Multi-objective Design Exploration and Its Application to Algorithmic Optimisation in Real-time Robot Vision

Luigi Nardi, PhD

Software Performance Optimisation group

@Stanford

January 19th 2017

In collaboration with:

B. Bodin, M Z. Zia, H. Wagstaff, E. Vespa, S. Saeedi, G. S. Shenoy, M. K. Emani, J. Mawer, A. Nisbet, M. Luján, B. Franke, G. Riley, M. F. P. O'Boyle, A. J. Davison, P. H. J. Kelly and S. Furber

MANCHESTER 1824 The University of Manchester

The University of Edinburgh

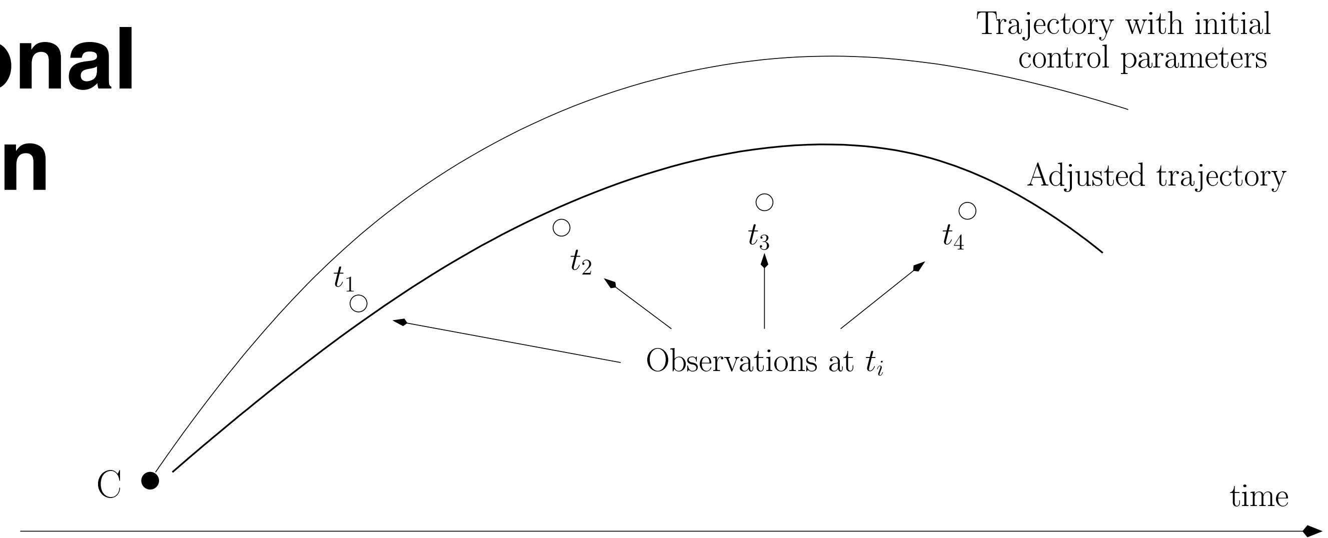
Imperial College London

EPSRC Engineering and Physical Sciences Research Council

PAMELA

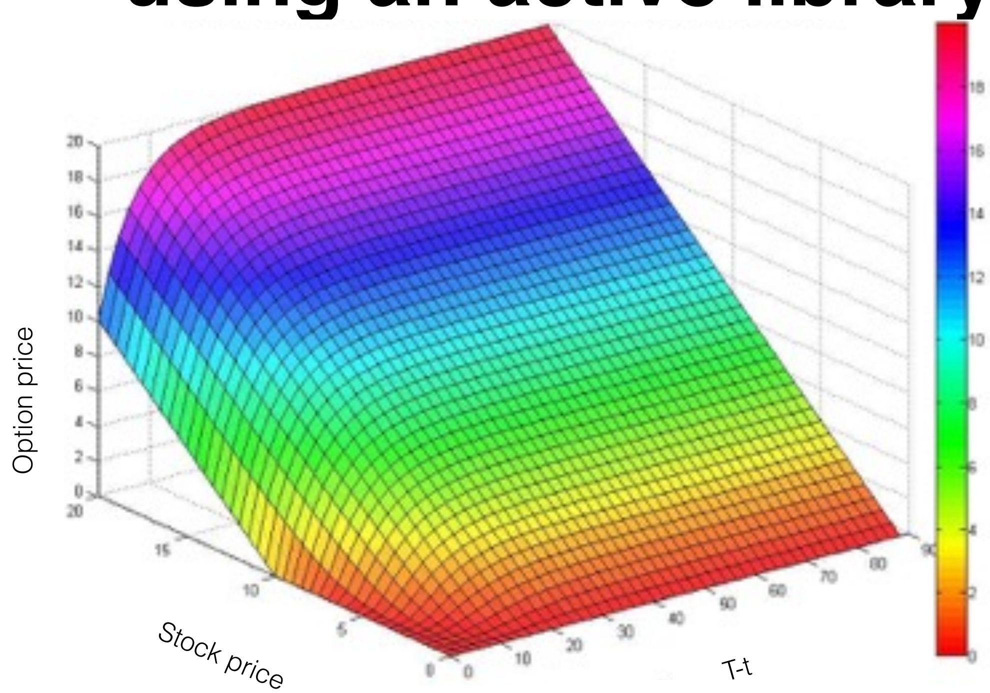
Where I come from: three application domains

1) DSL for Variational data assimilation



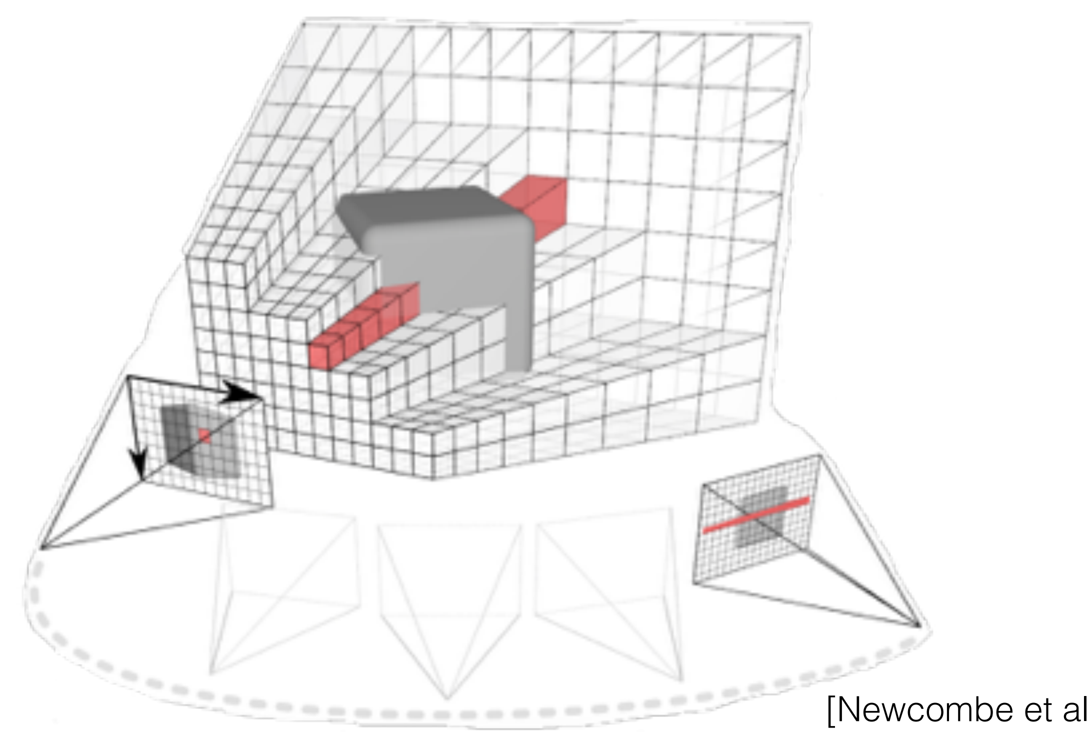
Ph.D. Computer/computational science, LOCEAN and CEDRIC labs

2) HPC computational finance using an active library



Permanent researcher, R&D at Murex S.A.S.
Murex Analytics (MACS) group

3) Computer vision

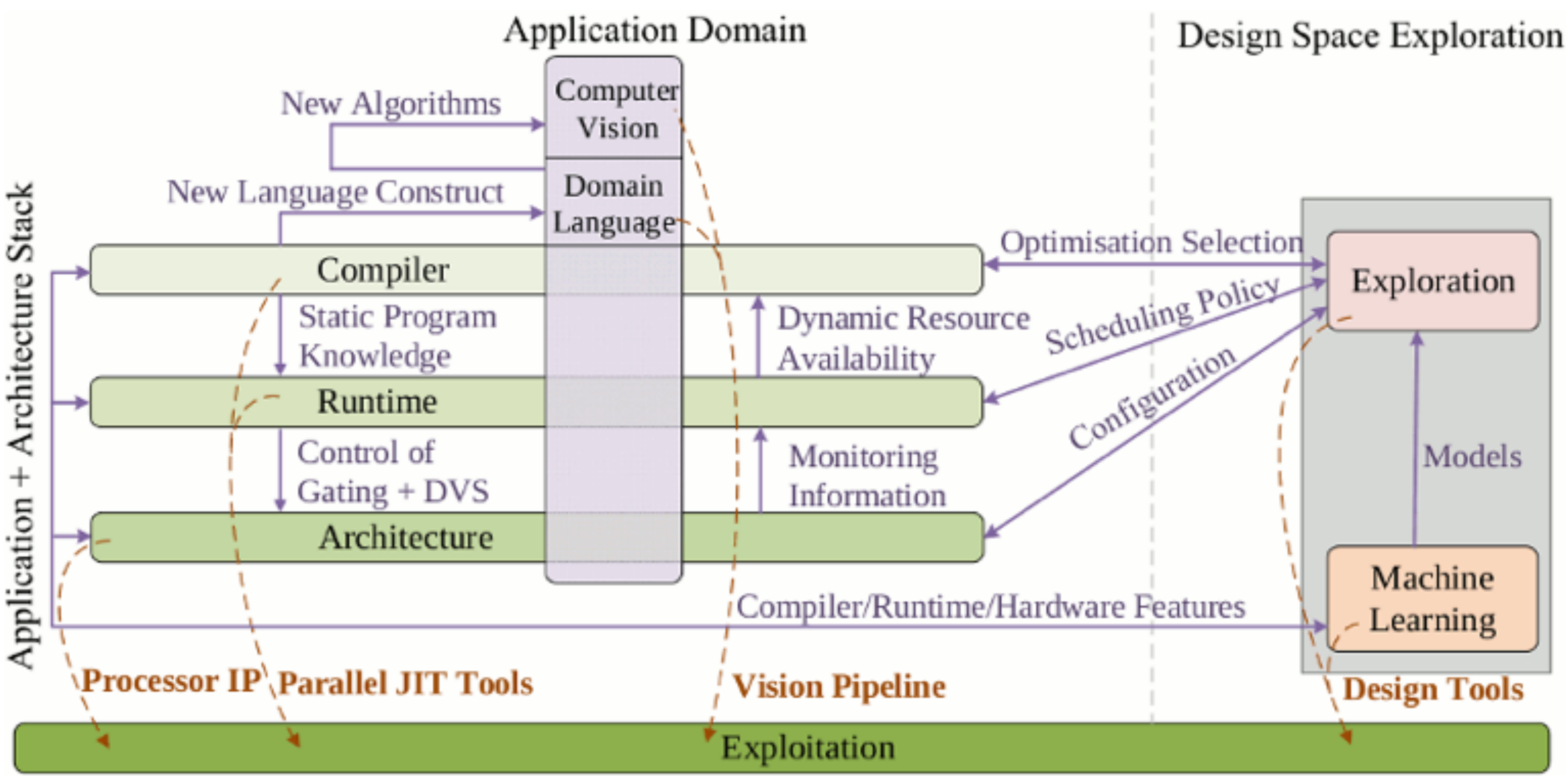


[Newcombe et al., ICCV 2011]

Research Associate, Imperial College London
Software Performance Optimisation group

PAMELA project

Panoramic Approach to the Many-corE LAndscape - from application to end-device: a holistic approach



Main paper behind this talk

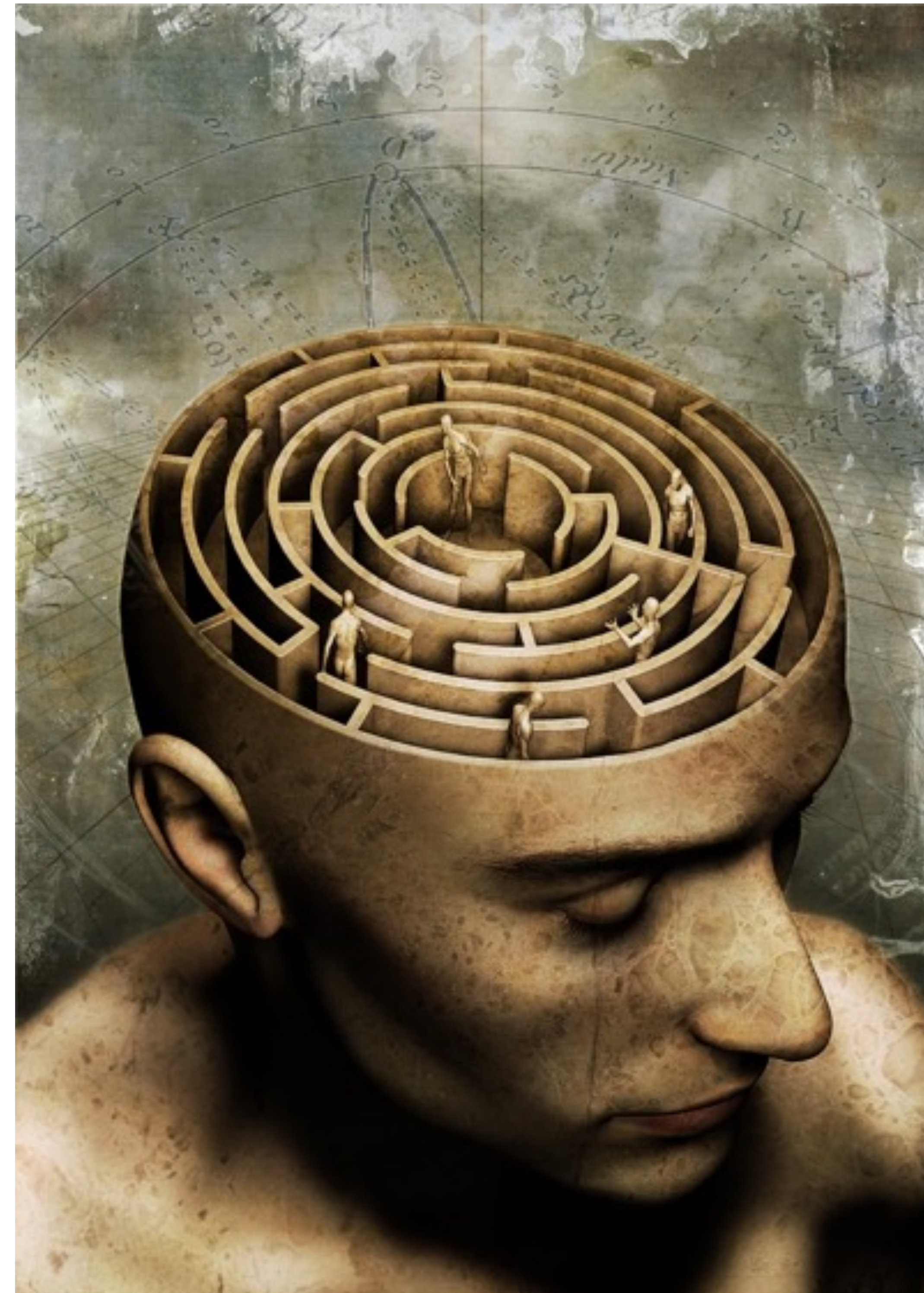
- Haifa, September 2016:
  International Conference on
  Parallel Architectures and Compilation Techniques (**PACT**)

- Paper title:
  Integrating algorithmic parameters into benchmarking and design space exploration in dense 3D scene understanding

- Authors:
  B. Bodin, **L. Nardi**, Zia Zeeshan, H. Wagstaff, G. S. Shenoy, M. Emani, J. Mawer, C. Kotselidis, A. Nisbet, M. Lujan, B. Franke, Paul H. J. Kelly, M. O'Boyle

Outline

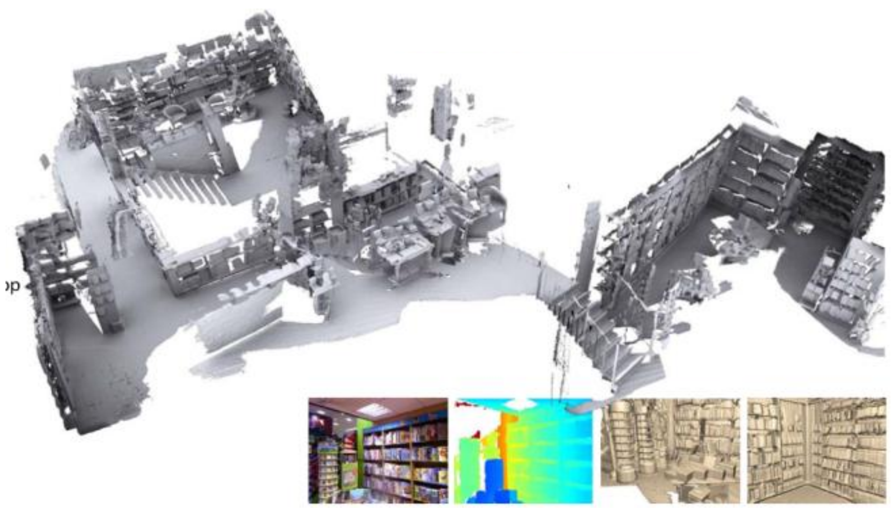
- **The SLAM application, a brief introduction**
- Benchmarking methodology
- Space exploration of algorithmic and implementation design choices



The three R's of vision: Spectrum of Computer Vision Research

Reconstruction

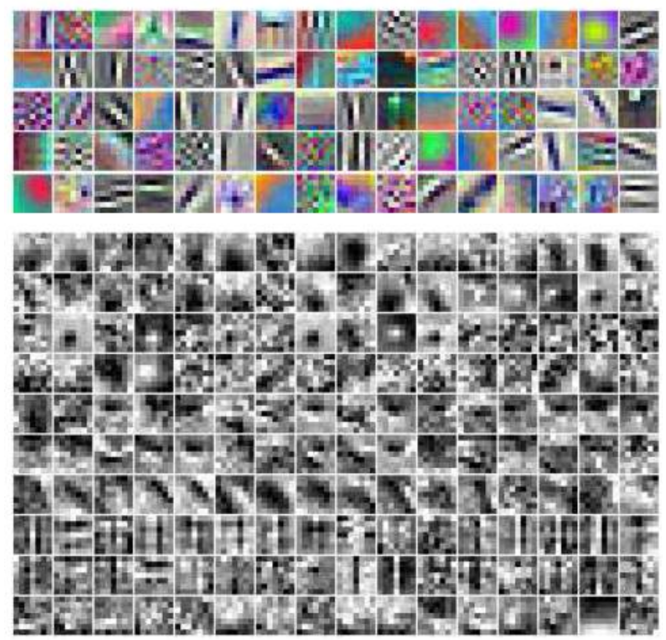
Scalable Kinect Fusion (2013)

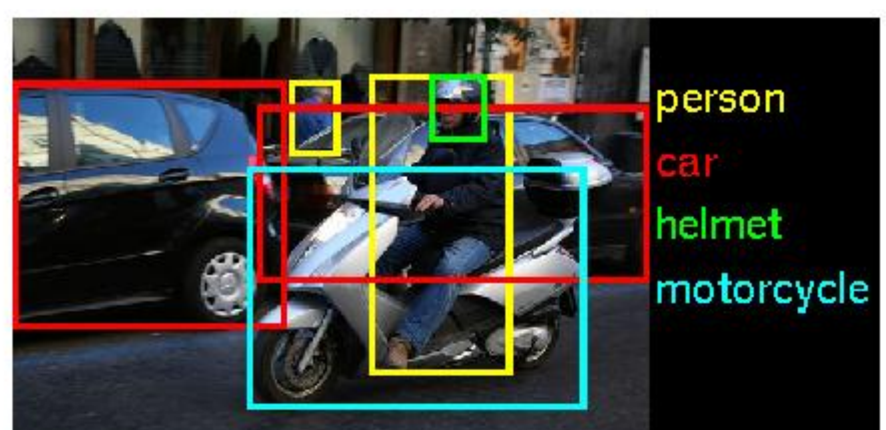


Building Rome on a cloudless day (2010)

Recognition

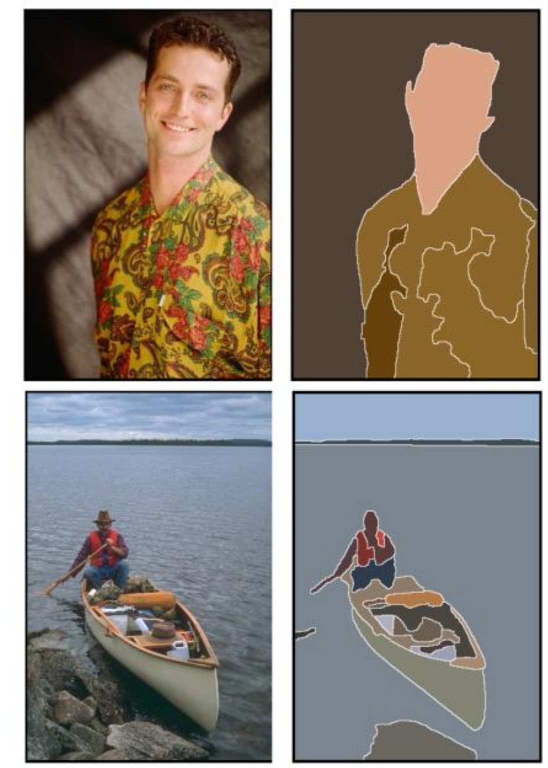
Deep learning for scalable Object class detection (2014)





Reorganisation or Grouping

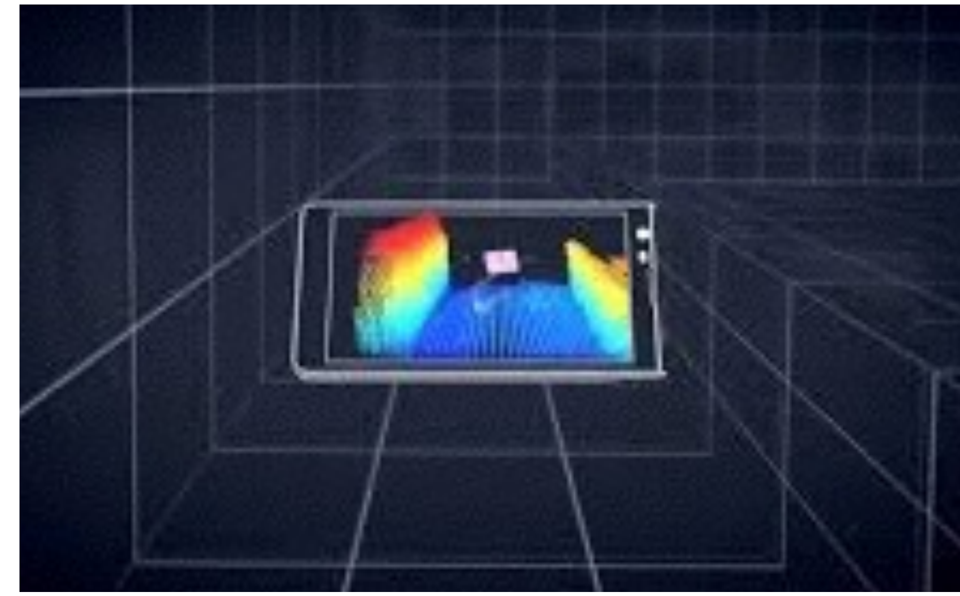
Contour detection and segmentation (2011)



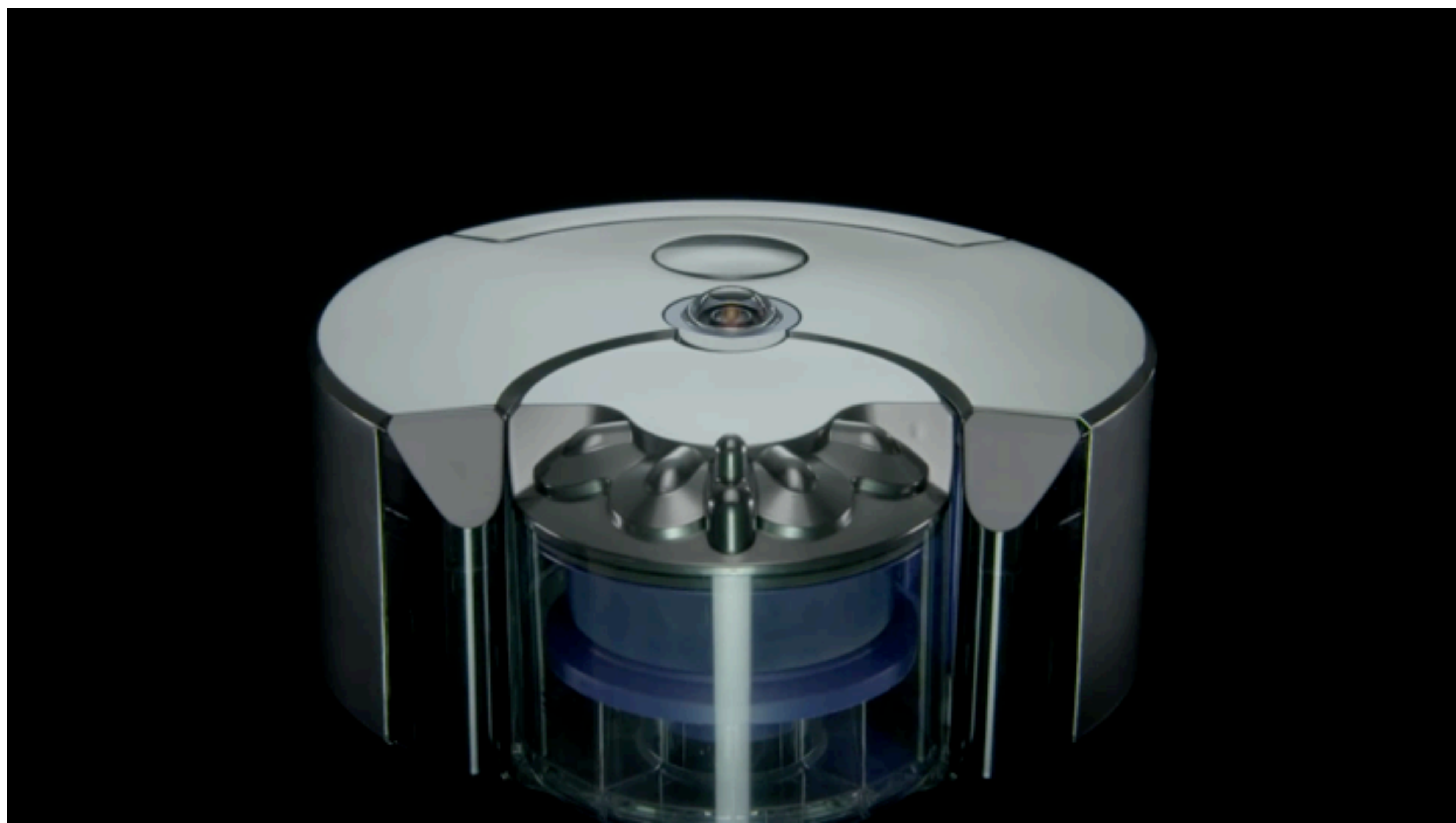
Credit: Zeeshan Zia

Simultaneous localisation and mapping (SLAM)

Build a coherent world representation and localise the camera in real-time



Sparse SLAM



Video:
[Dyson 360 Eye]

Dense SLAM

Video: **KinectFusion**
[Newcombe et al. ISMAR 2011]

Simultaneous localisation and mapping (SLAM)

Build a coherent world representation and localise the camera in real-time

Dense SLAM

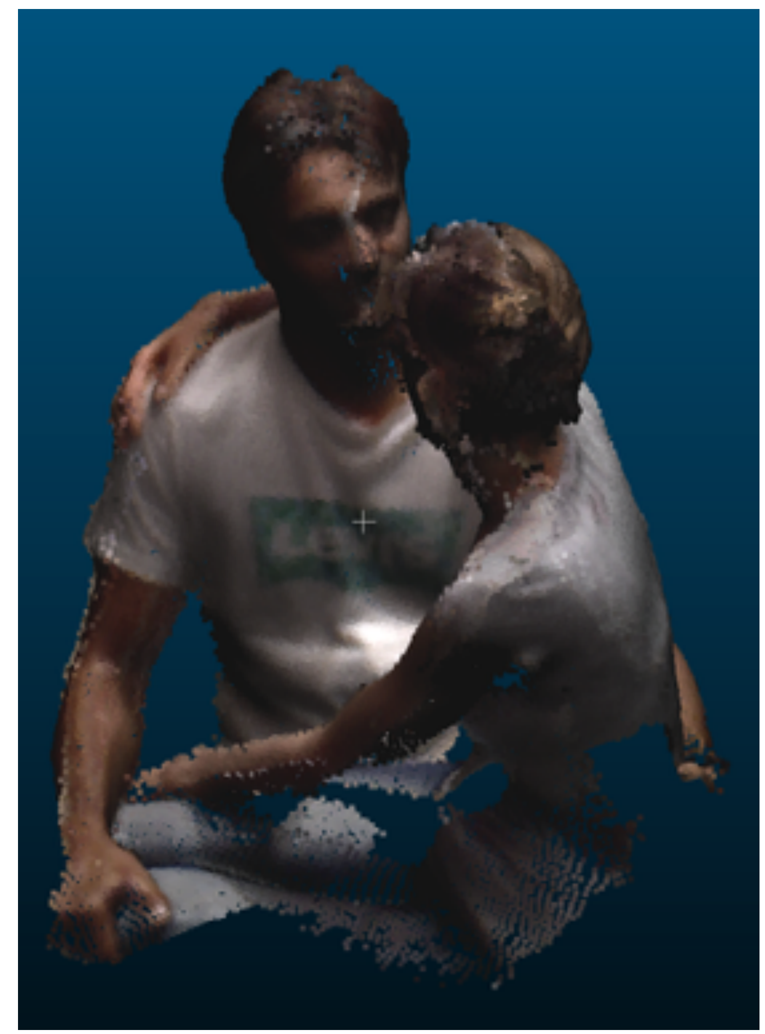
In this talk I will focus on two dense algorithms:

- KinectFusion [Newcombe et al. ISMAR 2011]
- ElasticFusion [Whelan et al. RSS 2015]

Jesse Clayton (NVIDIA)
3D reconstruction

Applications, e.g.:

- Robotics
- Autonomous driving
- 3D printing
- Augmented reality
- Telepresence

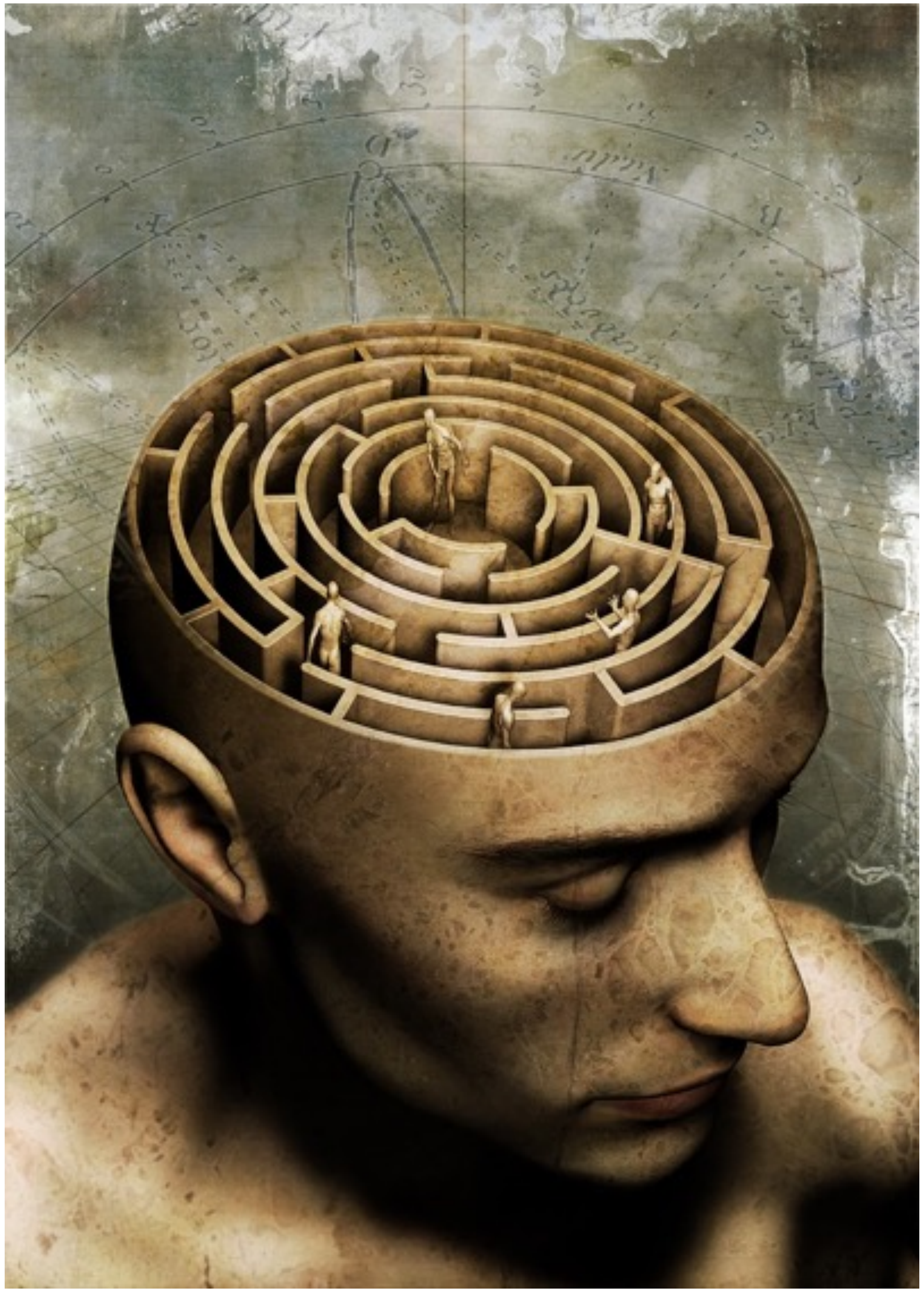


Daniele and Daniela
3D reconstruction

Pictures featured using ElasticFusion

Outline

- The SLAM application, a brief introduction
- **Benchmarking methodology**
- Space exploration of algorithmic and implementation design choices



What is “Performance”?

1. In several domains performance is **execution time**
2. In some domains performance is **accuracy**
3. What about **energy**?
4. But also memory consumption, temperature, robustness, etc.

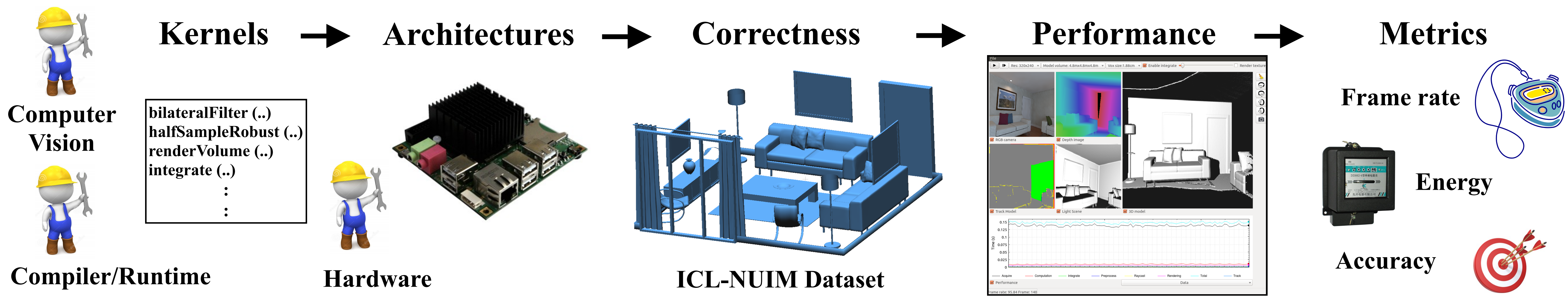
A modern system evaluation considers multiple metrics:

$$Performance = \begin{bmatrix} Runtime \\ Energy \\ Accuracy \end{bmatrix}$$

This defines a multi-objective optimisation problem: trade-off

Holistic approach to SLAM performance:

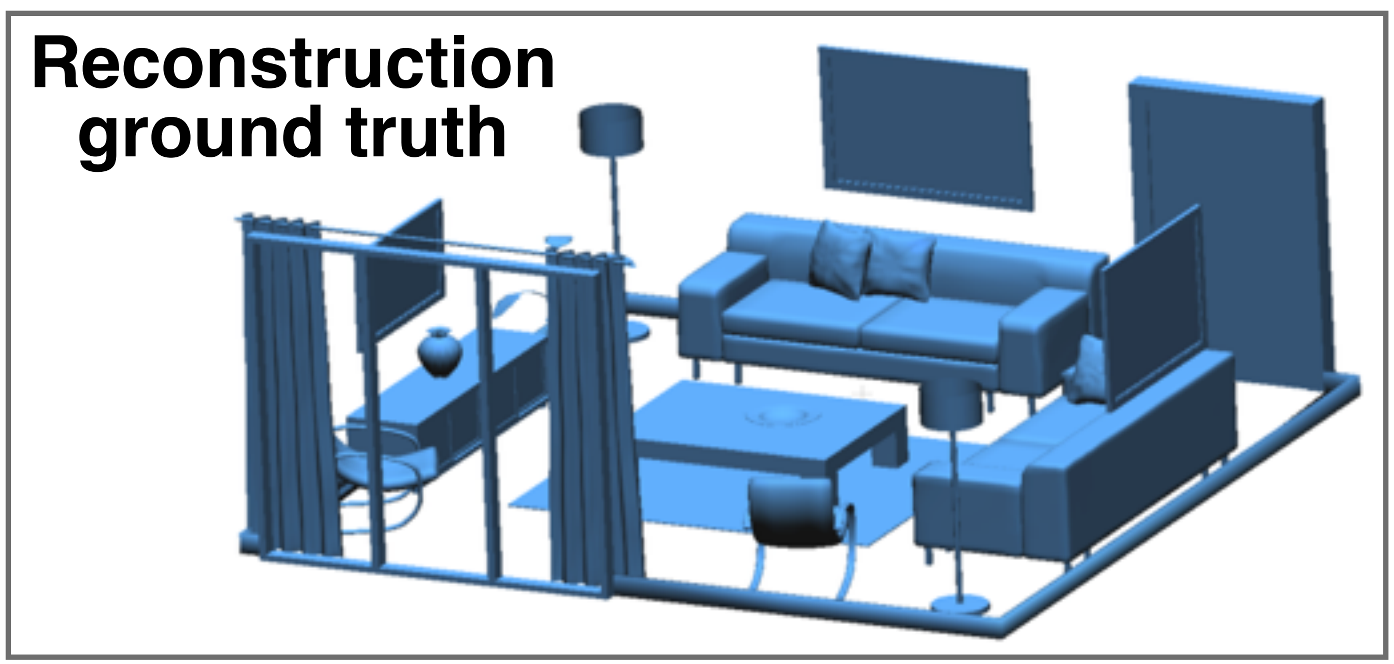
SLAMBench

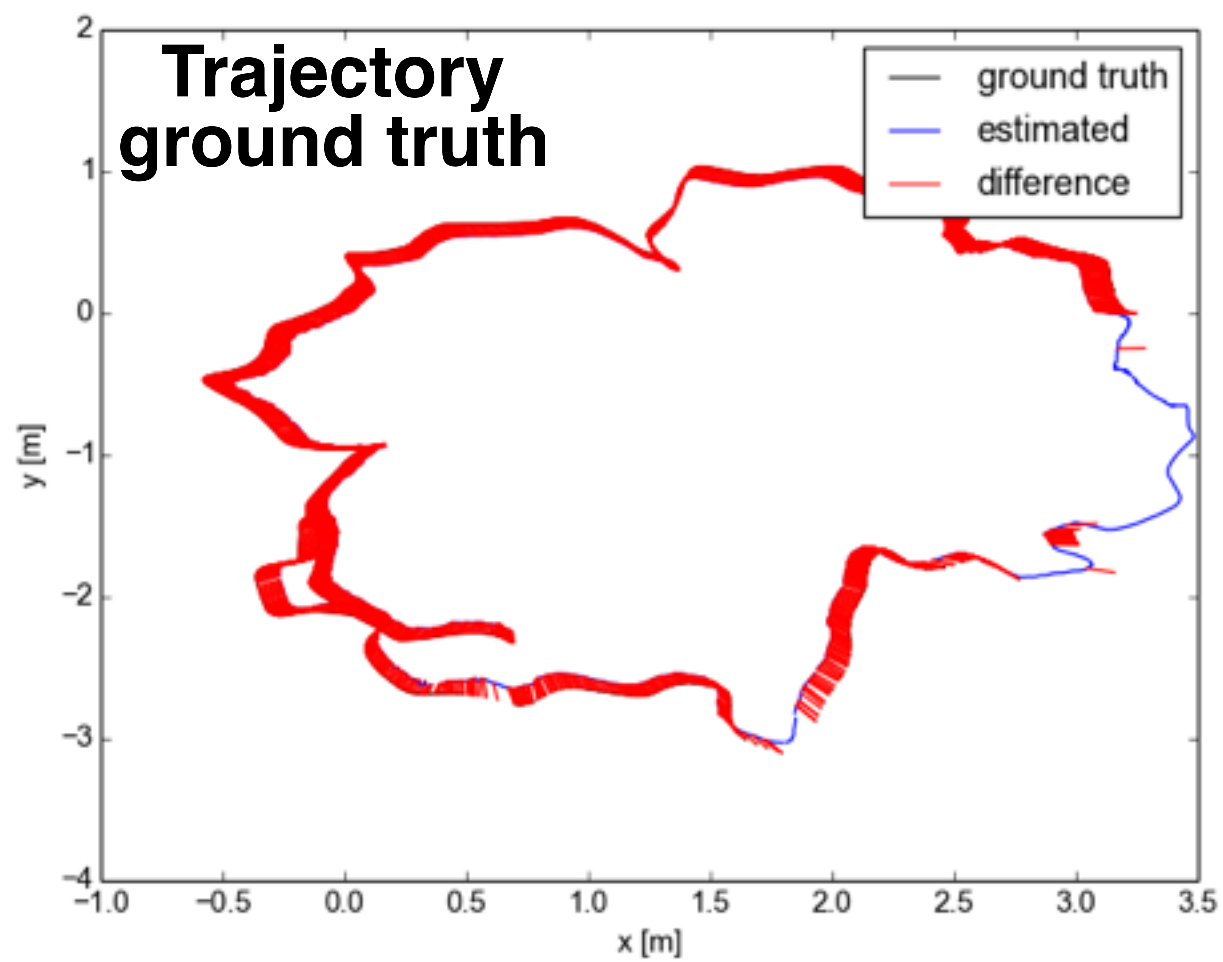


A publicly-available <u>benchmarking framework</u> for quantitative, comparable and validatable experimental research to investigate trade-offs in performance, accuracy and energy consumption of a SLAM system

Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM **(ICRA 2015)**

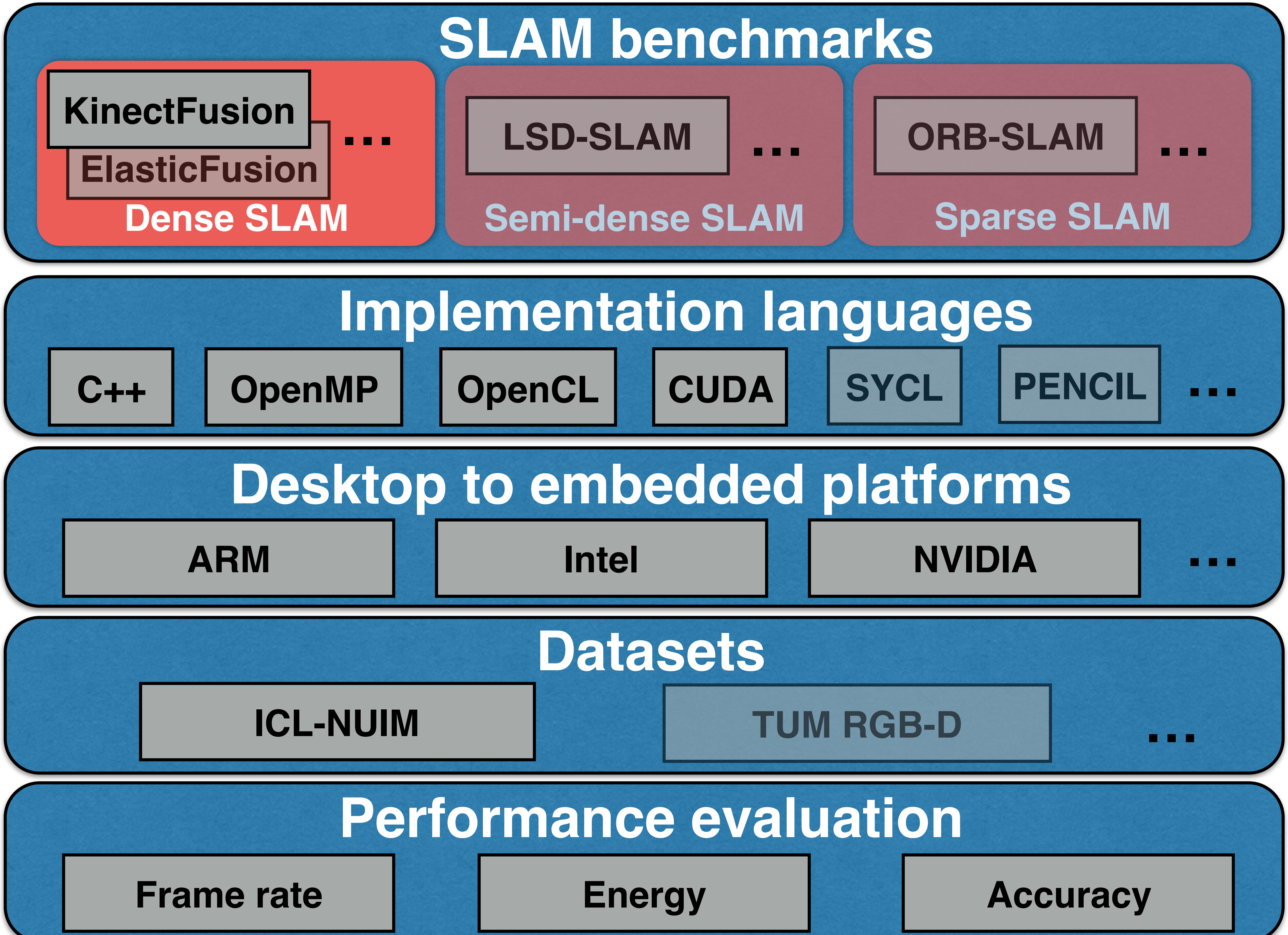
ICL-NUIM dataset





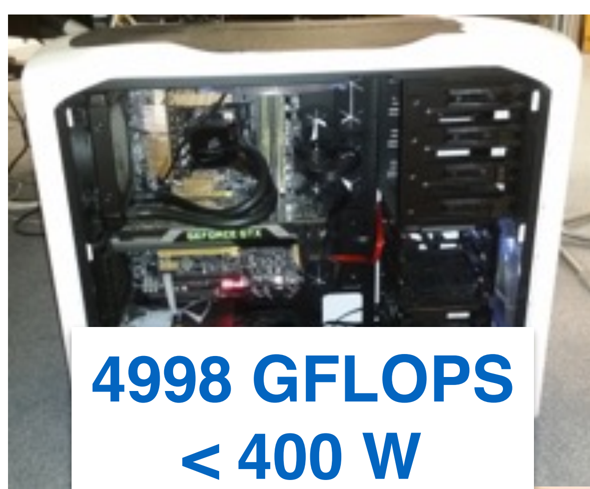
- ICL-NUIM synthetic dataset [Handa et al. 2014]
- 880 RGB-D frames at 30 FPS
- Absolute trajectory error (ATE) based on ground truth

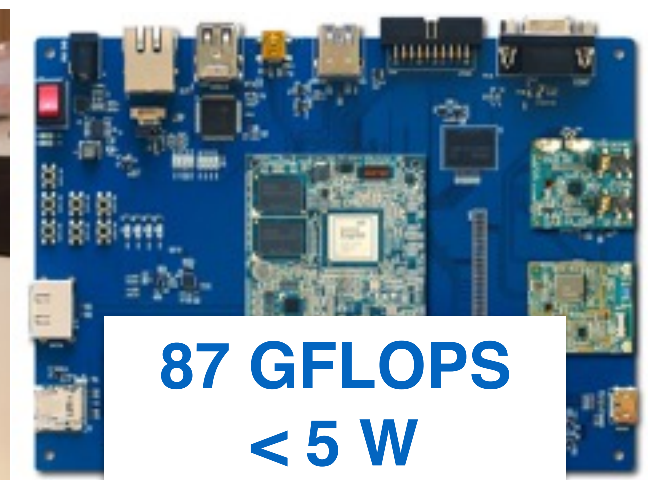
SLAMBench framework



[Nardi et al., 2015]

Machines	TITAN	GTX870M	TK1	ODROID	Arndale
CPU	Intel	Intel	ARM	ARM	ARM
CPU name	i7 Haswell	i7 Haswell	NVIDIA 4-Plus-1	Exynos 5422	Exynos 5250
CPU GFLOPS	448	307	74	80	27
CPU cores	4	4	4 + 1	4 + 4	2
GPU	NVIDIA	NVIDIA	NVIDIA	ARM	ARM
GPU name	TITAN	GTX 870M	Tegra K1	Mali-T628	Mali-T604
GPU GFLOPS	4500	2520	330	60 + 30	60





Platforms

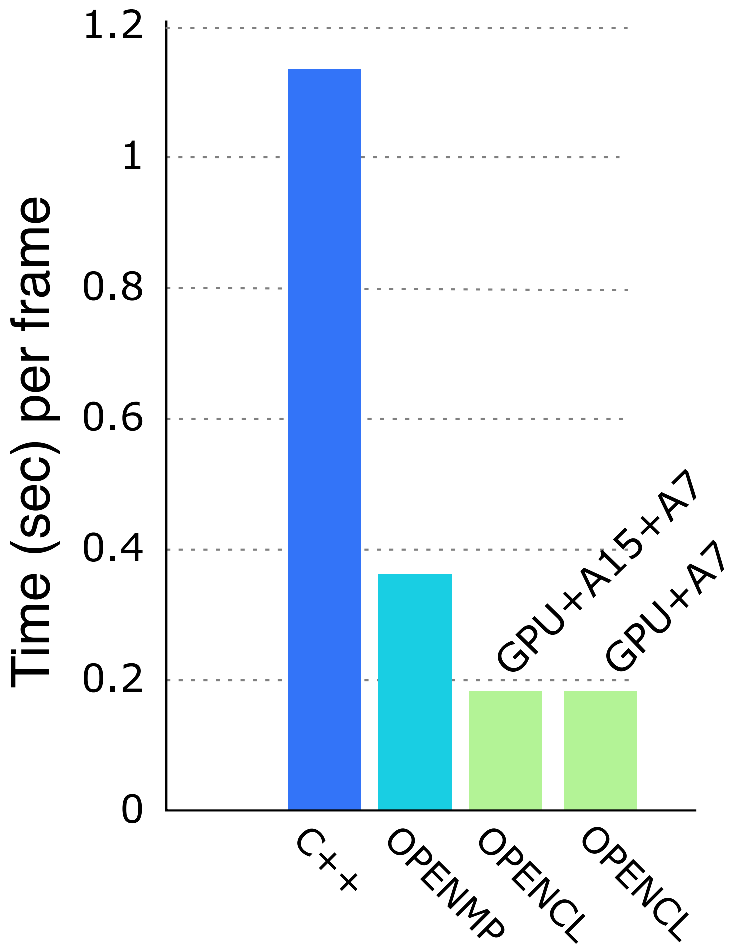
[Nardi et al., 2015]

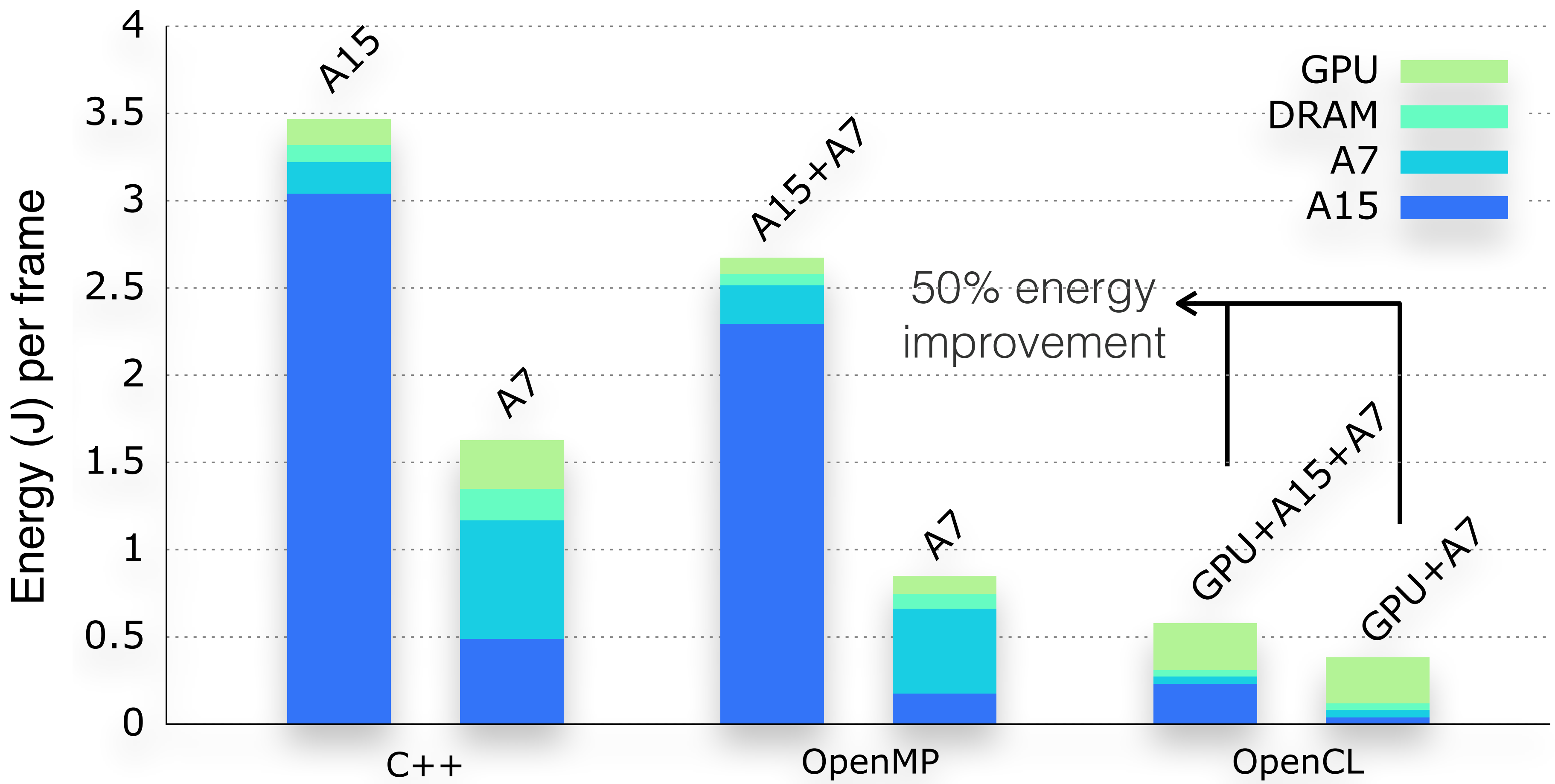
"Performance" on SLAMBench

- Runtime/energy/accuracy measurements
- Accuracy provided via absolute trajectory error (ATE)

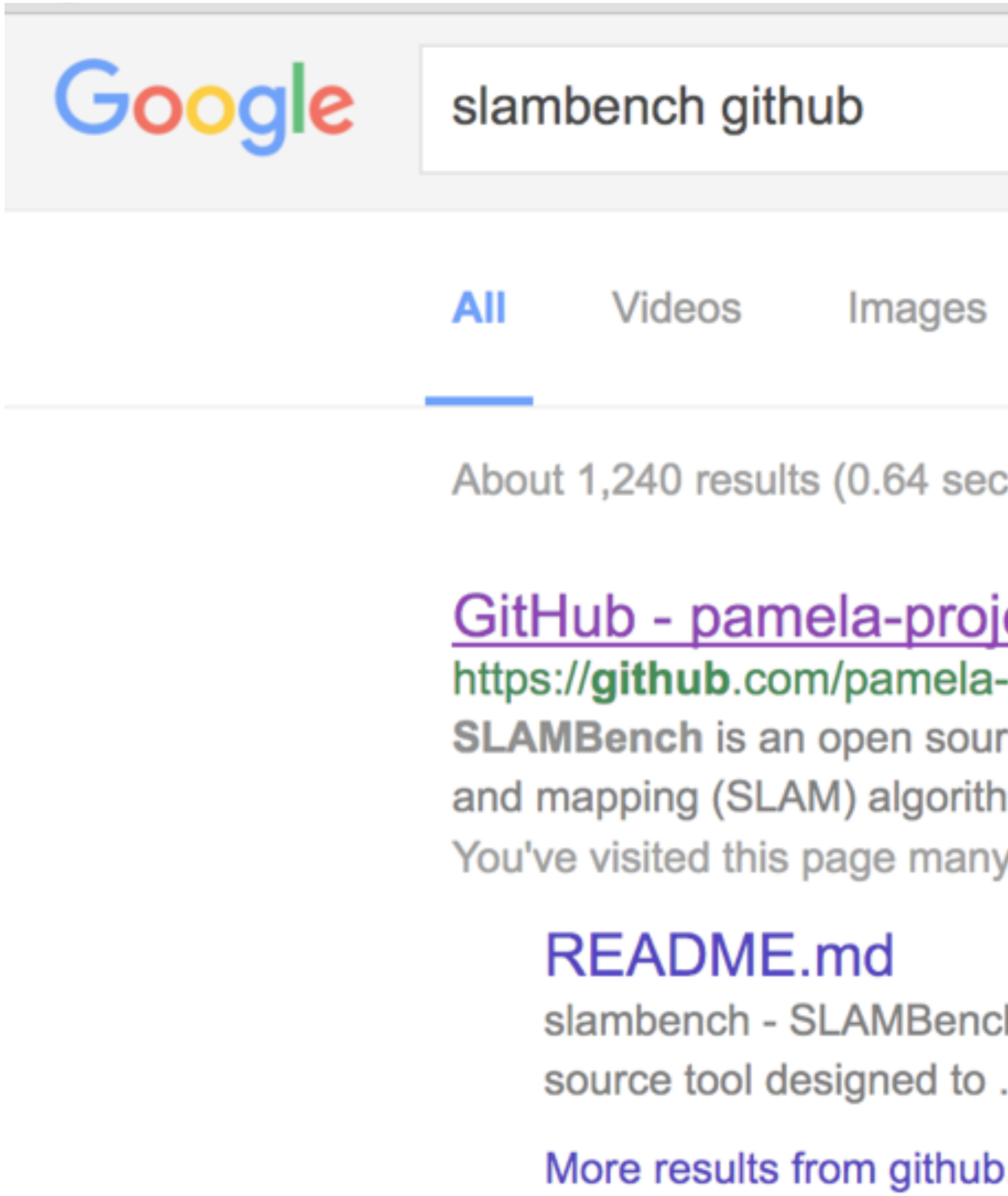
ATE in cm	
C++	2.06
OpenMP	2.06
OpenCL	2.01

Machine	CPU	CPU name	CPU GFLOPS	CPU cores	GPU	GPU name	GPU GFLOPS	TDP Watts
Hardkernel ODROID-XU3	ARM A15 + A7	Exynos 5422	80	4 + 4	ARM	Mali-T628	60 + 30	10



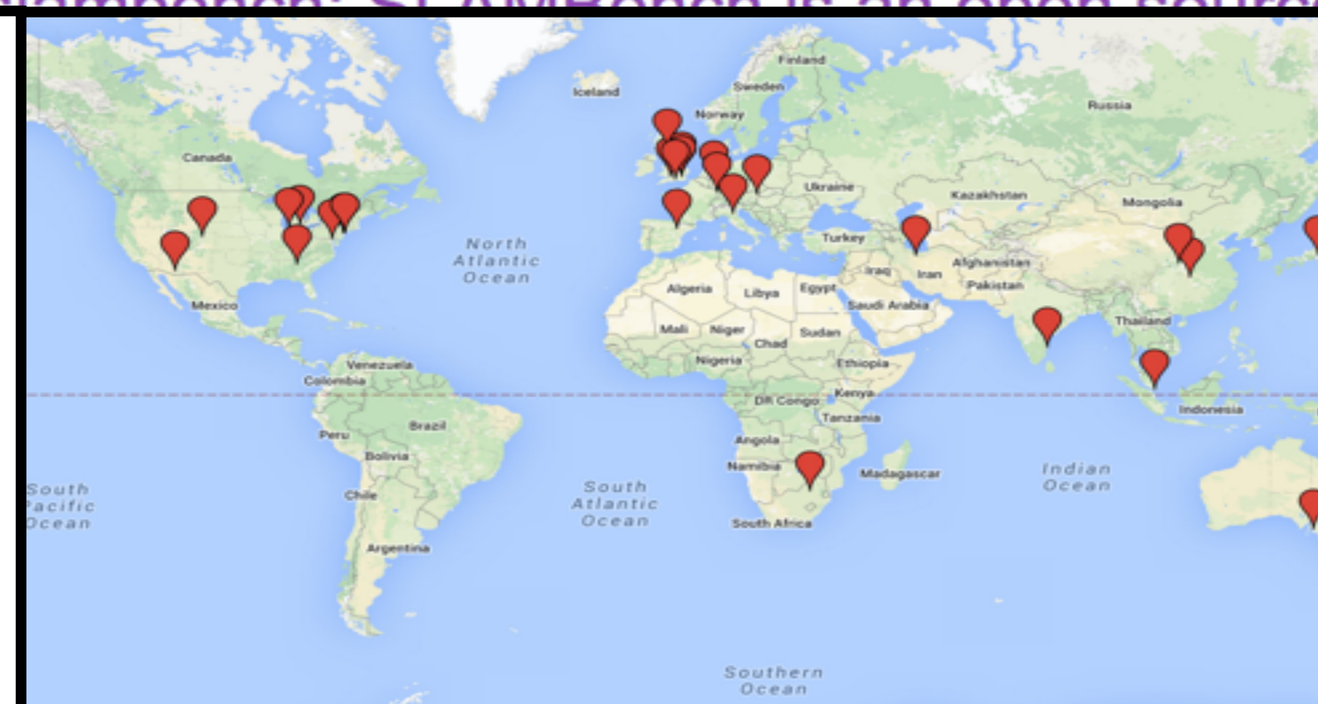


[Nardi et al., 2015]



Company users

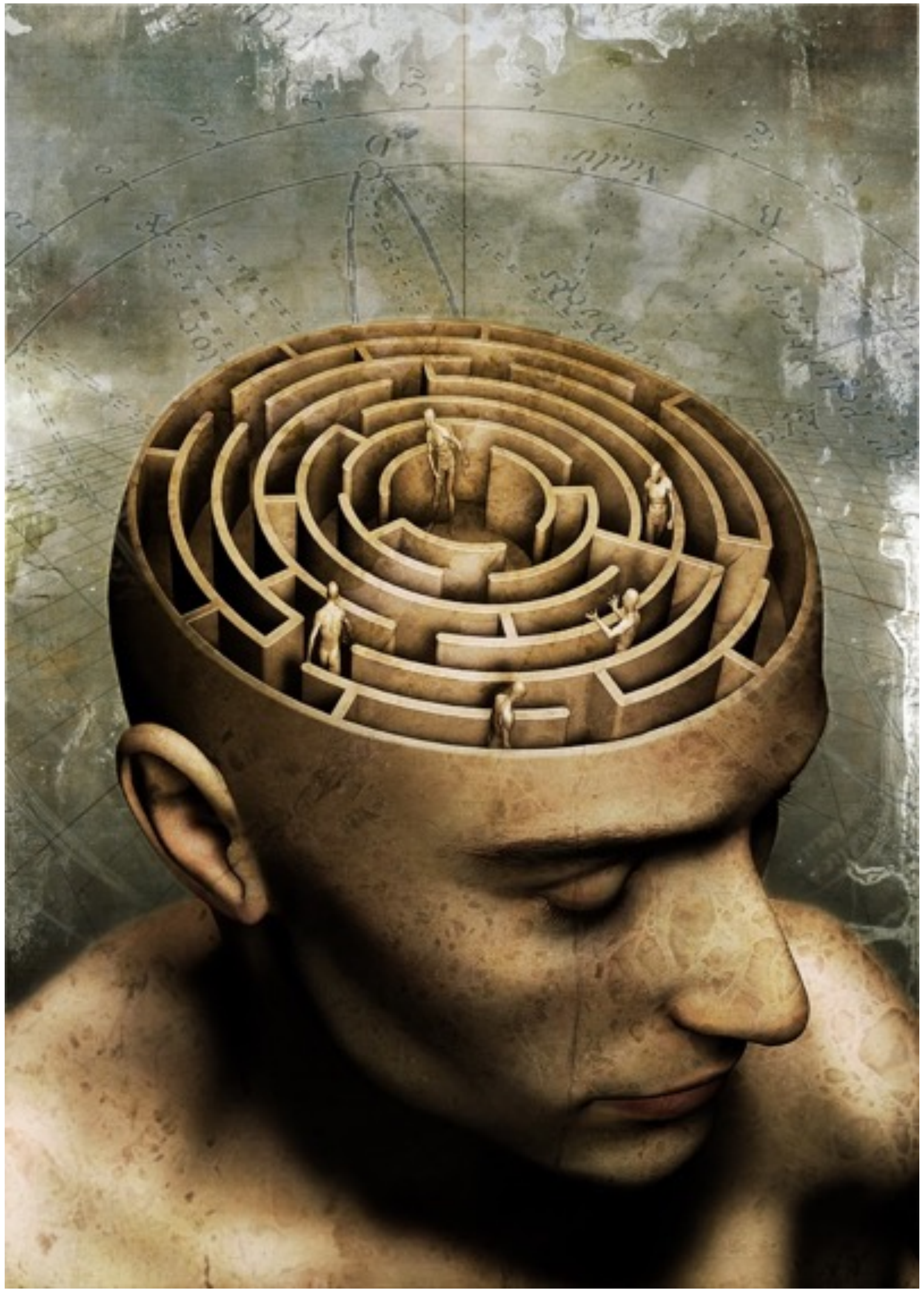
Academic users



Publicly released 13/11/2014
(1400+ downloads)

Outline

- The SLAM application, a brief introduction
- Benchmarking methodology
- **Space exploration of algorithmic and implementation design choices**



What is the optimisation space?

Configuration parameters:

Co-design space

Space 1

1. Algorithmic:
   - Application-specific parameters
   - Minimisation methods
   - Early exit condition values

Space 2

2. Compilation:
   - opencl-params: -cl-mad-enable,-cl-fast-relaxed-math, etc.
   - LLVM flags: O1, O2, O3, vectorize-slp-aggressive, etc.
   - Local work group size: 16/32/64/96/112/128/256
   - Vectorisation: width (1/2/4/8), direction (x/y)
   - Thread coarsening: factor (1/2/4/8/16/32), stride (1/2/4/8/16/32), dimension (x/y)

Space 3

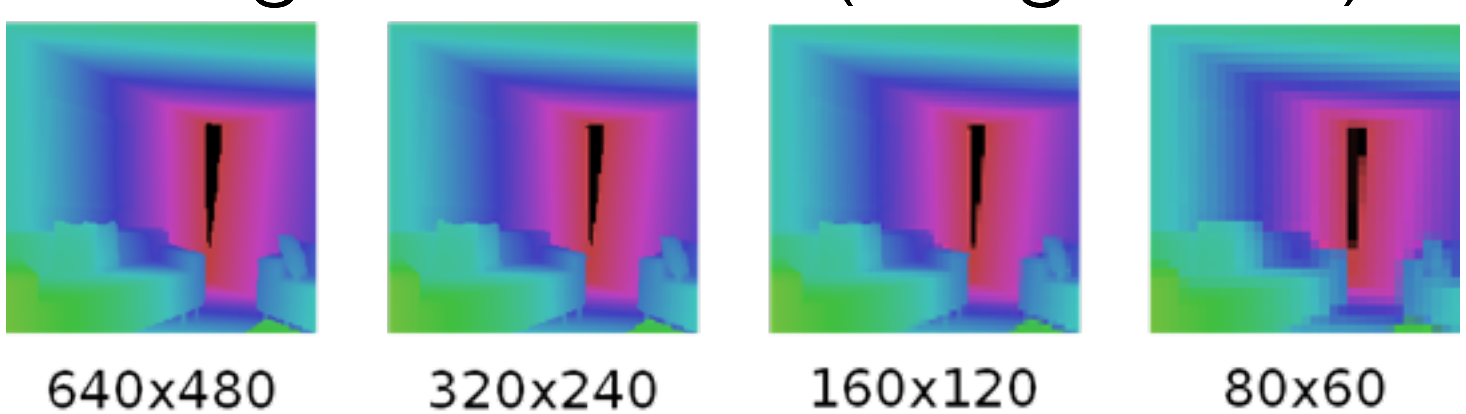
3. Architecture:
   - GPU frequency: 177/266/350/420/480/543/600/DVFS
   - # of active big cores: 0/1/2/3/4
   - # of active LITTLE cores: 1/2/3/4

Warning: huge spaces, impossible to run exhaustively

KinectFusion algorithmic features

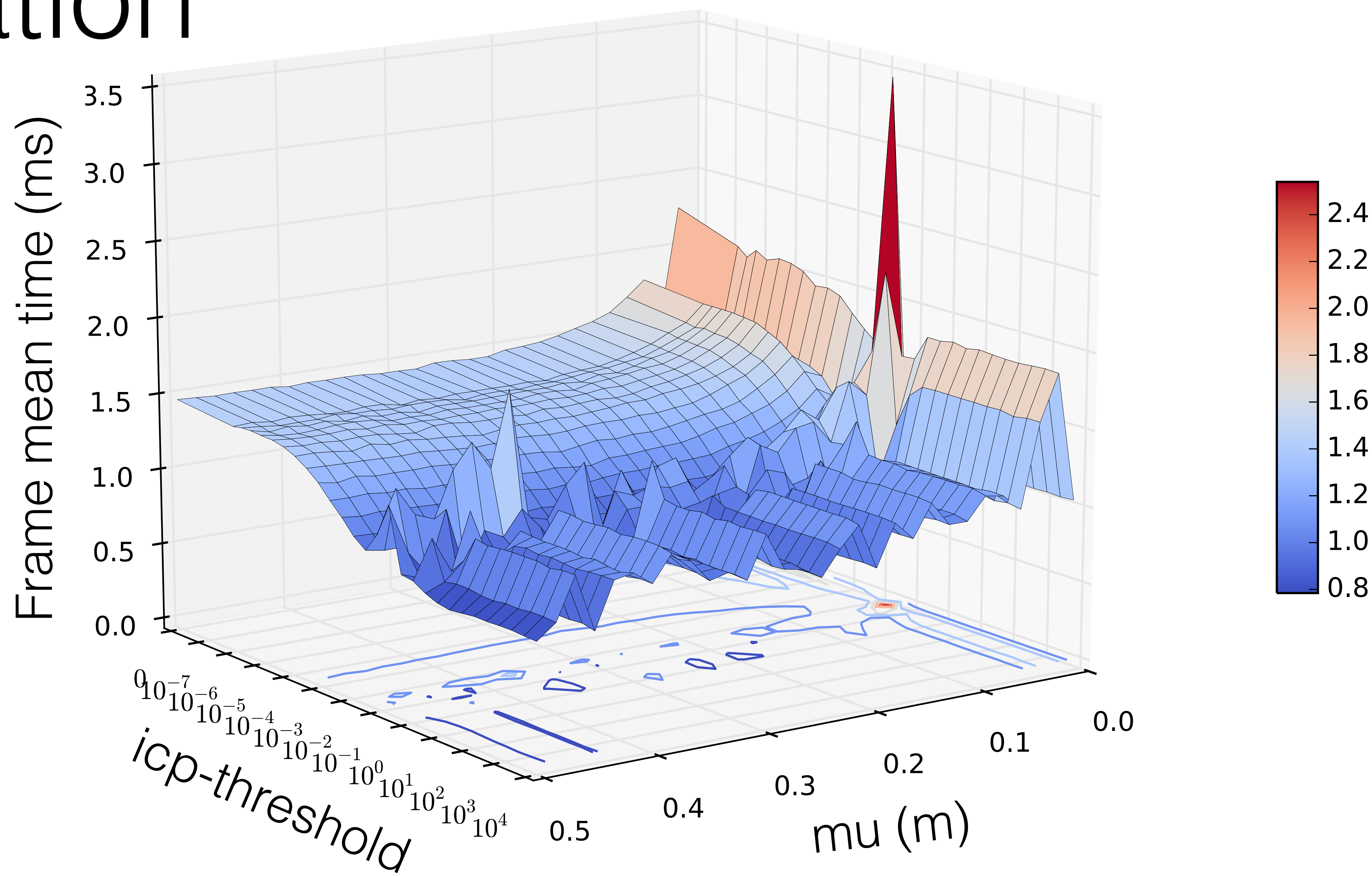
Features	Ranges
Volume resolution	64x64x64, 128x128x128, 256x256x256, 512x512x512
μ distance	0 .. 0.5
Pyramid level iterations (3 levels)	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Image resolution (image ratio)	1, 2, 4, 8
Tracking rate	1, 2, 3, 4, 5
ICP threshold	10^{-6} .. 10^{2}
Integration rate	1 .. 30

Image resolution (image ratio)



Different algorithmic features for **ElasticFusion**

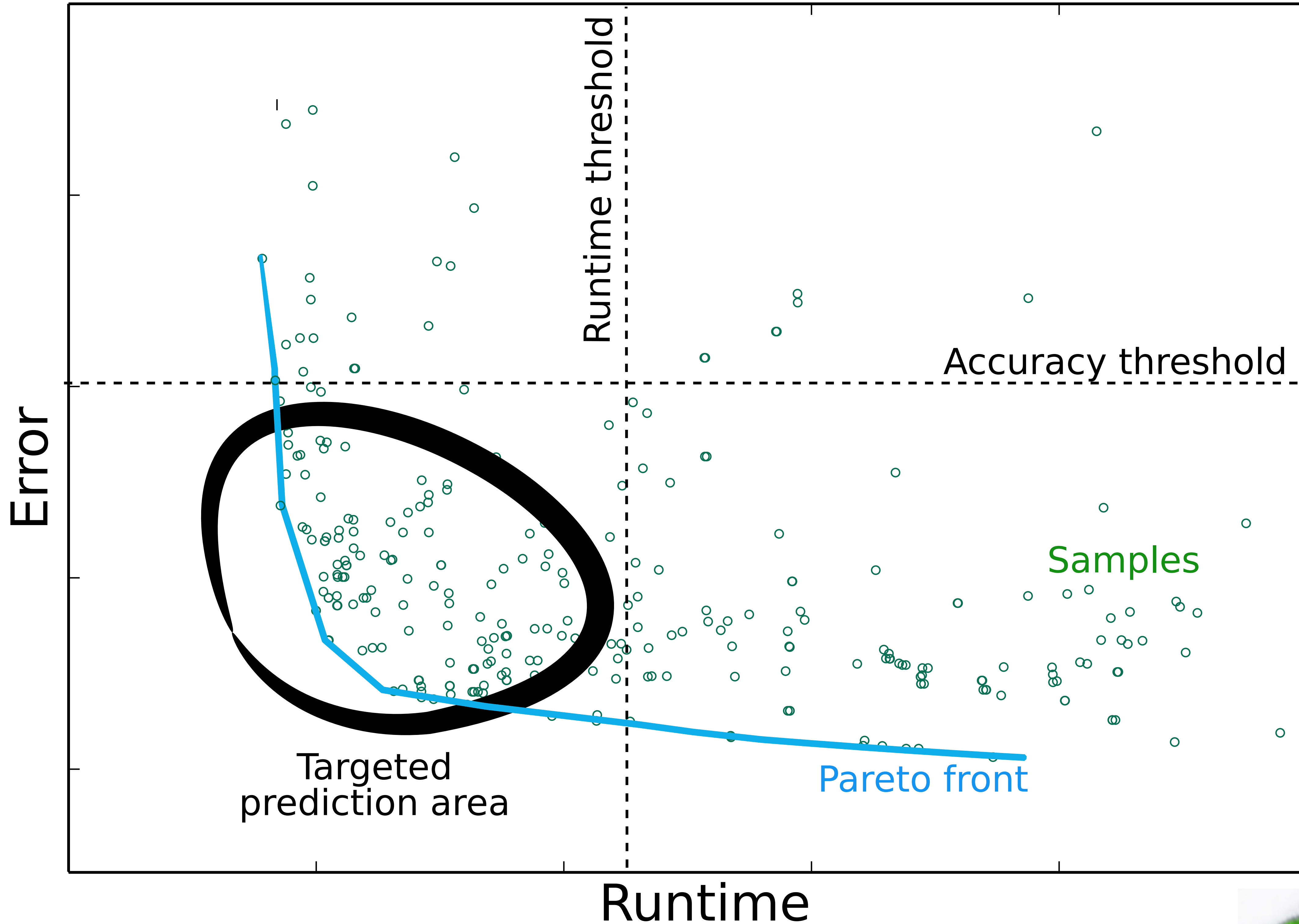
Motivation



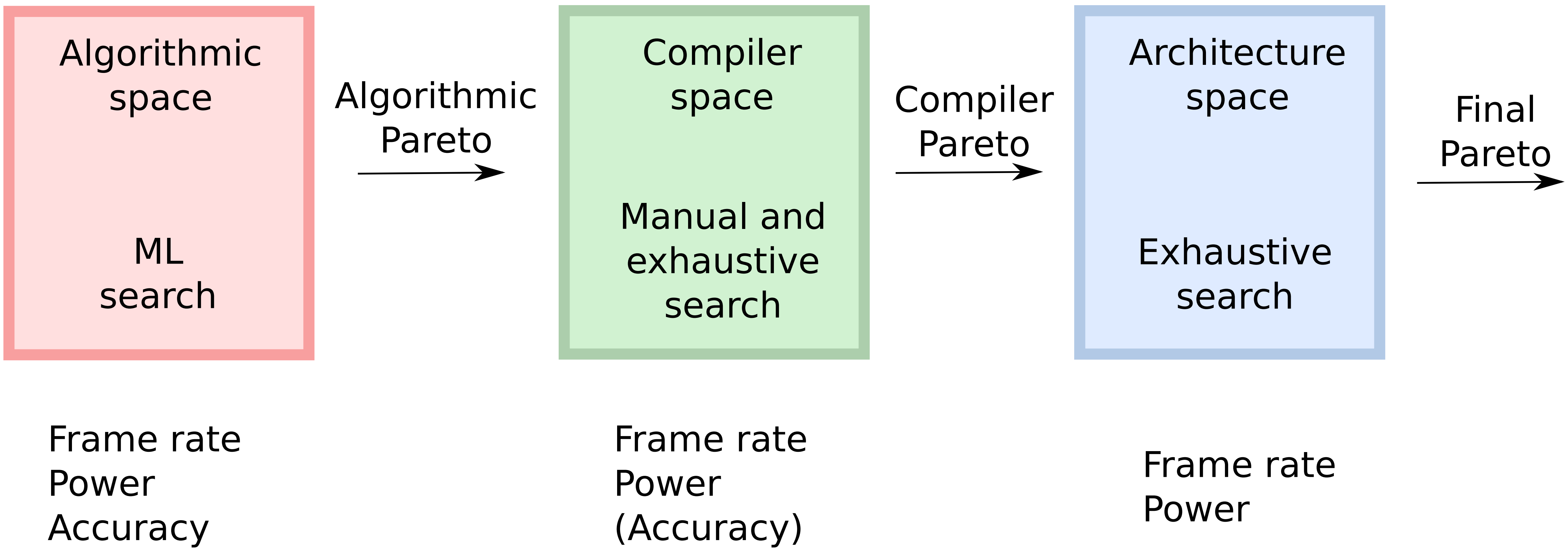
- KinectFusion runtime response surface: non-linear, multi-modal and non-smooth
- Optimal **algorithm configurability** enables better performance and better accuracy of the computation

Integrating Algorithmic Parameters into Benchmarking and Design Space Exploration in 3D Scene Understanding **(PACT 2016)**

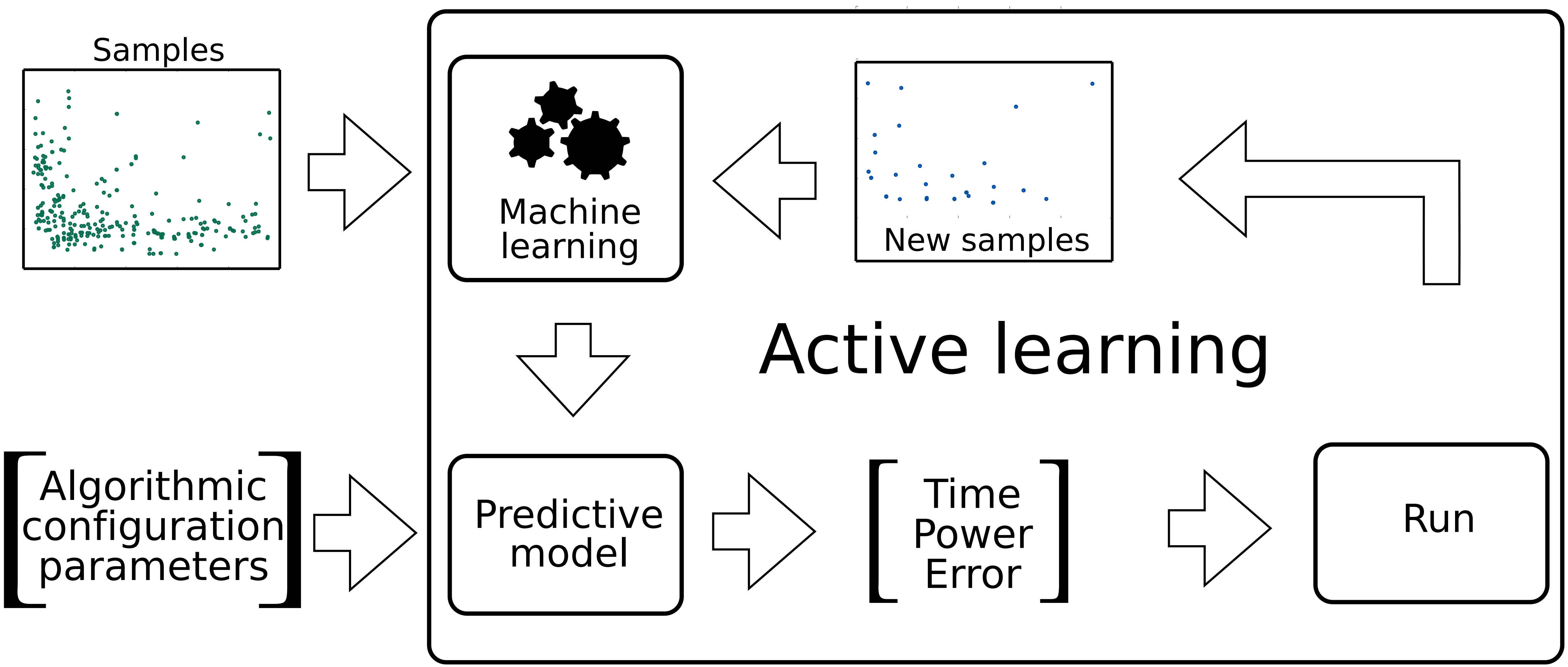
Exploration goal



Incremental exploration approach



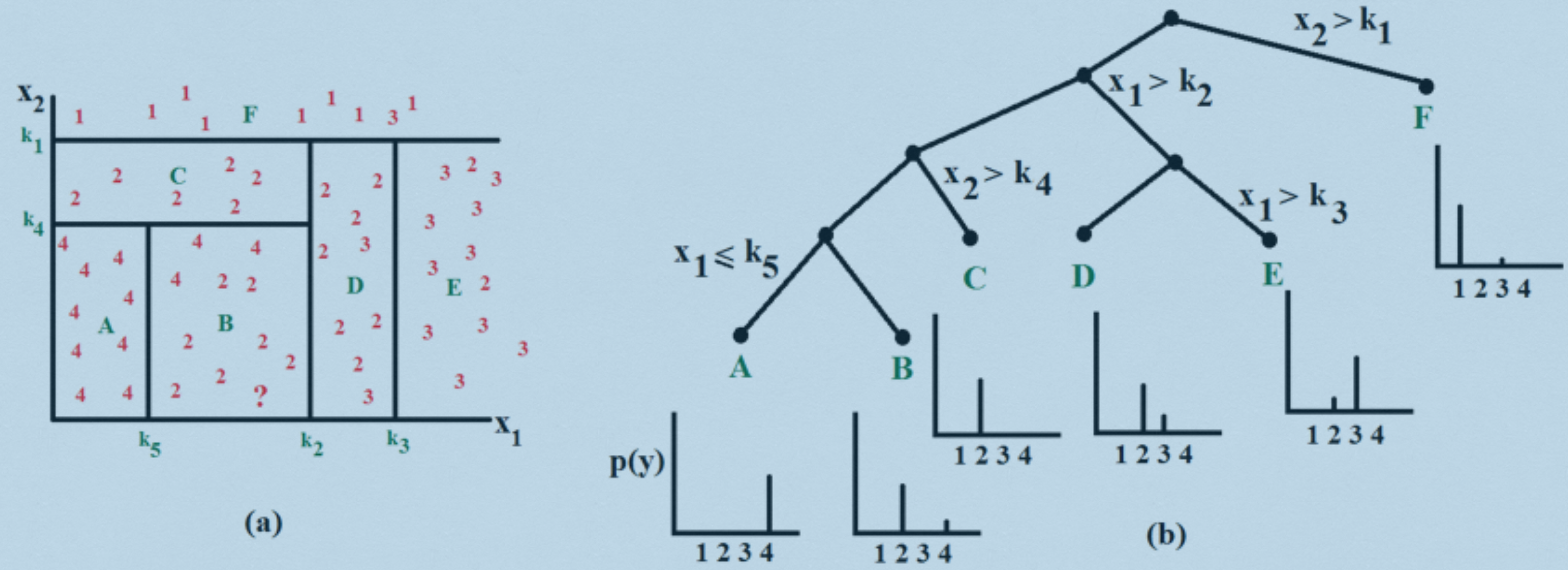
Algo design-space exploration (DSE)



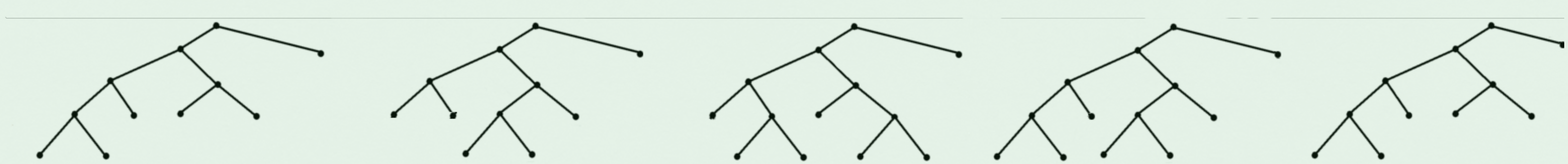
Luigi Nardi - Imperial College London

Machine learning methods used

Decision Tree



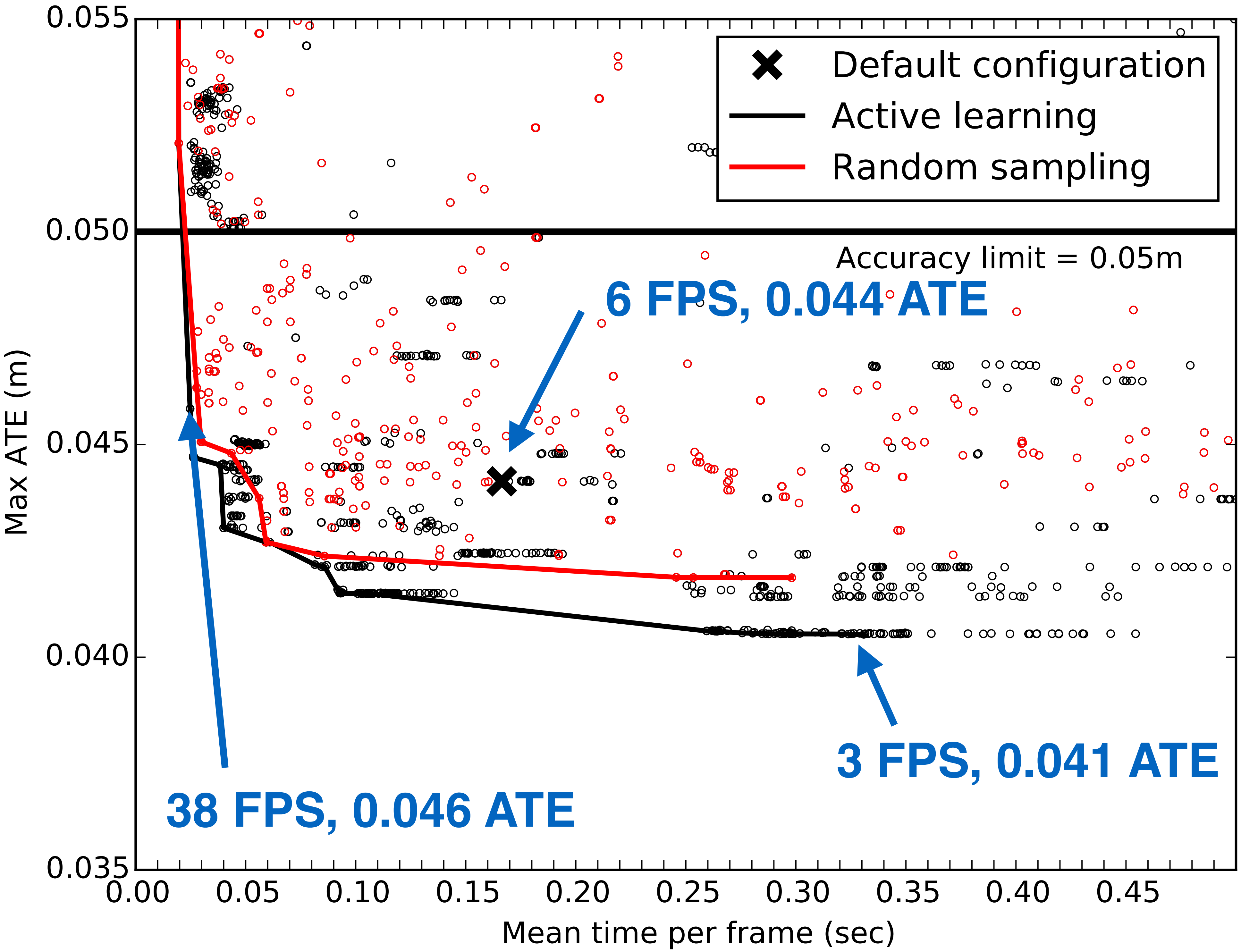
Random Forest



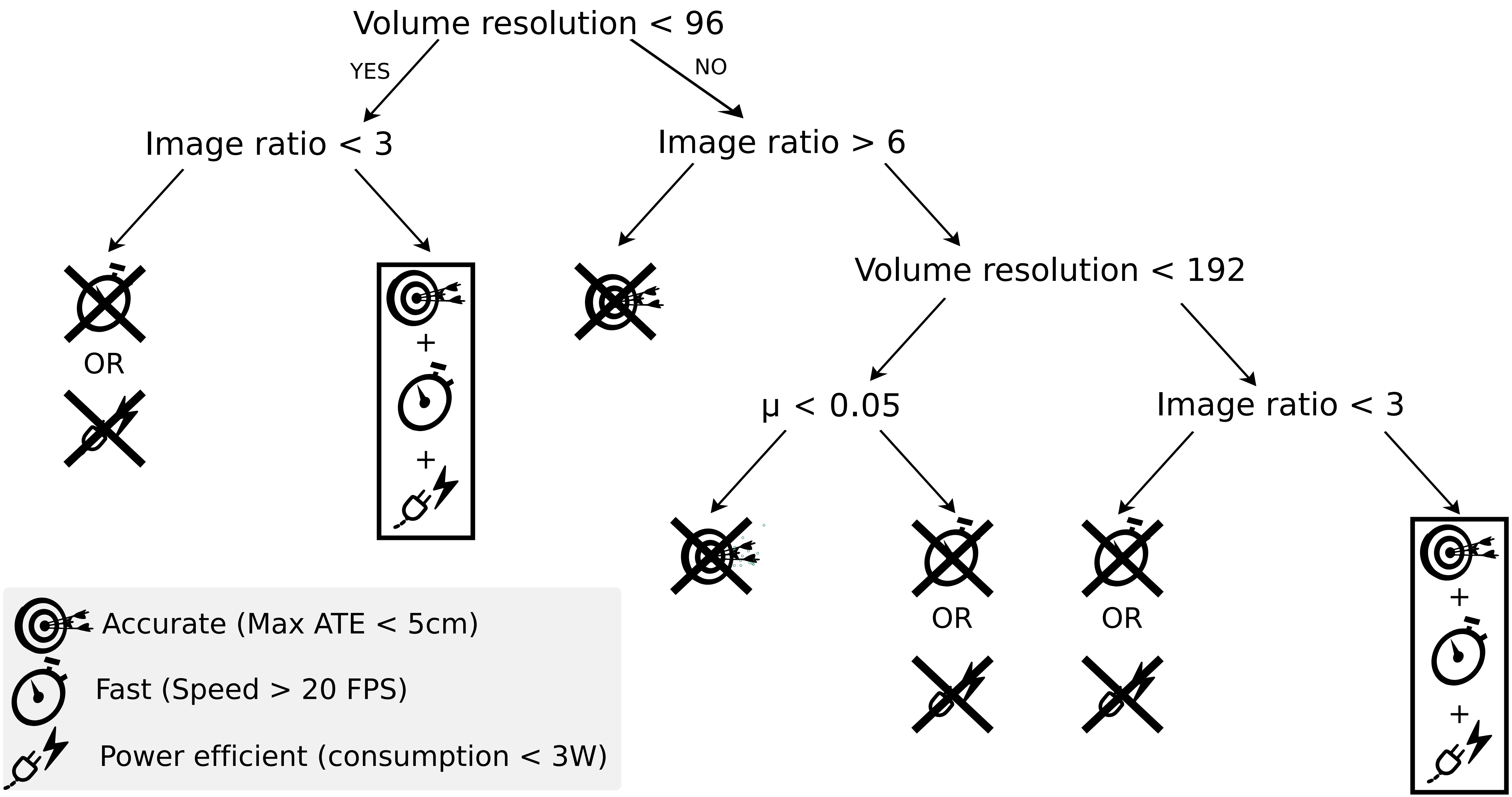
Luigi Nardi - Imperial College London

Results KinectFusion algo DSE error/runtime

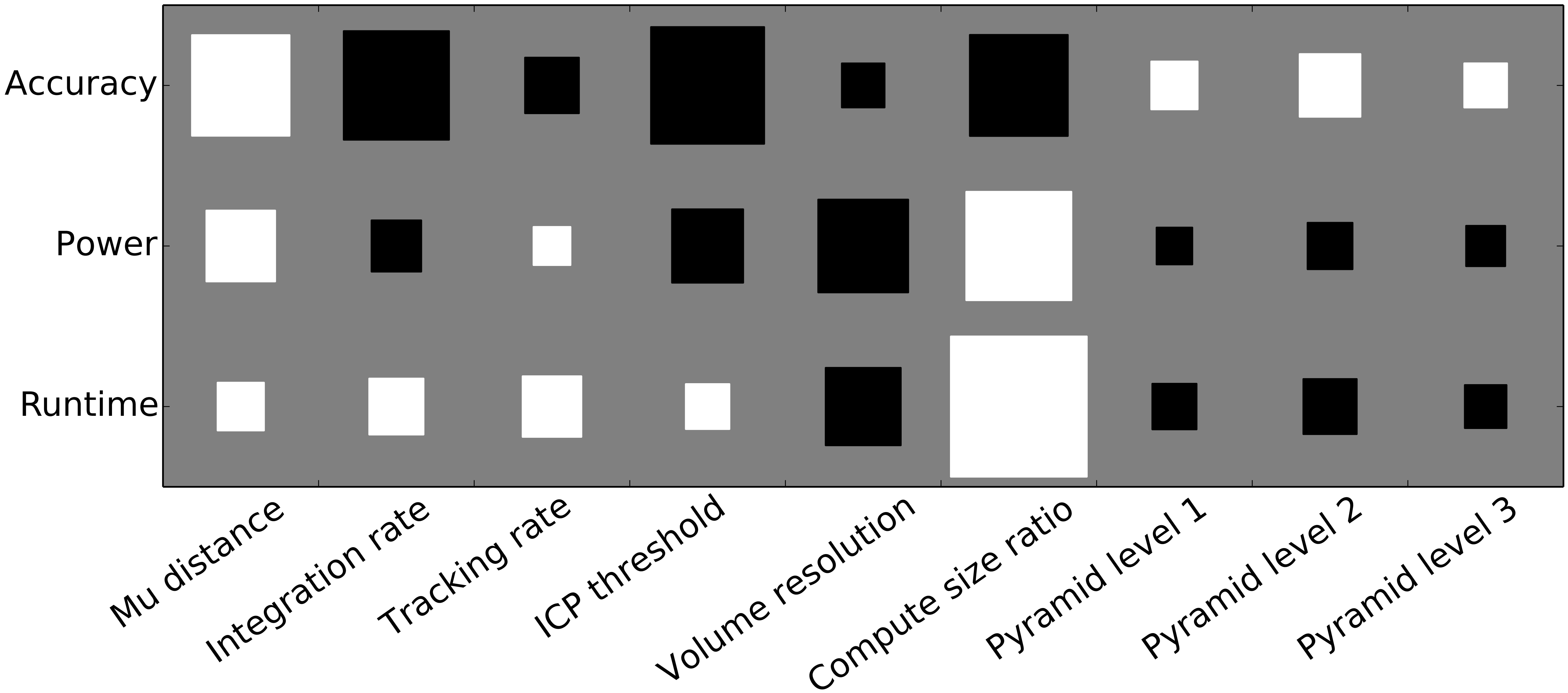
Machine	Type	CPU	CPU name	CPU cores	GPU	GPU name
Hardkernel ODROID-XU3	Embedded	ARM A15 + A7	Exynos 5422	4 + 4	ARM	Mali-T628



Predominant algorithmic features

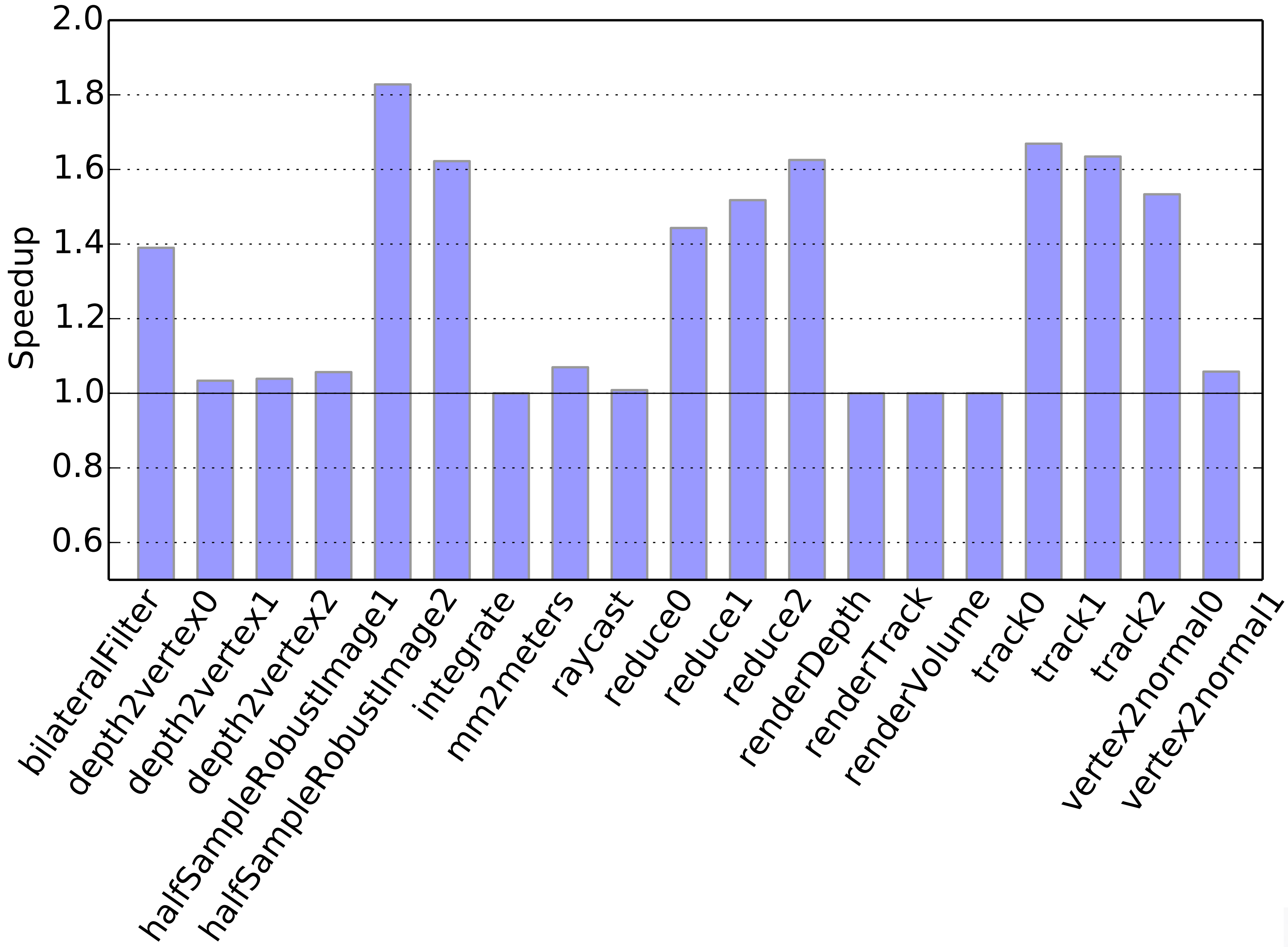


Hinton correlation diagram on algorithmic features

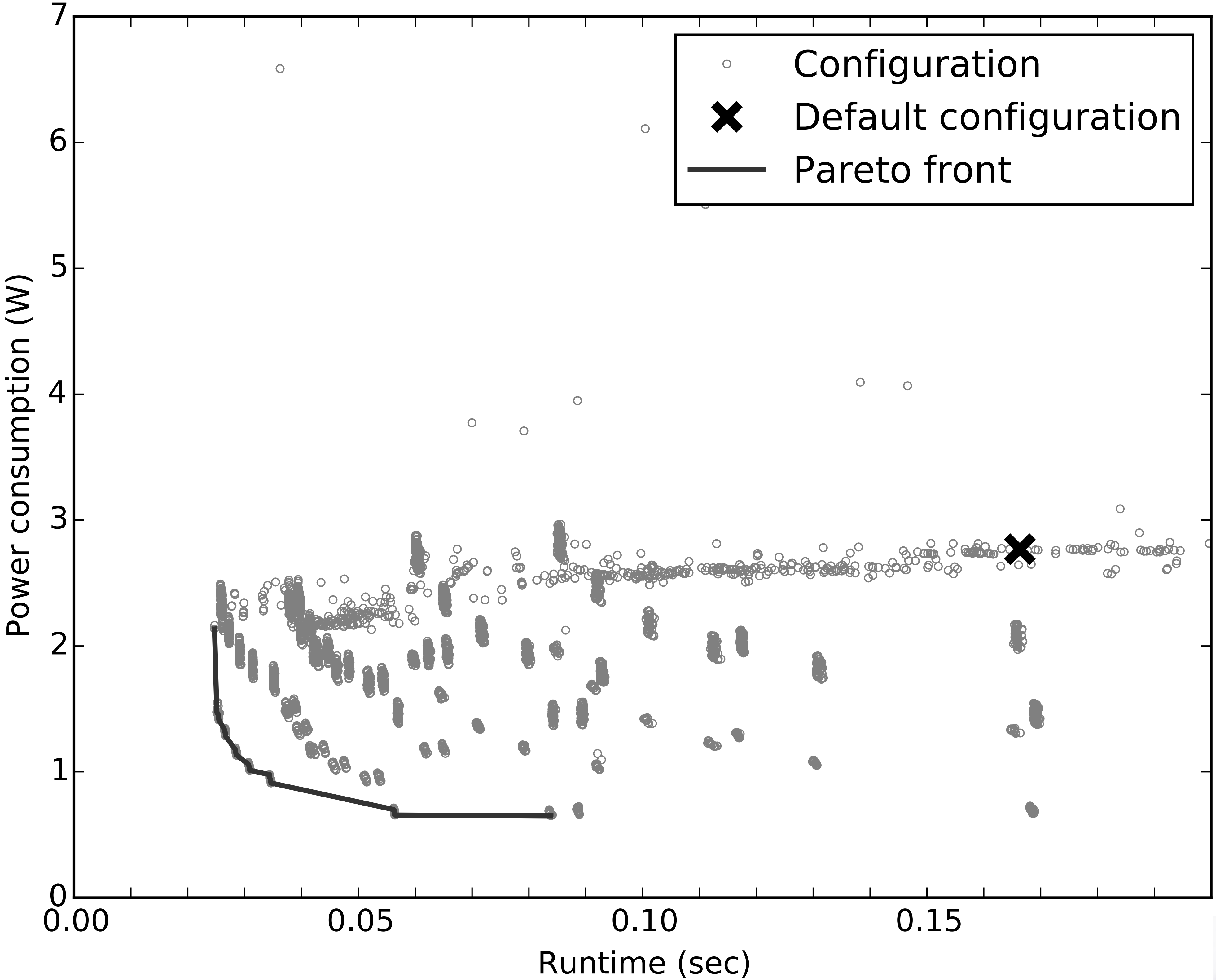


- Impact of algorithmic parameters (x-axis) on the performance metrics (y-axis) on the ODROID-XU3 platform.
- Bigger squares indicate a higher correlation.
  <u>White square</u>: a parameter which when increased improves the metric
  <u>Black square</u>: a parameter which when increased worsen the metric

Results KinectFusion compiler DSE speedup



Results KinectFusion architecture DSE power/runtime



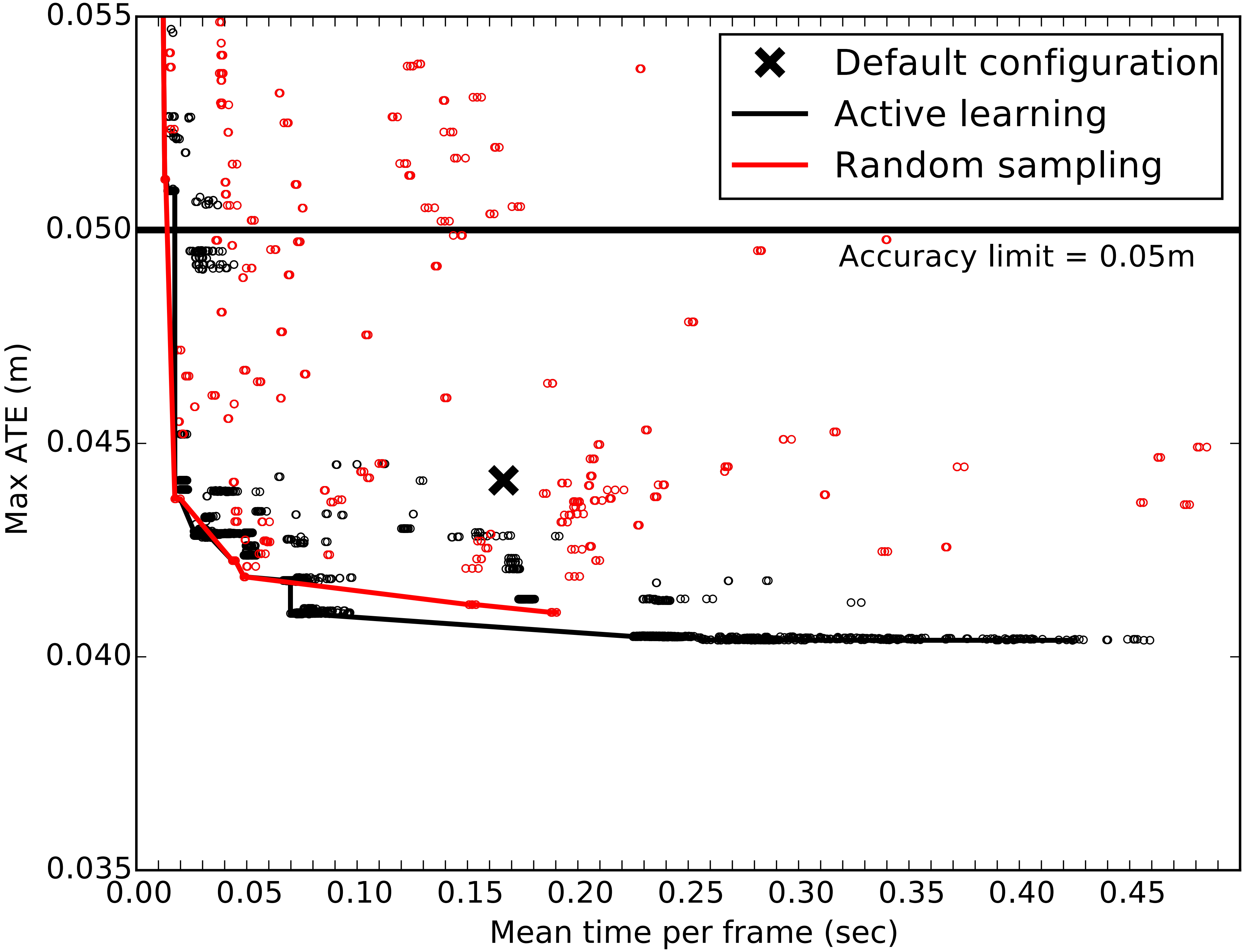
DSE final result

Constraint	Runtime (FPS)	Max ATE (cm)	Power (Watts)
Default	6.03	4.41	2.77
Best runtime	39.85	4.47	1.47
Best accuracy	1.51	3.30	2.38
Best power	11.92	4.45	0.65
Power $<$ 1W	29.09	4.47	0.98
Power $<$ 2W	39.85	4.47	1.47
FPS $>$ 10	11.92	4.45	0.65
FPS $>$ 20	28.87	4.47	0.91
FPS $>$ 30	32.38	4.47	1.01

- Most of the improvement comes from the algorithmic space
- KinectFusion real-time on a popular embedded device
- **Enabling auto-tuning at the domain-specific language (DSL) level**

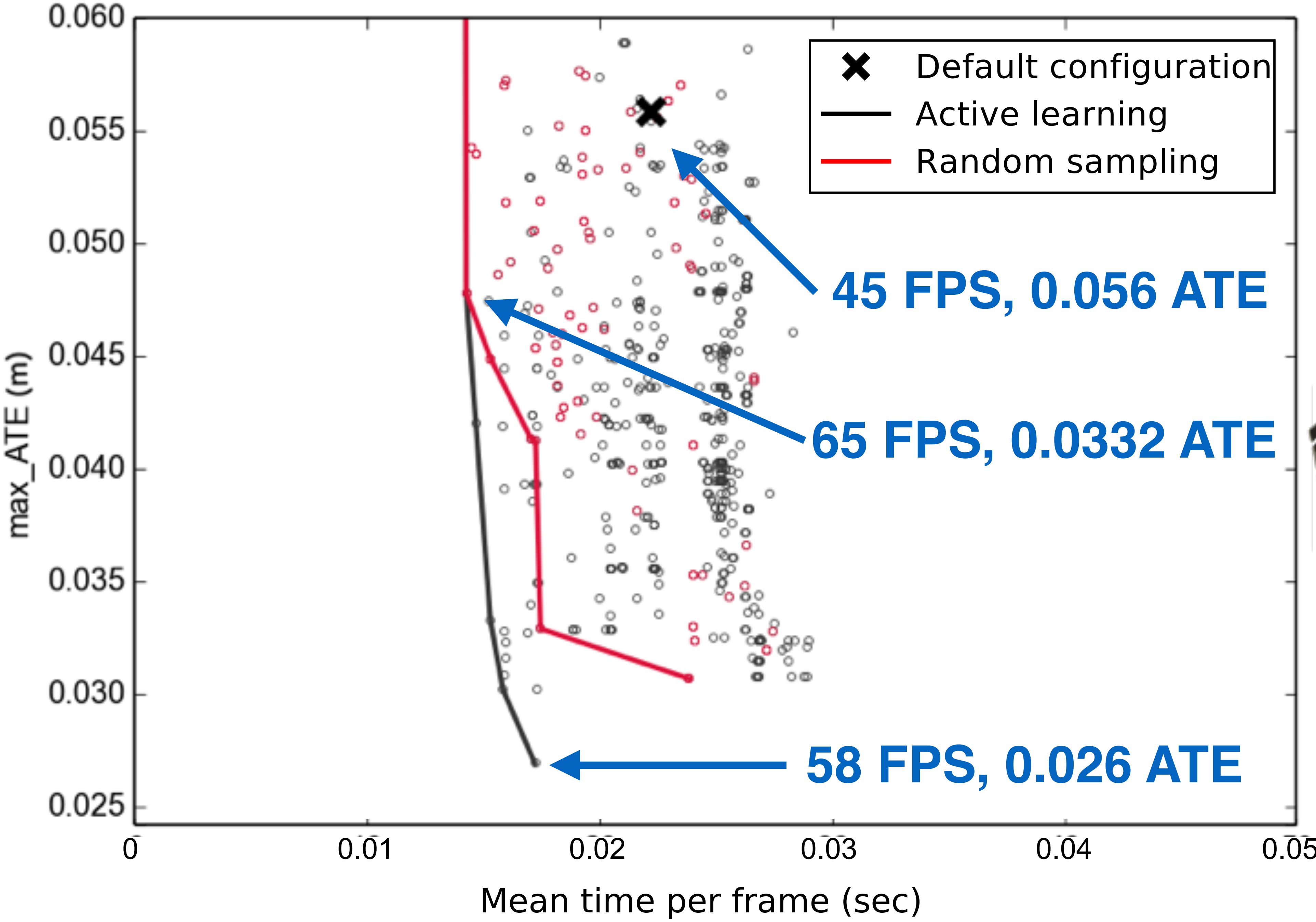
Results KinectFusion algo DSE error/runtime

Machine	Type	CPU	CPU name	CPU cores	GPU	GPU name
ASUS T200TA	Detachable laptop	Intel Silvermont	Atom Z3795	4	Intel	HD Graphics

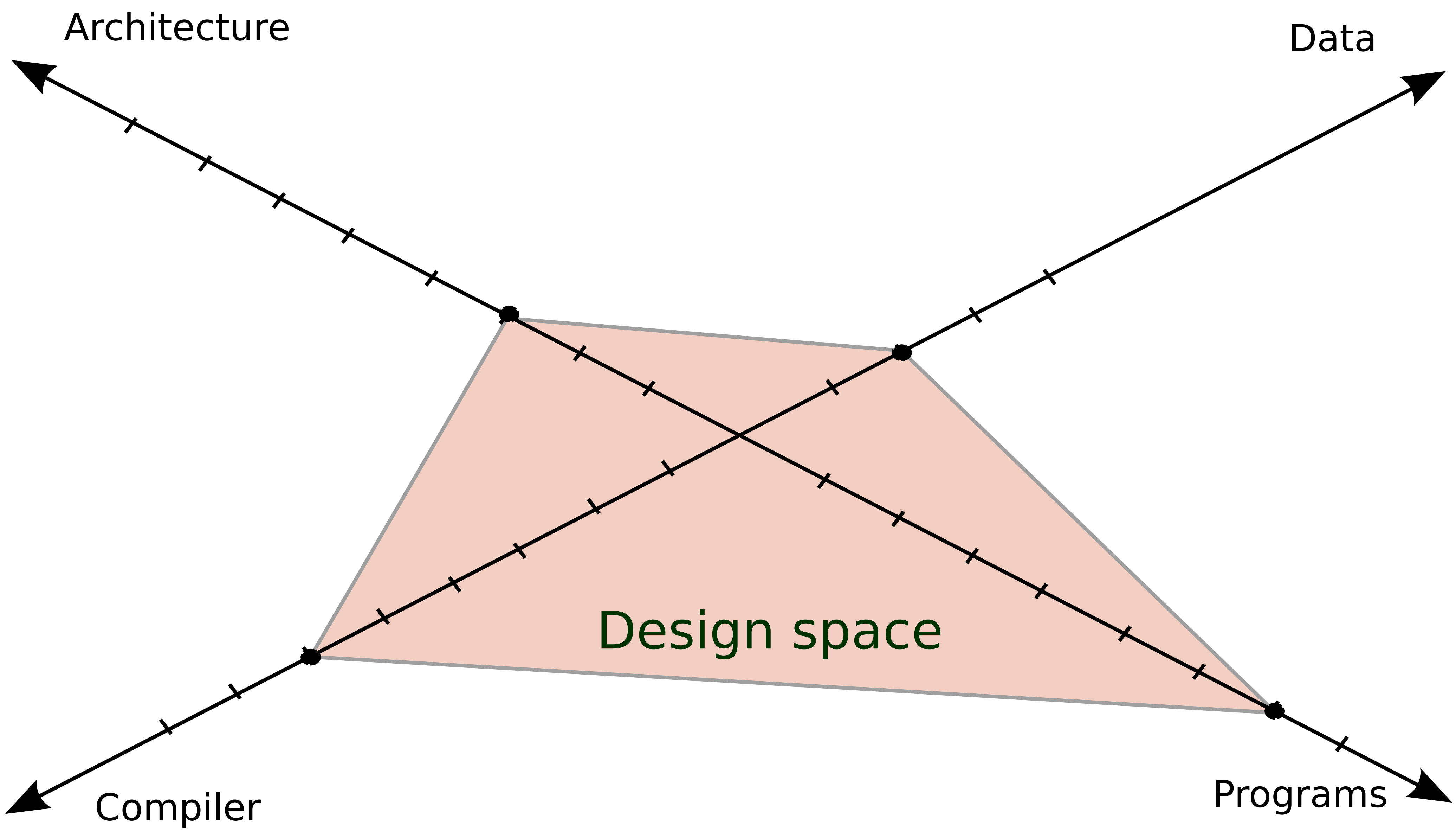


Results **ElasticFusion** algo DSE error/runtime

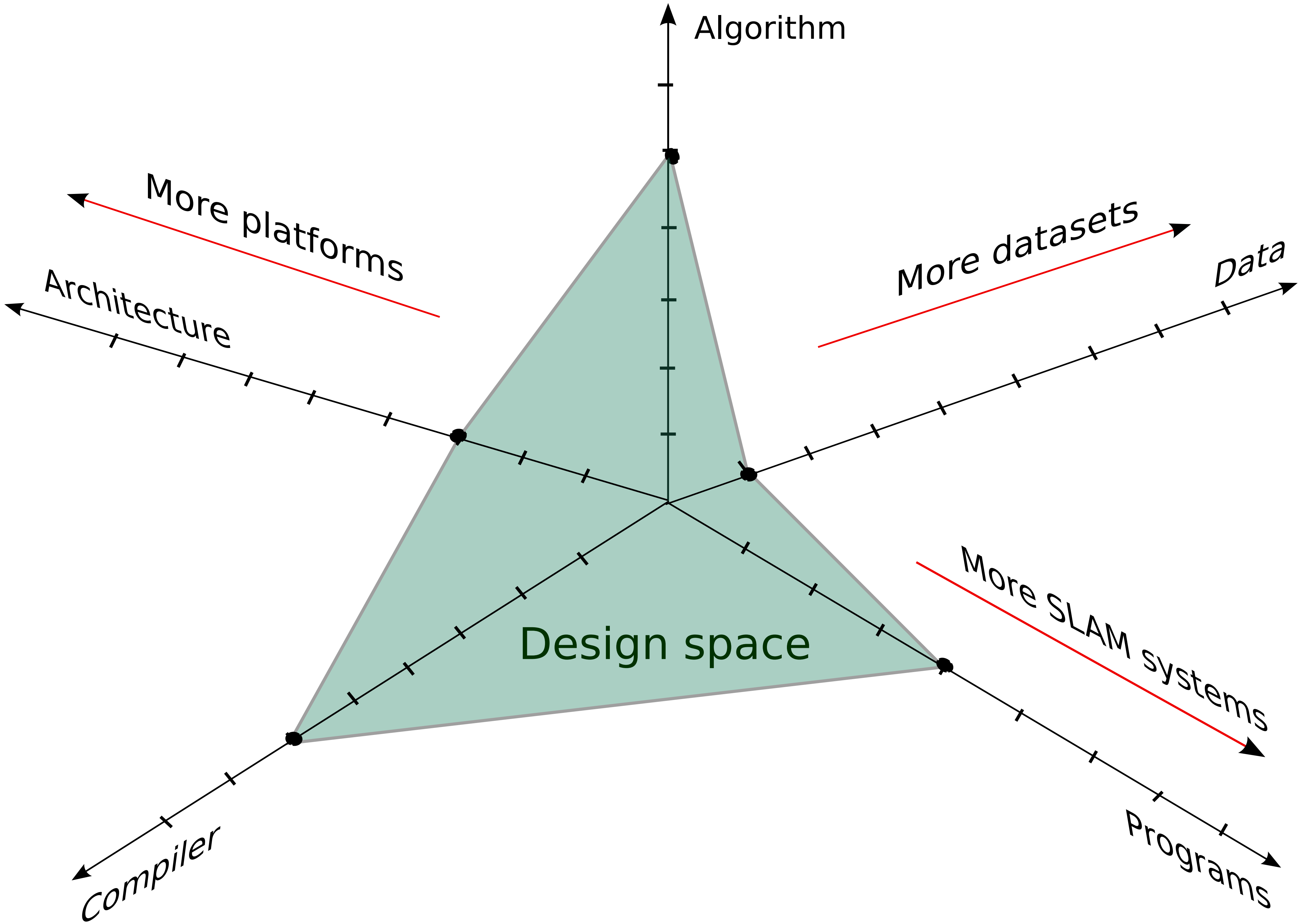
Machine	Type	CPU	CPU name	CPU cores	GPU	GPU name
NVIDIA/Intel	Desktop	Intel Ivy Bridge	E5-1620	8	NVIDIA	GTX 780 Ti



DSE the big picture I

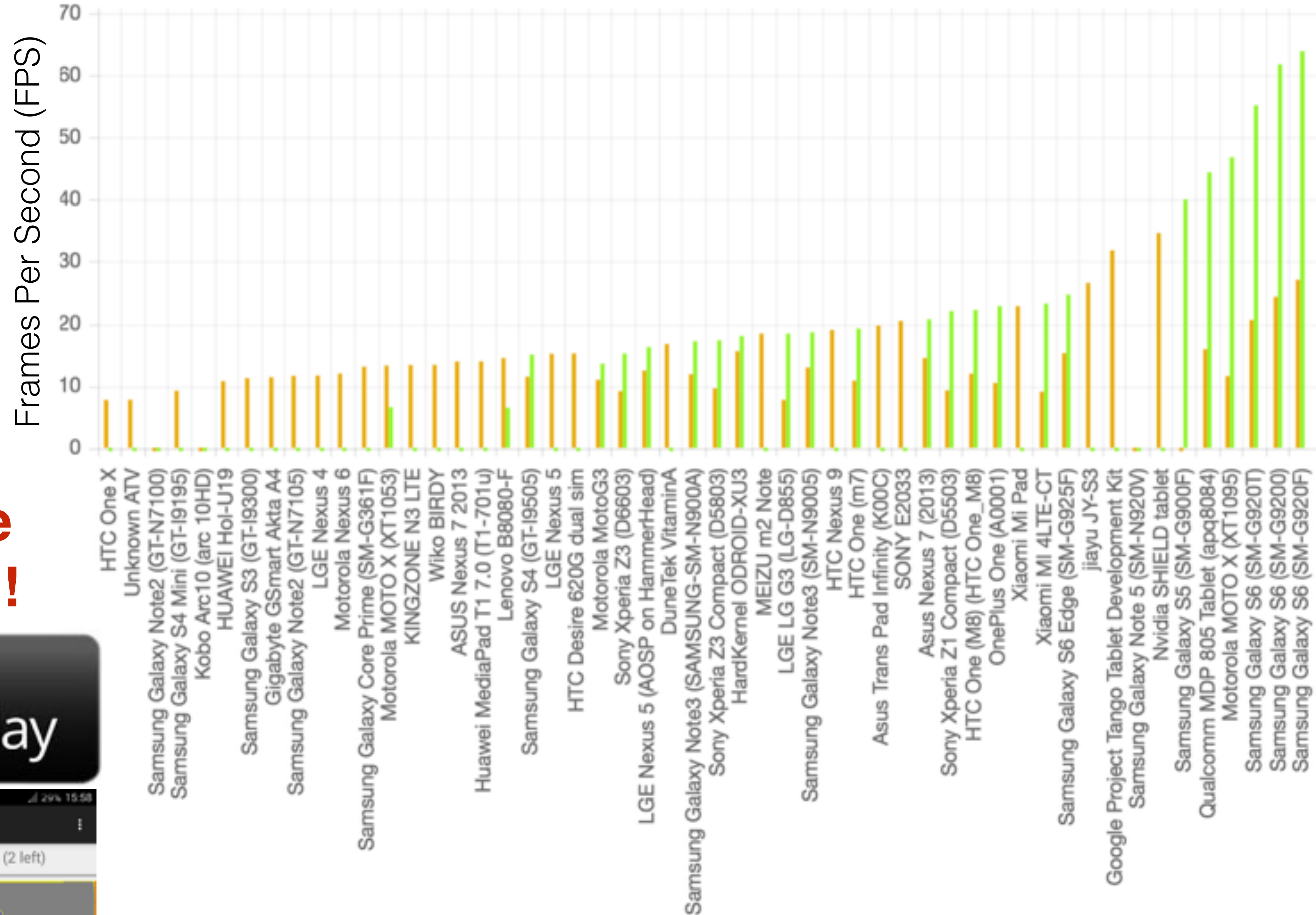


DSE the big picture II



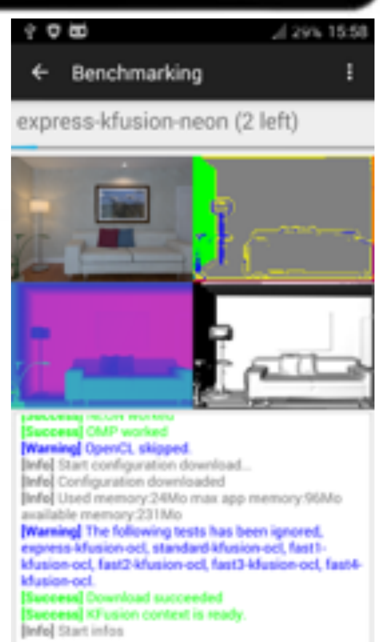
Luigi Nardi - Imperial College London

Crowdsourcing mobile Android SLAMBench



Get it now,
And see where
your device is!!

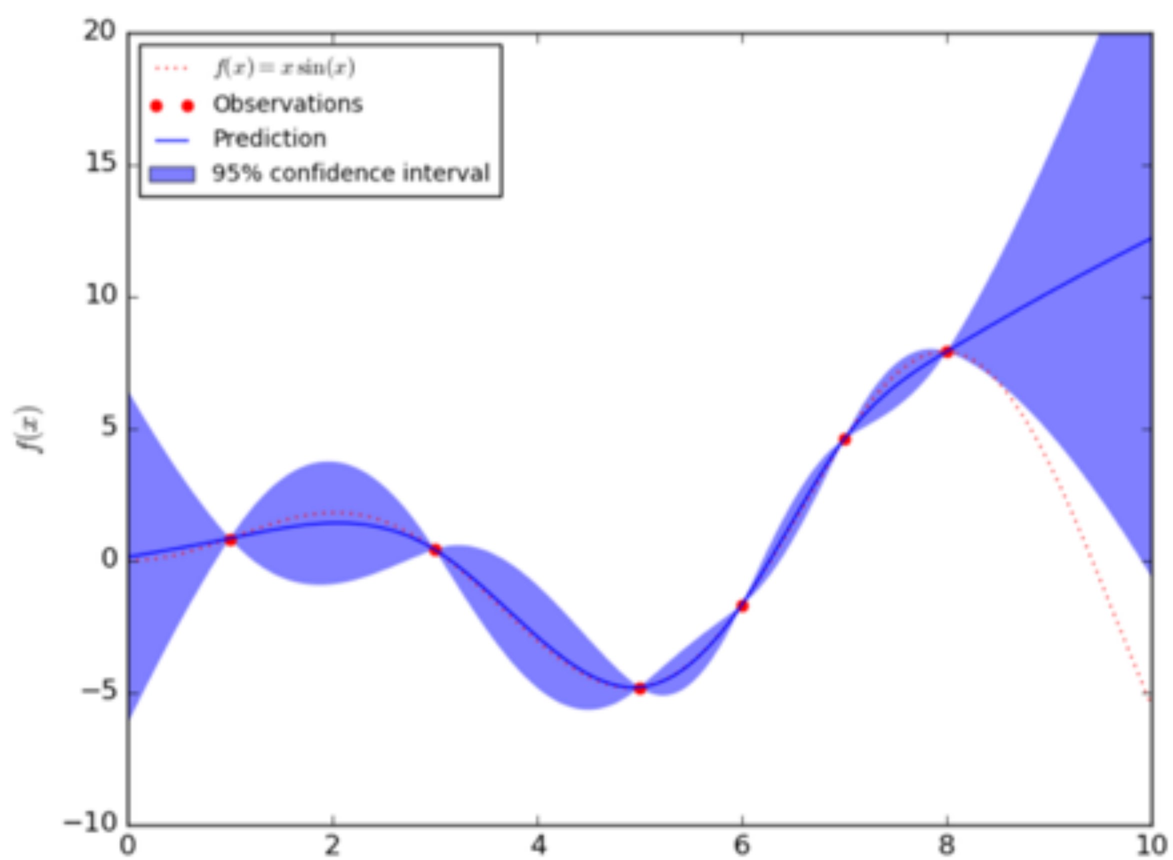
Get it on Google play

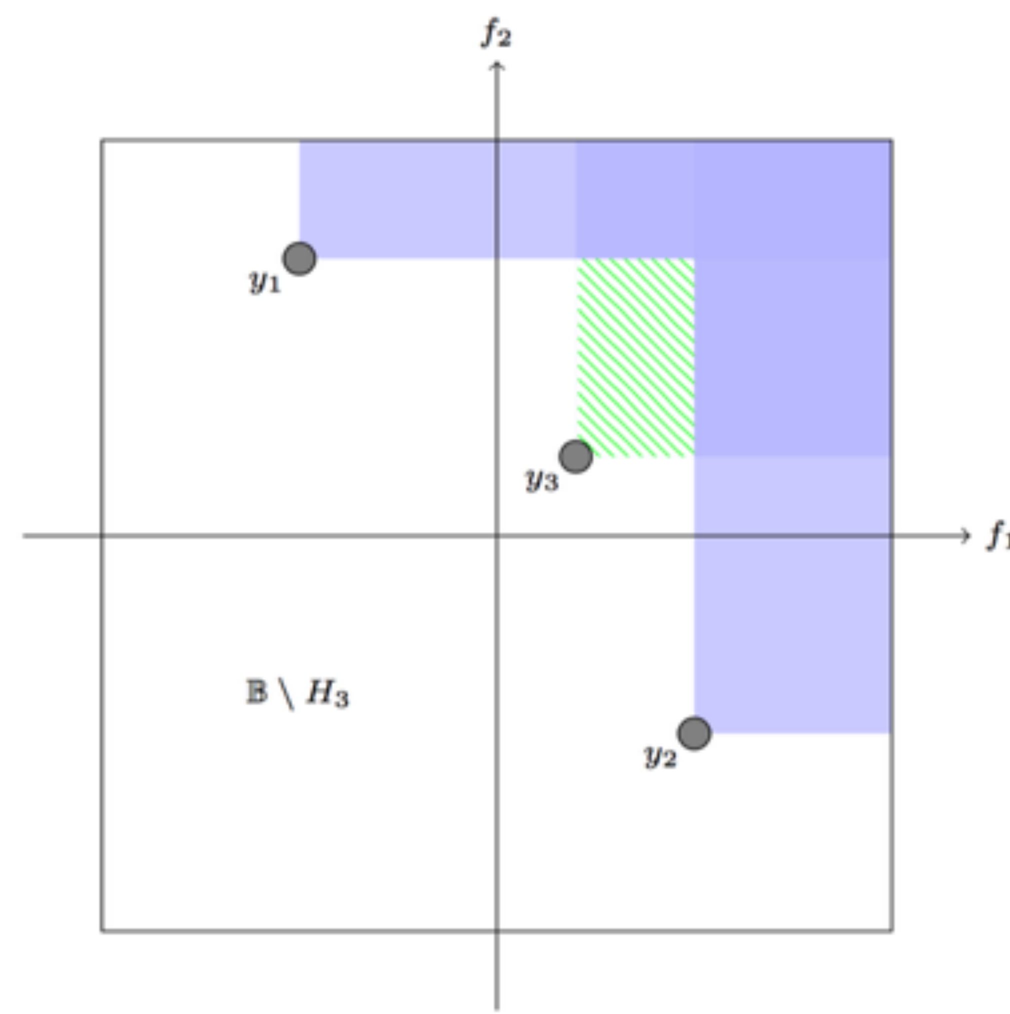


- It runs a set of configurations on the available languages on your device
- Then shows the best achieved result

Future work

- Gaussian Processes (GPs) to reduce the total time to create the Pareto
- Global optimisation of Black-Box functions
- Multi-objective optimisation: Expected HyperVolume Improvement (EHVI)
- Transfer learning, e.g. learn on INTEL and deploy on ARM
- Application of this methodology to other application domains
- Application to DSL back-end optimisation





A Bayesian approach to constrained single- and multi-objective optimization, Paul FELIOT, Julien BECT and Emmanuel VAZQUEZ, 2015

Conclusion - take away messages

1. Building tools to explore the performance landscape for SLAM solutions
2. Generalisation to other applications
3. Multi-objective optimisation: speed/power/accuracy
4. Semantic accuracy check is very powerful:
   - enables non bit-wise accuracy check
   - aggressive approximate computing and auto-tuning
5. Pareto maps how configurations should be adapted when objectives change - static and dynamic
6. Large improvement over default configuration

References I

- [Nardi et al. 2015] L. Nardi, B. Bodin, M. Z. Zia, J. Mawer, A. Nisbet, P. H. J. Kelly, A. J. Davison, M. Luján, M. F. P. O'Boyle, G. Riley, N. Topham, and S. Furber. "Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM." Submitted, arXiv:1410.2167, 2015.
- [Newcombe et al. ICCV 2011] R. A. Newcombe, S. J. Lovegrove and A. J. Davison. "DTAM: Dense tracking and mapping in real-time." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.
- [Rusinkiewicz and Levoy 2001] S. Rusinkiewicz, and M. Levoy. "Efficient variants of the ICP algorithm." 3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on. IEEE, 2001.
- [Chen et al. 2013] J. Chen, D. Bautembach, and S. Izadi, Scalable real-time volumetric surface reconstruction, in ACM Trans. Graph., 2013.
- [Newcombe et al. ISMAR 2011] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. "KinectFusion: Real-time dense surface mapping and tracking." 10th IEEE Int. Symp. on Mixed and augmented reality (ISMAR), 2011.
- [Handa et al. 2014] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A Benchmark for RGB-D Visual Odometry, 3D Reconstruction and SLAM. IEEE Int. Conf. on Robotics and Automation, ICRA 2014.
- [Reitmayr] G. Reitmayr. KFusion github 2011. https://github.com/GerhardR/kfusion
- [Curless and Levoy 1996] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In Proc. Computer graphics and interactive technique. ACM, 1996.
- [Whelan et al. 2012] T. Whelan, M. Kaess, M. Fallon,H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended kinectfusion. 2012.
- C. Jiawen, D. Bautembach, and S. Izadi. "Scalable real-time volumetric surface reconstruction." ACM TOG, 2013.
- Frahm, Jan-Michael, et al. "Building Rome on a cloudless day." Computer Vision–ECCV 2010. Springer Berlin Heidelberg, 2010.
- Erhan, Dumitru, et al. "Scalable object detection using deep neural networks." Proceedings of the IEEE CVPR. 2014.

References II

- Arbelaez, Pablo, et al. "Contour detection and hierarchical image segmentation." IEEE Pattern Analysis and Machine Intelligence, 2011.
- [Ogilvie 2014] Ogilvie, William, et al. "Fast automatic heuristic construction using active learning." Proceedings of the Workshop on Languages and Compilers for Parallel Computing (LCPC’14). 2014.
- [Siegmund 2015] Siegmund Norbert et al. "Performance-influence models for highly configurable systems”, submitted FSE 2015.
- [Guo 2013] Guo, Jianmei, et al. "Variability-aware performance prediction: A statistical learning approach." Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference on. IEEE, 2013.
- [Grewe 2011] Grewe, Dominik et al. "A static task partitioning approach for heterogeneous systems using OpenCL." Compiler Construction. Springer Berlin Heidelberg, 2011.
- [Kurek 2013] Kurek, Maciej, Tianchi Liu, and Wayne Luk. "MULTI-OBJECTIVE SELF-OPTIMIZATION OF RECONFIGURABLE DESIGNS WITH MACHINE LEARNING." 2nd Workshop on Self-Awareness in Reconfigurable Computing Systems (SRCS’13). 2013.
- [Balaprakash 2013] Balaprakash, Prasanna, Robert B. Gramacy, and Stefan M. Wild. "Active-learning-based surrogate models for empirical performance tuning." Cluster Computing (CLUSTER), 2013 IEEE International Conference on. IEEE, 2013.
- [Vespa 2015] Vespa Emanuele. "Sparse voxelization of dense volumetric reconstruction with automated analysis of scene reconstruction quality.” M.Res. thesis, Imperial College London, 2015.

Copyrights

- Author: unknown. Microsoft Kinect camera. [Image]. Retrieved from http://channel9.msdn.com/Series/KinectSDKQuickstarts/Understanding-Kinect-Hardware
- Author: Dyson Ltd. Dyson 360 Eye. [Video]. Retrieved from https://www.youtube.com/watch?v=OadhuICDAjk
- Author: Google Inc. Google Tango project. [Image]. Retrieved from http://blogthinkbig.com/en/project-tango-googles-mobile-kinect/
- Author: unknown. Audi autonomous car. [Photograph]. Retrieved from http://www.wired.com/2010/06/audis-robotic-car-looks-hot-in-old-school-livery/
- Author: ExtremeTech. Google Shaft robot. [Photograph]. Retrieved from http://www.extremetech.com/extreme/173318-google-wins-darpas-robotics-challenge-wonders-if-it-was-a-good-idea-to-turn-down-future-military-contracts
- Author: HardKernel. ODROID-XU3 board. [Photograph]. Retrieved from http://www.hardkernel.com/main/products/prdt_info.php?g_code=G135235611947
- Author: PC Specialist Ltd. Vortex series laptop. [Photograph]. Retrieved from https://www.pcspecialist.co.uk/forums/showthread.php?23366-My-new-beast-15-6-quot-Vortex-III
- Author: Arndale.org. Arndale board. [Photograph]. Retrieved from http://www.arndaleboard.org/wiki/index.php/Main_Page
- Author: Unknown. Chip. [Image]. Retrieved from https://cajalesygalileos.wordpress.com/2013/06/23/un-chip-ultrasensible-identifica-15-cepas-de-gripe/
- Author: Unknown. Eye. [Image]. Retrieved from http://gallery.digitalculture.asu.edu/?/interactive-environments/computer-vision/
- Author: Unknown. Compiler. [Image]. Retrieved from http://d3q6qq2zt8nhwv.cloudfront.net/107/large-icon.png