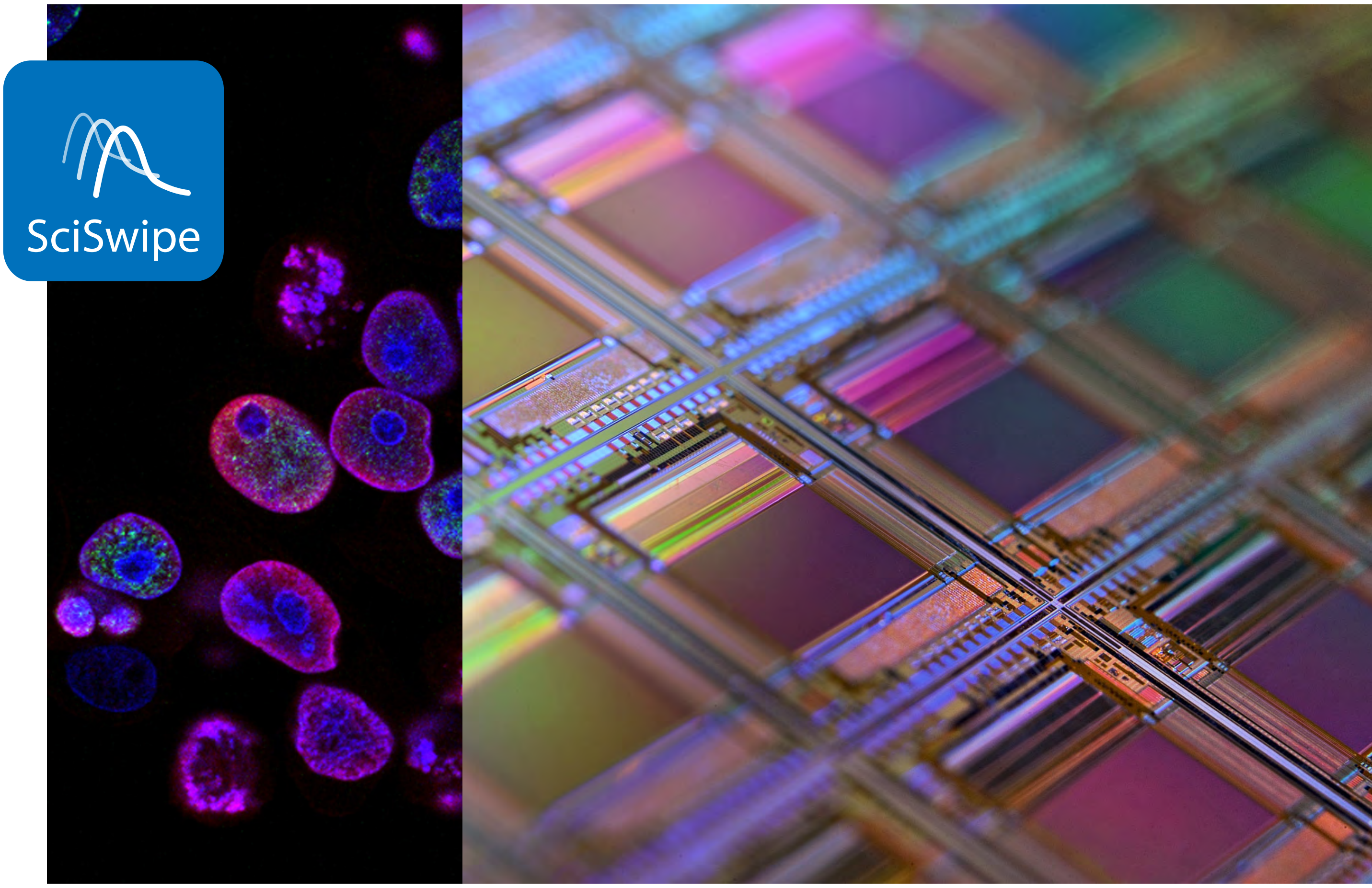


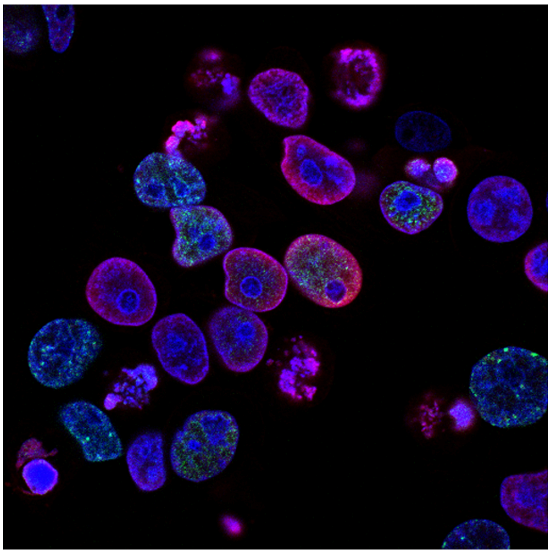
Unlock The Value Of Complex, Large-Scale Data

A white paper for the life sciences on how to enhance the analysis of complex, large-scale data using human-algorithm synergies.

www.sciswipe.com

Summary

When we sequence human genomes, we generate complex, large-scale data - billions of bytes of sequences of characters. This is just one of the types of data that researchers across academia and industry generate in-house and derive from secondary sources on a daily basis. While MRI image data can enable the detection of a tumor, Next Generation Sequencing (NGS) data may hold clues to pinpoint mutations that contribute to the development of a particular disease. Although major manual efforts as well as computational advances have led to large collections of segmented images across disciplines, the complexity and nature of medical images hinders similar developments in the healthcare domain. Such data necessitates a costly, manual expert analysis, making analysis hard to scale, time-consuming, and potentially error-prone. Along these lines, certain types of data that do not rely on expert analysis are unfit for the processing through computational methods employed to date, be it based on statistical analyses or machine learning, as information is lost in the high level of noise. Particularly in situations where no prior information is available, state-of-the-art computational approaches and in-depth analysis face compatibility issues. SciSwipe scalably and generically transforms both medical images and omics data to visualizations that are optimized for humans to detect patterns in. Human pattern detection is algorithmically supported to provide speed, scale, and reliability. Outsourcing specific sets of tasks to humans and to algorithms in synergistic software pipelines enables the exploration of complex data at unprecedented depth and scale. The novel algorithms that enable these synergies set the stage to uncover fundamental biological mechanisms and novel therapeutics.



SciSwipe develops and deploys software that seamlessly connects human pattern recognition abilities with computational speed and accuracy to mine complex, large-scale data.

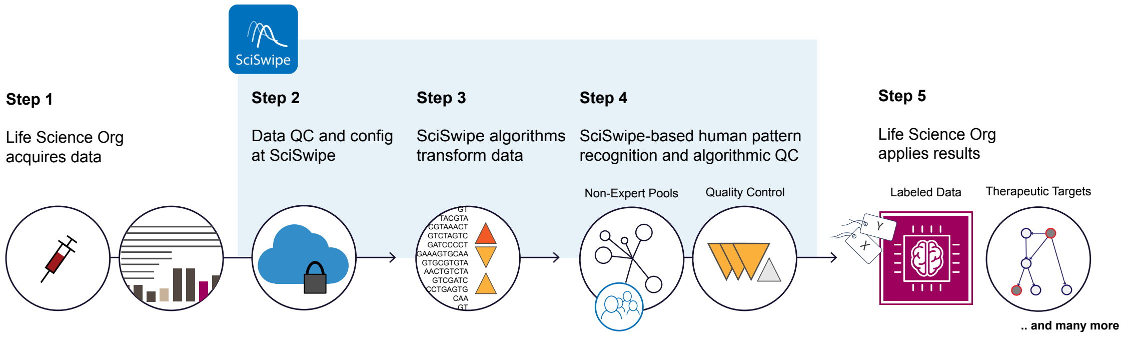


Figure 1: Synergies at scale: a secure workflow to process complex, large-scale data through human-algorithm synergistic software pipelines. Biological samples are obtained and data is generated, or acquired data is re-purposed (Step 1). Following QC (Step 2), data is transformed and analyzed (Step 3), after which the results are processed to meet the input requirements for downstream applications, and placed in a biologically or clinically meaningful context (Step 4). Downstream applications make use of the results of the workflow (Step 5), e.g. the generation of labeled data for supervised machine learning, or the discovery of novel molecular targets for therapy. Abbreviations: QC, Quality Control.

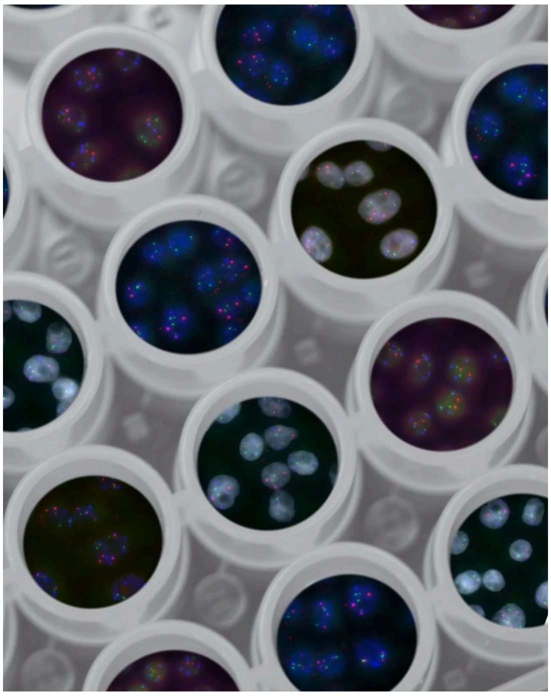
Complex, large-scale data holds unexplored potential

In June 2000, the Human Genome Project[1] made available to the public an initial draft sequence of the human genome. The human genome counts an approximate 3.2 billion base pair (bp) of DNA. The achievement formed the prelude to the era of whole-genome and large-scale sequencing, in which genomic sequences of organisms were published one after another and high-throughput measurements made their way to laboratories, shedding light on human genome biology and enabling genome perturbation and measurement at scale. Along these lines, the acquisition of NGS data has exponentially increased in both academia and industry. Data can be acquired in-house or from secondary sources, such as public repositories, licensed repositories, or from partners in primary healthcare. In May 2021, the largest public repository for sequencing data, the Sequence Read Archive[2], counted more than 43 petabytes of data and a continuation of its exponential increase. Similar trends in the acquisition of data are observed throughout systems biology, for instance in the study of the proteome and metabolome. Novel methods to assess, at scale, the different steps of interpretation and modification of biological molecules that occur in the cell or organism are developed at a regular pace. These trends signal that the scale of data acquisition will continue to rapidly expand.

The significant growth of available data resulted in an acceleration of the development of analytics technologies and techniques. Open-source software such as Bioconductor[3] has had a major impact on the field and associated repositories are continuously maintained and improved to account for increasing data collection, computational power and more recently, numerous efforts to integrate multiple types of data. These exciting advances have opened

[1] https://www.genome.gov/human-genome-project
[2] https://www.ncbi.nlm.nih.gov/sra
[3] https://www.bioconductor.org

new routes to improve patient care through prevention, prediction and early diagnosis of disease. **At the same time, the complexity of the data, privacy and consent regulations and the societal and technical complications associated with complete automation have hindered comprehensive breakthroughs in the study of complex, large-scale patient data.** In the section “Artificial intelligence in medicine relies on recycled labeled data”, we discuss hurdles in the analysis of medical image data. In the section “Low signal-to-noise ratios hinder omics data analysis”, we use challenges in the analysis of NGS data to exemplify those in the broader omics field.



Patient data complexity hinders the in-depth extraction of information, even with state-of-the-art computational algorithms.

Artificial intelligence in medicine relies on recycled labeled data

Although deep learning has been around since the early 1990s, the necessary increase of computational power held back its wide acceptance until recent years. Rapid technological advances and state-of-the-art algorithms, both public and private, have shaped the field of machine learning and deep learning in particular. Three features have contributed to its success across industries:

- the ability to quickly prototype and prove initial value by training smaller models on a personal computer
- models can be readily scaled using commercially available cloud-based solutions such as Amazon Web Services[4] and Google Cloud Platform[5]
- trained models can be reproduced and results can be shared by making use of virtualization tools such as Docker[6] to containerize the software

With the increase of computational power and the availability of large labeled datasets for certain domains such as computer vision and Natural Language Processing (NLP), deep learning has become the method of choice to perform complicated tasks involving pattern recognition. Examples range from object recognition in images to text translation. Such tasks are performed through supervised machine learning methods, where a computer learns a model by using input data (features) and the associated output data (labels). For example, in an object recognition model the input data would be the image and the associated label would be the object present in the image, such as a tumor in the consideration of an MRI scan. The model consists of trainable parameters (weights) that are optimized by making use of

[4] https://aws.amazon.com
[5] https://cloud.google.com
[6] https://www.docker.com/

the labeled data in the training phase. If training is successful, the model has learned to recognize patterns in the data which it will use to make predictions of the label based on just the features in what is called the inference phase. If the model is able to generalize, it is able to do this well for data it has not seen before, resulting in real-world applicability. Depending on the data type, quality and extent to which the model can be generalized, supervised machine learning has been applied to a diverse set of use cases with varying amounts of success.

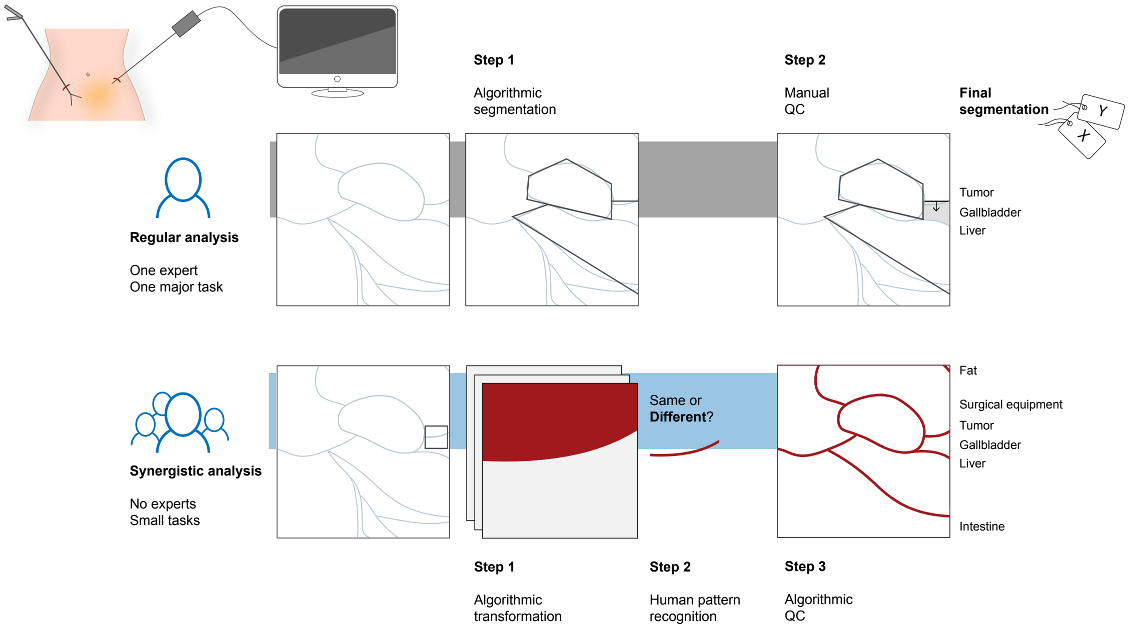


Figure 2: A simplified view of how an expert can analyze a complex image (top), and how, with the proper algorithmic transformations, a crowd can analyze the same data through multiple simpler images (bottom). Cartoons depict image data derived through laparoscopic surgery in which tissue in the patient abdomen is manipulated via small incisions. Computational algorithms transform the data (bottom, Step 1). Users of the SciSwipe platform can then inspect the visualizations to recognize patterns, and inform the system (Step 2). SciSwipe algorithms perform QC and update the metadata and labels (Step 3). Taken together, the right synergistic setup between non-experts and algorithms allows for increased speed, scale and a deeper analysis than would be possible with a human expert. Here, image segmentation and subsequent training of machine learning models enable robotic surgery.

One of the most important drivers for successfully training a deep learning model is the amount of labeled data available. Although anyone with access to the internet regularly labels simple images for Google when reCAPTCHA-prompted, not all data can be labeled using this approach. For example, medical data such as surgically obtained human tissue images are notoriously difficult to label, oftentimes even for an expert in the field. Furthermore, data might contain information that can allow to identify the individual from which the data was derived. Therefore, not all of us can or should access medical data. **As a result, the research community relies on a relatively small amount of labeled datasets that are**

recycled, thereby vastly limiting the performance of deep learning-based models. The use of recycled data limits the user in terms of the type of data, the amount of data, and the labels and label quality. Possible resolutions to this problem which are being pursued by academia and industry, are investing in the expertise of the human data labeler and making the labeling process simpler by algorithmically pre-processing the data. However, making the labeling process scalable and generic for many different datasets has proven to be difficult.

SciSwipe makes the accurate labeling of complex hard-to-label data in the life sciences sector scalable, while being generic enough to apply to many different datasets. Enabling the labeling of these datasets (Fig. 2) will result in dramatic improvements of performance of machine learning-based models applied to the medical domain.

Genetic variation in the human genome may or may not underlie disease

Differences in the DNA from one human to another in the form of point mutations (single nucleotide polymorphisms, SNPs), insertions, deletions and larger structural variation can underlie disease and disease predisposition. In a textbook example of how genetic variation may lead to dramatic phenotypic effects, each of a current 382 of 466 annotated variants in the cystic fibrosis transmembrane conductance regulator (CFTR) gene encoding an epithelial chloride ion channel gives rise to cystic fibrosis[7]. At the same time, over 2.000 variants of the CFTR gene have been identified.

The genetic difference between individuals is thought to be an approximate 6 million bp, which roughly corresponds to 1 in 1.000 bp. At the same time, not every change appears to have an effect. We can readily sequence the genomes of individuals that are affected by a disease, but it is painstakingly difficult to pinpoint the underlying genetic mutations that contribute to or directly underlie the disease. Best practices and even standardization have arisen for the initial steps in the processing of the raw data. Subsequent computational steps to further analyze the data, however, can yield divergent results. In part, the complexity is in the combinatorial effect of differences, where combinations of mutations together contribute to or lead to a disease phenotype. In addition, each individual will have mutations that may or may not contribute to a disease, and some of these mutations that in fact do not contribute will be shared among individuals, for instance through ancestry or geography. To be able to see less obvious, but therapeutically more interesting anomalies, confounding shared characteristics must be excluded from these samples. Some studies attempt to eliminate confounding effects by sampling from genetically similar populations. However, this is not always possible nor desirable, for instance when considering rare diseases. Others attempt to exclude these effects using dimensionality reduction methods such as principal component analysis.

The multi-factorial problem that arises from the combination of mutations, confounding features, but also noise is, at best, computationally expensive to solve. The current solution to focus on the most prominent hits has led to biologically meaningful results. At the same time, there is an increasing awareness that the complexity must be addressed in order to gain a deeper understanding of human disease and boost target discovery for therapeutic strategies.

Low signal-to-noise ratios hinder omics data analysis

After a model has successfully been trained using labeled data, it will perform well for the specific task it was trained for. However, it may not be able to generalize to other tasks as its learned weights were not optimized for those tasks. Another problem is that the grouping of data may leave room for interpretation. Consider grouping samples based on phenotypic characteristics to enable the consideration of the effect of mutations (see inset "Genetic variation in the human genome may or may not underlie disease"). Assigning the respective labels to non-binary, non-numerical conditions such as major depressive disorder (MDD) or inflammatory bowel disease (IBD) may in fact not be descriptive enough to separate populations and tease out contributing mutations. Along these lines, binary classifications – "cancer" or "healthy" – may too rigidly stratify groups, missing out on for instance mutations present in healthy patients that do not yet, but might eventually, give rise to cancer. There can be many genetic changes that have no observable phenotypic effect until the patient reaches a certain stage or threshold. Combinatorial studies that include single cell lineage tracing or similar frontier technologies shed light on such consecutive occurrences, but are not applicable at scale.

In recent years, methods have been reported for the analysis and exploration of complex, large-scale data that do not rely on labeled data and rather make use of unsupervised machine learning, meaning only features and no labels are available in the dataset. It has, however, been proven difficult to adapt existing supervised machine learning approaches to perform well in the unsupervised setting. Purely unsupervised machine learning approaches such as clustering and dimensionality reduction for data visualization are generally only able to provide a limited understanding of the data. This limitation holds especially true when the data has a low signal-to-noise ratio. Noise is inherent to biology, but is also a result of technical artefacts at the sample acquisition stage, throughout sample preparation prior to data acquisition, during data acquisition (e.g. NGS), and in initial data processing. Automation is hindered by the necessity to tweak parameters to account not only for low signal-to-noise, but for variable ratios that relate to the different types of data assessed.

The significant advances in research and diagnostic hardware must be met with innovative computational analytics solutions to enable precise and accurate care.

A concrete example of how noise may drown out the true signal present in biological data can be found in studies aimed at determining the pathological contribution of individual

[7] The Clinical and Functional TRanslation of CFTR (CFTR2); available at http://cftr2.org; 24 September 2021

molecular variations in human DNA (see inset “Genetic variation in the human genome may or may not underlie disease”). These variations are at the basis of diagnosis, prevention of disease onset and treatment of individuals that harbor them. Contributions can be numerically assessed by computational methods, of which genome-wide association studies (GWAS) form an unbiased approach, although features with which the variants are associated are defined in a potentially biased way. In GWAS, sets of single point mutations observed in the studied DNA, called single nucleotide polymorphisms (SNPs), are captured from large sample sizes of human participants. GWAS attempt to link the occurrence of single or combinations of SNPs to a specific disease using computational methods. A crucial disadvantage of GWAS is that rare mutations remain undetected, as mutations with a high frequency are captured more readily. Methods that provide increased statistical strength, such as rare variant association studies (RVAS), might capture such mutations. However, they either build on the premise of strong effect sizes, which do not always appear to be present, or must rely on predefined assumptions, thereby running the risk of missing signals that do not satisfy the made assumptions. The main challenge is to characterize novel mutations with little to no prior information.

Taken together, there is a limited availability of labeled data and an inherently low and variable signal-to-noise ratio across different types of biological data. **The constraints on re-purposing trained models to new tasks and on using unsupervised methods to gain insight in low signal-to-noise ratio data have vastly limited the applicability of existing machine learning approaches.** This especially holds true in the healthcare domain. **SciSwipe detects signals hidden in omics data by synergistically leveraging the human innate ability to visually detect patterns and the speed and scalability of computational algorithms.** Pattern detection is highlighted in the inset “Humans excel at pattern recognition”. Increasing our ability to unmask true signals in noisy data (**Fig. 3**) will dramatically advance the study of large-scale biological datasets available in the life sciences sector, improving our understanding of fundamental biological mechanisms in disease and enabling potential cures.

Humans excel at pattern recognition

The human brain is accustomed to pattern recognition as a result of evolutionary pressure. We are wired to assess even complex situations at high speed given the correct presentation. Everyday applications such as commercials, movie scenes, and simplification to pictograms in, for instance, public spaces make use of our ability to visually correctly assess features based on prior knowledge, which may include context. In data science, intuitive data visualization includes presentation formats such as heatmaps and clustering visualization methods like t-distributed stochastic neighbor embedding (t-SNE). Interactive data visualization and analysis platforms in essence apply creative filters that leave known or anticipated values out and highlight important features of the data, that guide but do not overly interpret for the viewer. It is important to realize that such visualization formats and analyses use common assumptions across human viewers and anticipated values in data sources. If used incorrectly, this may lead to displaying patterns that are in no sense beneficial to the understanding of the underlying data. In the section “Human and algorithm synergies require a supporting framework”, we highlight how any result must thus be embedded into an extensive pipeline. Ideally, this pipeline provides biological context, setting the stage for targeted measurements that place results into a statistical framework, and enable experimental validation at a feasible scale.

Human and algorithm synergies require a supporting framework

In a regular bioinformatics workflow, data is pre-processed, explored through iterative rounds of knowledge- or problem-guided computation, and statistical calculations are performed to support the manual assessment. An underexplored alternative enabler of insight generation is a combination of human pattern recognition at scale seamlessly integrated with algorithmic pre- and post-processing of the data – human-algorithmic synergistic software pipelines. In a step-by-step approach, clear alternating parts are played by the human and the machine. To enable these synergies, at some point in the workflow, data must be presented in a dramatically simplified form to the human participant. This necessitates a certain creativity and realization of both how humans recognize patterns and how algorithms can interact with or follow up on human actions. The general idea is that where a human will be quick and able to recognize an item by association, place an item in a context, a computer can support this process by providing and processing numerical information as soon as such an appointment is made, for example a numerical answer to the question of how closely two items match. This is critical when dealing with large-scale data, a large amount of results and the necessity to provide a ranking in order to be useful as an output.

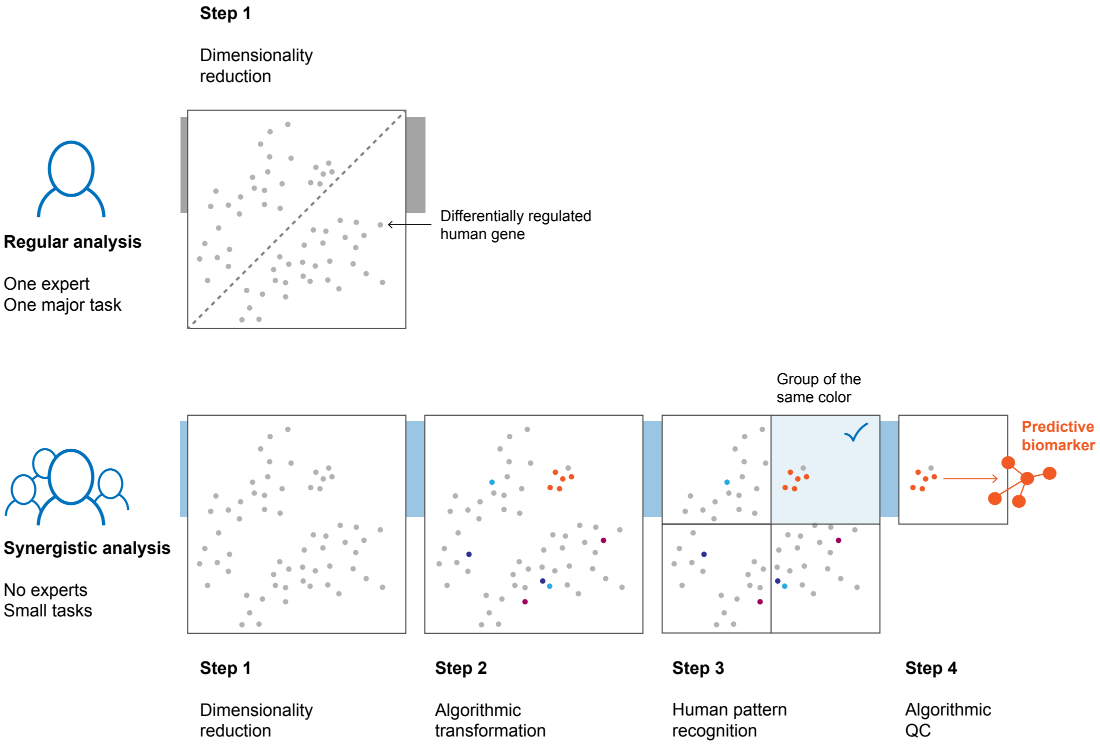


Figure 3: A simplified view of how an expert can generate a complex graph (top), and how, with the proper algorithmic transformations, a crowd can analyze the same data through multiple simpler images (bottom). Here, fold changes in human gene expression from healthy and diseased individuals are compared, with the aim to derive a combination of genes that can predict disease in undiagnosed individuals. Taken together, the right synergistic setup between non-experts and algorithms allows for increased speed, scale and a deeper analysis than would be possible with a human expert.

The best-of-both-worlds scenario in which human and computational capabilities are seamlessly integrated opens the possibility to explore more than what is currently possible (**Fig. 3**). We identify five prerequisites for human-algorithmic synergistic software pipelines at scale.

- The actions for which human participants are required must be boiled down to the lowest effort possible. This involves the transformation of complex data into simple visualizations amenable to human pattern recognition, but also the supporting infrastructure. The infrastructure should seamlessly integrate human and algorithmic interaction.
- Human participants must be stratified into carefully curated pools in order to allow for different levels of expertise to be put to use as efficiently as possible.
- The management of human participants and participant pools must be impeccable, assuring participants can perform the task for which their input is needed.
- Systems must be in place to assess, monitor and report the accuracy of human actions.
- There must be a thorough understanding and numerical reasoning as to which human actions are required and at which scale in order to meet output criteria (e.g. statistical significance).

For the latter, this means the data provider and the data scientist need to have a plan available for how the data is generated, which problems are to be expected and how labels or actions will lead to model training and validation and/or insight generation and validation. Preferably, it should be known beforehand with which accuracy the analysis should take place in order to be able to generate insights into the data with the desired statistical significance. Subsequently, the resulting numbers can be used to determine the necessary scale of the operation.

Computational algorithms transform complex data to images optimized for human pattern recognition, setting the stage to improve diagnostics and develop novel therapeutics.

SciSwipe is a software platform for synergistic analysis

SciSwipe develops and deploys software that seamlessly combines human pattern recognition abilities and computational speed and accuracy. We enable data analysis for customers in the life sciences that produce or own data that

- is large in number of data entries – depending on complexity, this means 1-100k or more entries
- is complex, typically requiring expert review and annotation
- has an inherent low signal-to-noise ratio

Our proprietary software encompasses algorithmic data pre- and post-processing, data transformation, software-guided human pattern recognition and algorithmic quality control of human actions (**Fig. 1**). The cloud-based platform includes software-based tools for the validation and interpretation of results. The custom-developed software architecture allows for a lean, flexible and scalable setup. We build to scale to both novel data transformation algorithms in order to accommodate distinct types of research data and to thousands of users. The setup also allows for the handling of customer data independent of storage technology and location – SQL or NoSQL data sources that are either stored in the cloud or on-premises – thereby anticipating to relieve a major barrier of entry. In-house, accredited expertise enables us to design and implement industry-standard robust and fully disclosed security and privacy measures.

Data is processed and managed in a secure workflow

Complex, large-scale datasets in the life sciences are privacy-sensitive when generated based on patient samples. In addition, the datasets may be sensitive from a corporate perspective, as they can provide information on the strategy of the company or on proprietary methods used to acquire the data. As such, it is important to not only ensure the confidentiality of data for the sampled individual, but also the proprietary nature of such data for corporations and other organizations. Particularly the regulations encompassing privacy-sensitive data are in constant evolution and differ between countries. As an example, genetic data can be traced back to its sampled source even when anonymized and should thus involve measures that prevent unauthorized access. In the near future, additional types of data in the life sciences that are generated based on human subjects may be considered to be privacy sensitive – omics fingerprints for example – and will likely necessitate increased consideration prior to direct transfer, cloud-based sharing or submission to public repositories. The transformation of these datasets and the review of the transformed data by non-expert humans require access to different modalities of the underlying data and should take these constraints into consideration. Taken together, in any large-scale setup to transfer, transform and analyze privacy-sensitive or confidential data, protocols need to be enforced to ensure safe data handling and destruction in an end-to-end workflow. We identify the following measures that should be undertaken in data transfer, data transformation and data analysis (**Fig. 1**) to realize a secure workflow:

Data transfer Data is anonymized at the source. Consultation is done on-premises or in the cloud, in which data is transferred to dedicated systems, over dedicated connections. Reports and data are matched on site and/or transferred over a secure connection.

Data transformation Initial transformation is done on site or in dedicated systems. Note that the algorithmic transformation of the data before analysis effectively anonymizes the data.

Data analysis Transformed data is snippeted to such an extent, that no privacy-sensitive or corporate confidential data can be derived. Snippets never overlap to the extent where connections can be made. Participants never receive or are able to connect a full set of transformed snippets.

The measures that we take ensure meeting and fostering corporate, proprietary and privacy-related sensitivity requirements.

SciSwipe enables labeling and in-depth data analysis at scale

Although major manual efforts as well as computational advances in image segmentation have led to large collections of segmented images across disciplines, the complexity and nature of medical images hinders similar developments in the healthcare domain. As a result, the research community relies on a relatively small amount of labeled datasets that are recycled, thereby vastly limiting the performance of deep learning-based models. Along these lines, yet in distinct analytical efforts, the low signal-to-noise ratio in omics data has vastly limited the applicability of computational approaches and the depth at which the data is analyzed, particularly in situations where no prior information is available.

SciSwipe enables data labeling and data exploration at scale for the major types of complex, large scale data in the life sciences. Our synergistic human-algorithmic software pipelines scalably and generically transform complex large-scale data to visualizations that are optimized for humans to detect patterns in, thereby finding hidden signals in the underlying data. This unique approach to abstract complexity allows leveraging synergies between algorithms and non-expert participants. Ultimately, we enable accurate, scalable analysis at a significantly higher depth in comparison to current strategies. We provide a dedicated, cloud-based platform and have implemented robust and fully disclosed security and privacy measures that meet best practices in industry, as well as traceability measures that ensure that SciSwipe meets present day requirements for medical software. Our platform can handle raw data in image or numerical format and is currently available for the following applications:

- **Image Segmentation** of medical images at high speed and accuracy. We enable rapid image segmentation at scale, in which the customizable accuracy can be tailored to meet specific machine learning algorithm input demands
- **Target Discovery** based on Next Generation Sequencing (NGS) data. We enable target discovery based on novel algorithms for the exploration of NGS data. We deliver a feature set with features that can be employed directly as combinatorial genetic marker or used for applications, for instance in lead generation
- **Haplotyping** based on NGS or Third Generation Sequencing (TGS) data. We automate the process to produce results at high confidence and at scale. Our algorithms are equipped to handle targeted and whole-genome sequencing data from different sequencing platforms

We unlock the value of complex, large-scale data: SciSwipe sets the stage for you to leverage the power of state-of-the-art machine learning and to gain insight in fundamental biological mechanisms in disease. Ultimately, we enable you to develop the next-generation treatments for patients with unmet needs.

SciSwipe
Palo Alto, CA, USA

Looks good?
www.sciswipe.com/contact

www.sciswipe.com

Photo’s by the National Cancer Institute,
Laura Ockel, and TheStandingDesk.com.