Source Mining

Statische Analyse mit Columbo

Jens-D. Doll
Context IT
www.cococo.de

Bio

- 1975ff Studium der Informatik/Mathematik
- 1980 Entwurf eines Fortran 77-Compilers
- 1980 erste Idee zu Columbo
- 1981 Entwicklung einer REXX-ähnlichen Sprache
- 1982 Adaption eines COBOL-Compilers
- 1993ff verschiedene Projekte mit COBOL
- 1998 Einführung eines Repositories
- 2001 Reverse Engineering Projekt
- 2004 Konzept für Columbo
- 2005 Prototyp von Columbo vorhanden

Grundlagen

- Ralf Lämmel et al „COBOL-Grammatik“
- F.W. Schröer et al „Gentle und Accent“
- Glynn Winskel „Formal Semantics“
- Cousot, P u. R. „Abstract Interpretation“
- Uwe Schöning „Theoretische Informatik“
- A. Steger „Diskrete Mathematik“
- ...

COBOL COBOL COBOL?

- COBOL ist auch objektorientiert
- Vererbung
- Polymorphismus
- Persistenz
- Die Sprache ist wortgewaltig

Langfristiges Ziel

- Übersetzung von Programmeigenschaften in die **natürliche Sprache** für Anwender
- Definition eines **Satzes von Funktionen** zur qualitativen Beschreibung von Software

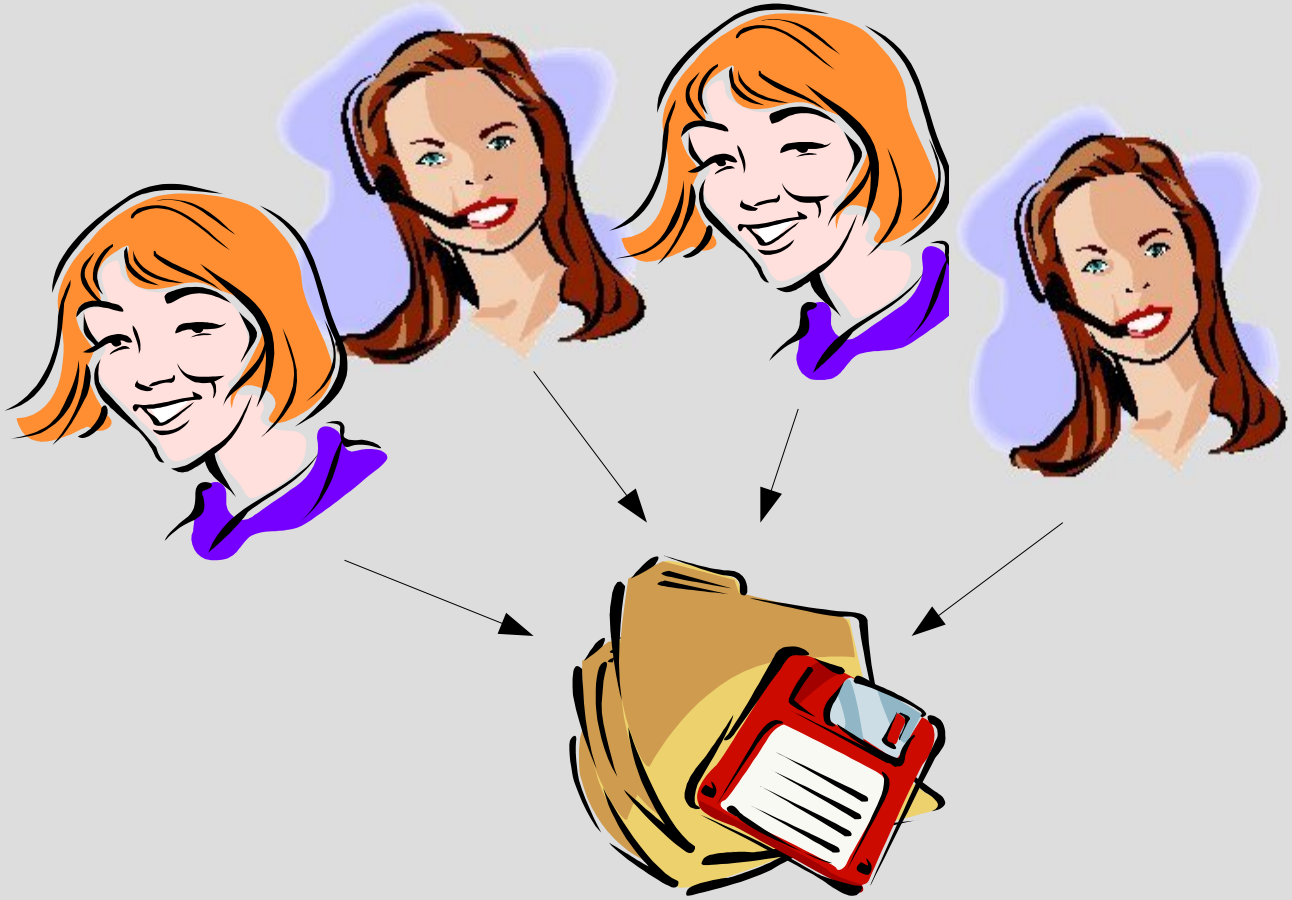
Entwicklung von Software

- Fachbereich stellt Anforderungen
- IT-Abteilung erstellt System
- Fachbereich nimmt fertige Software ab

(aber man weiß nicht so genau,
ob die Anforderungen erfüllt werden)

Programmierung

Programmierer erstellen Software

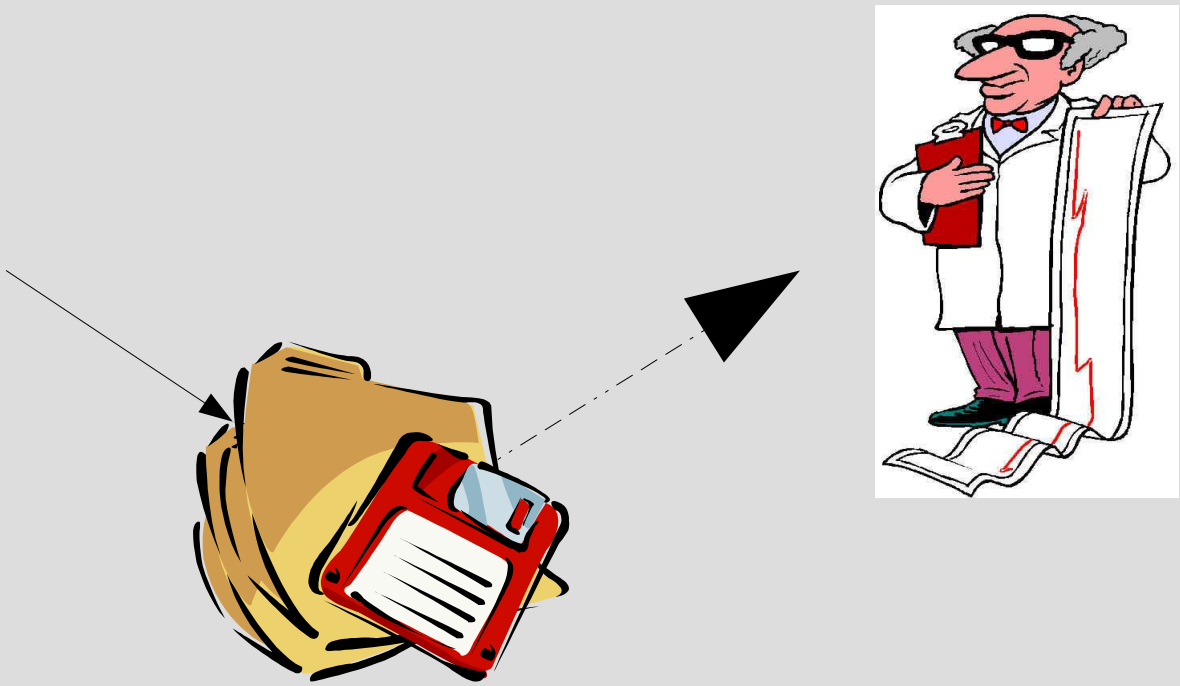


Quellcode A ist sauber

```
 identification division.
  program-id. exp.
  author. "JD".
  date-written. 25.10.2004.
  date-compiled.
  data division.
  working-storage section.
  77 a pic 9(4).
  77 c pic 9(4).
  linkage section.
  77 i   pic 9(4).
  77 n   pic 9(4).
  77 r   pic 9(4).
  procedure division using in r.
* computes i to the power of n
     move 1 to a c.
     perform until n <= c
       multiply a by i
       add 1 to c
     end-perform .
     move a to r.
    end-program exp.
```

Abnahme

Fachbereich testet und nimmt Software ab

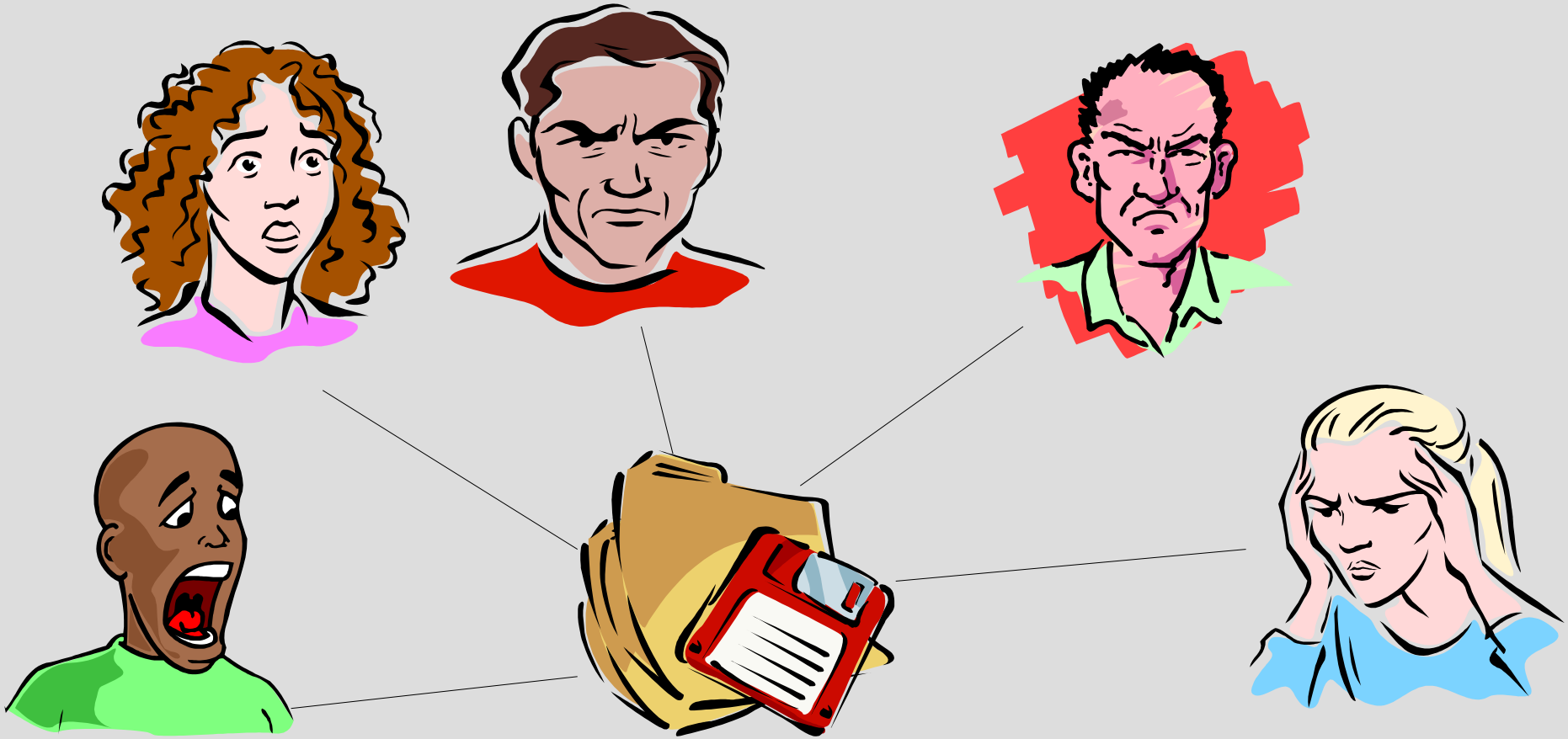


Monate und Jahre danach

- Fachbereich fordert Änderungen
- IT-Abteilung ändert Software
- und so weiter ...

Erweiterung und Wartung

Verschiedene Entwickler ändern die Software



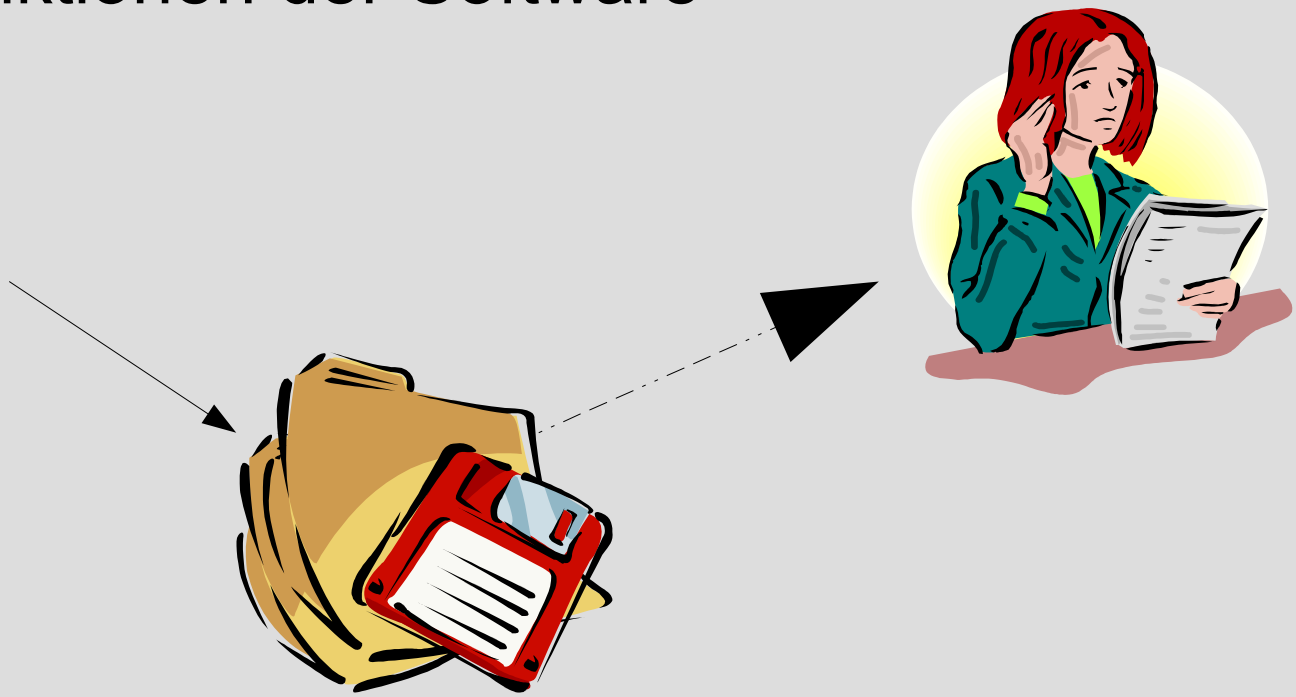
Quellcode B ist nicht mehr erkennbar

```
       identification division.
        program-id. exp-coded.
        author. "JD".
        date-written. 25.10.2004.
        date-compiled.
        data division.
        working-storage section.
       ******************************************************
       * Umbenennung von Franz
       ******************************************************
        77 achtung     pic 9(4).
        77 coffein     pic 9(4).
        77 dieter      pic 9(4).
        77 moechte     pic 9(4).
        77 gruenen     pic x(28).
        77 tee         pic 9(4).
        78 kaffee      value 0.
        78 wasser      value 1.
        78 whiskey     value 2.
        linkage section.
        77 innovation      pic 9(4).
        77 neuigkeiten     pic 9(4).
        77 richtunggebend pic 9(4).
        procedure division using innovation neuigkeiten
                                 richtunggebend.
       * also computes r = i to the power of n
       * but redundantly
       * i -> innovation
       * n -> neuigkeiten
       * r -> richtunggebend
       ******************************************************
       * Verstreuung von Armin
       ******************************************************
```

```
        A11    move kaffee to dieter
        A11    move kaffee to tee
               move dieter to achtung coffein
               add  wasser to achtung
               add  wasser to coffein
               move wasser to dieter
       ******************************************************
       * Klausulierung von Arno
       ******************************************************
        033      perform until coffein
                           > neuigkeiten + kaffee
       ******************************************************
       * Indirektion von Norbert
       ******************************************************
        N22        move  wasser to tee
        N22        move innovation      to moechte
        N22        multiply achtung by moechte
        N22        add tee       to coffein
        N22        move     wasser to tee
                   subtract wasser from  tee
                 end-perform.
                 move achtung to richtunggebend.
       ******************************************************
       * Verschleierung von Ute
       ******************************************************
                 if tee < - whiskey * dieter
                   subtract wasser from  tee
                 end-if.
               end-program  exp-coded.
```

Dilemma des Fachbereichs

Unternehmen ist unsicher über die Funktionen der Software



Fachbereiche

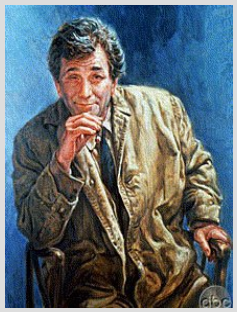
- Fachbereiche sprechen
  natürlich,
  fachspezifisch,
  aber nicht formal
- sie erwarten
  Qualität
- sie möchten
  ihre Geschäftsregeln wiederfinden

Lösung der Aufgabe

- Columbo ist ein Übersetzer **von formalen Sprachen** in natürliche Sprachen
- Columbo erzeugt nicht operationale , sondern **denotationale Semantik**
- Daraus wird **natürliche Sprache** generiert

Aufgabe für Columbo

Columbo erstellt Nachdokumentation der Software



Columbo 1.0

Columbo-Ergebnis A

```
 identification division.
 program-id. exp.
 author. "JD".
 date-written. 25.10.2004.
 date-compiled.
 data division.
 working-storage section.
 77 a pic 9(4).
 77 c pic 9(4).
 linkage section.
 77 i  pic 9(4).
 77 n  pic 9(4).
 77 r  pic 9(4).
 procedure division using i n r.
* computes i to the power of n
    move 1 to a c.
    perform until n <= c
      multiply a by i
      add 1 to c
    end-perform.
    move a to r.
  end-program exp.
```

```
                Analyseergebnis für exp
            ------------------------

Funktion
--------

A) Das Programm (r) = exp (i, n)

    ermittelt r = (i**n)


B) oder das Programm

    ermittelt r = 1.


Komplexität
-----------

  Das Verfahren hat folgendes Zeitverhalten

     O(n).


Randbedingungen
---------------

  Dies gilt nur, wenn die Variablen im Intervall

    [0 .. 10000]

  liegen.
```

Columbo-Ergebnis B

```
       Identification division.
       program-id. exp-coded.
       author. "JD".
       date-written. 25.10.2004.
       date-compiled.
       data division.
       working-storage section.
      ******************************************************
      * Umbenennung von Franz
      ******************************************************
       77 achtung    pic 9(4).
       77 coffein    pic 9(4).
       77 dieter     pic 9(4).
       77 moechte    pic 9(4).
       77 gruenen    pic x(28).
       77 tee        pic 9(4).
       78 kaffee     value 0.
       78 wasser     value 1.
       78 whiskey    value 2.
       linkage section.
       77 innovation     pic 9(4).
       77 neurigkeiten   pic 9(4).
       77 richtunggebend pic 9(4).
       procedure division using innovation neurigkeiten
                                richtunggebend.
      * also computes r = i to the power of n
      * but redundantly
      * i -> innovation
      * n -> neurigkeiten
      * r -> richtunggebend
      ******************************************************
      * Verstreuung von Arm in
      ******************************************************
A11    move kaffee to dieter
A11    move kaffee to tee
       move dieter to achtung coffein
       add  wasser to achtung
       add  wasser to coffein
       move wasser to dieter
      ******************************************************
      * Klausulierung von Arno
      ******************************************************
O33    perform until coffein
                   > neurigkeiten + kaffee
      ******************************************************
      * Indirektion von Norbert
      ******************************************************
N22      move  wasser to tee
N22      move innovation    to moechte
N22      multiply achtung by moechte
N22      add tee      to coffein
N22      move     wasser to tee
         subtract wasser from tee
       end-perform.
       move achtung to richtunggebend.
      ******************************************************
      * Verschleierung von Ute
      ******************************************************
         if tee < - whiskey * dieter
           subtract wasser from tee
         end-if.
       end-program exp-coded.
```

Analyseergebnis für exp-coded

Funktion

A) Das Programm (richtunggebend) = exp-coded (innovation, neurigkeiten)

ermittelt richtunggebend = (innovation**neurigkeiten)

B) oder das Programm

ermittelt richtunggebend = 1

Komplexität

Das Verfahren hat folgendes Zeitverhalten

O(n).

Randbedingungen

Dies gilt nur, wenn die Variablen im Intervall

[0 .. 10000]

Liegen.

Vorteile

Vorteile

- Unterstützt die Entwicklung
- Objektiviert die Qualität
- Schafft Sicherheit

Nachteil

- Erzeugt Leistungsdruck

Konkurrenzprodukte

- Revolve
- Polyspace
- DRDA
- ...

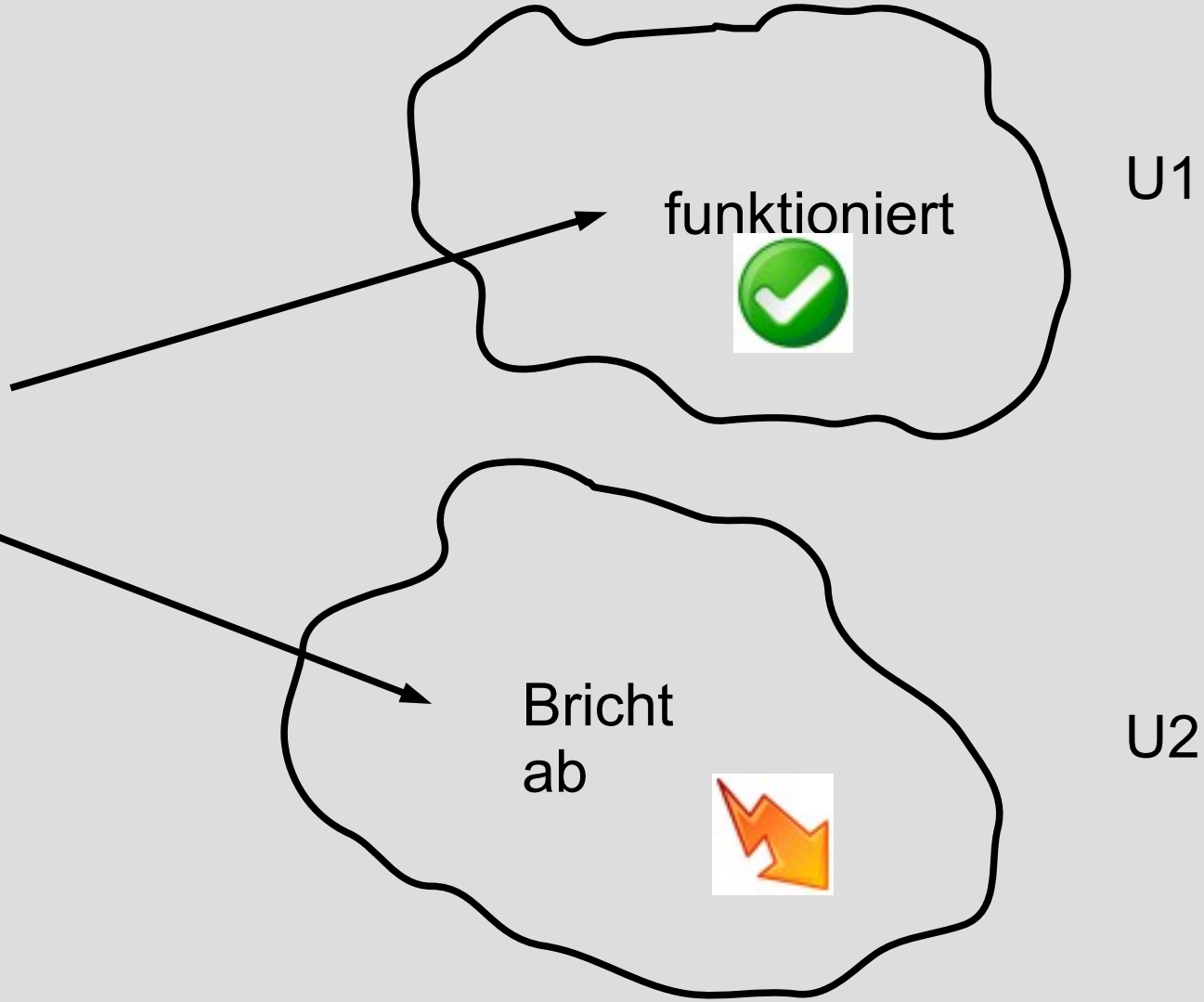
Statische Analyse

- Undefinierte Variablen
- Division durch 0 etc.
- Indexüberlauf
- Unreachable Code
- Constant Propagation
- Impact Analysis
- Control Flow
- Data Flow
- Style Check
- Anweisungsebene (sehr wenige)
- Automatischer Selbsttest
- ...

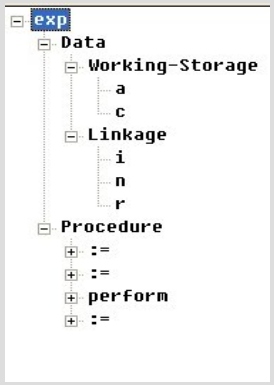
Umgebungen

```
identification division.
program-id. exp.
author. "JD".
date-written. 25.10.2004.
date-compiled.
data division.
working-storage section.
77 a pic 9(4).
77 c pic 9(4).
linkage section.
77 i   pic 9(4).
77 n   pic 9(4).
77 r   pic 9(4).
procedure division using i n r.
* computes i to the power
of n
  move 1 to a c.
  perform until n <= c
    multiply a by i
    add 1 to c
  end-perform.
  move a to r.
end-program exp.
```

Quelle, verifiziert oder validiert

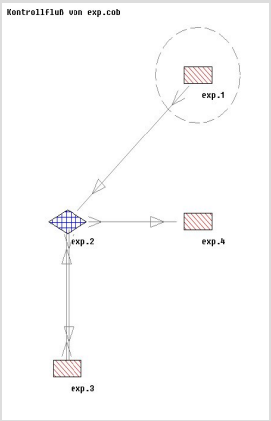


Sichten auf die Quelle



Navigator

```
identification division.
program-id. exp.
author. "JD".
date-written. 25.10.2004.
date-compiled.
data division.
working-storage section.
77 a pic 9(4).
77 c pic 9(4).
linkage section.
77 i  pic 9(4).
77 n  pic 9(4).
77 r  pic 9(4).
procedure division using i n r.
* computes i to the power
of n
  move 1 to a c.
  perform until n <= c
    multiply a by i
    add 1 to c
  end-perform.
  move a to r.
end-program exp.
```

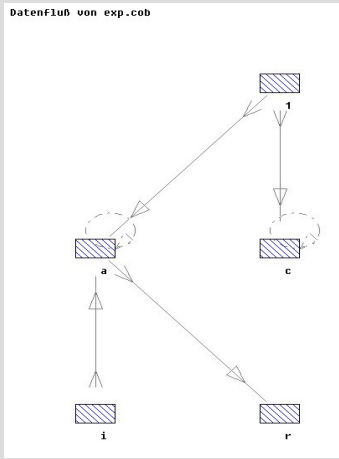


CFA

```
Datei "exp.cob"
==============

Prozedur exp
-----------
Statistik
  Anzahl Zeilen:                20
  Anzahl Kommentare:             5 %
  Anzahl Statements:             5
  Anzahl Felder:                 5
  Analysezeit:                   0 msec
  Speicherbedarf:             5509 kB
Komplexität
  Zyklomatische Komplexität:     2
  McCabe  Maß:                   2
Aufwand
  Halstead Maß:                100
  Function Point Aufwand:      0.8 PM
```

Statistik

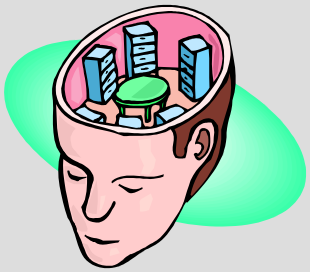


DFA

Semantik

```
identification division.
program-id. exp.
author. "JD".
date-written. 25.10.2004.
date-compiled.
data division.
working-storage section.
77 a pic 9(4).
77 c pic 9(4).
linkage section.
77 i   pic 9(4).
77 n   pic 9(4).
77 r   pic 9(4).
procedure division using i n r.
* computes i to the power
of n
  move 1 to a c.
  perform until n <= c
    multiply a by i
    add 1 to c
  end-perform.
  move a to r.
end-program exp.
```

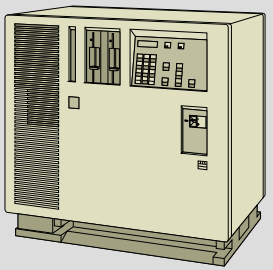
denotational



axiomatisch

$$e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!} + i \sum_{n=1}^{\infty} \frac{(-1)^{n-1} x^{2n-1}}{(2n-1)!} = \cos x + i \sin x.$$

operational



Sprachklassen

Prozedural (A)
COBOL
JAVA
C++
Assembler
Maschinensprache

Mengenorientiert (B)
Prolog
SQL
PL(1-2)
~Standard ML

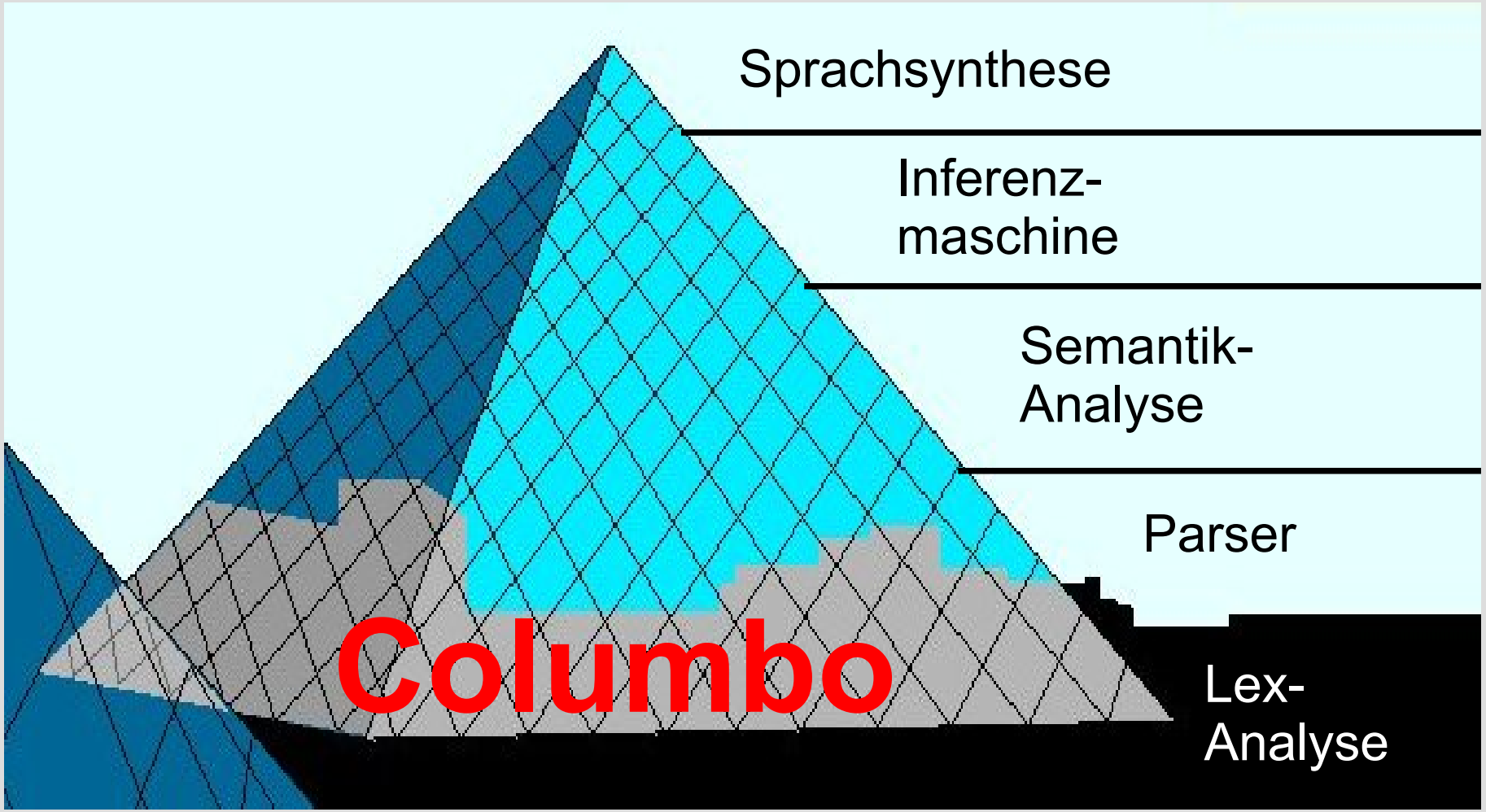
COLUMBO soll von A nach B übersetzen

Qualitätssicherung

Columbo wird mittels dieser Methoden geprüft:

- Automatischer Selbsttest
- Manueller Test
- Style Tests
- Code Review
- Decompilation
- ...

Architektur von Columbo



Nächste Schritte

- Prototyp zum Produkt machen
  Arrays, Pointer, Datentypen,
  Dateien, SQL, ...
- Aufteilung in
  Free, Professional und Enterprise Version
- Qualitätssicherung durchführen
- Pilotanwender oder -kunden finden

Timeline

– Ende 2005	Prototyp
– in 2006	Free Version
– in 2006	Web-Interface
– Ende 2006	Professional Version

Ziel

Verifikationssoftware
Perfekte Dokumentation

Fragen:
- Mission (im)possible ?
- wie sichert man hier die Qualität ?