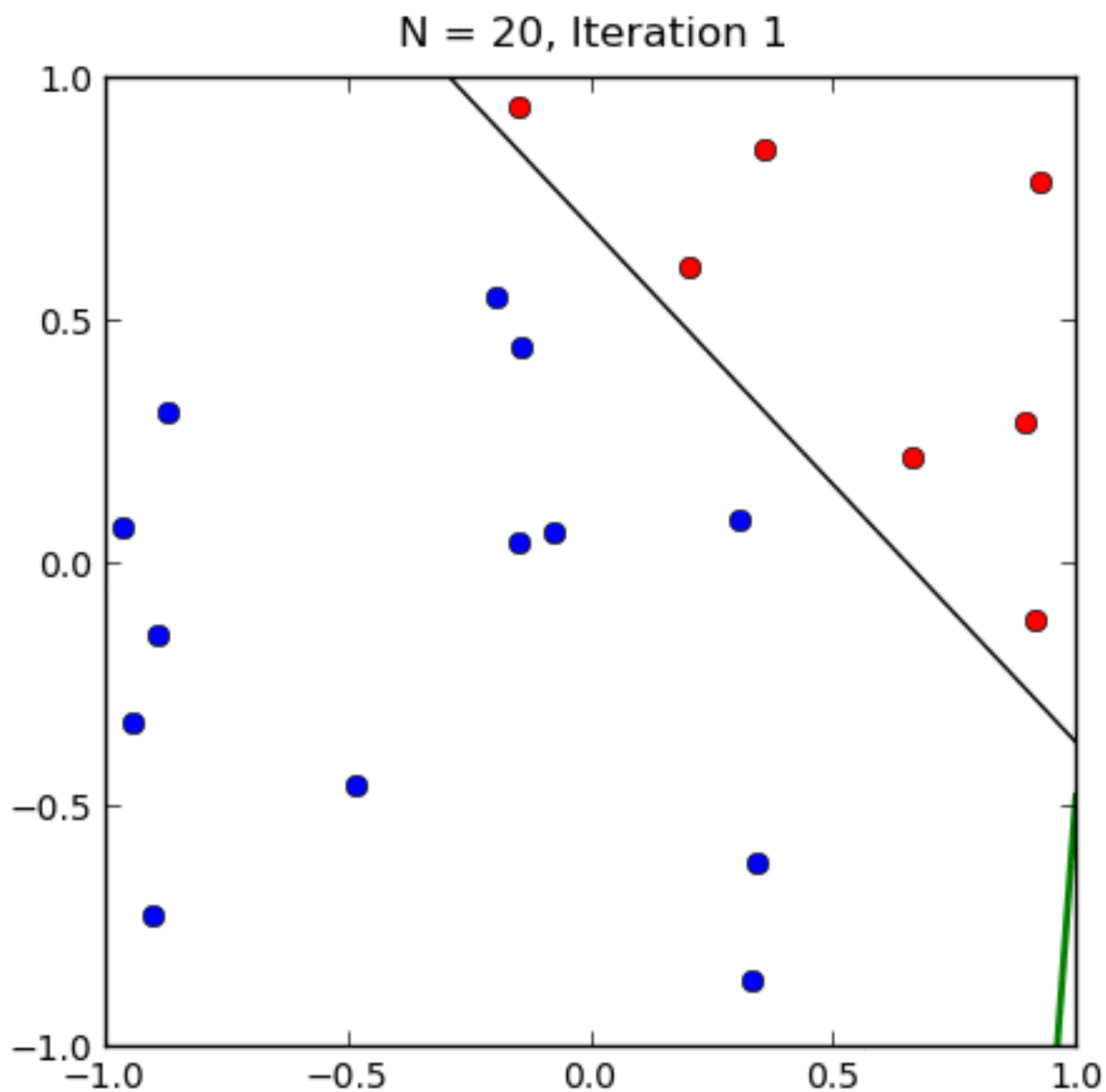
New Weight 1 = -7.5 + (2 x 9) = 10.5

Supervised Learning: Perceptron



https://datasciencelab.wordpress.com/2014/01/10/machine-learning-classics-the-perceptron/

Supervised Learning: Perceptron (Animation)

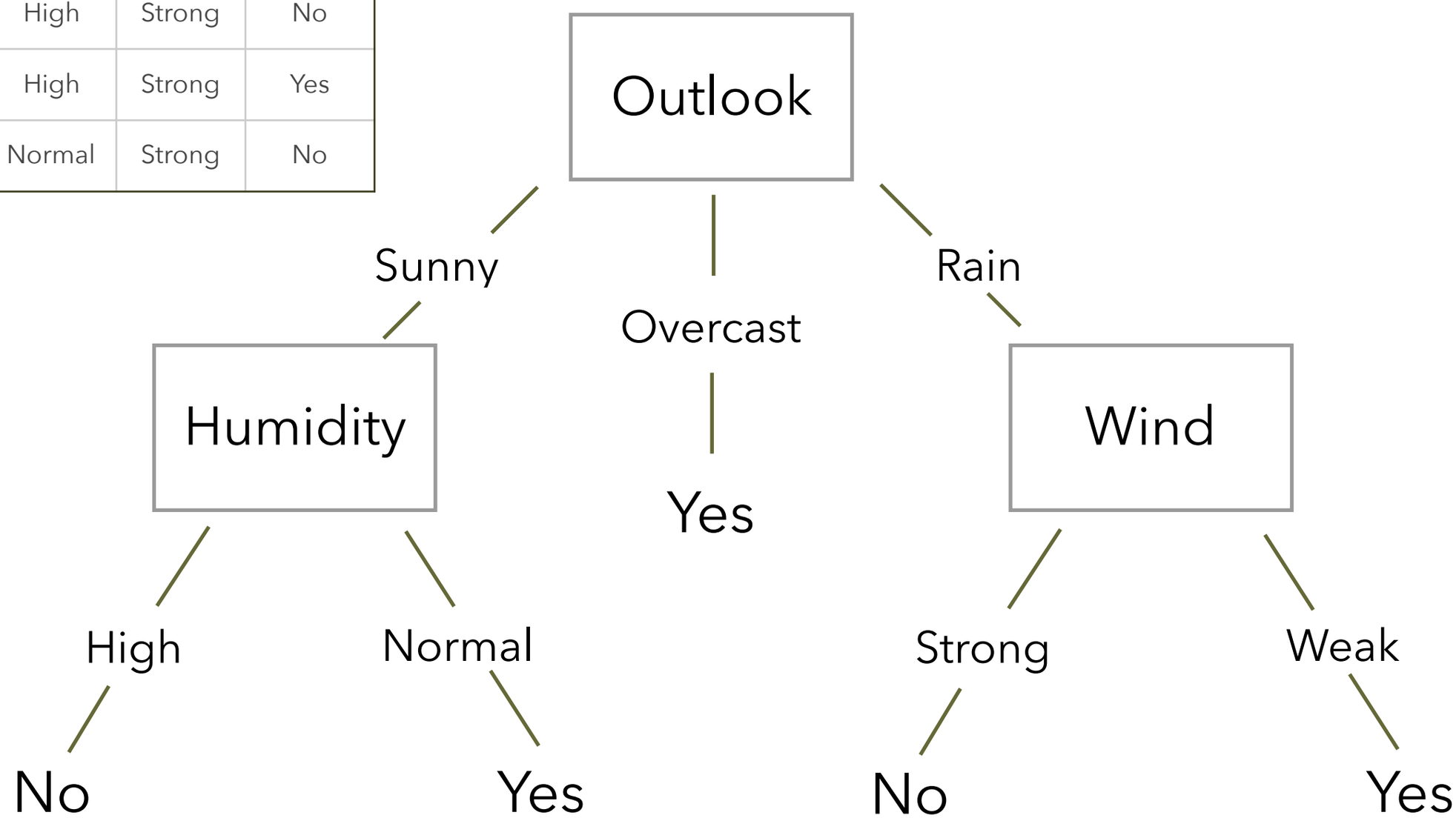


https://datasciencelab.wordpress.com/2014/01/10/machine-learning-classics-the-perceptron/

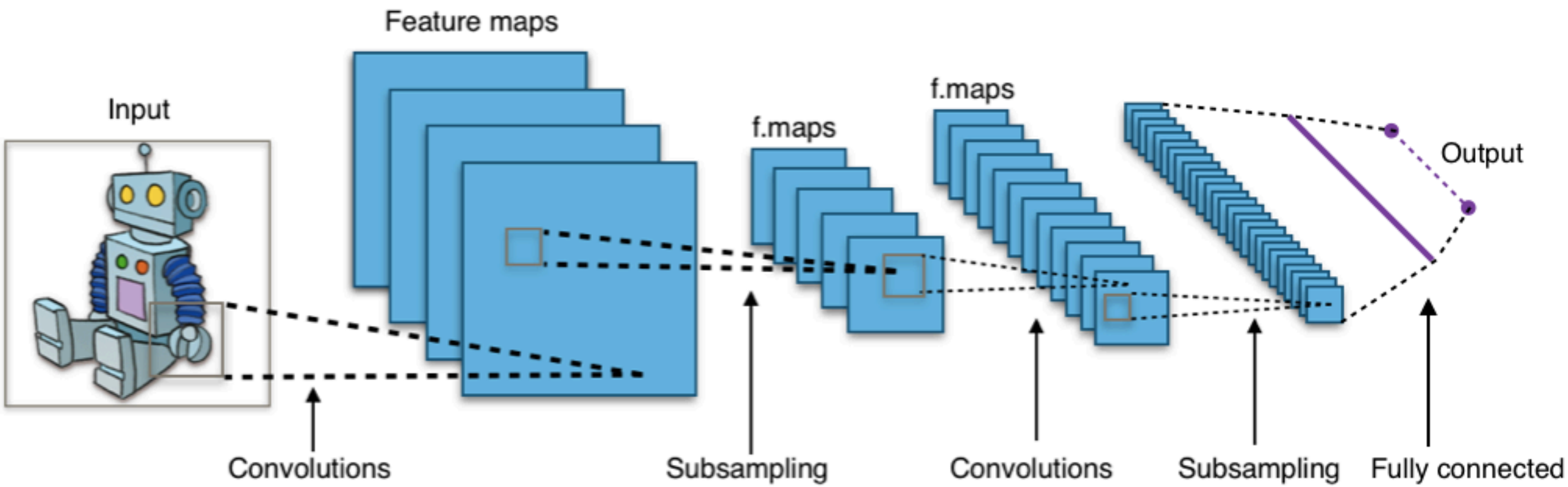
Supervised Learning: Decision Tree

Day	Outlook	Temp	Humidity	WInd	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Mild	High	Strong	Yes
D4	Rain	Cool	Normal	Strong	No

Play tennis?



Supervised Learning: Deep Neural Networks



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

Unsupervised Learning

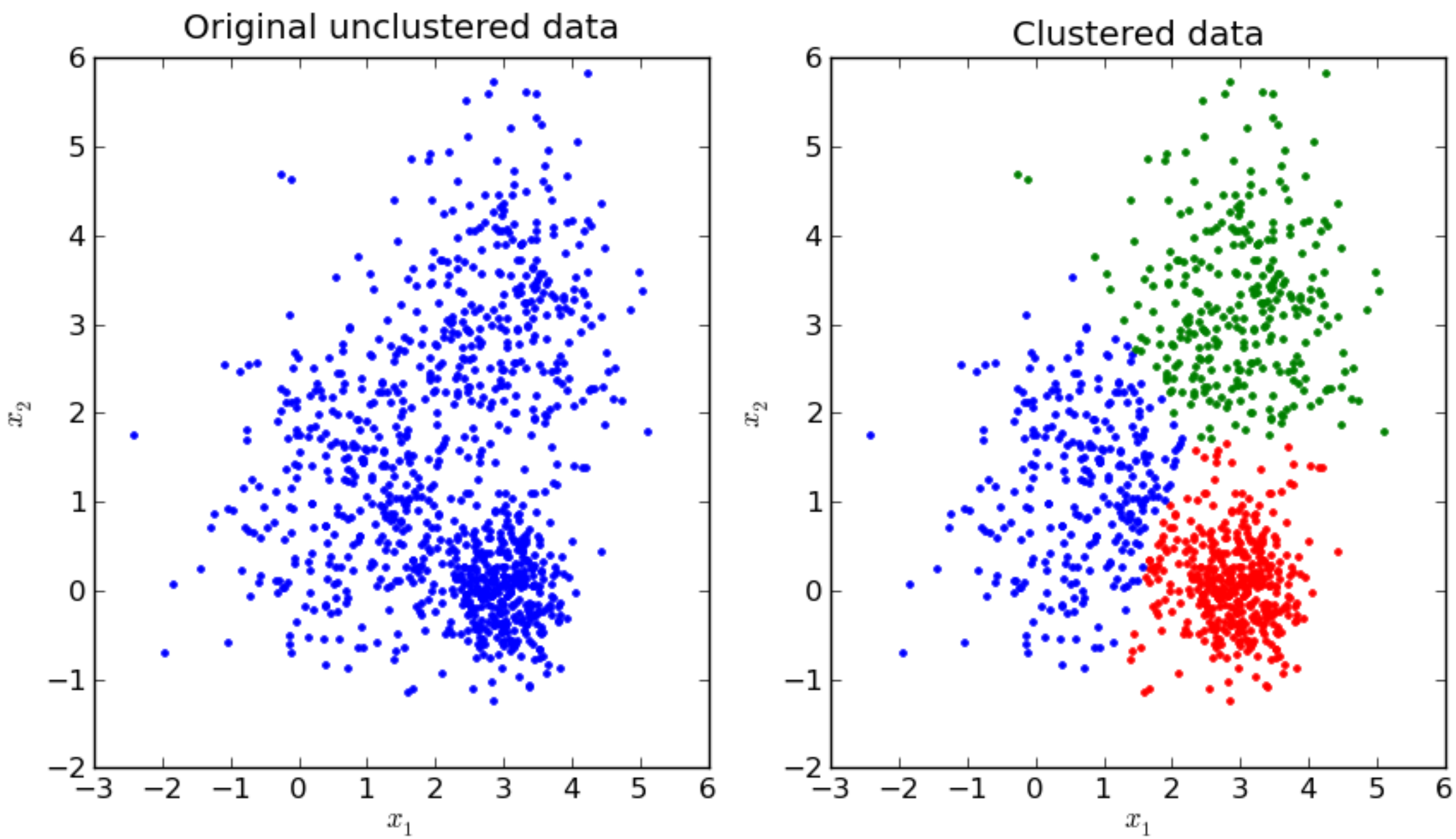
$$f(x)$$

Given *x*, find a function f that gives a compact description of *x*.

Unsupervised Learning: Algorithms

- Clustering
  - *k*-Means Clustering
  - Hierarchical Clustering
- Recommendation
- etc.

Unsupervised Learning: *k*-Means

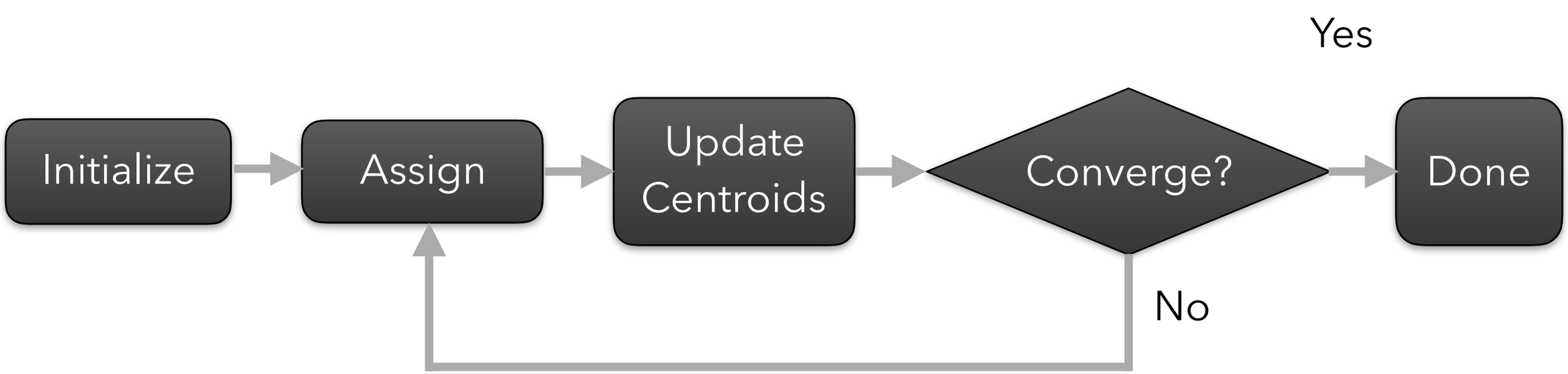


http://stackoverflow.com/questions/24645068/k-means-clustering-major-understanding-issue/24645894#24645894

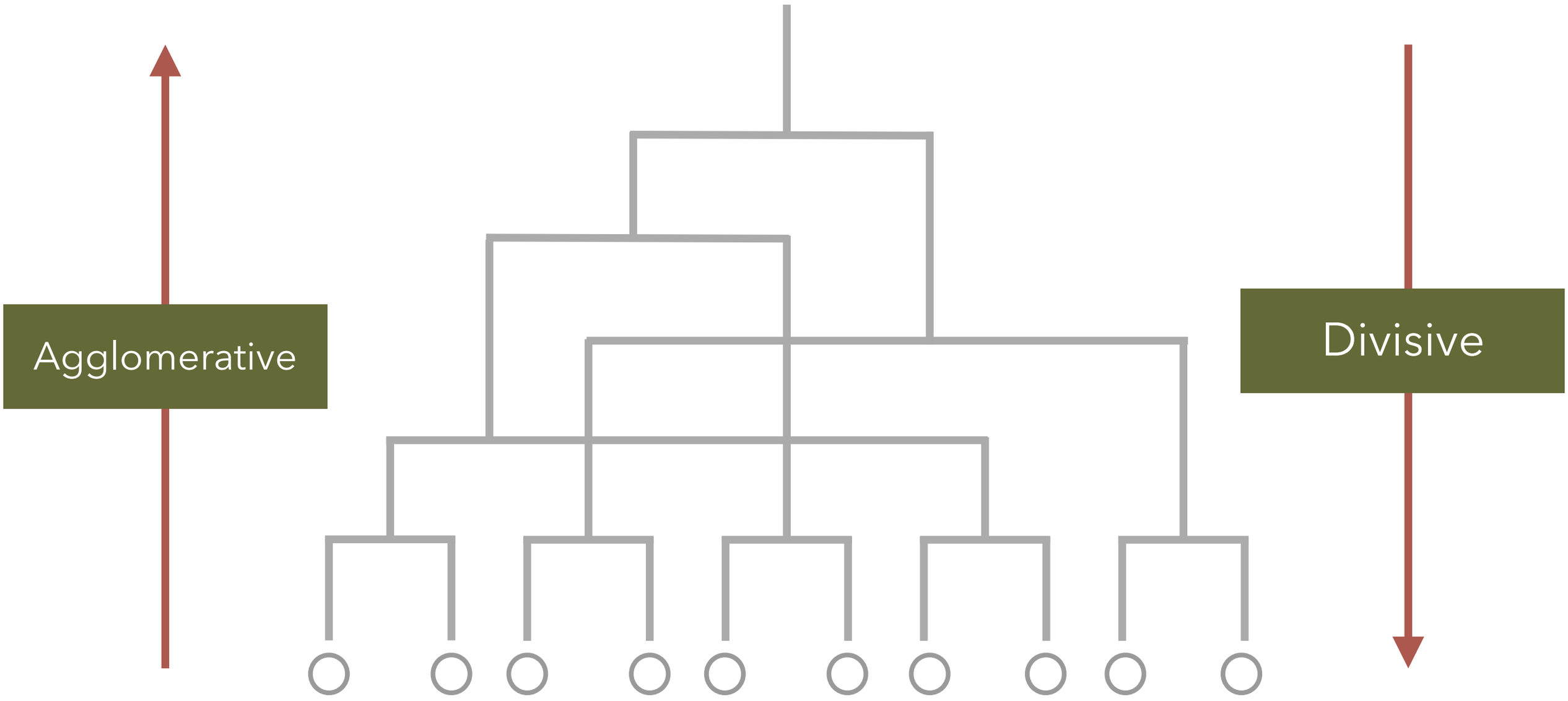
Let's really try *k*-Means!

Exercise: 1D *k*-Means Clustering

- Given these items: {2, 4, 10, 12, 3, 20, 30, 11, 25}
- Given these initial centroids: $m_1 = 2$ and $m_2 = 4$
- Find me the final clusters!



Unsupervised Learning: Hierarchical Clustering



Unsupervised Learning: Recommendation

amazon.com®

NETFLIX

Building Recommender Systems

where machine learning and search synergize!

http://www.kdnuggets.com/2015/12/more-data-science-humor-cartoons.html

Information Overload

http://edition.cnn.com/2010/TECH/02/03/content.overload/

"The Web, they say, is leaving the era of search and entering one of discovery. What's the difference? **Search** is what you do when you're looking for something. **Discovery** is when something wonderful that you didn't know existed, or didn't know how to ask for, finds you."

–CNN Money, "The race to create a 'smart' Google"

Recommender Problem

$C := \{\text{users}\}$

$S := \{\text{recommendable items}\}$

$u :=$ utility function measuring the usefulness of item s to user c

$u : C \times S \rightarrow R$

where $R := \{\text{recommended items}\}$

For each user c choose the items s that maximize u

$$\forall c \in C, s'_c = \text{argmax}_u(u(c, s)), s \in S$$

We predict items (or ratings for items)
that the user may be interested in.

Simplest Recommender System

Just return items **not** seen or reviewed by a given user.. ;)

01-naive-engine

Can we make it better?

How?

- This is where machine learning starts to meet search.
- One approach is that we group users with similar preferences (Clustering).
- We can group similar items too, but we need to have knowledge of those items.
- Items reviewed by users within a cluster tend to score the same way a given user does.

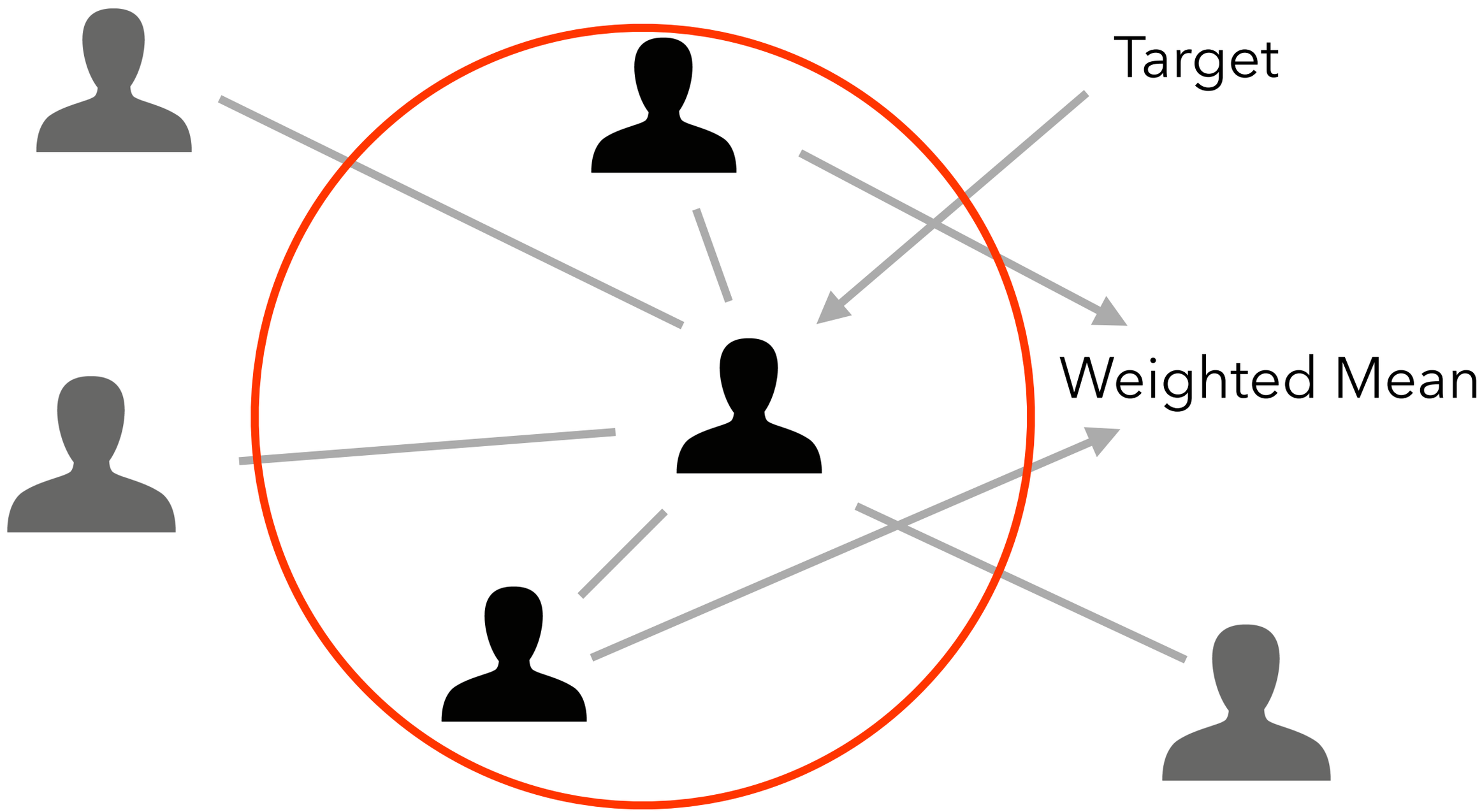
02-naive-engine-using-clustering

Is that it?

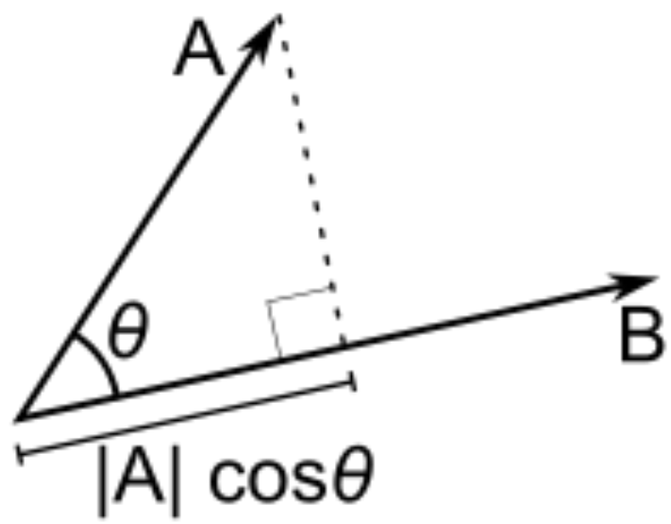
Approaches to Recommendation

- (User-Based | Item-Based) Collaborative Filtering
- Content-Based Recommendation
- Latent Factor Models: Matrix Factorization
- etc.

User-Based Collaborative Filtering



Cosine Similarity Measure



$$\cos(\vec{A}, \vec{B}) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

http://www.gettingcirrius.com/2010/12/calculating-similarity-part-1-cosine.html

Should I recommend "The Last Which Hunter" to Roofimon? (User-Based)

	The Hunger Game	Warcraft The Beginning	The Good Dinosaur	The Last Witch Hunter
Kan	5	4	1	3
Roofimon	5	4	3	?
Juacompe	1		3	3
John	4		1	

Suppose

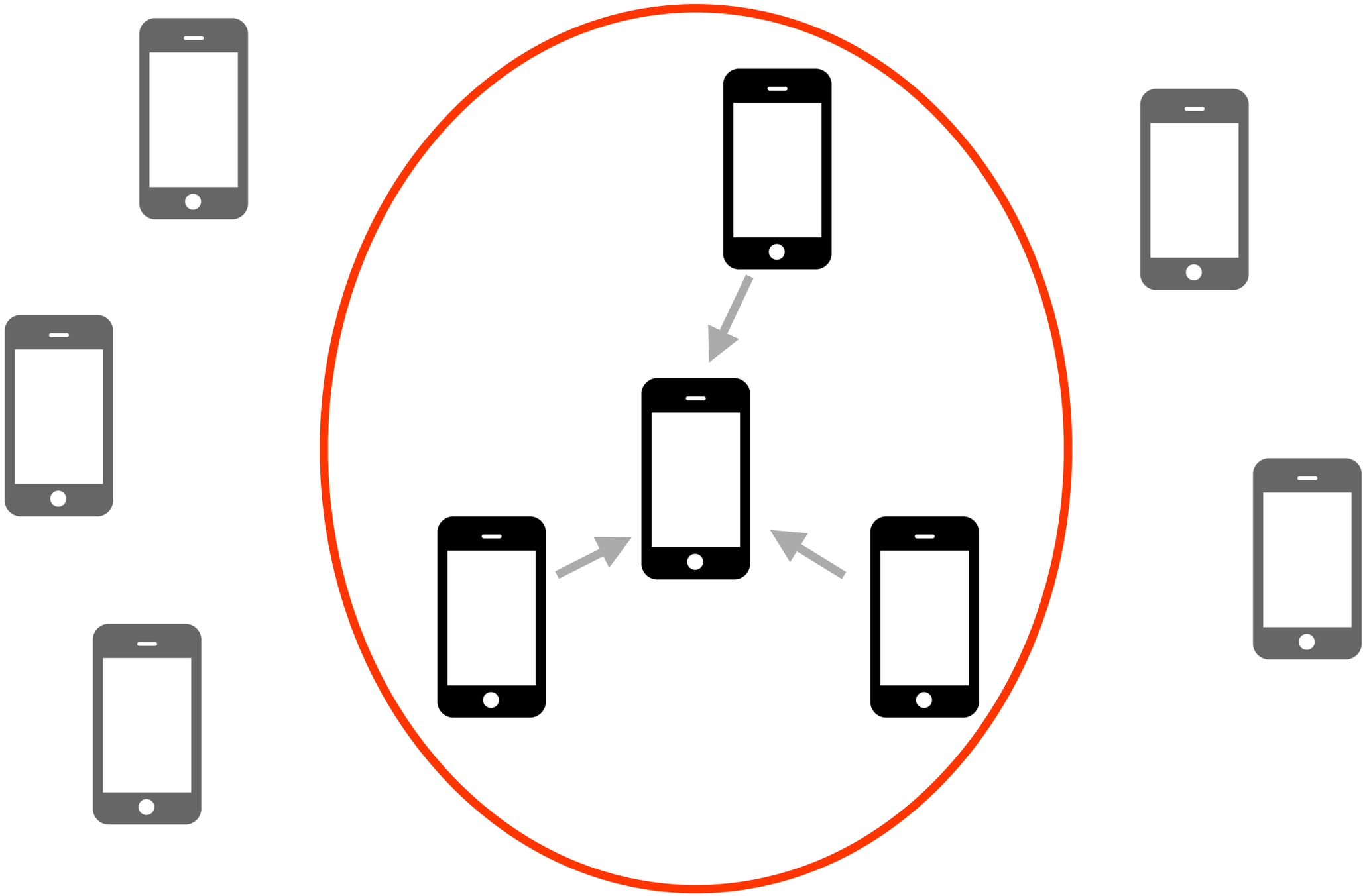
$$\mathrm{sim}(R, K) = 0.8$$

$$\mathrm{sim}(R, J) = 0.5$$

$$\frac{(0.8 \times 3) + (0.5 \times 3)}{0.8 + 0.5} = 3$$

03-user-based-collaborative-filtering

Item-Based Collaborative Filtering



Should I recommend "The Last Which Hunter" to Roofimon? (Item-Based)

	The Hunger Game	Warcraft The Beginning	The Good Dinosaur	The Last Witch Hunter
Kan	5	4	1	3
Roofimon	5	4	3	?
Juacompe	1		3	3
John	4		1	

Suppose

$$\mathrm{sim}(TLWH, THG) = 0.6$$
$$\mathrm{sim}(TLWH, WTB) = 0.7$$
$$\mathrm{sim}(TLWH, TGD) = 0.8$$

$$\frac{(0.6 \times 5) + (0.7 \times 4) + (0.8 \times 3)}{0.6 + 0.7 + 0.8} = 3.905$$

04-item-based-collaborative-filtering

Content-Based Recommendation

- We recommend items based on information on the content of items rather than on other users' interactions.
- Recommend items similar to those a given user likes.
- Content can be title, description, genre, or year of movies.
- Common approach is text-based.
  - Term Frequency times Inverse Document Frequency (TF-IDF)

TF-IDF

Given a collection of N documents.

Let f_{ij} = number of times word i appears in document j.

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

Suppose word i appears in n_i of the N documents. IDF for a term (word) defined as follows.

$$IDF_i = \log(\frac{N}{n_i})$$

TD-IDF can be calculated as follows.

$$TF \cdot IDF_{ij} = TF_{ij} \times IDF_i$$

TF-IDF Example

	angeles	los	new	post	times	york
d1	0	0	1	0	2	1
d2	0	0	1	1	0	1
d3	2	1	0	0	1	0

IDF values for the terms:

angeles $\log(\frac{3}{1}) = 0.477$

los $\log(\frac{3}{1}) = 0.477$

new $\log(\frac{3}{2}) = 0.176$

post $\log(\frac{3}{1}) = 0.477$

times $\log(\frac{3}{2}) = 0.176$

york $\log(\frac{3}{2}) = 0.176$

TF-IDF Example (cont.)

	angeles	los	new	post	times	york
d1	0	0	(1/2)*0.176	0	(2/2)*0.176	(1/2)*0.176
d2	0	0	(1/1)*0.176	(1/1)*0.176	0	(1/1)*0.176
d3	(2/2)*0.477	(1/2)*0.477	0	0	(1/2)*0.176	0

TF-IDF Example (cont.)

	angeles	los	new	post	times	york
d1	0	0	0.088	0	0.176	0.088
d2	0	0	0.176	0.176	0	0.176
d3	0.477	0.2385	0	0	0.088	0

Cosine Similarity Measure (cont.)

	angeles	los	new	post	times	york
d1	0	0	0.088	0	0.176	0.088
d2	0	0	0.176	0.176	0	0.176
d3	0.477	0.2385	0	0	0.088	0

Query: "new new times"

	angeles	los	new	post	times	york
q	0	0	(2/2)*0.176	0	(1/2)*0.176	0

$$\cos(\vec{d1}, \vec{q}) = \frac{(0.088 \times 0.0176) + (0.176 \times 0.088)}{\sqrt{0.088^2 + 0.176^2 + 0.088^2} \times \sqrt{0.176^2 + 0.088^2}}$$

05-content-based-engine

Latent Factor Models: Matrix Factorization

	The Hunger Game	Warcraft The Beginning	The Good Dinosaur	The Last Witch
Kan	5	4	1	3
Roofimon	5	4	3	-
Juacompe	1	-	3	3
John	4	-	1	-

Ideas behind Matrix Factorization

- There should be some latent (hidden) features that determine how a user rates an item.
- Two users would give high ratings to a movie if they both like its genre.

Ideas behind Matrix Factorization (cont.)

$$A = \begin{array}{c} \begin{array}{cccc} A_1 & & A_3 & \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 8 & 12 \\ 3 & 6 & 7 & 13 \end{bmatrix} \end{array}$$

$$A_1 = 1 \cdot A_1 + 0 \cdot A_3$$

$$A_2 = 2 \cdot A_1 + 0 \cdot A_3$$

$$A_3 = 0 \cdot A_1 + 1 \cdot A_3$$

$$A_4 = 2 \cdot A_1 + 1 \cdot A_3$$

$$X = \begin{bmatrix} 1 & 3 \\ 2 & 8 \\ 3 & 7 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$XY = \begin{bmatrix} 1 & 2 & 3 & 5 \\ 2 & 4 & 8 & 12 \\ 3 & 6 & 7 & 13 \end{bmatrix}$$

Should I recommend "The Last Which Hunter" to Roofimon? (Matrix Factorization)

	The Hunger Game	Warcraft The Beginning	The Good Dinosaur	The Last Witch Hunter
Roofimon	5	4	3	?

User	Scary	Kiddy
Roofimon	2	5

Movie	Scary	Kiddy
TLWH	3/4	1/4

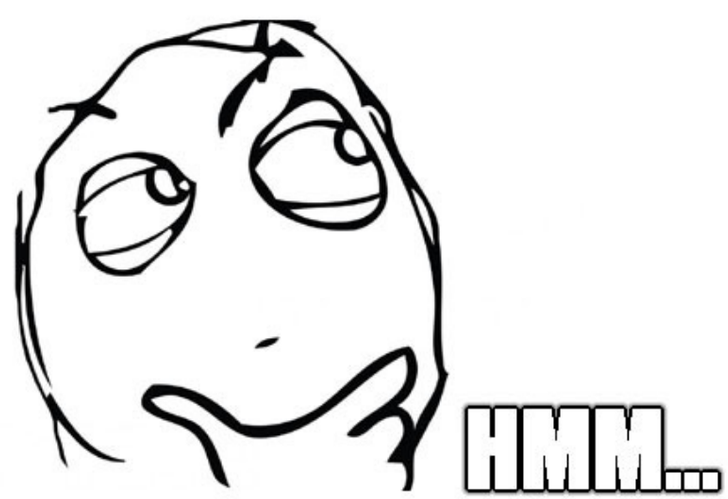
$$(2 \times 3/4) + (5 \times 1/4) = 2.75$$

06-matrix-factorization

Recap: Recommendation

1. Naive engine
2. User-based collaborative filtering
3. Item-based collaborative filtering
4. Content-based engine
5. Matrix factorization

Supervised vs. Unsupervised?



Reinforcement Learning

$$y = f(x)$$

$$z$$

Given *x* and *z*, find a function f that generates *y*.

Policy

State

Action

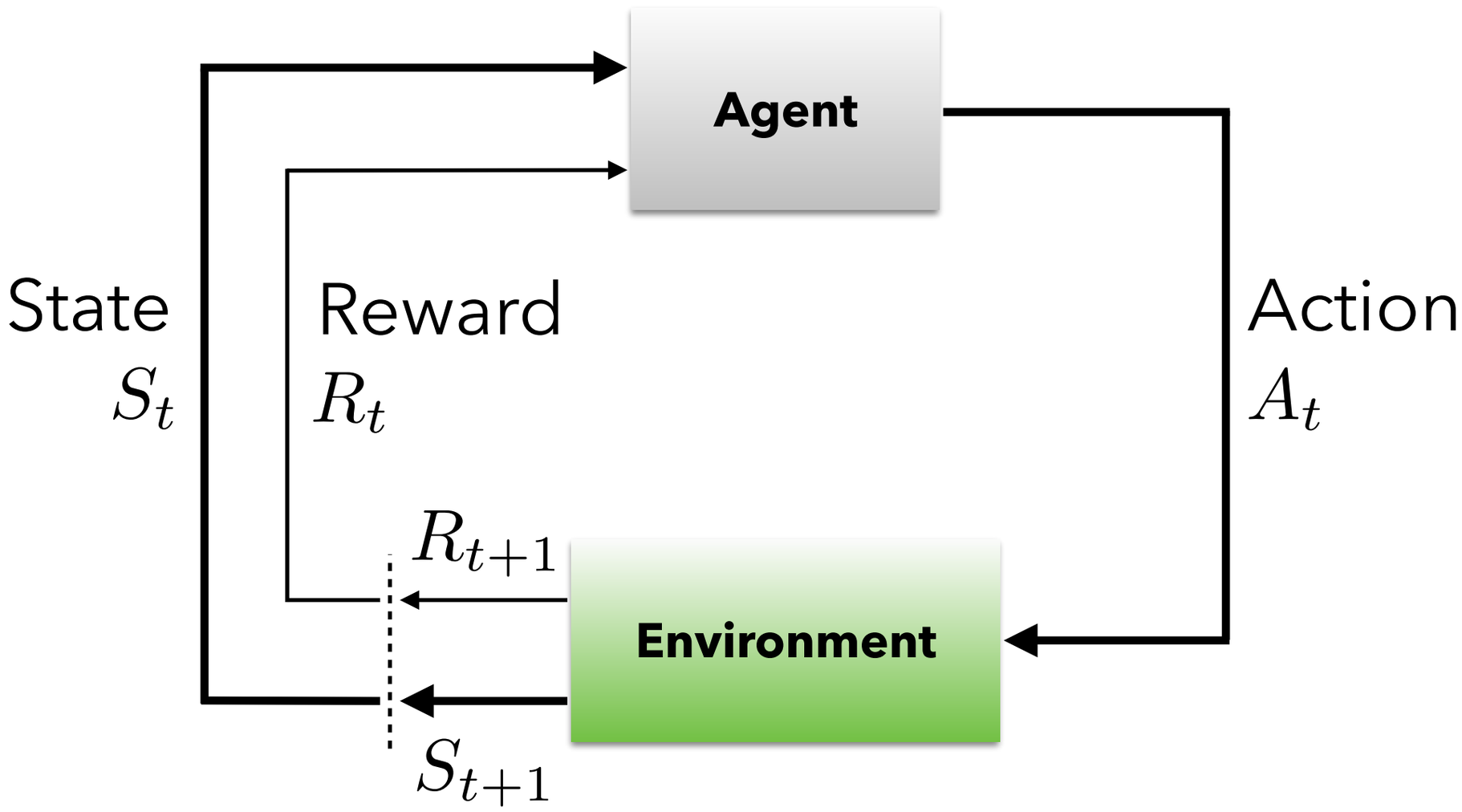
$$y = f(x)$$

$$z$$

Reward

Given *x* and *z*, find a function f that generates *y*.

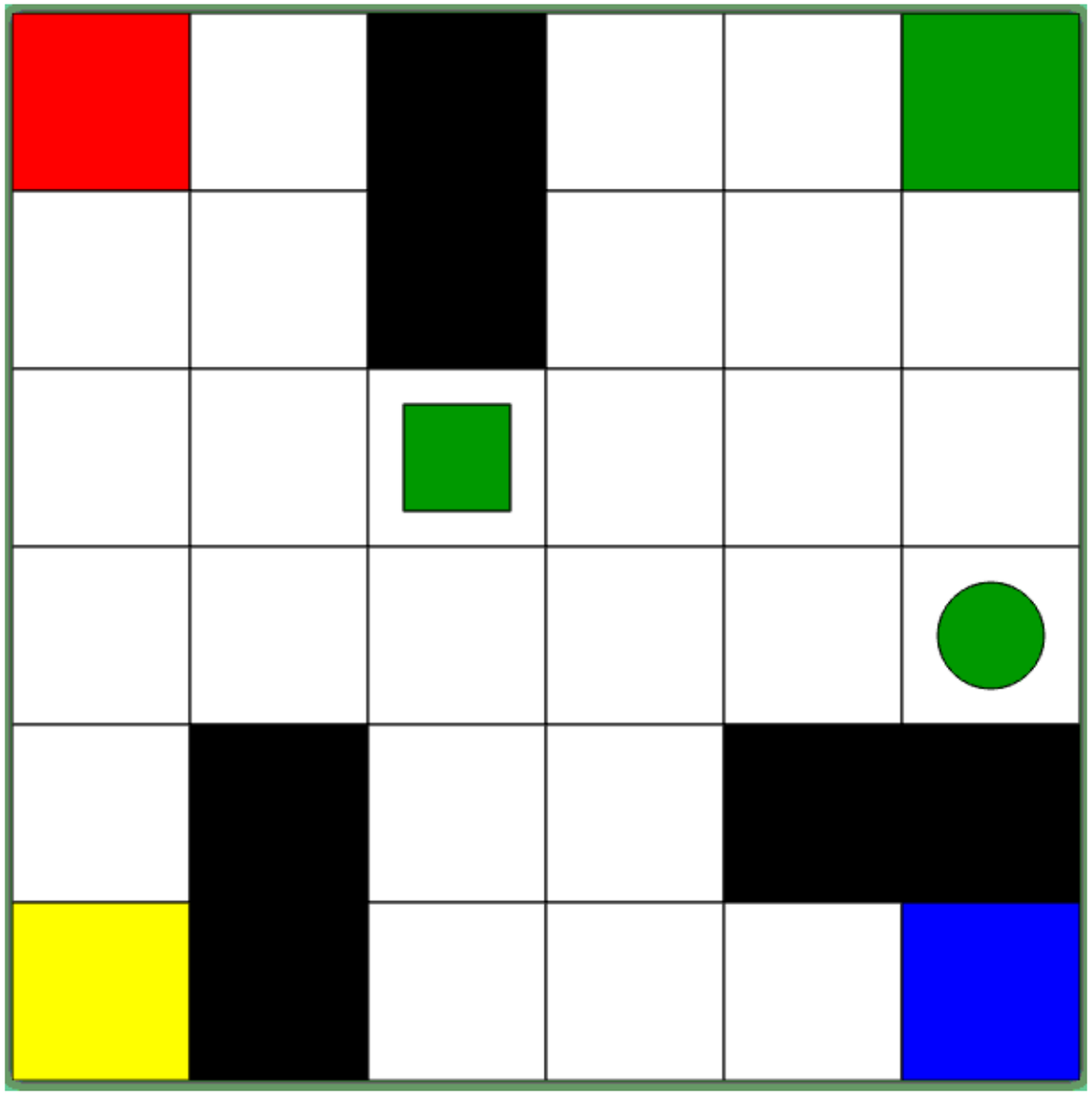
Agent-Environment Interaction



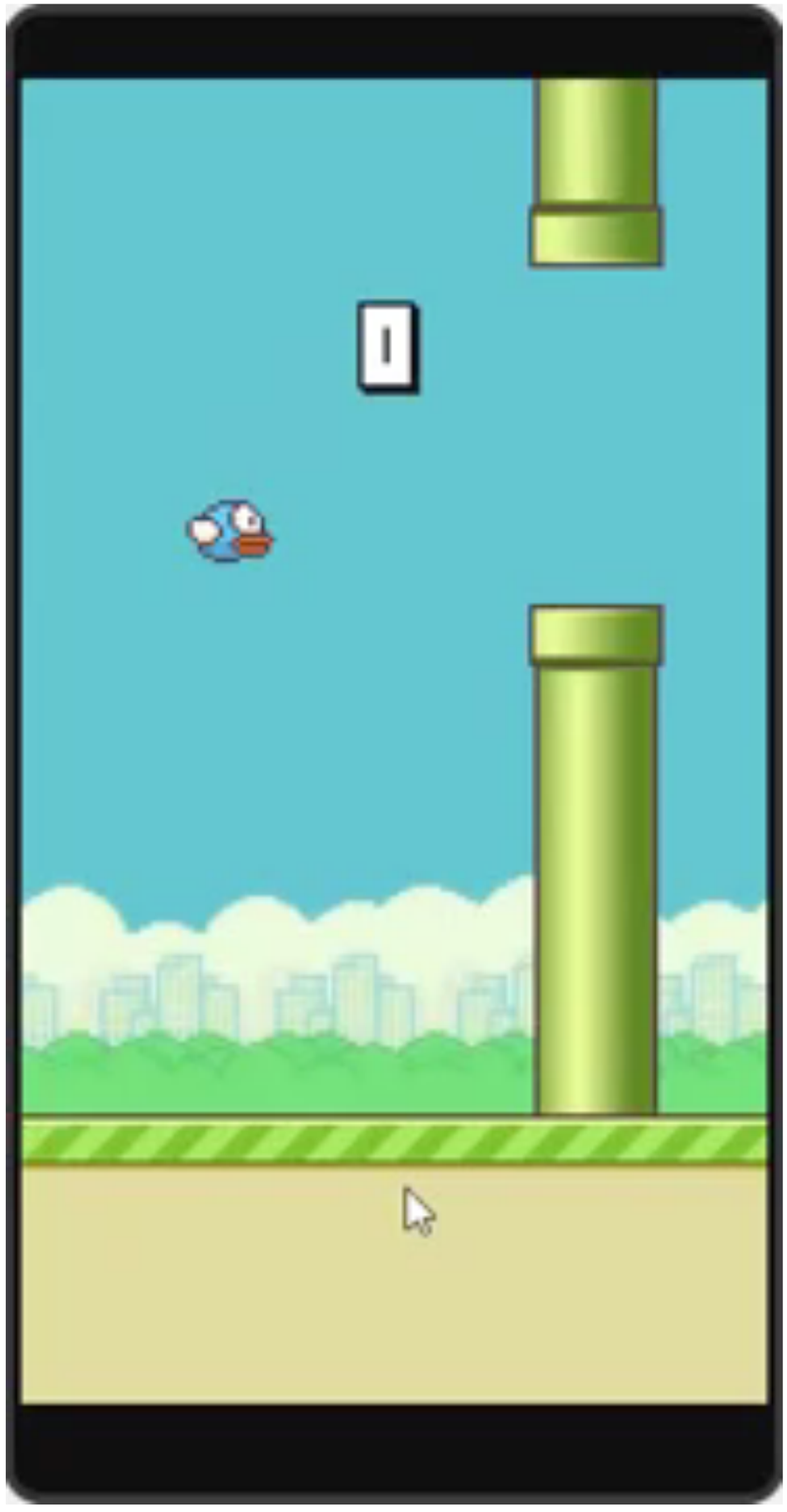
Policy

- The learning agent's way of behaving at a given time
- A mapping from perceived states of the environment to actions to be taken when in those states
- A simple lookup table

Mystery Game



https://github.com/guiferviz/rl_udacity



Flappy Bird Hack using Reinforcement Learning

http://sarvagyavaish.github.io/FlappyBirdRL/

Robot Motor Skill Coordination with EM-based Reinforcement Learning

Petar Kormushev, Sylvain Calinon, and Darwin G. Caldwell

Italian Institute of Technology

Google DeepMind's Deep Q-learning playing Atari Breakout

https://www.youtube.com/watch?v=V1eYniJ0Rnk

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play. Without any lookahead search, the neural networks play Go at the level of state-of-the-art Monte Carlo tree search programs that simulate thousands of random games of self-play. We also introduce a new search algorithm that combines Monte Carlo simulation with value and policy networks. Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs, and defeated the human European Go champion by 5 games to 0. This is the first time that a computer program has defeated a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.

All games of perfect information have an optimal value function, $v^*(s)$, which determines the outcome of the game, from every board position or state s, under perfect play by all players. These games may be solved

policies[13–15] or value functions[16] based on a linear combination of input features.

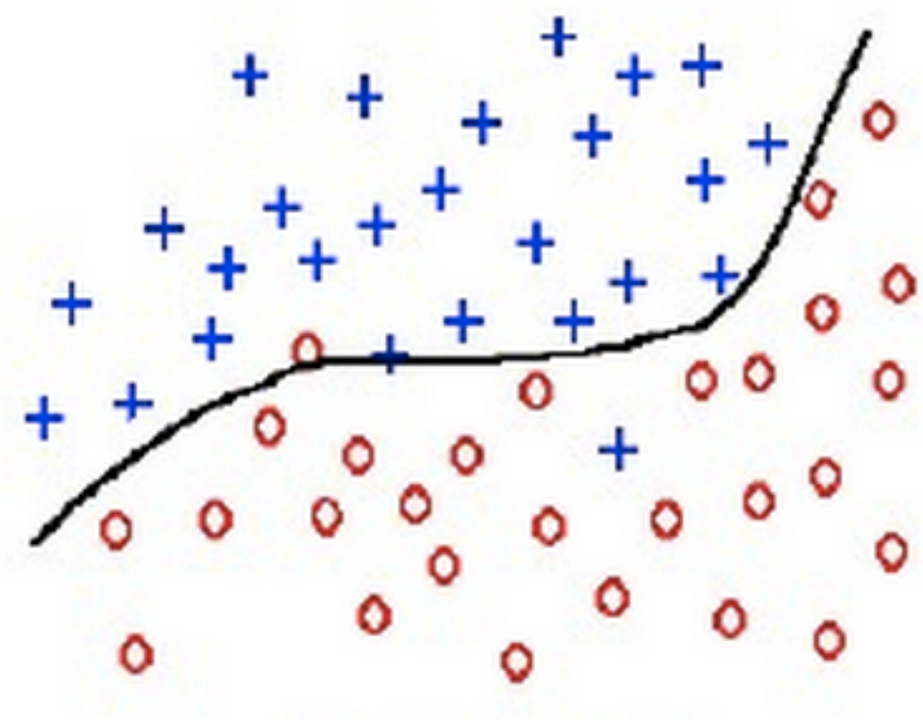
Recently, deep convolutional neural networks have achieved unprec-

How do we **validate**, **select**, and **evaluate** a model?

Model Validation

https://500px.com/photo/65907417/like-a-frog-trapped-inside-a-coconut-shell-by-ellena-susanti

Overfitting (High Variance)

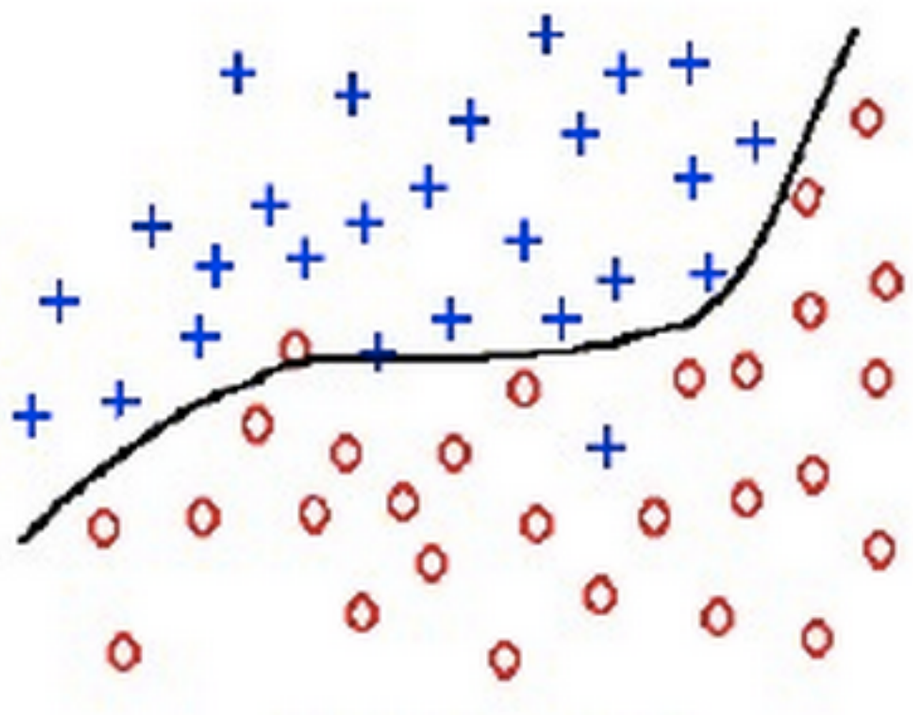


Normal fit

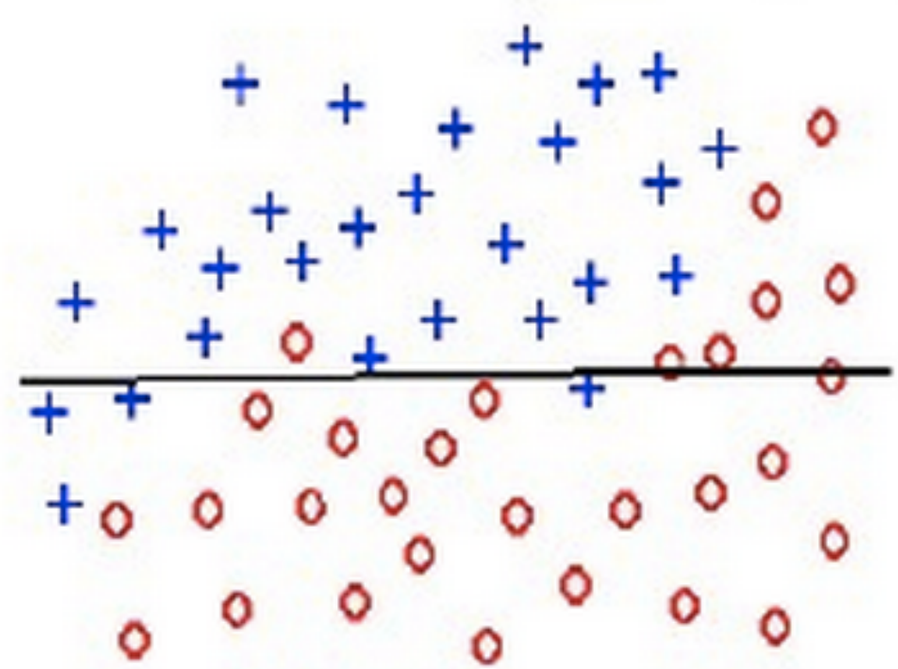
Overfitting

http://blog.csdn.net/love_tea_cat/article/details/25972921

Underfitting (High Bias)



Normal fit



Underfitting

http://blog.csdn.net/love_tea_cat/article/details/25972921

How to Avoid Overfitting and Underfitting

- Adding more data does **NOT** always help!
- Recommend to
  - find a good set of features;
  - run cross validation;
  - use regularization when overfitting is found.

Model Selection

Selecting a Model

- Run **cross validation** to find the best set of parameters for the model.

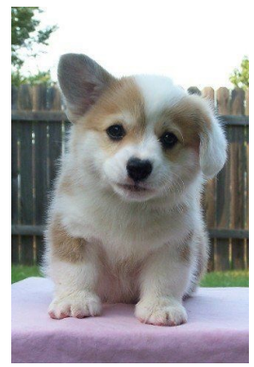
Model Evaluation

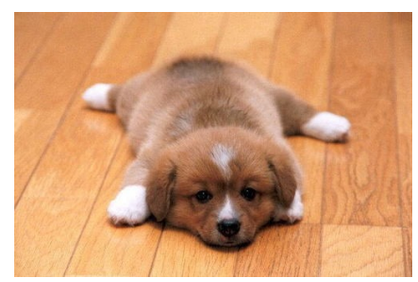
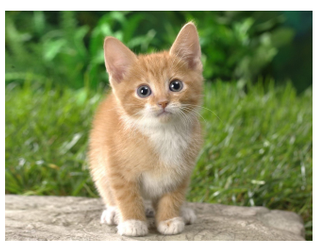
Evaluating a Model with Metrics

Let's just focus on these (for now)!

- Accuracy
- True Positive, False Positive, True Negative, and False Negative
- Precision and Recall
- F1 Score
- etc.

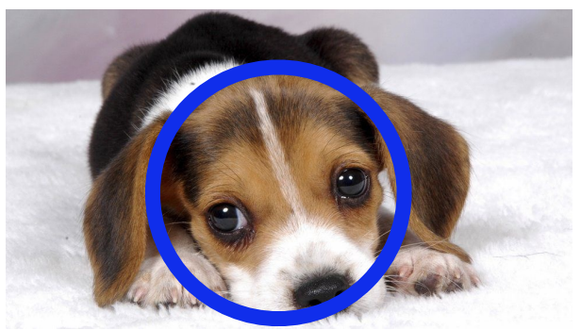
Let’s Try to Evaluate This!

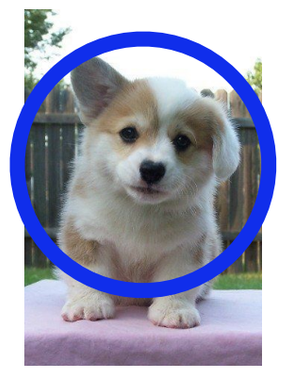


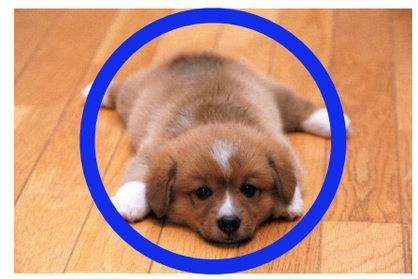



Give me Cats!

4 True Negatives







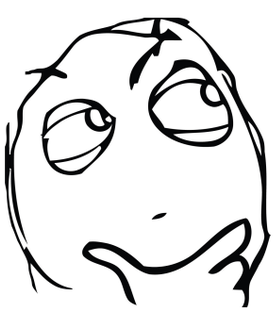
2 False Negatives

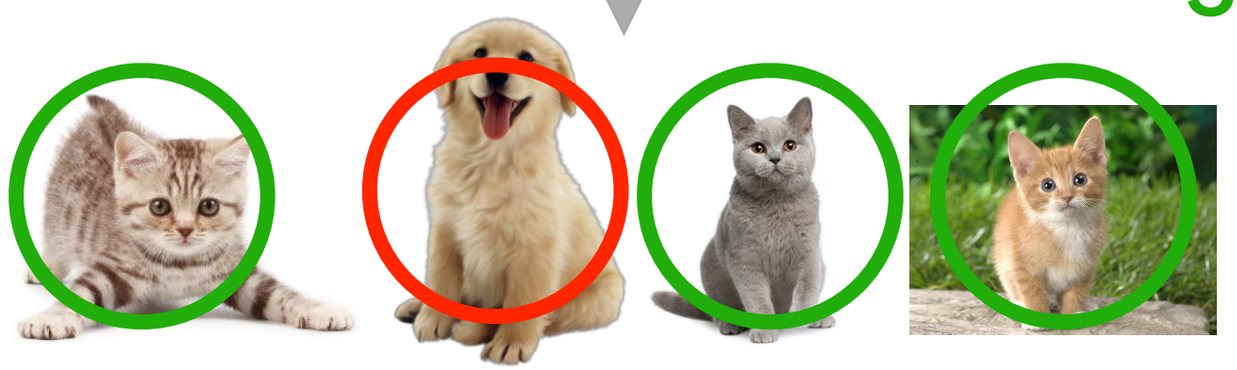
System

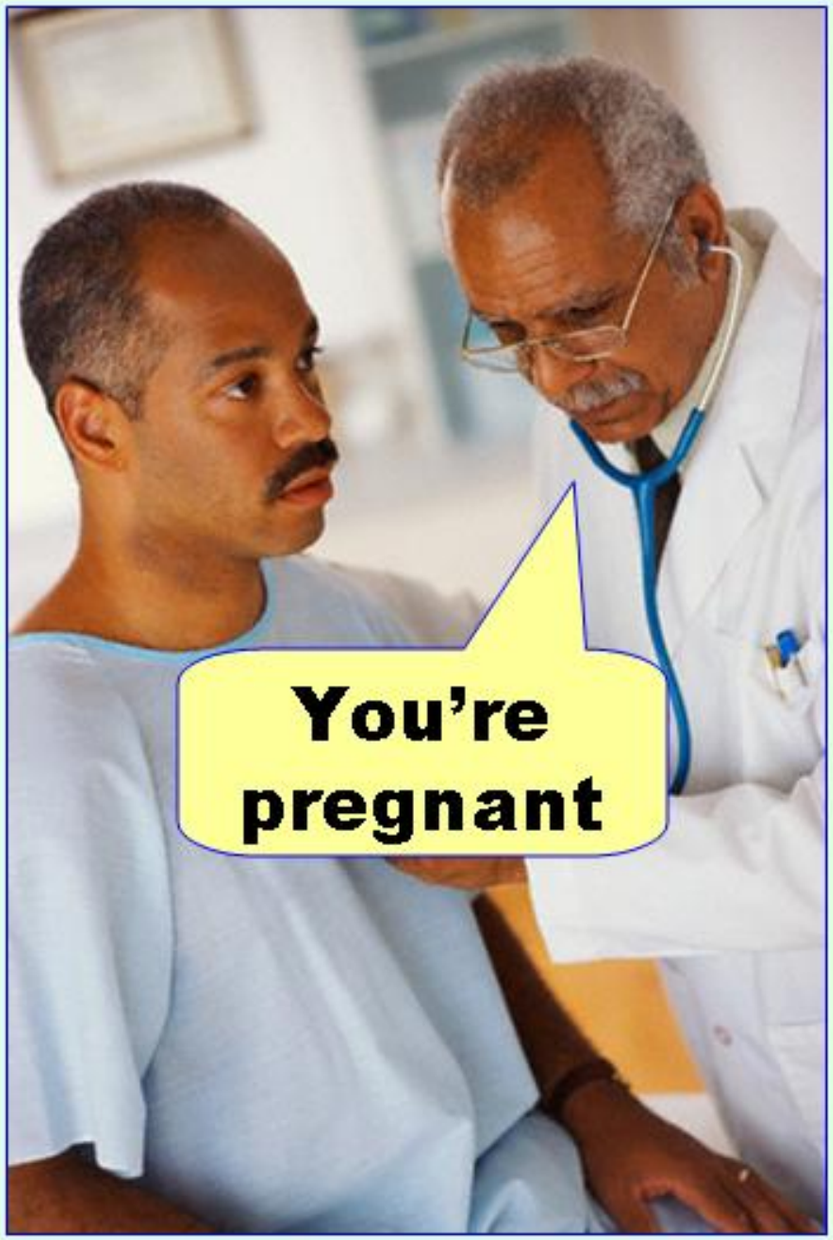
1 False Positive

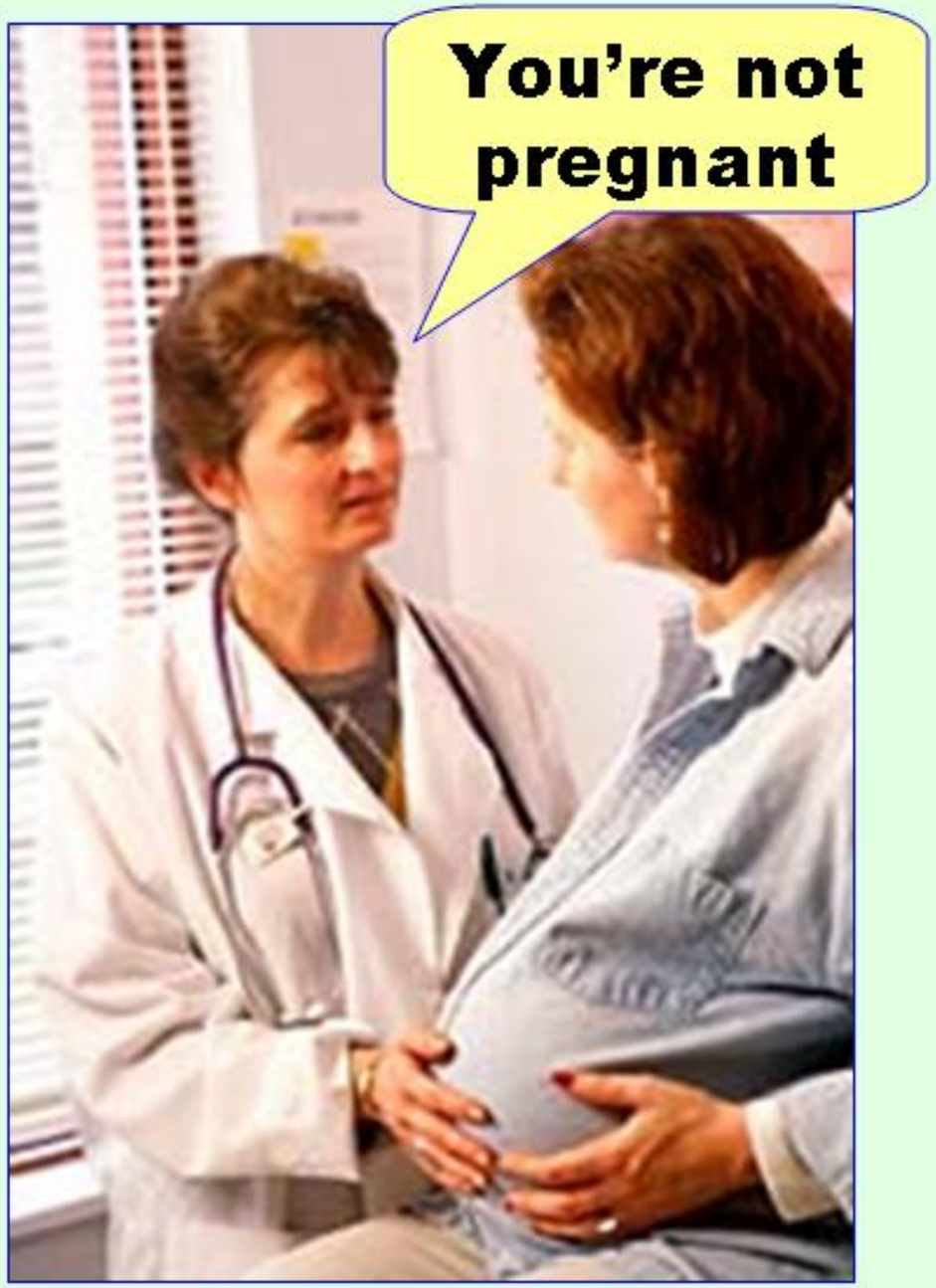
3 True Positives

User



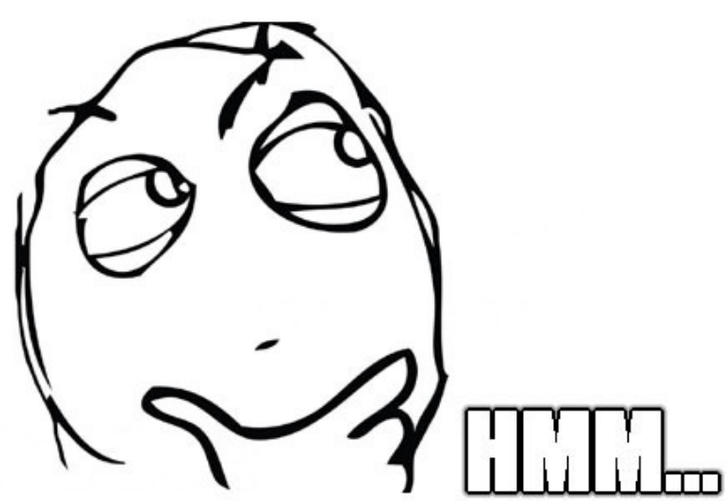






http://www.timvanderzee.com/error-control-p-values-versus-bayes-factors/

False Positive or False Negative?



Metrics Summary

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

accuracy (ACC)

$$ACC = \frac{TP + TN}{P + N}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Informedness = Sensitivity + Specificity - 1

Markedness = Precision + NPV - 1

Sources: Fawcett (2006) and Powers (2011).[44][45]

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

specificity (SPC) or true negative rate (TNR)

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

precision or positive predictive value (PPV)

$$PPV = \frac{TP}{TP + FP}$$

negative predictive value (NPV)

$$NPV = \frac{TN}{TN + FN}$$

fall-out or false positive rate (FPR)

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - SPC$$

false discovery rate (FDR)

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

miss rate or false negative rate (FNR)

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP}$$

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

Course Outline

1. ✅ Data Science and Tools
2. Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

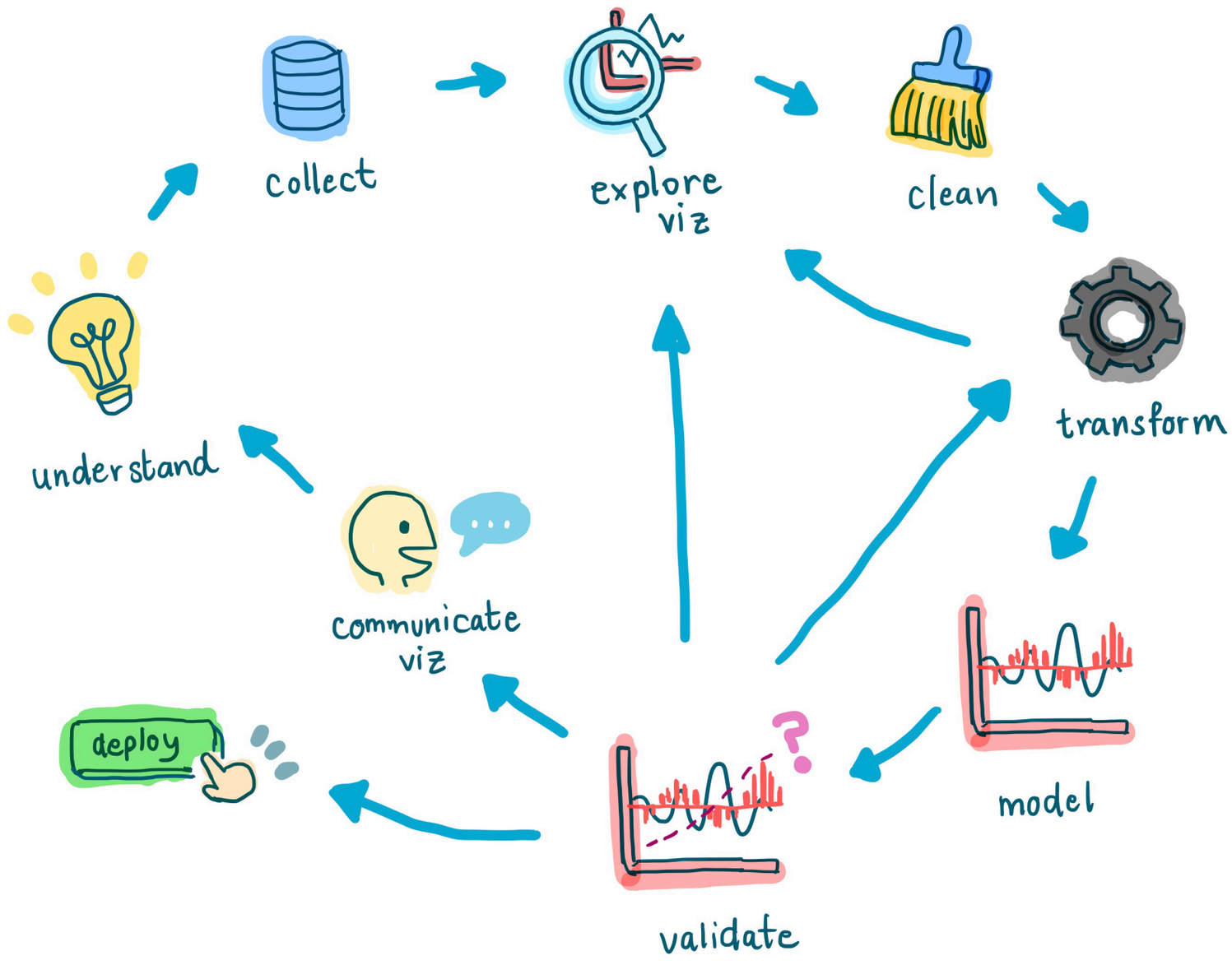
Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

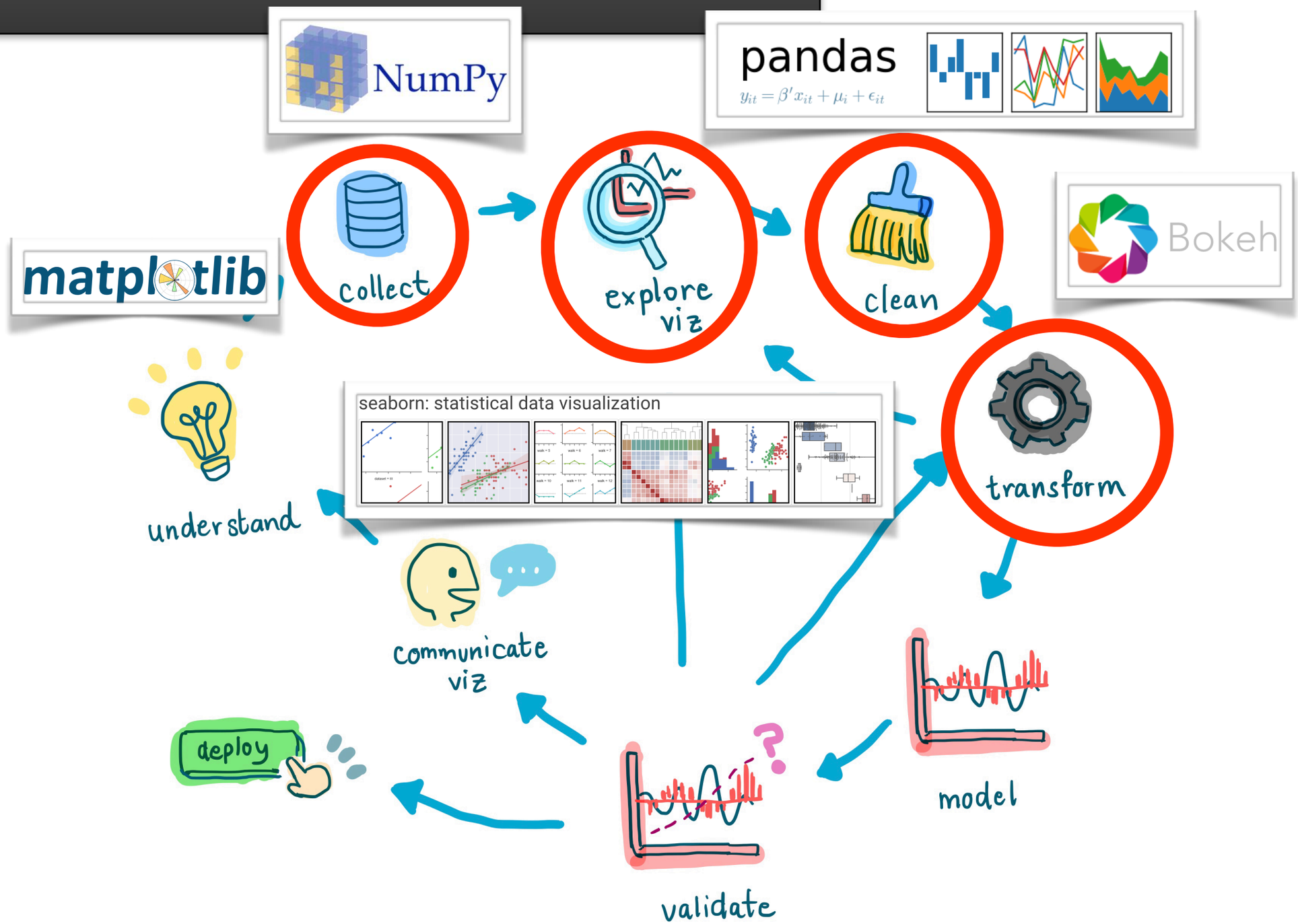
https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

How do we apply the **ML process**?

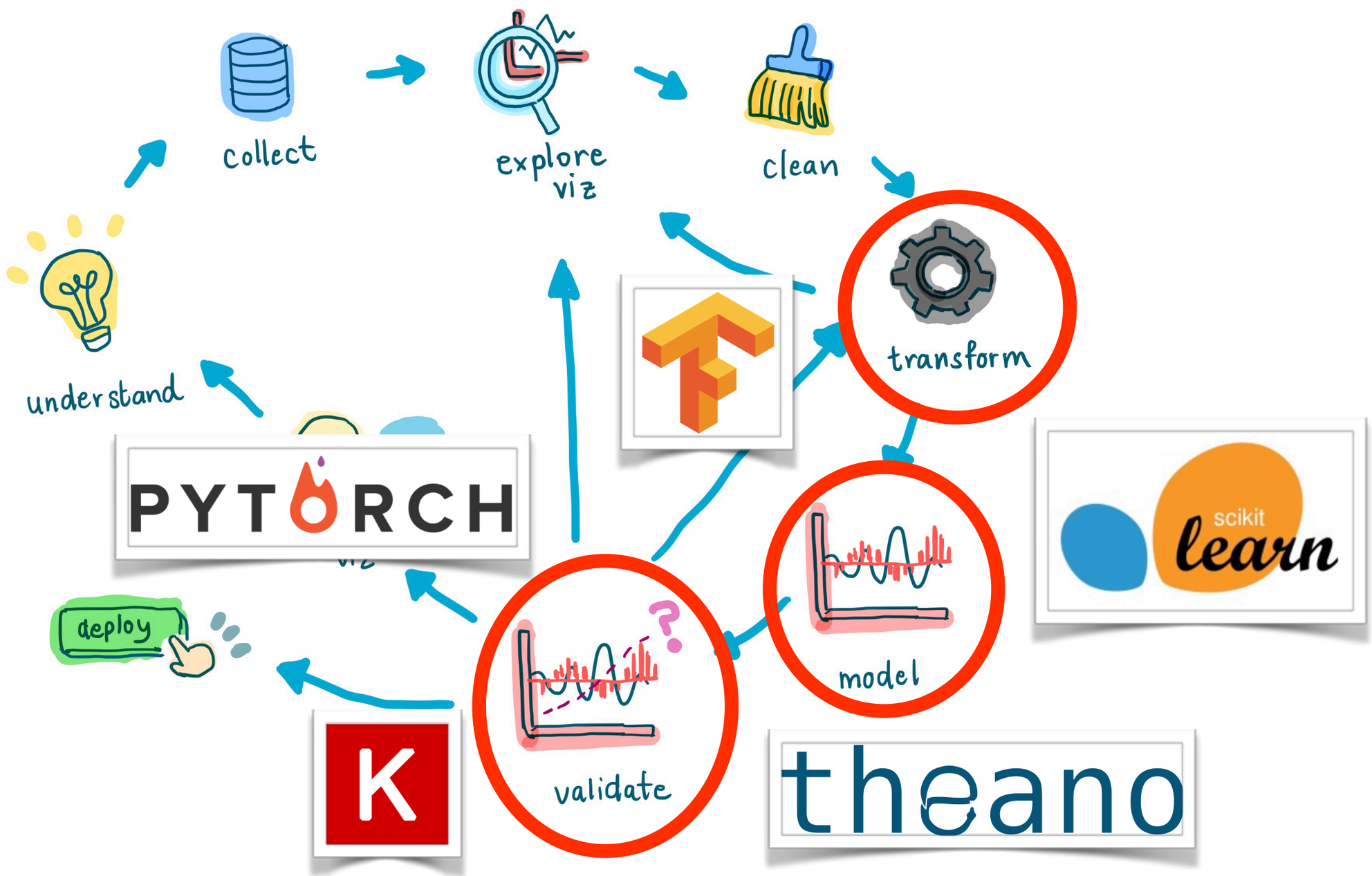
Machine Learning Process



Machine Learning Process

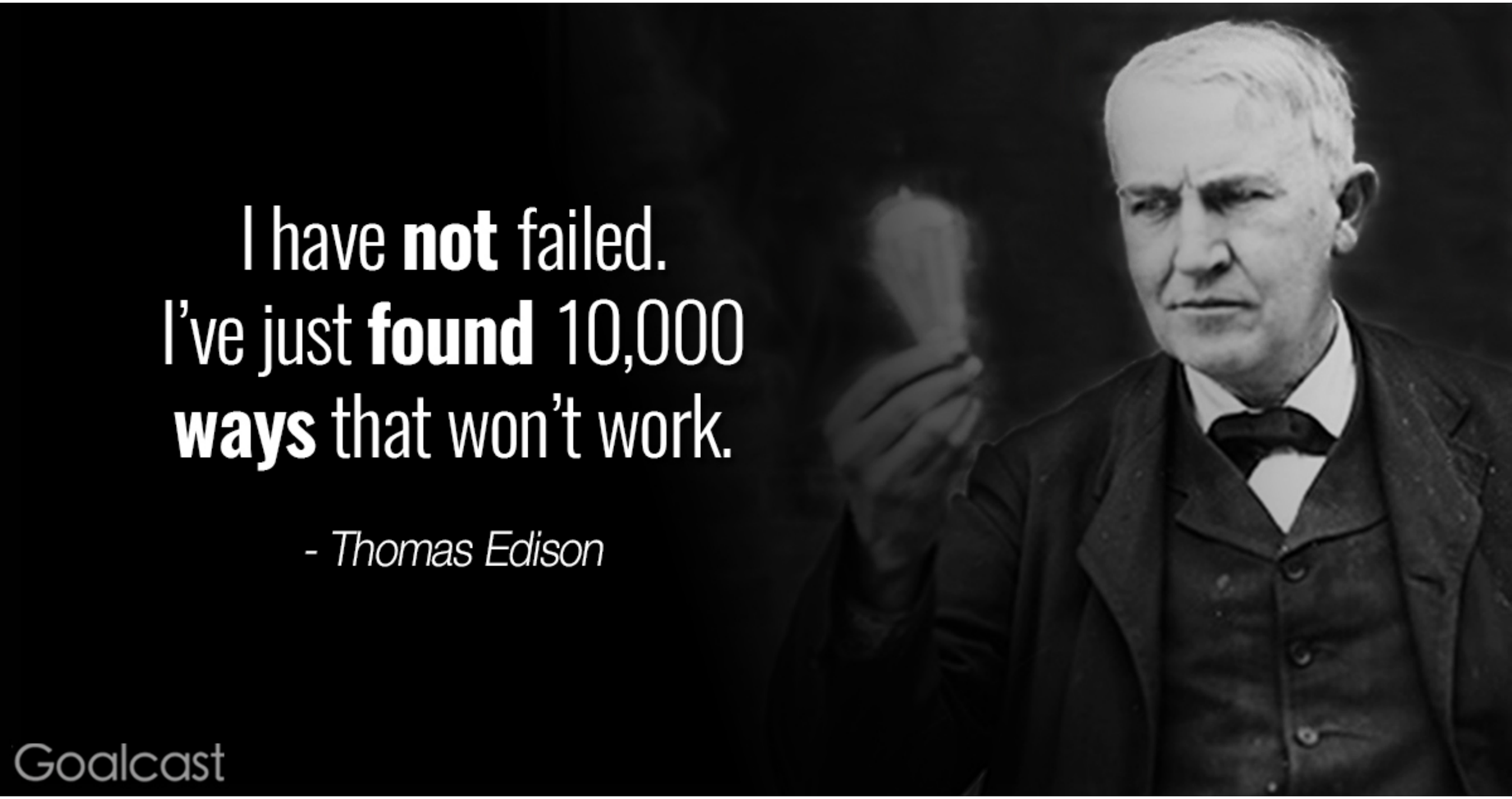


Machine Learning Process



Tools mentioned in the Process

- NumPy - http://www.numpy.org/
- Pandas - https://pandas.pydata.org/
- Bokeh - http://bokeh.pydata.org/en/latest/
- Matplotlib - https://matplotlib.org/
- Seaborn - https://seaborn.pydata.org/
- Scikit-Learn - http://scikit-learn.org/
- TensorFlow - https://www.tensorflow.org/
- PyTorch - http://pytorch.org/
- Theano - http://deeplearning.net/software/theano/
- Keras - https://keras.io/


I have **not** failed.
I've just **found** 10,000
ways that won't work.
- *Thomas Edison*
Goalcast

Cautions

- Curse of dimensionality
- Correlation does NOT imply causation.
- Learn many models, not just ONE.
- More data beats a cleaver algorithm.
- Data alone are not enough.

If two models fit the data equally well,
choose the **simpler** one.

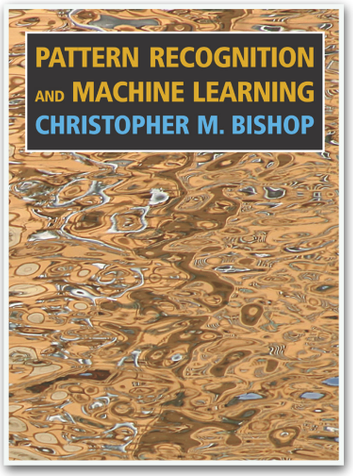
ML system cannot predict stuff
it doesn’t know about.

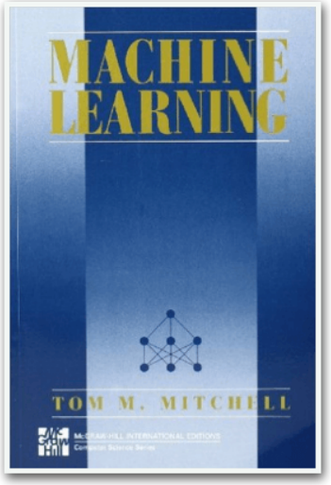
Garbage In - Garbage Out

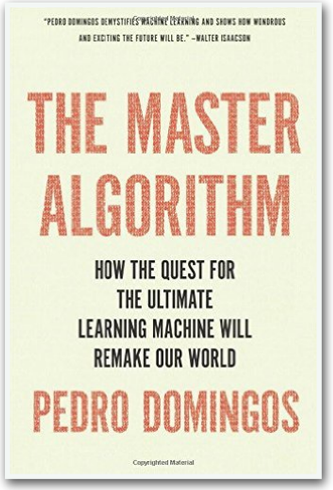
http://blog.marksgroup.net/2013/05/zoho-crm-garbage-in-garbage-out-its.html

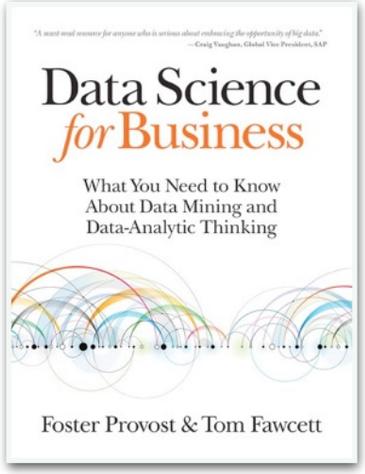
Resources?

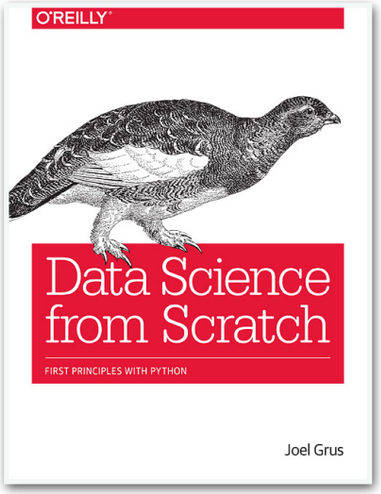
Recommended Books

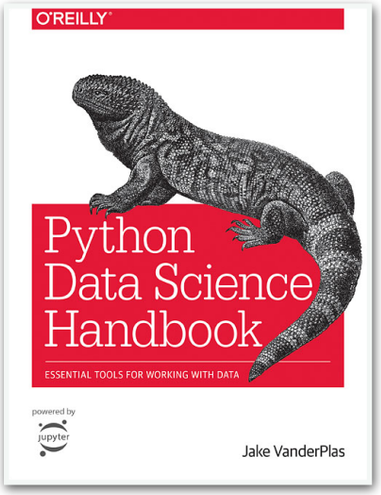




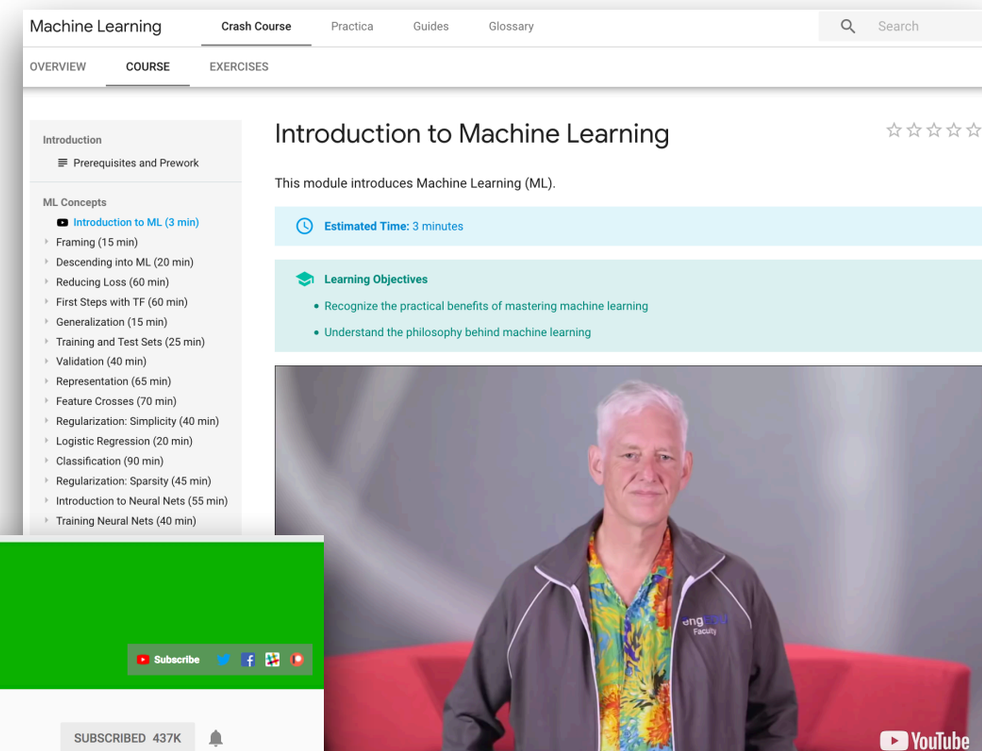


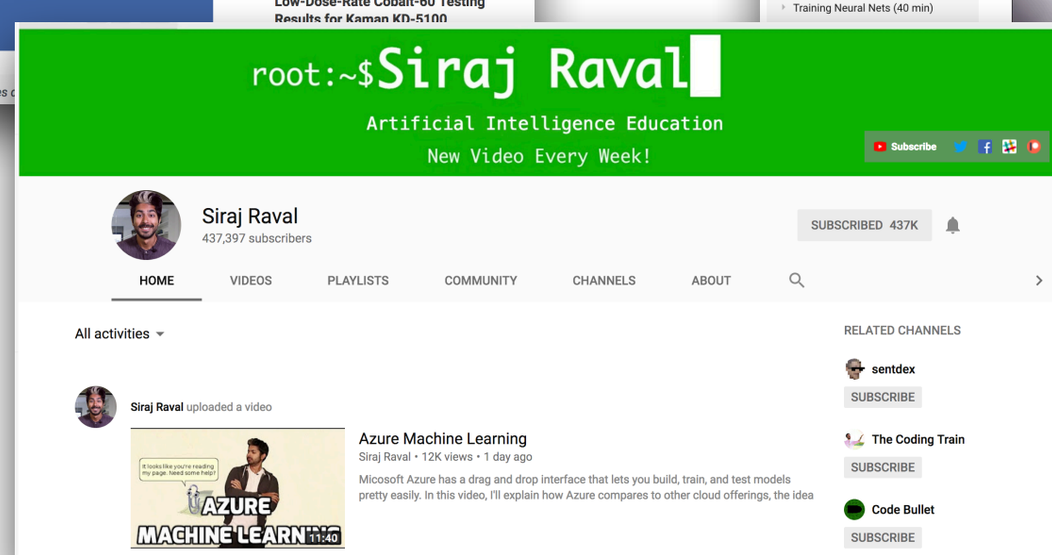






Recommended Videos





DataTau new | threads | comments | leaders | submit

1. Data Science in Entertainment and Media (dominodatalab.com)
   2 points by AnnaOnTheWeb 2 hours ago | discuss
2. Building Scalable Data Science pipelines with Luigi, Apache Spark,Flask (github.com)
   7 points by nischalhp 1 day ago | 2 comments
3. Building a Forex Data Pipeline w/ Python, Airflow, Postgres & Redis (marksblogg.com)
   13 points by marklit 2 days ago | 1 comment
4. Take a Break from Deep Learning, Think About Causality Instead (yanirseroussi.com)
   4 points by brohrer 1 day ago | discuss
5. List of IPython (Jupyter) Notebooks by Peter Norvig (norvig.com)
   8 points by pmigdal 2 days ago | discuss
6. Introduction to Stats and Basics of Maths for Data Science - The Hacker's Way (github.com)
   11 points by rouse 3 days ago | 2 comments
7. Sharding Postgres with semi-structured data and its performance implications (citusdata.com)
   3 points by craigkerstiens 1 day ago | discuss
8. Datatau is now among top blogs by traffic (analyticscosm.com)
   2 points by datameer 1 day ago | discuss
9. Understanding Gaussian Processes with 3D shapes (unibas.ch)
   13 points by izahg 6 days ago | discuss
10. Gradient Boosting: explained in 3d (github.io)
    18 points by axelr 7 days ago | 1 comment
11. The 10 Algorithms Machine Learning Engineers Need to Know — Medium (medium.com)
    12 points by Anon84 8 days ago | 2 comments
12. What factors can increase your data science salary? (springboard.com)
    4 points by Rogerh91 4 days ago | discuss
13. Predicting what user reviews are about with LDA (github.com)
    9 points by yippieale 7 days ago | discuss

http://www.datatau.com/

Research at Facebook

Search publications, blog posts and more

Home Research Academic Programs **Publications** Blog People

Publications

Year Month Forum

Learning to Refine Object Segments

Pedro O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, Piotr Dollar

ECCV · October 10, 2016

In this work we propose to augment feedforward nets for object segmentation with a novel top-down refinement approach.

Facebook AI Research (FAIR), Artificial Intelligence, Computer Vision, Applied Machine Learning, Machine Learning

Metric Learning with Adaptive Density Discrimination

Oren Rippel, Manohar Paluri, Piotr Dollar, Lubomir Bourdev

ICLR · May 2, 2016

Distance metric learning approaches learn a transformation to a representation space in which distance is in correspondence with a predefined notion of similarity.

Facebook AI Research (FAIR), Artificial Intelligence, Applied Machine Learning, Machine Learning, Computer Vision

Improved Arabic Dialect Classification with Social Media Data

Fei Huang

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing · September 17, 2015

Arabic dialect classification has been an important and challenging problem for Arabic language processing, especially for social media text analysis and machine translation. In this paper we propose...

Research Groups

Applied Machine Learning

Computational Photography

Connectivity Lab

Core Data Science

Facebook AI Research (FAIR)

Systems Infrastructure

User Experience Research

Networking Group

Research Areas

Artificial Intelligence

Computer Vision

Connectivity

Data Science

Economics and Computation

https://research.facebook.com/publications/machinelearning/

Research at Google

Search

Home Publications People Teams Outreach Blog Work at Google

Machine Intelligence

Research at Google is at the forefront of innovation in Machine Intelligence, with active research exploring virtually all aspects of machine learning, including deep learning and more classical algorithms. Exploring theory as well as application, much of our work on language, speech, translation, visual processing, ranking and prediction relies on Machine Intelligence. In all of those tasks and many others, we gather large volumes of direct or indirect evidence of relationships of interest, applying learning algorithms to understand and generalize.

Machine Intelligence at Google raises deep scientific and engineering challenges, allowing us to contribute to the broader academic research community through technical talks and publications in major conferences and journals. Contrary to much of current theory and practice,

567 Publications

A Simple and Efficient Method to Handle Sparse Preference Data Using Domination Graphs: An Application to YouTube

Shumeet Baluja

ICCS 2016, 2302–2311

Adversarial Autoencoders

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow

International Conference on Learning Representations (2016) (to appear)

http://research.google.com/pubs/MachineIntelligence.html

UDACITY

coursera

edX

The 7 Best Data Science and Machine Learning Podcasts

Learn the basics and keep up with the latest news in data science, machine learning and artificial intelligence by listening to these great podcasts

https://medium.com/swlh/the-7-best-data-science-and-machine-learning-podcasts-e8f0d5a4a419#.94hla5goy

rushter / data-science-blogs

Watch 260 | Unstar 2,708 | Fork 430

Code | Pull requests 0 | Projects 0 | Pulse | Graphs

A curated list of data science blogs

275 commits | 1 branch | 0 releases | 39 contributors

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

rushter Update opml file — Latest commit a1c86cb 23 days ago

README.md	Sort by blog name	23 days ago
data-science.opml	Update opml file	23 days ago
frequency.py	Add blog post frequency script	a year ago
generate_opml.py	Fix opml generator	a year ago

README.md

https://github.com/rushter/data-science-blogs

Data science blogs

A curated list of data science blogs

- A Blog From a Human-engineer-being http://www.erogol.com/ (RSS)
- Aakash Japi http://aakashjapi.com/ (RSS)

kaggle™

Join the Communities!

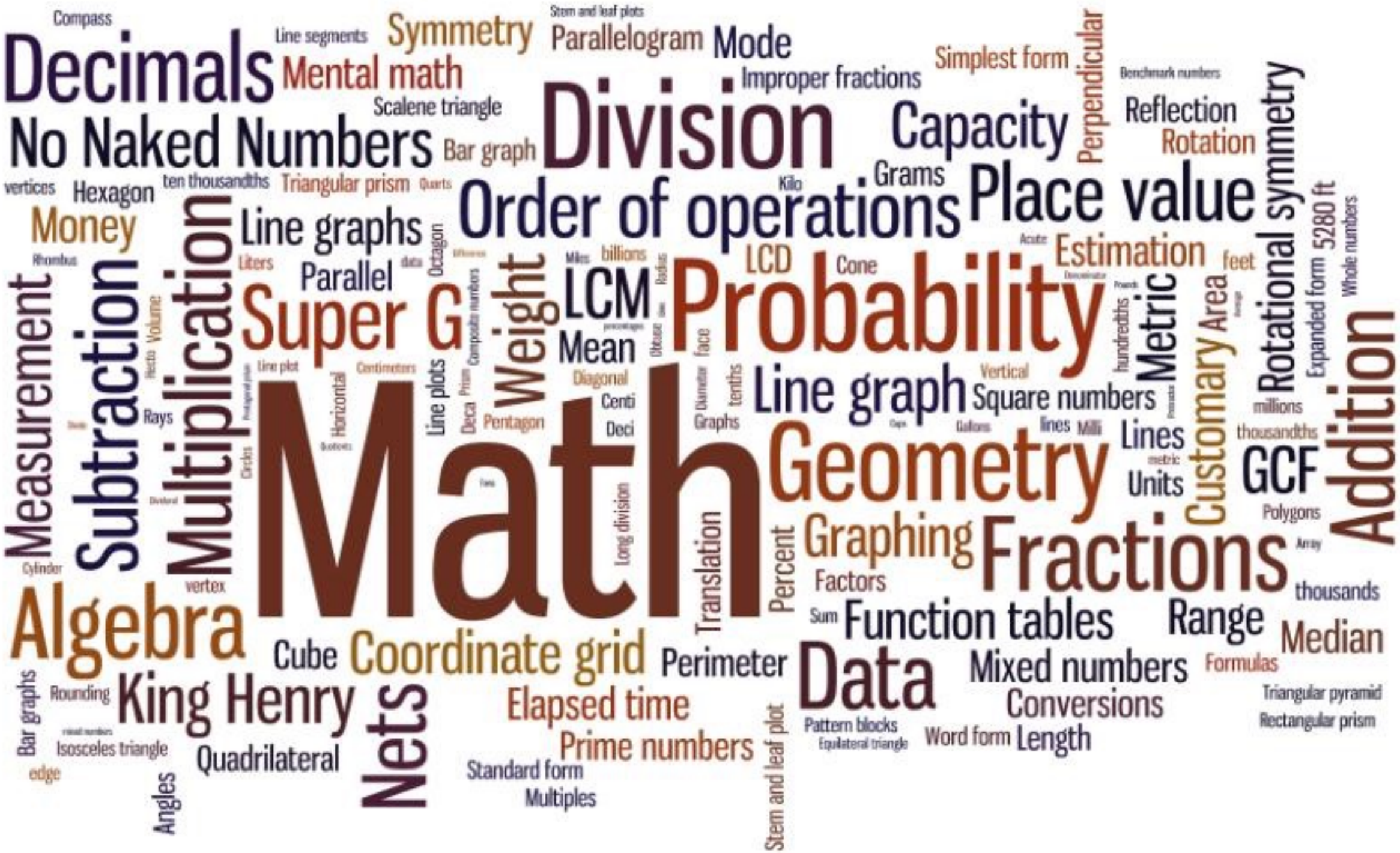
http://www.meetup.com/DataScienceTH/

https://www.facebook.com/groups/DataCouncilAIBangkok/

https://www.facebook.com/datasciencechill/

http://www.meetup.com/bangkokml/

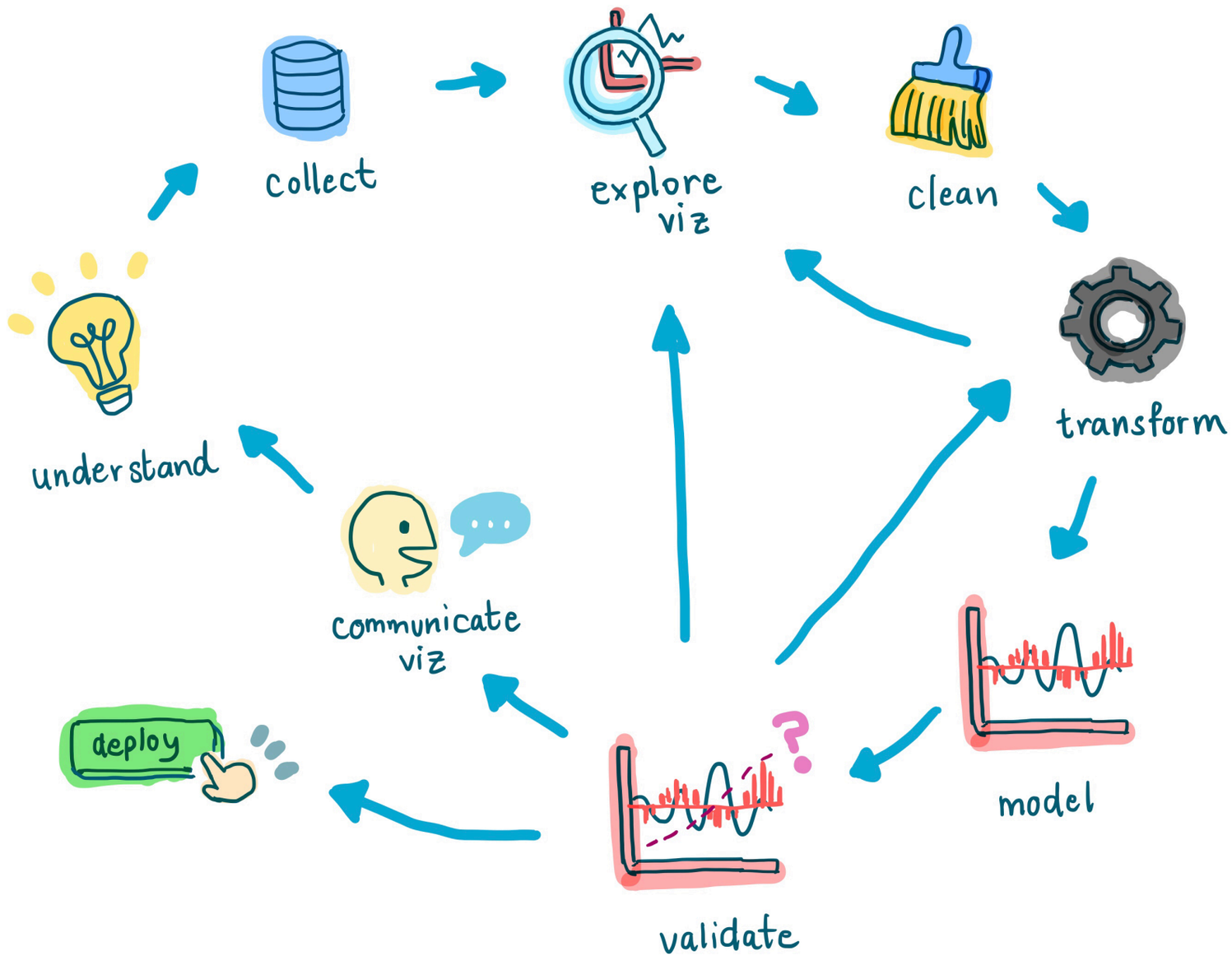
https://www.facebook.com/datarockie/



http://www.barnstable.k12.ma.us/domain/210

https://twitter.com/robinhanson/status/803302419249725440

Machine Learning Process



Course Outline

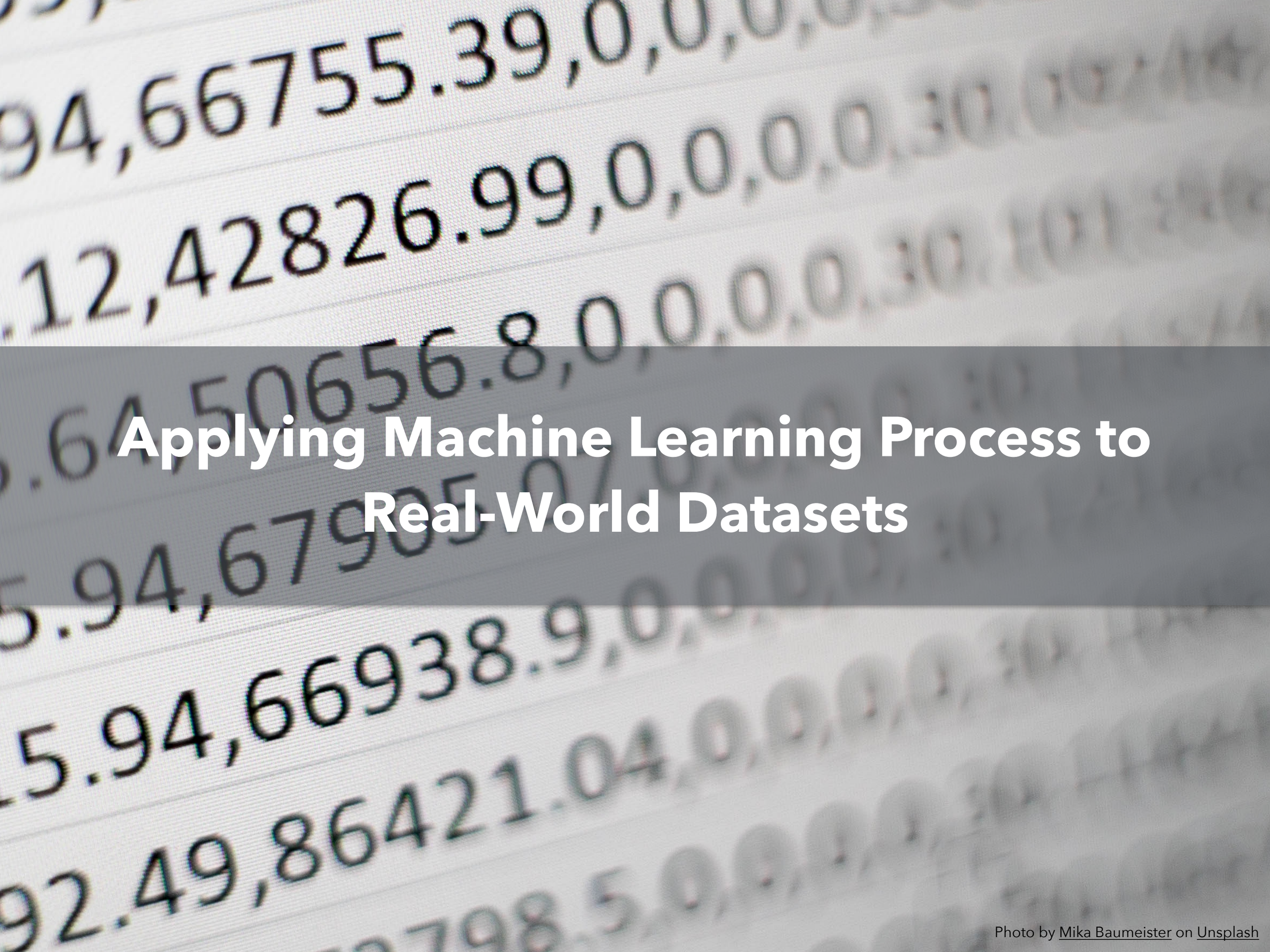
1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning



Applying Machine Learning Process to Real-World Datasets

Photo by Mika Baumeister on Unsplash

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. ✅ Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Machine Learning Model Deployment

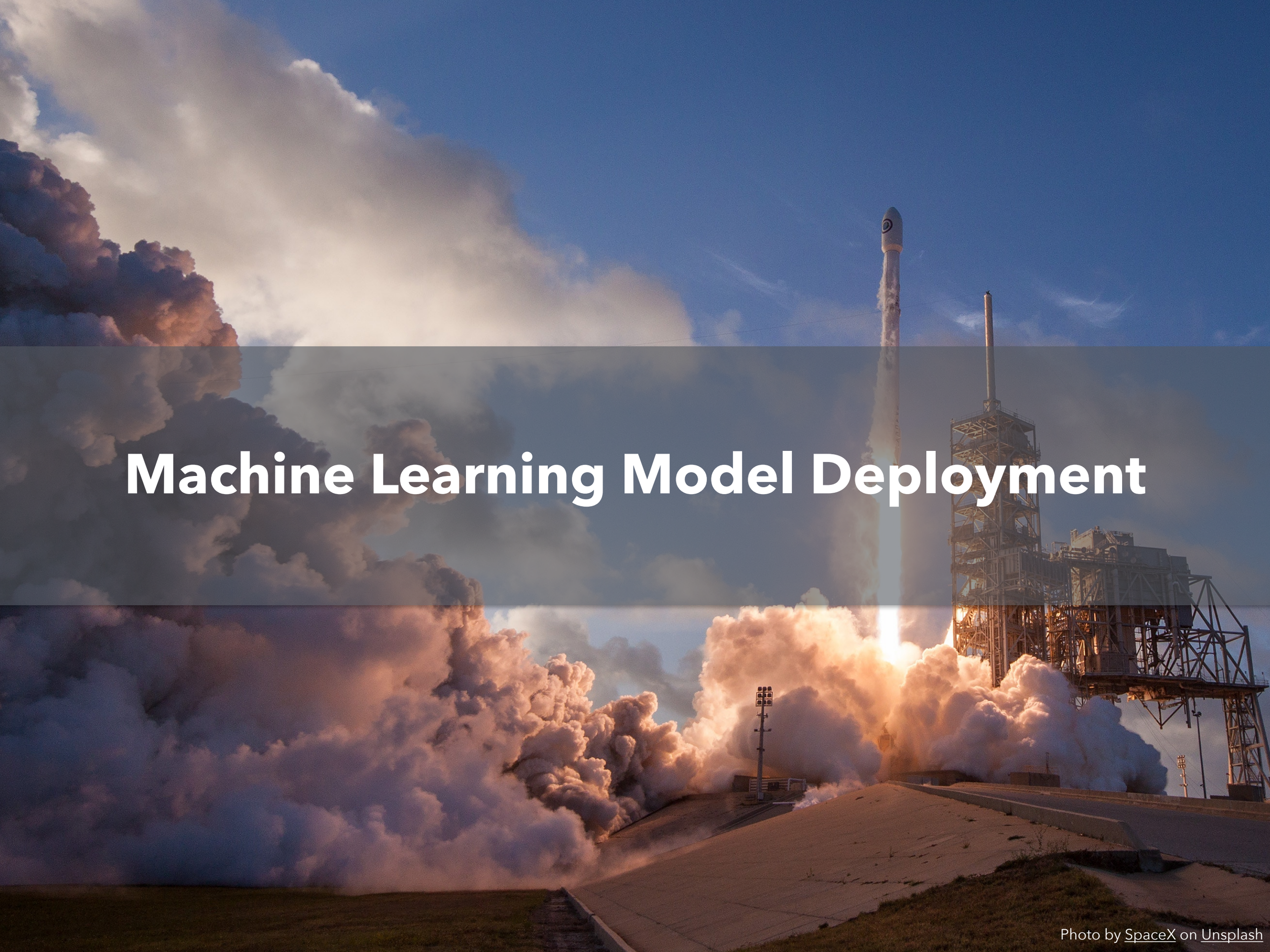


Photo by SpaceX on Unsplash

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. ✅ Applying Machine Learning Process to Real-World Datasets
5. Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. ✅ Applying Machine Learning Process to Real-World Datasets
5. ✅ Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Brief Intro to Deep Learning

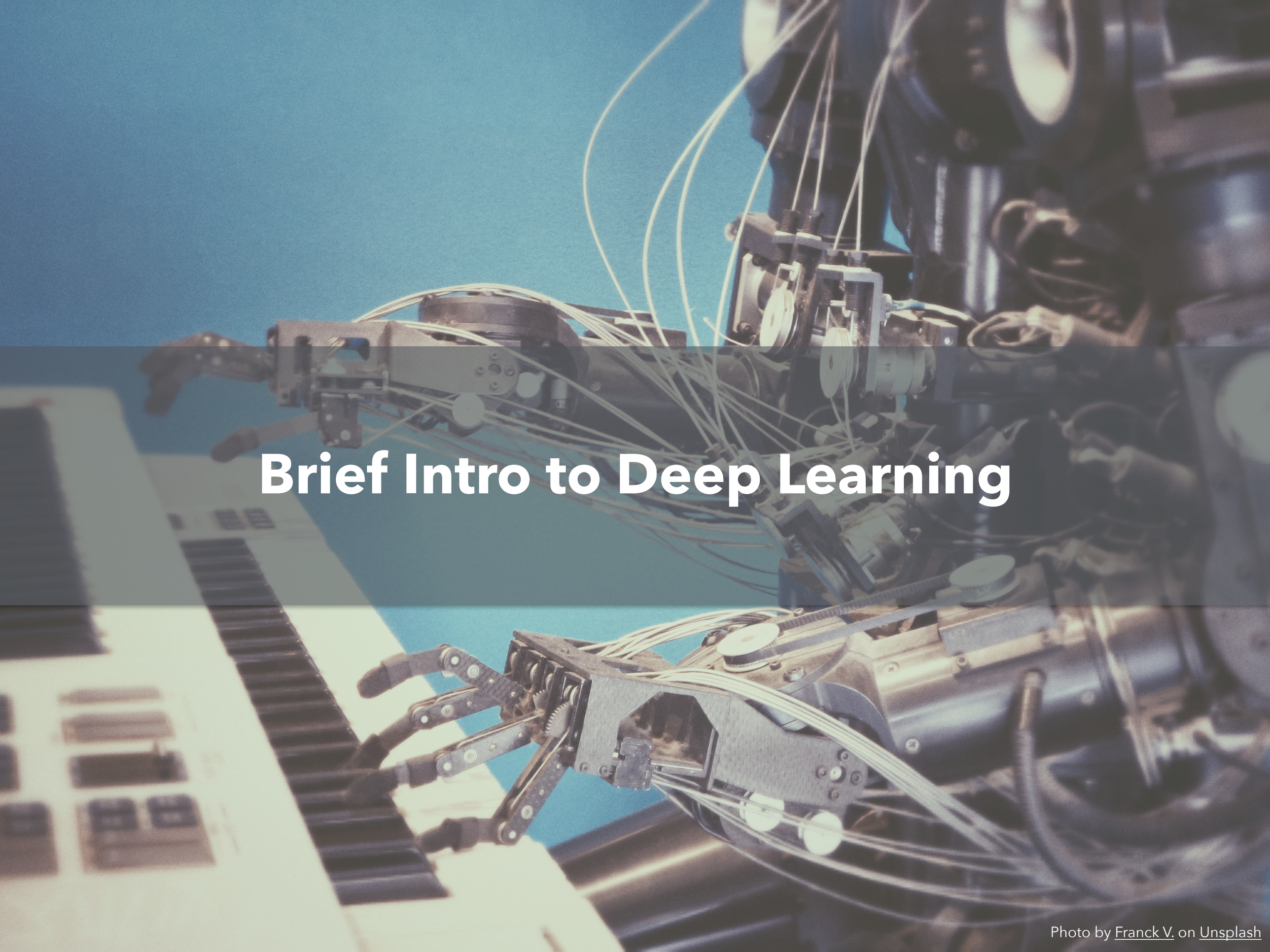
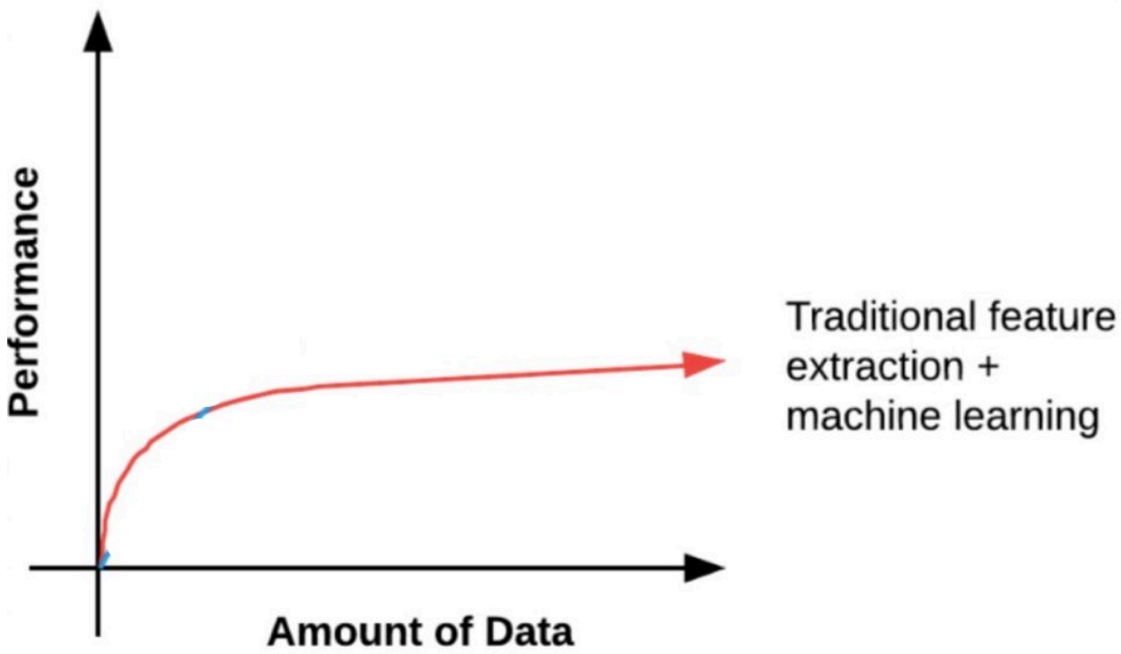
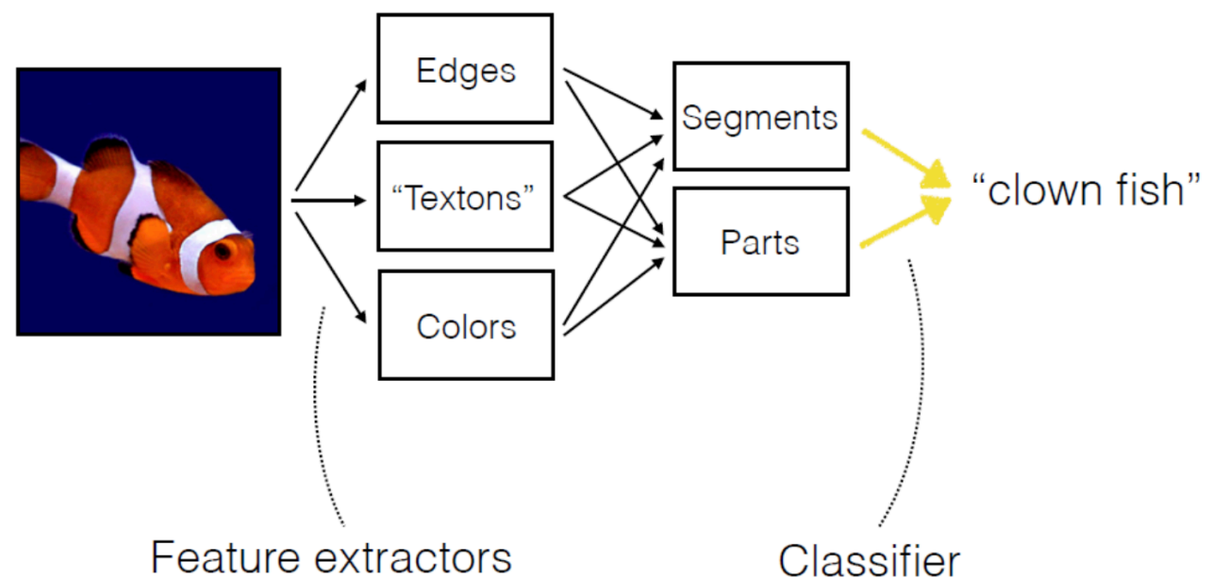


Photo by Franck V. on Unsplash

Traditional Machine Learning

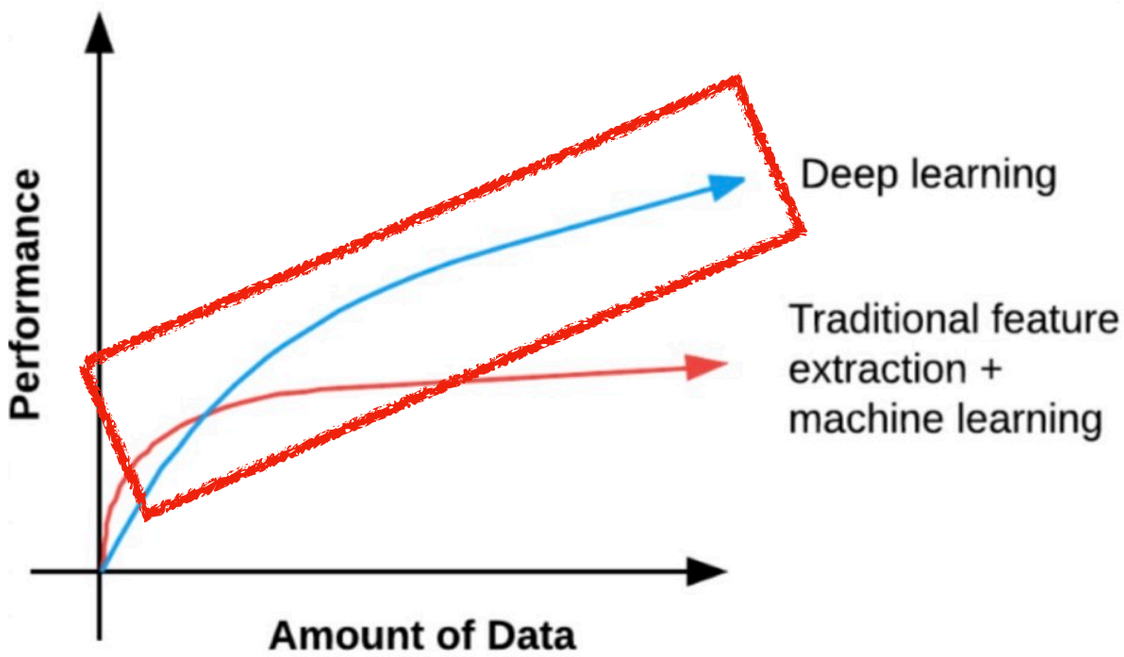


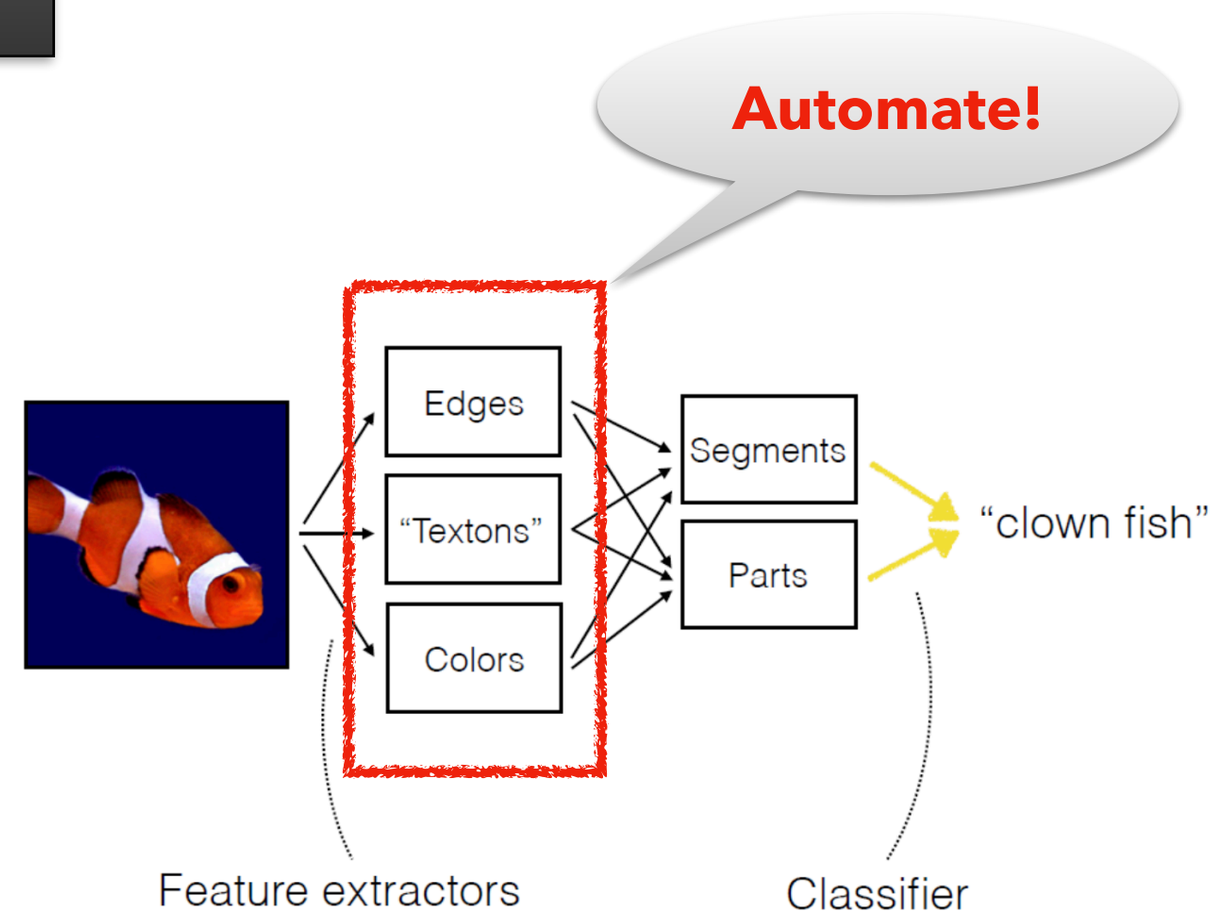
Deep Learning for Computer Vision with Python by Dr. Adrian Rosebrock



http://i-systems.github.io/HSE545/machine%20learning%20all/16%20Deep%20learning/CNN.html

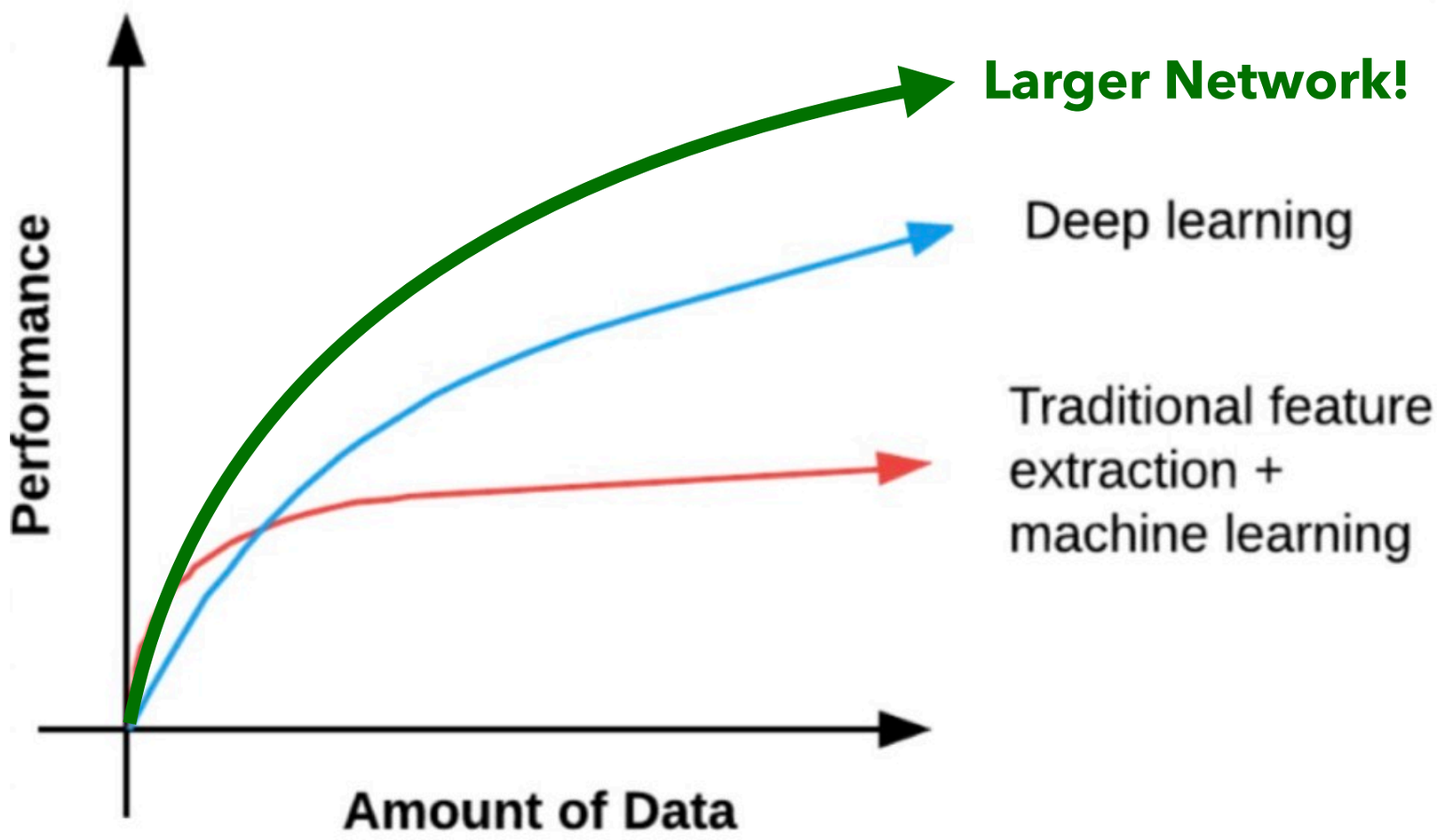
Deep Learning





Deep Learning for Computer Vision with Python by Dr. Adrian Rosebrock

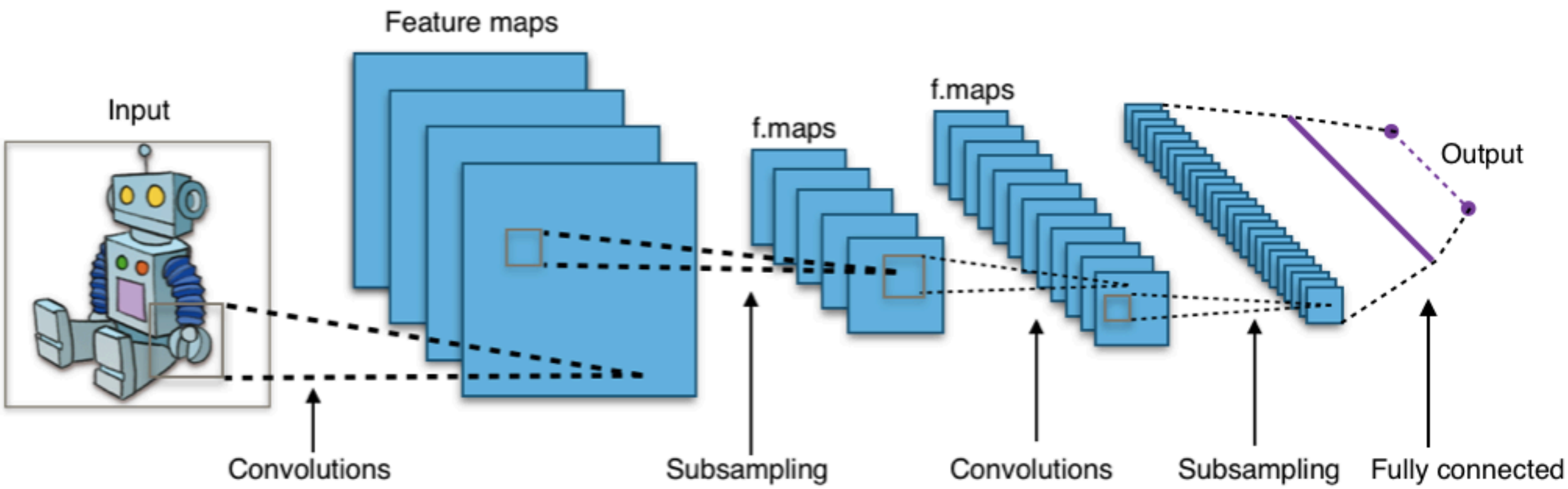
Deep Learning with Larger Network



Deep Learning for Computer Vision with Python by Dr. Adrian Rosebrock

Convolutional Neural Networks (CNNs)

A Class of Deep Neural Networks



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

About 3 Types of Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

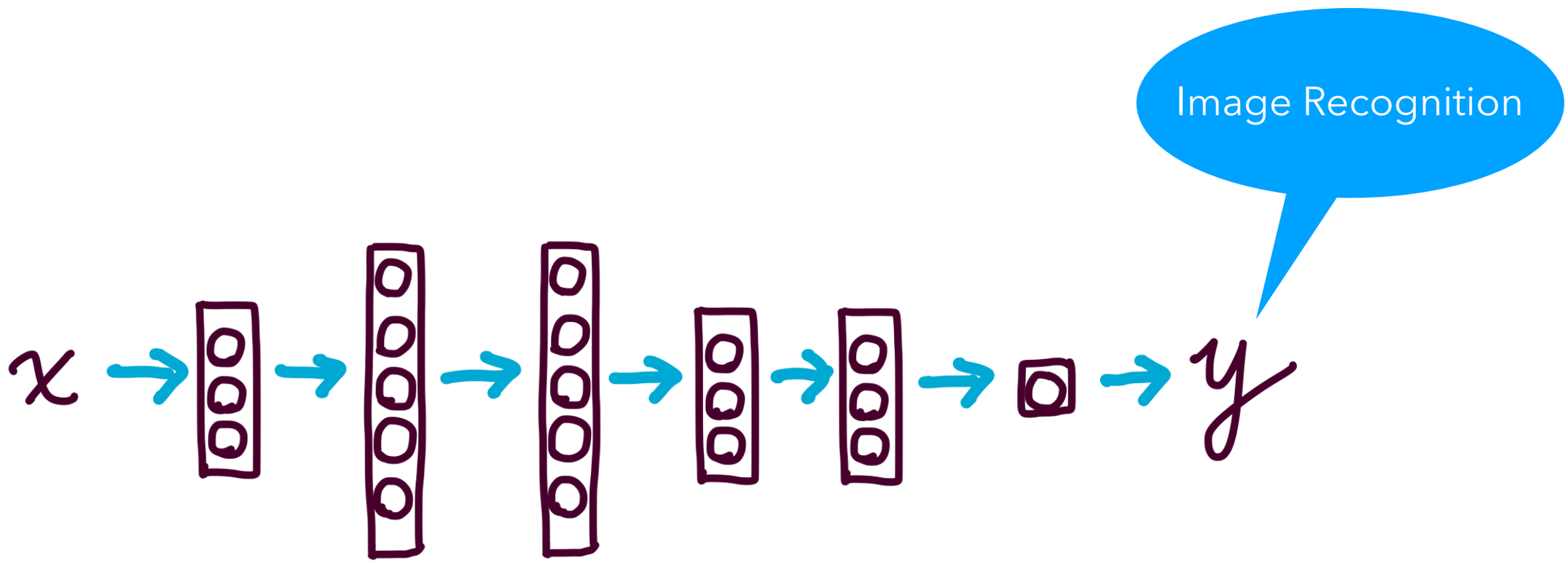
One Type of Learning Dominating the World

- **Supervised Learning**
- Unsupervised Learning
- Reinforcement Learning

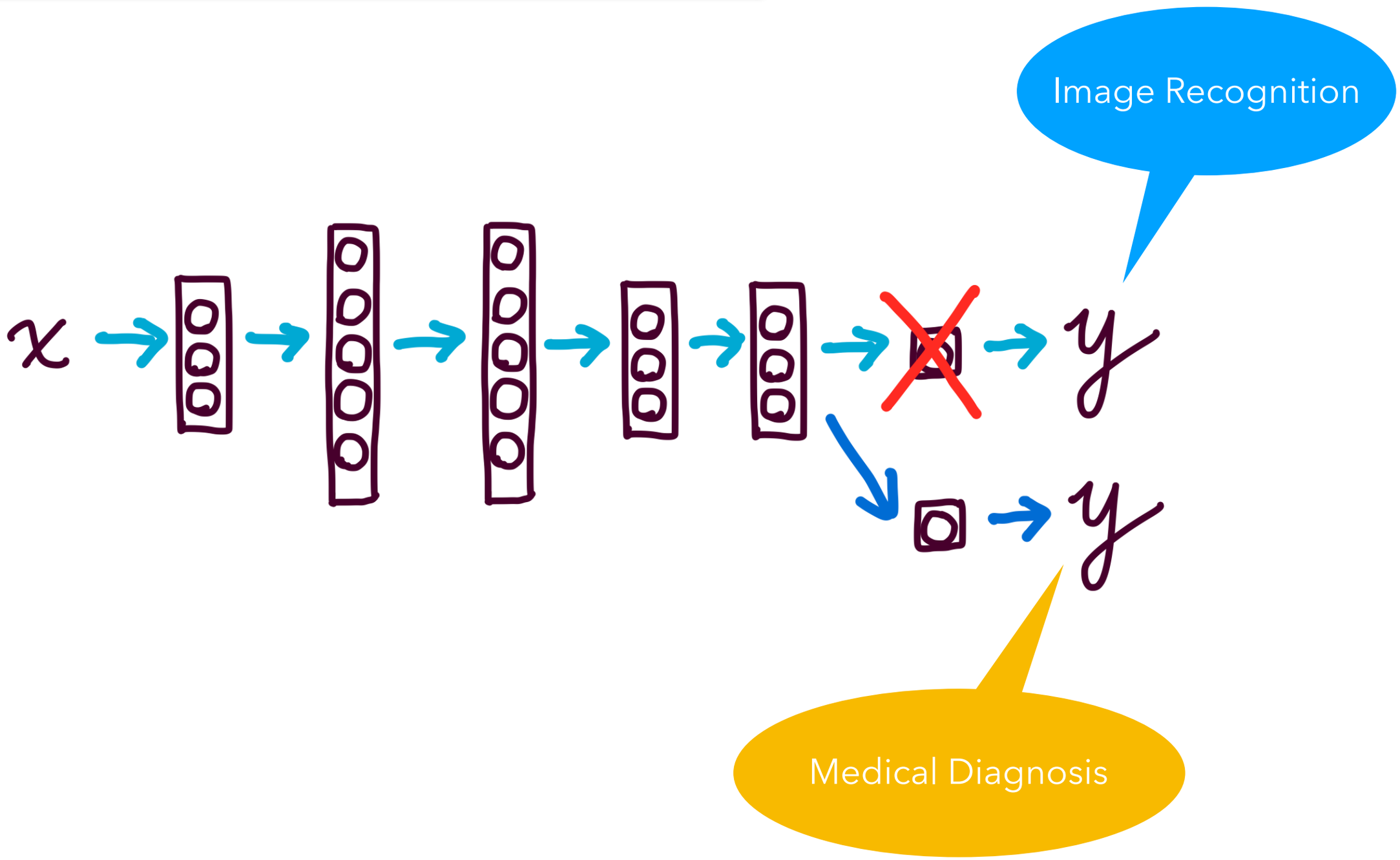
Two More Interesting Types of Learning

- **Supervised Learning**
- Unsupervised Learning
- Reinforcement Learning
- **Transfer Learning**
- **Adversarial Learning**

Transfer Learning



Transfer Learning (cont.)



Adversarial Learning

Stats and Bots Following

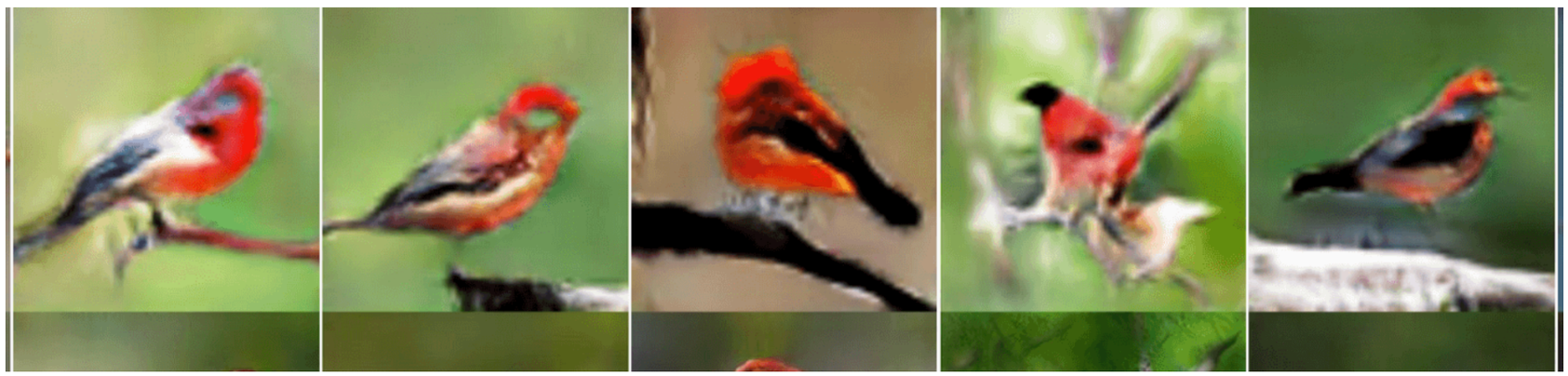
HOME DATA SCIENCE ANALYTICS STARTUPS BOTS DESIGN SUBSCRIBE | TRY STATSBOT FREE

Applause from Luba Belokon, Sakares Saengkaew, and 116 others

Anton Karazeev Follow
Researcher at MIPT
Aug 17 · 8 min read

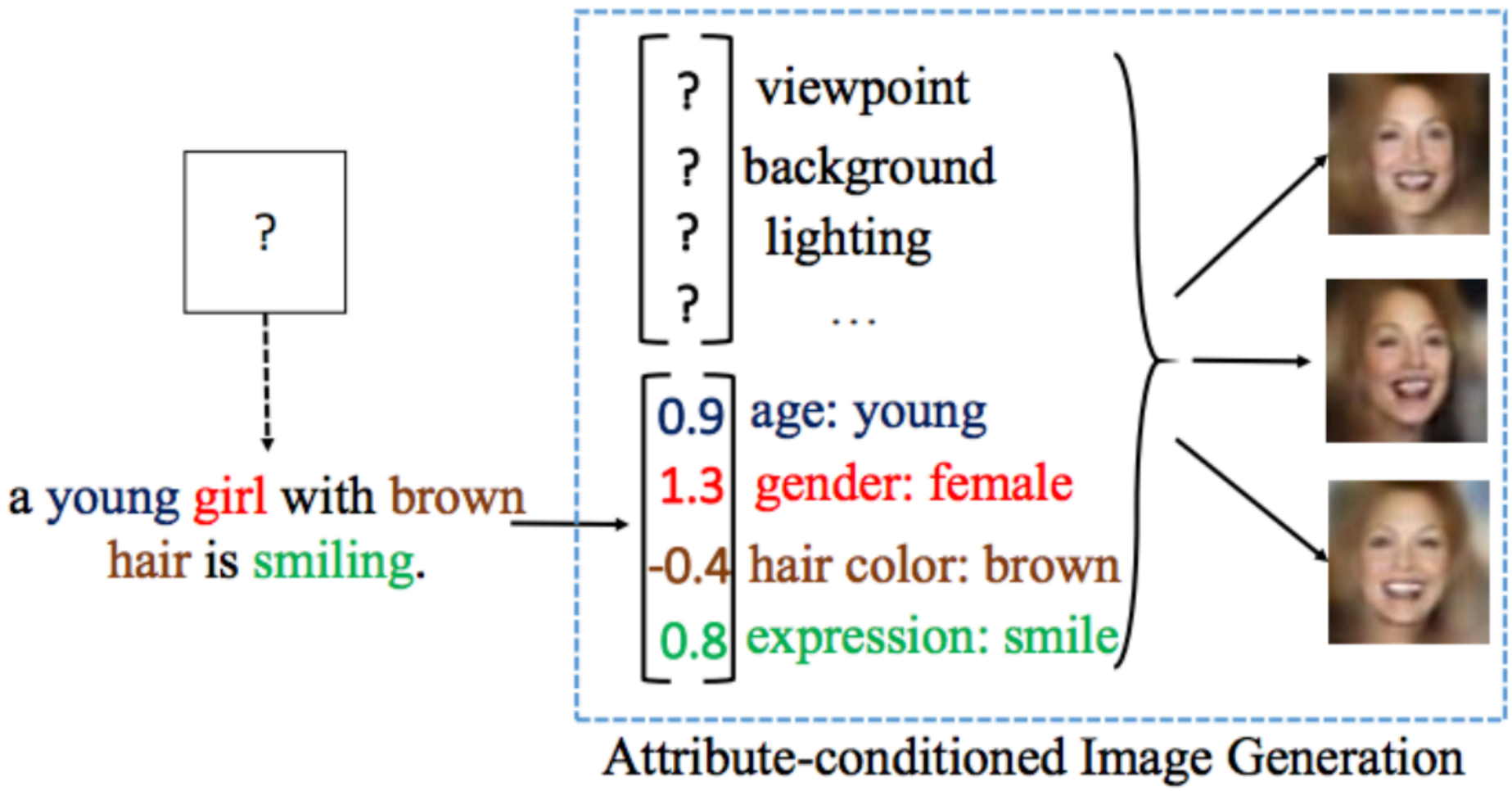
Generative Adversarial Networks (GANs): Engine and Applications

How generative adversarial nets are used to make our life better



https://blog.statsbot.co/generative-adversarial-networks-gans-engine-and-applications-f96291965b47

Example of Adversarial Learning Application



Attribute2Image: Conditional Image Generation from Visual Attributes by Yan et al.

https://arxiv.org/pdf/1512.00570v2.pdf

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. ✅ Applying Machine Learning Process to Real-World Datasets
5. ✅ Machine Learning Model Deployment
6. Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning

Course Outline

1. ✅ Data Science and Tools
2. ✅ Intro to Machine Learning
3. ✅ Machine Learning Process
4. ✅ Applying Machine Learning Process to Real-World Datasets
5. ✅ Machine Learning Model Deployment
6. ✅ Brief Intro to Deep Learning

https://www.re-work.co/blog/machine-intelligence-junhyuk-oh-university-michigan-reinforcement-learning