The Art of Categorization

A Reader

With contributions by Maroua Gaddur, Gilles Goosen, Orkan Kaan Pişkin, Danae Lankhorst, Seungju Lee and Lisa Ziebermayr

I have, in a way, been working with data as part of my artistic practice for my entire career. I have always tried to collect, to sort, to organise. But when I started working with machine learning in 2016, it became a central part of my practice. Data became an important tool and medium that I use.

A dataset is fundamental to making machine learning or artificial intelligence work. What is contained in the dataset becomes the knowledge that an algorithm has in order to create its world. The datasets need to be large, so that the algorithm has enough information to make inferences. They also need to be cleaned and standardised so that they become useable. Although datasets are primarily designed only to be read by machines—as seen in TF records or a NumPy array[1] where they appear as incomprehensible numbers to humans—datasets are also designed and made in the world, by people. Because of this, they play an important role in my artistic practice.

HITL (Human-in-the-loop) Methods of operating generative AI include 'Human-out-the-loop (HOTL)' and 'Human-in-the-loop (HITL).' Unlike HOTL, which focuses on allowing machines to do all learning and decision-making, HITL means human intervention to allow computers to make the right decisions when there is not yet much data available for machine learning. In this process, humans choose the most necessary and important data to improve their models, allowing them to improve machine learning beyond a random sampling process. This can create a continuous feedback loop and produce better results as the algorithm learns each time. Different types of datasets are required depending on the algorithm, where the HITL approach can be utilised for different types of data labelling processes. In terms of creation/criticism using generative AI, it leaves room for artists to intervene in the occurrence of intentional labelling errors in order to derive artistic messages or achieve objectives in the process of artistic work. Through this, it is possible to examine how artistic judgement and decision-making are made in the relationship between generative AI and artists, and how there are differences in the results. Seungju Lee

I do not think of datasets as something cold, sterile or algorithmic but as something incredibly human—each piece of data is a trace of someone's life, something that has existed. Part of my role as an artist is reconstructing what these traces might mean and understanding them in different ways. Early image-generation techniques needed a pre-constructed or self-constructed dataset. This meant that for each project, I had to construct my own dataset by hand, either by creating the content myself or by carefully selecting and editing it. Doing so gave me a very direct, intimate connection to it. Constructing my own datasets inverts the way in which large datasets are usually built: by mechanical turks and with imagery that has been scraped from the internet.

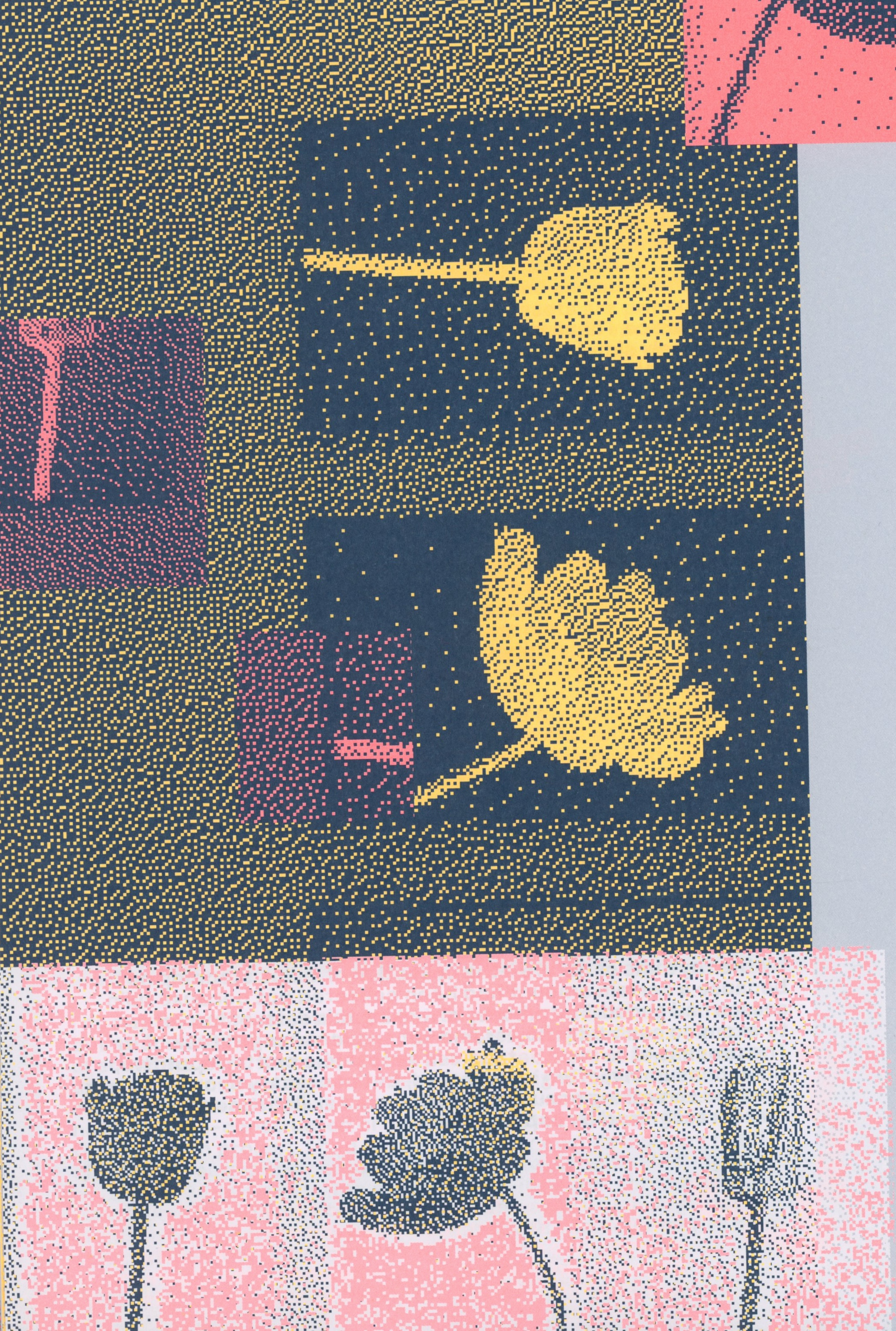
Crowdsourcing The practice of outsourcing small, discrete tasks, often referred to as Human Intelligence Tasks (HITs), to a distributed and flexible online workforce in exchange for payment. Crowdsourcing involves a large group of participants contributing and is not limited to online activity. Crowdsourcing is cost efficient, fast, improves quality and flexibility among other things. Crowdsourcing can be used in AI to label, clean, and test data, creating the illusion that AI is primarily independent from humanity. *Gilles Goosen*

In my practice, I have worked extensively with flowers. Something as seemingly simple as a flower actually contains many layers and narratives. Although flowers are often considered kitsch or feminine within the context of art, each flower has hidden histories of power, religion, science, capitalism, and colonialism. When working with flowers in the context of machine learning, particularly datasets, there are obvious parallels with herbariums: both make decisions as to what is important enough to be recorded; both are systems of organising knowledge; and both strip away the original context in which the object, or photograph, in question was found. A flower that features in many of my works is the tulip, which is a deliberate choice. It allowed me to make connections between speculation and value through tulip mania, one of the first speculation bubbles that took place in the seventeenth century. The parallels were also visual: in art history, tulips feature prominently in Dutch still lifes (fig. 1a, 1b), so-called 'vanitas' paintings that illustrated that beauty and treasure are only fleeting.

Based on information in the thousands of images of tulips that were in the dataset, the GAN[2] I worked with constructed an image not of a real tulip, but what it thinks a tulip should be. This is similar to how seventeenth-century still life paintings were created: they do not depict actual displays of flowers but imagined bouquets, based on the knowledge, drawings and sketches of the painter. It would be impossible for some of the bouquets to exist in reality, as the flowers shown bloom at different times of the year.

To create the dataset, I took ten thousand, or a myriad, of photographs of tulips over the course of three months whilst on residency in the Netherlands in 2018 (figs. 2a, 2b, 2c). This number was driven by the rhythms of nature—the collection of tulips stopping not because a certain amount had been collected but because tulip season had ended. The process of making this dataset forced me to examine each tulip flower and subsequent image I took of it. The process of making datasets is almost like craft—repetitive, time-consuming, often unauthored, but necessary in order to produce something beautiful. And there is a skill to it, something that is recognised by copyright law.

Fauxtomation Coined by writer and artist Astra Taylor, the term combines the word 'faux,' meaning fake or false, with 'automation' to express the phenomenon of how work accomplished through genuine human effort is falsely perceived as automated. It is



highlighting a general misconception surrounding a task. This could be a result of big tech companies seamlessly executing these tasks. The consumer assumes basic tasks are taken over by AI, leaving them with no shame when using applications that cannot operate without the human labour big tech companies hide. Gilles Goosen

Therefore, each photograph was carefully selected, as part of an iterative process, to produce the type of result that I desired.[3] This process is in direct opposition to a dataset like ImageNet (a canonical dataset often used in computer vision), where there are only around 1,000 tulips out of its fourteen million images, all of which have been anonymously found, categorised and labelled. In contrast, I had a direct connection to the objects I was documenting. The process was very physical—buying, moving, stripping hundreds and hundreds of flowers. It is easy to forget in the digital age that information is physical and that things that are seen on a screen once started out in the real world and those things have a weight and a mass.

The other important part of creating this dataset was the act of categorising the photographs. Computability is often accomplished by categorisation. In order for a particular image to be retrieved or understood by the model, it needs to have a label. I originally created this dataset as a 'working object'[4] for a very specific purpose: to create a GAN-generated moving image piece of single tulip stems, where the shape and appearance of the tulip would be controlled by the price of bitcoin (figs. 3, 4, 5a).

Labelling Labelling is a method to classify and describe something in short sentences and words. In psychology, 'labelling theory' posits that the identity and behaviour of an individual can be influenced by classification. By labelling, a subject is put under a narrow worldview and is subjected to negative labelling when existing outside of the perceived norm. When a person is labelled as a 'criminal,' their behaviour, appearance and actions become 'criminal' by default. With AI datasets using labelling to sort and train the systems, they are inherently biased towards a certain norm. Can datasets be developed without labelling or can we use these biased systems to gain insight to our own biases and adapt accordingly? Lisa Ziebermayr

In order for this to happen, I needed to provide labels so that I could effectively 'map out' the latent space and control the path that was taken, linking it to the price of the cryptocurrencies. Much of my time was spent creating these entirely subjective labels and categories (fig. 5b)—the colour of a tulip, for example, the state of the flower, the stripiness of the petal. Even something as simple as a tulip is difficult to put into discrete categories: is it white or pale pink? Is it orange or yellow? Is it a bud or has it just started to bloom? Because machine learning is so heavily dependent on language and categoriion, there is always a human decision somewhere along the chain using it. These decisions are not absolute correct things. What erceive as an orange flower could very well be seen as yellow to meone else. Herein lies the issue concerning datasets: implicit s becomes virtually unavoidable. However, in using photographs t I have taken I am able to impose greater control over the eaning of the images and how they eventually are processed by e algorithm. The resulting errors or assumptions made are mine d mine alone.

t this dataset did not remain a 'working object.' It became an twork in and of itself. *Myriad (Tulips)* is an installation of the otographs that make up the dataset, each one with a handwritten pel underneath with some of the categories that were used (fig. 6). choosing to create a physical installation of the dataset, aspects out the data can be comprehended in a way that might not occur herwise. The installation of all of the photographs is over 50 uare metres, giving an overwhelming sense of the time, money d effort that goes into constructing a dataset. Even installing the ork is a laborious process, with each photograph carefully affixed e by one with magnets to a specially painted black wall to form a eemingly precise grid. Up close, however, slants and errors come to view, evoking the arduous human labour behind machine arning and also its imperfection.

Data The word data is the plural form of the Latin word datum '(thing) given.' The first English use of the word, in the 1640s, meant 'a fact given or granted.' It was first used to mean 'transmissible and storable computer information' in 1946. In essence, data are a collection of observations; the things given. Since the start of the digital age, data has become an entity unto itself more than a collection of observations. The arrival of computers has turned 'data' into the digital version of sand or rain; basically uncountable and easier to refer to as a singular mass. We no longer really seem to know the difference between data and information. Data is a representation. Like any representation, it is incomplete, imperfect, subject to interpretation. It cannot be understood separate from the frame in which it was collected. Everyone sees the world through a particular lens, so collected data is a part of that. But data is not seen as subjective; data are no longer the things given, no longer our observations of the world. Now, data is a rigid entity. Danae Lankhorst

Last year, I returned to work with the same dataset. But this time I used the latest technology in machine learning to create *The Black Tulip*, which is an update to the tulip series that continues to explore speculation and value. Echoing Dumas' exploration of rarity in the novel of the same name, true black tulips have historically been almost impossible, creating a speculative desire for one. Because black tulips do not exist, I had to create imagined ones. I finetuned a text-to-image diffusion model to make a dataset of these impossible black tulips, which then were used to train another GAN.

Text-to-image diffusion models require orders of magnitude more data to train: DALL-E was trained on 400 million pairs of image-captions; Midjourney on a 100 million. These types of models are being described as using the entire internet as its base. The AI-generated tulips in this new dataset are weirder, uncanny and more artificial (fig. 7) than those in *Mosaic Virus* (2019). The later echo the single stems and precise placements of my design, the former twist and double and shine and pull in ways that I never intended. Petals become leaves, and flowers stretch to fill the entire frame. Stems disappear. These are the traces of these unknown images that crop up in the results that it produces for me: black tulips in shapes and positions and configurations that were never in my original dataset but now have made it into this body of work.

I have also tried to explore these newer technologies in another, newer body of work called *Synthetic Iris Dataset*: a series of prints and handwritten charts based on 150 iris flowers generated using the same prompt across different text-to-image generators (fig. 8), of varying degrees of 'iris-ness' (figs. 9a, 9b).

Synthetic Media Synthetic media is an umbrella term for all content that is artificially generated, modified or manipulated by AI and algorithms. This includes deep fakes, speech synthesis, image generation, etc. Synthetic raises many ethical questions since it can spread misinformation, resulting in the loss of credibility or trust in certain sources (for example, fake news, fraud, propaganda) and can even harm communities/individuals (for example, revenge porn, fraud, theft). On the other hand, synthetic media can generate content that does not exist yet, which could be beneficial or even inspiring, for example, the use of synthetic voices. This synthetic medium can aid people with speech or hearing disabilities, translating text into speech with the possibility of choosing a voice that resonates most with the user, resulting in more independence of speech. Maroua Gaddur

In many ways, this project is an inverse of the dataset I created using tulips: rather than me trying to teach a machine what should be in the image based on my own photographs, this is trying to parse what a machine sees in an image generated by a machine. The original iris dataset was created in the 1930s by Ronald Fisher, abstracting 150 irises found in the same field in Canada into a series of numbers (fig. 10).

It is now widely used as a test case for many statistical classification techniques in machine learning and is included in the package scikit-learn.[5] Each time it is downloaded, so is the iris dataset. Thousands of little iris datasets exist now on thousands of computers.

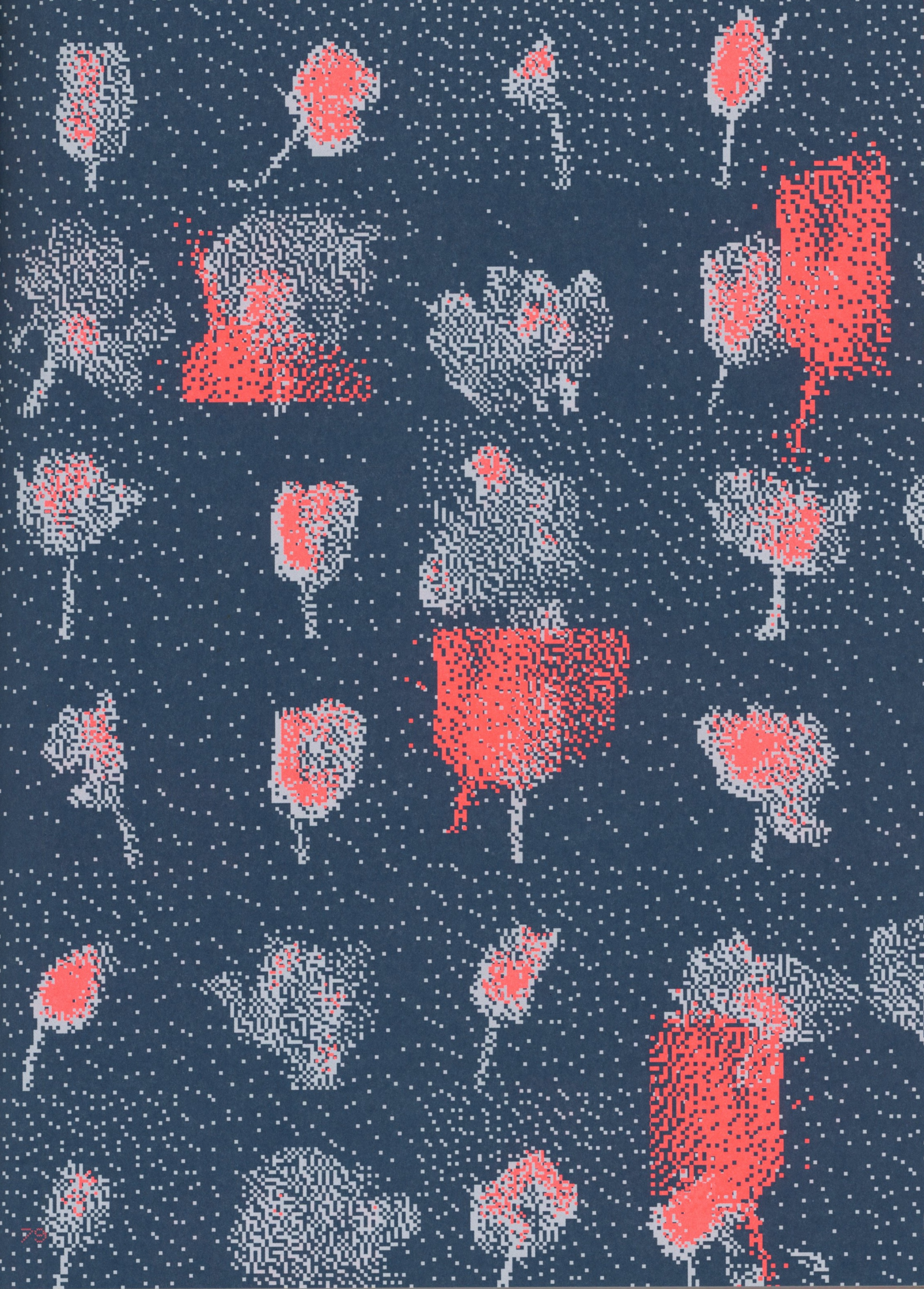
The work comes in several parts: the first part is a series of handwritten charts that describe the attributes of synthetic flowers as seen by a computer vision model and gives a score of how confident it is. Some of these appear reasonable—flower, petal—while others

Pattern Matching Pattern matching in the context of artificial intelligence refers to the process of identifying and recognising specific structures or sequences within data based on predefined templates, rules, or patterns. This concept is fundamental in many AI applications, including natural language processing, computer vision, machine learning, and expert systems. In Computer Vision, these patterns could inhabit trying to recognise faces in images, for example. The predefined templates/rules would then be extracting features like eyes, nose and mouth arrangements to identify a face. Pattern matching enables AI systems to make sense of complex and unstructured data by finding meaningful associations and regularities within it. Since some patterns or sequences may be ambiguous, it can be challenging to accurately interpret them. Misinterpretation of data can lead to incorrect conclusions or actions, which makes it difficult in building accurate models, especially when relationships between variables are not well understood. Is it then really ethically considerate to automate crucial decisions based on pre-designed algorithms?
Orkan Kaan Pişkin

are less so—sky, event. It is also clear that language has muddied things in the dataset at some point. Iris has two meanings in English: the flower and the coloured centre of the eye. Violet indicates both a pigment and the flower. Eyes and colours keep being seen by the API where there are none.

The second part of *Synthetic Iris Dataset* is a series of prints with the colours of the 150 synthetic flowers generated, as determined by a computer vision API (fig. 12). The problem of the communication of colour has always been difficult. Swedish botanist Carl Linnaeus openly opposed using colour in describing plants because of its inherent subjectivity. Something that I keep returning to explore is this idea of haziness around the definitions and perceptions of colour and the way that material form must be translated into language. Some of the colours that were found using image classification in part 1 (fig. 11) do not seem to appear here (fig. 12) ('magenta,' 'electric blue'). This is partially due to the model mis-seeing what is in the image, but it is further heightened by the translation of the RGB colours into CYMK, and the inherent difficulty of rendering the digital accurately in the real world.

I am still building new floral datasets for new projects—growing obscure versions plants in my garden that cannot be easily found on the internet, trying to capture every bloom in a particular season—and translating and exploring them using the latest technologies. The speed and accuracy of generating images and how they are commodified and used has proved more rapid and malleable than I could have foreseen even a few years ago. It seems certain to mutate into newer forms, and I will be there, unpicking it, in my own small way.



1a. Ambrosius Bosschaert the Elder, *Bloemstilleven*, 1614.

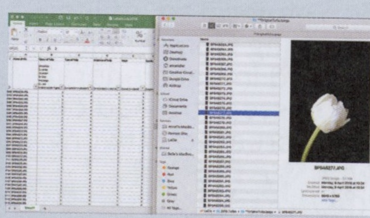
1b. Christoffel van den Berghe, *Still Life with Flowers in a Vase*, 1617.

2a. Documentation of the process of creating the dataset of tulips.

2c. Documentation of the process of creating the dataset of tulips.

2b. Documentation of the process of creating the dataset of tulips.

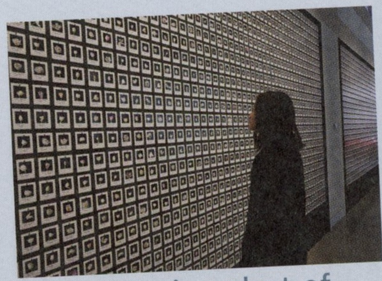
5a. Process of making and organising physical photographs.



4. Process of creating labels for dataset.

3. Anna Ridler, still from *Mosaic Virus*, 2018.

5b. Process of making and organising physical photographs.



6. Installation shot of *Myriad (Tulips)*.

7. Sample of the dataset of the black tulips.

8. Screenshot of the results of the prompt.

9a. The most 'iris-y' iris according to the computer vision API.

9b. The least 'iris-y' iris, according to me.

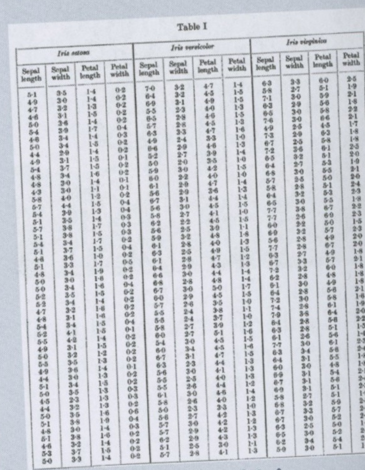
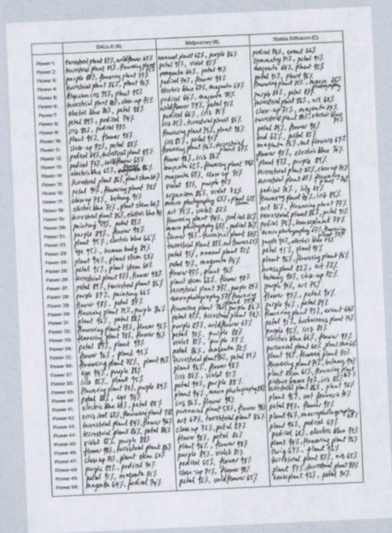


Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

10. The original Iris Dataset.



11. Synthetic Iris Dataset, part 1.

12. Synthetic Iris Dataset, part 2.