Planes, Trains, Apples, and Oranges

Reproducible Results and Fair Comparisons in Localization Research

Pat Pannuto

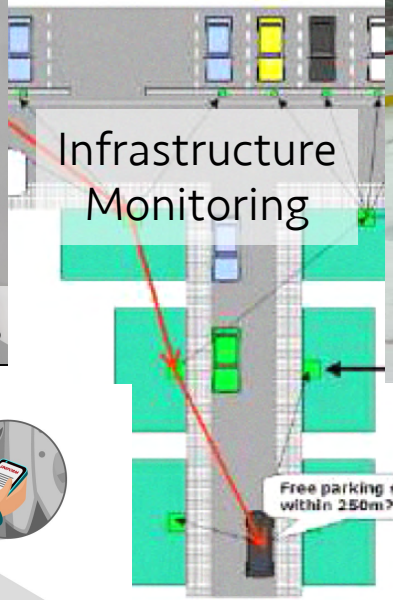
UC Berkeley

CPS-IoTBench'19

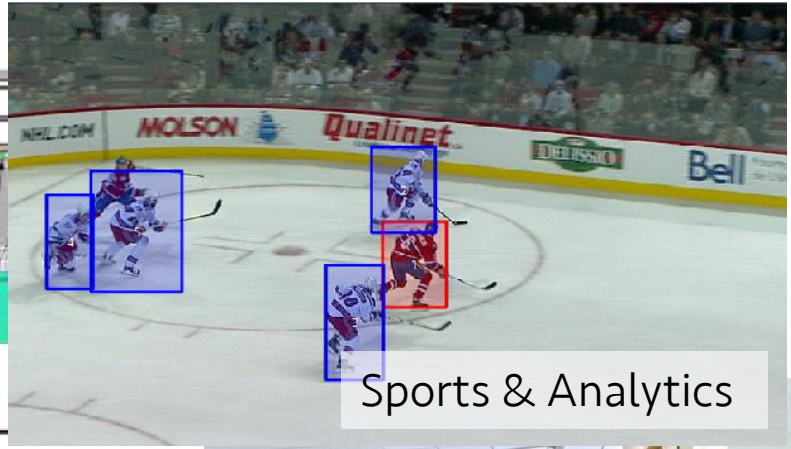
April 15, 2019

Localization empowers current and future technologies

Autonomous Factories, Warehouses



Infrastructure Monitoring



Sports & Analytics

Guided Tours

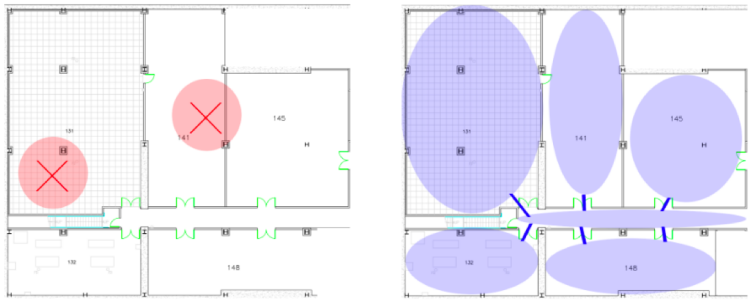
Augmented Reality

But localization means many things to many people

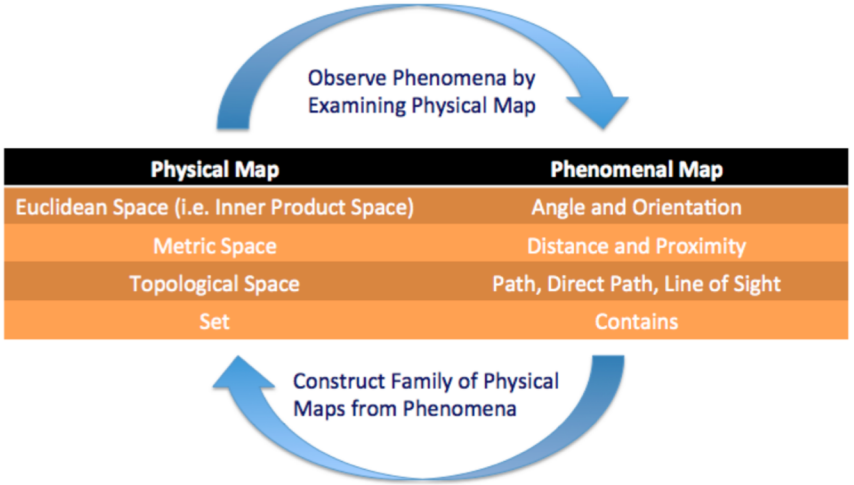
Semantic Localization

Localize by significance rather than absolute position in space.

Geographic **Semantic**



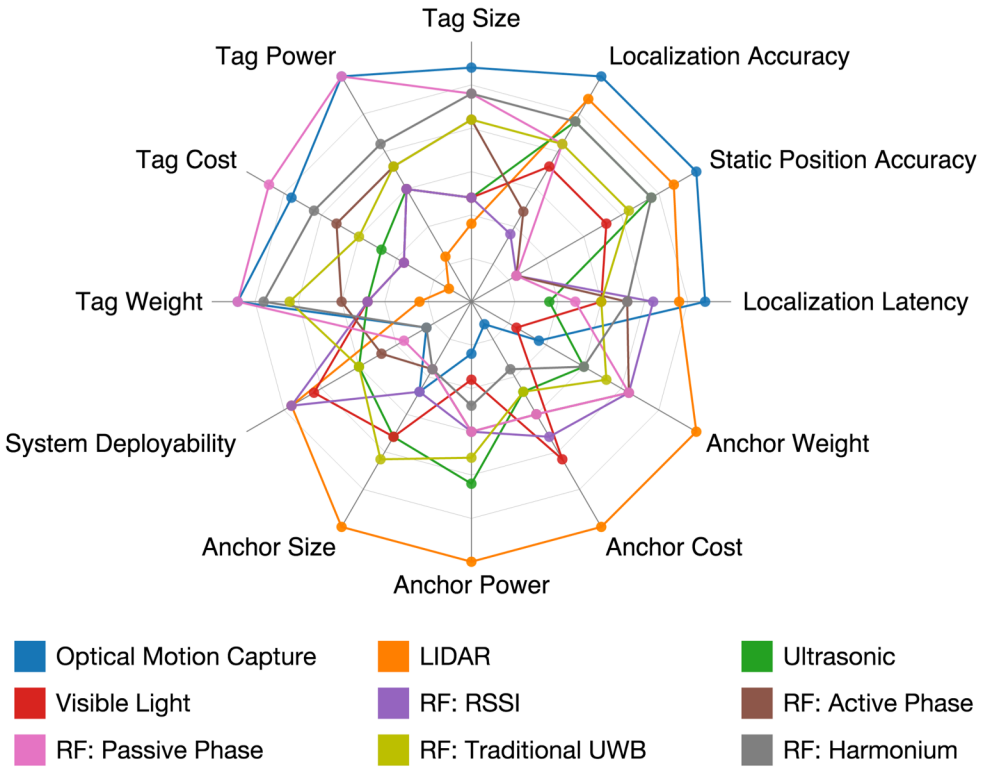
Matthew Weber, Self-Organizing Semantic Localization, TerraSwarm (2013)



Matthew Weber, Edward Lee, A Model for Semantic Localization, IPSN (2015)

Which leads to a wide array of considerations for localization systems

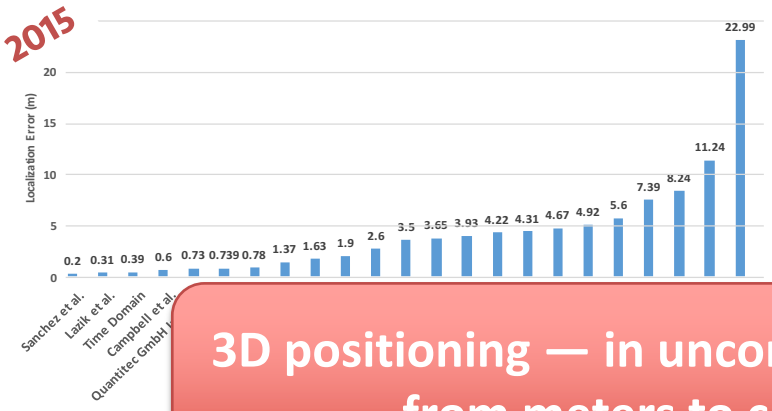
- Here are 12 dimensions
  - Covering 9 technologies
    - Each have several implementations
  - (bigger is better) →
- No one technology will suit all applications
  - What does it mean to localize a person to 1cm?
  - Motion? Through-wall?

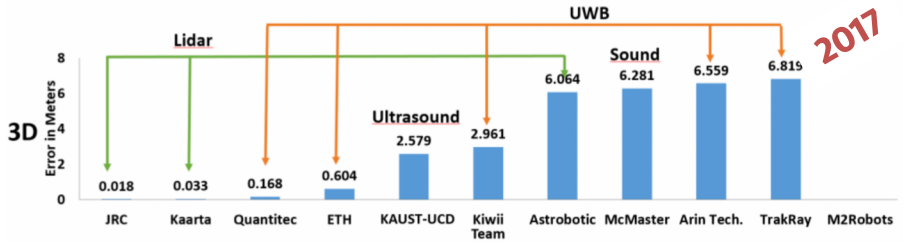


What happens with a shared benchmark?

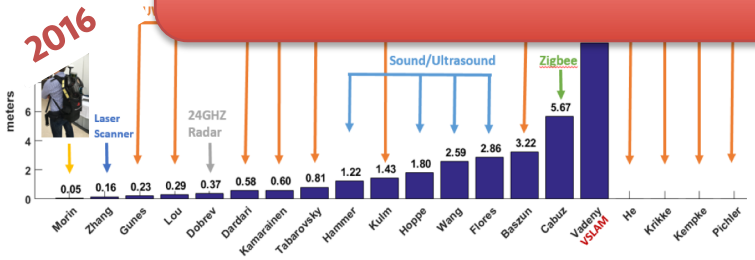
Case Study: The Microsoft Indoor Localization Competition

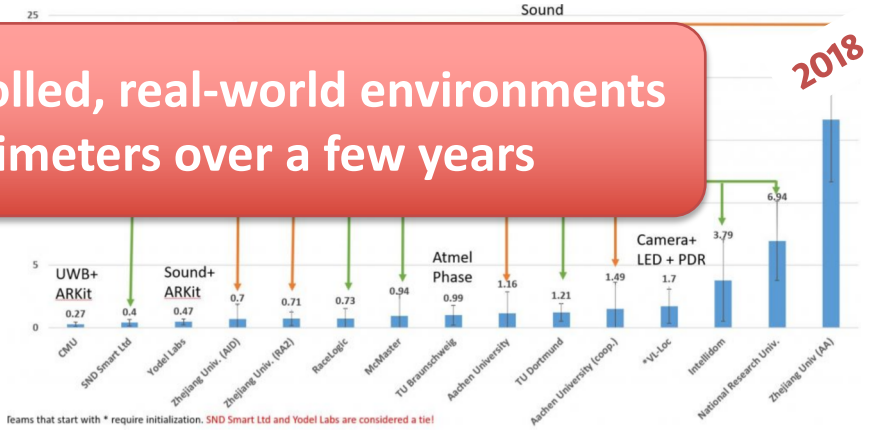
- "solved" accuracy, and a little bit deployability





3D positioning — in uncontrolled, real-world environments — from meters to centimeters over a few years





Competitions for every possible combination do not scale

- Results in fragmentation
  - How do you “fairly” compare systems with different application requirements?

Mautz, Rainer. "Indoor positioning technologies." (2012).

Fragmentation is exacerbated by fragmentation of the community

- Very non-exhaustive list of venues with recent interesting work
  - MobiCom'18: Session 7: Where are U Now? Localization and Motion Tracking
  - NSDI'19: Session: Wireless Applications [1 Localization, 1 Tracking paper]
  - SenSys'18: Session IV: Lost [3 Localization papers]
  - IPSN'19: Session 1: Location tracking
  - SIGCOMM'18: Session 3: Wireless Links [1 Localization paper]
  - IPIN'18 (International Conference on Indoor Positioning and Indoor Navigation)
  - ICL-GNSS'19 (International Conference on Localization and GNSS)
  - [and these are just more systems-focused venues…]

Because the community is fragmented, no meta-analysis of results and norms

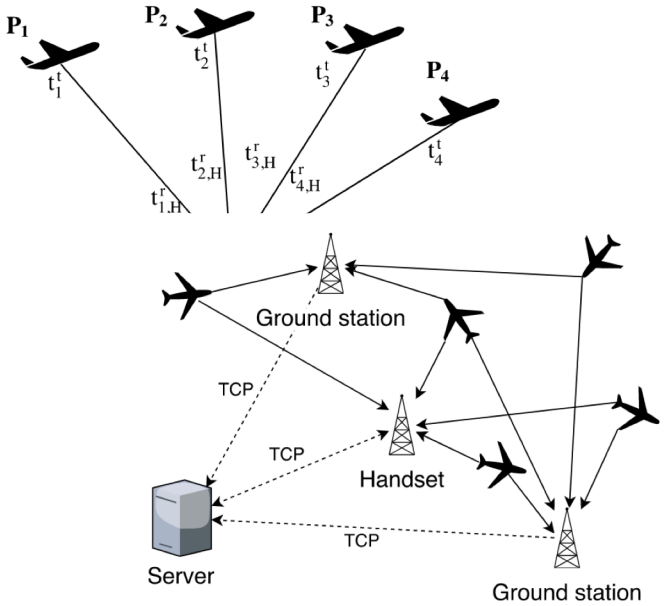
- Though efforts like the Indoor Localization Competition show people are willing to come together for common goals
- Opportunity for CPS-IoTBench?

Outline

- Reproducibility
  - Planes: Scenarios that cannot be (easily) replicated
  - Trains: How to design ground truth
- Comparisons
  - Apples & Oranges: How to quantitatively compare different architectures?

Sometimes things in the physical world only happen just that way, just that once

- Idea: ADS-B signals from planes are plentiful and strong – indoor GPS??



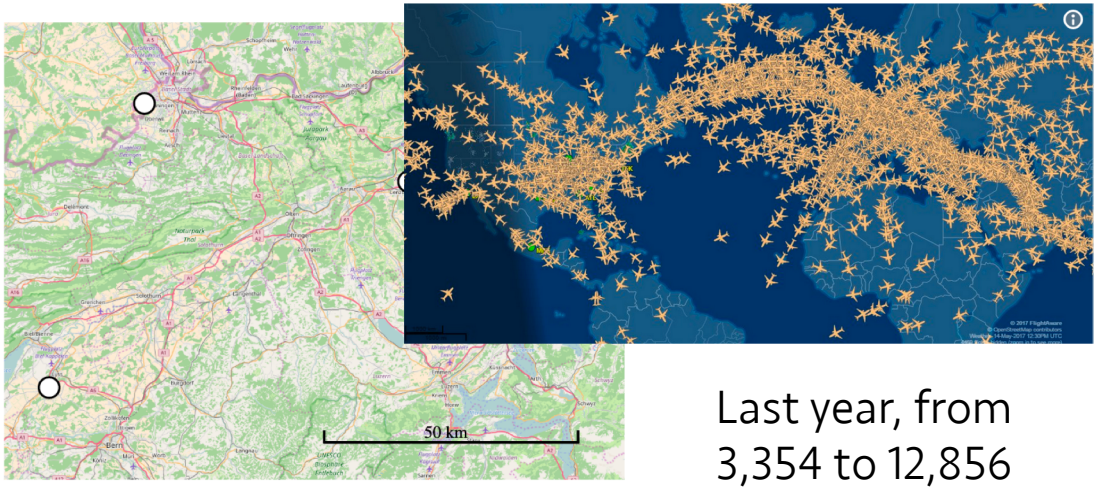


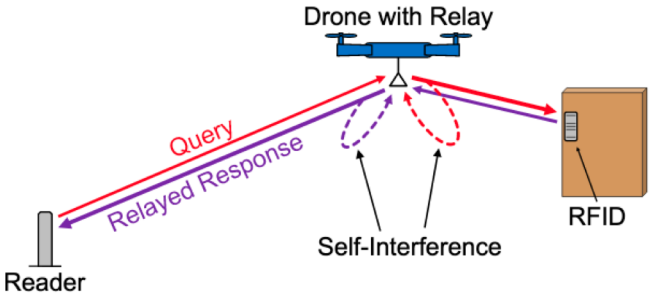
Figure 6: Location of the ground stations (white circles) and a handset (black circle) for the evaluation. The ground stations span over a region approximately 110 km in diameter.

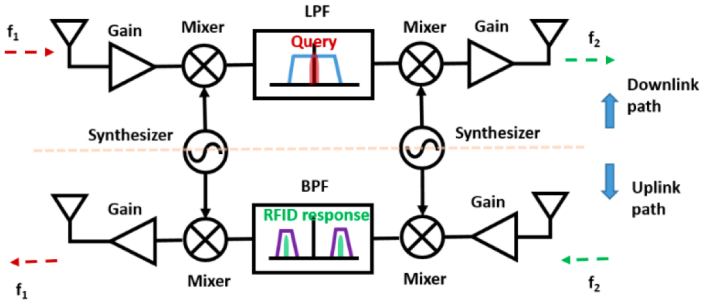
Last year, from 3,354 to 12,856 planes in the sky

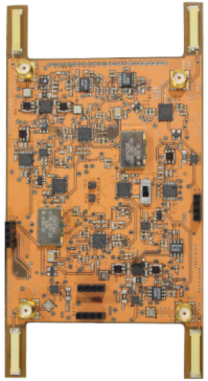
Manuel Eichelberger, Kevin Luchstinger, Simon Tanner, Roger Wattenhofer, Indoor Localization with Aircraft Signals, SenSys 2017

Sometimes interesting ideas require interesting hardware

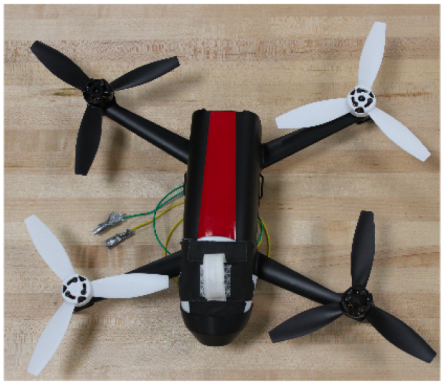
- Idea: Resolve RFID range/coverage with relay on a drone







(a) RFly's relay circuit

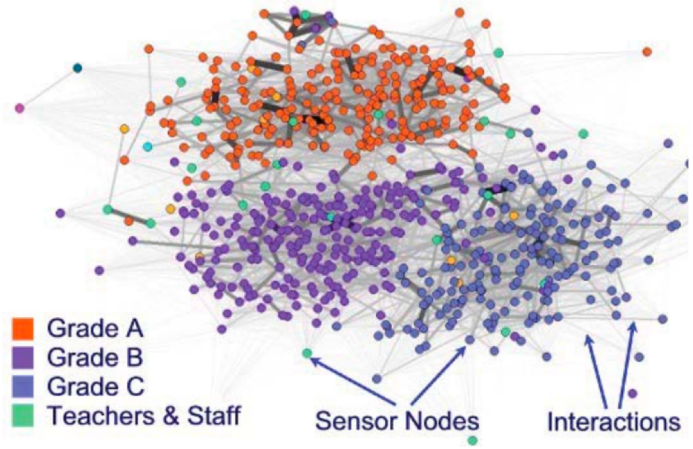


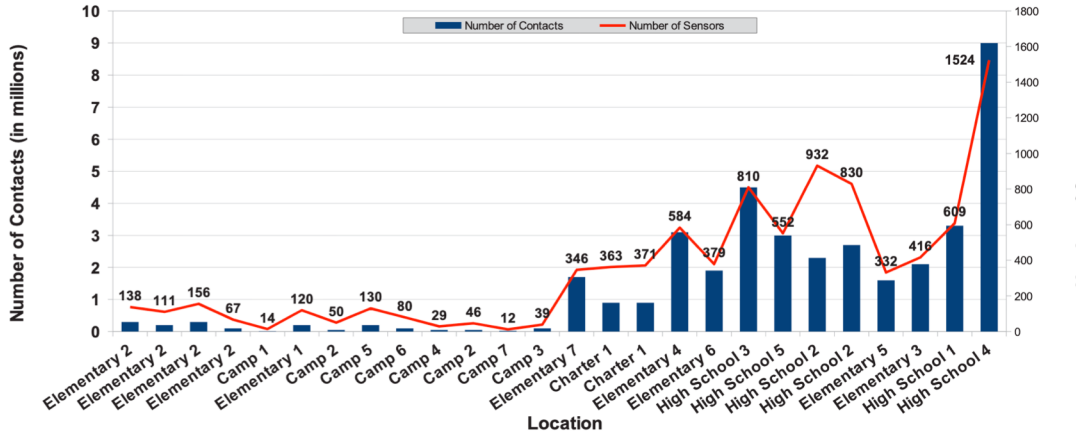
(b) Bebop2 drone

Yunfei Ma, Nicholas Selby, Fadel Adib, Drone Relays for Battery-Free Networks, SIGCOMM'17

Sometimes interesting work comes from data that is not easily shared

- Finding people is interesting, protecting their privacy is hard





Kyeong T. Min, Andrzej Forys, Anh Luong, Enoch Lee, Jon Davies, Thomas Schmid, WRENSys: Large-Scale, Rapid Deployable Mobile Sensing System

Should a *discussion* of reproducibility be required?

- “Future Work” – how and by whom?
- Are irreproducible results intrinsically bad?
- Internally reproducing experiments…
  - Run trials until you get the good looking graph, discard the rest
  - Can we create a reviewer checklist?

Are datasets the answer?

- Shameless plug: DATA workshop at SenSys again this year
  - Expanded scope: "The collection and use of data", what makes datasets useful?
- Dataset release enables post-hoc benchmarking

- Slocalization
  - Have ~36 GB of data traces still around
    - Maybe half of which ended up in the paper…
    - And there are graphs where the data is gone
  - Interesting problems in the data, but unlabeled
    - Non-deterministic reward!

Outline

- Reproducibility
  - Planes: Scenarios that cannot be (easily) replicated
  - Trains: How to design ground truth
- Comparisons
  - Apples & Oranges: How to quantitatively compare different architectures?

Ground truth is the localization chicken and egg problem

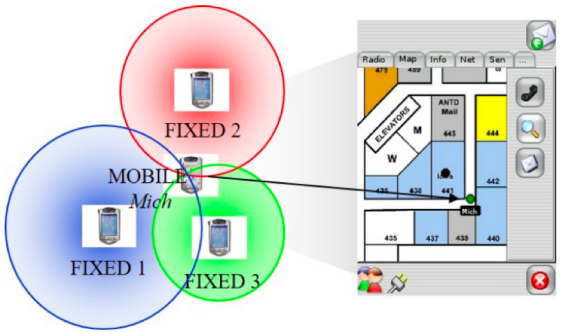
- If you build a better localization system than anything that has come before, how do you evaluate it?

For a metrology problem, let's look to the metrology experts!

- New standard dataset for "infrastructure-free" systems

NIST

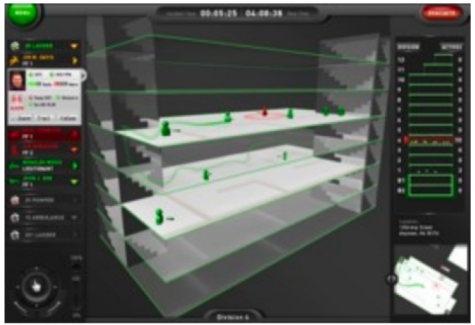
PerfLoc: Performance Evaluation of Smartphone Indoor Localization Apps



For a metrology problem, let's look to the metrology experts!

- ISO/IEC 18306: 2016
  - identifies appropriate performance metrics and test & evaluation scenarios for localization and tracking systems, and it provides guidance on how best to present and visualize the T&E results.

Testing of Indoor Localization and Tracking Systems (LTSs)



Standardized test environments lag the leading edge of research ideas

- There is a gap between new technique and 1,300+ point measurement
- Good for capstone research, product development

Is a FlockLab of localization plausible?

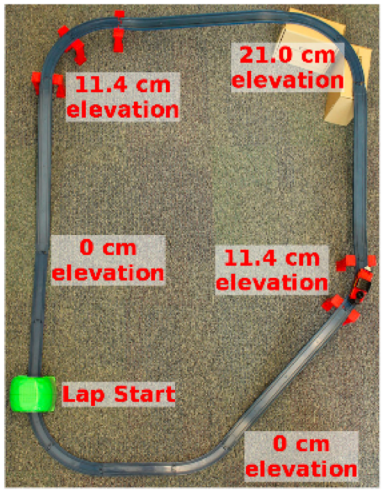
- Something lighter weight than NIST testbed
- Do novel physical layer enhancements make this impossible?
  - Hardware to testbed versus testbed to hardware?
  - Or do SDRs save the day?

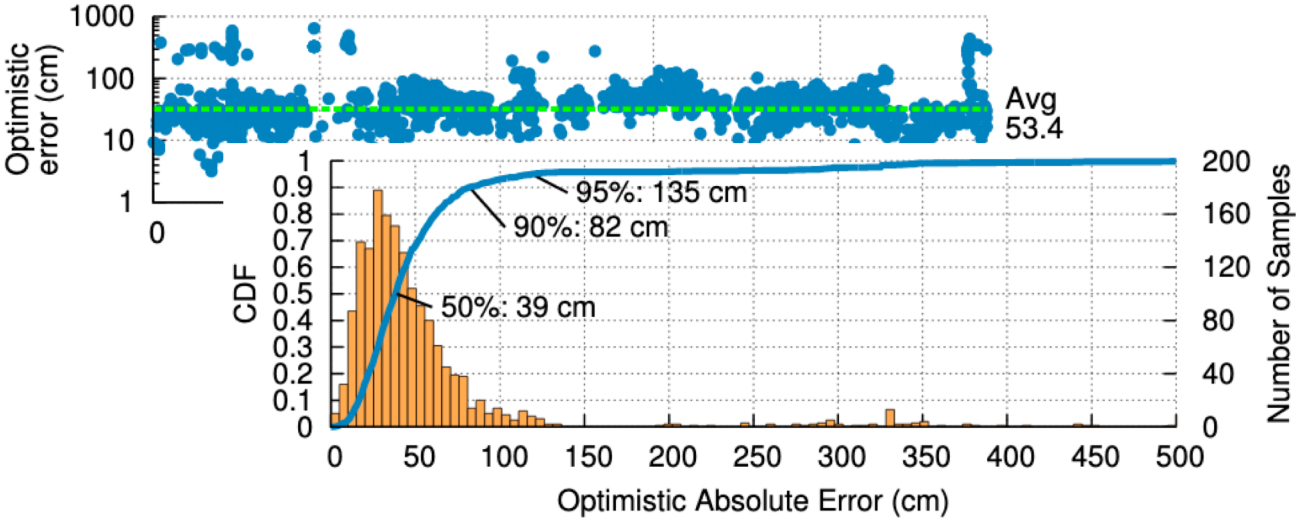
What are we doing today?

- In the absence of a standard testbed…

Capturing accurate ground truth can be challenging

- *As the exact position of the tag in space and time is unknown when each sample is taken, we compute the **optimistic error**, that is the minimum distance from a Harmonia location estimate to the nearest point on the track.*

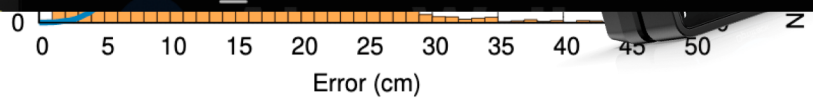




Benjamin Kempke, Pat Pannuto, and Prabal Dutta, Harmonia: Wideband Spreading for Accurate Indoor RF Localization, HotWireless'14

We have solutions, but they can be prohibitively expensive

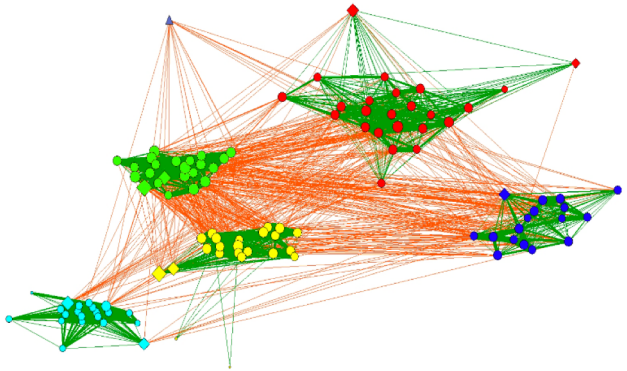
```
TIME_FIX = 3
if EXP == 'out.1':
.    SEC_FIX = 14*60-24.2
if EXP == 'out.2':
.    SEC_FIX = -24*60*60 - 30*60 - 8
if EXP == 'out.3':
.    SEC_FIX = -24*60*60 - 30*60 - 32
if EXP[:5] == 'out.4':
.    SEC_FIX = -24*60*60 - 30*60 - 81.5
if EXP == 'out.5':
.    SEC_FIX = -24*60*60 - 35*60 - -40
if EXP == 'out.6':
.    SEC_FIX = -24*60*60 - 83*60 - -13
if EXP == 'out.7':
.    SEC_FIX = -24*60*60 - 87*60 - -17
```

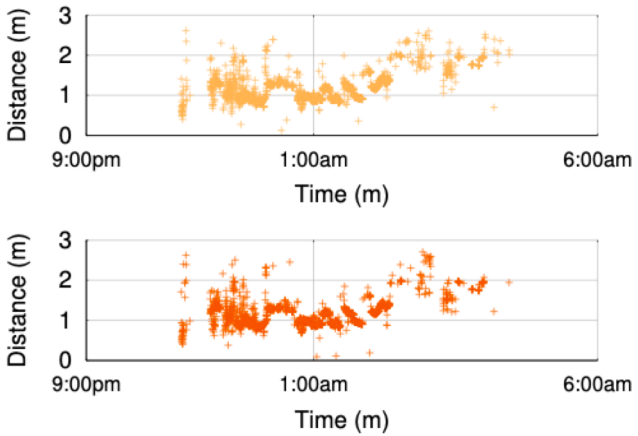


Benjamin Kempke, Pat Pannuto, and Prabal Dutta, Harmonium: Asymmetric, Bandstitched UWB for Fast, Accurate, and Robust Indoor Localization, IPSN'16

We have no idea how to handle evaluations in infrastructure-free scenarios

- What do you do when you cannot instrument evaluation spaces?
  - State-of-the-art: Internal consistency and satisfying intuitions





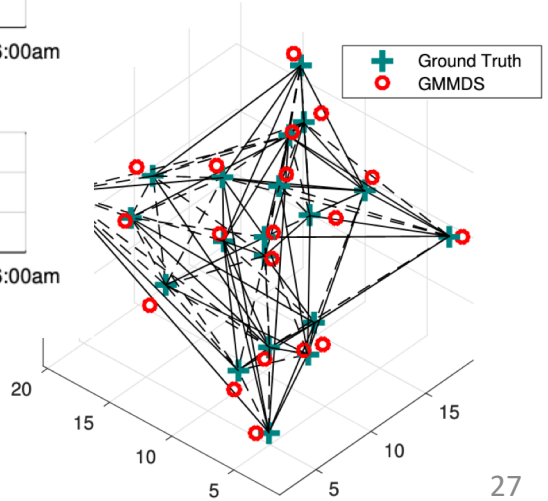


Figure 5: This represents a human contact network. Dots represent individuals wearing a WREN. Lines represent a contact interaction with another node. The different colored dots represent students in different classrooms. This is a snapshot of several classrooms and their student interactions during the day.

Experimental design is being under-reported

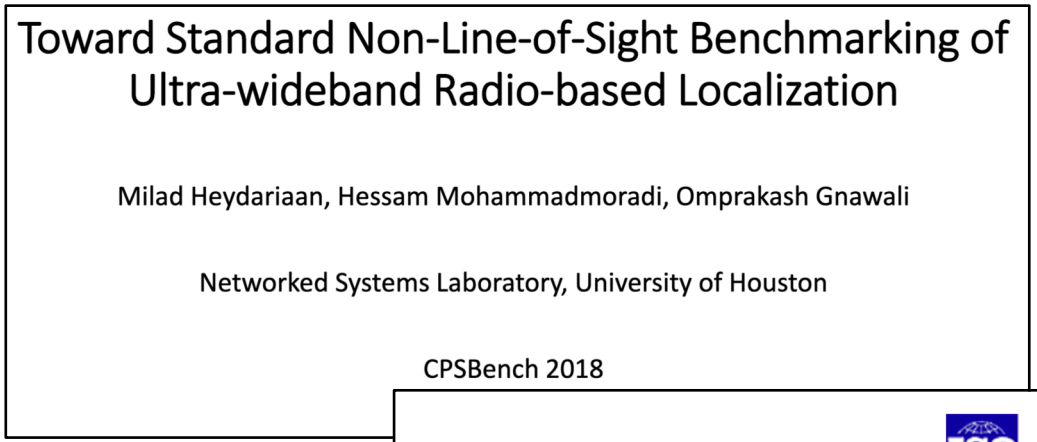
- One of the original motivations for the Indoor Localization Competition
- Reviewers of localization papers should critically analyze ground truth
  - ”We used an expensive good system” is not enough!

Outline

- Reproducibility
  - Planes: Scenarios that cannot be (easily) replicated
  - Trains: How to design ground truth
- Comparisons
  - Apples & Oranges: How to quantitatively compare different architectures?

Step 1: We all need to agree on the underlying language

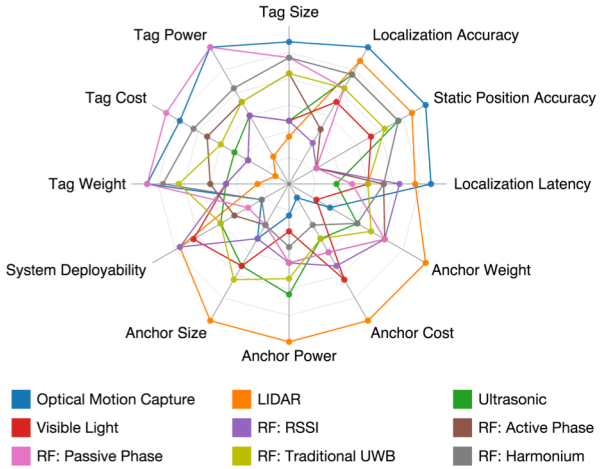
- Bottom-up and top-down efforts here



Related work sections are the art of choosing a different subset of metrics

- No longer just the binary ‘this metric is important’

Mautz, Rainer. "Indoor positioning technol



6 RELATED WORK

6.1 Indoor Localization

Much of the research on indoor localization focuses on providing accurate localization, for instance to room level or even sub-meter accuracy. The cost factors to get so accurate are

(I) the installation of dedicated **I**nfrastructure, like for instance one beacon in each building up to several in each room;
(T) a **T**raining or initialization phase to gather data which is necessary for the subsequent localization;
(E) the usage of **E**xpensive user equipment.

Ultrasound. (I) In contrast to WiFi based localization, which is infrastructure free, ultrasound based methods require dedicated hardware. However, ultrasound systems are relatively inexpensive

Light. (T,E) The most accurate results in the Microsoft Indoor Localization Competition are achieved by laser- and camera-based methods.[13] The best system achieves an accuracy of 5 cm using

Bluetooth. (T,I) Another type of signal used for indoor localization is Bluetooth. Bluetooth is similar to WiFi in that both systems share the 2.4 GHz frequency band. Compared to WiFi, which can

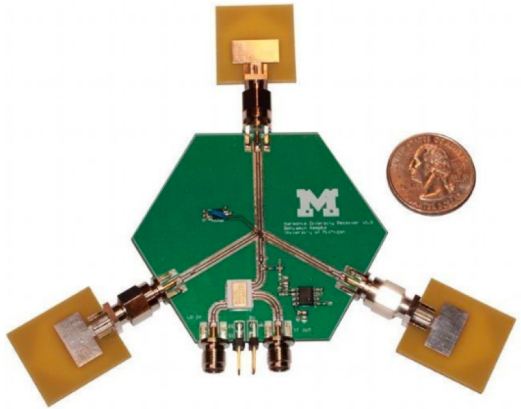
Abstracts demand “one-number” performance

- People “need” a quick handle for comparison
  - Needs to be novel mechanism, and it needs to be better result
  - And it’s in your interest for this number to be better than previous numbers
- Median, 90th, 99th, worst-case?
- Application dependent?
  - Then how do we compare??

Newer innovations becoming increasingly cross-layer

- Systems papers as a disservice?
  - Improve physical layer measurement AND processing layer technique
    - (Particle/Kalman filtering etc)
  - And what's the physical layer anyway?

These techniques, however, rely on non-static environments and measure *changes* in target position but either blindly preserve a static initial offset or retroactively learn true position after several seconds of motion. Furthermore, these systems rely on point-to-point state to feed models that predict viable motion paths to reject outliers and smooth estimates. Such application-specific optimizations are complementary and could also be applied to raw Harmonium estimates to further improve accuracy, but also require that any direct comparisons respect the difference between what is presented.



Datasets are an opportunity to decouple?

- But not everything is always collected
- Slocalization
  - Direct physical channel ~1536 MB/sec baseband [decimated on FPGA]
  - "Raw" IQ data logged ~100 GB
  - Processed to recover CIRs ~100 MB
  - Processed to recover locations ~100 kB
- Bigger data is often harder to use, generalize, and share

Growing the scope of CPSBench?
Venue for (ir)reproducible results?

- Great intro-to-a-new-area type of work
  - Community should value and provide a venue to publish reproduction efforts
- Particularly interesting to "reproduce" in new physical spaces
  - Or otherwise challenge understated assumptions
- Many other communities have or are growing similar things
  - ISCA + Workshop on Deduplicating, Deconstructing, and Debunking (14 years!)
  - ICLR + Workshop on Reproducibility in Machine Learning (3 years)
  - IEEE RAM + Short replication articles (r-articles) (2 years)

Conclusion / Looking toward the next session…
What role does CPS-IoTBench have moving forward?

- Venue has the potential to be authority for CPS evaluations
  - Need to allow new ideas to have imperfect evaluations! (Within reason….)
    - Research != product, “Prove it’s possible”
  - Let science take its course and develop corrections
- Service to the community
  - Validation of prior work
  - Resolution of evaluation metrics for new physical-world ideas

patpannuto.com @patpannuto