Machine Intelligence Garage

Machines for Machine Intelligence

Research Report 2018

CATAPULT Digital

CONTENTS

APPENDICES

Digital Catapult is a technology innovation centre that unlocks digital growth in the UK economy. It works with companies of all sizes to transform their businesses by accelerating the practical application of digital innovation. We bridge the gap between research and industry, finding the right technologies to solve problems, increase productivity and unearth new opportunities.

We are eager to collaborate with existing initiatives - public, private, open source - that share our mission to build solutions that help machine intelligence startups reach their potential. We want to hear from startups who want to access our programme or have ideas about other ways that we can assist.

Please contact us at
hello@migarage.ai
www.migarage.ai

FOREWORD

From Alan Turing to DeepMind, the UK boasts a rich history and exciting present in machine learning and AI research and development, by academia as well as industry.

Jérôme Pesenti and I were recently appointed to author the independent report for the UK government titled: 'Growing the Artificial Intelligence Industry in the UK'.[1] We believe that the UK is well situated to take a lead role in the AI revolution and have identified numerous opportunities, along with the challenges which need to be addressed.

The report highlights recommendations around access to training data, computation facilities and skills, as well as enhancement of academic research and finally support for the demand side of AI and machine learning technology.

We were surprised at first that access to computation power is amongst the most significant barriers to innovation around Machine Learning and AI. Even more so considering the ever decreasing costs and increasing availability of cloud computing. Nevertheless, the demand and cost for computation power for state-of-the-art machine learning models is rapidly increasing.

The 'Machines for Machine Intelligence' report provides original research and insight into this barrier. It is great to see Digital Catapult's Machine Intelligence Garage programme looking to help startups address and overcome the challenge , and we hope to see cross fertilisation with other organisations such as the Alan Turing Institute that will be looking into similar challenges for the academic community.

Dame Wendy Hall,
DBE, FRS, FREng

EXECUTIVE SUMMARY

UK GDP could be around 10% higher in 2030 as a result of Artificial Intelligence (AI)[2], making AI the largest commercial opportunity for the British economy. According to an Accenture report, AI could generate the equivalent of an additional £630 billion by 2035[3]. A great business opportunity, and one that is accentuated by the fact that a recently published, HMG industrial strategy[4] has emphasised the importance of machine learning and AI and highlights their potential to disrupt many traditional industries by automating workflows.

That said, data now supports the fact that there are a great many UK Machine Intelligence startup companies that will miss out on this opportunity, simply because of the lack of access to high quality training data and computation power and in some cases expertise around it.
So why is this happening?

Digital Catapult's 'Machines for Machine Intelligence' report reviews the current computational landscape for machine intelligence startups, identifies areas of need and possible areas of intervention; and looks ahead to possible needs of the future.

Based on original research carried out by Digital Catapult including a survey, desk research and interviews, the main result of the study indicates that more than half of the surveyed startups report being computation-constrained, across multiple technology approaches, stages and geographical locations. Computation-constrained means that progress is slowed because of the cost of, or lack of access to, computation power. While cost of cloud computing is decreasing, the computation demands of machine learning models are currently increasing much faster. Naturally, these costs start to accumulate before technology proof-of-concept, before startups are funded, or achieve revenue. Potentially they limit what startups can do, their ambitions, and competitiveness.

The report also demonstrates that startups would find it useful to be able to experiment with new hardware platforms and access relevant expertise and concludes with recommendations of ways that Digital Catapult can support the UK's machine intelligence startups. By collaborating with existing commercial and national centres of expertise, Digital Catapult aims to engage startups in a programme of activities that will directly overcome computational barriers to success:

Machines

- Access to equipment for R&D and proof-of-concept
- Access to cloud and high-performance computing partners for continuous development, scaling, and deployment
- Experimentation with new hardware and platforms

Intelligence

- A repository of information, discussion and advice on computation, system and software topics
- Benchmarking and prediction activities to evaluate hardware options for different workloads
- Practical workshops on topics such as distributed computing, cluster management and job scheduling, and heterogeneous computing

Our ambition is to provide a full-suite of tailored assistance that will promote UK machine intelligence startup success.

INTRODUCTION

All technology startups face challenges. However, Machine Intelligence startups may face additional challenges. These are normally around access to quantities of high quality training data, access to the computation power and systems expertise needed to train machine learning models on that data, and challenges around adoption of the technology, which requires cultural and business process changes.

We use 'Machine Intelligence' to cover a number of aspects of Artificial Intelligence including machine-, deep- and reinforcement-learning, computer vision, natural language processing, simulation, robotics, and symbolic AI.

Digital Catapult has written previously about talent and data network effects[6]. Access to specialist computation facilities and the expertise to harness them is becoming an increasingly important factor for success of Machine Intelligence companies and is the focus of this report.

> *"Developing deep learning models is a bit like being a software developer 40 years ago. You have to worry about the hardware and the hardware is changing quite quickly… Being at the forefront of deep learning also involves being at the forefront of what hardware can do."*
>
> ***Phil Blunsom, Oxford University and DeepMind*** [7]

DEEP LEARNING IS COMPUTATIONALLY EXPENSIVE

Some Machine Intelligence tasks require startling amounts of computation power. For example:

- DeepMind noted that AlphaZero used '5,000 first-generation Tensor Processing Units (TPUs) to generate self-play games and 64 second-generation TPUs to train the neural networks' to achieve 'superhuman level of play in the games of chess and shogi (Japanese chess) as well as Go'.[8] That's around 500 petaOP/s of compute, greater than the World's top ten supercomputers combined (see appendix A1.1)
- Baidu Research used a 11 petaFLOP/s GPU supercomputer of 1500 GPUs to study deep learning scaling[9], and they remark that, 'This experiment would cost over $2 million USD'[10] if performed on a cloud service

THE NEED FOR TECHNICAL EXPERTISE IS RISING

As the size of neural networks increases to tackle new challenges this gives rise to a computation expertise gap as these large models need significant system engineering and code optimisation to train and deploy effectively and efficiently.

COMPUTATION RESOURCE CHOICE IS EXPANDING

Computation choice for Machine Intelligence workloads is starting to become much more complex. In recent years, deep learning on GPUs has driven progress and engineers have used a simple rule-of-thumb: buy a GPU with as much memory as you can afford[11]. In 2018 a number of new choices will be available and there will be a consequent need for expertise to understand the engineering, performance, cost, and power implications for specific machine intelligence workloads.

MACHINE INTELLIGENCE TECHNIQUES ARE EVOLVING

Machine Intelligence workloads are evolving, and it's likely that the intelligence solutions of the future will look very different to the ones available today. New computational capabilities will enable new fields to flourish, and new research avenues will have different computational requirements.

IMPLICATIONS FOR STARTUPS

Startups do not typically have the funds to enjoy unfettered access to substantial compute resource and expertise, nor do they enjoy priority access to new hardware.

'Since many of the interesting machine learning papers now regularly require 100s or even 1000s of CPU/GPUs for replication - what strategies are realistically left for startups, public institutions and individuals to do meaningful research in ML?'[12]

This report reviews the current computational landscape for startups, identifying areas of need, and possible areas of intervention; and it looks ahead to what might be the needs of the future. It is based on original research carried out by Digital Catapult including a survey, desk research and interviews.

The main result of the survey is that more than half of the surveyed startups report being computation-constrained, across multiple technology approaches, startup stages and geographical locations. In addition, startups would find it useful to be able to experiment with new hardware platforms and access relevant expertise.

This report continues by supporting the case for computation by describing recent advances in machine intelligence and their computational 'cost' (section 2); next is a description of Digital Catapult's original research into startup computation needs (section 3) and ends with conclusions and future directions (section 4). Additional information on benchmarking and performance modelling and a hardware glossary are presented in the appendices.

LITERATURE REVIEW

This section describes recent advances in machine intelligence, with emphasis on their computational 'cost', to support the case for startups being computation constrained.

3.1A
A ZOOM IN ON DEEP LEARNING

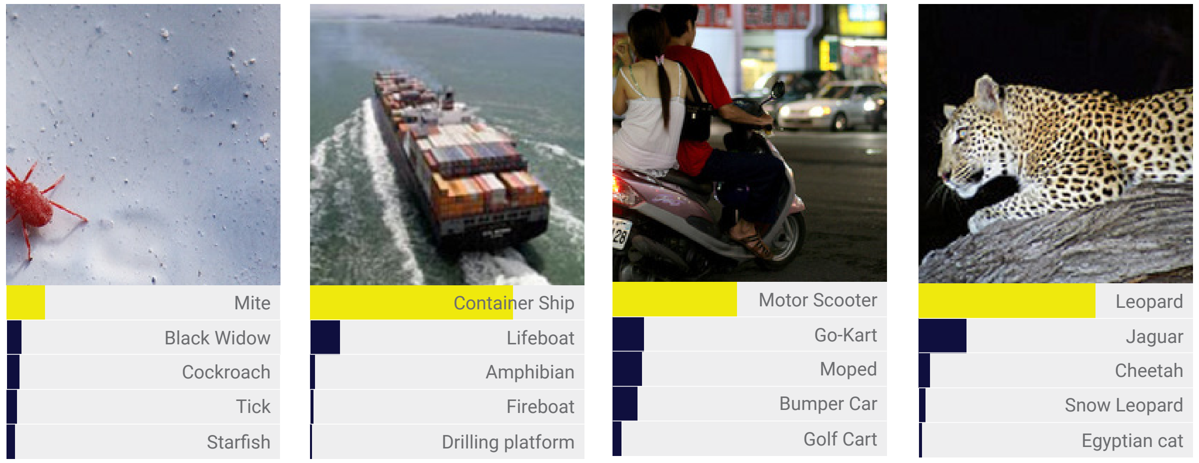
What is it about Machine Intelligence that leads to demanding computation requirements? To answer, we start by placing focus on deep learning since that has been behind the recent renaissance in AI (we will return to other techniques later). In particular, consider many-layered artificial neural networks, a.k.a. deep neural networks.

Engineering a deep neural network to solve a particular task requires four main computational elements:

- Carrying out data pre-processing and augmentation tasks
- Training the deep learning model
- Storing and adapting the trained model for efficient deployment
- Engineering the scale-out system for deploying the model

Of these, training is where peak computation is often encountered. Training a deep neural network involves a forward pass in which data is passed through the network, and a backward pass in which the network weights are updated. There can be millions of training examples (the more the better[13]), and millions of model parameters to update (bigger models have correlated with better performance[14]). Training a single model typically involves many passes (epochs) through the data. For an example see Section 3.1b.

Training a single model, once, may not break the bank. However, it's obviously not guaranteed that the model will converge to a solution, or that the solution is a good one. Researchers conduct multiple experiments to find a network architecture that works and then search for the best solution on it (hyperparameter optimisation). The process restarts every time there is a need to implement new ideas or train on newly available data. This approach of architecture engineering and tuning is experimental, iterative and continues to evolve.



▲

Figure 1
Example images and label predictions Figure taken from 'ImageNet Classification with Deep Convolutional Neural Networks' [17]

3.1B EXAMPLE: TRAINING A DEEP NEURAL NETWORK FOR THE IMAGENET LARGE SCALE VISUAL RECOGNITION CHALLENGE (ILSVRC)[15]

ILSVRC is an annual challenge that has attracted a great deal of research time and effort, and in which deep neural networks have had prodigious (now better-than-human[16]) success. The training data for the image recognition challenge consists of **1.28m images** (like the ones shown in Figure 1).

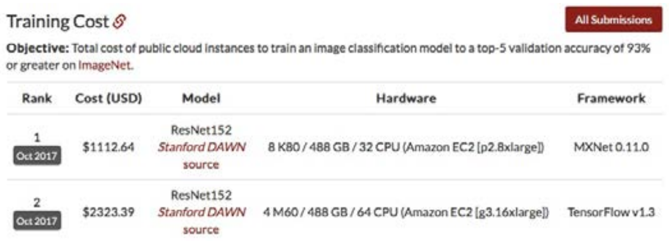
Figure 2 compares notable image classifiers trained on ILSVRC data, plotting algorithm performance against (a) the computational demand (on the left) of processing **a single image** (the vertical dotted line indicates **1 billion** multiply-add operations, and some models use more than **30 billion**) and (b) model size (on the right - note that models can have tens of millions of parameters).

This allows us to calculate that an approximate **minimum** computation requirement for **training one model, once**, for this competition is of the order **exaFLOP** (10^{18} floating-point operations - see appendix A1.2).

How this translates to cost depends on the hardware, software and pricing model, but for illustration, see figure 3 below, which captures the cost of training the ResNet-152[19] model using public cloud GPU instances. These are relatively modest amounts for one training run, perhaps. But a sweep of training hyper-parameters would require such a training run to be repeated many times to find the best model. It's easy to see how costs escalate with architecture engineering and tuning.

▼

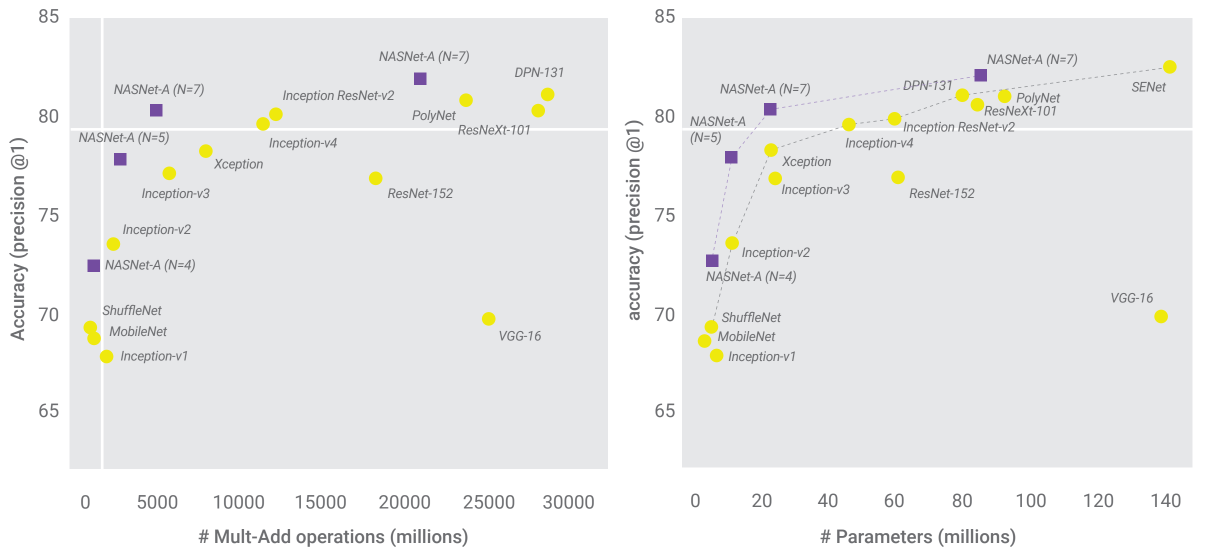
Figure 3
Screengrab from DAWNBench[20] showing a lowest cost of $1,112 to train to 93% accuracy



Training Cost

Objective: Total cost of public cloud instances to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet.

Rank	Cost (USD)	Model	Hardware	Framework
1 Oct 2017	$1112.64	ResNet152 *Stanford DAWN* source	8 K80 / 488 GB / 32 CPU (Amazon EC2 [p2.8xlarge])	MXNet 0.11.0
2 Oct 2017	$2323.39	ResNet152 *Stanford DAWN* source	4 M60 / 488 GB / 64 CPU (Amazon EC2 [g3.16xlarge])	TensorFlow v1.3



▲

Figure 2
Comparison of notable image classifiers trained on ILSVRC data. Reproduced from 'Learning Transferable Architectures for Scalable Image Recognition' [18]

In practice, huge amounts of computation continue to be thrown at this one problem. In 2012, it took 'between five and six days to train on two GTX 580 3GB GPUs'[21]. Now, although more than one research group has trained an ILSVRC model in an hour, this came with significant hardware budgets: 256 NVIDIA P100 GPUs[22], and 512 Intel Xeon Phi Processor 7250[23] respectively. It's hard to imagine readily renting this kind of resource, and to purchase it would need an investment of millions of dollars.

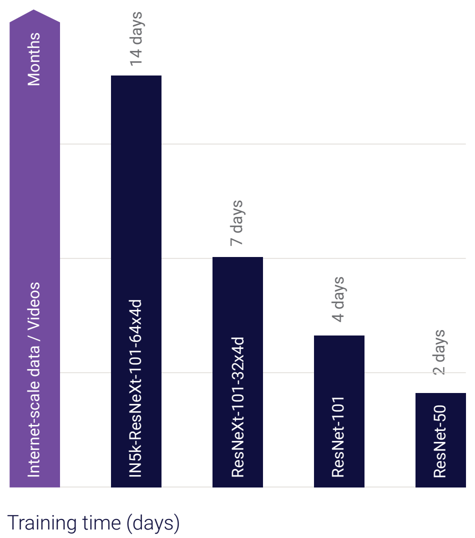
Moreover, the computation required to compete in the ILSVRC challenge may be dwarfed by real-world applications. Figure 4 notes that training on internet-scale data or videos is likely to take months. Innovating in other domains and with other types of data is no less time-consuming and expensive.

> *"Five years ago, it took more than a month to train a state-of-the-art image recognition model on the ImageNet dataset. Earlier this year, Facebook demonstrated that such a model could be trained in an hour. However, if we could parallelize this training problem across the world's fastest supercomputers (~100 PFlops), it would be possible to train the same model in under a minute."*

Quote from the website of the NIPS 2017 Workshop: Deep Learning At Supercomputer Scale [24]

▶

Figure 4
Illustration of slide from a talk given by Priya Goyal, Facebook Research, at the NIPS workshop on Deep Learning At Supercomputer Scale on 9th December 2017 [25]



3.2A
ARCHITECTURES FOR DEEP LEARNING APPLICATIONS

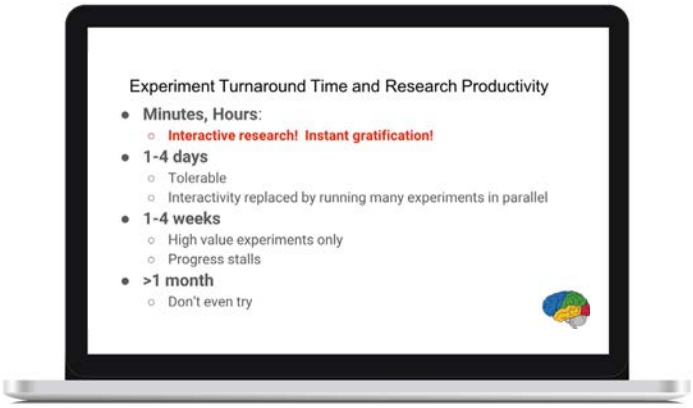
The computational requirements of cutting-edge deep learning research are growing. Applications like voice translation and speech recognition use models with hundreds of millions of parameters, requiring **tens or hundreds of exaFLOPs** to train.

Since the introduction of hidden units artificial neural networks have doubled in size roughly every 2.4 years. This growth is driven by faster computers with larger memory and by the availability of larger datasets. Larger networks are able to achieve higher accuracy on more complex tasks. This trend looks set to continue for decades.[26]

This comes with attendant engineering challenges. First, if a single model takes days or weeks to train, the pace of innovation is slowed or even halted (Figure 5). Hence significant effort is expended on reducing training time, by distributing training effort across compute nodes and optimising algorithms.

Secondly, specialist skills are required to optimise the resulting models and systems for efficient and cost-effective deployment. For example, the utility of language translation or speech recognition that happens after a delay is low, but if latency can be reduced to human-like response times, that's extremely useful. See section 3.2b for examples.

> *"Tremendous amounts of data are funnelled through our machine learning pipelines, and this creates engineering and efficiency challenges far beyond the compute nodes."*
> ***Facebook Research.***[28]



Experiment Turnaround Time and Research Productivity

- **Minutes, Hours:**
  - Interactive research! Instant gratification!
- **1-4 days**
  - Tolerable
  - Interactivity replaced by running many experiments in parallel
- **1-4 weeks**
  - High value experiments only
  - Progress stalls
- **>1 month**
  - Don't even try

▲
Figure 5
Slide from 'Building Intelligent Systems with Large Scale Deep Learning' talk given by Jeff Dean, Google Brain, to the YC AI group in summer 2017 [27]

These engineering hurdles present particular challenges for startups, who have neither unlimited compute resource nor the expertise to build the infrastructure for Machine Intelligence research and deployment pipelines.

> *"You need to build systems to run very large, demanding jobs at scale, and to do this in an easy-to-use way so your researchers can conduct as many experiments as they desire. These parts are not commoditized and when you move into AI systems that require larger and larger models, the expertise required to make the infrastructure grows. It doesn't diminish."*
> ***Jack Clark, OpenAI*** [29]

Naturally, these costs (in resource and time) start to accumulate before technology proof-of-concept, before startups are funded, or achieve revenue. Potentially they limit what startups can do, their ambitions, and competitiveness.

3.2B
EXAMPLES OF DEPLOYED DEEP LEARNING APPLICATIONS

Speech Recognition

Baidu's 'Deep Speech 2' automatic speech recognition results were exciting: it was able to beat human transcription performance in some cases, and to do so equally well for English and Mandarin Chinese speech (including accents and in noisy environments). However, training a single model required tens of exaFLOPs which would take 3-6 weeks to execute on a single GPU .

By deploying optimisation techniques typically found in High Performance Computing (HPC) and distributing the training task over multiple GPUs, Baidu cut training down to 3-5 days, enabling it to iterate more quickly. Substantial effort then went in to optimising the trained model so that Baidu could deploy Deep Speech at low cost, low latency, and high throughput. In other words, to meet the requirements of serving interactive applications at scale.[30]

Speech Synthesis

Google DeepMind first revealed its WaveNet speech synthesis research in 2016. At that time, it was far too slow to be used in real-world applications. One year of research advances and engineering effort later, a new improved version of WaveNet was announced[31]. This new version is **1000x** faster than the original, taking **50 milliseconds** to create one second of speech – that's fast enough for consumer products. WaveNet is now deployed in Google Assistant to generate much more natural US English and Japanese voices.

Even so, this is only possible because the WaveNet voice interface is running on Google's proprietary Tensor Processing Unit (TPU). Breakthroughs like WaveNet require such enormous amounts of computation – for both training and deployment – that Google designed its own specialist hardware (no mean feat itself).[32]

Machine Translation

Once the stuff of sci-fi wish-fulfilment [33], (near-)instant language translation is now reality. For example, Google switched last year from statistical machine translation to Neural Machine Translation (NMT) to take advantage of the better accuracy that deep learning makes possible. This switch needed deep learning, yes, but just as important was the significant engineering investment needed to work out how to train a deep neural network on the very large data sets that are necessary, and how to build a system fast and accurate enough for real-world use [34]. The paper that Google wrote to describe how they overcame these challenges highlights how computationally expensive NMT models are to train and deploy. Training a basic NMT model takes **6 days using 96 GPUs**. At inference, 'low latency translation [is] difficult, and high volume deployment computationally expensive.'[35]

A subsequent analysis of NMT architecture parameters undertaken by Google required, 'over **250,000 GPU hours** on the standard WMT English to German translation task.'[36]

Like WaveNet, interactive production deployment of the translation service involves Google's TPU specialist hardware.[37]

3.3 HARDWARE CONSIDERATIONS

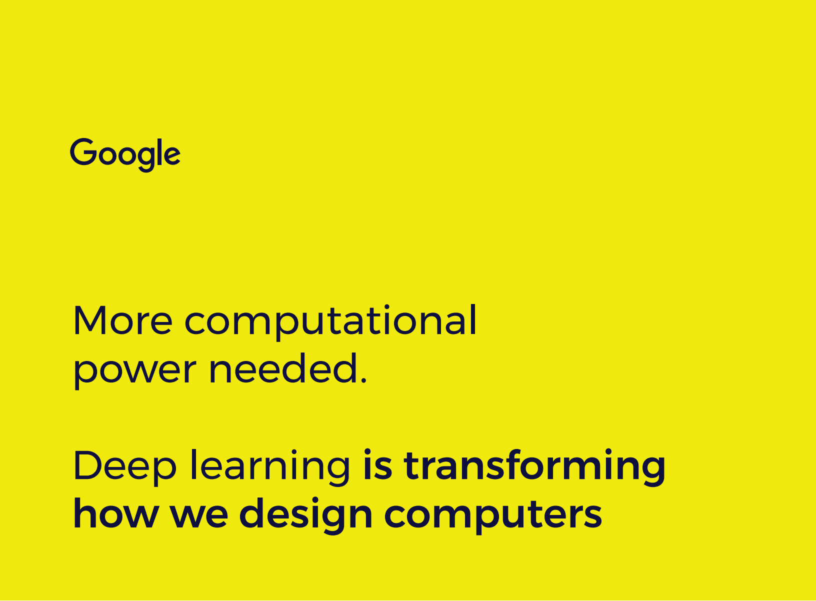
We noted in Section 3.2 the system complexity involved in building and deploying Machine Intelligence solutions. One other aspect to mention is hardware choice. This is becoming a much more meaningful question than it has been. Today, NVIDIA has a dominant position in the deep learning market with its GPU and CUDA platform.

> *"We see NVIDIA as a major beneficiary of the 4th Tectonic Shift in Computing, where serial processing (x86) architectures give way to massively parallel processing capabilities as the next wave of connected devices approach 10b units by 2022"*
> ***Mark Lipacis, Jefferies & Co.***[38]

In fact, a simple rule of thumb for deep learning developers choosing hardware has been to choose the GPU with the most memory that they can afford.

> *"To be honest, there isn't that much choice. We do look at the memory on the cards."*
> ***Tobias Rijken, Kheiron Medical***

However, hardware innovation is starting to address increased demand (Figure 6) and 2017 saw multiple new hardware options for Machine Intelligence workloads announced. These range from initiatives from major chip vendors such as Intel[39] and AMD[40]; top tech companies like Google[41], Tesla[42] and Apple[43] building their own chips; and startups such as the UK's celebrated Graphcore developing new types of processor to meet the demands of Machine Intelligence workloads.



Google

More computational power needed.

Deep learning **is transforming how we design computers**

Figure 6
Slide from the 'Machine Learning for Systems and Systems for Machine Learning' talk given by Jeff Dean, Google Brain, at the NIPS workshop on Systems for Machine Learning on 8 December 2017[45]

At the same time, commercial cloud providers offer access to a range of appropriate hardware for rent, along with software and system options, and engineering expertise that can assist startups in building and deploying scalable and robust solutions.

Each of these options has implications for the performance, cost, energy consumption and engineering effort of a startup's experiments, and these trade-offs are non-trivial to evaluate (see appendix A.2). With more unusual technology options such as neuromorphic[46] and quantum[47] computing peeping over the horizon, choosing the right architecture will become even more complex.

3.4 FUTURE DIRECTIONS

The application examples chosen earlier in this section might be grouped as perceptual tasks. Improving and applying these techniques to new domains will continue to yield useful results, products, and services.

The next frontier for Machine Intelligence is applications that require higher-level cognitive functions such as **planning, reasoning, knowledge-abstraction, and decision-making.**

This distinction between processing and reasoning is important. Powerful deep learning architectures, such as ResNets, are highly capable visual processors, but they may not be the most appropriate choice for reasoning about arbitrary relations.[48]

These tasks are likely to require progress in many research fields (for a more detailed analysis, see Francois Chollet's blog on 'The Limitations of Deep Learning' [49]). From a computational point of view, the prolific success enjoyed by deep learning approaches in recent years would not have been possible without concurrent advances in GPU hardware (specifically via the data parallelism that GPU architecture permits). It is likely that continued hardware and systems innovation will allow other disciplines to flourish in the same way (see Box 1).

Likewise, despite this report's focus on deep learning, surprisingly few commercial products and services rely upon it. Nor is that likely to change – solutions that combine multiple techniques (such as combinations of deep learning and rules-based or relational approaches) seek to combine the advantages of each. Flexible, heterogeneous computation platforms may be required for these 'hybrid' models. Some of the UK Tier 2 HPC centres (see later) have been designed with this in mind.

With Machine Intelligence technologies still in their relative youth, the techniques that enjoy success and market prominence today may not be the same ones that power the products of tomorrow.

> *"Max Planck said, 'Science progresses one funeral at a time.' The future depends on some graduate student who is deeply suspicious of everything I have said. … My view is throw it all away and start again… I suspect that means getting rid of back-propagation."*
> ***Geoff Hinton, Google.*** [51]

Graphical models are an intuitive way of representing and visualising the relationships between many variables, combining uncertainty with logical structure. However, querying a graphical model is computationally intractable ('every type of inference in graphical models is NP-hard or harder'[52]) so approximate inference techniques are used instead. These approximation algorithms themselves can be quite computationally intensive. A 10x faster compute platform therefore permits 10x faster sampling or iteration.

Reinforcement learning (RL) is the problem of getting an agent to act in the world so as to maximise its rewards, or in other words, to make 'rational decisions'. Two strategies for approaching RL problems are trial and error (e.g. Q-learning) and search (e.g. evolution strategies)[53]. They involve conducting experiments over possible actions or model parameters. Here, faster compute allows faster experimentation and exploration of more complex scenarios.

Box 1: Examples of machine intelligence disciplines that could flourish with computational innovation

The UK's breadth and depth of machine intelligence expertise puts it in a good position to lead on complex systems that combine learning, memory and reasoning. It needs the right environment to turn promising research into valuable products and services.

As mentioned in the introduction, we hypothesise that barriers around access to computation power and expertise around it slow down innovation, particularly for startups. Digital Catapult conducted extensive market research to support and elaborate on this hypothesis. The methodology and results of the study are described in the next section.

MARKET RESEARCH: METHODOLOGY AND RESULTS

In order to better understand constraints around computation power Digital Catapult conducted extensive market research, with design informed by desk research and detailed interviews with startups representing a range of stages, technology and application areas. The survey was promoted through Digital Catapult's five local centres (in Brighton, Sunderland, Belfast, Bradford and London) and on social media.

4.1 SURVEY RESPONDENTS

We estimate[54] that there are 600 machine intelligence SMEs in the UK, of which around 400 of are startups[55], and around 270 of those are at an early (angel- or seed-investment) stage.[56]

In all, we reached 81 startups via 13 in-depth interviews, and 68 survey responses. The startups were located all around the UK (the location was unclear for 3 of them) but showed a strong skew towards the south-east, in particular London. This reflects the actual location skew that we encounter (Figure 7).

It's obvious that the level of computation required by a startup is highly dependent on workload. The extent to which compute is seen as a 'constraint' may depend on the level of funding/revenue. Therefore, we asked respondents to give details about their stage of development as well as the Machine Intelligence algorithms and tasks being developed or used, in case these highlighted certain use-cases over others.

4.2 COMPUTATION-CONSTRAINT

It seems reasonable to suppose that very early stage companies are more likely to be computation-constrained than late ones (because of lack of cash) and there is evidence of this, which we can see by plotting computation-constraint by company stage, as illustrated in Figure 8:

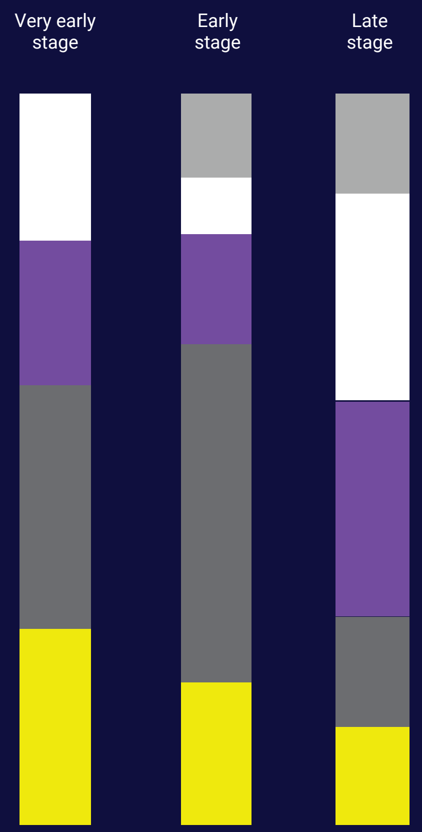


Figure 7
Map of survey respondent locations

Figure 8
Computation-constraint plotted by company stage

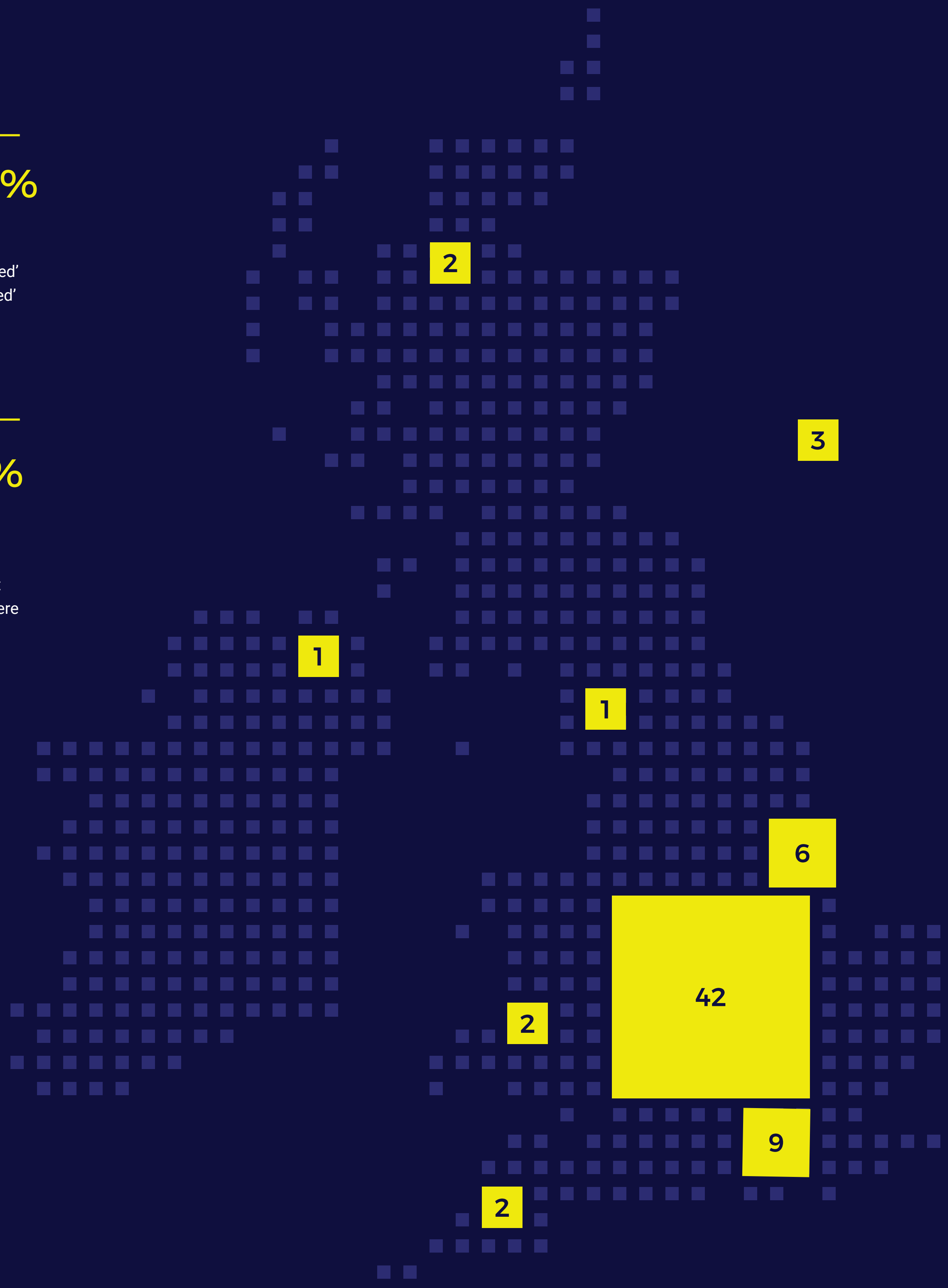
● Strongly Disagree
● Disagree
● Neutral
● Agree
● Strongly Agree

60%

of startups 'agreed' or 'strongly agreed' that they were computation-constrained.[57]

22%

Another 22% disagreed, whilst the remainder were neutral.



4.3
EFFECT OF DIFFERENT MACHINE INTELLIGENCE ALGORITHMS

It was no surprise to find that machine learning is used by nearly all (93%) of the survey respondents. However, many respondents were using simulation (19%) and symbolic (26%) approaches as part of their solution (Figure 9).

Reinforcement learning practitioners were most likely to feel compute-constrained (79%) but the next-most constrained were users of symbolic approaches (78%) (Figure 10). This may reflect a resurgence of 'hybrid' approaches (all of the companies that used symbolic AI techniques also used other approaches), and/or may reflect related tasks such as indexing of large corpuses, data pre-processing and knowledge graph caching that startups told us they struggled to do efficiently.

70% of startups using image/video or speech/audio training data said that they were compute-constrained. Less obviously, three quarters of survey respondents working with behavioural data described themselves as compute-constrained. We speculate that these tasks may involve a lot of data, or they may need to model and store long-range dependencies, or do inference over streamed data.

Figure 9 ▼
Approaches used by surveyed companies

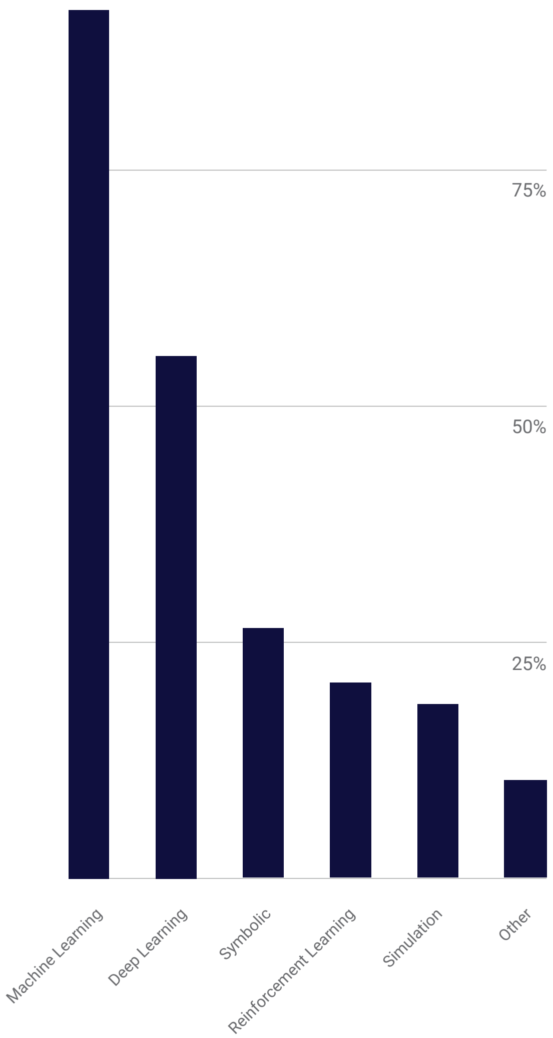


Figure 10 ▼
Compute-constraint by algorithmic approach

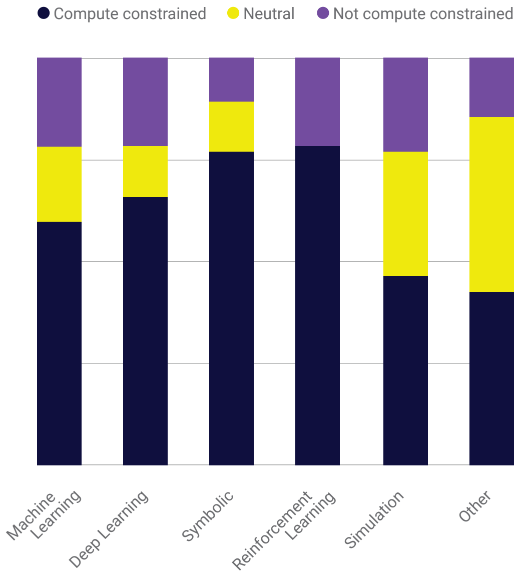
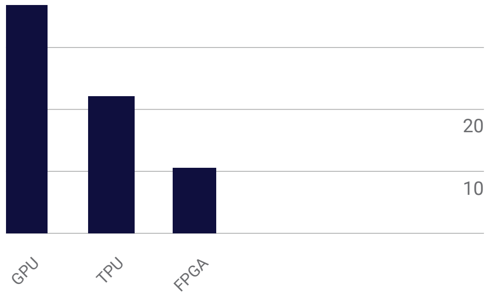


Figure 11 ▼
Use of specialist cloud compute instances



4.4
HARDWARE USED

We saw in section 3.2b how Google is using its own TPUs to make low-latency deployment of its speech synthesis and translation products possible. Others, like Microsoft, are using FPGAs for efficient inference[58]. Heterogeneous computing, where more than one type of processor or core may be required to build efficient implementations (e.g. as in CPU and GPU coprocessors) is an essential offering of commercial cloud providers, and a potential advantage of UK HPC – but it can be non-trivial to engineer. We therefore asked startups about what hardware they bought, what they rented, and why.

Rental

We asked startups using cloud providers about the hardware they rented. We did not ask about the various other services, storage, and support options offered by cloud providers. Most companies used cloud compute, and 40 companies used 'specialist' compute instances. Unsurprisingly, nearly all (95%) of those companies were using GPUs. Three companies were using Google's TPU (for R&D, not deployment) and three used FPGAs (Figure 11). This reflects the relative maturity/availability of these options in the cloud.

Some startups have access to cloud 'credits' via cloud providers' startup programmes. Machine Intelligence startups told us that they were likely to be much heavier users of the credits than other startups who in general use only a small proportion of the credits available to them (e.g. to host their website). Credits are typically only available through select Venture Capital and Accelerator partners. In fact, part of the motivation for a Machine Intelligence startup to join an accelerator may be to take advantage of this benefit.

These are the reasons startups told us they chose to rent hardware:

Availability – Poor local connectivity can mean that owned hardware cannot be accessed reliably remotely
Reliability – Owned hardware sometimes overheats and needs to be restarted, or breaks and needs to be replaced
Scalability – Cloud services allow users great flexibility to scale computation up and down as their needs change
Support – Cloud providers' support their customers to resolve issues relating to computational infrastructure and advise on relevant options for their requirements
Up-to-date – Protection from hardware obsolescence
Overhead – Less engineering and management overhead than owned hardware – someone else deals with the set-up, power, noise, security, space, and air-conditioning requirements

▸ **Figure 13**

68%

of startups told us that they would like to be able to access a GPU or GPU cluster owned by Digital Catapult

▸ **Figure 14**

65%

would like to be able to try newly available hardware before renting or purchasing it

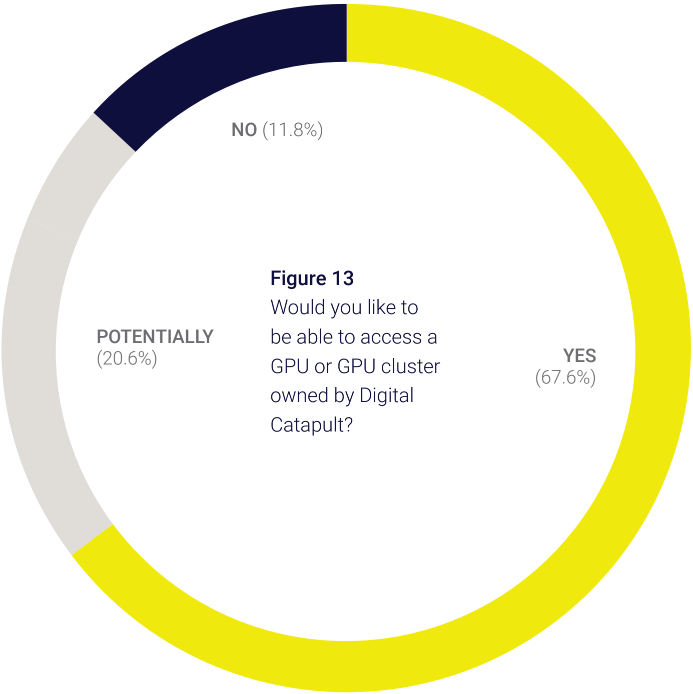


Figure 13
Would you like to be able to access a GPU or GPU cluster owned by Digital Catapult?

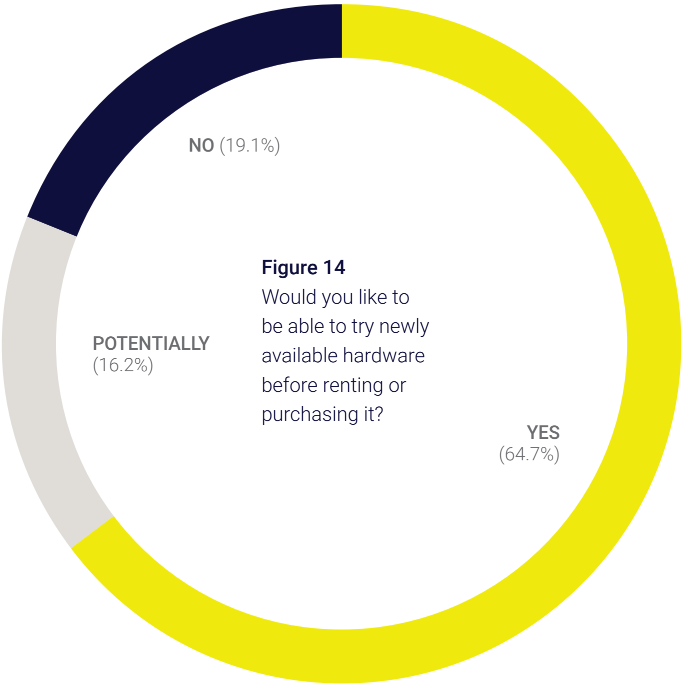


Figure 14
Would you like to be able to try newly available hardware before renting or purchasing it?

PURCHASE

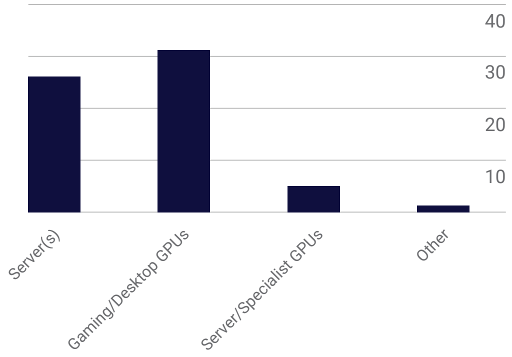
47 (69%) of respondents had bought their own hardware for R&D and/or deployment. Of these, 42 told us what they'd bought (Figure 12). Startups told us that they bought hardware for the following reasons:

Price
- Buying can be cheaper than renting.

Data
- Data upload speeds to the cloud could be too slow
- Expense of storing data in, and downloading data from, the cloud
- Control and compliance issues around data privacy

Choice
- No ability to rent gaming / desktop GPUs and a lag before the very latest enterprise-level GPU instances are available to rent[59]

Figure 12 ▾ Hardware bought



4.5
FOCUS ON UK HIGH PERFORMANCE COMPUTING FACILITIES

National High Performance Computing (HPC) infrastructure in the UK consists of three tiers of resource, from the largest capability machines (Tier 1) to experimental and customised architectures (Tier 2) and local university resources (Tier 3). In 2017, the Tier 2 infrastructure was refreshed with investment of a total of £20m in six new centres[60], including:

- Peta-5 (now CSD3[61]) facility (led by University of Cambridge)
- JADE the Joint Academic Data Science Endeavour[62] (led by University of Oxford and hosted at STFC Hartree Centre)
- Cirrus[63] (led by University of Edinburgh)

These centres provide diverse computing architectures including high throughput and GPU computing, supported by local expertise. The facilities naturally prioritise academic users, but industrial collaboration and use is encouraged[64]. 63% of survey respondents said that they would be interested in being able to access UK HPC facilities (Figure 15).

Potential issues that were noted included:

- Cost
- Ease of managing and provisioning efficient access to the resource
- Compatibility with commonly used tools and software

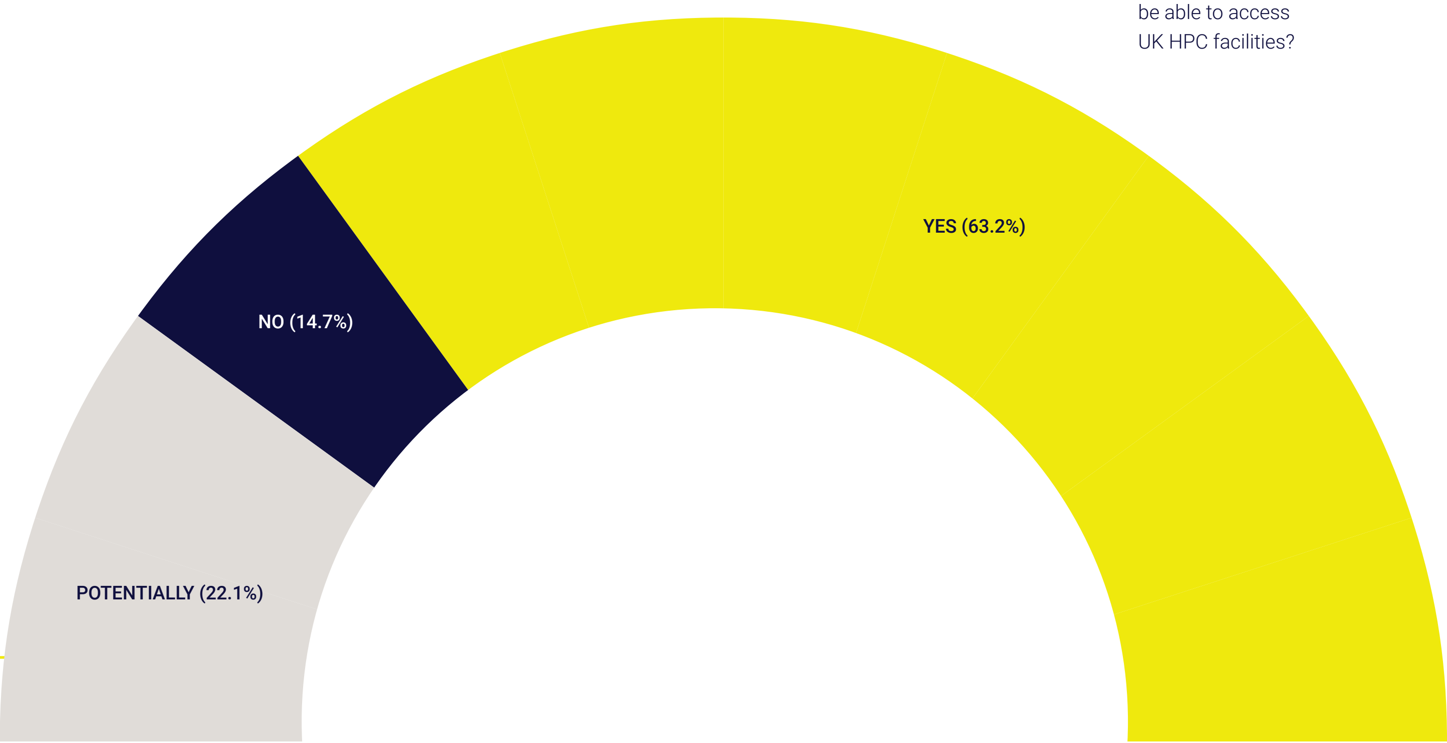


Figure 15
Would you like to be able to access UK HPC facilities?

What is a container?
Containers are a solution to the problem of how to get software to run reliably when moved from one computing environment to another

It is becoming easier to port code through HPC support of 'containerisation' (see Box 2). HPC may offer other benefits such as efficient data pre-processing and expertise in optimising code for heterogeneous compute nodes or in distributing models over multiple nodes.

What is a container?
'Containers are a solution to the problem of how to get software to run reliably when moved from one computing environment to another...

...Put simply, a container consists of an entire runtime environment: an application, plus all its dependencies, libraries and other binaries, and configuration files needed to run it, bundled into one package. By containerizing the application platform and its dependencies, differences in OS distributions and underlying infrastructure are abstracted away.'[65]

Examples of container solutions are Docker[66], Rkt[67], Shifter[68] and Singularity[69]

Box 2: Containers

4.6
OTHER REQUIREMENTS

We asked startups what else they would find useful (Figure 16). Unsurprisingly, data was the most in-demand item, specifically access to new public (or private, sandboxed) datasets for training machine learning models. There was also great interest in contractual templates for data sharing. Data sharing agreements must be keeping many a lawyer busy at the moment, since each startup has to 'reinvent the wheel' when it negotiates with its customers around data sharing for model training.

Establishing a template with common principles and options, which complies with relevant regulation, has the potential to save all parties time and reassurance. Indeed, this is a central recommendation of the recent independent report 'Growing the artificial intelligence industry in the UK' by Dame Wendy Hall and Jérôme Pesenti[70], as well as the Touchpaper[71] initiative. A number of startups commented on the need for software and system expertise, both for training and deployment of machine learning models. Two common issues were the time consumed in setting up on new infrastructure, and the engineering effort required to automate data and experiment pipelines and ensure maximum utilisation of hardware.

> *"The bottleneck is around software and system expertise – it's not easy setting up in the cloud, using docker and kubernetes for the first time (although of course cloud providers have support and there is much to read)."*
> ***Pyry Takala, True AI***

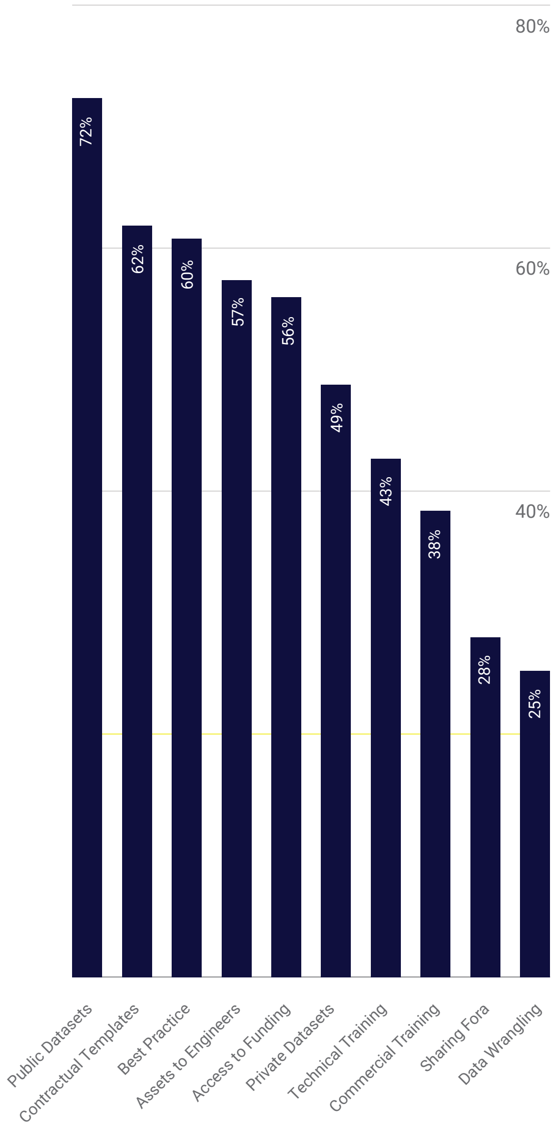
"If you don't have workflow in place, you're not utilising the machine fully. It's all very well being able to spin up a GPU instance and train a model, but if it finishes in the middle of the night and needs human intervention to start a new job that's a waste."
Tobias Rijken, Kheiron Medical

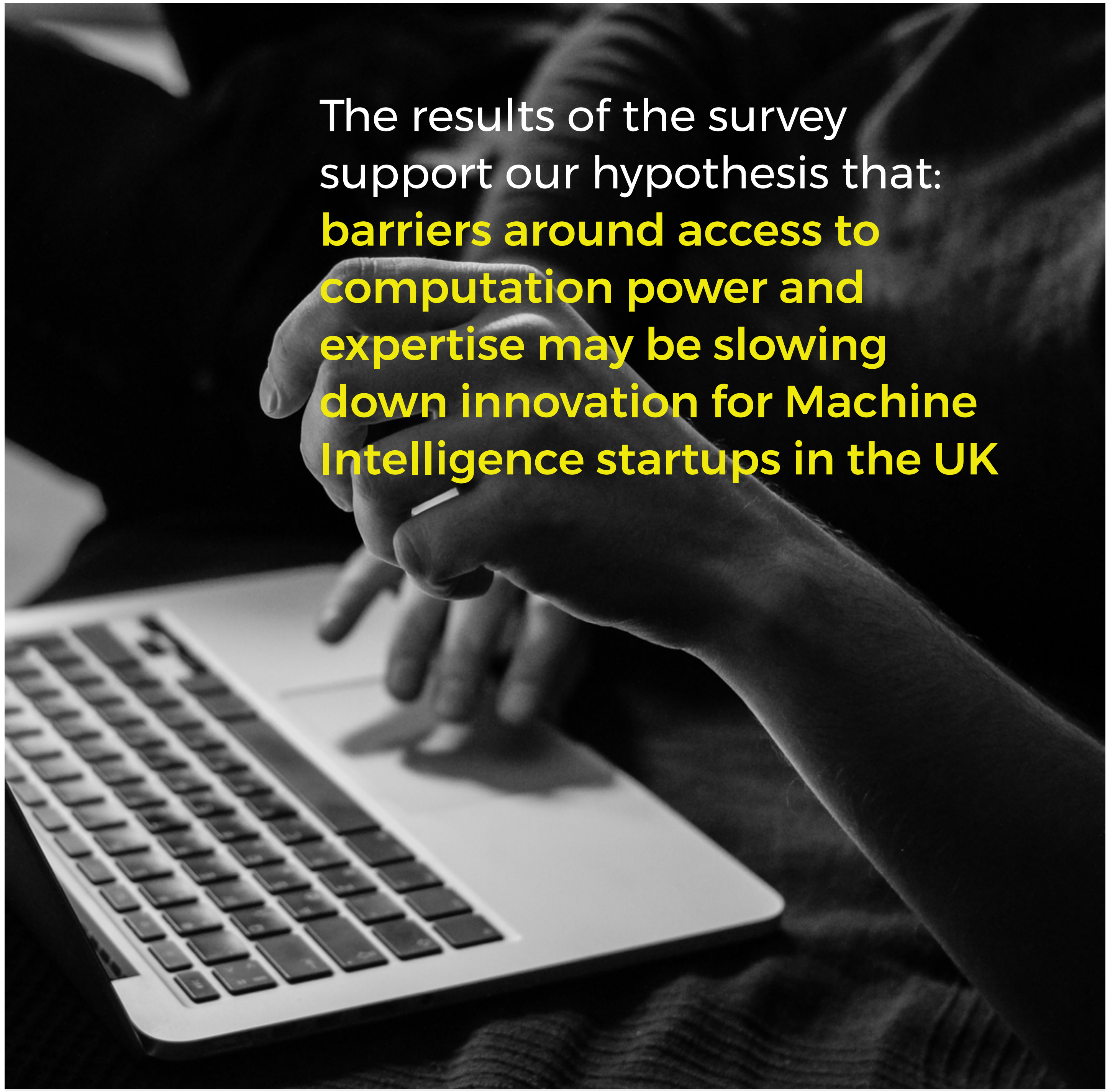
Startups additionally listed the familiar 'shopping' items of the resource-constrained: office space, access to customers, and access to other functions such as marketing. A co-working space would have the additional benefit of being a physical location for sharing of relevant experience and best practice, supplemented with guidance provided online – the third-most sought after listed item.

4.7 Summary

The results of the survey support our hypothesis that barriers around access to computation power and expertise may be slowing down innovation for Machine Intelligence startups in the UK. However, no possible solution we could offer is 'one-size-fits-all', and both Machine Intelligence and computation solutions are rapidly evolving, so any intervention needs to be targeted and flexible. In the next section, we propose what these might be, and future directions.

Figure 16 ▼
What would you like Digital Catapult to provide?





The results of the survey support our hypothesis that: **barriers around access to computation power and expertise may be slowing down innovation for Machine Intelligence startups in the UK**

CONCLUSIONS AND FUTURE DIRECTIONS

This research identified specific areas where Digital Catapult could collaborate with startups and existing commercial and public expert partners to overcome computational barriers to success.

5.1 MACHINES

For R&D and proof of concept

The costs of computation can accumulate very quickly for early-stage Machine Intelligence startups, and are difficult to defer to post-financing or revenue. Digital Catapult could provide access to compute that is:

- Not commonly available to rent from the commercial sector
- For short and meaningful experimentation and static/offline training
- To enable proof of concept and reach new product milestones

For continuous development, deployment, and scaling

Through partnerships with commercial and public computational resource providers, Digital Catapult could assist qualifying startups to access these services on preferential terms. The services should be tailored to Machine Intelligence startup requirements and allow them to scale their development and deployment pipelines, with expert support. Service providers may value access to, and feedback from, this fast-growing community.

Experimentation with new hardware

Digital Catapult could provide a hub for companies of all sizes to learn more about and try new hardware as it becomes available. Matching hardware innovators with startups solving interesting problems could represent a ‘win-win’ proposition.

5.2 INTELLIGENCE

Information repository

Digital Catapult could provide an information repository on topics including getting up and running in the cloud, data security and compliance, and data and experiment pipeline management. In addition Digital Catapult should signpost startups to other sources of expertise and specialist services.

Benchmarking and performance modelling[72]

Given the burgeoning costs of Machine Intelligence workloads, it is vital for startups to be able to understand the tradeoffs involved in using different computer architectures and service options. Digital Catapult could contribute to community efforts to benchmark and provide performance modelling tools.

Practical workshops and expertise

In collaboration with partners and collaborators, Digital Catapult could provide training on topics such as distributed training, cluster management and job scheduling, and heterogeneous computing; and develop in-house expertise to assist startups with common computation-related problems.

In the future, we expect that demand for other service innovations will emerge. As Machine Intelligence startups mature, their requirements will change. As the Machine Intelligence landscape evolves, the computational workloads will too. To keep up with these developments, Digital Catapult could conduct similar market research projects periodically.

To conclude, we identified computation as a potential barrier to Machine Intelligence startups' ability to realise ambitious business plans, and we conducted research to ascertain where computation-constraints were being felt and what possible interventions could alleviate them. We found that whilst the sheer cost of compute is a barrier for many startups, there are also significant challenges around systems expertise.

To conclude, we identified computation as a potential barrier to Machine Intelligence startups' ability to realise ambitious business plans, and we conducted research to ascertain where computation-constraints were being felt and what possible interventions could alleviate them. We found that whilst the sheer cost of compute is a barrier for many startups, there are also significant challenges around systems expertise.

APPENDICES

To supplement the paper we provide three appendices. In the first, we give an explanation of how we derived two computational cost claims we made earlier in the report. The second appendix contains a discussion around the need for hardware benchmarking and performance modelling, and highlights some existing initiatives in this space. The third appendix contains a hardware glossary (limited to the terms used in this report), along with some explanatory notes regarding current relevant hardware architectures and their uses.

A.1
WORKINGS

In this appendix we briefly explain our assumptions and calculations relating to two claims we make in the report about the computational cost of particular Machine Intelligence workloads.

A1.1

Comparing the compute budget for DeepMind's AlphaZero and the world's supercomputers

- TPU1: 'The heart of the TPU is a 65,536 8-bit MAC matrix multiply unit that offers a peak throughput of 92 TeraOp/s (TOPS) and a large (28 MiB) software-managed on-chip memory'[73]
- TPU2 (4 chips per unit) '180 teraflops of computation, 64 GB of HBM memory, 2400 GB/s mem BW'[74]
- AlphaZero 'Training proceeded for 700,000 steps (mini-batches of size 4,096) starting from randomly initialised parameters, using 5,000 first-generation TPUs to generate self-play games and 64 second-generation TPUs to train the neural networks'[75] => 5000x92 + 64x180 = 471,520 TeraOp/s
- Top 10 supercomputers (November 2017) => 389,083 TeraFLOP/s for top ten
- But we are not comparing the same precision operations, so take it with a pinch of salt. At the silicon core, the mixed-precision (16.32b) multiply-add operations becoming ubiquitous for machine intelligence require about 1/16th the silicon area and energy of the double-precision (64b) multiply-add operations which dominate legacy HPC

A1.2

Calculating the operations required to train a single ImageNet classifier

Our assumptions:

- Training (forward and backwards passes) needs c.3 times as many operations as the forward pass alone (forward + backward + weight updates on each of the parameters)
- Each multiply-add operation is 2 FLOPs (floating-point operations)

Examples of computation for image classifiers trained on the 1.28m images from ImageNet:

- The 16-layer VGG-16[76] network from Oxford University requires ~31 billion FLOPs per image for one forward pass, and was trained over 74 epochs. Total **8.8 exaFLOPs**
- The 50-layer ResNet-50[77] network from Microsoft Research requires ~8 billion FLOPs per image for one forward pass, and was trained over 120 epochs. Total **3.7 exaFLOPs**
- The 76-layer Inception-v4[78] network from Google Brain requires ~25 billion FLOPs per image for one forward pass, and was trained over 160 epochs. Total **15.4 exaFLOPs**

A.2
BENCHMARKING AND PERFORMANCE MODELLING

Startups that want to select appropriate computational infrastructure and to budget for computational expenses need to balance the need for fast results with cost, power and engineering considerations.

Estimating the performance of hardware for a startup's particular workload can be non-trivial. A common approach is to calculate the number of operations in the algorithm and compare that with hardware peak performance specifications (given in floating-point operations per second, FLOP/s). However, this can be an oversimplification, since it ignores memory and communication constraints which can result in significant degradation of performance[79]. So, FLOP/s can only get one so far.

Similarly, one can extrapolate from the academic literature, where authors sometimes report on how long their experiments take. But unless this is reported alongside the exact hardware and software used, configuration details, and algorithmic choices (e.g. batch size) this is an exercise in false precision.

'We have an idea of the performances by trying multiple servers on the Cloud, but the cost is so high that we would benefit from analytics on hardware.' Survey respondent

Benchmarking and performance modelling are community efforts to illuminate this complex space:

Benchmarking

> *"Deep learning developers and researchers want to train neural networks as fast as possible. Right now we are limited by computing performance,... The first step in improving performance is to measure it..."*
>
> ***Greg Diamos, senior researcher at Baidu's Silicon Valley research lab[80]***

A number of attempts have been made to establish reference workloads to test hardware so that relative performance can be assessed. These include Deepbench[81], Deepmark[82], DAWNBench[83] and Fathom,[84] and are focused on performance measured against time (or cost), but energy efficiency is of increasing importance:

> *"The problem with GPUs is they're very power intensive, ... We had to put a moratorium on people putting big GPU clusters in their offices. ... [software developers need more efficient hardware that] doesn't require us to build power stations next to the data center."[85]*
>
> ***Andrew Moore, dean of Carnegie Mellon's computer science school.***

62%

of survey respondents said they would use third-party benchmarks if they were more comprehensive

62%

of survey respondents said they would use analytical methods to estimate the time and cost of using different hardware, frameworks, and service-providers, if they were more readily available

More detailed discussion about benchmarking metrics can be found via MIT's 'Tutorial on Hardware Architectures for Deep Neural Networks.'[86]

Performance Modelling

Benchmarking initiatives are important, but it can be difficult to extrapolate from the test problems to a specific workload, since the workloads may be quite different or the performance achieved may have been due to very specific parameter, software and implementation choices. In addition, they may not cover all hardware options.

That is why attempts to predict hardware performance against any workload are being developed. One such open source initiative is called Paleo[87], 'an analytical performance model for exploring the space of scalable deep learning systems. By extracting computational requirements carried by neural network architectures and mapping them to the design space of software, hardware, and communication strategies, Paleo can effectively and accurately model the expected scalability and performance of a putative deep learning system.'[88]

To be truly useful, such tools need to work for algorithm descriptions in all relevant frameworks and for all relevant hardware, and this is a huge task for any such open source project.

62% of survey respondents said they would use analytical methods to estimate the time and cost of using different hardware, frameworks, and service-providers, if they were more readily available.

In practical terms, startups also have to estimate the engineering overhead relating to the ease of getting up and running on different architectures and considerations such as framework support. The total cost of ownership may be difficult to accurately calculate; similarly total cloud costs will depend on provider, choice of instances, billing-unit, data costs and pricing model (on-demand, spot-pricing, block-buying etc).

Digital Catapult could support benchmarking and performance modelling initiatives by capitalising on its vendor-agnostic positioning to run experiments and contribute to open source libraries. It could collaborate with partners to develop expertise in optimising set-up routines, code and systems, and make this information widely available through its website and workshops.

A.3
HARDWARE GLOSSARY

CPU
Central Processing Unit

GPU
Graphical Processing Unit

IPU
Intelligence Processing Unit

TPU
Tensor Processing Unit

Core
The word ‘core’ here is used to describe a processor capable of running a programme, which is the convention used by every computing and silicon company except the GPU manufacturers who use the word ‘core’ in their GPU marketing to refer to the (much larger) number of floating-point units.

FPGA
Field Programmable Gate Array. A highly flexible hardware device which requires its processing function to be configured once at power-up, after which it is fixed in contrast to a processor which executes a software program to determine its function dynamically as it runs.

Floating-Point Unit
A floating-point unit is the arithmetic execution part of a computer system designed specifically to carry out operations (such as +, -, x, /, sqrt) on floating-point numbers.

HPC
High Performance Computing. The traditional computing workloads of supercomputers. Examples include high energy physics, weather forecasting, seismology and drug discovery. These workloads contrast with Machine Intelligence particularly in their requirement for high precision arithmetic calculations.

FLOP
A floating-point operation is any mathematical operation or assignment that involves floating-point numbers, but is typically quoted as twice the number of fused multiply-add (FMA) operations. Such FMAs are the basis of almost all tensor calculations performed in Machine Intelligence. In this usage a count of FLOPs is comprised of half multiplications and half additions. Not all FLOPs involve numbers of the same size (in bits) – the IEEE Standard 754-2008 for Floating-Point Arithmetic defines a 64-bit ‘double’ precision common in HPC, a 32-bit ‘single’ precision common in graphics, and a 16-bit ‘half’ precision becoming common in Machine Intelligence.

FLOP/s
Floating-point operations per second, a measure of computer performance (commonly but confusingly also written as FLOPS).

Arithmetic Precision
See FLOP

Computer systems have frequently complemented the CPU with special purpose accelerators for intensive tasks, most notably graphics, but also sound, video, etc. Over time various accelerators have appeared that have been applicable to AI workloads.[89]

The GPU, TPU and IPU are accelerators of Machine Intelligence workloads and work in concert with a CPU. Each of these accelerators uses the CPU as a conduit to storage and network resources. The CPU can of course operate alone, and the line between CPUs and coprocessors is being blurred by the integration of ever more potent vector units into the CPU – the ultimate limit on this trend will be power-density of individual chips, which is expected to favour specialisation. All silicon computing platforms, including all processors relevant to Machine Intelligence, are power limited.[90]

A modern CPU, GPU, IPU or TPU contains many cores, each with one or more execution units capable of arithmetic on each element of a vector of numbers in parallel.

- The largest Xeon Phi CPU from Intel has 72 cores, each with a 512bit wide arithmetic path which operate on vectors of 32 16bit floating-point numbers[91]
- The largest Volta GPU from NVIDIA has 320 cores, each with a 1024bit wide arithmetic path which can operate on vectors of 64 16bit floating-point numbers[92]
- The upcoming Colossus IPU from Graphcore has 1,216 cores each with a 64bit narrow arithmetic path which can operate on vectors of 4 16bit floating-point numbers[93]
- The second generation TPU from Google has 2 cores each with a 4096bit wide arithmetic path which operates on vectors of 128 32bit floating-point numbers (using 16bit precision internally)[94]

These numbers themselves do not reflect the performance available, firstly because all silicon computation is power limited and these different devices consume different amounts of power, with multiple chips often connected together to form a cluster. For example, Google’s card has 4 TPU chips in an undisclosed power envelope likely to be somewhat greater than a GPU, and Graphcore’s card has 2 IPU chips in the same power envelope as a single GPU. Secondly, as mentioned in Appendix A2, performance is critically dependent on the bandwidth available to local memory – the ability to feed the floating-point units with data.

So far, the most advanced application of FPGAs in Machine Intelligence is the Microsoft Brainwave in which Intel FPGAs are configured to form arrays of processors which can then run programs; these processors are called ‘soft DNN processing units (DPUs)[95]. The flexibility provided by FPGAs allows such soft processors to use different data sizes for different calculations, but such configurability carries an extreme efficiency cost of 1-2 orders of magnitude compared with the hardened design of execution units in a processor.

FOOTNOTES

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

[2] Consultancy UK: £232 billion AI market is UK's largest economic opportunity, 4 July 2017

[3] https://newsroom.accenture.com/subjects/technology/artificial-intelligence-poised-to-double-annual-economic-growth-rate-in-12-developed-economies-and-boost-labor-productivity-by-up-to-40-percent-by-2035-according-to-new-research-by-accenture.htm

[4] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/industrial-strategy-white-paper-web-ready-version.pdf

[5] We use 'computation' to mean calculations and related processes undertaken by computers and computer systems

[6] Barriers: scaling UK machine learning companies https://www.digitalcatapultcentre.org.uk/machine-learning-barriers/

[7] Quote from 2016/17 lecture series 'Deep Learning for Natural Language Processing', Lecture 6 'Hardware and Software for NLP' (audio) http://media.podcasts.ox.ac.uk/comlab/deep_learning_NLP/2017-01_deep_NLP_6_nvidia_gpus.mp4

[8] D. Silver, T. Hubert, J. Schrittwieser et al., 'Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,' 2017

[9] J. Hestness, et al, 'Deep Learning Scaling is Predictable, Empirically,' 2017.

[10] Slide 6, 'Deep Learning scaling is predictable (empirically)', Greg Diamos December 9, 2017 'https://supercomputersfordl2017.github.io/Presentations/scaling-is-predictable.pdf (retrieved on 3rd January 2018)

[11] OpenAI's mailing address, given (with tongue-in-cheek) in Jack Clark's Import AI newsletter is, 'Many GPUs, Oakland, California 94609'

[12] Tweet on 19 December 2017 by Samim https://twitter.com/samim/status/943154745816084485

[13] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era,' 2017

[14] C. Sun, K. Murphy et al., 'Speed/accuracy trade-offs for modern convolutional object detectors,' 2017

[15] O. Russakovsky, J. Deng et al., 'ImageNet Large Scale Visual Recognition Challenge'. IJCV, 2015

[16] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition,' 2015

[17] A. Krizhevsky, I. Sutskever, and G. Hinton, 'Imagenet classification with deep convolutional neural networks,' 2012.

[18] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, 'Learning Transferable Architectures for Scalable Image Recognition', 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition,

[20] http://dawn.cs.stanford.edu/benchmark/ (retrieved 14 Dec 2017)

[21] A. Krizhevsky, I. Sutskever, and G. Hinton, 'Imagenet classification with deep convolutional neural networks,' 2012.

[22] P. Goyal, P. Dollár, R. Girshick, et al., 'Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour,' 2017.

[23] Y. You, Z. Zhang, C. Hsieh, J. Demmel, '100-epoch ImageNet Training with AlexNet in 24 Minutes,' 2017

[24] https://supercomputersfordl2017.github.io (retrieved 3 Jan 2018)

[25] Slides available from https://supercomputersfordl2017.github.io (retrieved 3rd January 2018)

[26] Deep Learning by Ian Goodfellow, Aaron Courville and Yoshua Bengio, 2016. Section 1.2.3 p21 http://www.deeplearningbook.org

[27] https://blog.ycombinator.com/jeff-deans-lecture-for-yc-ai/ (retrieved 3 January 2018)

[28] K. Hazelwood, S. Bird et al., 'Applied Machine Learning at Facebook : A Datacenter Infrastructure Perspective.' 2017

[29] https://medium.com/initialized-capital/the-public-policy-implications-of-artificial-intelligence-1df075c49755 (retrieved 3 January 2018)

[30] D. Amodei, R. Anubhai et al., 'Deep Speech 2: End-to-End Speech Recognition in English and Mandarin', 2015.

[31] https://DeepMind.com/blog/WaveNet-launches-google-assistant/

[32] https://www.blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

[33] Babel fish, the universal instant translator from Douglas Adams, 'The Hitchhiker's Guide to The Galaxy', 1979.

[34] https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

[35] Y. Wu, M. Schuster et al., 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation', 2016

[36] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, 'Massive Exploration of Neural Machine Translation Architectures,' 2017.

[37] https://research.googleblog.com/2016/09/a-neural-network-for-machine.html

[38] Quoted in http://markets.businessinsider.com/news/stocks/nvidia-stock-price-set-to-dominate-the-forth-tectonic-shift-in-computing-2017-7-1002161013 (retrieved 3 January 2018)\[39] https://www.intelnervana.com/intel-nervana-neural-network-processors-nnp-redefine-ai-silicon/

[40] https://instinct.radeon.com/en/product/mi/radeon-instinct-mi25/

[41] https://www.blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

[42] https://www.theregister.co.uk/2017/12/08/elon_musk_finally_admits_tesla_is_building_its_own_custom_ai_chips/

[43] https://www.wired.com/story/apples-neural-engine-infuses-the-iphone-with-ai-smarts/

[44] https://www.graphcore.ai/posts/big-names-in-machine-intelligence-join-graphcores-new-30-million-funding-round

[45] http://learningsys.org/nips17/assets/slides/dean-nips17.pdf (retrieved 3 January 2018)

[51] Artificial intelligence pioneer says we need to start over https://www.axios.com/ai-pioneer-advocates-starting-over-2485537027.html

[52] Graphical Models in a Nutshell https://ai.stanford.edu/~koller/Papers/Koller+al:SRL07.pdf

[53] Page 8 of Reinforcement Learning: An Introduction by Richard S. Sutton, Andrew G. Barto, 1998; Evolution Strategies as a Scalable Alternative to Reinforcement Learning, https://blog.openai.com/evolution-strategies/

[54] The difficulty in counting is both definitional (what constitutes a 'machine intelligence' 'startup') and data-constrained (proxy data for 'machine intelligence' and 'startup' may not be readily available).

[55] Estimate from amalgamated Digital Catapult, Crunchbase (https://www.crunchbase.com) and AngelList (https://angel.co/) data.

[56] MMC Ventures reports that the rate of incorporation of machine intelligence startups has accelerated in recent years, with one new company incorporated every 5 days and that two thirds are still at the earliest (seed or angel funded) stages of their journey (https://www.mmcventures.com/wp-content/documents/The_State_of_AI_2017_Inflection_Point.pdf).

[57] Constrained by cost/access which slows the company's progress, as opposed to insufficient speed (latency-constrained). Latency may not be solvable by additional compute, as per the speech and translation examples noted in section 3.2b which required algorithmic and engineering, not raw-compute, solutions.

[58] https://www.wired.com/2016/09/microsoft-bets-future-chip-reprogram-fly

[59] GPUs may be marketed for desktop/gaming or server/specialist markets. The difference in price can be substantial. Some startups told us that they are quite happy with gaming GPU performance but that these are not usually available to rent in the cloud.

[60] https://www.epsrc.ac.uk/research/facilities/hpc/tier2/

[61] http://www.hpc.cam.ac.uk/CSD3

[62] http://www.jade.ac.uk

[63] http://www.cirrus.ac.uk

[64] https://www.epsrc.ac.uk/files/research/tier2hpcstrategy/

[65] What are containers and why do you need them? https://www.cio.com/article/2924995/software/what-are-containers-and-why-do-you-need-them.html (Retrieved 4 January 2018)

[66] https://www.docker.com

[67] https://github.com/rkt/rkt/

[68] https://github.com/NERSC/shifter/

[69] http://singularity.lbl.gov

[70] https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

[71] A free toolkit to help make startup and corporate relationships flourish https://touchpaper.org/. Digital Catapult is a founding member

[72] See appendix A.2

[73] N. P. Jouppi et al., 'In-Datacenter Performance Analysis of a Tensor Processing Unit,' 2017.

[74] 'Machine Learning for Systems and Systems for Machine Learning' talk given by Jeff Dean, Google Brain, at the NIPS workshop on Systems for Machine Learning on 8 December 2017, available here http://learningsys.org/nips17/assets/slides/dean-nips17.pdf

[75] D. Silver et al., 'Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm,' 2017.

[76] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition,' 2015.

[77] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition,' 2015.

[78] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, 'Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,' 2016.

[79] For example, Baidu publishes hardware performance benchmarks for a suite of machine intelligence computation kernels called DeepMark, characterising the deployment of machine intelligence services in Baidu's datacentres. Results for December 2017 show that Nvidia's latest V100 GPU delivers less than 25% of its headline (125Tflop) performance in 83% of the convolution tests and 61% of the matrix-multiply tests, at the relevant mixed floating-point precision. These two test classes represent just the heavy-lifting arithmetic kernels of typical CNN and RNN usage respectively. Across the whole application the disparity between peak and actual flops will be even greater, and this disparity is not unique to GPUs as a hardware platform https://github.com/baidu-research/DeepBench

[80] http://research.baidu.com/baidu-research-announces-new-open-source-deep-learning-benchmark/ (retrieved 5 January 2018)

[81] https://github.com/baidu-research/DeepBench

[82] https://github.com/soumith/convnet-benchmarks/issues/101

[83] http://dawn.cs.stanford.edu/benchmark/

[84] https://github.com/rdadolf/fathom

[85] https://www.cnbc.com/2017/09/19/ai-chip-startups-graphcore-and-mythic-raising-big-venture-rounds.html (retrieved 5 January 2018)

[86] http://www.rle.mit.edu/eems/wp-content/uploads/2017/06/Tutorial-on-DNN-7-of-9-Benchmarking-Metrics-for-DNN-Hardware.pdf

[87] https://github.com/TalwalkarLab/paleo

[88] H. Qi, E. R. Sparks, and A. Talwalkar, 'Paleo: a Performance Model for Deep Neural Networks,' 2017.

[89] https://en.wikipedia.org/wiki/AI_accelerator

[90] A. Padram, S. Richardson et al, 'Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era,' 2016

[91] https://www.intel.com/content/www/us/en/products/processors/xeon-phi/xeon-phi-processors.html (retrieved 6 January 2018)

[92] NVIDIA Tesla V100 GPU has 80 Streaming Multiprocessors containing 4 processors each http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf (retrieved 6 January 2018)

[93] 'Scalable Silcon Compute' talk given by Simon Knowles, Graphcore, at the NIPS workshop on Deep Learning At Supercomputer Scale on 9 December 2017 available here: https://supercomputersfordl2017.github.io/Presentations/SimonKnowlesGraphCore.pdf

[94] 'Machine Learning for Systems and Systems for Machine Learning' talk given by Jeff Dean, Google Brain, at the NIPS workshop on Systems for Machine Learning on 8 December

[95] Accelerating Persistent Neural Networks at Datacenter Scale' talk given by Eric Chung & Jeremy Fowers of Microsoft at Hot Chips 22 August 2017, available here: https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.60-NeuralNet1-Pub/HC29.22,622-Brainwave-Datacenter-Chung-Microsoft-2017_08_11_2017.pdf

This research report was co-authored by Machine Intelligence team members:

Libby Kinsey, Machine Learning Consultant
libby@projectjuno.ai

Dr Anat Elhalal, Head of Technology
& AI+ML Lead Technologist
anat.elhalal@digicatapult.org.uk

We are eager to collaborate with existing initiatives - public, private, open source - that share our mission to build solutions that help machine intelligence startups reach their potential. We want to hear from startups who want to access our programme or have ideas about other ways that we can assist.

Please contact us at
hello@migarage.ai
www.migarage.ai

CATAPULT Digital

Digital Catapult
101 Euston Road
London
NW1 2RA
0300 1233 101
www.digicatapult.org.uk