
CASE FOR OPEN DATA IN MACHINE LEARNING

Janu Verma

Hike, New Delhi

@januverma

About Me

- Data Scientist at Hike, New Delhi — machine learning, recommendation systems, NLP, network analysis.
 - Previously :
 - IBM Research, New York
 - Cornell University
 - Kansas State University
 - Cambridge University
 - Researcher in machine learning , data visualization, and mathematics.
 - Public Speaking, Teaching ML and data science.
 - Advise Startups on ML, data science, and hiring & team building.
-

Open Data

- Data that can be freely accessed, used, modified and shared by anyone for any purpose - subject only, at most, to requirements to provide attribution or share alike license.
 - Specifically, requires that the data be
 - **Legally open:** that is, available under an open (data) license that permits anyone freely to access, reuse and redistribute.
 - **Technically open:** that is, that the data be available for no more than the cost of reproduction and in machine-readable form.
 - Source: Open Data Handbook (opendatahandbook.org/)
-

Open Data In The Wild

- Government data on demographics, policy planning and outcomes etc. (<https://www.data.gov/>)
 - Amazon Open Data Registry (<https://registry.opendata.aws/>)
 - Google BigQuery Public Datasets (<https://cloud.google.com/bigquery/public-data/>)
 - FRED Economic Data (<https://fred.stlouisfed.org/>)
 - [Archive.org Datasets](#) - The Dataset Collection consists of large data archives from both sites and individuals.
 - Enigma Public - world's broadest collection of public data. (<https://public.enigma.com/>)
 - [Archive-it from Internet Archive](#) - web archiving service for cultural heritage on the web
-

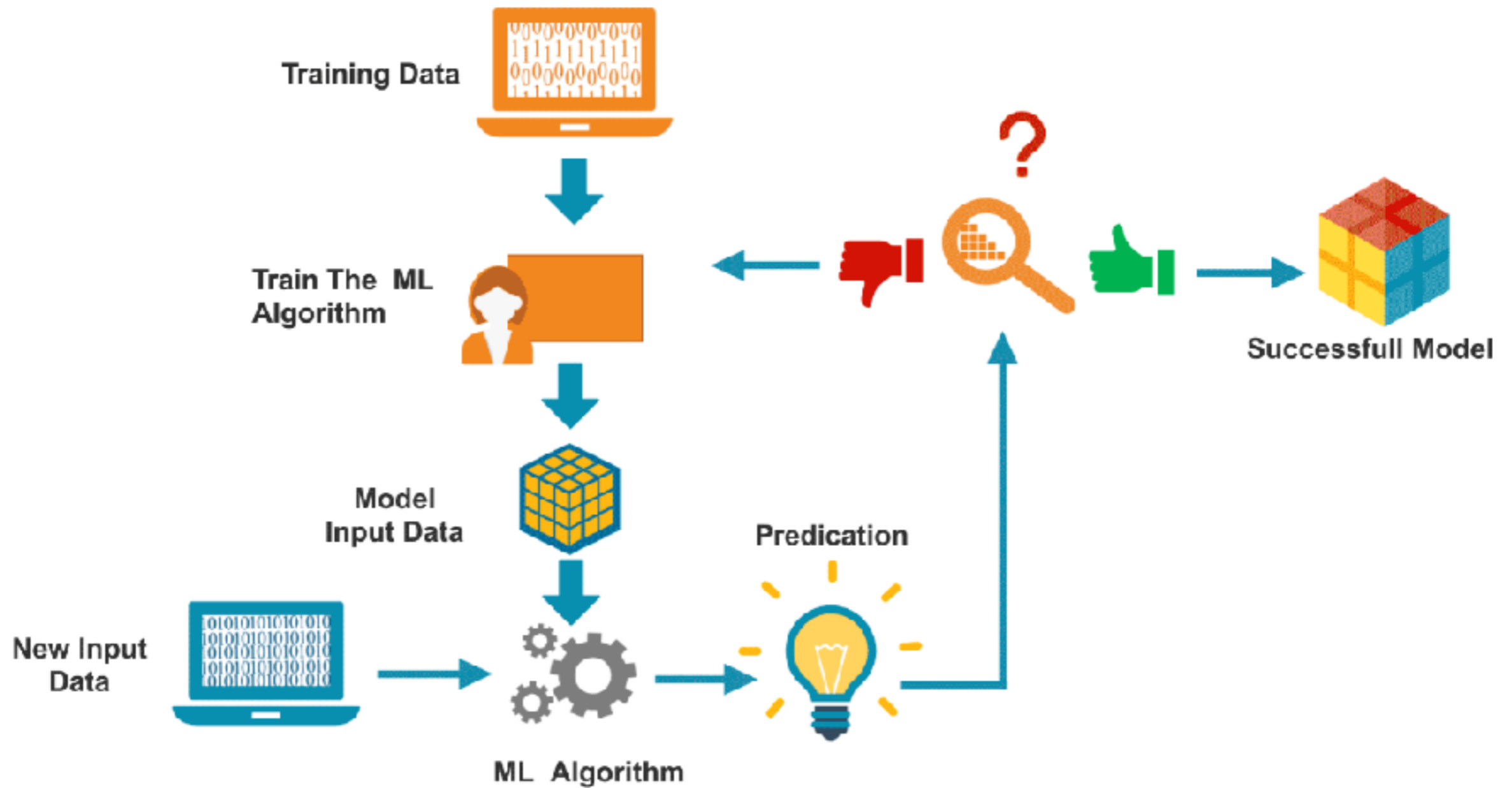
Why Open Data?

- Transparency
 - Finnish *tax tree* and British *where does my money go* show how citizen's tax money is being spent by the govt.
 - folketsting.dk track activity in parliament and the law making processes
 - Democratization of data : Everyone a data scientist.
 - findtoilet.dk which showed all the Danish public toilets.
 - vervuilingsalarm.nl warns you with a message if the air-quality in your vicinity is going to be > the threshold tomorrow.
 - Where you can walk your dog in NYC.
 - In Germany and the UK, find places to live, taking into account the duration of your commute to work, housing prices, and how beautiful an area is.
 - Innovation in products and services - Companies being founded and new/improved products due to open data.
 - Google Translate uses the enormous volume of open EU documents.
-

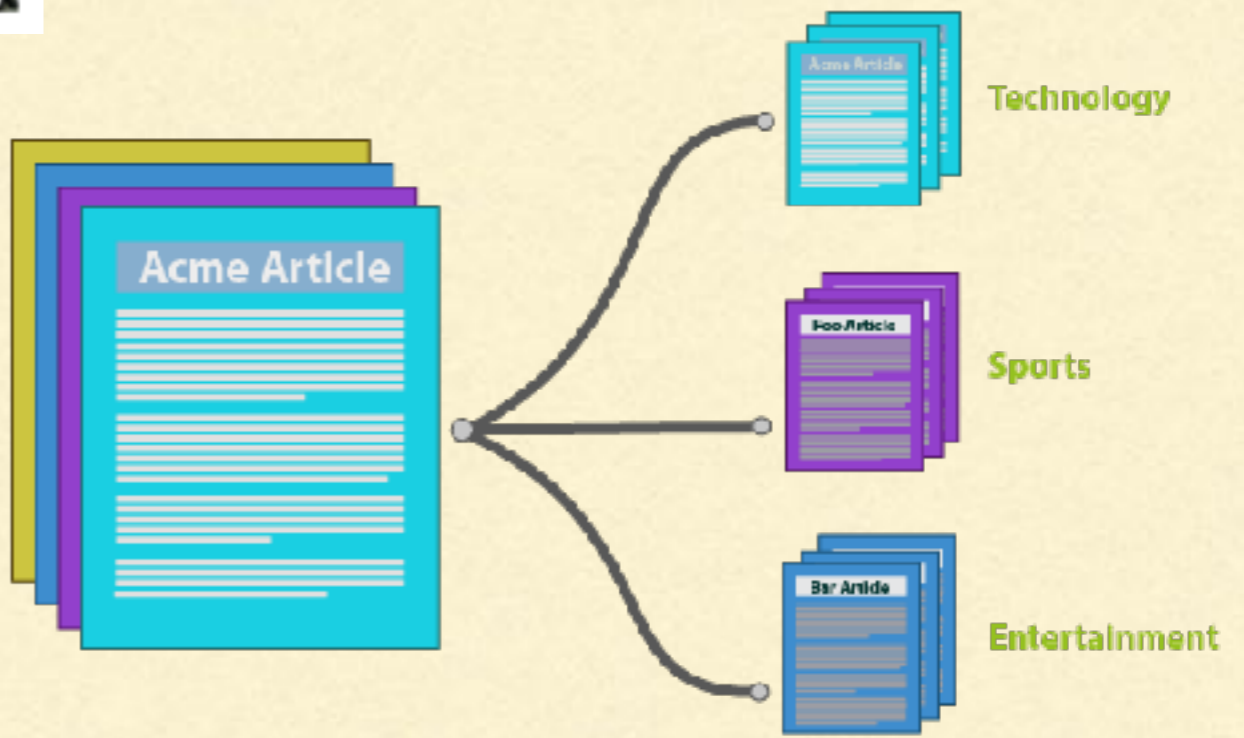
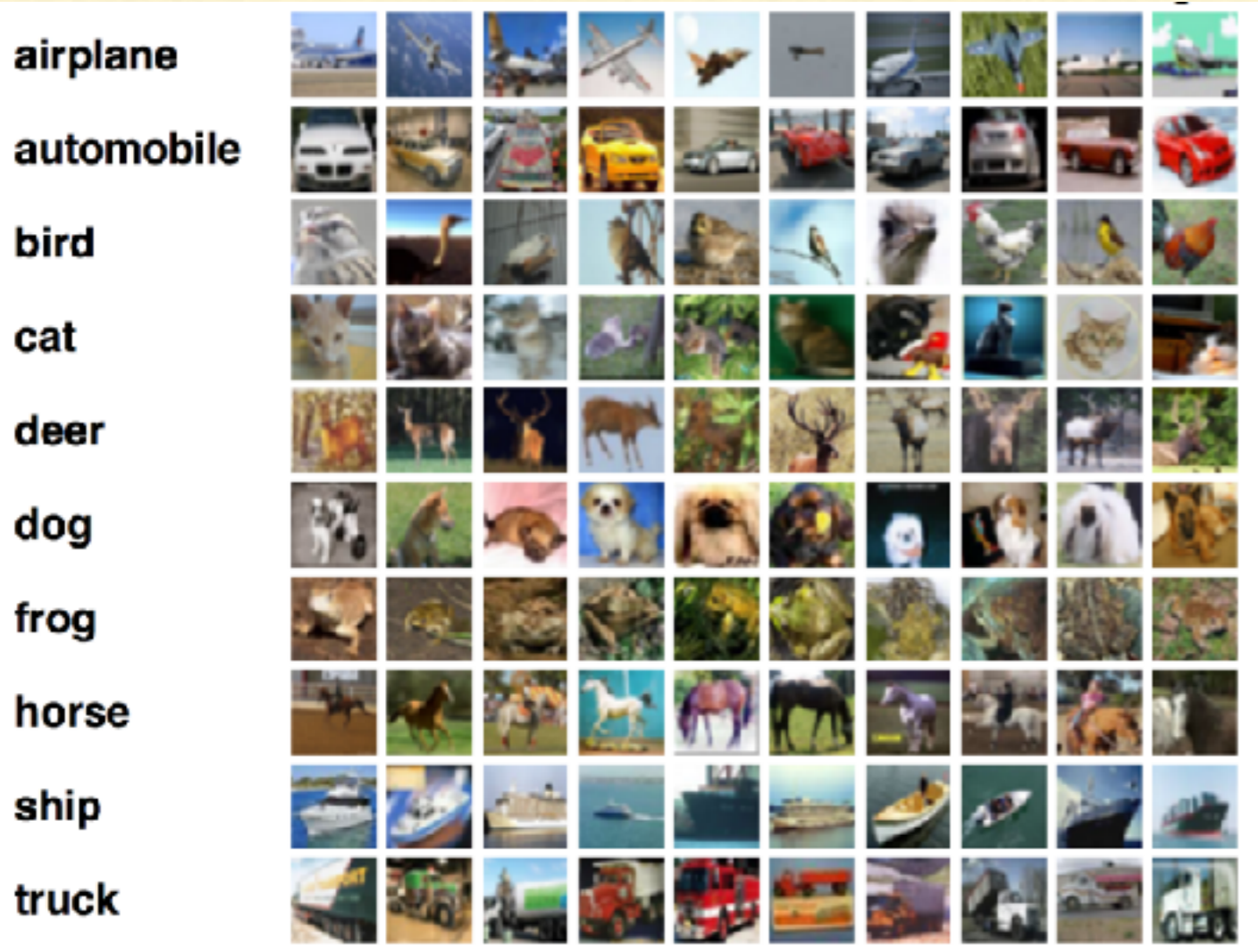
Machine Learning

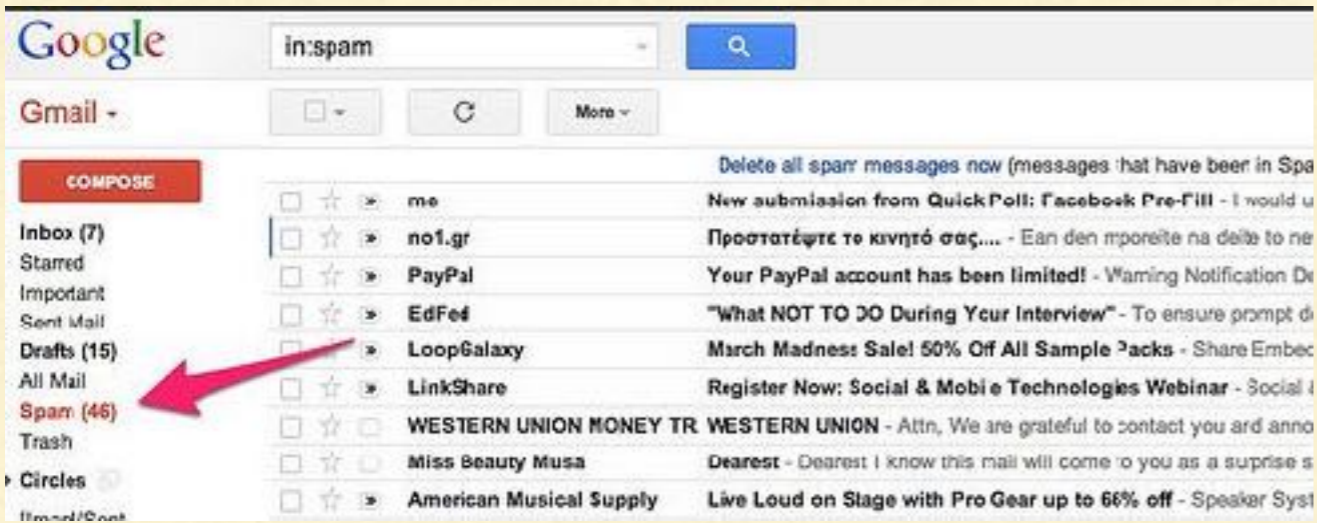
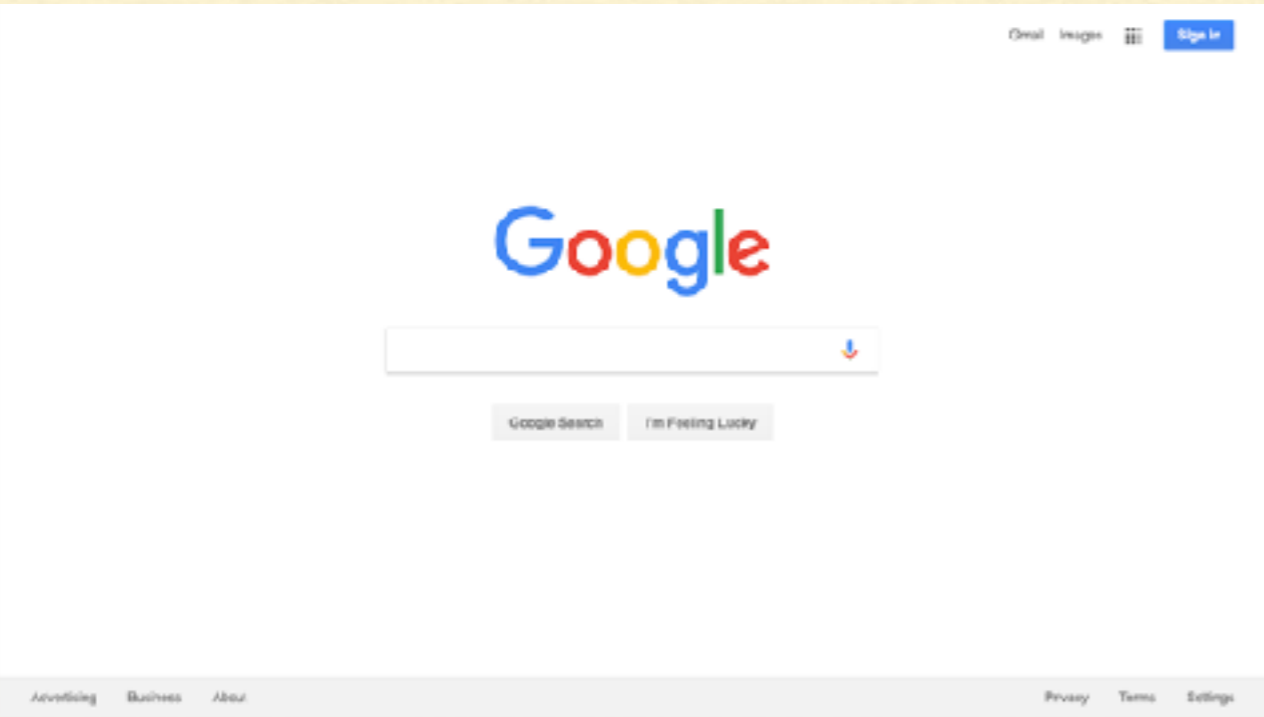
- **Machine learning** is a field of computer science that designs systems with the ability to automatically learn and improve from experience without being explicitly programmed.
- Computer systems that access data and use statistical/mathematical techniques to learn patterns and make inference based on probabilistic responses.
- *Supervised learning* involves providing example inputs and respective outputs to the the program, which 'learns' to make predictions on the outputs for new, unseen inputs.
- e.g. a classification system to categorize the images as cats or dogs.





Machine Learning Pipeline
















People You May Know

Add people you know as friends and connect with public profiles you like.

 Lala Lalabs Add as friend	 Brian Crecente Add as friend	 Brian Ashcraft Add as friend
 justin bieber Add as friend	 Camile Gozon Add as friend	 Karla Danielle Beger Add as friend
 Taylor-Alison Swifty Add as friend	 Adam Rifkin Add as friend	 Luke Plunkett Add as friend

Who to follow - Refresh - View all

	Andrew Robb  @AndrewRobb... Followed by EssentialView and ot... Follow	
	ABC South East SA @abcsouthe... Followed by LGA South Australia ... Follow	
	Ian Shuttleworth @Shutts10 Followed by Jason McConnell an... Follow	

Browse categories - Find friends



A screenshot of a social media post. The main image shows a group of about ten people posing together indoors. Below the image is a list of tagged users:

- Nicholas Carlson (Me) - Salinas, California - Davidson
- Nick Livelli - Santa Teresa, California - American
- Nick Denton - Greater Media - New York, New York
- Nicole Schumacher - Coxsack - New York, New York
- Nick Blum - The New York Times Company - San Francisco, California
- Angela Nicht - Inverness - San Francisco, California
- Nicholas Salvo - Riverdale - Brooklyn, New York
- Erwan Michalec

On the right side of the post, there is a comment section with the following text:

Nicholas Carlson
January 21 via Instagram

SAE http://instagram.com/...

Done Tagging | Add Location | Edit

Like | Comment | Help | Low Power | Share | Edit

Danielle Lacombe and Owen Thomas like this.

Owen Thomas Needs.
January 20 at 6:35pm - likes - 0

Write a comment...

See more

Maziar Kazerooni likes Owen's...

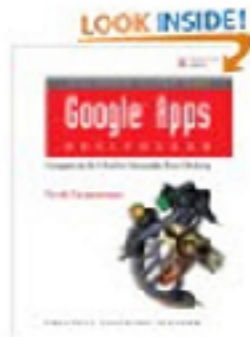
Dom's a's like

Justin Smith likes 1 page

Target a's like

Recommended for You

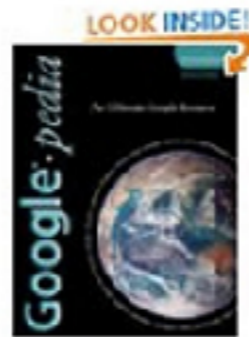
Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)



[Google Apps Administrator Guide: A Private-Label Web Workspace](#)



[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

House of Cards
★★★★★ 2013 TV-MA 1 Season 100 55

Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

Because you watched Orange Is the New Black

Because you watched Red Lights

NETFLIX

Breaking Bad

HOUSE of CARDS

the L word

LAGUE

ALWAYS SUNNY IN PHILADELPHIA

New Girl

SOULS

ELECTRIC MIST

FASTENEN

Open Data for Machine Learning

Algorithms are free, data is proprietary

Open Data In Machine Learning

- Lack of large data sets for effective machine learning, especially deep neural networks.
 - Deep learning was ready in 80's but success came only in this decade.
 - Advancement of ML research
 - Great progress in last 5 years with the availability of open data.
 - Check your smartphones!!
 - Democratization of machine learning - Google v/s a small company.
 - Remember, data is proprietary. Only distinguishing factor, other than talent (?)
 - Education
 - Teaching practical machine learning requires real-world open data.
-

Open Data In Machine Learning

- Accessibility
 - Everyone a ML researcher or practitioner.
 - Reproducibility and reusability of ML
 - Can I verify your claims ?
 - Benchmarking models
 - Trained models re-use & *transfer learning*.
 - ML based companies
 - Require data sets to train models.
 - Lot of work for ML in legal, finance, socio-economics, healthcare etc.
-

Traditional Data Sets

- University of California at Irvine (UCI) repository



- 440 data sets in a nice searchable UI for empirical evaluation of ML algorithms.
 - Created in 1987, widely used by students, educators, and researchers all over the world.
 - Top 100 most cited papers in all of computer science, > 1000 citations.
 - Iris species data, Cars Evaluation, Ozone Level, Wine Quality, Human activity Recognition using smartphones etc.
 - New datasets are added regularly.
 - Datasets are included in ML software : *sklearn*, *R*
-

Traditional Data Sets

- UCI and similar data are very small (~1000 examples).
 - Not very representative of the real world.
 - Not much data on real-world applications e.g. images, speech, text, stock prices etc.
 - ML for products require a large amount of real-world data.
 - Proper benchmarks for ML models.
 - Test of scalability of ML.
 - Also, Deep learning!!!!
-

Making data available leads to great innovation in both
research and product.

Netflix Challenge

- A million dollars to improve Netflix's recommendation by 10%.
 - Based on previous user ratings, predict user ratings of new movies.
 - Data:
 - `<user_id, movie_id, timestamp, rating>`
 - 100,480,507 ratings that 480,189 users gave to 17,770 movies
 - Training data: 99,072,112 ratings
 - Probe set: 1,408,395 ratings
 - Qualifying set: 2,817,131 ratings split into test and quiz test.
 - Improve the RMSE by 10%.
 - Ran for 3 years from 2006 to 2009.
-

Netflix Prize



[Home](#) |
 [Rules](#) |
 [Leaderboard](#) |
 [Update](#)

Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Researchers usually have the luxury of choosing datasets, and of having more information about that data. In the Netflix contest, the contestants were forced to apply all algorithms to the same set of frustratingly uneven real-world data.

“Because people had to use a fixed dataset, they needed to deal not only with the advantages of a particular method, but also the weaknesses of it, you could not escape it.”

– Ces Bertini, The Ensemble

Netflix Challenge

- Invaluable contribution to the research community.
 - Changed the field of machine learning by reviving interest in real-world problems. Revolutionized recommendation systems research and production.
 - Many great techniques were developed for challenge e.g. collaborative filtering by kNN, SVD, neural network etc.
 - Model ensembles.
 - Collaborative research.
 - Benchmark for recommendation systems even today.
 - Hugely changed the recsys conference.
-

Baseline Models

- Overall average rating for each movie on the training data (with the probe subset removed) to make the predictions
 - RMSE on training set -1.0104
 - RMSE of probe set 1.0528
 - RMSE on quiz set 1.0540
 - Analysis of Variance (ANOVA) techniques :
 - Rating for a user to be the average rating of that user.
 - Rating of a movie to be the average rating of that movie.
 - Two-factor model where rating is made by combining both the above approaches.
-

KNN

- Rating for a user is given by the average rating of the users 'similar' to it.
 - Rating for an item is given by the average rating of the items 'similar' to it.
 - Collaborative filtering.
 - Notion of Similarity - Jaccard index, cosine etc.
 - Improvements over the basic idea. Using global weights, regularisation, combining with ANOVA models.
-

SVD

- If A is a $m \times n$ matrix, the Singular Value Decomposition (SVD) is defined as the matrix factorization:

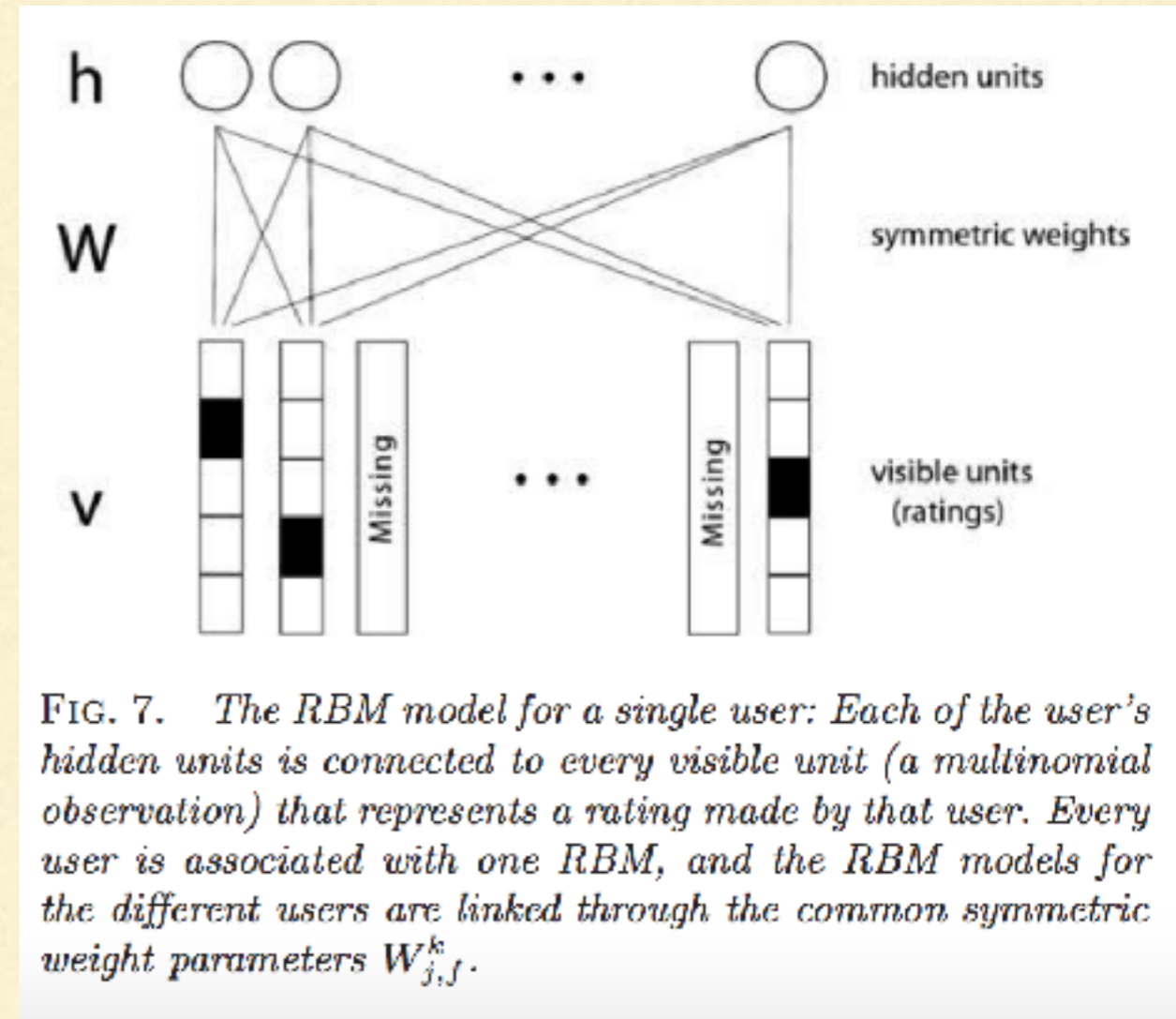
$$A = UDV'$$

- Where U and V are such that their columns are orthonormal basis for AA' . D is a diagonal matrix, entries equal to square root of eigenvalues of AA' , called Singular Values.
 - Given an SVD of A , the *Eckart–Young Theorem* states that, for a given $k < \min(m, n)$, the best rank k reconstruction of A , in the sense of minimizing the *Frobenius norm* of the difference, is $U_k D_k V_k'$, where U_k is the $m \times k$ matrix formed from the first k columns of U , V_k is the $n \times k$ matrix formed by the first k columns of V , and D_k is the upper left $k \times k$ block of D .
 - Such a reconstruction is a good approximation of the matrix.
-

-
- If A is a matrix containing some data e.g. rows are data vectors, one can use SVD to project the vectors onto a low dimensional space. PCA, dimensionality reduction.
 - Taking A to be the user-item interaction matrix, each row corresponds to a user's ratings for all the movies (columns). Such a matrix is very sparse, since a user has rated only a very small fraction of movies.
 - SVD reconstruction provides a way to get dense representation of A .
 - Filled entries can be used as recommendation for unseen movies.
 - Many variations, improvements over the standard SVD.
-

RBM

- Restricted Boltzman Machines (RBM) used by Salakhutdinov, Mnih and Hinton for Netflix recommendation. ML@Toronto team.
- A neural network consisting of one layer of visible units, and one layer of invisible ones; there are no connections between units within either of these layers, but all units of one layer are connected to all units of the other layer.
- While the Netflix qualifying data omits ratings, it does provide implicit information in the form of which movies users chose to rate; this is particularly useful for users having only a small number of ratings in the training set.
- RBM models can incorporate such implicit information in a relatively straightforward way. This is a key strength of RBM models



Winning Ideas

- Ensemble of SVD, kNN, RBM, ANOVA.
 - Incorporate temporal information. Modifications SVD \longrightarrow SVD++.
 - The large number (often millions) of parameters in these models make them prone to overfitting, affecting the accuracy of the prediction process. Ideas like parameters shrinking, regularization.
 - Collaboration of three teams - BellKor, Big Chaos, Pragmatic theory.
-

NETFLIX CHALLENGE

TABLE 1
RMSE values attained by various methods

Predictive model	RMSE	Remarks and references
$\hat{r}_{i,j} = \mu$	1.1296	RMSE on probe set, using mean of training set
$\hat{r}_{i,j} = \alpha_i$	1.0688	Predict by user's training mean, on probe set
$\hat{r}_{i,j} = \beta_j$	1.0528	Predict by movie's training mean, on probe set
$\hat{r}_{i,j} = \mu + \alpha_i + \beta_j$, naive	0.9945	Two-way ANOVA, no interaction
$\hat{r}_{i,j} = \mu + \alpha_i + \beta_j$ "Global effects"	0.9841	Two-way ANOVA, no interaction
Cinematch, on quiz set	0.9657	Bell and Koren (2007a, 2007b, 2007c)
Cinematch, on test set	0.9514	As reported by Netflix
kNN	0.9525	Target is to beat this by 10%
"Global" + SVD	0.9174	Bell and Koren (2007a, 2007b, 2007c)
SVD	0.9167	Bell and Koren (2007a, 2007b, 2007c), on probe set
"Global" + SVD + "joint kNN"	0.9167	Bell and Koren (2007a, 2007b, 2007c), on probe set
"Global" + SVD + "joint kNN"	0.9071	Bell and Koren (2007a, 2007b, 2007c), on quiz set
Simon Funk	0.8982	Bell and Koren (2007a, 2007b, 2007c), on quiz set
TemporalDynamics + SVD++	0.8914	An early submission; Leaderboard
Arkadiusz Paterek's best score	0.8799	Koren (2009)
ML Team: RBM + SVD	0.8789	An ensemble of many methods; Leaderboard
Gravity's best score	0.8787	See Section 6; Leaderboard
Progress Prize, 2007, quiz	0.8743	November 2007; Leaderboard
Progress Prize, 2007, test	0.8712	Bell, Koren and Volinsky (2007a, 2007b, 2007c)
Progress Prize, 2008, quiz	0.8723	As above, but on the test set
Progress Prize, 2008, test	0.8616	Bell, Koren and Volinsky (2008), Toscher and Jahrer (2008)
Grand Prize, target	0.8627	As above, but on the test set
Grand Prize, runner up	0.8572	10 % below Cinematch's RMSE on test set
Grand Prize, runner up	0.8553	The Ensemble, 20 minutes too late; on quiz set
Grand Prize, runner up	0.8567	As above, but on the test set
Grand Prize, winner	0.8554	BellKor + BigChaos + PragmaticTheory, on quiz set
Grand Prize, winner	0.8567	As above, but on the test set

Selected RMSE values, compiled from various sources. Except as noted, RMSE values shown are either for the probe set after fitting on the training data with the probe set held out, or for the quiz set (typically from the Netflix Leaderboard) after fitting on the training data with the probe set included.

Impacts Of Netflix

- The Netflix challenge was unusual for the breadth of the statistical problems it raised and illustrated, and for how closely those problems lie at the frontiers of recent research.
 - Still a benchmark for recommendation system research. — recsys conference.
 - Large scale recommendation for real-world.
 - How to deal with non-convexity of the optimisation problem ?
 - Can SVD be extended to the non-convex regime ?
 - Can better algorithms be devised for fitting RBM models, for having them converge to global optima, and for deciding on early stopping for regularization purposes?
 - Require deeper understanding of issues and methods surrounding penalization, shrinkage and regularization.
 - General questions about bagging, boosting and ensemble methods, as well as of the trade-offs between model complexity and prediction accuracy.
-

Other Recommendations

- Movielens (<https://grouplens.org/datasets/movielens/>)
 - Million song (<https://labrosa.ee.columbia.edu/millionsong/>)
 - Amazon product data (<http://jmcauley.ucsd.edu/data/amazon/links.html>)
 - last.fm music recommendation (<http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html>)
 - Book recommendation (<http://www2.informatik.uni-freiburg.de/~cziegler/BX/>)
 - Food rating dataset (<http://archive.ics.uci.edu/ml/datasets/Entree+Chicago+Recommendation+Data>)
-

Deep Learning

- Deep learning has seen tremendous success in past few years with State-of-The-Art performance in many real-world applications.
 - All examples mentioned earlier are dominated by deep learning models.
 - But training a decent deep learning model requires large amount of data.
 - Success of deep learning can be attributed to development of great computation resources (e.g. nvidia, distributed systems) and availability of huge amount of data (Google).
 - Deep learning tools existed, but data was missing to actualise the full potential.
 - Only big companies have data large enough for deep learning.
 - *Again: Algorithms open, data proprietary.*
-

Open Data for Images

MNIST

- One of the most popular dataset for image analysis. Contains hand-written digits for recognition.
 - 60k training examples and 10k in test set.
 - Early success of neural networks.
 - Established efficacy of Convolution neural networks (ConvNets) in image recognition.
 - LeNet-5 : Deep ConvNet by Yan LeCun in 1998. Used by several banks to reconginze hand-written numbers of checks.
 - Many new deep ConvNet architectures have been proposed to improve performance on this dataset.
 - SOTA : Dynamic Routing Between Capsules (Sabour, Frost, Hinton) Nov 2017
-

IMAGENET



The data that transformed AI research—and possibly the world

IMAGENET

- Data that changed everything. Big time!!
 - Fei Fei Li and team in 2009.
 - Previous datasets didn't capture the variability of the real world.
 - Even identifying pictures of cats was infinitely complex.
 - **WordNet:** a hierarchal structure for the English language. Like a dictionary, but words would be shown in relation to other words rather than alphabetical order.
 - **Idea:** WordNet could have an image associated with each of the words, more as a reference rather than a computer vision dataset.
-



- S: (n) Eskimo dog, husky (breed of heavy-coated Arctic sled dog)
 - direct hypernym / inherited hypernym / sister term
 - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
 - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
 - S: (n) canine, canid (any of various furred mammals with nonretractile claws and typically long muzzles)
 - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) "terrestrial carnivores have four or five clawed digits on each limb"
 - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
 - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
 - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
 - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
 - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
 - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
 - S: (n) living thing, animate thing (a living (or once living) entity)
 - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
 - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
 - S: (n) physical entity (an entity that has physical existence)
 - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

The ImageNet hierarchy derived from WordNet.

cf: http://www.image-net.org/papers/imagenet_cvpr09.pdf

IMAGENET

- WordNet contains approximately 100,000 phrases and ImageNet has provided around 1000 images on average to illustrate each phrase.
 - It consisted of 3.2 million labelled images, separated into 5,247 categories, sorted into 12 subtrees like “mammal,” “vehicle,” and “furniture.”
 - Originally published as a poster in CVPR, with our attracting much fanfare.
 - “There were comments like ‘If you can’t even do one object well, why would you do thousands, or tens of thousands of objects?’” - Jia Deng, co-creator of Imagenet.
-

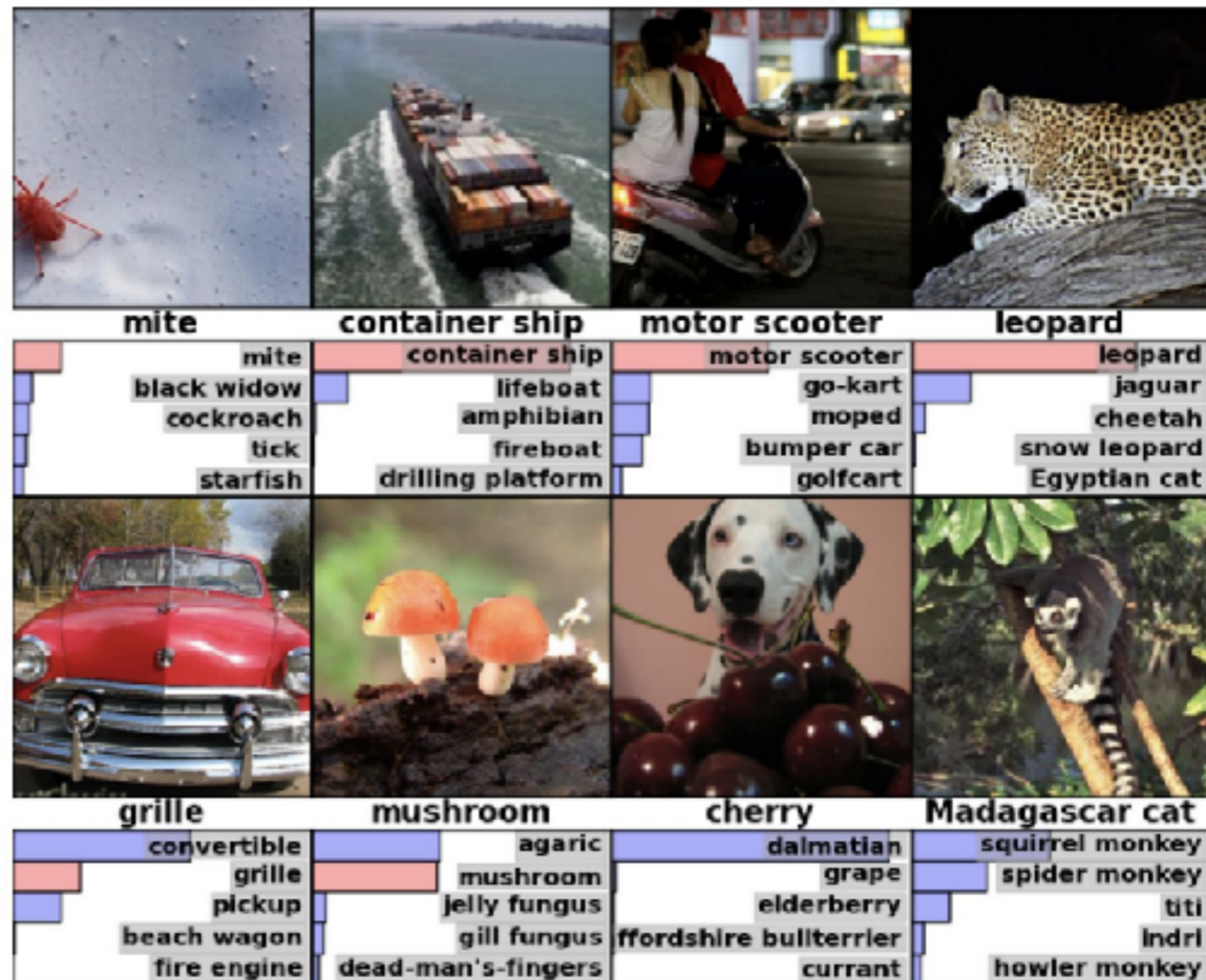
If data is the new oil, it was still dinosaur bones in 2009.

IMAGENET Competition

ImageNet Challenge

IMAGENET

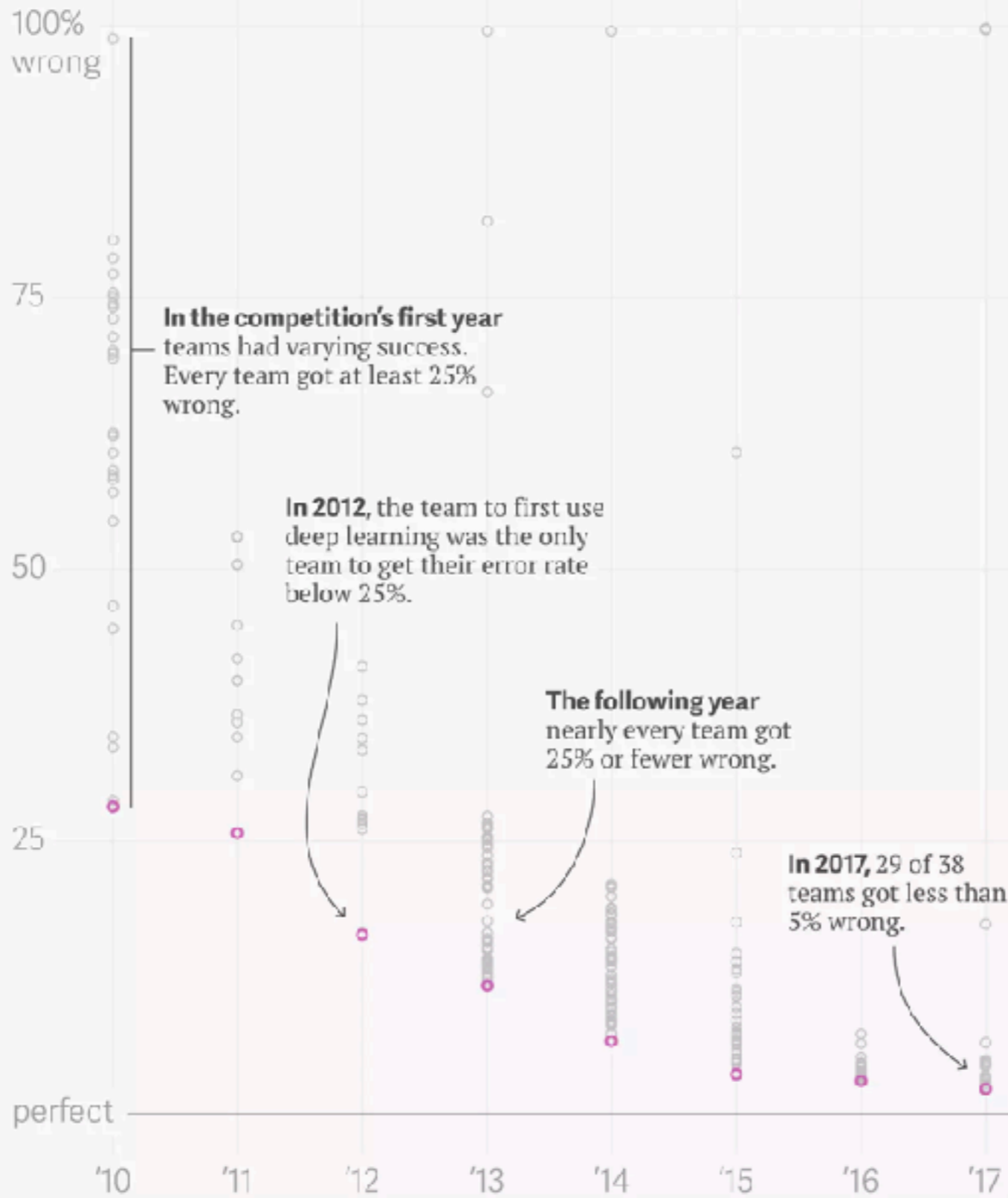
- 1,000 object classes (categories).
- Images:
 - 1.2 M train
 - 100k test.



IMAGENET Competition

- The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) based on the data in Imagenet opened in 2010.
 - Soon became a benchmark for how well image classification algorithms fared against the most complex visual dataset assembled at the time.
 - Algorithms performed better when trained on Imagenet.
 - Competition ran for 8 years.
 - In 2012, the deep neural network submitted by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton performed 41% better than the next best competitor, demonstrating that deep learning was a viable strategy for machine learning
 - Accuracy went from 25% to < 5% through the course of ILSVRC.
-

ImageNet Large Scale Visual Recognition Challenge results



IMAGENET

- If the deep learning boom we see today could be attributed to a single event, it would be the announcement of the 2012 ImageNet challenge results.
 - Deep learning research exploded.
 - Imagenet went from a poster on CVPR to benchmark of most of the presented papers today.
 - “It was so clear that if you do a really good on ImageNet, you could solve image recognition,” - Ilya Sutskever
 - Without Imagenet, the deep learning revolution would have been delayed.
 - After LeNet-5 for reading handwritten cheques, deep ConvNets (and Hinton?) needed a much bigger data to be useful in the real world.
-

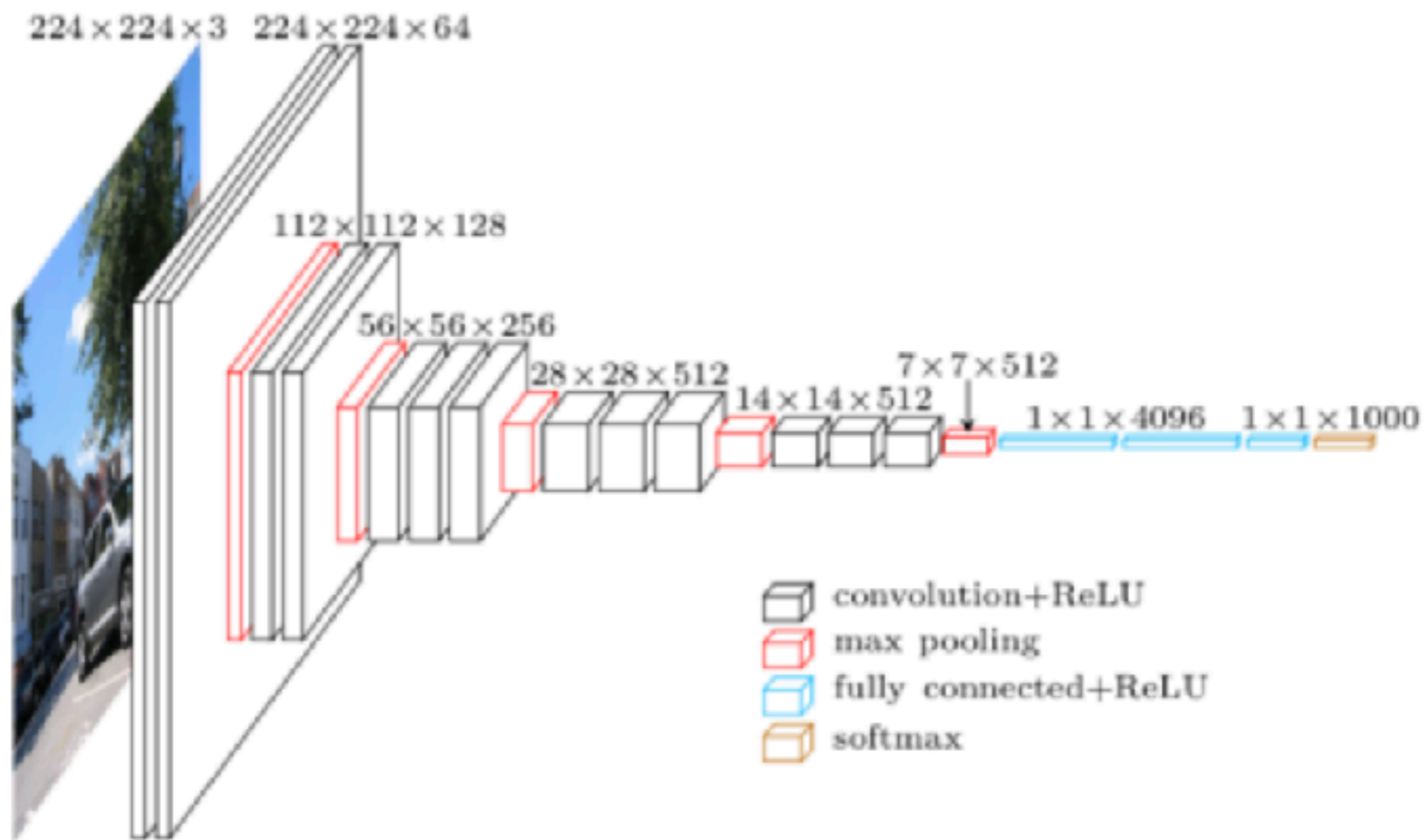
Impacts of IMAGENET

AlexNet

- ILSVRC 2012.
 - First successful use of deep convnet for large scale image classification. Possible because of large amounts of labelled data from ImageNet as well as computations on 2 GPUs.
 - **ReLU** non-linearity activation functions, finding that they performed better and decreased training time relative to the *tanh* function. The ReLU non-linearity now tends to be the default activation function for deep networks.
 - **Data augmentation** techniques that consisted of image translations, horizontal reflections, and mean subtraction. They techniques are very widely used today for many computer vision tasks.
 - **Dropout** layers in order to combat the problem of overfitting to the training data.
 - Proposed style of having **successive convolution** and **pooling layers**, followed by **fully-connected** layers at the end is still the basis of many state-of-the-art networks today.
-

Clarifai

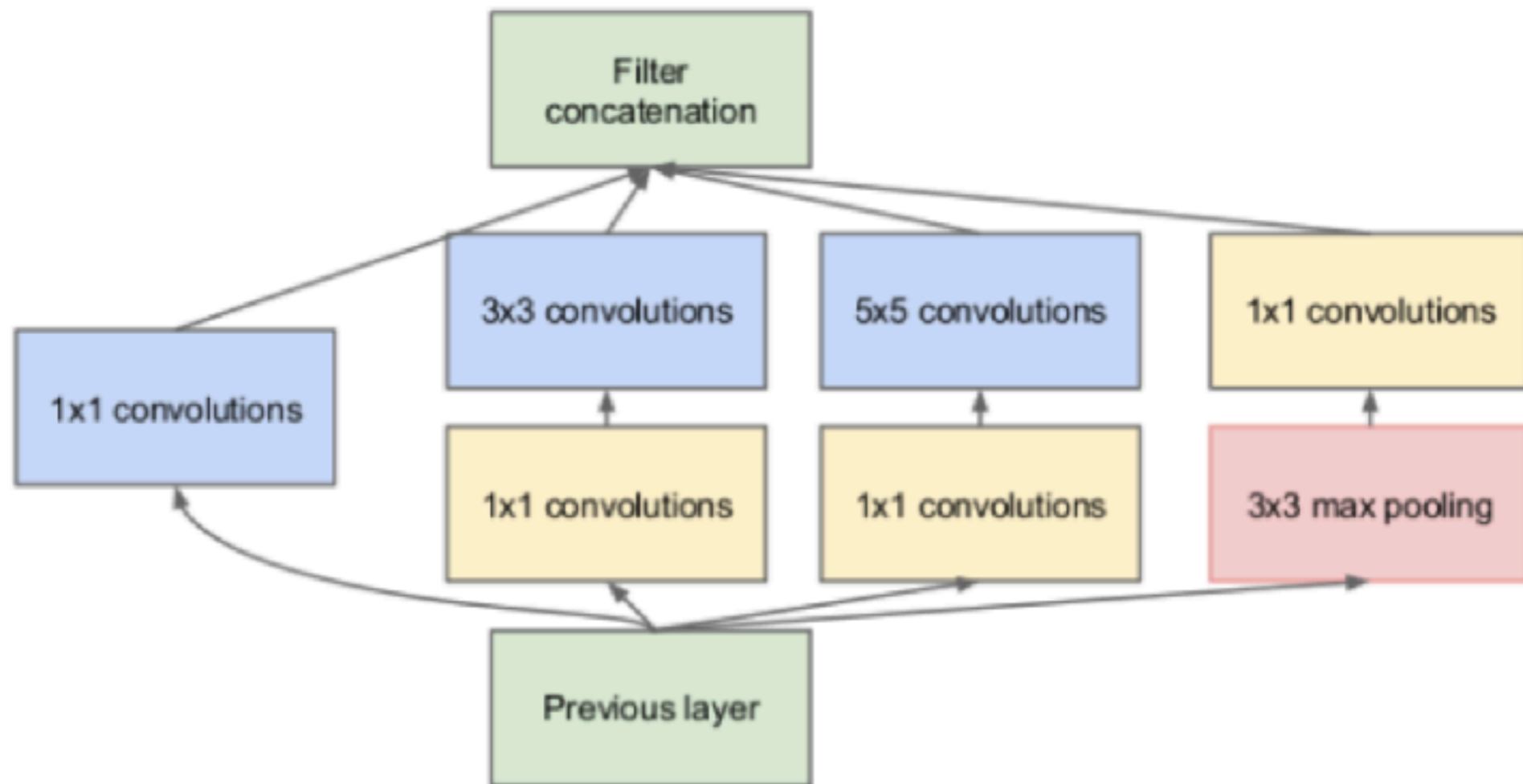
- ILSVRC 2013.
 - Matthew Zeiler, a PhD student at NYU in 2013 and Rob Fergus won the 2013 competition.
 - Matthew Zeiler built Clarifai based off his 2013 ImageNet win, and is now backed by \$40 million in VC funding.
 - “widely seen as one of the most promising [startups] in the crowded, buzzy field of machine learning.” (**Forbes**)
 - Also, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks by Yan LeCun and team at NYU.
-



VGGNet Architecture

VGG

- Visual Geometry Group @ Oxford, UK.
 - **Idea:** You didn't really need any fancy tricks to get high accuracy. Just a deep network with lots of small 3x3 convolutions and non-linearities will do the trick!
 - Two successive 3x3 convolutions has the equivalent receptive field i.e. the pixels it sees as a single 5x5 filter, and 3 3x3 filter ~ a 7x7 filter. Same stimulation of pixels with added benefits of smaller filters.
 - Decrease in the number of parameters.
 - Using a ReLU function in-between each convolution introduces more non-linearity into the network which makes modeling decision function better.
 - As the spatial size of the input volumes at each layer decrease (as a result of the pooling layers), the depth of the volumes increase since need more discriminative features to use for accurate classification.
 - New kind of data augmentation: scale jittering.
-

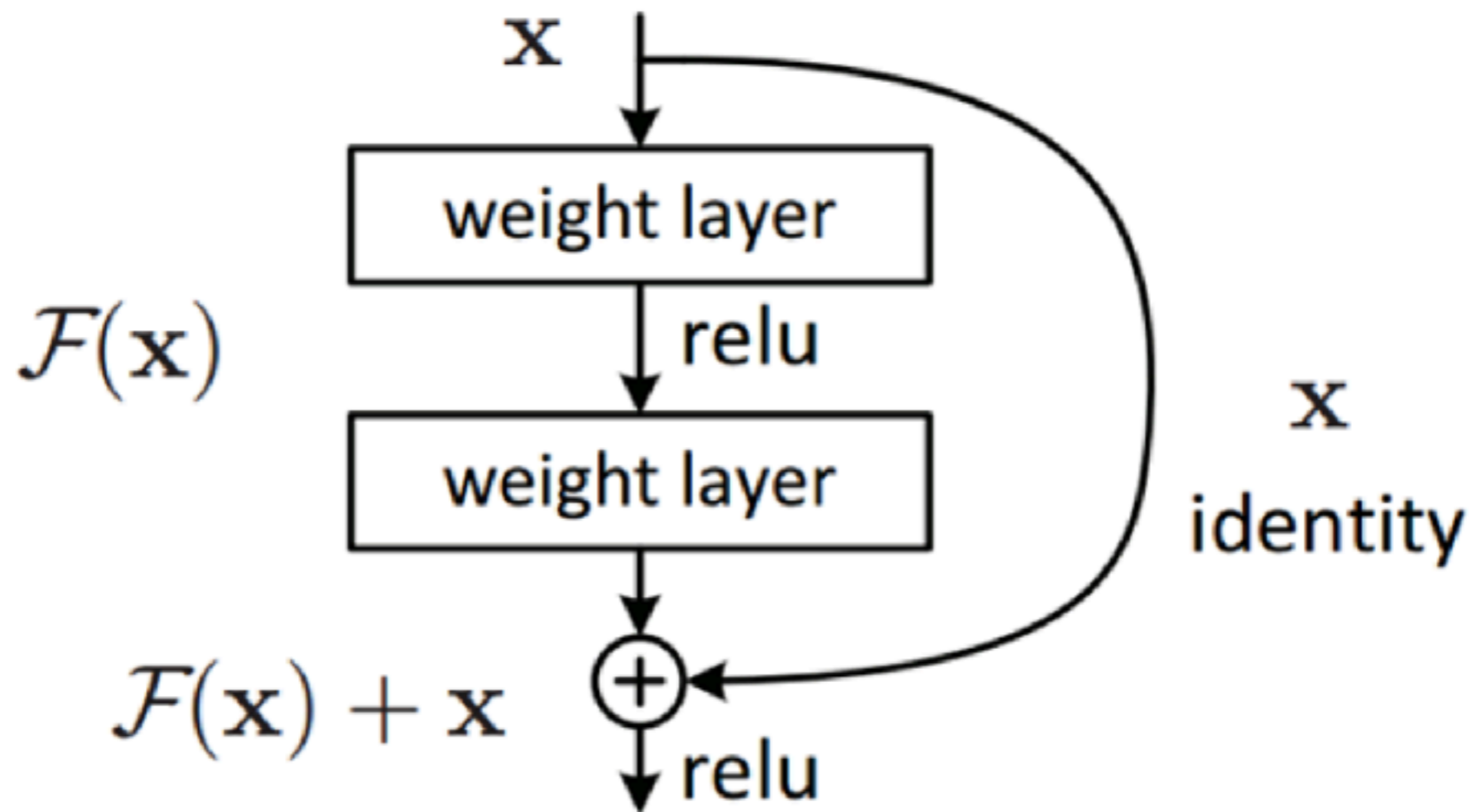


Inception Module from GoogLeNet

GoogleLeNet and Inception

- First to really address the issue of computational resources along with multi-scale processing.
 - Through the use of 1×1 convolutions before each 3×3 and 5×5 , the inception module reduces the number of feature maps passed through each layer, thus reducing computations and memory consumption.
 - The inception module has 1×1 , 3×3 , and 5×5 convolutions all in parallel. The idea behind this was to let the network decide, through training what information would be learned and used.
 - GoogLeNet was one of the first models that introduced the idea that CNN layers didn't always have to be stacked up sequentially. The authors of the paper showed that you can also increase network width for better performance and not just depth.
-

Skipping over with a shortcut: ResNet



A residual block

Residual block from ResNet

ResNet

- The ResNet architecture was the first to pass human level performance on ImageNet.
 - Main contribution of *residual learning* is often used by default in many state-of-the-art networks today.
 - A naive stacking of layers to make the network very deep won't always help and can actually make things worse.
 - The idea is that by using an additive skip connection as a shortcut, deep layers have direct access to features from previous layers. This allows feature information to more easily be propagated through the network. It also helps with training as the gradients can also more efficiently be back-propagated.
 - The first “*ultra deep*” network, where it is common to use over 100–200 layers.
-

Impacts of IMAGENET

- **Transfer learning:** researchers soon realized that the weights learned in state of the art models for ImageNet could be used to initialize models for completely other datasets and improve performance significantly.
 - Achieving good performance with as little as one positive example per category.
 - Pre-trained ImageNet models have been used to achieve state-of-the-art results in tasks such as
 - object detection
 - semantic segmentation
 - human pose estimation
 - video recognition.
 - Applications in domains where the number of training examples is small and annotation is expensive. (e.g. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition)
-

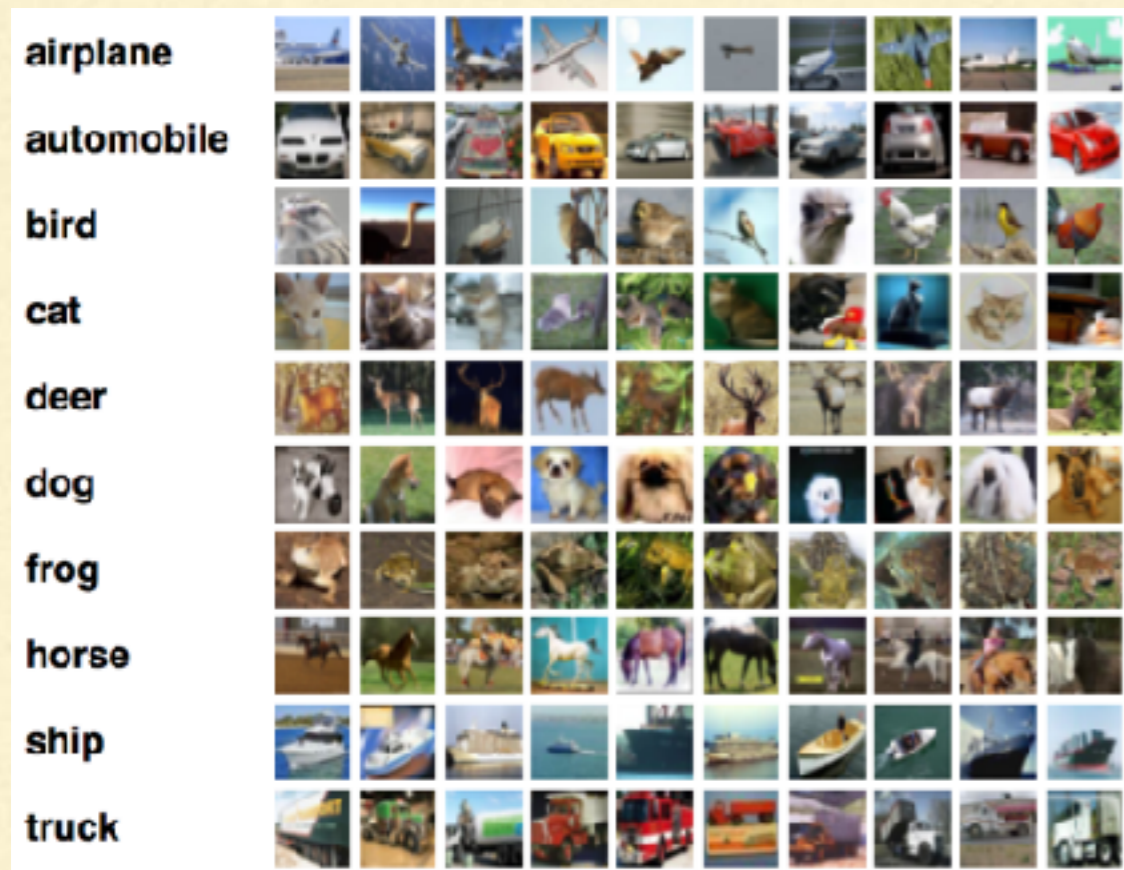
“One thing ImageNet changed in the field of AI is suddenly people realized the thankless work of making a dataset was at the core of AI research. People really recognize the importance the dataset is front and center in the research as much as algorithms.”

– Fei Fei Li, Creator of ImageNet and Chief Scientist Google Cloud.

Beyond IMAGENET

- Google released the Open Images Database, containing 9 million images in 6000 categories.
 - YouTube-8M Dataset - to accelerate research on large-scale video understanding, representation learning, noisy data modeling, transfer learning, and domain adaptation approaches for video.
 - Visual Genome - A knowledge database to connect structured image concepts to language.
 - Fine-tuned ImageNet models are pushing the state of the art in many traditional image datasets like CIFAR, PASCAL etc.
-

CIFAR



- For image classification. Beyond digits.
 - 60k color images with 10 classes.
 - Piggy-banking on the success of IMAGENET and deep convnets.
 - SOTA : ShakeDrop regularization 2018
-

| Method | Reg | Cos | Fil | Depth | #Param | CIFAR
-10 (%) | CIFAR
-100 (%) |
|---|-----|-----|-----|-------|--------|------------------|-------------------|
| Coupled Ensemble
(Dutt et al., 2017) | | | | 118 | 25.7M | *2.99 | *16.18 |
| | | | | 106 | 25.1M | *2.99 | *15.68 |
| | | | | 76 | 24.6M | *2.92 | *15.76 |
| | | | | 64 | 24.9M | *3.13 | *15.95 |
| | | | | - | 50M | *2.72 | *15.13 |
| | | | | - | 75M | *2.68 | *15.04 |
| | | | | - | 100M | *2.73 | *15.05 |
| ResNeXt
(Xie et al., 2017) | | ✓ | | 26 | 26.2M | +3.58 | - |
| | | | | 29 | 34.4M | - | +16.34 |
| ResNeXt + Shake-Shake
(Gastaldi, 2017) | SS | ✓ | | 26 | 26.2M | *2.86 | - |
| | | | | 29 | 34.4M | - | *15.85 |
| ResNeXt + Shake-Shake + Cutout
(DeVries & Taylor, 2017b) | SS | ✓ | CO | 26 | 26.2M | *2.56 | - |
| | | | | 29 | 34.4M | - | *15.20 |
| PyramidNet
(Han et al., 2017b) | | | | 272 | 26.0M | *3.31 | *16.35 |
| | | ✓ | RE | 272 | 26.0M | 3.42 | 16.66 |
| PyramidDrop
(Yamada et al., 2016) | RD | | | 272 | 26.0M | 3.83 | 15.94 |
| | RD | ✓ | RE | 272 | 26.0M | 2.91 | 15.48 |
| PyramidNet + ShakeDrop
(Proposed) | SD | | | 272 | 26.0M | 3.41 | 14.90 |
| | SD | | RE | 272 | 26.0M | 2.89 | 13.85 |
| | SD | ✓ | | 272 | 26.0M | 2.67 | 13.99 |
| | SD | ✓ | RE | 272 | 26.0M | 2.31 | 12.19 |



1

Upload photo

The first picture defines the scene you would like to have painted.



2

Choose style

Choose among predefined styles or upload your own style image.



3

Submit

Our servers paint the image for you. You get an email when it's done.



Open Data for Text and NLP

World Embeddings

- Word embeddings are a representation of words in a natural language as vectors in a continuous vector space where semantically similar words are mapped to nearby points.
 - Assumption: Words that appear in the same context are semantic closer than the words which do not share same context.
 - Essentially, we *embed* words in a vector space. And the weight of the word is distributed across many dimensions which capture the semantic properties of the words.
 - Train a neural network on a large corpus of text data e.g. wikipedia dump.
-

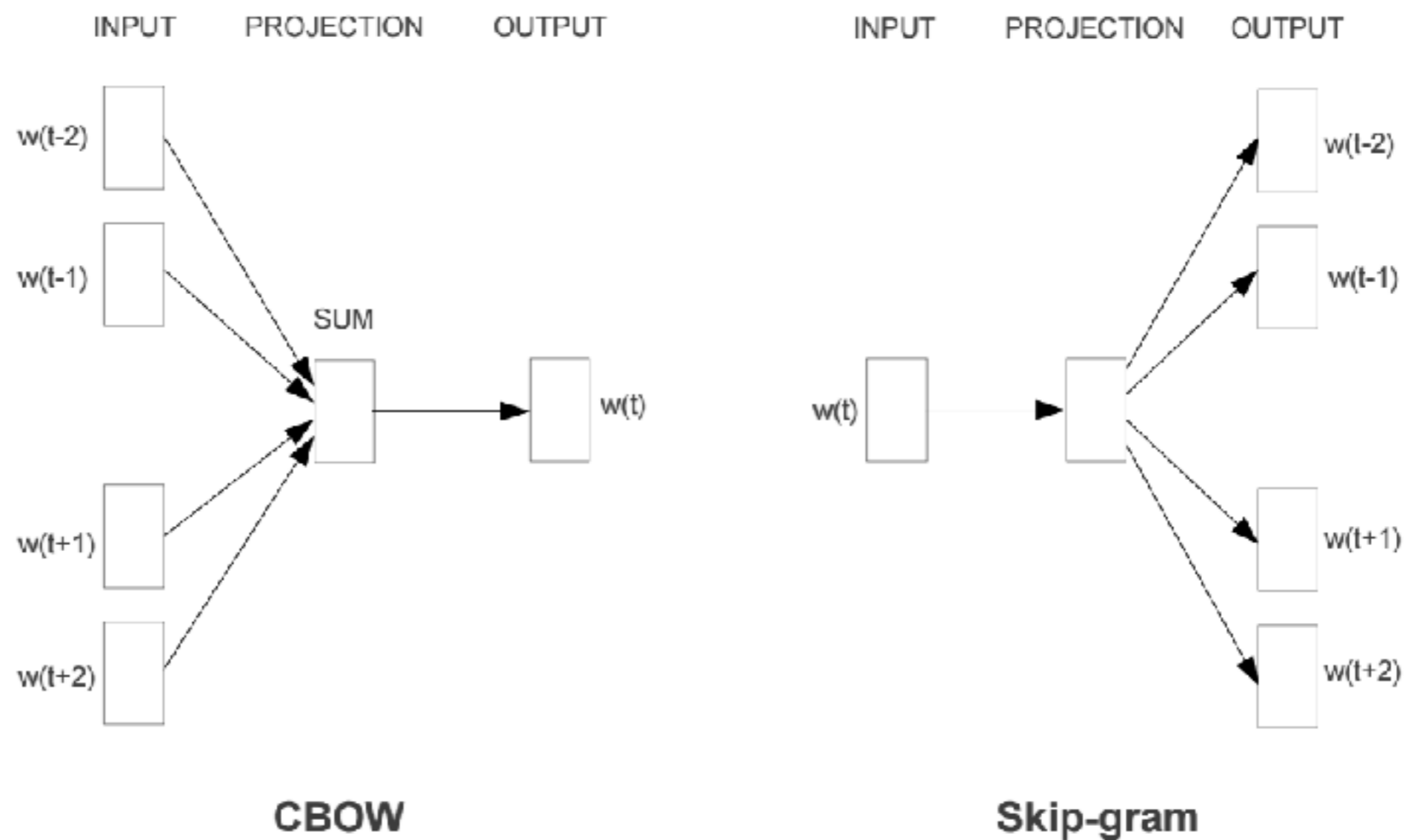


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Open word embeddings

- Word2Vec is an implementation of word embeddings by researchers at Google Brain in 2013.
 - Google released 300 dimensional word-vectors for 3M words and phrases, trained on Google News Data that has 300B tokens.
 - Another open implementation is GloVe: Global Vectors for Word Representation.
 - Trained on wikipedia dump, common crawl, twitter.
 - 50, 100, 300 dimensional vectors.
 - Facebook FastText
 - Word vectors for 157 languages trained on Wikipedia and Crawl
 - The words and phrases in these public data are fairly representative of the the language (e.g. English)
-

A Somewhat surprisingly, it was found that similarity of word representations goes beyond simple syntactic regularities. Using a word offset technique where simple algebraic operations are performed on the word vectors, it was shown for example that $\text{vector}(\text{"King"}) - \text{vector}(\text{"Man"}) + \text{vector}(\text{"Woman"})$ results in a vector that is closest to the vector representation of the word Queen.

–Mikolov et al, Google

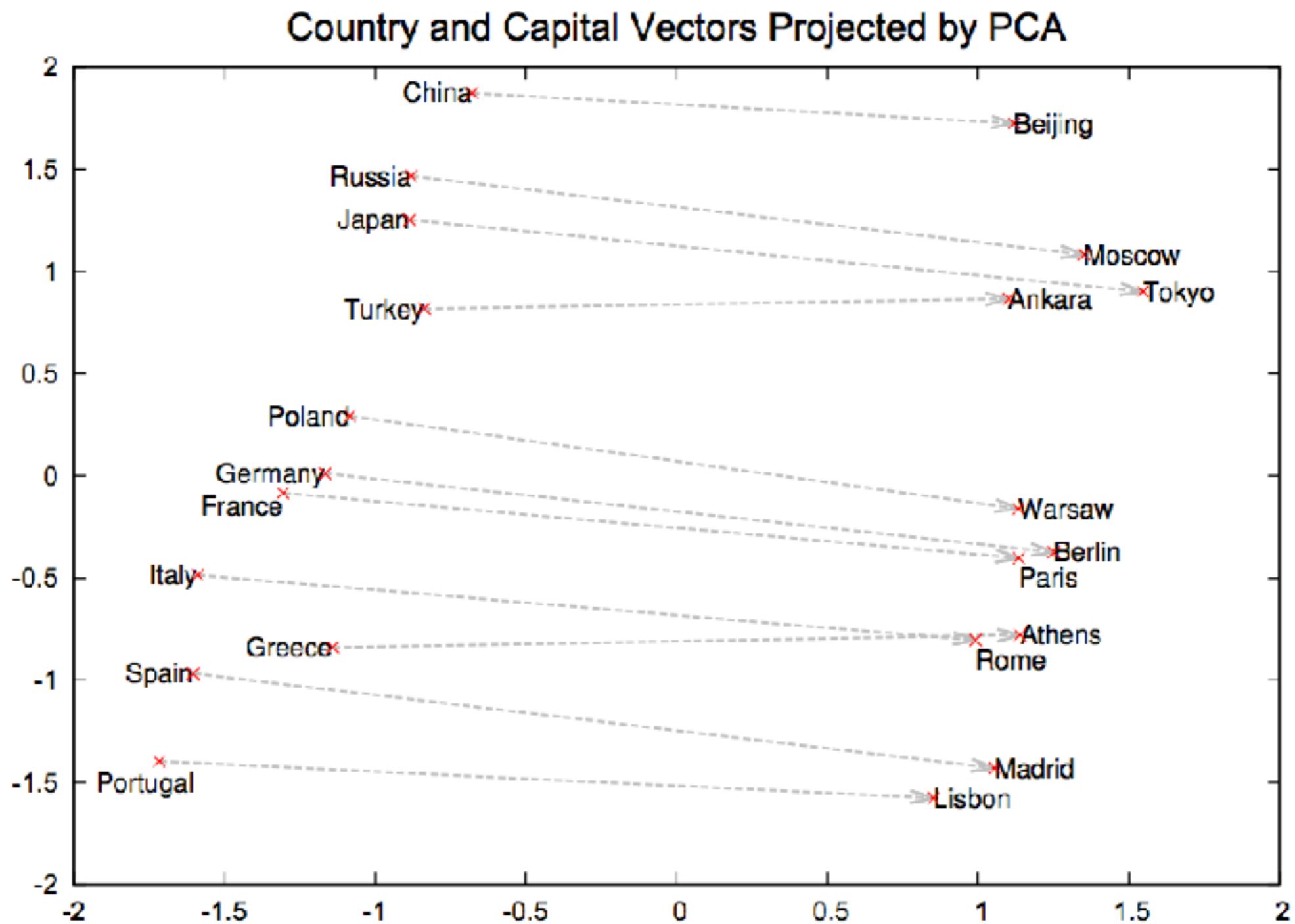


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

| | | | | |
|------------------|-------------------|------------------------|-----------------|----------------------|
| Czech + currency | Vietnam + capital | German + airlines | Russian + river | French + actress |
| koruna | Hanoi | airline Lufthansa | Moscow | Juliette Binoche |
| Check crown | Ho Chi Minh City | carrier Lufthansa | Volga River | Vanessa Paradis |
| Polish zolty | Viet Nam | flag carrier Lufthansa | upriver | Charlotte Gainsbourg |
| CTK | Vietnamese | Lufthansa | Russia | Cecile De |

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

| | | | |
|---------------------|-----------------------|---------------|---------------------|
| Newspapers | | | |
| New York | New York Times | Baltimore | Baltimore Sun |
| San Jose | San Jose Mercury News | Cincinnati | Cincinnati Enquirer |
| NHL Teams | | | |
| Boston | Boston Bruins | Montreal | Montreal Canadiens |
| Phoenix | Phoenix Coyotes | Nashville | Nashville Predators |
| NBA Teams | | | |
| Detroit | Detroit Pistons | Toronto | Toronto Raptors |
| Oakland | Golden State Warriors | Memphis | Memphis Grizzlies |
| Airlines | | | |
| Austria | Austrian Airlines | Spain | Spainair |
| Belgium | Brussels Airlines | Greece | Aegean Airlines |
| Company executives | | | |
| Steve Ballmer | Microsoft | Larry Page | Google |
| Samuel J. Palmisano | IBM | Werner Vogels | Amazon |

Table 2: Examples of the analogical reasoning task for phrases (the full test set has 3218 examples). The goal is to compute the fourth phrase using the first three. Our best model achieved an accuracy of 72% on this dataset.

The distance between similar words is low:

```
dist(vecs[wordidx["puppy"]], vecs[wordidx["dog"]])
```

0.27636240676695256

```
dist(vecs[wordidx["queen"]], vecs[wordidx["princess"]])
```

0.20527545040329642

And the distance between unrelated words is high:

```
dist(vecs[wordidx["celebrity"]], vecs[wordidx["dusty"]])
```

0.98835787578057777

```
dist(vecs[wordidx["kitten"]], vecs[wordidx["airplane"]])
```

0.87298516557634254

Impacts of word2vec

Text Classification

- Text classification is totally revived by wordvecs. Instead of traditional approaches like tf-idf, we use the pre-trained wordvecs as input to a text classification model e.g. SVM or logistic regression.
 - Word embeddings pre-trained on large amounts of unlabeled data via algorithms such as word2vec and GloVe are used to initialize the first layer of a neural network, the rest of which is then trained on data of a particular task. e.g. CNN or LSTM for text classification. (Standard for document classification).
 - *Very Deep Convolutional Networks for Text Classification* - Facebook AI Research 2017. SOTA in many text classification tasks including 20 News Dataset and Amazon review sentiment classification. Input is pre-trained wordvecs.
-

Machine Translation

- An interesting and natural application is in machine translation. Socher et al 2013 show that we can learn to embed words from two different languages in a single, shared space. In this case, we learn to embed English and Mandarin Chinese words in the same space.
 - Lot of work by Facebook Research on machine translation using embeddings without a parallel corpora.
 - In [Attention Is All You Need](#) from Google AI, the Transformer, a novel neural network architecture based on a self-attention mechanism was proposed which is believed to be particularly well suited for language understanding. SOTA on machine translation tasks.
-

Search And Discovery

- Replaced the traditional methods like Latent Semantic Hashing in search engines.
 - Query matching in search engines is facilitated by word embeddings.
 - Classification of pages are relevant or non-relevant for a query.
-

Other Applications

- Word vector based techniques are state of the art in :
 - Language Modeling
 - Named Entity Recognition
 - Sentiment Analysis
 - Topic Modeling
 - Short text (Twitter) classification
 - Text Summarization
-

Text Data Resources

- 20 Newsgroup - News articles belonging to 20 categories like politics, legal, finance, sports etc.
 - IMDB Reviews - 25k highly polar movie reviews for training and another 25k for testing.
 - Yelp Reviews - 5,200,000 reviews, 174,000 business attributes, 200,000 pictures and 11 metropolitan areas.
 - SQAD - **S**tanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles.
 - Billion Words - a standard training and test setup for language modeling experiments.
-

Beyond Word Vectors

- Though pre-trained wordvecs have been immensely influential, they have a major limitation: they only incorporate previous knowledge in the first layer of the model and the rest of the network still needs to be trained from scratch. **UNLIKE IMAGENET.**
 - Initialize a computer vision model with ImageNet params only in the first layer. This can see edges, but fail to capture higher-order features.
 - We need higher-order features for NLP also. Else the models will still require a large amount of data to train. There is no fine-tuning.
 - Core aspect of language understanding requires modeling complex language phenomena such as compositionality, polysemy, anaphora, long-term dependencies, agreement, negation, and many more.
-

ImageNet of NLP ??

IMAGENET OF NLP

- ELMo, ULMFiT, and the OpenAI transformer : demonstrated that pre-trained language models can be used to achieve state-of-the-art results on a wide range of NLP tasks.
 - **Idea:** pre-train a language model on a large corpus of data.
 - Pre-training language model was talked about earlier also, but it remained unclear whether a single pre-trained language model was useful for many tasks.
 - The above methods demonstrated empirically how pre-training a LM performs.
 - All three methods employed pre-trained language models to achieve state-of-the-art on a diverse range of tasks in Natural Language Processing, including text classification, question answering, natural language inference, coreference resolution, sequence labeling, and many others. In some cases, these improvements ranged between 10-20% better than the state-of-the-art on widely studied benchmarks
-

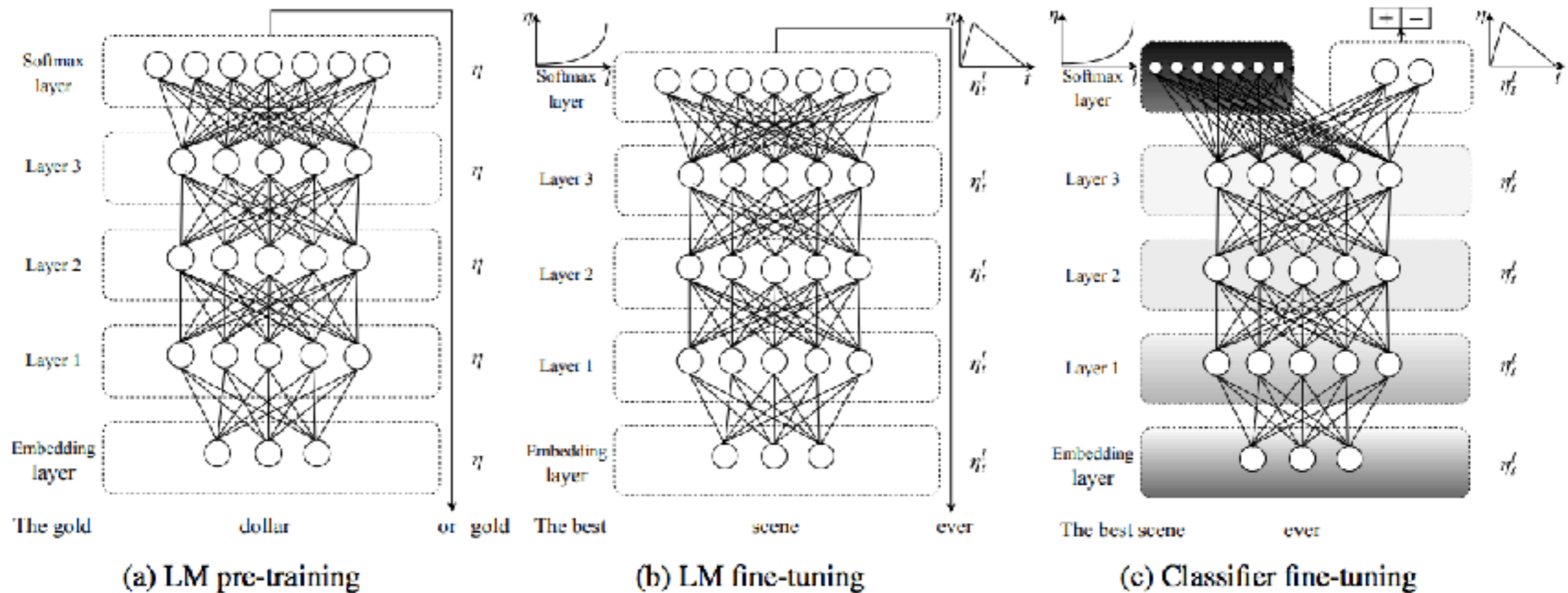


Figure 1: ULMFiT consists of three stages: a) The LM is trained on a general-domain corpus to capture general features of the language in different layers. b) The full LM is fine-tuned on target task data using discriminative fine-tuning (*Discr*) and slanted triangular learning rates (STLR) to learn task-specific features. c) The classifier is fine-tuned on the target task using gradual unfreezing, *Discr*, and STLR to preserve low-level representations and adapt high-level ones (shaded: unfreezing stages; black: frozen).

It is very likely that in a year's time NLP practitioners will download pre-trained language models rather than pre-trained word embeddings.

Thanks
@januverma

Further

- Kaggle
 - Data for healthcare - NIH data (<https://www.analyticsindiamag.com/the-new-nih-dataset-and-ai-may-revolutionise-lesion-detection/>)
-