International Journal of Robotics and Automation

Kosmos Publishers

Research Article

Int Jr Rob and Auto Engg: IJARE-105

How can Supervised Machine Learning (ml) Algorithms be enhanced in improving their predictive power by requiring proper feature selection prior to training; thus, Advancing our understanding of still Obscurely regulated complex Phenomena, such as the gradual Physiological decline due to aging or cancer?

Thomas Hahn[1*], Daniel Wuttke[2], Philip Marseca[3], Richard Segall[4], Neha Gupta[5], Ankush Sharma[6,7], Rawad Hodeify[6,7], Md. Sahidul Islam[8], Hidayat Ur Rahman[9], Ana Lara-Rodriguez[10], Fusheng Tang[1]

[1]University of Arkansas at Little Rock, Little Rock, AR, USA
[2]Wuttke Technologies, Palo Alto, CA, USA
[3]SAP, Newtown Square, PA, USA
[4]Arkansas State University, Jonesboro, AR, USA
[5]Lamar University, Beaumont, TX, USA
[6]Erasmus University Medical Centre, Rotterdam, The Netherlands
[7]American University of Ras Alkhaimah, United Arabic Emirates
[8]University of Rajshahi, Rajshah, Bangladesh
[9]Lahore Leads University, Lahore, Pakistan
[10]Universidad de los Andes, Bogota, Colombia

[*]**Corresponding author:** Thomas Hahn, University of Arkansas at Little Rock, Little Rock, AR, 2811 Fair Park Blvd., Little Rock, AR, 72204, USA. Tel: +13182433940; Email: TFHahn@UALR.edu

Citation: Hahn T, Wuttke D, Marseca P, Segall R, Gupta N, et al. (2018) How can Supervised Machine Learning (ml) Algorithms be enhanced in improving their predictive power by requiring proper feature selection prior to training; thus, Advancing our understanding of still Obscurely regulated complex Phenomena, such as the gradual Physiological decline due to aging or cancer?. Int Jr Rob and Auto Eng: IJRAE-103.

Received Date: 12 September, 2018; **Accepted Date:** 27 September, 2018; **Published Date:** 8 October, 2018

Abstract

Background:

Incomplete feature selection causes reproducible prediction errors. Such prediction errors are the consequences of still Imperatively Hidden Objects (IHOs). They are called IHOs because – even though they cannot be directly observed or measured – they affect the predicted outcome. Such IHOs are often the hidden causes. Their consequences are exposed by reproducible prediction errors. IHOs are challenging because their exact impact on the observation of interest cannot be fully considered until they have been fully uncovered. But, since feature selection must be completed before any supervised machine learning (ML) algorithm can be properly trained, such kind of IHOs must be discovered before feature selection can be fully completed.

Results:

To distinguish IHOs from one another and from background noise, the measuring methods and their surrounding experimental environment must be varied. Independent features can be considered as singlets, e.g. the impact of transcription, length of poly-(A)-tails and ribosomal coverage on protein abundance, because they don't depend on one another. However, codons must be considered as atomic triplets only. When the prediction can be made by different methods in different dimensions feature selection can be considered as fully completed, e.g. when transcription can be fully predicted either by considering the trajectories of time series plots or Transcription Factor Binding Site (TFBS) distributions, Transcription Factor (TF) ratios and TFs abundances. New features can be discovered by creating external experimental conditions, which cause any until then correctly ML prediction algorithm to fail, because it indicates another still unknown dimension by which IHOs differ from one another and their background noise environment. Strategies for uncovering IHOs are discussed.

Conclusion:

Focusing on IHO discovery can speed up our scientific progress because each newly uncovered IHOs should be added to the selected features for training new ML algorithms because IHOs prevents better understanding of complex phenomena, such as aging and cancer.

1. Introduction

Our key observation for this discovery was that we humans are very bias pertaining to what kind of information (i.e. seeing, hearing, feeling, tasting and smelling) we pay attention but disregard the rest. That is why we don't get to see the whole picture and hence, remain stuck in a confusing state of mind - sometimes for many years - because we tend to neglect to consider almost all information outside the spectrum of our very subjective and limited range of sensory perception. Since we humans tend not to be systematic in selecting the data sources and dimensions of information (i.e. feature selection) without even being aware of it, we need the help of less bias artificial intelligence (AI).

The fact that we cannot perceive this information does not make it any less important or relevant for our lives. This was realized while trying to get a better understanding of the regulation of the aging process. The problem is that there are no omics datasets to test new aging hypotheses. This implies that nobody before us seemed to have felt that collecting transcriptome, proteome, metabolome and epigenetic data every 5 minutes throughout the entire lifespan of the yeast, would be worth the effort. We are totally capable of generating the data needed to advance in our understanding of aging and many other complex and still obscure phenomena, but the wet-lab scientists, who design our biological and medical experimental studies, don't seem to be even aware of missing something very important.

A good example is the magnetic field. We humans tend to ignore it because we cannot feel it. But, nevertheless, it affects our lives. It can cure depression. It can cause involuntary muscle twitching. Some birds use it to navigate the globe on their seasonal flight migration.

We are concerned about that there are other fundamental phenomena similar to the magnetic field, of which none of us is aware yet, because so far we have not tried to look for similar imperceptible information carrying dimensions.

For example, spiders, ants and bets are blind. However, visible light affects their lives regardless whether or not they have a concept of vision, since they have never experienced it. There could be other information carrying dimensions that – like the light for the spiders, ants and bets – is an imperatively hidden object (IHO), although it affects our lives so profoundly that we cannot understand aging and many other complex phenomena without considering such kind of information as well. That is why we recommend using AI to reduce our observational bias.

Often, scientific progress has been made by accident. The means that a mistake, which changed the otherwise constant experimental environment in such a way that an unexpected result or observation was the consequence, was what helped us unexpectedly to finally make some progress. That is why we propose to intentionally vary external experimental conditions, methods, measurements, study designs, etc., to discover new features, which affect the outcome, much sooner.

The theories about the impact of still imperatively hidden objects (IHO) below might be challenging to understand, but it is worth trying, because if it succeeds, it will fundamentally improve our experimental methods and scientific study design

2. How can the inherent challenges posed by hidden objects be adequately addressed and eventually overcome?

2.1. Is proper, correct and exhaustive feature selection for training machine learning algorithms already possible even before all imperatively hidden objects/factors/dimensions (IHO/F/D), which are required for correctly conceptualizing aging and many other complex phenomena, are fully discovered?

This section is about imperatively hidden objects (IHOs) and the need for new concept discoveries, which are needed for subsequently selecting all necessary features (i.e. proper feature selection), which are required for fully understanding aging, cancer and many other still incompletely understood complex phenomena. Humans are very bias in choosing their method of conducting experimental measurements or make observations without being aware of it. What percentage of the entire electromagnetic wave spectrum can we perceive? No more than 5% for sure. But the changes, of which we must be aware, before we can understand aging, are most likely much more distinct outside our narrow sensory window because our sensory limitations did not affect the evolution of aging in any way. For example, humans can only hear part of the sound an elephant makes because humans cannot hear such low frequencies as the elephant can. This tends to prevent the full understanding of the elephants' communication options. Humans cannot distinguish such low sound frequencies from background noise, i.e. environment, because they cannot perceive the low elephant sound frequencies from being different from the background environment. However, without considering those imperatively hidden factors we cannot fully understand elephant communication. Therefore, humans tend to miss cellular processes, which can only be distinguished from background noise outside the electromagnetic wavelength interval, for which humans have evolved sensory organs, i.e. eyes, ears and skin. The mechanism by which the tongue and nose operate is of an entirely different dimension because they cannot sense any wavelength.

For example, before magnets were discovered, they remained for us an imperatively hidden object (IHO) because we could not even suspect them in any way. But still, just because we lack any senses for perceiving any kind of magnetism does not stop it from affecting our lives. Only after we discovered the consequences of the forces, which the magnetic field has on some metals, prompted us to search outside the limited window, within which we can sense differences in wave length. Magnetic fields could affect life in many positive ways because they are used to treat major depressive disorder and cause involuntary muscle contraction. But has anybody even thought of measuring the magnetic field of a cell or brain, which I expect to be strong enough for us to measure with sensitive devices? Since any electric current causes a perpendicular radiating magnetic field, it can be hypothesized that the weak magnetic field is pulse-like and depends on the temporal pattern by which neurons fire action potentials. The changes in the magnetic field of a cell is expected to be enriched for the cellular component membranes because they have proton pumps and maintain an electric gradient to produce adenosine triphosphate (ATP). But what if changes in this magnetic field are causing us to age? Then we could stop the aging process by any intervention, which sets our cellular magnetic field pattern back to its youthful benchmark. It is suspected that the reason for our rudimentary understanding of the aging process is caused by us missing such kind of imperatively hidden objects (IHOs), which are required for making the essential key observations without which aging cannot be fully explained. A magnetic field as a concept, which exists, regardless weather we are aware of it. There may be many more other hidden concepts, which we must develop correctly, before we can reverse aging.

2.2. Analogies to aid in the understanding of the concept of Imperatively Hidden Objects (IHOs)

Let's say that an immortal interstellar alien highly intelligent out-of-space extraterrestrial critter has landed on Earth. Let's imagine that he can only perceive wave lengths within the limits of the magnetic field. Then we humans would not even notice this out of space interstellar visitor because he/she remains an imperatively hidden object (IHO) that we cannot even suspect. Let's say this interstellar species has not evolved a body or anything to which our senses are sensitive. Let's say that this life can be fully defined by irregularities within the magnetic field. But this interstellar critter can perceive us humans because our magnetic field disrupt the homogeneity of the background environment and must therefore be something other than background noise. Let's say that this immortal interstellar critter can perceive and process all the magnetic fields on Earth. Could he maybe develop the concept of siblings or parents on its own? Is the magnetic field of relatives more similar to each other than expected by chance? It is very likely because humans vary a lot in their neuronal wiring architecture. Hence, each human could be defined by the pattern of his/her neuronal action potential firing pattern. This inevitably causes a very weak unique perpendicularly acting electromagnetic field that cannot be detected by our instruments. Therefore, instead of humans, we should use the giant squid as model organism to understand the relationships between life, aging and changes in magnetic field because it has the thickest neuron. Therefore, it must fire stronger action potentials than our human neurons. This will inevitably cause a stronger perpendicularly acting electromagnetic field, which may be strong enough to be detected by our instruments.

Let's say that this interstellar critter wants to use machine learning to predict the risk of any particular university student in the USA for having to return home after graduation, because they lost their immigration status and could not find a job, which would have made them eligible for one year Optional Practical Training (OPT). Let's say that this interstellar critter has no concept of aging and that his most important goal is to develop a classifier by developing a new machine learning algorithm, which can predict in advance the risk that any particular student is facing to no longer been allowed to reside in the United States. Let's say that accomplishing this objective has the same meaning and importance to this critter as for us the cure of aging and elimination of death. What should he do? He cannot talk. No human even suspects him. He could start using supervised machine learning by observing thousands of students to find out what those students share, which are forced to leave, or what they lack compared to citizens, who are always welcome here. We hypothesize that no matter how clever and sensitive to irregular interruption of the homogenous electromagnetic field, which is the only dimension, in which he can sense the presence of humans and any other form of life, he has no chance to understand the risk factors for being forced to leave America after graduation, because they are an imperatively hidden concept (IHC) to this critter, since he cannot even suspect them in any way. However, without developing the right concepts in advance, this critter can never discover the risk factors for having to leave the USA after graduation.

The same applies to aging. We are still missing essential concepts without which we cannot fully understand it. But even if somebody by chance could detect the magnetic irregularities caused by this foreign extraterrestrial critter, he/she could never suspect that it is highly intelligent. This means that even if we measured a cell across the entire wavelength spectrum and could clearly detect its presence, we would never suspect it to have any kind of intelligence, because we would consider the anomalies in the magnetic field as background noise. Our visiting interstellar critter has a similar problem. He cannot develop the essential concepts, without which he could never develop a machine learning algorithm to predict all the correct risk factors, which impair the chances for somebody to be allowed to keep residing in the US while not full time enrolled. As long as this critter has no concept of "country", e.g. the USA, he has absolutely no chance to discover nationalities, because even if he could figure out the nationality of everyone, it would make no sense to him. But words like "American" "German", "French" or "Indian" cannot make any sense to this critter as long as the concept of "country" remains an imperatively hidden object for him. How can somebody be considered "German" or "American" as long as the concept of Germany or USA are still lacking? One can only be German if Germany exists. Without at least suspecting the concept of a country, e.g. Germany, there is absolutely no way to discover the required concept of citizenship. Unfortunately, without determining the feature "citizenship" no machine learning algorithm could learn to make correct predictions. .The same applies to aging. We are still lacking so many essential concepts without which aging can never be understood.

For example, as long as the concept of a ribosome is lacking, we have no way of understanding the changes in the relative abundance ratio of mRNA and proteins. We may have some initially success with building a model to predict protein abundance and concentration because it is about 70% similar to the transcriptome. However, according to Janssens et al. (2015) [1], this similarity declines with age and is a driver of replicative aging in yeast. But no matter how many training samples we use to train our predictor, it must fail, unless we have developed a mental concept of a ribosome. We believe we face a similar predicament with understanding the causes and regulation of epigenetic changes over time with advancing age, despite being able to measuring them so clearly that we can use them to determine the biological age. But unfortunately, as long as we lack any concept, by which epigenetic changes could be connected to other cellular processes, we cannot understand how epigenetic changes are regulated. Before we could correctly conceptualize the role and scope of the ribosome we had no way to explain the mechanisms by which mRNA and protein abundance is linked. But even after we conceptualized the role of the ribosome correctly any machine learning algorithm to predict protein concentration would inevitably fail as long as we lack the correct concept of the poly-AAA-tail. Similarly, there are still lots of imperatively hidden concepts, factors, dimensions or objects, which we cannot suspect because we cannot perceive them, which prevent us from fully understanding aging. However, the fact that our current observations fail to fully explain aging, indicate the presence of imperatively hidden factors of which we can see the consequences without being able to detect their causes. But since every consequence must have a cause, any unexplained consequence indicates the presence of imperatively hidden imperceptible factors (IHIF) without which we cannot succeed to improve our feature selection. As explained in the student immigration example, only when selecting the correct feature, e.g. citizenship, the risk for being asked to leave America by the federal government can be fully understood and hence, can be predicted much better. Could we convince some of our readers of the high likelihood of the presence of imperatively hidden factors, which we cannot perceive yet as being distinctly different from their environment and from one another?

3. Conclusions and proposed responses/adaptations of our study design

3.1. What is the rate-limiting bottleneck, which limits our research progression, and why?

The current bottleneck in defeating aging is not addressed by further improving our machine learning algorithms and increasing the training samples, but instead, we must focus on improving proper feature selection first. The main contribution of this conceptual research towards defeating aging is to predict features, measurement types and intervals between measurements, which could show the actions of aging much clearer than the features, which have been currently selected to stop aging and defeat death. Now it is up to wet-lab scientists to test our aging hypotheses. But even if all of them can be ruled out, the possibilities, by which the mechanism of aging could function, would be reduced. This would leave us with fewer hypotheses left to test. Since the options we have for fully understanding the aging process are large - but yet finite - any crazy appearing – no matter highly unlikely seeming - hypothesis, which can be ruled out, brings us a tiny step closer to immortality. The reason why we claim that correct feature selection, but not the gradually improving performance of our machine learning algorithms, is the current bottleneck, which is holding us back from improving our understanding of the aging process, is that our machine learning algorithms have been improving gradually over time, but our feature selection methods have not.

The fact that we cannot find any data for measuring the yeast transcriptome in five-minute intervals for more than 3 out of the average 25 replications, which is considered the average wild type (WT) yeast replicative lifespan, indicates that nobody has seriously suspected that we could at least observe the effects of the aging mechanism by selecting new periodic features, such as period length, temporal phase shift or amplitude, which only make sense if we replace our linear with a periodic concept of life. However, this requires us to change our concepts about life to be driven by linearly acting trends to cyclical periodically acting trends in order to expand our feature selection options to periodic quantities, such as period length, temporal phase shift, amplitude or oscillation pattern, which would have been impossible to imagine when holding on to the old linear concept. In this case – although we could clearly measure the period length - we could not detect it as a feature affected by aging until we explicitly define, select and measure this new feature, e.g. the period length, temporal phase shift, amplitude or oscillation pattern. That is why rapid concept discovery is so important. We are worried about that there could be many still hidden dimensions, which are very similar to the magnetic field that we cannot yet anticipate. But we must first associate information from these kinds of magnetic-field-resembling still imperatively hidden dimensions with aging before we can understand aging. Since we humans have observational tunnel vision, which is mostly limited to the dimensions of our sensations, we must use artificial intelligence, because for it all the different dimensions and the features, which define them, are more equal. Only if we can make people understand this, we will have a chance to collectively survive. That is why it's so important to get this published because otherwise it won't be taken seriously. It is possible to provide proof-of-principle that we still very naive and observation bias humans would have missed important relevant features, if we would not have let artificial intelligence (AI) define possible aging-relevant features in a much more systematic and less bias manner. For us bias humans to create a much less bias AI, we must be able to look at life from many different ridiculous-seeming perspectives because that is what we expect our aging-features-selecting AI to accomplish. Below is an example how this could work.

4. AI Could Drive The Below Described Continuously Ongoing Emerging Random Evolution Mimicking Procedure Aimed At Discovering Unpredictable Means To Survive, i.e. AI Could Power The Random Operation "Unpredictable Survival"!

A major threat is that we are aging much faster than we can reverse it. We are still very far away from inferring, which information is most likely relevant for reversing aging that we MUST take an undirected random method, which is based on trial and error, to counteract this problem because we do not have any better alternatives.

Every day lots of new pairs of information are added to the web. Anything, which defines at least two indivisible pieces of information as a value pair indicating a specific instance, can be ingested by machine learning algorithms. Therefore, we should start developing independently working software, which keeps crawling the net for any instance defined by at least two informational units as input data. Then, even though this software cannot infer the meaning of any of the event-defining information pairs, it can use their values in predicting pretty much any other combination of paired information and try to predict any pair with any other pair. This would allow for discovering even weak correlations and dependencies much sooner than when exclusively selecting features manually in our traditional way based on logic reasoning. Although logic reasoning and highly directed and targeted manipulations are good to have, it takes us way too much time until our understanding and concepts of new correlations has developed far enough to contribute to logically driven data feature selection and data manipulations.

This continuously web-crawling software keeps adding anything, which could either serve as input our output value for any kind of supervised machine learning process. When this software can predict any random feature by whatever means it can possibly think of, it will let us know so we can check whether this could possibly make sense. We need to improve the Natural Language Processing (NLP) and semantic recognizing ability of this randomly feature adding software so that it can combine the same informational components into a single unit feature. But nevertheless, just like evolution randomly mistakes in grouping the same information component into a single indivisible feature, variations in the groupings of informational components, which must be predicted all at once, could turn out to be a good thing. For example, considering all transcription factor binding sites (TFBS)-associated information into a single informational group, may allow for the most accurate prediction rate; but only when our random model contains all input features, which we need to define any possible informational dimensions, which is needed to sufficiently define all the parameters/features that could belong to the TFBS dimension.

For example, if our feature-hungry crawler has not yet discovered that TFBS binding is a co-operative rate and not a Boolean process, it would fail. However, if it could learn to predict time series plots only based on the Boolean value indicating whether a particular transcription factor (TF) could possibly bind to a promoter, but disregard the number and order of the TFBS for the same TF in the promoter of one gene, it could still predict time series plots well enough to raise its prediction power far above the threshold at which we would take a look at it. Although this old model is still imperfect, it has value to get it as soon as possible, instead of waiting until our crawler has found all input to parameter to assign a value to all possible dimensions of the TFBS domain. This would actually speak in favor of allowing our prediction crawler to randomly vary any specific dimension of any domain, which is suited for training supervised machine learning algorithms, because the fewer the number of dimensions, which make up any domain, the fewer input components (i.e. features) are required for building a model, which is based on randomly considering grouped information components.

Currently, most of us are not aware of the artificial imperative limitations resulting from letting humans have the monopoly on deciding, which dimensions can be grouped together to form a meaningful instance for input or output to train a supervised model. It is likely that smaller domains consisting of fewer dimensions or larger domain combining more dimensions could be better for understanding and predicting.

But, although there are so many humans on this planet, our thinking, understanding, conceptualizing, imagining and applying our intuitive preferences for including very specific dimensions into an indivisible input or output instance without even worrying about possible alternatives since our perceptions and understanding of life's concepts are far to similar within our species. But, nevertheless, the way in which our senses, perceptions, imaginations, concepts and partial understandings of any phenomenon intuitively select the dimensions to a larger domain, which most of us would never even consider to predict in parts or as a very small dimension of a much larger super-domain, is only one out of very many possible options for combining any number of specific dimensions into a domain from which any number of input or output instances can be formed.

One could imagine a domain as a column in a data frame, which – like a gene – can have any number of columns i.e. its dimensions, which must be considered like a single instance in their combination, because it is lacking the option to consider only a few of its columns or combining some of with columns from an entirely different and unrelated table. Good examples are time series plots. Human tend to be bias and prefer to define the gene expression time series trajectories by mRNA distance measures at each time point. This may sound obvious, but is this the best way for conceptualizing the temporal expression signature for each gene?

Our colorful time series plots have much more meaning and can carry much more informational value as well as a more meaningful concept for imaging, comparing and analyzing gene specific temporal signatures. However, although they look very pretty and are a good way to get a first impression about the similarities between two time series trajectories, they are not well suited to find out whether the plots for the genes, which belong to the same gene ontology term, are indeed more correlated to each other than to the rest of the genome. But imagine, how many more options you would had, if you were not a human, because then you would not limit your dimensions for defining your domains to only those you can easily mentally visualize and imagine. A computer can randomly extract and try out any combination, subset or superset of dimensions without tending to be limited to those dimensions that can easily be conceptualized as a picture.

Unsupervised machine learning algorithms, which never get tired to randomly define an indivisible domain by any combination of dimensions, might have much more luck to uncover still imperatively hidden objects/factors (IHO/F) than the entire observationally and perceptionally very bias world population of homo Sapiens, which tends to prefer familiar analytical methods, to which it can most easily relate without much regards for, whether the most convenient and intuitively-seeming analytical methods, measurements, selected features and research procedures are truly best suited for solving the very specific scientific problem at hand. Even professionally very successful scientists, experimentalists, researchers and data analysts tend to search for the best problem to match their analytical skills, experiences and preferred methods of measuring rather than choosing the best set of research procedures for overcoming a very specific scientific challenge. AI won't suffer from this human methodical bias if trained properly.

5. How can the initially still Imperatively Hidden Objects/Features (IHO/F) be uncovered and subsequently made available for proper, exhaustive feature selection followed by optimizing the training set for the newly developed supervised machine learning algorithms?

This fifths chapter focuses on answering the pivotal question:

How to naively discover new essential – but initially still imperatively hidden – objects (IHO), which are defined by their initially hidden features, which must be uncovered for proper complete feature selection and subsequent supervised training of novel adaptive machine learning algorithms, to unravel the mysteries of aging and many other poorly understood complex phenomena, which initially depended on hidden objects (i.e. hidden causes of which only their consequences could be observed)?

5.1. Introduction to feature discovery to train supervised machine learning algorithms for artificial intelligence (AI) applications

5.1.1. Feature discovery and selection for training supervised machine learning algorithms: – An analogy to building a two story house:

Imagine a building named “Aging”. It consists of two stories: the ground floor, which is called “feature selection”, and the second floor, which is called “developing, optimizing and training the machine learning algorithm”.

Before any machine learning algorithm can be trained properly, feature selection must be perfected and completed. Otherwise, the machine learning algorithm may learn irrelevant things caused by the ambiguity, which is due to missing features. This poses a high risk for overfitting, i.e. superior performance on the training set but poor performance on the testing set. Not much time and efforts should be invested into optimizing, training and improving the machine learning algorithm until all features are properly selected. As long as feature selection is incomplete one must focus on finding the missing features instead of tuning the algorithm.

In other words, using our building analogy, here is the most important advice: Do not try to complete and perfectionate the 2nd floor called “training, tuning and optimizing the machine learning algorithm”, before you are certain that the ground floor, i.e. “feature selection”, has been fully and properly completed. If this is not the case, one must focus on discovering the missing features for the training samples first. Lots of research has been dedicated to perfectionate algorithms before completing feature selection. Therefore, our algorithms have gradually improved whereas our feature selection has not. It’s like the waterfall model. The previous step (i.e. feature selection) must be fully completed before the subsequent step (i.e. developing and training the supervised machine learning algorithm to make correct predictions) can be started.

5.2. How can missing/hidden features be discovered?

If the machine learning algorithm cannot be trained to make perfect predictions, it indicates that essential data input features are still lacking. When the predicted values fail to match with the observed measurements, despite tuning the algorithm, it means features selection is not yet completed. This is the case when the error between predicted and observed values approaches an asymptote, which is not equal zero. The prediction error is most likely caused by a still hidden object. This hidden object is the cause of the error. But we cannot see the hidden cause yet. However, we can see its consequence, i.e. the error. But since every consequence must have a cause, we must start looking for it.

5.3. Example of a speculative scenario concerning predicting protein folding part 1

Let us take protein folding prediction as an example. Only about 30% of the predicted folding patterns are correct. We must then go back to the last step, at which our prediction still matched reality. As soon as we get deviations we must scrutinize the deviating object, because - most likely - it is not a single object, but instead, 2 or more objects, which to us look still so similar that we cannot yet distinguish between them. In order for an object to no longer remain hidden, it must have at least one feature by which it differs from its background environment and all other objects.

5.4. Example of a speculative scenario about discovering the molecules we breathe

As soon as one feature is found by which 2 objects differ from one another, we must identify them as distinct. Let us take air as an example. Air in front of background air still looks like nothing. Even though air is a legitimate object, it remains hidden as long as its surrounding background, to which it is compared, is also air. Air has no feature that could distinguish it from background air. Legitimate real objects, which lack any feature, by which they could be distinguish from their background and/or other objects, are called imperatively hidden objects (IHO) because no kind of measurement can help to distinguish them in any way as an object, which is something other than its background. If objects like air, uniform magnetic field or gravitational force are omnipresent and uniformly distributed they remain imperatively hidden objects because we have no phase that is not the object, which would allow us to distinguish it as an object. An object before a background of the same object remains an imperatively hidden object unless we find an instance, which varies from the object, maybe in strength, because we need something to compare it with to identify a difference.

The only way by which an omnipresent uniform hidden object can be discovered is if there is some variation in strength or it can become totally absent (e.g. gravity). Otherwise, it remains imperatively hidden because it cannot be distinguished from itself, the environment or other objects. Therefore, in order to uncover imperatively hidden objects, we must intentionally induce variation in the surrounding environment, measurements and methods until new features, by which the object can be distinguished from its environment and/or other objects, can be discovered.

If we can advance our conceptual understanding of air being not the same as its background because of wind, which causes a variation in resistance by which air slows down our motion, air still looks like air. But we know that air consists of at least four very different groups of objects, i.e. 20% oxygen, 78% nitrogen, 1.5% helium and 0.5% carbon dioxide. Now these four objects are no longer imperatively hidden but they appear like a single object. When trying to predict molecular behavior we will get errors because what we think of as one object is actually at least four. By looking, sonar sounding, radiating, magnetizing or shining light on air, we cannot distinguish the four objects from one another yet. But if we start cooling them down gradually, suddenly we can distinguish them from one another by their different freezing temperatures.

5.5. Which features would be best to learn the difference between different oxygen species?

Let us assume that the element oxygen has been discovered and that the method by which atoms are distinguished from one another is to count their protons. Still, puzzling observations, which cannot be predicted by relying on proton numbers alone, will be encountered as soon as ozone (O3), molecular oxygen (O2) and free oxygen radicals are in our air sample. To get the most predictive power and highest F-score, investigators tend to try to optimize prediction by trying to find a method to predict the most common outcome. Accordingly, in this oxygen example, researchers tend to develop an algorithm, which is best to predict bimolecular oxygen (O2), because it is most abundant among all oxygen molecules (i.e. ozone (O3), bimolecular oxygen (O2) and the negatively charged oxygen radical (O-). The error rate under the assumption that there is no difference between oxygen molecules would be equal to ozone/bimolecular oxygen + oxygen ions/molecular oxygen. In order to distinguish between these different kinds of oxygen the electron/proton ratio, the different charge distribution on the molecular surface, molecular weight, molecular volume, the arrangements of chemical bonds, and the position of the oxygen atoms relative to one another within the same molecule, could be added to our training data as newly selected features in order to distinguish between the different oxygen species. But let us assume that we are still naïve and cannot measure the needed features yet, how could we go about discovering the missing/hidden features?

In general, varying the features of the input training data for training a supervised machine algorithm, the learning steps, the inert environment and the methods of measurement must be selected based on intuition due to lack of any better alternatives. For AI to correctly determine the overall electrical charge of an oxygen molecule, AI needs the number of protons and electrons as input data. Unfortunately, if the instruments for detecting protons, electrons and neutrons are lacking, we can see the effect of the still hidden factor, i.e. electron/proton ratio, on the overall molecular charge, but its reason still remains a hidden mystery. In this case, investing time to discover electrons, neutrons and protons, is much wiser than trying to tweak the parameters after the error rate has reached its asymptote, because even if this improves prediction, there is a big risk of over-fitting since AI is basing its decisions on features, which actually have no effect on the overall molecular charge. But instead of using the electron/proton ratio as input features, the molecular size of the different oxygen species, would also work for training our AI-molecular charge predictor. Electron/proton ratio (i.e. a simple fraction) and molecular size (i.e. a volume measured in cubic nanometers) are different dimensions; yet both of them can express the same event, i.e. electric charge. Therefore, both could be used to train AI on predicting the molecular charge correctly. If, like in the example above, the in reality observed outcome can be perfectly predicted in at least two different dimensions, then it is reasonable to believe that all hidden factors have been discovered. The relationship between electron/proton ratio and molecular volume is about the same as between transcription factor binding sites (TFBS) and the trajectories of gene expression time series plots.

To support our hypothesis that we are still many concepts away from understanding and manipulating aging, the example of the mesentery can be used. It took humanity until early 2017 to discover its function and group it together differently according to the new discoveries. We have seen this organ for a long time, yet still, it remained a hidden factor for us, that it is indeed an organ, but it could not be recognized as an organ because its functions were still unknown [2]. This is what is meant by imperative hidden object (IHO).

5.6. A speculative example how nitrogen could be discovered using this method of intuition-driven hidden object discovery procedure, which relies on randomly varying background and object features, until they differ in at least one feature, by which they can be told apart from one another and from other similar-looking objects.

Looking, sonar sounding, radiating, magnetizing, light shining and changing temperature are considered variations in measuring methods. Below -80 degree Celsius the feature aggregate state for nitrogen is liquid, which differs from the remaining still gaseous objects. Therefore, we can conclude that the liquid must be a separate object from the gaseous. Thus, we have found a feature by which it differs from the rest. Hence, we took away the imperative hidden nature by changing environmental conditions, i.e. temperature, until nitrogen looked different from the rest. There was no data, which could have told us in advance that gradually lowering the temperature would expose features of difference between the objects. That is why we must become creative in finding ways by which we can vary environmental conditions and measurement methods until hidden objects differ from one another or their environment in at least one feature, which we can measure. However, until we cool it down to – 80 degree Celsius, nitrogen remains an imperative hidden object unless researchers can find other means to make one of nitrogen's features to stick out from its background and other visible objects. If cooling does not work, nitrogen could be isolated from the other gases by boiling it.

5.7. Challenges encountered in distinguishing between similar objects

Imagine a small brown ball laying in front of an equally brown wall. When looking at it, we can see a brown object, which looks like a wall. However, we cannot see that there is an equally colored ball laying in front of the wall as long as the light is dim and their colors are perfectly matching when looking at the wall. By looking at both objects, they appear to be only one. Even though a brown wall is a legitimate visible object, it serves as background camouflaging the equally colored brown ball in front of it. Thus, 2 different objects are mistaken into one object.

However, a bat, who navigates very well by sonar sounding reflection (i.e. by echo-looting), has no problem to distinguish between ball and wall, no matter how equally colored they look, as long as the ball is some distance in front of the wall. This is an example how changes in measurement dimensions, e.g. substituting visual with echo-looting perception may allow switching over to another feature, i.e. sound reflection, to distinguish between optically indistinguishable objects.

However, on the other hand, this example can also demonstrate the limitations of this environmental and measurement method/dimension variation approach. Let's assume that scientists figured out a way to make the bat more intelligent than humans. Unfortunately, no matter how clever the bat may become, it may never understand the concept of reading, writing and the benefit of a computer screen, since it cannot extract anything in all three cases, because when sonar-sounding and echo-looting a screen, all letters, pictures, figures and graphs remain imperatively hidden objects, if explored by sonar-sound reflection only. It would even be challenging to explain the bat the concept of a screen because it cannot imagine different colors.

This shows again that we must remain flexible with our observational measuring techniques, because if we don't vary them profoundly enough, we may fail in the discovery of still hidden objects. The naïve observer can only discover by trial and error. Lucky, that we humans have developed devices to measure differences in dimensions, for which we lack inborn sensory perception. But we also must use our different measuring devises to collect data, make observations and explore innately hidden dimensions, if we fail to discover differences between at least one feature of our hidden object of interest and other similar-looking objects as well as in at least one feature from its surrounding environment, if we cannot make such kinds of distinction within the limitations for our relatively small innate sensory sensitivity range.

The good news is that any two distinct objects must vary from one another and their environmental background by at least one feature because otherwise they could not be different objects. The challenge is to discover at least one feature, by which hidden objects differ in at least one situation from one another and their camouflaging environmental background.

This is the conceptual foundation, according to which anyone, who can observe in all possible dimensions, must eventually by systematically applying trial and error alone encounter conditions, which allow to expose difference in at least one feature based on which any object can be discerned from its environment and other objects by at least one feature under at least one environmental condition. That is why AI can play a very valuable role in systematically iterating through no matter how many options as long as the total number of combinations for condition and observation variations remains finite.

Numerical calculations, evaluations, comparisons or rankings are not required as long as qualitative distinction allows for at least a Boolean decision, i.e. the hidden object differs or does not differ from its environment and other objects in at least one feature under one set of experimental conditions. True or false, yes or no, is enough for succeeding in uncovering previously imperatively hidden objects.

5.8. Why are data and numbers only crutches?

Data and numbers are only crutches, on which most of us depend on way too much in determining their next steps. We tend to refuse exploring new options of variations unless data points us to them. But the naïve imperative hidden object discoverer has no data or information to infer that changing temperature would expose a feature for distinction, but shining light, radiating, sonar sounding, fanning and looking would not. The new feature hunter must simply use trial and error. He must follow his intuition because directing data is lacking. It would not help him to perfection ate the resolution of his measuring techniques as long as he has not changed the environmental conditions such that objects differ in at least one feature from one another and/or their surrounding environment. In such cases heuristic search options are lacking. There is no method that tells how and what to vary in order to expose new features for novel distinctions between formerly hidden objects.

5.9. What are the great benefits of highly speculative hypothetical assumptions?

It is not justified to refuse considering even the most speculative hypothetical theory or assumption for testing because what is the alternative? The alternative is not to vary features. But without it, no improvements in feature selection are possible. It is still better to have a 0.00001% chance of the most speculative hypothetic assumption to change the conditions such that a distinguishing feature gets exposed. Any hypothetic and speculative hypothesis – no matter how unlikely it will be true – is better than the status quo, because it implies changes in feature selection, which is always better than keeping the status quo regarding selected training features. That is why even highly speculative hypothetical assumptions and theories – as long as they do not internally contradict themselves - should not be frowned upon; but instead, they should be very seriously tested. Even if most of them will eventually get disproven, it means progress. Any ruled out hypothesis is progress, because it is a discovery about how aging is not regulated. This excludes many options by giving an example for a way by which aging cannot be manipulated.

5.10. Why are diversity in research parameters, methods, features and workforce, surrounding environment essential for rapid scientific progress?

Instead of getting discouraged, researchers and students should be encouraged and rewarded for developing and testing speculative hypothetical assumptions, because they require inducing variations, which have the potential to expose an - until then still hidden - object, which could be identified at least by one distinguishable feature. If the data-driven approach directs our attention for a specific condition in a certain direction, we have been lucky. But if not, we must not stop varying conditions, just because no data can give us directions.

5.11. What are the best research methods?

Numbers and calculations are only one out of many tools to uncover imperative hidden objects. They tend to work well when available. But that does not mean that we should refuse exploiting alternative discovering methods, such as intuition, imagination, visions, dreams, trends, internal visualizations, analogies and other irrationally rejected feature discovering methods, which are completely number independent. We should explore these non-data-driven numerically independent methods of variation determinations for directional environmental changes at least as seriously as the numeric ones. Otherwise, we unnecessarily deprive ourselves to keep making progress in feature selection in the absence of numerical data. Of course intuition and data should not contradict one another. But no option should be discarded because it is erroneously believed as being “too hypothetical or speculative”.

5.12. Why is the discovery of the magnetic field such an essential analogy for explaining the challenges in feature discovery, feature selection and feature engineering and representation learning, which Artificial Intelligence (AI) faces during supervised machine learning?

For example, in order to uncover the magnetic field from a hidden to an observable object, it takes lots of trial and error variation of the kind described above. One must have magnets and iron objects before one can observe the consequences of the initially still hidden magnetic field (object). Once we have the consequences, we can use the feature and measurement variation methods analog to those outlined above, to hunt for the still hidden causes, i.e. hidden factors/objects. There could be many more imperatively hidden objects (IHO) like the magnetic field, which we cannot sense and hence still know nothing about, even though they could profoundly affect our lives. The magnetic field is a good analogy to effectively communicate the possibility that many similar dimensions are still awaiting their discovery.

5.13. Protein folding example part 2

Let us apply these variation methods to protein folding prediction. If our prediction accuracy is only 30%, we must scrutinize the product, because most likely, it is NOT one – but at least two or more – different objects, which still only look the same within our current observational space.

Apparently, although they all look like the same protein, they obviously cannot be the same object, because they differ in one very significant function-determining feature, i.e. their overall three-dimensional folding. This makes them actually imperatively different objects. Objects are considered to be imperatively different, when it has become impossible to devise a method of measurement or distinction that could erroneously still mistake them as only one object.

In case of proteins, we are unnecessarily limiting ourselves to the dimension “protein” because actually the low folding prediction accuracy implies that – despite them sharing the same primary amino acid sequence – they must be considered as different versions of a protein, which differ in their feature “three-dimensional folding” from one another. If objects differ in at least one feature, they must no longer be considered as the same, but as distinctly different objects.

Why are proteins not treated like RNA? For RNA it is explicit that there are many kinds, with very specific functions and which therefore, cannot be substituted for one another. For example, we distinguish between mRNA, tRNA, rRNA microRNA, etc.

Similarly, assuming that there are 3 protein folding states, we should develop a new nomenclature, which can distinguish between the alpha, beta and gamma folding state.

5.14. Why could evolution favor unpredictable protein folding patterns for the same primary amino acid sequence?

As we know protein folding affects protein function. So what could have been the evolutionary advantages that caused proteins with different folding patterns to evolve? Here, we have no data. All we have is our intuition and insights, which work surprisingly well, if we will stop refusing to develop and apply this insight-based methods much more readily and confidently and stop considering them as to be inferior, less valuable and reliable than data-driven predictions. If one can tell a difference between two objects, they can no longer be the same, but instead, they must be accounted for as two distinct objects, which should no longer be considered as one. E.g. a blue and a red dice are two different objects of the kind dice, but they can never be the same objects when observed by the human eye. These are then inferred as imperatively different objects (IDO). The same applies to proteins of different folding shapes even more so; because not only do they differ in their feature 3D-folding, but also in their feature “function”. Hence, they can no longer be considered as one of the same kind.

As it seems to be the case for all initially hidden objects, we seem to observe the consequence (i.e. differences in three-dimensional protein folding) before its still hidden cause. To find the cause there must be a hidden object or transition during translation, which makes identical amino acids sequences to fold up in different ways after translation. Where could this be useful?

5.15. A highly speculative – but nevertheless still valuable – hypothesis regarding the potential evolutionary benefits of several protein folding states

For example, too high concentration of geronto-proteins shortens lifespan. But too low concentration of the same gerontogenes (i.e. genes, if knocked out, extend lifespan) could interfere with maintaining life-essential functions. That is why their concentration must remain within a very narrow window, too narrow to be adhered to transcriptional regulation alone. There are too many variables, which can affect how much protein is getting translated from a certain mRNA concentration. Hence, instead of a front-end (i.e. transcriptome) we need a back-end, i.e. protein folding dependent functional adjustment. Such kind of a much more sensitive and much more autonomously functioning enzymatic reaction speed adjustment mechanism could work like this:

If the substrate concentration is high, the nascent protein can bind a substrate molecule in its active site even before it has detached from the ribosome. But while still in the process of getting translated, no activator can bind to the protein’s alosteric binding site. This time is long enough for the protein-substrate complex to function thermodynamically like one molecule and fold to reach its lowest energetic state. After the protein detaches from the ribosome, the co-factor can bind, the protein cuts its substrate, but it remains locked in the same folding state as it was when it still formed a molecular folding unit with its substrate.

If protein concentration becomes toxically high the yeast wants to turn off all the mRNA coding for the protein, which is about to rise to toxic levels. Degrading mRNA takes too long. It is much easier to figure out a way to make the excess toxic proteins to fold into an enzymatic inactive state. This can easily be achieved, because enzymatic over-functioning enzymes can quickly process the still remaining substrates in the cytoplasm or at the endoplasmatic reticulum (ER); hence, turning enzymatic functions off in less time. This is a quick way to lower cytotoxic protein concentration. This causes the nascent protein to fail in binding a substrate while still getting translated. The missing substrate causes this protein to have a different lowest energy state and accordingly folds in a different way as if it had bound its substrate. But this is an enzymatic non-functional folding arrangement. The co-factor cannot bind to its alosteric site. Thus, the upwards trends of the toxically rising protein is already getting reversed. But to make this directional concentration change even stronger, the enzymatic inactive folding state allows for a repressor co-factor to bind at its other alosteric binding site. This causes this protein to change its conformational shape again so that it can use its now exposed DNA-binding domain to bind to the promoter of exactly the same gene, which is coding for it; thus, inhibiting its further transcription by functioning as repressor but only in its non-functional conformational folding state.

5.16. Intuition/Hypothesis-driven vs. data-driven research approach

In the example above, no quantitative numerical, but instead, only qualitative observations – based on intuition alone – have been reported to develop a completely legitimate hypothesis, because it can be tested. And no matter how highly speculative and therefore unlikely this hypothetical example differential protein folding hypothesis above may seem, in order to test it, the environmental, measurement and observational methods must most likely be varied in many ways across many dimensions, which would remain hidden without our measuring devices. This increases the odds for discovering an environmental condition and measurement method functional pair by chance alone, which allows a totally unexpected distinctive feature to emerge. That is exactly what we intend.

Different scientists have different gifts, which are great in some, but completely worthless, in other situations. Most researchers tend to feel much more comfortable in employing data-driven numerically reproducible analytical methods. However, a few like me, enjoy intuition-based prediction of hypothetical theoretically possible scenarios resulting from imaginations from situational visions, which are still a much better option than trial and error under circumstances when any kind of numerically reproducible data is completely lacking. As we have seen already, this is a very beneficial and eventually effective approach to rapidly uncover initially imperative hidden object by systematically AI-based multidimensional combinatorial observational dimensionality variation until eventually at least one hidden feature, which defines its hidden object, gets exposed; thus, uncovering the until then still hidden object as being imperatively different from its environment and other objects.

Unfortunately, since we tend to favor numerical over intuitional based prediction methods, dimensions within which we can qualitatively, but not quantitatively, distinguish from one another, remain underexplored or even completely ignored, because no researcher dares to admit that his conclusions are not based on solid numbers.

5.17. What is the ideal life cycle of a new machine learning algorithm?

There is always a need for better algorithms. As we discover more relevant features, according to the methodology described in the previous chapter, we indeed need better and more comprehensive algorithms to account for them. So we will use trial and error and hopefully also some intuition and parameter tuning to improve our F-score. We will again approach an error asymptote, which is greater than zero eventually. But even if we get perfect prediction, this should not be our main final objective, but only means to an end to unravel still imperatively hidden objects. Our work is not done when we have reached perfect prediction, although it implies proper feature selection. But we are never satisfied. As soon as we have the ideal machine learning solution, we want to create conditions, which will cause our algorithm to fail. Why? The reason why we are interested in forcing our algorithm to fail is because we want to explore situations when the assumptions of our algorithm are no longer met. For such kind of situations, we will have more or different essential features, which must account for new circumstances, connectional innovations and perceptional changes in perspectives adequate for addressing a more complex situation, which has previously not yet been considered.

For example, when this research was started in August of 2016, it was still erroneously believed that there are only three kinds of aging regulating genes, i.e.:

1.) Lifespan-extending genes (i.e. aging suppressors)
2.) Lifespan-shortening genes (i.e. gerontogenes)
3.) Genes, which do not affect lifespan.

Dr. Matt Kaeberlein’s lab kindly provided lifespan data for most of the possible gene knockout mutants. Caloric Restriction (CR) extended lifespan in wild type (WT), but shortened it in Erg6 and Atg15 knockouts. The generalization that CR is a lifespan extending intervention suddenly no longer held true for both of our knockouts. Tor1 and Sch9 knockouts lived about as long as WT during CR. Hence, on normal 2% glucose media (YEPD), they are functioning like aging-suppressor genes, but during CR, they are functioning like non-aging genes. This would have caused every machine learning algorithm, which only assumes that an intervention can shorten, lengthen or have no change on lifespan, to inevitably fail, if the genotype feature is not given as part of the training data too. This causes genotype and intervention to become an imperative pair, whose members must not be considered in isolation, when training a more predictive machine learning algorithm.

Let's say that our machine learning algorithm was only trained on WT data to classify into three broad categories, i.e. lifespan extending, shortening or not changing interventions. Then CR would always extend lifespan. But if we – instead of WT – apply CR to the Atg15 knockout – its lifespan shortens through CR. Our algorithm would fail because it was not trained on knockout data. This kind of predictive failure is not at all a bad thing - but instead a blessing in disguise, because it is teaching us that apart from the feature intervention, there is also the feature genotype, which affects lifespan and which must be considered together with genotype like an indivisible unit-pair of atomic data, whose components must never be evaluated in isolation. We only could notice it because our only WT data trained AI imperatively failed on predicting the impact of CR on Atg15 knockouts. From then onwards we know that for correct prediction genotype and intervention must be given together as a pair to train our artificial intelligence (AI). This allows us to establish that apart from intervention, genotype is another essential feature for correctly predicting lifespan. So far, we only trained our AI on glucose media. Since it was the same for all the training sets this feature was not yet essential as long as it could only take on the same value. But when testing it on galactose, tryptophan or methionine deficient media our algorithm will imperatively fail again because now we need to consider a triplet as one piece of information, i.e. intervention, genotype and media. Only if we train our AI on indivisible triplet unit pairs it can succeed. We just have shown how intentionally creating variations in the condition can reveal new hidden objects but only when a naively perfectly working AI suddenly starts failing. But without naïve AIs to have failed we could have never discovered this new feature. Hence, causing perfectly scoring AIs to fail is a very good method of choice for discovering new features.

However, if everything is true as described here, why was our attention never drawn by a single peer-reviewed paper, which looked at these issues from a similar perspective? The protein folding prediction provides plenty of regulatory scenarios, which can be hypothesized and subsequently tested. For example, we know that the speed of translation depends on the charged tRNA ratios in the cytoplasm and at the endoplasmatic reticulum (ER) as well as on mRNA binding affinity to its translating ribosomes.

For example, we know that three tryptophans in a row cause translation to stop prematurely since the concentration of tryptophan-charged tRNAs is too low for continuing translation on time. Using our newly derived machine learning feature selection, methodical and environmental toolboxes, we would assumed that we can see a consequence, i.e. premature translation abortion, for which we must now start looking for the still hidden cause. However, in this particular case, the obscure reason for the abortion is not even hidden, because the mRNA nucleotides coding for the three tryptophans can be clearly and easily measured and observed. But this tryptophan triplet, i.e. these 3 identical - yet still distinct - objects, started to form a kind of conceptual super-object possessing completely novel properties/features that none of its three individual units posses even in small parts on their own. This totally unrelated qualitatively completely novel dimension, which is totally lacking in any of its parts, has a gain-of-novel-function effect; i.e. it terminates translation. Hence, these three termination causing tryptophans form a new shapeless super-object, on a whole different level/dimension, which cannot be accounted for by simply adding up the properties of the three tryptophans individually. Their mode of action to stop translation is of a much different nature and mode of action than their complementary codon-based translational mRNA/tRNA binding. The three tryptophans possess a new quality that cannot be distributed to each single tryptophan member alone.

It is kind of like we humans, who keep adding a lot of dimensionless, shapeless and physical matter-independent virtual features, based on which we distinguish between each other, which may be hard for AI to grasp. E.g., based on our first, middle and last name, SSN, citizenship, job, family role, etc., we make big differences between ourselves, which affect lifespan. Unfortunately, AI could not discover those, unless it can add the feature to perceive spoken and written communication. This is the only way, by which our virtual self-imposed physical dimensionless features, can be distinguished from one another.

5.18. Why will feature discovery never stop?

The new feature discovering cycle will never end because as soon as we think we have got it to work, we hope to succeed in creating an exception, which causes are newly trained AI to fail, since this allows us to discover even another new relevant feature.

We started out with the feature “lifespan intervention (e.g. CR vs. YEPD) and discovered the no longer hidden objects/features “genotype” and “food media type”. The next Kaeberlein yeast lifespan dataset had features like temperature, salinity, mating type, yeast strain, etc., which also affect lifespan. Now for one loss-of-function mutant, we could have more than 10 different reported lifespans. This would make the concept of purely aging-suppressing gene or geronto-gene obsolete. This, in turn, would raise the number of components, which must be considered together as an indivisible atomic unit, of which none of its parts must be considered in isolation, to consist of already seven components that must be given with every supervised input training sample for our AI. If this trend keeps growing like that, then the number of components, which form a single data-point like entry, keeps growing by one new component for every new feature discovered/added. But would this not cause our data points to become too clumsy? But even if it does, for every new feature, which we decide to consider, our indivisible data unit must grow by one component. However, this would mean that 10 essential features would create data points of 10 dimensions. If we keep driving this to the extreme, when considering 100 new features, then we have 100 dimensional data points. But this would almost connect everything we can measure into a single point. This would put away with independent features because their dimensions will all get linked together.

From this chapter we can conclude that the best AIs are those, which fail in a way that allows us to discover a new feature for subsequent feature selection.

5.19. How many still hidden concepts are still separating us from reversing aging?

But how many of such kind of essential key features breakthrough discoveries are we still away from solving and reversing aging? The lack of progress in extending lifespan by more than 50% indicates serious problems with feature selection. What does it take to make our experimental life scientists to please understand this essential feature selection concept through variation and to consider it in their experimental design? When this concept was published online in October of 2017, it became the most read material from UALR and has remained the most read contributions from the Information Science Department. Since then, the contributions to this topic have engaged more than 350 readers per week on www.ResearchGate.net. Occasionally, more than 200 readers were counted on a single day. The contributions to this topic received 20 recommendations last week. Only because of the strong encouragement and conceptual validation by researchers with very good reputations and impressive peer-reviewed publication track record, who took the time to answer conceptual questions at www.ResearchGate.net and www.Academia.edu, caused the necessary gain in self-confidence, which is needed for spending lots of time on working and revising this absolutely non-mainstream manuscript, since its authors are convinced that this is the only way to raise our chances as a social species to excel scientifically and improve methodically to accelerate our overall knowledge discovery rate and research efficiency to accomplish true immortality and permanent rejuvenation into feeling forever young, healthy, strong, energetic, optimistic and goal-driven within our potential reach of achieving this dream (i.e. currently a still deeply imperatively hidden object) within the upcoming 2 decades, if the recommendations outlined in this and other related manuscripts in preparation are not only widely considered but enthusiastically implemented, applied and further enhanced by all stakeholders. This manuscript intends to change the way research is conducted by minimizing the time periods during which everyone feels confused by providing highly effective guidance for overcoming the limitations posed by still imperatively hidden objects/features/factors/causes/elements.

5.20. What kinds of datasets are needed to reverse engineer aging?

The best scenario would be to measure every 5 minutes through the entire yeast’s lifespan its transcriptome, proteome, metabolome, microbiome, epigenetic, lipodome, automatic morphological microscope pictures, ribogenesis, ribosomal foot printing, DNA chip-chip and DNA chip-seq. analysis, speed of translation, distribution and ratios between charged tRNA in cytoplasm, length of poly-AAA-tail, non-coding RNA binding, autonomous replicating regions (ARS), vacuolar acidity, autophagy, endocytosis, proton pumping, chaperon folding, cofactors, suppressors, activators, etc.

5.21. What is the temporal alignment rejuvenation hypothesis?

Temporal alignment between genes, which must be co-expressed together, and genes that must never get co-expressed together, e.g. sleep and wakefulness genes, is getting gradually lost with advancing age. Generally during preschool age, it felt that there was almost no time gap between falling asleep in the evening and waking up refreshed the next morning. Ten hours of time seemed to pass by every night like an eye-blink.

Unfortunately, middle aged adults no longer feel this way. They are aware that their sleeping time is spread out over many hours. This is because their temporal alignment between their initially i.e. during early childhood still perfectly co-expressed sleeping and wakefulness, genes, which MUST never be co-expressed simultaneously with one another, because sleeping and wakefulness are mutually exclusive since they inhibit one another very profoundly, is getting gradually lost due to increased stress-levels during advancing age.

One can either sleep or be awake but not both simultaneously. Unfortunately, it is unhealthy to partially sleep while remaining partially awake at the same time. This prevents the biological processes and molecular functions, which require restful REM sleep, from rejuvenating older individuals. This is because not all wakefulness genes are turned off at night time anymore while not all sleeping genes are turned on together when they need to work together during sleep, like a single physiological functional unit, similar to all Gene Ontology (GO) Term member genes, in order to provide all the life-essential benefits of sleeping properly.

Similarly, during day time not all sleeping genes are turned off properly while – at the same time – not all wakefulness genes are turned on completely during daytime. This interferes with mental clarity, focus, concentration and results in forgetfulness. It has already been scientifically proven that the circadian rhythm is a very important marker of biological age and highly indicative of the remaining lifespan ahead. This applies uniformly for humans, animals, model organisms, such as mice, flies, worms, fish and even yeast.

The circadian rhythm is essential in balancing the need to perform well during the day while properly recovering during nights. If this is true, then aging related decline could be reversed simply by realigning the temporal regulation of gene expression pattern to what they were during the first years of elementary school. Only after proper sleep, anyone can perform at his/her desired peak during daytime in researching, working discovering, studying, experimenting, writing, driving, biking, exercising, computing, etc.

6. Example how uncovering hidden objects/features, feature discovery, feature selection and subsequent supervised training of a machine learning algorithm could work?

Perfect prediction by computation optimization must be impossible as long as essential input training data features for supervised machine learning are still lacking. Therefore, we emphasize not to invest much effort into algorithmic computational optimization before the feature discovery process allows selecting all needed features through variation before realistic predictions can be achieved by enhancing computations and parameters. This simple progression from completed feature selection to optimized algorithm is essential for more rapid discoveries. That is why we are still wondering why we have seen so many papers focusing on improving computational predictions without any kind of prior considerations about having properly completed the absolutely essential exhaustive feature selection process, without which no subsequent computations can lead to satisfactorily predictions, which are reasonably consistent with our experimental observations and measurements.

We are worried about being the only author team, to whom the preceding writings above make sense. We expected much more enthusiasm, excitement and optimism about very likely accelerating our hidden feature and object discovery rate by first focusing on uncovering still hidden objects and features through diverse variations of conditions, procedures, methods, techniques and measurements, followed by the exhaustive selection for all relevant needed features, followed by designing, developing, combining and optimizing the computational steps of our machine learning algorithm until our predictions match our experimentally obtained observations. Once this major machine learning objective has been achieved we have reached its final status beyond which we cannot improve it unless we can generate conditions, which cause our previously perfectly predicting machine learning algorithm to obviously fail, because this is an absolute prerequisite for discovering more relevant essential features to be selected reflecting more complex and higher feature dimensionality and complexity, as we have encountered while trying to lay the conceptual framework for permanently reversing all adverse effects of aging.

Any newly discovered essential input data feature inevitably causes a rise in the dimensionality of input data components, which must be considered together but never in isolation. For example, if we train with 100 input features, our input variable must consist of exactly 100 components or parts, which together form a new level of single measuring points, which tend to be much different in controlling their manipulations and their overall effects from anything, which could possibly get anticipated, when trying to add up the effects of its 100 parts to a new total. This new total tends to consist of many different dimensions and often refers to completely unrelated kind of data than when combining all 100 components consisting of exactly the same input values for every of their 100 variables, but by considering all 100 components like a single indivisible unit of measurement points, which often results in completely different kinds of unrelated seeming properties/features, which are not even closely reflecting the results, which we'd obtain if we executed each dimension on its own in isolation and in sequential order.

For example, stopping translation prematurely by three consecutive tryptophans has a much different impact, i.e. stopping translation prematurely, than when translating each of the tryptophan in isolation separated by other amino acids, since this causes the nascent polypeptide chain to grow.

Each tRNA charged with tryptophan, which complements mRNA triplets, causes the peptide to grow by a single tryptophan, which gets added to it. So when you try 2 tryptophans, then the polypeptide grows by two amino acids. But when you try 3 consecutive tryptophans, then - counter-intuitively - instead of the expected growth by 3 amino acids - translation prematurely stops. Stopping translation prematurely is of a much different dimension, level, effect and data kind, then when keep adding more amino acids to the growing peptide chain. If we consider the effect of complementary binding of a tRNA to its mRNA codon, our peptide grows by one amino acid; any charged tRNA adds another amino acid of one of 20 categorical values. Normally no amino acid can cause the translation to stop prematurely, not even two amino acids as a pair. But three amino acids, as an indivisible triplet, which must be considered as a new single value, requiring all three tryptophans to be sequentially present, like a single indivisible unit data block, which must NOT be divided into smaller groups other than triplets, because only triplets, but no pair or singlet, can stop translation prematurely.

Another example is predicting overall cellular protein composition. It depends on how many mRNA strands coding for a particular protein are in the cytoplasm. There is proportionality between number of cytoplasmic mRNA strands and total protein abundance. Therefore, if the cell needs to double protein abundance it could double transcription and keeps everything else the same. But a much better and less step intensive, more economic way of doubling protein concentration is to double the length of the poly-(A)-tail. Extending the length of the poly-(A)-mRNA-tail may require about 100 additional adenines whereas doubling transcription requires about at least 500 - instead of only 100 - new nucleotides in addition to all needed transcriptional modification steps with their elaborate synthesis machinery.

If the dividing yeast must raise its lipid synthesis by more than 10-fold during the short M-phase, it could increase transcription by a factor of 10, it could make the poly-(A)-mRNA-tail 10 times longer, or it could synthesized 10 times more new ribosomes to increase the ribosomal translation by a factor of 10 simply by reducing the distance of free uncovered mRNA nucleotides between adjacent ribosomes translating running down the same mRNA strand. If more than one ribosome is translating the same mRNA strand simultaneously, it is called a poly-ribosome or polysome. Hence, having 10 times more ribosomes binding to the same mRNA strand at the same time increases translation by a factor of 10, without needing any additional transcription.

Above we have given three easy examples to get 10 times more proteins. Although all 3 methods have the same final result, i.e. 10 times more proteins, their mode of action, their required essential features, their dimensions and their minimally required parts, which must be considered like a single value, are totally different.

If the cell employs all three options simultaneously it can raise protein abundance by a factor of 1,000 during the short only 20 minutes long M-phase. The relevant essential input feature for poly-(A)-mRNA-tail-based input to predict protein increase is simply the number of adenines added to the tail. The only essential selected feature is a simple integer numeric not requiring any dimensional specifications since only adenines can be added. But we should note that unit of the required feature is number of adenines.

However, when increasing transcription the input feature is number of mRNA strands. Note that the number of mRNA strands cannot be directly converted into number of added poly-(A)-adenines. Synthesizing an additional mRNA strand affects protein abundance by a different mechanism and amount than adding an extra adenine to the tail. There is probably a way to experimentally figure out how many more adenines must be added to the tail to increase protein abundance by the same factor as synthesizing an additional mRNA strand.

The input feature for ribosomal coverage is an integer of the unit ribosome. Adenine, mRNA strand and ribosome are different feature dimensions. We could now experimentally figure out how many additional mRNA strands need to be transcribed to increase protein abundance by the same amount as adding a single new ribosome. Then we could figure out how many adenines have the same effect as a ribosome and how many adenines have the same effect as an additional mRNA strand and how many mRNA strands have the same effect as a ribosome on overall protein concentration increase. This will give us a nice conversion table. This gives us fixed adenine to mRNA strand, fixed adenine to ribosome and fixed mRNA strands to ribosome ratios based on which we can make meaningful predictions in each of the different dimensions, which contribute to protein abundance by completely different and unrelated modes of actions, i.e.

1.) Adding an adenine,
2.) Transcribing a mRNA strand,
3.) Synthesizing a ribosome.

To simplify assuming that translation rate can only be affected by varying length of poly-(A)-tail on mRNA, transcription and ribosome synthesis rate, which essential features do we need to train our machine learning algorithm?

Answer: We need 3 input features. i.e.:

1.) Number of adenine of the dimension adenine
2.) Number of mRNA strands of the dimension mRNA strands
3.) Number of ribosome of the dimension ribosome.

For each of these three dimensions we will get an integer input value. Based on our previously experimentally determined calibrated conversion table between the translation rate affects of the input features, i.e. namely adenine, mRNA strand and ribosome, we can predict total protein abundance. The total protein abundance should not be affected by whether or not we are considering adenines, mRNAs and ribosomes in isolation and sequentially or combination of triplets or combinations of twin pairs because each of these three dimensions and their mode of action can function totally independently from one another.

An example for the pair or triplet unit concept is given below.

For any wild type (WT), caloric restriction (CR) is a lifespan extending intervention compared to YEPD (normal yeast growth media). As long as this holds true always CR is a reliable way to extend lifespan. We could train a machine learning algorithm, which predicts lifespan only based on the input values, CR or YEPD assuming we have WT. This is a very simple binary classifier. As soon as we got it to work we want to cause it to fail. To accomplish this we vary lots of features, e.g. protein, genotype, temperature and whatever else we are capable of. We will keep doing this until we find an instance, where our algorithm predicts an extension whereas our observation shows a shortage in lifespan.

If we find such a case, we must scrutinize the yeast cell, for which CR truly extended lifespan and compared it with the other cell, for which CR suddenly shortened lifespan. The fact that the same manipulation, i.e. CR has opposite effects on both still very similarly looking phenotypes, must make us understand that both yeast cells must no longer be considered to be an object of the same kind. The fact that CR affected both instances of yeast in opposite ways makes it imperative that both yeast cells must differ in at least a single feature from one another.

Our next task is to compare both cells until we find the defining and distinguishing difference(s) between them because without such kind of a difference CR could not have opposite effects on two exactly identical instances of yeast cell. After carefully checking we notice that both yeast cells are fully identical to one another, except for their Atg15 sequence, in which they differ. This was a very simple example of essential input feature discovery. Let’s assume that we are conducting such kind of feature discovery before having formed a concept of a gene. In this case, we have 2 Atg15 sequences. For the first one, CR extends lifespan, but for the second one exactly the same kind of CR shortens lifespan by at least 7 replications. This discovery causes our concept about lifespan extending interventions to become obsolete because of a single example, where a difference in Atg15 nucleotide sequences causes CR to shorten lifespan. When we look at protein abundance we can easily see that the lifespan of the Atg15-less yeast gets shortened by CR whereas the lifespan of the yeast with Atg15 protein (i.e. WT) gets extended exactly by the same kind of CR. We have succeeded in finding a reproducible causal relationship, which is causing our single dimensional input feature, CR, or YEPD to fail every time when the phenotype lacks the Atg15 protein. We have just discovered a new feature! Congratulations!!! Whatever difference or change causes our old machine learning algorithm to reproducibly fail in the same manner by the same distinct input parameters that the dimension in which they differ from one another must be our newly discovered feature, which we must include in our feature selection before we can return to retrain the machine learning algorithm based on both instead of only one feature.

As long as we only had a single genotype, i.e. WT, the input-feature genotype was not needed because it was the same for all the encountered instances of yeast cell. Since the genotype was always the same, i.e. WT, the object or feature "genotype" remained still hidden because it could not be used to distinguish between yeast instances as long as in all cases the genotype is WT. In this particular case the object "gene" itself, may not be hidden, because it consists of physical matter, which we can measure but as long as this feature was always WT, it could not show up as a feature unless it can take at least two distinct values. As soon as we discovered that the visible object genome differed in their feature Atg15 protein present or absent, we must recognize that we must provide our algorithm with a second data dimension, because CR shortens life in Atg15 knockout while lengthen it in WT. We have discovered the first example, in which the visible object Atg15-coding gene could take two distinct values, either knockout or WT. This puts us into a good position to proceed with gene based feature discovery until we succeed in knocking out Erg6. Again, for the Erg5 knockout CR shortens lifespan to less than 10 replications whereas it lengthens lifespan by about 4 replications for WT. CR can no longer be considered a lifespan extending intervention because - in order to train an algorithm with supervised learning on predicting lifespan effects we must provide now 2 dimensions, i.e. 2 training input features, i.e. glucose concentration and genotype. In this example the 2 components, i.e. glucose and genotype, must be considered as an indivisible informational pair. When considering any of its 2 components in isolation proper supervised learning and correct prediction are impossible. Only by considering both components (dimensions) together like a single indivisible measurement point, allows for proper input feature selection.

Let’s assume all measurements described above were performed at 30 degree Celsius. As long as we only have a single and always the same value for temperature, we can measure it, but it remains a hidden feature, until we start to vary it. Let’s say heating up to 40 degree Celsius generally shortens WT lifespan due to the heat shock induced improper folding experience, but that for some knockouts raising the temperature from 30 to 40 actually increases lifespan. This will result in a three dimensional input vector.

Important: It was recently discovered that when the different input features cannot be converted into one another by a kind of conversion table as we had for adenine, mRNA and ribosome and hence yield the same results regardless whether we consider the features separately and in sequence or together as a single indivisible unit because the mechanisms of actions for each dimension don't depend on one another and can be performed completely independently without having to share any scares resources.

However, when we have selected input features/dimensions, which can never be converted in one another by a simple experimentally obtained proportion ratio table, as it is clearly the case for CR vs. YEPD, knockout vs. WT, or 30 vs. 40 degree Celsius, then our three-dimensional input variable, which consists of a glucose, genotype and temperature component, both of which are Boolean variables in this example, all three components must be considered together like a single indivisible unit consisting of components, which must never be evaluated in isolation from one another, and they must form a single value providing a unit input feature value, in order to make proper lifespan predictions.

It is very similar to our numeric system. Let’s say we have the numbers 1, 2 and 3. This is another example for case above where 123 is not the same as processing 1, 2 and 3 in isolation and sequentially. If we have a three-digit number, we must always consider all three digits at a time to make good predictions, similar to predicting glucose, genotype and temperature always together without ever splitting it apart into its single components.

Now let’s say we have A, B, and C, it does not matter in which component order I consider these three letters and whether I process them as a single unit or one after the other, the result should always be the same unless they can form different Words.

Feature discovery, feature selection and machine learning algorithm training and tuning must be performed as three discrete, separate steps, life the waterfall model, which requires that the previous step must be fully completed before the subsequent step can be started. Since improper feature selection has held us unnecessarily long back because somehow the main influencers, who have the biggest impact on the way research is conducted and studies are designed, seemed to have overlooked the fact that computational model prediction can only work if all of the needed input features have been selected properly. But this realization is so basic, fundamental, obvious and self-explanatory that it would be common sense to follow it implicitly. But from my literature searchers I remember many papers discussing the effects of variations in calculation on predictive outcome. But I cannot remember any paper, except for my NLP poster last summer, where feature selection was explicitly used as a method to improve predictive outcome.

The main problem is that I am almost blind and can only read a paper per day. Actually, I can type this much faster than I can read it. This could mean that feature selecting papers are also plentiful, but that I did not get the chance yet to discover them. However, I am almost certain that nobody has ever used my concept of hidden objects and features and written out detailed examples for feature discovery, feature selection, algorithm training leading to almost perfect prediction. If perfect prediction has been accomplished, then we are actively searching for conditions, which cause our new predictor to fail. Such kind of failure is caused by a still undiscovered difference between two objects, which have been considered to be exactly the same until a reproducible difference of the same treatment makes these two instances 2 separate imperatively distinct objects, which must no longer be substituted for one another because they must differ in at least one single feature from one another. We must compare all aspects of such objects until we discover the feature that allows us to distinguish between them. This new feature, by which they differ, is a new essential feature, which must be added to the set of selected features before any kind of adequate predictions are possible by tuning our machine algorithm with another new input feature and dimension of input training variable.

7. About the need to centralize our genomic research efforts to focus on creating complete multi-dimensional datasets to win the War on Aging

Are we ready to win the war on aging? Do we already have the weapons to defeat death?

In the 1960s a lot of resources and research was directed to fight the “War on Cancer” with initially very primitive tools. However, in contrast to us today, the researchers in the 1960s aimed to accomplish their objective to test every substance or compound about its anti-cancer effects. In contrast to us, anti-cancer-researchers did not have the luxury of data and tools to which we have access today. Back then, nobody dared to imagine proposing a new data-driven-research approach. Lacking any data for making inferences about the carcinogenic or cancer-killing effect, they had nothing to rely on, except for their intuition and the commitment to systematically test all the compounds, which they managed to generate. Yet, they gradually succeeded until this very day. Each new compound tested functions like a feature of the object cancer.

Basically, what our parent's generation did intuitively without being aware of it was to vary feature selection every time a chemical compound failed to show promising anti-carcinogenic effects. They implicitly agreed that selecting their anti-cancer agents by random chance alone. They saw no point testing the same compound twice after it had failed once. But today, at least part of our research community seems wanting to stick to their old proven methods and keep analyzing the same features over and over again despite having failed more than 10 times in the past already. We, i.e. the species of homo Sapiens, would have had developed the necessary and sufficient tools for understanding aging much better, if we had taken the same approach as the researchers in the 1940s, who developed the first two nuclear missiles in Los Alamos.

Ironically, the Fuehrer Adolf Hitler, who caused the death of more than 50,000,000 people, caused more fear, resistance and counterattacks than the 100 times faster killing mechanism of aging, which inevitably results in death. Our planet is home for more than 7,000,000,000 people. This means that Adolf Hitler is responsible for the death of a little less than 1% of the total human population on Earth. However, when comparing Adolf Hitler's killings with the mortality rate due to aging and death, Hitler looks almost harmless, because aging kills at least 100 times faster and 100 times more people than Adolf Hitler's entire inhuman World War II. In contrast to today, during World War II, there was a widespread common implicit consensus that every measure to stop Hitler's killing machinery is worth the effort. We must therefore, conceptualize the "Mechanisms of Aging" as being at least 100 times eviler, dangerous and deadly than Adolf Hitler and his World War II was. Aging is a 100 times faster killing machine than the Nazis. But why gave humans so much attention to Adolf Hitler, who is still 100 times less harmful than death? This shows how irrationally most instances of Homo sapiens make their decisions. That would never happen if the priorities were set by Artificial Intelligence (AI).

At least back in the 1940s, the government took the initiative to bring as many bright researchers as it could find to Los Alamos, New Mexico, USA. Their only task was to keep trying and researching until the first two nuclear missiles were waiting to execute their deadly missions in the two Japanese cities of Hiroshima and Nagasaki. It may be true that death may be 100 times harder to defeat than the Nazis. Unfortunately, only an extremely small minority appears to be seriously disturbed, concerned and worried about stopping Aging from eventually killing all of us inevitably even 100 times faster than the Nazis.

But unfortunately, most people seem to be too complacent and stuck in their old obsolete concepts that they don't even consider opposing death. In World War II there was a central command, which was capable of focusing all resources and bright minds on accomplishing the most urgently perceived objective, i.e. to build the first nuclear bombs to speed up the victory against Japan. If research would have been as decentralized as it is today, where very small groups of researchers struggle to duplicate, triplicate and even redundantly reproduce each other's works under slightly varying conditions, which unfortunately, makes their data incompatible for combined common data analysis.

Imagine UALR, UAMS, ASU, Louisiana Tech University, University of New Orleans, Tulane, Harvard, MIT, Yale, Princeton, etc., were each assigned to build the first nuclear bomb today, i.e. 73 years later than it was actually built. Even with today's much higher technical capabilities, no research group or university could succeed on its own all alone, because governments can deploy necessary resources, which no university or research entity ever could.

8. What have the atomic bomb and anti-aging research in common?

To invent the nuclear bomb, a critical mass of resources is needed at a single location. Similarly, to stop aging from being 100 times as deadly as Hitler, one needs an even 100 times more concentrated focus of energy, HR, material and equipment to win out in the end. Today, 73 years after the first two nuclear missiles were fired; no American Legal Entity could deploy the necessary and sufficient resources to build the first nuclear bomb. Unfortunately, nobody thought about keeping this much more effective centralized research structure from World War II to fight the War on Cancer and Aging.

Unfortunately, this meant 73 years of only suboptimal slow progress. Since researchers accomplished 73 years ago, which we cannot accomplished today with our decentralized research funding structure, it means that if we had at least kept parts of the centralized structure for large projects, we could have accomplish technological, medical and social objectives, which may not be available for any of us for the next 73 years if this trend persists. Imagine, you could travel 73 years, i.e. an entire human lifespan, into the future, how much more technical, medical, lifespan extending, rejuvenation, entertainment and other options would be available in 2091, which nobody could even dream of back in 2018?

Maybe 73 years from now immortality is already reality. Unfortunately, we – who are living today – are the victims of the complacency and indecisiveness of decision makers because we must pay for it with our lives since most of us won’t be alive in 73 years anymore. Since the government does not seem to be inclined to assert the same leading role as in World War II, researchers must act on their own to bundle their resources together and focus them on rapidly defeating aging. In America alone, we have more than 20 yeast, worm, fly, mouse, mosquitoes, E-coli, HIV, cancer, Alzheimer, Parkinson’s, Diabetes, etc. labs.

If the top 20 labs would dedicate all their grant money towards generating a master-dataset of the highest quality at the highest temporal resolution with the maximum –omit dimensions, not exceeding intervals of 5 minutes between measurements for a lifetime of the model organism, we could probably figure out how epigenetic changes are brought about and interact with other cellular components, functions and processes.

No university and no lab can achieve this mammoth milestone on its own. Therefore, our data is much more incomplete than it could have been if – maybe even for only a month - all disease and life-extension researchers would gather in Los Alamos to produce master-omics datasets of as many species as we can, including humans.

Ironically, the total number of experiments, funds, other lab resources, etc, needed for creating master-omics-datasets for each species is far below our current spending. It is better to have one high-dimensional –omics wild type (WT) time series dataset, spanning the entire lifespan with extremely high temporal resolution of less than 5 minutes between time points than hundreds of smaller low-dimensional –omics datasets.

Unfortunately, the multitude of much smaller and less-dimensional datasets produced, when considering all decentralized research teams together, is at least 100 times worse than having a single high-quality master-dataset, which everyone can use. This has the advantage that all –omics disciplines/dimensions would be measured by exactly the same methods, under the same environmental and experimental conditions, and temporally properly aligned well enough for discovering much more causal relationships and interactions between cellular processes and matter without which we have no chance of defeating aging and cancer in our lifetime.

The problem is that people refused to do it unless they are forced to. How can it help me to have access to hundreds of microarray datasets when I cannot consider them together because of differences in their data acquisition methods, reaction environments, media, growth conditions, etc.? This makes the timely proper integration of –omics data from different dimensions practically almost impossible.

I just realized today, while for the very first time outlining the global war on aging in writing, that we could have been technologically 73 years ahead of today. In my dissertation I intend to describe methods to speed up hidden feature discoveries by almost randomly varying methods, conditions, genotypes, phenotypes, etc. until new initially still hidden features emerge. Research must be much more centralized. It is sufficient to have one expert group for each technique, skill or method in the nation, which can travel to campuses and train students, faculty and Principle Investigators (PIs) in the latest techniques, methods or skills.

Currently, we have a lot of graduate students, who know some programming, tool usage, bioinformatics pipelines, analytical and modeling tools. Unfortunately, since those graduate students had to struggle a lot on their own to figure everything out, they cannot be expected to be perfect even by the time they graduate. Computational training could take place remotely and hands-on lab training could be performed by the mobile expert team, which is ready to train newcomers on demand.

Since it almost does not matter how many people attend a webinar or participating remotely in a presentation type Power Point Lecture, the NSF and FDA could gradually transition to the Global Science Foundation (GSF) and The Global Institute of Health (GIH), respectively. This strategy would be much more efficient than implying nuclear threats because nations invited to webinars are much less inclined to threaten war. This is the end of my description and dream about mutually shared implicit insights, which would have allowed for an even higher productivity than during World War II; thus, it would have raised our chances for succeeding in escaping aging and death.

9. The Evolution of Aging

Aging could be regulated by the interplay between many different kinds of data-dimensions, all of which provide a fraction of information and dependencies, which must be manipulated in such a way that our evolved internal suicide clock, which is most likely driven by our developmental genes, can not only be stopped but also reversed, because our lives should no longer depend on a kind of evolution, which selects for mechanisms that cause our lifespan to be finite.

Long time ago, back in the RNA world, evolution could not select against an individual RNA strand without adversely affecting its replication rate. Because back then, everything, which helped the RNA strand to withstand degradation and stressors, also helped its replication. Hence, there was no distinction between the individual and the replication-relevant material, since both were exactly identical and therefore, they could not be separated.

But now evolution can select against individual parents without adversely affecting any relevant aspect of replication. As long as the entire individual was completely composed of exactly the same matter, which was essential for replication, e.g. an individual RNA strand, there was - by default - no aging at all - but instead - only replication.

Aging could only evolve in the protein world because then not all the physical matter, of which the parents consisted, was essential for replication anymore. Only this distinction allowed evolution to select for active killing programs, which are most likely driven either directly by actively programmed destruction mechanisms, e.g. apoptosis, or indirectly by neglecting to maintain, repair and restore essential functions, e.g. chaperone-aided protein-folding, peroxisome degradation, or maintaining the steepness of the needed proton-, salinity-, ion- and nutrient-gradients across membranes because our evolved in-built suicide clock killed faster than those life-essential processes declined enough for posing a threat on life.

The life-cycle, i.e. the time span from birth to death, seems to be very similar to the cell cycle because it appears to consist of long phases of relative stability and little change interrupted by short periods of rapid changes, which can be as drastic as metamorphosis in species, like worms, flies or frogs, but which nevertheless can be found to a lesser extend in all species. The periodic interval pattern of changes is too similar across members of the same species to be solely the result of the much more randomly acting wear and tear process alone.

Women, for example, lose their ability to have children between 50 and 60 years of age. This low variation makes it impossible for this loss of function being caused by wear and tear alone. The same applies to the lifespan. Its variation between members of the same species is way too small for claiming that its length is determined by wear and tear alone. Therefore, I believe that it is likely that there is actually an actively regulated and well timed transition mechanism, which works similar to cell cycle checkpoints, from old age into death.

Such kinds of questions are of interest to me and they keep crossing my mind when analyzing time series datasets because they could help to elucidate the mechanisms of aging. And we must understand them before we can effectively disrupt them.

We need to start thinking about initiating mechanisms similar to targeted and directed, i.e. intelligently designed and goal-driven evolution, which is aimed at maintaining and restoring all life-essential processes or substituting them accordingly, if they cannot be maintained in the way they have initially evolved. We need to become fast enough that

- if we see a particular approach to fail - we'll still have enough time for quickly developing much better alternatives for preventing the otherwise unavoidable -seeming aging-induced decline, which would inevitably kill us.

Please let me know if you want to contribute towards completing this project. If you do, please add your name and institution to the very beginning of this writing right underneath the first major heading.

Please direct questions, comments, suggestions and recommendations, i.e. especially how to get this published in a peer-reviewed scientific bioinformatics journal or how to find conferences, which offer travel grants, to which I could present it either as talk or poster to me by any of the communication options listed below. Thanks a lot in advance for your interest, time, help and assistance.

10. Funding acknowledgment

This project was supported by the Arkansas INBRE program, with a grant the National Institute of General Medical Sciences, (NIGMS), P20 GM103429 from the National Institutes of Health.

References

1. Janssens GE, Meinema AC, González J, Wolters JC, Schmidt A, et al. (2015) Protein biogenesis machinery is a driver of replicative aging in yeast. Weis K, ed. eLife: 4:e08527.

2. Coffey JC, O'Leary DP (2016) The mesentery: structure, function, and role in disease, The Lancet Gastroenterology and Hepatology 1: 238-247.