Why Small Data is Essential for Advancing AI

Small data refers to data that humans can observe
and process personally. Information about any
individual in a crowd is small data. The same information,
collected for all of the people in a crowd is big data.

Everything was small data before we had big data. The scientific discoveries of the 19th and 20th centuries were all made using small data. Darwin used small data. Physicists made all calculations by hand, thus exclusively using small data. And yet, they discovered the most beautiful and most fundamental laws of nature. Moreover, they compressed them into simple rules in the form of elegant equations. Einstein championed this with E=mc^2 . Although it's estimated that perhaps 65% of the hundred biggest innovations of our time are really based on small data, current AI developments seem to focus mostly on big data, forgetting the value of observing small samples.

This is not another article comparing small data to big data. This article is about the role of small data in the future of AI. Efforts have already begun in this direction. Although the current mantra of deep learning says "you need big data for AI", more often than not, AI becomes even more intelligent and powerful if it has the capability to be trained with small data. Some AI solutions that rely only on small data outperform those working with big data. Some other AI solutions use big data to learn how to take advantage of small data.

Small data turns big data into a tool, and takes over
the main role of defining the specifics of an AI system.

Although it appears that everyone wants to use deep learning in referenced frameworks, in some cases we don't really need deep learning to create top performing AI. And, although there's a tendency to train AI using as much data as possible, and mixing real data with synthetic data, in many cases problems can't be solved with big data alone. Some important conclusions can only be drawn from small data. For example, we can gain precise insight

into personality, habits and motivations of an individual *only* with small data. Small data can be packed with meaning derived from that individual's unique personal data. AI must be capable of extracting this meaning. Massive datasets obscure this meaning and machine learning methods typically applied to those datasets simply average out this information. A big challenge ahead of us is creating AI approaches that are capable of finding value in small data.

> With big data you can afford to have noise that you
> filter out. With small data everything is meaningful.

Small data can be packed with meaning derived from that individual's unique personal data. Small data can be analyzed [by humans](). Also, a single human normally only generates small data. Therefore, any product or service optimized for individual humans must work with small data. Who wants to be treated like an average person or a cluster of people? Ultimately, AI needs to be tailored to an individual person and should learn from feedback coming solely from that one person. Imagine the potential of AI that can leverage the power of both big and small data simultaneously. This is possible, and we will illustrate how, but first we'll describe an application of AI that relies on small data *exclusively* to provide unprecedented personalization. Then, we'll describe new technology that uses big data in order to train AI to subsequently learn *exclusively* from small data.

## AI That Relies on Small Data *Exclusively* to Provide Unprecedented Personalization

Ultimately, AI is about mastering knowledge, not processing data. It involves giving a machine the knowledge needed to perform a task. A very specific form of knowledge is required to personalize a product for an individual human being. An example is personalizing a pair of shoes for a specific person. Although luxury footwear brands are selective and exclusive by definition and have a desperate need for small-data, the emphasis has been on big data. Sales is big data. Marketing is big data. Trends are big data. Even individual product recommendations involve bigger data than is ideal. Using small data, one can move from average-based analytics to immediate decision making, from multicast to individualization, from generalization to personalization.

Product fit is typically elaborated through simulations. Individualized simulation cannot be achieved through big data because the simulation model needs to incorporate customer feedback on an individual basis. This is a small data job.

> Any product or service optimized for individual
> humans must work with small data. Who wants to
> be treated like an average person or a cluster of people?

Milan based ELSE Corp has developed a patent-pending method based on hybrid AI for "best fit" detection on an individual level. Their method is built on small data and a proprietary machine learning algorithm based on customer feet scans reconstructed in 3D, predictive fitting metadata, and individual fit results of actual shoes tried on by the customer.
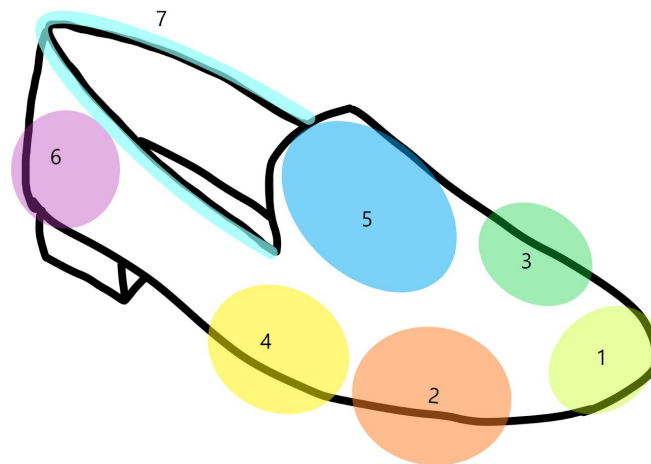
Instead of using the traditional "top down" mapping from big data to a person, the company has developed a "bottom up" approach. Their algorithm incorporates each individual customer's physical fitting feedback. Instead of creating micro-clusters from big data, this method builds a separate algorithm for each individual customer. This small data approach entails knowing exactly what data one is operating on. This could not be achieved with traditional deep learning or other types of advanced statistics algorithms. ELSE's AI does not acquire knowledge through deep learning and big data. Instead, the data is inserted into the AI directly by humans, much like in GOFAI.

ELSE's virtual fitting service is called MySize.shoes. A customer enters information such as foot width, toe shape and arch type. This metadata is considered "predictive fitting classifiers" by the algorithm. Next, the customer's feet are scanned via a 3D scanning device which creates 3D models of their feet. Then, ELSE's AI converts the customer's input from their real fitting, such as the tightness or looseness of each fitting zone, into a suitable small data based recommendation of a pair of shoes from stock with the best fit characteristics, individualized specifically for that customer. The same approach could be used to generate optimal data to be used to produce made-to-measure shoes.
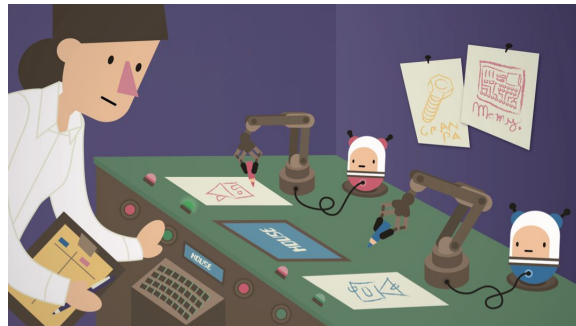
AI Kindergarden: learning from small data

Humans only learn from small data.
For a human, it is sufficient to see one
car in order to recognize all cars afterwards.
Humans do not need millions of examples
of cars in order to reliably recognize cars.

Using Big Data to Train AI to Subsequently Learn *Exclusively* From Small Data

Deep learning usually requires big data. In fact, the biggest limitation of deep learning is the need for large amounts of data. If a deep learning project fails, it's most often due to a lack of sufficiently large quality data set. The reason deep learning needs a lot of data is because this machine learning method is highly eclectic. It can learn anything—and because it can learn anything—it needs a lot of data to learn *something*. In contrast, specialized learners can learn from smaller amounts of data. One example of a semi-specialized learner is the general linear model: unlike deep learning that can learn all kinds of non-linearities, general linear model can deal only with linear

relationships between variables. However, such a linear model requires only a fraction of data compared to what a deep learning model would require if we give it the same type of data.

An unavoidable fact about specialized learners is that they need to be limited to a certain domain of learning. This is due to the no free lunch theorem. Specialized learners are capable of learning from small data because they possess correct inductive biases. Inductive biases represent knowledge about the world in which the learning will take place and are present in the model even before the training has begun. In other words, the model already needs to know how to extract meaning from a certain type of data. Only if a machine learning model has a sufficient amount of such knowledge, it can successfully learn from small data.

> Small data is human-centric. Specialized
> knowledge for human learning is in human DNA.

For example, humans cannot learn any random language. We can only learn certain structures of languages. They must have nouns, verbs, etc. In contrast, deep learning software can learn any random language, those that humans could never master. But deep learning needs a million times more examples than a human needs to learn a human language. Moreover, after training on all of these millions of examples, deep learning programs still do a poor job compared to humans when it comes to understanding what is being said by a language. Just try having a conversation with Siri/Cortana/Alexa, each of which is trained with millions of examples that far exceed the amount of language data that a human person can possibly process in a lifetime. Yet, humans are much more fluent users of that language.

> How do we make machines that can learn from
> small data, and how can we make DL that learns
> from a small number of data points and still
> extracts all the necessary non-linearities?

Frankfurt based startup RobotsGoMental has developed technology that turns any deep learning solution into a learner from small data. The company has developed technology that creates specialized deep learners. Such specialized learners cannot learn anything. They are limited to a certain type of problem—a domain. But then they can learn all kinds of non-linearities even from small data; this is because they know exactly which relations in data are relevant for that type of problem and which should be ignored. That way, they become experts for a given domain. Within that domain, they learn rapidly—like humans. It is possible to create expert learners for any industrial application such as voice recognition, face recognition, predictive maintenance, churn prediction, time series, etc. Apparently, there is a huge market for this technology: The market for traditional deep learning technology is already big. By the very nature of the world in which we live, the market for such expert deep learners must be much, much bigger. In fact, this may be the future of 99% of deep learning.

Conclusions

AI created directly from big data is just a prelude to what's coming. The future of AI is AI that can master small data. Such AI is possible and already exists. Small-data AI can be created both 'by hand' and through machine learning performing an initial training from big data. The optimum is combining human expertise and machine driven training applied simultaneously to insert small data expertise into machines. Humans have always excelled

using small data. We are now just beginning to build intelligent machines that can leverage small data to help us enhance civilization to levels never seen before. Mastering small data is essential for advancing AI.

We are honored that this article was published in MIT Technology Review Italia National Edition, August 6, 2018

[Gli Small Data sono importanti per avanzare l'IA](#)

INTELLIGENT MACHINES

Gli Small Data sono importanti per avanzare l'IA

Con Small Data si intendono dati che gli esseri umani possono osservare e elaborare personalmente. Laddove il CV di un individuo è Small Data, un database con milioni di CV è Big Data.

Di **Margaretta Colangelo, Danko Nikolić e Andrey Golub** il 06-08-18

Possiamo ottenere una visione precisa della personalità, delle abitudini e delle motivazioni di

This article was written by Margaretta Colangelo, Danko Nikolić, PhD, and Andrey Golub, PhD. No part of this article may be used without their express written permission.

[Margaretta Colangelo](#) Margaretta is Managing Partner at Deep Knowledge Ventures an investment fund focused on DeepTech with investments in AI, advanced biomedicine, and Longevity. is President of U1 Technologies. U1 provides the communications infrastructure for trading platforms used by some of the world's top multinational investment banks. Margaretta is Co-founder of Jthereum. Margaretta is an advisory board member at Robots Go Mental and is an External Advisor at ELSE Corp the two companies mentioned in this article. She is based in San Francisco. [@realmargaretta](#)

[Danko Nikolić, PhD](#) is a [brain and mind scientist](#) and an AI practitioner. He is Chief Data Officer and Head of AI at [savedroid](#) AG an award-winning German FinTech company specialized in AI based saving technology to democratize cryptocurrencies. He is also founder and CEO of [Robots Go Mental](#). His work on brain research at Max Planck led him to develop the theory of [practopoiesis](#) and the concept of [ideasthesia](#). Danko is an honorable professor of psychology at the University of Zagreb and is based in Frankfurt, Germany. [@dankonikolic](#)

[Andrey Golub, PhD](#) is a AI-driven design evangelist, tech entrepreneur, technology consultant, experienced R&D Manager, contract professor of Fashion Retail strategies and now Co-founder and CEO of [ELSE Corp](#), a B2B Cloud SaaS platform for [Virtual Retail](#) and 3D Commerce. Andrey is based in Milan, Italy. [@aVg](#)