

Good practices for reproducible computer-aided research

Konrad Hinsen

Centre de Biophysique Moléculaire, Orléans, France
Synchrotron SOLEIL, Saint Aubin, France

12 June 2024

Practices

Wikipedia

A best practice is a method or technique that has been generally accepted as superior to alternatives because it tends to produce superior results. Best practices are used to achieve quality as an alternative to mandatory standards. Best practices can be based on self-assessment or benchmarking. Best practice is a feature of accredited management standards such as ISO 9000 and ISO 14001.

Practices

Wikipedia

A best practice is a method or technique that has been generally accepted as superior to alternatives because it tends to produce superior results. Best practices are used to achieve quality as an alternative to mandatory standards. Best practices can be based on self-assessment or benchmarking. Best practice is a feature of accredited management standards such as ISO 9000 and ISO 14001.

What I know and teach:

- Practices that work pretty well.

Practices

Wikipedia

A best practice is a method or technique that has been generally accepted as superior to alternatives because it tends to produce superior results. Best practices are used to achieve quality as an alternative to mandatory standards. Best practices can be based on self-assessment or benchmarking. Best practice is a feature of accredited management standards such as ISO 9000 and ISO 14001.

What I know and teach:

- Practices that work pretty well.
- Superior to alternatives? I don't know.

Practices

Wikipedia

A best practice is a method or technique that has been generally accepted as superior to alternatives because it tends to produce superior results. Best practices are used to achieve quality as an alternative to mandatory standards. Best practices can be based on self-assessment or benchmarking. Best practice is a feature of accredited management standards such as ISO 9000 and ISO 14001.

What I know and teach:

- Practices that work pretty well.
- Superior to alternatives? I don't know.
- **Good** rather than **best** practices.

Practices

Wikipedia

A best practice is a method or technique that has been generally accepted as superior to alternatives because it tends to produce superior results. Best practices are used to achieve quality as an alternative to mandatory standards. Best practices can be based on self-assessment or benchmarking. Best practice is a feature of accredited management standards such as ISO 9000 and ISO 14001.

What I know and teach:

- Practices that work pretty well.
- Superior to alternatives? I don't know.
- **Good** rather than **best** practices.
- Mostly *not* **good enough** practices.

Science vs. tech

Science

- Science = Curiosity + Critical Thinking
- Requires epistemic humility

Science vs. tech

Science

- Science = Curiosity + Critical Thinking
- Requires epistemic humility

Tech

- Marketable products
- Rapid change, fashion, tech churn

Science vs. tech

Science

- Science = Curiosity + Critical Thinking
- Requires epistemic humility

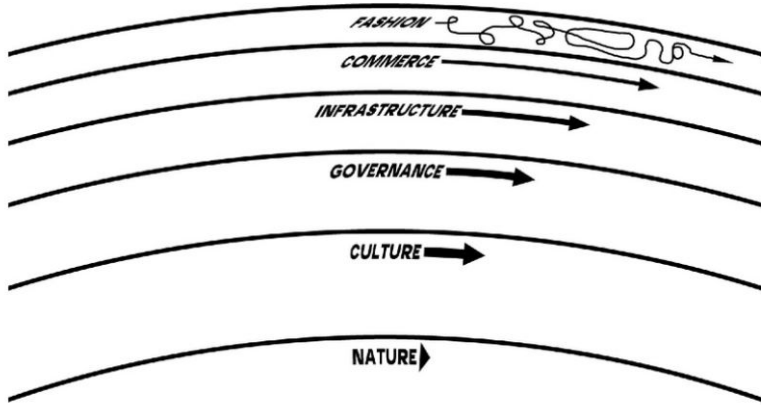
Tech

- Marketable products
- Rapid change, fashion, tech churn

“This is the original sin of software dev: it’s a pop culture where we’re trained to accept gossip as evidence.”

Baldur Bjarnason, “Trusting your own judgement on ‘AI’ is a huge risk”

Pace layers



The order of a healthy civilization. The fast layers innovate; the slow layers stabilize. The whole combines learning with continuity.

Stewart Brand, "Pace Layering: How Complex Systems Learn and Keep Learning"

nature

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

[nature](#) > [technology features](#) > article

TECHNOLOGY FEATURE | 01 May 2023

The sleight-of-hand trick that can simplify scientific computing

Computational environments and the tools to manage them can help researchers to deliver code that is reproducible, documented and shareable.

[Jeffrey M. Perkel](#)



Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research

Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research
 - Three sessions since 2018

Good practices for reproducible research

Two MOOCs on reproducible computational science:

① **Reproducible research: methodological principles for transparent science**

- Beginner level, for anyone doing research
- Three sessions since 2018
- Currently: long-term session until December 2025

Good practices for reproducible research

Two MOOCs on reproducible computational science:

① Reproducible research: methodological principles for transparent science

- Beginner level, for anyone doing research
- Three sessions since 2018
- Currently: long-term session until December 2025
- Sign up whenever you want, progress at your own pace

Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research
 - Three sessions since 2018
 - Currently: long-term session until December 2025
 - Sign up whenever you want, progress at your own pace
- ② **Reproducible Research II: Practices and tools for managing computations and data**
 - Advanced level, for computational science

Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research
 - Three sessions since 2018
 - Currently: long-term session until December 2025
 - Sign up whenever you want, progress at your own pace
- ② **Reproducible Research II: Practices and tools for managing computations and data**
 - Advanced level, for computational science
 - First session in 2024

Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research
 - Three sessions since 2018
 - Currently: long-term session until December 2025
 - Sign up whenever you want, progress at your own pace

- ② **Reproducible Research II: Practices and tools for managing computations and data**
 - Advanced level, for computational science
 - First session in 2024
 - Second session open *now*, until September 10

Good practices for reproducible research

Two MOOCs on reproducible computational science:

- ① **Reproducible research: methodological principles for transparent science**
 - Beginner level, for anyone doing research
 - Three sessions since 2018
 - Currently: long-term session until December 2025
 - Sign up whenever you want, progress at your own pace

- ② **Reproducible Research II: Practices and tools for managing computations and data**
 - Advanced level, for computational science
 - First session in 2024
 - Second session open *now*, until September 10

<https://www.fun-mooc.fr>

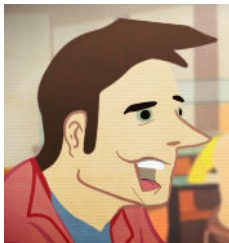
Alice and Bob meet at a conference



MOOC “Reproducible Research II: Practices and tools for managing computations and data”

The challenge

I have computed the equilibrium distance between the ligand and the active site of our pet protein. **It's 0,9 nm.**



I have computed the same distance, but **I find 1,1 nm.**

Investigating

Uhhh... Well... I will look at your code,
and you look at mine. Let's meet again
tomorrow.



OK!

If you don't want to share your code...

PHYSICS TODAY

HOME

BROWSE▼

INFO▼

RESOURCES▼

JOBS

DOI:10.1063/PT.6.1.20180822a

22 Aug 2018 in Research & Technology

The war over supercooled water

How a hidden ~~coding error~~ fueled a seven-year dispute between two of condensed matter's top theorists.

Ashley G. Smart

A.G. Smart, Physics Today, 2018

Next day

I couldn't compile your code. Look at this error message!



It works for me! You use Debian 12? I still run Debian 9.
The good news: I managed to run your code. But I get **0,8 nm**.

I use libode version 3.4. The documentation says it must be compiled with gcc 10 or later. You probably have an older gcc.



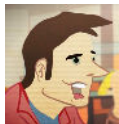
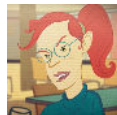
Uhhh... Well... I will install a virtual machine with Debian 12, and you with Debian 9. Shall we meet again in a week?

OK!



A week later

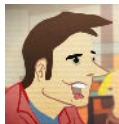
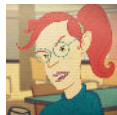
Under Debian 9, I managed to run your code. I get 1,1 nm, like you do. But I don't understand why! **Your code is unreadable.**



Under Debian 12, **your code yields 0.85 nm for me.** That's not your value of 0,9 nm. Nor 1,1 nm as I get using my method. I don't understand why!

A week later

Under Debian 9, I managed to run your code. I get 1,1 nm, like you do. But I don't understand why! **Your code is unreadable.**



Under Debian 12, **your code yields 0.85 nm for me.** That's not your value of 0,9 nm. Nor 1,1 nm as I get using my method. I don't understand why!

How can Alice and Bob proceed?

Computational rep. . . bility

Reproducibility

obtaining *identical* results using the same input data, computational steps, methods, code, etc.

Quality control: checks for complete documentation of a calculation.

Computational rep...bility

Reproducibility

obtaining *identical* results using the same input data, computational steps, methods, code, etc.

Quality control: checks for complete documentation of a calculation.

Replicability

obtaining *consistent* results across studies aimed at answering the same scientific question.

Scientific validation: checks for robustness of scientific methods.

Computational rep...bility

Reproducibility

obtaining *identical* results using the same input data, computational steps, methods, code, etc.

Quality control: checks for complete documentation of a calculation.

Replicability

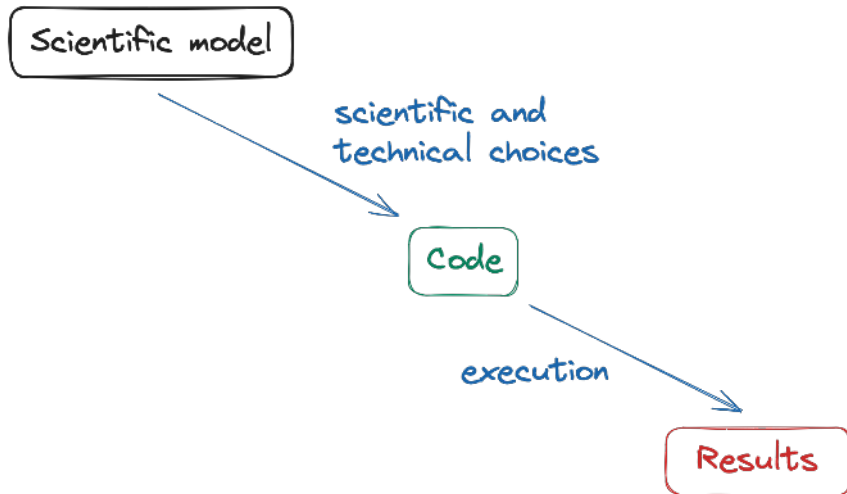
obtaining *consistent* results across studies aimed at answering the same scientific question.

Scientific validation: checks for robustness of scientific methods.

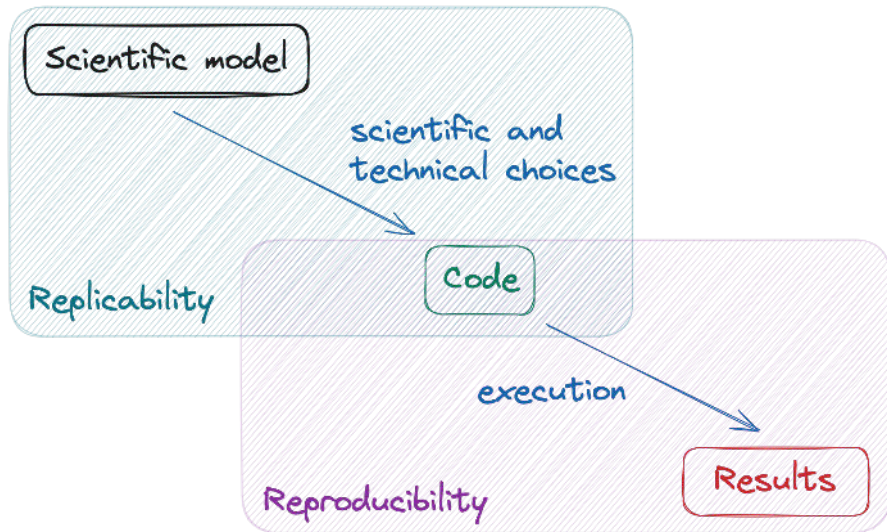
Alice and Bob

- **Neither** Alice nor Bob **can replicate** the other's result: $0,9\text{nm} \neq 1,1\text{nm}$
- Alice **can reproduce** Bob's result: $1,1\text{nm}$
- Bob **cannot reproduce** Alice's result: $0,85\text{nm} \neq 0,9\text{nm}$

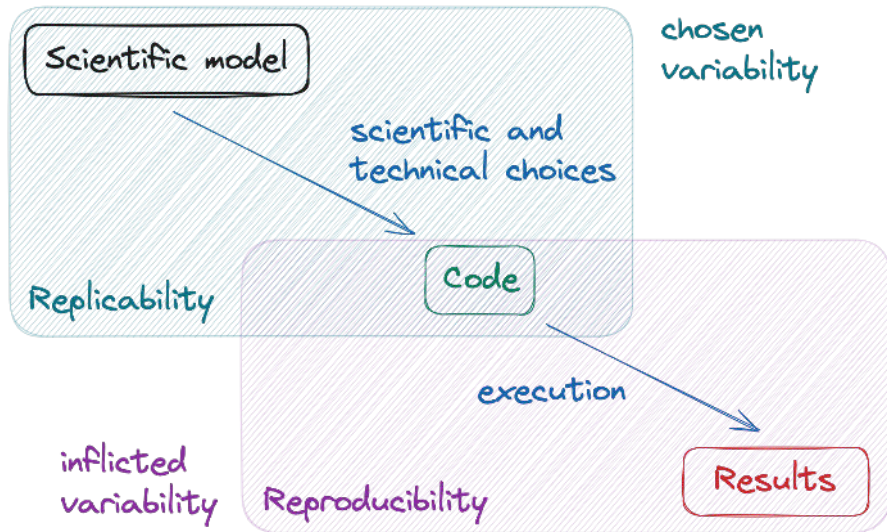
Model, code, results



Model, code, results



Model, code, results



Reproducibility

- Archive all the ingredients of a computation.

How to rep. . .

Reproducibility

- Archive all the ingredients of a computation.

Replicability

- Document all steps in sufficient detail for a human reader.
- Make each step inspectable and modifiable.

How to rep. . .

Reproducibility

- Archive all the ingredients of a computation.

Replicability

- Document all steps in sufficient detail for a human reader.
- Make each step inspectable and modifiable.

Obstacles

We adopt tools and practices from the software industry,
which doesn't care about reproducibility or replicability.

Reproducibility in theory

Observational data

- Record with provenance metadata
- Archive in a repository (Zenodo, ...)
- Assign a unique and stable identifier

Human choices: source code, parameters, ...

Computed results

Reproducibility in theory

Observational data

Human choices: source code, parameters, ...

- Evolution managed by version control
- Archive on Software Heritage
- Compute an intrinsic identifier

Computed results

Reproducibility in theory

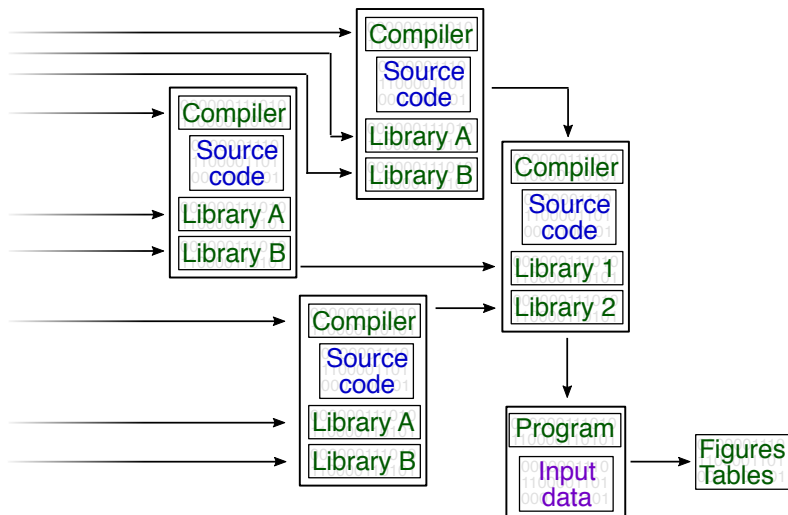
Observational data

Human choices: source code, parameters, ...

Computed results

- Record and archive
 - the identifiers of the inputs (including the code)
 - a computed intrinsic identifier
 - an identifier for the hardware
- *Ensure reproducibility of inputs*

Reproducible data



What kind of data is this?



Protein Data Bank in Europe
Bringing Structure to Biology



Examples: [hemoglobin](#), [B1](#)

PDBe > 1iee

STRUCTURE OF TETRAGONAL HEN EGG WHITE LYSOZYME AT 0.94 Å FROM CRYSTALS GROWN BY THE COUNTER-DIFFUSION METHOD

Source organism: *Gallus gallus*

Primary publication:
 [Structure of tetragonal hen egg-white lysozyme at 0.94 Å from crystals grown by the counter-diffusion method.](#)

[Sauter C, Otálora F, Gavira JA, Vidal O, Glegé R, García-Ruiz JM](#)
Acta Crystallogr D Biol Crystallogr **57** 1119-26 (2001)
PMID: 11468395 

X-ray diffraction
0.94Å resolution

Released: 08 Aug 2001
DOI: [10.2210/pdb1iee/pdb](#)

Model geometry 
Fit model/data 



PDB entry 1IEE

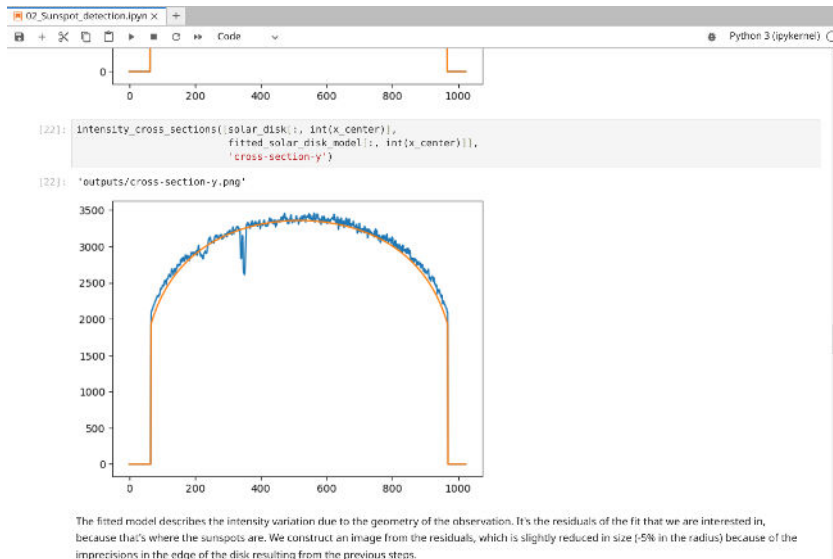
What kind of data is this?

```
data_1IEE
#
_entry.id 1IEE
#
loop_
_citation.id
_citation.title
_citation.journal_abbrev
_citation.journal_volume
_citation.page_first
_citation.page_last
_citation.year
_citation.journal_id_ASTM
_citation.country
_citation.journal_id_ISSN
_citation.journal_id_CSD
_citation.book_publisher
_citation.pdbx_database_id_PubMed
_citation.pdbx_database_id_DOI
primary 'Structure of tetragonal hen egg-white lysozyme at 0.94 Å from crystals grown by the counter-diffusion method.' 'Acta Crystallogr
1 'A SUPERSATURATION WAVE OF PROTEIN CRYSTALLIZATION' 'To be published' ? ? ? ? ? ? 0353 ? ? ?
```

What kind of data is this?

```
ATOM 1 N N . LYS A 1 1 ? 1.982 9.243 10.078 1.00 15.75 ? 1 LYS A N 1 1 UNP P00698 19 K
ATOM 2 C CA . LYS A 1 1 ? 1.020 9.618 9.045 1.00 14.41 ? 1 LYS A CA 1 1 UNP P00698 19 K
ATOM 3 C C . LYS A 1 1 ? 1.178 11.114 8.817 1.00 12.96 ? 1 LYS A C 1 1 UNP P00698 19 K
ATOM 4 O O . LYS A 1 1 ? 1.187 11.831 9.821 1.00 12.59 ? 1 LYS A O 1 1 UNP P00698 19 K
ATOM 5 C CB . LYS A 1 1 ? -0.406 9.250 9.429 1.00 15.03 ? 1 LYS A CB 1 1 UNP P00698 19 K
ATOM 6 C CG . LYS A 1 1 ? -1.474 9.753 8.488 1.00 15.51 ? 1 LYS A CG 1 1 UNP P00698 19 K
ATOM 7 C CD . LYS A 1 1 ? -2.850 9.311 8.916 1.00 16.30 ? 1 LYS A CD 1 1 UNP P00698 19 K
ATOM 8 C CE A LYS A 1 1 ? -3.891 9.779 7.917 0.30 17.64 ? 1 LYS A CE 1 1 UNP P00698 19 K
ATOM 9 C CE B LYS A 1 1 ? -4.056 9.849 8.212 0.70 17.10 ? 1 LYS A CE 1 1 UNP P00698 19 K
ATOM 10 N NZ A LYS A 1 1 ? -5.176 9.075 8.187 0.30 18.64 ? 1 LYS A NZ 1 1 UNP P00698 19 K
ATOM 11 N NZ B LYS A 1 1 ? -4.104 9.497 6.764 0.70 20.94 ? 1 LYS A NZ 1 1 UNP P00698 19 K
ATOM 12 N N . VAL A 1 2 ? 1.274 11.553 7.578 1.00 13.42 ? 2 VAL A N 1 2 UNP P00698 20 V
ATOM 13 C CA . VAL A 1 2 ? 1.236 12.984 7.236 1.00 12.17 ? 2 VAL A CA 1 2 UNP P00698 20 V
ATOM 14 C C . VAL A 1 2 ? -0.191 13.287 6.793 1.00 11.61 ? 2 VAL A C 1 2 UNP P00698 20 V
ATOM 15 O O . VAL A 1 2 ? -0.637 12.897 5.710 1.00 14.60 ? 2 VAL A O 1 2 UNP P00698 20 V
ATOM 16 C CB . VAL A 1 2 ? 2.253 13.384 6.203 1.00 14.64 ? 2 VAL A CB 1 2 UNP P00698 20 V
ATOM 17 C CG1 . VAL A 1 2 ? 2.130 14.888 5.878 1.00 14.71 ? 2 VAL A CG1 1 2 UNP P00698 20 V
ATOM 18 C CG2 . VAL A 1 2 ? 3.658 13.095 6.678 1.00 15.27 ? 2 VAL A CG2 1 2 UNP P00698 20 V
ATOM 19 N N . PHE A 1 3 ? -0.975 13.933 7.653 1.00 10.72 ? 3 PHE A N 1 3 UNP P00698 21 F
ATOM 20 C CA . PHE A 1 3 ? -2.341 14.317 7.330 1.00 10.11 ? 3 PHE A CA 1 3 UNP P00698 21 F
ATOM 21 C C . PHE A 1 3 ? -2.382 15.387 6.266 1.00 10.93 ? 3 PHE A C 1 3 UNP P00698 21 F
ATOM 22 O O . PHE A 1 3 ? -1.554 16.281 6.242 1.00 11.20 ? 3 PHE A O 1 3 UNP P00698 21 F
ATOM 23 C CB . PHE A 1 3 ? -3.055 14.861 8.573 1.00 10.47 ? 3 PHE A CB 1 3 UNP P00698 21 F
ATOM 24 C CG . PHE A 1 3 ? -3.664 13.785 9.465 1.00 10.38 ? 3 PHE A CG 1 3 UNP P00698 21 F
ATOM 25 C CD1 . PHE A 1 3 ? -2.926 13.039 10.362 1.00 11.55 ? 3 PHE A CD1 1 3 UNP P00698 21
```

What kind of data is this?



Today's computational infrastructure

Identifiers

- unstable: file names, URLs, ...
- imprecise: version numbers

Provenance tracking

- not supported

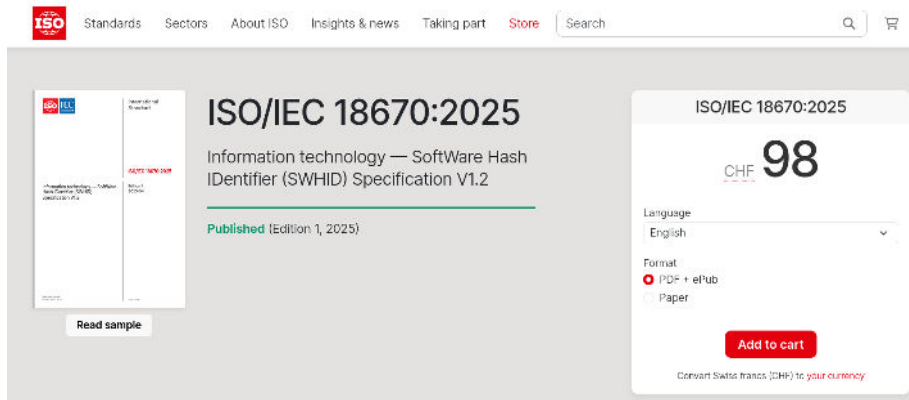
Documentation

- separate from code

Intrinsic identifiers

- **Computed** from the data, rather than **assigned** to it
- Example: checksum
- Well-studied problem in cryptography
- Used by git, Nix/Guix, Software Heritage, cryptocurrencies, ...

Software Hash Identifier



The screenshot shows the ISO website's product page for ISO/IEC 18670:2025. The top navigation bar includes links for Standards, Sectors, About ISO, Insights & news, Taking part, Store, and a search bar. The main content area features a thumbnail of the standard's cover on the left, which includes the ISO and IEC logos and the text 'Information technology — Software Hash Identifier (SWHID) Specification V1.2'. To the right of the thumbnail is a large heading 'ISO/IEC 18670:2025' followed by the subtitle 'Information technology — SoftWare Hash Identifier (SWHID) Specification V1.2'. Below this, it states 'Published (Edition 1, 2025)'. On the far right, there is a pricing section showing 'ISO/IEC 18670:2025' with a price tag of 'CHF 98'. Below the price, there are dropdown menus for 'Language' (set to English) and 'Format' (with radio buttons for 'PDF + ePub' and 'Paper'). A red 'Add to cart' button is positioned below the format options. At the bottom of the pricing section, a note says 'Convert Swiss francs (CHF) to your currency'.

Read it for free at: <https://www.swhid.org/>

Replicability in theory

No simple recipes!

- Make your code readable
- Make your code durable
- Provide documentation
- Provide examples and test cases

Replicability in theory

No simple recipes!

- Make your code readable
- Make your code durable
- Provide documentation
- Provide examples and test cases

Good but not good enough technology:

- Literate programming (1984)
- Notebooks (1988)
- My own: [HyperDoc](#) (2025)

Take-home messages

- Today's best practices aren't good enough.
- Adopting tech practices isn't always good for science.
- Short term: Learn to live with the tools and practices we have.
- For the details: [follow the MOOC!](#)
- Long term: design, develop, and maintain **digital infrastructure for science.**