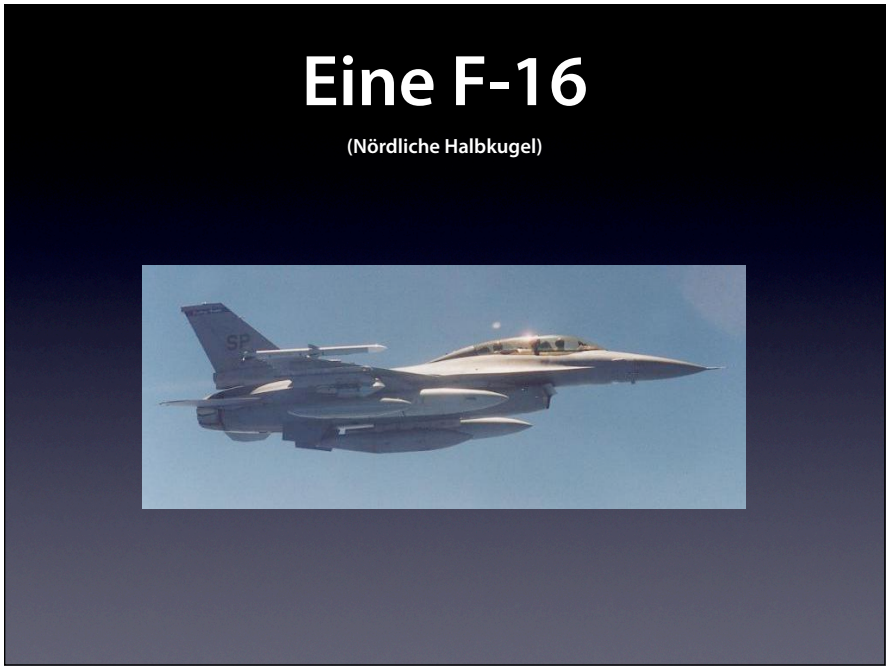
Programme, die Programme prüfen

Jeder Programmierer kennt die Situation: Ein Programm läuft nicht so, wie es soll. Ich stelle Techniken vor, die automatisch

(a) die Ursachen eines Fehlverhaltens finden - indem wir genau die Aspekte isolieren, die das Zustandekommen eines Fehlers verursachen;
(b) Programmfehler finden - indem wir Anwendungen systematisch und vollautomatisch testen; und
(c) vorhersagen, wo in Zukunft Fehler auftreten werden - indem wir maschinell lernen, welche Code- und Prozesseigenschaften bisher mit



An F-16 fighter plane on the northern hemisphere.
Why the northern hemisphere, you ask?

Because this is what an F-16 on the southern hemisphere would look like. (BTW, interesting effect if you drop a bomb :-)

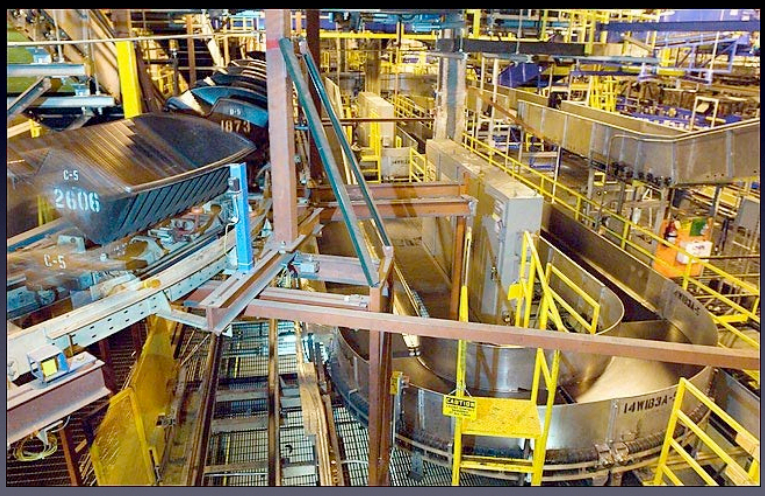
From risks.digest, volume 3, issue 44:
o Since the F-16 is a fly-by-wire aircraft, the computer keeps the pilot from doing dumb things to himself. So if the pilot jerks hard over on the joystick, the computer will instruct the flight surfaces to make a nice and easy 4 or 5 G flip. But the plane can withstand a much higher flip than that. So when they were 'flying' the F-16 in simulation over the equator, the computer got confused and instantly flipped the plane over, killing the

F-16 Fahrgestell

From risks.digest, volume 3, issue 44:
o One of the first things the Air Force test pilots tried on an early F-16 was to tell the computer to raise the landing gear while standing still on the runway. Guess what happened? Scratch one F-16. (my friend says there is a new subroutine in the code called 'wait_on_wheels' now...) [weight?]

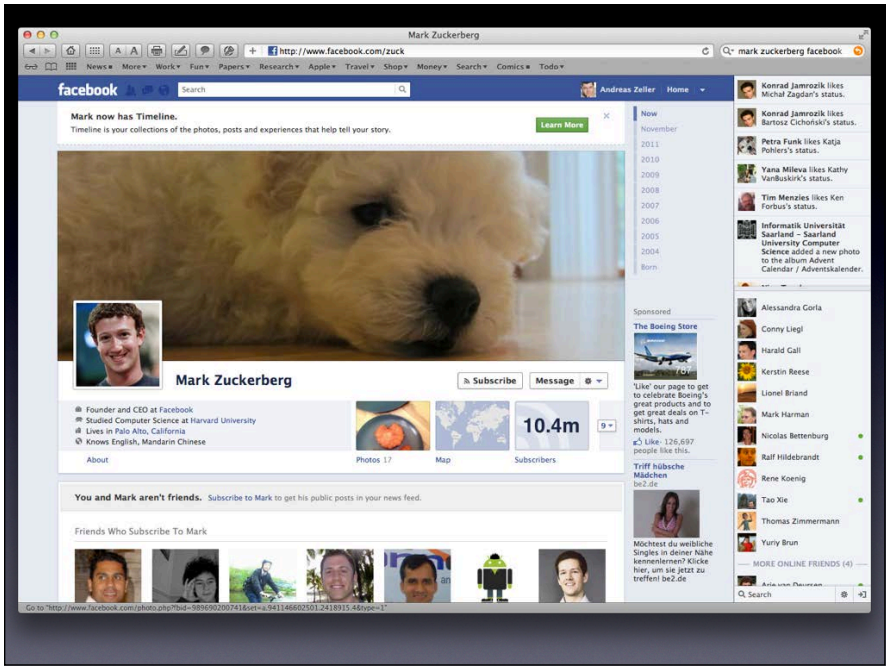
(Folklore has it that the programmer checked the height above sea level rather than the height above ground - AZ)

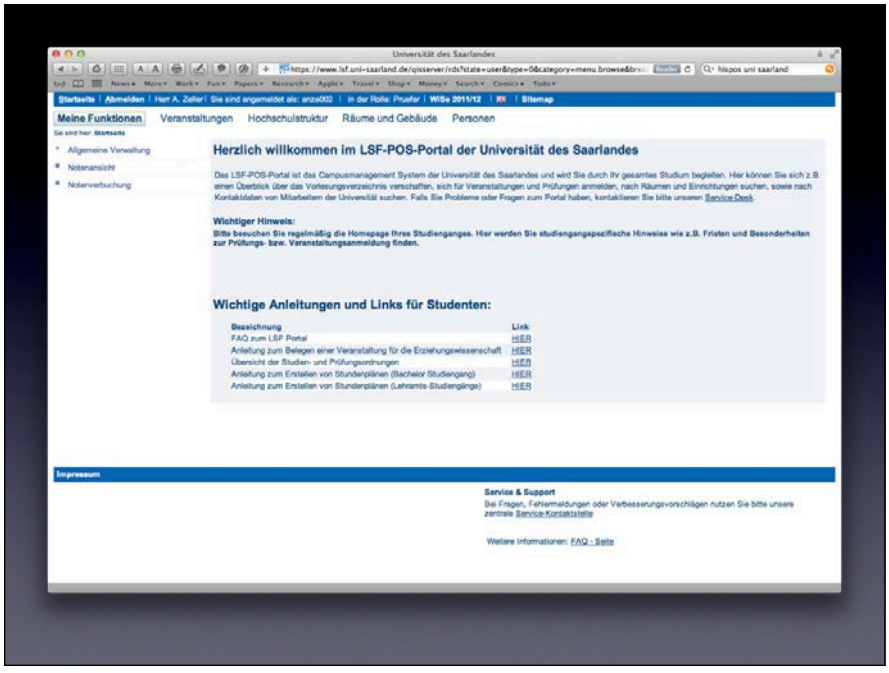
Flughafen Denver



What camera crews depicted was truly a disaster; carts jammed together, damaged luggage everywhere, some bags literally split in half, and the tattered remains of clothing strewn about causing subsequent carts to derail. Finally, adding insult to injury, half the luggage that survived the ordeal ended up at the wrong terminal.

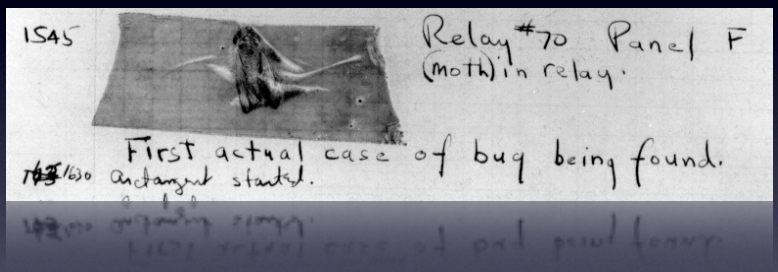
http://www.aeroxp.org/2009/01/lesson-on-infinite-loops/
http://www.youtube.com/watch?v=fYTJ9v2vsaE





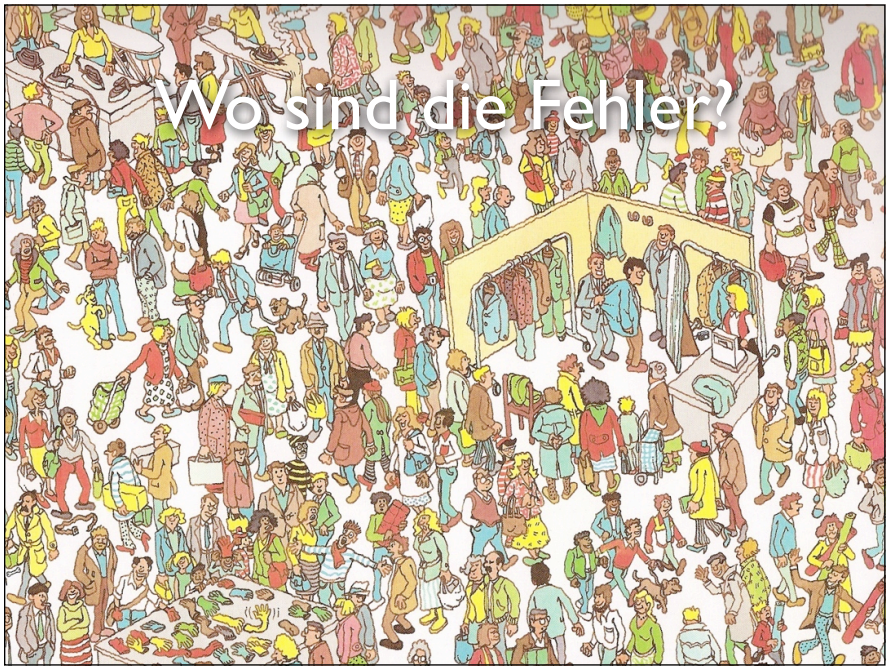
Der erste Bug

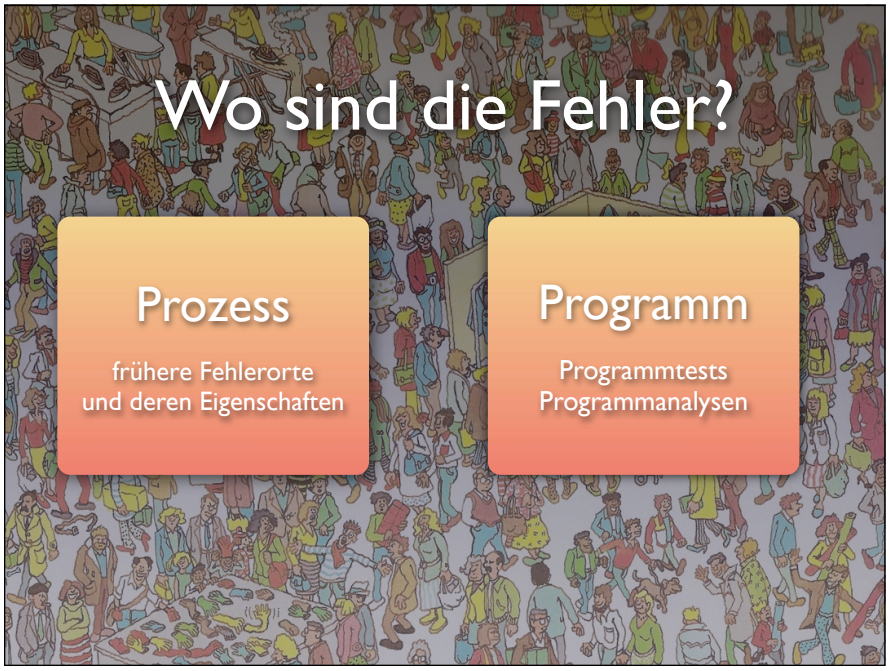
9. September 1947

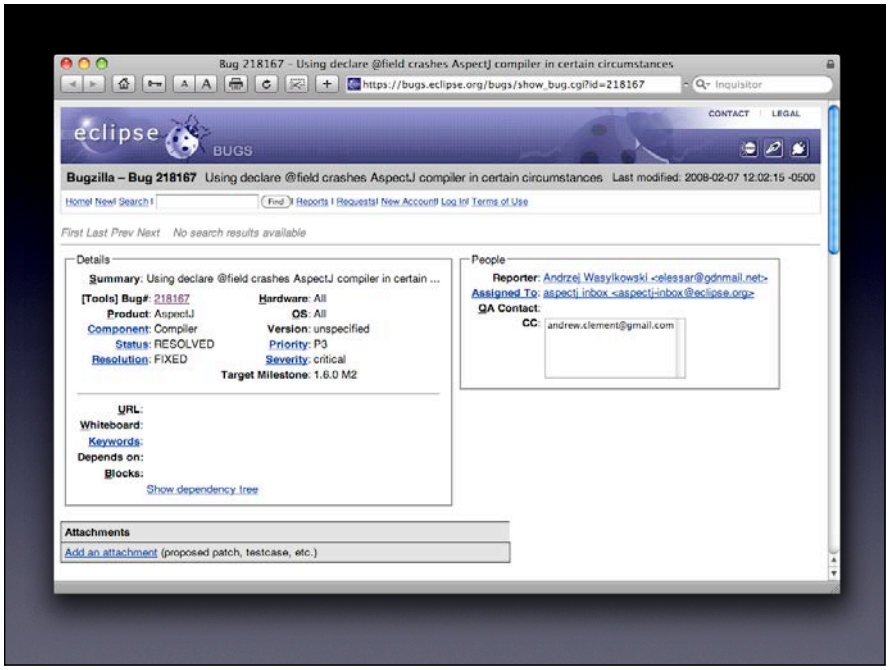


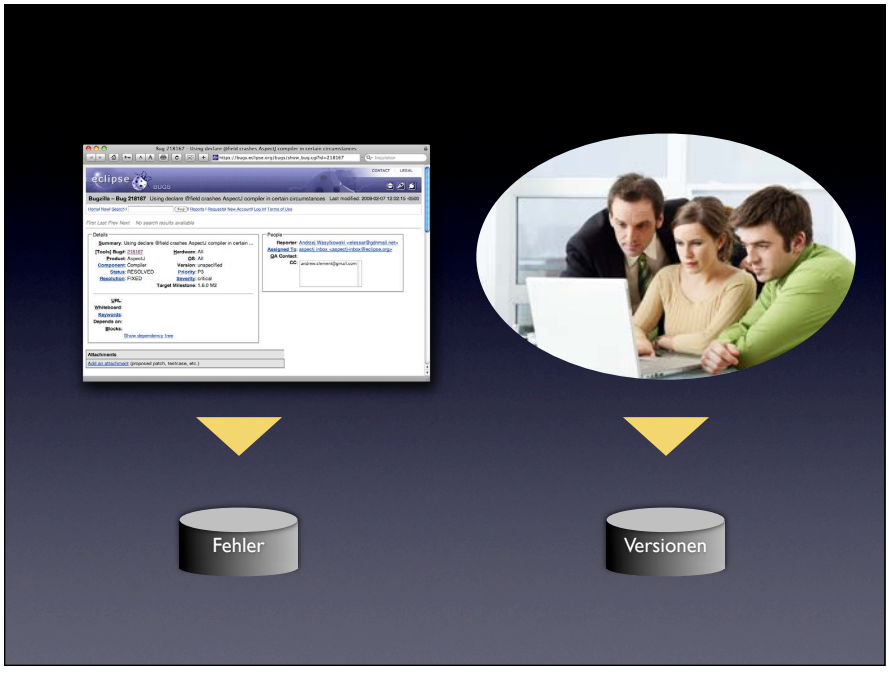
Retrieved by a technician from the Harvard Mark II machine on September 9, 1947.

Now on display at the Smithsonian, Washington

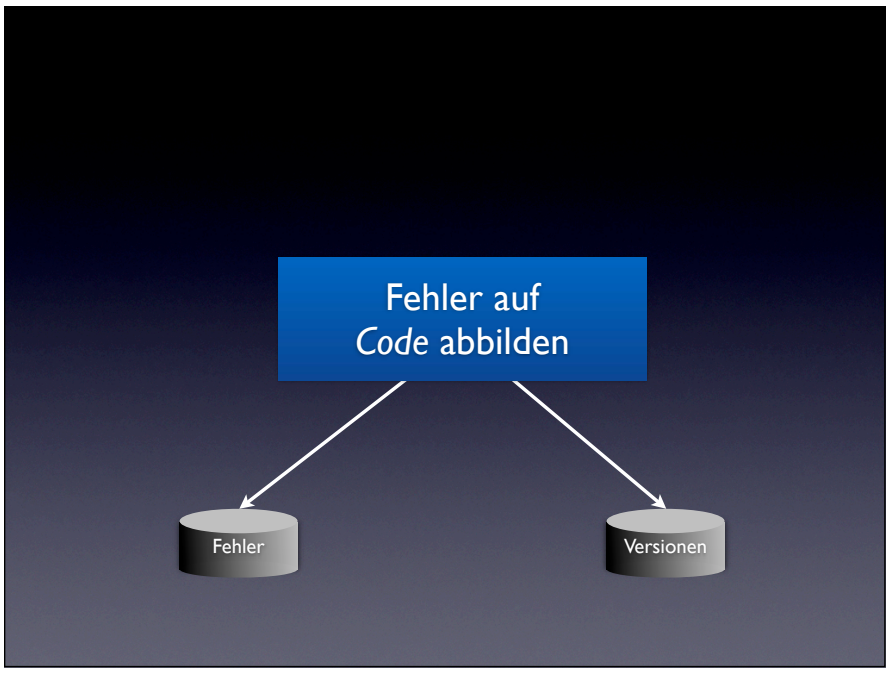




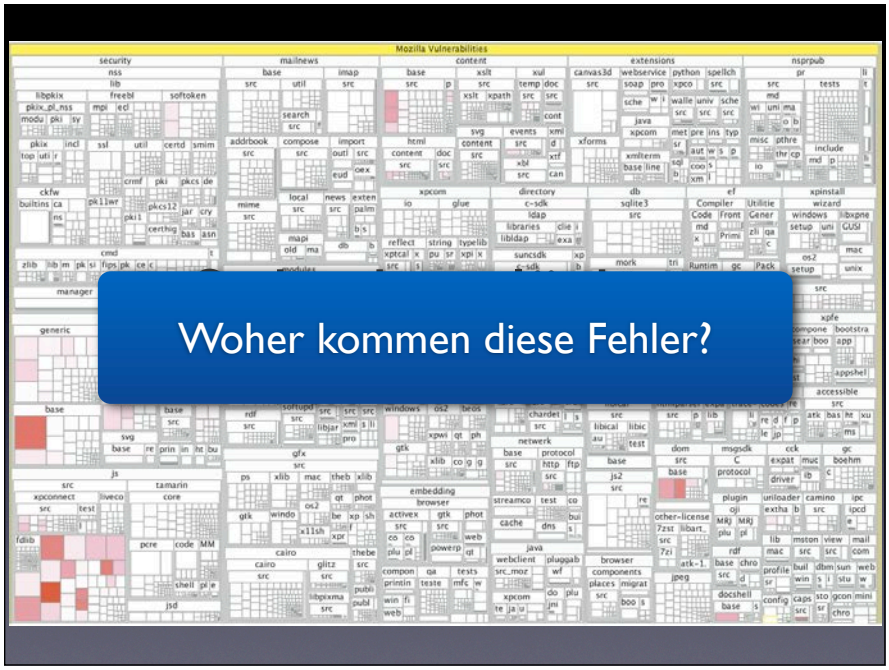


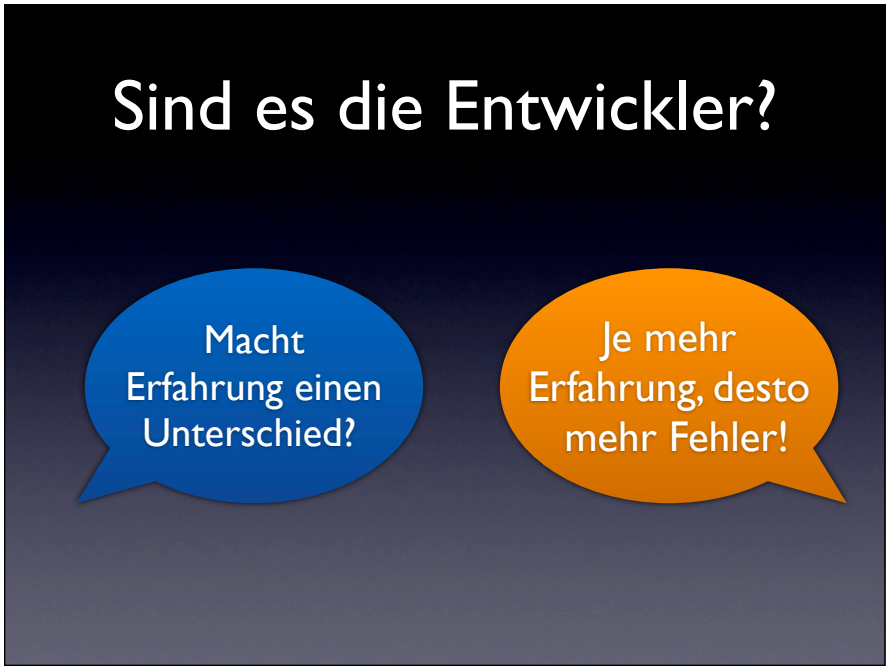


Such software archives are being used in practice all the time. If you file a bug, for instance, the report is stored in a bug database, and the resulting fix is stored in the version archive.

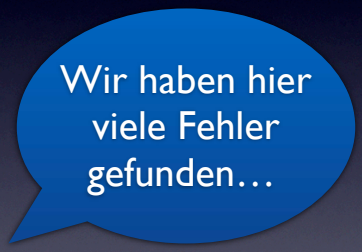


These databases can then be mined to extract interesting information. From bugs and changes, for instance, we can tell how many bugs were fixed in a particular location.

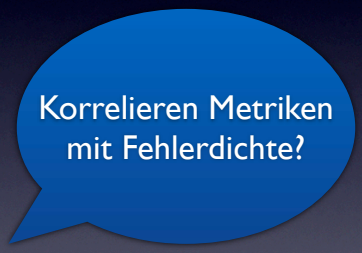




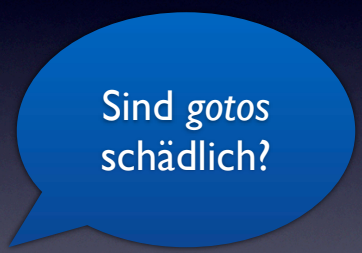
Oder die Geschichte?



Wie steht es mit Metriken?



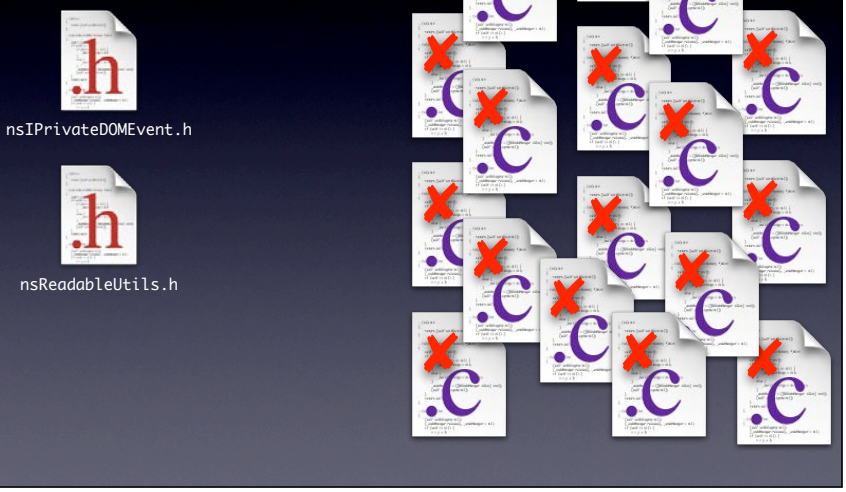
Programmiersprache?



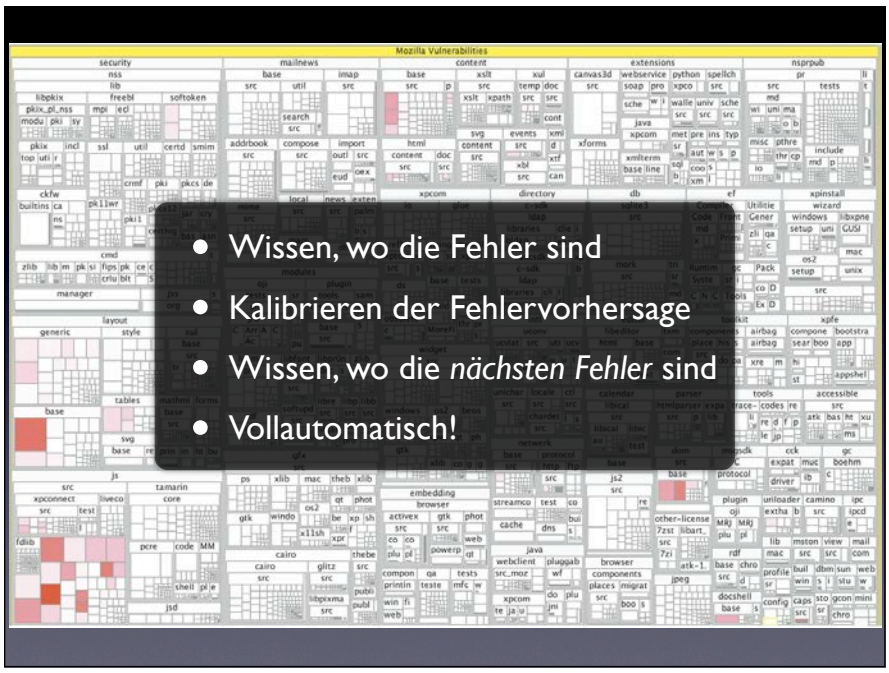
Ok. Problembereich?

Welche Tokens sind relevant?

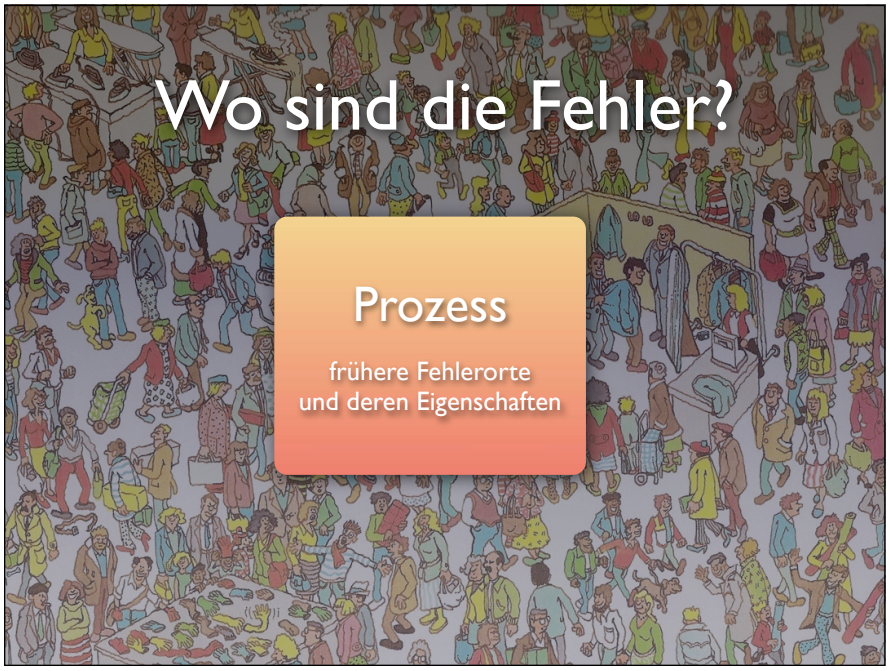
import • extends • implements



Vorhersage	Komponente	Tatsache
1	nsDOMClassInfo	3
2	SGridRowLayout	95
3	xpcprivate	6
4	jsxml	2
5	nsGenericHTMLElement	8
6	jsgc	3
7	nsISEnvironment	12
8	jsfun	1
9	nsHTMLLabelElement	18
10	nsHttpTransaction	35



Sogar das heute journal hat einen Bericht gebracht – seitdem bin ich bekannt aus Funk und Fernsehen :–)



Edgar Degas: The Rehearsal. With a rehearsal, we want to check whether everything will work as expected. This is a test.

Again, a test. We test whether we can evacuate 500 people from an Airbus A380 in 90 seconds. This is a test.

And: We test whether a concrete wall (say, for a nuclear reactor) withstands a plane crash at 900 km/h. Indeed, it does.

We can also test software this way. But software is not a planned linear show – it has a multitude of possibilities. So: if it works once, will it work again? This is the central issue

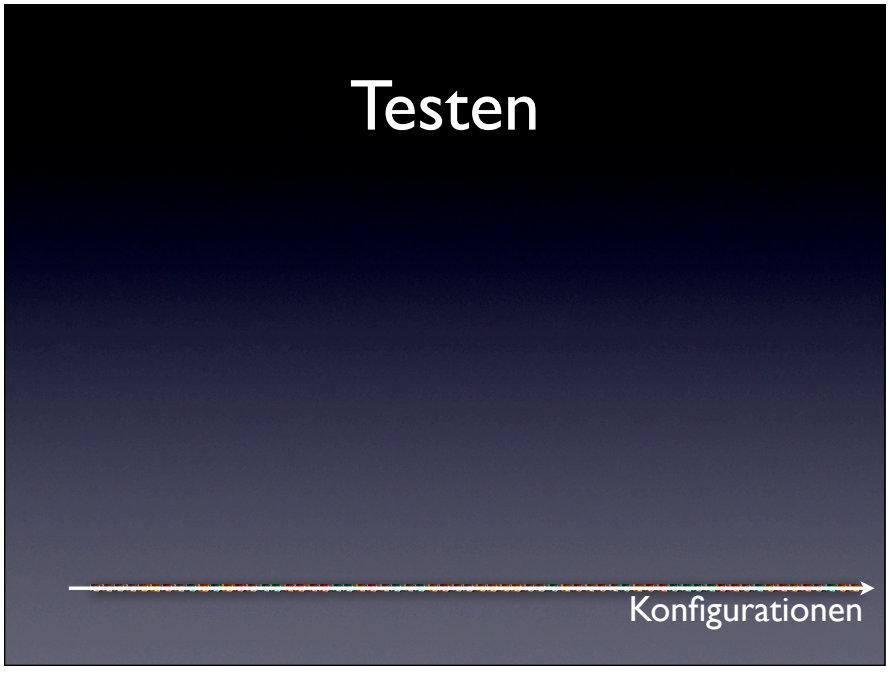
We can also test software this way. But software is not a planned linear show – it has a multitude of possibilities. So: if it works once, will it work again? This is the central issue

The problem is: There are many possible executions. And as the number grows…

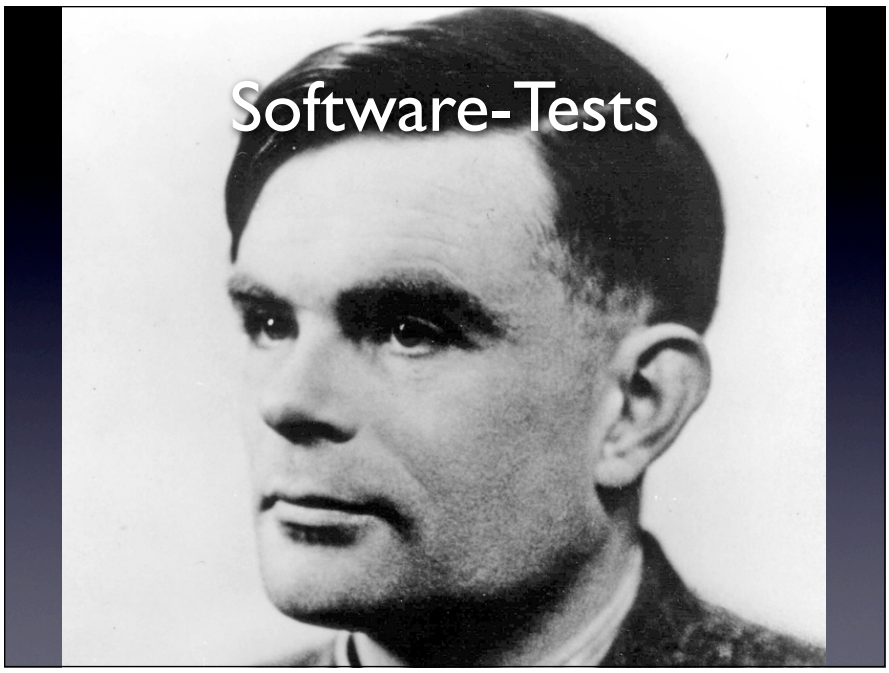
and grows…

and grows…

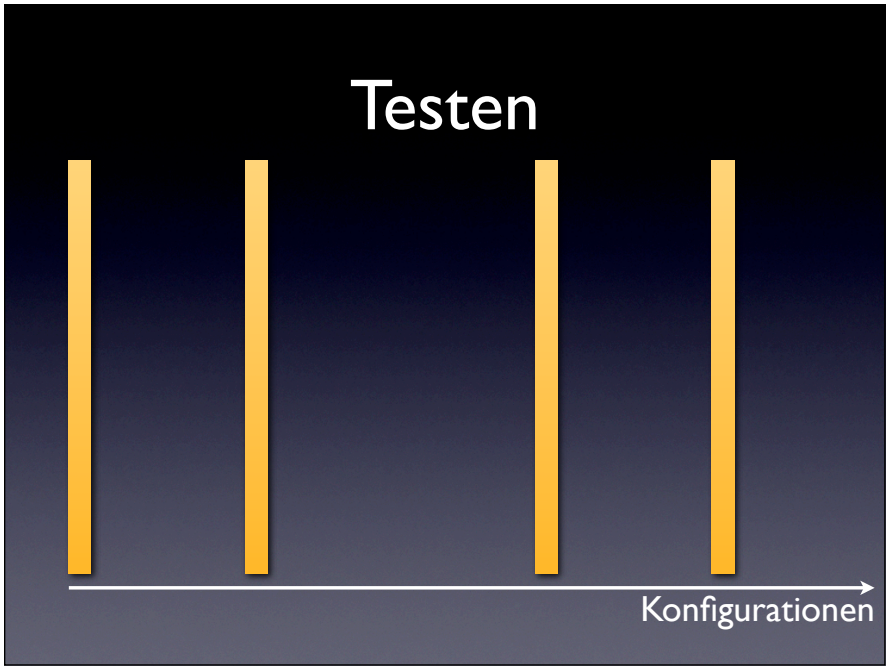
and grows…



…you get an infinite number of possible executions, but you can only conduct a finite number of tests.

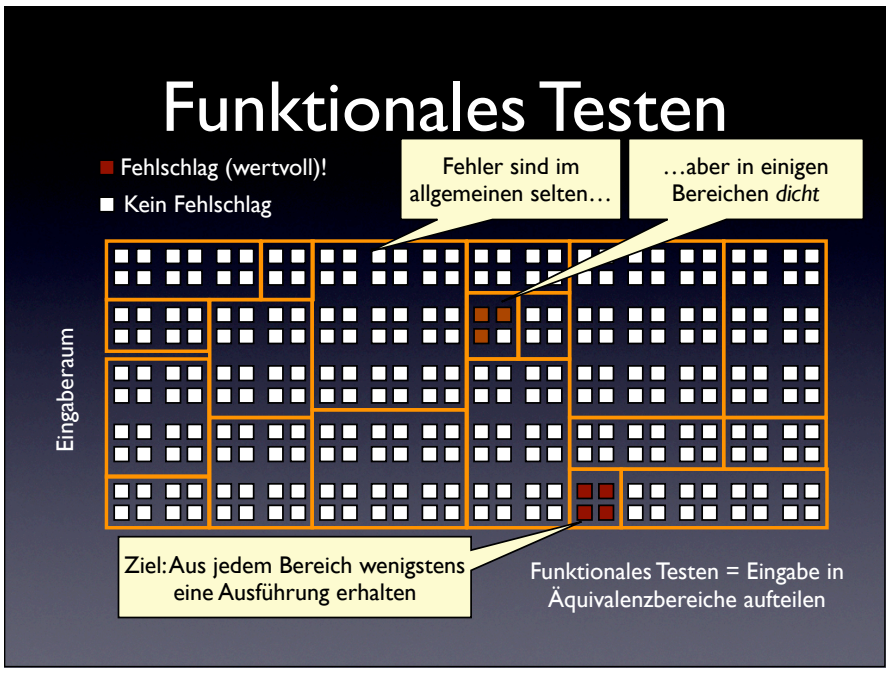


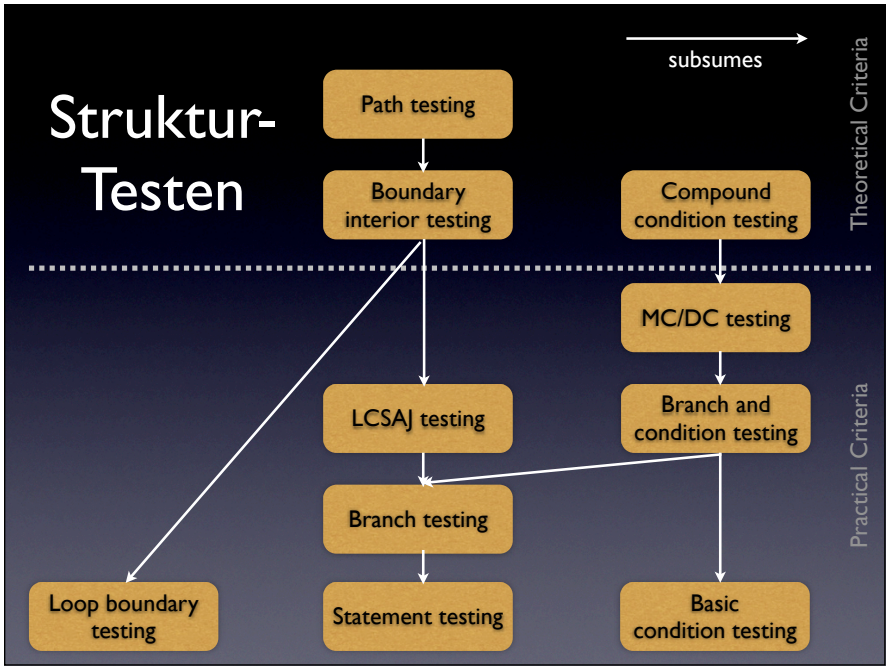
…and this was something the first testers also needed to realize.

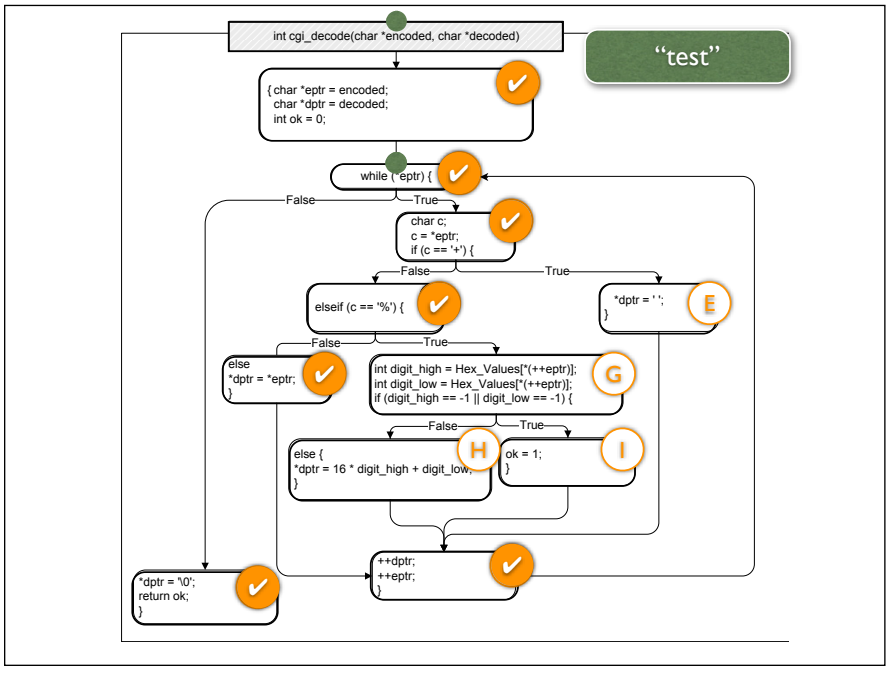


With testing, you pick a few of these configurations – and test them.

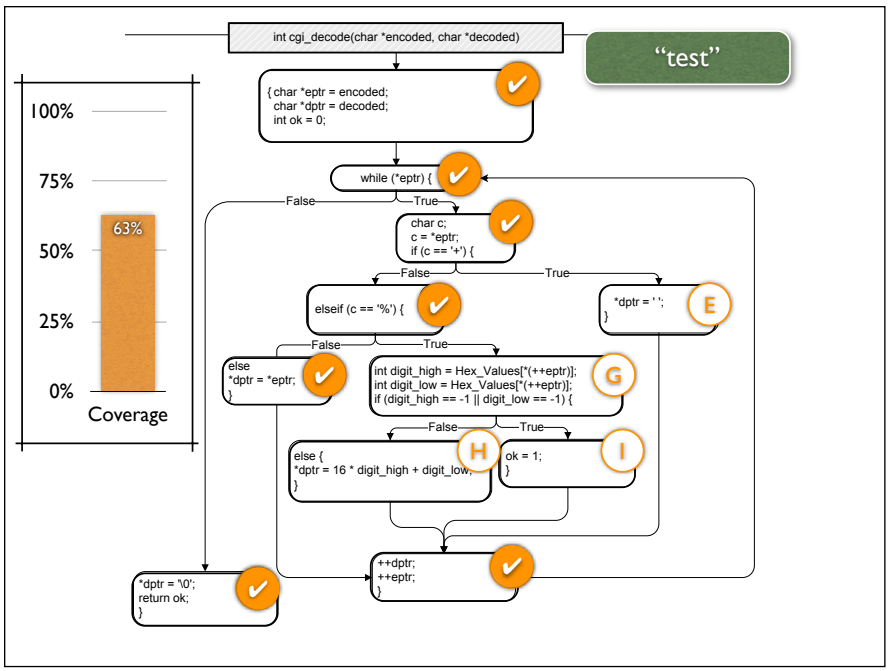
So, how can we cover as much behavior as possible?



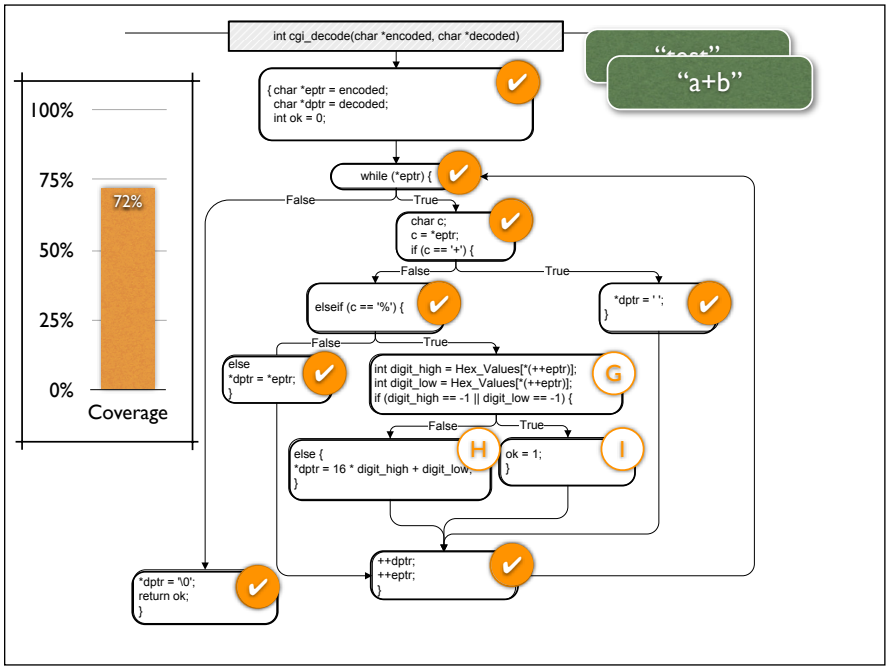




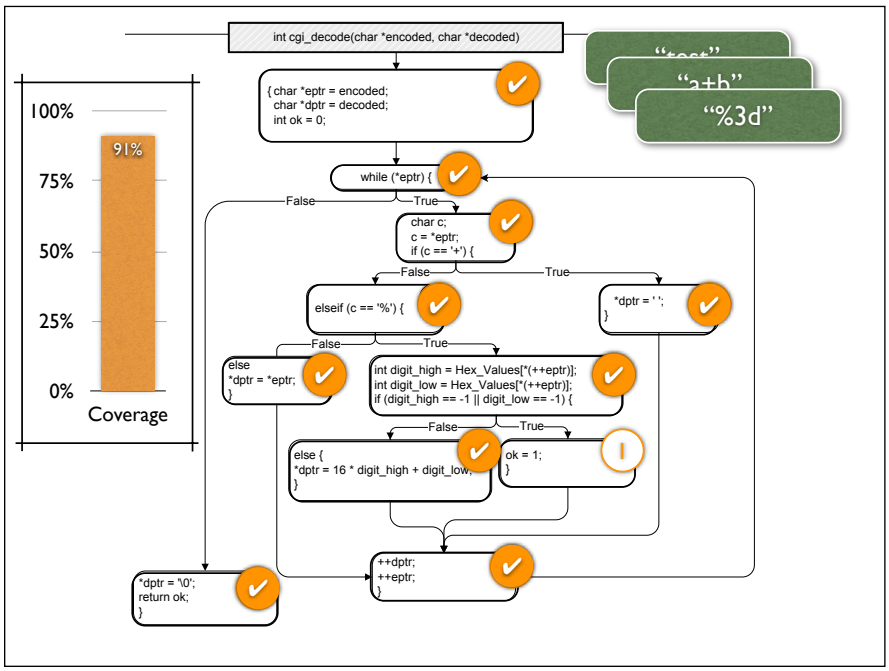
While the program is executed, one statement (or basic block) after the other is covered – i.e., executed at least once – but not all of them. Here, the input is "test"; checkmarks indicate executed blocks.

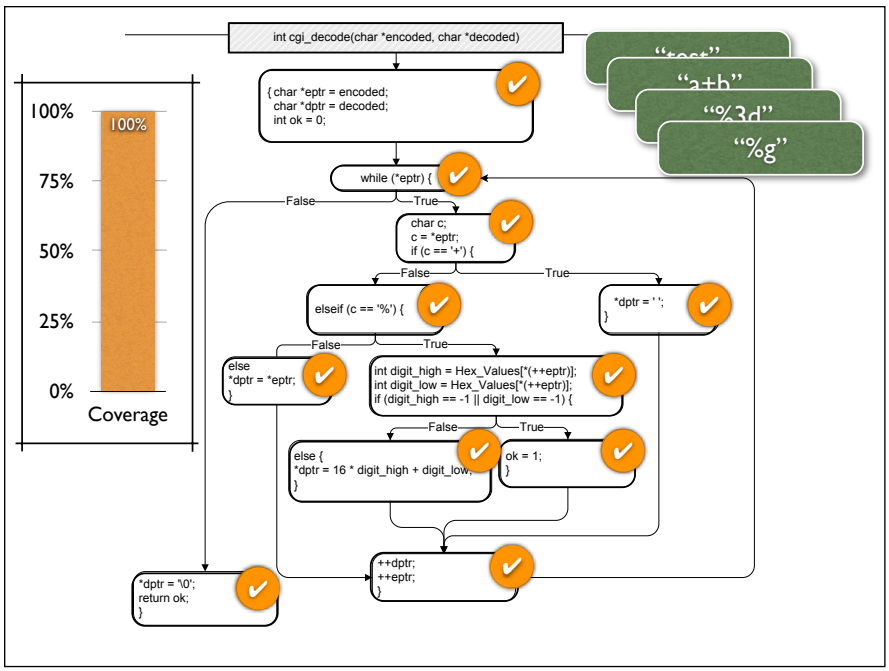


The initial coverage is 7/11 blocks = 63%. We could also count the statements instead (here: 14/20 = 70%), but conceptually, this makes no difference.

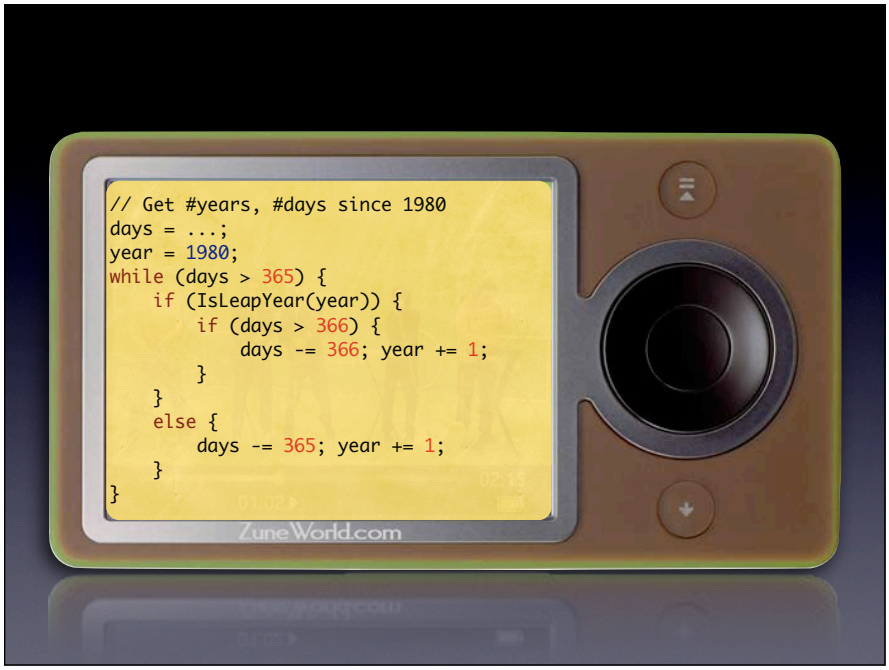


and the coverage increases with each additionally executed statement…





… until we reach 100% block coverage (which is 100% statement coverage, too).

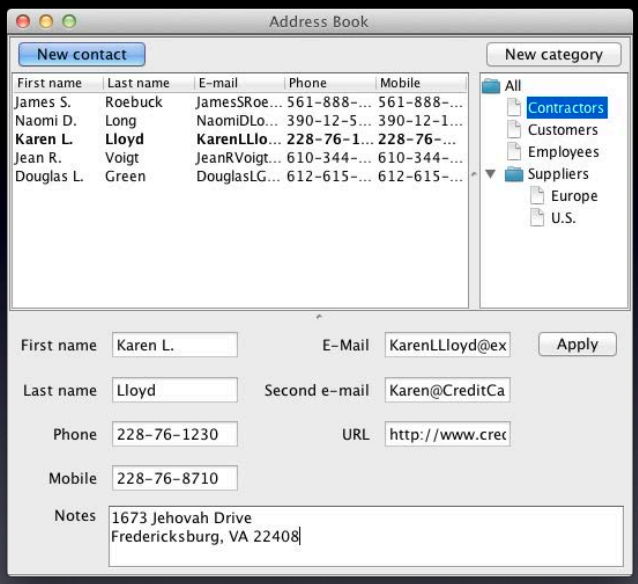


http://www.aeroxp.org/2009/01/lesson-on-infinite-loops/
http://www.youtube.com/watch?v=fYTJ9v2vsaE

All these techniques attempt to find the needle in the haystack…

Automatisierung

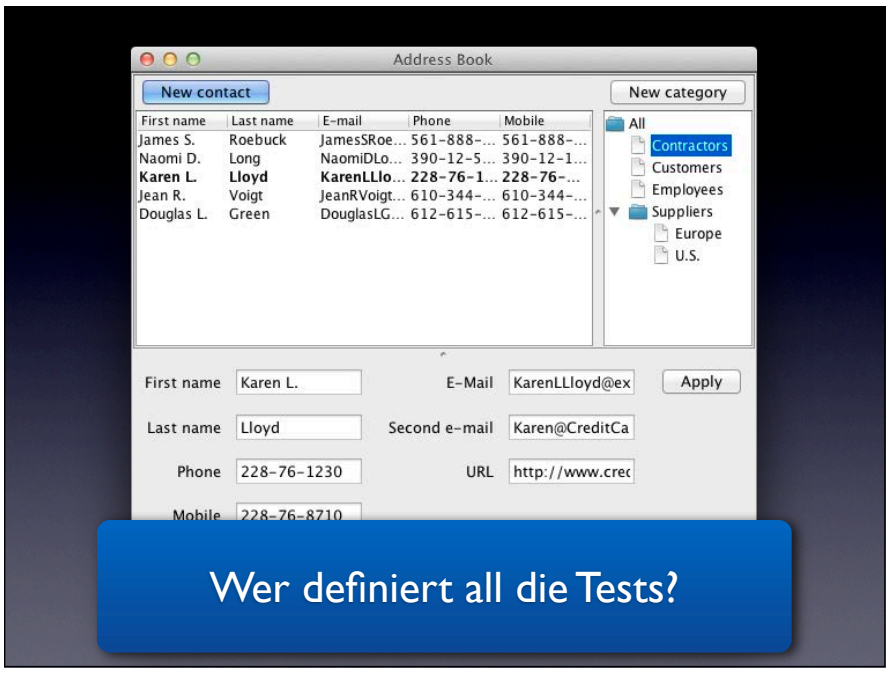
- Automatisches Ausführen
- Automatisches Generieren
- Automatisches Prüfen



As an example, here's the Addressbook program: a simple Java application which manages a set of contacts that can be entered, searched, and grouped into

Capture + Replay

- Wir können Tastatur und Maus *aufzeichnen*
- …und nach Belieben wieder *abspielen*!



As an example, here's the Addressbook program: a simple Java application which manages a set of contacts that can be entered, searched, and grouped into

Zufallstesten

```
public class RandoopTest0 extends TestCase {
 ...

 public void test8() throws Throwable {
  if (debug) System.out.printf("%nRandoopTest0.test8");

  AddressBook var0 = new AddressBook();
  EventHandler var1 = var0.getEventHandler();
  Category var2 = var0.getRootCategory();
  Contact var3 = new Contact();
  AddressBook var4 = new AddressBook();
  EventHandler var5 = var4.getEventHandler();
  Category var6 = var4.getRootCategory();
  String var7 = var6.getName();
  var0.addCategory(var3, var6);
  SelectionHandler var9 = new SelectionHandler();
  AddressBook var10 = new AddressBook();
  EventHandler var11 = var10.getEventHandler();
```

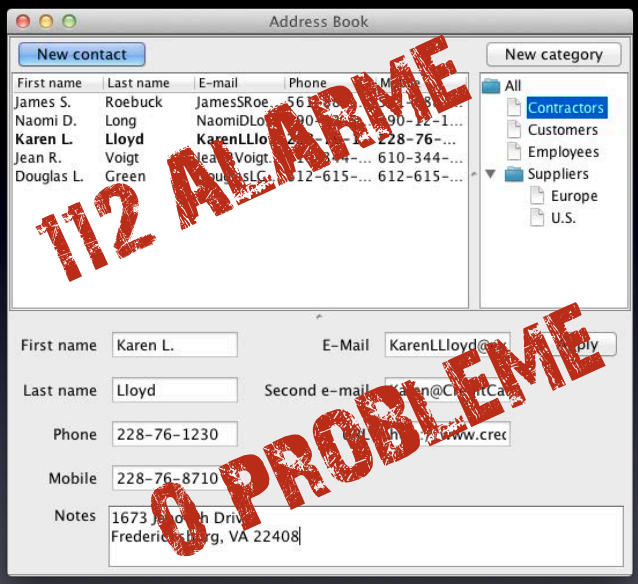
Here's a test case generated by Randoop. It's >200 lines long…

```
  AddressBook var65 = new AddressBook();
  EventHandler var66 = var65.getEventHandler();
  Category var67 = var65.getRootCategory();
  Contact var68 = new Contact();
  Category[] var69 = var68.getCategories();
  var65.removeContact(var68);
  java.util.List var71 = var65.getContacts();
  AddressBook var72 = new AddressBook();
  EventHandler var73 = var72.getEventHandler();
  Category var74 = var72.getRootCategory();
  EventHandler var75 = var72.getEventHandler();
  SelectionHandler var76 = new SelectionHandler();
  actions.CreateContactAction var77 = new actions.CreateContactAction(var72, var76);
  boolean var78 = var77.isEnabled();
  AddressBook var79 = new AddressBook();
  EventHandler var80 = var79.getEventHandler();
  Category var81 = var79.getRootCategory();
  String var82 = var81.getName();
  var77.categorySelected(var81);
  Category var85 = var65.createCategory(var81, "hi!");
  String var86 = var85.toString();
  Category var88 = var0.createCategory(var85, "exceptions.NameAlreadyInUseException");
 }
```

… and in the end, it fails. What do you do now?

Zufallstesten

- Einfach zu realisieren…
- …aber erzeugt viele unsinnige Tests!



The catch is: There's never more than one addressbook! So the Randoop test makes little sense, because it violates an implicit precondition. When testing the Addressbook classes, Randoop detects * 112 failures. However, all of them are false, pointing to an error in the generated test case rather than the application itself, which has *0 problems.

Ein Fehlalarm

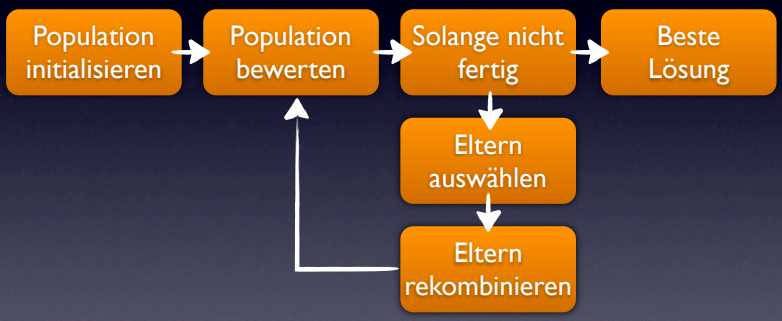
```
public class RandoopTest0 extends TestCase {
  public void test8() throws Throwable {
    if (debug) System.out.printf("%nRandoopTest0.test8");

    AddressBook a1 = new AddressBook();
    AddressBook a2 = new AddressBook();
    Category a1c = a1.createCategory(a1.getRootCategory(), "a1c");
    Category a2c = a2.createCategory(a1c, "a2c");
  }
}
```

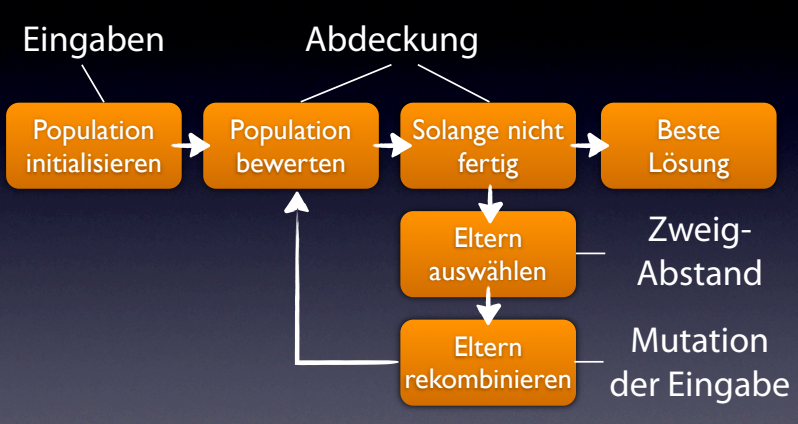
A simplified version of the above. If you use two address book objects and make one's category depend on one the other, it'll crash.

System-Tests

- Erzeuge Tests für die Bedienoberfläche
  - Jede Eingabe ist korrekt
  - Keine Fehlalarme

Genetische Algorithmen



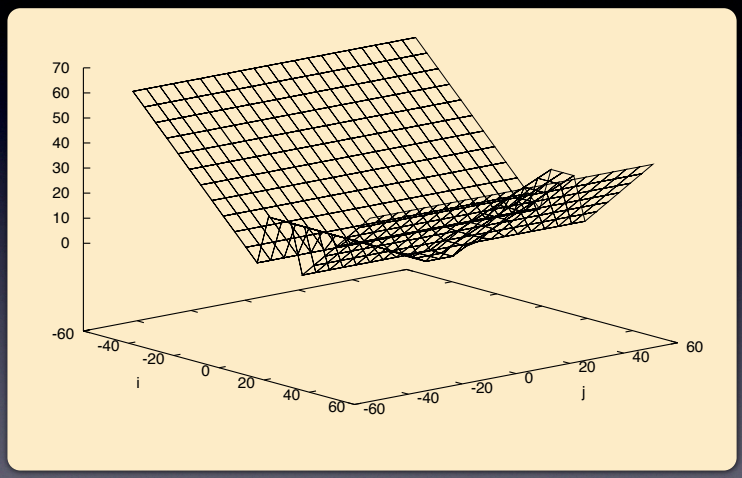


Zweigabstand

```
void landscape_example(int i, int j) {
  if (i >= 10 && i <= 20) {
    if (j >= 0 && j <= 10) {
      // target statement
      // ...
    }
  }
}
```

Wie dicht sind wir an diesem Prädikat?

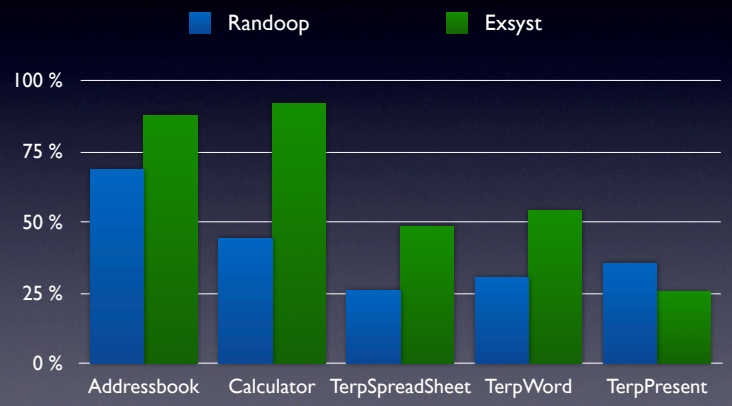
Suchlandschaft



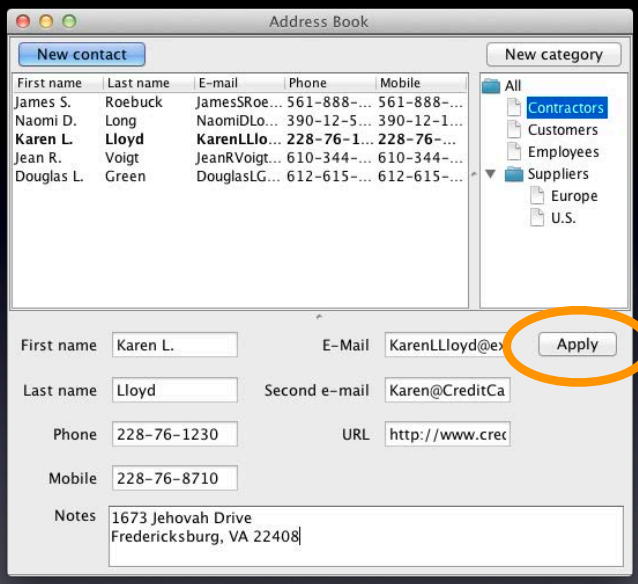
What I'm going to demo you now is our prototype called EXSYST, for Explorative SYStem Testing. EXSYST takes a Java program with a graphical user interface, such as our Addressbook example. It then generates user inputs such as mouse clicks or keystrokes and feeds them into the program. What you see here is EXSYST clicking and typing into the address book program; at the top, you see the statement coverage achieved so far. (Normally, all of this takes place in the background, so you don't see it, and it is also much much faster).

At first, these inputs are completely random, as you can see in these initial

Erzielte Abdeckung

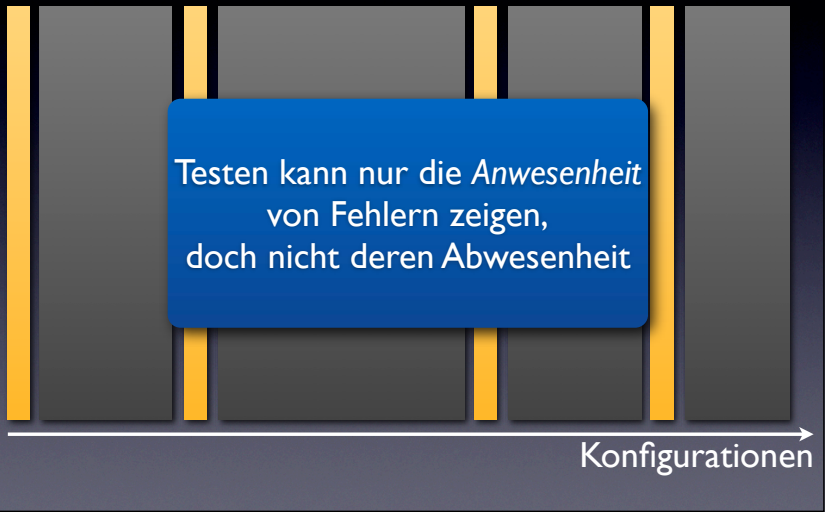


The results are clear. Although it's going through the GUI, EXSYST achieves a far higher coverage than Randoop. Here are the results for * Addressbook and *** three more



EXSYST found failures in all five programs which can be invoked with a few simple inputs. In AddressBook, for instance, if you press the Apply button without

Dijkstras Fluch



But still, testing suffers from what I call Dijkstra's curse – a double meaning, as it applies both to testing as to his famous quote. Is there something that can find the

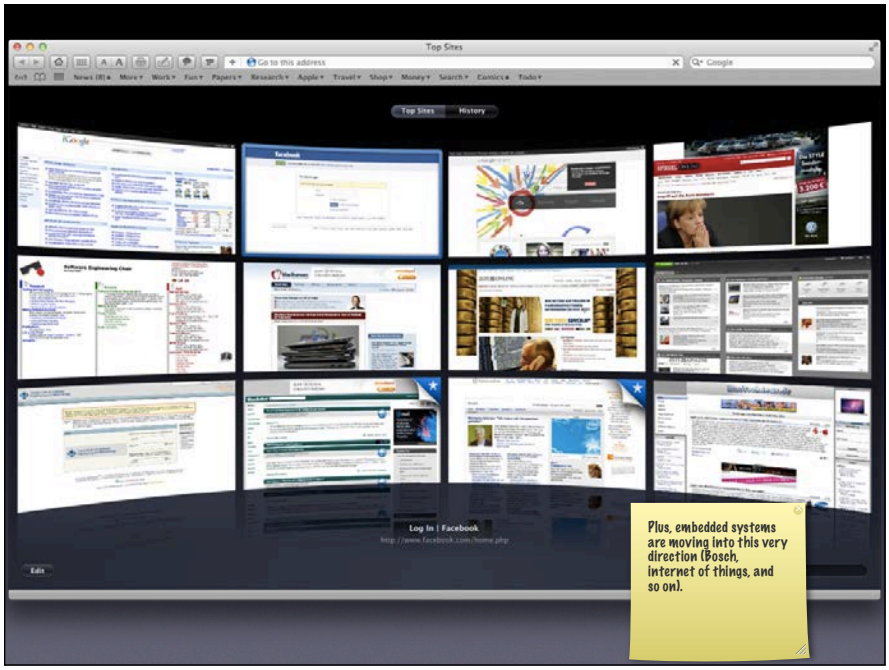
Areas missing might be: the operating system, the hardware, all of the world the system is embedded in (including humans!)

Außer Kontrolle

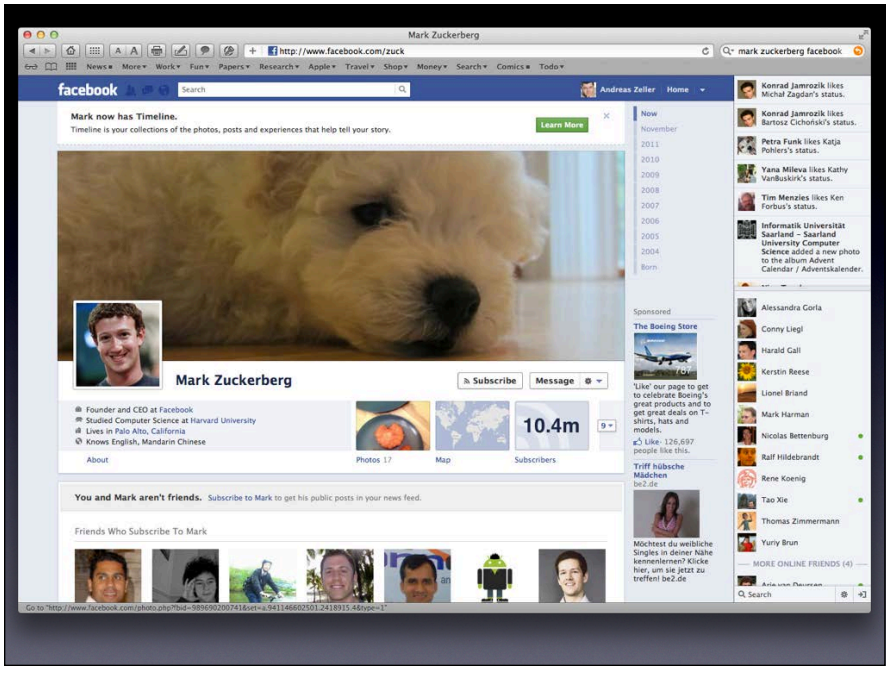
Moderne Programme sind nicht mehr zu beweisen:

- Mehrere Sprachen
- Obskurer oder nicht verfügbarer Code
- Verteilte Aufrufe

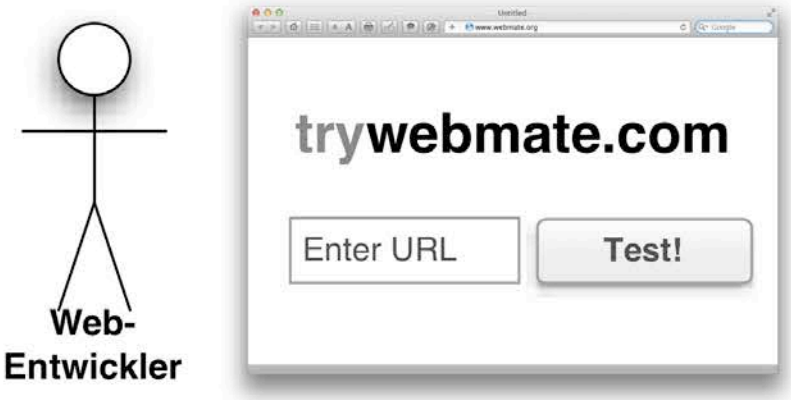
There's a reason why verification today focuses on embedded systems – because that's the only area where we can still assume we have everything under control!



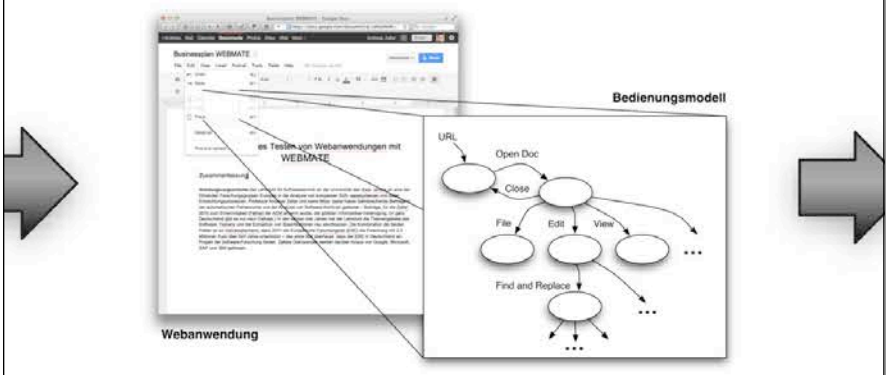
Well, everyone has. You start a browser, you have it all. None of this is what program analysis can handle these days. We're talking scripts, we're talking distributed, we're talking amateurs, we're talking security.



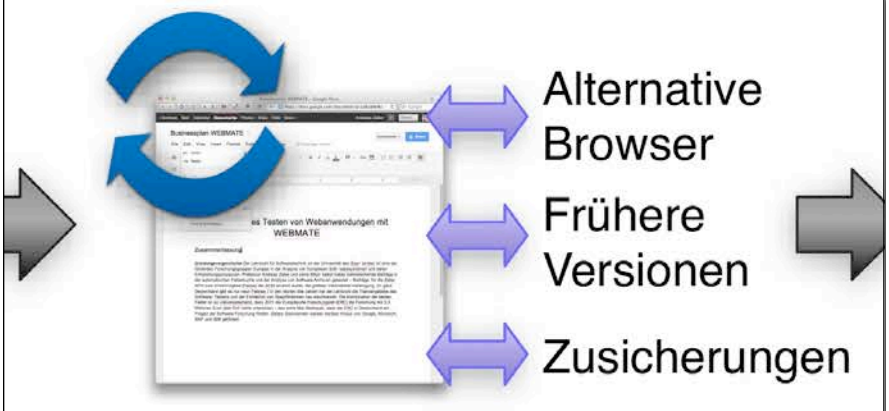
Webmate



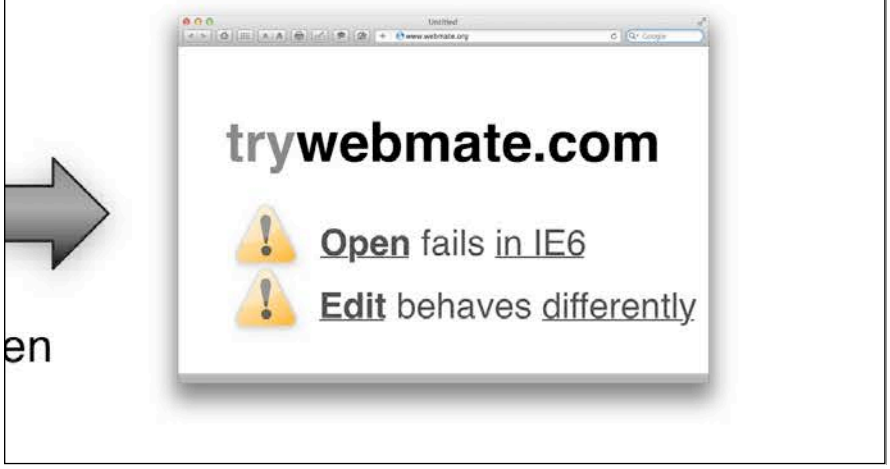
Webmate



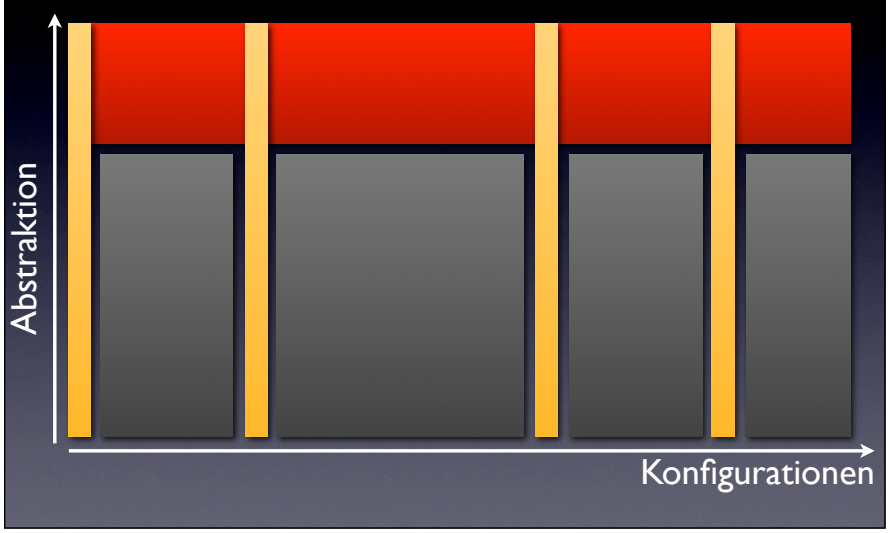
Webmate



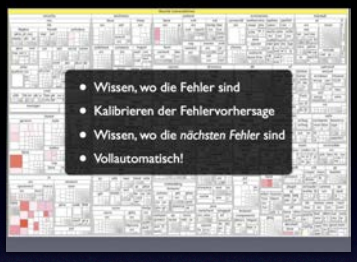
Webmate



Das Beste aller Welten



We might not be able to cover **all** abstraction levels in **all** configurations, but we can do our best to cover as much as possible.



Programme, die Programme prüfen