

# The RAWSEEDS benchmarking toolkit

How we did it!

---

*Matteo Matteucci, [matteo.matteucci@polimi.it](mailto:matteo.matteucci@polimi.it)*

# About myself

---

- Associate professor at Politecnico di Milano
  - Robotics
  - Cognitive Robotics
  - Machine Learning
- Main research interests
  - Robot vision/perception
  - Machine learning
  - Benchmarking and performance evaluation





# Why Benchmarking?

- Robocup Lisbon 2004 (left), Bremen 2006 (right)



# We need a Benchmark ...

---

- “*Defining a standard benchmark for mobile service robots*” (The RoSta wiki – 2008)
  - Benchmark:
    - A standard by which something is evaluated or measured
    - A surveyor's mark made on some stationary object and shown on a map as a reference point
  - Benchmarking:
    - To measure the performance of an item relative to another similar item in an impartial scientific manner. (<http://en.wiktionary.org/wiki/benchmark>)



# Good Experimental Methodologies

---

- “*General Guidelines for Robotics Papers using Experiments*” (March 2008 DRAFT)
  - Is it an experimental paper?
  - Are the system assumptions/hypotheses clear?
  - Are the performance criteria spelled out explicitly?
  - What is being measured and how?
  - Do the methods and measurements match the criteria?
  - Is there enough information to reproduce the work?
  - Do the results obtained give a fair and realistic picture of the system being studied?
  - Are the drawn conclusions precise and valid



# Experiences to imitate

- Other fields in Computer Science had paved the way:

- Machine Learning @ UCI
- Stereo vision @ Middlebury
- Performance Evaluation of Tracking and Surveillance
- PASCAL (object recognition database)
- ...



The screenshot shows the UCI Machine Learning Repository website. The header features the UCI logo and the text "Machine Learning Repository" and "Center for Machine Learning and Intelligent Systems". Navigation links include "About", "Citation Policy", "Donate a Data Set", and "Contact". A search bar is present with a "Search" button and a "View ALL Data Sets" link. The main content area welcomes visitors and provides information about the repository's services, including a searchable interface and links to the "About page", "Citation Policy", and "donation policy". It also mentions support from various organizations and collaborations. The footer is divided into three columns: "Latest News" with a list of recent updates, "Newest Data Sets" with a list of newly added datasets, and "Most Popular Data Sets (hits since 2007)" with a list of popular datasets and their hit counts.

Latest News:	Newest Data Sets:	Most Popular Data Sets (hits since 2007):
<b>06-25-2007:</b> Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope	<b>03-04-2008:</b>  <a href="#">Mammographic Mass</a>	<b>12224:</b>  <a href="#">Iris</a>
<b>04-13-2007:</b> Research papers that cite the repository have been associated to specific data sets.	<b>02-29-2008:</b>  <a href="#">Forest Fires</a>	<b>9853:</b>  <a href="#">Adult</a>
<b>04-09-2007:</b> Three new data sets have been added: Poker Hand, Call2 Building People Counts, Dodgers Loop Sensor.	<b>06-01-2007:</b>  <a href="#">UJI Pen Characters</a>	<b>7659:</b>  <a href="#">Breast Cancer Wisconsin (Diagnostic)</a>
<b>09-08-2006:</b> The Beta site has been launched.	<b>05-01-2007:</b>  <a href="#">MAGIC Gamma Telescope</a>	<b>7172:</b>  <a href="#">Wine</a>
<b>09-01-2006:</b> SPECTF.test has been modified by the donor.	<b>01-01-2007:</b>  <a href="#">Poker Hand</a>	<b>6766:</b>  <a href="#">Poker Hand</a>
<b>08-28-2006:</b> PHP faceted browse has been implemented.		
<b>08-23-2006:</b> The metadata fields for each data set in the Repository have been filled out.		

# Experiences to imitate


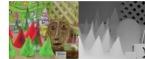


- Other fields in Computer Science had paved the way:

- Machine Learning @ UCI
- Stereo vision @ Middlebury
- Performance Evaluation of Tracking and Surveillance
- PASCAL (object recognition database)
- ...

[vision.middlebury.edu](#)  
[stereo](#) • [mview](#) • [MRF](#) • [flow](#)

[Stereo](#) • [Evaluation](#) • [Datasets](#) • [Code](#) • [Submit](#)

### Middlebury Stereo Datasets

	<a href="#">2001 datasets</a> - 6 datasets of piecewise planar scenes [1] (Sawtooth, Venus, Bull, Poster, Barn1, Barn2)
	<a href="#">2003 datasets</a> - 2 datasets with ground truth obtained using structured light [2] (Cones, Teddy)
	<a href="#">2005 datasets</a> - 9 datasets obtained using the technique of [2], published in [3, 4] (Art, Books, Dolls, Laundry, Moebius, Reindeer, Computer, Drumsticks, Dwarves)
	<a href="#">2006 datasets</a> - 21 datasets obtained using the technique of [2], published in [3, 4] (Aloe, Baby1-3, Bowling1-2, Cloth1-4, Flowerpots, Lampshade1-2, Midd1-2, Monopoly, Plastic, Rocks1-2, Wood1-2)

**How to cite our datasets:**  
We grant permission to use and publish all images and disparity maps on this website. However, if you use our datasets, we request that you cite the appropriate paper(s): [1] for the 2001 datasets, [2] for the 2003 datasets, and [3] or [4] for the 2005 and 2006 datasets.

**References:**  
[1] D. Scharstein and R. Szeliski. [A taxonomy and evaluation of dense two-frame stereo correspondence algorithms](#). *International Journal of Computer Vision*, 47(1/2/3):7-42, April-June 2002.  
[2] D. Scharstein and R. Szeliski. [High-accuracy stereo depth maps using structured light](#). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, volume 1, pages 195-202, Madison, WI, June 2003.  
[3] D. Scharstein and C. Pal. [Learning conditional random fields for stereo](#). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, MN, June 2007.  
[4] H. Hirschmüller and D. Scharstein. [Evaluation of cost functions for stereo matching](#). In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, MN, June 2007.



# Experiences to imitate

- Other fields in Computer Science haD paved the way:

- Machine Learning @ UCI
- Stereo vision @ Middlebury
- Performance Evaluation of Tracking and Surveillance
- PASCAL (object recognition database)
- ...

vision.middlebury.edu  
stereo • mview • MRF • flow

Stereo Evaluation • Datasets • Code • Submit

Middlebury Stereo Evaluation - Version 2  
New features and main differences to version 1  
Submit and evaluate your own results

Open a new window for each link

Algorithm	Avg. Rank	Error Threshold = 1				Sort by nonocc				Sort by all				Sort by disc			
		Triakuba dataset				Venus dataset				Teddy dataset				Cones dataset			
		nonocc	all	disc	all	nonocc	all	disc	all	nonocc	all	disc	all	nonocc	all	disc	all
CostSfM [111]	26.9	4.72	6.08	20.3	1.41	2.48	18.5	8.18	15.9	23.8	1.91	10.2	11.8				
ReliabilityDP [131]	27.9	1.35	3.39	7.25	2.35	3.48	12.2	2.82	16.9	19.5	12.2	15.9	19.7				
TreeDP [8]	28.6	1.92	2.84	9.96	1.41	2.10	7.74	15.9	23.9	27.1	10.0	18.3	18.9				
GC [14]	29.3	1.94	4.12	9.39	1.79	3.44	8.75	16.5	25.0	24.9	7.70	18.2	15.3				
DP [18]	32.9	4.12	5.04	12.0	10.1	11.0	21.0	14.0	21.6	20.6	10.5	19.1	21.1				
PhaseBased [31]	34.2	4.26	6.53	15.4	6.71	8.16	26.4	14.5	23.1	25.5	10.8	20.5	21.2				
SSD+MF [14]	34.6	5.23	7.07	24.1	1.74	5.16	11.9	16.5	24.8	32.9	10.6	19.8	26.3				
STCA [14]	35.8	7.10	9.63	27.8	8.19	9.58	40.3	15.8	23.2	37.7	9.80	17.8	28.7				
SG [14]	36.3	5.02	7.22	12.2	9.44	10.9	21.9	19.9	28.2	26.3	11.0	22.8	22.3				
PhaseDiff [23]	37.0	4.82	7.11	16.3	8.34	9.76	26.0	20.0	28.0	29.0	12.8	28.5	27.5				
Infection [10]	37.4	7.55	9.54	28.9	4.41	5.53	31.7	17.7	25.1	44.4	14.3	21.3	38.0				

References

[1] D. Scharstein and R. Szeliski. *A taxonomy and evaluation of dense two-frame stereo correspondence algorithms*. IJCV 2002.  
a. SSD + min filter (i.e., shiftable windows), window size = 21.  
b. Dynamic programming, similar to Bobick and Ittlice (ICV 1999).  
c. Scanline optimization (D optimization using horizontal smoothness terms).  
d. Graph cuts using alpha-beta sweeps (Boykov, Veksler, and Zabih, PAMI 2001).  
[2] V. Kolmogorov and R. Zabih. *Computing visual correspondence with occlusions using graph cuts*. ICCV 2001.  
[3] V. Kolmogorov and R. Zabih. *Multi-camera scene reconstruction via graph cuts*. ECCV 2002.  
[4] M. Beyer and M. Gehrig. *Accelerated stereo algorithm using image segmentation and global visibility constraints*. ICIP 2004.  
[5] L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. *High-quality video view interpolation using a layered representation*. SIGGRAPH 2004.  
[6] H. Hirschmüller. *Accurate and efficient stereo processing by semi-global matching and mutual information*. CVPR 2005, PAMI 30(2): 328-341, 2008.  
[7] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum. *Symmetric stereo matching for occlusion handling*. CVPR 2005.  
[8] O. Veksler. *Stereo correspondence by dynamic programming on a tree*. CVPR 2005.  
[9] P. Mordohai and G. Medioni. *Stereo vision using tensor voting within the tensor voting framework*. PAMI 28(6): 968-982, 2006.  
[10] G. Olague, F. Fernández, C. Pérez, and L. Lutton. *The infection algorithm: an artificial epidemic approach for dense stereo correspondence*. Artificial Life, 2005.  
[11] R. Brackley, M. Hund, and B. Mertschinger. *Stereo vision using cost relaxation with 3D support regions*. Image and Vision Computing New Zealand (IVCNZ), 2005.



# Here it comes RAWSEEDS

---

- EU Funded Project in the VI Frame Program (1<sup>st</sup> November 2006 to July 2009)
- A ***Specific Support Action*** to collect and publish a (S)LAM benchmarking toolkit
- Involved Institutions:
  - Politecnico di Milano (Italy – Coordinator)
  - Università di Milano-Bicocca (Italy – Partner)
  - University of Freiburg (Germany – Partner)
  - Universidad de Zaragoza (Spain – Partner)



# Why Benchmarking SLAM?

---

- Benchmarking of a robotic application might be complex and hard to tackle as a whole
- The SLAM community was already establishing a “dataset” culture for algorithms evaluation
- Simultaneous Localization And Mapping could have been one of the easiest activity to benchmark in robotics ...
  - We can establish proper metrics for SLAM
  - The community agrees on the use of such metrics
  - The community appreciate the effort for using it
  - ...





# What about simulation?

---

- “*Towards Quantitative Comparisons of Robot Algorithms: Experiences with SLAM in Simulation and Real World Systems*” (Balaguer et al. - Benchmarking @ IROS 2007)
  - Simulators can be available for free (almost)
  - Ground Truth is perfect and easy to collect ;-)
  - Experiments are "easy" to replicate
- Seems the solution for benchmarking problems, “however real life differs from simulation”
- Useful in the lifecycle of a scientific idea, but robots eventually get real ...

# Benchmarking Beyond Radish

---

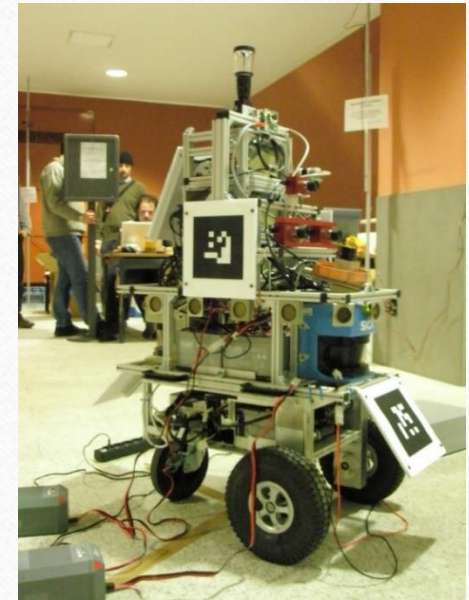
- RAWSEEDS toolkit fosters publishing of:
  - Extended multi-sensor data sets for the testing of systems on real-world scenarios from different sensor perspectives
  - Benchmarks and methodologies for quantitative evaluation and comparison of algorithms (and eventually sensors)
  - Off-the-shelf algorithms, with demonstrated performances, to be used for bootstrapping and comparison.

[www.rawseeds.org](http://www.rawseeds.org)



# RAWSEEDS Sensor Suite

- Use of an extensive sensing suite
  - B/W + Color cameras (monocular)
  - Stereo cameras (SVS by Videre)
  - LRFs (SICK 2D & Hokuyo)
  - Omnidirectional camera (V-Stone)
  - GPS and RTK-GPS (Outdoor GT)
  - Other proprioceptives (e.g., odometry, Inertial Measurement Unit)
- Sensors synchronized and acquired at maximum frequency allowed by onboard PCs



# Issue #1: Design of the Datasets

---

- Defined relevant scenarios beforehand
  - Indoor scenarios: offices, halls, corridors, flat and non-flat walls, doors & passages, windows, horizontal floors, ramps, stairs, elevators, and several pieces of furniture.
  - Outdoor scenarios where the robot moves in the open between buildings and the obstacles are comparable with those found along urban roads.
  - Mixed scenarios with parts surrounded by walls and parts located in the open.
- Different acquisition setups
  - Static and Dynamic environments (i.e., people walking around)
  - Different lighting conditions (i.e., natural daylight & artificial light)



# Indoor Locations in Bicocca



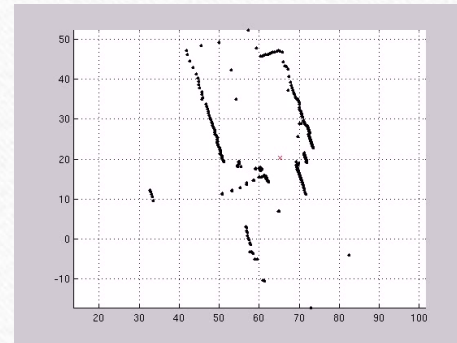
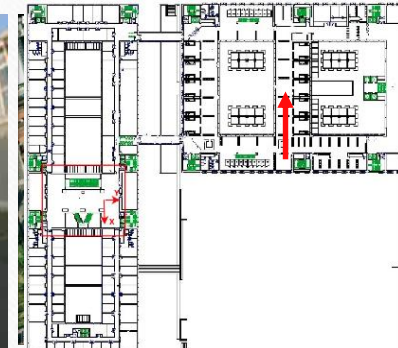
# Outdoor and Mixed Locations in Bovisa





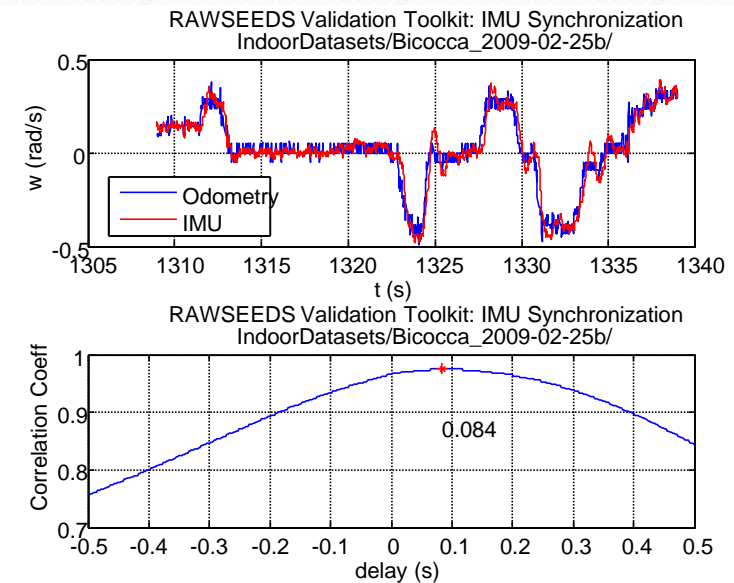
# 11 Datasets Collected

- Indoor
  - 1 static lamps + 1 static daylight
  - 1 dynamic lamps + 2 dynamic daylight
- Outdoor:
  - 2 static + 1 dynamic
- Mixed:
  - 2 static + 1 dynamic



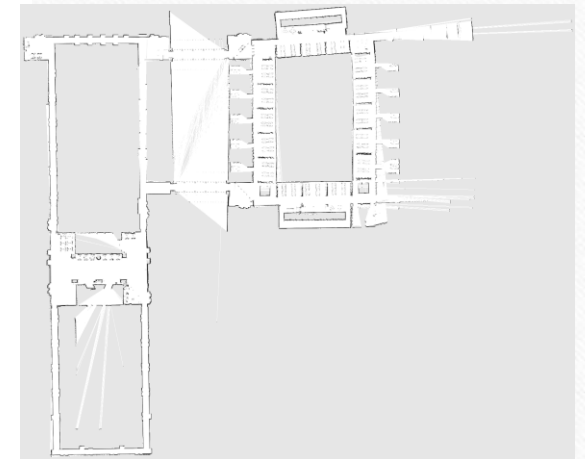
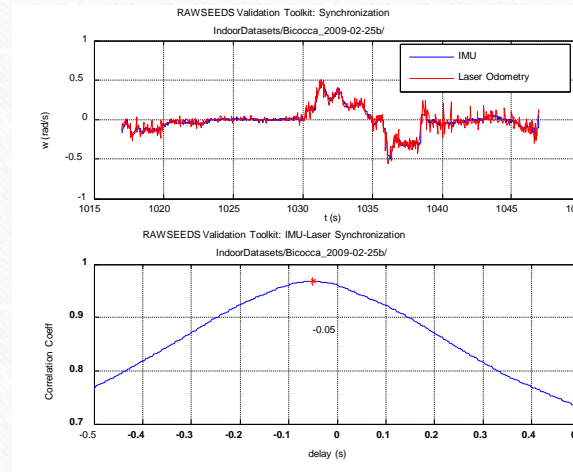
# Issue #2: Are the data any good?

- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays



# Issue #2: Are the data any good?

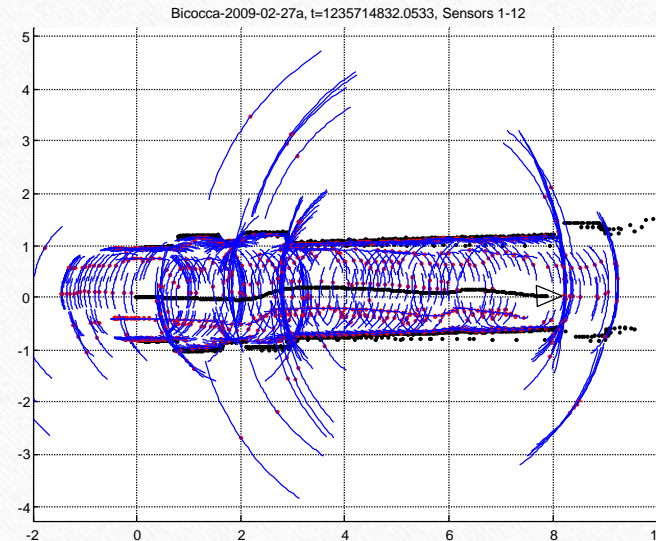
- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - LASERS checked for overlap





# Issue #2: Are the data any good?

- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - SONAR checked by visual inspection



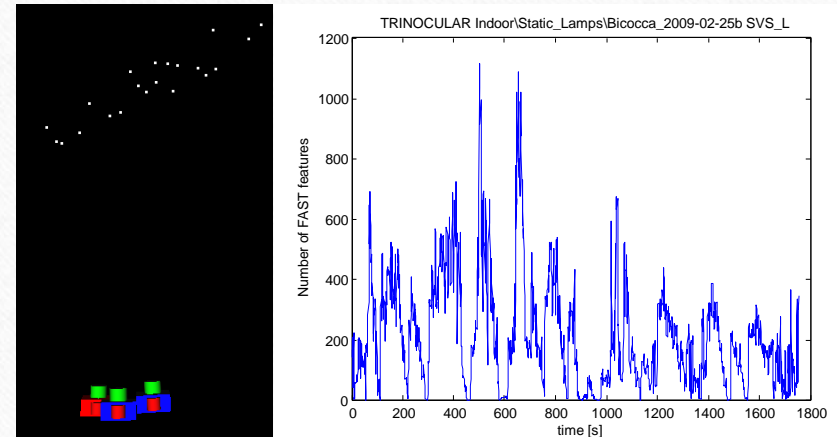
# Issue #2: Are the data any good?

- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - SONAR checked by visual inspection
  - MONOCULAR checked for features



# Issue #2: Are the data any good?

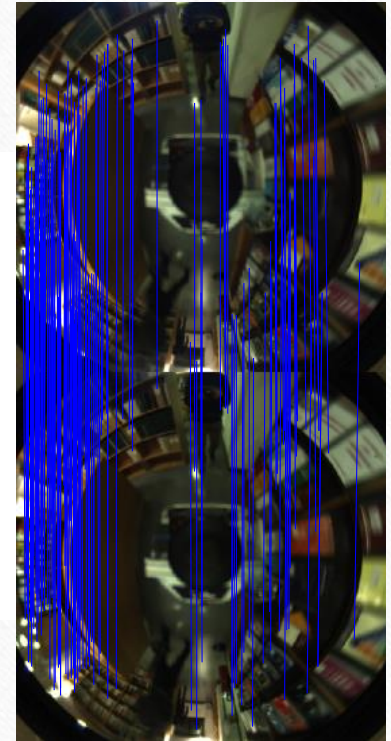
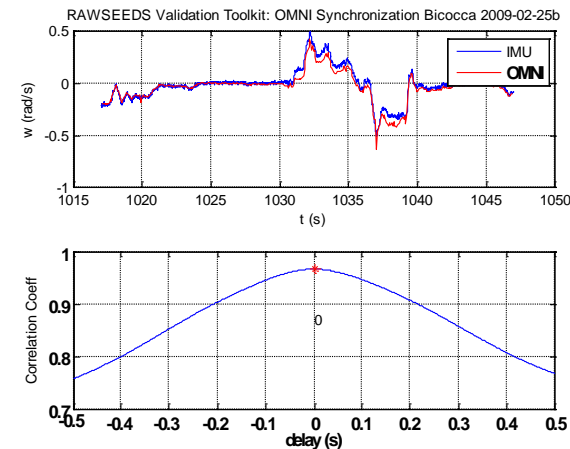
- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - SONAR checked by visual inspection
  - MONOCULAR checked for features
  - TRINOCULAR checked also for calibration





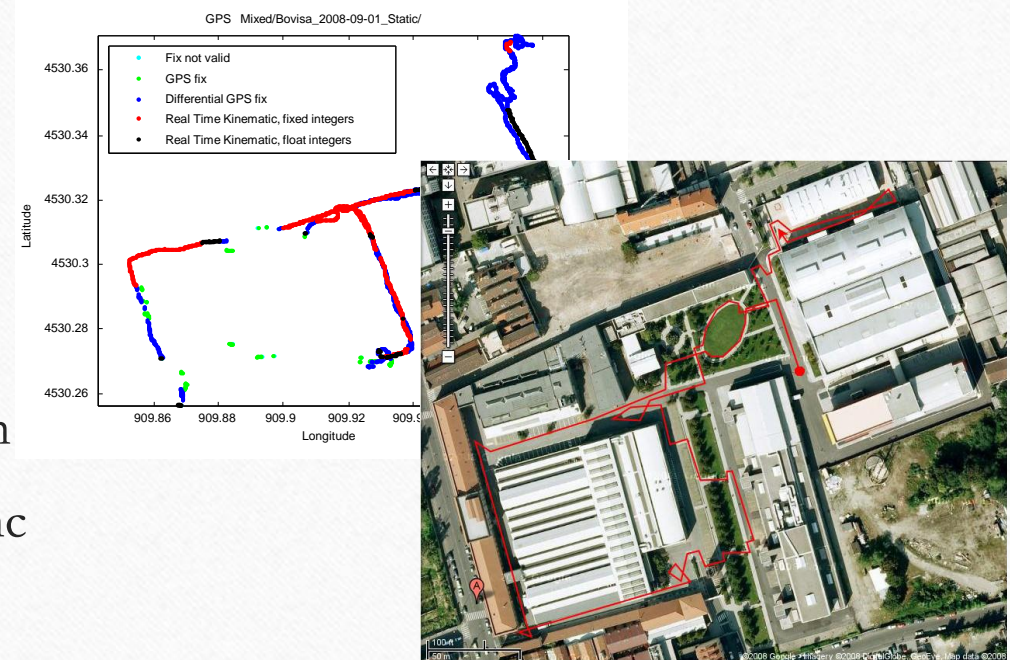
# Issue #2: Are the data any good?

- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - SONAR checked by visual inspection
  - MONOCULAR checked for features
  - TRINOCULAR checked also for calibration
  - PANORAMIC checked for features and sync



# Issue #2: Are the data any good?

- Independent evaluation of the data quality by Zaragoza partner
  - IMU used as time reference
  - ODOMETRY checked for delays
  - SONAR checked by visual inspection
  - MONOCULAR checked for features
  - TRINOCULAR checked also for calibration
  - PANORAMIC checked for features and sync
  - GPS checked for quality and coverage





# Issue #3: How do we evaluate SLAM?

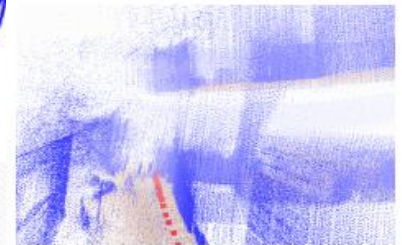
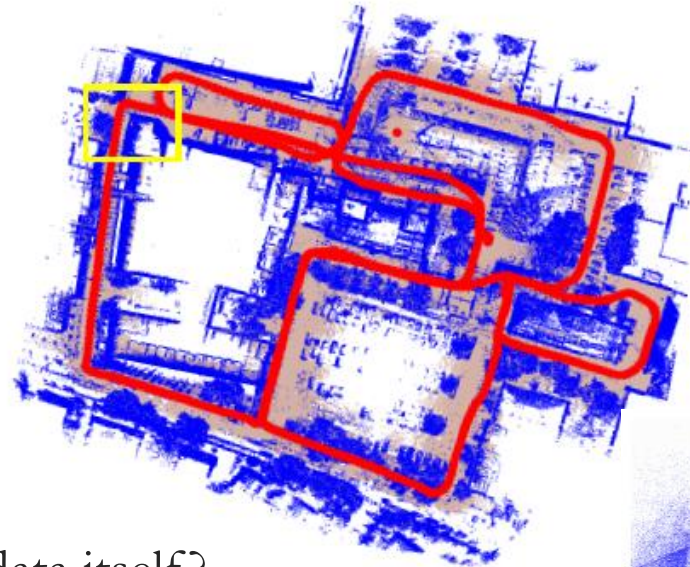
---

- A SLAM benchmark needs to assess the performance of a SLAM algorithm
  - Quantitative measures of map/path quality, w.r.t. ground truth
  - Performance variation as map size grows
  - How realistic/pessimistic/optimistic is the estimation error
  - Large loop recognition and closure
  - ...
- Clearly no single measure, we need a set of measures + *ground truth!*



# A Trick for Generating Ground Truth

- “*Benchmarking Urban 6D SLAM*” (Wulf et al. – Benchmarking Workshop @ IROS 2007)
  - Highly accurate RTK-GPS receivers can not be used in outdoor urban areas
  - Surveyed maps can be obtained from the national land registry offices
  - Monte Carlo Localization can be used with such accurate maps to estimate ground truth positioning from the data and a manual supervision step to validate the MCL results.
- Isn't there a solution which does not use the data itself?



# RAWSEEDS Ground Truth Setup

- Two GT Collection Systems
  - Outdoor: RTK (Real Time Kinematic) GPS
  - Indoor: vision-based (*GT-vision*) and LRF-based (*GT-laser*)





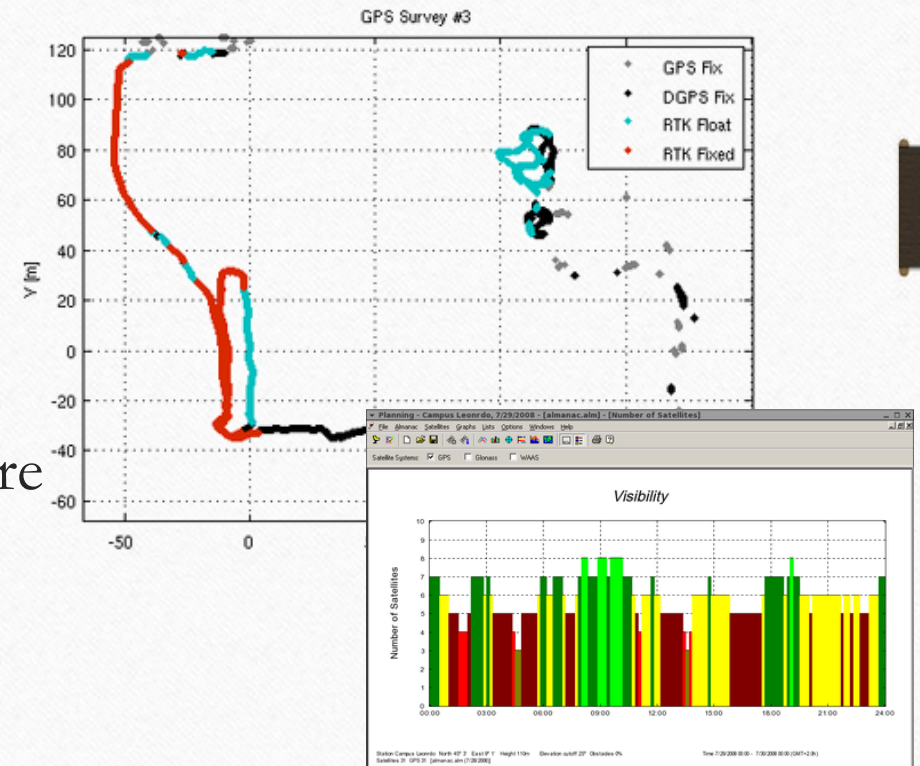
# Outdoor GT: RTK GPS

- Two GPS receivers (fixed + mobile)
- Radio link between the receivers
- Pros: no drift, (somehow) easy setup, high positioning precision
- Cons: does not operate indoors, costly hardware extremely sensible to obstacles, performance varies widely over time and space



# Outdoor GT: RTK GPS

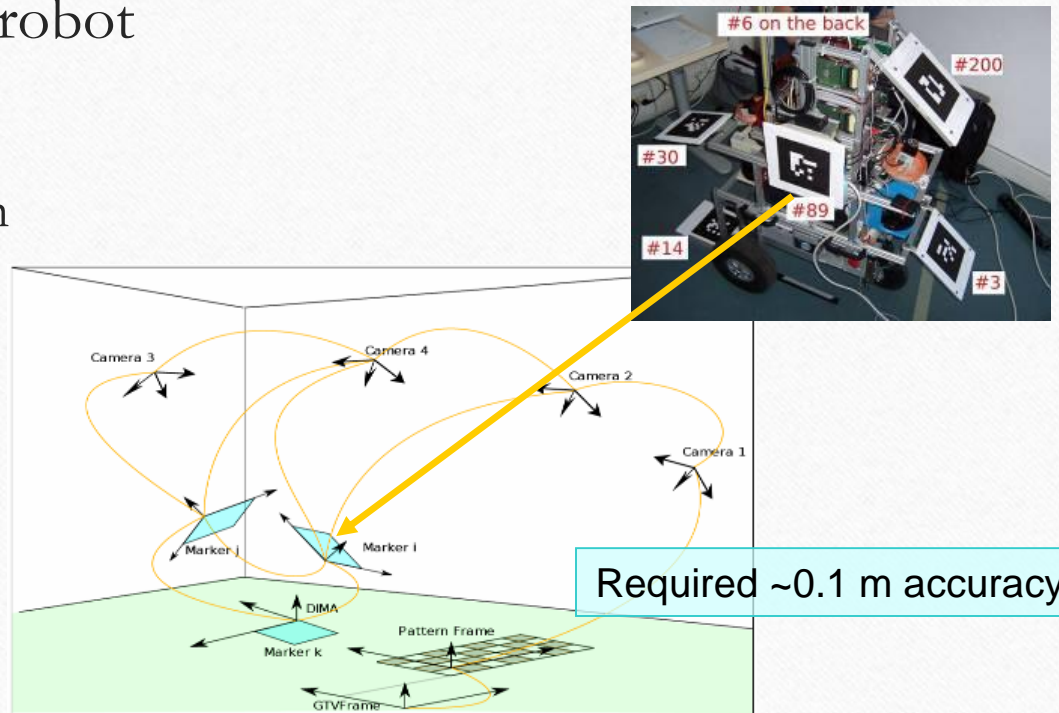
- Two GPS receivers (fixed + mobile)
- Radio link between the receivers
- Pros: no drift, (somehow) easy setup, high positioning precision
- Cons: does not operate indoors, costly hardware extremely sensible to obstacles, performance varies widely over time and space





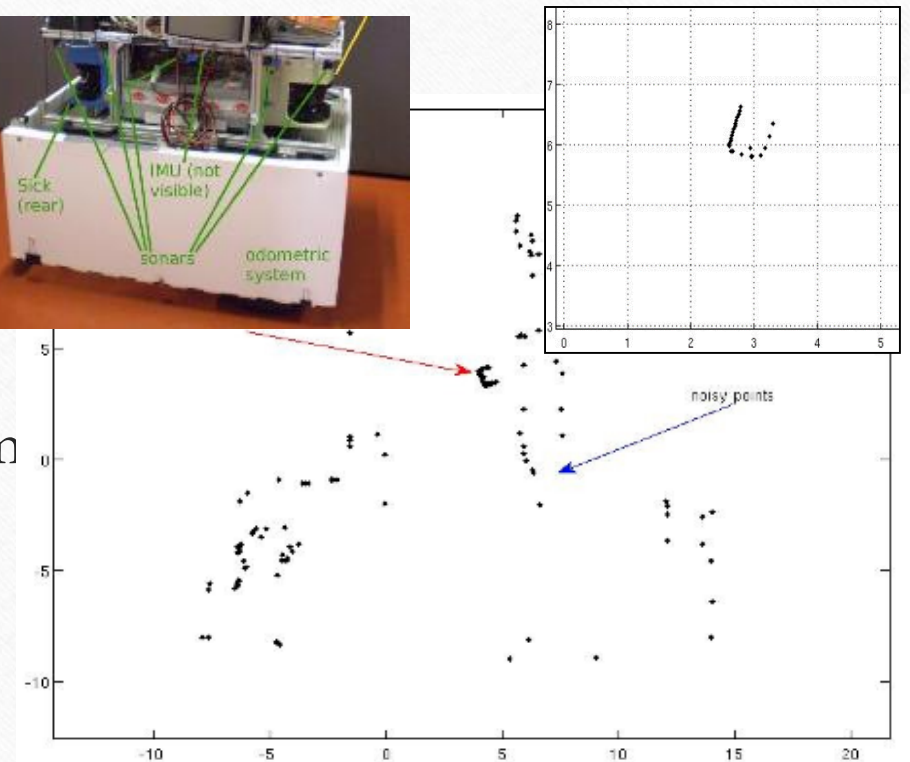
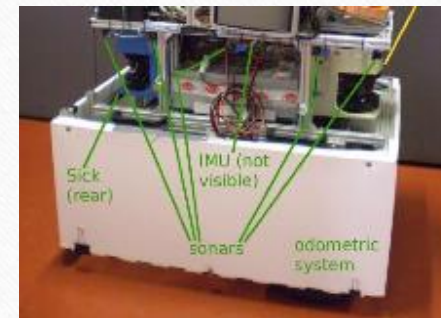
# Vision and Laser Indoor GT System

- Use a camera network to localize the robot
  - Good: Independent sensor
  - Bad: Requires (painful) setup/calibration
  - Doubt: Might not be accurate enough



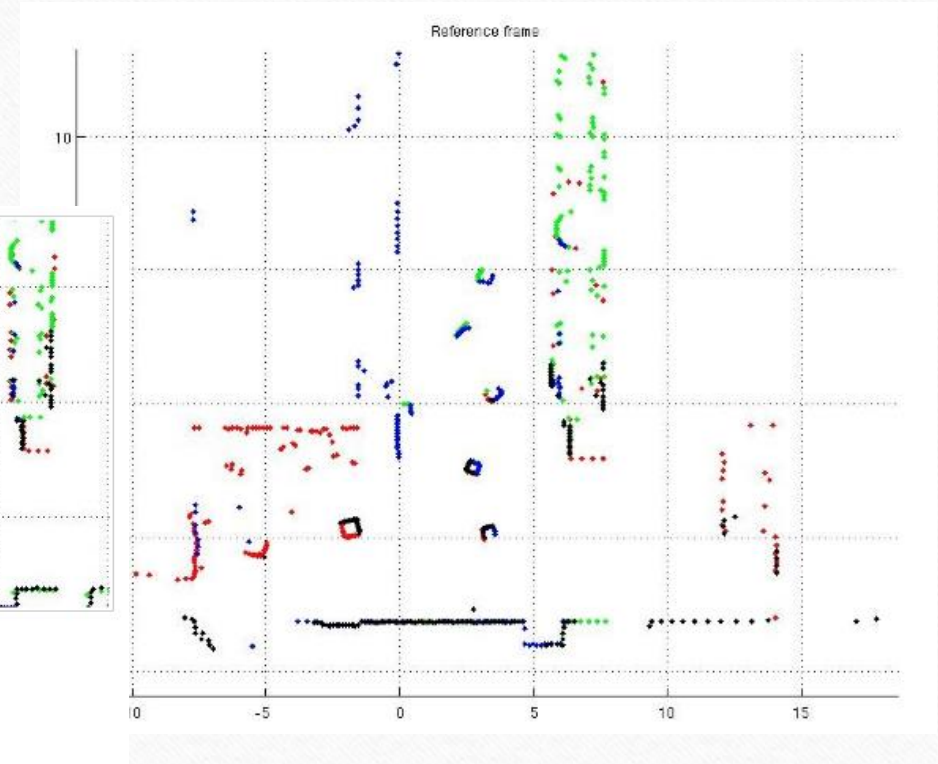
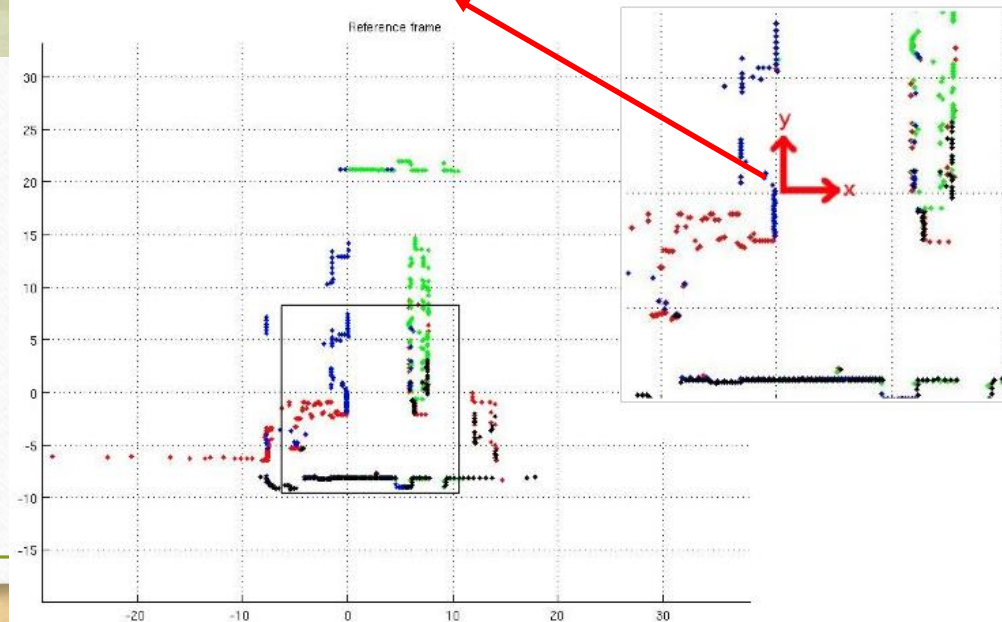
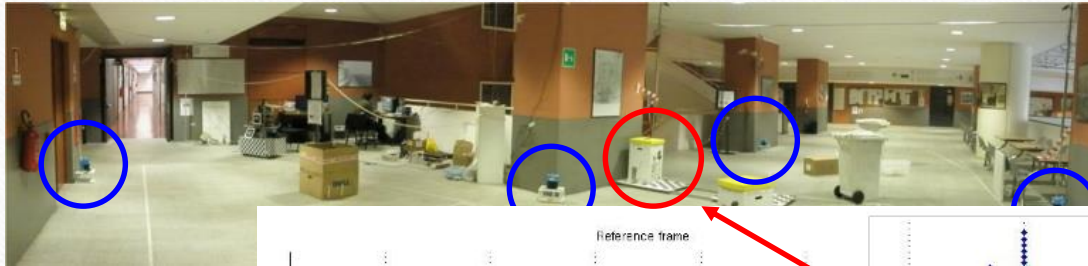
# Vision and Laser Indoor GT System

- Use a camera network to localize the robot
  - Good: Independent sensor
  - Bad: Requires (painful) setup/calibration
  - Doubt: Might not be accurate enough
- Improve accuracy by an (offboard) laser system
  - 4 sick laser-scanners in the Vision GT are
  - robot localization with ICP in the overall scan

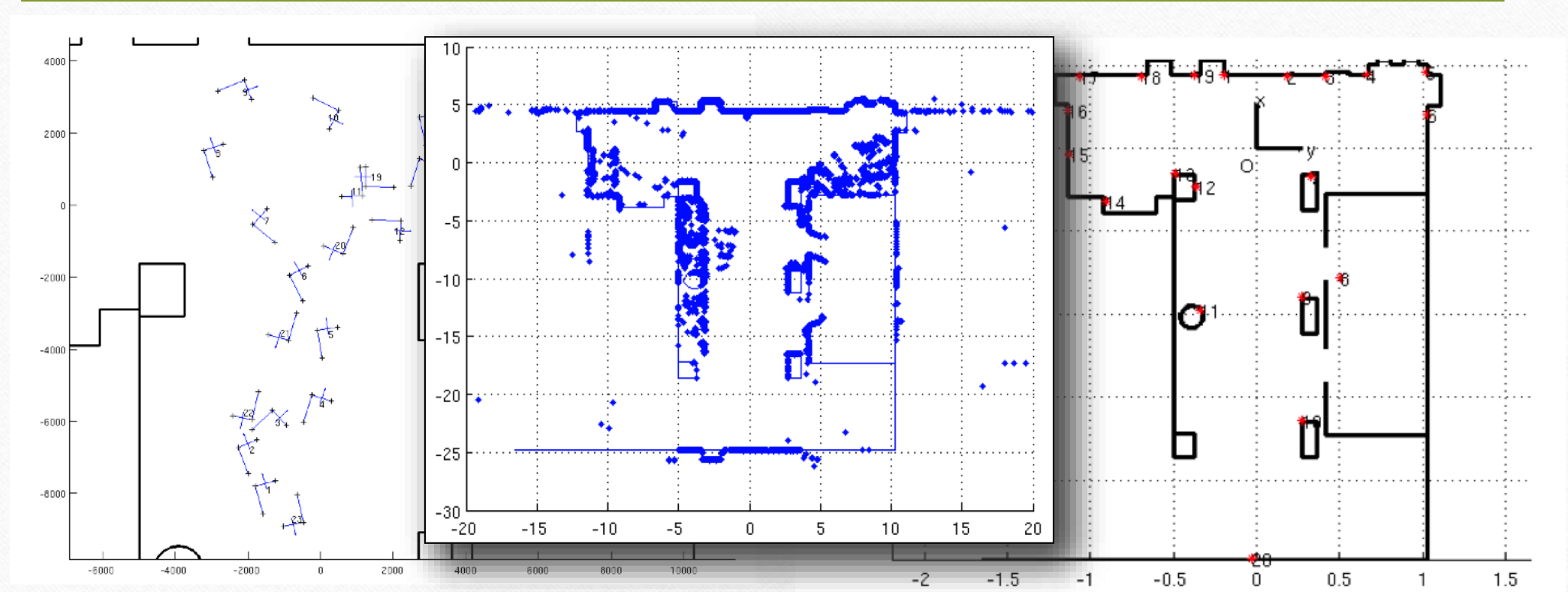




# Issue #3.1: Indoor GT Systems Alignment

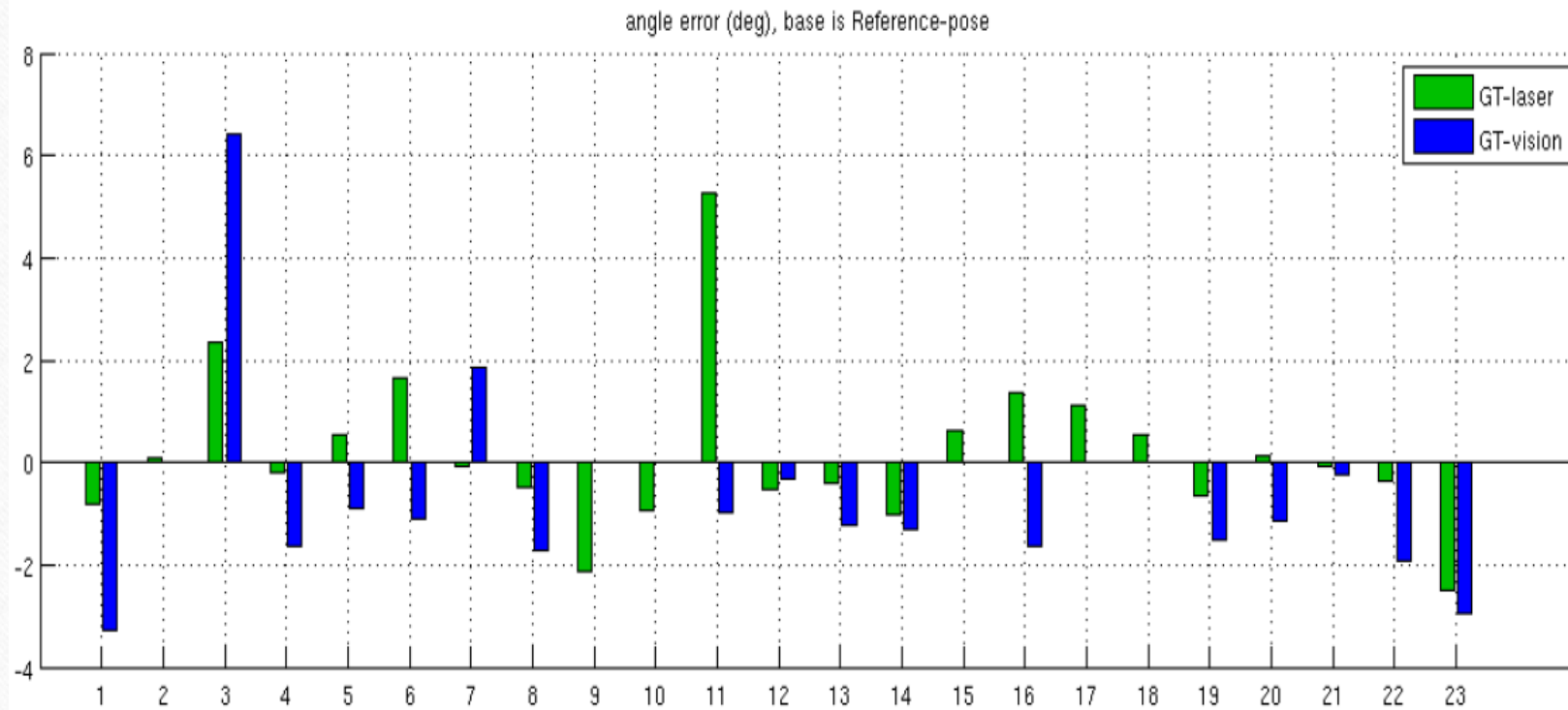


## Issue #3.2: Indoor GT Validation



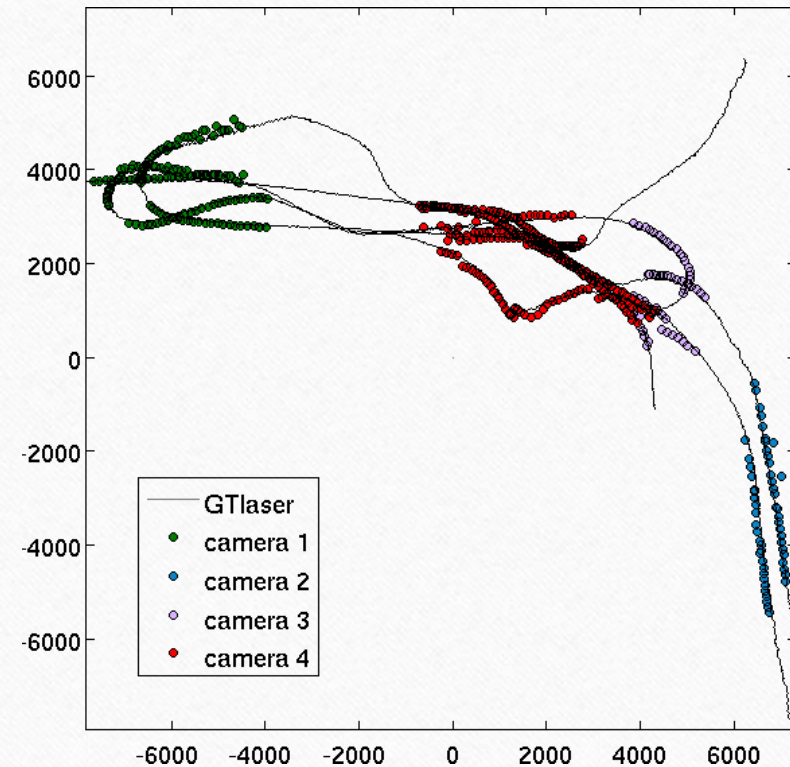


## Issue #3.2: Indoor GT Validation



## Issue #3.2: Indoor GT Validation

- Vision GT
  - $112 \pm 90\text{mm}$  in position
  - $-0.8 \pm 2.16$  degs in orientation
- Laser GT
  - $20 \pm 11\text{mm}$  in position
  - $0.15 \pm 1.56$  degs in orientation
- Overall Accuracy
  - $19 \pm 11\text{mm}$  in position
  - $-0.12 \pm 1.56$  degs in orientation





# Issue #4: Is it any useful?

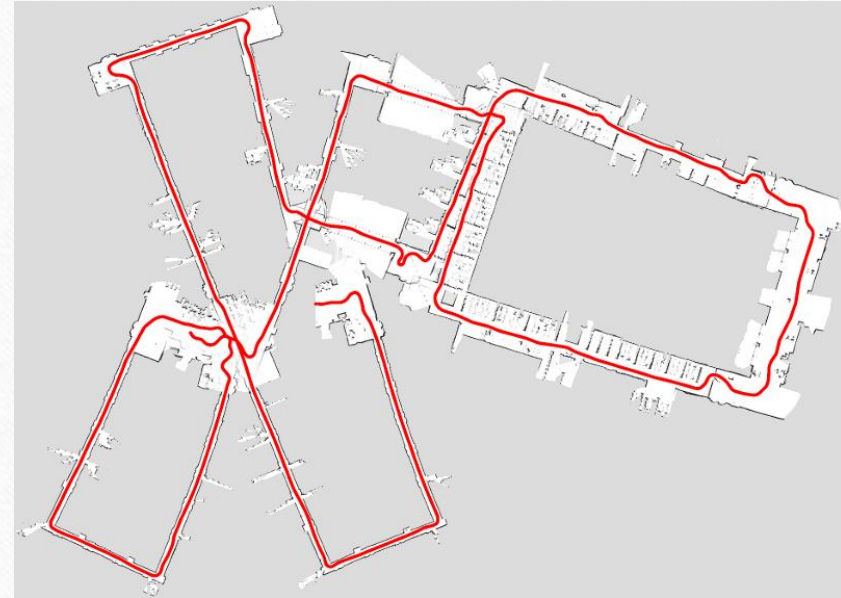
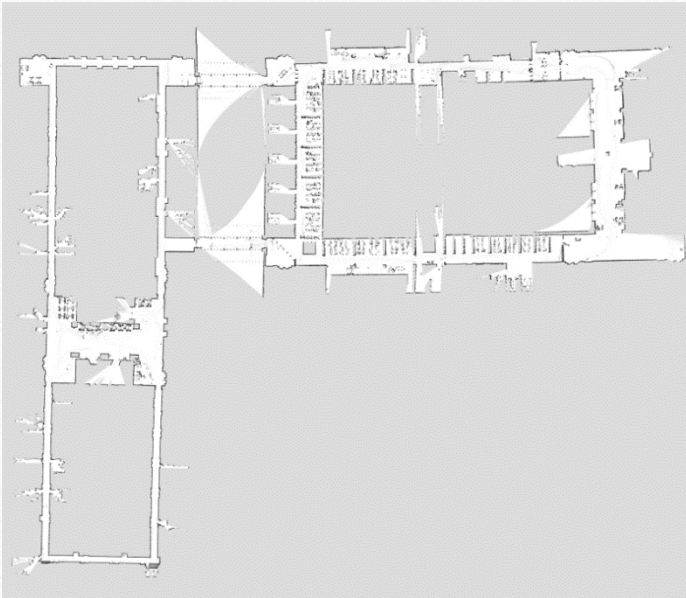
---

- Ready to use solutions from the partner used to validate the benchmark
  - Laser Based
    - Scan-matching [ALUFR]
    - Rao-Blackwellized Particle Filters [ALUFR]
    - Graph-based SLAM [ALUFR]
  - Vision Based
    - Monocular and Stereo SLAM [UNIZAR]
    - Trinocular SLAM [UNIMIB + POLIMI]

# Laser Based SLAM (indoor)

---

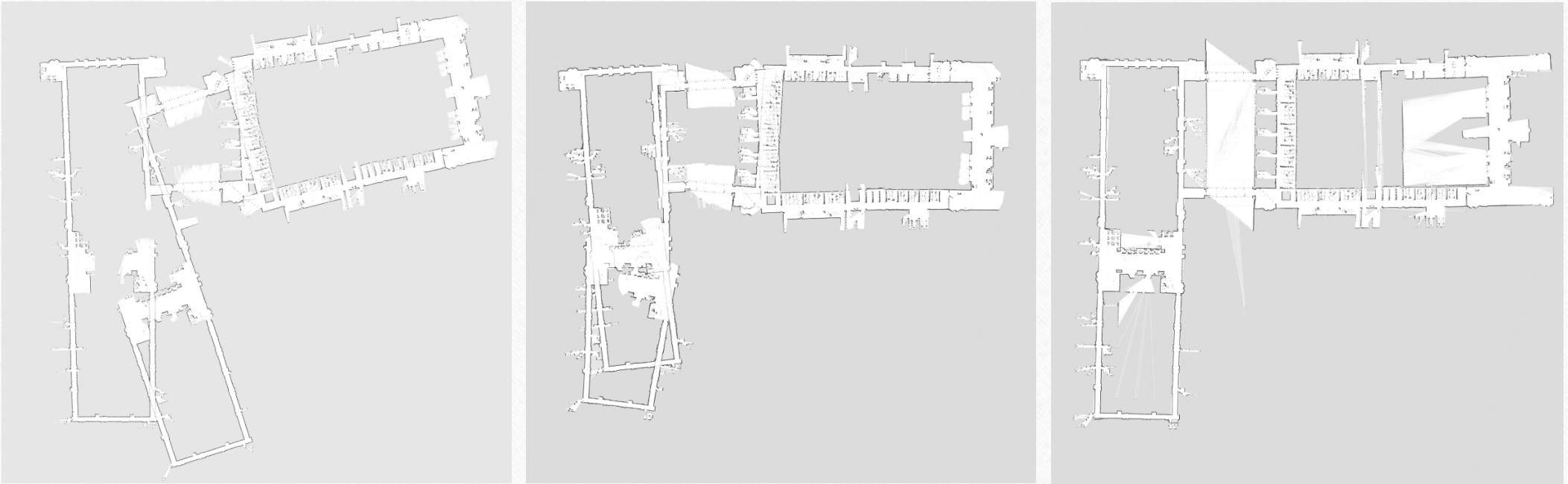
- Map ground truth obtained by manual alignment (left) and odometry (right)





# Laser Based SLAM (indoor)

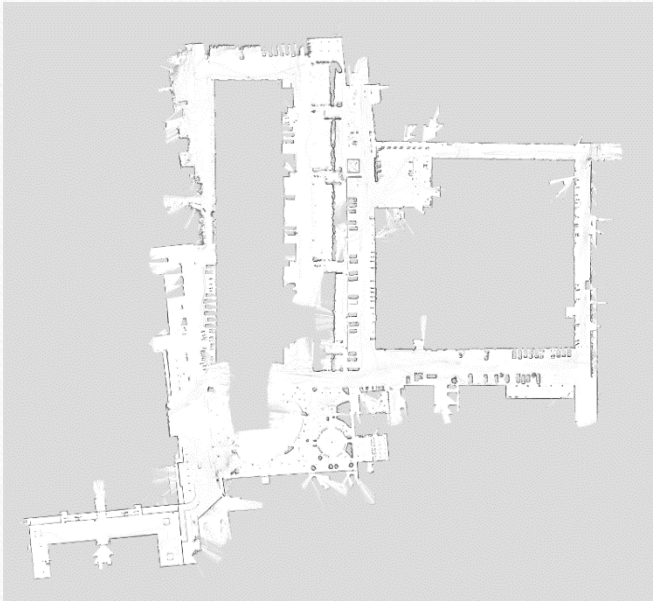
- Metrics capture the expected improvements (vasco, rbpf, graph-mapper)



# Laser Based SLAM (outdoor)

---

- Map ground truth obtained by manual alignment (left) and odometry (right)

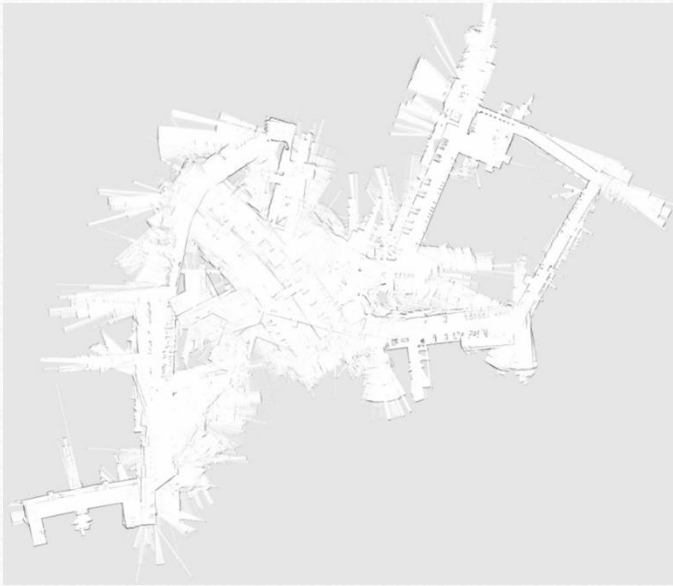




# Laser Based SLAM (outdoor)

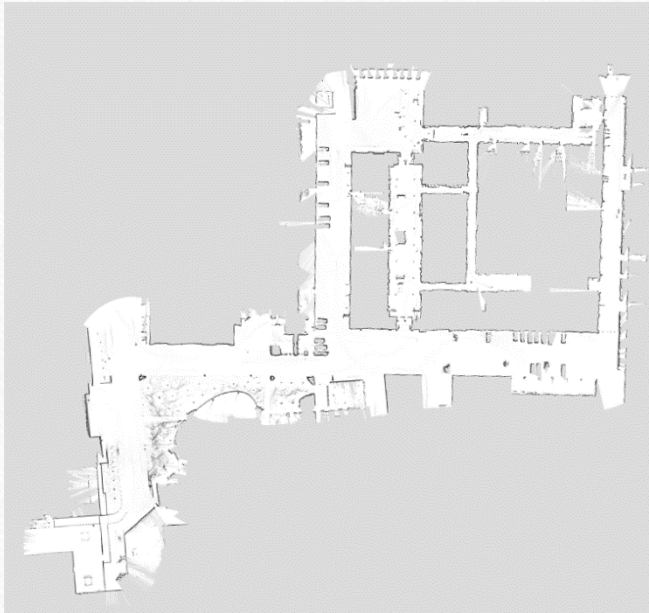
---

- Metrics capture the expected improvements (vasco, rbpf, graph-mapper)



# Laser Based SLAM (mixed)

- Map ground truth obtained by manual alignment (left) and odometry (right)

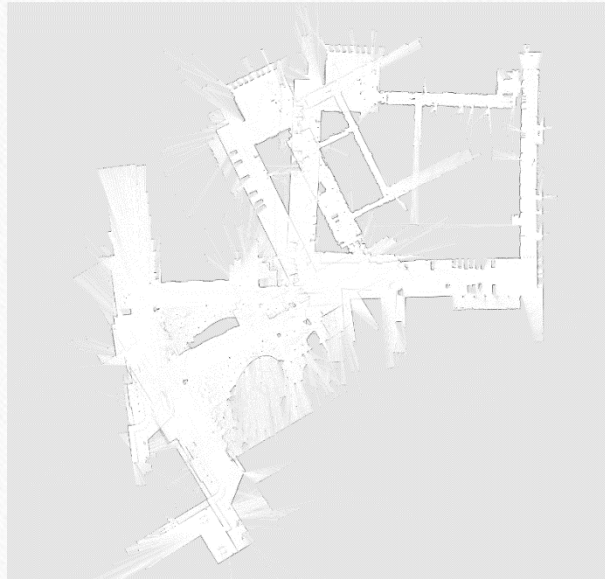




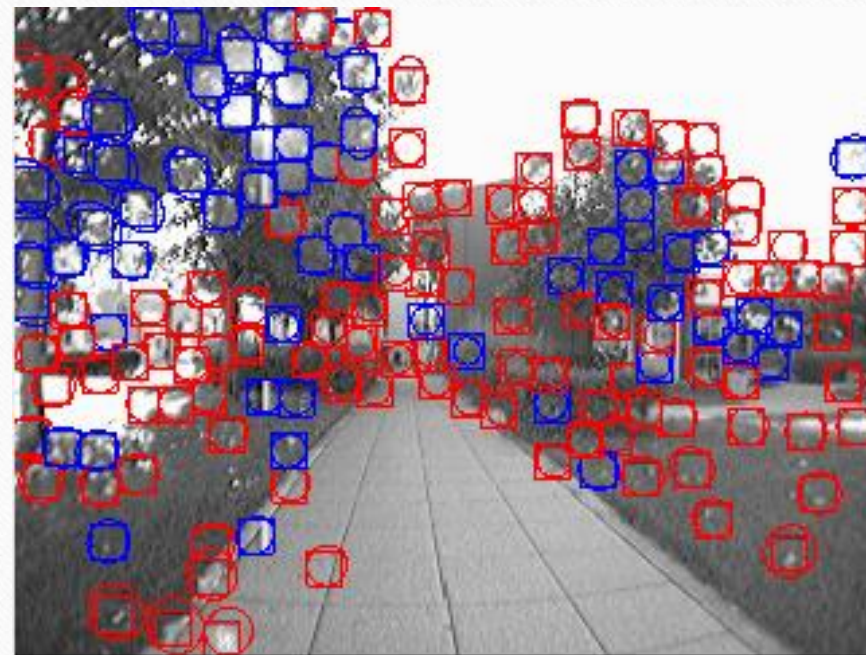
# Laser Based SLAM (mixed)

---

- Metrics capture the expected improvements (vasco, rbpf, graph-mapper)



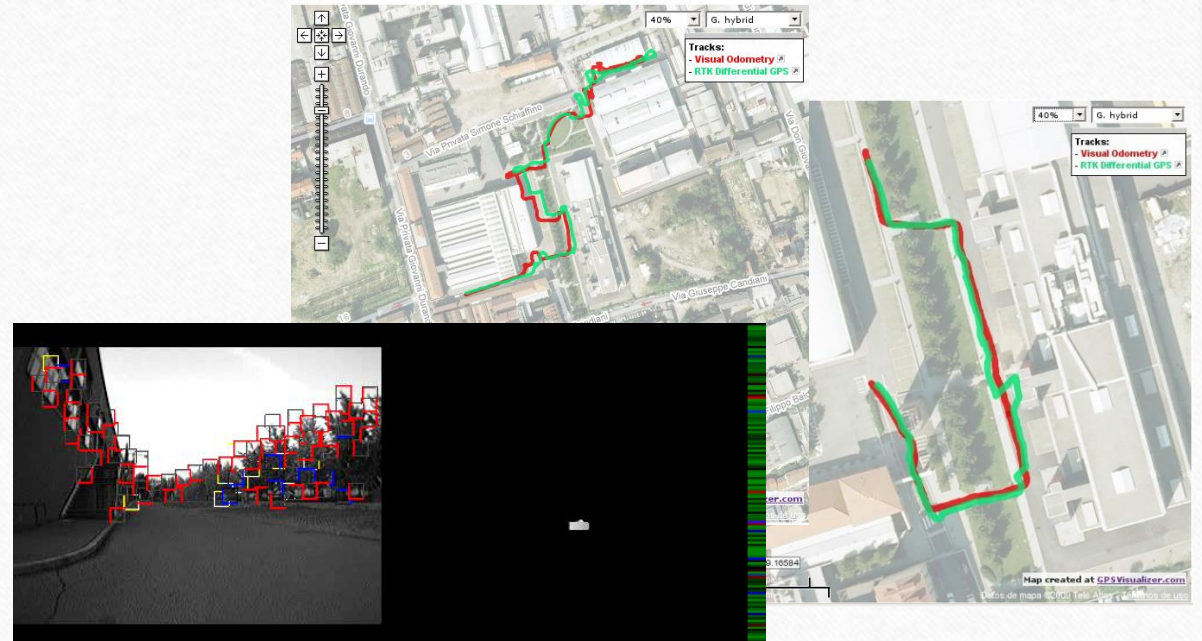
# Monocular SLAM





# Monocular SLAM Results

- 153m trajectory (5400 frames), 650m trajectory (24180 frames)
  - Low error ( $\sim 1\%$  of the trajectory)
  - Longest trajectories ever using filtering-based visual estimation
  - Near real-time processing ( $\sim 1$  second per frame)
  - Efficient spurious search based on RANSAC



# Conclusions & Seeds for Discussion

- The RAWSEEDS benchmarking toolkit still available!
  - Multisensorial datasets with ground truth
  - Well defined benchmarks with metrics
  - Off-the shelf solutions to compare with
- What's after RAWSEEDS?
  - ~~More solutions were expected!~~
  - ~~More problems were welcome!~~
  - Different uses for the same data
  - More datasets
    - One platform is there, but collection costs!
    - Other platform datasets (e.g., UAV, cars, ...)
- SLAM is a small step, let's benchmark systems and control loops ...

