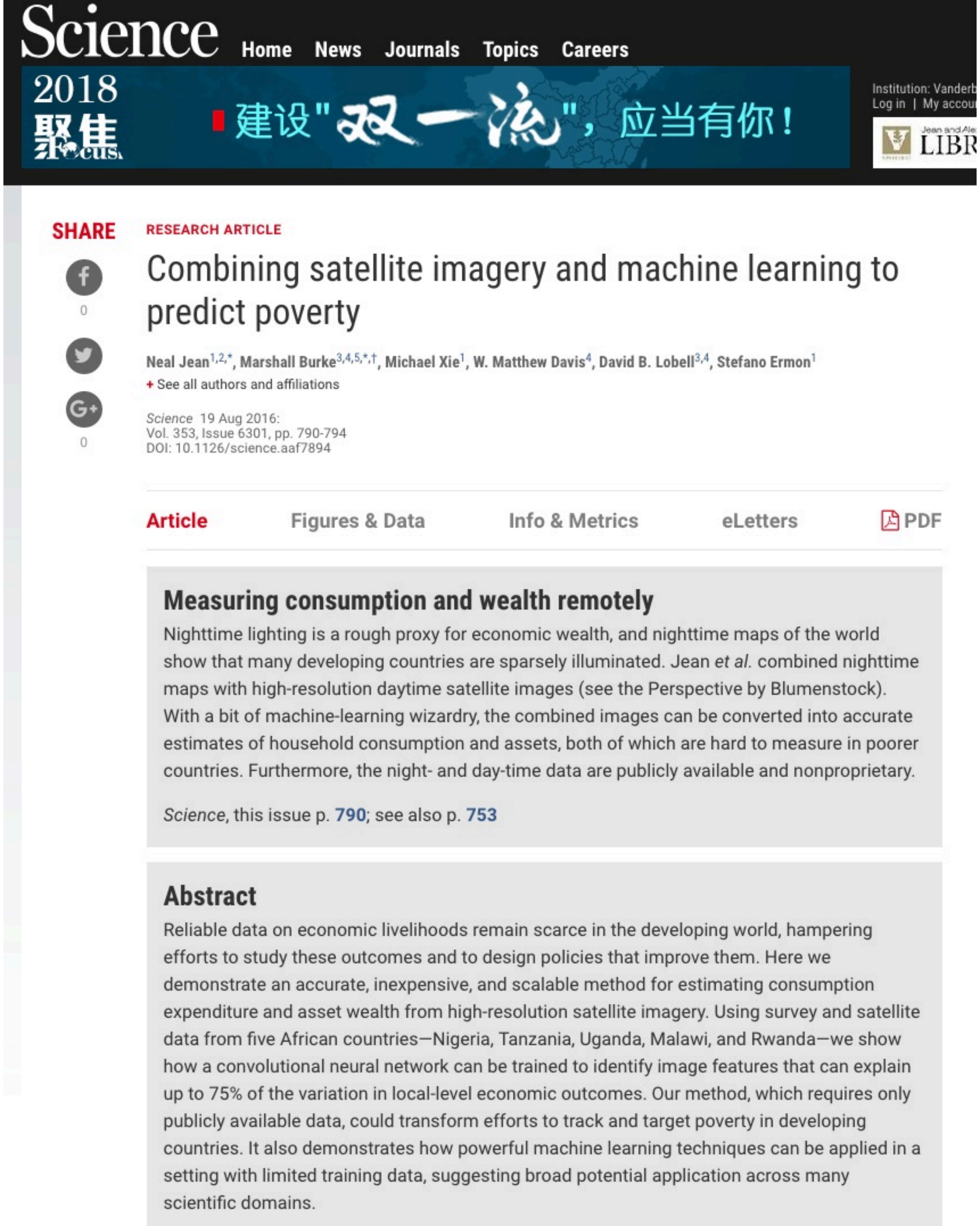
http://science.sciencemag.org/content/353/6301/790



RESEARCH ARTICLE

Combining satellite imagery and machine learning to predict poverty

Neal Jean[1,2,*], Marshall Burke[3,4,5,*,†], Michael Xie[1], W. Matthew Davis[4], David B. Lobell[3,4], Stefano Ermon[1]

+ See all authors and affiliations

Science 19 Aug 2016:
Vol. 353, Issue 6301, pp. 790-794
DOI: 10.1126/science.aaf7894

Article

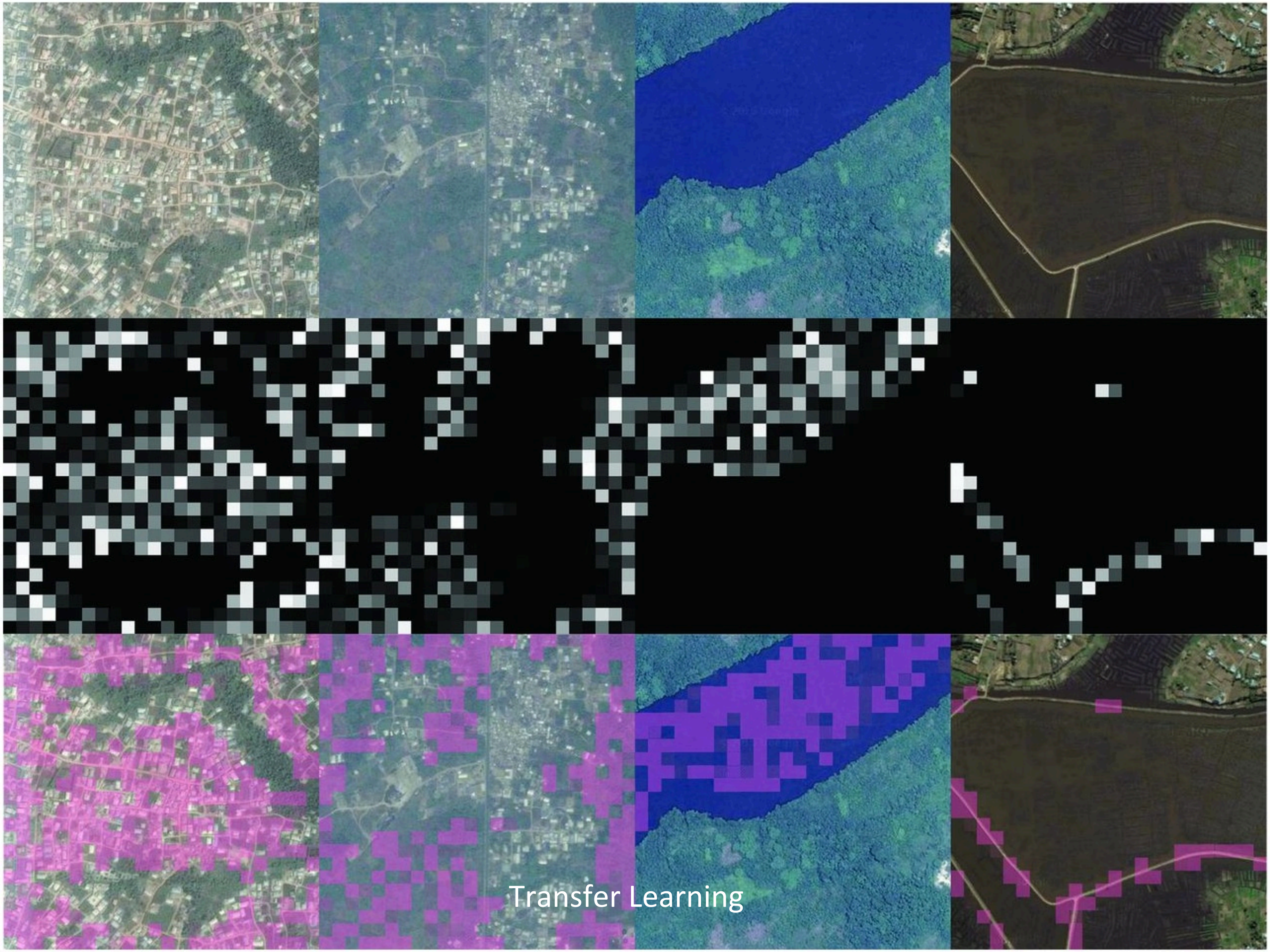
Measuring consumption and wealth remotely

Nighttime lighting is a rough proxy for economic wealth, and nighttime maps of the world show that many developing countries are sparsely illuminated. Jean *et al.* combined nighttime maps with high-resolution daytime satellite images (see the Perspective by Blumenstock). With a bit of machine-learning wizardry, the combined images can be converted into accurate estimates of household consumption and assets, both of which are hard to measure in poorer countries. Furthermore, the night- and day-time data are publicly available and nonproprietary.

Science, this issue p. 790; see also p. 753

Abstract

Reliable data on economic livelihoods remain scarce in the developing world, hampering efforts to study these outcomes and to design policies that improve them. Here we demonstrate an accurate, inexpensive, and scalable method for estimating consumption expenditure and asset wealth from high-resolution satellite imagery. Using survey and satellite data from five African countries—Nigeria, Tanzania, Uganda, Malawi, and Rwanda—we show how a convolutional neural network can be trained to identify image features that can explain up to 75% of the variation in local-level economic outcomes. Our method, which requires only publicly available data, could transform efforts to track and target poverty in developing countries. It also demonstrates how powerful machine learning techniques can be applied in a setting with limited training data, suggesting broad potential application across many scientific domains.



Transfer Learning

IBM Watson - Machine Learning Writ Large

IBM Watson is a cloud-based machine learning platform designed to enable others to develop software and systems which use machine learning, and has spawned myriad applications. The original Watson which defeated former Jeopardy winners uses machine learning to make medical diagnoses. Since then, the "IBM Watson Machine Learning" service has been deployed for several other applications, such as assessing the validity of automobile insurance claims.

I choose this example in particular, as the particular concept of machine-learning as a service allows innovators with limited machine-learning domain knowledge to expand the applications of the field. This evolution in the industry will only serve to increase the ubiquity of machine learning, and allow advancements in Machine Learning theory to affect a large swath of products simultaneously. For example, an IBM Machine Learning researcher's development, once included in the libraries of Watson Machine Learning, can immediately be used by dozens of other firms and products.

Reference: "IBM Watson Machine Learning - IBM Watson and Cloud Platform Learning Center." The DeveloperWorks Blog, 18 July 2016, developer.ibm.com/clouddataservices/docs/ibm-watson-machine-learning/.

Machine learning as a service
Personal data mining

Google Translate as Supervised learning

This article presents Google Translate's fielded machine learning application. It describes the origin of Google Brain, the premier AI and Machine Learning research group at Google, and how it picked up research on machine translation done by researchers. Open research had established that you can represent a language pretty well in a mere thousand or so dimensions. Eventually, Google recognized that neural networks along with the established theoretical foundations suggested that it could overhaul Google translate, which was then using "phrase-based statistical machine translation." (known by that because the time the system gets to the next phrase, it doesn't know what the last one was). Using context-enriching neural networks based on Google-scale data, over the course of an entire year Google engineers were able to develop an at-scale, fast translation service based on super-vised machine learning.

Gideon Lewis-Kraus, "The Great A.I. Awakening", New York Times, Dec. 4, 2016.
https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Using an existing model to bootstrap learning
Learning deviations from the model's prediction

Udacity supervised machine learning course
https://www.youtube.com/watch?v=Ki2iHgKxRBo&list=PLAwxTw4SYaPl0N6-e1GvyLp5-MUMUjOKo

This online course by Udacity is an extensive introduction into the field of supervised learning. It covers a large variety of topics including decision trees, regression, nearest neighbors, support vector machines, boosting, bayes classification, and much more. The course also covers lots of subjects that apply generally to supervised learning and machine learning such as training and test set selection, cross validation, gradient descent and probability. The course is very thorough, with lots of examples, worked problems, and frequent quizzes that test your understanding of the material. In my opinion, the instructors are very clear and easy to understand. I highly recommend going through the videos if you have the time.

Machine learning online courses

Supervised ML for Credit Card Fraud

In the research paper "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study" (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8424696), authors Dhankhad, Mohammed, and Far compare the performance of various supervised machine learning models on detection of credit card fraud. The authors find that the top three performing models are a stacking meta-classifier, random forest, and XGBoost. However, the performance of each model varied slightly based upon which metric was used for evaluation (e.g. recall or precision) and also how balanced the dataset was initially. I was surprised at how marginal some of the improvements were using more advanced machine learning techniques, as simply logistic regression scored .94 on recall and precision while the random forest classifier and stacking meta-classifiers scored only .01 points higher on both of those metrics (.95). I think that this illustrates that the need for more advanced machine learning models may depend on the initial problem, as simple regression models may suffice for certain well-defined problems, and that advances can instead be made more easily by collecting better data or engineering better features.

Baseline algorithms in evaluation of new algorithms
Rapidly diminishing returns
Different ways of building a forest

Supervised/Unsupervised/Semi-Supervised ML
Link: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Description:

This website takes you through the definitions and differences of supervised, unsupervised and semi-supervised machine learning. Supervised learning has two main types which are classification ("this car is red, that care is blue") and regression(attempting to reach a discrete or continuous correct value). Unsupervised learning is split into clustering(grouping data based on similarities) and association(uncovering rules that govern the data). Semi-supervised machine learning is somewhere in between the two, where only some of the data has been labeled. This is where most of machine learning lies in actual practice. This is a very introductory, yet super useful reference on different kinds of machine learning and what they can be useful for.

Combining learning paradigms
Biswas

MarI/O - Machine Learning in Videogames
https://www.youtube.com/watch?v=qv6UVOQ0F44

This video describes an application called MarI/O, which is a program designed to teach a computer how to play Super Mario World. The program uses neural networks and genetic algorithms to construct better paths through the level as the agent learns. Given enough repetitions, the computer can learn a path through any Super Mario World level. The computer literally guesses input combinations at the start - it has no knowledge of the world around it yet. Over time, the computer can learn which combinations can get it the farthest through the level, and modify the best paths to find even better ones. I think it's amazing that we can teach machines to play games designed for humans.

Learning from lots of simulations like Alpha Go

Cancer classification using machine learning to analyze gene expression data
Reference: https://bura.brunel.ac.uk/bitstream/2438/3013/1/TanGilbertNZ2003.pdf

Traditionally, cancer classification has taken place at the phenotypic level, which considers only the morphological appearances of tumors. The algorithm described in the paper attempts to employ ensemble learning (a supervised machine learning algorithm) to perform cancer classification at the genotypic level. Given a gene expression dataset, the algorithm constructs a decision tree using the C4.5 decision tree algorithm. In the resulting decision tree, internal nodes represent the genes within the supplied dataset, branches represent conditions for gene expression, and leaves represent the decision outcome (either 'is tumor tissue' or 'is normal tissue'). I believe this study presents a particularly suitable application of supervised machine learning. Historically, machine learning has performed well in biological fields requiring analysis of enormous data samples to explain complex relationships – in this case, the relationships between the expression of multiple genes. This application of supervised learning requires the algorithm to have little understanding of biological theory and also provides explanations for certain predictions by way of decision trees.

Need for explanations in medicine a motivation for DTs

Google's Crowdsourcing for Data Labeling
URL: https://crowdsource.google.com/
This website lists out the different domains in which Google enlists the Internet community (or "crowdsources") to help label data that is in turn used to improve machine learning algorithms. A great example of this is getting users to provide keywords for various images, which in turn helps improve Google Image Search. Google has gamified this experience for users who spend their time labeling data, thus making an otherwise tedious task seem more enjoyable. While reading various articles about this, I also discovered an interesting post that wondered whether data labeling is the new "blue-collar" job, which I thought was an interesting perspective on this issue.

Crowdsourcing data labeling

OpenAI Dota 2 Bot Beats Professionals

Haridy, Rich. “AI Beats the World's Best Gamers after Just Two Weeks of Learning.” New Atlas - New Technology & Science News, New Atlas, 14 Aug. 2017, newatlas.com/open-ai-dota2-machine-learning/50882/.

OpenAI, a non-profit artificial intelligence research company, has developed a bot that, after only two week of machine learning, is able to beat professional players in a 1v1 duel in the competitive video game, DotA 2. Greg Brockman from OpenAI says that, "The rules of Dota are so complicated that if you just think really hard about how the game works and try to write those rules down, you're not even gonna be able to reach the performance of a reasonable player." Unlike Chess and Go, the amount of options that the player has at every moment in time makes DotA 2 impossible to solve with an adversial search tree. Rather, OpenAI is able to create the world's best DotA 2 player through trial-and-error machine learning, by running an extreme amount of game simulations until it is able to tell which action trees lead to success.

Safe General AI

https://openai.com/

Machine Learning

Supervised machine learning help find connections given data with research and study outcomes. There are thousands of data and statistics on the internet. It is arduous job for a man to find related data and connect with expected outcomes. By high frequency of computing, using supervised machine learning, programs are able to find relationships between tremendous amount of data with given outcomes quickly and correctly. There are binary classification, multiclass classification, and regression provided by supervised machine learning. Overfitting and underfitting are the major problems of supervised machine learning, greatly lowering efficiency of programs. Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns. Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms.

https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Relationship to computational creativity – exploitation and exploration

Recurrent Neural Network Handwriting Generation Demo
This demo from the University of Toronto writes in handwriting what is typed into the text box, and allows users to select from a number of handwriting styles. The handwriting system uses long short-term memory units in recurrent neural networks in order to generate the sequence of black pixels which we interpret as handwriting. I found this demo particularly interesting because after being trained on a large amount of handwriting, the system predicts where the next pixel will be placed in relation to the current pixel to form a letter it has been trained to recognize. More explicitly, instead of training the classifier to make more classifications, the system is trained to create original content which would be classified properly given its previously trained classifier. I think this flipping of the classifer's purpose makes this demo fairly unique and a very tactile demonstration of a system trained using supervised learning.

Here is a link to the demo: http://www.cs.toronto.edu/~graves/handwriting.html

Here is a link to an explanation of recurrent neural networks and long short-term memory: https://arxiv.org/abs/1308.0850

Learning sequences

Supervised Machine Learning in Spam Detection

An example of supervised machine learning, which has been fielded and improved over many years is spam detection. There are 3 primary techniques for machine learning spam detection: Random Forest, k-Nearest Neighbors, and Support Vector Machines. Random forest works by generating many decision trees, which work together to form a forest of trees. This method works well when working with small datasets, which require complex classification. K nearest neighbor works by comparing new inputs to the existing dataset to find the most similar category of inputs and classifying the input as the same as the most similar neighbors. This runs real time and requires a distance computation for each input, which can be costly for large datasets. Support Vector machines rely on decision planes rather than trees and are very effective for less noisy datasets.

Source: https://dev.to/matchilling/comparison-of-machine-learning-techniques-in-email-spam-detection--2p0h

Comparing methods

Alexander Reed
Data is playing an ever-larger role in healthcare treatment today. Vanderbilt Medical Center has a Bioinformatics Lab dedicated to researching and applying uses of data analytics and machine learning to healthcare. One specific use case is the prediction of patient discharge and readmission dates (a project I am continuing to work on with Dr. Fabbri). Here, we are looking at various data sets, many of which are dirty and discontinuous, to help train our models to best predict potential patient discharge and readmission dates. Thus, if we are able to best train our model to recognize certain outcomes, we can help maximize the hospital's efficiency.

https://hiplab.mc.vanderbilt.edu/people/malin/Papers/Negative_AMIA2016.pdf

Postoperative complications
Cancer tumor risk identification

PayPal Fraud Detection

Dominique Carbone posted Oct 2, 2018 9:01 PM Subscribe
https://www.infoworld.com/article/2907877/machine-learning/how-paypal-reduces-fraud-with-machine-learning.html

This article describes the ways in which PayPal uses machine learning to prevent fraud and other crimes through an interview with the company's Senior Director of Risk Sciences, Dr. Hui Wang. It goes on to describe the three types of machine learning the company uses, which include linear, neural network, and deep learning algorithms. To sustain these algorithms, PayPal collects huge amounts of data about their customers, and uses all three together to produce the best results. What I found most interesting about this article was the writer's discussion with Dr. Wang about the potential dangers of the future of artificial intelligence. Wang states that she doesn't worry that machines will replace humans because machines cannot find the data on its own, and does not have the oversight to decide which information is important.

Coursera: Machine learning and Reinforcement Learning in Finance
https://www.coursera.org/specializations/machine-learning-reinforcement-finance

This specialization of Coursera introduces students to the application of supervised machine learning and reinforcement learning in the finance industry. The first part of the course focuses on the mathematical foundations of ML, such as linear regression, overfitting and model capacity, etc., and basic software packages such as DataFlow, TensorFlow, etc. Armed with these skills, students are able to perform basic analysis of market data. Neural networks, gradient descent optimization and other more advanced topics are discussed sequentially. A project aimed at predicting credit spreads is accomplished with the aid of decision tree methods and support vector machines(SVM). Additionally, the course pairs ML with the Markov Decision Process and the Black-Scholes-Merton model to price an option, evaluate risk and perform hedging.

To me, this specialization is a very comprehensive and rewarding introduction to the applications of ML in finance. Supervised machine learning and artificial intelligence are playing a more and more important role in the finance industry, especially in quantitative trading and high-frequency trading, where it helps companies conduct transactions over millions of dollars per minute.

Supervised Learning for Businesses

Supervised learning has great prospects in digital marketing and online driven sales. The process of using Internet will leave detailed footprints for analysis.We have three main process connecting the marketing with the supervised learning: collecting data, labeling data and improving the accuracy of the prediction. Collecting and resembling data correctly requires considerable business skills and capital investment. Not only do we need to set up a consistent and correct data collection mechanism, but also have to ensure that variables are related to prediction. And prediction performance is a challenging. For example,the face-id technology used in iPhoneX. Although Apple claims that face-id uses machine learning to adapt image recognition about changing human appearance, whether you wear glasses or a beard. Although Apple thinks they can achieve 100 per cent accuracy, we have to understand ML solutions does not always give the right answer.

The linkhttps://www.altexsoft.com/blog/business/supervised-learning-use-cases-low-hanging-fruit-in-data-science-for-businesses/

Machine Learning – Coursera

Machine learning online course is one of the most famous online courses offered on Coursera. The instructor is Andrew Ng, who is an adjunct professor in Stanford University and co-founded Google Brain. The courses aimed to teach students the most effective machine learning techniques, and gain practice implementing them and getting them to work. The main language used throughout the course is MATLAB, which is easy to use for beginners but hard to apply to other use cases. All the supervised learning topics include parametric/non-parametric algorithms, support vector machines, kernels, neural networks.

https://www.coursera.org/learn/machine-learning

Top 15 MOOC on Matlab:
https://engineering.vanderbilt.edu/news/2018/vanderbilts-online-matlab-course-is-a-top-mooc/

Michael Fitzpatrick and Akos Ledeczi

Classifying RNA riboswitches

The NIH last year published a paper describing the application of supervised machine learning for classifying RNA elements called riboswitches. They are interested in "understand[ing] their mechanisms of action and us[ing] them in genetic engineering." The paper details how they test several different machine learning algorithms, including sequential minimal optimization and Naïve Bayes. These algorithms were checked for categories like accuracy and sensitivity. This study found that Multilayer Perceptron algorithm was best for classifying the riboswitches. I found the portion of the paper I read to be pretty interesting (though the full paper is locked by account access), and it was good to acquainted with machine learning techniques that haven't been covered in class yet, even if I don't fully understand the science behind riboswitches.

Source: https://www.ncbi.nlm.nih.gov/pubmed/27040116

Smart Deduplification
source: https://dedupe.io/

Dedupe.io is an open-source application that uses supervised machine learning to detect and remove complex duplicates from an excel sheet or database. Dedupe uses a fragment of the dataset for training: the user is prompted with an example of two potential duplicates and is asked to classify whether these data are indeed duplicates. Once the user decides that enough training has been done Dedupe will build a model for detecting complex duplicates and will classify all remaining potential duplicates in the test set (total dataset - training examples).

Machine Learning in Computer Vision

Anne Zou posted Oct 2, 2018 7:29 PM Subscribe

Computer vision is an application of supervised machine learning in which convolutional deep neural networks can be trained to construct internal representations of the world and identify them in visual imagery. The convolutional network is trained using labeled examples (i.e. supervised learning) for features it should learn to detect, and is made up of multiple layers of neurons designed to be shift invariant (i.e. translating an image should not affect the result). I find it very interesting that this design was actually based on the organization of the animal visual cortex. As our understanding of neuroscience and psychology improves, we come up with more and better methods of mimicking human thinking and perception for our machines.

Source: http://yann.lecun.org/exdb/publis/pdf/lecun-iscas-10.pdf

Music Recommendation

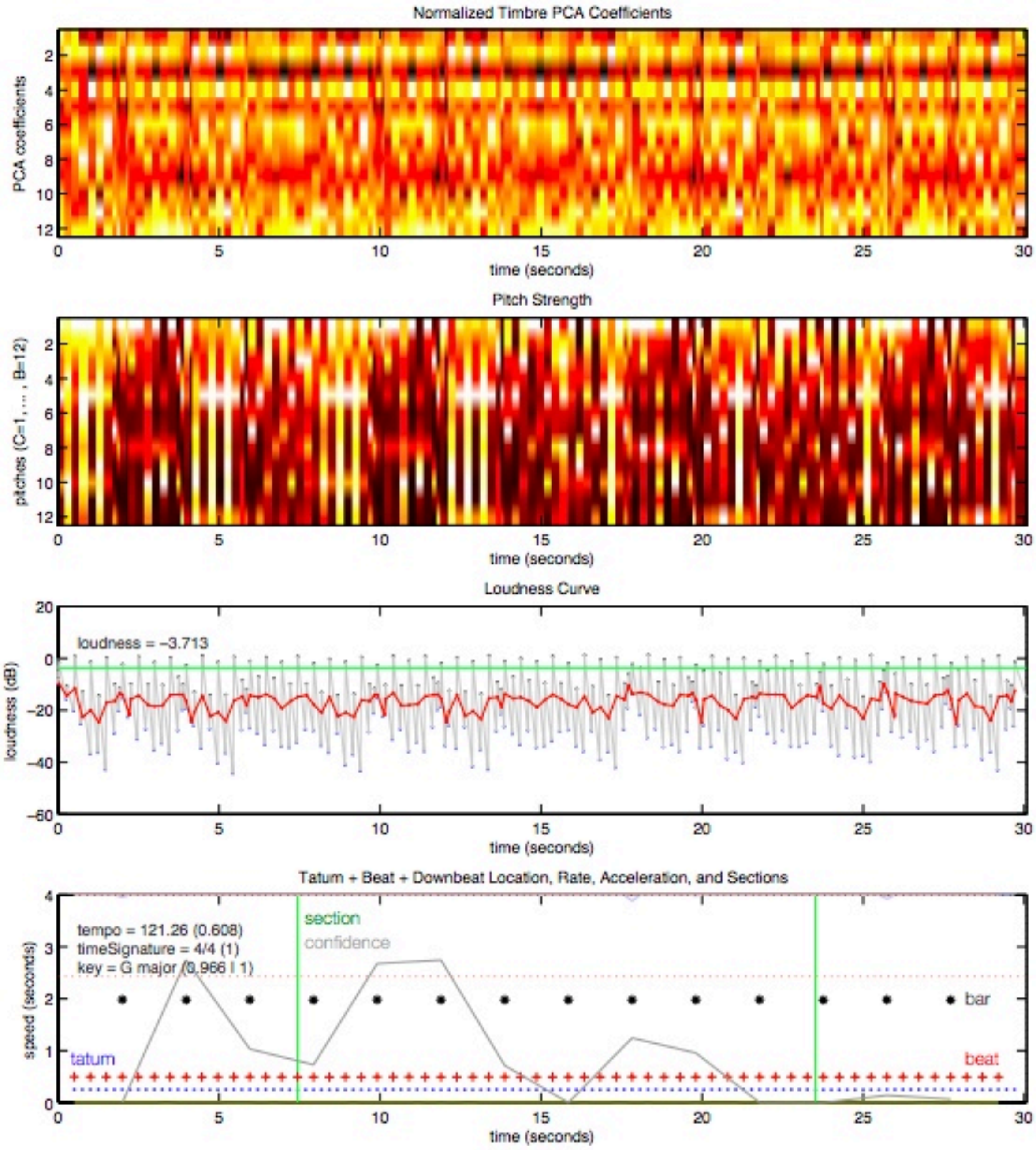


Image source: Tristan Jehan & David DesRoches, via The Echo Nest.

https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe