Robot Competitions Kick Innovation in Cognitive Systems and Robotics

Matteo Matteucci (POLIMI)

Benchmarking - Learnings from RoCKIn

Benchmarking - Learnings from RoCKIn

- RoCKIn Project in a nutshell (more to come later)
  - Design and execute two robot benchmarking competitions
  - Involve as much as possible the general public, the academic community and industries
- Current stage of the project (just a sketch)
  - Camps to involve participants are ongoing (Jun 2013, Jan 2014)
  - Dates and locations have been selected (Nov 2014, Nov 2015)
  - Competition rules will be presented @ERF (TBC)
- Benchmarking activities
  - Inspire the design of competitions to allow for benchmarking
  - Design suitable metrics for the competitions (not done yet)
  - Apply such metrics during the competitions (not done yet)
  - Compare results after the two competitions (not done yet)

RoCKIn Lesson #0: Never let someone else decide the title of your talk!

Today's Special …

- The RoCKIn Project
  - Benchmarking through competitions
  - Set up scientific robot competitions
- Benchmarking through Competitions
  - Challenges vs. Competitions
  - Competitions - Benchmarking – Experiments
- On the Benchmarking of Modules and Systems
  - A functional reference platform for benchmarking
  - Benchmarking functionalities
  - Benchmarking systems

Robot Competitions Kick Innovation in Cognitive Systems and Robotics

Pedro U. Lima (IST-ID), Daniele Nardi (UNIROMA1), Gerhard Kraetzschmar (BRSU), Rainer Bischoff (KUKA), **Matteo Matteucci (POLIMI)**, Graham Buchanan (INNO**)**

The RoCKIn Project

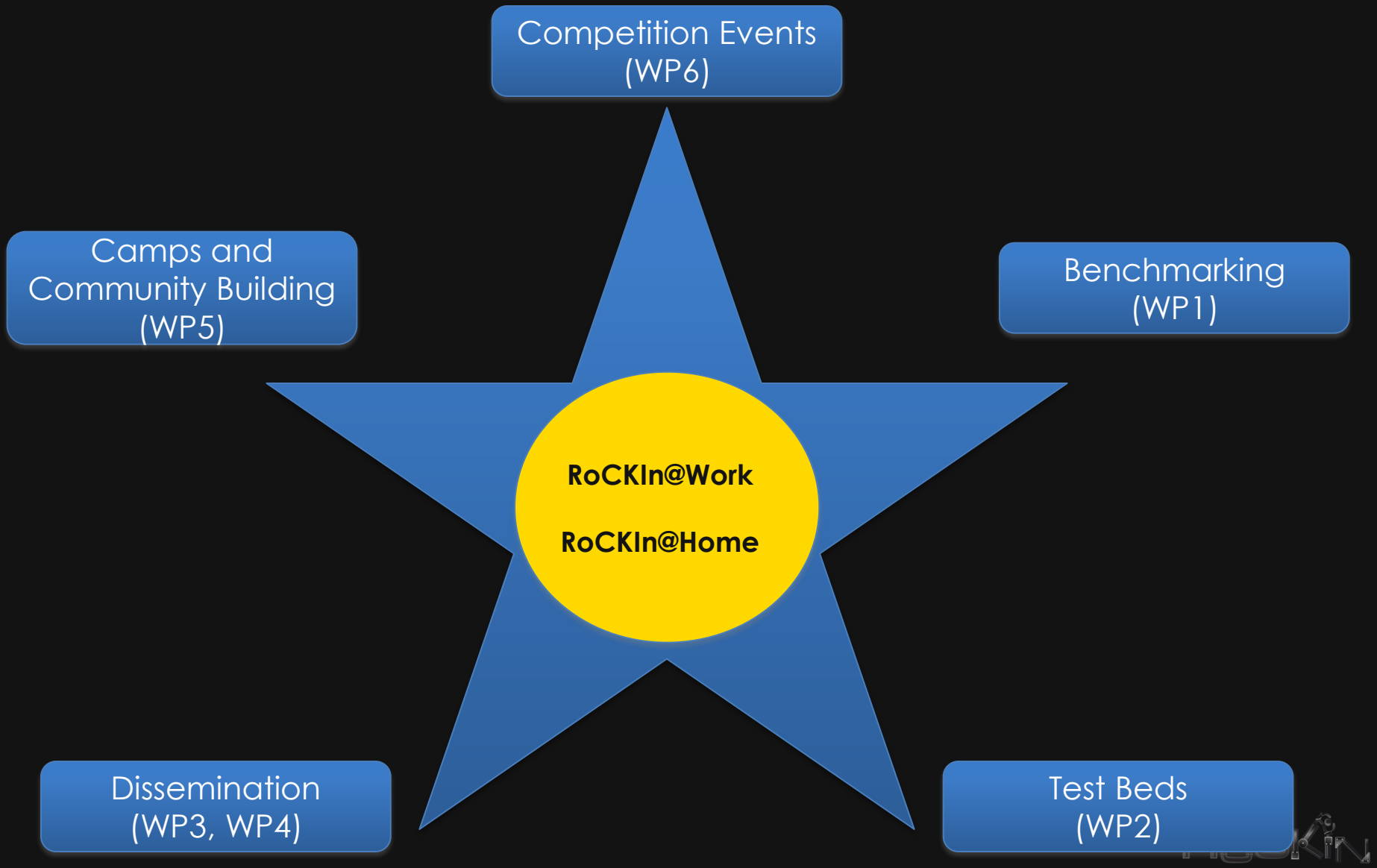
Competitions lead to…

- Innovation
  - Competitions are a powerful means to foster progress in R&D and to introduce best practices
  - Best practices in relevant domains lead to technology transfers
  - thus, competitions can be seen as catalyst for smarter, more dependable robots.
- Focused R&D
  - research challenges derived from real-world problems
  - development of commonly accepted testbeds & benchmarks
  - experimental validation of state-of-the-art research
- Better awareness of new technologies among citizens
- Higher attractiveness of scientific and engineering disciplines
  - primary / high school education
  - university level education

The RoCKIn idea

- Build upon the well-established infrastructure of **RoboCup** competitions **plus**:
  - Introducing the **networked robot systems** (multiple robots, multiple sensors and devices in the environment)
  - adding further **natural interaction** between robots and humans, based on cognitive systems and principles
  - reducing the number and importance of subjective evaluation *vs* **objective evaluation in the competitions**
  - revising the evaluation criteria of tests such that the overall measure combines the **quality of subsystems** and the success in performing the **overall task**
  - applying more care in the design and revision of the competition rules for **better comparison across years**
  - **lowering the entry barrier for new teams**

Our Coordinator favorite picture ...



RoCKIn@Work

- Innovative robot applications in industry that:
  - Work interactively with humans
  - Have reduced initial programming requirements
  - Have enhanced physics simulation capabilities

- Contribute to the continued commercial competitiveness of European industry

RoCKIn@Home

- Socially beneficial domestic service robots that
  - Have enhanced networking and cognitive abilities
  - Support the impaired and the elderly
- Contribute to an improved quality of life for the population of Europe

Forthcoming camps and related events

- **RoCKIn Camp**
  - 26-30 January, 2014 – Rome, Italy
  - Small groups working together to build upon 2013 introductory event
  - 30+ applications from 15 countries, including non-EU
  - Aim is to educate, advance current tech and develop robots appropriate for RoCKIn competitive events
  - Support from IEEE RAS Summer School Program
- **RoCKIn Field Exercise**
  - Planned for 2015
  - Supports the final stage of team development
  - Opportunity to test creations in a real world situation
- **Forums and workshops**

The RoCKIn consortium

Advisory Board Members:
Adam Jacoff, NIST, USA
Bill Smart, Oregon State University, USA
Bruno Siciliano, University of Naples Federico II, Italy
Jon Agirre Ibarbia, Tecnalia, Spain
Manuela Veloso, Carnegie-Mellon University, USA
Oskar von Stryk, Technical University of Darmstadt, Germany
XiaoPing Chen, University of Science and Technology of China, China

Experts Board:
Alessandro Saffiotti, Örebro University, Sweden
Herman Bruyninckx, University of Leuven, Belgium
Tijn van der Zant, University of Groningen, The Netherlands

SEVENTH FRAMEWORK PROGRAMME

Competition events

- **2014:**
  - Cité de L'Espace, Toulouse, 26-30 Nov. 2014 (European Robotics Week)

- **2015:**
  - Planned for European Robotics Week, Lisbon, end of November 2015

Yet another robot competition?

- Quite a number of Robotics Competitions exists already:
  - The DARPA Grand Challenge
  - Robocup Soccer / @Home / @Work
  - ICRA 2013 Robot Challenge
  - FIRST Lego Competitions
  - Robot Cleaning Competition
  - …

http://robots.net/rcfaq.html#LNK034

- They (or most of them) lack "scientificity"
  - Sometimes it is not clear the scientific question they answer
  - Their results cannot be used as benchmarking tools
  - Their results are not replicable
  - They are not "benchmarking competitions"

Robot Competitions Kick Innovation in Cognitive Systems and Robotics

Matteo Matteucci (POLIMI)

Benchmarking Competitions

Challenges vs Competitions

- A Competition is
  - something like a sporting event where there can only be one winner (excluding ties); the winner is determined as a function of relative position
  - about ranking and comparing participants
- A Challenge is
  - an event where there can be multiple winners because winning is determined as a function of achievement
  - about reaching a possibly ambitious objective

The Marathon Example

- Winning a specific marathon is a competition. The winning time of any one race has no bearing on the outcome of other races.
- Finishing a marathon is a challenge – any runner will congratulate any other on the accomplishment of running that 26.2 mile race.

Some challenges criticism …

- A Competition is
  - something like a sporting event where there can only be one winner (excluding ties); the winner is determined as a function of relative position
  - about ranking and comparing participants
- A Challenge is
  - an event where there can be multiple winners because winning is determined as a function of achievement
  - about reaching a possibly ambitious objective

Both contribute to the advancement of Robotics

- RoboCup has been the stating point of the Kiva System
- The Darpa Grand Challenge has boosted autonomous cars research and market

RoCKIn Lesson #1: Agree on a common terminology!

Hey look ma! No hands!

"one-time demonstrations of robot performance (e.g., grand challenges or other competitions) in robotics are one way of comparing the performance of robots, but they do not necessarily prove that one's robotics research is consistently better or worse than another lab's. Furthermore, unless the robots are specifically designed to test the effectiveness of particular aspects of robots (e.g., quadruped vs. biped), then these competitions do not necessarily offer generalizable solutions for future robotics research projects."

Leila Takayama (Google[x], formerly at Willow Garage)
"Towards a Science of Robotics: Goals and Standards for Experimental Research",
RSS 2009 Workshop on Good Experimental Methodology in Robotics

Gotcha! Everything boils down to Experimental Research!
So lets change the question!

Competition and Experiments

Can Competitions be treated as scientific experiments?

"Challenge and competition events in robotics provide an excellent vehicle for advancing the state of the art and evaluating new algorithms and techniques in the context of a common problem domain. [...] treat competitions and challenges as repeatable experiments."

Monica Anderson, Odest Chadwicke Jenkins, and Sarah Osentoski
Recasting Robotics Challenges as Experiments, IEEE Robotics & Automation Magazine, June 2011, 10-11

Experiments vs. Competitions

- How do experiments and competitions (usually) differ?
  - An experiment should be repeatable, while a competition is usually held once and it is not aimed at being repeated under exactly the same conditions
  - An experiment should be reproducible, while the specifications of competitions are often (probably intentionally) vague
  - An experiment evaluates a specific hypothesis while a competition usually evaluates general abilities
  - An experiment describes the whole system while in competitions the systems are not necessarily known
  - An experiment is aimed at explaining why a result has been obtained, a competition often provides only a ranking of competitors.
  - Competitions push to development of solutions, experiments to exploration of phenomena and sharing of results

Competition as Experiments

Competitions should aim at becoming benchmarks adopting a scientific approach (in goals and methods)

"Scientific" means able to increase scientific and technological knowledge by using rigorously experimental method

The experimental method suggest experiments to be designed to allow for:

- Comparison
- Reproducibility / repeatability
- Justification / explanation

What Makes an Experiment

Comparison: to know what has been already done in the field, to avoid the repetition of uninteresting experiments, and to get hints on promising issues to tackle.

Reproducibility and repeatability: they are related to the idea that scientific results should be severely criticized to be confirmed; reproducibility is the possibility for independent scientists to verify the results of a given experiment by repeating it with the same initial conditions, instruments and techniques; repeatability is the property of an experiment that yields the same outcome from a number of trials performed at different times and/or in different places.

Justification and explanation: it is not sufficient to collect as many precise data as possible, but it is also necessary to look for an explanation, namely all the experimental data should be interpreted in order to derive the correct implicaions that lead to the conclusion.

Benchmarking Competitions

Competitions often lack scientific grounding

- They do not apply the "scientific method" to allow comparison, reproducibility and repeatability, justification and explanation
- As for justification and explanation, they produce a ranking, but few insights on the motivations for this ranking
- Their results cannot be used as benchmarking tools

The Benchmarking through Competition Challenge

"Design competitions to make them more scientifically grounded and suitable as benchmarks"

Competitions as Experiments

- Minimum requirements: reproducibility and repeatability should be guaranteed
  - **Reproducibility** is the possibility to verify, in an independent way, the results of a given experiment. It refers to the fact that other experimenters, different from the one claiming for the validity of some results, are able to achieve the same results, by starting from the same initial conditions, using the same type of instruments, and adopting the same experimental techniques.
  - **Repeatability** concerns the fact that a single result is not sufficient to ensure the success of an experiment. A successful experiment must be the outcome of a number of trials, performed at different times and in different places.

Why Benchmarking Competitions?

- Robotic competitions have positive effects
  - They are appealing (people like to compete)
  - They take place with regularity and precise timing
  - They showcase current state of the art in research / industry
  - They switch the focus from specific subsystems towards complete systems and highlight the influence of integration
  - They promote critical analysis of experiments out of labs
  - They share among participants the cost and effort of setting up complex experimental installations

- Benchmarking has some drawbacks, but we need it
  - It is time consuming
  - It has a small return on investment
  - Not suitable tools available
  - It is not sexy

Non-Robotic Scientific Competitions

- Scientific Competitions treated as (paper) experiments:
  - Machine Learning and Pattern Recognition (e.g., Kaggle)
  - Computational Intelligence in Games (e.g., CIG)
  - Information Retrieval (e.g., TREC)
  - Computer Vision (e.g., PETS)

- Most of them have nowadays reached the level of
  - Defining proper metrics to measure significant aspects of the scientific result (e.g., F-measure)
  - Having different testbeds/tasks/scores (with different features) used to avoid overspecialization (e.g., background subtraction)
  - Investigating general features of the tasks and testbed used to design new competitions from an application perspective

RoCKIn Competitions

- Design of RoCKIn Benchmarking Competition to
  - be relevant w.r.t. relevant scientific challenges
  - consider distributed sensing and multi robot systems
  - advance the state of the art of robot competitions
  - result in systems which are general w.r.t. the relevant scientific challenges (i.e., not just a specific robot with a specific sensor in a specific setting)
  - engage scientific & industrial community in participating
  - involve people and general public in a successful event
  - allow meaningful comparison of heterogeneous systems
  - provide quantitative evaluation metrics (numeric / pass or fail)
  - be a specific way of performing experimental evaluation
  - be reproducible and repeatable

RoCKIn Lesson #2: benchmarking competition design is much more complex than expected!

Do not forget RoCKIn Lesson #1 !!!

- Lets agree on the use of the following terms :
  - ***Task***: an activity (or set of), usually requiring some functionalities, a robot system is required to perform in a (benchmarking) competition (e.g., "bring me the glasses")
  - ***Functionality***: one of the basic abilities a robot is required to possess in order to perform a task and thus be subjected to a benchmarking experiment (e.g., self-localization, and grasping)
  - ***Metric***: a precisely defined, quantitative criterion to assess one or more aspects of the performance of a robot system in the context of the execution of a task
  - ***Testbed***: a physical installation which sets a platform for scientific and technological experimentation in the context of a benchmarking competition by including the elements of the environment that the participating robot systems interact with

Functional and task benchmarks

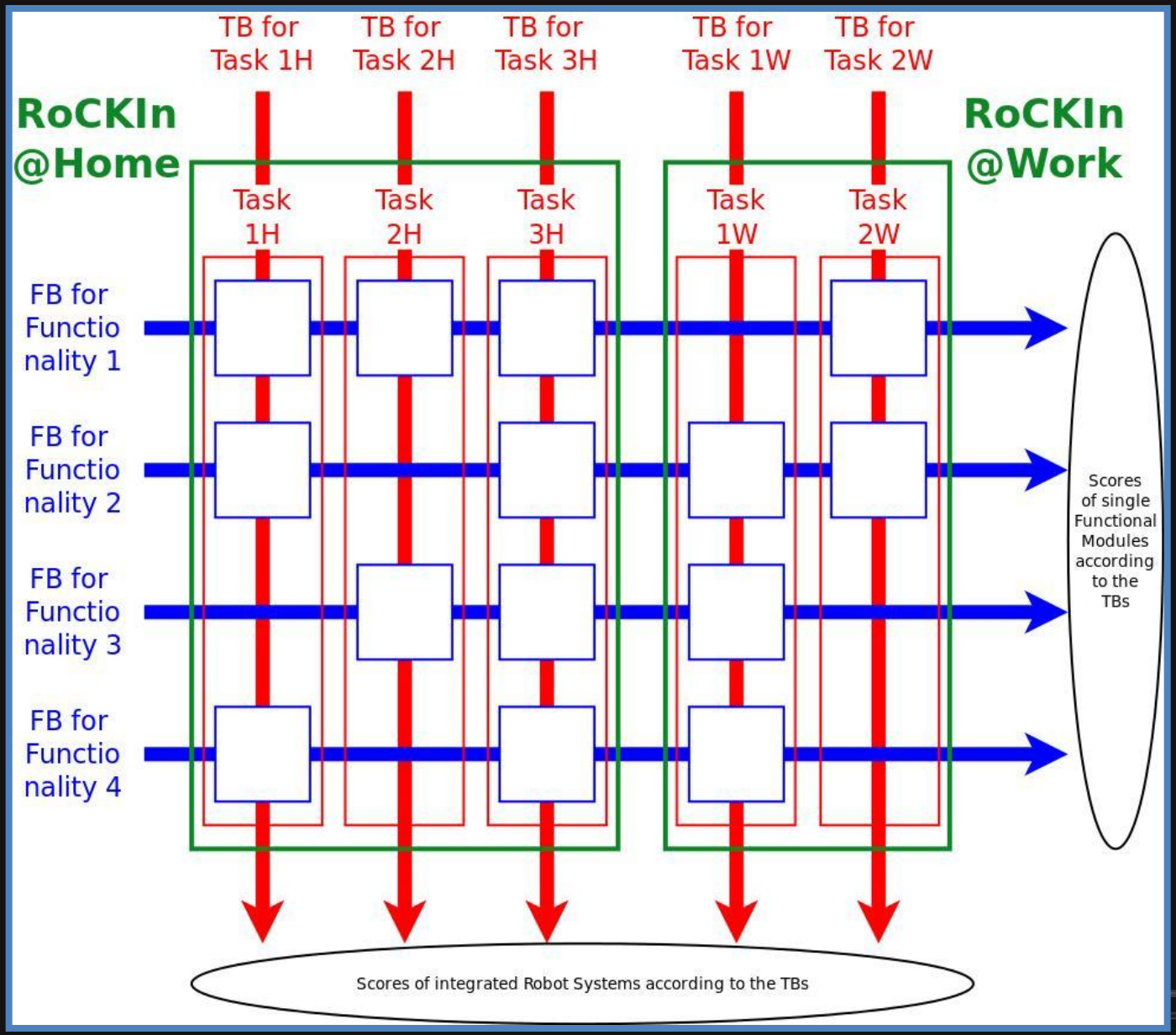
- Competitions can challenge robots at two different levels (ability vs capability in SRA jargon?)
  - Task Level: evaluation of whole systems on a specific task (e.g., the "bring me a beer" tasks)
  - Functionality Level: evaluation of modules implementing, in a general manner, functionalities required by the competition tasks (e.g., grasping and manipulation)

- Benchmarking competitions should allow independent evaluation at both levels
  - To encourage participation of people interested in specific aspects of robotics (e.g., object recognition)
  - To evaluate at what extent the Interplay among modules is relevant (e.g., the precision in positioning before grasping)

Do not forget RoCKIn lesson 1 !!!

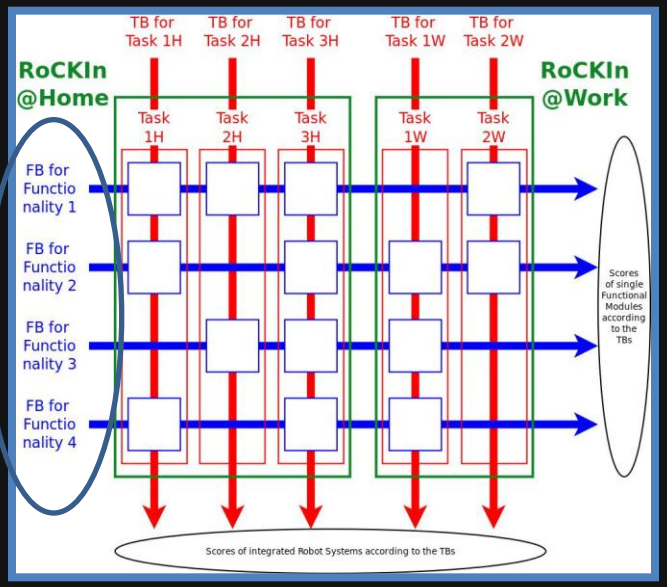
- Lets agree on the use of the following terms (continued):
  - ***Benchmarking experiment (Benchmark):*** the composition of a task to be performed by a robot system and the testbed where the robot system performs the test plus a set of metri5cs to evaluate those (task + testbed + metrics)
  - ***Functional benchmark:*** benchmark aiming at evaluating the quality and effectiveness of a specific functional module of a robot system in the context of one or more scenarios
  - ***Task benchmark:*** benchmark aiming at evaluating the quality of the overall execution of a task by a robot system in the context of a single scenario and technological experimentation in the context of a benchmarking competition by including the elements of the environment that the participating robot systems interact with

My favorite RoCKIn picture



One example (not from RoCKIn yet)

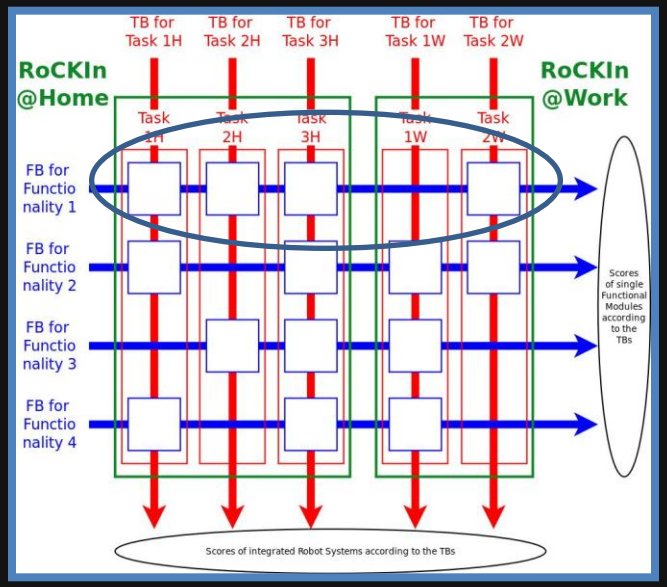
- Functionalities 1 to 4 are available to the robot
  - Autonomous navigation;
  - Object recognition;
  - Grasping and manipulation;
  - Processing of voice commands.



- *Task Benchmark 2H*: "The robot is provided with a map of the environment. It must enter the testbed, navigate through it to reach an object located in a predefined position, and pick it up".
- *Task Benchmark 1W*: "The robot is located in a specified pose in front of a table. Over the table are located randomly (but according to suitable specifications) 5 identical mugs which differ only in their color. The robot receives a voice command from a human, specifying the color of a mug, then it picks up the required mug".

One example (not from RoCKIn yet)

- Functionalities 1 to 4 are available to the robot
  - Autonomous navigation;
  - Object recognition;
  - Grasping and manipulation;
  - Processing of voice commands.



- *Functional Benchmark 2*: "Recognize 10 objects, randomly selected out of all possible objects from RoCKIn@Home and RoCKIn@Work databases. Category, size , position, and color have to be returned."
- *Functional Benchmark 3*: "Grasp and lift firmly10 different objects, randomly selected out of all predefined objects from RoCKIn@Home and RoCKIn@Work, in a given working space. The pose of each object is sent to the robot at the beginning of the test."

A new quest …

- Design the competitions to
  - Allow people interested in specific functional benchmark to participate only to those
  - Stimulate people to tackle both functional/module and task/system benchmarks
- Provided that we are able to have
  - **Functionality benchmarks :** investigate the performance of a specific module in a deeper and more general way with respect to task-level benchmarks.
  - **Task-level benchmarks:** evaluate whole-system performance over a limited set of situations, taking into account all system modules as well as their interaction.

Q: Is it possible, by jointly combining those to acquire information about higher-level properties of the system, such as *quality of system integration or interaction issues among modules*?

A: I have no clue about the answer ☹

RoCKIn "Episode I" and related tasks (TBC)

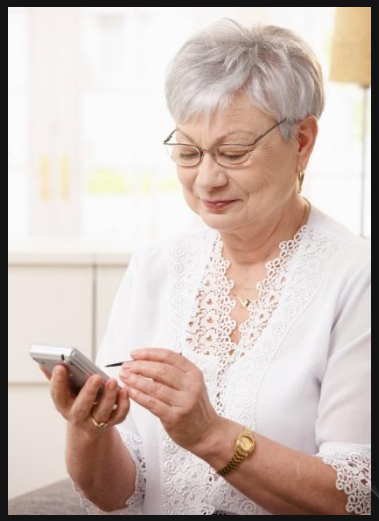
User Story: Granny Annie

Granny Annie is a nice but slightly seasoned lady. Luckily, she could get into a new program, sponsored by her health & social security insurances, by which elderly people are supplied with household and elderly care robots to assist in managing and mastering their daily lives

- User task 0: "Welcome the robot"

Granny Annie is waking up and today she feels a bit tired because she has not slept very well. Still a number of tasks need to taken care of. The home robot will help her in all these tasks.

- User task 1: "Cater for bedroom comfort"
- User task 2: "Handle the home pets"
- User task 3: "Find the reading glasses"
- User task 4: "Welcome a visitor"



Toward a Functional Reference Platform

Different functionalities different benchmarks:

- Some capabilities of robots can be measured and benchmarked "externally" (e.g., position of picked-and-placed objects)
- Some capabilities of robots require "internal" data to be measured and benchmarked (e.g., accuracy of maps used for path or trajectory planning)

A Functional Reference Platform (FRP) for benchmarking

- aims at defining the capabilities of robots required by benchmarks
- aims at identifying capabilities that require "internal" data to be benchmarked and the interfaces for getting these data

What functionalities for Granny Annie? (TBC)

	welcome	bedroom	pets	glasses	visitor
Task Planning	(X)	(X)	X	X	X
Autonomy[1]	(X)	(X)	X	X	X
Geometric Mapping (or SLAM)	X	-	-	-	-
Semantic Mapping (or SLAM)	X	-	-	-	-
Self-localization	-	X	X	X	X
Path planning (mobile base)	(X)	X	X	X	X
Path following (mobile base)	-	X	X	X	X
Object recognition	(X)	X	X	X	-
Object state perception	-	X	X	-	X
Object tracking	-	-	X	-	-
Face detection	-	-	-	-	X
Face recognition	-	-	-	-	X
Path planning (arm)	-	(X)	X	X	-
Path following (arm)	-	X	X	X	-
Grasp planning	-	(X)	X	X	-
Grasp execution	-	X	X	X	-
Operate physical devices (e.g., doors, switches)	-	X	X	-	X
Input from humans through Data	-	X	-	X	-
Input from humans through Gesture	(X)	-	-	X	-
Input from humans through Speech	X	X	X	X	X
Output to humans through Display	-	-	-	X	X
Output to humans through Speech	(X)	X	X	X	X
Interactive communication with humans	-		-	X	X
Data exchange with physical devices	-	-	-	X	-
...					

Defining functional benchmarks (TBC)

- A functional benchmark is defined to evaluate whether, and possibly, to what extent a functional module provides a given functionality to a robot system

- In RoCKIn functional benchmarks involve four elements
  - *Description*: general description of the functionality, its aims, and the expected ability provided to the robot
  - *Input/output*: information available to the module implementing the functionality when executed and the expected outcome
  - *Benchmarking data*: the data needed to perform a rigorous evaluation of the functional module during the benchmark
  - *Metrics*: set of metric to evaluate the outcome of a functional module in an objective way

Functional benchmark: self-localization

- *Description:* being able to estimate the robot's own pose with respect to a reference frame in a map while moving through it
- *Input/Output:* 3DoF pose estimate(s) with respect to a known fixed reference frame in the given known map
- *Benchmarking data:* sequence of poses estimated by the robot during a path, ground truth measurement of the sequences of the poses of the robot during its movement
- *Metrics:* time required to self-localized from an unknown pose, average and maximum pose error on a given path, average and maximum pose error on a path of a given length, relative position error, …

Functional benchmark: object recognition

- *Description:*
- *Input/Output:*
- *Benchmarking data:*
- *Metrics:*

Functional benchmark: grasping

- *Description*:
  - Obtain (actively) a physical contact of the end effector with an object to make it possible transport it (e.g., to lift it and it does not fail)
- *Input/Output*:
  - End effector model, object shape/model, pose of the object, pose of end effector, contact sensing device, configuration of end effector and its trajectory, pose of the object on the end effector grasping forces
- *Benchmarking data*:
  - Depends on the class of grasping (type of gripper, type of grasping power/precision/…)
  - Pose/forces of the object on the end effector
- *Metrics*:
  - Time to complete
  - Precision in obtaining the desired configuration
  - Forces applied to the object
  - Human evaluation of politeness/safety/…

Functional benchmark: human-robot inter.

- *Description:*
- *Input/Output:*
- *Benchmarking data:*
- *Metrics:*

Functional benchmark: autonomy

- *Description:*
- *Input/Output:*
- *Benchmarking data:*
- *Metrics:*

Let me thank you all for …

- … having invited me to talk at this workshop
- … having spent so much time in listening and discussing with me about benchmarking through competitions
- … having shared with me you ideas and views about challenges and competitions
- … having shared with me your ideas and knowledge about task and functional benchmarking
- … having helped me in filling up the missing slides for next week RoCKIn project review meeting

RoCKIn Lesson #3: Robotic benchmarks definition is a complex task, it should be a community effort.

Robot Competitions Kick Innovation in Cognitive Systems and Robotics

Matteo Matteucci (POLIMI)

Benchmarking - Learnings from RoCKIn