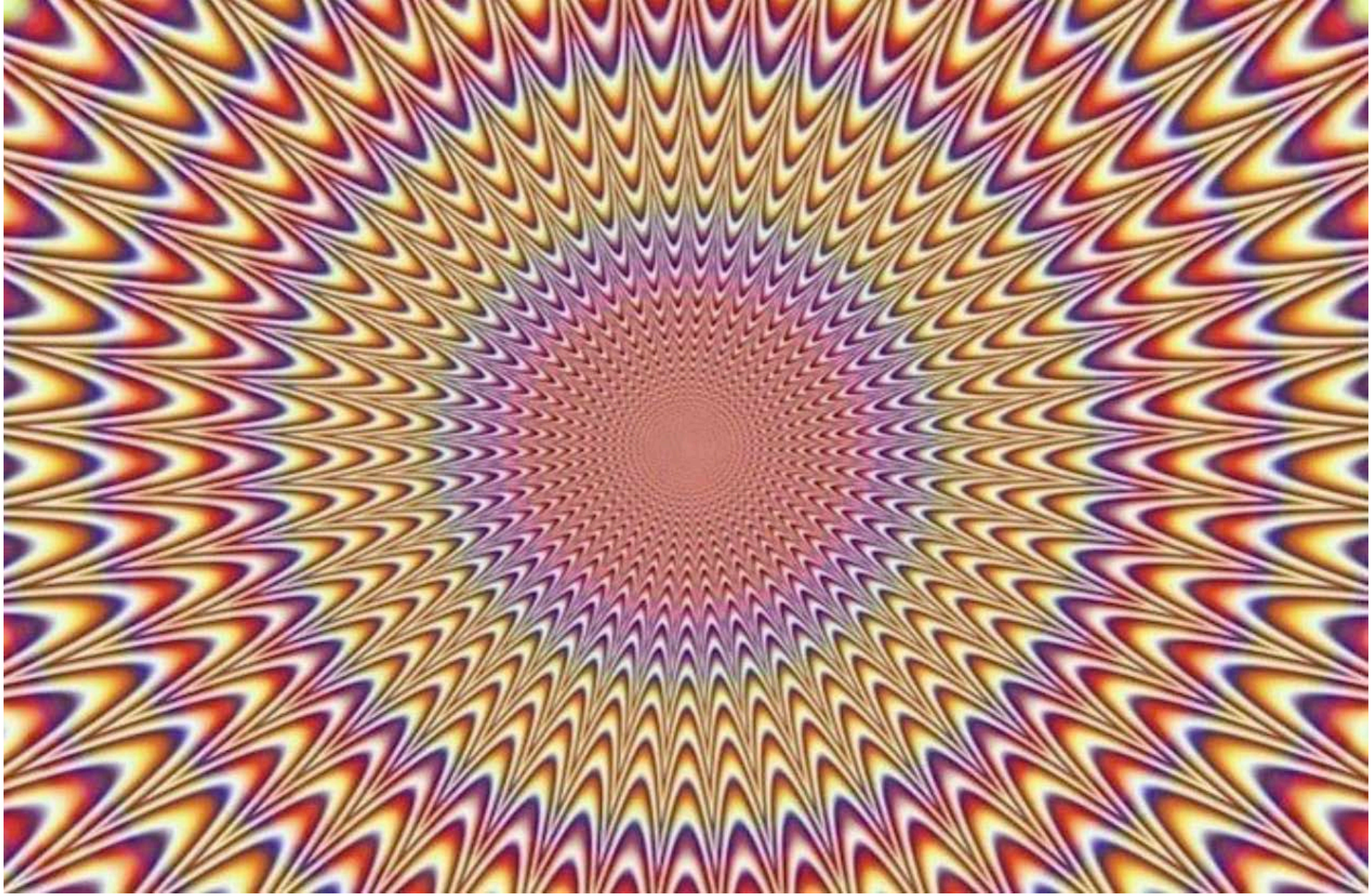


[This is a static image]

16.485: VNAV - Visual Navigation for Autonomous Vehicles

Luca Carlone

Lecture 27: Research Directions in SLAM

Today

https://arxiv.org/abs/1606.05830

1

(Past,)Present, and Future of Simultaneous Localization And Mapping: Towards the Robust-Perception Age

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif,
Davide Scaramuzza, José Neira, Ian Reid, John J. Leonard

Abstract*—Simultaneous Localization And Mapping (SLAM) consists in the concurrent construction of a model of the environment (the *map*), and the estimation of the state of the robot moving within it. The SLAM community has made astonishing progress over the last 30 years, enabling large-scale real-world applications, and witnessing a steady transition of this technology to industry. We survey the current state of SLAM and consider future directions. We start by presenting what is now the *de-facto* standard formulation for SLAM. We then review related work, covering a broad set of topics including robustness and scalability in long-term mapping, metric and semantic representations for mapping, theoretical performance guarantees, active SLAM and exploration, and other new frontiers. This paper simultaneously serves as a position paper and tutorial to those who are users of SLAM. By looking at the published research with a critical eye, we delineate open challenges and new research issues, that still deserve careful scientific investigation. The paper also contains the authors' take on two questions that often animate discussions during robotics conferences: *Do robots need SLAM?* and *Is SLAM solved?

***Index Terms*—Robots, SLAM, Localization, Mapping, Factor**

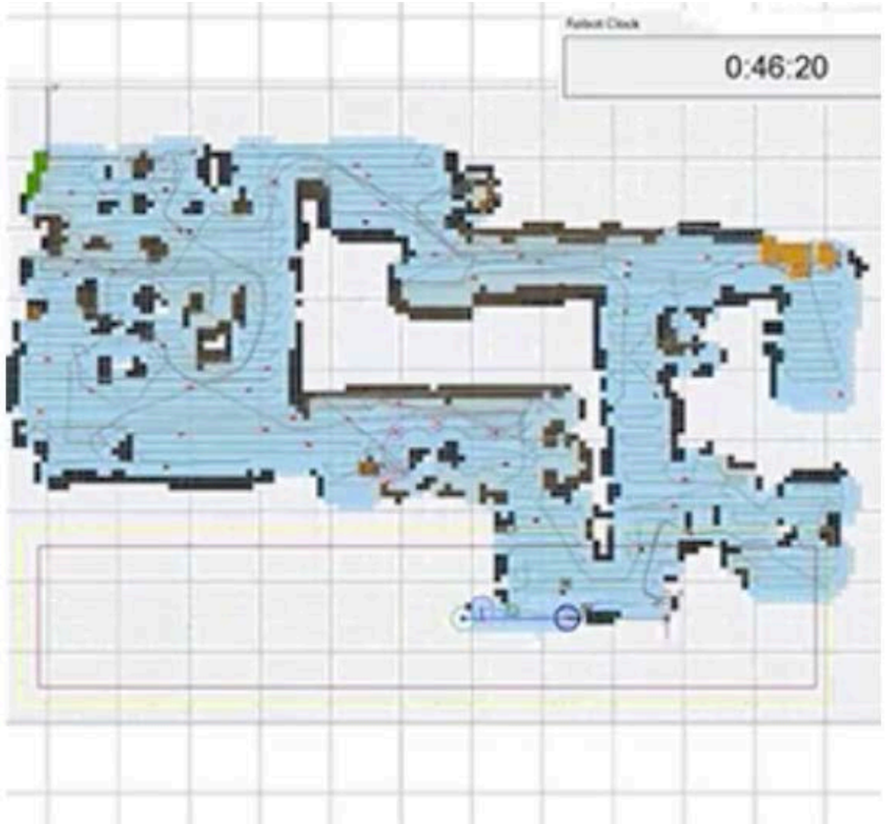
I. Introduction

SLAM comprises the simultaneous estimation of the state of a robot equipped with on-board sensors, and the construction of a model (the *map*) of the environment that the sensors are perceiving. In simple instances, the robot state is described by its pose (position and orientation), although other quantities may be included in the state, such as robot velocity, sensor biases, and calibration parameters. The map, on the other hand, is a representation of aspects of interest (e.g., position of landmarks, obstacles) describing the environment in which the robot operates.

The need to use a map of the environment is twofold. First, the map is often required to support other tasks; for instance, a map can inform path planning or provide an intuitive visualization for a human operator. Second, the map allows limiting the error committed in estimating the state of the robot. In the absence of a map, dead-reckoning would

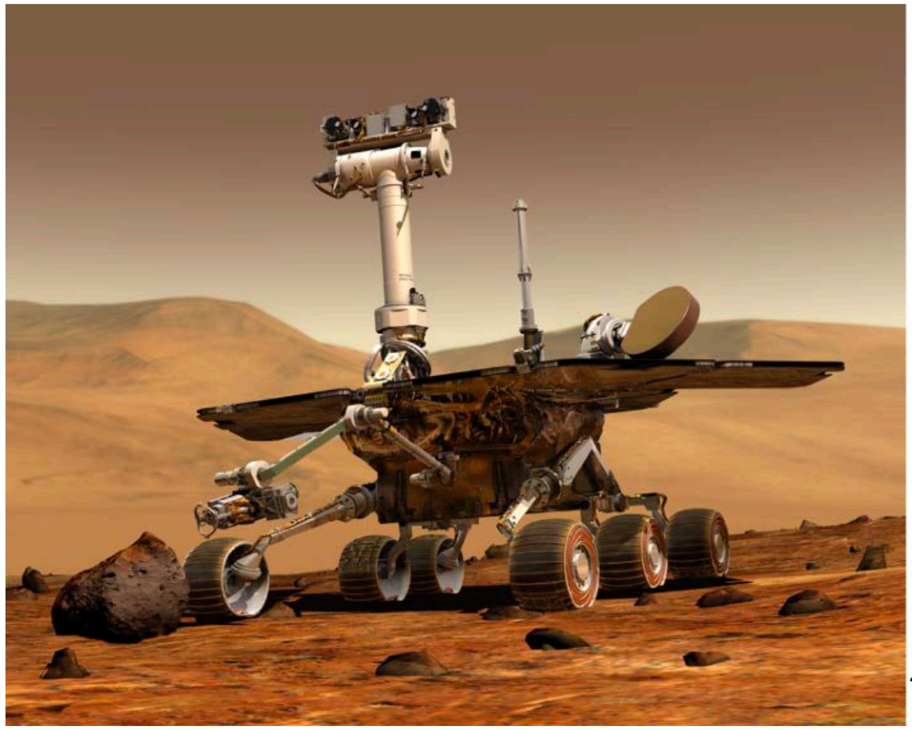
SLAM & SfM: Engineered Solutions / Applications

Roomba 980
Vacuum Cleaner



Kuka's Navigation Solution

Mars
Rovers
(VO)



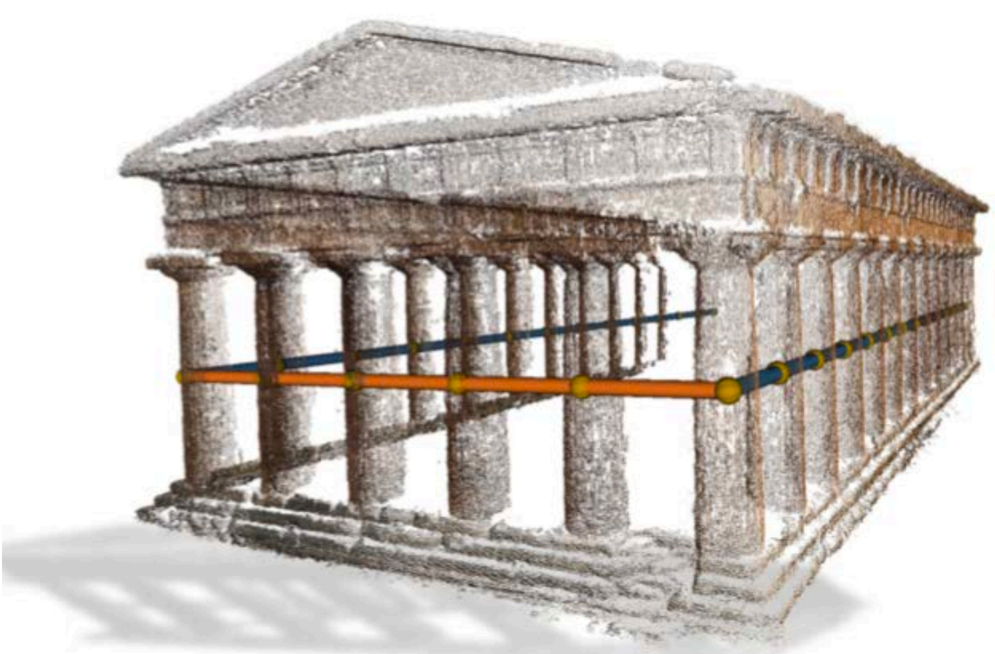
Source: public domain.

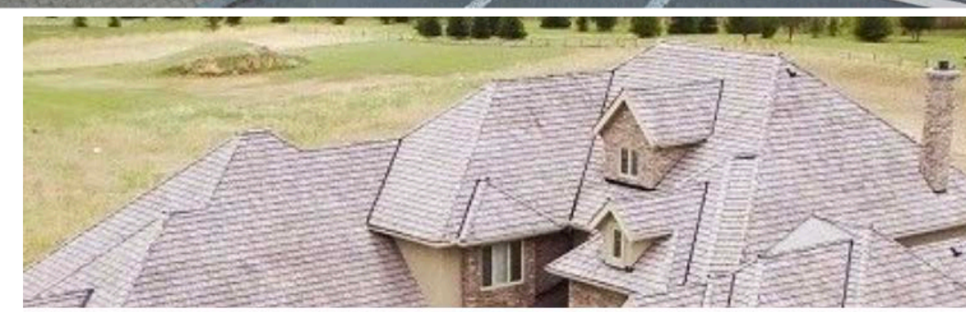
SLAM & SfM: Engineered Solutions / Applications

Precision agriculture

Google street view

Monitoring of historical sites





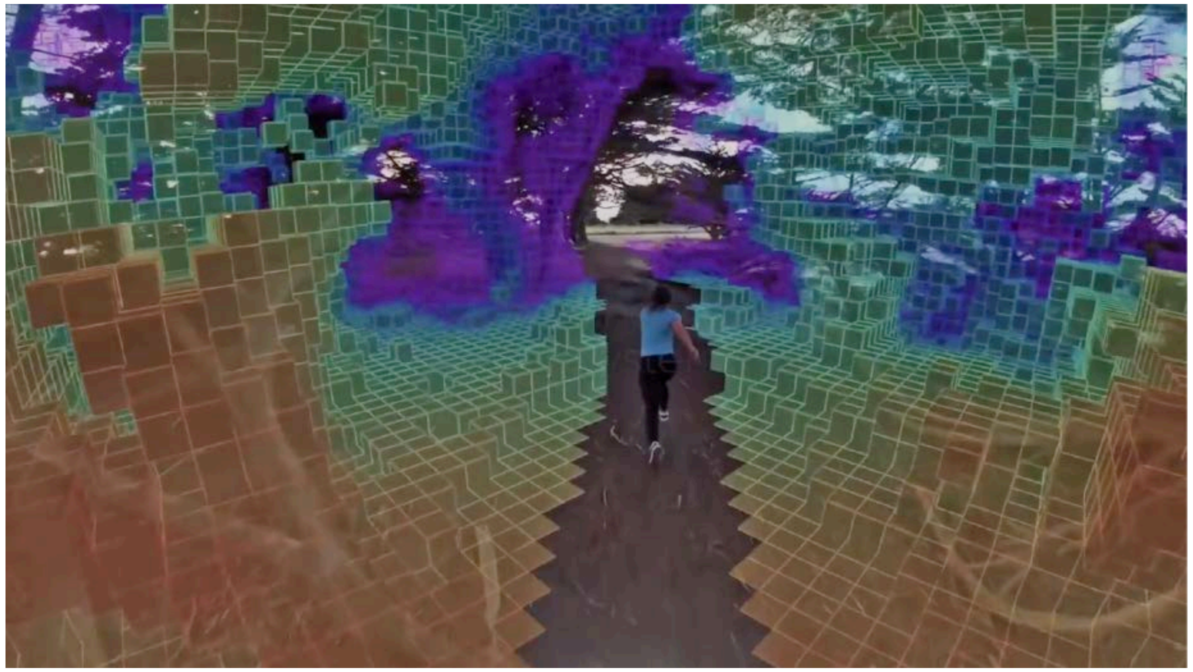
Selling high-end property with drone mapping

USE CASES   REAL ESTATE   PIX4DMAPPER   3D MODELING
MAPPING
22 OCTOBER 2015

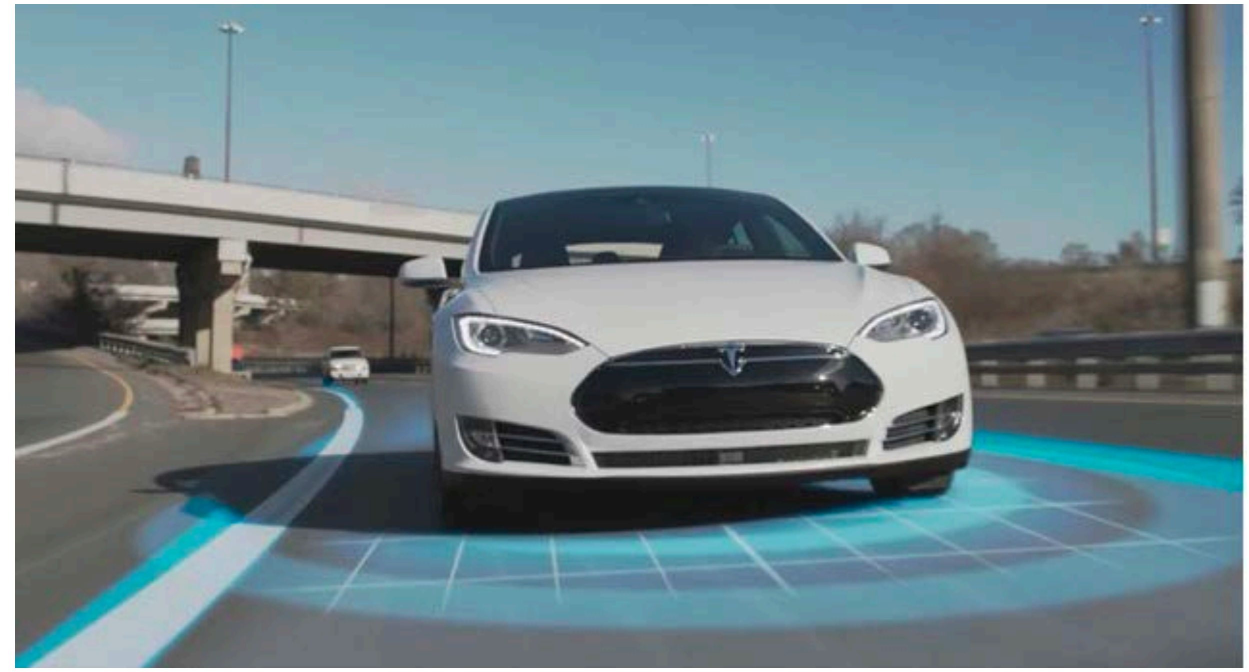
Pix4Dmapper is used to create a comprehensive 3D reconstruction of a luxury house for potential real estate clients.

SLAM & SfM: Engineered Solutions / Applications

Skydio R1 drone



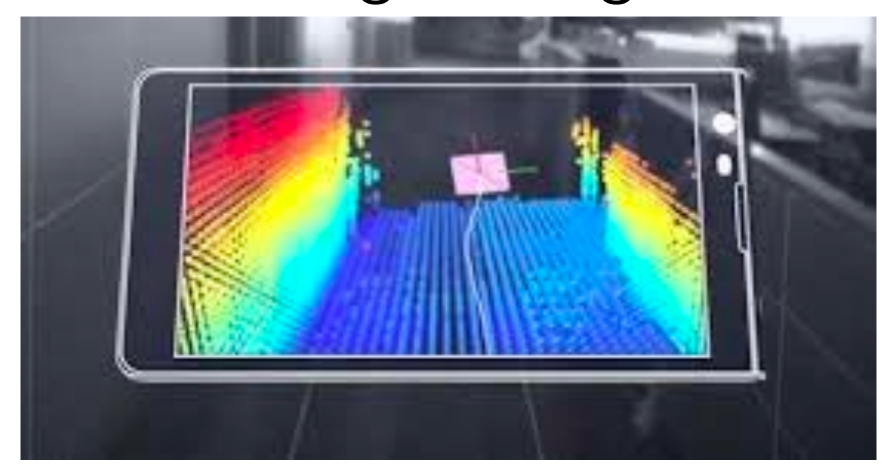
Tesla's autopilot



Oculus Rift Goggles

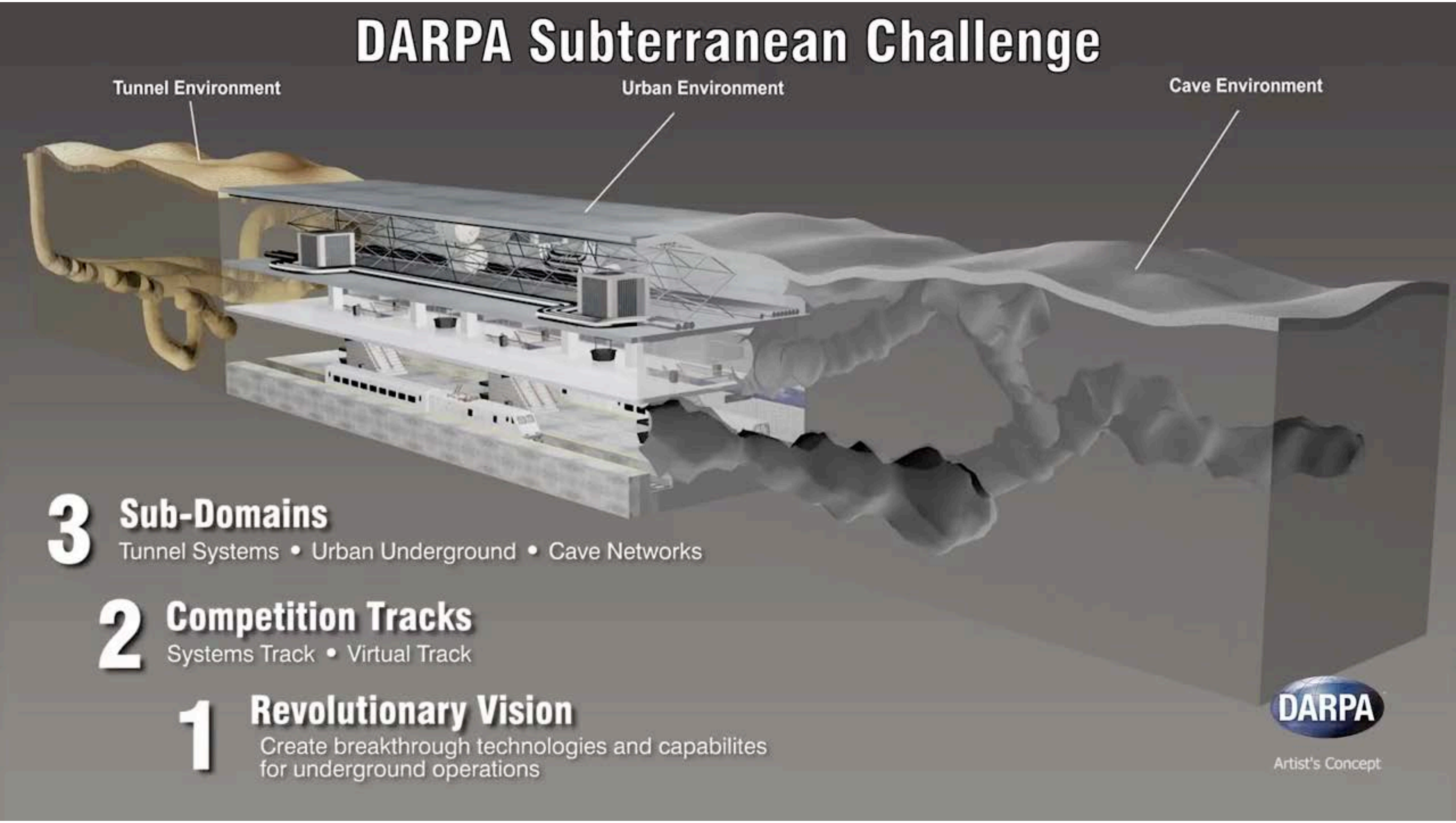
Pokemon Go

Google Tango



Reinvented as ARCore in 2017

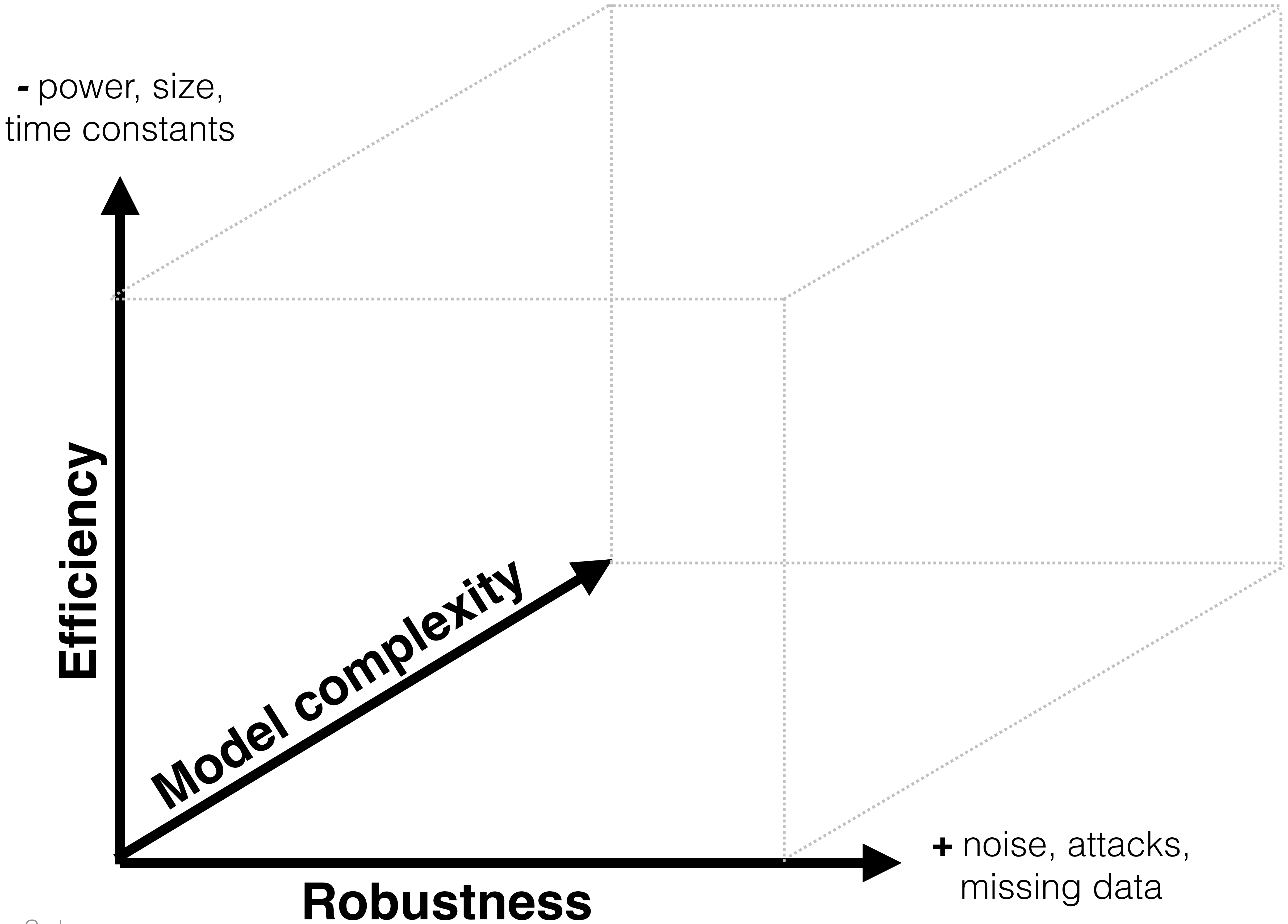
SLAM & SfM: Engineered Solutions / Applications



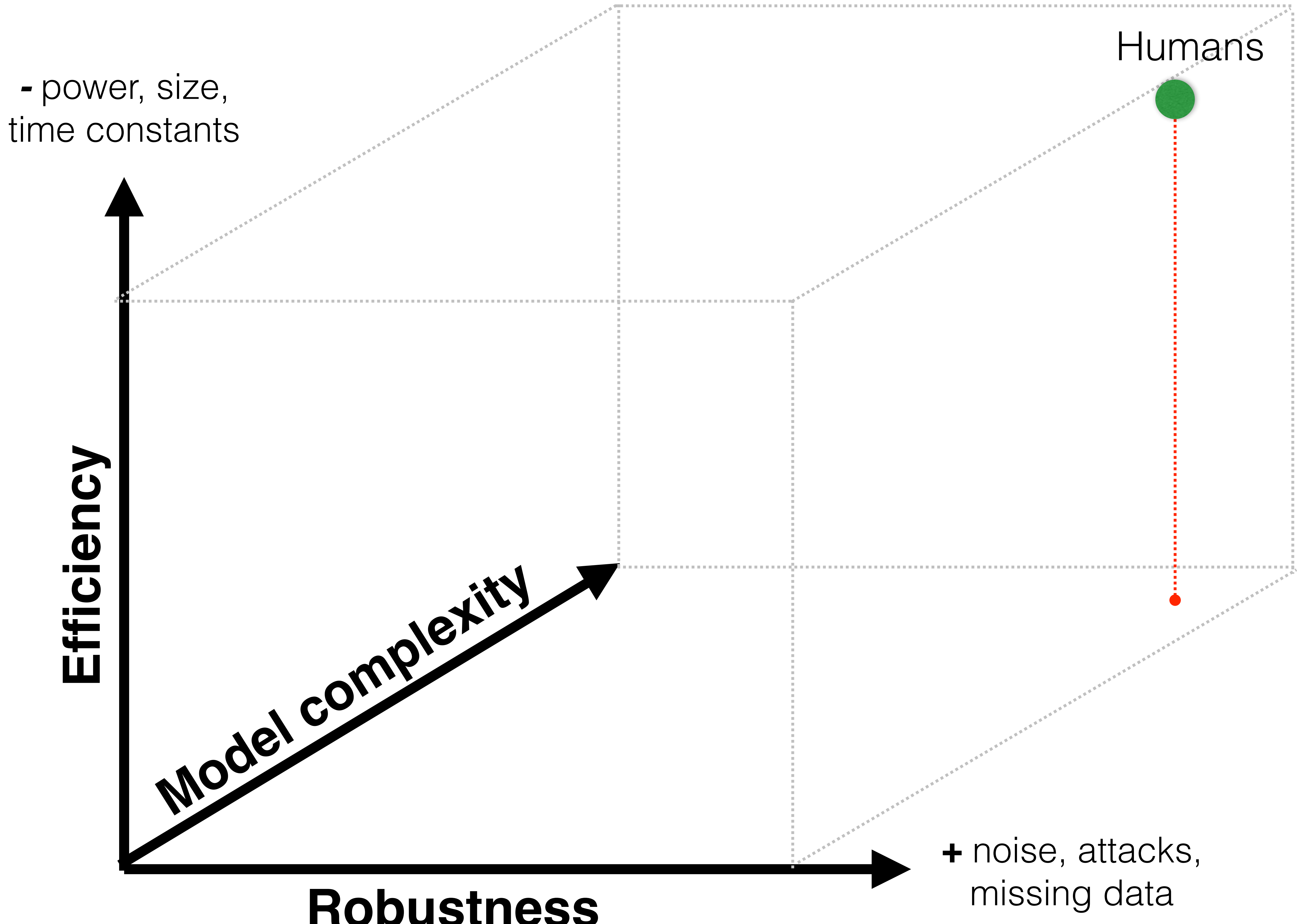
DARPA Subterranean Challenge

Source: public domain

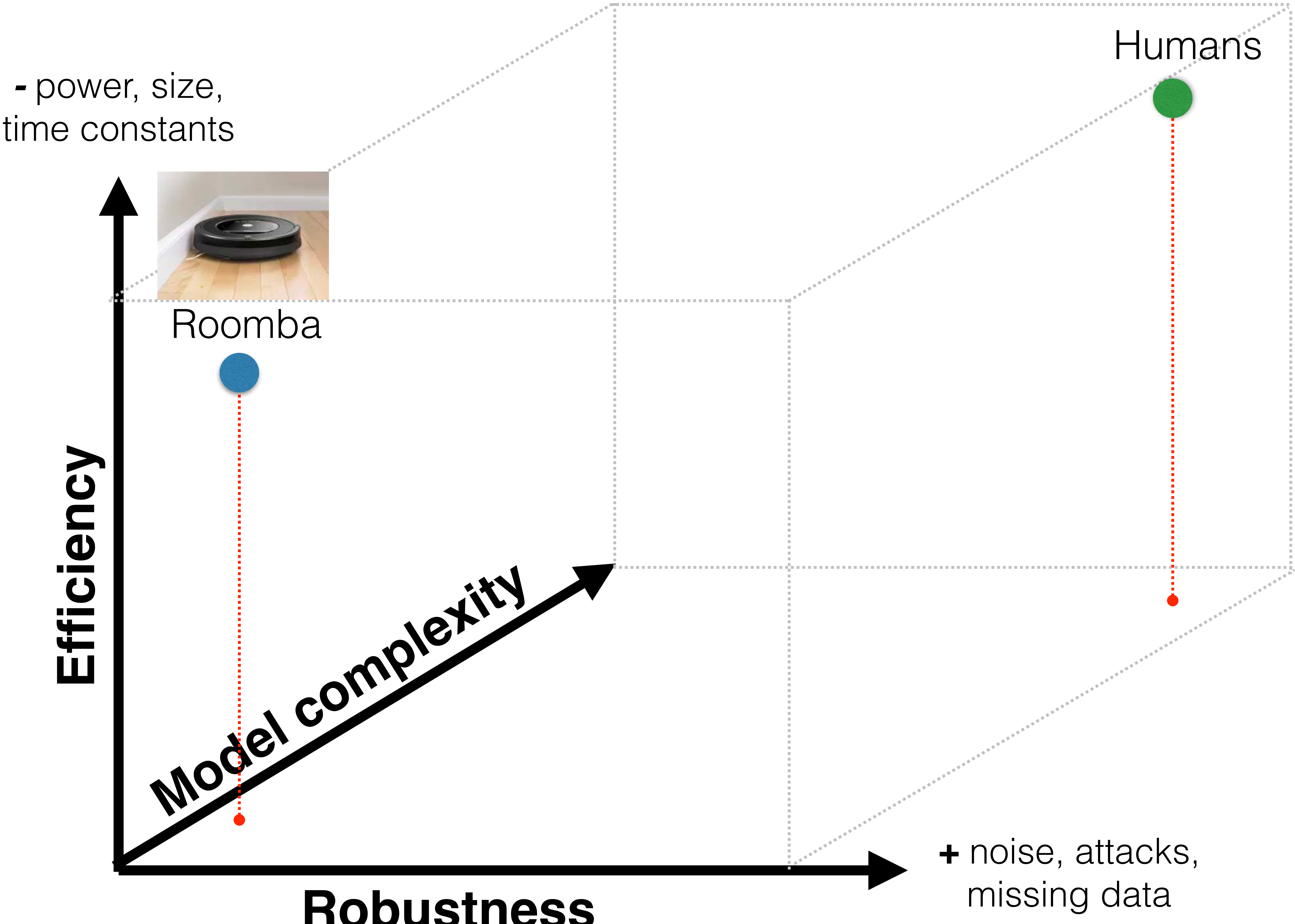
Axes of complexity



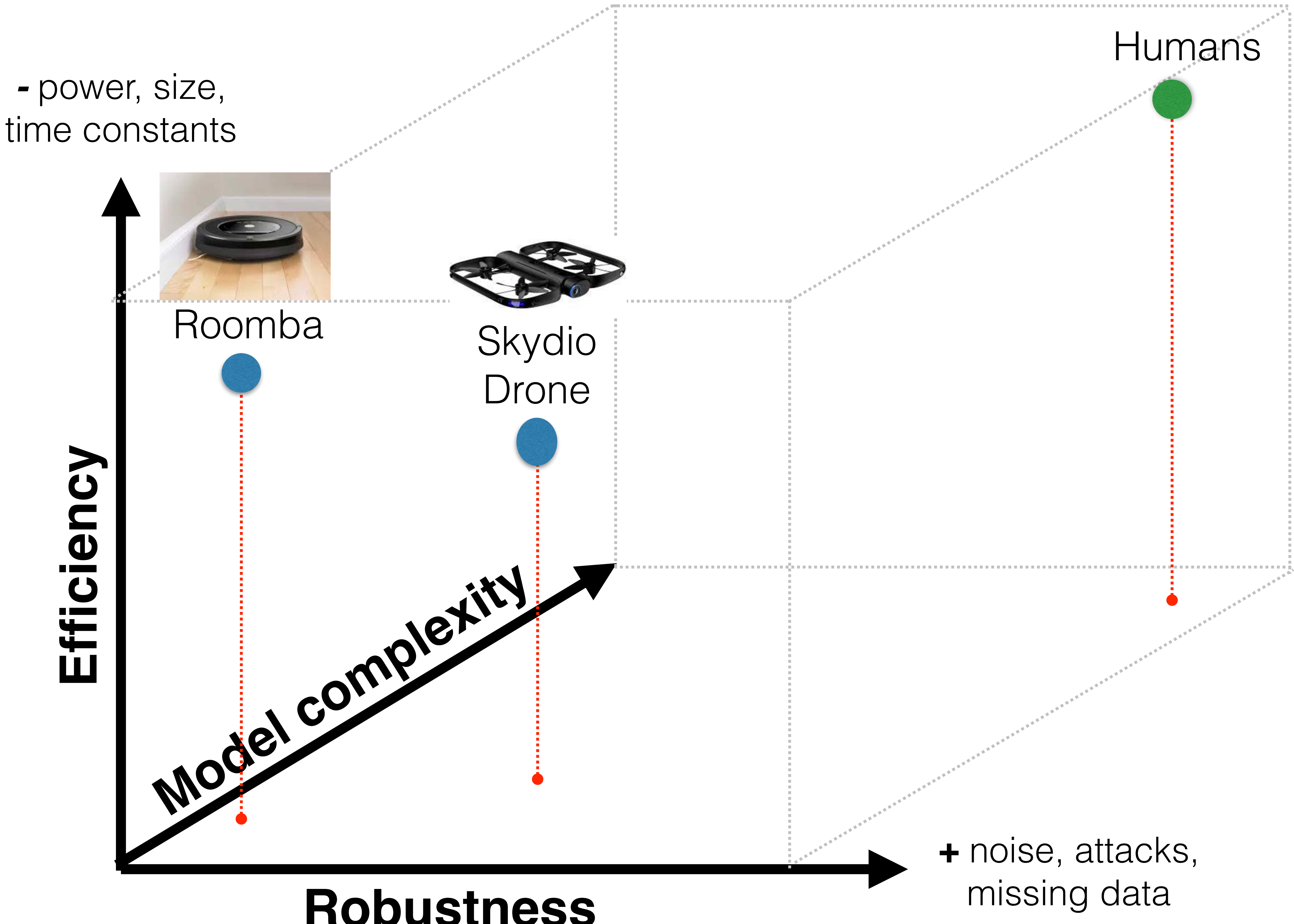
Axes of complexity



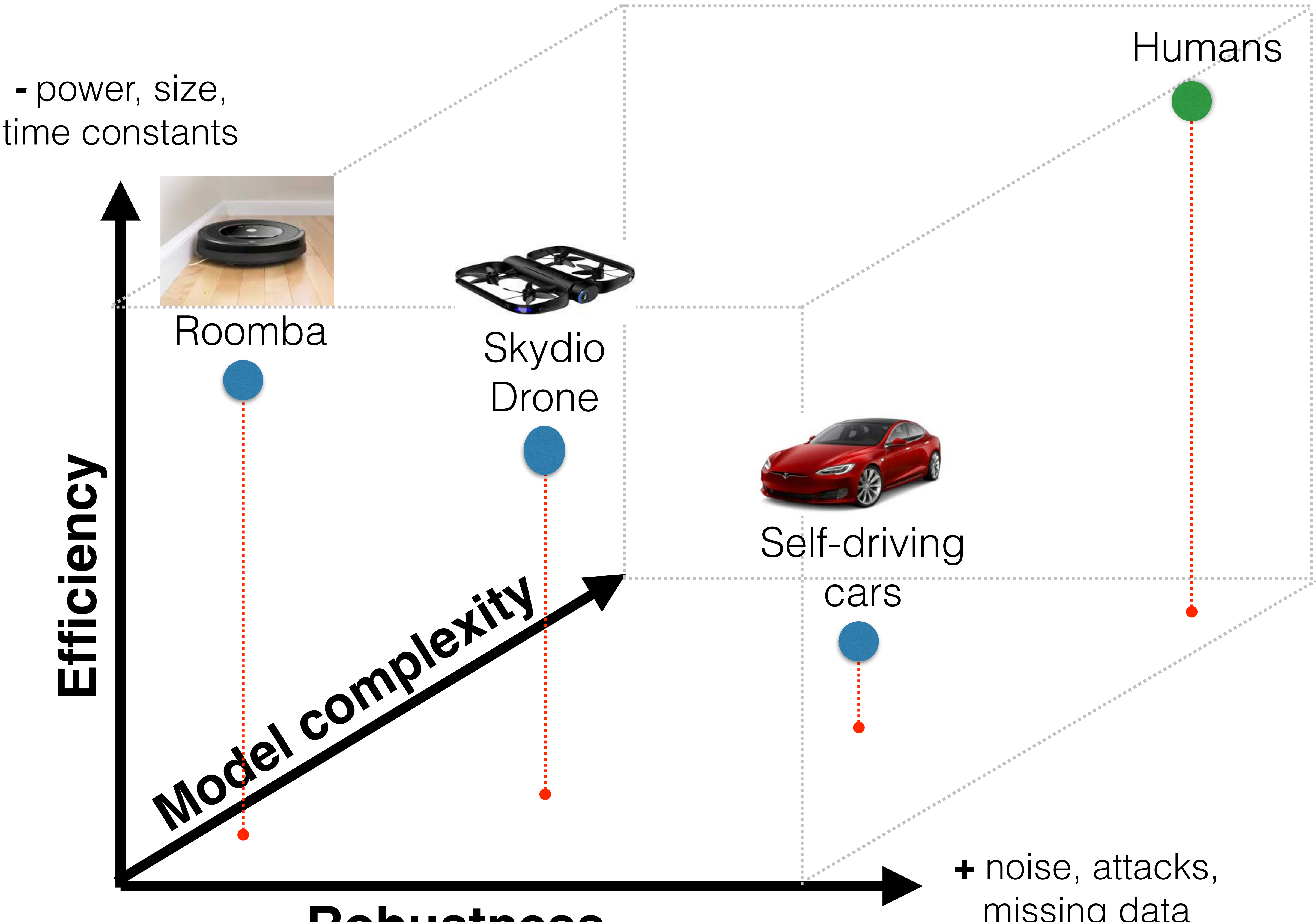
Axes of complexity



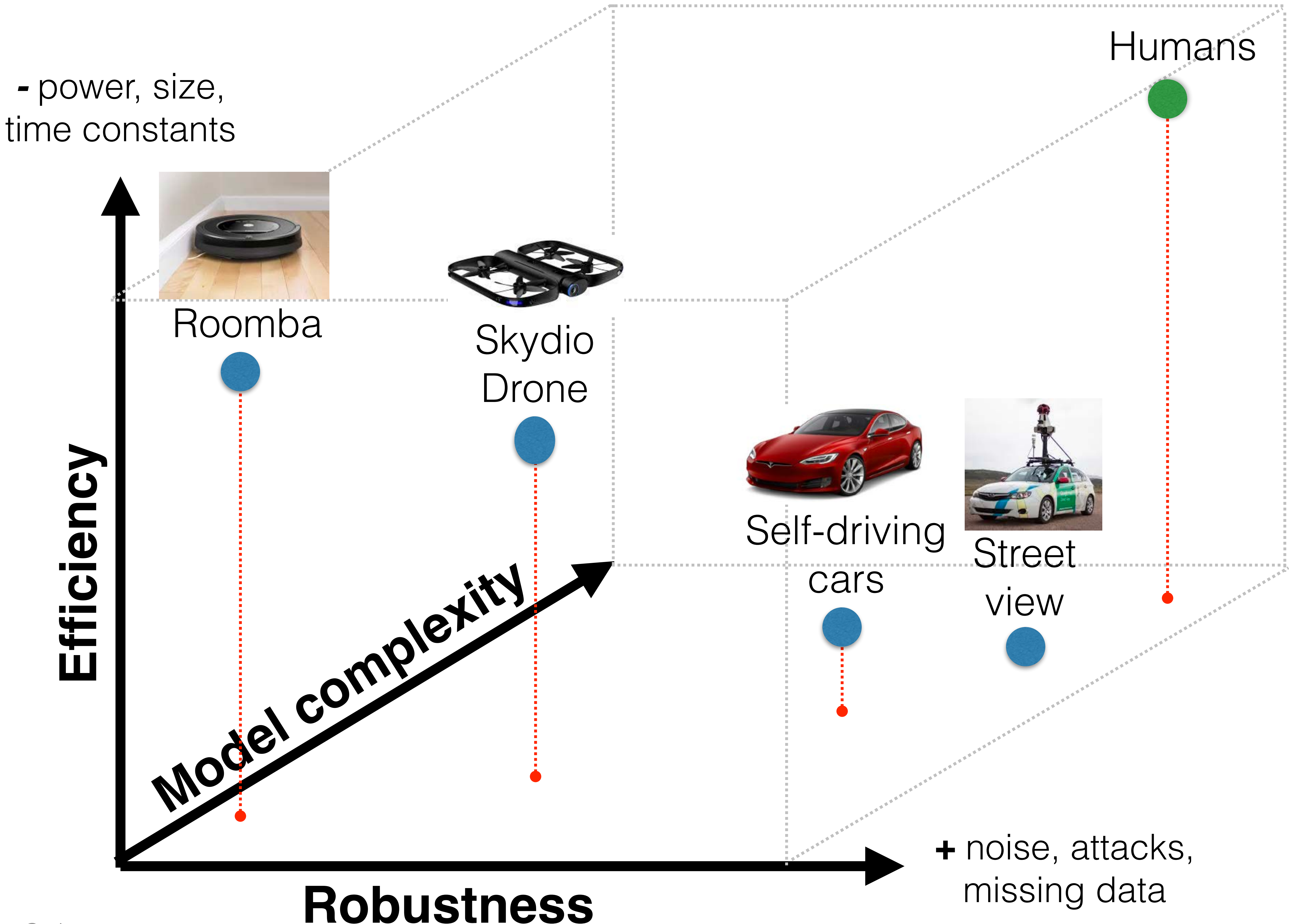
Axes of complexity



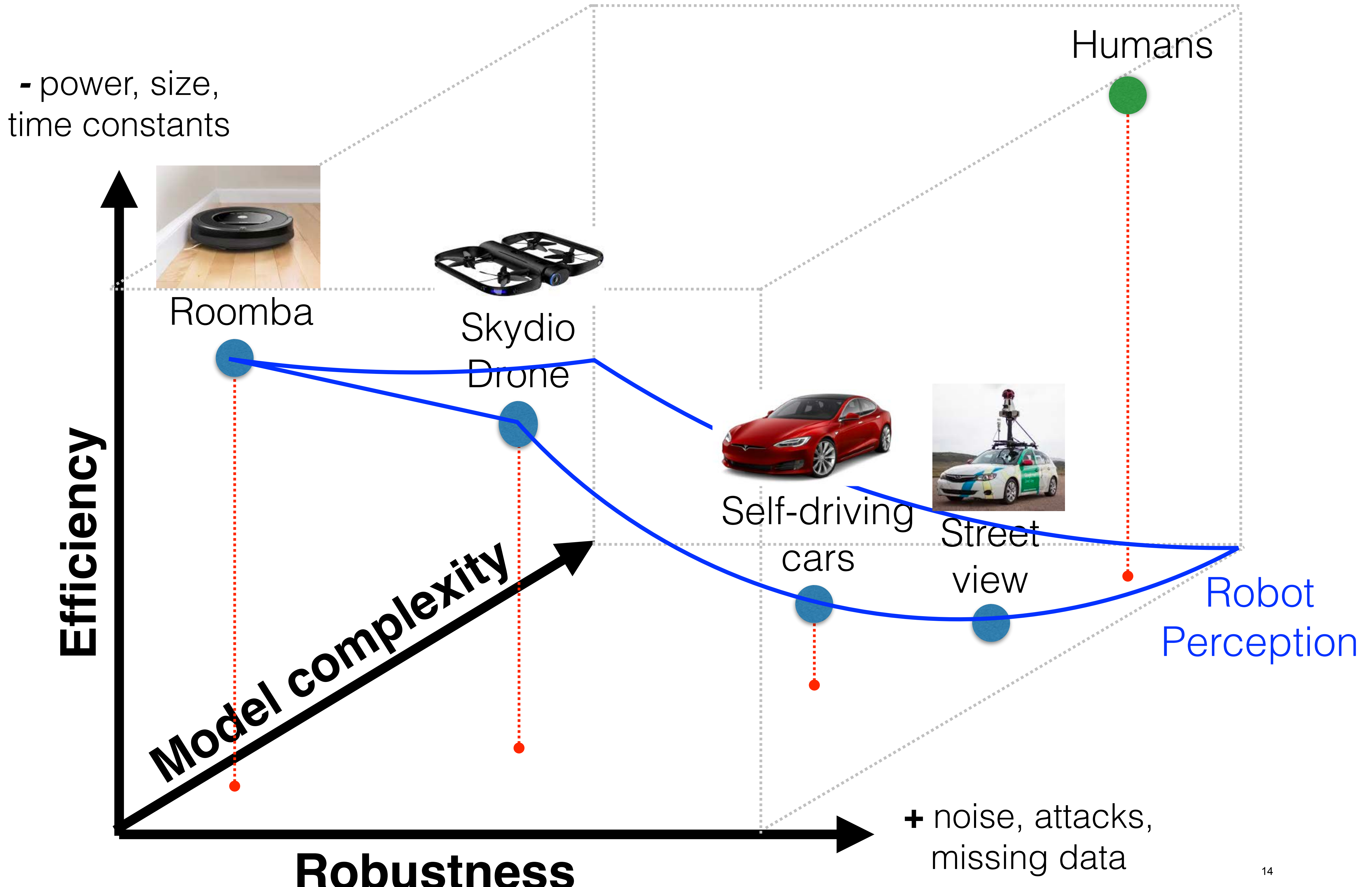
Axes of complexity



Axes of complexity

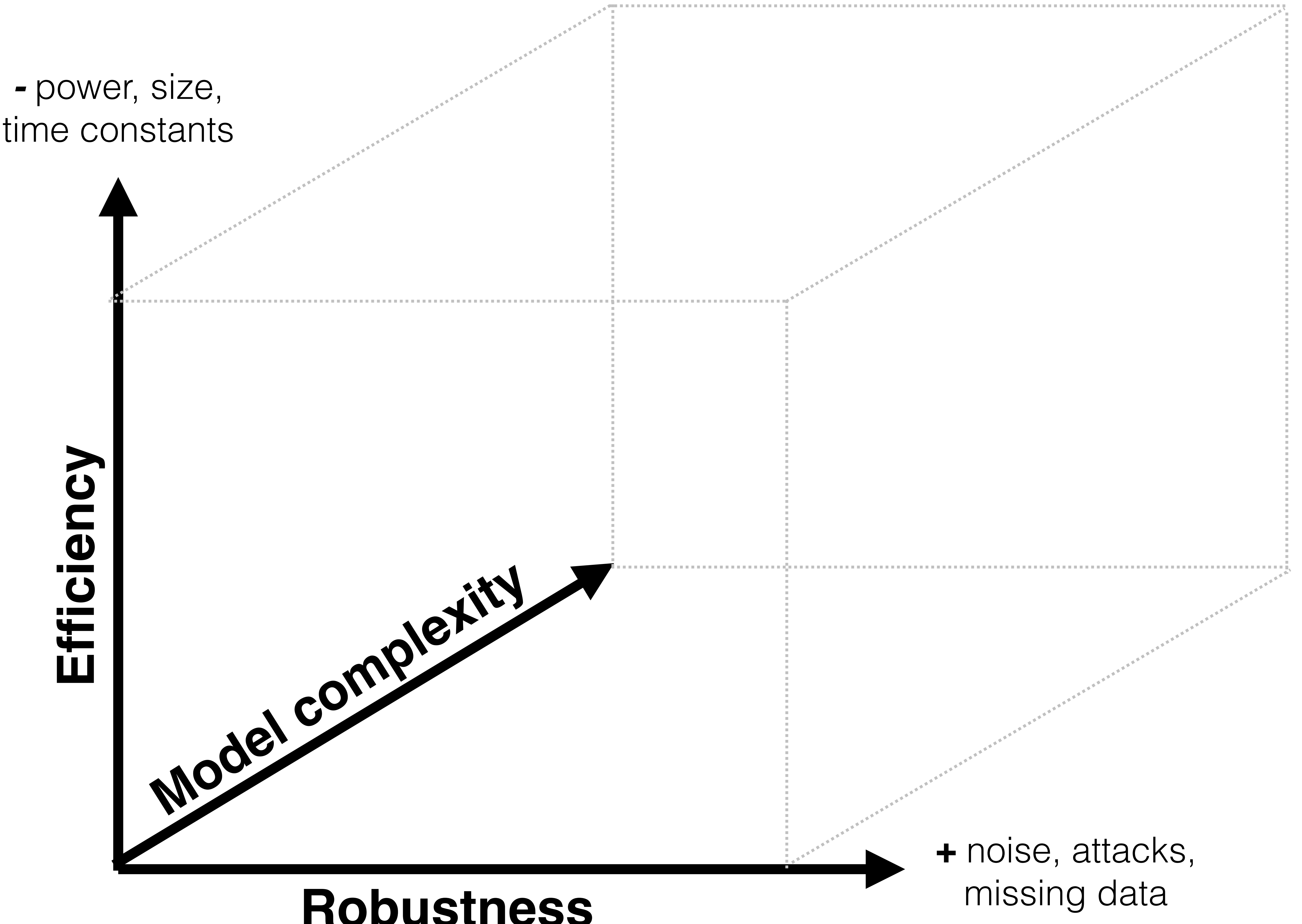


Axes of complexity

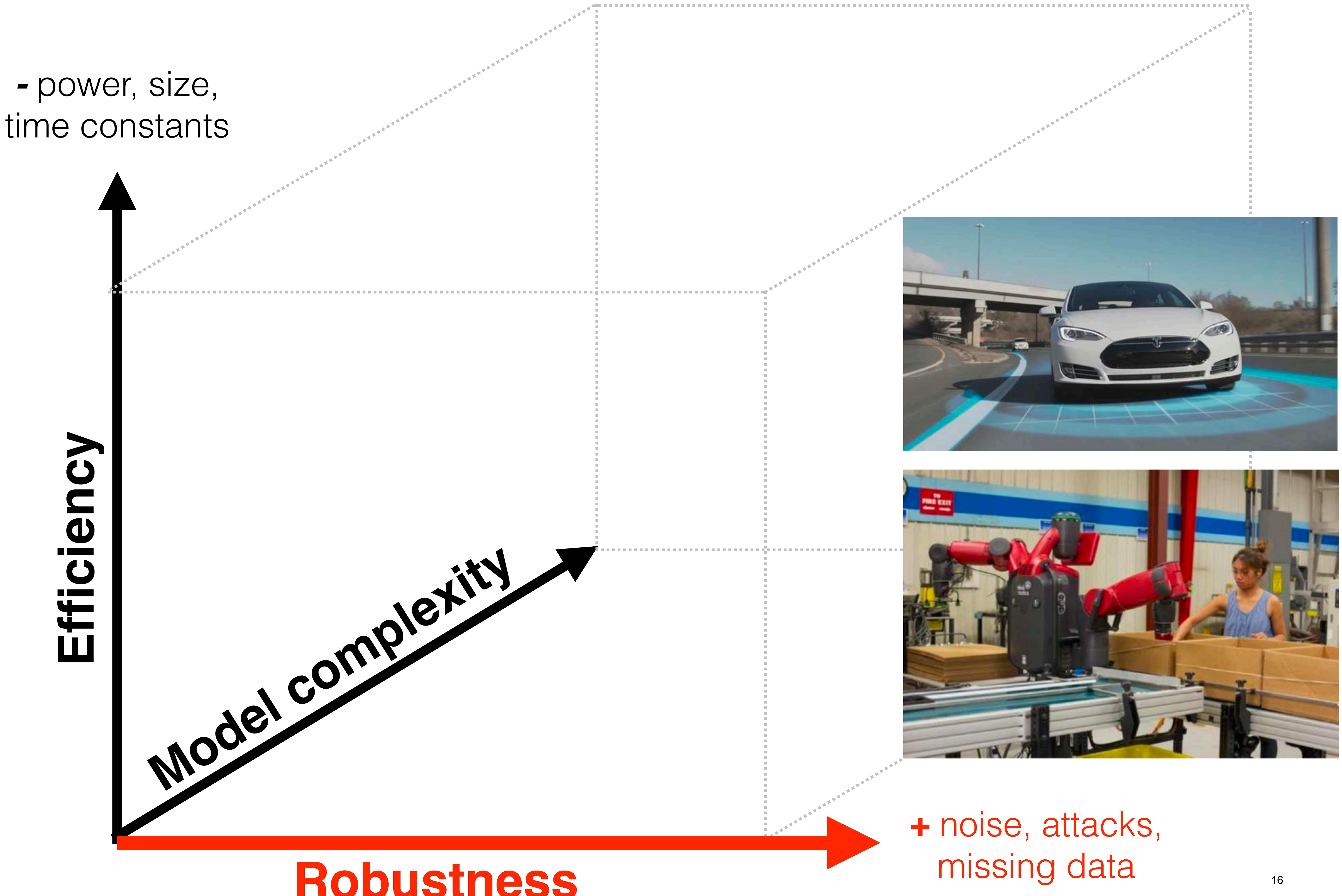


Luca Carlone

Active Research Directions



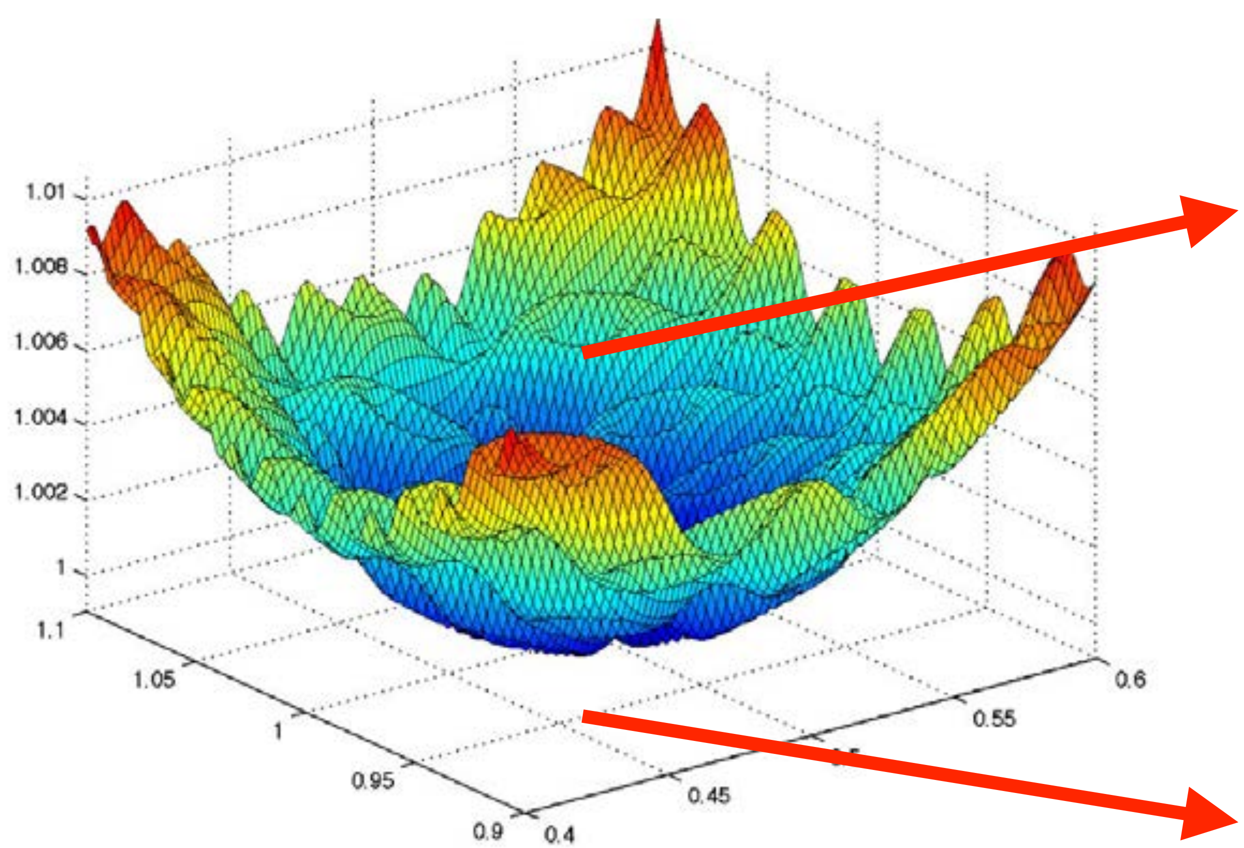
Active Research Directions

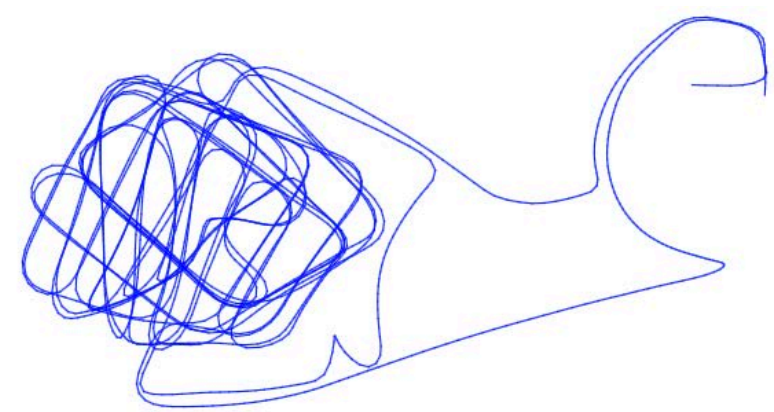


Robustness to noise

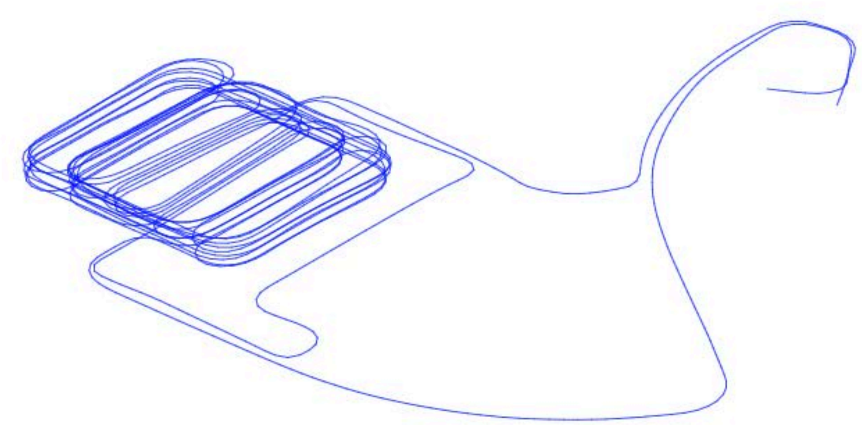
$$\min_{\substack{\{p_i \in \mathbb{R}^3\} \\ \{R_i \in \mathrm{SO}(3)\}}} \sum_{(i,j)\in\mathcal{E}} \frac{1}{\sigma_p^2} \left\|\bar{p}_{ij} - R_i^\mathsf{T}(p_j - p_i)\right\|^2 + \frac{1}{\sigma_R^2} \left\|\bar{R}_{ij} - R_i^\mathsf{T} R_j\right\|_\mathsf{F}^2$$

non-convex optimization





Suboptimal critical point



Optimal estimate

Iterative methods (e.g., gradient descent) may get stuck into bad minima

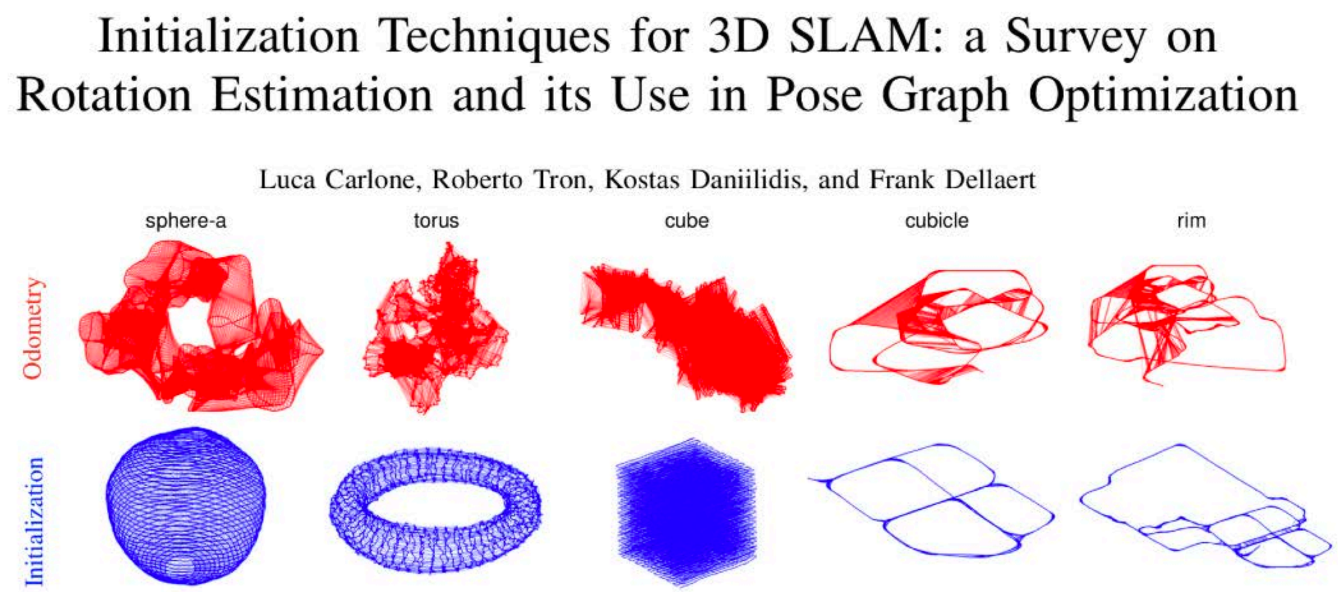
Initial guess gets worse when noise is large

17

Robustness to noise (convergence)

- **Analysis:** number of minima,
  basin of attraction of iterative solvers (Gauss-Newton),
  factors impacting quality of solution

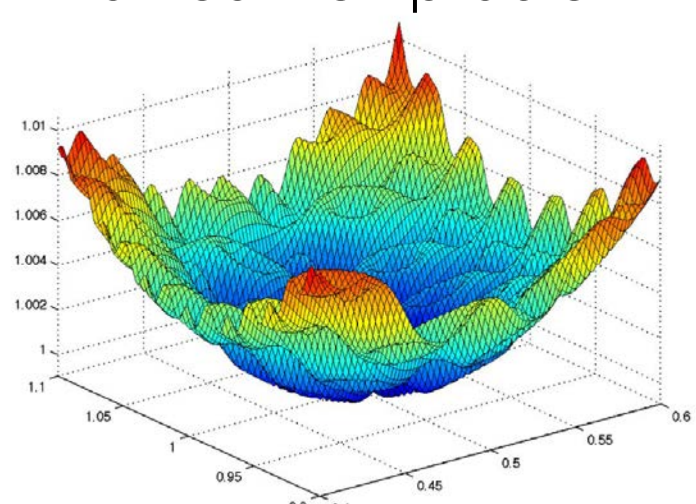
- **Initialization Techniques**


Initialization Techniques for 3D SLAM: a Survey on Rotation Estimation and its Use in Pose Graph Optimization

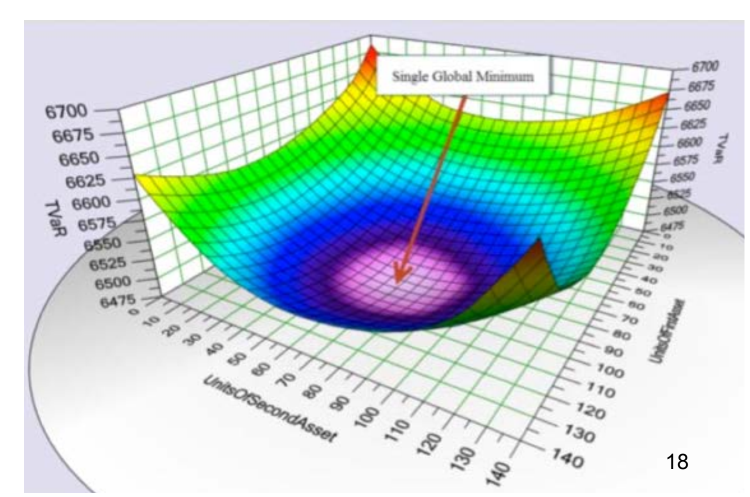
Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert

- **Global Solvers**

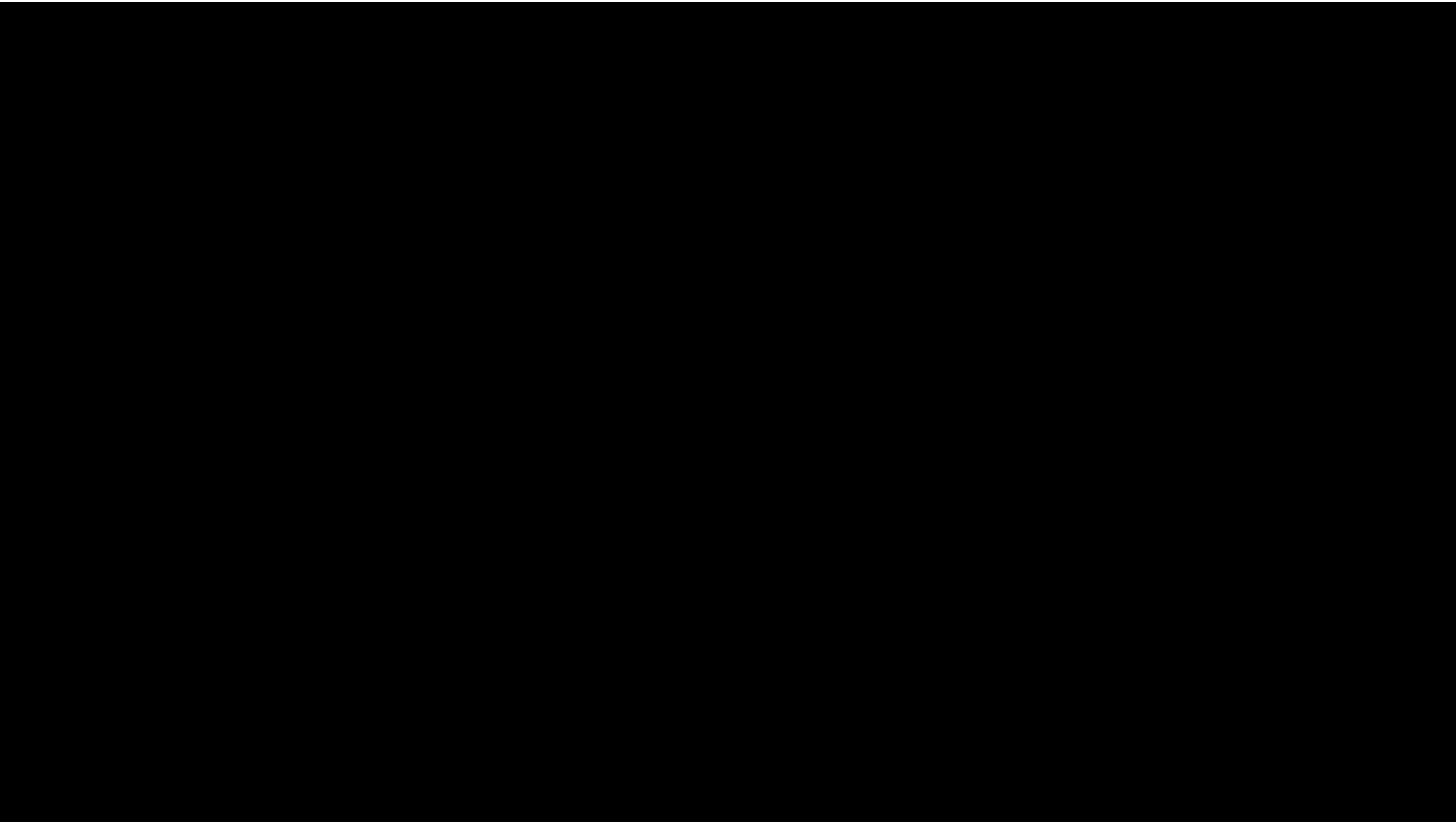
non convex problem



convex relaxation

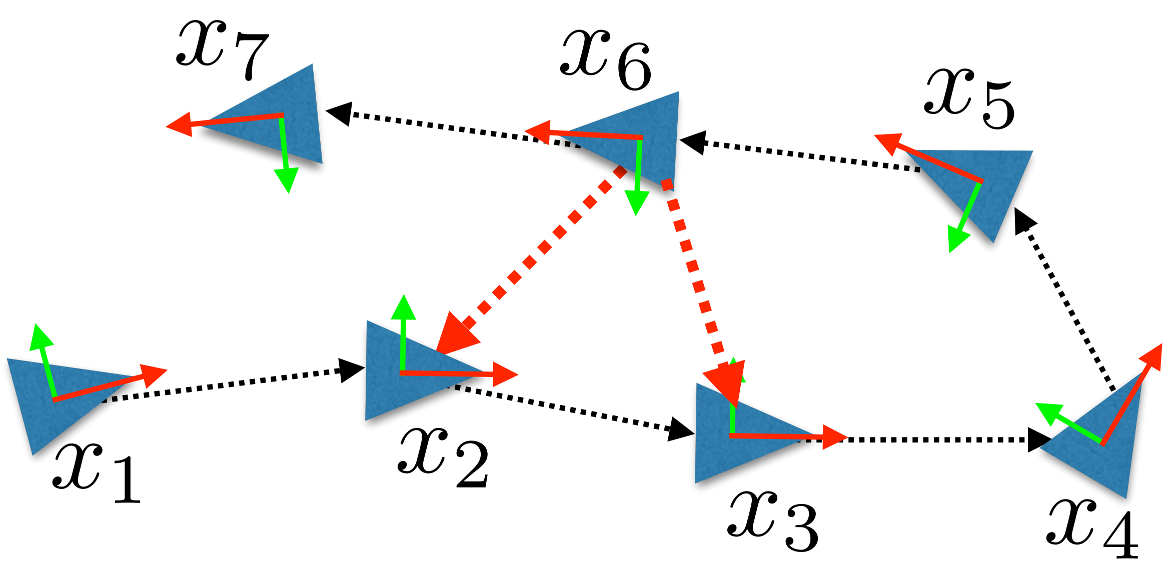


What if place recognition fails?



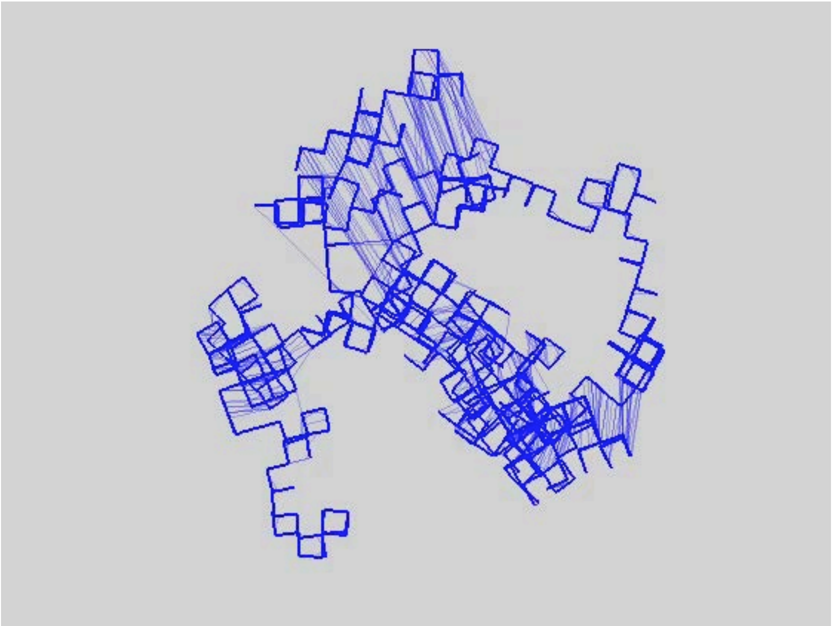
19

What if place recognition fails?

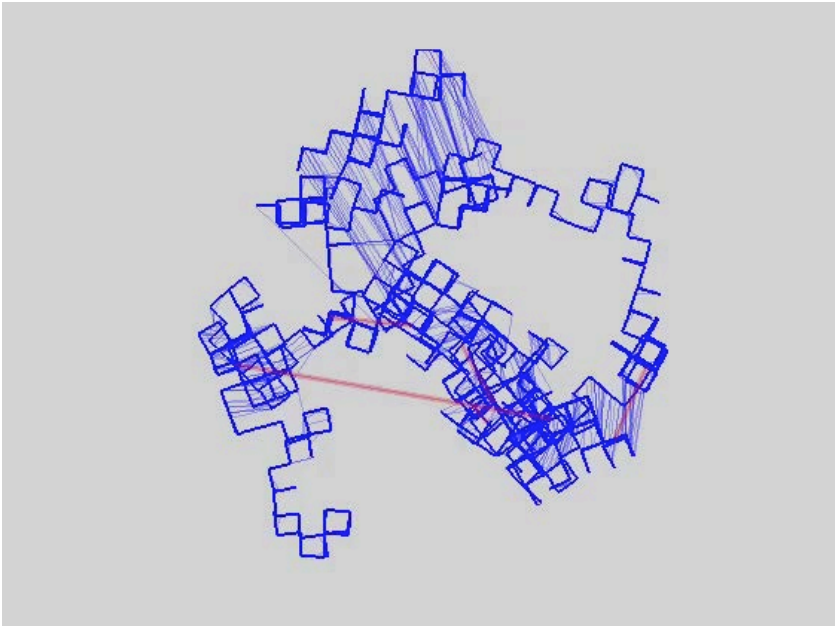


outliers: completely incorrect measurements (Perceptual Aliasing)

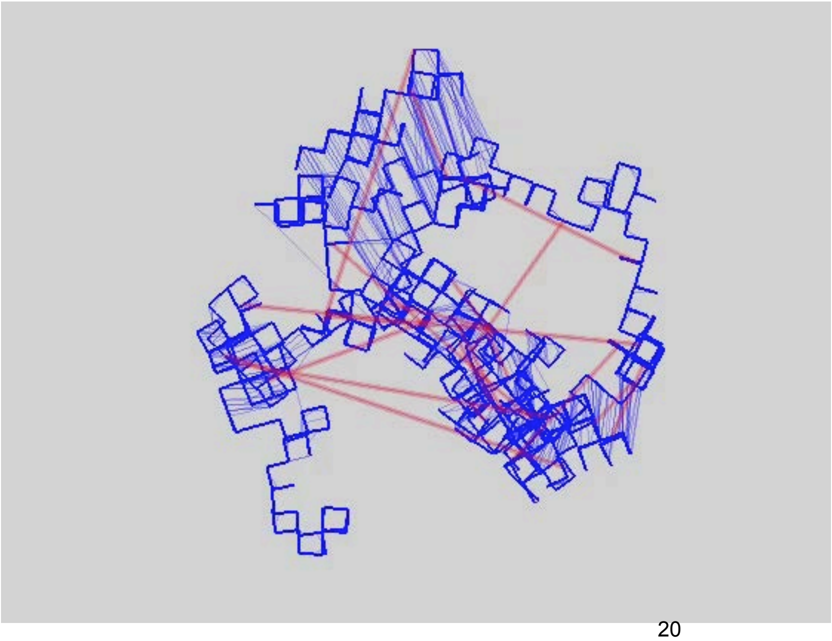
0 outliers



5 outliers

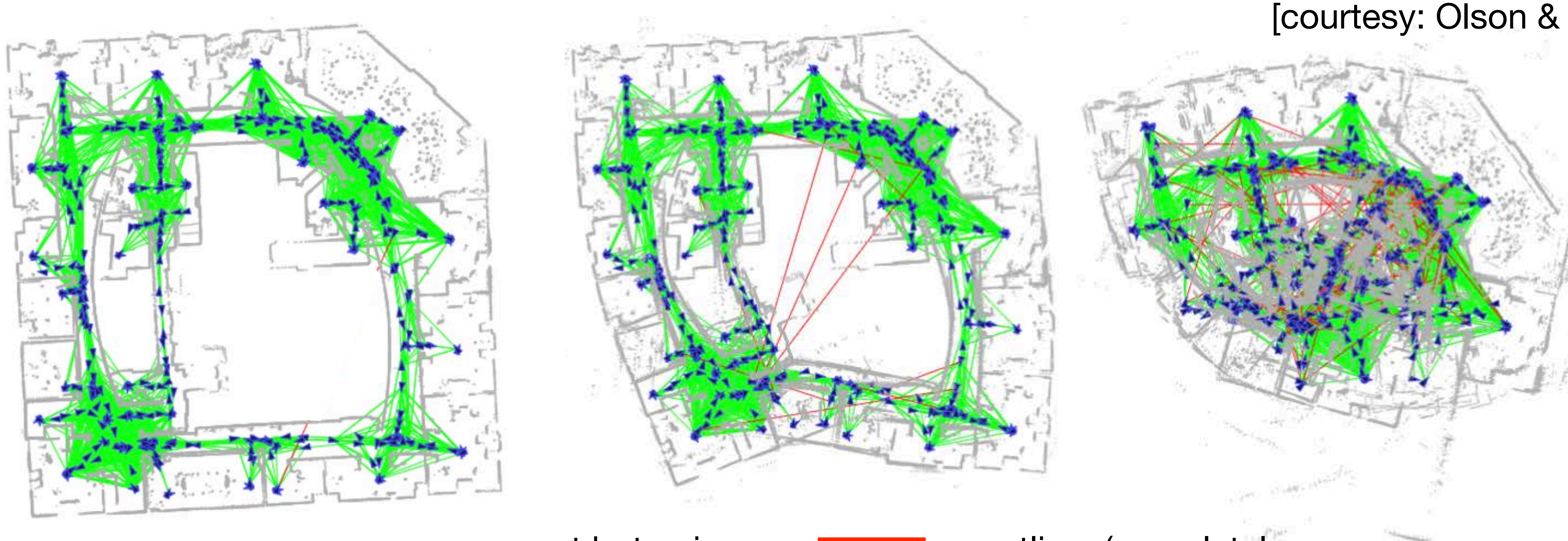


20 outliers



Robustness to outliers

[courtesy: Olson & Agarwal]



correct but noisy measurements

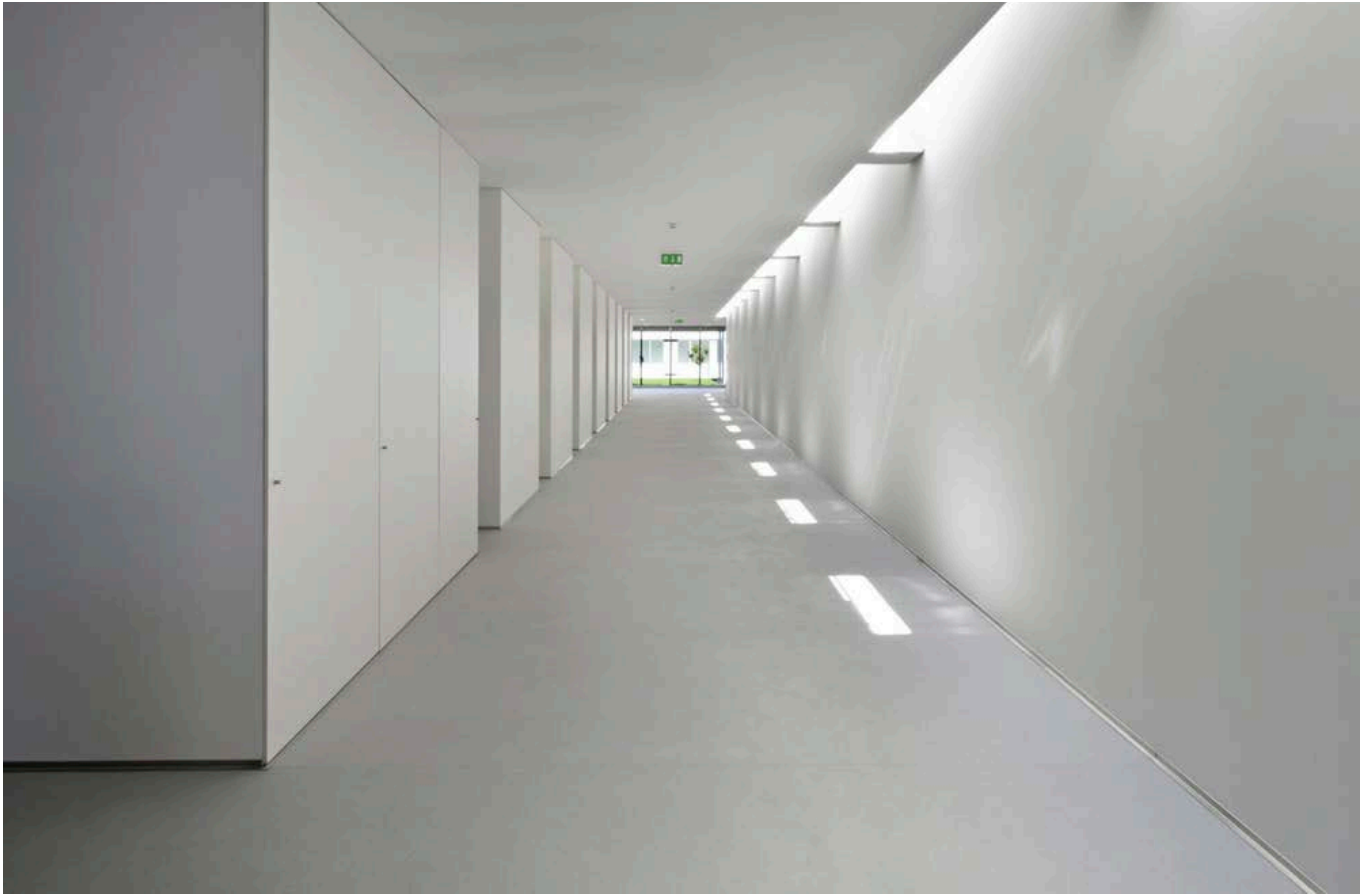
outliers (completely wrong measurements)

least-square SLAM estimators catastrophically fail if outliers are not carefully handled

"Google employs a small army of human operators to manually check and correct the maps" [Wired]

Robustness to dynamic scenes

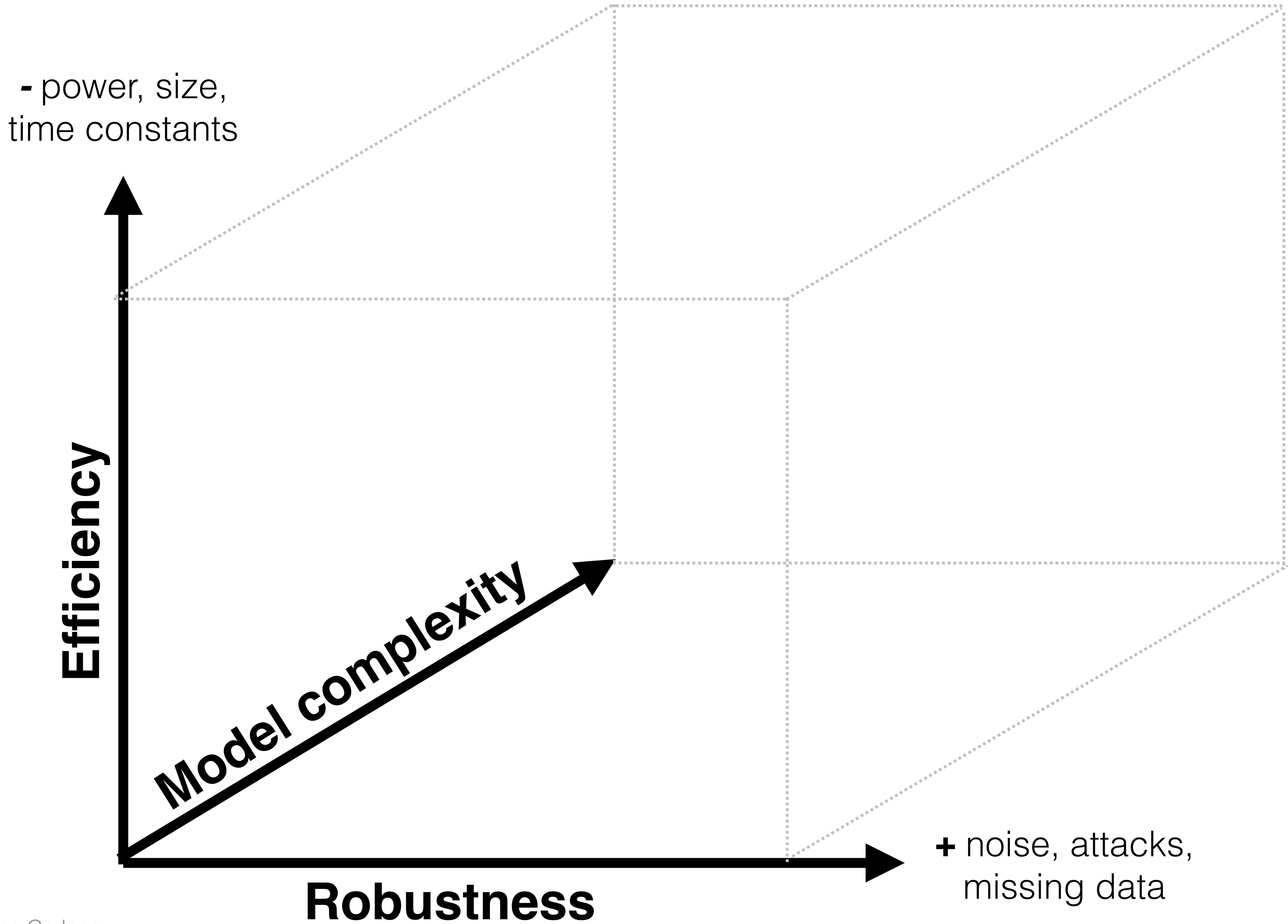
Robustness to missing data



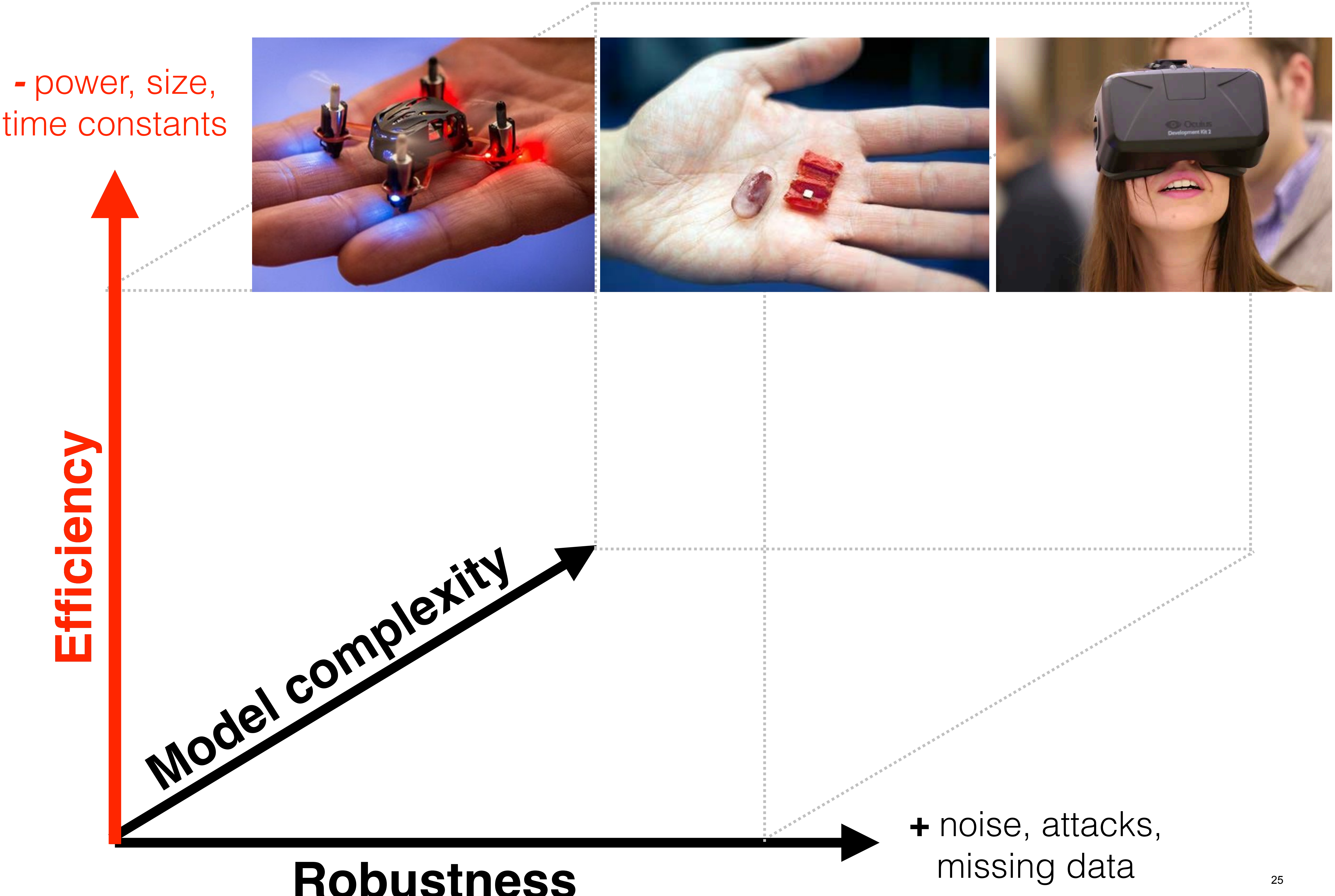
[Zhang, Kaess, Singh, On Degeneracy of Optimization-based State Estimation]

23

Active Research Directions



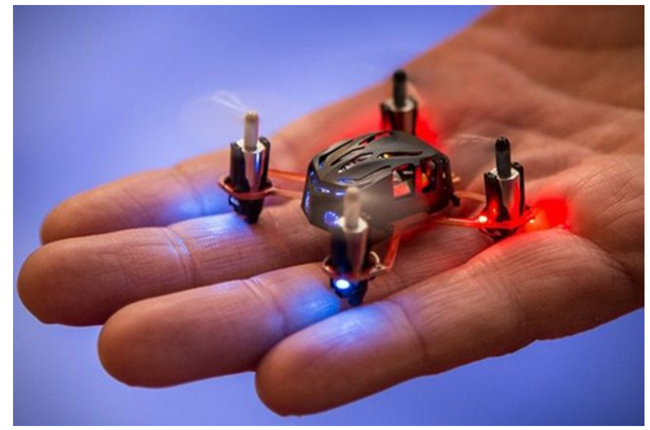
Active Research Directions



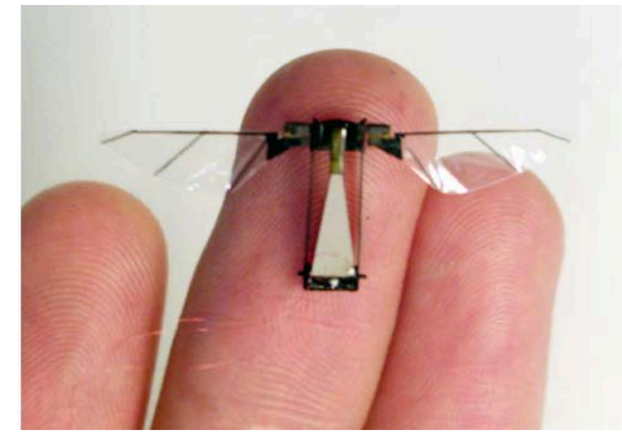
Efficiency and Miniaturization

>10 W

< 5 W



< 200 mW

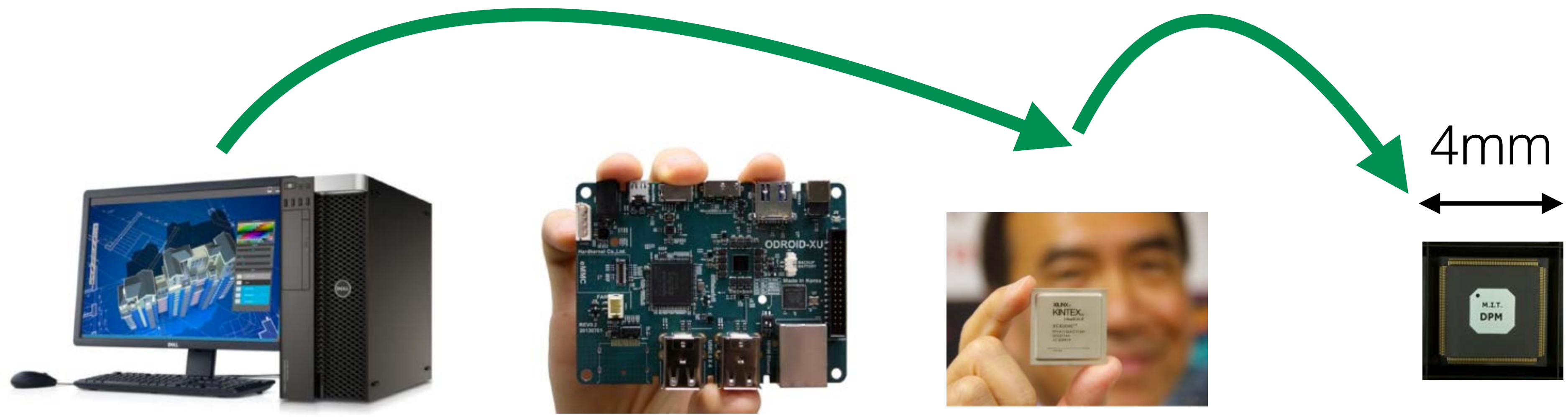


Human vision

Machine vision

Data stream	10^8 - 10^9 bits/second	$5 \cdot 10^8$ bits/second (stereo)
Performance	parse scene: 13ms	object detection: 22ms (GPU) SLAM: >100ms
Power	20W	250 W (Titan X GPU)

Algorithms-and-hardware co-design



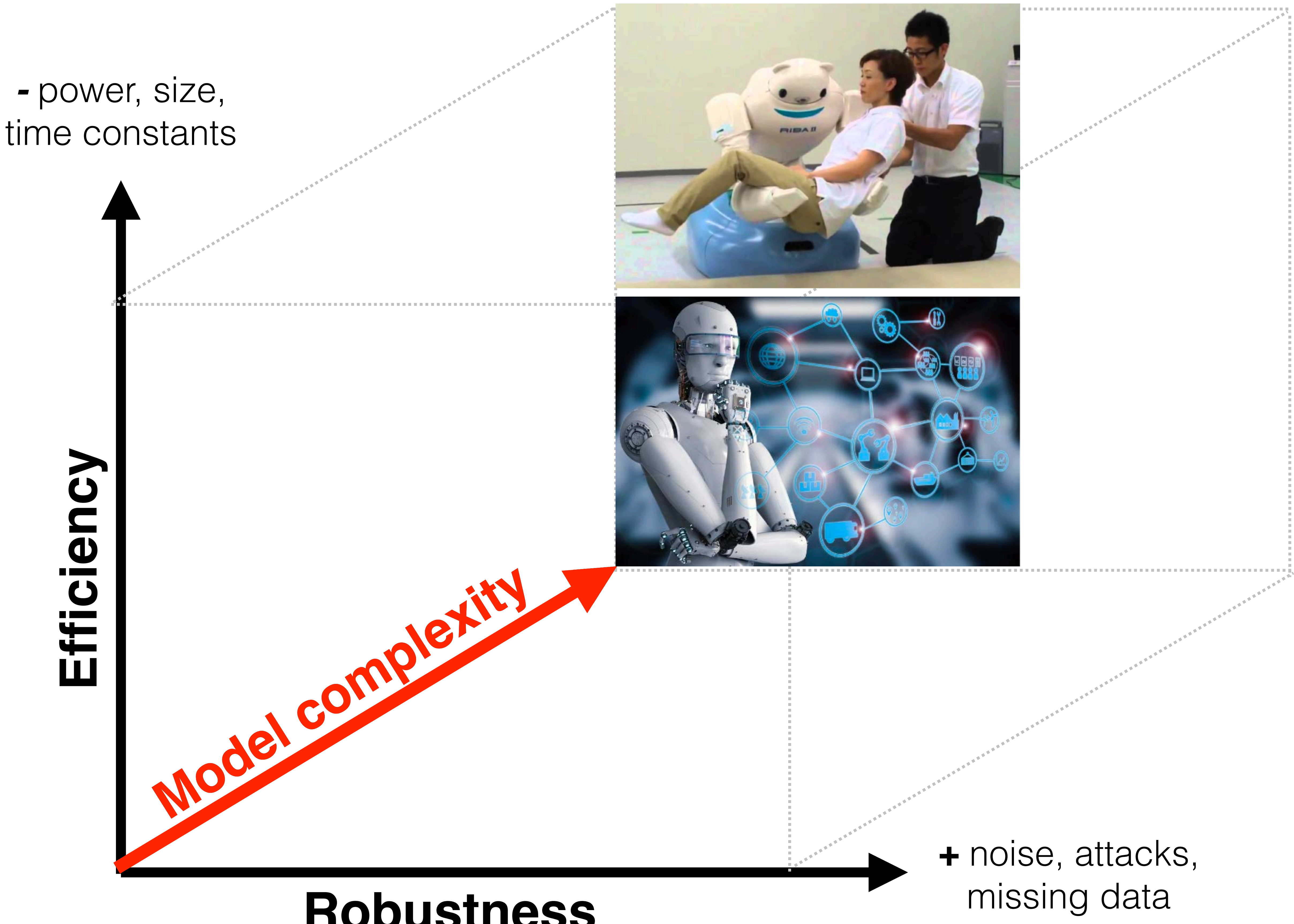
	Desktop CPU	Embedded CPU	Xilinx FPGA	Navion chip
latency image processing	50 ms	200ms	50 ms	**22ms**
latency MAP estimation	80 ms	400ms	200ms	**30ms**
power	26.1 W	2.33 W	1.46 W	**24mW**
accuracy	16cm	16cm	19cm	23cm

http://navion.mit.edu/

Efficiency and Miniaturization

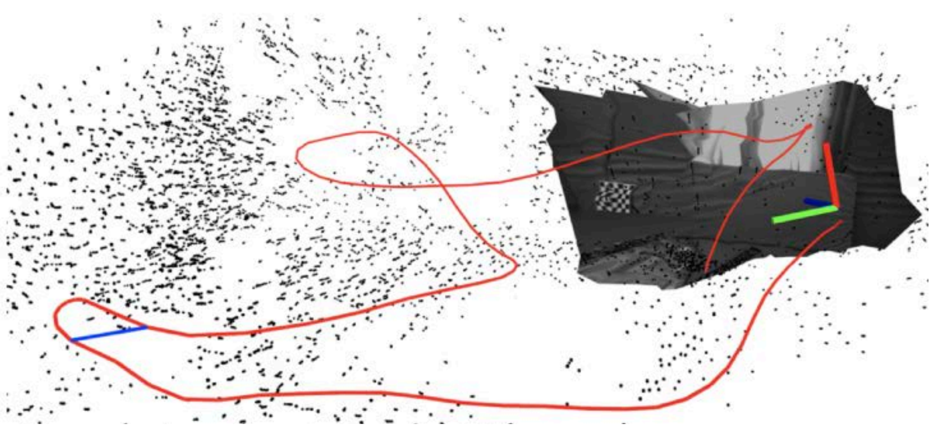
28

Active Research Directions

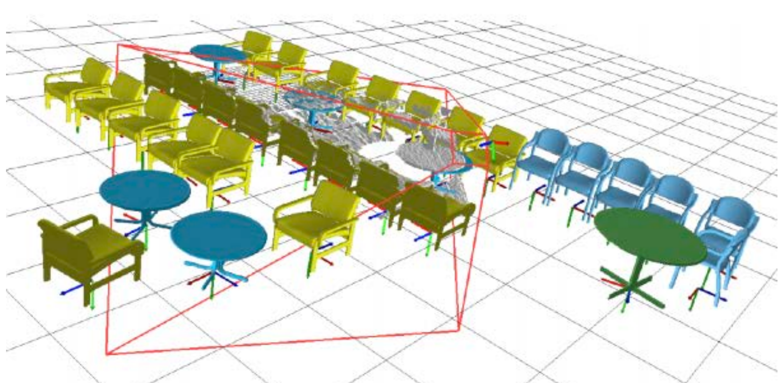


Mind the Gap with Human Perception

Sparse or Dense Point Clouds

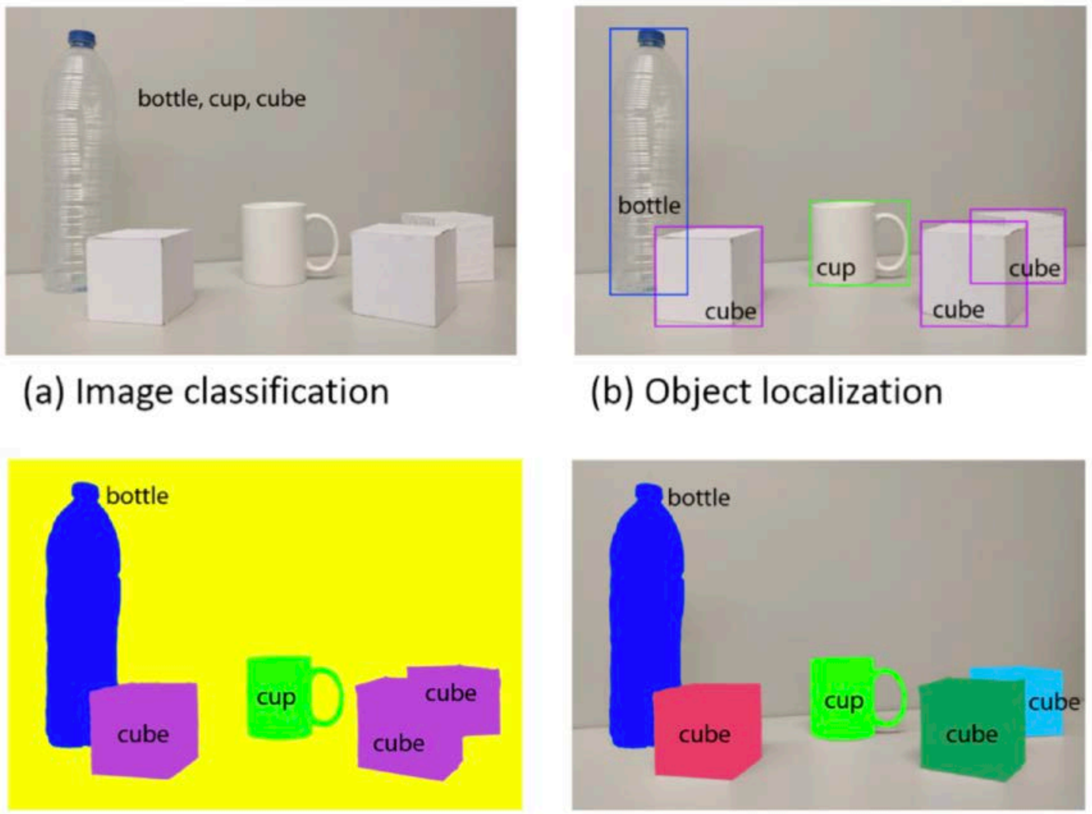


Object-based Maps [Salas-Moreno, CVPR'13]



.. lines, voxels, meshes

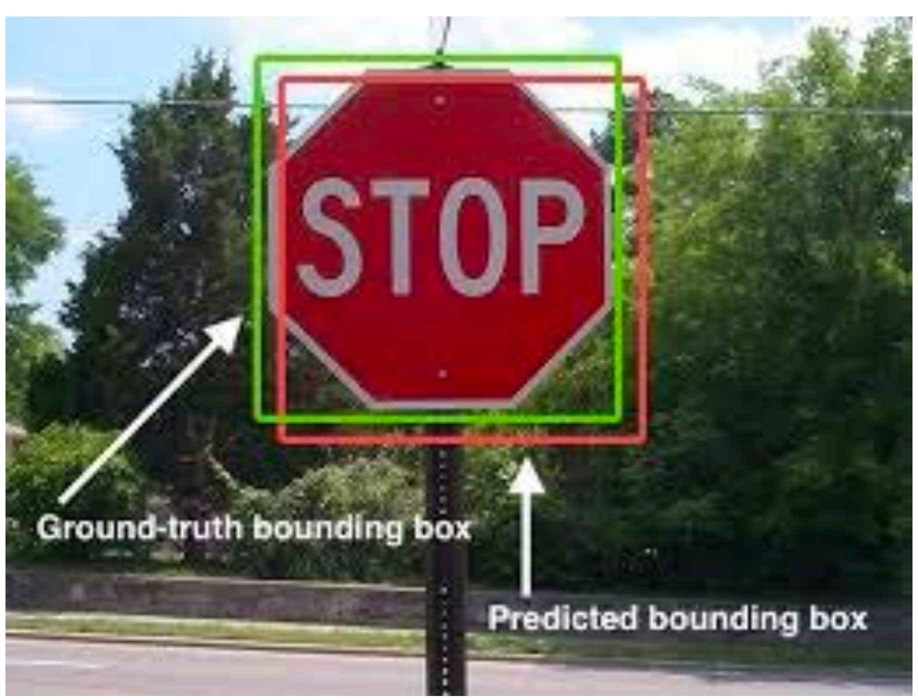
High-level Understanding: Opportunities

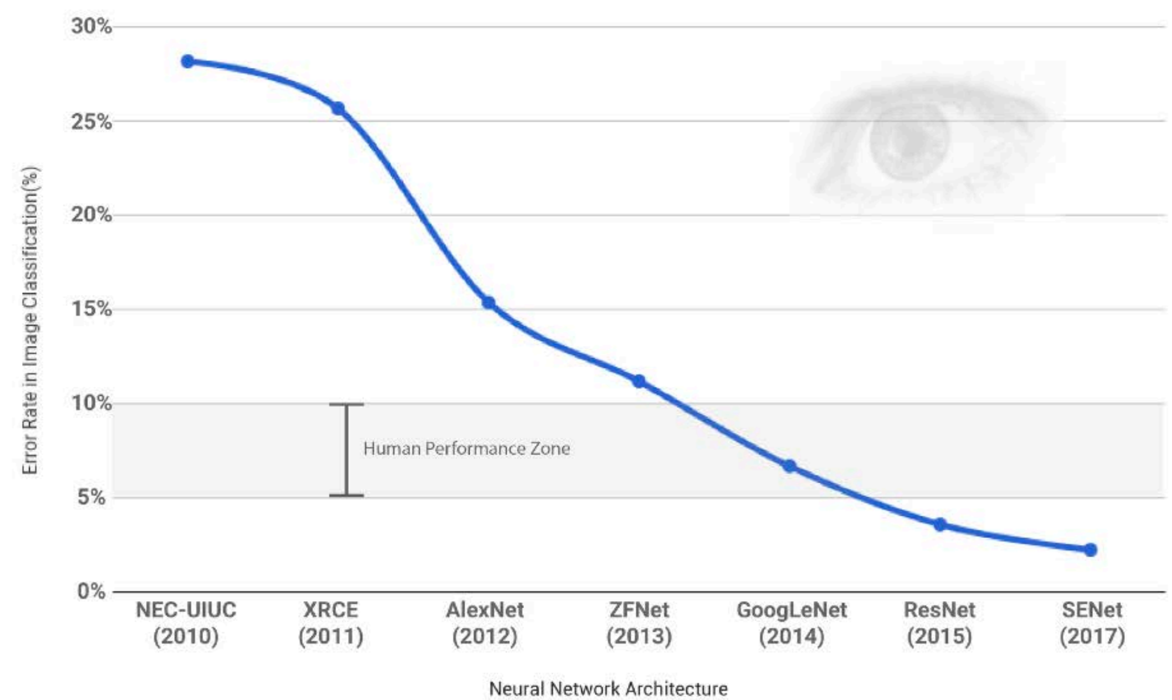


(a) Image classification (b) Object localization

(c) Semantic segmentation (d) Instance segmentation

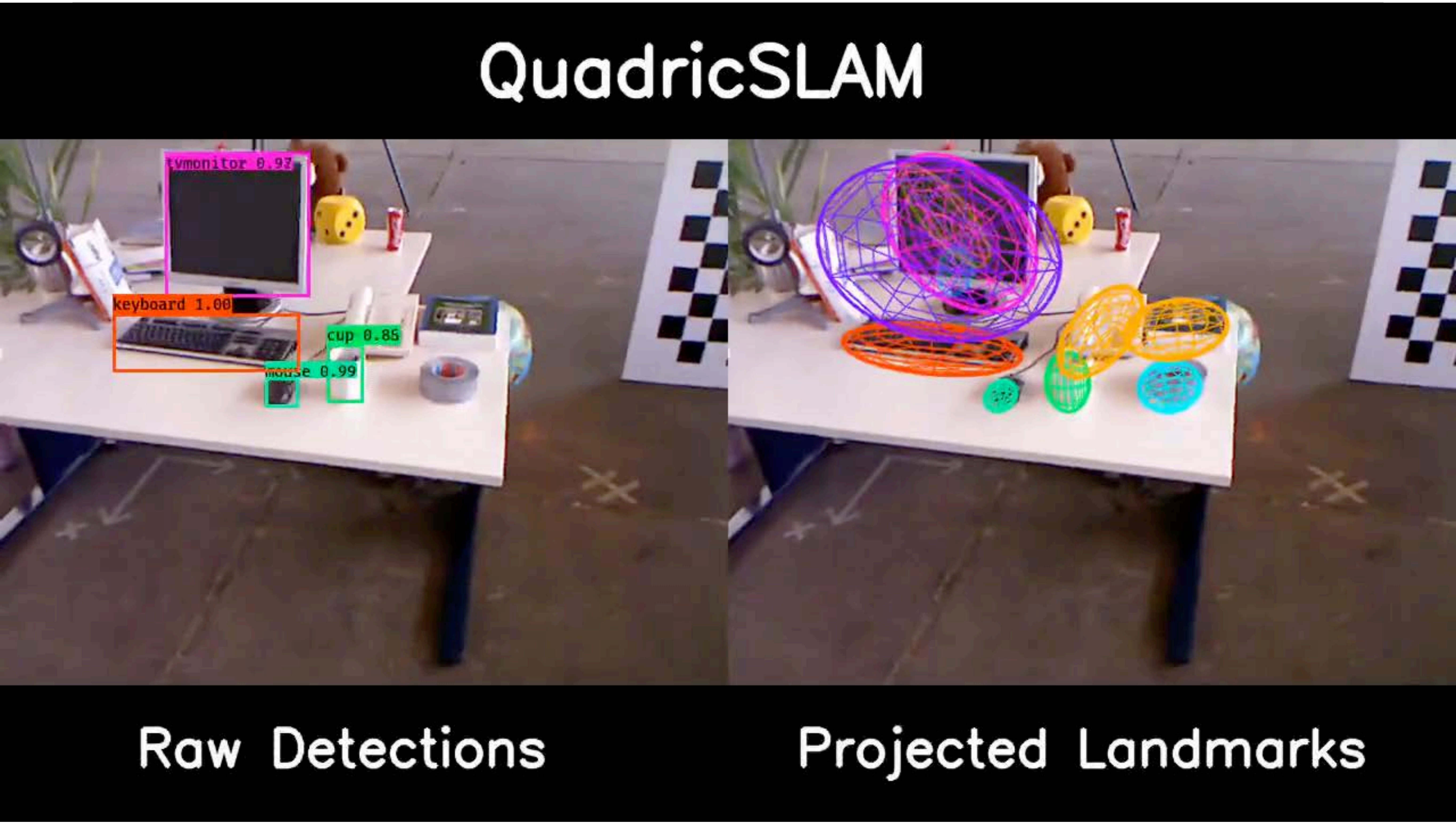
[Garcia-Garcia et al., 2017]





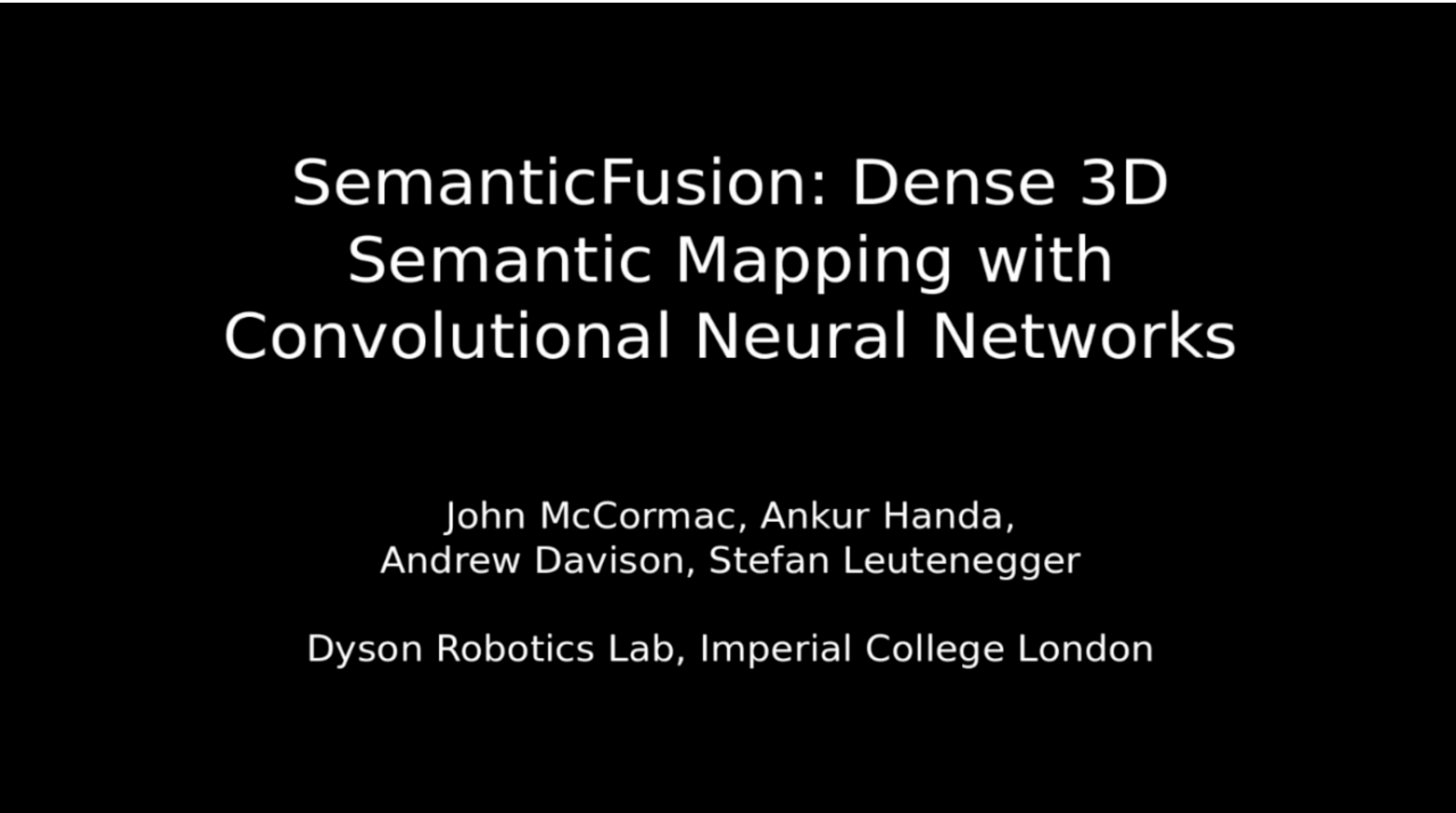
The deep learning revolution!

Sparse Object-level SLAM



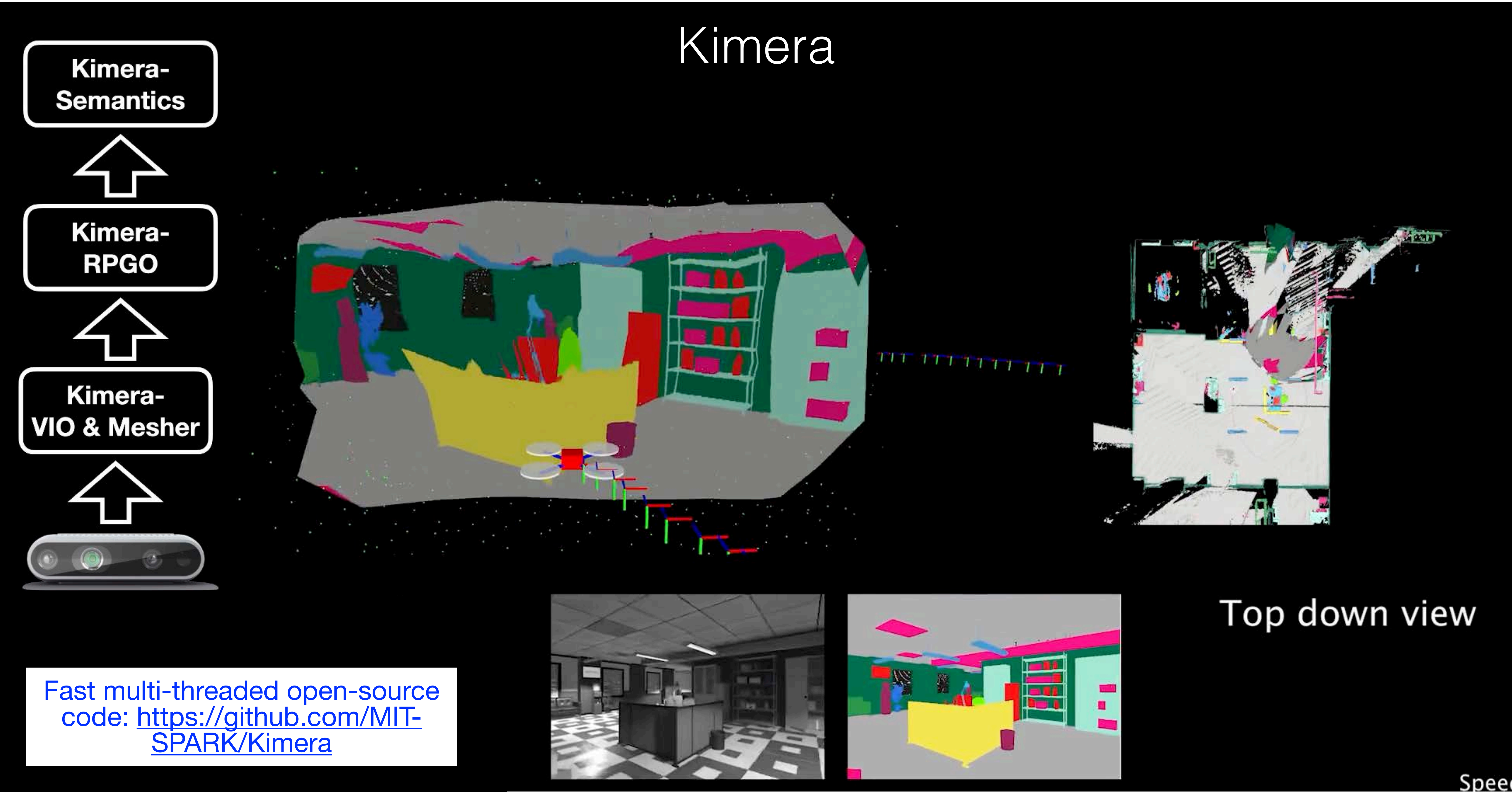
[Sunderhauf and Milford, 2017]

Dense Metric-Semantic SLAM on GPU



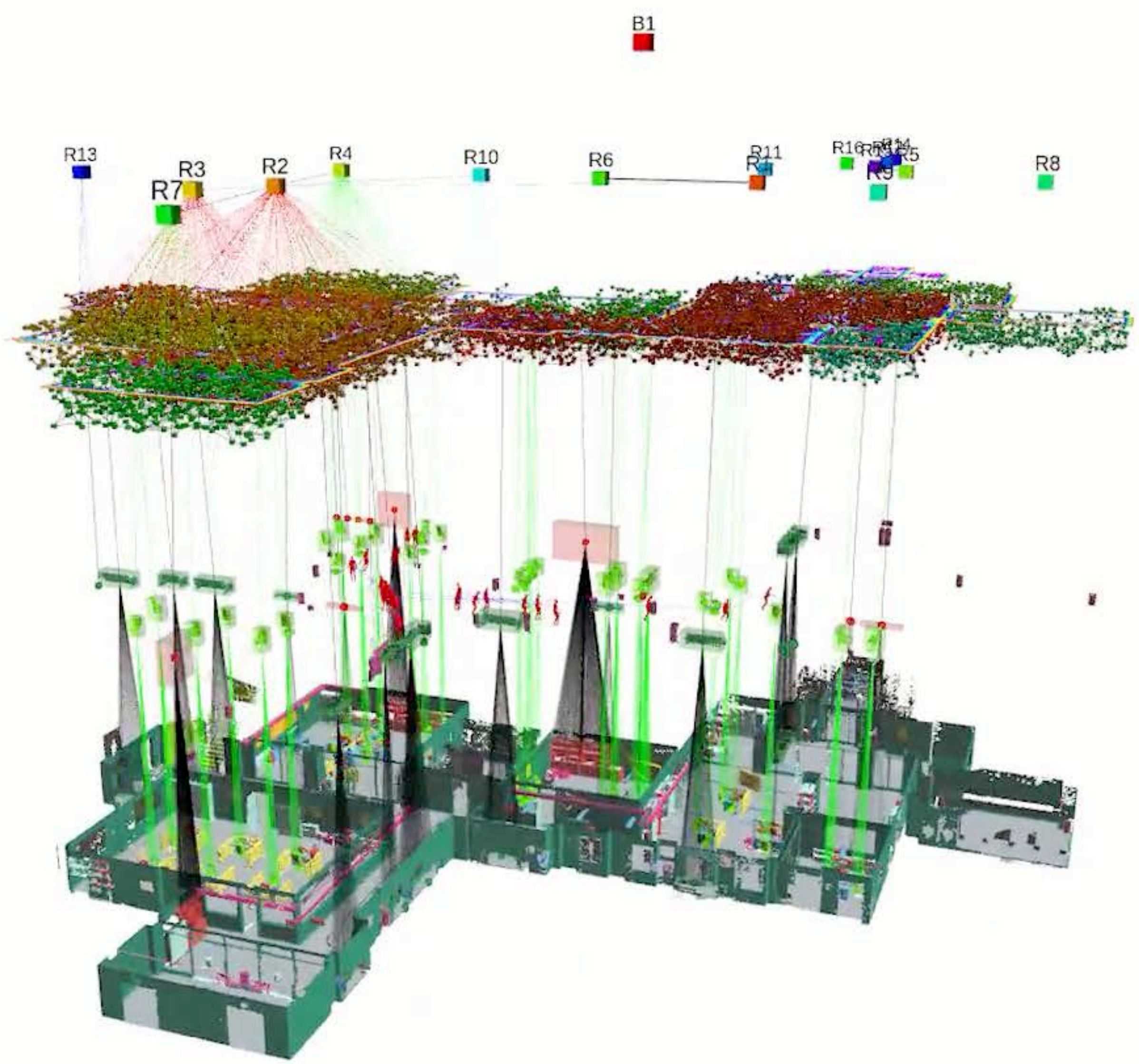
[McCormac et al., SemanticFusion]

Dense Metric-Semantic SLAM on **C**PU



Rosinol, Abate, Chang, Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. ICRA 2020.

High-level Understanding: 3D Scene Graphs



Directed graph, where:

- **nodes** are *spatial concepts* (i.e., concepts grounded in 3D)
- **edges** represent spatio-temporal relations between concepts (e.g., agent "i" in room "j" at time t)

We present a unified representation for actionable spatial perception: 3D Dynamic Scene Graphs (DSGs)

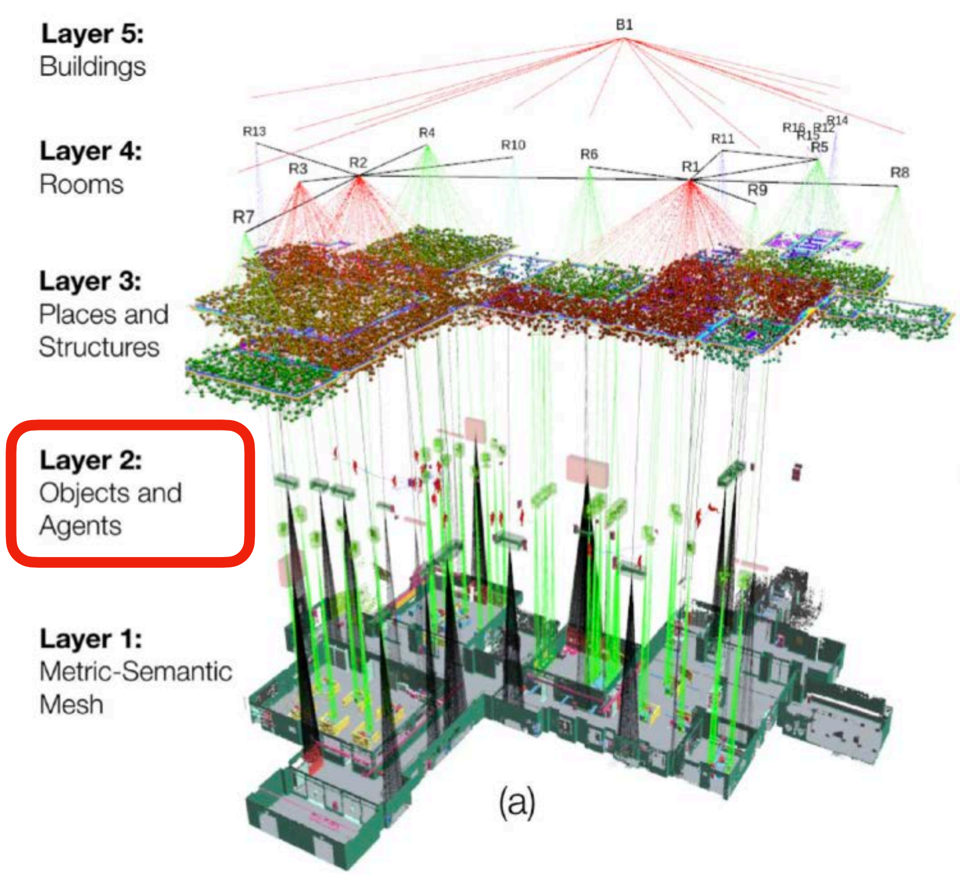
[Armeni et al., 3D scene graph: A structure for unified semantics, 3D space, and camera. ICCV'19]
[Rosinol et al., 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans, RSS'20]

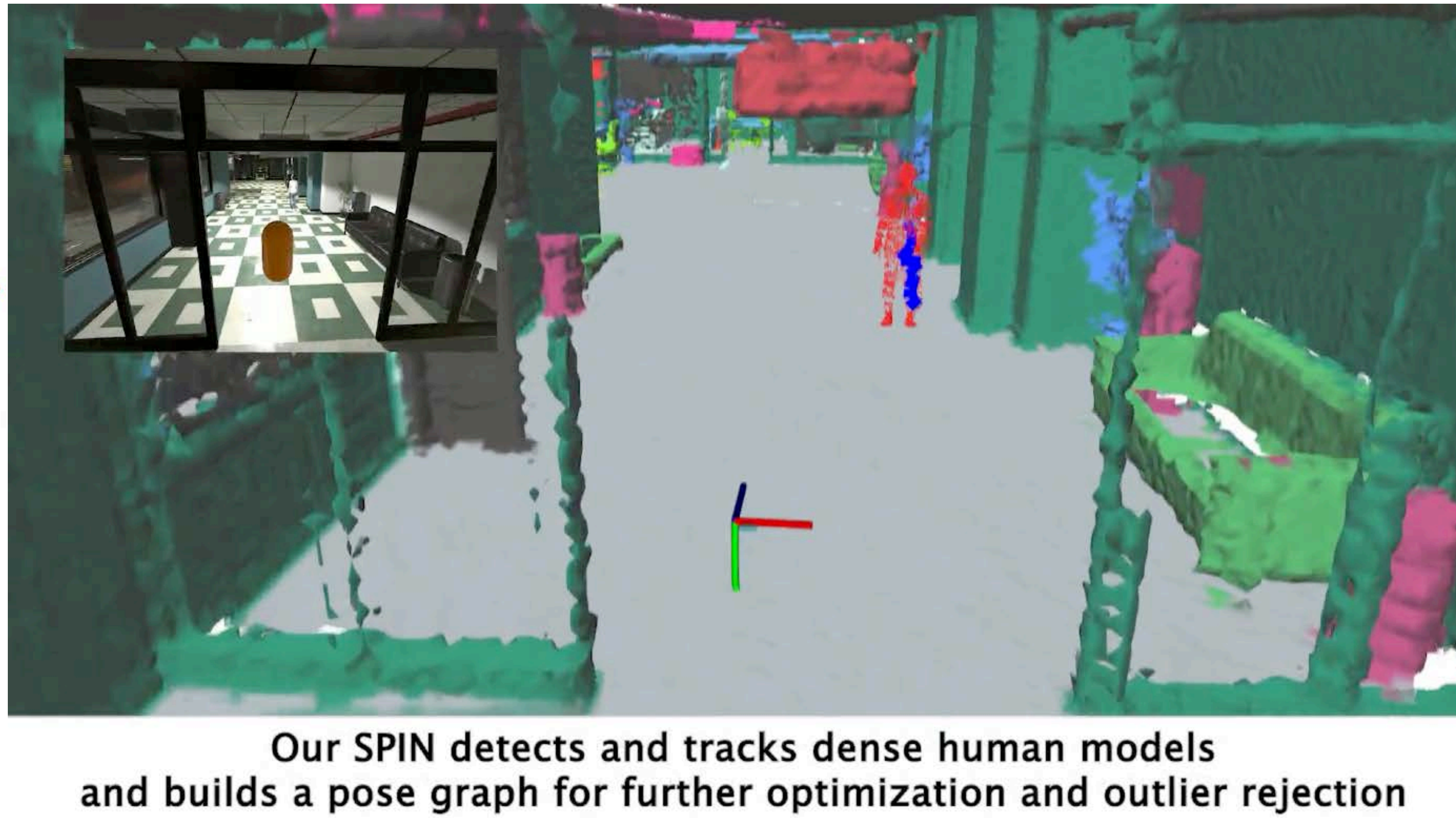
High-level Understanding: 3D Scene Graphs

- From SLAM algorithms to a **Spatial Perception engINe (SPIN)**, that infers geometry, semantics, a hierarchy of high-level spatial concepts and their relations

[Rosinol, Gupta, Abate, Shi, Carlone, 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans, RSS'20]

Layer 2: Objects and Agents

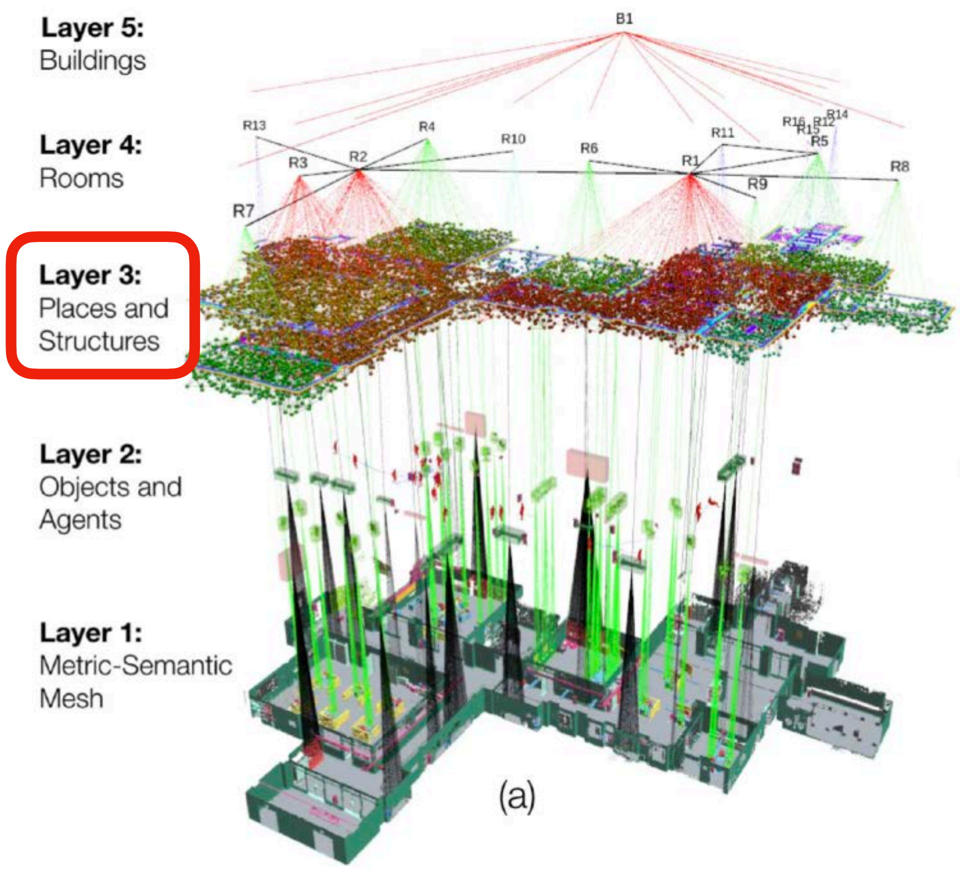




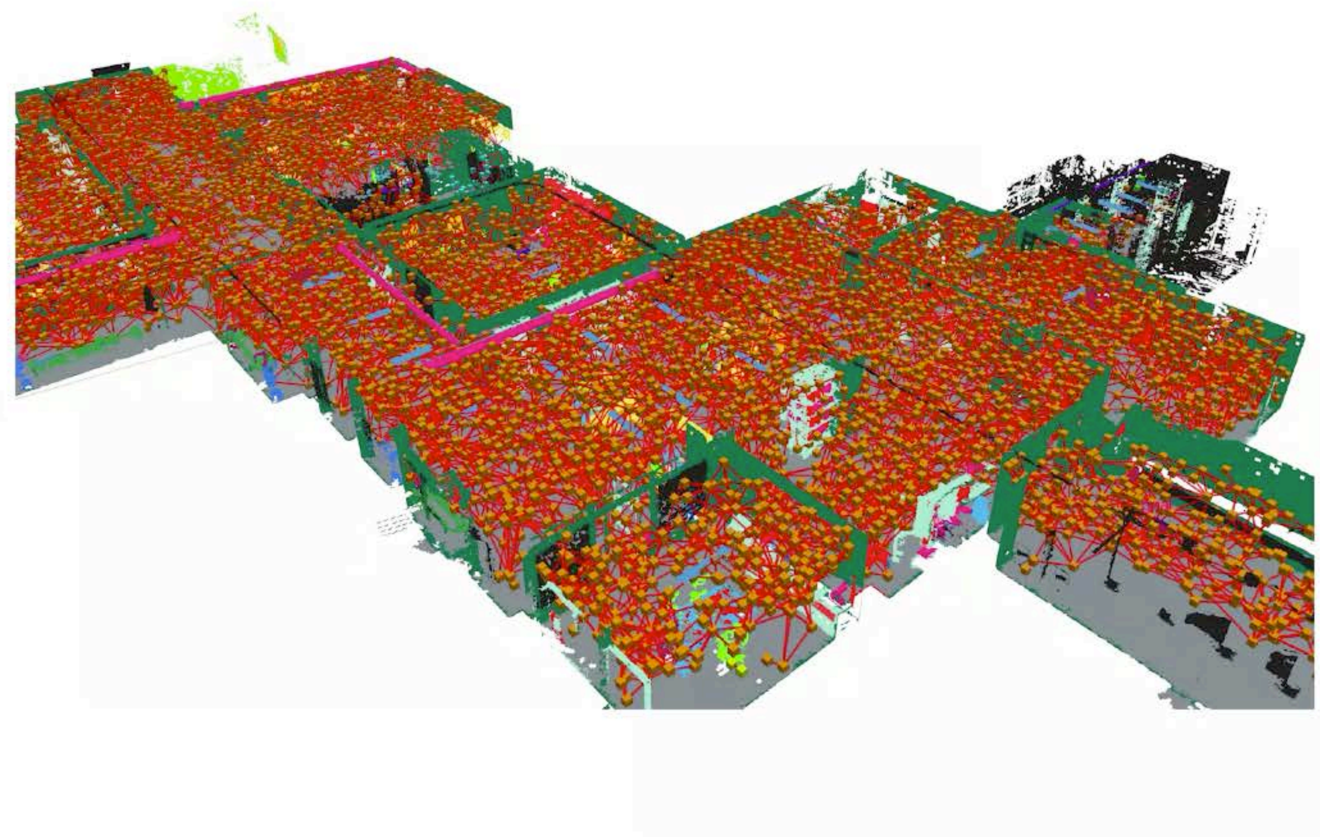
- **Humans**:
  - 3D dense shape reconstruction from monocular images [2]
  - Robust Pose Graph Optimization to track human poses over time
- **Objects**:
  - Euclidean clustering (when shape is unknown)
  - TEASER++ (when shape is known)

[1] Rosinol et al., 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans, RSS'20.
[2] Kolotouros, Pavlakos, Daniilidis, Convolutional mesh regression for single-image human shape reconstruction, CVPR'19.

Layer 3: Places and Structures



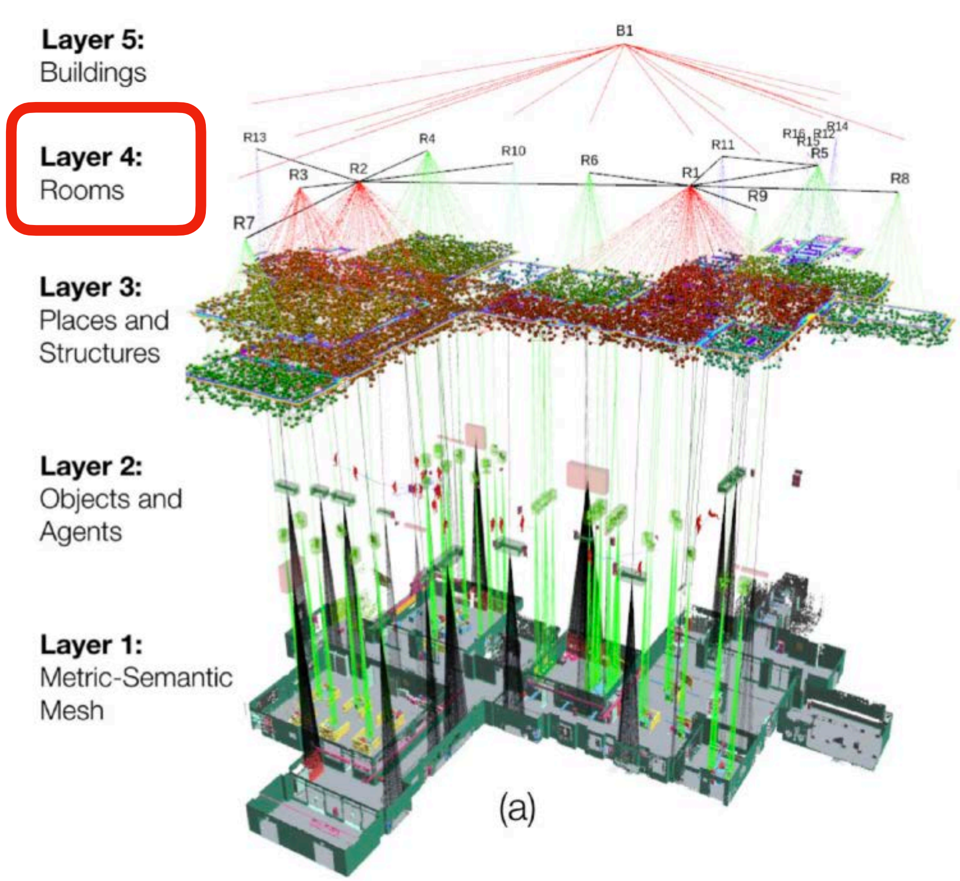
A global path planner (A*) uses Layer 3 (topological graph) for fast large-scale path planning, while the local path planner adjusts the trajectory to previously unseen obstacles and to the dynamics of the robot.



- **Places**: obstacle-free locations in the map, such that there is line-of-sign between pairs of nodes (suitable for fast path planning), using [2]
- **Structures**: separators between free space (walls, ground floor, ceiling)

[1] Rosinol et al., 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans, RSS'20.
[2] Oleynikova, Taylor, Siegwart, Nieto, Sparse 3D topological graphs for micro-aerial vehicle planning, IROS'18.

Layer 4: Rooms



We cluster the places in the environment into different rooms, obtaining an actionable representation for navigation and planning

- **Rooms**:
  - extracted from graph of places using graph clustering
- **Remark**: traversability described at the level of rooms, places, and in the mesh: this is a “feature”, rather than a “bug” (-> hierarchical planning)

[1] Rosinol et al., 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans, RSS’20.

Active Research Directions