Big data, data collection, aggregation, analysis and visualization

Lasse Berntzen

Buskerud and Vestfold University College

NOTE

- The original presentation included movie clips that are not accessible in this .pdf file.

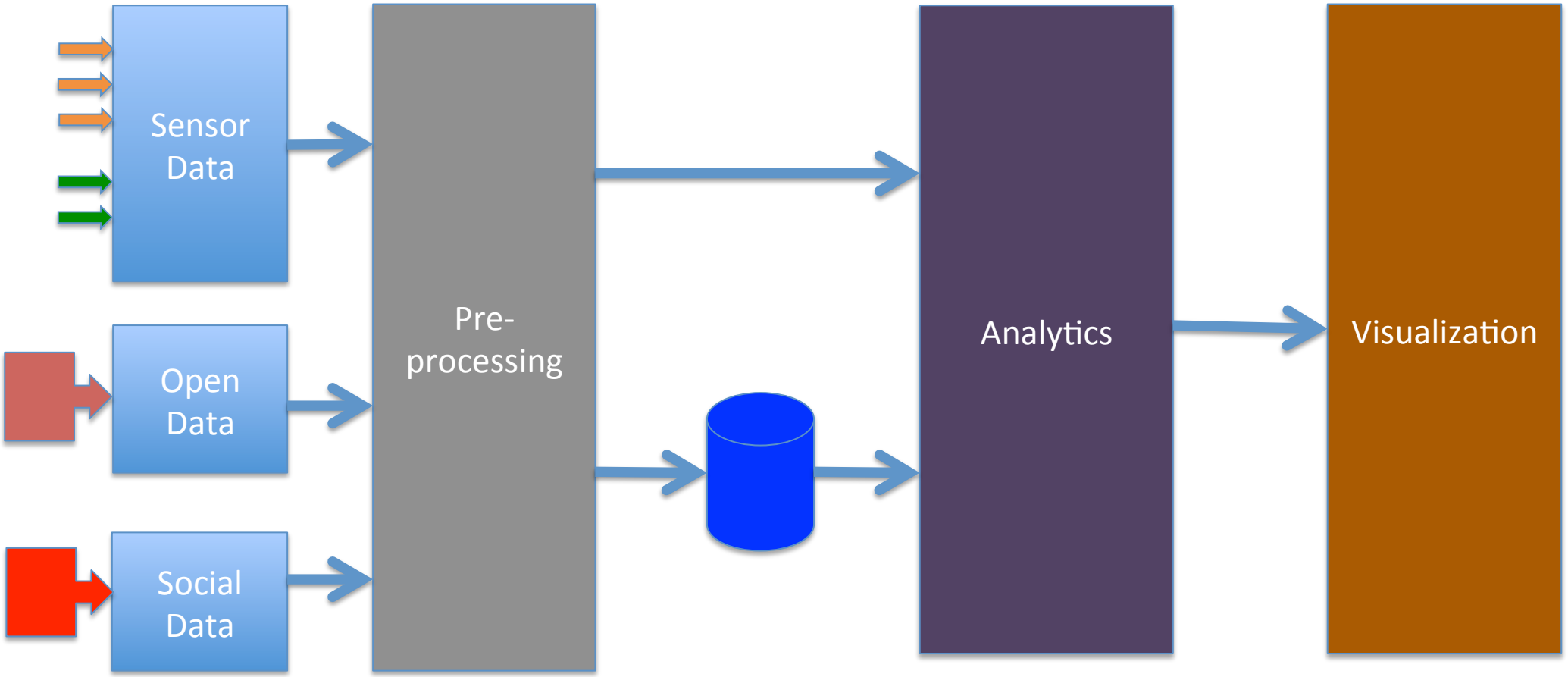
About myself

- Professor at Buskerud and Vestfold University College, Norway, Dept. of Business and Management
- Research group: Innovation & Management
- Since 2004, research on:
  - e-democracy, e-participation, e-government
  - Public sector innovation
  - “Smart cities”
- Teaching:
  - .NET
  - Mobile solutions
  - e-Commerce
  - e-government and e-participation

Outline

- Experience based
- A model to show the context
- The eGovMon project
- Mining social data
- Sensors as a source for big data
- Using sensor data

A conceptual model



”BIG DATA” SOURCES

“Big Data” Sources

- Social Data
  - From WEB 2.0 applications like social networks (Facebook, LinkedIn, Twitter, Blogs, Flickr).
- Open Data
  - Repositories of data sets accessible for everyone.
  - Open government initiatives
- Sensor data
  - Can produce endless streams of (real-time) data

Data may be a lot of things

- Structured data
  - Tables, comma separated files, XML
- Unstructured data
  - Social media content

Structured Data

- Structured data is easy to handle.
  - Easy to parse
  - Easy to put into database tables
  - Easy to analyze

Unstructured Data

- More difficult to process for various reasons
  - Syntax is not semantics
  - Need to understand the context
    - A sentence may have a complete different meaning in another context
  - Computers are not good at understanding complex language
    - Example: Social media and use of irony
  - The more precise, the better
    - Example: Newspaper content vs. social media
  - But we can always extract useful information
    - Compare to traffic analysis

Case: Citizens Insight System

- In 2008, I worked with colleagues from University of Athens and Westminster Business School to extract information from a reporting system implemented by the municipality of Polichni in Greece.
- The reporting system allowed citizens to report problems related to municipal services.
- Our team focused on extracting information about reports related to road deficiencies and garbage collection

Research Problem

- The reporting system became very popular.
- The municipality was overwhelmed by the response from citizens, and soon understood that they needed help to extract relevant issues.
- This was the rationale for the research. How could we help create a list of most important issues to be dealt with?

Components

- Tokenizer - divides the text into tokens i.e., words, sentences, paragraphs etc
- Linguistic Parser – adds linguistic information about the text under examination
- Database Lookup – collection of algorithms that add - related to the text - information that is stored in a database i.e., synonyms, hyponyms, hypernyms, standard expressions, locations, names etc.

Ontology and analyzis

- We created an ontology to see how terms fit together.
- We used state-of-the art artificial intelligence techniques to analyze the reports.

Thomas Mavroudakis, Panos Hahamis, Lasse Berntzen, Nodas Douzinas, Markos Hanna, Nikos Farfas, Niki Krikeli and Charalampos Karanikas (2009) An Exploration of the Citizens Insight System: A Case Study. Proceedings of the 9th European Conference of e-Government (ECEG) held in London, Academic Conferences Ltd.

Results

- Citizens submitted reports in natural language.
- But such reports are pretty informal in its nature.
- We set up three performance measures:
  - Precision: the fraction of documents retrieved that is relevant to the user's information need.
  - Recall: the fraction of documents that is relevant to the query that is successfully retrieved.
  - Fall-out: the proportion of non-relevant documents retrieved, out of all non-relevant documents available.

Results

- 1000 reports analyzed.
  - 300 were related to road deficiencies
    - System returned 310 reports, of which 198 was true
    - Precision: 64%, Recall: 66%, Fallout: 11%
  - 400 were related to waste collection
    - System returned 340 reports, of which 211 was true
    - Precision: 62%, Recall: 53%, Fallout: 13%

Conclusion

- Did we succeed?
- Not easy to analyze such reports, but we managed to get a list of issues to be investigated by the municipality.
- The case shows the problem of analyzing informal reports.
- The results would probably be much better for newspaper articles.

Informal Language

- Example: “The Norwegian Railway System is just great”
- Comment from customer as response to:
  - The railway system was halted for a day due to problem with the signal system.
- Irony is not easy to detect.

Analyzing Social Media

- It is much more than content.
- There is so much metadata out there.
  - Traffic analysis.
  - Relations. Friends and followers. Influence.

Analyzing Social Media

- Understanding sematics is hard, but we can do other things with content.
  - Look for keywords (something we are interested in)
  - Extract all words and measure how many times each word are used. (We can then find emerging topics)
  - Look for addresses (geolocations). Student project: Get addresses from documents, run them through geolocation service (Google), and make links on a map.

Mining Social Media

- Twitter and Facebook have API´s.
- I have used the Facebook API to extract information on Facebook presence of municipalities.
- Two problems encountered:
  - API changes happens (and your application stops working)
  - Limited number of requests

For a detailed discussion on all aspects of mining social media, I recommend the following book:: Russel, Matthew A. Mining the Social Web. O´Reilly, 2011

THE SEMANTIC WEB
(AND APPLIED SEMANTIC NETWORKS)

Semantic Web

- World Wide Web – mostly documents for human beings (HTML).
- Sematic Web is about data that can be manipulated automatically.
- Semantic Web is an enhancement to the World Wide Web. It makes the World Wide Web more useful.

Semantic Web

- Tim Berners Lee (2001)
- Computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically.
- The resulting infrastructure will spur the development of automated Web services such as highly functional agents.
- Ordinary users will compose Semantic Web pages and add new definitions and rules using off-the-shelf software that will assist with semantic markup.

Semantic Web in Daily Use

- Tagging systems
- E.g. Digg, del.icio.us
- Flickr, YouTube
- Twitter

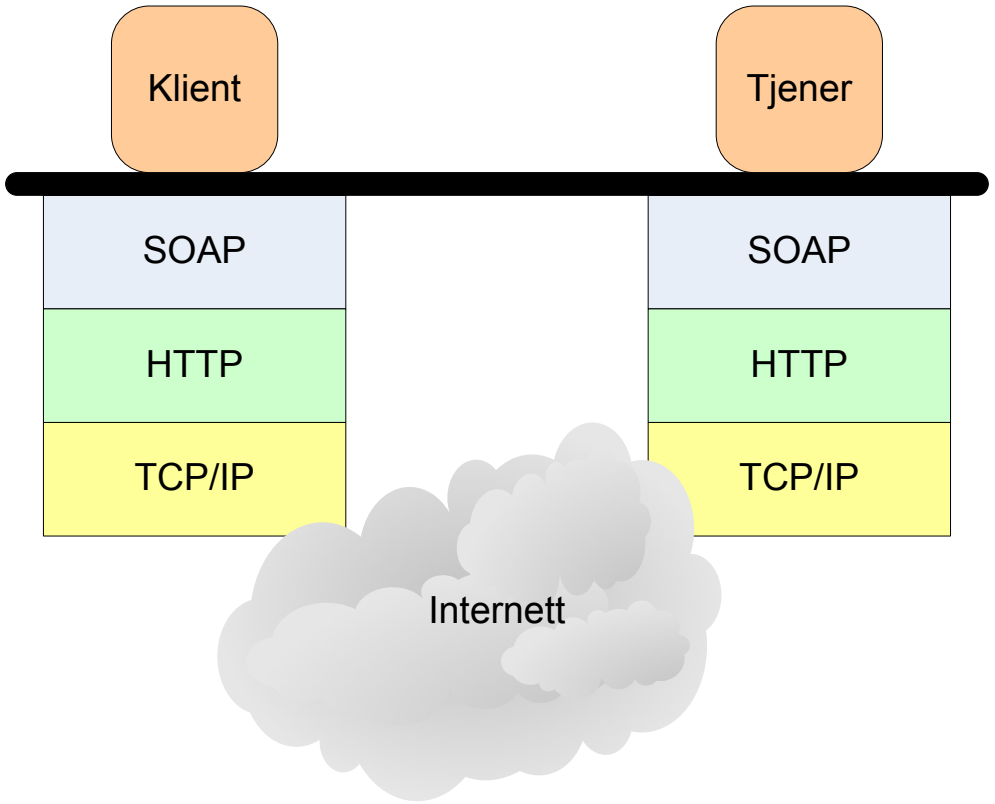
Why Semantic Web?

- Amount of data (Big Data)
- Problem of outdated pages

Technologies

- XML
  - XML Schema
  - XSL/XSLT (eXtensible Stylesheet Language)
- Web services
  - SOAP (Simple Object Access Protocol)
  - REST

Web Services



Web Services

- Web Service Description Language
- Shows methods and parameters
- A program can look for relevant methods
  - (But not often used in practice)

Ontologies

- RDF: Resource Description Framework
- OWL: Web Ontology Language

RDF

- The Resource Description Framework (RDF) is a flexible data model for resources described as objects and the relations among them.
- It provides a simple semantics for this data model, and these data models can be represented in XML syntax;
- RDF schema is a vocabulary for describing properties and classes of RDF resources, with semantics for hierarchies of such properties and classes;

OWL

- The Ontology Web Language (OWL) intends to provide a language suitable for describing the classes and relations inherent to Web documents.
- OWL has more facilities for expressing meaning than XML, RDF and RDF-S.

What is e-Government

- The use of ICT within government to provide better services to its citizens
- Improve government efficiency and quality
  - Externally
  - Internally

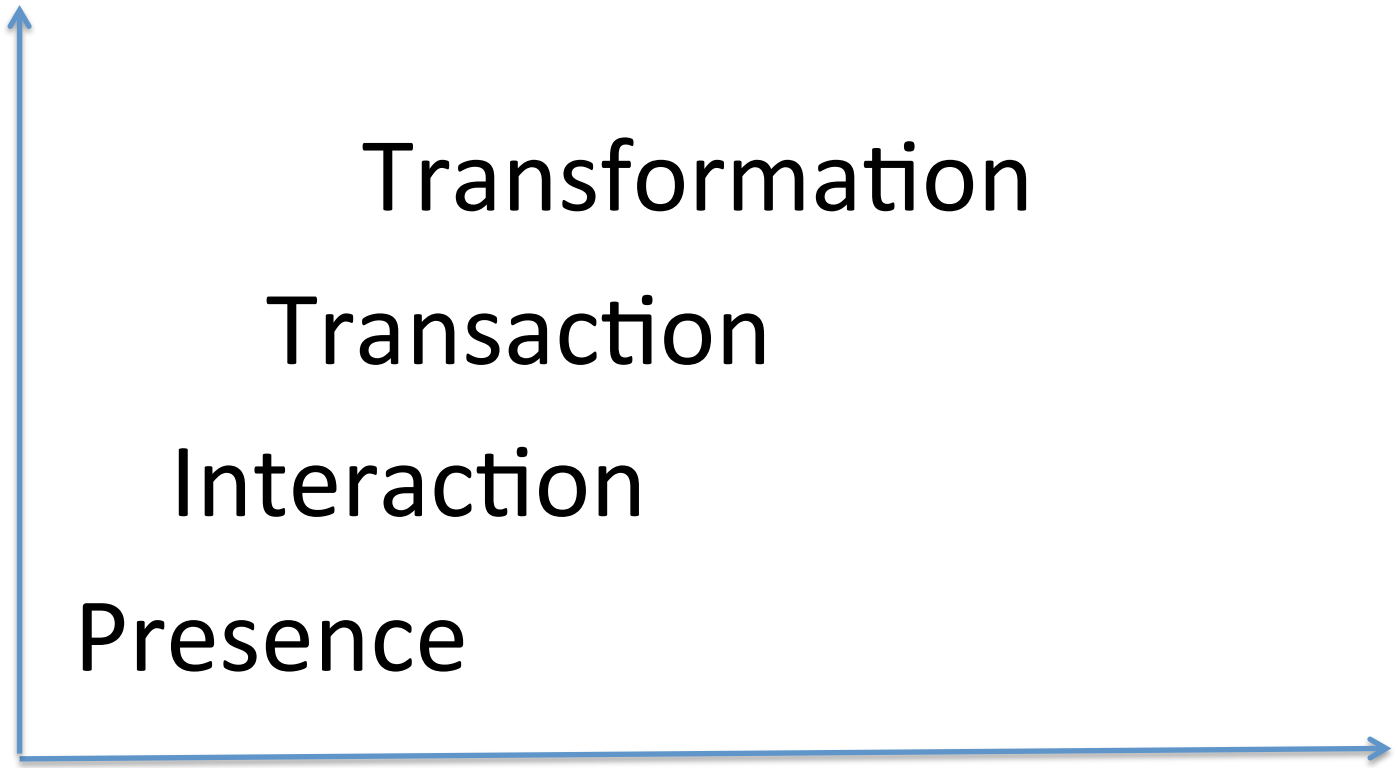
Early e-Government

- Provide electronic services for citizens 24 hours / 7 days a week
- Self-service
- Transactions through forms
  - Applying for Kindergarten
  - Tax return statements through Internet
- Technology, not organization

e-Government today

- Focus on backoffice integration
- Process engineering
- Multiple channel service delivery
- Portals and customization
- Organization, not technology

Maturity models



Presence

- Static web pages
- Information on services and organization
- Newsletters
- Contact information

Interaction

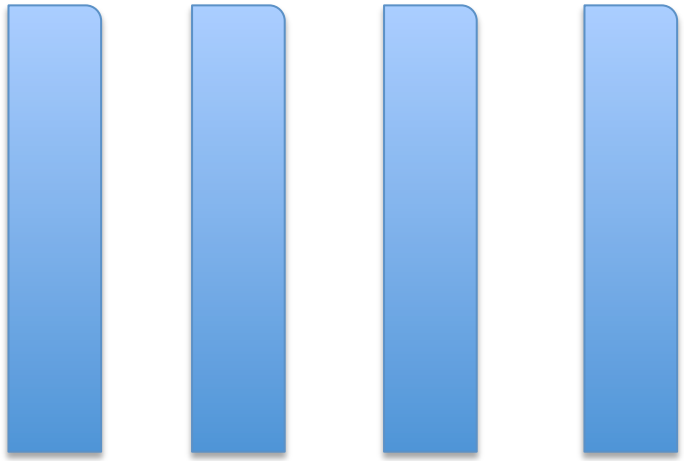
- Electronic forms
- Citizens can submit electronic forms
- Cross agency linking

Transaction

- Digital signatures
- Authentication
- Electronic payments

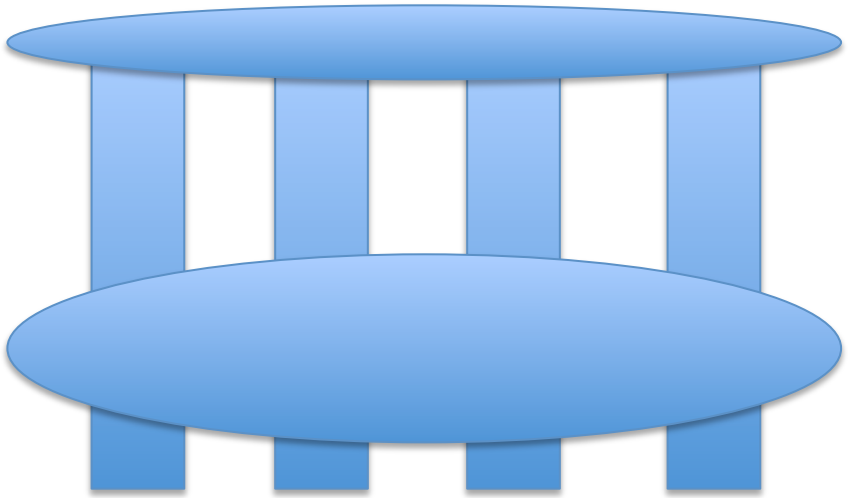
Often seen: Silos

- Each ministry, each directorate and agency have their own systems.
- Little willingness to share.



Transformation

- Integration of services
- Vertical integration
- Horizontal integration
- One-stop solution



Applications

- To address the problems of seamless integration and interoperability, stakeholders in e-government are experimenting with:
  - Semantic Web
  - Web Service Technologies
  - Service Oriented Architecture (SOA)
- As means for achieving integration and inter-operation in the service transformation phase.

Example: date of birth

- Used in many different contexts
- Stored by different technologies
- In different formats
- Over long time
- ‘citizen date of birth’ must be independent of the information systems and technologies that persist and present it: we must communicate understanding in a way that transcends technology, purpose, language, and time.

Research directions

- Semantic Web research directions that are relevant to e-government include:
  1. Social networking;
  2. Service composition and collaborative workflow;
  3. Security and trust;
  4. Automated collection and processing information;
  5. Adaptive information delivery.

Semantic interoperability

- Make sure that information exchanged can be understood by applications that are not made for handling this specific information,
- Semantic interoperability makes it possible for computer systems to combine received information with other information and process it in a meaningful way.

Semicolon

- Project funded by Research Council of Norway.
- Main goal:
- Develop and test ICT-based methods, tools and metrics to obtain faster and cheaper semantic and organizational interoperability both within and with public sector.
- http://www.semicolon.no

Semicolon

- Semicolon (2007-2010)
  - Semantic interoperability
  - Organizational interoperability
- Semicolon II (2010-2013)
  - Legal aspects
  - Political aspects
  - Collaboration platforms
- Measuring interoperability

Semicolon

- Presentation on some issues of interoperability:
- http://www.semicolon.no/wp-content/uploads/2013/10/2013-10-14-verdikt-erlend.pdf

Semantic technology

- Semantic technology facilitates the use of data from different sources.
- Three important public registers:
  - Persons
  - Organizational entities (businesses, organizations)
  - Properties

Another example

- ECRN European Civil Registry Network
- Funded by European Commission
- www.ecrn.eu
- Not focus on semantic interoperability, but it is a good case where semantic interoperability would make sense.

Search engine

- eGovMon
- Research project funded by Research Council of Norway
- Made experimental search engine for municipalities
- Feature: Semantic relations
- Help user find what the user are looking for
- Example: Translate synonyms, provide alternatives

SOME EUROPEAN PROJECTS

SemanticGov

- Semantic Web Services
- Across agency and national borders

SemanticGov

- In the SemanticGov project a fully transactional semantic portal has been developed, which:
  - helps citizens and businesses discover the public services which address their needs and personalize these services according to their profiles through a user-friendly, self-explanatory interface that offers them the necessary guidance throughout the process
  - informs citizens and businesses about their eligibility for a specific public service
  - provides complete and well-structured information for the identified service
  - allows citizens and businesses to invoke public services that are available online.

Terregov

- Implement government processes that invoke services (eProcedures, access to existing legacy information systems and databases) from multiple administrations;
- Make these government processes available to other administrations as eGovernment services;
- Support civil servants involved in such eGovernment processes in getting a clear knowledge of the processes and of the services in order to act as a knowledgeable front-end to citizens (providing advices, identifying the most adequate services, launching the processes for specific citizen cases).

Terregov

- Web Services and eGovernment Processes to combine flexible eGovernment interoperable services in end-to-end process workflows.
- Semantic enrichment eGovernment Services to enable Web Services to discover each other on a semantic basis.
- Support to Civil Servants to enable civil servants to focus on the added value of the service delivered to Citizens - increasingly acting as advisers.

OntoGov

- (2003-2006)
- The overall objective of OntoGov is to develop, test and validate a semantically-enriched (ontology-enabled) platform that will facilitate the consistent composition, re-configuration and evolution of e-government services.

OntoGov

- More specifically, OntoGov aims to:
- 1)Define a high-level generic ontology for the e-government service lifecycle (i.e. covering all the phases from definition and design through to implementation and reconfiguration of e-government services) that will provide the basis for designing lower-level domain ontologies specific to the service offerings of the participating public authorities;

OntoGov

- 2)Develop a semantically-enriched platform that will enable public administrations to model the semantics and the processes of their e-government service offerings at different levels of abstraction; easily and consistently re-configure their e-government services; and knowledge-enrich the provision of e-government services to citizens and businesses;

OntoGov

- 3)Pilot and evaluate the OntoGov platform in three public administrations in three different European countries. The evaluation of the project results will not be limited merely to the technical evaluation; rather, it will take into account both organisational and social aspects of the project.

Access eGov

- *Access to e-Government Services Employing Semantic Technologies (*01.01.2006-28-02-2009)
- Central idea is to ensure that the meaning of the shared and exchanged information is captured, formalized, and understood in the same way by stakeholders and applications.
- RDF, OWL
- Three pilots: Getting married (Germany), Obtaining a building permission (Slovakia) and Establishing an enterprise (Poland)

Conclusion

- Lots of demos
- Integration is progressing, but slowly
- Not only technical issues
- Organizational, legal, résistance to change
- Most interesting developments happen in the area of open data
- Big opportunities for making new and innovative applications.

OPEN DATA

Open Data

- In 2011, the Norwegian government opened a portal for open data.
- The ministries and government agencies asked to submit data sets to the portal.

Current Status

- As of today, 558 data sets have been submitted.
- The most popular categories is shown in the list:

- Public Administration: 178
- Municipalities and regions: 69
- Economy: 35
- Transport and Communication: 35
- Food, fishing and agriculture: 31
- Health and care: 26
- Education and research:26
- Environment: 24

Open Data

- Open Data is growing, and can be combined with other data sources to make new, exiting applications.
- Open data is not only a government thing. Also companies and private persons may contribute.

Case #1: The Digital Inn

- Established by the Norwegian National Archive Services as an extension of a information retrieval service called the ”Digital Archive". The “Digital Archive” stores archive material as images, transcribed texts and databases, and makes such material available through the Internet.

”The Digital Archive”

- Content is of particular interest to historians and genealogists, and include:
  - Censuses
  - Parish records
  - Military service records
  - List of emigrants

”The Digital Archive”

- Some material is transcribed, but today even more material is stored as images.
- Transcribed material may be searched, but not images
- Transcription is a very time-consuming process

”The Digital Archive”

- The Digital Archive made a strategic decision to open their infrastructure to individuals and voluntary organizations registering parish records and other historical content as digital information.
- This is what is called ”The Digital Inn”.
- You get a room and fill it with your own belongings..

”The Digital Archive”

- This is one good example on how to consider citizens as a resource.
- The individual contributions are shared with others through a public infrastructure.

Case #2: The Map Hostel

- Based on the same ideas.
- Let government provide infrastructure, and invite citizens and organizations to provide map data.

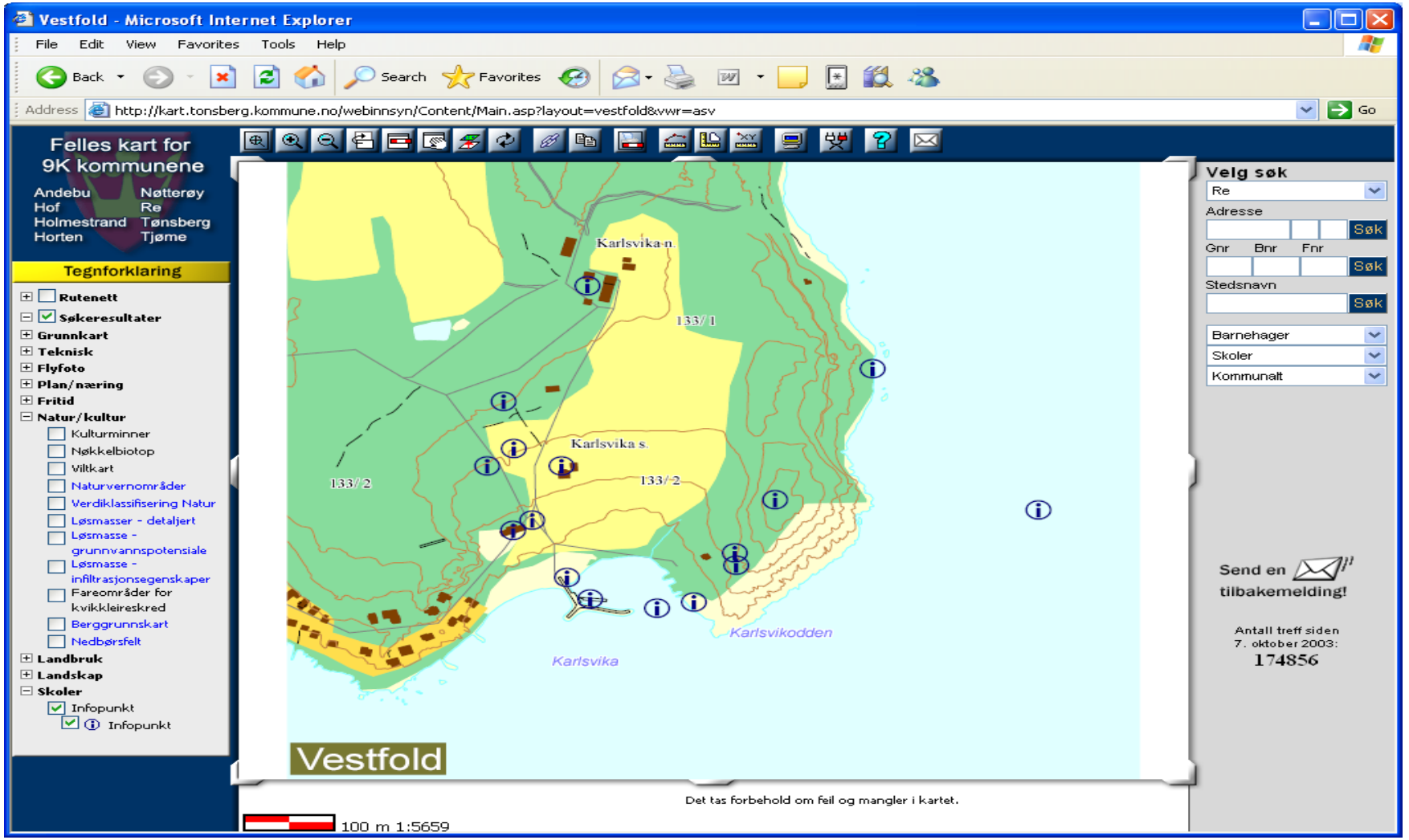
The rationale

- Map services on Internet is flourshing
  - Google maps
  - Microsoft
- But applications are limited, since
  - Maps provided lack details
  - There are limitations on what data can be stored
  - ”Street maps”

Shared maps

- My research group has worked on shared maps since 2005.
- At that time we collaborated with municipalities in order to enhance their maps with new information elements.

Trial experiment



Shared Maps

- Since then we have got Google Maps and Bing. These services are much easier to use, and to enhance.

Make your own map

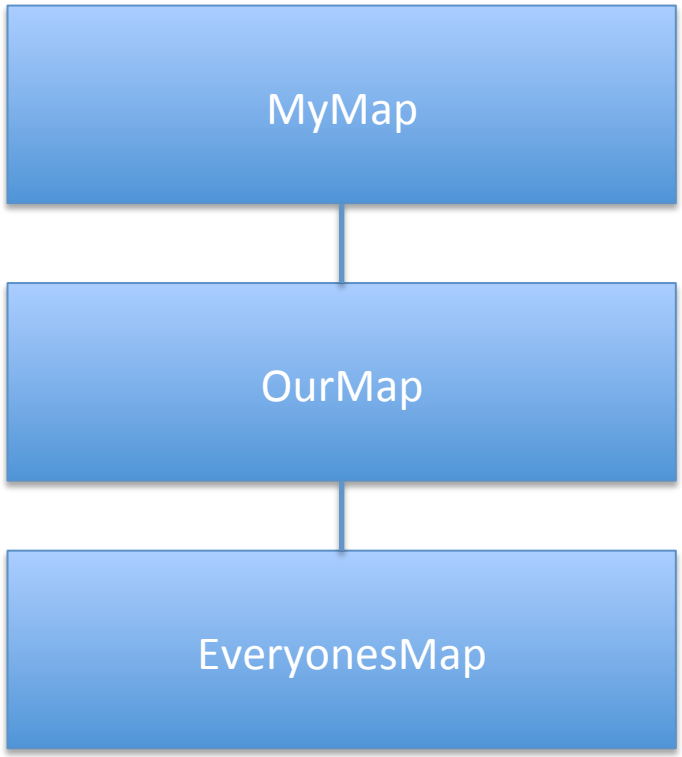
- Student project.
- Used mobile device (Windows based) for registration.

Use of mobile units

- New mobile phones with built-in GPS facilitates in-the-field data collection.
- Have developed application to send GPS-data and message to web server through a web service.
- Have developed application to import from TrackStick.
- Are now implementing on other devices using external (Bluetooth) GPS (better solution).

Hierarchy

- First level: **MyMap** (personal data or data being collected)
- Second level: **OurMap** (data shared by closed group or members of an organization)
- Third level: **EveryonesMap** (data accessed through a public catalog)



Applications

- Sports clubs
- Registration of places of historical interest
- Orienteering
- Geology
- Schools
- Applications are endless

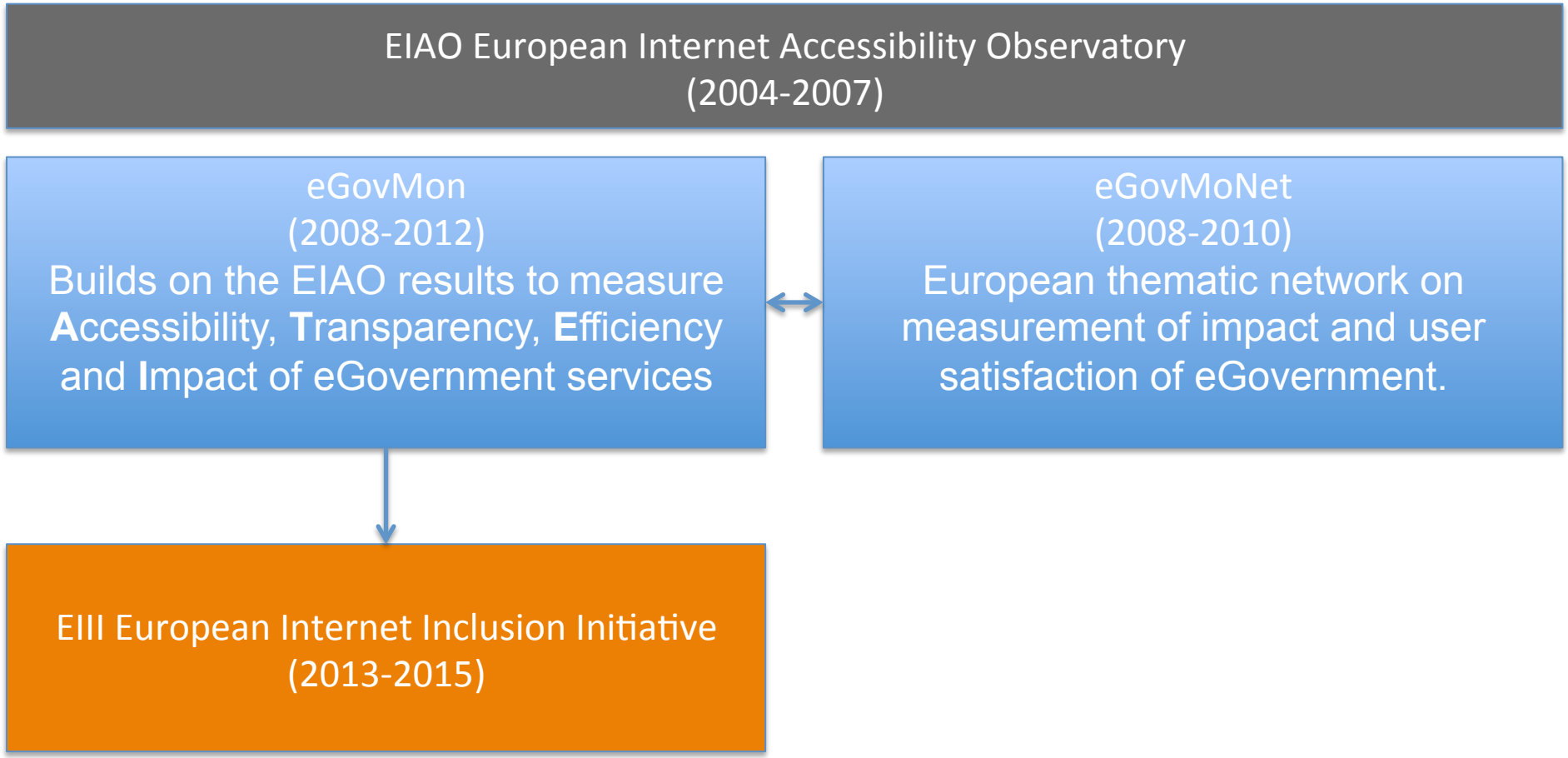
One approch to collect ”big” data

EGOVMON

eGovMon

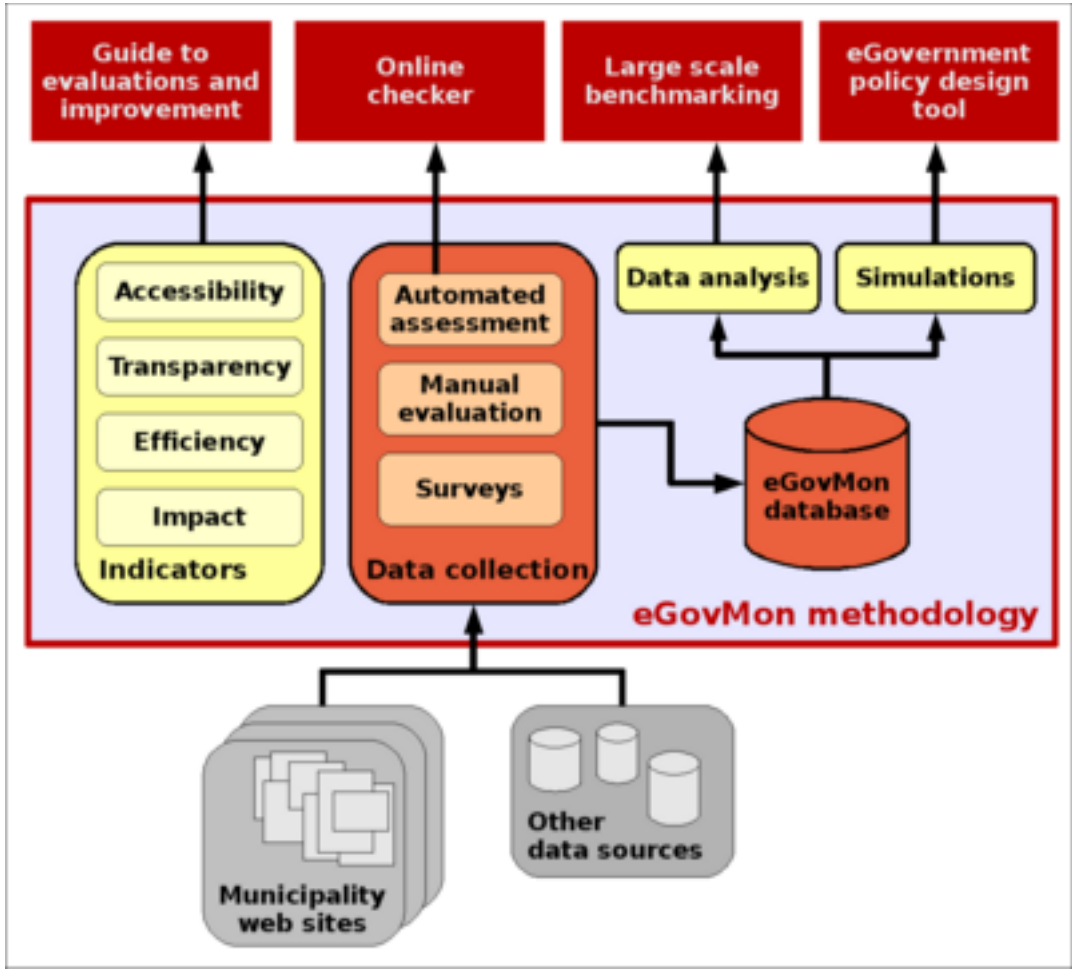
- Research project financed by the Research Council of Norway to automate assessment of public web sites.
- Four pillars:
  - Accessibility
  - Transparency
  - Efficiency
  - Impact

Background


EIAO European Internet Accessibility Observatory
(2004-2007)
eGovMon
(2008-2012)
Builds on the EIAO results to measure Accessibility, Transparency, Efficiency and Impact of eGovernment services
eGovMoNet
(2008-2010)
European thematic network on measurement of impact and user satisfaction of eGovernment.
EIII European Internet Inclusion Initiative
(2013-2015)

Methodology

- Develop indicators
- Collect data
- Analyze data
- Perform simulations



Observations

- What was successful:
  - Indicators were developed for all four pillars
  - Data mining were used for accessibility and transparency
  - Accessibility is mostly technical and easy to assess
  - Transparency is harder, but we managed through various techniques to achive very good assessments
  - We did some very nice things on analysis and visualization

Accessibility

- Mostly technical rules
- Web Content Accessibility Guidelines
- Most rules can be checked automatically

How to find transparency elements

- Site crawler
- All municipalities have web addresses with the following format:
- <municipality name>.kommune.no

Two approaches

- Machine learning
- Identify relevant subsystems

Step one

- Obtained list of municipality names from Statistics Norway
- Comma-separated values
- Constructed URL´s based on the municipality names
- Crawled all web sites to check is response was OK
- A few fixes was need, mostly because of municipalities with same name or special characters (æ,ø,å)
- We now had a complete list of valid URL´s

Step two

- Identify portals used.
- We have four large vendors and a few smaller vendors
- Found patterns to identify the four large. After running the crawler, we had a much smaller list of non-identified portal vendors.
- Started with the first, identified the vendor, found a pattern to identify the vendor, and ran the crawler again.
- After some iterations, we had identified all vendors.

Use of Pattern Matching

- The following example shows the use of pattern matching in practice.
- ACOS is one of the big vendors of document handling systems. A module makes political documents available online.
- We found two distinct patterns which shows the presence of the transparency module.

Pattern Matching (1)

```
private void acosPatterns(DataRow d)
    {
        WebClient a1 = new WebClient();
        a1.Headers.Add("User-Agent", "Mozilla/5.0 (Windows NT 6.2; rv:9.0.1) Gecko/20100101 Firefox/9.0.1");
        byte[] b1; string url1;
        url1 = d["url"].ToString();
```

Pattern Matching (2)

```
 try {
    b1 = a1.DownloadData("http://" + url1 + "/innsyn.aspx?response=arkivsak_detaljer");
    string res = convertByteArrayToString(b1);
    if (res.Contains("innsyn.aspx")){
       d["innsynsmodul"] = "ACOS 1 WebSak transparency module found at http://" + url1 + "/innsyn.aspx";
       return;
    } else { Console.WriteLine("Found Acos, but verification failed"); }
}
catch (System.Net.WebException ex) { Console.WriteLine("EXCEPTION: " + ex.Message); }
```

Pattern Matching (3)

```
Try {
    b1 = a1.DownloadData("http://" + url2 + "/");
    string res = convertByteArrayToString(b2);
    if (res.Contains("/wfinnsyn"))
    {
        d["innsynsmodul"] = "ACOS 2 transparency module found on http://" + url2;
        return;
    } else { Console.WriteLine("Found Acos, but verification failed"); }
}
catch (System.Net.WebException ex) { Console.WriteLine("EXCEPTION: " +
ex.Message); }
} // END acosPatterns END
```

DEMO

- http://checkers.eiii.eu
- The checkers shown here are based on developments within the eGovMon project. They are further developed as part of the EIII project described earlier.

THE USE OF SENSORS

Sensors

- Analog
- Digital
- Sensors with built-in logic
- The I^2C interface
- Vision and sound

Analog sensors

- The sensor delivers analog values
- These values need to be translated into digital values to be processed
- ADC – Analog to Digital Converter
- Example: Temperature sensor

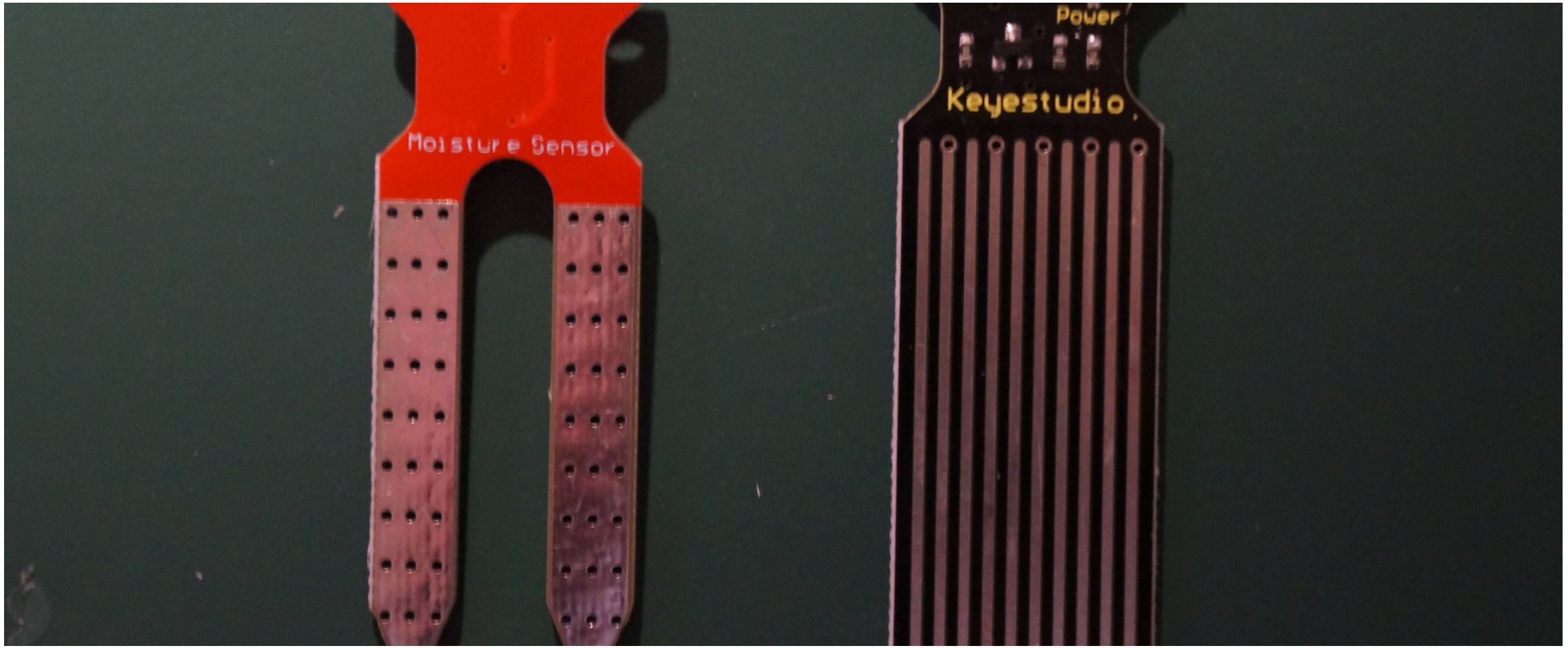
Digital sensors

- Provides a digital value (on/off) or a set of digital values
- No need to convert values
- Example: Switches

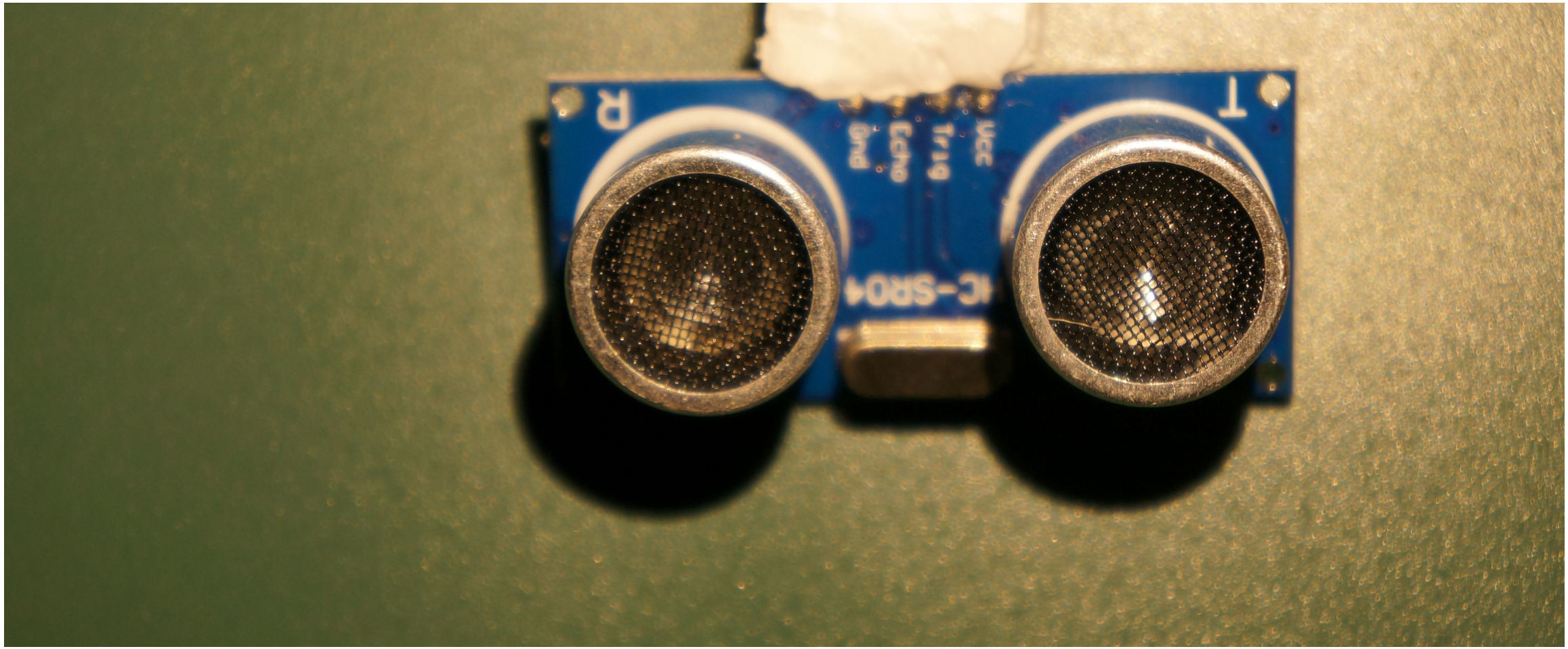
”Intelligent” sensors

- The sensor contains built-in electronics to do conversion
- Often, such sensors supports polling
- I^2C – a bus that uses two wires for communication

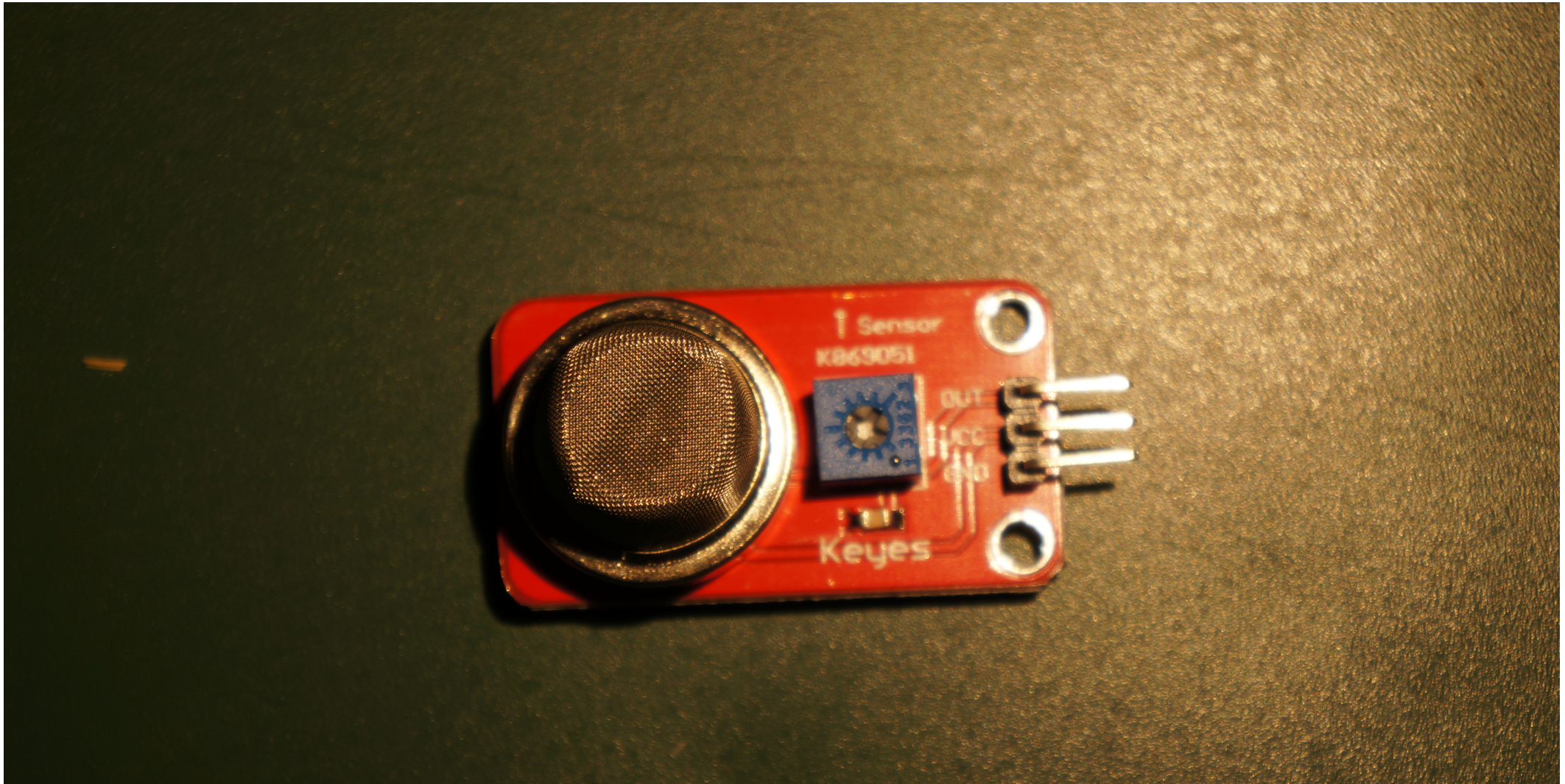
Humidity and Water Sensors



Ultrasound Distance Measurement Sensor



Gas Sensor



I^2C

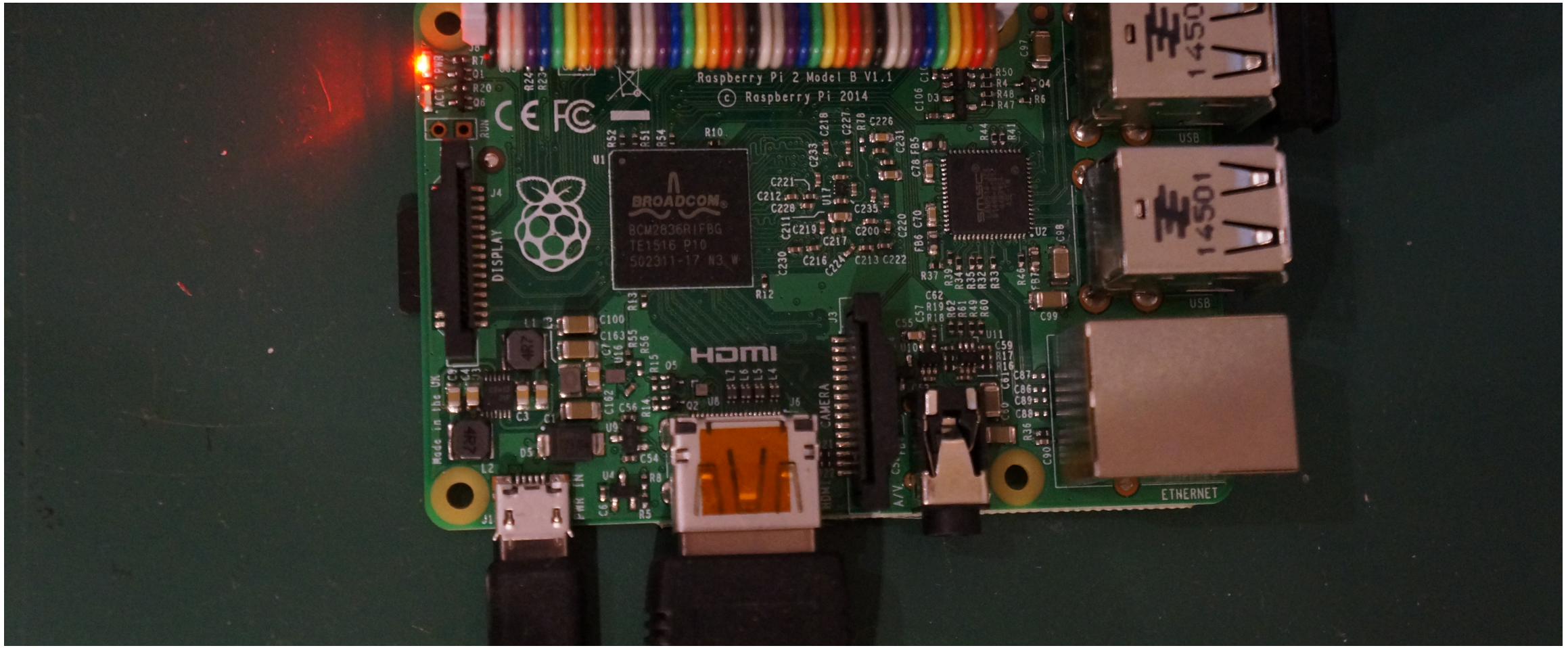
- The processing unit sends a signal to select the sensor
- The sensor responds with values

Temperature sensors

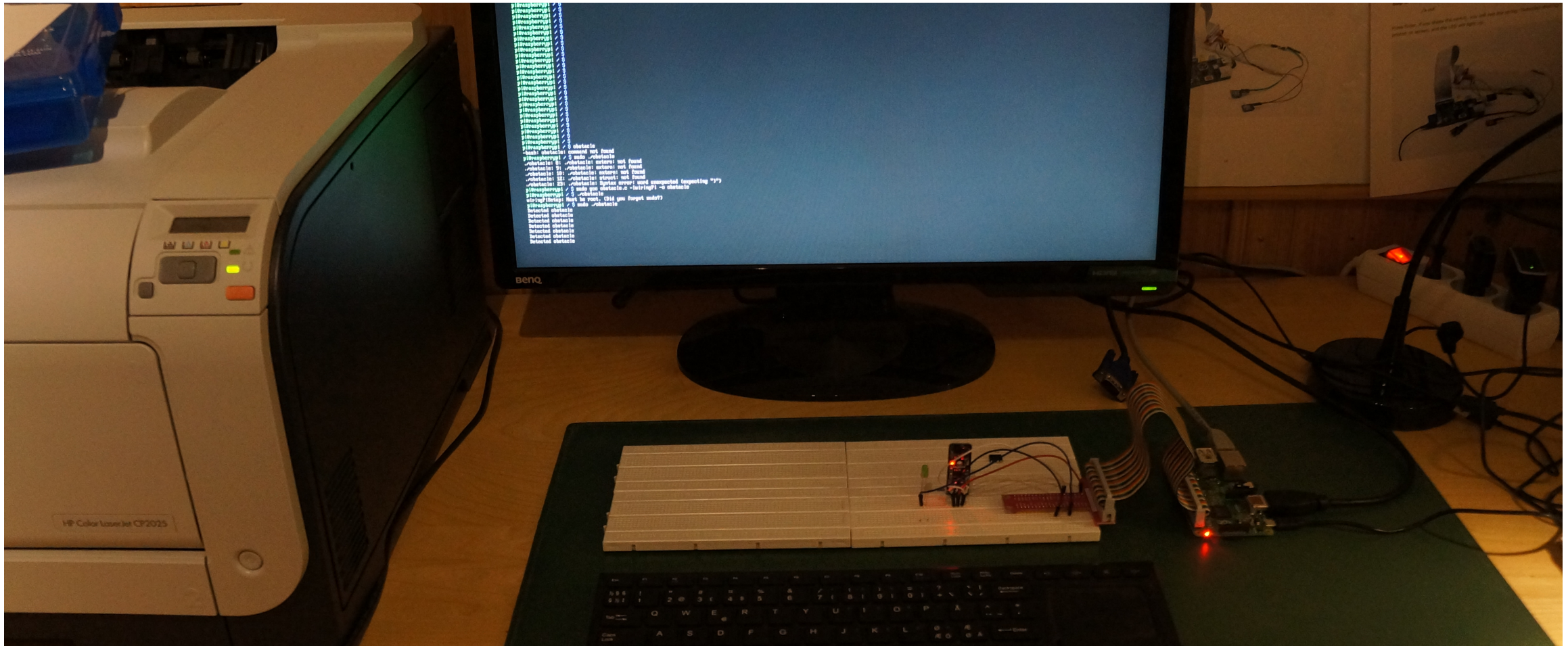
Raspberry Pi

- Cheap processing unit
- On-board Ethernet
- Memory card
- Video and HDMI
- Audio
- Four USB ports

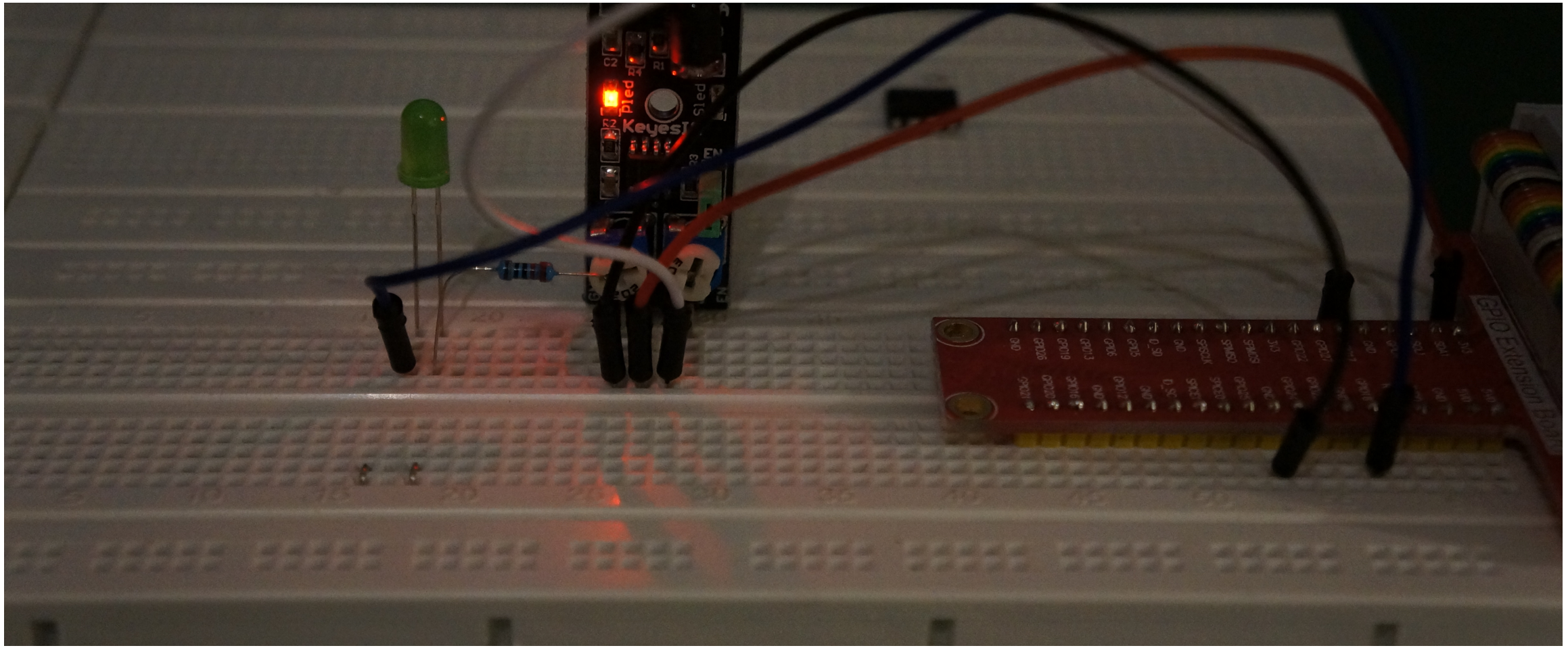
Raspberry Pi



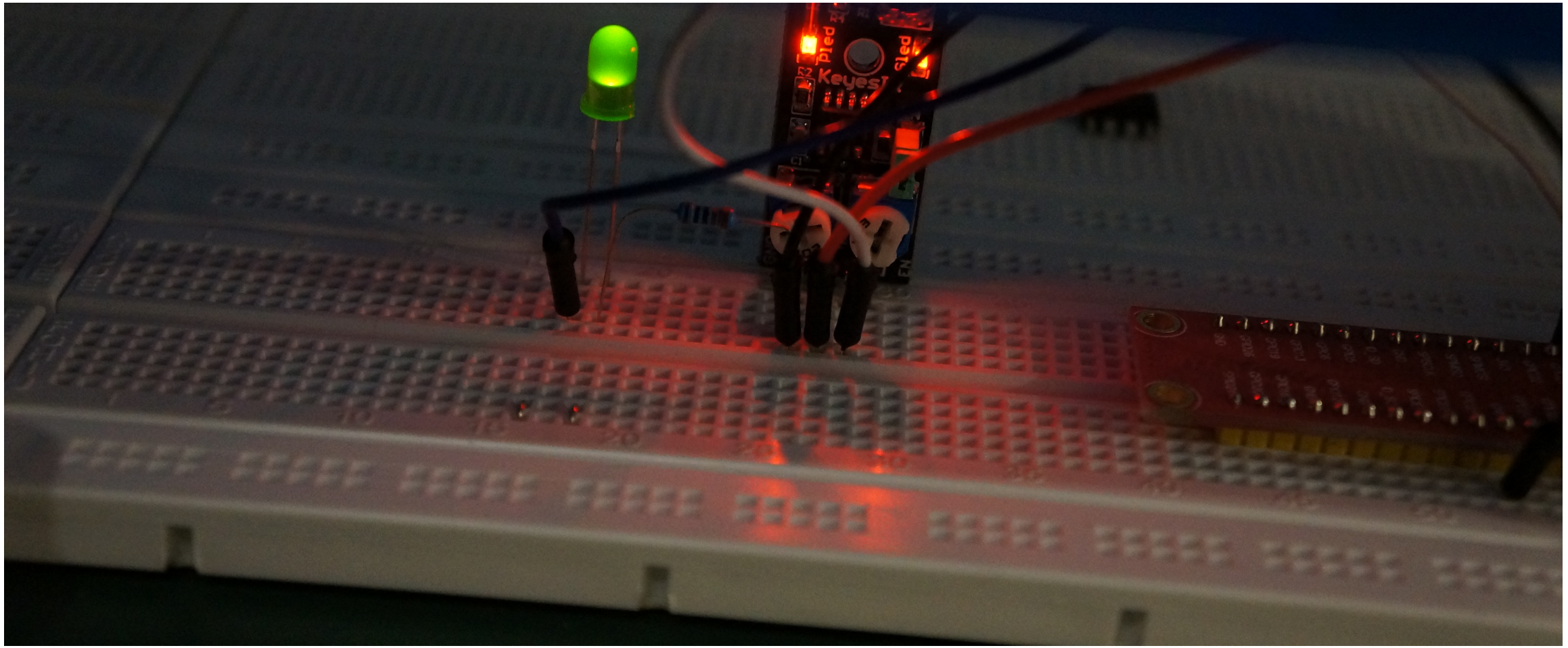
Raspberry Pi Development Environment



Example: Obstacle sensor



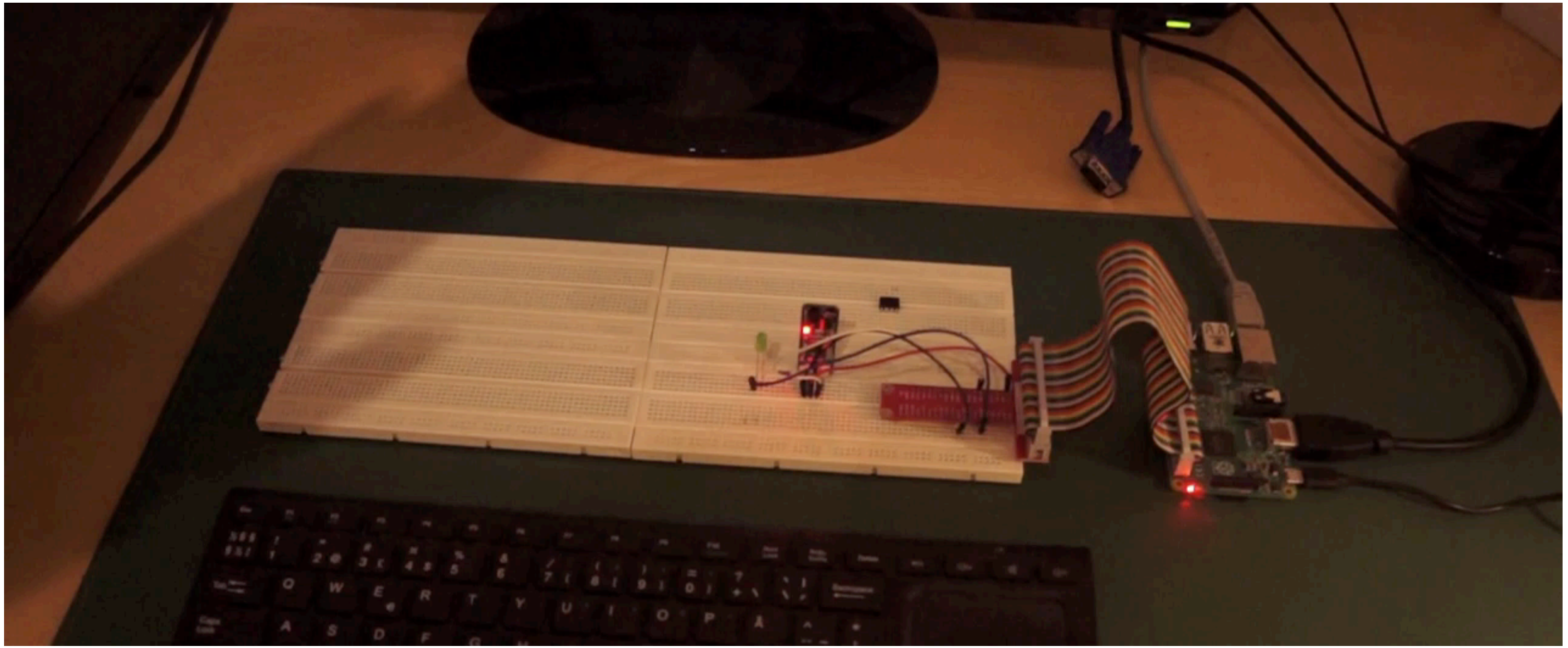
Example: Obstacle sensor



Raspberry Pi Example

```
#include <wiringPi.h>
#include <stdio.h>
#include <stdlib.h>
#define ObstaclePin 0
#define LedPin 1
int main(void){ if(wiringPiSetup() == -1){
   printf("Setup failed");  return 1; }
   pinMode(ObstaclePin, INPUT);
   pinMode(LedPin, OUTPUT);
   while(1){  if(digitalRead(ObstaclePin)==LOW){
     delay(25);
     if(digitalRead(ObstaclePin)==LOW){
       printf("Detected obstacle\n");
       digitalWrite(LedPin,HIGH);
       delay(2000);
       digitalWrite(LedPin,LOW);
       }
     }
   }
return 0;
}
```

Demo: Obstacle Sensor

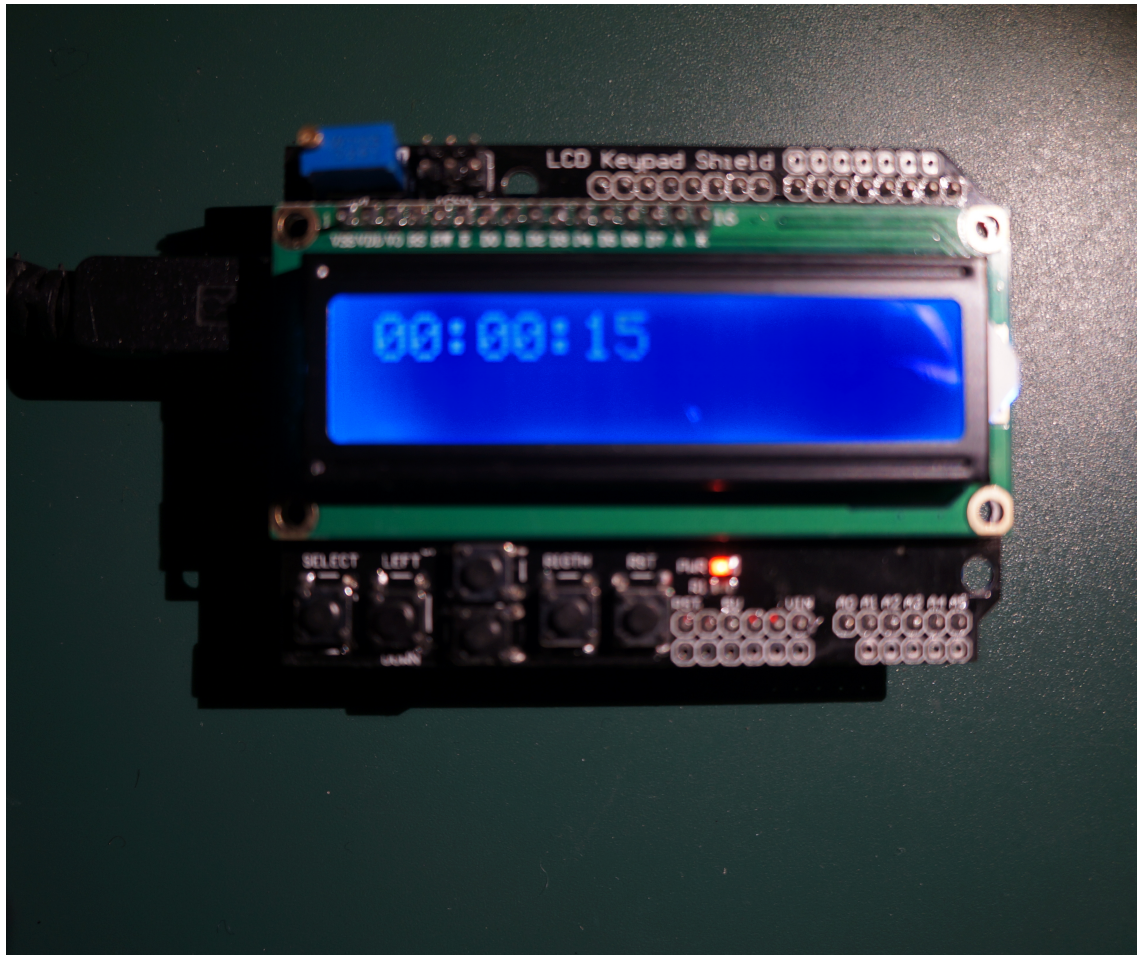


Arduino

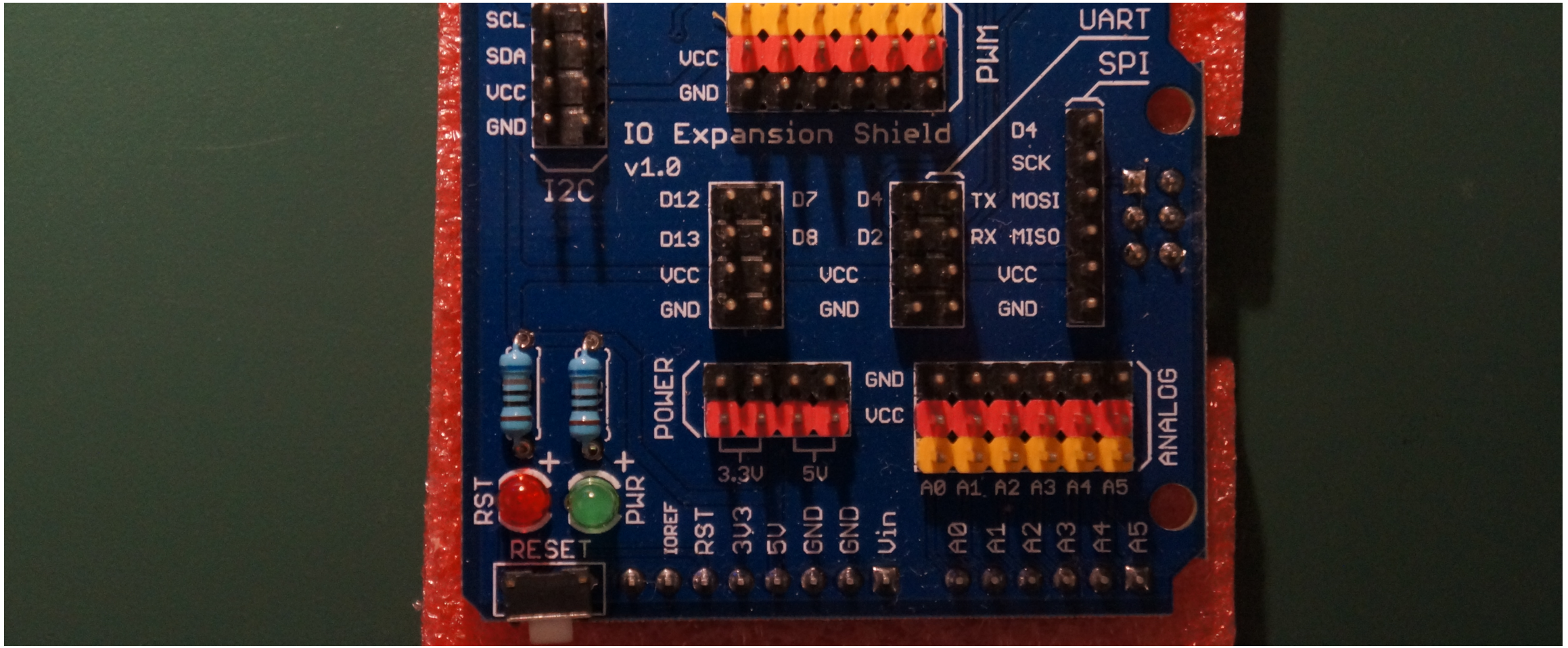
- Cheap, open source processing unit
- Supports both analog and digital inputs
- Arduino can be enhanced by shields
- A shield is a circuit board that is put on top of the Arduino

Arduino

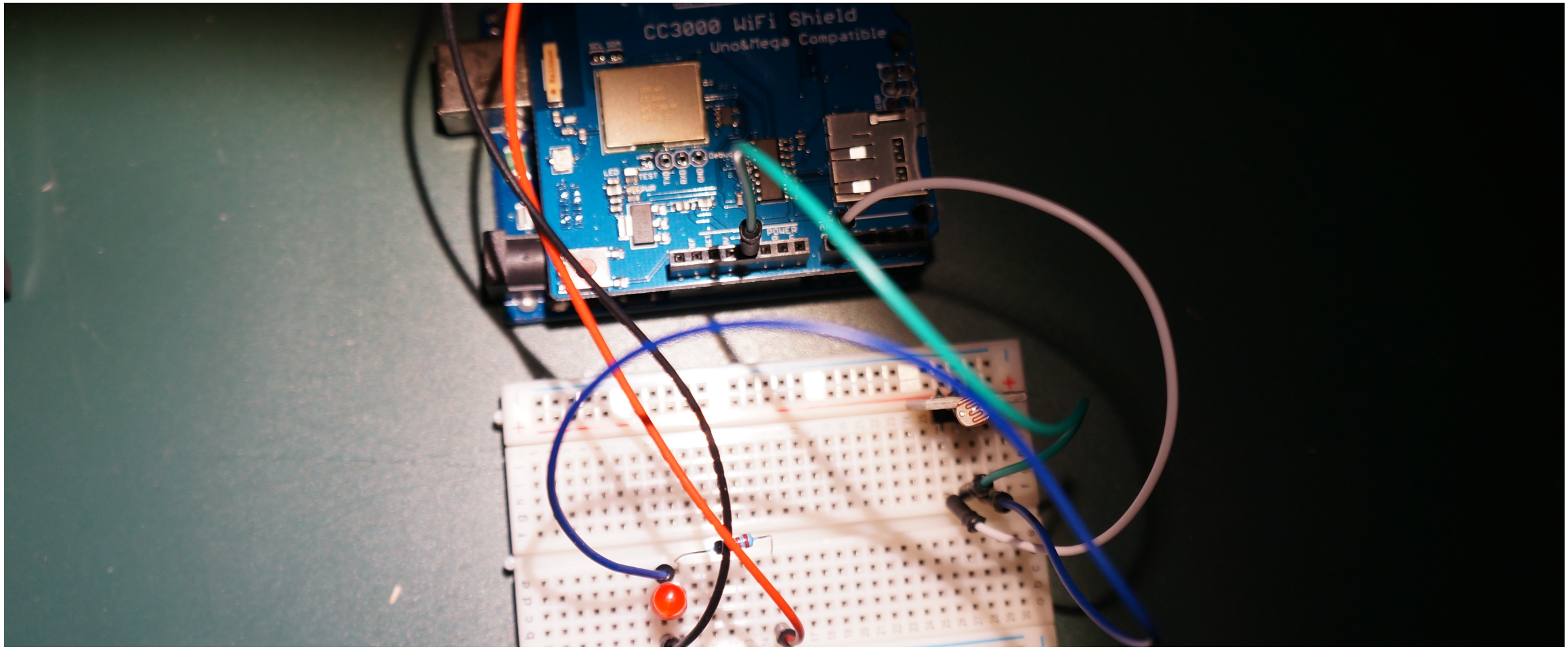
- This is one example of a shield providing a LCD display and six buttons.
- This was used to implement a stopwatch application.



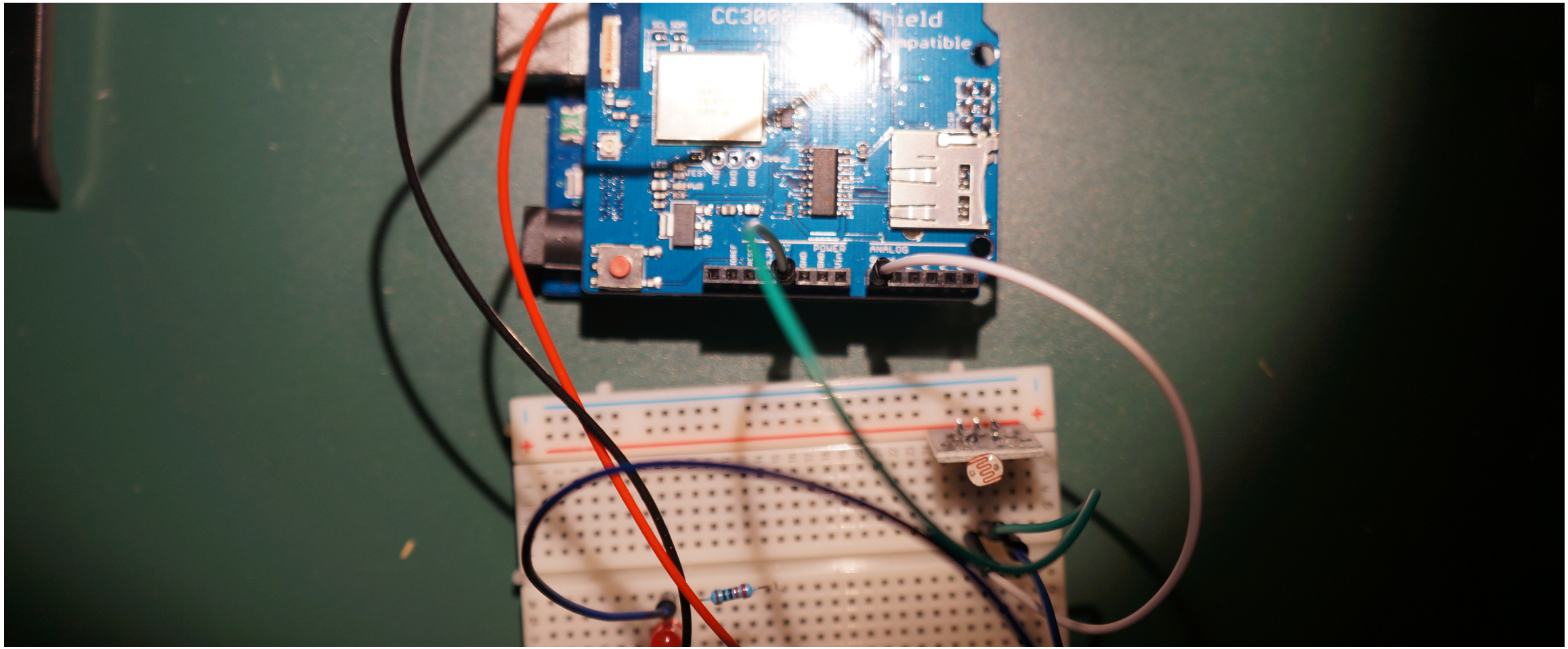
Shield


SCL
SDA
VCC
GND
I2C
VCC
GND
PWM
IO Expansion Shield
v1.0
UART
SPI
D4
SCK
TX MOSI
RX MISO
VCC
GND
D12
D13
VCC
GND
D7
D8
VCC
GND
D4
D2
POWER
GND
VCC
3.3V
5V
ANALOG
A0 A1 A2 A3 A4 A5
RST
PWR
RESET
IOREF
RST
3V3
5V
GND
GND
Vin
A0
A1
A2
A3
A4
A5

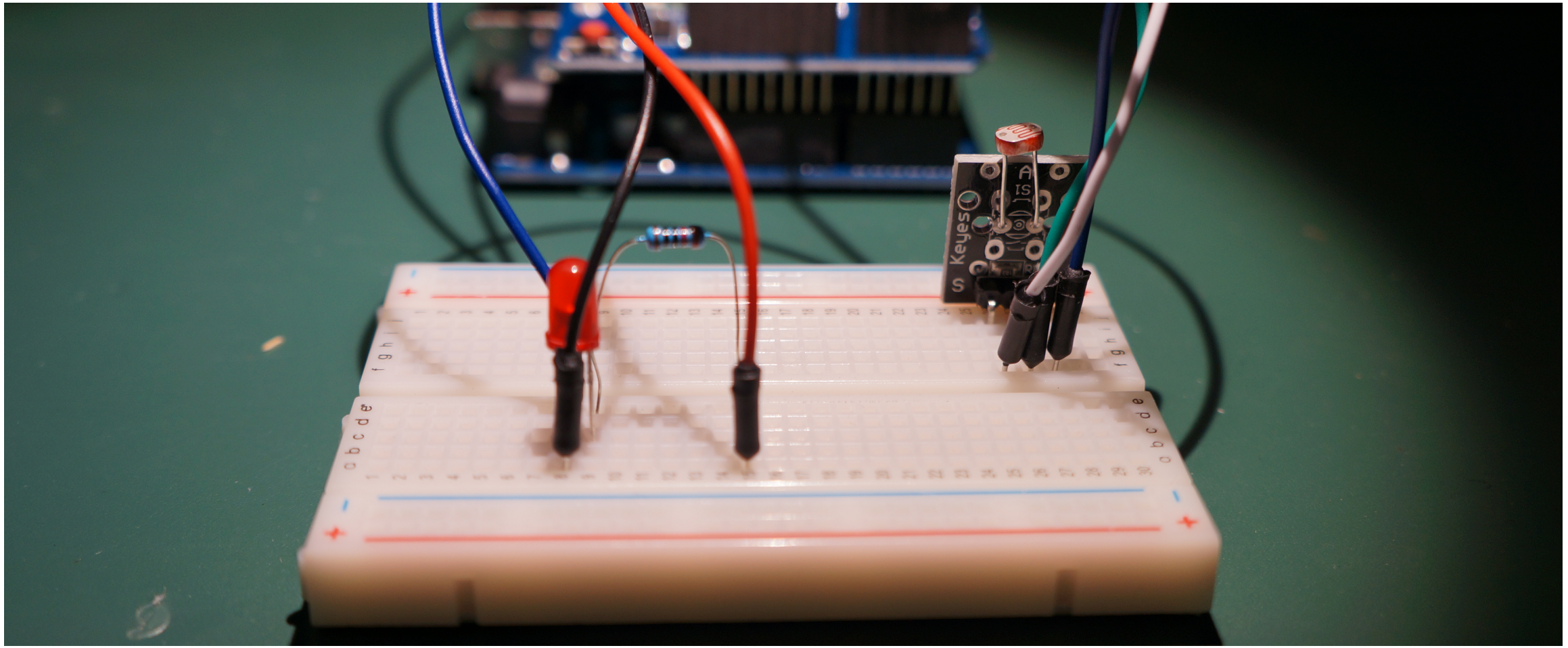
Demo: Arduino + LAN Shield



Using a Light Sensor



Using a Light Sensor



Live Demo

- Arduino Uno
- Lan Shield
- Photoresistor
- LED

- Network library provides connectivity for a very small web server. The web server can respond to commands (light the led), and provide information on light level.

Mobile (GSM) Unit

ANALYZIS / ANALYTICS

Analytics

- We have all these data from social media, open data sources and sensors.
- How do we proceed?
  - Machine learning
  - Neural networks
  - Statistical techniques (R is very popular)

For more information on analytics, I recommend to look at PacktPubs series of books.
http://www.pactpub.com

VISUALIZATION

Why Visualization?

- ”A picture is worth a thousand words”.
- Refers to visualization.
- Absorb large amounts of data quickly.

Visualization can be done in numerous ways

- Charts (bar charts, pie charts)
- 2D and 3D
- Maps

Dashboard

- Presenting information from many sources on the same screen.
- Examples: Tableau
- http://get.tableau.com

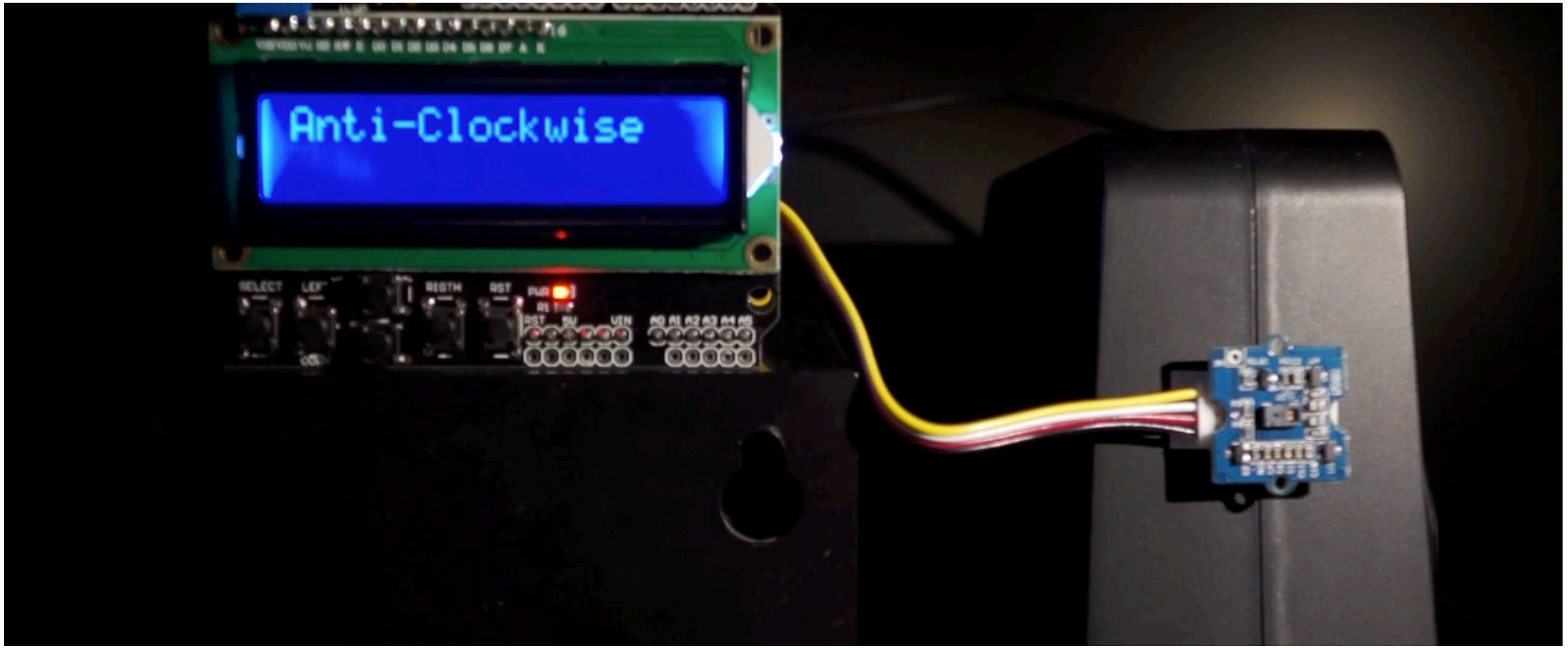
Visualization

- SimSam, our laboratory for collaboration and simulation

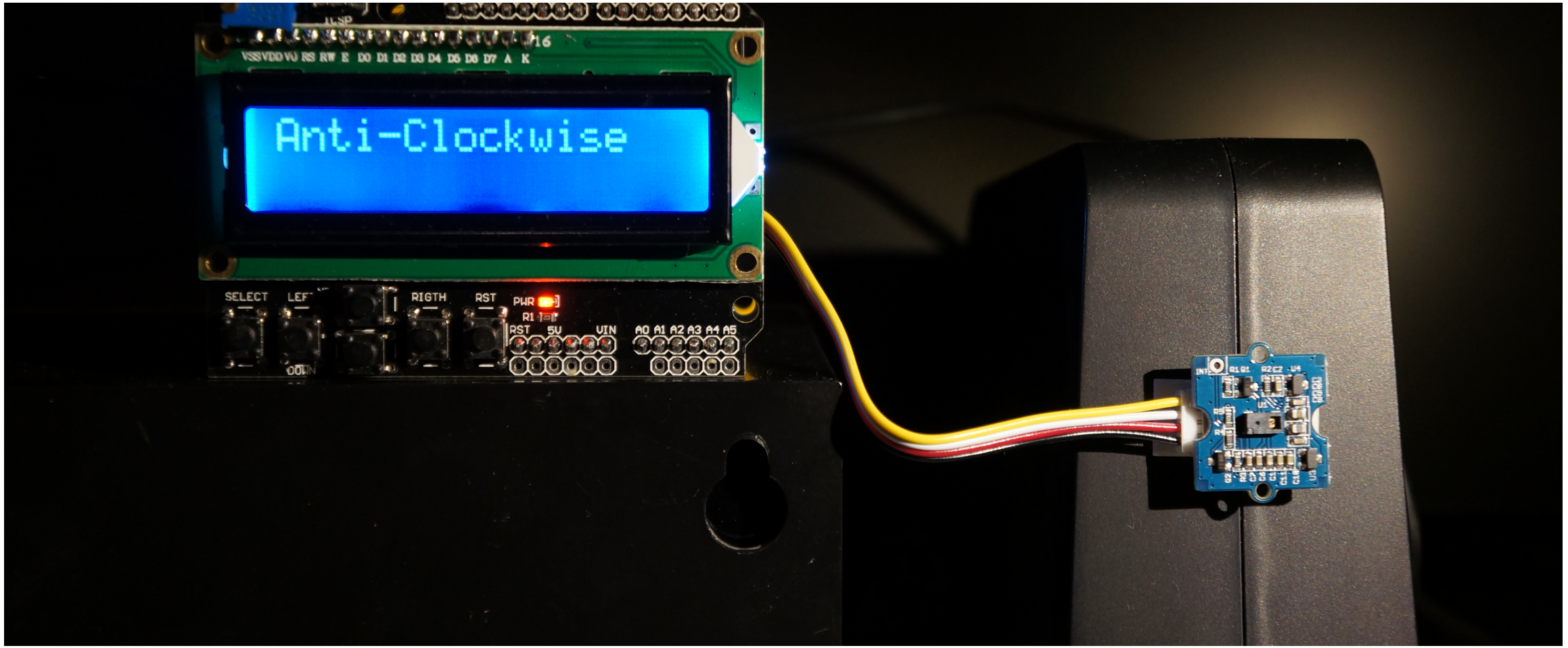
Sensors and visualization

- Gestures
- Apple, mobile phones

Gesture Sensor



Arduino with LCD shield and gesture sensor



Making Better Decisions

- We need to:
  - Collect data from many sources
  - Convert to information elements
  - Store and analyze information
  - Visualize the information to give better background for decisions
  - Make the decisions

- Case: Situation room
  - Handle incidents by informed decision making
  - Can be used also in private secor/big corporations

THANK YOU FOR LISTENING!