Opportunities and Risks of MDSE: experience with Derric, a DSL for Digital Forensics

Tijs van der Storm

@tvdstorm / storm@cwi.nl

CWI

Centrum Wiskunde & Informatica

MDSE Promise

MDSE Reality?

Tesla Motors' Devastating Design Problem

http://jalopnik.com/5887265/tesla-motors-devastating-design-problem

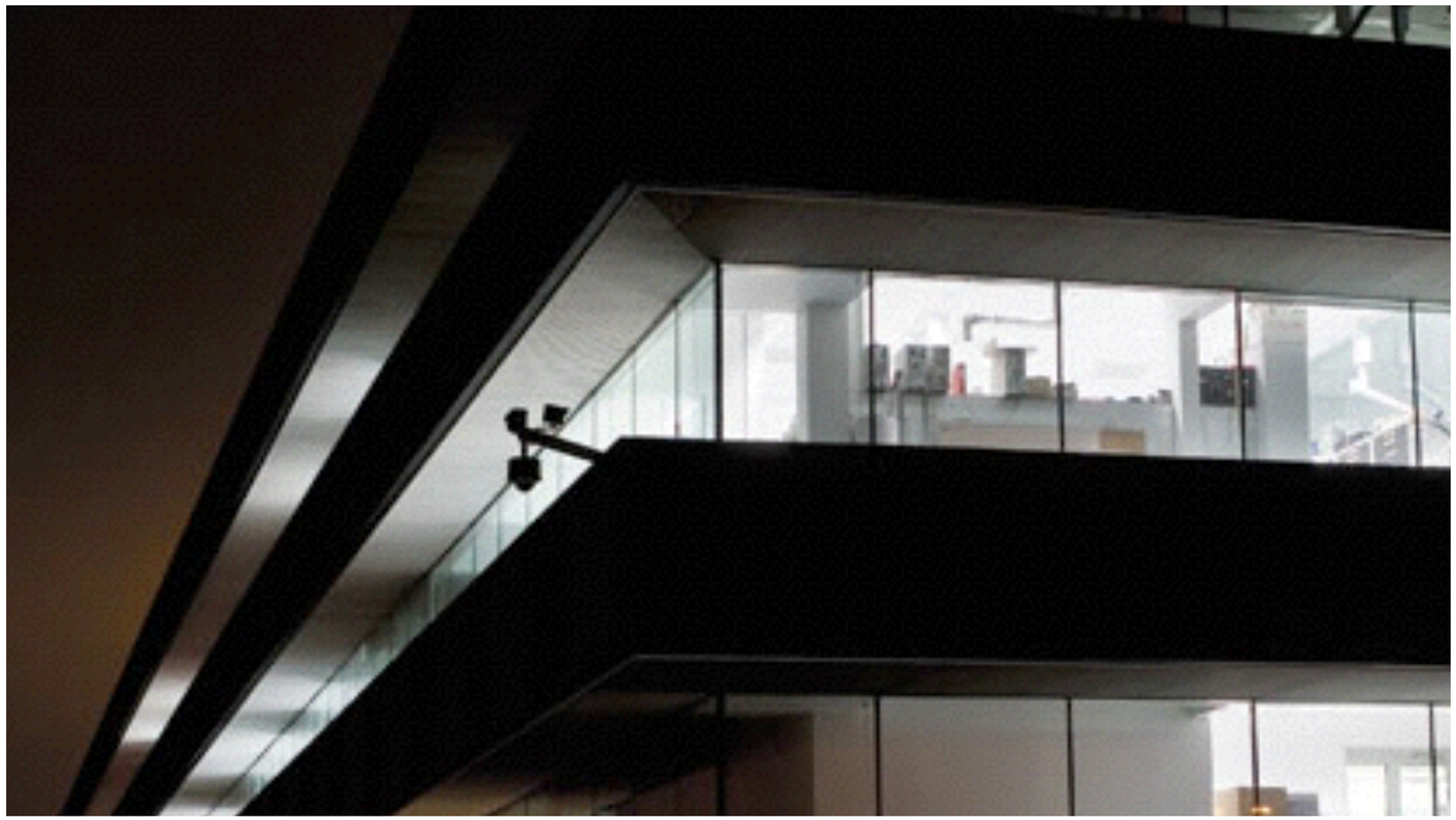
About us: SWAT

- Me: Tijs van der Storm
- SoftWare Analysis and Transformation
- Centrum Wiskunde & Informatica (CWI)
  - Rascal Meta Programming Language
    - http://www.rascal-mpl.org
  - => Language workbench for DSLs

CWI
Centrum Wiskunde & Informatica

Digital forensics

- From Wikipedia:
- "Digital forensics is a branch of forensic science encompassing the recovery and investigation of material found in digital devices, …"



Netherlands Forensic Institute ("CSI Holland")

Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie

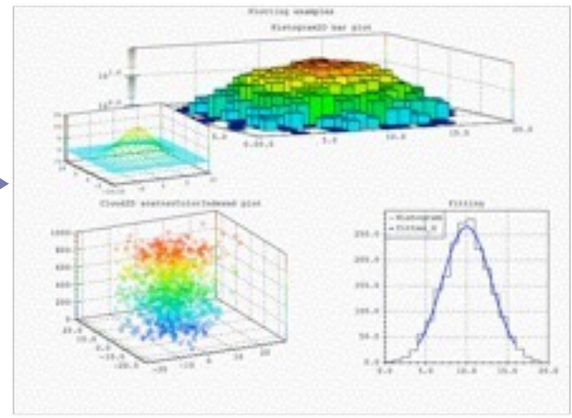
Typical process

Acquisition

Securing the data

Recovery

Turning data into information



Analysis

Finding relevant information

Problems in forensic data recovery

- High variability in devices, file systems etc.
- Off-the-shelf software inadequate
- Frequent, just-in-time customization
- Data in tera-byte range

Derric

- Key insight: it's all about file formats
- Derric is a DSL for binary file format description
- Implemented in Rascal

http://www.derric-lang.org http://www.rascal-mpl.org

```
format jpeg
extension jpeg jpg jfif

unit byte
size 1
sign false
type integer
endian big
strings ascii

sequence
  SOI
  ([APP0JFIF APP0JFXX?] [APP1 APP2?])
  !(SOI APP0JFIF APP0JFXX EOI)*
  EOI

structures
SOI {
  marker: 0xFF, 0xD8;
}

APP0JFIF {
  marker: 0xFF, 0xE0;
  length: lengthOf(rgb) + (offset(rgb) -
              offset(length)) size 2;
  identifier: "JFIF", 0;
  version: size 2;
  units: 0 | 1 | 2;
  xdensity: size 2;
  ydensity: size 2;
  xthumbnail;
  ythumbnail;
  rgb: size xthumbnail * (ythumbnail * 3);
}
```

```
APP0JFXX {
  marker: 0xFF, 0xE0;
  length: size 2;
  identifier: "JFXX", 0;
  thumbnailformat: 0x10 | 0x11 | 0x13;
  thumbnaildata: size length - (offset(thumbnaildata)
                  - offset(length));
}

Segment {
  marker: 0xFF;
  identifier: 0xD0..0xD7 | 0xDB..0xDC | 0xDF | 0xF0..0xFD;
  length: size 2;
  data: size length-(lengthOf(length));
}

SOS = Segment {
  identifier: 0xDA;
  compressedData:
     jpegdata(huffmantable=DHT.data,
              quantizationtable=DQT.data,
              terminator=0xFFD9+0xFFC4+0xFFDA,
              terminatorsize=16,
              includeterminator="false");
}

APP1 = Segment { identifier: 0xE1; }
APP2 = Segment { identifier: 0xE2; }
DQT = Segment { identifier: 0xDE; }
DHT = Segment { identifier: 0xC4; }
DRI = Segment { identifier: 0xDD; }
SOF0 = Segment { identifier: 0xC0; }
SOF2 = Segment { identifier: 0xC2; }
APPX = Segment { identifier: 0xE0..0xEF; }
COM = Segment { identifier: 0xFE; }

EOI { marker: 0xFF, 0xD9; }
```

- Opportunities
  - raise level of abstraction
  - separation of concerns
  - domain-specific optimization
- Risks
  - it's not better (functional/non-functional)
  - short-term domain analysis (myopia)
  - over-engineering / accidental complexity

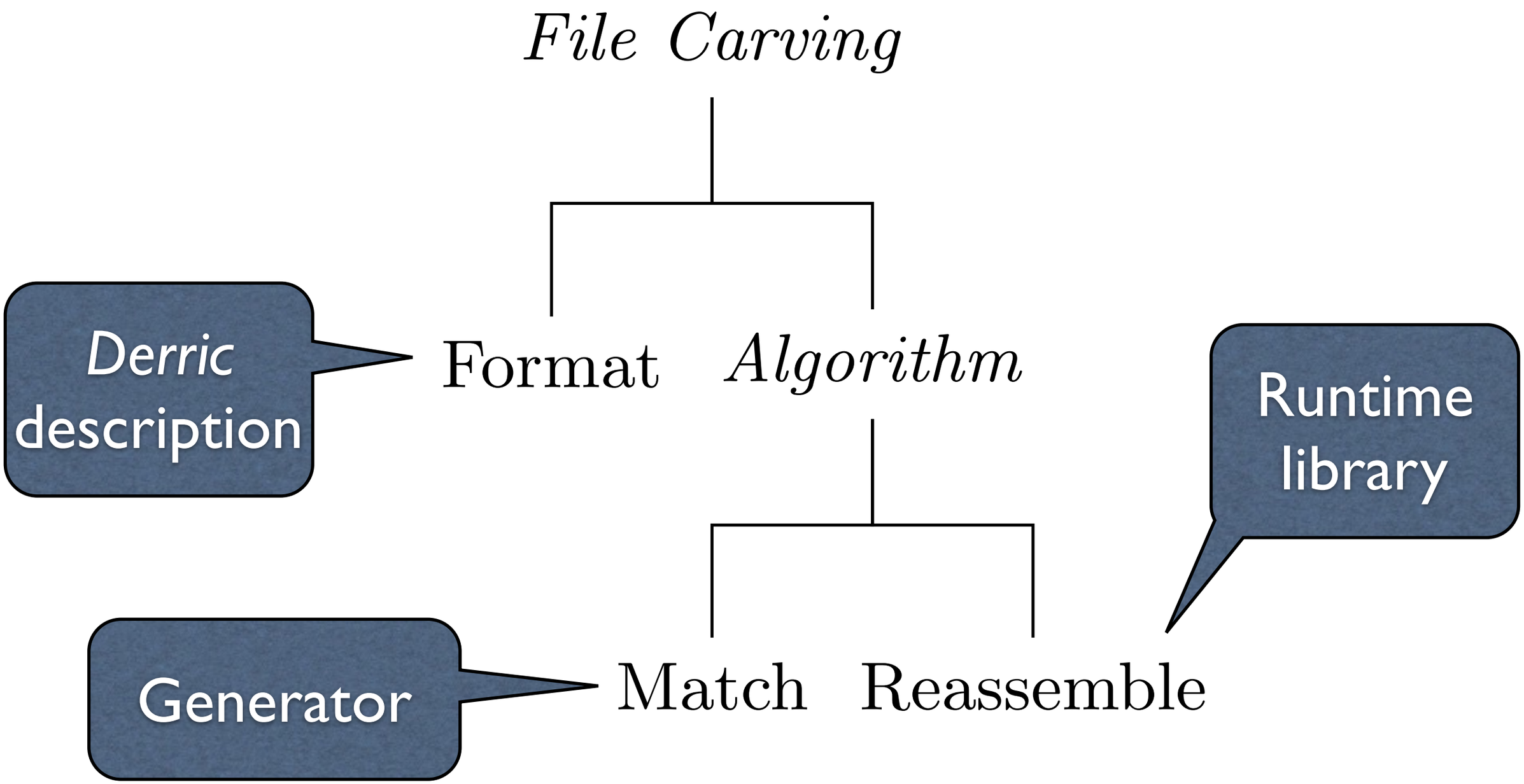
How we exploited them

How we mitigated them

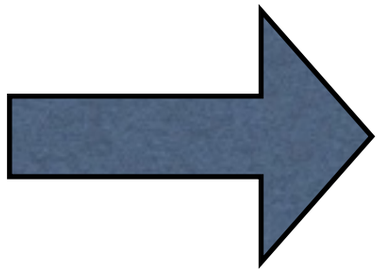
Raise level of abstraction

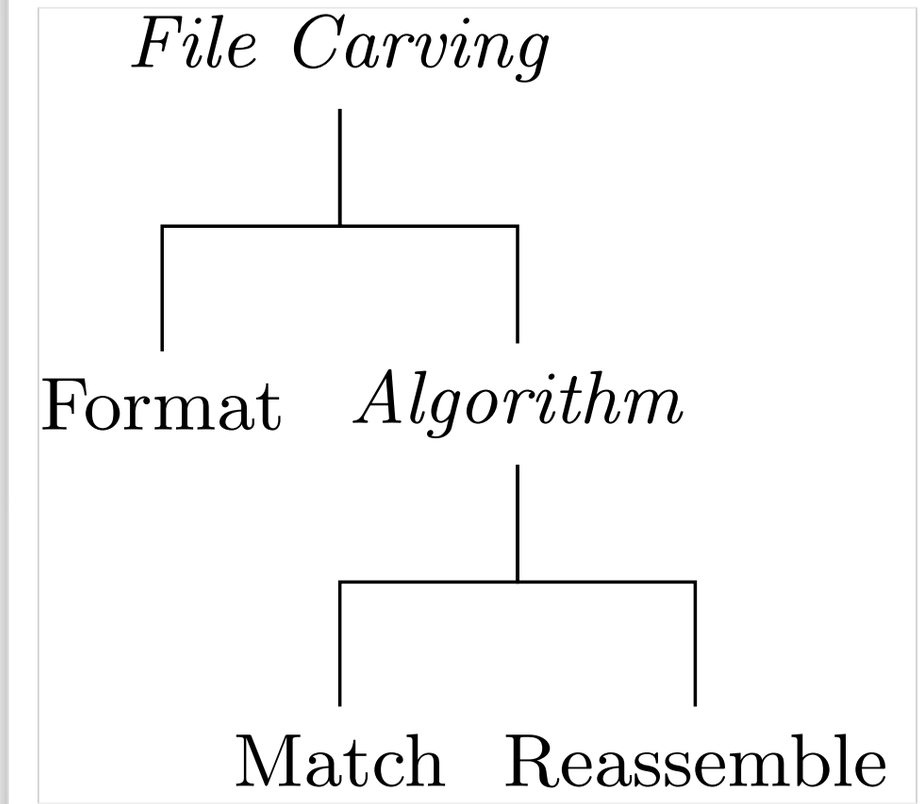
Component	Language	Size (SLOC)
JPEG description	Derric	92
Derric syntax	Rascal	52
Code generator	Rascal	510
Runtime	Java	372
Total		*1026*

Separation of concerns



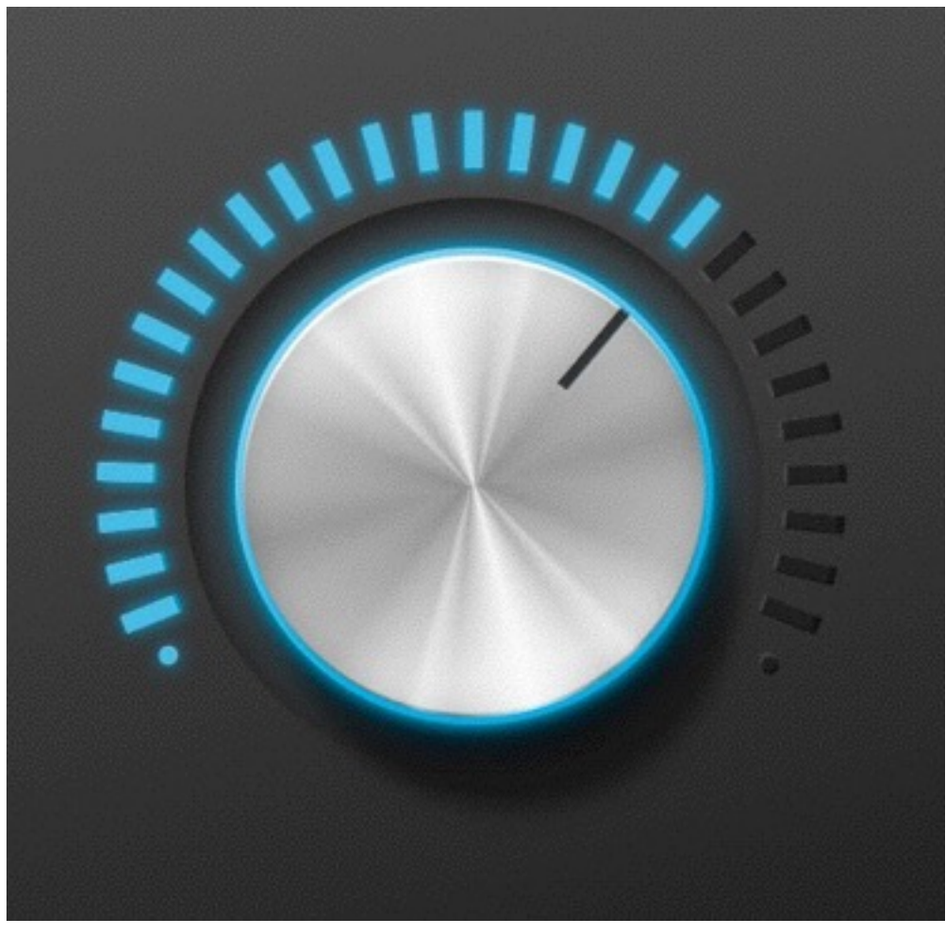
Separation of concerns





Domain-specific optimization

Performance



Accuracy

Model transformations

```
format jpeg
extension jpeg jpg jfif

unit byte
size 1
sign false
type integer
endian big
strings ascii

sequence
  SOI
  ([APP0JFIF APP0JFXX?] [APP1 APP2?])
  !(SOI APP0JFIF APP0JFXX EOI)*
  EOI
```

Fast and imprecise

...

Slow but precise

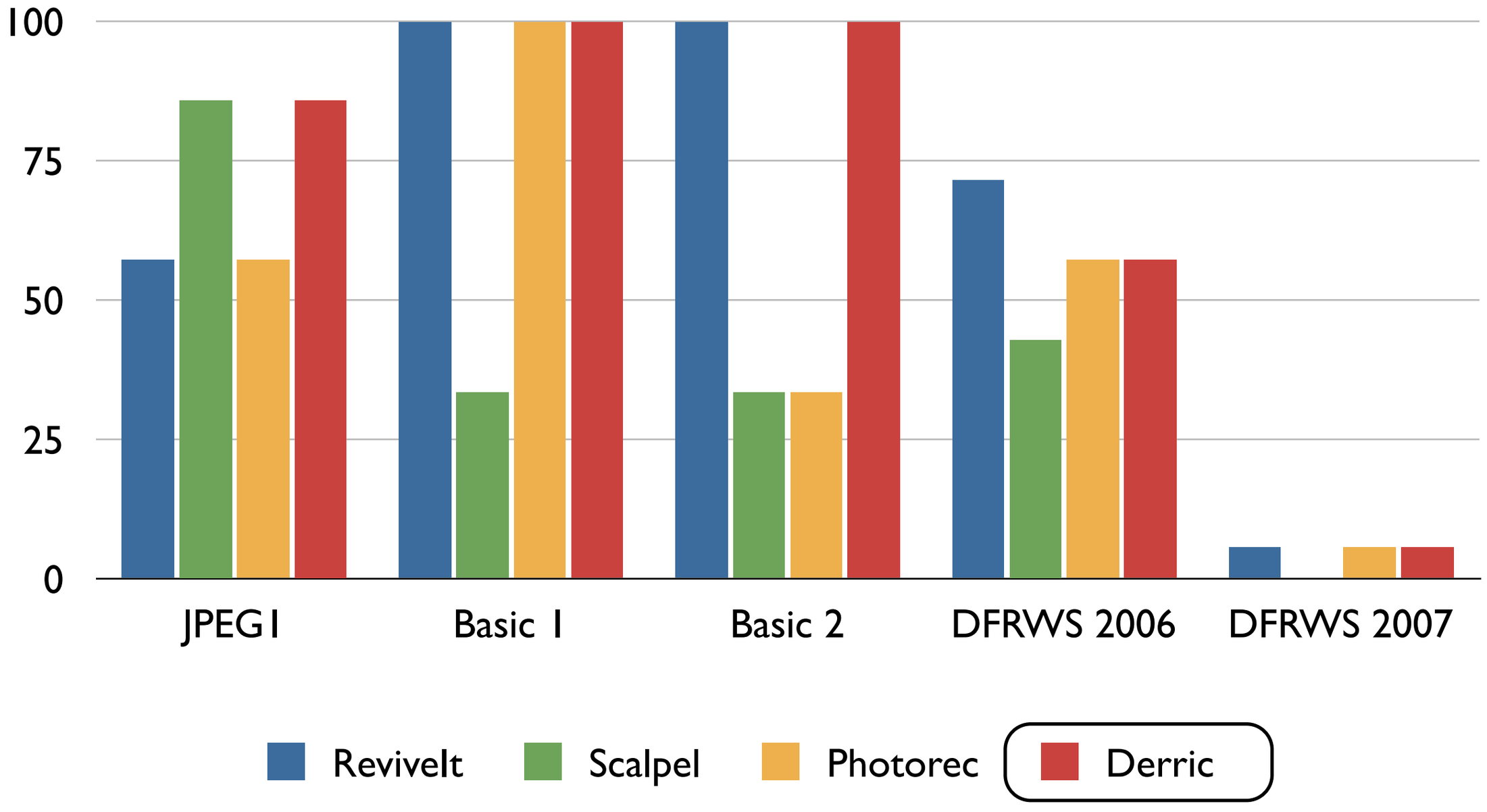
- Opportunities
  - raise level of abstraction
  - separation of concerns
  - domain-specific optimization
- Risks
  - it's not better (functional/non-functional)
  - short-term domain analysis (myopia)
  - over-engineering / accidental complexity

Is it worth it?

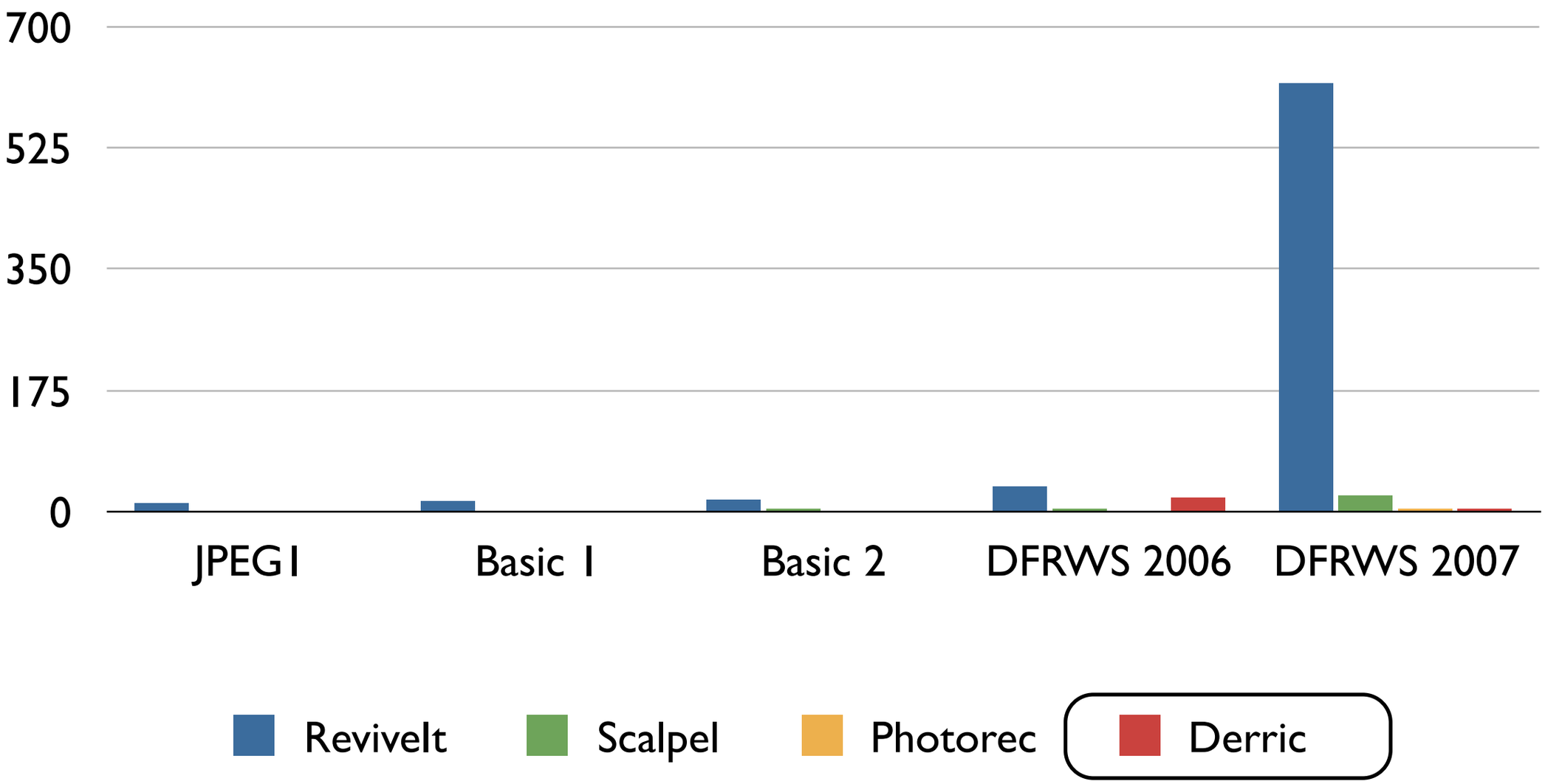
- Compare to existing, open-source file carvers
  - **ReviveIt**
  - **Scalpel**
  - **Photorec**
- Standard, publicly available test images
  - **Digital Forensics Tool Testing Images**
  - **DFRWS Forensic Challenges**

J. van den Bos and T. van der Storm, *Bringing Domain-Specific Languages to Digital Forensics,* (ICSE'11 SEIP)

Recovered files (%)

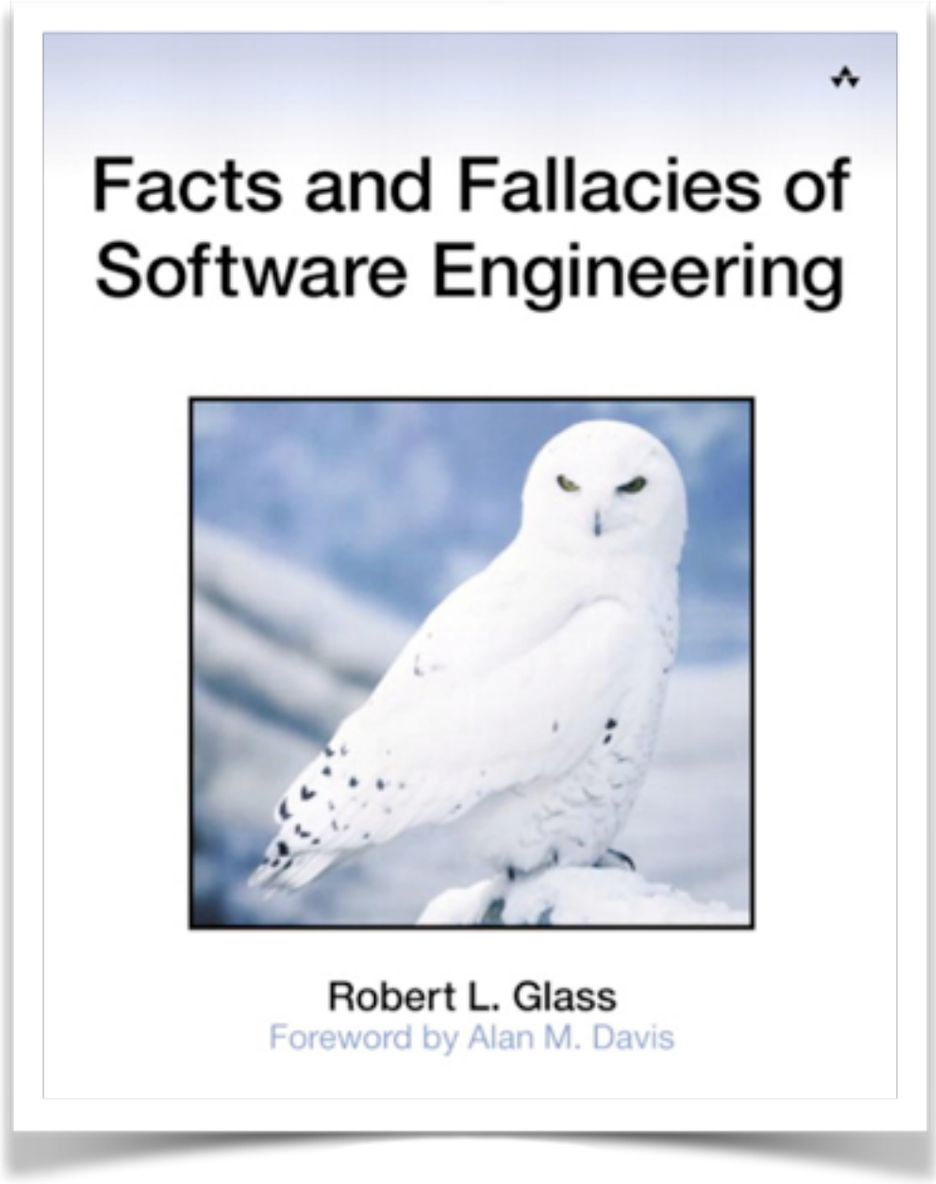


Runtime performance (s)



Long term perspective

- Fact 45:
- Better methods lead to more maintenance, not less.



Approach

- Run Derric derived matchers on **all** images of Wikipedia (JPG, PNG, GIF)
- Record failures
- Fix the models
- Repeat until no more failures
- Classify edits in complexity classes

A Case Study in Evidence-Based DSL Evolution

Jeroen van den Bos[1,2] and Tijs van der Storm[1]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[2] Netherlands Forensic Institute, Den Haag, The Netherlands
jeroen@infuse.org, storm@cwi.nl

European Conference on Modelling foundations and Applications (ECMFA'13)

Over-design

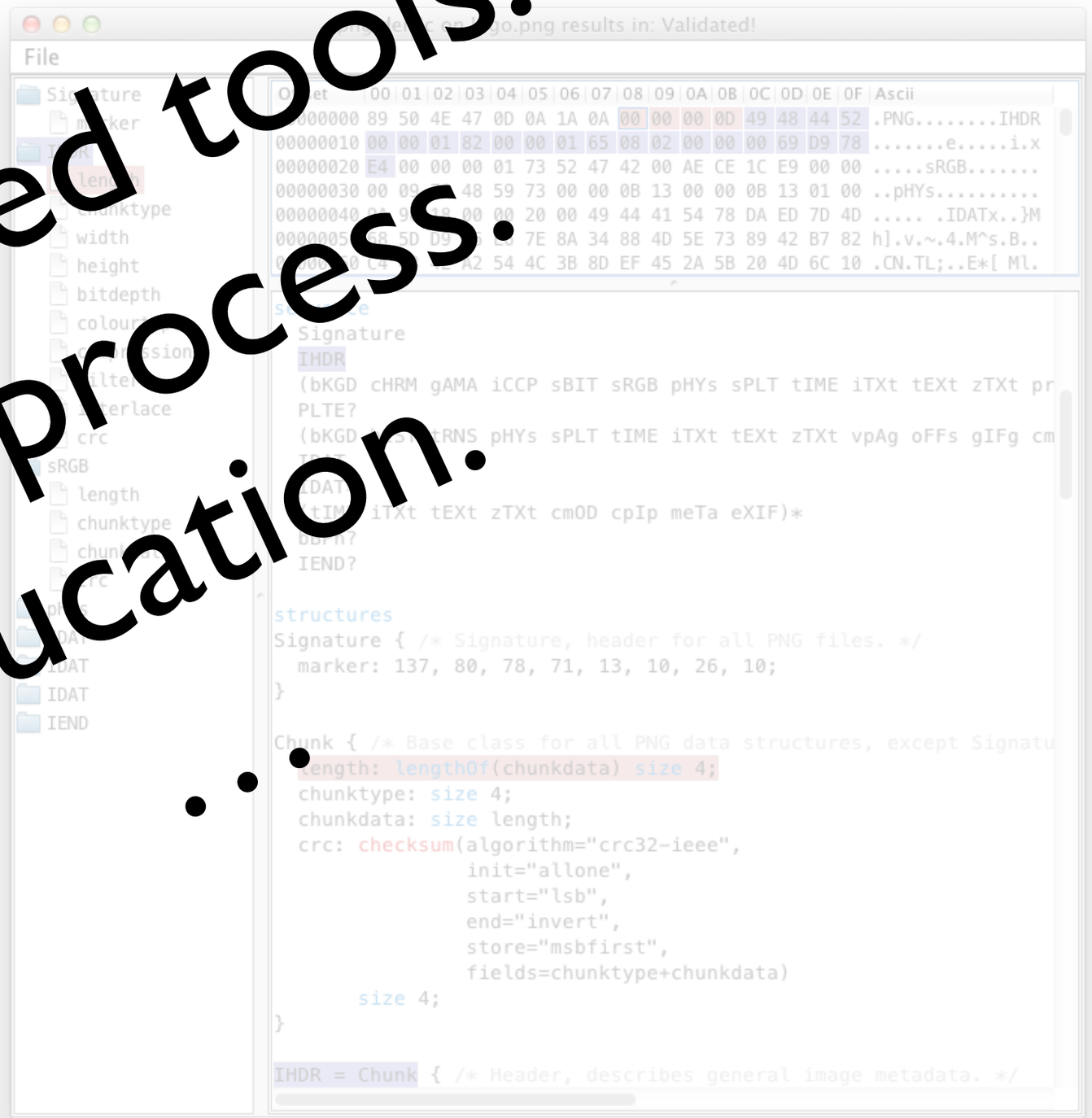
```
format jpeg
extension jpeg jpg jfif

unit byte
size 1
sign false
type integer
endian big
strings ascii

sequence
  SOI
  ([APP0JFIF APP0EXIF? APP1 APP2?])
  !(SOI APP0JFIF APP0EXIF EOI)*
  EOI

structures
SOI {
  marker: 0xFF, 0xD8;
}

APP0JFIF {
  marker: 0xFF, 0xE0;
  length: lengthOf(rgb) + (offset(rgb) -
              offset(length)) size 2;
  identifier: "JFIF", 0;
  version: size 2;
  units: 0 | 1 | 2;
  xdensity: size 2;
  ydensity: size 2;
  xthumbnail;
  ythumbnail;
  rgb: size xthumbnail * (ythumbnail * 3);
}
```



Specialized tools. Build process. Education. ...

Currently: internal DSL

- Embedding in plain Java
- Interpreter
- Practically no dependencies
- Just as good…

MDSE: a balancing act

- Opportunities
  - Raise your level of abstraction
  - Separate what can't be separated
  - Domain-specific analysis and optimization
- Risks
  - It actually doesn't make a difference
  - Short-term domain analysis (myopia)
  - Over-engineering and accidental complexity


Is Derric a success?
A: Yes
B: It depends
C: No
D: N/A

MDSE: a balancing act

- storm@cwi.nl / @tvdstorm
- http://www.cwi.nl/research-groups/Software-analysis-and-transformation
- http://www.rascal-mpl.org

CWI
Centrum Wiskunde & Informatica

Nederlands Forensisch Instituut
Ministerie van Veiligheid en Justitie