

CSC and Machine Learning

UEF Summer School on Machine Learning Applied to Bioinformatics and Speech Technology, August 2017

Markus Koskela
markus.koskela@csc.fi



CSC – Finnish expertise in ICT for research, education, culture and public administration

ICT Solutions for Brilliant Minds

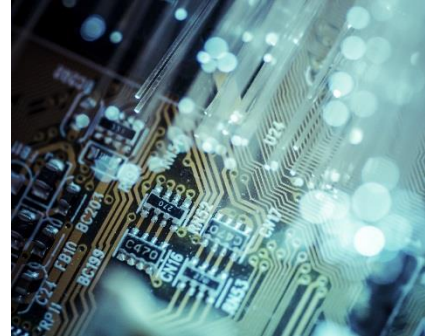


CSC is a Finnish centre of expertise in ICT

that provides ICT expert services at an internationally high level of quality

for research, education, culture, public administration and enterprises,

to help them thrive and benefit society at large.



Non-profit
state enterprise
with special
tasks



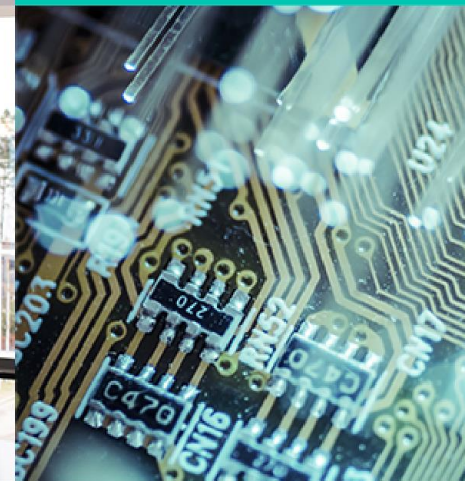
Turnover
in year 2015
36.8M€



Headquarters
in Espoo,
datacenter
in Kajaani,
Finland



Owned
by state
(70%)
and all Finnish
education higher
institutions (30%)



Circa
290
employees
in year 2016

CSC's Services



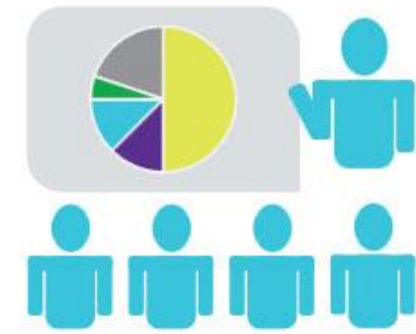
Scientific Computing and
Software



Funet Network Services



Identity and Access
Management



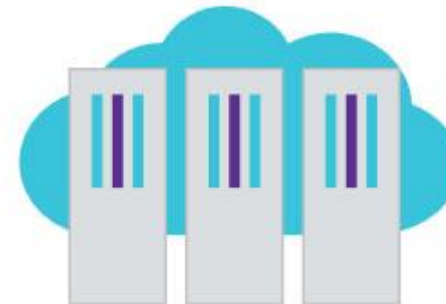
Training services



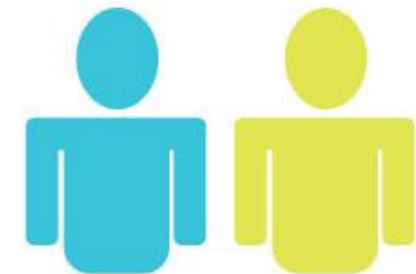
Research Information
Management



Education Management and
Student Administration Services

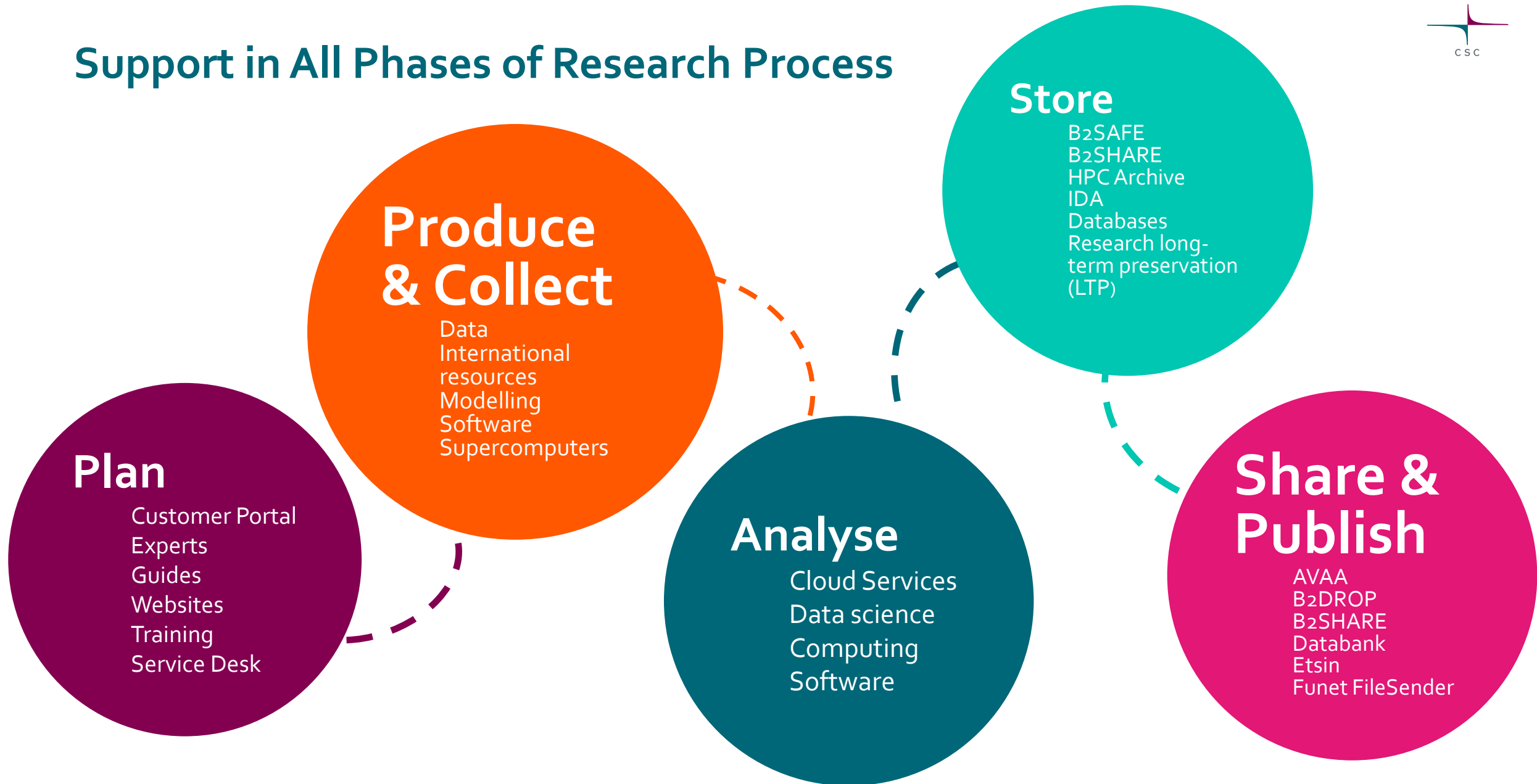


Datacenter and Capacity
Services



Consultation and Tailored
Solutions

Support in All Phases of Research Process



Computing services

- Supercomputer (Sisu)
- Supercluster (Taito)
- Grid (FGCI)
- Cloud services (cPouta, ePouta)
- Accelerated computing (GPUs)
- International resources
 - Extremely large computing (PRACE)
 - Nordic resources (NEIC)

Main computing capacity: Sisu and Taito



	Sisu (Phase 2)	Taito (Phase 2)
Availability	2014-	2015-
CPU	Intel Haswell 2 x 12 cores 2.6 GHz, Xeon E5-2690v3	Intel Haswell and Sandy Bridge, 2 x 12 and 2 x 8 cores, 2.6 GHz, Xeon E5- 2690v3 and E5-2670
Interconnect	Aries	FDR IB
Cores	40512	9768+9216
RAM/node	64 GB	64/128/256/1536 GB
Tflops	1688	515
GPU nodes	-	50
Disc space (shared)	4 PB	4 PB

CSC datacenter in Kajaani

- 3 000 m² (option to 4 000 m² additional datacenter space)
- Redundant green power scalability up to hundreds of MW, based on customer need. Existing power capacity: 10 MW (redundant)
- Local and competent partner network guarantees rapid scalability and secure operations
- State-of-the-art datacenter technology (modularity of the datacenter, easy expansion, free air cooling all year round) delivering world-class eco-efficiency and zero carbon footprint. Annual pPUE 1,03 (2015)
- Also traditional water cooling datacenter facilities are available for certain supercomputer type of services
- High-end availability on both power supply, cooling and core network connectivity



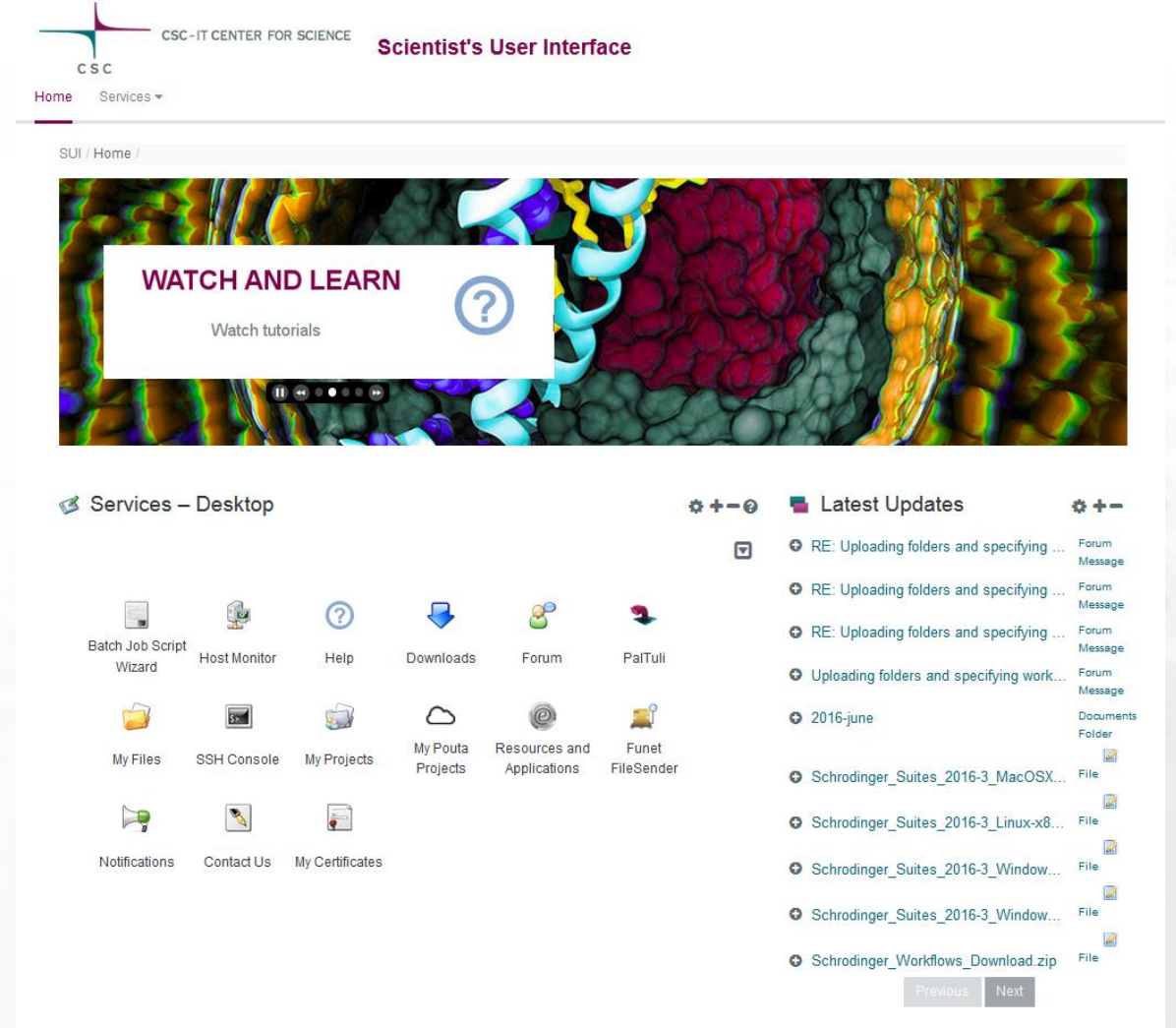
Scientist's User Interface (SUI)



WWW-portal for all CSC users – <https://sui.csc.fi>

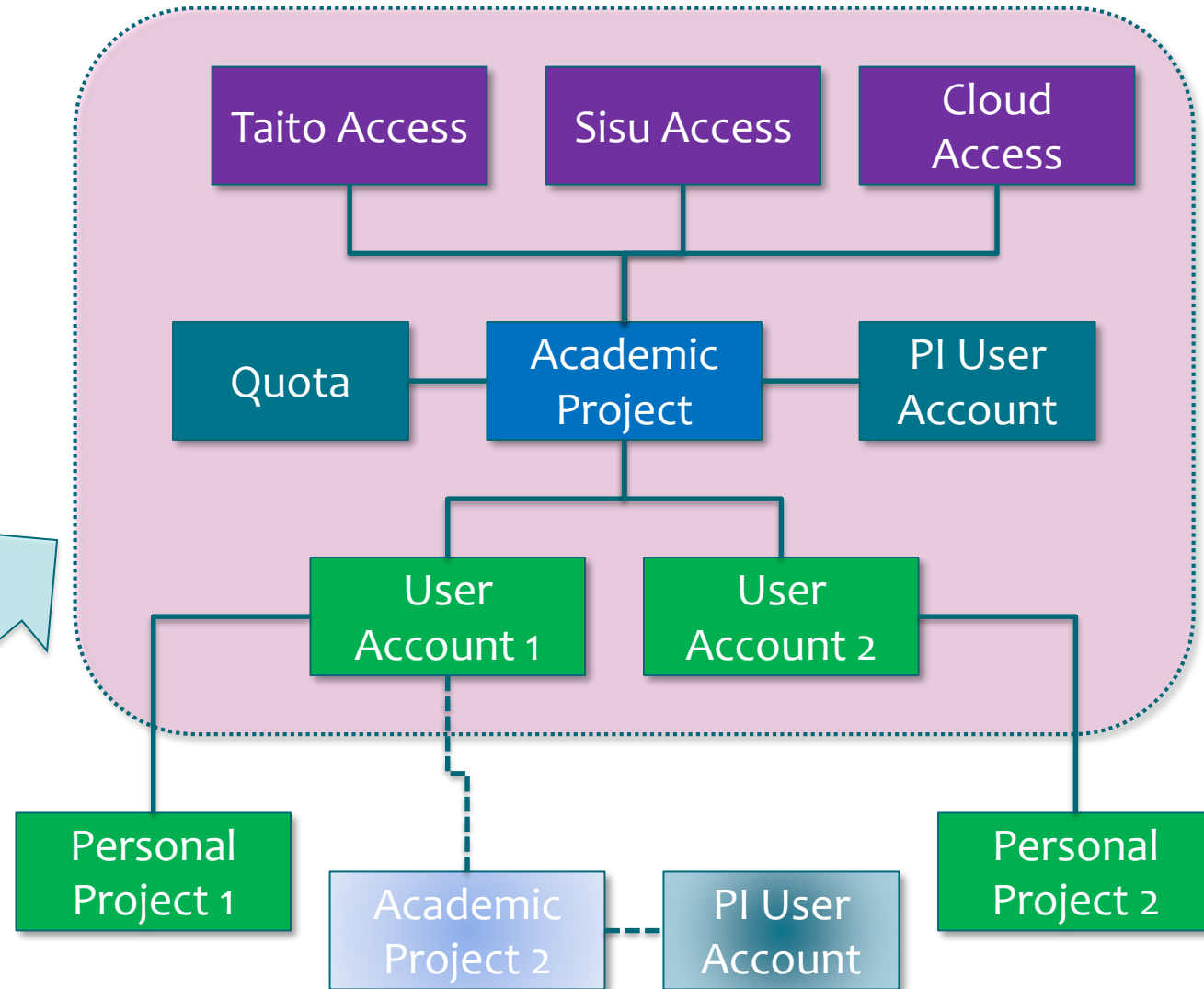
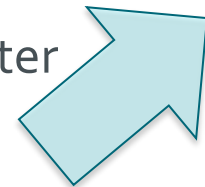
- Sign up as customer
- **Reset your password**
- Manage your account
- Apply for an Academic project
- Apply for computing services

- Access your data
- Download material
- Watch videos
- Monitor hosts
- Personalize your use
- Message board
- + more



Getting an Account: The process

- Register to get a User Account
 - You get a *Personal Project*
 - This is not for CPU time
- **Ask PI** to apply for an (or invite to an existing) Academic Project
 - PI logs in SUI and invites by your email
 - Set as an accountable project
 - You can belong to many projects
- **Ask PI** to apply for a Service e.g. Taito cluster access
 - Accept Terms of Use (link via email)
- Resources are managed at Academic Project level
 - Quota, Services, Members per project
- Keep your personal details up to date



1. Register: User account

- <https://research.csc.fi/csc-guide-getting-access-to-csc-services>
- Login via HAKA authentication to SUI <https://sui.csc.fi>
 - There you find the Registration functionality "*Sign Up*"
 - If you don't have a HAKA account, see [web page](#)
- This will get you an initial computing quota
 - Sending computation job consumes processor cores
 - User gets a Personal Project with 1'000 billing units (500 core-hours) and access to Taito cluster.
 - It is just for piloting, not for large jobs and you cannot apply for additional computing quota or services
 - Don't use this project as your "Billing Project"

2. Apply for an Academic Project

- Professors and PIs can apply for an Academic Project.
 1. Login via HAKA authentication to SUI <https://sui.csc.fi>
 2. From eService menu Resources and Applications tool
 3. Fill the application form for the Academic project
- <https://research.csc.fi/csc-guide-projects-and-resource-allocation>
- You will get 10000 Billing Units by default
- PI Invites users to join an Academic Project
 - SUI → My Projects → Select Project → Edit Project → Invite Member

3. Apply access for a Service

- Only an Academic Project (its PI) can apply access to Service *i.e.* not a Personal Project
- Principal Investigator of an Academic Project can apply for access to Sisu, cPouta and IDA storage Services in SUI
 - <https://sui.csc.fi/group/sui/resources-and-applications>
- In SUI's menu: eService – Resources and Applications

Data analytics and machine learning



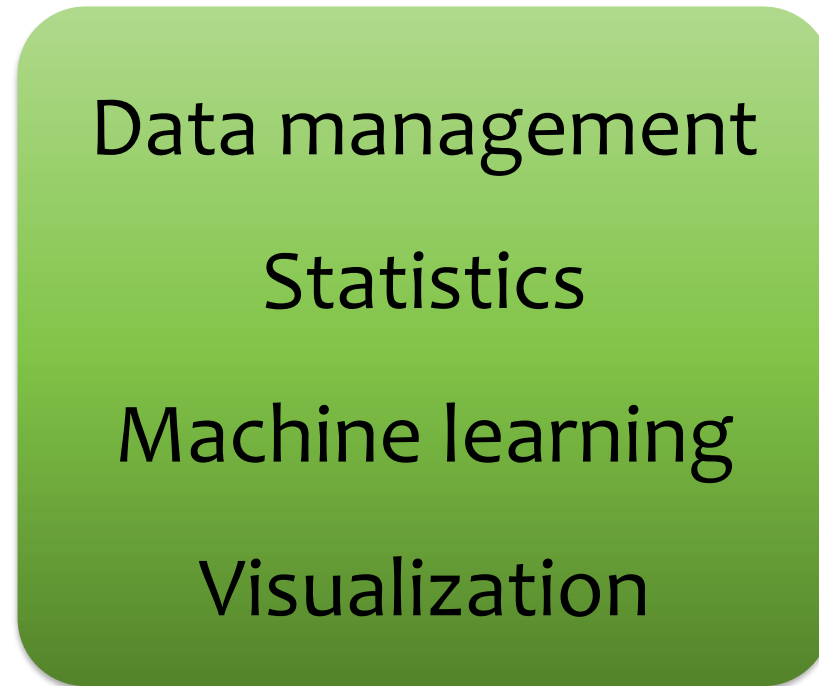
Data analytics at CSC

- Traditionally CSC has been the hub for high-performance computing in Finland
- In previous ~5 years importance of data driven services has increased rapidly
 - Big data and data science activities for research
 - Major data storage provider (scientific data, media archives)
 - Business intelligence and data analytics services for the public sector
- Data analytics group established in 2014

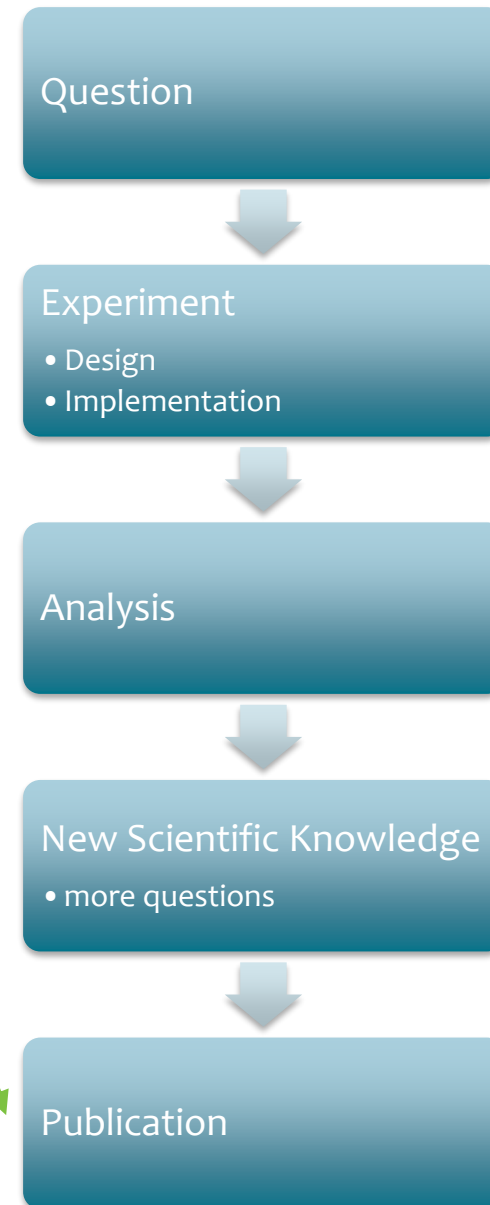
Data analytics services

- Analytics software (R, Python, Spark, Hadoop, ...)
- Web interface for analytics (Notebooks, <https://notebooks.csc.fi/>)
- Software suites (Chipster, Mylly)
- Machine learning environments (TensorFlow, Theano, Keras, Torch, Caffe, ...)

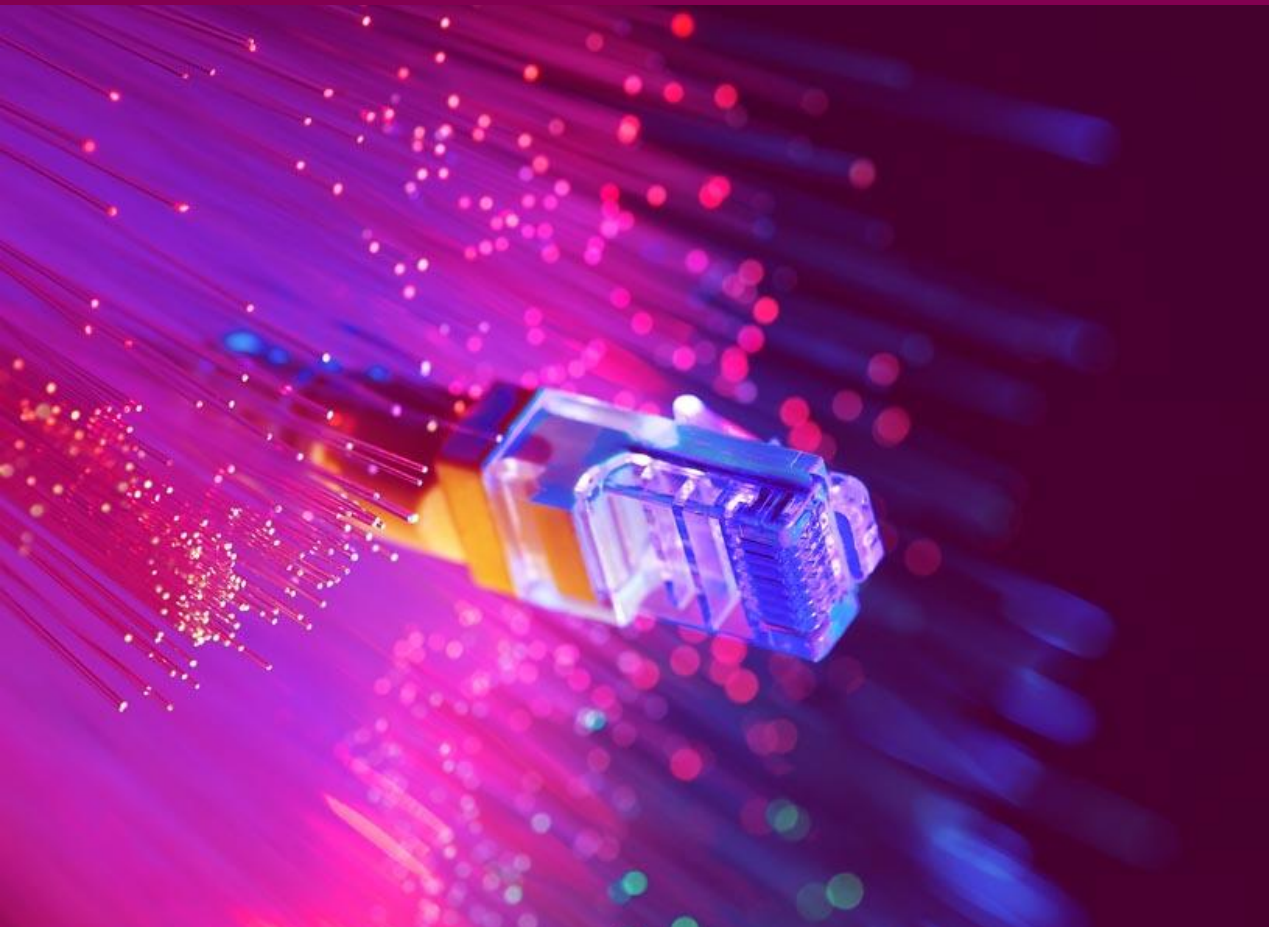
CSC expert support



[mailto: servicedesk@csc.fi](mailto:servicedesk@csc.fi)



GPU computing

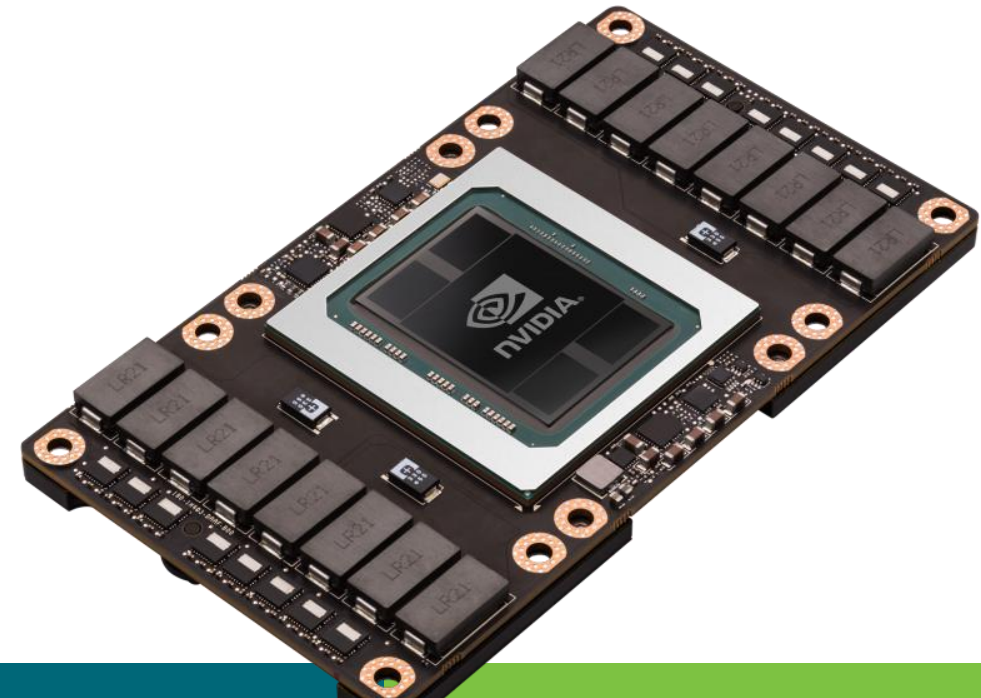


Graphics processing units (GPUs)

- The free lunch is over (for CPUs)
- CPUs are optimized for latency whereas GPUs are optimized for throughput
- CSC's new GPU servers:

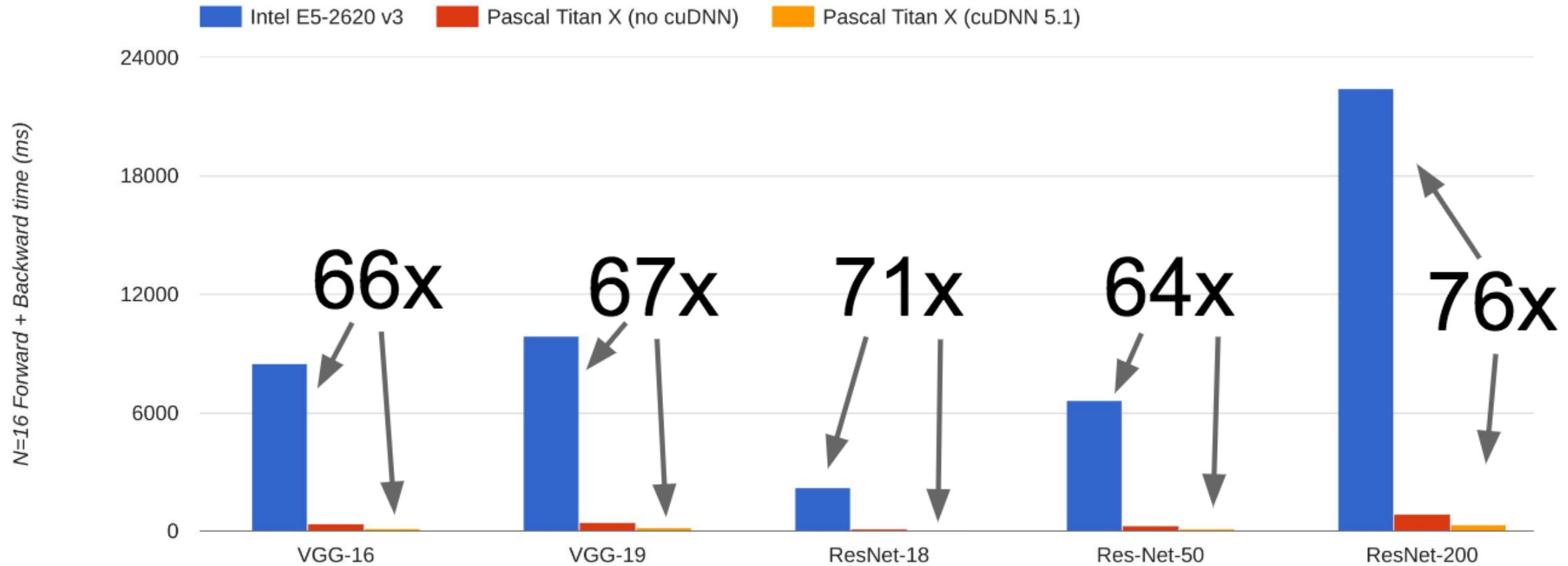


	#cores	clock speed	memory
2 x Xeon CPUs	2 x 14	2.40 / 3.30 GHz	512 GB
4 x P100 GPUs	4 x 3584	1.328 / 1.48 GHz	4 x 16 GB



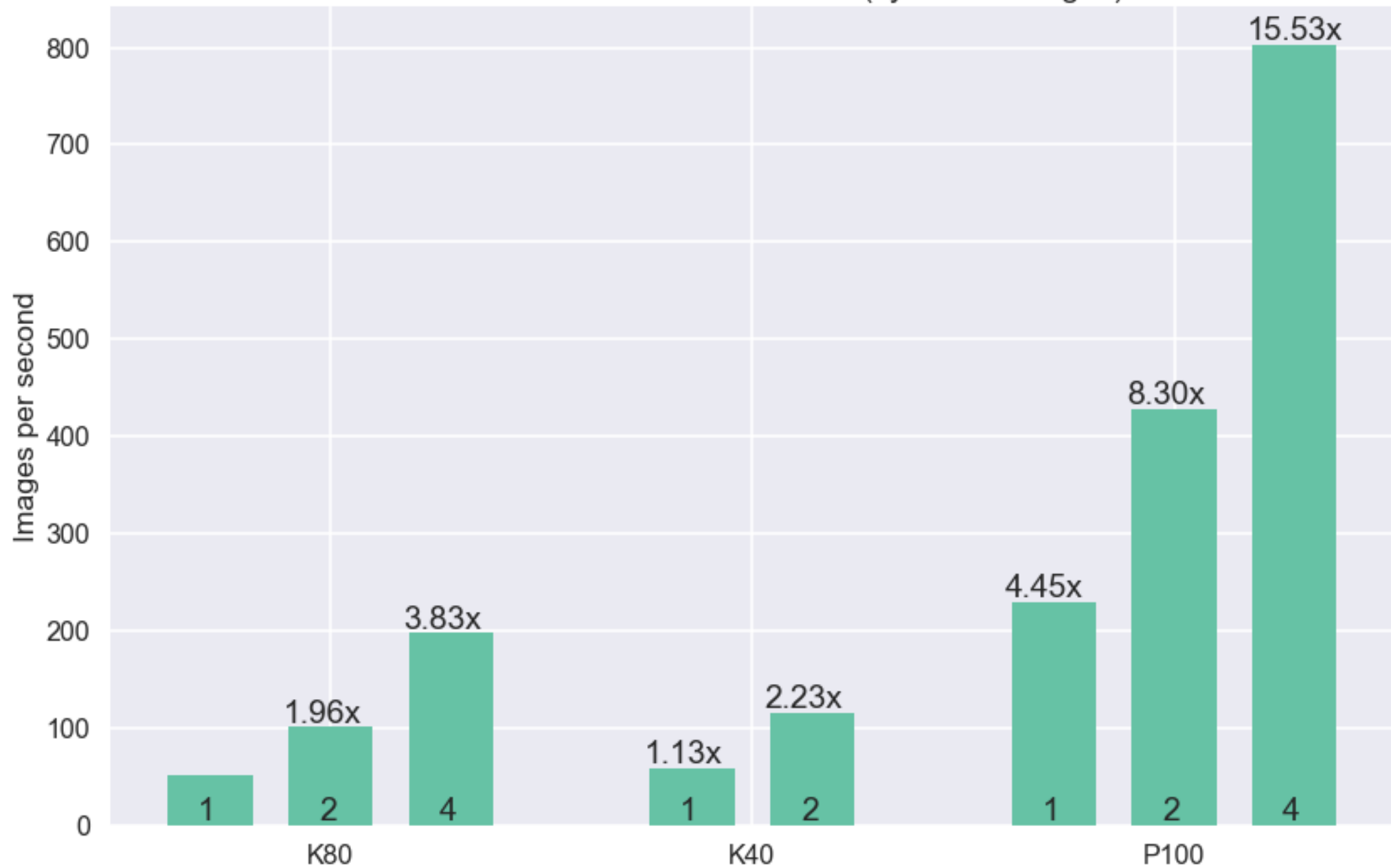
CPU vs GPU in practice

(CPU performance not well-optimized, a little unfair)



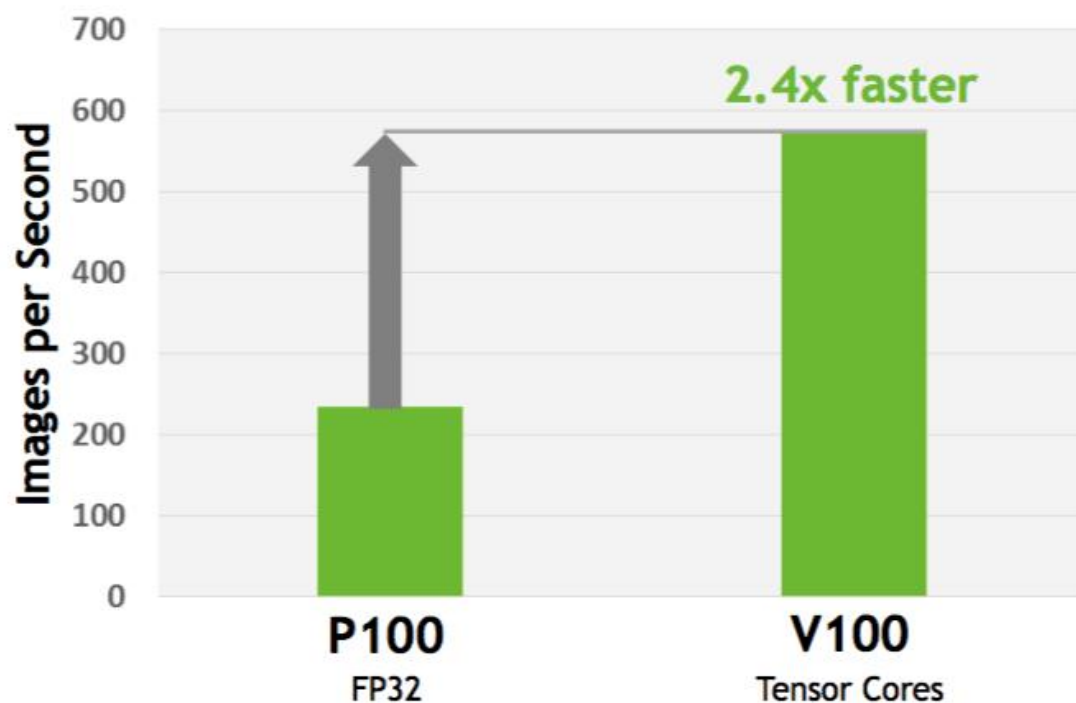
Data from <https://github.com/jcjohnson/cnn-benchmarks>

TensorFlow resnet50 benchmark (synthetic images)



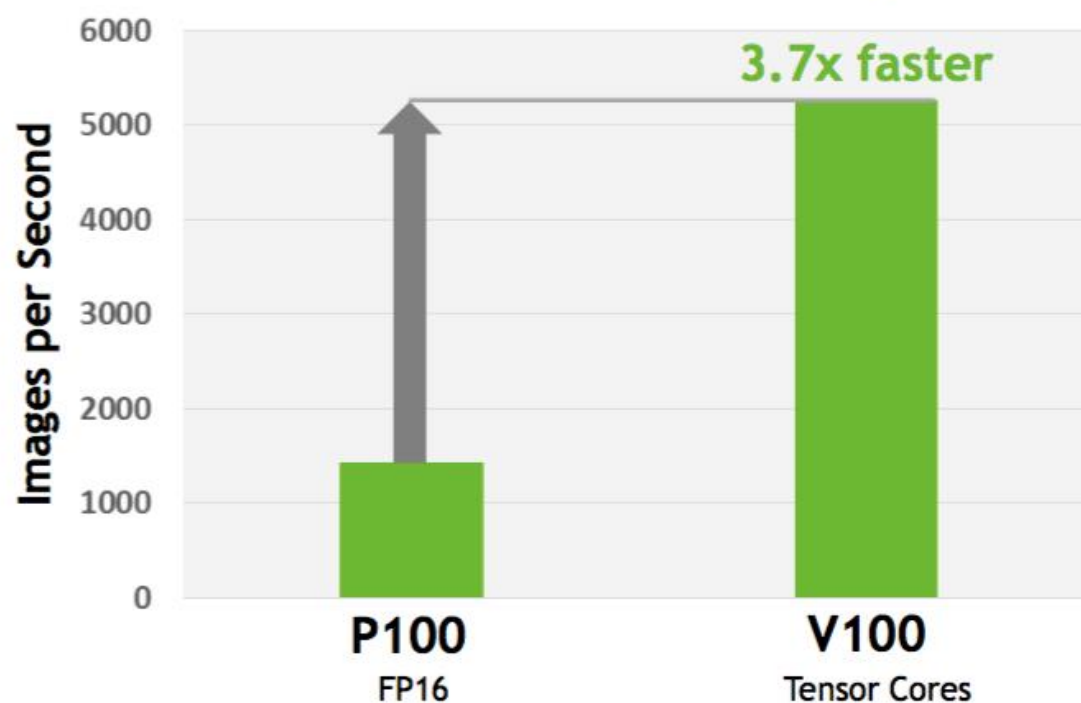
VOLTA: A GIANT LEAP FOR DEEP LEARNING

ResNet-50 Training



ResNet-50 Inference

TensorRT - 7ms Latency



Running batch jobs at CSC

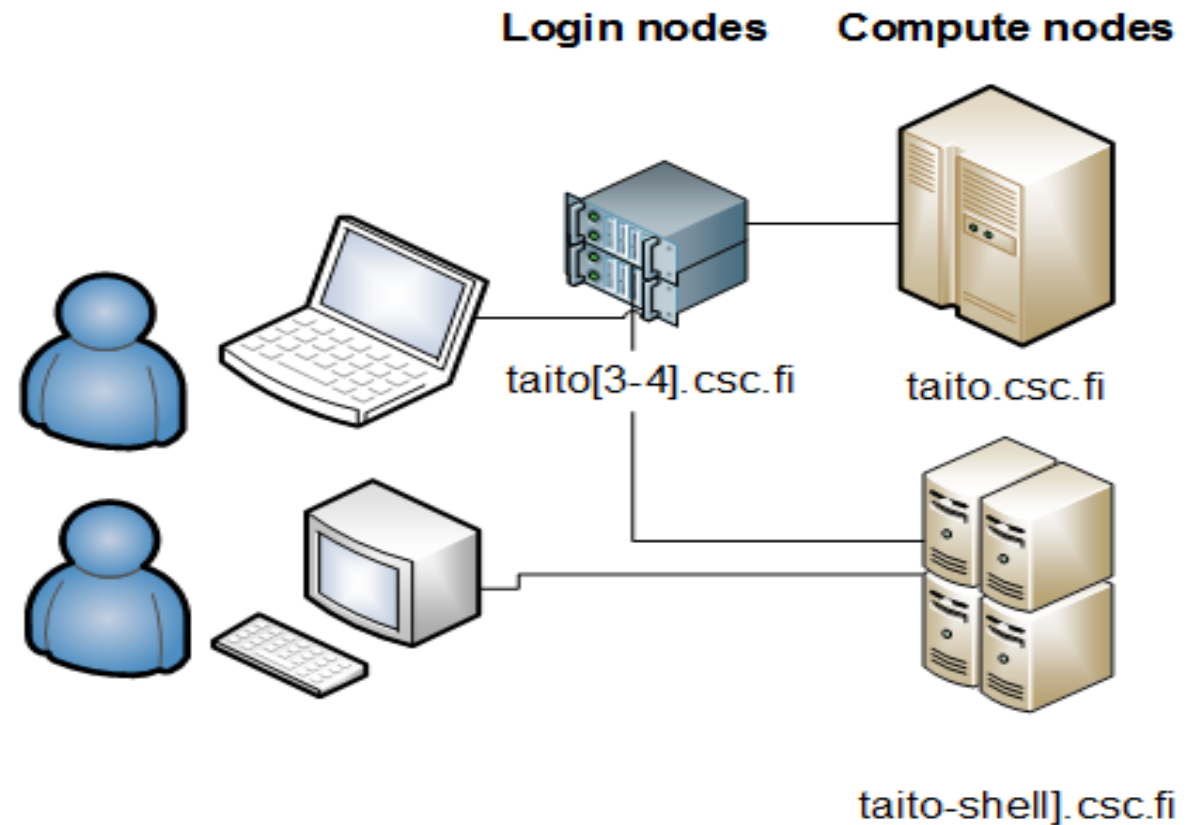


Compute nodes are used via queuing system

Do not use login nodes for heavy computation!

```
$ sbatch job_script.sh
```

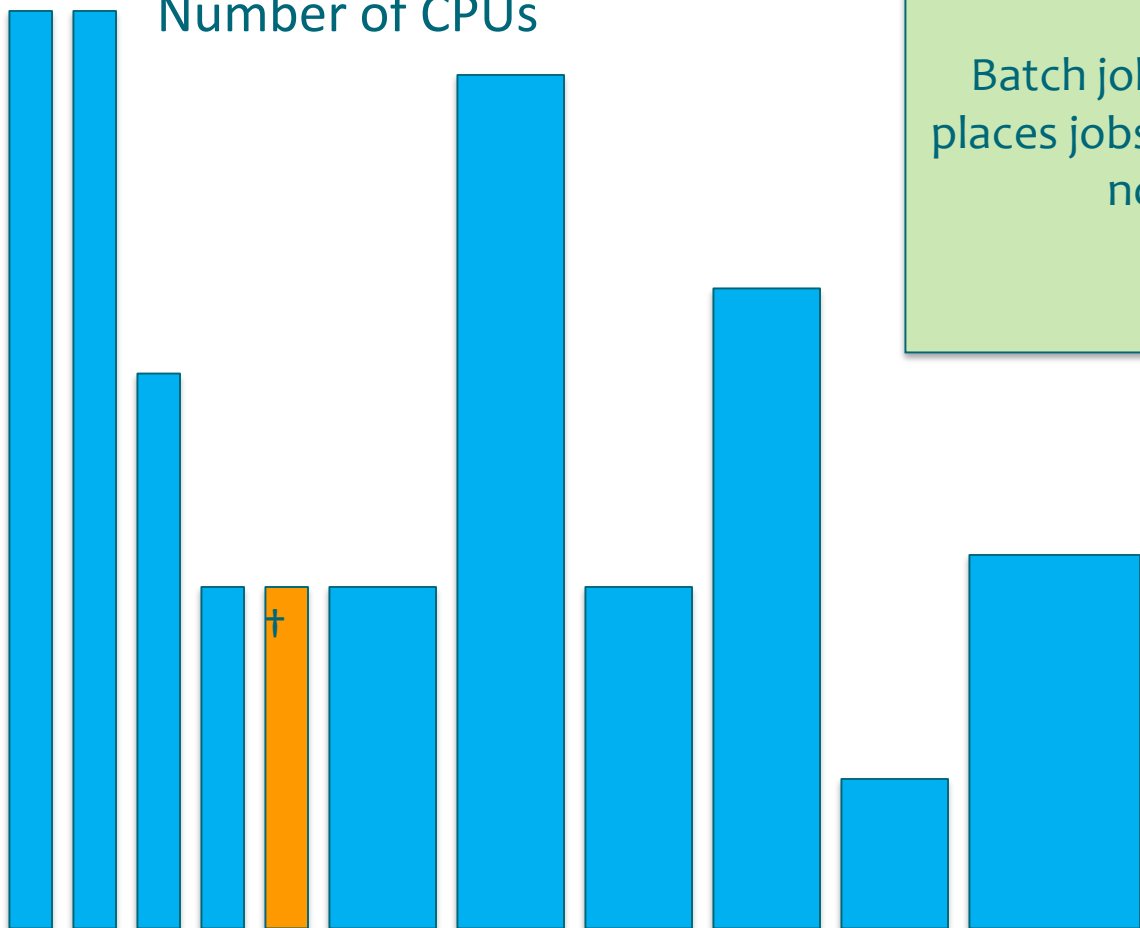
```
$ ./my_prog &
```



Batch system

time

Number of CPUs



Batch job scheduler
places jobs on compute
nodes

Compute
node 1

Compute
node 2

Compute
node 3

† Big memory demand

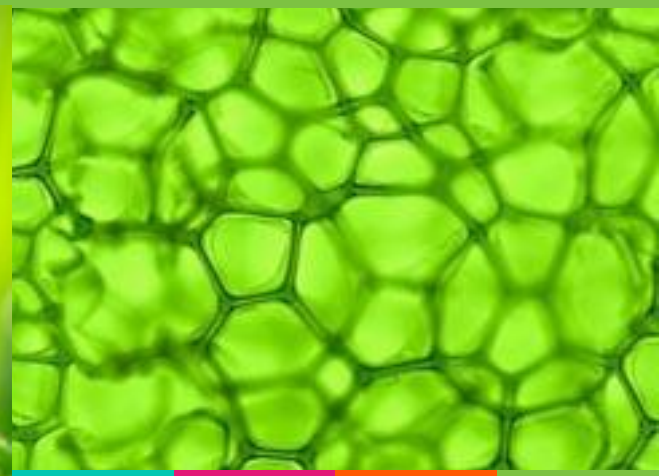
Individual batch jobs

Example serial batch job script on Taito

```
#!/bin/bash -l
#SBATCH -J myjob
#SBATCH -e myjob_err_%j
#SBATCH -o myjob_output_%j
#SBATCH --mail-type=END
#SBATCH --mail-user=a.user@foo.net
#SBATCH --mem-per-cpu=2000
#SBATCH -t 02:00:00
#SBATCH -n 1
#SBATCH -p serial
#SBATCH --constraint=snb

module load myprog
srun myprog -option1 -option2
```

Notebooks



Upcoming: notebooks.csc.fi

- Web-based interface to data analytics tools
 - Jupyter Notebooks (Python), R Studio, Spark, TensorFlow
- User authentication using Finnish university accounts (Haka)
 - **no CSC account needed!**
- Especially for teaching, self-studying, and running small analyses
- Low barrier to entry service for data analysis services and tools
- GPU support planned
- Pilot system used on this course: **pb.csc.fi**

Notebooks

A tool for provisioning ephemeral private cloud resources.

haka

Login

Login

Sign in

If you don't have access to your password, you can [reset it here](#)

Blueprints

Jupyter Datascience (ephemeral) - System.default

Description: This is the ephemeral data science blueprint that will lose it's contents after it is destroyed.

Lifetime: 5h 0m

➤ Launch new

Jupyter ML Blueprint - System.default

Description: Jupyter with Python Machine Learning Libraries: scikit-learn, scipy, Theano, Keras, pandas. This resource is ephemeral. All data on it will be destroyed after the resource expires. Download your results! Apurva do we sanitize our inputs?

Lifetime: 3h 0m

State	Name	Lifetime	Access	Actions
-------	------	----------	--------	---------



pb-terry-
the-black

2h 59m

Open in
browser

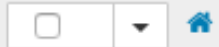
✕ Destroy

➤ Launch new



[Files](#)[Running](#)[Clusters](#)

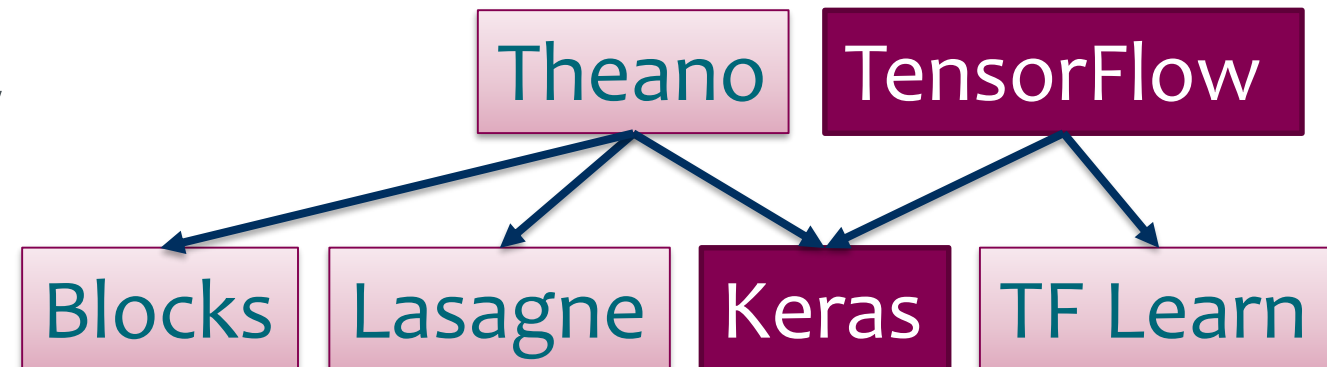
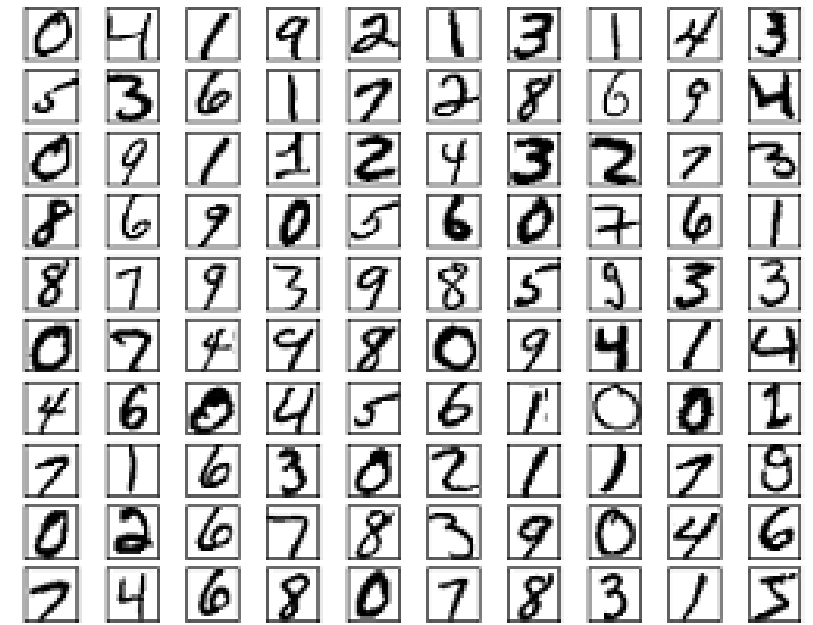
Select items to perform actions on them.

[Upload](#)[New ▾](#)

Notebook list empty.

Hands-on introduction to Notebooks

- Setup: Python notebook in a VM using Pouta Blueprints
 - Go to <https://pb.csc.fi/> and login using HAKA
 - Launch new “Jupyter ML testing (EXPERIMENTAL)” notebook
 - Wait for “Open in browser” to appear, click it when it does
 - Open the notebook “**keras-mnist-mlp.ipynb**”
- Overview
 - Keras (<https://keras.io>) on top of TensorFlow
 - MNIST digit classification
 - MLP (and CNN) training



Hands-on introduction to Taito-GPU

- Go to:

<https://github.com/CSCfi/machine-learning-scripts/tree/master/courses/uefml2017>

- or:

<https://github.com/CSCfi> => machine-learning-scripts => courses => uefml2017

Learning diary

Pros and cons of the Notebooks and Taito-GPU approaches to data analysis?

Focus on your personal perspective, which one do you prefer or would more likely use in your upcoming projects.



CSC – IT Center for Science Ltd

+3589 457 2821 (service requests)
+3589 457 2001 (call center/ contacts)

servicedesk@csc.fi

www.csc.fi



<https://www.facebook.com/CSCfi>



<https://twitter.com/CSCfi>



<https://www.youtube.com/c/CSCfi>



<https://www.linkedin.com/company/csc---it-center-for-science>