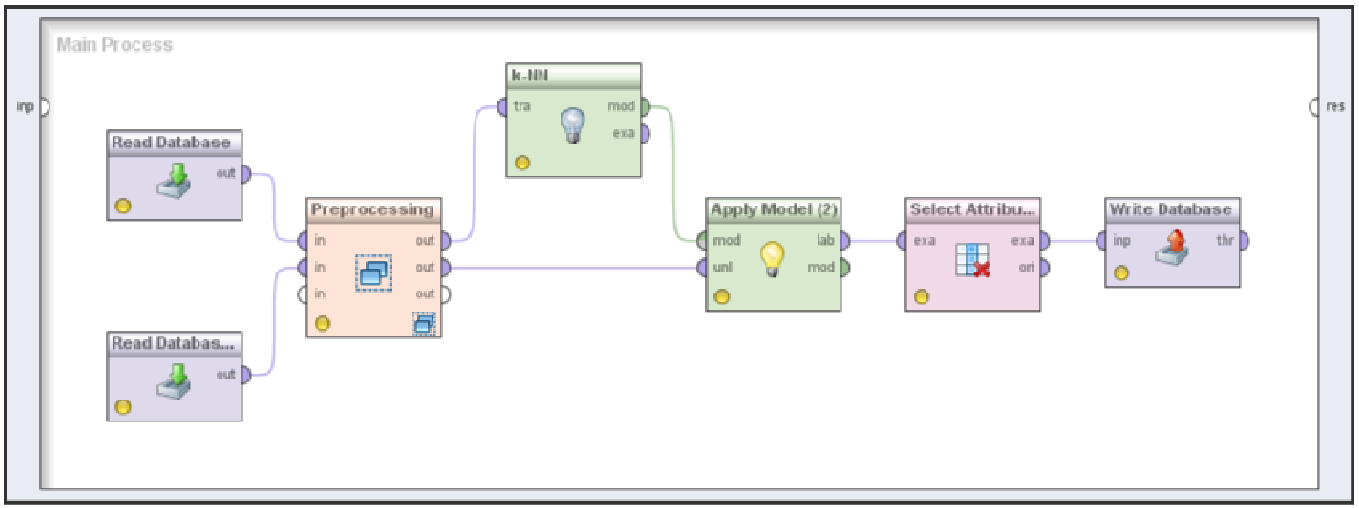
Meta-Data Mining v Interaktivní Evoluci

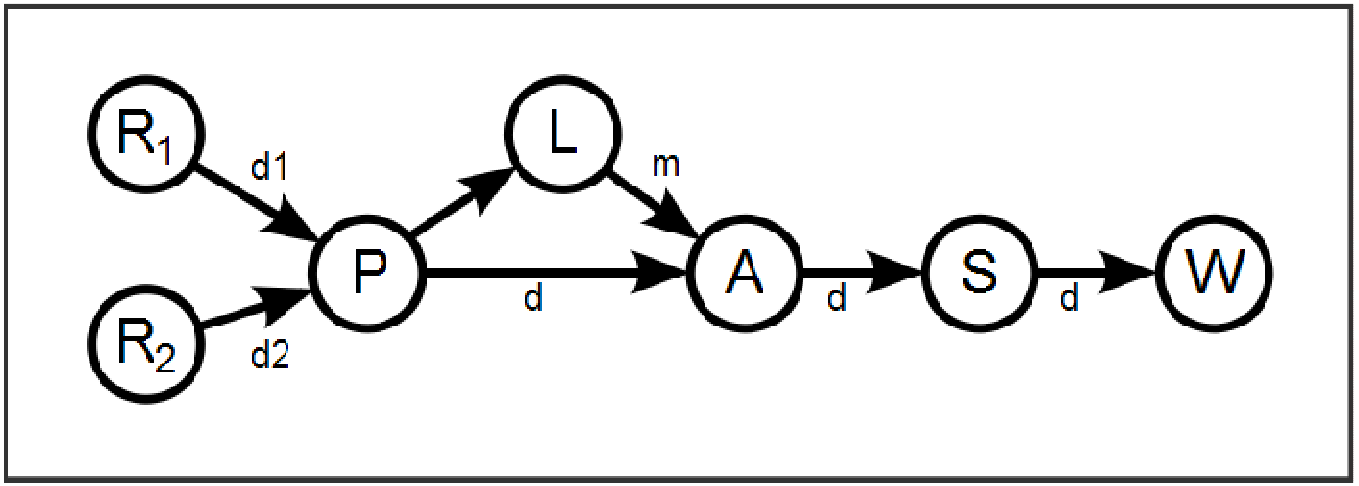
Tomáš Řehořek

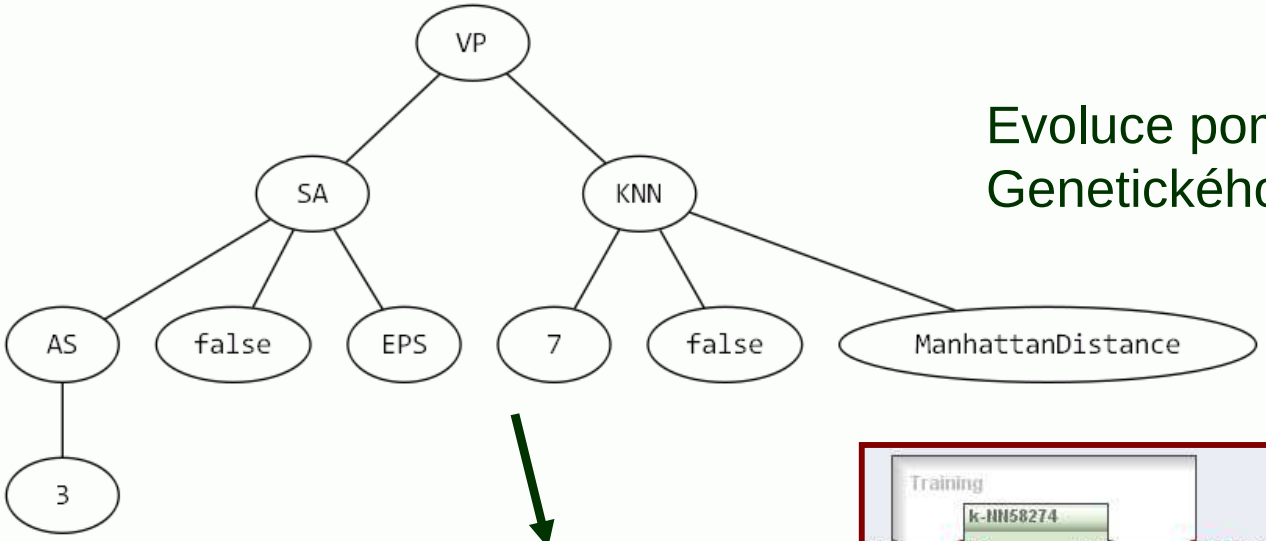
Tomáš Řehořek

- 2006 – 2009: FEL STM, obor SW inženýrství
  - především návrh webových aplikací
  - zlomový bod: Y36VD s Pavlem Kordíkem
  - BP: JavaScriptová grafická knihovna, Animace evoluce GAME sítě
- 2009 – 2011: FEL OI, obor Umělá Inteligence
  - DP: Evoluce procesů
    - Automatická konstrukce klasifikátoru v RapidMineru
    - Interaktivní evoluce pro data mining

Evoluce klasifikátoru pro RapidMiner



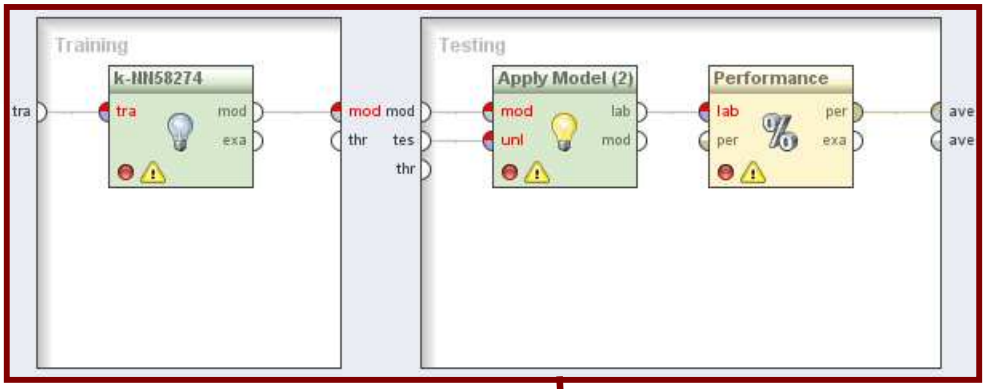


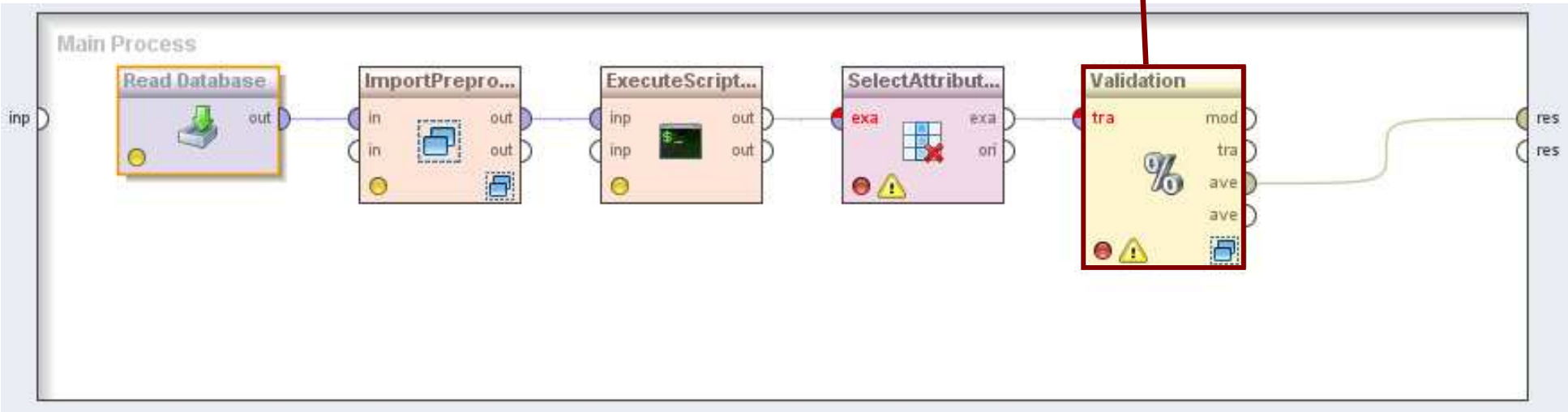


Evoluce pomocí
Genetického Programování

XML

Fitness = klasifikační přesnost
(cross-validace)





Léto 2011: Doporučovací engine pro nangu.TV

- Analýza a návrh **doporučovacího systému** pro IPTV platformu nangu.TV
- SQL dump o velikosti 11 GB
- Jsou k dispozici trojice

  (uživatel, film, počet shlédnutí)
- Pro daného uživatele máme doporučit film, který ještě neviděl, ale mohl by se mu líbit
- De facto variace Netflix Prize

User id	Movie id	Count
65869	265	**1**
66575	748	**1**
75008	7911	**1**
68715	8097	**1**
78464	11043	**1**
72535	1488	**1**
73075	1009	**2**
70225	4296	**1**
65106	616	**1**
72811	5174	**1**
70392	5197	**2**
75307	5877	**1**
73255	6967	**1**
78509	13540	**2**
74762	4783	**1**
66487	684	**1**
72434	2057	**1**
72342	5538	**1**
78540	5930	**1**
70402	1980	**1**

User id	Movie id	Count
65963	257	**2**
75837	5892	**1**
72388	1173	**1**
75334	2167	**1**
76310	8041	**1**
66112	5614	**1**
77406	5876	**1**
562	558	**1**
78509	9576	**1**
76063	13389	**2**
77477	5708	**1**
73660	5755	**2**
79069	12659	**1**
76459	7291	**1**
70731	1331	**1**
72889	4951	**2**
71893	1496	**2**
75307	6837	**1**
69057	1495	**2**
66095	5930	**1**

User id	Movie id	Count
71693	7884	**1**
562	844	**1**
76430	2386	**2**
65455	4896	**2**
68880	5947	**1**
72089	14348	**1**
76387	7143	**1**
73369	12220	**1**
78509	4806	**3**
4886	1015	**1**
69724	3375	**1**
69742	760	**3**
67643	3366	**1**
76114	1306	**1**
72342	9841	**1**
77636	9316	**1**
72915	12976	**1**
78351	9804	**1**
79532	12954	**1**
73372	759	**1**

… a 53,412 dalších

Problémy s datasetem

- Standardní Netflix reprezentace: matice Uživatelé × Filmy
- Přibližně 5000 uživatelů a 5000 filmů
  - Dataset má 5000 atributů!
- Původní snaha: Převod na Netflix
- Zjištění: Data mají principiální nedostatky a Netflix metody zatím nejsou použitelné
- Prozatímní řešení: Asociační pravidla

Asociační pravidla v nangu.TV

- Vstup: matice binárních hodnot
  Uživatelé × Filmy

- Výstup: pravidla ve tvaru

{ Movie#7, Movie#11 } ==60%==> { Movie#5 }

nangu.TV: První výsledky

Association rule #1: support=***0.006347***, confidence=***0.603774***
{ Elephants Dream [id:6, zanr: N/A] }
=>
{ Big Buck Bunny [id:415, zanr: Komedie] }

Association rule #1508: support=***0.005157***, confidence=***0.787879***
{ Imperium - Mafie v Atlantic City (7) [id:9796, zanr: N/A] }
=>
{ Imperium - Mafie v Atlantic City (3) [id:9148, zanr: N/A],
Imperium - Mafie v Atlantic City (5) [id:9251, zanr: N/A] }

Association rule #1509: support=***0.005157***, confidence=***0.838710***
{ Imperium - Mafie v Atlantic City (5) [id:9251, zanr: N/A] }
=>
{ Imperium - Mafie v Atlantic City (3) [id:9148, zanr: N/A],
Imperium - Mafie v Atlantic City (7) [id:9796, zanr: N/A] }

Intermezzo: DM v jiných vědách

- Biologie – analýza chování organismů
- Chemie – vývoj molekul
- Ekonomie, Meteorologie, Geologie, Astronomie, Informatika, Medicína…

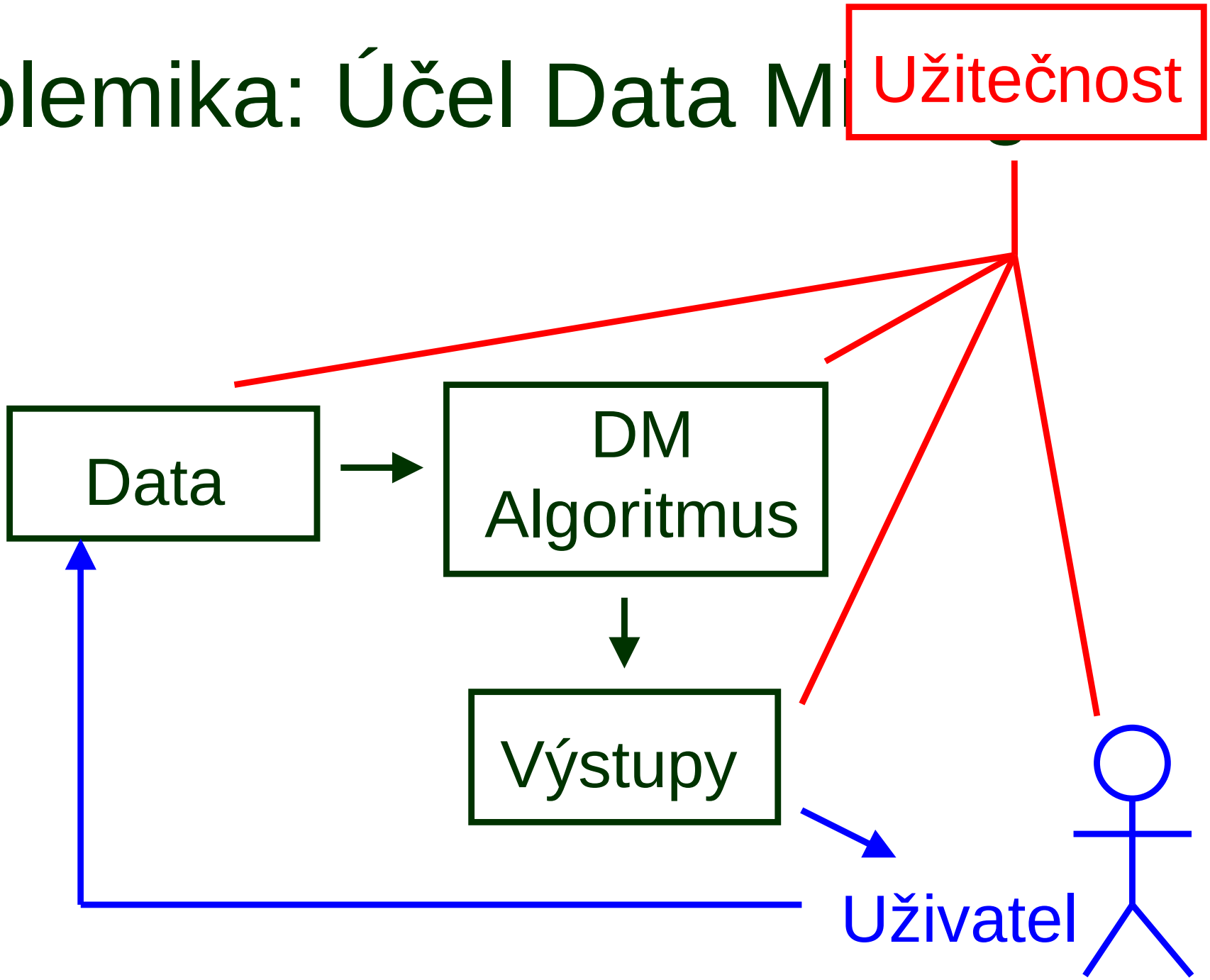
…and the hat still goes deeper ☺

Velký potenciál pro spolupráci s obrovským množstvím oborů!

Intermezzo: DM v jiných vědách

- Problémy:
  - nutnost přizpůsobování DM metod
  - nepřeberné množství možných aplikací
- Jak zpřístupnit DM širšímu okruhu uživatelů?
  - každý obor vyžaduje specifické pohledy na specifická data

Polemika: Účel Data Mi



Hypotéza

- Data Mining je kolekce uživatelů, dat, nakonfigurovaných algoritmů a jejich výstupů takových, že DM algoritmus (včetně jeho konfigurace) na daných datech vrátí výstupy, které mají pro uživatele v danou dobu velkou informační hodnotu.

Příklady: Meteorolog

- **Data**: kolekce hodnot naměřených na různých zeměpisných souřadnicích v různých časech
- **Algoritmus**: Predikce oblačnosti a hustoty srážek dne 9. 9. 2011, 16:00 v okolí FIT
- **Výstupy**: Graf srážek a oblačnosti viditelný na mapě

Příklady: Biolog

- **Data**: kolekce záznamů o pavoucích, plošticích, jejich barvách a interakcích v čase
- **Algoritmus**: Detekce korelací s velkou absolutní hodnotou
- **Výstup**: Korelace mezi barvou ploštice a počtem útoků pavouka v rostoucím čase

Boom Internetu

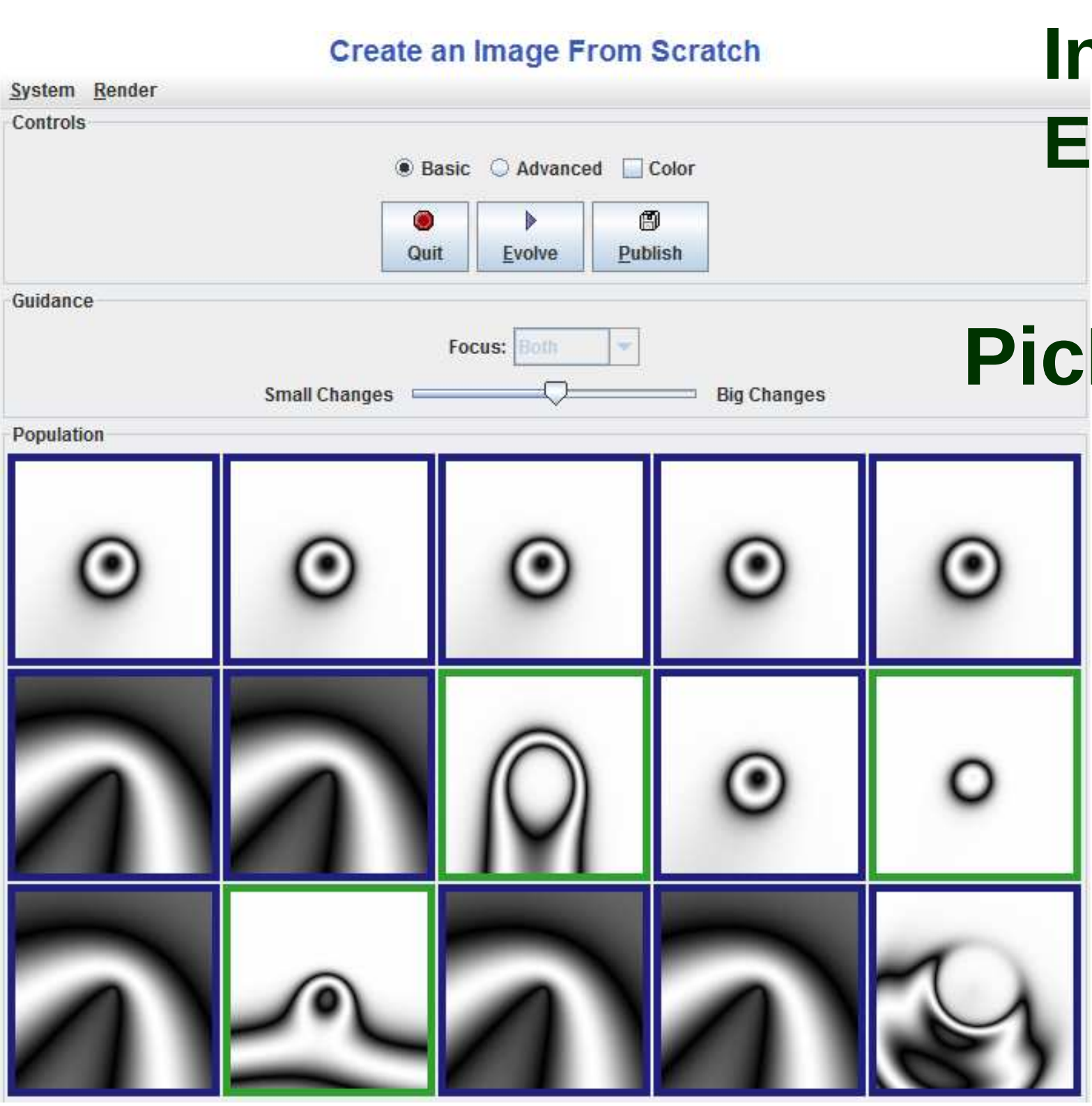
- Moderní internetové technologie umožňují vytvoření distribuovaného „biologického počítače“ ☺
- Wikipedia, Open Source: vznik extrémně rozsáhlých a propracovaných produktů

Kolaborativní Interaktivní Evoluce

- Běh evolučního algoritmu, kde fitness určuje uživatel – člověk
- Populární příklad: PicBreeder.org od Kena Stenleyho

Interactive Evolution

PicBreeder



by

Jimmy Secretan

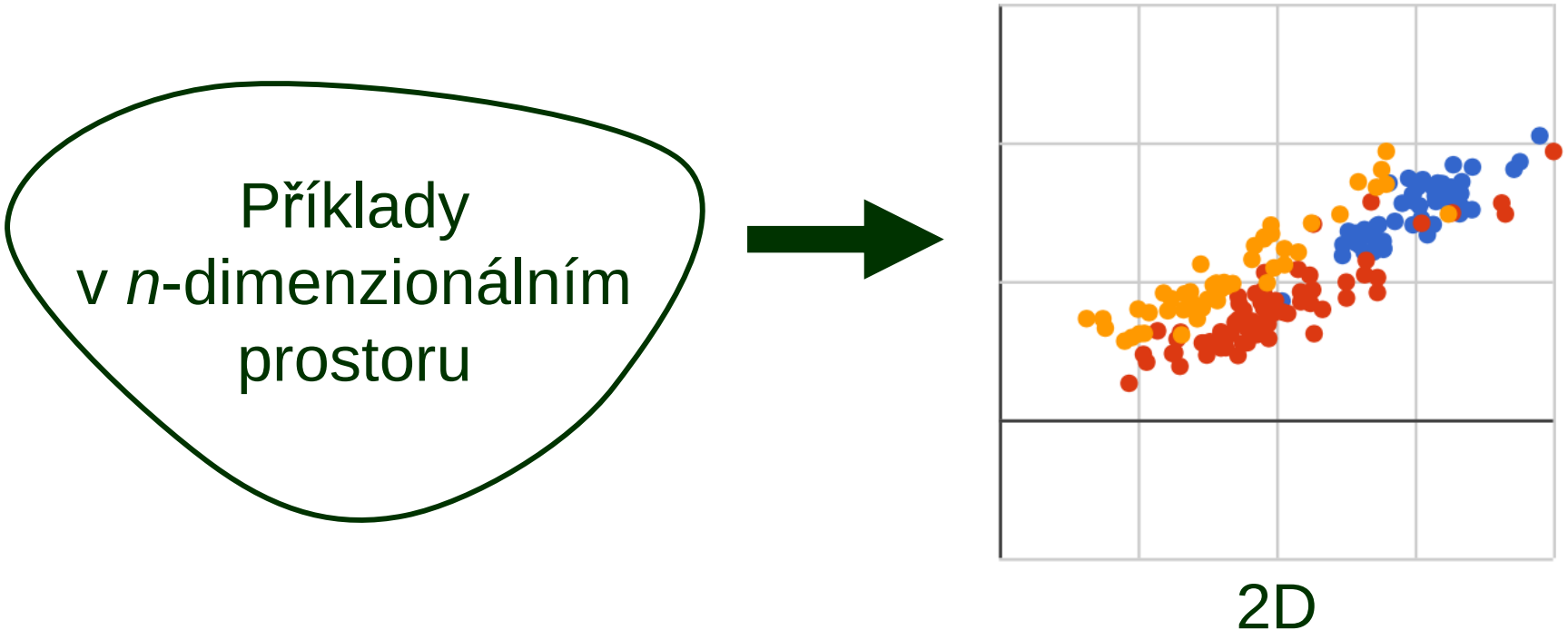
Kenneth Stanley

And after 75 generations ...

... you eventually get something interesting

Experiment: Projekce dat

- Transformace data-setu do 2D $f : \mathbb{R}^n \rightarrow \mathbb{R}^2$



- Podobné PCA, Sammonově projekci atd.

Interactive Evolution: Kernel-PCA Projection - Mozilla Firefox

Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda

http://localhost:8080/rm-web/EvolverServlet

Google.com (in English)

Interactive Evolution: K...

Interactive Evolution: Linear transformation

Experiment setup

Problem: Linear transformation **Data:** Wine

Population size: 7 **Tournament size:** 3

Mutation rate: 40 **Crossover rate:** 40 **Duplication rate:** 20

Feedback sample size: 10 **Feedback sample type:** random

A priori fitness: constant=1, only feedback taken into account

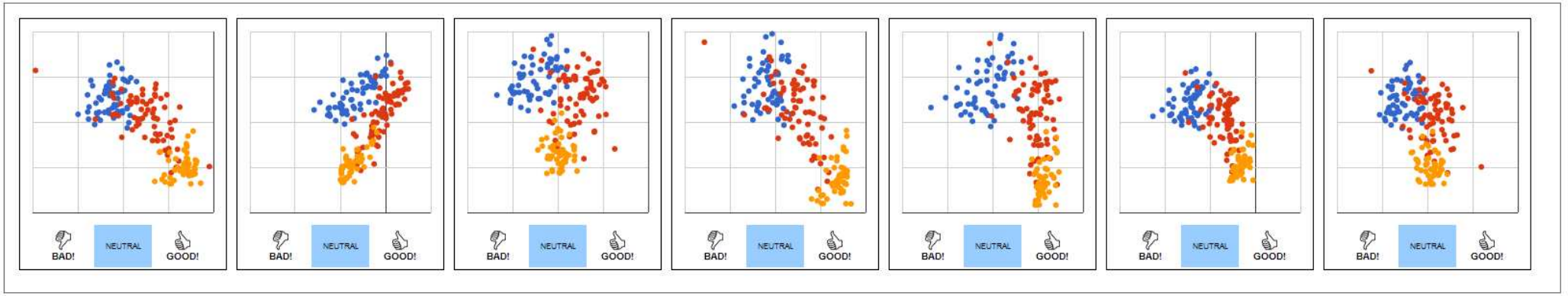
Negative feedback effect: Fitness = infinity

Positive feedback effect: Fitness = 0

Initial coefficient value: 1

Initialize Reset

Candidate solutions



Sumbit & Iterate

Nápad: Optimalizace s heuristikou

- V každém kroku:
  1. Nabídni uživateli aktuální řešení,
  2. Posbírej zpětnou vazbu,
  3. Ulož poznatek do databáze,
  4. Na základě dosavadních poznatků
     v databázi nabídni kandidátské řešení
     s nejvyšší očekávanou fitness.

Heuristika: Meta-Data Mining

- **Data**: kolekce pětic:
  - uživatel,
  - datový soubor, metadata,
  - algoritmus a jeho konfigurace,
  - výsledky,
  - koeficient užitečnosti (dodává fyzický uživatel).
- **Algoritmus**: Detekce shluků uživatelů, kteří mají podobné zájmy
- **Výstup**: Prediktivní model uživatelů doporučující pětice s vysokou očekávanou užitečností

Utopie?

- Doporučovací systém, kde:
  - Uživatelé kladou problémy,
  - Odborníci dodávají data-miningová řešení,
  - Uživatelé hodnotí různé konfigurace algoritmů,
  - Systém doporučuje uživateli vhodný algoritmus pro daná data na základě zkušeností.

Děkuji za pozornost ☺