Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Probabilistic plausibility for surface data

EUMETNET STAC AQC workshop, March 2019

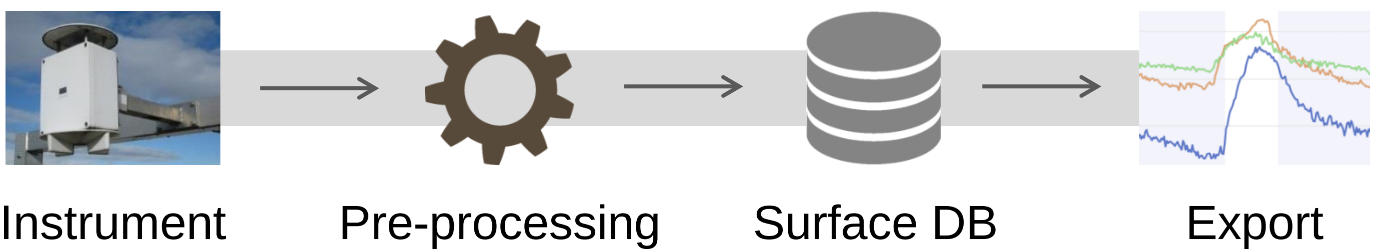
Christian Sigg, Deborah van Geijtenbeek,
Markus Abbt and Marc Musa

christian.sigg@meteoswiss.ch

In a nutshell

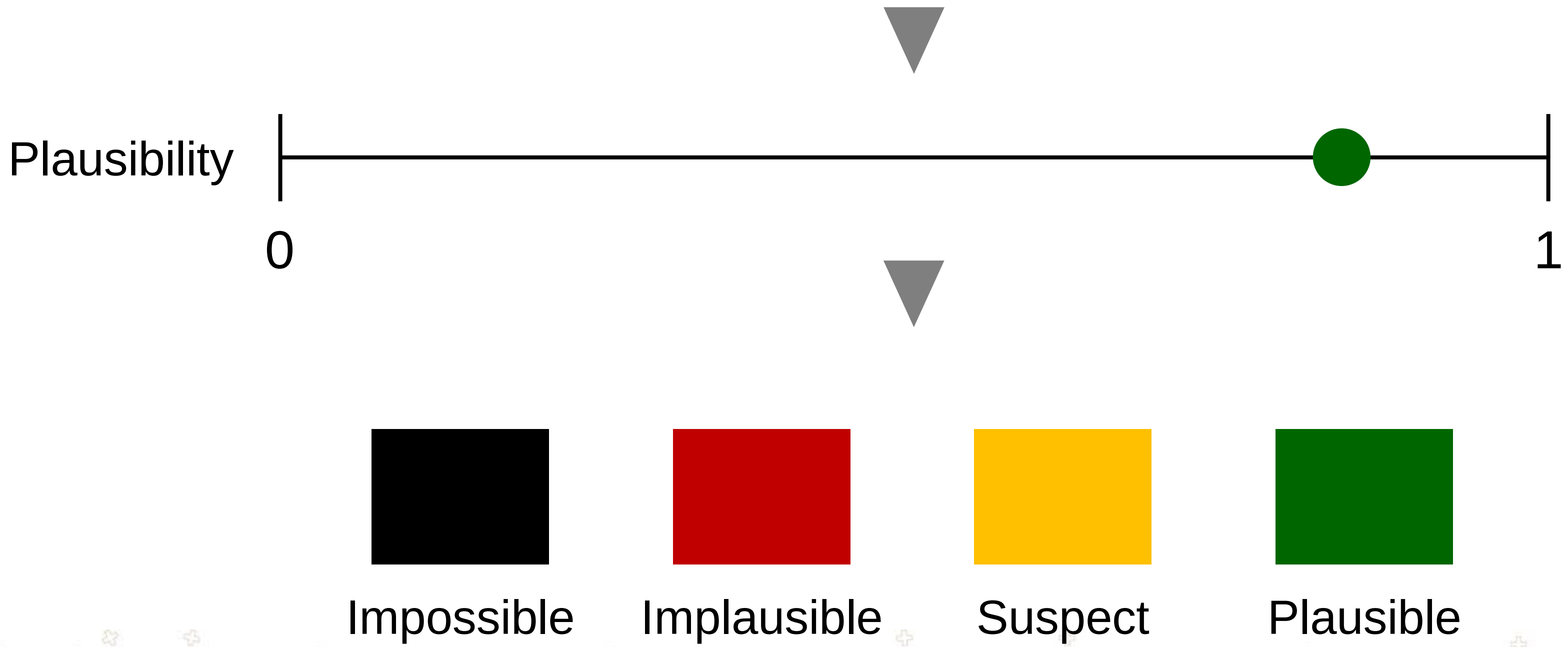
Probabilistic plausibility adresses two challenges:

1. How to **combine** quality information (QI) generated by multiple independent quality control (QC) systems along the data processing chain



In a nutshell

2. How to provide a **summary** of the QI that is simple, well-defined and relevant to the user

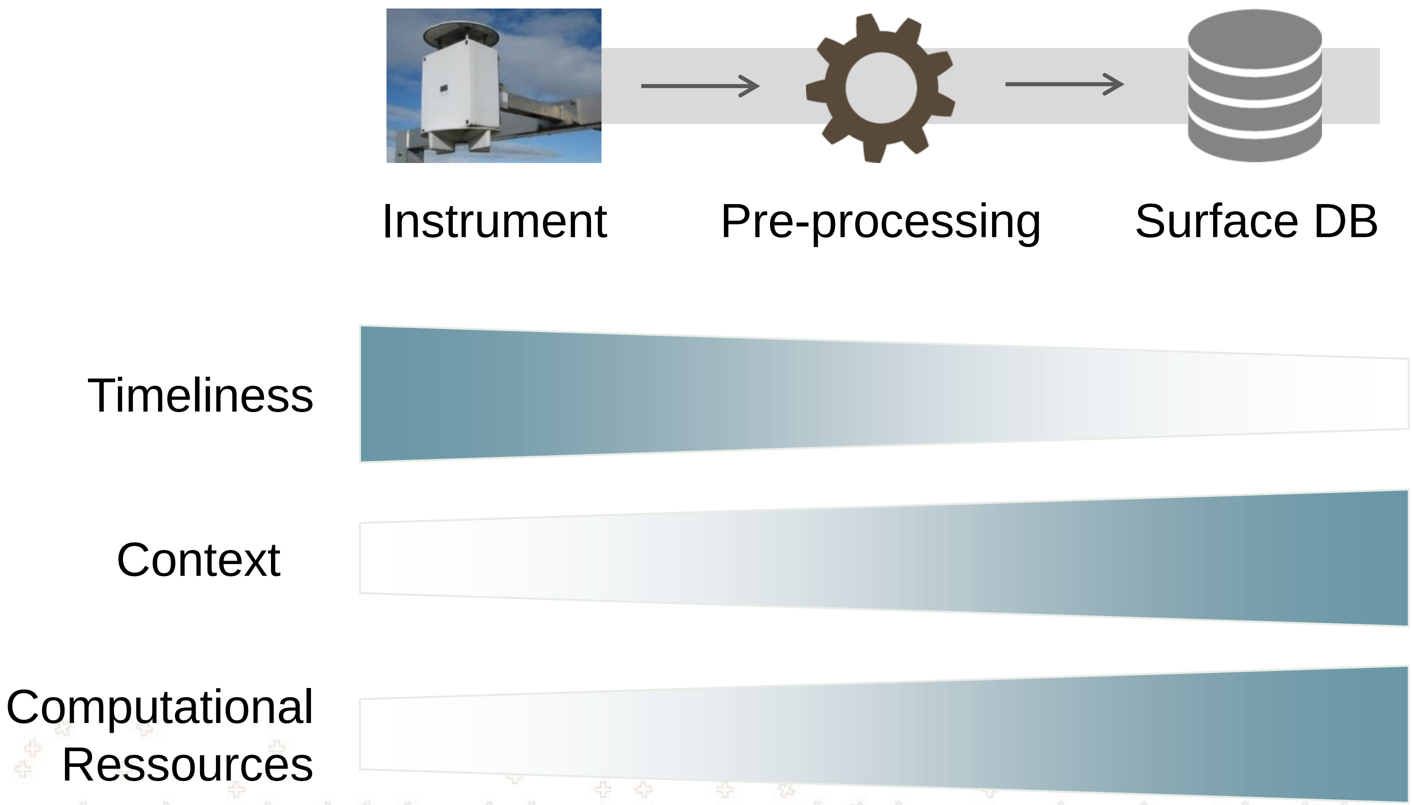


Outline

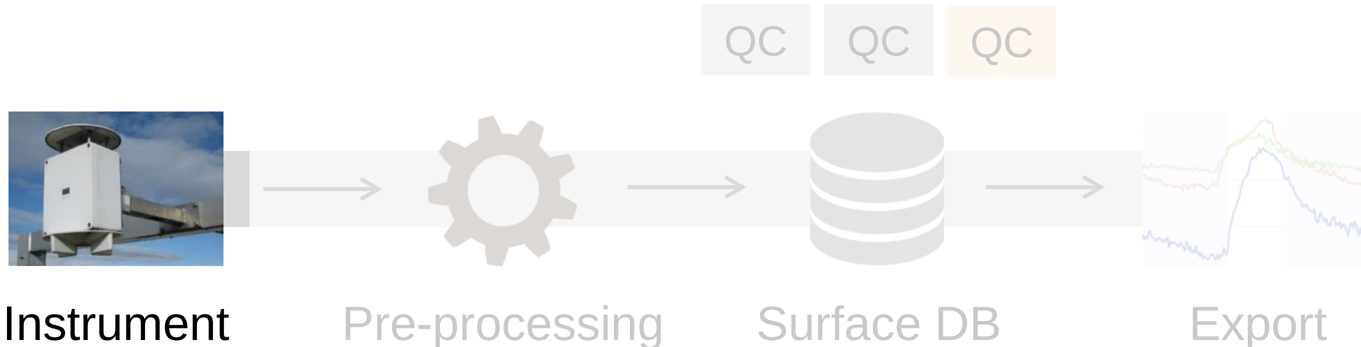
- QC along our data processing pipeline
- Our former QI
- Probabilistic plausibility
- Summarizing QI for the user
- Discussion

QC along data processing pipeline

Why do QC in multiple stages? **Trade-off**:

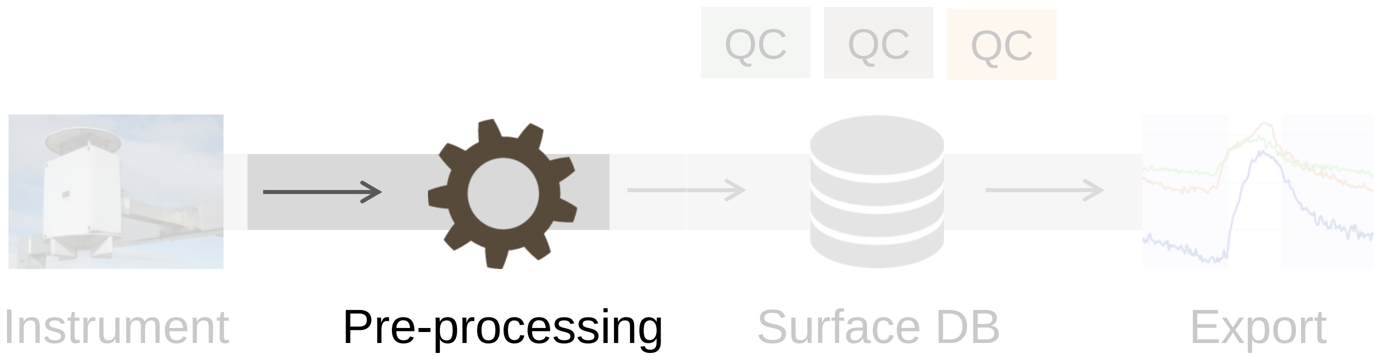


QC along data processing pipeline



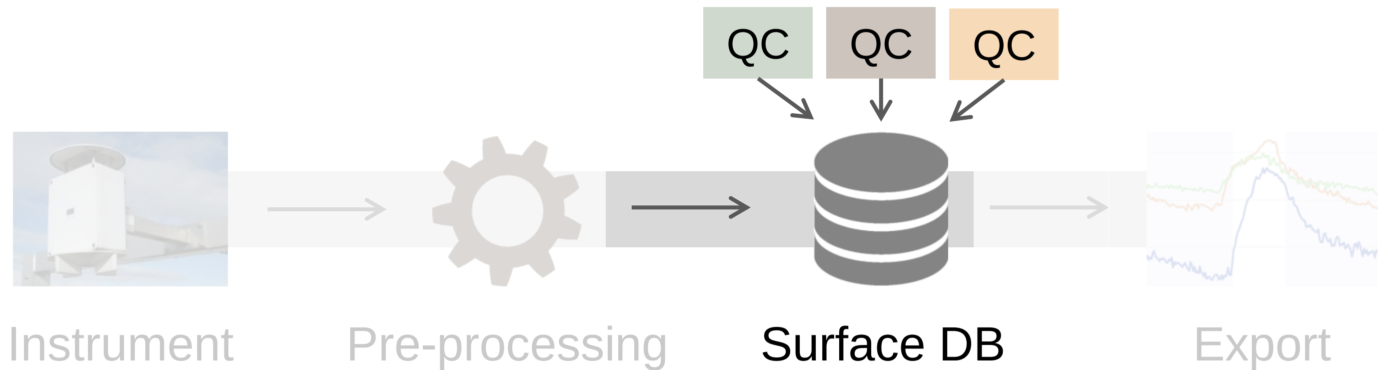
- “Smart” instruments: self-monitoring in the firmware
- QC status codes sent along with measurements

Pre-processing and import



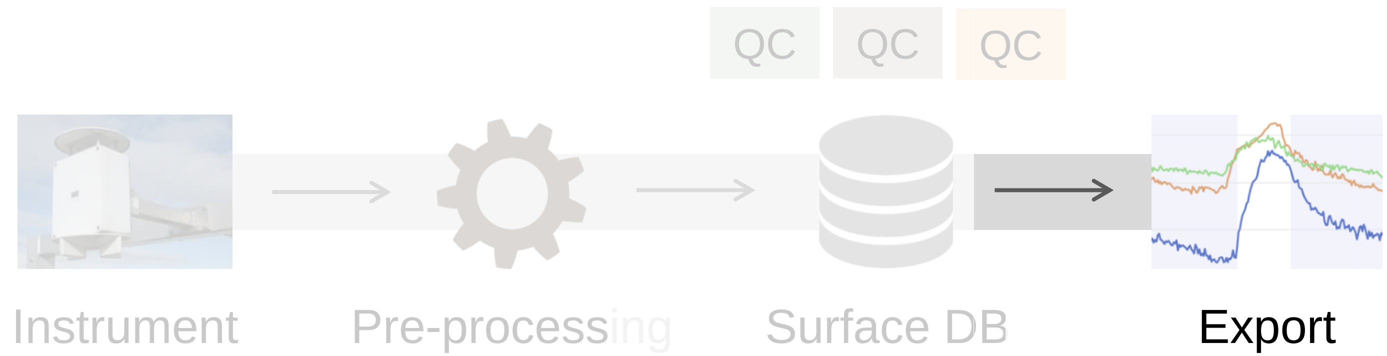
- Computation of derived quantities
- Real-time “hard” tests (e.g. physical limits): failed values are suppressed

Storage and QC



- Multiple independent QC systems, acting on single data representation
- Testing methods: climatological limits, extreme value rankings, spatio-temporal consistency, statistical models
- Testing frequencies: from hourly, daily to yearly

Export



- Filtering based on available QI
- Export of QI along measurement data

Outline

- QC along our data processing pipeline
- Our former QI
- Probabilistic plausibility
- Summarizing QI for the user
- Discussion

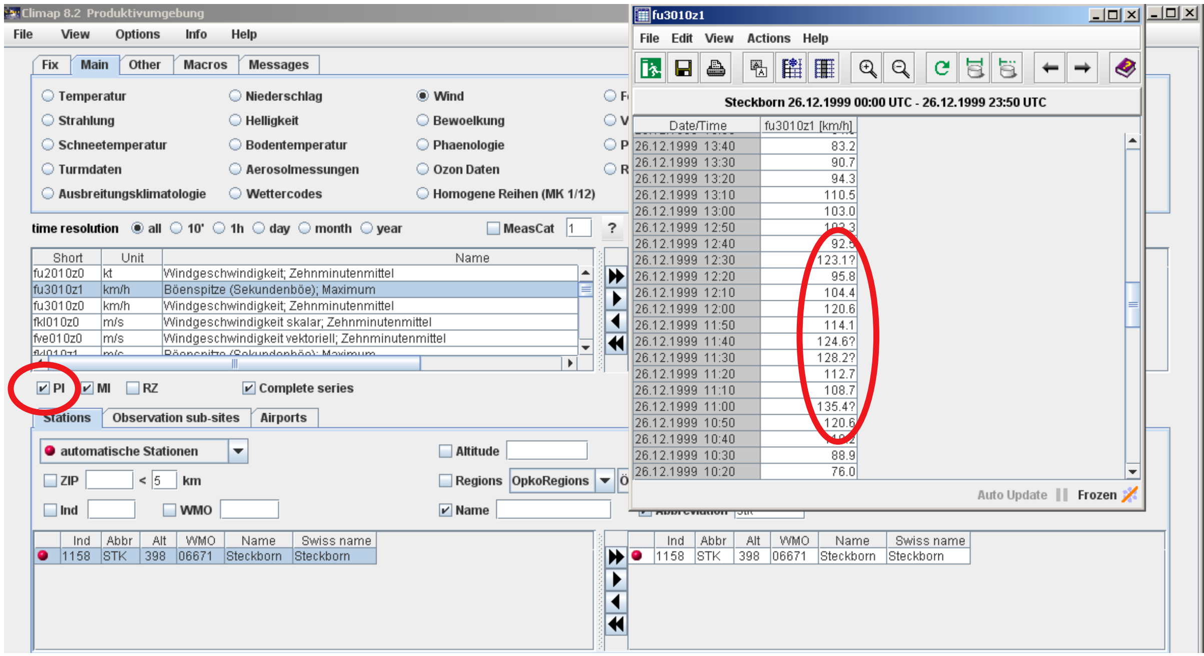
Former relational data model

Parameter	Instrument	Reference time	Value	P	C	I	S
rre150z0	11356	05.01.2019 23:20	-23	Y	N	N	N
rre150d0	9838	03.02.2019	5.5	N	N	N	Y
tre200s0	20324	01.02.2019 08:10	11.5	N	Y	Y	N

Non-extensible bit mask of categories:

- **P**hysical limit exceeded
- **C**limatological limit exceeded
- **I**nconsistent to another parameter
- **S**patially inconsistent

QI export



- Physically impossible values are suppressed
- Logical OR of plausibility bits optionally displayed as «?»

Discussion

+ Straightforward data model:

- Summarize test outcomes into categories
- Store in bit mask, right along measurement

– Categories combine test outcomes with different evidence-strength:

- Sensitivity and specificity varies greatly among tests
- Only tests from **P** category have a known evidence-strength

– Categorical QI is hard to integrate in customer application → categories beyond **P** have rarely been used

Outline

- QC along our data processing pipeline
- Our former QI
- **Probabilistic plausibility**
- Summarizing QI for the user
- Discussion

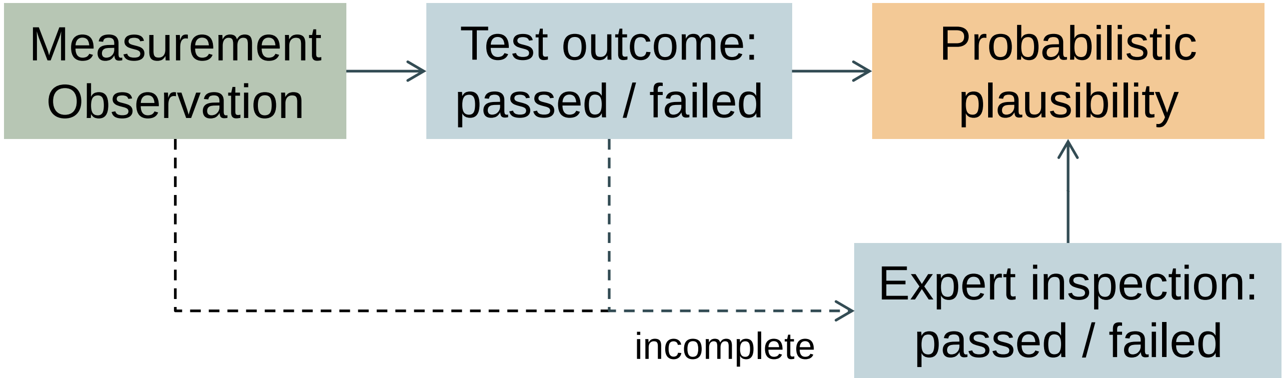
Definition of plausibility

A measurement is *plausible* if it is **confirmed** during expert inspection.

A measurement is *implausible* if it is **corrected or suppressed** during expert inspection.

- Expert treatment is the reference
- Expert inspection is incomplete: measurements are assumed to be plausible unless they are explicitly implausible

Probabilistic plausibility



1. Store test outcomes and expert inspections (both *failed* and *passed*)
2. Compute *probabilistic plausibility:* chance that measurement would pass expert inspection, given all test outcomes

Outcomes of automated tests

- Automated tests emulate expert inspection
- But they are incomplete and create false alarms:

	Measurement plausible	Measurement implausible
Test passed	True negative (TN)	False negative (FN)
Test failed	False positive (FP)	True positive (TP)

Goal: Test outcomes should contribute to plausibility according to the strength of their evidence.

Prior plausibility $p(q)$

Prior $p(q)$ → Test likelihood $p(t|q)$ → Posterior $p(q|t)$

Probabilistic plausibility **before** automated testing and inspection:

$$p(q = 1) = 1 - p(q = 0)$$

$q \in \{1,0\}$: measurement is *plausible* or *implausible*

Example: $p(q = 1) = 0.99$ corresponds to 1 in 100 chance that measurement would fail expert investigation.

Estimating the prior plausibility

Estimated by simple counting:

$$\hat{p}(q = 1) = 1 - \frac{|\mathcal{I}|}{|\mathcal{M}|}$$

$\mathcal{M}$: set of all tested measurements

$\mathcal{I} \subseteq \mathcal{M}$: implausible measurements

Subjective estimates are also possible in case of insufficient data.

Test likelihood $p(t|q)$

Prior $p(q)$ $\rightarrow$ $\boxed{\text{Test likelihood } p(t|q)}$ $\rightarrow$ Posterior $p(q|t)$

Likelihood of test outcome given the plausibility of the measurement:

$$p(t|q)$$

$t \in \{1,0\}$: test outcome *passed* or *failed*

- *Failed* outcomes **decrease** plausibility
- *Passed* outcomes **increase** plausibility

Probabilistic plausibility $p(q|t)$

Prior $p(q)$ $\rightarrow$ Test likelihood $p(t|q)$ $\rightarrow$ Posterior $p(q|t)$

Plausibility **after** automated testing and/or inspection:

$$p(q|t)$$

Posterior probability computed from *prior* and *test likelihood* using **Bayes' rule**:

$$p(q|t) \propto p(t|q)p(q)$$

Combining multiple test outcomes

Naive Bayes assumption: Test outcomes are conditionally independent

$$p(t_1, t_2|q) = p(t_1|q)p(t_2|q)$$

→ Update posterior plausibility whenever a new test outcome is available

Posterior $p(q|t_1)$ becomes prior for next update:

$$p(q) \rightarrow \boxed{\begin{array}{c}\text{Likelihood}\\ p(t_1|q)\end{array}} \rightarrow p(q|t_1) \rightarrow \boxed{\begin{array}{c}\text{Likelihood}\\ p(t_2|q)\end{array}} \rightarrow p(q|t_1, t_2) \rightarrow \ldots$$

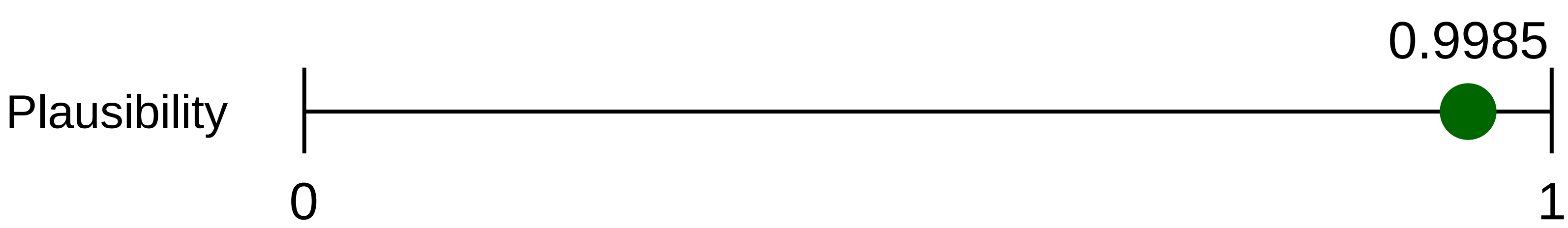
Calculation example

- Plausibility of automated air temperature measurements (2 m above ground)
- 10 min granularity → 144 measurements per day
- Probabilities estimated from one year of data (values rounded)

Estimated prior

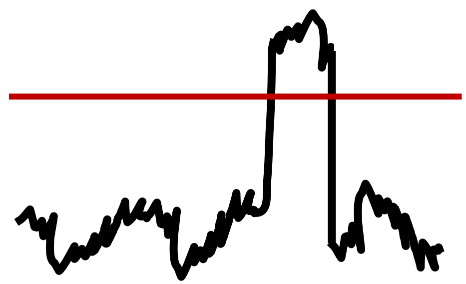
Prior $\hat{p}(q)$ → Test likelihood $\hat{p}(t_1|q)$ → Posterior $\hat{p}(q|t_1)$

About 1 in 670 measurements is implausible
→ 1 implausible measurement per 4.7 days per instrument.



Physical limit test

Prior $\hat{p}(q)$ → Test likelihood $\hat{p}(t_1|q)$ → Posterior $\hat{p}(q|t_1)$

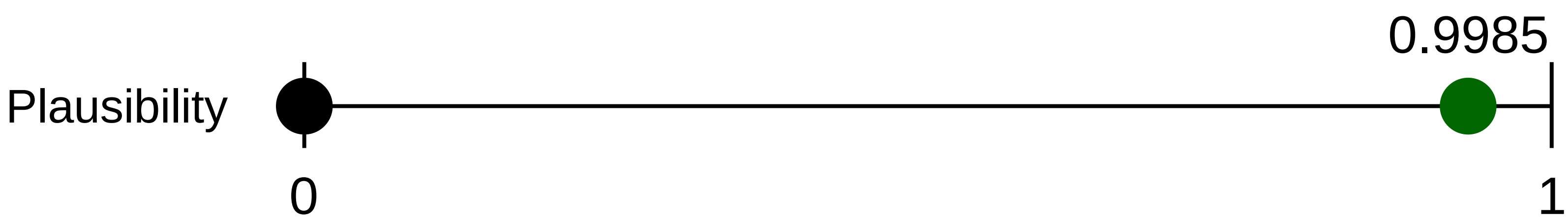


- By definition, test has 100 % specificity (no false positives)
- 22 % of all implausible values exceed physical limit

Posterior plausibility

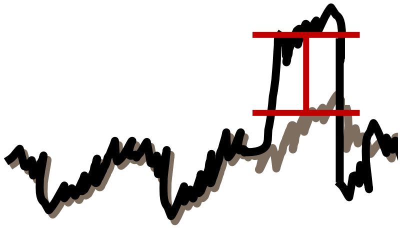
Prior $\hat{p}(q)$ → **Test outcome** $t_1 = 0$ → Posterior $\hat{p}(q|t_1 = 0)$

Measurement fails physical limit test, $t_1 = 0$:



Consistency test

Prior $\hat{p}(q)$ → Test likelihood $\hat{p}(t_2|q)$ → Posterior $\hat{p}(q|t_2)$



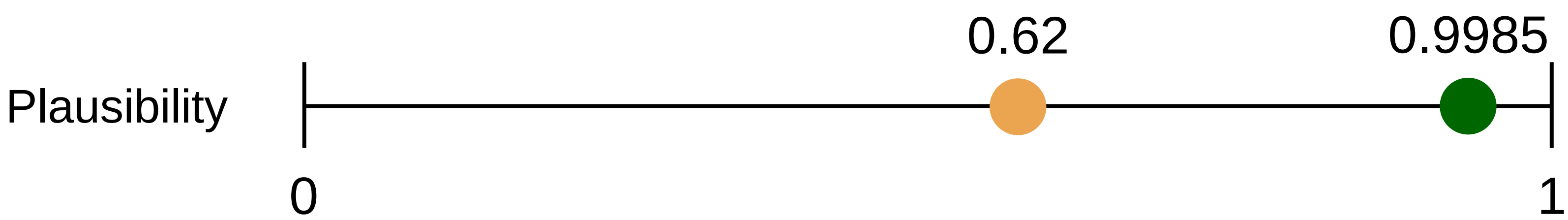
Limit of absolute difference to redundant measurement:

- 0.014 % false positive rate
- 5.7 % of implausible measurements fail consistency test

Estimated posterior

Prior $\hat{p}(q)$ → **Test outcome** $t_2 = 0$ → Posterior $\hat{p}(q|t_2 = 0)$

Measurement fails consistency test, $t_2 = 0$:



Combining test outcomes

$\hat{p}(q) \rightarrow$ Likelihood $\hat{p}(t_2|q)$ $\rightarrow \hat{p}(q|t_2) \rightarrow$ Likelihood $\hat{p}(t_3|q)$ $\rightarrow \hat{p}(q|t_2, t_3)$

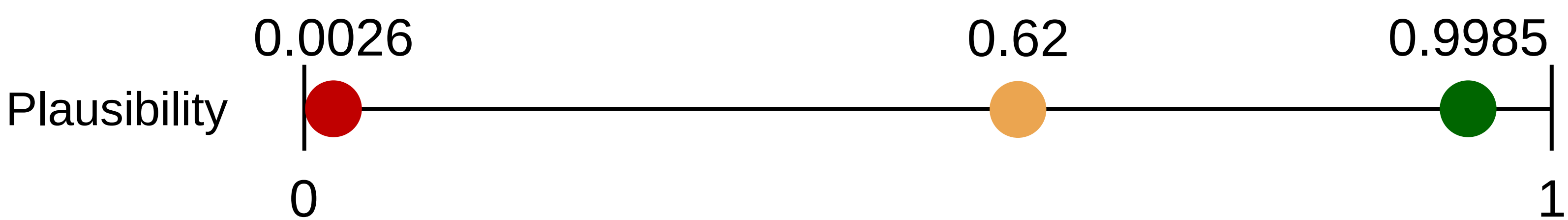
Minimum variability test:

- 0.0015 % false positive rate
- 1.2 % of implausible measurements fail minimum variability test

Combining test outcomes

$\hat{p}(q) \rightarrow$ Outcome $t_2 = 0$ $\rightarrow \hat{p}(q|t_2 = 0) \rightarrow$ Outcome $t_3 = 0$ $\rightarrow \hat{p}(q|t_2 = 0, t_3 = 0)$

Measurement fails both the consistency test, $t_2 = 0$, and the minimum variability test, $t_3 = 0$:



Expert inspection

$\hat{p}(q) \rightarrow$ $\boxed{\text{Likelihood } \hat{p}(t_2|q)}$ $\rightarrow \hat{p}(q|t_2) \rightarrow$ $\boxed{\text{Likelihood } p(t_e|q)}$ $\rightarrow p(q|t_2, t_e)$

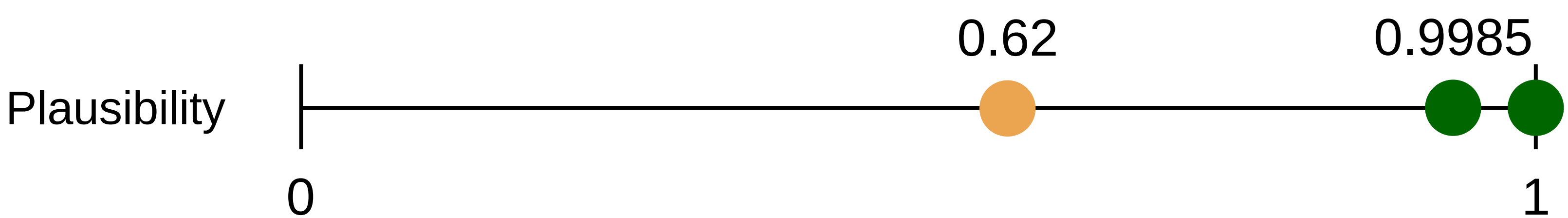
Model expert inspection as another test t_e with:

- 100 % specificity (no false positives)
- 100 % sensitivity (finds all implausible values)

Expert corrects false positive

$$\hat{p}(q) \rightarrow \boxed{\begin{array}{c}\text{Outcome}\\ t_2 = 0\end{array}} \rightarrow \hat{p}(q|t_2 = 0) \rightarrow \boxed{\begin{array}{c}\text{Outcome}\\ t_e = 1\end{array}} \rightarrow \underline{p(q|t_2 = 0, t_e = 1)}$$

Measurement fails the consistency test, $t_2 = 0$, but is confirmed by the expert, $t_e = 1$:



Discussion

+ Outcomes contribute according to their evidence:
  - Test outcomes increase or decrease plausibility
  - Accumulate weak evidence of several test outcomes into strong evidence
+ Combine outcomes from independent QC systems:
  - Incorporate new test outcomes whenever they arrive
  - Re-calculate plausibility using Naive Bayes

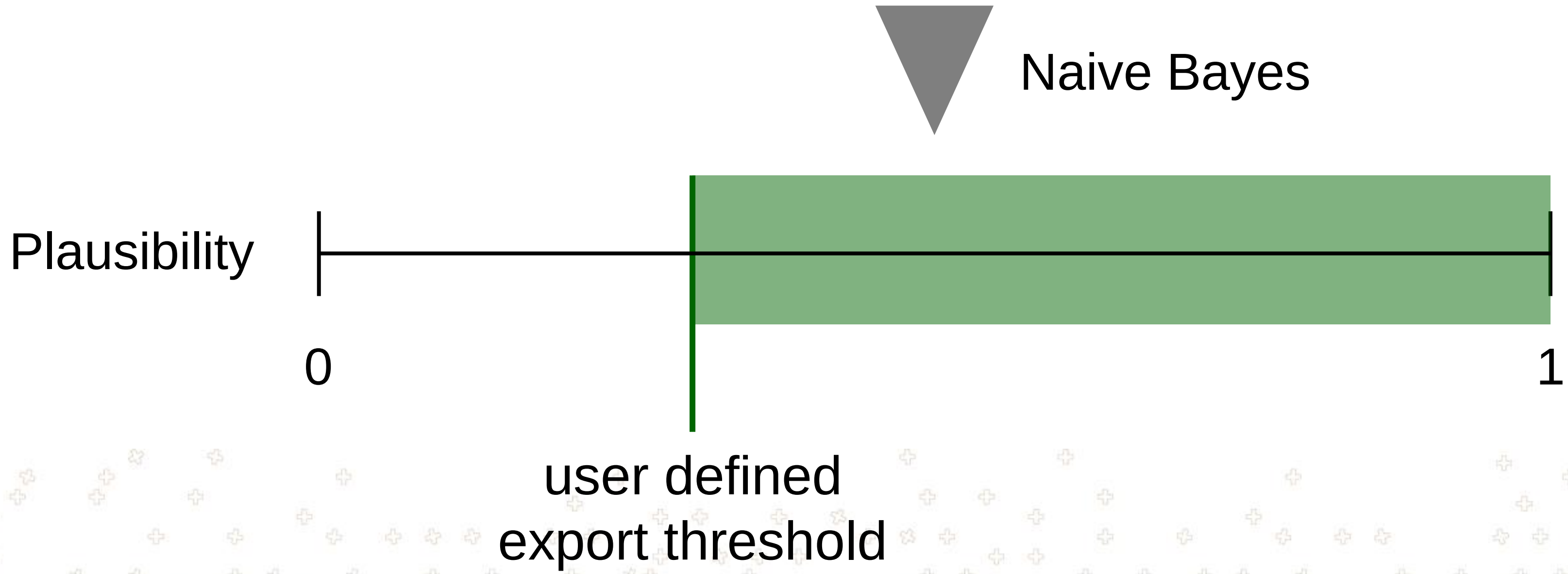
– Cannot store outcomes in fixed-length bitmask
– Computation necessary to obtain plausibility

Outline

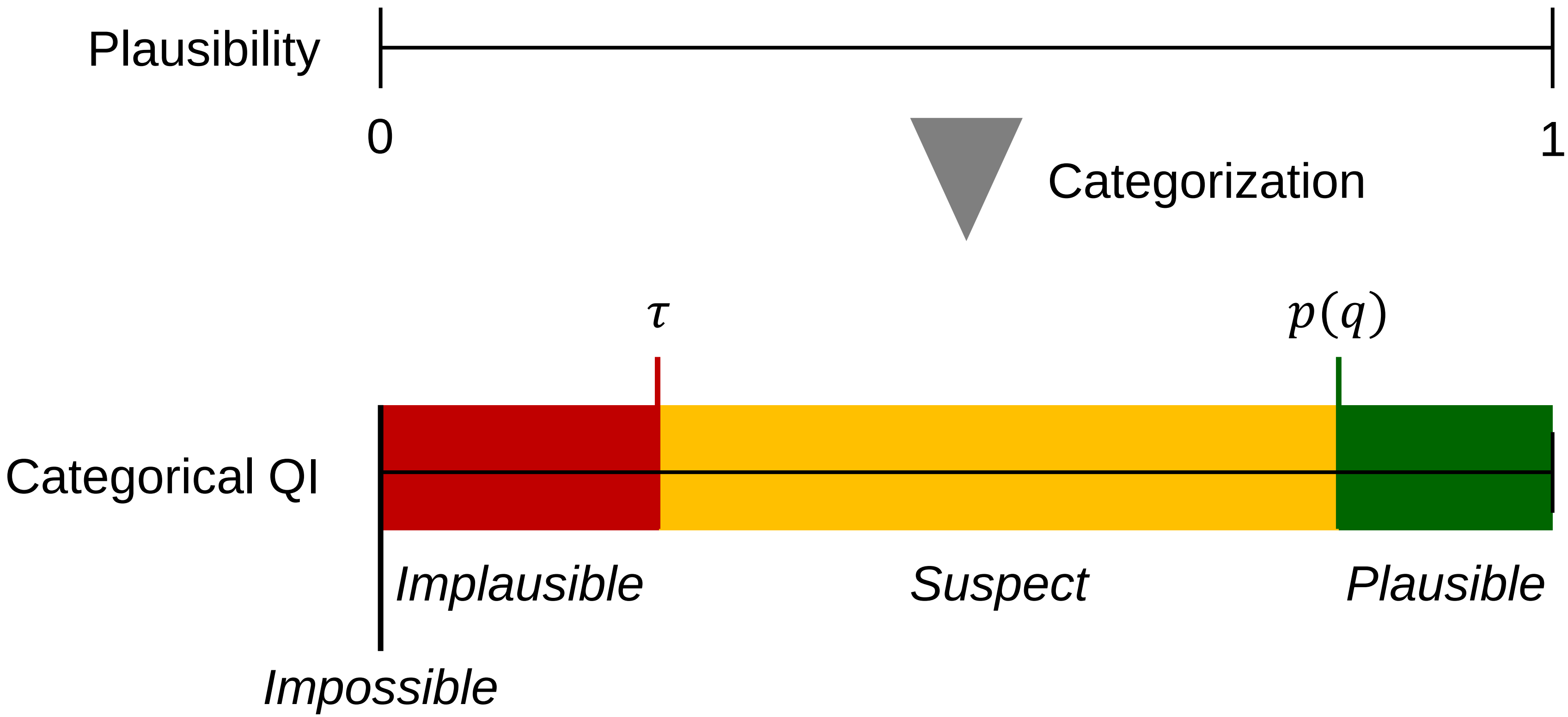
- QC along our data processing pipeline
- Our former QI
- Probabilistic plausibility
- Summarizing QI for the user
- Discussion

Quantitative QI summary

Measurement	Test	Passed
4614406274	8	N
4614406274	112	Y
4614406274	236	Y



Categorical QI summary



Implausible: strong evidence against measurement
→ e.g. automated substitution with interpolated value

Outline

- QC along our data processing pipeline
- Our former QI
- Probabilistic plausibility
- Summarizing QI for the user
- Discussion

Practical concerns

Probabilistic plausibility scales to size of our surface DB (currently 21 billion records)

Storage:

- Unknown or irrelevant test outcomes can be safely omitted (no effect on computation of posterior)

Computation:

- Posterior calculated by multiplication of few terms
- New tests and whole QC systems can be introduced without recomputing existing posterior probabilities

Practical concerns

Inference:

- Prior and test likelihoods estimated by simple counting of proportions
- Conditional independence assumption of Naive Bayes works well, even when it is not satisifed exactly

Summary

Probabilistic plausibility:

- Quantitative representation of data quality
- Combines prior information, multiple outcomes from automated tests and expert inspection
- Accumulates weak into strong evidence
- Derive simple categorical QI with well-defined meaning
- Efficient computation, scales to our surface DB

Experts are humans...

Likelihood $p(t_e|q)$ assumes that experts are perfect, but mistakes happen → probabilistic plausibility is recomputed whenever expert treatments change.

Alternative, given the necessary ressources:

1. Have multiple experts inspect the same data
2. Define plausibility using majority vote
3. Compute average expert likelihood $\hat{p}(t_{\bar{e}}|q)$

Estimated prior

Prior $\hat{p}(q)$ → Test outcome t_1 → Posterior $\hat{p}(q|t_1)$

About 1 in 670 measurements is implausible
→ 1 implausible measurement per 4.7 days per instrument.

Estimated **prior** $\hat{p}(q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
	0.9985	1.5E-3

Physical limit test

Prior $\hat{p}(q)$ → Test outcome t_1 → Posterior $\hat{p}(q|t_1)$

- By definition, test has 100 % specificity (no false alarms)
- 22 % of all implausible values exceed physical limit

Estimated **likelihood** $\hat{p}(t_1|q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_1 = 1$	1	0.78
Failed $t_1 = 0$	0	0.22

Estimated posterior

Prior $\hat{p}(q)$ → Test outcome t_1 → Posterior $\hat{p}(q|t_1)$

Estimated **posterior** $\hat{p}(q|t_1)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_1 = 1$	0.9988	1.2E-3
Failed $t_1 = 0$	0	1

Compared to estimated prior $\hat{p}(q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
	0.9985	1.5E-3

Consistency test

Limit of abs. difference to redundant measurement:

Likelihood $\hat{p}(t_2|q)$:

	Plausible $q=1$	**Implausible** $q=0$
Passed $t_2=1$	0.99986	0.943
Failed $t_2=0$	1.4E-4	0.057

- 0.014 % false positive rate
- 5.7 % of implausible measurements fail consistency test

Consistency test

Posterior $\hat{p}(q|t_2)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_2 = 1$	0.9986	1.4E-3
Failed $t_2 = 0$	0.62	0.38

Compared to prior $\hat{p}(q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
	0.9985	1.5E-3

Combining test outcomes

Likelihood of consistency test $\hat{p}(t_2|q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_2 = 1$	0.99986	0.943
Failed $t_2 = 0$	1.4E-4	0.057

Likelihood of minimum variability test $\hat{p}(t_3|q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_3 = 1$	0.999981	0.988
Failed $t_3 = 0$	1.9E-5	0.012

Combining test outcomes

Posterior $\hat{p}(q|t_2, t_3)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed, Passed	0.99862	1.38E-3
Failed, Failed	2.6E-3	0.9974

Compared to posterior $\hat{p}(q|t_2)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_2 = 1$	0.9986	1.4E-3
Failed $t_2 = 0$	0.62	0.38

Expert corrects false positive

Before expert inspection: $\hat{p}(q = 1|t_2 = 0) = 0.62$

Model expert inspection as test with **likelihood** $\hat{p}(t_e|q)$:

	Plausible $q = 1$	**Implausible** $q = 0$
Passed $t_e = 1$	1	0
Failed $t_e = 0$	0	1

After expert inspection:

Plausibility: $\hat{p}(q = 1|t_2 = 0, t_e = 1) = 1$