Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology

Moritz Lürig[1]; Seth Donoughe[2]; Erik I. Svensson[1]; Arthur Porto[3,4]; Masahito Tsuboi[1]
[1] Department of Biology, Lund University, Lund, Sweden
[2] Department of Molecular Genetics and Cell Biology, University of Chicago, Chicago, US
[3] Department of Biological Sciences, Louisiana State University, Baton Rouge, US
[4] Center for Computation & Technology, Louisiana State University, Baton Rouge, US

Abstract

For centuries, ecologists and evolutionary biologists have used images such as drawings, paintings and photographs to record and quantify the shapes and patterns of life. With the advent of digital imaging, biologists continue to collect image data at an ever-increasing rate. This immense body of data provides insight into a wide range of biological phenomena, including phenotypic trait diversity, population dynamics, mechanisms of divergence and adaptation, and evolutionary change. However, the rate of image acquisition frequently outpaces our capacity to manually extract meaningful information from images. Moreover, manual image analysis is low-throughput, difficult to reproduce, and typically measures only a few traits at a time. This has proven to be an impediment to the growing field of phenomics - the study of many phenotypic dimensions together. Computer vision (CV), the automated extraction and processing of information from digital images, provides the opportunity to alleviate this longstanding analytical bottleneck. In this review, we illustrate the capabilities of CV as an efficient and comprehensive method to collect phenomic data in ecological and evolutionary research. First, we briefly review phenomics, arguing that ecologists and evolutionary biologists can effectively capture phenomic-level data by taking pictures and analyzing them using CV. Next we describe the primary types of image-based data, review CV approaches for extracting them (including techniques that entail machine learning and others that do not), and identify the most common hurdles and pitfalls. Finally, we highlight

recent successful implementations and promising future applications of CV in the study of phenotypes. In anticipation that CV will become a basic component of the biologist's toolkit, our review is intended as an entry point for ecologists and evolutionary biologists that are interested in extracting phenotypic information from digital images.

From phenotypes to phenomics

Faced with the overwhelming complexity of the living world, most life scientists confine their efforts to a small set of observable traits. Although a drastic simplification of organismal complexity, the focus on single phenotypic attributes often provides a tractable, operational approach to understand biological phenomena, e.g. phenotypic trait diversity, population dynamics, mechanisms of divergence and adaptation and evolutionary change. However, there are also obvious limitations in how much we can learn from studying small numbers of phenotypes in isolation. Evolutionary and conservation biologist Michael Soulé was one of the first to demonstrate the value of collecting and analyzing many phenotypes at once in his early study of the side-blotched lizard (*Uta stansburiana*; (Soulé, 1967)); reviewed in Houle et al. (2010)). While doing so, he defined the term "phenome" as "the phenotype as a whole" (Soulé, 1967). Phenomics, by extension, is the comprehensive study of phenomes. In practice, this entails collecting and analyzing multidimensional phenotypes with a wide range of quantitative and high-throughput methods (Bilder et al., 2009). Given that biologists are now attempting to understand increasingly complex and high dimensional relationships between traits (Walsh, 2007), it is surprising that phenomics still remains underutilized (Fig. 1), both as methodological approach and as an overarching conceptual and analytical framework (Houle et al., 2010).

Phenomic datasets are essential if we are to understand some of the most compelling but challenging questions in the study of ecology and evolution. For instance, phenotypic diversity can fundamentally affect population dynamics (Laughlin et al., 2020),

community assembly (Chesson, 2000) and the functioning and stability of ecosystems (Hooper et al., 2005). Such functional diversity (Petchey and Gaston, 2006) is ecologically extremely relevant, but can be hard to quantify exactly, because organisms interact with their environment through many traits of which a large portion would need to be measured (Villéger et al., 2008; Blonder, 2018). Moreover, natural selection typically does not operate on single traits, but on multiple traits simultaneously (Lande and Arnold, 1983; Phillips and Arnold, 1999), which can lead to correlations (Schluter, 1996; Sinervo and Svensson, 2002; Svensson et al., 2021) and pleiotropic relationships between genes (Visscher and Yang, 2016; Saltz et al., 2017). Phenotypic plasticity, which is increasingly recognized in mediating evolutionary trajectories (Pfennig et al., 2010), is also an inherently multivariate phenomenon involving many traits and interactions between traits, so it should be quantified as such (Morel-Journel et al., 2020). Put simply: if we are to draw a complete picture of biological processes and aim to understand their causal relationships at various levels of biological organization, we need to measure more traits, from more individuals and a wider range of different species.

High dimensional phenotypic data are also needed for uncovering the causal links between genotypes, environmental factors, and phenotypes, i.e. to understand the genotype-phenotype map (Houle et al., 2010; Orgogozo et al., 2015). The advent of genomics - high throughput molecular methods to analyze the structure, function or evolution of an organism's genome in parts or as a whole (Church and Gilbert, 1984; Feder and Mitchell-Olds, 2003) - has already improved our understanding of many biological phenomena. This includes the emergence and maintenance of biological diversity (Seehausen et al., 2014), the inheritance and evolution of complex traits (Pitchers et al., 2019), and the evolutionary origin of key metabolic traits (Ishikawa et al., 2019). Thus, accessible molecular tools have lowered the hurdles for discovery-based genomic research and shifted the focus away from the study of observable organismal traits and phenotypes

towards their molecular basis. However, a similar “moonshot-programme” for the phenotype, i.e. an ensemble of phenomics methods that matches genomics in their comprehensiveness, is still lacking (Freimer and Sabatti, 2003). The growing mismatch in how efficiently molecular and phenotypic data are collected may hamper further scientific progress in ecological and evolutionary research (Houle et al., 2010; Orgogozo et al., 2015; Lamichhaney et al., 2019).

Following previous calls for phenomic research programmes (Bilder et al., 2009; Houle et al., 2010; Furbank and Tester, 2011), some recent studies have collected phenotypic data with high dimensionality and on a massive scale, for example, in plants (Ubbens and Stavness, 2017), animals (Cheng et al., 2011; Kühl and Burghardt, 2013; Pitchers et al., 2019) and microbes (Zackrisson et al., 2016; French et al., 2018). All of these studies use some form of image analysis to quantify external (i.e. morphology or texture) and internal phenotypes (e.g. cells, bones or tissue), or behavioral phenotypes and biomechanical properties (e.g. body position, pose or movement). Such data represents phenomics in a narrow sense: the collection of (external, internal, behavioral) phenotypic data on an organism-wide scale (Houle et al., 2010). In addition, many biologists also use image analysis to detect presence and absence of organisms (e.g. within a population, community or environment; e.g. by means of camera traps or satellite images), or to identify species (by experts or algorithms). While species monitoring and taxonomic identification constitutes an important and rapidly growing discipline on its own (Norouzzadeh et al., 2018; Wäldchen and Mäder, 2018; Høye et al., 2020), this review focuses on the extraction of phenotypic data from digital images as a key methodological approach for the study of phenomes (Houle et al., 2010).

Previous work has supplied us with an immense body of image data that has provided insight into a wide range of biological phenomena, yet when biologists manually extract phenotypes from images for phenomic-scale research, they confront several main

bottlenecks (Houle et al., 2003; Gerum et al., 2017; Ubbens and Stavness, 2017). A major constraint when working with large amounts of images (~1000 or more) is processing time and cost. Manual extraction of phenotypic data from images is slow and it requires trained domain experts whose work is extremely expensive. Moreover, the collection of such metrics in a manual fashion entails subjective decisions by the researcher, which may make it prone to error, and certainly makes reproducibility difficult. Last, manually measured traits tend to be low-dimensional measurements of higher dimensional traits. For example, external colour traits, such as human eye colour phenotypes, are often scored as discrete categories (e.g. brown vs blue phenotypes), whereas *pixel* level information (number of brown vs. blue pixels) can provide a continuous phenotypic metric (Liu et al., 2010). Such quantitative, high-dimensional data can provide insight into previously hidden axes of variation, and may help provide a mechanistic understanding of the interplay of phenotypes, their genetic underpinnings, and the environment.

In this review we extol *computer vision* (CV - for a definition of terms in *italics* see Box 1), the automatic extraction of meaningful information from images, as a promising toolbox to collect phenotypic information on a massive scale. The field has blossomed in recent years, producing a diverse array of computational tools to increase analytic efficiency, data dimensionality, and reproducibility. This technological advancement should be harnessed to produce more phenomic datasets, which will make our conclusions and inferences about biological phenomena more robust. We argue that CV is poised to become a basic component of the data analysis toolkit in ecology and evolution, enabling researchers to collect and explore phenomic-scale data with fewer systematic biases (e.g. from manual collection). Our review is intended to provide an entry point for ecologists and evolutionary biologists to the automatic and semi-automatic extraction of phenotypic data from digital images. We start with a general introduction to CV and its history, followed by some practical considerations for the choice of techniques based on the given data, and

finish with a list of some examples and promising open-source CV tools that are suitable for the study of phenotypes.

The structure of digital images

A two dimensional image is an intuitive way to record, store, and analyze organismal phenotypes. In the pre-photography era, ecologists and evolutionary biologists used drawings to capture the shapes and patterns of life, later to be replaced by analog photography, which allowed for qualitative assessment and simple, often only qualitative analysis of phenotypic variation. With the advent of digital photography, biologists could collect phenotypic data at unprecedented rates using camera stands, camera traps, microscopes, scanners, video cameras, or any other instrument with semiconductor image sensors (Goesele, 2004; Williams, 2017). Image sensors produce two-dimensional raster images (also known as bitmap images), which store incoming visible light or other electromagnetic signals into discrete, locatable picture elements - in short: *pixels* (Fig. 2)(Fossum and Hondongwa, 2014). Each pixel contains quantitative phenotypic information that is organized as an array of rows and columns, whose dimensions are also referred to as “pixel resolution” or just “resolution”. An image with 1000 rows and 1500 columns has a resolution of 1000 x 1500 (= 1 500 000 pixels, or 1.5 megapixels). The same applies for digital videos, which are simply a series of digital images displayed in succession, where the frame rate (measured as frames per second = fps) describes the speed of that succession.

On the pixel level, images or video frames can store variable amounts of information, depending on the *bit depth*, which refers to the number of distinct values that a pixel can represent (Fig. 2). In binary images, pixels contain information as a single bit, which can take exactly two values - typically black or white (2^1 values = 2 intensity values). Grayscale images from typical consumer cameras have a bit depth of 8, thus each pixel can take a value between 0-255 (2^8 values = 256 intensity values), which typically represents a level of

light intensity, also referred to as pixel intensity. Colour images are typically composed of at least three sets of pixel arrays, also referred to as channels, each of which contain values for either red, green or blue (RGB; Fig. 2). Each channel, when extracted from an RGB image, is a grayscale representation of the intensities for a single colour channel. Through the combination of pixel values at each location into triplets, colours are numerically represented. Today the industrial standard for colour images is 24-bit depth, in which each colour channel has a bit depth of 8 and thereby can represent 256 colours (Fig. 1). Thus, 24-bit RGB images can represent over 16 million colour variations in each pixel (2^{24} = 256 x 256 x 256 = 16 777 216 intensity values), which already greatly surpasses the estimated 2.28 millions of colour variations that humans can perceive (Pointer and Attridge, 1998).

Today, high resolution image sensors are an affordable way to store externally visible phenotypic information, like colour and shape. However, advanced image sensors can also combine information from different spectra other than the visible light, like infrared radiation, which can be used to quantify individual body temperatures. With thermal image sensors, biologists can estimate body surface temperatures, which are correlated with internal (core) body temperatures (Tattersall and Cadena, 2010), particularly in small animals like insects (Tsubaki et al., 2010; Svensson et al., 2020). Thermal imaging, or thermography, offers new opportunities for ecophysiological evolutionary research of how animals cope with heat or cold stress in their natural environments (Fig. 5B) (Tattersall et al., 2009; Tattersall and Cadena, 2010; Svensson and Waller, 2013). Fluorescence spectroscopy is another way to quantify phenotypes in high throughput and with high detail. For example, plate readers typically used in microbial and plankton research, can combine light in the visible spectrum with images containing information of cell fluorescence or absorbance to an “image stack” (Roeder et al., 2012). Image stacks and the inclusion of multiple spectral channels provide a promising avenue of research towards capturing a more complete representation of the phenotype (Fig 4A) (Hense et al., 2008; Di et al., 2014).

A brief introduction to computer vision

CV is an interdisciplinary field at the intersection of *signal processing* and *machine learning* (Fig. 4) (Mitchell, 1997), which is concerned with the automatic and semiautomatic extraction of information from digital images (Shapiro and Stockman, 2001). CV-based extraction of phenotypic data from images can include a multitude of different processing steps that do not follow a general convention, but can be broadly categorized into preprocessing*, segmentation, and* measurement (Fig. 3). These steps do not depict a linear workflow, but are often performed iteratively (e.g. preprocessing often needs to be adjusted according to segmentation outcomes) or in an integrated fashion (e.g. relevant data can already be extracted during preprocessing or segmentation).

Preprocessing: Preparing an Image for Further Processing

Independent of how much care has been taken during image acquisition, preprocessing is an important step to prepare images for the CV routines to follow. There is a wealth of image processing techniques that can be applied at this stage, such as transformations to reduce or increase noise (e.g. gaussian blur) or enhance contrast (e.g. histogram adjustment). Images can also be masked or *labeled* as a way to filter the image so that subsequent steps are applied to the intended portions of each image. Defining the appropriate coordinate space (i.e. pixel-to-mm ratios) is also part of preprocessing. Finally, certain machine learning techniques such as *deep learning* require an enormous amount of data, which may require data augmentation: the addition of slightly modified copies of existing data or the addition of newly created synthetic data (Shorten and Khoshgoftaar, 2019). Overall, preprocessing tasks are highly specific to the respective study system, image dataset or computer vision technique, and may initially require some fine-tuning by the scientist to ensure data quality, which, however, can typically be automated afterwards.

Segmentation: separation of "foreground" from "background"

The central step in any phenotyping or phenomics related CV pipelines is the segmentation of images into pixels that contain the desired trait or *feature* (*foreground*) and all other pixels (*background*). In its most basic form, segmentation of grayscale images can be done by simple signal processing algorithms, such as a *threshold (Zhang and Wu, 2011)* or *watershed (Beucher, 1979)*. Similarly, *feature detection* algorithms examine pixels and their adjacent region for specific characteristics or key points, e.g. whether groups of pixel form edges, corners, ridges or blobs (Rosten and Drummond, 2006). Videos or multiple images of the same scene provide an additional opportunity for segmentation: foreground detection can detect changes in image sequences to determine the pixels of interest (e.g. a specimen placed in an arena, or animals moving against a static background), while subsequent background subtraction isolates the foreground for further processing (Piccardi, 2004). Finally, *object detection* describes the high level task of finding objects (organisms, organs, structures, etc.) in an image, which is typically addressed through classical machine learning or *deep learning (*see section "A History of Computer Vision Methods") (LeCun et al., 2015; Heaton, 2020; O'Mahony et al., 2020). In classical machine learning, features have to be first engineered or extracted from a *training* dataset using *feature detectors*, then used to train a classifier, and finally applied to the actual dataset (Mitchell, 1997). Deep learning algorithms are a family of machine learning methods based on artificial *neural networks* that "learn" what constitutes the object of interest during the training phase (LeCun et al., 2015; Heaton, 2020). With sufficient training using labeled images (and in some cases unlabelled images - see Box 2), deep learning-powered object detection algorithms can be highly accurate and often greatly outperform pre-existing object recognition methods (Krizhevsky et al., 2012; Alom et al., 2018) - in some cases even human experts, for example, when identifying species (Buetti-Dinh et al., 2019; Valan et al., 2019; Schneider et al., 2020b). Each of these approaches has advantages and limitations, which mostly depend on the noise level within

the images, the size of the dataset, and the availability of computational resources (see section “Practical considerations for CV” and Fig. 5).

Measurement: extraction of phenotypic data

CV can retrieve a multitude of phenotypic traits from digital images in a systematic and repeatable fashion (see Table 1 and Table 2). In the simplest case, CV may measure traits that are established in a given study system, such as body size (e.g. length or diameter) or colour (e.g. brown phenotype vs blue phenotype). In such cases, switching from a manual approach to a semi- or fully automatic CV approach is straightforward, because the target traits are well embedded in existing statistical and conceptual frameworks. The main benefits from CV are that costly manual labor is reduced and that the obtained data becomes more reproducible, because the applied CV analysis pipeline can be stored and re-executed. However, just as manual measurements require skilled personnel to collect high quality data, great care needs to be taken when taking images so that their analysis can provide meaningful results (also see section “Image Quality: Collect Images That Are Maximally Useful”*).* It is also possible to increase the number of dimensions without much extra effort and without discarding the traditionally measured traits (Table 1). For example, in addition to body size, one could extract body shape traits, i.e. the outline of the body itself (i.e. contour coordinates of the foreground), and texture (i.e. all pixel intensities within the foreground). Such high dimensional traits can be directly analyzed using multivariate statistics, or transformed into continuous low dimensional traits, such as continuous shape features (circularity or area), texture features (colour intensity or variation, pixel distribution), or moments of the raw data (Table 1).

A history of computer vision methods

The field of CV is now close to celebrating its 6th decade. It first emerged in the late 1950s and early 1960s, in the context of artificial intelligence research (Rosenblatt, 1958). At the

time, it was widely considered a stepping-stone in our search for understanding human intelligence (Minsky, 1961). Given its long history, a wide-variety of CV techniques have emerged since its inception, but they all contain variations of the same basic mechanism. CV is, from the methodological standpoint, the process of extracting meaningful features from image data and then the use of such features to perform tasks, which, as described above, may include classification, segmentation, recognition, detection, among others. In this section, we will not aim at presenting an all-encompassing review of all CV methods, but rather to identify the major trends in the field and highlight the techniques that have proved useful in the context of biological research. It is worth noting that even classical CV approaches are still routinely used in the modern literature, either in isolation or, most commonly, in combination with others. In a large part, methodological choices in CV are highly domain-specific (see section *Practical considerations for computer vision,* Fig. 5, and Fig. 4).

First wave - Hand-crafted features

The first wave of CV algorithms is also the closest one to the essence of CV, namely, the process of extracting features from images. Starting with the work of Larry Roberts, which aimed at deriving 3D information from 2D images (Roberts, 1963), researchers in the 1970s and 1980s developed different ways to perform feature extraction from raw pixel data. Such features tended to be low-level features, such as lines, edges, texture or lighting, but provided us with the initial basic geometric understanding of the data contained in images. A notable example of such algorithms is the watershed algorithm. First developed in 1979 (Beucher, 1979), the watershed algorithm became popular in biological applications in the 1990s, being initially used to quantify elements and extract morphological measurements from microscopic images (e.g., (Bertin et al., 1992; Rodenacker et al., 2000)). This algorithm treats images as a topographic map, in which pixel intensity represents its height, and attempts to segment the image into multiple separate ‘drainage basins’. Certain

implementations of the watershed algorithm are still routinely used in signal processing (Fig. 4), and can be effectively used to process biological images such as those obtained through animal or plant cell microscopy (McQuin et al., 2018). Other initial low-level hand-crafted approaches that achieved popularity include the Canny and Sobel filters (edge detectors; (Canny, 1986; Kanopoulos et al., 1988)) and Hough transforms (ridge detection; (Duda and Hart, 1972)).

Another approach that gained popularity in the CV literature in the early 1990s was principal component analysis (PCA). In a PCA, independent, aggregate statistical features are extracted from multidimensional datasets. These can be used, for example, in classification. One of the most notable uses of PCA in the context of CV was the eigenfaces approach (Turk and Pentland, 1991). Essentially, Turk and Pentland (1991) noted that one could decompose a database of face images into eigenvectors (or characteristic images) through PCA. These eigenvectors could then be linearly combined to reconstruct any image in the original dataset. A new face could be decomposed into statistical features and further compared to other known images in a multidimensional space. Similar pioneering approaches emerged in the context of remote sensing research, in which spectral image data was decomposed into its eigenvectors ((Bateson and Curtiss, 1996; Wessman et al., 1997)). PCA has notably found many other uses in biology (e.g, (Ringnér, 2008).

In the late 1990s and early 2000s, Scale Invariant Feature Transform (SIFT) (Lowe, 1999, 2004) and Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) were developed. Both SIFT and HOG represent intermediate-level local features that can be used to identify keypoints that are shared across images. In both approaches, the first step is the extraction of these intermediate-level features from image data, followed by a feature matching step that tries to identify those features in multiple images. Finding keypoints across images is an essential step in many CV applications in biology, such as object detection, landmarking (Houle et al., 2003), and image registration (Mäkelä et al., 2002).

These intermediate-level features have several advantages over the lower-level features mentioned above, most notably the ability to be detected in a wide-variety of scales, noise and illumination. Another key aspect of SIFT and HOG features is that they are generally invariant to certain geometric transformations, such as uniform scaling and simple affine distortions.

Second wave - Initial machine-learning approaches

While the use of hand-crafted features spurred much of the initial work in CV, soon it became apparent that without image standardization, those low- and intermediate-level features will often fall short of producing sufficiently robust CV algorithms. For example, images belonging to the same class can often look very different and the identification of a common set of shared low-level features can prove to be quite challenging. Consider, for instance, the task of finding and classifying animals in images: two dog breeds can look quite different, despite belonging to the dog class (e.g. Chihuahua vs. Bernese mountain dog). As such, while the initial feature-engineering approaches were essential for the development of the field, it was only with the advent of machine-learning that CV acquired more generalizable applications.

Machine learning algorithms for CV can be divided in two main categories (but see Box 2): supervised and unsupervised (Geoffrey Hinton, Terrence J. Sejnowski, 1999). Unsupervised algorithms attempt to identify previously unidentified patterns on unlabeled data. In other words, no supervision is applied to the algorithm during learning. While it can be argued that PCA was one of the first successful unsupervised learning algorithms applied directly to CV, here we group PCA with “first wave” tools due to its use as a feature extractor. Other unsupervised learning algorithms commonly used in CV include clustering techniques, such as k-means (Lloyd, 1982) and gaussian mixture models (GMM) (Reynolds and Rose, 1995). Clustering algorithms represented some of the first machine learning approaches for CV. Their aim is to find an optimal set of objects (or components) that are

more similar to each other than to those in other sets. This type of approach allowed researchers to find hidden patterns embedded in multidimensional data, proving useful for classification and segmentation tasks. For example, GMM has been extensively used to classify habitat using satellite image data (Zhou and Wang, 2006), to segment MR brain images (Greenspan et al., 2006), and classification of animals from video (Edgington et al., 2006), to name a few.

However, it is in the supervised domain that machine learning for CV has been most successful (Heileman and Myler, 1989). In supervised learning approaches, the user supplies labeled training data in the form of input-output pairs (Box 2). The ML algorithm iteratively "learns" a function that maps input into output for the labeled training data. Among the initial supervised approaches for CV, Support Vector Machines (SVM) were by far the most common approach (Cortes and Vapnik, 1995). Given a certain image dataset and their corresponding labels (e.g., classes in a classification task), SVMs find the feature space that maximizes the separation between the classes of interest (referred to as *hyperplane*). An essential aspect of SVMs is that such learned decision boundaries separating the classes can be nonlinear in the original feature space, allowing the model to separate classes that would not be separable by a purely linear technique (Cortes and Vapnik, 1995). Support vector machines have been widely used in ecological research, e.g. for image classification (Sanchez-Hernandez et al., 2007) and image recognition (Hu and Davis, 2005), among others.

Third wave - Ensemble methods

While SVMs were extremely successful in CV and spurred much of the supervised work that happened afterwards, it became clear by the early 2000s that single estimators often underperformed approaches combining the predictions of several independent estimators, an approach known as ensemble methods (Krogh and Others, 1996; Dietterich, 2000). Ensemble methods represent a slightly different philosophical approach to machine learning,

in which multiple models are trained to solve the same task and their individual results are combined to obtain an even better model performance. Several ensemble methods have been developed in the literature, but they are generally divided in two main families: bagging and boosting.

Bagging approaches combine several models that were trained in parallel through an averaging process (Bauer and Kohavi, 1999). Each underlying model is trained independently of the others based on a bootstrap resample of the original dataset. As a consequence, each model is trained with slightly different and (almost) independent data, greatly reducing the variance in the combined model predictions. A classical example of bagging approach is the random forest algorithm (Breiman, 2001), in which multiple learning trees are fitted to bootstrap resamples of the data and posteriorly combined through mean averaging (or majority vote). In biology, bagging approaches have been used for environmental monitoring (Mortensen et al., 2007), sample identification (Lytle et al., 2010), among others.Boosting, on the other hand, combines learners sequentially rather than in parallel (Bauer and Kohavi, 1999). Among boosting algorithms, gradient boosting (Friedman, 2000) is one of the most widely used in CV. In gradient boosting, models are combined in a cascade fashion, such that a downstream model is fitted to the residuals of upstream models. As a consequence, while each individual model in the cascade is only weakly related to the overall task, the combined algorithm (i.e., the entire cascade) represents a strong learner that is directly related to the task of interest (Friedman, 2000). Since this approach, if unchecked, will lead the final model to overfit the training data, regularization procedures are usually applied when using gradient boosting.

Fourth wave – Deep learning

Deep learning approaches are, at the time of this writing, the state-of-the-art in CV and have recently become more accessible through the community-wide adoption of code-sharing practices. *Deep learning* refers to a family of machine learning methods based on

hierarchical artificial neural networks, most notably, *convolutional* neural networks (CNN). Networks with dozens or hundreds of *hidden layers* (i.e. deep neural networks) allow for the extraction of high-level features from raw image data (LeCun et al., 2015). While they have only recently become widespread, the history of artificial neural networks is at least as old as the field of CV itself. One of first successful attempts in the study of artificial neural networks was the perceptron (Rosenblatt, 1958), a computer whose hardware design was inspired by neurons, and which was used to classify a set of inputs into two categories. This early work, while successful, was largely restricted to linear functions and therefore could not deal with non-linearity, such as XOR functions (Minsky et al., 1969). As a consequence, artificial neural network research remained rather understudied until the early 80s when training procedures for multi-layer perceptrons were introduced (i.e., backpropagation) (Rumelhart and McClelland, 1987). Even then, multi-layer approaches were computationally taxing and the hardware requirements represented an important bottleneck to research in neural network based CV, which remained disfavored compared to much lighter approaches, such as SVMs.

When compared to the hand-crafted features that dominated the field for most of its history, neural networks learn features from the data itself, therefore eliminating the need for feature engineering (LeCun et al., 2015). In a large part, deep learning approaches for CV have only emerged in force due to two major developments at the beginning of the 21st century. On one side, hardware capability greatly increased due to high consumer demand for personal computing and gaming. On the other, there was a widespread adoption of the internet, leading to an exponential increase in data availability through shared image databases and labelled data. Today, deep learning is a general term that encompasses a wide-variety of approaches that share an architectural commonality of relying on training neural networks with multiple hidden layers (LeCun et al., 2015; O'Mahony et al., 2020). However, this superficial similarity hides a considerable array of differences between

different algorithms and one could say that the field of deep learning is as diverse as the domains in which CV is applied. In ecology and evolution, deep neural networks have been used for essentially any computer vision task, many of each can be seen in other parts of this review. We present some of the most relevant classes of deep learning approaches in Box 2.

Practical considerations for computer vision

Before taking images

Measurement Theory: Define Your Traits Thoughtfully

Defining meaningful phenotypes is deceptively challenging. Traditionally, biologists relied on intuition and natural history conventions to define phenotypes without quantitative verifications of their relevance for biological questions. When deciding what to measure, we suggest that researchers consider *measurement theory,* a qualitative formalization of the relationship between actual measurements and the entity that the measurements are intended to represent (Houle et al., 2011). In phenomics using CV, we recommend that researchers adhere to the following three principles: i) Ensure that the measurements are meaningful in the theoretical context of research questions. ii) Remember that all measurements are estimates. Measurements without uncertainties should always be avoided. iii) Be careful with units and scale types, particularly when composite values, such as the proportion of one measurement over another, are used as a measurement. Wolman (2006) and Houle et al. (2011) give details of measurement theory and practical guidelines for its use in ecology and evolutionary biology.

Image quality: collect images that are maximally useful

As a general rule of thumb, images taken for any CV analysis should have a *signal-to-noise* ratio (SNR) sufficiently high so that the signal (i.e. the phenotypic information) is detectable from the image background. High SNR can be achieved by using high resolution imaging

devices (e.g. DSLR cameras or flatbed scanners), ensuring that the object is in focus and always maintains the same distance to the camera (e.g. by fixing the distance between camera and object), and by creating a high contrast between object and background (e.g. by using backgrounds that are of contrasting colour or brightness to the organism or object). We recommend to iteratively assess suitability of imaging data early on in a project and adjust if necessary. This means taking pilot datasets, processing them, measuring traits, estimating measurement errors, and then updating the image collection process. Moreover, it is good practice to include a colour or size reference whenever possible (e.g. see Fig. 3). It helps researchers to assess if the image has sufficient SNR, increases reproducibility, and helps to evaluate measurement bias as we discuss in the next section.

On measurement error

Because conventional phenotyping methods are often time-consuming and depend on what is possible in a given period of time, biologists are rarely able to evaluate measurement errors and deal with them in downstream analyses. A major advantage of CV lies in its ability to assess the (in)accuracy of measurements easily. Formally, measurement inaccuracy is composed of imprecision and bias, corresponding to random and systematic differences between measured and true values, and can be expressed as the following relationship

$$\text{inaccuracy} = \text{imprecision} + \text{bias}^2$$

(Grabowski and Porto, 2017; Tsuboi et al., 2020). These two sources of errors characterize distinct aspects of a measurement: precise measurements may still be inaccurate if biased, and unbiased measurements may still be inaccurate if imprecise (Fig. 6). Measurement imprecision can be evaluated by the coefficient of variation (standard deviation divided by the mean) of repeated measurements. Bias requires a knowledge of true values.

We ultimately need to understand if a measurement is sufficiently accurate to address the research question at hand. Repeatability is a widely used estimator of measurement accuracy in ecology and evolutionary biology (Wolak et al., 2012), which in our notation could be expressed as

$$\text{repeatability} = 1 - \frac{inaccuracy}{total\ variance}$$

This expression clarifies that the repeatability depends both on measurement inaccuracy and total variance in the data. For example, volume estimates of deer antler from 3D photogrammetry have an average inaccuracy of 8.5%, which results in repeatabilities of 67.8-99.7% depending on the variance in antler volume that a dataset contains (Tsuboi et al., 2020). In other words, a dataset with little variation requires more accurate measurement to achieve the same repeatability as a dataset with more variation. Therefore, the impact of measurement error has to be evaluated in the specific context of data analysis.

One way to improve measurement precision is to repeat a measurement and take their mean as the representative measurement. For example, when measuring deer antler volume estimated from 3D photogrammetry (Fig. 5 Panel E) (Tsuboi et al., 2020), it was found that 70% of the total inaccuracy arose from the error in scaling arbitrary voxel units into real volumetric units. Therefore, by using the mean of two estimates obtained from two copies of an image that are scaled twice independently, the inaccuracy dropped to 5.5%. However, the opportunity to improve accuracy by repeated measurements is limited if a majority of error arises from the stored images themselves. For this reason, we recommend always taking repeated images of the same subject at least for a subset of data. This will allow evaluating the magnitude of error due to images relative to the error due to acquisition of measurements from images. If the error caused by images is large compared to the error

caused by data acquisition, it may be necessary to modify imaging and/or preprocessing protocol to increase SNR.

Assessing measurement bias requires separate treatments. When linear (length) or chromatic (color) measurements are obtained from images, it is a good general practice to include size and color scales as part of images to estimate bias as the difference between known values of imaged scales and measurements obtained through CV (i.e. the reference card in Fig. 3). Knowing the true value may be difficult in some cases, such as area or circularity of a trait (Hoffmann et al., 2018), since they are hard to characterize without a CV. When multiple independent methods to measure the same character exist, we recommend using them on sample data to determine the bias of one method relative to the other.

After taking images

Selecting a CV pipeline: As simple as possible, as complex as necessary

When using CV tools there are usually many different ways to collect a specific type of phenotypic information from images (Fig. 4). Therefore, one of the first hurdles to overcome when considering the use of CV is selecting the appropriate technique from among a large and growing set of choices. The continued emergence of novel algorithms to collect, process and analyze image-derived data may sometimes make us believe that any “older” technology is immediately outdated. Deep learning, specifically CNNs, is a prominent example of an innovation in CV that was frequently communicated as so “revolutionary” and “transformative” that many scientists believed it would replace all existing methods. However, despite the success of CNNs, there are many cases where they are inappropriate or unfeasible, e.g. due to small sample sizes, hardware or time constraints, or because of the complexity that deep learning implementations entail, despite many efforts to make this technology more tractable. We discourage readers from defaulting to using the newest technology stacks; rather, we suggest that researchers be pragmatic as to which is the

fastest and simplest way to get the phenotypic information of desire from any given set of images (Fig. 4, Table 2).

Begin by considering the size of a given image dataset, whether it is complete or whether there will be continued future additions, e.g. as part of a long term experiment or field survey. As a rough rule of thumb, if a dataset encompasses only a thousand images or fewer, consider it “small”; if a dataset has thousands to tens of thousands images, consider it “large” (see Fig. 4 for methodological suggestions for each case). The next assessment should be about the SNR in your images: images taken in the laboratory typically have a high degree of standardization, e.g. controlled light environment or background, and thus a very high SNR. Field images can also have a high SNR, for example, if they are taken against the sky or if the trait of question is very distinct from the background through bright colouration. If the dataset is “small” and/or has high SNR, it may not be necessary to use the more sophisticated CV tools; instead, signal processing, e.g. threshold or watershed algorithms, may already be sufficient for segmentation although typically some pre- and post processing is typically still required (e.g. blurring to remove noise, “morphology”-operations to close gaps, or masking false positives).

For large datasets, images with low SNR, or if the information of interest is variable across images (e.g traits are photographed from different angles or partially covered up), machine learning approaches are probably more suitable. In contrast to signal processing, where segmentation results are immediately available, all machine learning image analysis pipelines include iterative training and validation phases, followed by a final testing phase. Such a workflow can be complex to initiate, but pays off in the long run by providing results that become increasingly robust if more training data is supplied over time. Classic machine learning algorithms often require an intermediate amount of training data (500-1000 or more images) before they can produce satisfactory results (Schneider et al., 2020a). In this category, SVM or HOG algorithms are a good choice when areas of interest do not contrast

sufficiently from the surrounding area, for example, when automatically detecting landmarks (Porto and Voje, 2020a). Deep learning algorithms require much larger training datasets (minimum 1000 to 10000 images), but are less sensitive to noise and idiosyncrasies of the foreground. Thus, for large and continuously growing data sets, or for recurring image analysis tasks, deep learning has become the standard approach for segmentation (Sultana et al., 2020). Deeper networks may increase model accuracy, and thus improve the segmentation results, but have an increasing risk of overfitting the contained information - i.e. the model is less generalizable to input data. While the implementation of deep learning pipelines may require more expertise than other CV-techniques, they can be retrained and are typically less domain specific than classic machine learning pipelines (O'Mahony et al., 2020).

Recent examples of computer vision to collect phenomic data

"Phenomics" as a term has not yet gained widespread attention in the ecological and evolutionary biology research communities (Fig. 1), but many biologists are engaged in research programmes that are collecting phenomic data, even though it is not called as such. Some of them are already using automatic or semi-automatic CV to collect phenotypic data. Here we present small a selection of promising applications of CV to answer ecological or evolutionary research questions (points matching panels in Fig. 5):

i. **Shape and texture of resource competition traits -** Species diversity within ecological communities is often thought to be governed by competition for limiting resources (Chesson, 2000). However, the exact traits that make species or individuals the best competitors under resource limitation conditions are difficult to identify among all other traits. In this example, the phenotypic space underlying resource competition was explored by implementing different limitation scenarios for experimental phytoplankton communities. Images were taken with a plate reader that used a combination of visible light and fluorometry measurements (Hense et al.,

2008). The images were analyzed using signal processing, which allowed the rapid segmentation of several 1000 images by combining information from multiple fluorescence emission excitation spectra to an image stack. As a result, over 100 traits related to morphology (shape, size, and texture) and internal physiology (pigment content, distribution of pigments within each cell) were obtained at the individual cell level. (Gallego et al., unpublished data)

ii. **Thermal adaptation and thermal reaction norms** - Variation in body temperature can be an important source of fitness variation (Kingsolver and Huey, 2008; Svensson et al., 2020). Quantifying body temperature and thermal reaction norms in response to natural and sexual selection allows us to test predictions from evolutionary theory about phenotypic plasticity and canalization (Lande, 2009; Chevin et al., 2010). However, body temperature is an internal physiological trait that is difficult to quantify in a non-invasive way on many individuals simultaneously and under natural conditions. Thermal imaging is an efficient and non-invasive method to quantify such physiological phenotypes on a large scale and can be combined with thermal loggers to measure local thermal environmental conditions in the field (Svensson and Waller, 2013; Svensson et al., 2020).

iii. **Stochastically patterned morphological traits** - In contrast to homologous, landmark-based morphological traits, tissues also form emergent patterns that are unique to every individual. The arrangement of veins on the wings of damselflies is one such example. By measuring the spacing, angles, and connectivities within the adult wing tissue, researchers have proposed hypotheses about the mechanisms of wing development and physical constraints on wing evolution (Hoffmann et al., 2018; Salcedo et al., 2019).

iv. **Morphometrics and shape of complex structures** - Landmark-based morphometrics has become a popular tool used to characterize morphological

variation in complex biological structures. Despite its popularity, landmark data is still collected mainly through manual annotation, a process which represents a significant bottleneck for phenomic studies. However, machine-learning-based CV can be used to accurately automate landmark data collection in morphometric studies not only in 2D (McPeek et al., 2008; Porto and Voje, 2020b), but also in 3D (Porto et al., 2020).

v. **Volumes of morphologically complex traits**. Many topics in evolutionary ecology concerns investment of resources into a particular trait. However, measuring energetic investment, either as mass or volume of the target traits, has been challenging because many traits are morphologically complex, making it difficult to estimate investment from a combination of linear measurements. Photogrammetry is a low-cost and fast technique to create 3D surface images from a set of images. Using a simple protocol and a free proprietary software, Tsuboi et al. (2020) demonstrated that photogrammetry can accurately measure the volume of antler in deer family Cervidae. The protocol is still relatively low-throughput due primarily to the necessity of high number of images (> 50) per sample, but it allows extensive sampling (*sensu* (Houle et al., 2010)) of linear, area and volumetric measurements of antler structures.

Outlook

In this review we provided a broad overview of various CV techniques and gave some recent examples of their application in ecological and evolutionary research. We presented CV as a promising toolkit to overcome the image analysis bottleneck in phenomics. However, to be clear, we do not suggest that biologists discontinue the collection of univariate traits like body size or discrete colours. Such measures are undoubtedly useful, if they contain explanatory value and predictive power. Instead, we propose that CV can help to i) collect them with higher throughput, ii) in a more reproducible fashion, and to iii) collect additional

traits so we can interpret them in the context of trait combinations. We argue that CV is not bound to immediately replace existing methods, but it simply opens the opportunity to place empirical research of phenotypes on a broader base. We also note that CV based phenomics can be pursued in a deductive or inductive fashion. In the former case, scientists would simply conduct hypothesis driven research including a wider array of traits into causal models (Houle et al., 2011); in the latter, they would engage in discovery-based data-mining approaches that allow scientists to form hypotheses a posteriori based on the collected data (Kell and Oliver, 2004).

Although CV based phenomics provides new opportunities for many areas of study, we identify several specific fields that will profit most immediately from CV. First, evolutionary quantitative genetics will benefit tremendously from increased sample sizes that CV-phenomics entails, because the bottleneck of the field has been the difficulty in accurately estimating key statistics such as genetic variance covariance matrices and selection gradients. The recent discovery of tight matches between mutational, genetic, and macroevolutionary variances in Drosophilid wing shape (Houle et al., 2017) is exemplary of a successful phenomic project. Second, large-scale empirical studies of the genotype-phenotype map will finally become possible, because of the availability of high-throughput phenotypic data and analytical framework to deal with big data (Pitchers et al., 2019; Zheng et al., 2019; Maeda et al., 2020). Third, studies of fossil time-series will gain opportunities to document and analyze the dynamics of long-term phenotypic evolution with unprecedented temporal resolution (Brombacher et al., 2017; Liow et al., 2017). The ever-growing technology of CV indicates that these are likely a small subset of unforeseen future applications of CV phenomics in our field. Similar to the technological advancements in DNA-sequencing that have revolutionized our view of genomes, development and molecular evolution in the past decades, we anticipate that the way we look at phenotypic data will be changing in the coming years.

Just as CV is changing what it means to measure a trait, there is a complementary change in what can be considered scientific image data in the first place. Large, publicly available image datasets are fertile ground for ecology and evolutionary research. Such databases include both popular and non-scientific social media (e.g. Flickr or Instagram), but also quality-controlled and vetted natural history and species identification resources with global scope and ambitions (e.g. iNaturalist). Successful examples of how such public image databases can be useful are in studies aiming to quantify the frequencies variation of discrete traits, such as colour polymorphism frequencies in different geographic regions (Leighton et al., 2016). These manual efforts in mining available public image resources can potentially be replaced in the future using more automated machine learning or CV approaches. Similarly, the corpus of published scientific literature is full of image data that can be combined and re-analyzed in order to address larger-scale questions (Hoffmann et al., 2018; Church et al., 2019a, 2019b).

Previous calls for phenomics argued that, to make phenomics a successful endeavour, it has to be extensive, aiming at measuring many different aspects of the phenotypes, as well as intensive, aiming at characterizing each measurement accurately with large sample size and with high temporal resolution (Bilder et al., 2009; Houle et al., 2010; Furbank and Tester, 2011). We agree with this view, but we also emphasize that phenomics is nothing conceptually new in this respect. As discussed above, many researchers in our field have already adopted phenomic pipelines, i.e. they are collecting high-dimensional phenotypic data on a large scale, but they may not be using the term "phenomics". If so, what is the conceptual value and added benefit of explicitly studying phenomes? We argue that computer vision and other techniques will facilitate the rigorous quantification of phenomes in the same fashion as next generation genomics allows scientists to move away from a few markers of interest to simply reading all molecular data that is available. While it may not be possible to extract phenomes in a complete fashion

(Houle et al., 2010), a "phenomics-mindset" still gives us the opportunity to collect and analyze larger amounts of phenotypic data with virtually no extra cost. Taking images and analyzing them with computer vision enables biologists to choose freely between conducting conventional research of phenotypes, but with higher throughput and in a reproducible fashion, or to "harness the power big data" (Peters et al., 2014) for the study of high dimensional phenotypic data.

Acknowledgements

The publication of this study was funded through the Swedish Research Council International Postdoc Grant (2016-06635) to MT. ML was supported by a Swiss National Science Foundation Early Postdoc.Mobility grant (SNSF: P2EZP3_191804). EIS was funded by a grant from the Swedish Research Council (VR: grant no. 2016-03356). SD was supported by the Jane Coffin Childs Memorial Fund.

Authors contributions

ML conceived the idea for this review and initiated its writing. In the process, all authors contributed equally to the development and discussion of ideas, and to the writing of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., et al. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1803.01164.

Bateson, A., and Curtiss, B. (1996). A method for manual endmember selection and spectral unmixing. *Remote Sens. Environ.* 55, 229–243. doi:10.1016/S0034-4257(95)00177-8.

Bauer, E., and Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Mach. Learn.* 36, 105–139. doi:10.1023/A:1007515423169.

Bertin, E., Marcelpoil, R., and Chassery, J.-M. (1992). Morphological algorithms based on Voronoi and Delaunay graphs: microscopic and medical applications. in *Image Algebra and Morphological Image Processing III* (International Society for Optics and Photonics), 356–367. doi:10.1117/12.60655.

Beucher, S. (1979). Use of watersheds in contour detection. *Proceedings of the International Workshop on Image*. Available at: https://ci.nii.ac.jp/naid/10008961959/.

Bilder, R. M., Sabb, F. W., Cannon, T. D., London, E. D., Jentsch, J. D., Parker, D. S., et al. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience* 164, 30–42. doi:10.1016/j.neuroscience.2009.01.027.

Blonder, B. (2018). Hypervolume concepts in niche- and trait-based ecology. *Ecography* 41, 1441–1455. doi:10.1111/ecog.03187.

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/A:1010933404324.

Brombacher, A., Wilson, P. A., Bailey, I., and Ezard, T. H. G. (2017). The Breakdown of Static and Evolutionary Allometries during Climatic Upheaval. *Am. Nat.* 190, 350–362. doi:10.1086/692570.

Bruijning, M., Visser, M. D., Hallmann, C. A., and Jongejans, E. (2018). trackdem : Automated particle tracking to obtain population counts and size distributions from videos in r. *Methods Ecol. Evol.* 9, 965–973. doi:10.1111/2041-210X.12975.

Buetti-Dinh, A., Galli, V., Bellenberg, S., Ilie, O., Herold, M., Christel, S., et al. (2019). Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. *Biotechnol Rep (Amst)* 22, e00321. doi:10.1016/j.btre.2019.e00321.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-8, 679–698. doi:10.1109/TPAMI.1986.4767851.

Cheng, K. C., Xin, X., Clark, D. P., and La Riviere, P. (2011). Whole-animal imaging, gene function, and the Zebrafish Phenome Project. *Curr. Opin. Genet. Dev.* 21, 620–629. doi:10.1016/j.gde.2011.08.006.

Chesson, P. (2000). Mechanisms of Maintenance of Species Diversity. *Annu. Rev. Ecol. Syst.* 31, 343–366. doi:10.1146/annurev.ecolsys.31.1.343.

Chevin, L.-M., Lande, R., and Mace, G. M. (2010). Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biol.* 8, e1000357. doi:10.1371/journal.pbio.1000357.

Church, G. M., and Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 81, 1991–1995. doi:10.1073/pnas.81.7.1991.

Church, S. H., Donoughe, S., de Medeiros, B. A. S., and Extavour, C. G. (2019a). A dataset of egg size and shape from more than 6,700 insect species. *Sci Data* 6, 104. doi:10.1038/s41597-019-0049-y.

Church, S. H., Donoughe, S., de Medeiros, B. A. S., and Extavour, C. G. (2019b). Insect egg size and shape evolve with ecology but not developmental rate. *Nature* 571, 58–62. doi:10.1038/s41586-019-1302-4.

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.* 20, 273–297. doi:10.1023/A:1022627411411.

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. in *Proceedings of the First International Workshop on Multiple Classifier Systems* MCS '00. (Berlin, Heidelberg: Springer-Verlag), 1–15. Available at: https://dl.acm.org/citation.cfm?id=648054.743935 [Accessed December 7, 2020].

Di, Z., Klop, M. J. D., Rogkoti, V.-M., Le Dévédec, S. E., van de Water, B., Verbeek, F. J., et al. (2014). Ultra high content image analysis and phenotype profiling of 3D cultured micro-tissues. *PLoS One* 9, e109688. doi:10.1371/journal.pone.0109688.

Duda, R. O., and Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM* 15, 11–15. doi:10.1145/361237.361242.

Edgington, D. R., Cline, D. E., Davis, D., Kerkez, I., and Mariette, J. (2006). Detecting, Tracking and Classifying Animals in Underwater Video. in *OCEANS 2006*, 1–5. doi:10.1109/OCEANS.2006.306878.

Feder, M. E., and Mitchell-Olds, T. (2003). Evolutionary and ecological functional genomics. *Nat. Rev. Genet.* 4, 651–657. doi:10.1038/nrg1128.

Fossum, E. R., and Hondongwa, D. B. (2014). A review of the pinned photodiode for CCD and CMOS image sensors. *IEEE J. Electron Devices Soc.* 2, 33–43. doi:10.1109/jeds.2014.2306412.

Freimer, N., and Sabatti, C. (2003). The human phenome project. *Nat. Genet.* 34, 15–21. doi:10.1038/ng0503-15.

French, S., Coutts, B. E., and Brown, E. D. (2018). Open-Source High-Throughput Phenomics of Bacterial Promoter-Reporter Strains. *Cell Syst* 7, 339–346.e3.

doi:10.1016/j.cels.2018.07.004.

Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. in *Annals of Statistics* Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.9093 [Accessed December 7, 2020].

Furbank, R. T., and Tester, M. (2011). Phenomics--technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644. doi:10.1016/j.tplants.2011.09.005.

Gehan, M. A., Fahlgren, N., Abbasi, A., Berry, J. C., Callen, S. T., Chavez, L., et al. (2017). PlantCV v2: Image analysis software for high-throughput plant phenotyping. *PeerJ* 5, e4088. doi:10.7717/peerj.4088.

Geoffrey Hinton, Terrence J. Sejnowski (1999). *Unsupervised Learning: Foundations of Neural Computation*. MIT Press Available at: https://play.google.com/store/books/details?id=yj04Y0lje4cC.

Gerum, R. C., Richter, S., Fabry, B., and Zitterbart, D. P. (2017). ClickPoints: an expandable toolbox for scientific image annotation and analysis. *Methods Ecol. Evol.* 8, 750–756. doi:10.1111/2041-210X.12702.

Goesele, M. (2004). *New acquisition techniques for real objects and light sources in computer graphics*. Books on Demand Available at: http://download.hrz.tu-darmstadt.de/media/FB20/GCC/paper/Diss-Goesele-LossyCompression.pdf [Accessed February 2, 2021].

Grabowski, M., and Porto, A. (2017). How many more? Sample size determination in studies of morphological integration and evolvability. *Methods Ecol. Evol.* 8, 592–603. doi:10.1111/2041-210X.12674.

Greenspan, H., Ruf, A., and Goldberger, J. (2006). Constrained Gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Trans. Med. Imaging* 25, 1233–1245. doi:10.1109/tmi.2006.880668.

Hakim, A., Mor, Y., Toker, I. A., Levine, A., Neuhof, M., Markovitz, Y., et al. (2018). WorMachine: machine learning-based phenotypic analysis tool for worms. *BMC Biol.* 16, 8. doi:10.1186/s12915-017-0477-0.

Heaton, J. (2020). Applications of Deep Neural Networks. *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/2009.05673.

Heileman, G. L., and Myler, H. R. (1989). Theoretical and experimental aspects of supervised learning in artificial neural networks. Available at: https://dl.acm.org/citation.cfm?id=915701.

Hense, B. A., Gais, P., Jutting, U., Scherb, H., and Rodenacker, K. (2008). Use of fluorescence information for automated phytoplankton investigation by image analysis. *Journal of Plankton Research* 30, 587–606. doi:10.1093/plankt/fbn024.

Hoffmann, J., Donoughe, S., Li, K., Salcedo, M. K., and Rycroft, C. H. (2018). A simple developmental model recapitulates complex insect wing venation patterns. *Proc. Natl.*

Acad. Sci. U. S. A. 115, 9905–9910. doi:10.1073/pnas.1721248115.

Hooper, D. U., Chapin, F. S., III, Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., et al. (2005). EFFECTS OF BIODIVERSITY ON ECOSYSTEM FUNCTIONING: A CONSENSUS OF CURRENT KNOWLEDGE. *Ecol. Monogr.* 75, 3–35. doi:10.1890/04-0922.

Houle, D., Bolstad, G. H., van der Linde, K., and Hansen, T. F. (2017). Mutation predicts 40 million years of fly wing evolution. *Nature* 548, 447–450. doi:10.1038/nature23473.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866. doi:10.1038/nrg2897.

Houle, D., Mezey, J., Galpern, P., and Carter, A. (2003). Automated measurement of Drosophila wings. *BMC Evol. Biol.* 3, 25. doi:10.1186/1471-2148-3-25.

Houle, D., Pélabon, C., Wagner, G. P., and Hansen, T. F. (2011). Measurement and Meaning in Biology. *The Quarterly Review of Biology* 86, 3–34. doi:10.1086/658408.

Høye, T. T., Ärje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., et al. (2020). Deep learning and computer vision will transform entomology. *Ecology*, 760. doi:10.1101/2020.07.03.187252.

Hsiang, A. Y., Nelson, K., Elder, L. E., Sibert, E. C., Kahanamoku, S. S., Burke, J. E., et al. (2018). AutoMorph : Accelerating morphometrics with automated 2D and 3D image processing and shape extraction. *Methods Ecol. Evol.* 9, 605–612. doi:10.1111/2041-210X.12915.

Hu, Q., and Davis, C. (2005). Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. *Mar. Ecol. Prog. Ser.* 295, 21–31. doi:10.3354/meps295021.

Ishikawa, A., Kabeya, N., Ikeya, K., Kakioka, R., Cech, J. N., Osada, N., et al. (2019). A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science* 364, 886–889. doi:10.1126/science.aau5656.

Kanopoulos, N., Vasanthavada, N., and Baker, R. L. (1988). Design of an image edge detection filter using the Sobel operator. *IEEE J. Solid-State Circuits* 23, 358–367. doi:10.1109/4.996.

Kell, D. B., and Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* 26, 99–105. doi:10.1002/bies.10385.

Kingsolver, J. G., and Huey, R. B. (2008). Size, temperature, and fitness: three rules. *Evol. Ecol. Res.* 10, 251–268. Available at: http://www.evolutionary-ecology.com/abstracts/v10/2242.html.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 1097–1105. Available at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural

-networks.pdf.

Krogh, P. S. A., and Others (1996). Learning with ensembles: How over-fitting can be useful. in *Proceedings of the 1995 Conference*, 190. Available at: https://books.google.com/books?hl=en&lr=&id=ZkJrSots_SAC&oi=fnd&pg=PA190&dq=Sollich+P+and+Krogh+A+Learning+with+ensembles+How+overfitting+can+be+useful+Advances+in+Neural+Information+Processing+Systems+volume+8+pp+190-196+1996+)&ots=YkHQDTP9QY&sig=aHs8hk2oHxX848AQgqrXFH4dn-A.

Kühl, H. S., and Burghardt, T. (2013). Animal biometrics: quantifying and detecting phenotypic appearance. *Trends Ecol. Evol.* 28, 432–441. doi:10.1016/j.tree.2013.02.013.

Lamichhaney, S., Card, D. C., Grayson, P., Tonini, J. F. R., Bravo, G. A., Näpflin, K., et al. (2019). Integrating natural history collections and comparative genomics to study the genetic architecture of convergent evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374, 20180248. doi:10.1098/rstb.2018.0248.

Lande, R. (2009). Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. *J. Evol. Biol.* 22, 1435–1446. doi:10.1111/j.1420-9101.2009.01754.x.

Lande, R., and Arnold, S. J. (1983). The Measurement of Selection on Correlated Characters. *Evolution* 37, 1210–1226. doi:10.2307/2408842.

Laughlin, D. C., Gremer, J. R., Adler, P. B., Mitchell, R. M., and Moore, M. M. (2020). The Net Effect of Functional Traits on Fitness. *Trends Ecol. Evol.* 35, 1037–1047. doi:10.1016/j.tree.2020.07.010.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539.

Leighton, G., Hugo, P. S., Roulin, A., and Amar, A. (2016). Just Google it: assessing the use of Google Images to describe geographical variation in visible traits of organisms. 7. doi:10.1111/2041-210X.12562.

Le, V.-L., Beurton-Aimar, M., Zemmari, A., Marie, A., and Parisey, N. (2020). Automated landmarking for insects morphometric analysis using deep neural networks. *Ecol. Inform.* 60, 101175. doi:10.1016/j.ecoinf.2020.101175.

Liow, L. H., Di Martino, E., Krzeminska, M., Ramsfjell, M., Rust, S., Taylor, P. D., et al. (2017). Relative size predicts competitive outcome through 2 million years. *Ecol. Lett.* 20, 981–988. doi:10.1111/ele.12795.

Liu, F., Wollstein, A., Hysi, P. G., Ankra-Badu, G. A., Spector, T. D., Park, D., et al. (2010). Digital Quantification of Human Eye Color Highlights Genetic Association of Three New Loci. *PLoS Genetics* 6, e1000934. doi:10.1371/journal.pgen.1000934.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28, 129–137. doi:10.1109/TIT.1982.1056489.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1150–1157 vol.2.

doi:10.1109/ICCV.1999.790410.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* Available at: https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1023/B:VISI.0000029664.99615.94&casa_token=QZ343Z2JCdEAAAAA:mFB_O6l-9B2rwvVzDwW3vkDIWTIQ0qVpQwFMxhdkon0STu6nUtYN11qASAc4tskHr9hIr2PFptaiCL616-A.

Lürig, M. D. (2018). *phenopype - a phenotyping pipeline for python*. Available at: https://doi.org/10.5281/zenodo.3483222.

Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., et al. (2010). Automated processing and identification of benthic invertebrate samples. *J. North Am. Benthol. Soc.* 29, 867–874. doi:10.1899/09-080.1.

Maeda, T., Iwasawa, J., Kotani, H., Sakata, N., Kawada, M., Horinouchi, T., et al. (2020). High-throughput laboratory evolution reveals evolutionary constraints in Escherichia coli. *Nat. Commun.* 11, 5970. doi:10.1038/s41467-020-19713-w.

Mäkelä, T., Clarysse, P., Sipilä, O., Pauna, N., Pham, Q. C., Katila, T., et al. (2002). A review of cardiac image registration methods. *IEEE Trans. Med. Imaging* 21, 1011–1021. doi:10.1109/TMI.2002.804441.

McPeek, M. A., Shen, L., Torrey, J. Z., and Farid, H. (2008). The Tempo and Mode of Three-Dimensional Morphological Evolution in Male Reproductive Structures. *The American Naturalist* 171, E158–E178. doi:10.1086/587076.

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., et al. (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* 16, e2005970. doi:10.1371/journal.pbio.2005970.

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE* 49, 8–30. doi:10.1109/JRPROC.1961.287775.

Minsky, M, Papert, and S (1969). Perceptrons. Available at: http://papers.cumincad.org/cgi-bin/works/Show?_id=b029 [Accessed December 7, 2020].

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* 45, 870–877.

Morel-Journel, T., Thuillier, V., Pennekamp, F., Laurent, E., Legrand, D., Chaine, A. S., et al. (2020). A multidimensional approach to the expression of phenotypic plasticity. *Funct. Ecol.* Available at: https://sci-hub.do/https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2435.13667?casa_token=YZuBrxIYk0sAAAAA:Dcs3R_I3l0vQoqOJkD7O_CeQRpFum9ZziU42cttkX8r29V54hY_vah1L6Bu0OPOiin3cd0XZrxjLuX0.

Mortensen, E. N., Delgado, E. L., Deng, H., Lytle, D., Moldenke, A., Paasch, R., et al. (2007). Pattern Recognition for Ecological Science and Environmental Monitoring: An Initial Report. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.2058&rep=rep1&type=pdf

[Accessed February 10, 2021].

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., et al. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl. Acad. Sci. U. S. A.* 115, E5716–E5725. doi:10.1073/pnas.1719367115.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., et al. (2020). Deep Learning vs. Traditional Computer Vision. *Advances in Intelligent Systems and Computing*, 128–144. doi:10.1007/978-3-030-17795-9_10.

Orgogozo, V., Morizot, B., and Martin, A. (2015). The differential view of genotype-phenotype relationships. *Front. Genet.* 6, 179. doi:10.3389/fgene.2015.00179.

Petchey, O. L., and Gaston, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecol. Lett.* 9, 741–758. doi:10.1111/j.1461-0248.2006.00924.x.

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere* 5, art67. doi:10.1890/es13-00359.1.

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D., and Moczek, A. P. (2010). Phenotypic plasticity's impacts on diversification and speciation. *Trends Ecol. Evol.* 25, 459–467. doi:10.1016/j.tree.2010.05.006.

Phillips, P. C., and Arnold, S. J. (1999). Hierarchical Comparison of Genetic Variance-Covariance Matrices. I. Using the Flury Hierarchy. *Evolution* 53, 1506–1515. doi:10.2307/2640896.

Piccardi, M. (2004). Background subtraction techniques: a review. in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 3099–3104 vol.4. doi:10.1109/ICSMC.2004.1400815.

Pitchers, W., Nye, J., Márquez, E. J., Kowalski, A., Dworkin, I., and Houle, D. (2019). A Multivariate Genome-Wide Association Study of Wing Shape in Drosophila melanogaster. *Genetics* 211, 1429–1447. doi:10.1534/genetics.118.301342.

Pointer, M. R., and Attridge, G. G. (1998). The number of discernible colours. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 23, 52–54. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1520-6378(199802)23:1%3C52::AID-COL8%3E3.0.CO;2-2?casa_token=7LmOYEdzguoAAAAA:pg8hhYjlLpw6USMpbyZwgK8ZEt2mF6AtsPDy_82LdaU-15tzK6I-XheHSb_6ejjUumAQgwG8MJu_CT4.

Porto, A., Rolfe, S. M., and Murat Maga, A. (2020). ALPACA: a fast and accurate approach for automated landmarking of three-dimensional biological structures. *Cold Spring Harbor Laboratory*, 2020.09.18.303891. doi:10.1101/2020.09.18.303891.

Porto, A., and Voje, K. L. (2020a). ML-morph: A fast, accurate and general approach for

automated detection and landmarking of biological structures in images. *Methods Ecol. Evol.* doi:10.1111/2041-210x.13373.

Porto, A., and Voje, K. L. (2020b). ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. *Methods Ecol. Evol.* 11, 500–512. Available at: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13373.

Reynolds, D. A., and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Audio Speech Lang. Processing* 3, 72–83. doi:10.1109/89.365379.

Ringnér, M. (2008). What is principal component analysis? *Nat. Biotechnol.* 26, 303–304. doi:10.1038/nbt0308-303.

Roberts, L. G. (1963). Machine perception of three-dimensional solids. Available at: https://dspace.mit.edu/handle/1721.1/11589?show=full [Accessed December 7, 2020].

Rodenacker, K., Brühl, A., Hausner, M., Kühn, M., Liebscher, V., Wagner, M., et al. (2000). Quantification of biofilms in multi-spectral digital1 volumes from confocal laser-scanning microscopes. *Image Anal. Stereol.* 19, 151. doi:10.5566/ias.v19.p151-156.

Roeder, A. H. K., Cunha, A., Burl, M. C., and Meyerowitz, E. M. (2012). A computational image analysis glossary for biologists. *Development* 139, 3071–3080. doi:10.1242/dev.076414.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi:10.1037/h0042519.

Rosten, E., and Drummond, T. (2006). Machine Learning for High-Speed Corner Detection. in *Computer Vision – ECCV 2006* (Springer Berlin Heidelberg), 430–443. doi:10.1007/11744023_34.

Rumelhart, D. E., and McClelland, J. L. (1987). “Learning Internal Representations by Error Propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (MIT Press), 318–362. Available at: https://ieeexplore.ieee.org/document/6302929.

Salcedo, M. K., Hoffmann, J., Donoughe, S., and Mahadevan, L. (2019). Computational analysis of size, shape and structure of insect wings. *Biol. Open* 8. doi:10.1242/bio.040774.

Saltz, J. B., Hessel, F. C., and Kelly, M. W. (2017). Trait Correlations in the Genomics Era. *Trends Ecol. Evol.* doi:10.1016/j.tree.2016.12.008.

Sanchez-Hernandez, C., Boyd, D. S., and Foody, G. M. (2007). Mapping specific habitats from remotely sensed imagery: Support vector machine and support vector data description based classification of coastal saltmarsh habitats. *Ecol. Inform.* 2, 83–88. doi:10.1016/j.ecoinf.2007.04.003.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods* 9,

676–682. doi:10.1038/nmeth.2019.

Schluter, D. (1996). Adaptive radiation along genetic lines of least resistance. *Evolution* 50, 1766–1774. doi:10.1111/j.1558-5646.1996.tb03563.x.

Schneider, S., Greenberg, S., Taylor, G. W., and Kremer, S. C. (2020a). Three critical factors affecting automated image species recognition performance for camera traps. *Ecol. Evol.* 10, 3503–3517. doi:10.1002/ece3.6147.

Schneider, S., Taylor, G. W., and Kremer, S. C. (2020b). Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 44–52. Available at: http://openaccess.thecvf.com/content_WACVW_2020/html/w2/Schneider_Similarity_Learning_Networks_for_Animal_Individual_Re-Identification_-_Beyond_the_WACVW_2020_paper.html.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., et al. (2014). Genomics and the origin of species. *Nat. Rev. Genet.* 15, 176–192. doi:10.1038/nrg3644.

Shapiro, L. G., and Stockman, G. C. (2001). *Computer Vision*. Prentice Hall Available at: https://play.google.com/store/books/details?id=FftDAQAAIAAJ.

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 60. doi:10.1186/s40537-019-0197-0.

Sinervo, B., and Svensson, E. (2002). Correlational selection and the evolution of genomic architecture. *Heredity* 89, 329–338. doi:10.1038/sj.hdy.6800148.

Soulé, M. (1967). PHENETICS OF NATURAL POPULATIONS I. PHENETIC RELATIONSHIPS OF INSULAR POPULATIONS OF THE SIDE-BLOTCHED LIZARD. *Evolution* 21, 584–591. doi:10.1111/j.1558-5646.1967.tb03413.x.

Sultana, F., Sufian, A., and Dutta, P. (2020). Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/2001.04074.

Svensson, E. I., Arnold, S. J., Bürger, R., Csilléry, K., Draghi, J., Henshaw, J. M., et al. (2021). Correlational selection in the age of genomics. *Nature Ecology and Evolution*.

Svensson, E. I., Gomez-Llano, M., and Waller, J. (2020). Natural and sexual selection on phenotypic plasticity favour thermal canalization. *Cold Spring Harbor Laboratory*, 2020.06.11.146043. doi:10.1101/2020.06.11.146043.

Svensson, E. I., and Waller, J. T. (2013). Ecology and sexual selection: evolution of wing pigmentation in calopterygid damselflies in relation to latitude, sexual dimorphism, and speciation. *Am. Nat.* 182, E174–95. doi:10.1086/673206.

Tattersall, G. J., Andrade, D. V., and Abe, A. S. (2009). Heat exchange from the toucan bill reveals a controllable vascular thermal radiator. *Science* 325, 468–470. doi:10.1126/science.1175553.

Tattersall, G. J., and Cadena, V. (2010). Insights into animal temperature adaptations revealed through thermal imaging. *The Imaging Science Journal* 58, 261–268. doi:10.1179/136821910X12695060594165.

Tsubaki, Y., Samejima, Y., and Siva-Jothy, M. T. (2010). Damselfly females prefer hot males: higher courtship success in males in sunspots. *Behav. Ecol. Sociobiol.* 64, 1547–1554. doi:10.1007/s00265-010-0968-2.

Tsuboi, M., Kopperud, B. T., Syrowatka, C., Grabowski, M., Voje, K. L., Pélabon, C., et al. (2020). Measuring Complex Morphological Traits with 3D Photogrammetry: A Case Study with Deer Antlers. *Evol. Biol.* 47, 175–186. doi:10.1007/s11692-020-09496-9.

Turk, M., and Pentland, A. (1991). Eigenfaces for recognition. *J. Cogn. Neurosci.* 3, 71–86. doi:10.1162/jocn.1991.3.1.71.

Ubbens, J. R., and Stavness, I. (2017). Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks. *Front. Plant Sci.* 8, 1190. doi:10.3389/fpls.2017.01190.

Valan, M., Makonyi, K., Maki, A., Vondráček, D., and Ronquist, F. (2019). Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. *Syst. Biol.* 68, 876–895. doi:10.1093/sysbio/syz014.

Villéger, S., Mason, N. W. H., and Mouillot, D. (2008). New multidimensional functional diversity indices for a multifaceted framework in functional ecology. *Ecology* 89, 2290–2301. doi:10.1890/07-1206.1.

Visscher, P. M., and Yang, J. (2016). A plethora of pleiotropy across complex traits. *Nature Genetics* 48, 707–708. doi:10.1038/ng.3604.

Wäldchen, J., and Mäder, P. (2018). Machine learning for image based species identification. *Methods Ecol. Evol.* 9, 2216–2225. doi:10.1111/2041-210X.13075.

Walsh, B. (2007). Escape from flatland. *J. Evol. Biol.* 20, 36–8; discussion 39–44. doi:10.1111/j.1420-9101.2006.01218.x.

Weinstein, B. G. (2015). MotionMeerkat: integrating motion video detection and ecological monitoring. *Methods Ecol. Evol.* 6, 357–362. doi:10.1111/2041-210X.12320.

Wessman, C. A., Bateson, C. A., and Benning, T. L. (1997). Detecting fire and grazing patterns in tallgrass prairie using spectral mixture analysis. *Ecol. Appl.* 7, 493–511. doi:10.1890/1051-0761(1997)007[0493:dfagpi]2.0.co;2.

Williams, J. B. (2017). “Electronics Invades Photography: Digital Cameras,” in *The Electronics Revolution: Inventing the Future*, ed. J. B. Williams (Cham: Springer International Publishing), 243–250. doi:10.1007/978-3-319-49088-5_26.

Wolak, M. E., Fairbairn, D. J., and Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods Ecol. Evol.* 3, 129–137. doi:10.1111/j.2041-210X.2011.00125.x.

Wolman, A. G. (2006). Measurement and meaningfulness in conservation science. *Conserv.*

Biol. 20, 1626–1634. doi:10.1111/j.1523-1739.2006.00531.x.

Zackrisson, M., Hallin, J., Ottosson, L.-G., Dahl, P., Fernandez-Parada, E., Ländström, E., et al. (2016). Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale. *G3* 6, 3003–3014. doi:10.1534/g3.116.032342.

Zhang, Y., and Wu, L. (2011). Optimal Multi-Level Thresholding Based on Maximum Tsallis Entropy via an Artificial Bee Colony Approach. *Entropy* 13, 841–859. doi:10.3390/e13040841.

Zheng, J., Payne, J. L., and Wagner, A. (2019). Cryptic genetic variation accelerates evolution by opening access to diverse adaptive peaks. *Science* 365, 347–353. doi:10.1126/science.aax1837.

Zhou, X., and Wang, X. (2006). Optimisation of Gaussian mixture model for satellite image classification. *IEE Proceedings - Vision, Image and Signal Processing* 153, 349–356. doi:10.1049/ip-vis:20045126.

Figures

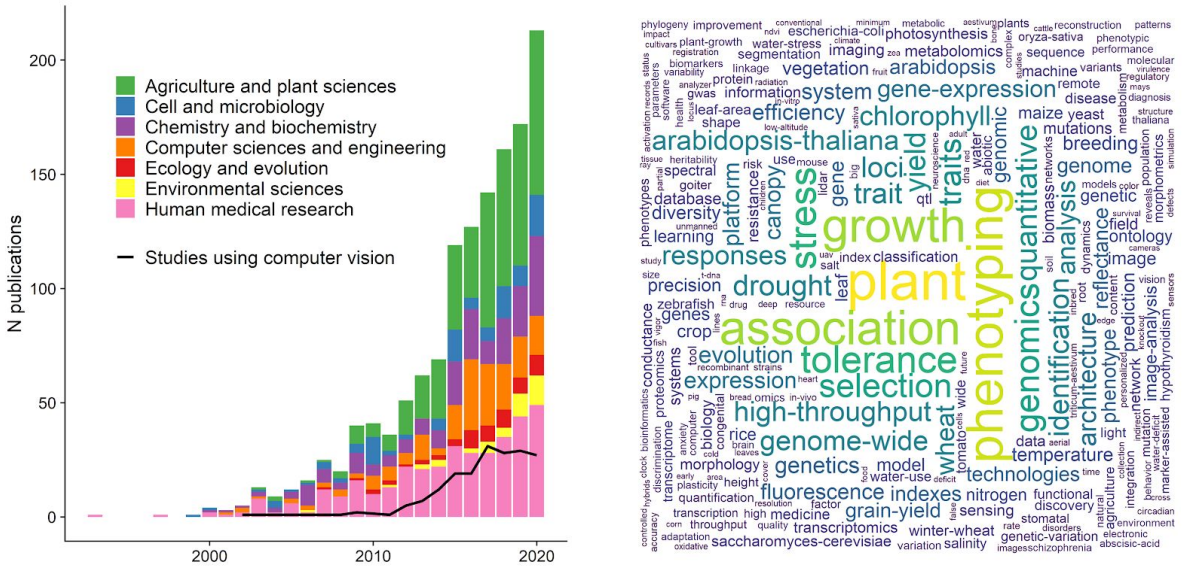

Figure 1 - The current state of phenomics research. Left panel: a literature survey using the search key “phenomic*" on in a Web of Science topic search (title, keywords, abstract) resulted in 1408 papers (on 01/03/2021). Here we show only papers published between 1990 and before 2021 (1343 papers) for better visual inference. Traditionally phenomics approaches are used in agricultural sciences and crop research to maximize yield, as well as in human medicine to study drug responses and disorder phenotypes. The black line denotes the studies that used computer vision or some sort of image analysis (acquired with a topic search among the retrieved papers using the strings "computer vision", "image analysis", "image based", "image processing"), indicating that only a small subset of the studies uses image analysis. Right panel: a word cloud that was constructed using the 500 most used keywords from the papers presented in the left panel.

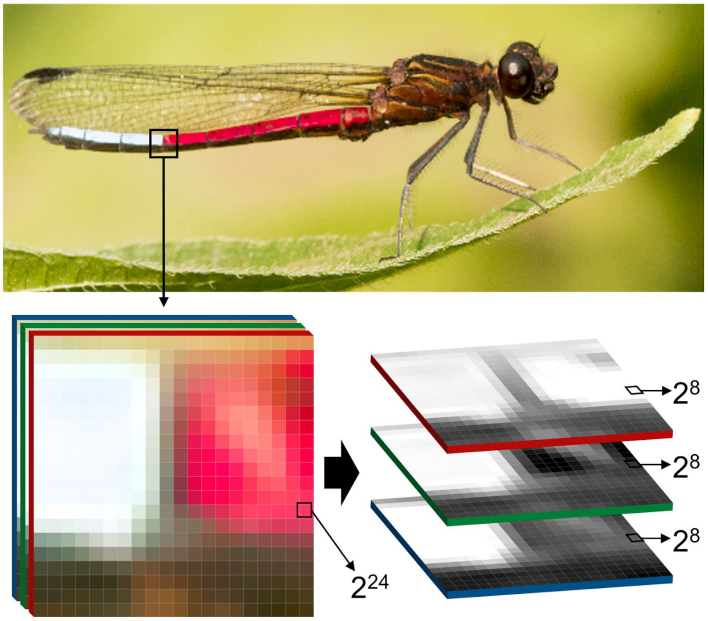

Figure 2 - The structure of digital images. Two-dimensional raster images, as produced by most commercially available cameras, are composed of three colour channels red, green, blue = RGB), each of which by itself is a grayscale image. The industrial standard for colour representation on the pixel level is 24 bit (2^{24} = 16 777 216 possible colour variations per pixel), which is achieved through additive mixing of each of the 8 bit channels ($2^8 + 2^8 + 2^8$). This enormous range of colour intensities among several million pixels is a potentially very high resolution representation of organismal traits, or the organism as a whole. Therefore, digital images are a useful medium for phenomics research, as they offer an inexpensive, memory efficient and standardizable way to capture, store and analyze complex phenotypes. The photograph shows a blue-tip jewel damselfly (*Chlorocypha curta*) in Cameroon (Africa) (image by Erik Svensson).

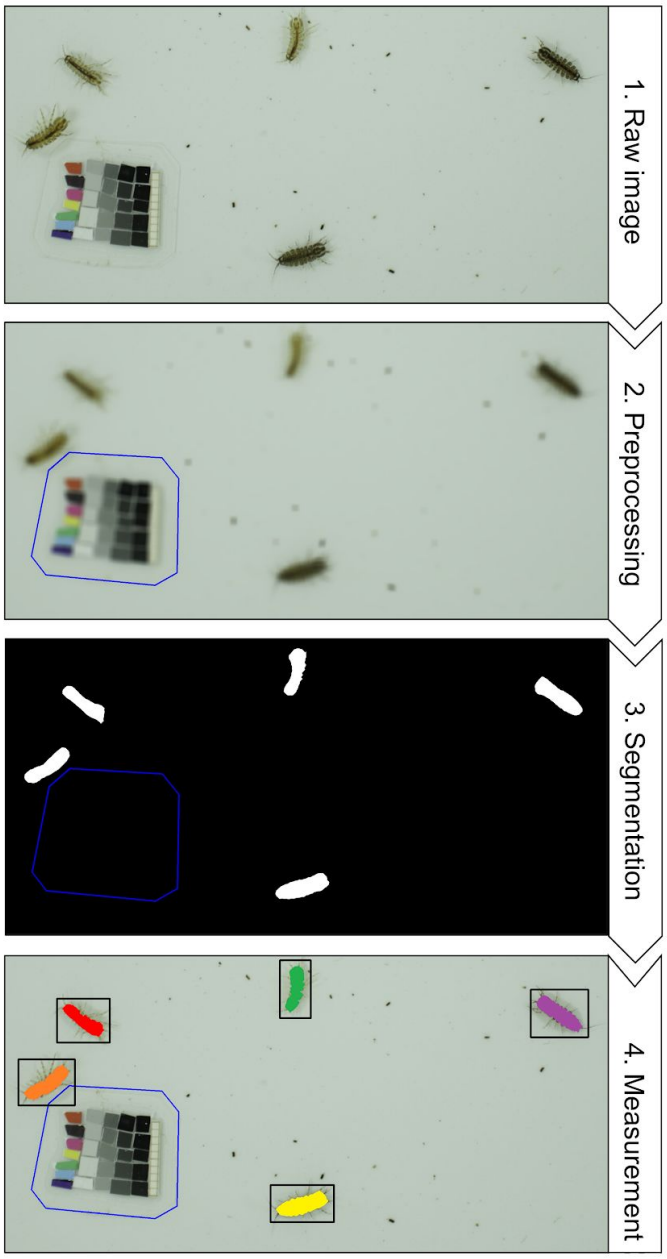

Figure 3 - A typical computer vision workflow using signal processing. 1) Raw image - The goal is to detect, count and measure freshwater isopods (*Asellus aquaticus*, image by Moritz Lürig) from the raw image that was taken under controlled laboratory conditions. 2) Preprocessing - The operating principle of most signal processing workflows is that the objects of interest are made to contrast strongly from all other pixels, meaning that images should have a high *signal-to-noise ratio* (SNR. In this specific case a high SNR is already present, because the isopods are much darker than the tray they are sitting on and much

larger than the fecal pellets and other detritus around them. To further increase the SNR, gaussian blur blends pixels in a given neighborhood (=*kernel* size), which effectively removes the smaller dark objects. The reference card gets excluded manually, and can be used to convert pixels to millimeters and to correct the colour space. 3) Segmentation - Using a thresholding algorithm all connected pixels that are above a specific grayscale value and larger than a specified area are designated foreground (white) and all pixels become background (black). The output from this step is referred to as a “binary mask”. 4) Measurement - Now the white pixels from the binary mask can be used to locate the areas of interest in the raw image and to extract information from them. Discrimination between multiple instances of the same class is referred to as instance segmentation.

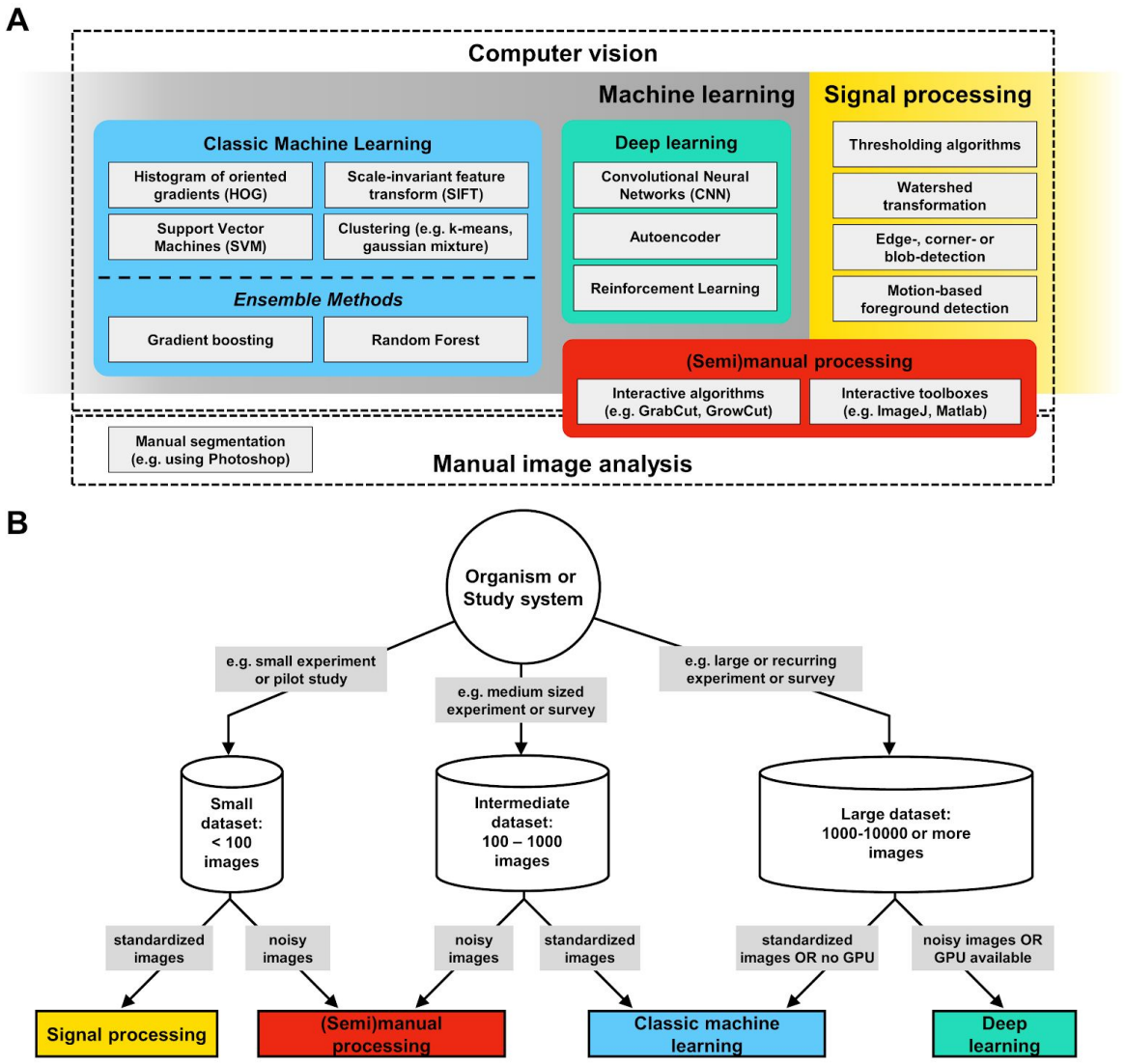

Figure 4 - Computer vision (CV) methods overview - which is the right one for my data? A) CV is a field at the intersection of machine learning and signal processing which is concerned with the automatic and semiautomatic extraction of information from digital images. B) Decision tree for CV methods: begin by considering the size of a given image dataset, whether it is complete, e.g. after an experiment, or whether there will be continued future additions, e.g. as part of a long term experiment or field survey. The next assessment should be about the signal-to-noise ratio (SNR) in your images: images taken in the laboratory typically have a high degree of standardization and thus a very high SNR, which makes them suitable for a signal processing approach. In contrast to signal processing,

where segmentation results are immediately available, all machine learning image analysis pipelines include iterative *training* and validation phases, followed by a final testing phase. Such a workflow can be complex to initiate, but pays off in the long run by providing segmentation results that become increasingly robust if more training data is supplied over time. Deep learning algorithms require large training datasets (several 1000s to 10000s) and a powerful graphics processing unit (GPU), but are less sensitive to noise and idiosyncrasies of the foreground. Thus, for large and continuously growing data sets, or for recurring image analysis tasks, deep learning has become the standard approach for segmentation.

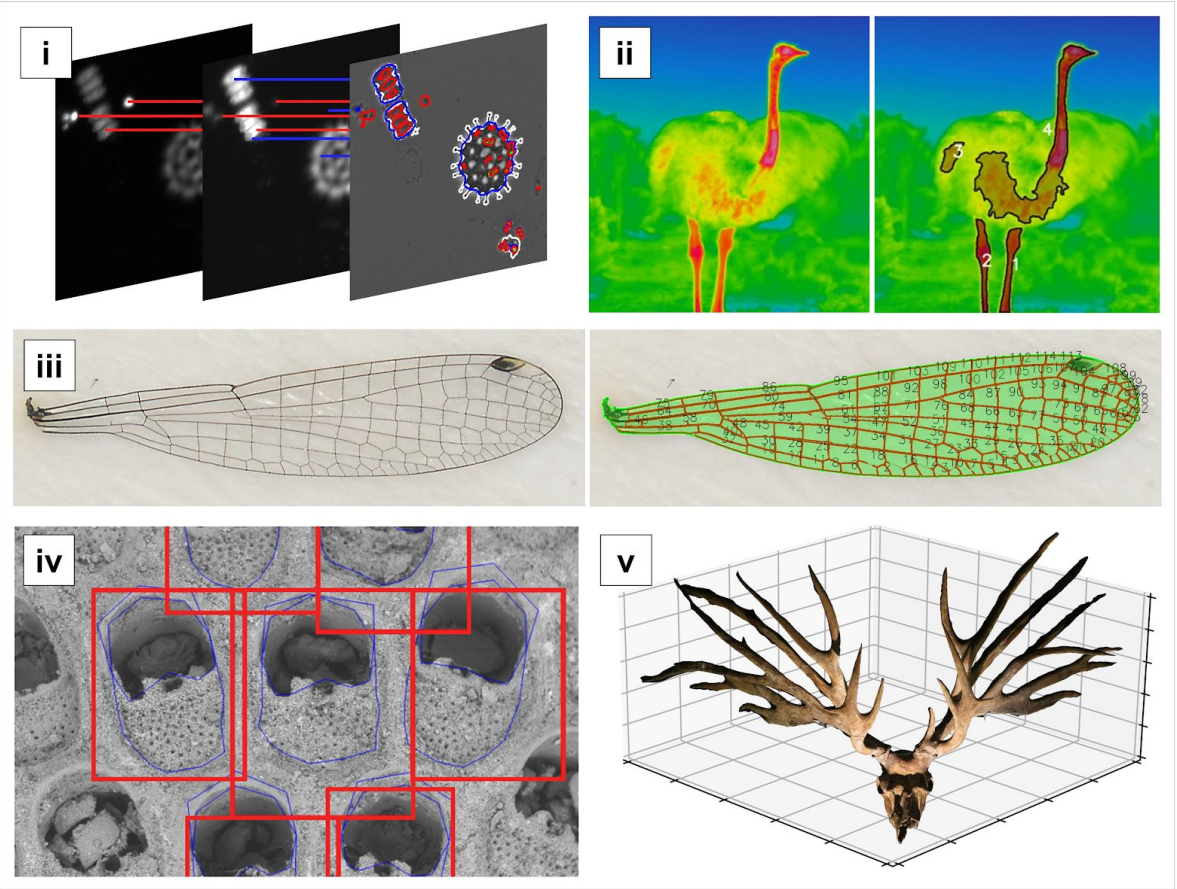

Figure 5 - Different types of high dimensional phenotypic data that are collected using a fully or semi-automatic computer vision approach. A) Morphology and fluorescence traits of phytoplankton communities are represented through a combination of shape features (e.g. circularity, perimeter length, area) and texture features (e.g. blob intensity and distribution within the cell) from images showing fluorescence intensity (images by Irene Gallego and Moritz Lürig). B) In ostriches (*Struthio camelus*), surface temperatures of bare body parts without feathers (necks and legs) are detected using signal processing (image by Erik Svensson). C) Signal processing approach that captures individual domains of a damselfly wing via thresholding (image by Masahito Tsuboi). D) Ensemble-based approach to shape prediction of individual zooids within a bryozoan colony (image by Arthur Porto) E) 3D image of the skull of extinct deer *Eucladoceros dicranios* from which we can measure linear, area, and volumetric measurements of antler features (image by Masahito Tsuboi).

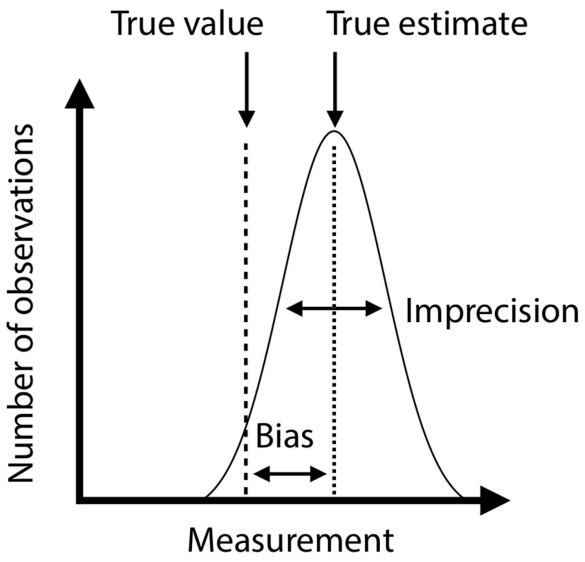

Figure 6 - Schematic illustration of bias and imprecision. X-axis represents phenotypic values and Y-axis represents number of observations. The gaussian curve shows the distribution of repeated measurements of the same specimen. Dashed line is the true estimate, and the variance of measurements around the true estimate is the imprecision. The true value may deviate systematically from the true estimate (long-dashed line). The difference between true estimate and true value is the bias.

Tables

Table 1 - Classes of phenotypic data. Depending on the research question, scientists define their phenotypes of interest using specific or abstract, low or high dimensional traits (see section “On measurement theory”). The human eye excels at rapidly recognizing externally visible phenotypes (e.g. benthic vs. limnetic morphotypes of fish), but has difficulties discerning what constitutes such phenotypes. Computer vision offers an objective way to collect any data type with high efficiency and reproducibility. For instance, by breaking down low dimensional traits (e.g. red vs. blue phenotype) into continuous low or high dimensional metrics (e.g. degree of red- or blueness), the decision of what constitutes a phenotype becomes more reproducible.

Trait type	**Low dimensional**	**High dimensional**
Specific / directly measurable	Size, discrete colour (“brown phenotype” vs. “blue phenotype”) and morphotype scoring (e.g. benthic vs limnetic)	Shape coordinates, texture maps, landmarks
Abstract / derived	Shape (e.g. circularity, area) and texture features (e.g. mean, SD, uniformity) , moments, principal components, hypervolumes	Matrices, activation maps

Table 2 - Select examples of recent open source computer vision libraries with a biology-context. Although typically first developed for a particular study system or organism (e.g. PlantCV or WorMachine), most CV applications apply techniques that are generally applicable to any type of phenotypic data contained in digital images.

Name	Year	Reference	Repository	Purpose	Application type	Description	Techniques
AutoMorph	2018	(Hsiang et al., 2018)	https://github.com/HullLab/AutoMorph	object detection and feature extraction	Python package	High throughput segmentation	Signal processing
ClickPoints	2017	(Gerum et al., 2017)	https://github.com/fabrylab/clickpoints	labelling, label evaluation	Python package	Interactive labelling tool	Signal processing
DeepMerkat	2018	(Weinstein 2015)	https://github.com/bw4sz/DeepMeerkat	object detection, classification	Python	Background subtraction and image classification for stationary cameras in ecological videos	Signal processing, deep learning
EB-Net	2020	(Le et al., 2020)	https://github.com/linhlevandlu/CNN_Beetles_Landmarks	keypoint and feature detection	Python	Insect morphometrics	Deep learning
ImageJ	2012	(Schindelin et al., 2012)	https://fiji.sc/; https://imagej.nih.gov/ij/download.html	multi purpose	standalone	Comprehensive, multi-purpose image processing library	Manual processing, signal processing, classic machine learning, feature extraction
ML-morph	2020	(Porto and Voje, 2020a)	https://github.com/agporto/ml-morph	landmark detection; geometric morphometrics	Python package	High throughput morphometrics	Classic machine learning, ensemble Methods

MotionMeerkat	2015	(Weinstein, 2015)	https://github.com/bw4sz/DeepMeerkat	motion tracking	Python package/ standalone	Deep learning driven motion detection	Signal processing, deep learning
Phenopype	2020	(Lürig, 2018)	https://github.com/mluerig/phenopype	object detection, feature extraction, motion tracking	Python package	Computer vision library with high throughput workflows	Signal processing
PlantCV	2017	(Gehan et al., 2017)	https://github.com/danforthcenter/plantcv	object detection and feature extraction; spectral analysis	Python package	Plant phenotyping library	Signal processing, classic machine learning
Scan-o-matic	2016	(Zackrisson et al., 2016)	https://github.com/Scan-o-Matic/scanomatic	object detection and feature extraction	Python package	Microbial phenotyping platform	Signal processing
Trackdem	2017	(Bruijning et al., 2018)	https://github.com/marjoleinbruijning/trackdem	motion tracking and blob counting	R package	Behavioral analysis pipeline	Signal processing
WingMachine	2003	(Houle et al., 2003)	https://www.bio.fsu.edu/~dhoule/Software/	keypoint and feature detection	standalone	Drosophila wing morphometrics	Signal processing, feature extraction
WorMachine	2018	(Hakim et al., 2018)	https://github.com/adamhak/WorMachineClient	object detection and feature extraction	Matlab	Integrated image processing and feature extraction	Signal processing, classic machine learning; deep learning

Boxes

Box 1 - Glossary of terms relevant for computer vision and machine learning in ecology and evolution used in this review. Terms in this list are printed in *italic* when first mentioned in the main text.

bit depth	number of values a pixel can take (e.g. 8 bit = 2^8 = 256 values)
computer vision	technical domain at the intersection of signal processing, machine learning, robotics and other scientific areas that is concerned with the automated extraction of information from digital images and videos.
convolution	mathematical operation by which information contained in images are abstracted. Each convolutional layer produces a feature map, which is passed on to the next layer.
deep learning	machine learning methods based on neural networks. supervised learning = algorithm learns input features from input-output pairs (e.g. labelled images). unsupervised = algorithm looks for undetected patterns (e.g. images without labelling)
feature	a measurable property or pattern. can be specific (e.g. edges, corners, points) or abstract (e.g. convolution via kernels), and combined to vectors and matrices (feature maps)
feature detection	methods for making pixel-level or pixel-neighborhood decisions on whether parts of an image are a feature or not
foreground	all pixels of interest in a given image, whereas the background constitutes all other pixels. the central step in computer vision is the segmentation of all pixels into foreground and background
hidden layer	a connected processing step in neural networks during which information is received, processed (e.g. convolved), and passed on to the next layer
kernel	a small mask or matrix to perform operations on images, for example, blurring, sharpening or edge detection. the kernel operation is performed pixel wise, sliding across the entire image.
labelling	typically manual markup of areas of interest in an image by drawing bounding boxes or polygons around the contour. can be multiple objects and multiple classes of objects per image. can also refer to assigning whole images to a class (e.g. relevant for species identification)
machine learning	subset of artificial intelligence: the study and implementation of computer algorithms that improve automatically through experience. (Mitchell 1997)
measurement theory	a conceptual framework that concerns the relationship between measurements and nature so that inferences from measurements reflect the underlying reality intended to be represented (Houle et al. 2011).
neural network	deep learning algorithms that use multi layered ("deep") abstractions of information to extract higher level features from input via convolution

object detection	methods for determining whether a pixel region constitutes an object that belongs to the foreground or not, based on its features
pixel	short for picture element; the smallest accessible unit of a digital raster image. Pixels have finite values (=intensities), e.g. 256 in an 8-bit grayscale image.
segmentation	the classification of all pixels in an image into foreground and background, either manually by labelling the area of interest, or automatically, by means of signal processing or machine learning algorithms. semantic segmentation = all pixels of a class, instance segmentation = all instances of a class
signal processing	technically correct: digital image processing (not to be confused with image analysis or image editing). subfield of engineering that is concerned with the filtering or modification of digital images by means of algorithms and filter matrices (kernels),
signal-to-noise ratio (SNR)	describes the level of the pixels containing the desired signal (i.e. the phenotypic information) to all other pixels. Lab images typically have a high SNR, field images a low SNR.
threshold algorithm	pixel-intensity based segmentation of images, e.g. based on individual pixel intensity (binary thresholding) or their intensity with respect to their neighborhood (adaptive thresholding). creates a binary mask which contains only black or white bixels
training data	representative image dataset to train a machine learning algorithm. can be created manually by labelling images, or semi-automatic by using signal processing for segmentation. can contain single or multiple classes
watershed algorithm	the segmentation of images by treating the pixels as a topographic map of basins, where bright pixels have high elevation and dark pixels have low elevation.

Box 2 - An overview of the main deep learning architectures and approaches.

Families of network topologies

A. **Deep convolutional network** - A large and common family of neural networks composed an input layer, an output layer and multiple hidden layers. These networks feature convolution kernels that process input data and pooling layers that simplify the information processed through the convolutional kernels. For certain tasks, the input can be a window of the image, rather than the entire image.
B. **Deconvolutional Network** - A smaller family of neural networks that perform the reverse process when compared to convolutional networks. It starts with the processed data (ie., the output of the convolutional network) and it aims to separate what has been convoluted. Essentially, it constructs upwards from processed data (e.g., reconstructs an image from a label).
C. **Generative Adversarial Network** - A large family of networks composed of two separate networks, a generator and a discriminator. The generator is trained to generate realistic data, while the discriminator is trained to differentiate between generated data from actual samples. Essentially, in this approach, the objective is for the generator to generate such realistic data that the discriminator cannot tell it apart from samples.
D. **Autoencoders** - A family of networks is trained in an unsupervised manner. The autoencoder aims to learn how to robustly represent the original dataset, oftentimes in smaller dimensions, even in the presence of noise. Autoencoders are composed of multiple layers, and it can be divided into two main parts: the encoder and the decoder. The encoder maps the input into the representation and the decoder uses the representation to reconstruct the original input.
E. **Deep Belief Network** - A family of generative networks that are composed of multiple layers of hidden units, in which there can be connections between layers but not within units within layers. Deep belief networks can be conceived as being composed of multiple simpler networks, where each subnetwork's hidden layer acts as a visible layer to another subnetwork.

Learning Classes

A. **Supervised Learning** - Training data is provided when fitting the model. The training dataset is composed of inputs and expected outputs. Models are tested by making predictions based on inputs and comparing them with expected outputs.
B. **Unsupervised Learning** - No training data is provided to the model. Unsupervised learning relies exclusively on inputs. Models trained using unsupervised learning are used to describe or extract relationships in image data, such as clustering or dimensionality reduction.
C. **Reinforcement Learning** - The learning process occurs in a supervised manner, but not through the use of static training datasets. Rather, in reinforcement learning, the model is directed towards a goal, with a limited set of actions it may perform, and model improvement is obtained through feedback. The learning itself occurs

exclusively through feedback obtained based on past action. This feedback can be quite noisy and delayed.

D. **Hybrid Learning Problems**

Semi-Supervised Learning - Semi supervised learning relies on training datasets where only a small percentage of the training dataset is labeled, with the remaining images having no label. It is a hybrid in between supervised and unsupervised learning, since the model has to make effective use of unlabeled data while relying only partially on labeled ones.

Self-Supervised Learning - Self supervised learning uses a combination of unsupervised and supervised learning. In this approach, supervised learning is used to solve a pretext task for which training data is available (or can be artificially provided), and whose representation can be used to solve an unsupervised learning problem. Generative adversarial networks rely on this technique to learn how to artificially generate image data.

Other learning Techniques

A. **Active Learning** - During active learning, the model can query the user during the learning process to require labels for new data points. It requires human interaction and it aims to being more efficient about what training data is used by the model

B. **Online Learning** - Online learning techniques are often used in situations where observations are streamed through time and in which the probability distribution of the data might drift over time. In this technique, the model is updated as more data becomes available, allowing the model itself to change through time.

C. **Transfer Learning** - Transfer learning is a useful technique when training a model for a task that is related to another task for which a robust model is already available. Essentially, it treats the already robust model as a starting point from which to train a new model. It greatly diminishes the training data needs of supervised models and it is, therefore, used when the available training data is limited.

D. **Ensemble Learning** - As mentioned in the main text, ensemble learning refers to a learning technique in which multiple models are trained either in parallel or sequentially and the final prediction is the result of the combination of the predictions generated by each component.