Benchmarking Robotics

Matteo Matteucci
Politecnico di Milano

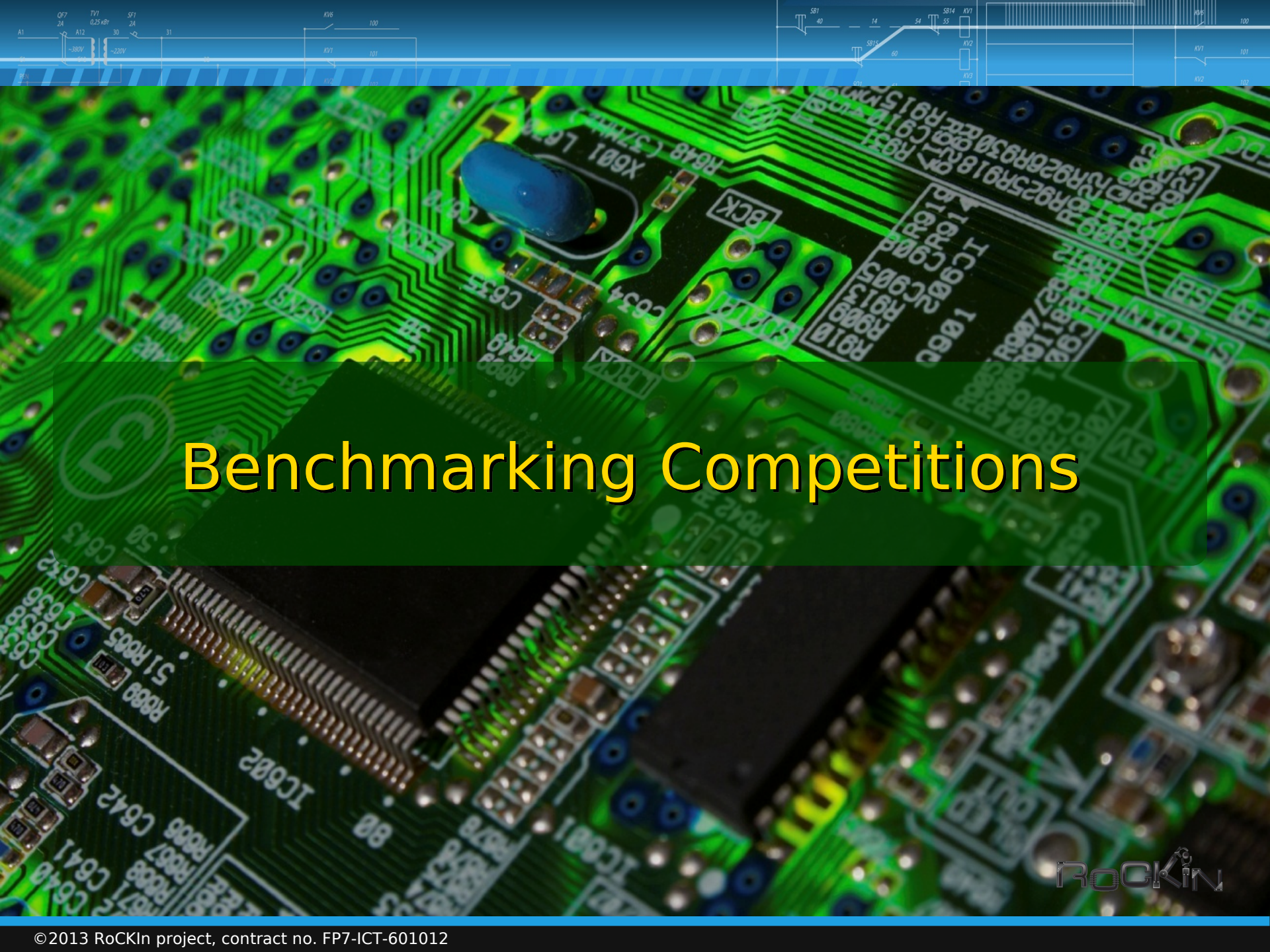
Benchmarking in Robotics

What is Benchmarking in Robotics?

“Objective performance evaluation of a robotic system or subsystem under controlled conditions”

Why Benchmarking in Robotics?

- It is a specific way of performing experimental evaluation
- It enables a comparison of different systems on a common, predefined, setting
- It provides a set of metrics (numerical scores / pass or fail / ranking / …) together with a proper interpretation to perform an objective evaluation
- It enables reproducibility and repeatability of experiments



Benchmarking Competitions

Benchmarking and Robotic Competitions

Quite a number of Robotics Competitions exists :

- The DARPA Grand Challenge
- Robocup Soccer / @Home / @Work
- ICRA 2013 Robot Challenge
- FIRST Lego Competitions
- Robot Cleaning Competition
- ...

How many of them in a year?

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

They have several positive effects, but they lack “scientificity” (e.g., their results cannot be used as benchmarking tools)

Competition and Experiments

Can Competitions be treated as scientific experiments (despite their obvious differences)?

"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 providing benchmarks by adopting a scientific approach (both 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 implications 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

“Designing competitions to make them more scientifically grounded and suitable as benchmarks”

Why Competitions to do Benchmarking?

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
- ...

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

RoCKIn Competitions are designed to

- be a specific way of performing experimental evaluation
- allow meaningful comparison of heterogeneous systems
- provide quantitative evaluation metrics (numeric / pass or fail)
- be reproducible and repeatable
- be relevant w.r.t. relevant scientific challenges
- 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)
- consider distributed sensing and multi robot systems
- advance the state of the art of robot competitions

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

RoCKIn@Work Practice

Exercise in Competition Design

Design a RoCKIn@Work competition, starting from the following scenario, to:

- Promote effective scientific and technological research advances of general application, not just engineering
- Capture the key elements of the scientific achievements required to successfully tackle the problems considered
- Allow for comparison, reproducibility and repeatability, justification and explanation

Manufacturing Logistics & Preassembly

- The environment consists of several workstations, in which either human workers or robots assemble a variety of different goods. A central scheduler gets orders for these goods and assigns them dynamically to the workstations.
- Once it is known where a product item will be assembled, the parts required for its assembly need to be available at the workstation sufficiently on time such that no delays in the production process are caused.
- Each type of product requires a different set of component items for the assembly phase. Some of these items, such as screws, nuts, and bolts, rings, etc. (commodity items) are used in all products, albeit in different numbers, while others are specific to the product type being assembled (specific items).
- All parts are kept in suitable storage containers and are delivered into adequately-sized boxes at the workstations The task to be solved by the robots fall into two categories: manufacturing logistics and preassembly.

RoCKIn Glossary

Scientific Contests

Scientific Challenge: wide-ranging scientific and technological problem (often stated with intentionally vague terms) that can only be solved in the long term, which is explicitly defined for the purpose of pushing forward and directing the state of the art in scientific and technological research. For instance the RoCKIn Scientific Challenges are:

- domestic service robots;
- innovative robot applications in industry.

Scientific Competition: contest associated to a Scientific Challenge, where a well-defined set of rules and regulations set the constraints that must be satisfied by the participating teams. The rules are designed to

- fairness and to promote effective scientific and technological research advances of general application, not just engineering
- capture some of the key elements of the scientific achievements required to successfully tackle the problems considered
- allow for comparison, reproducibility/repeatability, justification/explanation

Contests Scenario

Scenario: all the aspects of the context where a Benchmarking Competition takes place. Such aspects include physical settings, environmental features (e.g. lighting, dynamic and static elements, ...), events or sequences of events that may occur, presence of robots/people/objects, and so on.

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. A Testbed is not a completely passive element and it allows interaction, may include human beings and/or devices intended for human use, might be provided with data-collection systems to generate benchmarking information.

Parameter: a Parameter is an aspect of the specifications for a Testbed that can take different possible configurations over a specified, discrete or continuous, range. Parameters are used to introduce elements of controlled variability into a Testbed.

Scenario and Tesbed

Scenario: all the aspects of the context where a Benchmarking Competition takes place. Such aspects include physical settings, environmental features (e.g. lighting, dynamic and static elements, ...), events or sequences of events that may occur, presence of robots/people/objects, and so on.

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. A Testbed is not a completely passive element and it allows interaction, may include human beings and/or devices intended for human use, might be provided with data-collection systems to generate benchmarking information.

Parameter: a Parameter is an aspect of the specifications for a Testbed that can take different possible configurations over a specified, discrete or continuous, range. Parameters are used to introduce elements of controlled variability into a Testbed.

Tasks and Metrics

Task: operation or set of operations that a Robot System is required to perform in a Benchmarking Competition. These operations, their expected results, the way they must be executed, and the features of the environment where the operations occur can be specified more or less precisely

Subtask: a single operation or set of operations that a Robot System has to perform in order to execute a Task, but which by itself is not sufficient to achieve the final results of the Task

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. A Metric requires the application of a precisely defined algorithm to experimental data describing the execution of a Task, or comparison with the performance of some “reference” system or subsystem

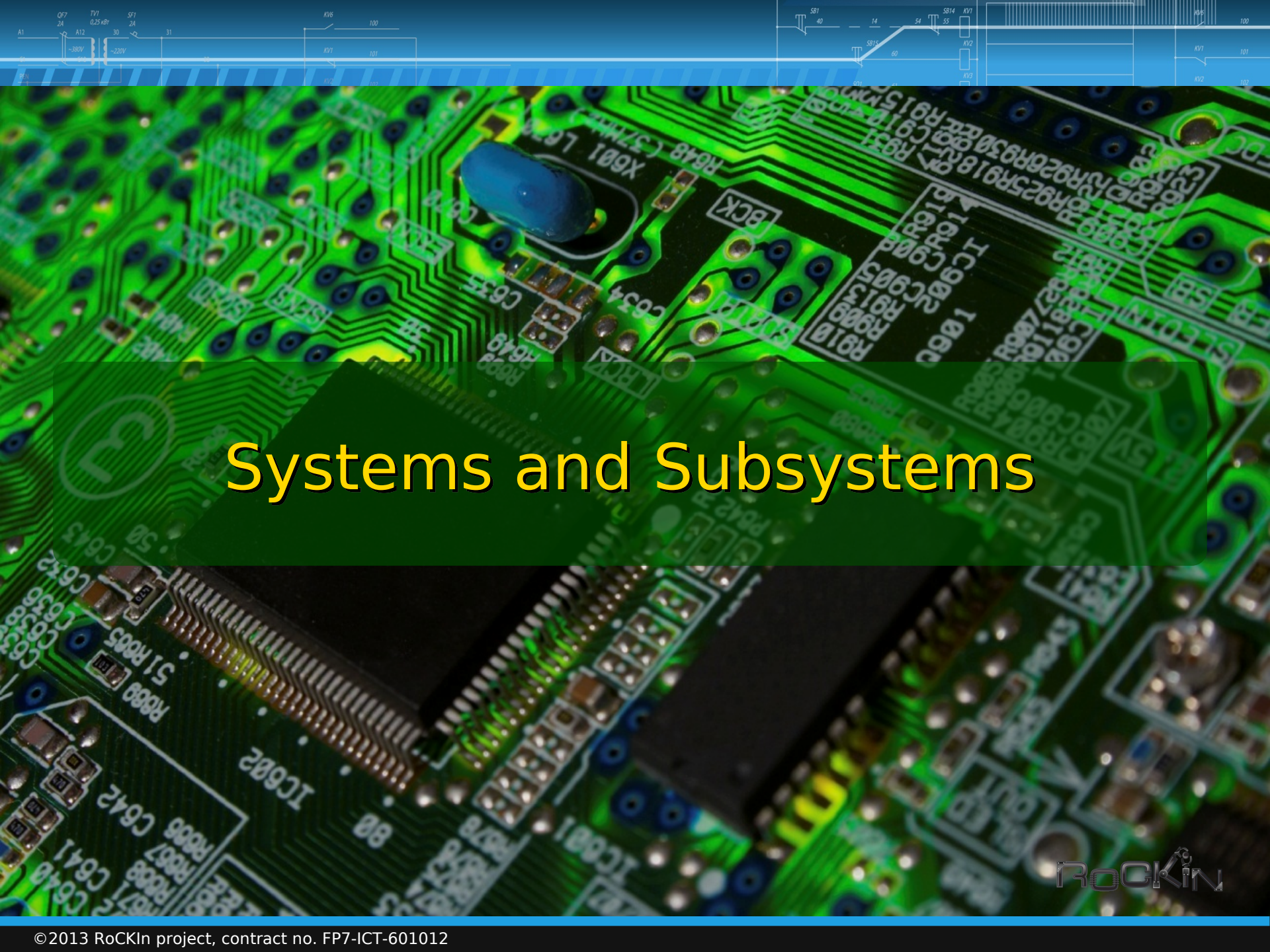
Benchmarking Competitions

Benchmarking: the process of evaluating the performance of a given Robot System according to a specified Metric. In the context of the RoCKIn project, benchmarking is performed through Benchmarking Competitions, the rules of which are oriented towards benchmarking objectives.

(Benchmarking) Competition: a Scientific Competition where the rules are designed in such a way that the rankings also take the role of measurements of the performance of participants, according to objective criteria.

(Benchmarking) Experiment: the composition of a Task or Subtask that has to be performed by a Robot System; the Testbed where the Robot System performs the test.

Benchmark: the union of one or more Benchmarking Experiments, a set of Metrics according to which the course and the outcome of the experiments will be evaluated.



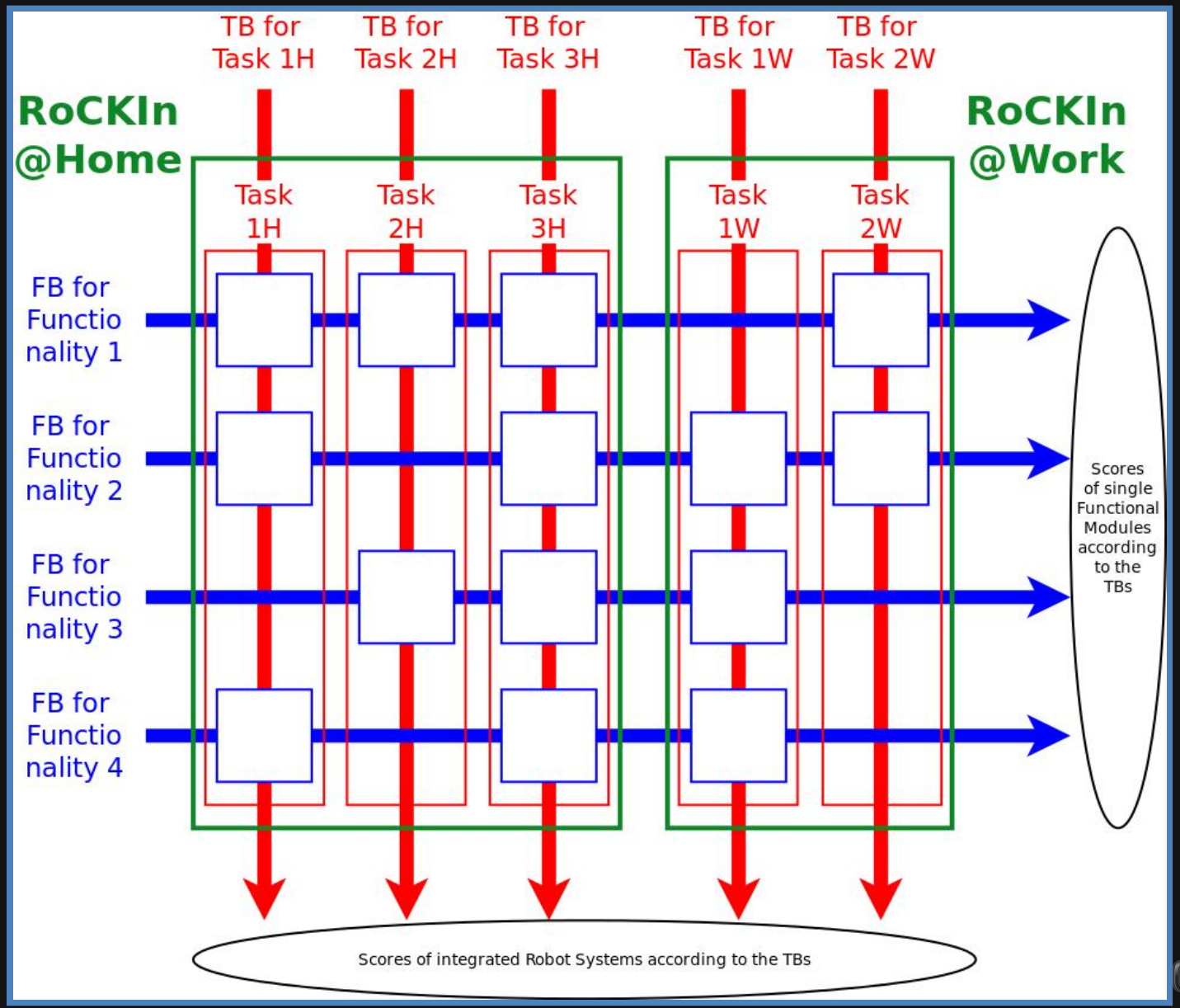
Benchmarking Robot Systems

Benchmarking competitions challenge robot systems at:

- *Task Level:* evaluation of whole systems on a specific task (e.g., the “bring me a beer” tasks)
- *Functionality Level:* evaluation of modules implementing capabilities (e.g., grasping and manipulation, navigation, HRI, etc.)

Benchmarking competitions can/should allow independent evaluation at both levels (Task and Functionality)

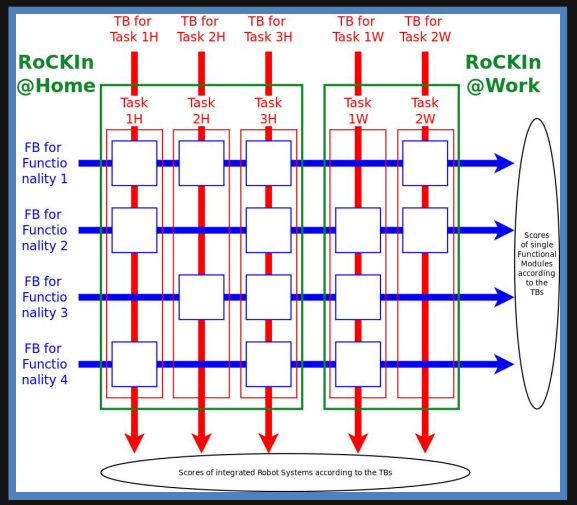
- To encourage participation of people interested in specific aspects of robotics (e.g., object recognition)
- To evaluate to what extent the interplay among modules is relevant (e.g., the precision in positioning before grasping)



RoCKIn “Hypothetical” Example (1/2)

Functionalities 1 to 4 (out of many):

1. Autonomous navigation
2. Object recognition
3. Grasping and manipulation
4. Processing of voice commands



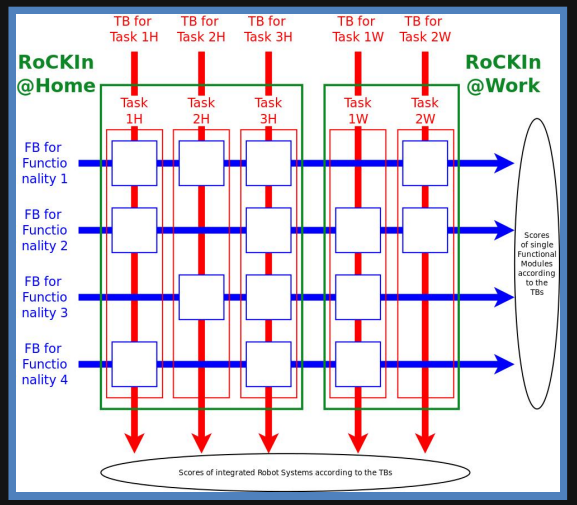
Task Benchmark 2H: “The Robot System is provided with a map of the environment. It must enter the Testbed, navigate through it to reach an object located in a predefined position, and pick it up”

Task Benchmark 1W: “The Robot System is located in a specified pose in front of a table. Over the table N mugs, which differ only in their color, are located randomly (according to suitable specifications). The Robot System must receive a voice command from a human, specifying the color of the mug to pick up, then pick up the required mug”

RoCKIn “Hypothetical” Example (2/2)

Functionalities 1 to 4 (out of many):

1. Autonomous navigation
2. Object recognition
3. Grasping and manipulation
4. 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 firmly 10 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”

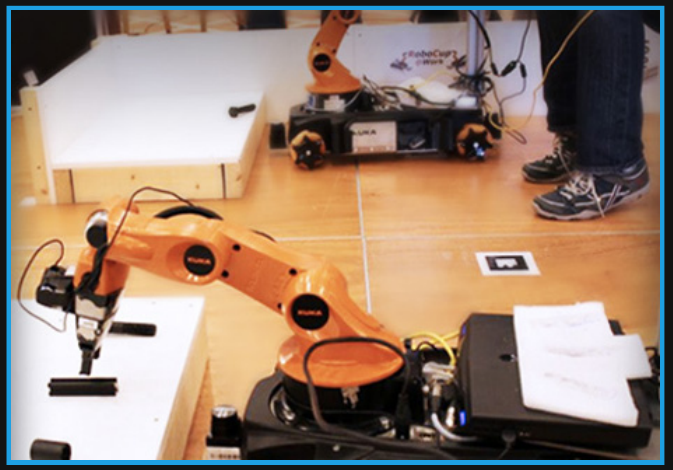
Benchmarking Modules & Systems

- **Task-level benchmarks** should evaluate whole-system functionality over a limited set of situations/tasks, taking into account all system modules as well as their interaction.

- **Functionality-level benchmarks** should be able to investigate the performance of a specific module in a deeper and more general way with respect to task-level benchmarks. To achieve this, they should be aimed at testing (only) one functionality under a range of different conditions, within the chosen scenario(s).

By jointly comparing their results, we acquire information about higher-level properties of the system, such as *quality of system integration or interaction issues among modules*.

RoboCup@Work Example (1)



Similar in spirit to RoCKIn

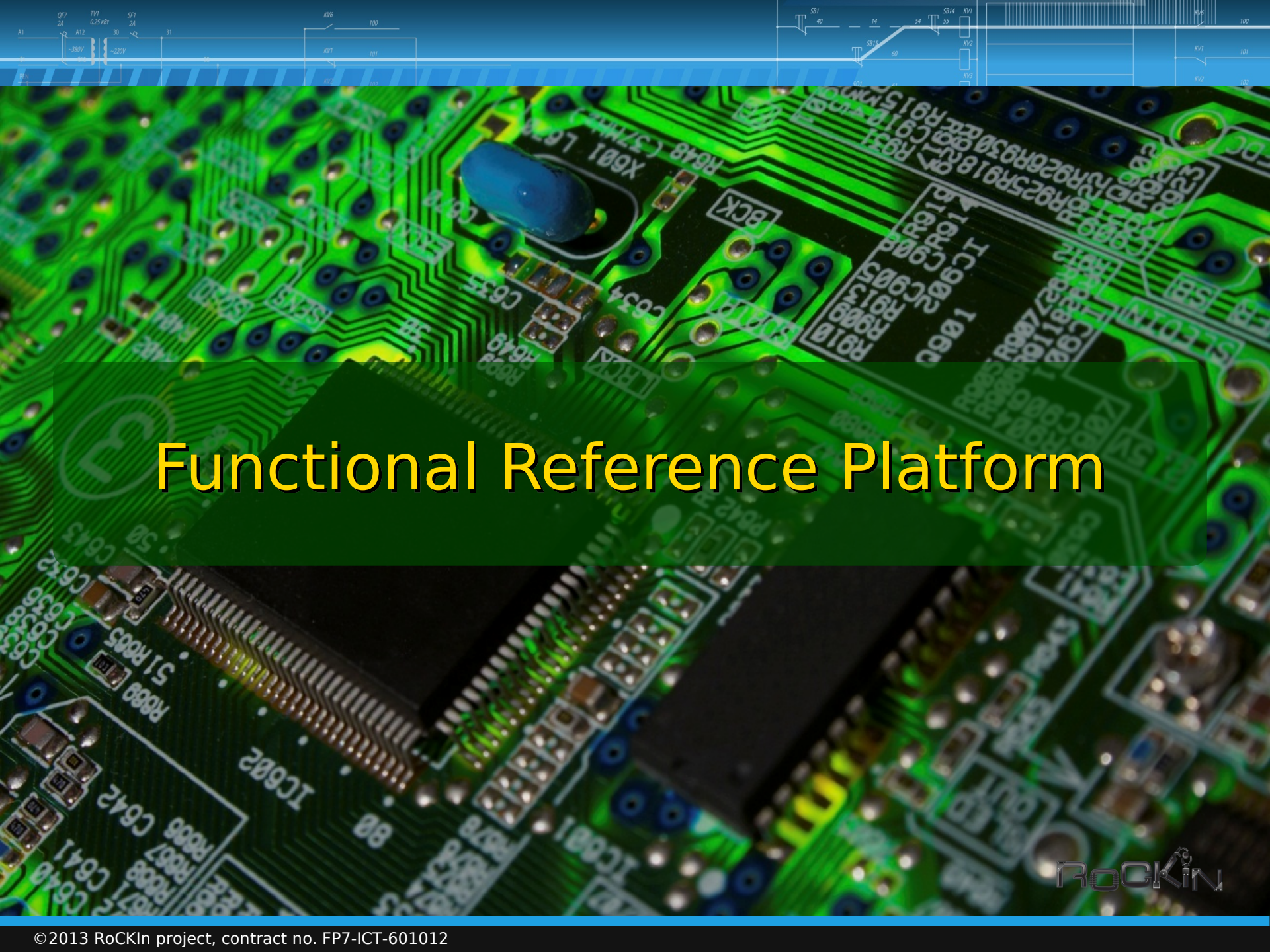
- Module-level and task-level tests
- Complete and precise specifications
- Scenario similar to RoCKIn@Work

RoboCup@Work Tests

- The *Basic Navigation Test* (BNT) is a Functional Benchmark to measure robot navigation capabilities
- The *Basic Manipulation Test* (BMT) is a Functional Benchmark to measure robot manipulation capabilities
- The *Basic Transportation Test* (BTT) is a Task Benchmark that requires both navigation and manipulation capabilities

RoboCup@Work Example (2)

- *BNT:* the robot will be sent a string containing a sequence of places to visit. The robot must enter the arena through a specific gate, move to the places specified in the string, in the order as specified by the string, orient itself according to the orientation given, pause its movement for the time in seconds as specified by the pause length, and finally leave the arena through the gate. Possible obstacle locations are provided.
- *BMT:* the robot has to execute a sequence of n grasp and place operations, possibly with base movement in between, which will, however, be short. The objective is to move a set of objects from one service area into another. The set of objects used in the test is known a priori. The placement of the objects in the origin is not known. A geometric constellation of the objects in the target area will be specified.
- *BTT:* the robot has to get several objects from source service areas and to deliver them to the destination service areas. The task specification consists of two lists: the first contains for each service area a list of manipulation object descriptions, the second contains for each destination area a configuration of manipulation objects the robot is supposed to achieve. Both lists comply with the formats defined by the BMT.



Functional Reference Platform

Motivations for a (RoCKIn) FRP

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

Required by Benchmarks?

- *Task planning functionality* is needed if the task is specified in a way, where the robot must itself decompose it into simpler activities, if it needs to determine which activities it must perform in order to achieve the goal of the task, or needs to determine the order in which these activities need to be performed
- *Path planning functionality* is needed when a task requires a mobile robot to move between different places in the environment and a path between these places is not known a priori, or may need to be modified due to the occurrence of obstacles
- *Grasp planning functionality* may be required if the task requires the robot to grasp and manipulate objects, and grasping cannot be achieved with e.g. purely haptic feedback
- *Visual object recognition* is required e.g. if the task requires the robot to fetch a particular object, which must be recognised in a scene where several objects appear at the same time in the image produced by the robot's camera.
- ...

What data for Benchmarking?

- Localization
  - Goal: determining the pose of a robot with respect to a map representing the external physical environment
  - Input: sensor data (e.g., laser scan and odometry), current map (possibly from mapping module)
  - Output: estimated pose of the robot in the map

- Path planning
  - Goal: determining a sequence of poses to safely move a robot from its current pose to a destination pose set by the user (safely = avoiding known obstacles)
  - Input: current pose (from localization module), current map (from mapping module), destination pose
  - Output: sequence of poses

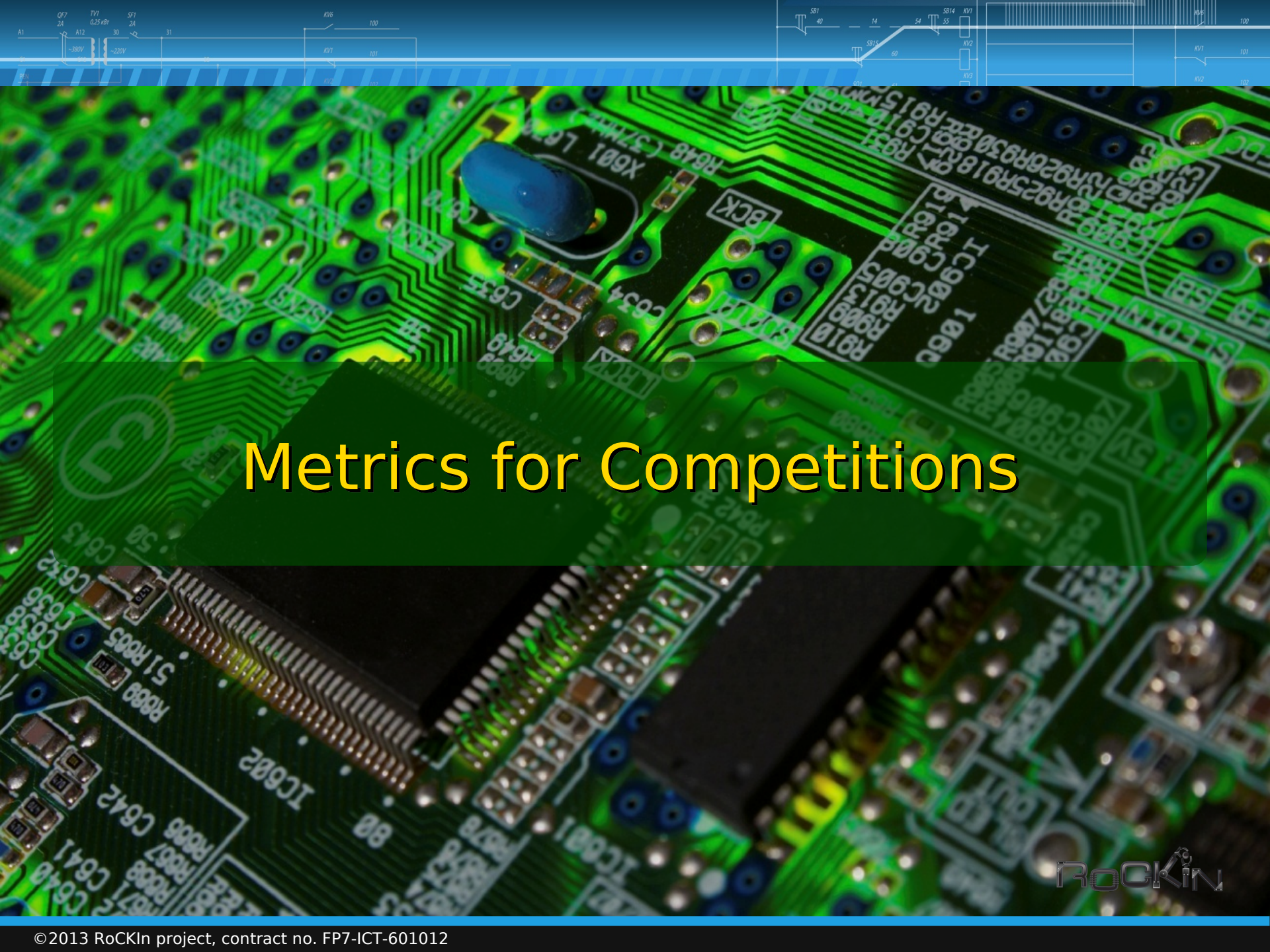
What data for Benchmarking?

- Mapping
  - Goal: constructing a map that describes the environment surrounding a robot (e.g., representing obstacles)
  - Input: sensor data (e.g., laser scan), current map, current pose (from localization module)
  - Output: updated map

Once you have the output what do you do with that?

- Estimated pose of the robot in the map
- Sequence of poses
- Updated map

What kind of output for Human Robot Interaction?
What is a suitable metric for it?



Mindsetting Examples (1/3)

Direct Match Competitions

- Tennis (ATP ranking for males; WTA ranking, for females): each competition (tournament) gives some points to the winner, to the runner-up, to the semi-finalists, and so on. The score of a player is obtained by summing up the points the player got in the tournaments played during the last year (no time discount, but expiration of points).
- Chess (ELO rating system): for some aspects similar to tennis, but more mathematically complex, the important issue is that the ranking of the adversarial is counted.
- Soccer (any European major league): fixed number of games, 3-1-0 points per game (win-draw-lost).
- Rugby (new rules Six Nations tournament): as soccer plus bonus for special situations (e.g., big difference in score).

Mindsetting Examples (2/3)

Subjective Score Based Competitions

- Figure Skating (ISU Judging System): points are awarded for each skating element; the sum of these points is the total element score (TES). Each element is also judged (12) for quality and execution (GOE). A randomized procedure selects 9 judges, then discarding the high and low value, and finally averaging the remaining 7.
- Diving: score considers three elements of the dive: approach, flight, and entry. Panels of five judges are assembled; highest and lowest scores are discarded and the middle three are summed and multiplied by the degree of difficulty.
- Snowboard: each run is scored on a scale of 0.1 to 10.0 by a panel of five judges. One judge scores the standardized moves, another scores amplitude (the height of maneuvers), one scores quality of rotations, and two score overall impression. Penalties are given, too.

Mindsetting Examples (3/3)

Objective Score Based Competitions

- Alpine Ski: competitors attempt to achieve the best time in four disciplines (slalom, giant slalom, Super G, and downhill). For every race points are awarded to the top 30 finishers; the racer with the most points at the end of the season in mid-March wins the Cup. Sub-prizes are also awarded in each individual race discipline
- Biathlon: ski as fast as possible, then hit a target the size of a half-dollar 50 meters away from a prone position and one the size of a coffee cup saucer from a standing position. For every missed target, ski a 150-meter penalty loop
- Show Jumping (horse): a complex combination of two types of penalties: jumping penalties and time penalties. In the past, 1/4 s. penalty for each second or fraction of a second over the time allowed. Since the early 2000s, changed into a different timing.

Issues in Competition Evaluation

Remember the list of competitions? Are they benchmarks?

- Competitions are usually aimed at defining a ranking (at a given moment) not benchmarks
- Scores may change over time to correct/adapt relevance with respect to competition intent, plus they are
  - Direct Match -> not repeatable, opponent dependent
  - Subjective Scoring -> not repeatable, biased judges
  - Objective Scoring -> hard to find relevant scores
- Objective scores (and clear defined testbeds) are the key aspects, but sometimes Subjective scores are unavoidable ...

RoCKIn@Home Practice

Exercise in Competition Analysis

Analyze one RoCKIn@Home competition starting from the following scenario:

- Identify all functionalities implicitly required by the task
- Decide which of those functionalities you are interested in benchmarking with this task and **why**
- Identify the kind of data you need to benchmark those functionalities
- Now you have the data … what do you do with them?

Bridge Round and Tea Party

Despite her age, Granny Annie is still diligently maintaining her social contacts. Twice a week, she hosts a couple of her friends for a round of bridge, which usually ends in a tea party. The robot is supposed to set the table for the bridge round.

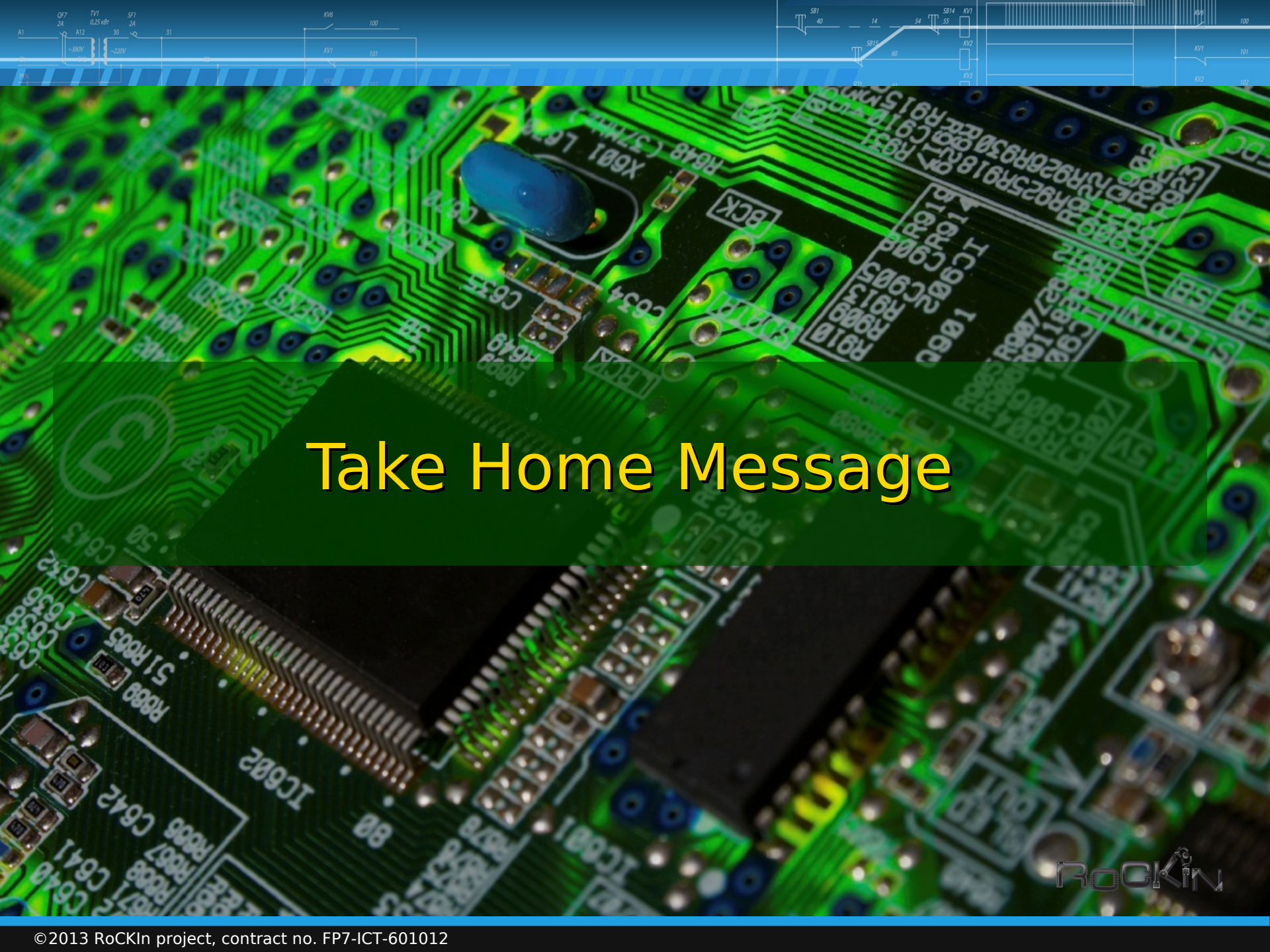
When the guests arrive, the robot needs to welcome them at the door, ask whether it can receive and store any items (umbrella, hat, hand bag), and guide them to the bridge table.

The robot serves drinks for the guests during the bridge round. When a glass or cup is emptied, the robot will ask if it should serve more.

Finally, just before tea time, the robot will set the table for the tea party, with freshly brewed tea, and serve pastries and light sandwiches it has ordered before from a delivery service.

The robot will prepare the items for the tea party on a cart, which it can push or pull from the kitchen towards the dining table.

Can't you borrow/reuse anything from RoCKIn@Work?



Last Before You Go!

Key points:

- The idea of competition/benchmark/experiment: limits and perspectives
- Benchmarking Competitions and what we actually want to benchmark
- The System vs. Subsystem evaluation and the FRP
- Difficulties in metrics definition (e.g., subjective metrics)

Caveats:

- Benchmarking is still an ongoing effort in robotics, so take all I have said as subject to change (in a few hours)
- The real difficulty in robotics benchmarking is that no one wants to do it

Benchmarking Robotics

Matteo Matteucci
Politecnico di Milano