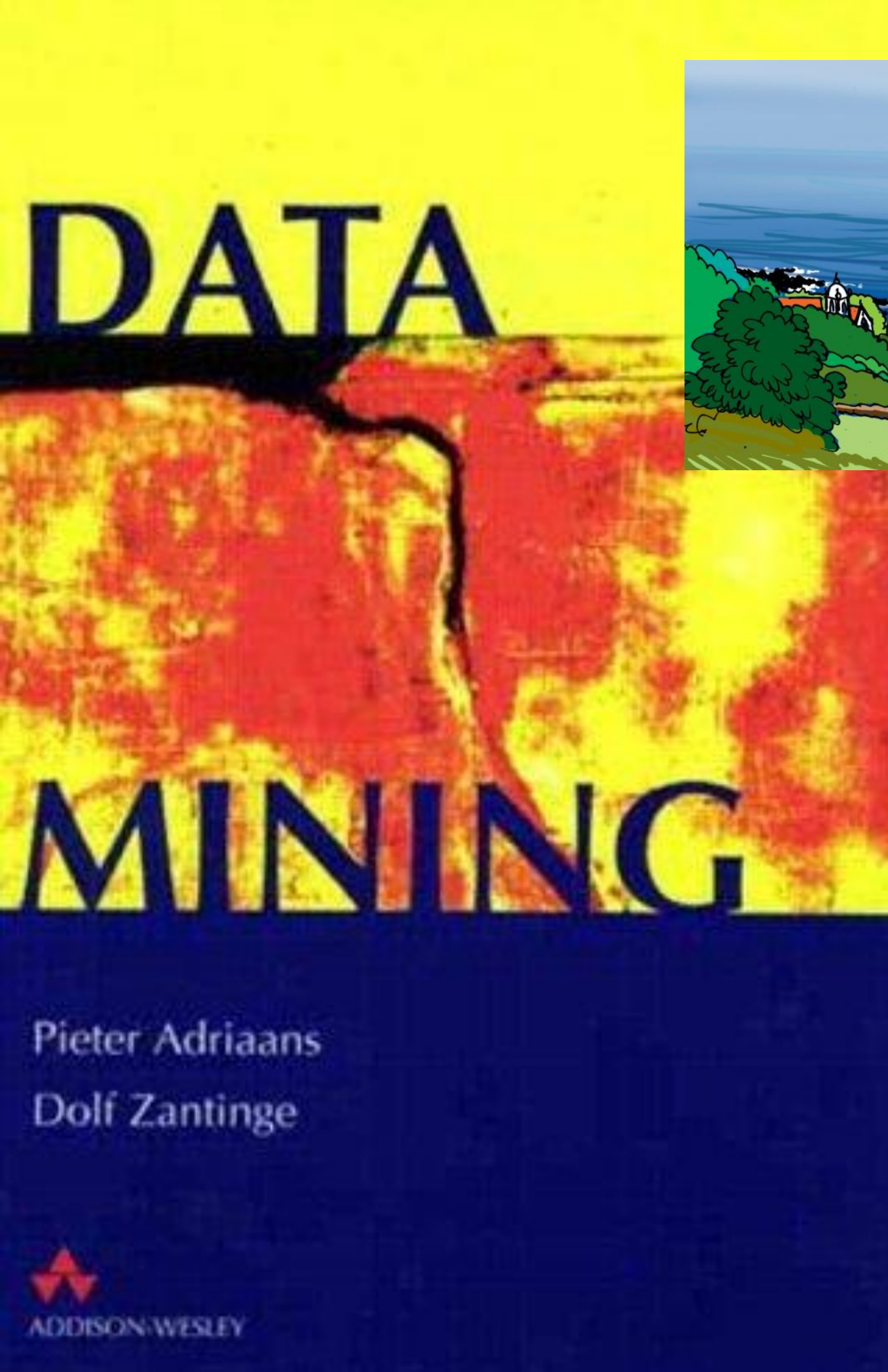


Looking Backward, Looking Forward

PTDM @ ICDM

Pieter Adriaans
Universiteit van Amsterdam

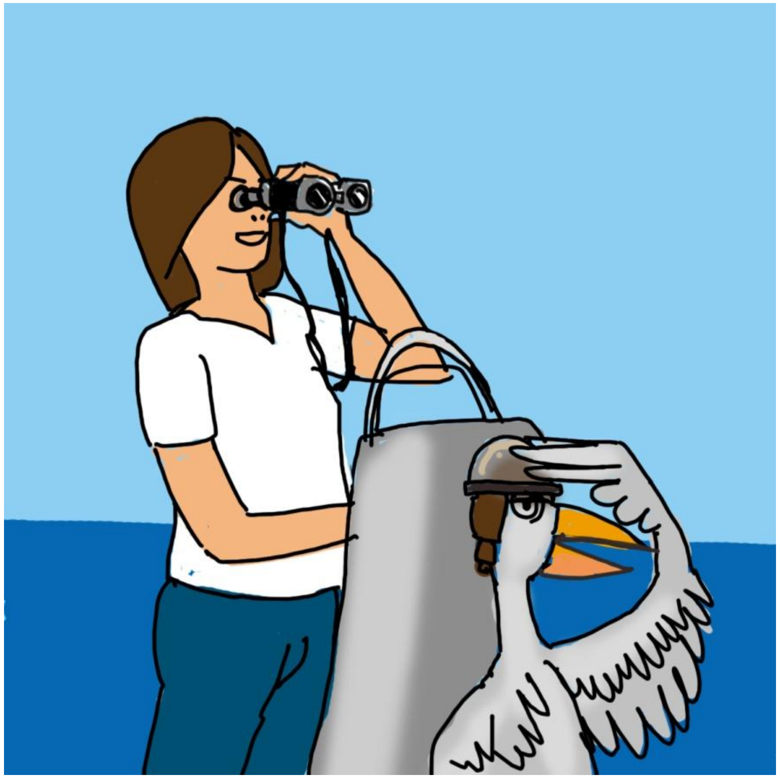

PEARSON Education
LOW PRICE EDITION
Data Mining
Pieter Adriaans
Dolf Zantinge
DATA MINING
Pieter Adriaans-Dolf Zantinge
Adatbányászat
TEXTS IN COMPUTER SCIENCE
An Introduction to Kolmogorov Complexity and Its Applications
Ming Li
Paul Vitányi
THIRD EDITION
Springer
Handbook of
THE PHILOSOPHY OF SCIENCE
PHILOSOPHY of INFORMATION
Edited by Pieter Adriaans
and Johan van Benthem
NH
DATA
Pieter Adriaans
Dolf Zantinge
Data Mining
Pieter Adriaans
Dolf Zantinge

Looking backward 1985-2012
More than 200 AI,ML,DM projects

- AI, Expert Systems, Prolog Data Base technology
(OBIS 1988, Captains, 1992)
- Data Mining: Workstation, Client Server, Mature ML research, large data bases
- Time dimension (ASM, ICT)
- Structured Data (GI, EMILE)
- Dynamic Systems (JSF, Robosail)
- E-science (Vl-e)
- Big Data (Commit, Data2Semantics)

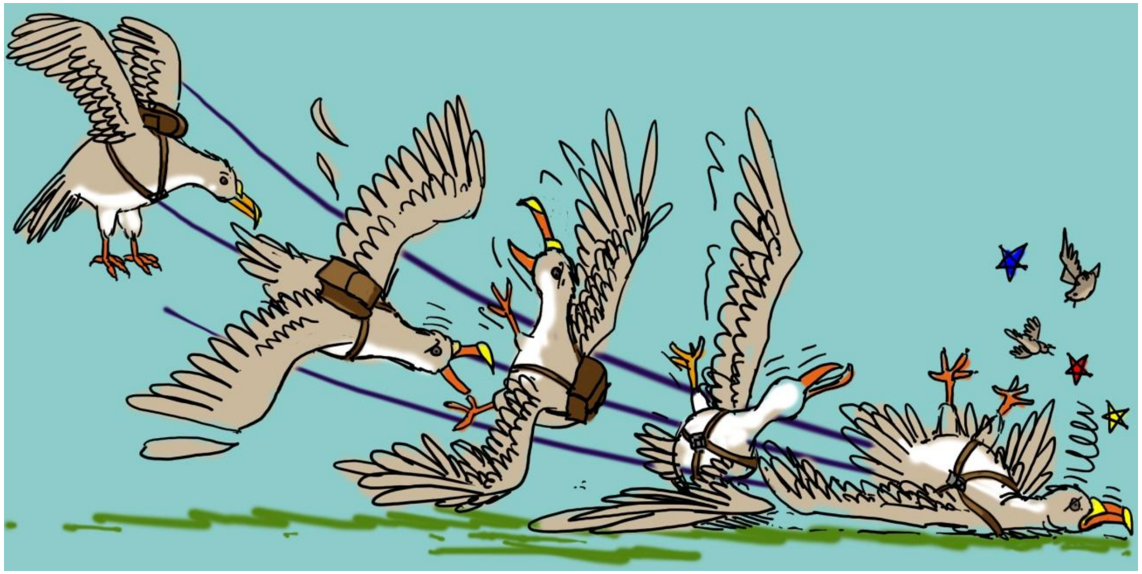
Looking forward

- Complexity measures: facticity
- Empirical incompleteness theories
- e-science methodology for the 21st century



1992: Captains

- Crew Availability planning KLM
- Oracle Database
- Prolog planning algorithms
- Genetic Algorithms for Bid Prediction

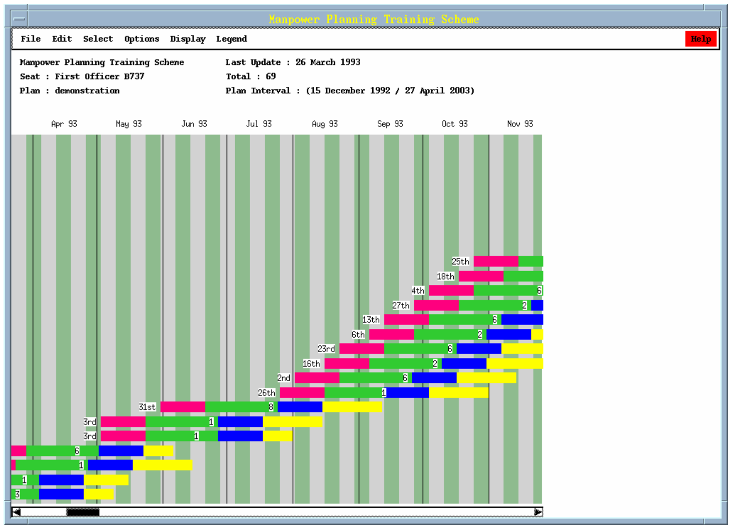


Plan board

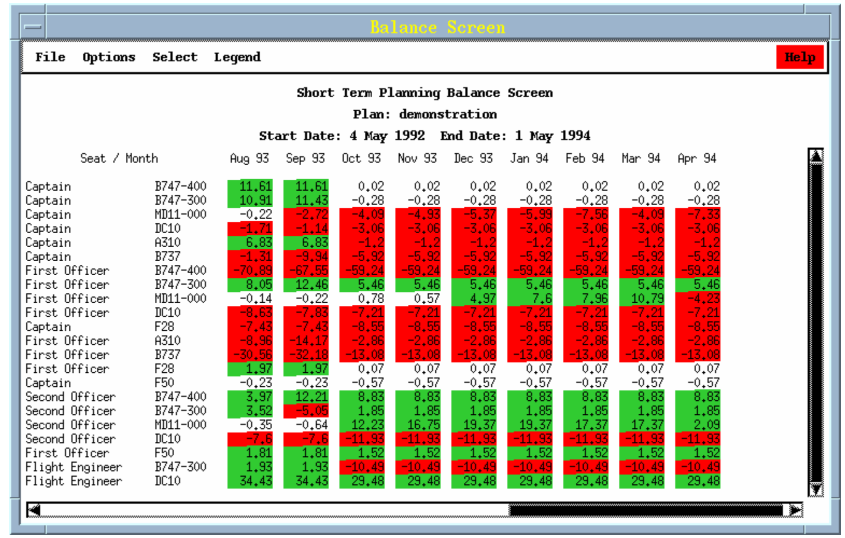
Seniority
Transitions
Horizontal/Vertical binding
Pilot bids (preferences)
Part Time
PlanBoard
File Edit Plan Options Query Legend Help
ACTUAL PLAN
75 KANIS 1
76 DUBOURCQ 1
77 SCHOEN 1
78 VOSKUIL 1
79 NAP 1
80 SPANJAART 1
81 SIERSMA 1
82 PLESMAN 2
83 WACHTER 1
84 HUIJGEN 2
85 SMIT 25
86 STUIK
87 WILLEMSE 7
88 PAYMANS 2
89 DELFT 2
90 SCHREUDERS 6
91 HOOGEBOOM 1
92 JONG 19
93 EYCK 3
94 RIJSWICK 3
112 SLOTEMA 1
113 KEUSS 1
114 MEULEN 6
115 WITKOP
116 VISSERS 5
117 POLL 2
118 BUYZER 1
119 VISSCHER 11
120 LAVERMAN
121 MULLER 1
122 ERBEN
123 BRUYN 9
124 VERWEY 2
125 BOSCH 10
126 TIMMERMANS 1
127 DYKSTRA 3
128 OLDENBURGER 1
129 GEYSENDORPHER 1
130 PFUNDT
131 VRIES 24
132 PLOMP 2
133 DUNKI JACOBS 1
134 ANDREA 1
135 HOOGENDOORN 1
136 VERWIEL 1
137 GILST 2
138 HOGERS 1
139 SIMOONS 5
140 OUD 1
141 BOEYENGA
142 WASSENAAR 1
143 WESDORP
144 WESTER 1
145 NOOD
146 FRANCHIMON
147 BAKKES 1
148 IDEMA 2

Manpower planning

Manpower Planning

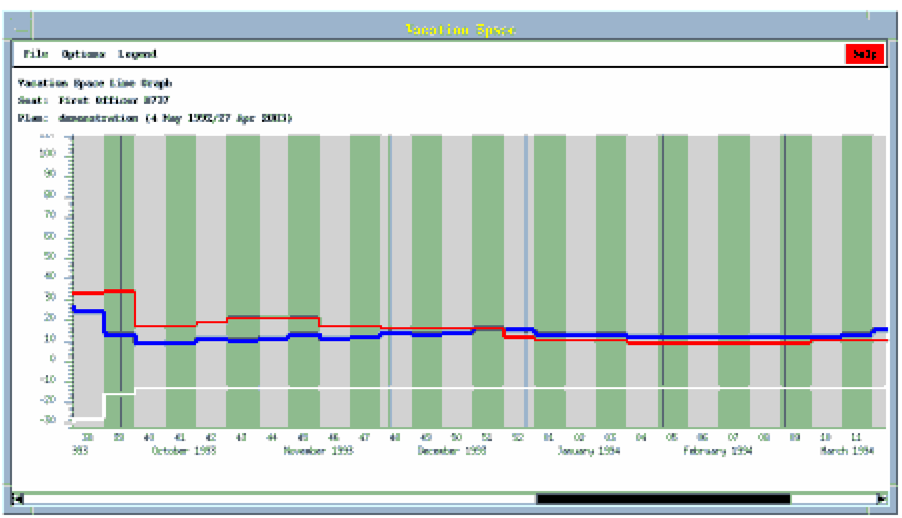


Seat Survey

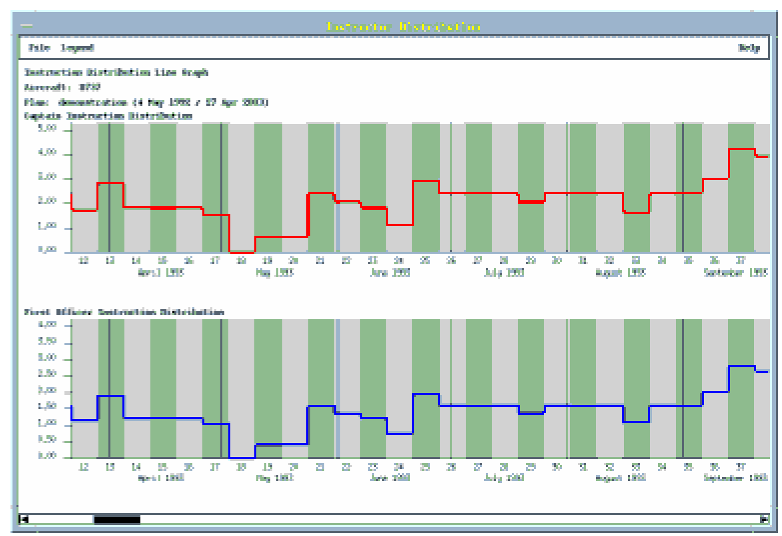


Seat / Month		Aug 93	Sep 93	Oct 93	Nov 93	Dec 93	Jan 94	Feb 94	Mar 94	Apr 94
Captain	B747-400	11.61	11.61	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Captain	B747-300	10.91	11.43	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28
Captain	MD11-000	-0.22	-2.72	-4.09	-4.93	-5.37	-5.99	-7.56	-4.09	-7.33
Captain	DC10	-1.71	-1.14	-3.06	-3.06	-3.06	-3.06	-3.06	-3.06	-3.06
Captain	A310	6.83	6.83	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2	-1.2
Captain	B737	-1.31	-9.94	-5.92	-5.92	-5.92	-5.92	-5.92	-5.92	-5.92
First Officer	B747-400	-70.89	-67.55	-59.24	-59.24	-59.24	-59.24	-59.24	-59.24	-59.24
First Officer	B747-300	8.05	12.46	5.46	5.46	5.46	5.46	5.46	5.46	5.46
First Officer	MD11-000	-0.14	-0.22	0.78	0.57	4.97	7.6	7.96	10.79	-4.23
First Officer	DC10	-8.63	-7.83	-7.21	-7.21	-7.21	-7.21	-7.21	-7.21	-7.21
Captain	F28	-7.43	-7.43	-8.55	-8.55	-8.55	-8.55	-8.55	-8.55	-8.55
First Officer	A310	-8.96	-14.17	-2.86	-2.86	-2.86	-2.86	-2.86	-2.86	-2.86
First Officer	B737	-30.56	-32.18	-13.08	-13.08	-13.08	-13.08	-13.08	-13.08	-13.08
First Officer	F28	1.97	1.97	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Captain	F50	-0.23	-0.23	-0.57	-0.57	-0.57	-0.57	-0.57	-0.57	-0.57
Second Officer	B747-400	3.97	12.21	8.83	8.83	8.83	8.83	8.83	8.83	8.83
Second Officer	B747-300	3.52	-5.05	1.85	1.85	1.85	1.85	1.85	1.85	1.85
Second Officer	MD11-000	-0.35	-0.64	12.23	16.75	19.37	19.37	17.37	17.37	2.09
Second Officer	DC10	-7.6	-7.6	-11.93	-11.93	-11.93	-11.93	-11.93	-11.93	-11.93
First Officer	F50	1.81	1.81	1.52	1.52	1.52	1.52	1.52	1.52	1.52
Flight Engineer	B747-300	1.93	1.93	-10.49	-10.49	-10.49	-10.49	-10.49	-10.49	-10.49
Flight Engineer	DC10	34.43	34.43	29.48	29.48	29.48	29.48	29.48	29.48	29.48

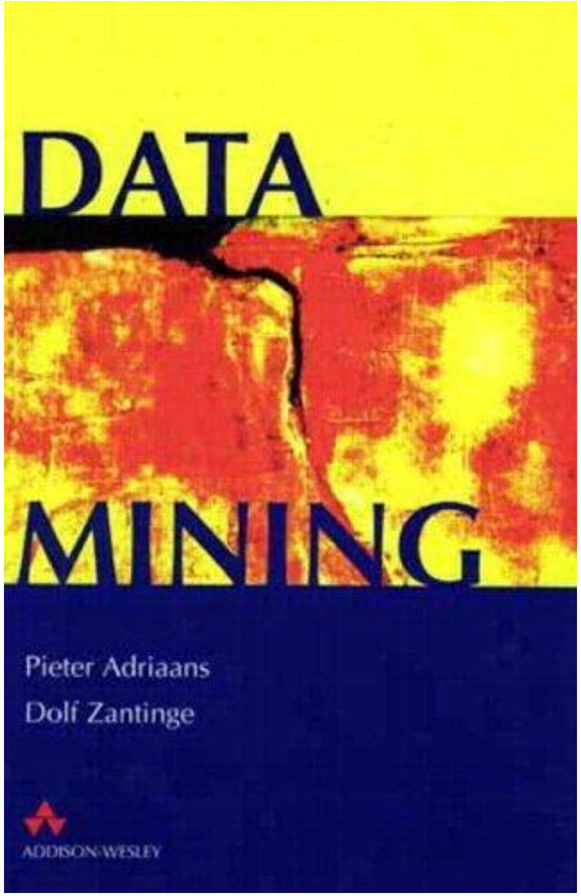
Vacation Distribution



Instruction Distribution



1996



The ancestors of Data Mining

Database

Query languages

4 Generation languages

Executive Information System

OLAP

Data Mining

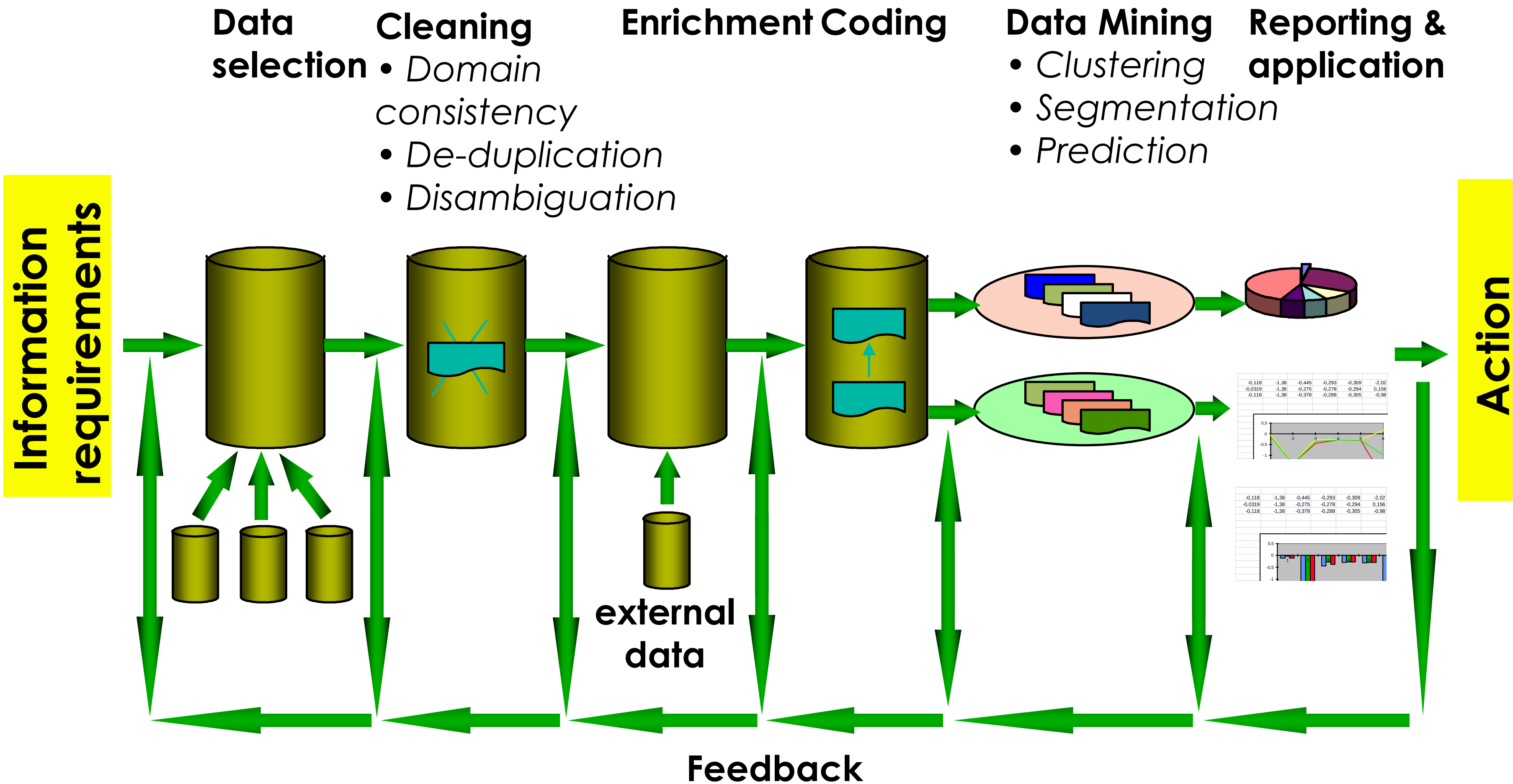
Next hype (?!)

KDD Definition

Knowledge Discovery in Databases is the non trivial extraction of implicit, previously unknown and potentially useful knowledge from data

(after Frawley Et al. 1991)

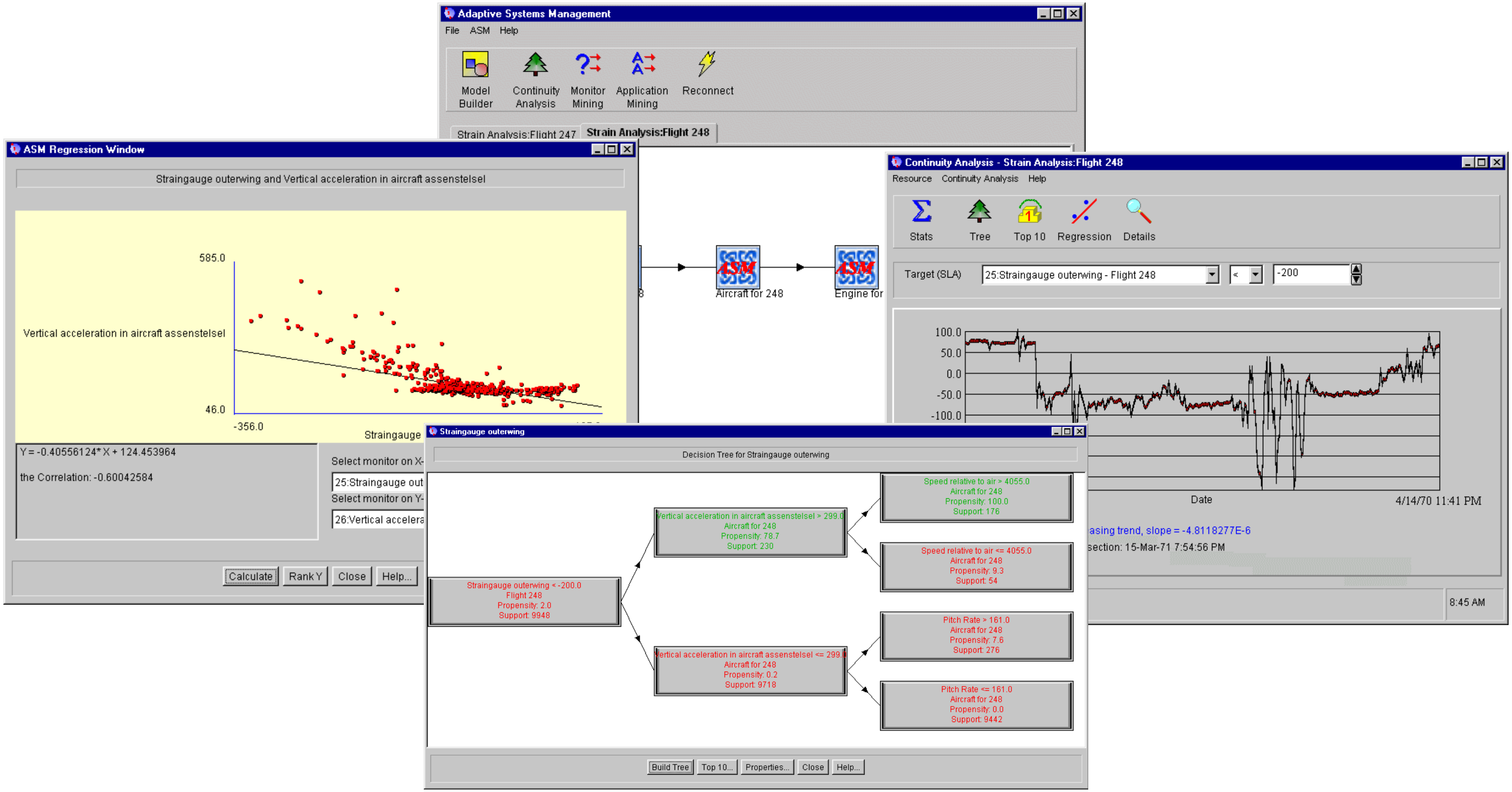
The KDD process



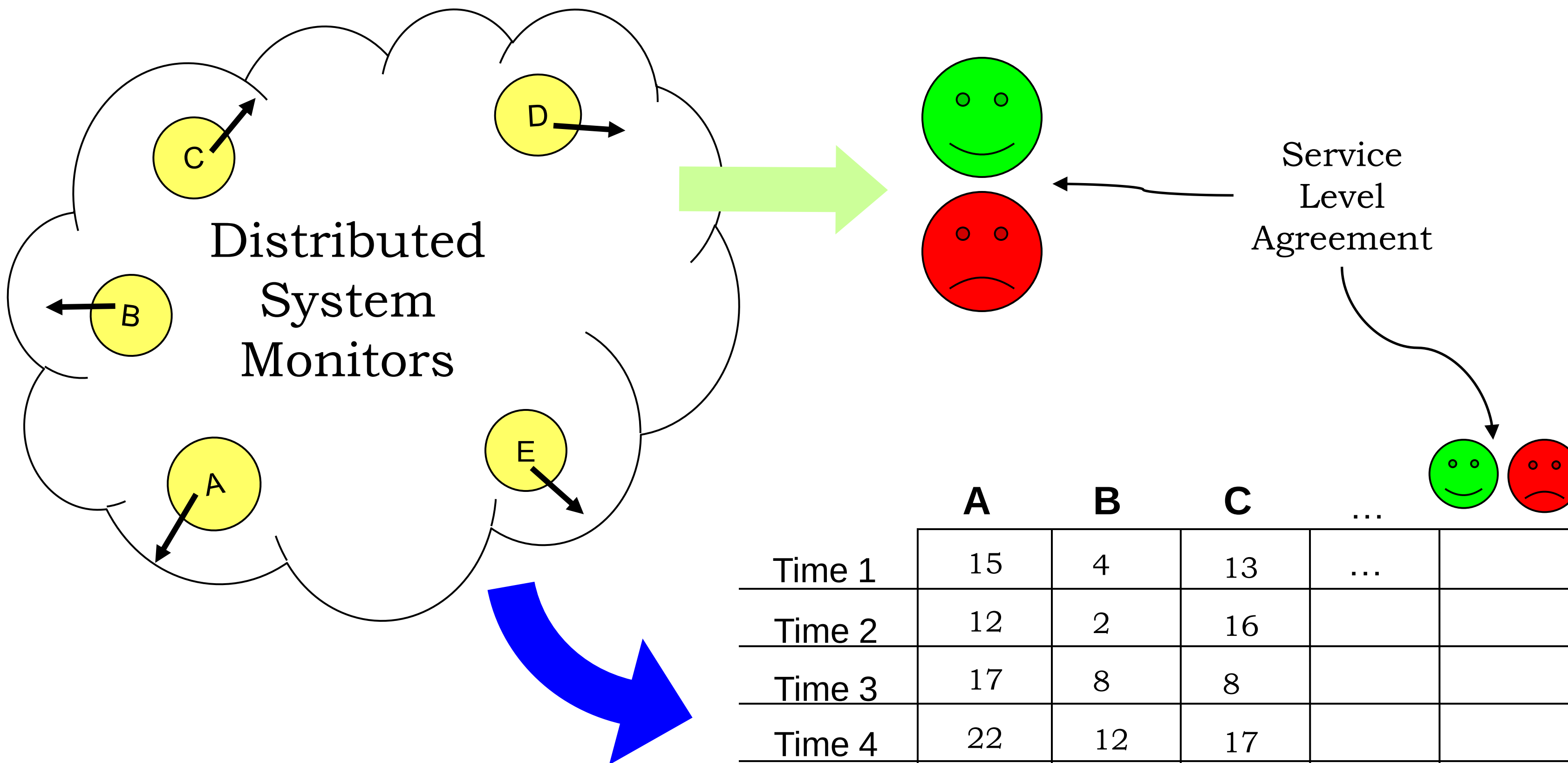
1997: Adaptive System Management

- the use of machine learning and data mining techniques to create self-learning models of the constantly changing IT-environment that allow us to predict the behavior of systems in the future and take timely automated remedial action to prevent system failure or decreased availability of the system.

Analysis of IT infrastructures using knowledge discovery



Monitoring in the system



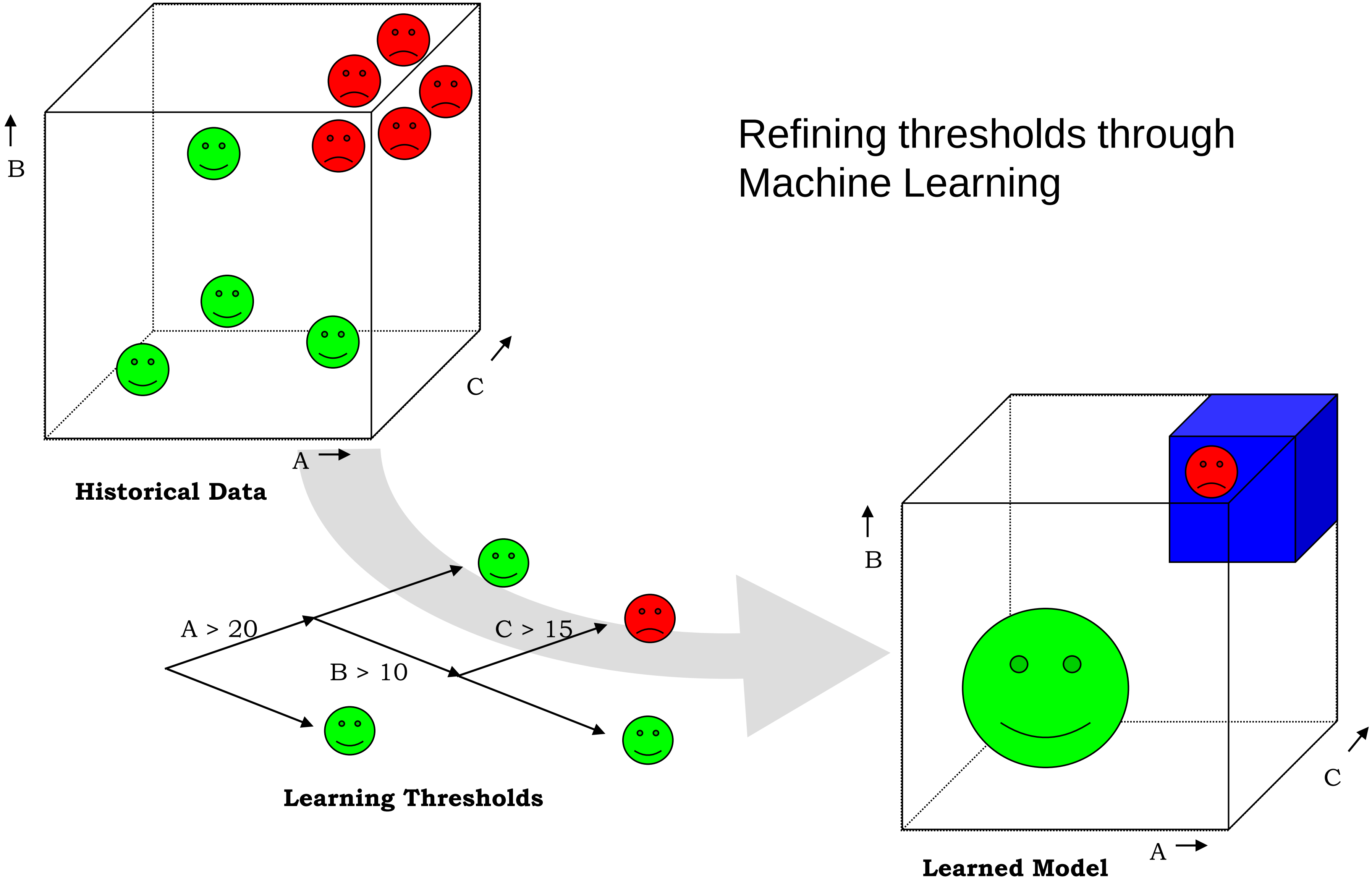
	A	B	C	…	
Time 1	15	4	13	…	
Time 2	12	2	16		
Time 3	17	8	8		
Time 4	22	12	17		
…	…	…	…		
Time k					

<u>Monitor Legend</u>

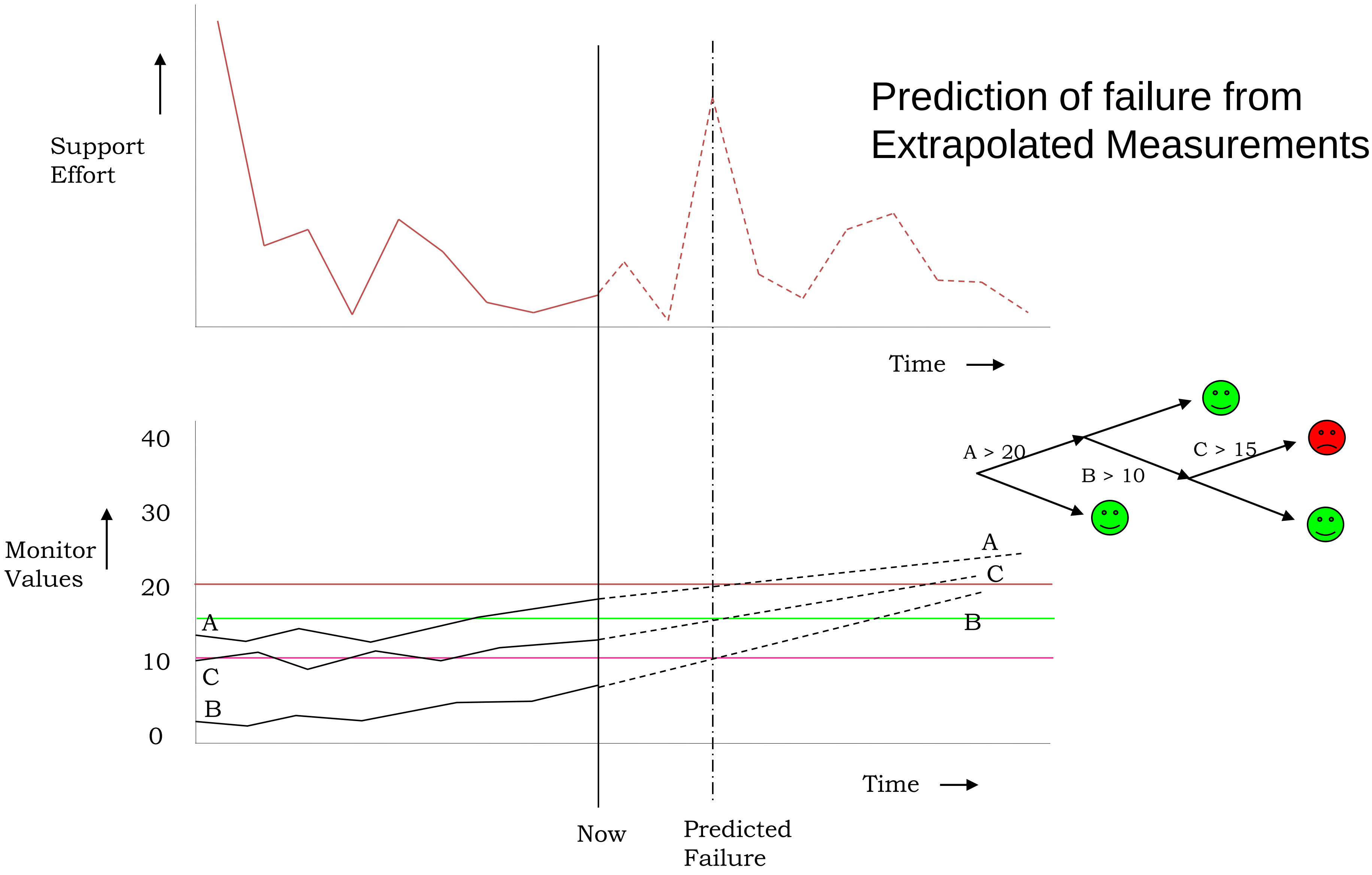
A: Nr. of Processes
B: Nr. of Failed Jobs
C: Login time Application
D: System Load
E: CPU Idle time
etc.

Refining thresholds

Refining thresholds through Machine Learning



Prediction of failure

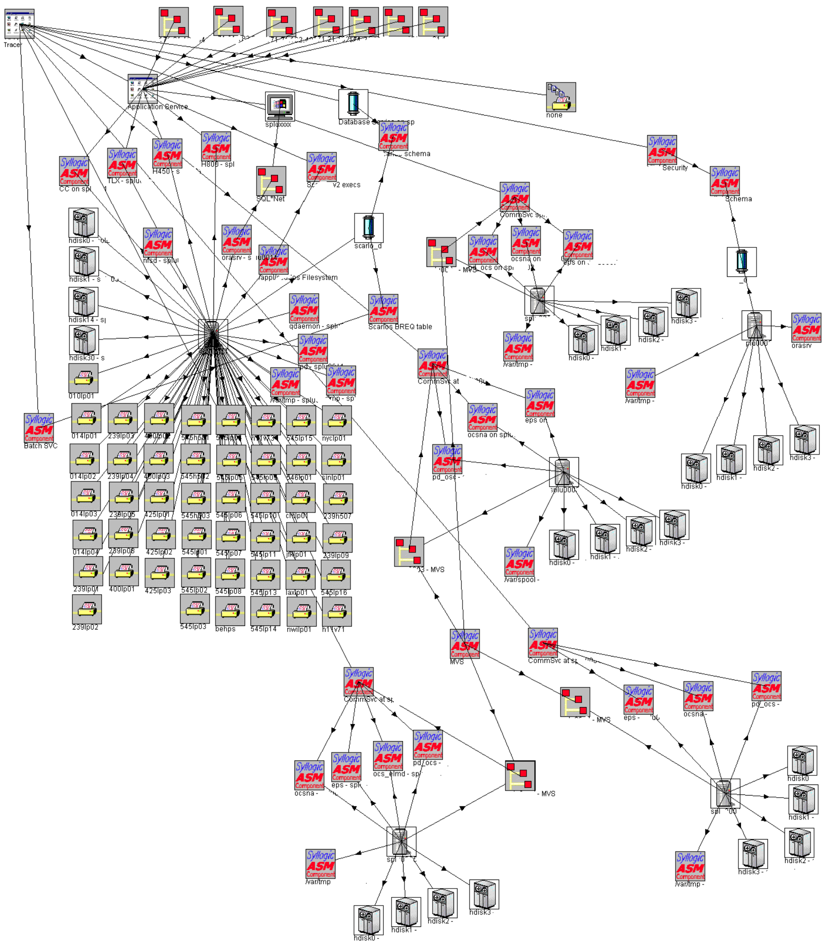


Patents

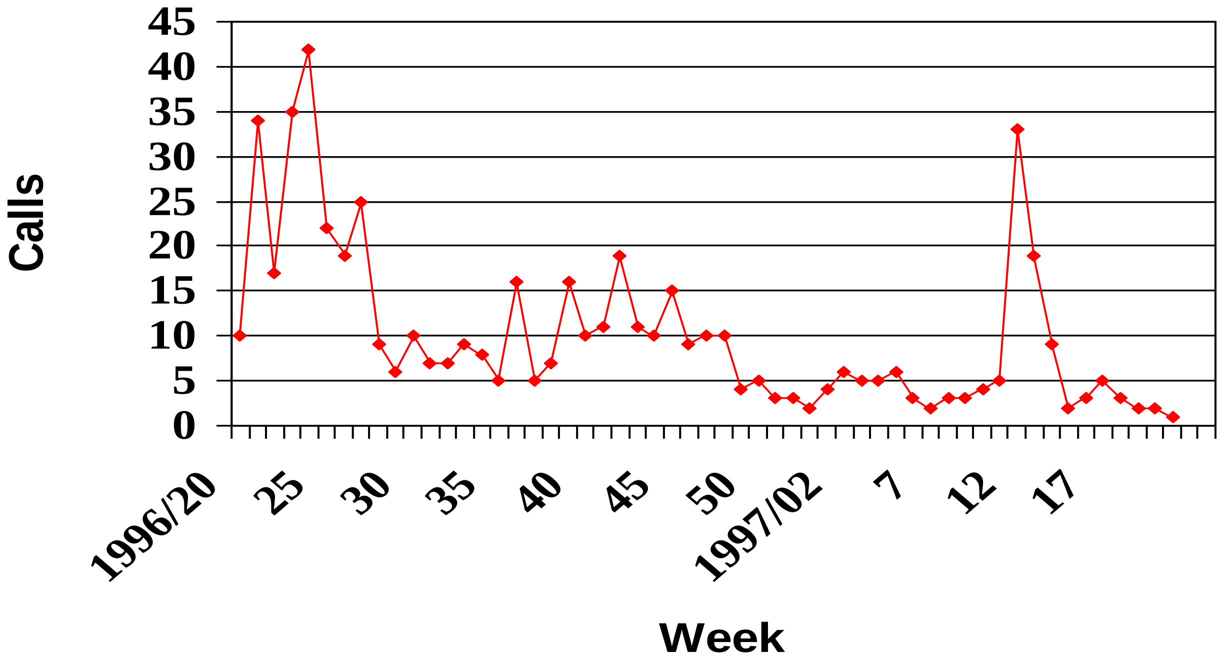
- Adriaans et al. United States Patent US 6,311,175 B1
  System and method for generating performance models of complex information technology systems.
- Adriaans et al. United States Patent US 6,313,390 B1
  A method for automatically controlling electronic musical devices by means of real-time construction and search of a multi-level data structure.
- Adriaans et al Unites States Patent US 6,393,387 B1
  System and method for model mining complex information management systems.

Experiments

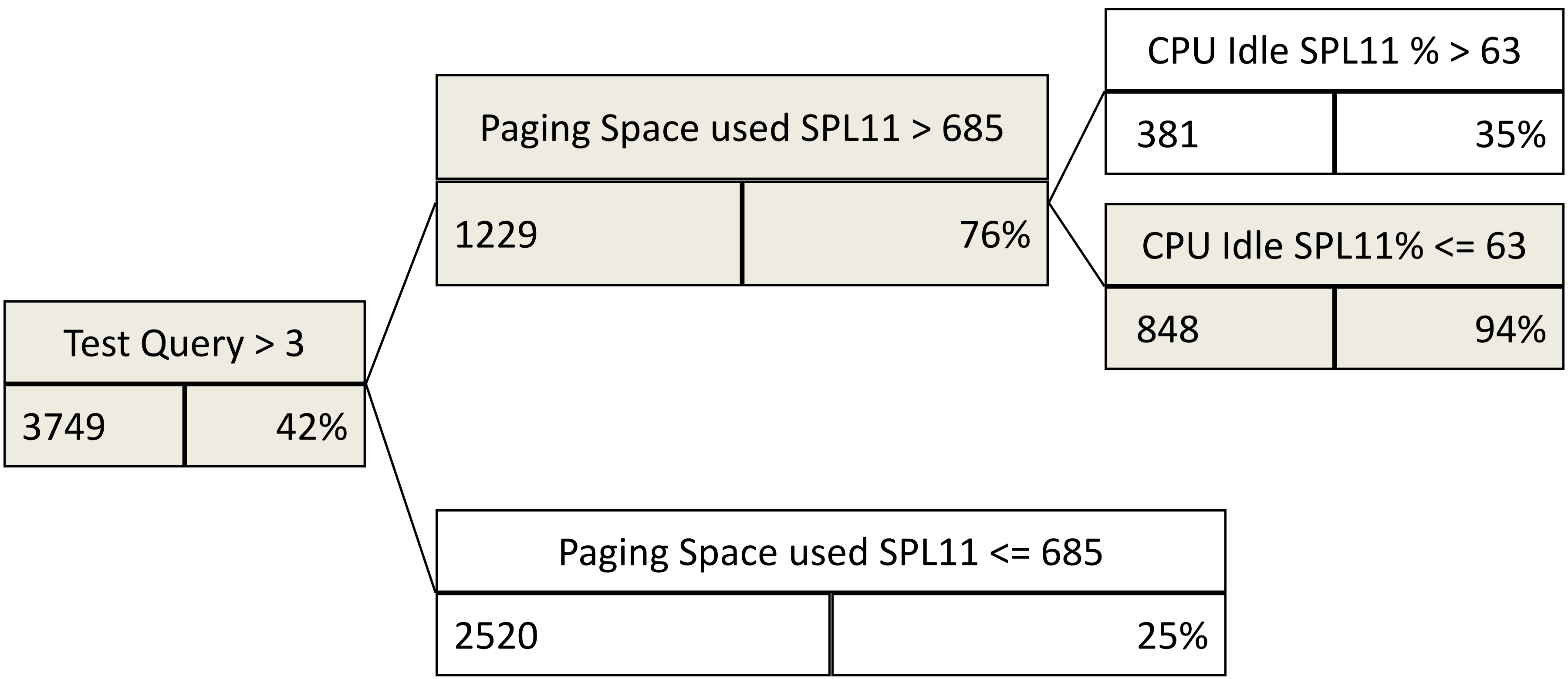
- 140 components
- 250 monitors
- 2 months
- 3500 snapshots



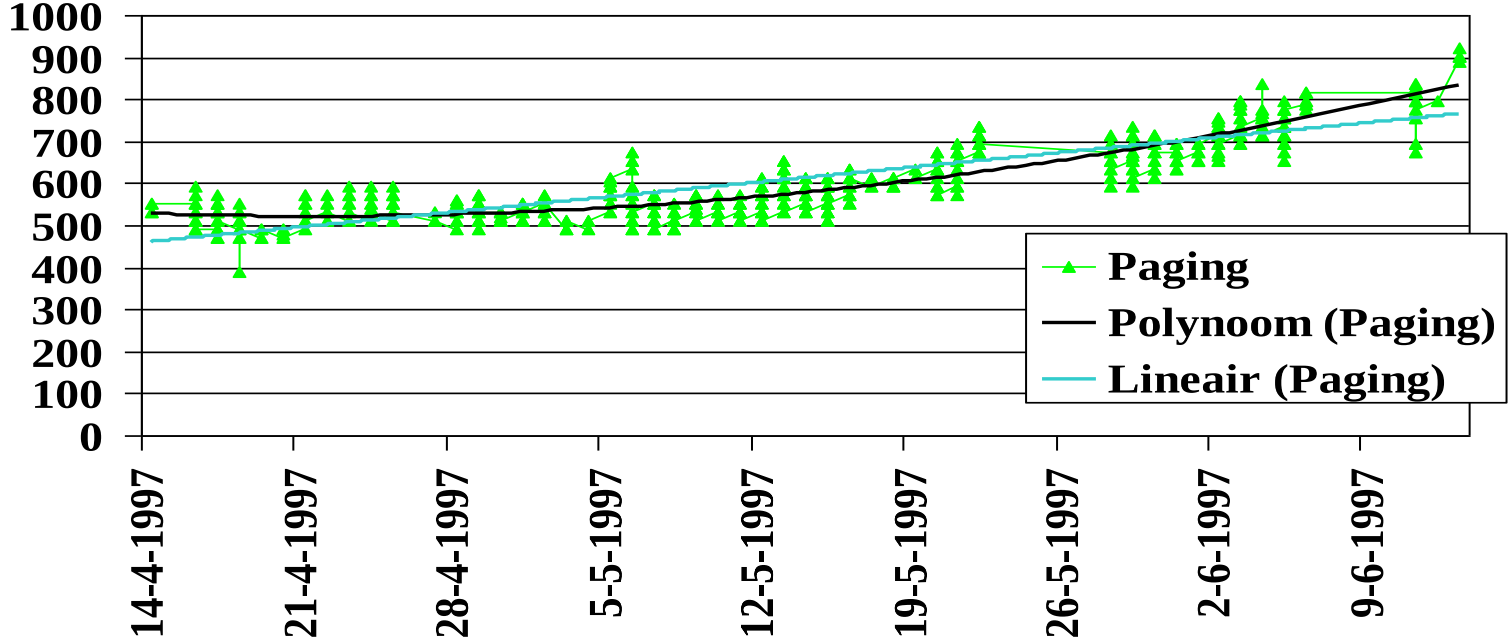
Helpdesk calls application



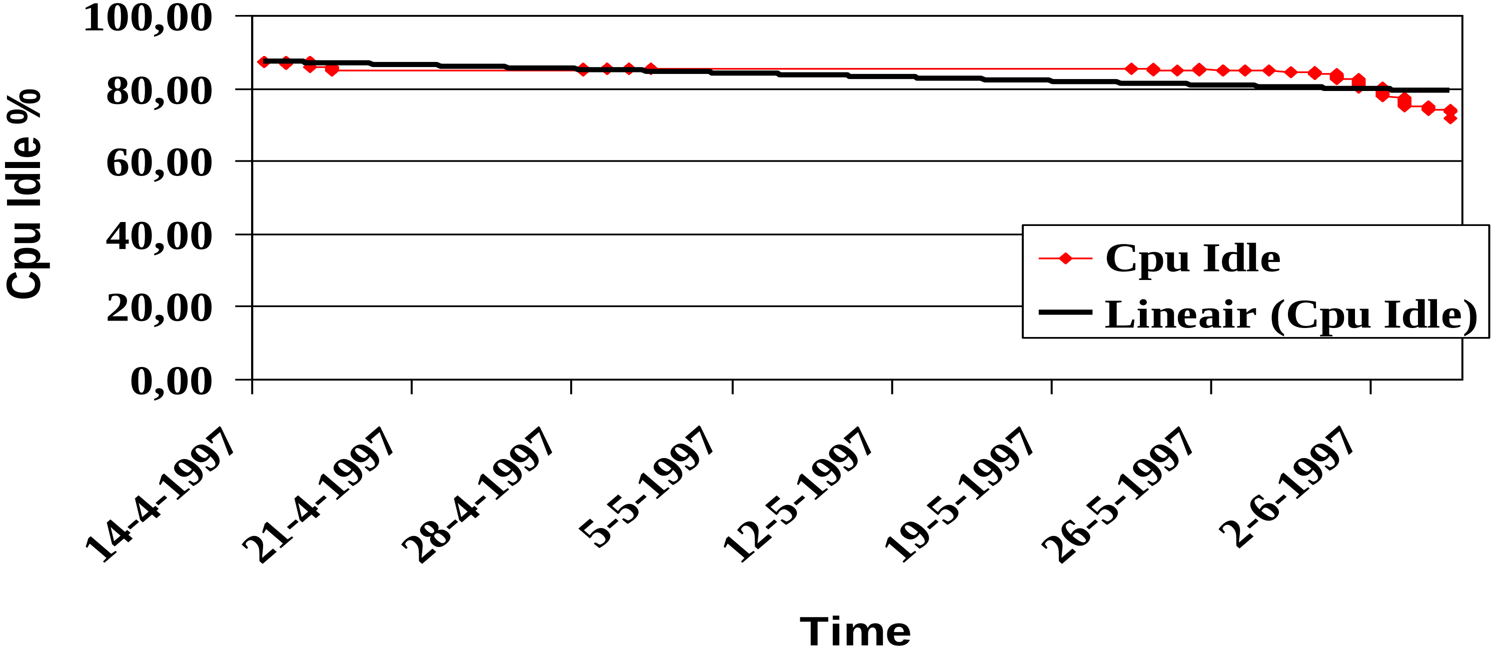
Decision tree Scarlos performance



Paging space develops polynomially



CPU Idle Trend



1998-2005

ROBOsail

Pole Balancing

State:

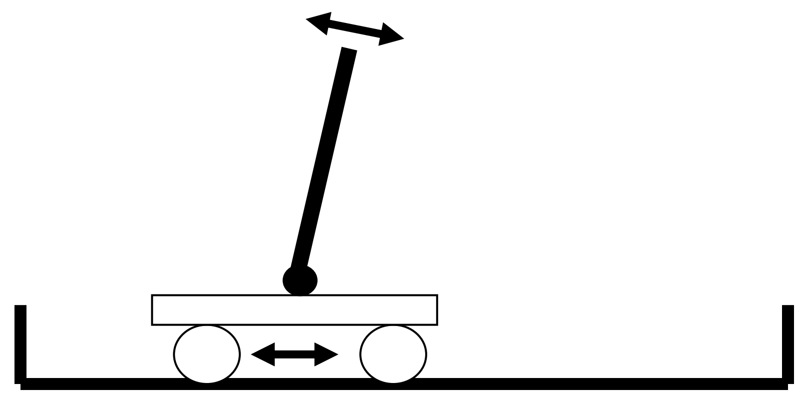
- Velocity of the cart
- Position of the cart
- Position of the stick
- Velocity of the stick

Action:

- Force on cart

Reward:

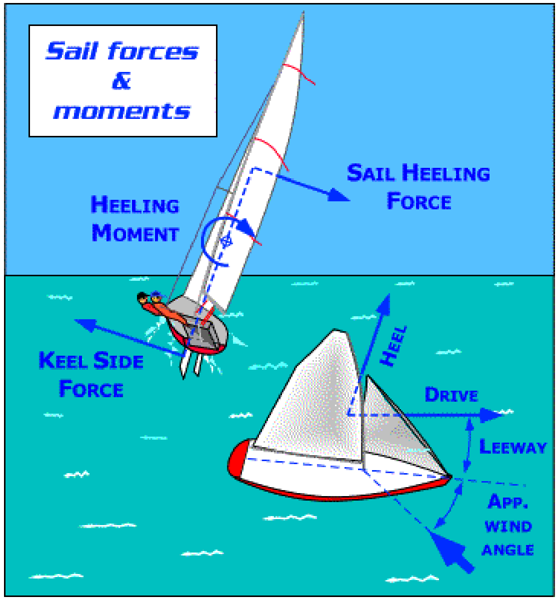
Stability of the pole



Ship Balancing

State:

- Velocity of the rudder
- Position of the rudder
- Course
- Log speed
- Apparent Wind speed
- Apparent wind angle
- Heel
- etc...

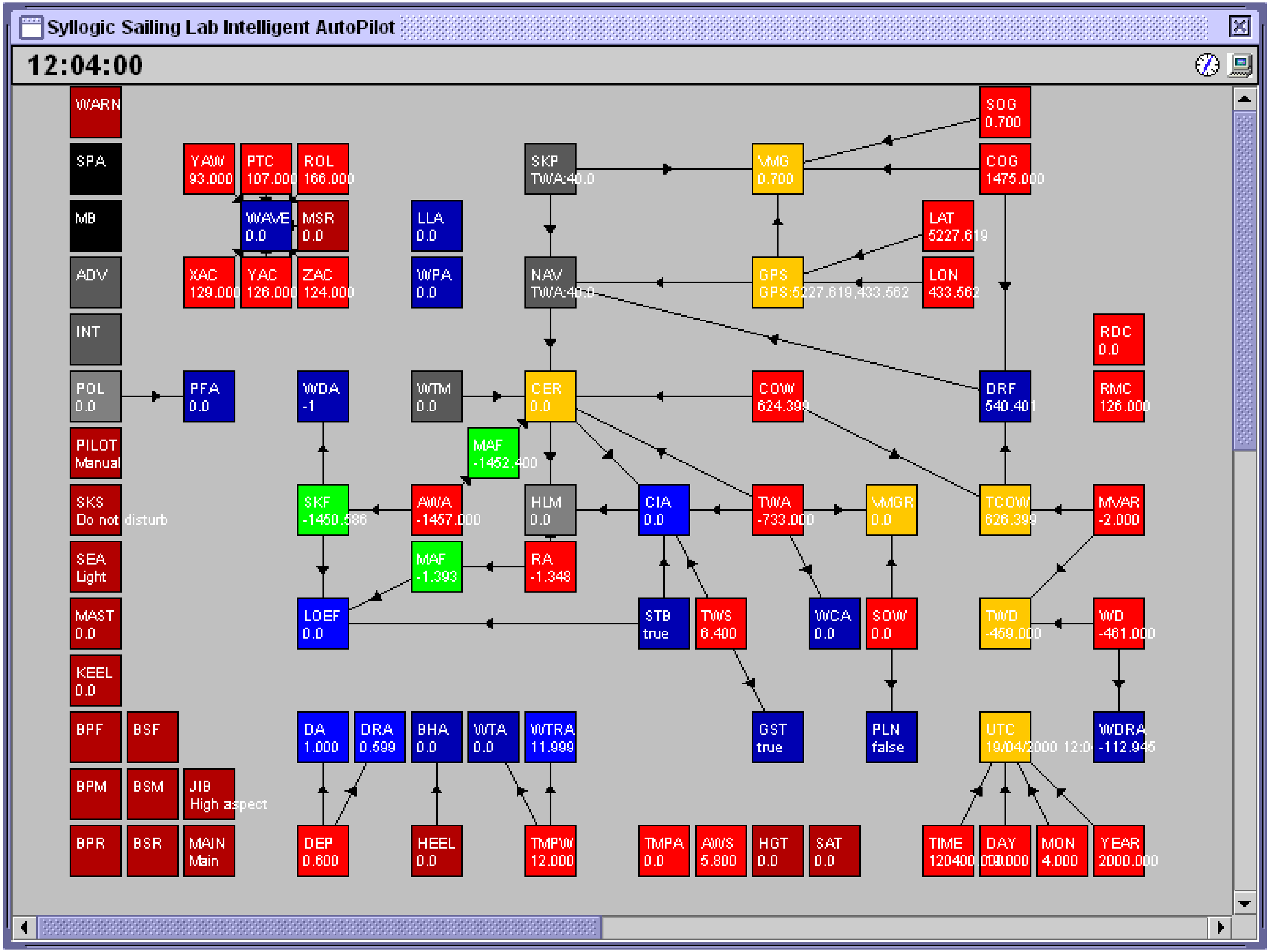


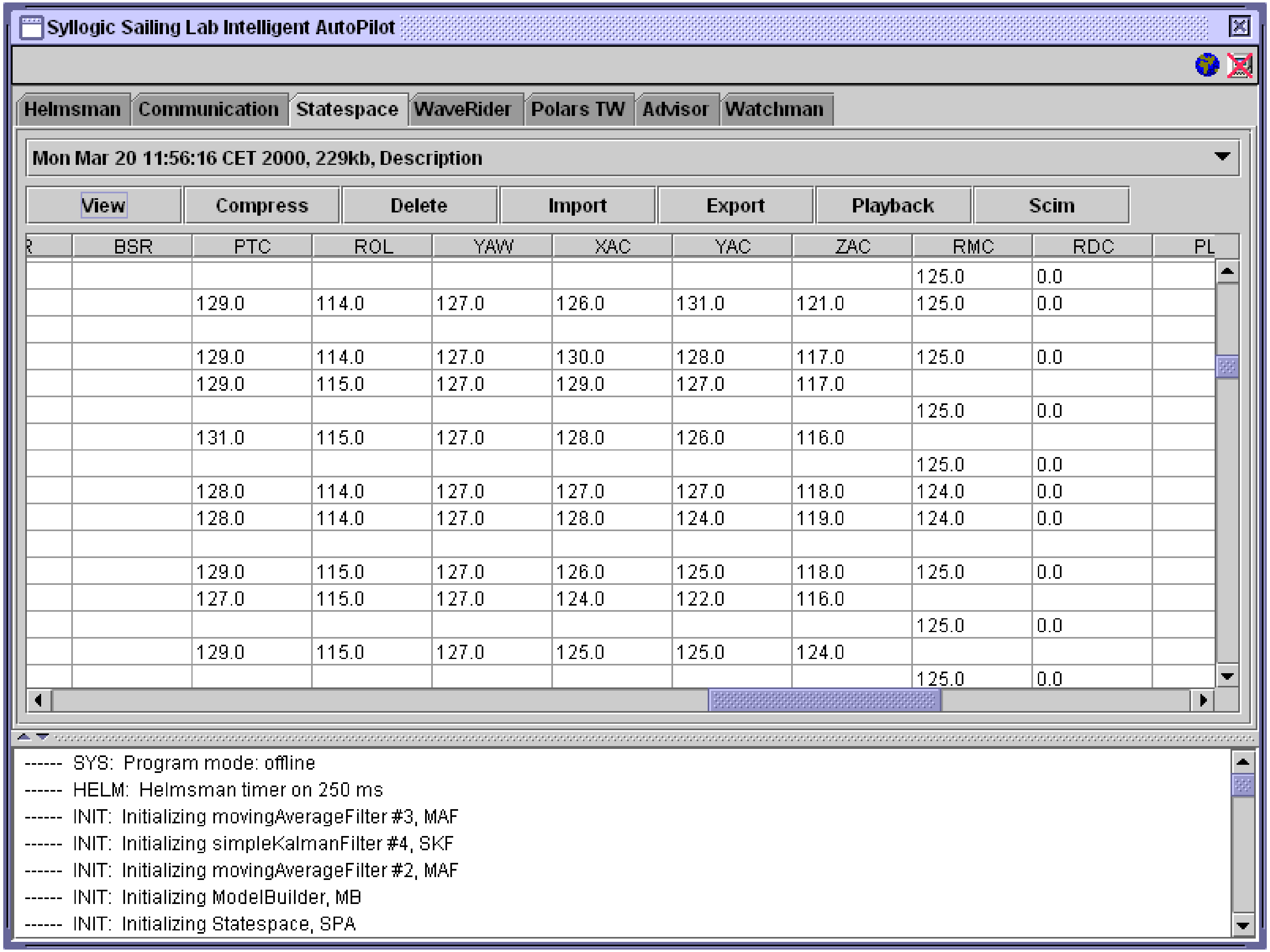
Action:

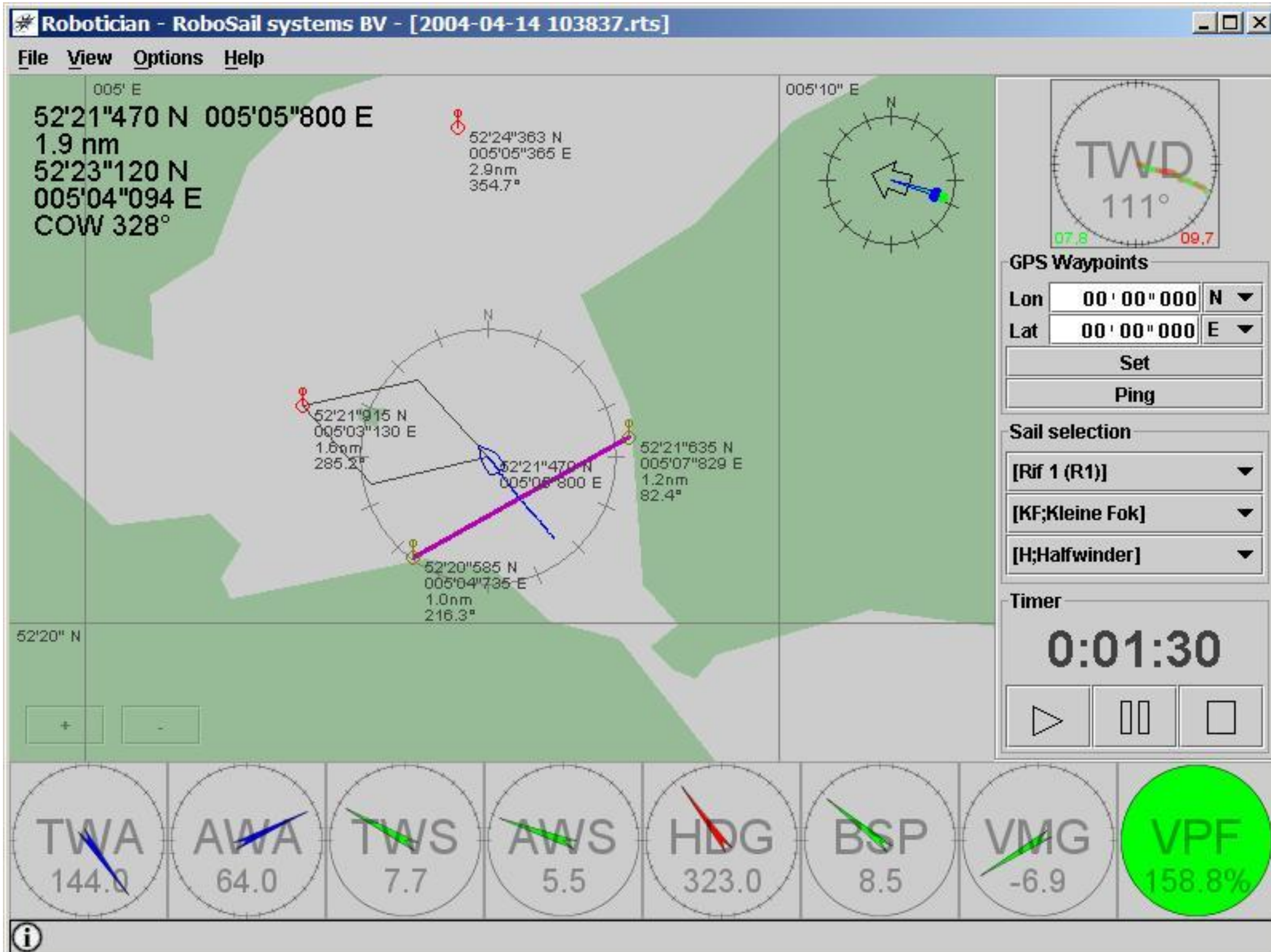
- Force Excercised on the rudder

Reward:

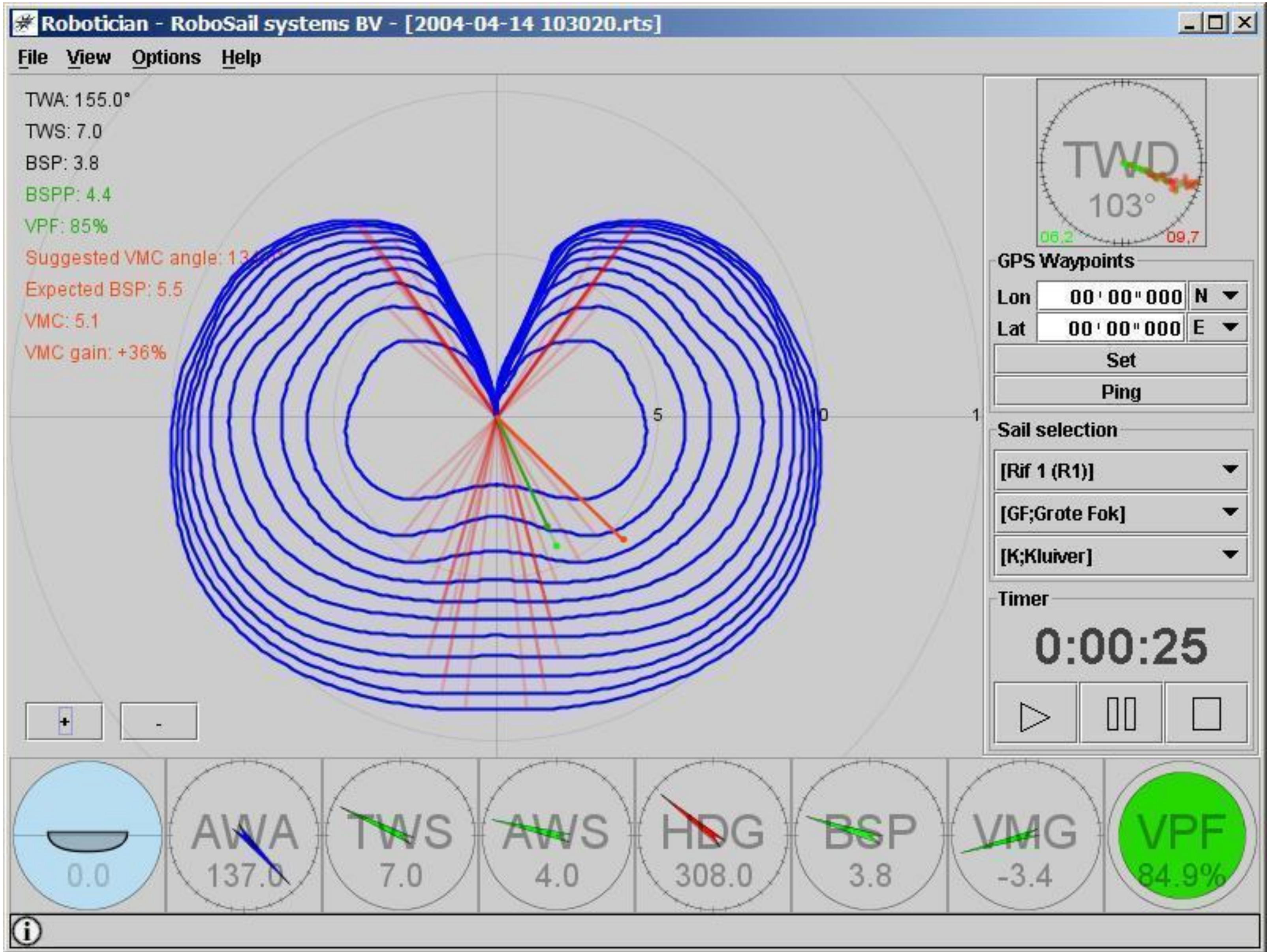
- Velocity of the yacht

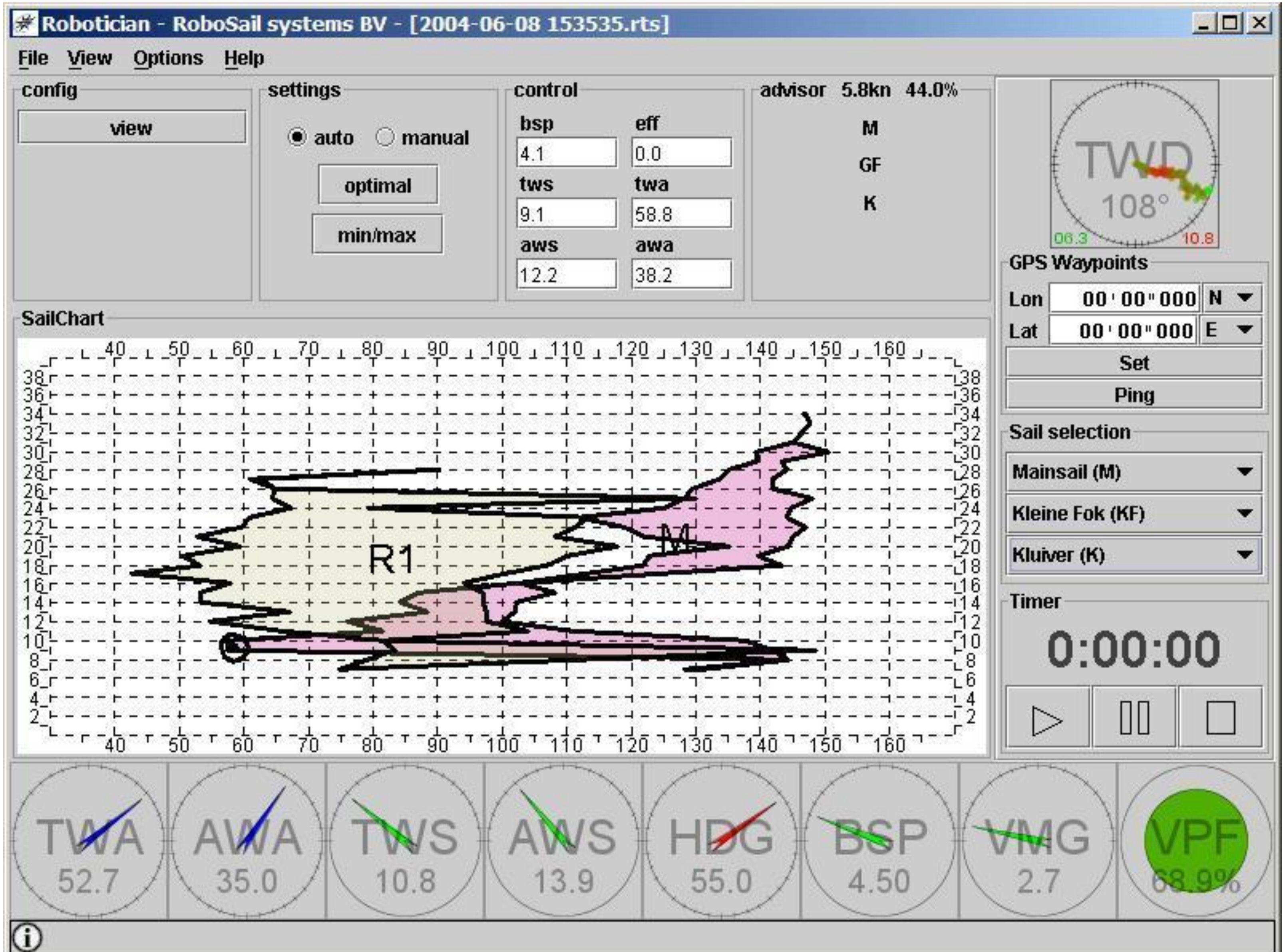

Syllogic Sailing Lab Intelligent AutoPilot
12:04:00
WARN
SPA
MB
ADV
INT
POL 0.0
PILOT Manual
SKS Do not disturb
SEA Light
MAST 0.0
KEEL 0.0
BPF
BSF
BPM
BSM
JIB High aspect
BPR
BSR
MAIN Main
YAW 93.000
PTC 107.000
ROL 166.000
WAVE 0.0
MSR 0.0
XAC 129.000
YAC 126.000
ZAC 124.000
PFA 0.0
WDA -1
SKF -1450.586
LOEF 0.0
LLA 0.0
WPA 0.0
WTM 0.0
MAF -1452.400
AWA -1457.000
MAF -1.393
SKP TWA:40.0
NAV TWA:40.0
CER 0.0
HLM 0.0
RA -1.348
CIA 0.0
STB true
TWS 6.400
GST true
VMG 0.700
GPS GPS:5227.619,433.562
COW 624.399
TWA -733.000
WCA 0.0
VMGR 0.0
SOW 0.0
PLN false
SOG 0.700
COG 1475.000
LAT 5227.619
LON 433.562
DRF 540.401
TCOW 626.399
TWD -459.000
UTC 19/04/2000 12:0
RDC 0.0
RMC 126.000
MVAR -2.000
WD -461.000
WDRA -112.945
DA 1.000
DRA 0.599
BHA 0.0
WTA 0.0
WTRA 11.999
DEP 0.600
HEEL 0.0
TMPW 12.000
TMPA 0.0
AWS 5.800
HGT 0.0
SAT 0.0
TIME 120400.000
DAY 19.000
MON 4.000
YEAR 2000.000


Syllogic Sailing Lab Intelligent AutoPilot
Helmsman
Communication
Statespace
WaveRider
Polars TW
Advisor
Watchman
Mon Mar 20 11:56:16 CET 2000, 229kb, Description
View
Compress
Delete
Import
Export
Playback
Scim
R BSR PTC ROL YAW XAC YAC ZAC RMC RDC PL
125.0 0.0
129.0 114.0 127.0 126.0 131.0 121.0 125.0 0.0
129.0 114.0 127.0 130.0 128.0 117.0 125.0 0.0
129.0 115.0 127.0 129.0 127.0 117.0
125.0 0.0
131.0 115.0 127.0 128.0 126.0 116.0
125.0 0.0
128.0 114.0 127.0 127.0 127.0 118.0 124.0 0.0
128.0 114.0 127.0 128.0 124.0 119.0 124.0 0.0
129.0 115.0 127.0 126.0 125.0 118.0 125.0 0.0
127.0 115.0 127.0 124.0 122.0 116.0
125.0 0.0
129.0 115.0 127.0 125.0 125.0 124.0
125.0 0.0
------ SYS: Program mode: offline
------ HELM: Helmsman timer on 250 ms
------ INIT: Initializing movingAverageFilter #3, MAF
------ INIT: Initializing simpleKalmanFilter #4, SKF
------ INIT: Initializing movingAverageFilter #2, MAF
------ INIT: Initializing ModelBuilder, MB
------ INIT: Initializing Statespace, SPA

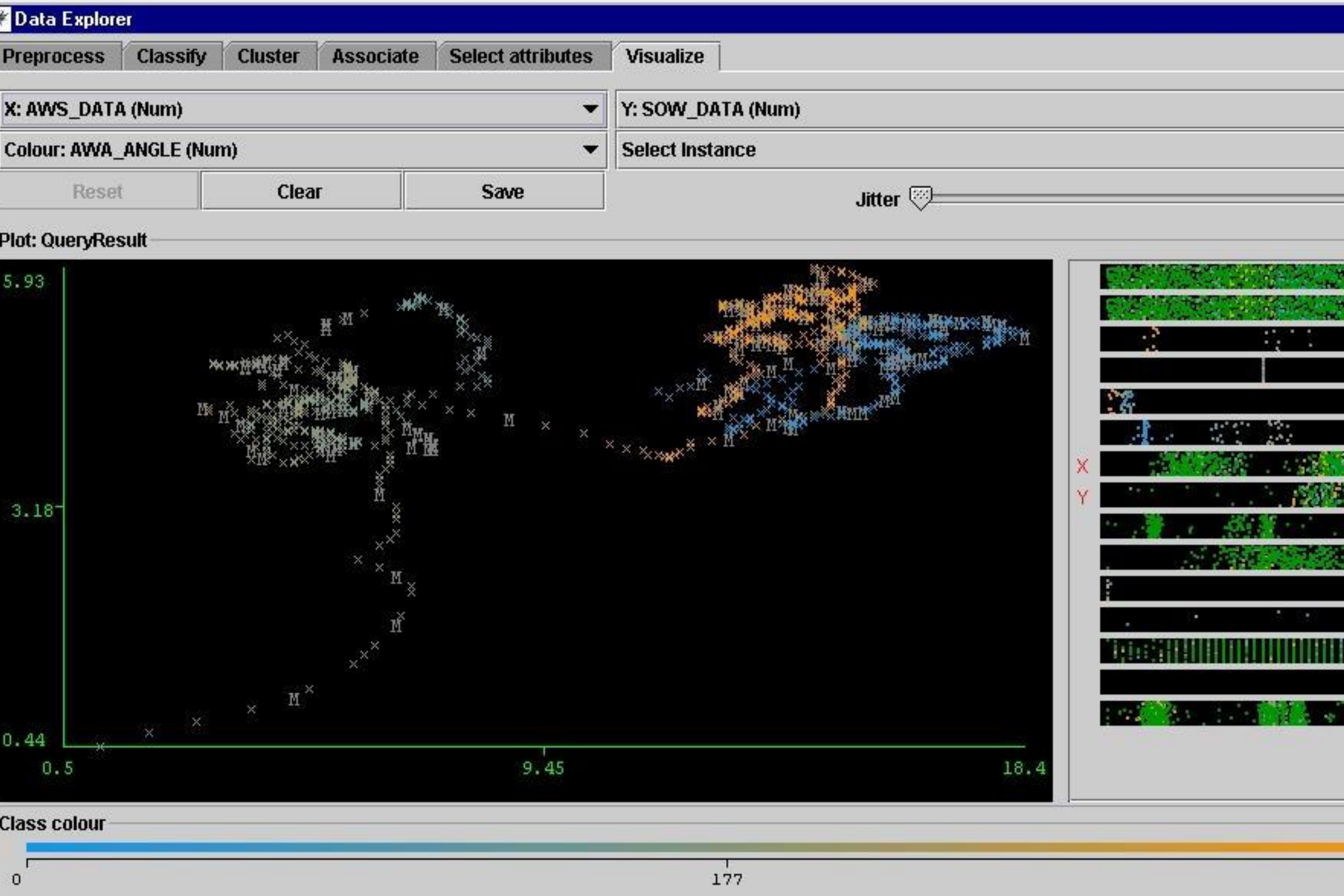


Robotician - RoboSail systems BV - [2004-04-14 103837.rts]
File View Options Help
005' E
005'10" E
52'21"470 N 005'05"800 E
1.9 nm
52'23"120 N
005'04"094 E
COW 328°
52'24"363 N
005'05"365 E
2.9nm
354.7°
52'21"915 N
005'03"130 E
1.6nm
285.2°
52'21"470 N
005'05"800 E
52'21"635 N
005'07"829 E
1.2nm
82.4°
52'20"585 N
005'04"735 E
1.0nm
216.3°
52'20" N
TWD
111°
07.8
09.7
GPS Waypoints
Lon 00 ' 00 " 000 N
Lat 00 ' 00 " 000 E
Set
Ping
Sail selection
[Rif 1 (R1)]
[KF;Kleine Fok]
[H;Halfwinder]
Timer
0:01:30
TWA 144.0
AWA 64.0
TWS 7.7
AWS 5.5
HDG 323.0
BSP 8.5
VMG -6.9
VPF 158.8%


Robotician - RoboSail systems BV - [2004-04-14 103020.rts]
File View Options Help
TWA: 155.0°
TWS: 7.0
BSP: 3.8
BSPP: 4.4
VPF: 85%
Expected BSP: 5.5
VMC: 5.1
VMC gain: +36%
5
TWD
103°
GPS Waypoints
Lon 00 ' 00 " 000 N
Lat 00 ' 00 " 000 E
Set
Ping
Sail selection
[Rif 1 (R1)]
[GF;Grote Fok]
[K;Kluiver]
Timer
0:00:25
+
-
0.0
AWA
TWS
7.0
AWS
4.0
HDG
308.0
BSP
3.8
VMG
-3.4
VPF
84.9%

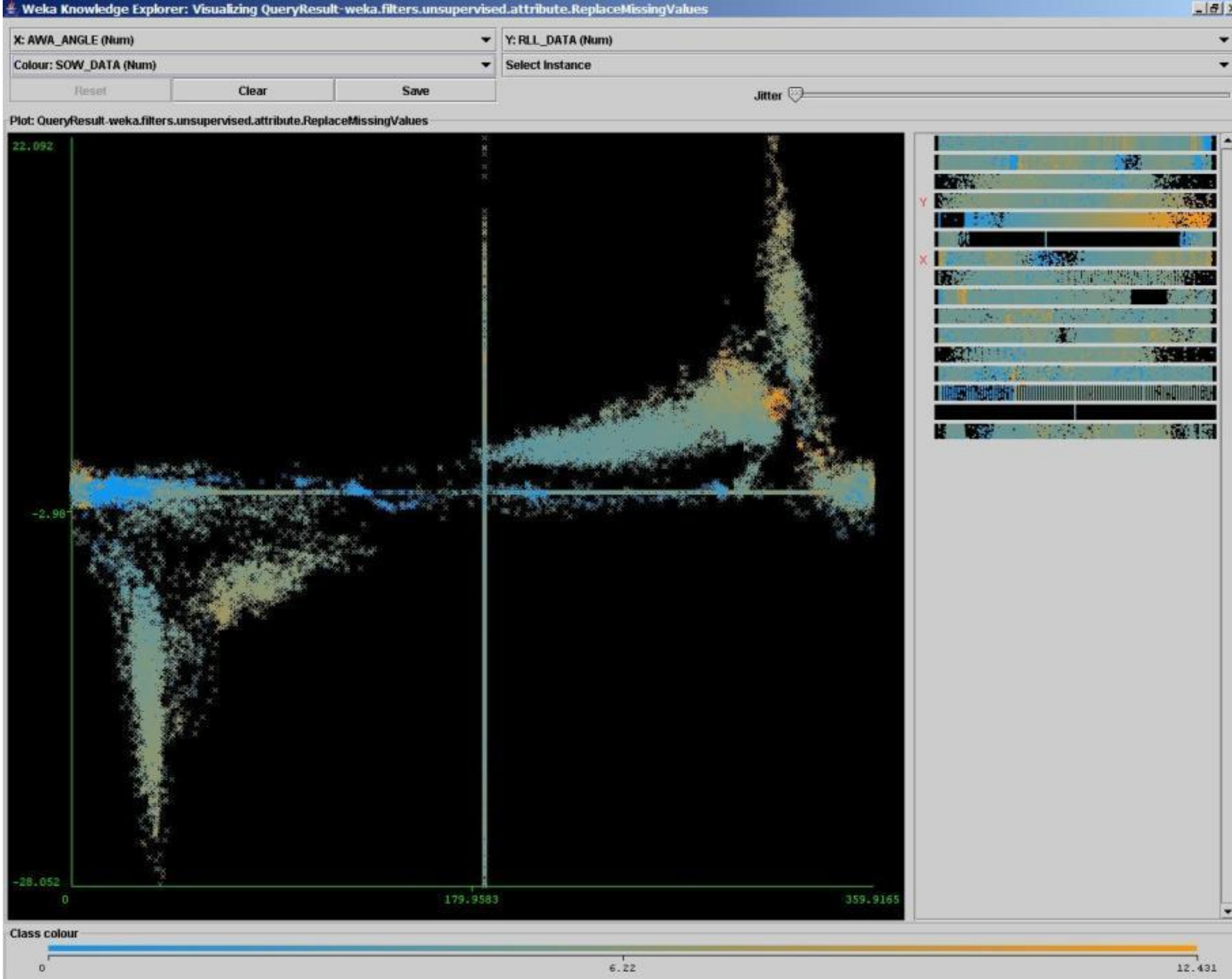

Robotician - RoboSail systems BV - [2004-06-08 153535.rts]
File View Options Help
config
view
settings
auto manual
optimal
min/max
control
bsp 4.1
eff 0.0
tws 9.1
twa 58.8
aws 12.2
awa 38.2
advisor 5.8kn 44.0%
M
GF
K
TWD
108°
06.3
10.8
GPS Waypoints
Lon 00' 00" 000 N
Lat 00' 00" 000 E
Set
Ping
Sail selection
Mainsail (M)
Kleine Fok (KF)
Kluiver (K)
Timer
0:00:00
SailChart
R1
M
TWA 52.7
AWA 35.0
TWS 10.8
AWS 13.9
HDG 55.0
BSP 4.50
VMG 2.7
VPF 68.9%

Mining for Sensor Calibration



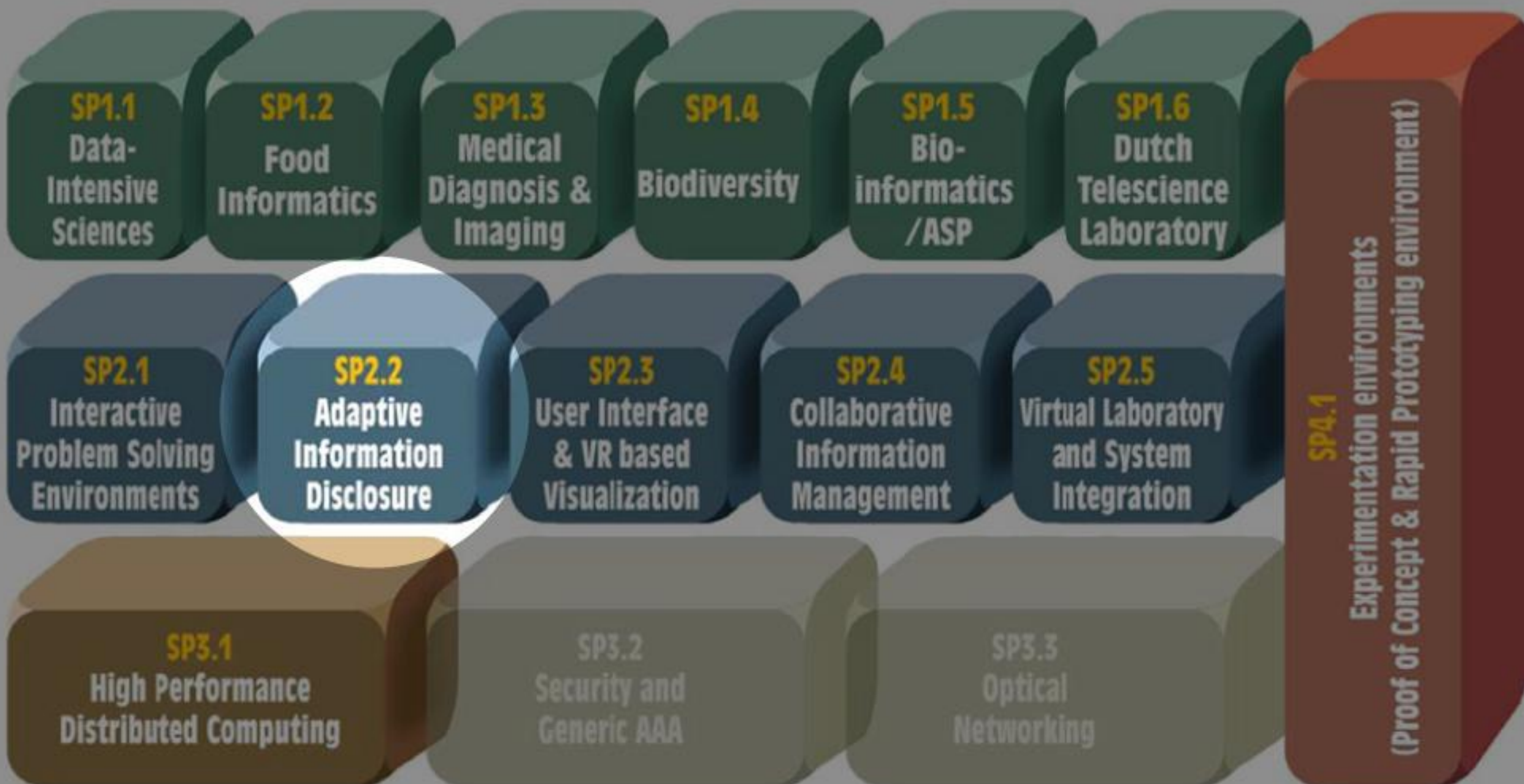
Mining for Sensor Calibration

- Possible causes:
  - Log sensor gives asymmetric data
  - Wind sensor gives asymmetric data
  - Boat itself is asymmetric
  - Boats with masts $>$ 25 m perform better over port-bow on Northern hemisphere

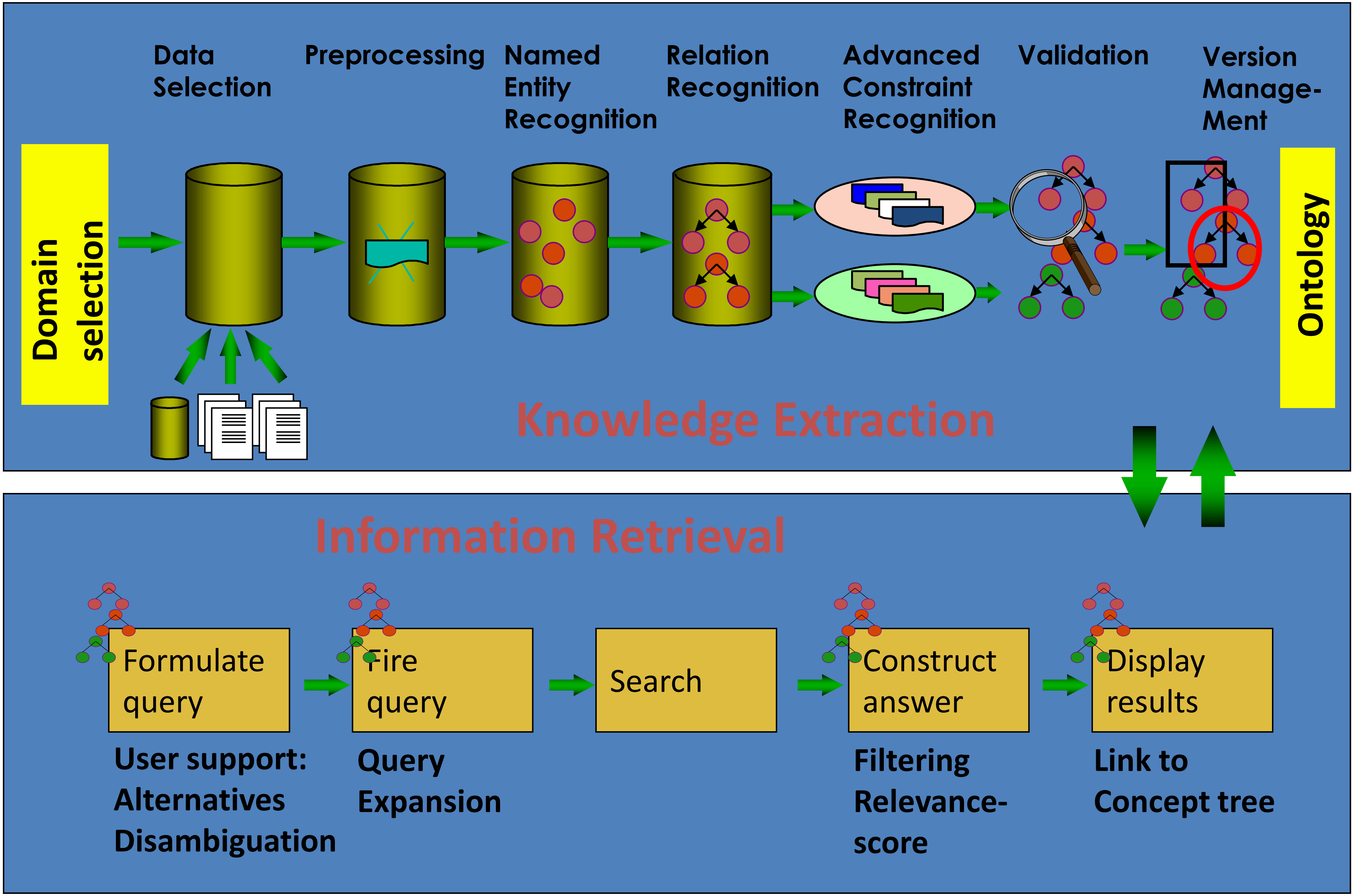

Weka Knowledge Explorer: Visualizing QueryResult-weka.filters.unsupervised.attribute.ReplaceMissingValues
X: AWA_ANGLE (Num)
Y: RLL_DATA (Num)
Colour: SOW_DATA (Num)
Select Instance
Reset
Clear
Save
Jitter
Plot: QueryResult-weka.filters.unsupervised.attribute.ReplaceMissingValues
22.092
-2.98
-28.052
0
179.9583
359.9165
Y
X
Class colour
0
6.22
12.431

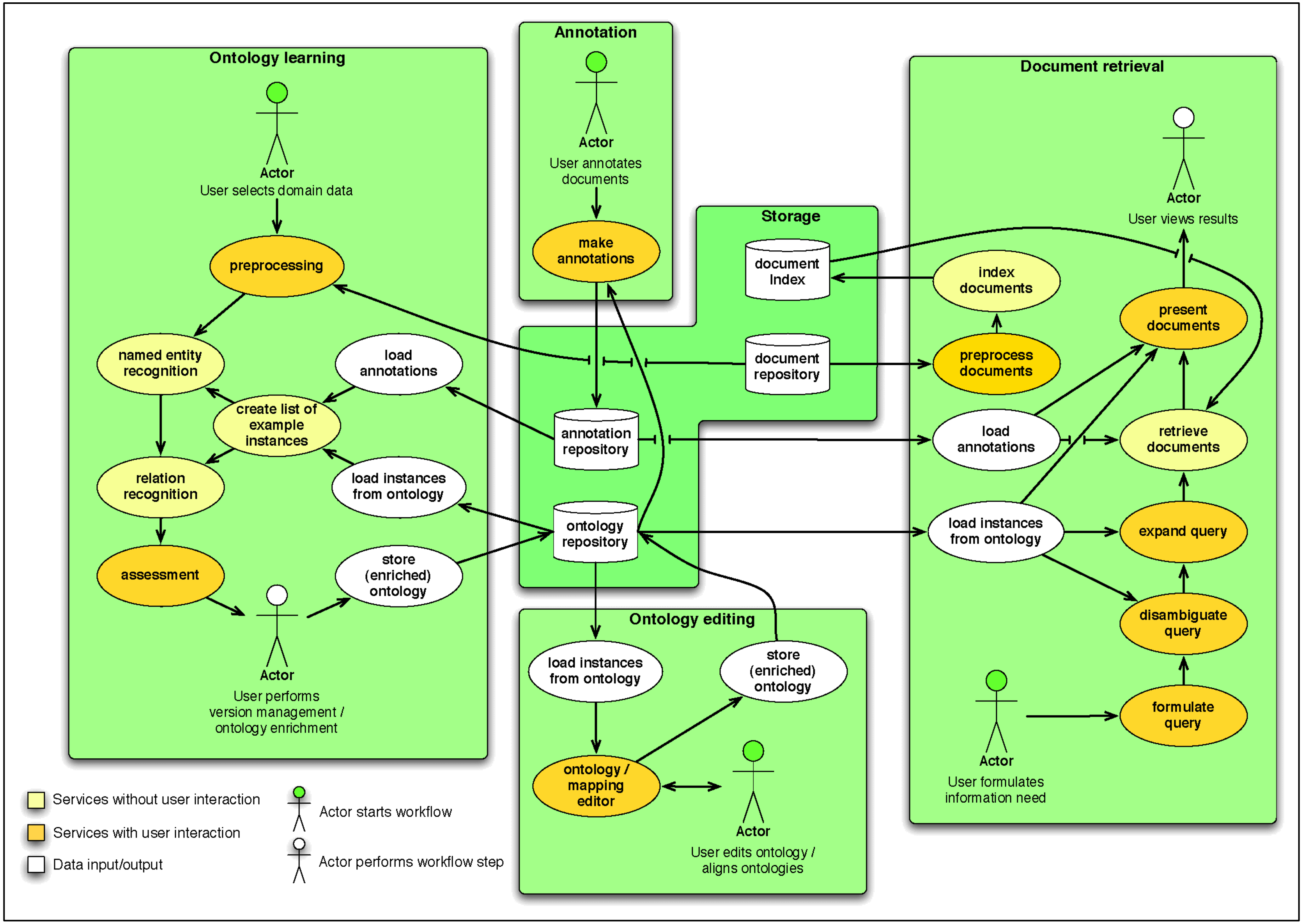
2003-2009 Adaptive Information Disclosure (AID)

participating in the VL-e (Virtual lav for e-Science) project



Adaptive Information Disclosure





The AIDA toolbox
for knowledge extraction and knowledge management
in a Virtual Laboratory for e-Science

2010-2015: D2S: From Data to Semantics
for data publishers

Challenges
How to **share** scientific data?
How to **access, analyse and interpret** the data?
How to **communicate and publish** results?

Domains
- Life-Sciences
- Humanities

End Users
- Elsevier
- Philips
- DANS

How do we speed up the transfer of scientific data, information, knowledge from a research paper into actionable form?

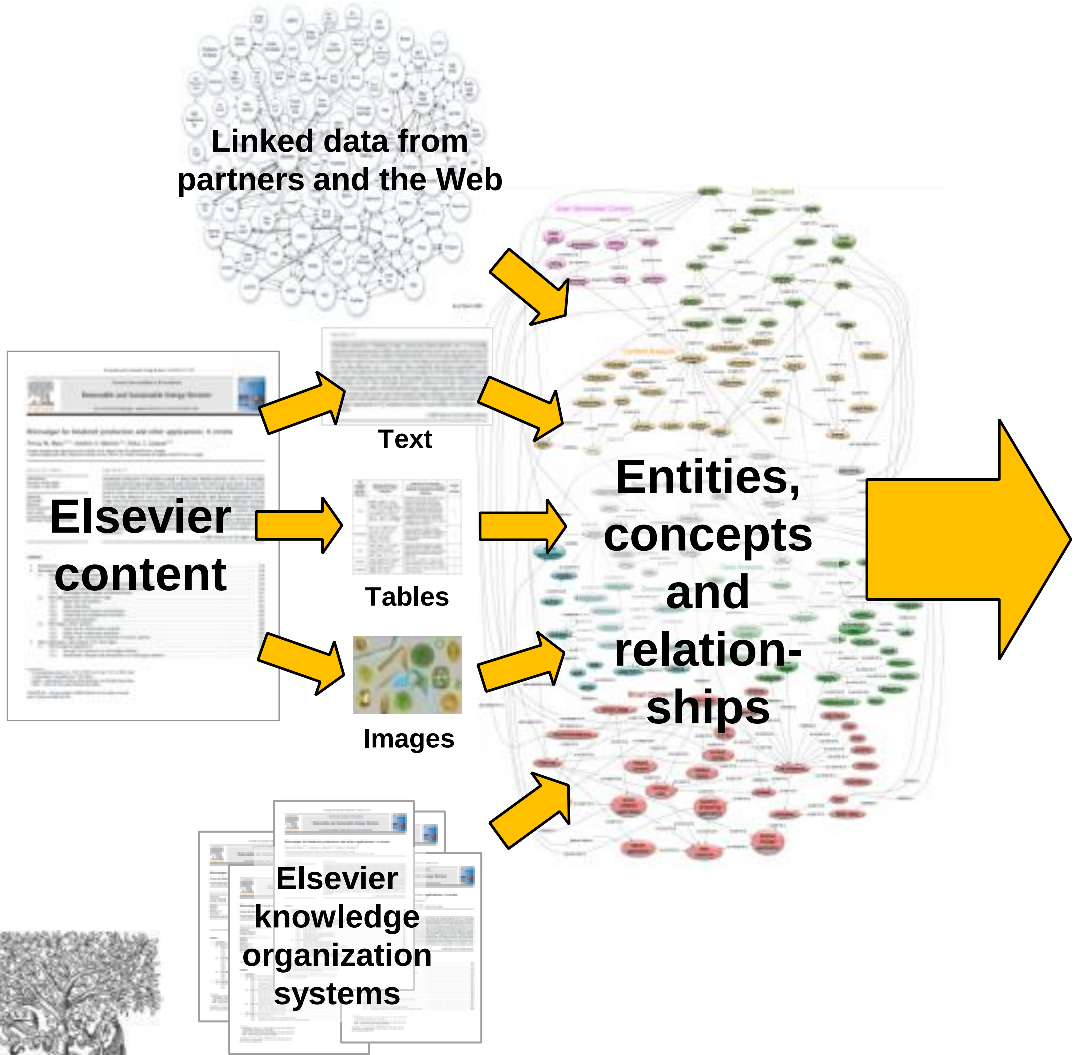
ELSEVIER

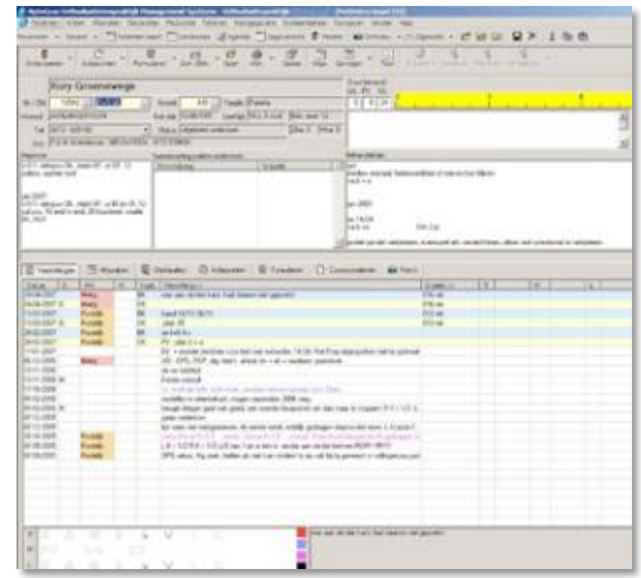
- Looking to become a provider of clinical rules
- Map clinical guidelines against recent papers
- Link to data sets of research paper

PHILIPS

- looking to quickly consume research findings into Clinical Decision Support systems
- current guideline update cycle is 5 years

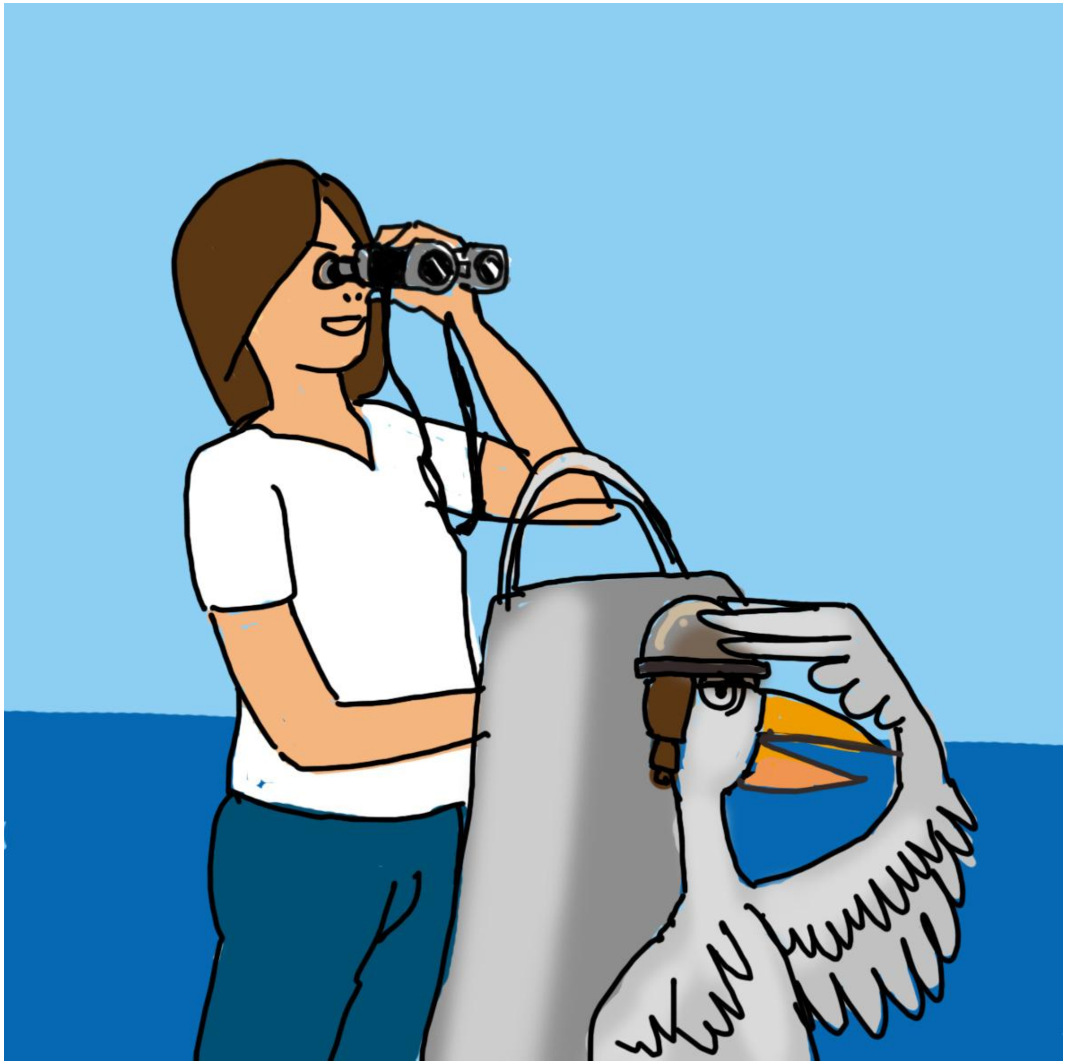
Smart content at Elsevier & Philips





Philips Hospital Information System

Looking Forward



EXPERT OPINION

Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that [illegible] rather than elementary par- [illegible] elegant math-

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-relate machine learning have been statistical speech re ognition and statistical machine translation. Th reason for these successes is not that these tasks a [illegible] than other tasks; they are in fact much har [illegible] document classification that netherlands eScience center

by SURF & NWO

Home News Agenda About us Projects Contact Case Studies

Optimizing Discovery in the Big Data Era

The **Netherlands eScience Center (NLeSC)** supports and reinforces multidisciplinary and data-intensive research through creative and innovative use of ICT in all its manifestations. To stimulate this enhanced Science (eScience) **NLeSC** works as a network organization focused on collaboration, with the aim to change scientific practice by making large-scale data analysis possible across multiple disciplines.

News

- ASCI course: A Programmer's Guide for Modern High-Performance Computing
- Dutch Research Consortium at SC12
- eScience climatology project to reap benefits of access to US supercomputers
- Carl-Christian Buhr of Neelie Kroes' cabinet visits NLeSC
- EU blijft streven naar universeel ID

Read more

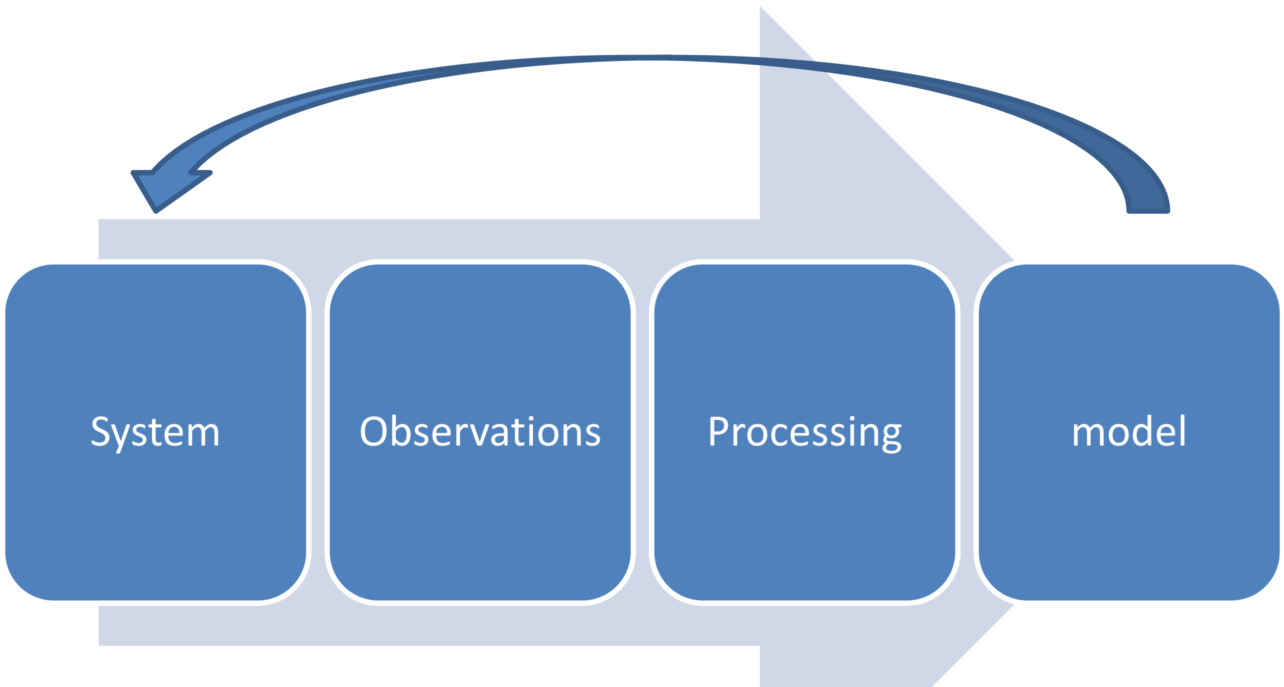
Agenda

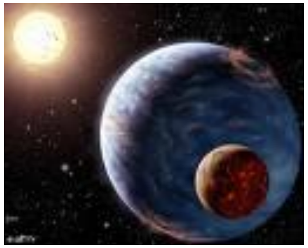
- Lustrum opleiding Bio-informatica 30 november 2012
- Data sharing workshop 14 December 2012
- e-BioGrid momentum 17 December 2012
- 27th SARA Superdag 19 December 2012
- Free webinar Text Mining: a new way to discover food knowledge 17 january 2013

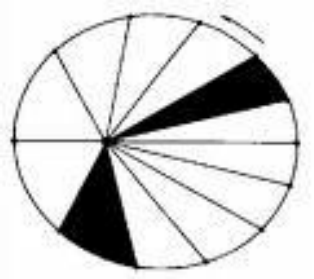
Read more

Case St

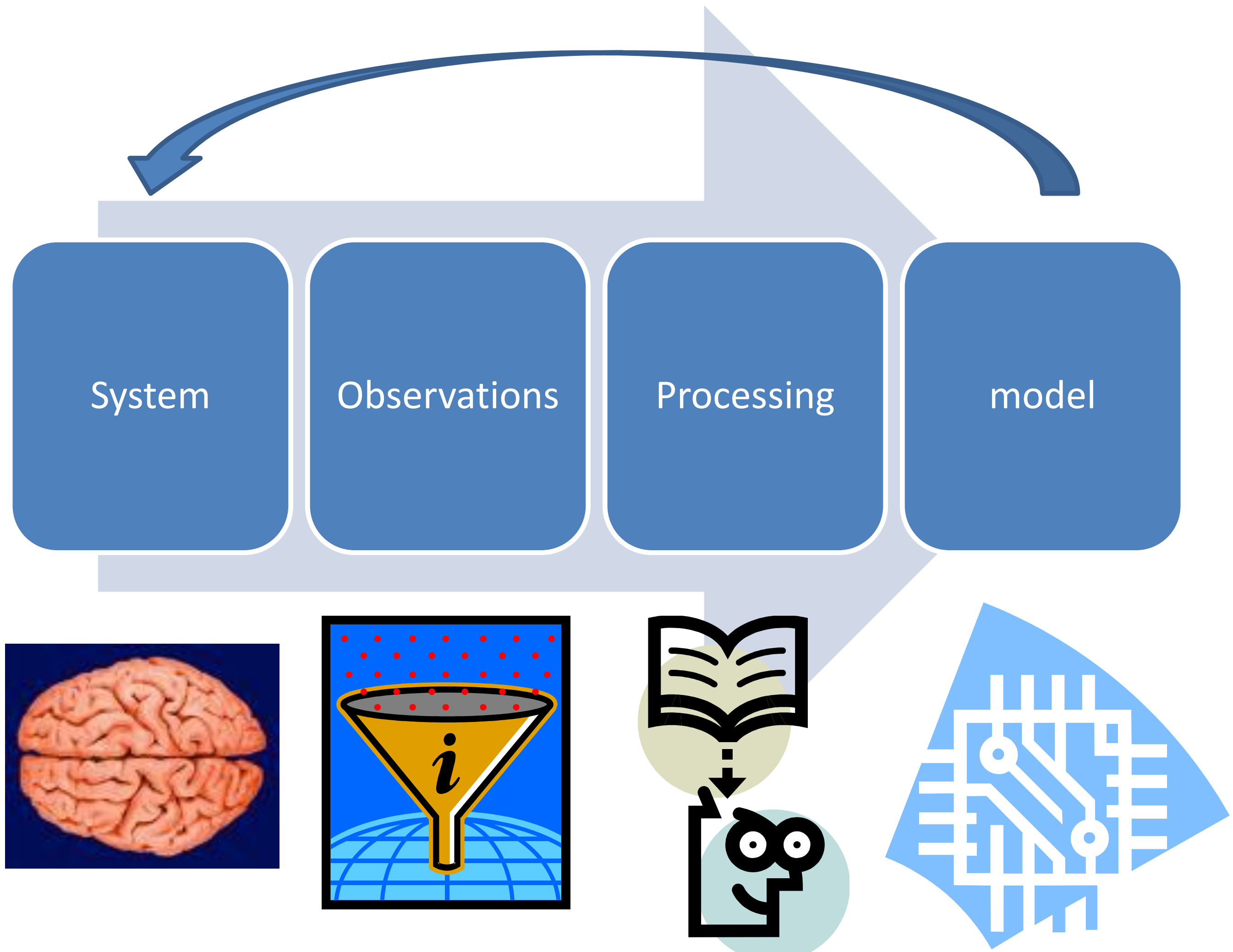
Research cycle







Research cycle



Science in the 21st century

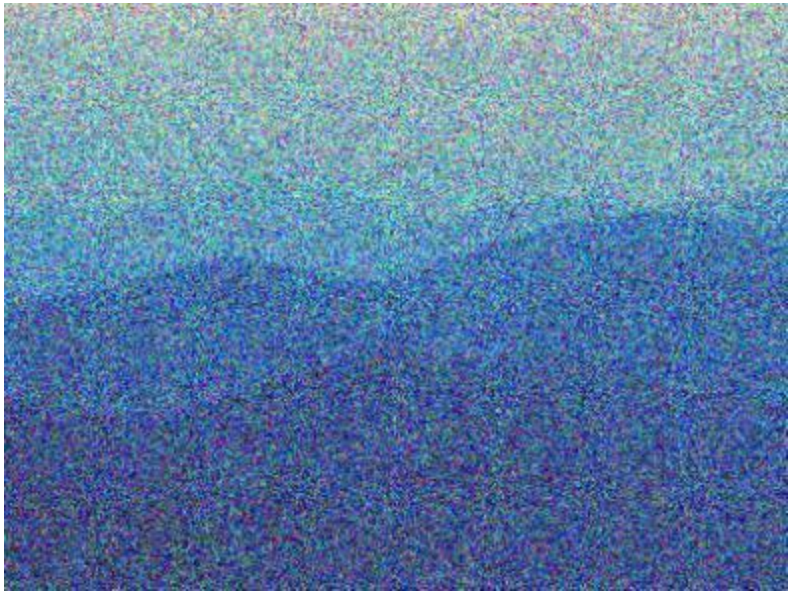
- Size and complexity of problems
  - Climate studies,
  - Human cognition,
  - The Cell
  - Elementary structure of matter,
  - Language
  - Social networks

Research objective facticity project

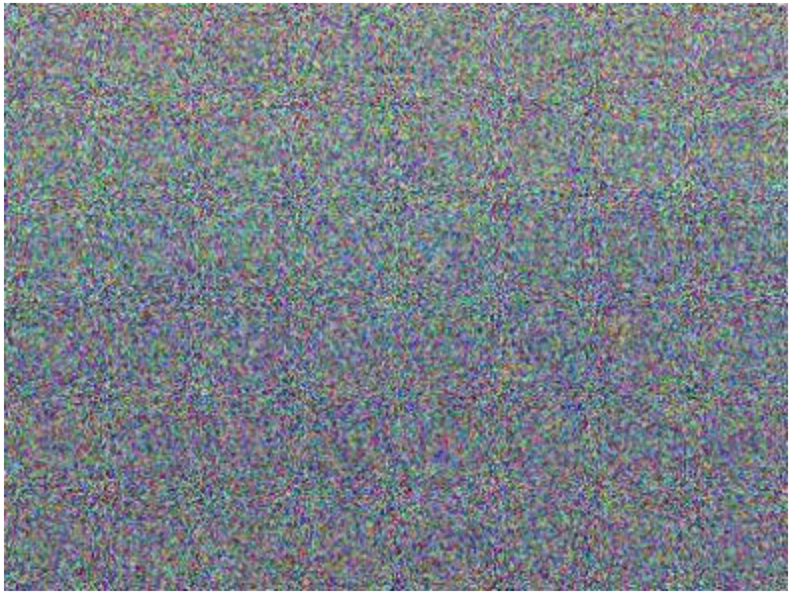
- Formulate a mathematical theory about an objective measurement of the amount of model information in a data set

25 % noise

50 % noise

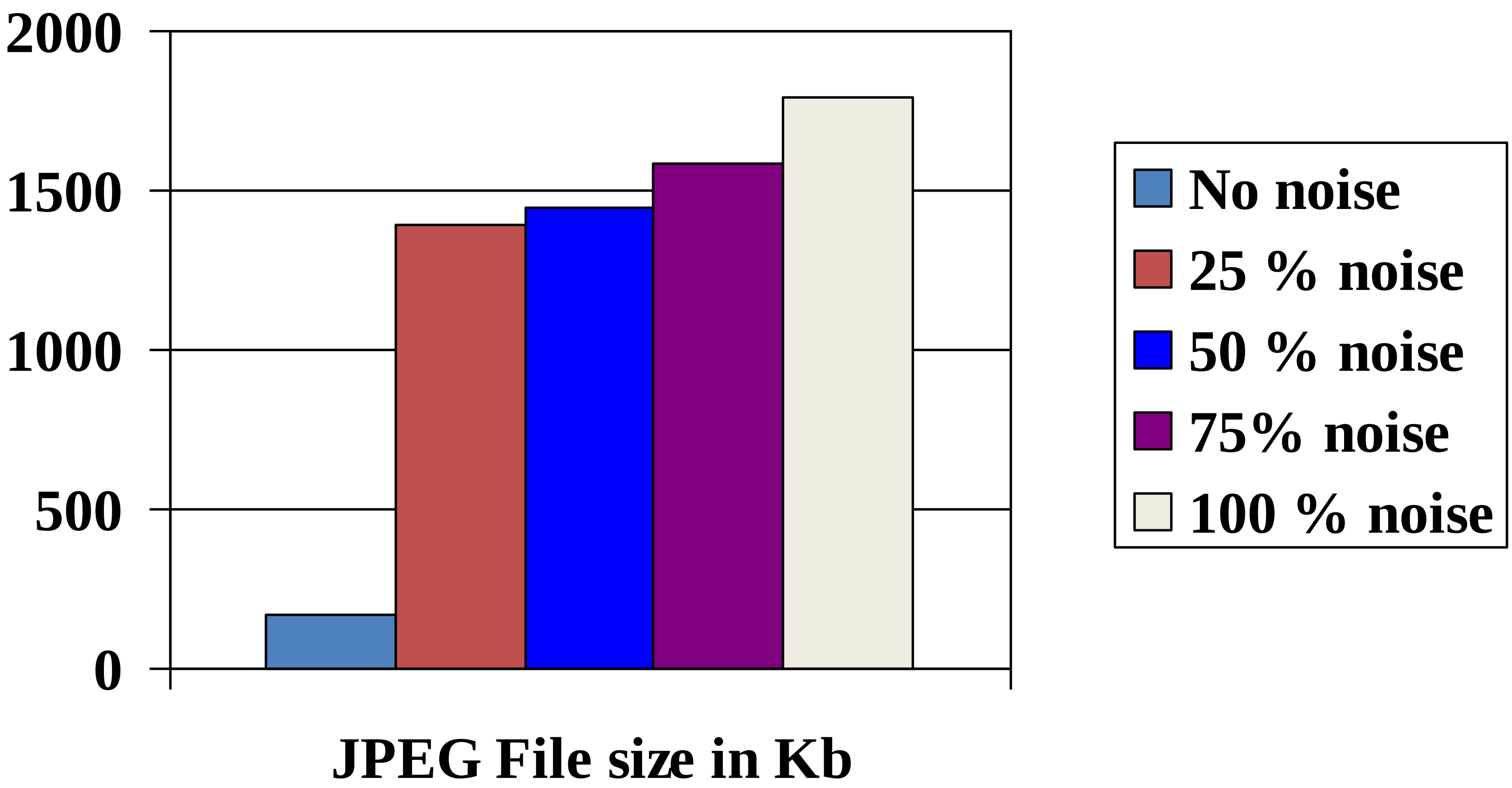


75 % noise



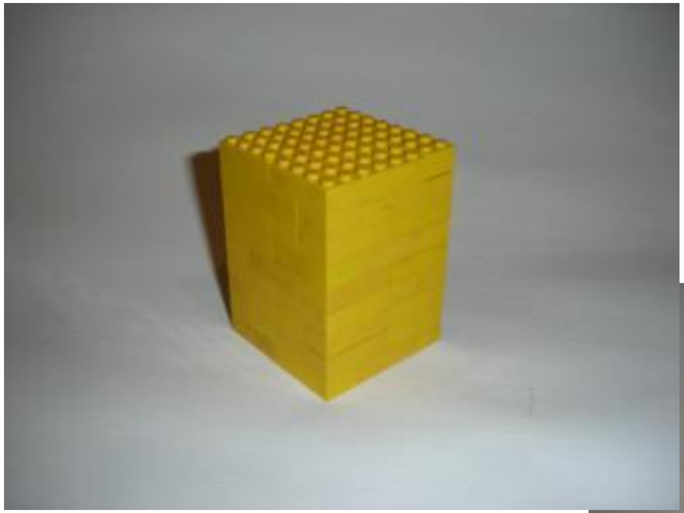
100 % noise

JPEG File size with noise added

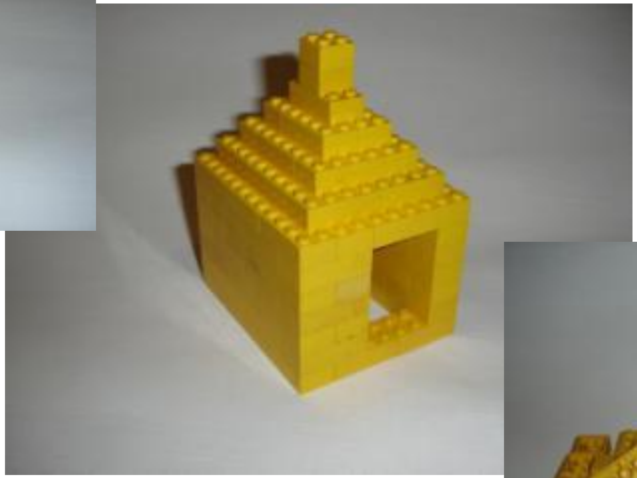


Between order and chaos: facticity

Order



Facticity

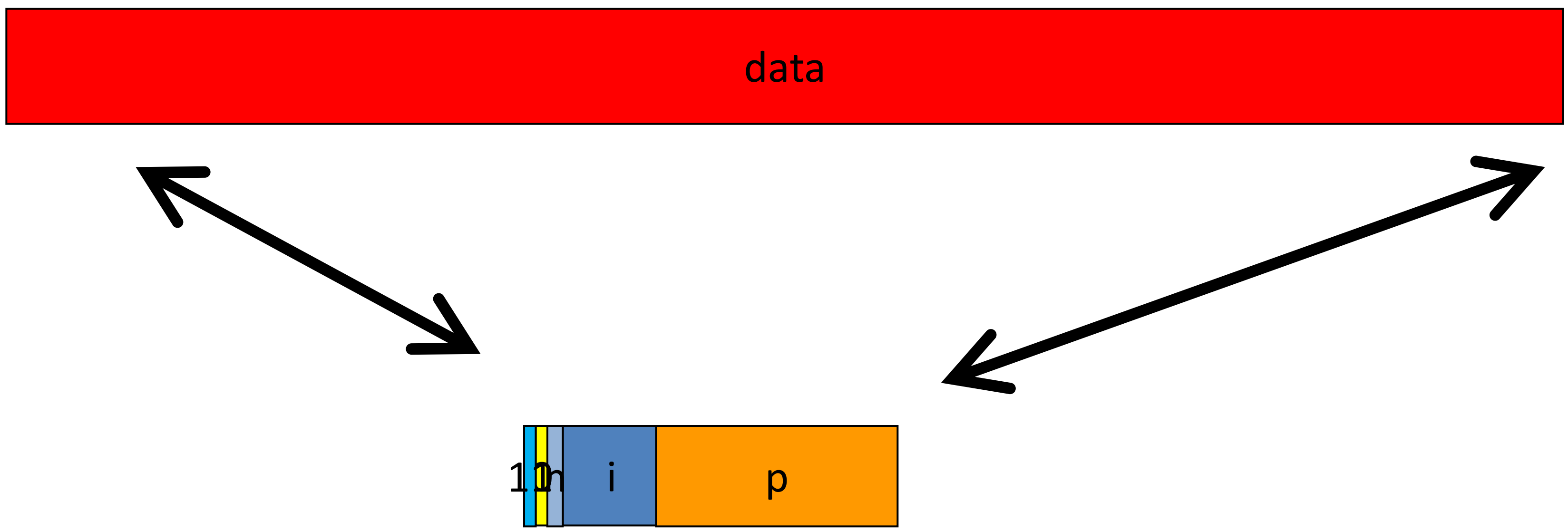


Chaos

Facticity (Questions)

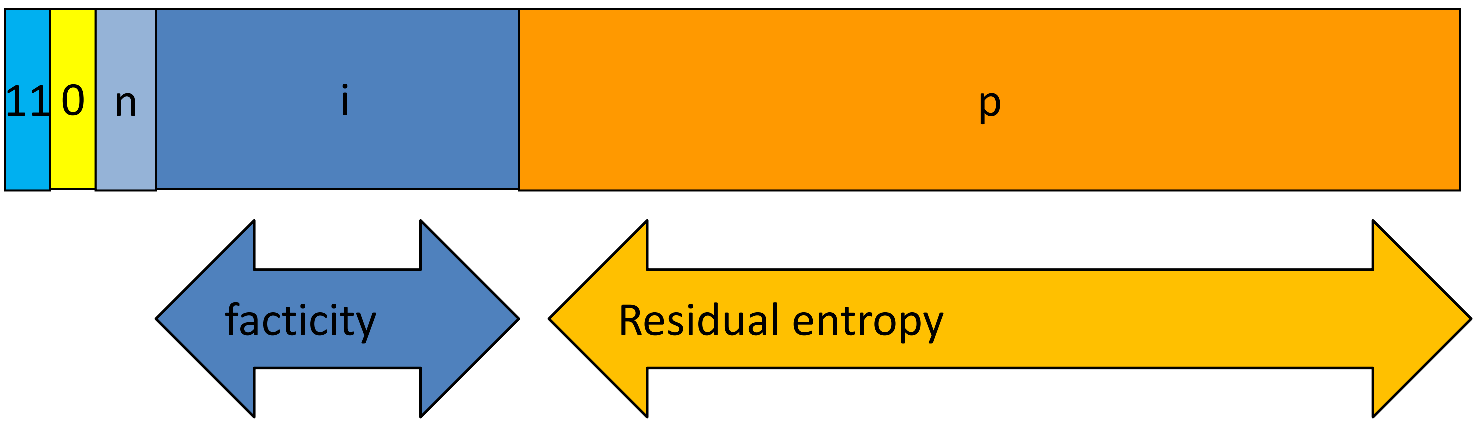
- Is there a universal model extraction method? (yes)
- Are there possibly more optimal models? (yes)
- Can we be sure that it is“definite”” i.e. all relevant model information is extracted and nothing more (tricky, but yes)
- Is facticity stable? (No, but more a feature then a bug)

Turing two-part code compression



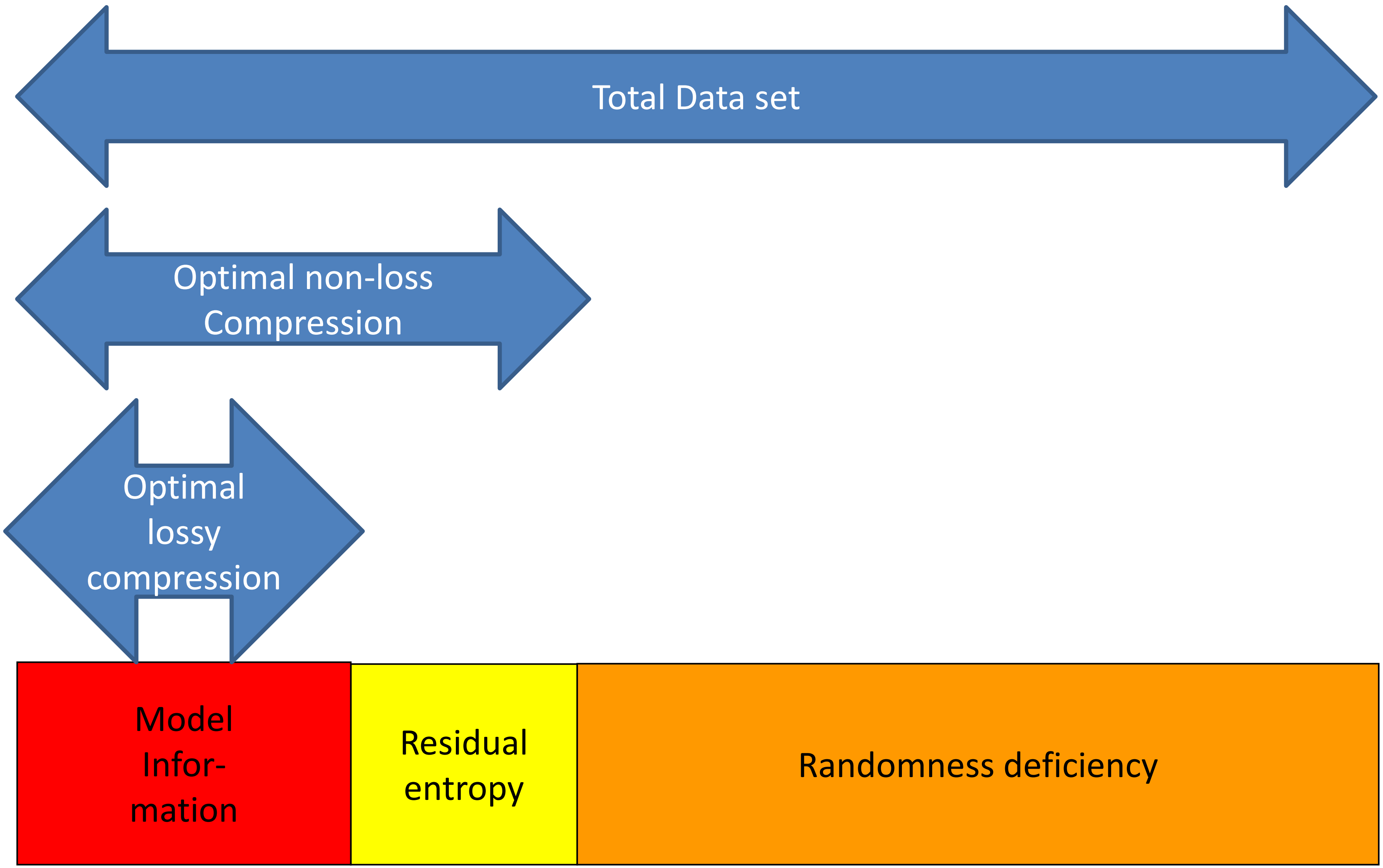
Definition of facticity: the amount of self-descriptive information in a dataset

$$\varphi^{U}(x) = \arg\min_{i} \{|\, i \,|: \{\min_{i,p}\{|\, \bar{i} \,| + |\, p \,|: U(\bar{i}p) = x\}\}$$



- The facticity ϕ(x) of a string x is the amount of self descriptive information in x.
- Facticity is definite!

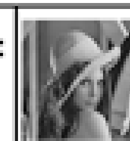
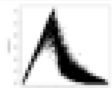
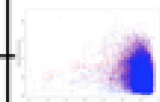
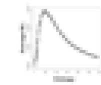
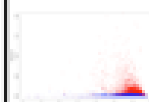
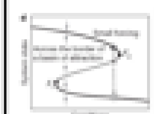
Crash Course Complexity Theory



existing complexity measures

- entropy estimators: (Shannon, Kolmogorov, VC-dimension)
- model complexity estimators (sophistication, computational depth, self-dissimilarity, facticity),
- network-analysis measures (structural graph-theoretic properties, centrality, betweennes, degree distributions)
- complex systems analysis (multi-scale analyses, robustness, dynamics)
- and their theoretical connections.

Atlas of Complexity: Comparative Analysis of Metrics & Datasets

Datasets / Metrics		1-D CA	2-D CA	Bit Strings	RNA sequence data	Protein sequence data	Human text	Microarray & GWAS	Networks: Concept Web		RANDOM.ORG
Shannon Entropy											
Mutual Information											
Crutchfield Complexity											
Local Entropy Variance (per Herman)											
Critical Slowing Down (per Scheffer)											
PCA (per Lude)											
Facticity (per Pieter)											
Zip compression											
MIC (per Yahya)											

NSURANCE
ROFITS are
HEALTH CARE
FOR AMERICA
NOW!
HEALTH CA
FOR AM
LARRY KING LIVE
February 26, 2010
30 Million Americans, 7,000 Rare Diseases
By Peter Saltonstall
Pres. & CEO of NORD - National Organization of
Rare Disorders
While the eyes of the nation are directed toward
CVZ: dure medicijnen voor drie zeldzame ziektes uit basispakket
CVZ
'De verhouding tussen kosten en baten is onacceptabel'
'Het doorslaggevende argument is de buitengewoon ongunstige kosteneffectiviteit'
80
ziekte van Pompe
Het College voor Zorgverzekeringen
drie medicijnen voor twee zeldzame
niet langer vergoeden uit het
Dat staat in twee conceptadviezen,
door Niels
DSM growth
1000
750
500
250
0
1940
1955
1970
1985
2000
year
pages
diagnoses

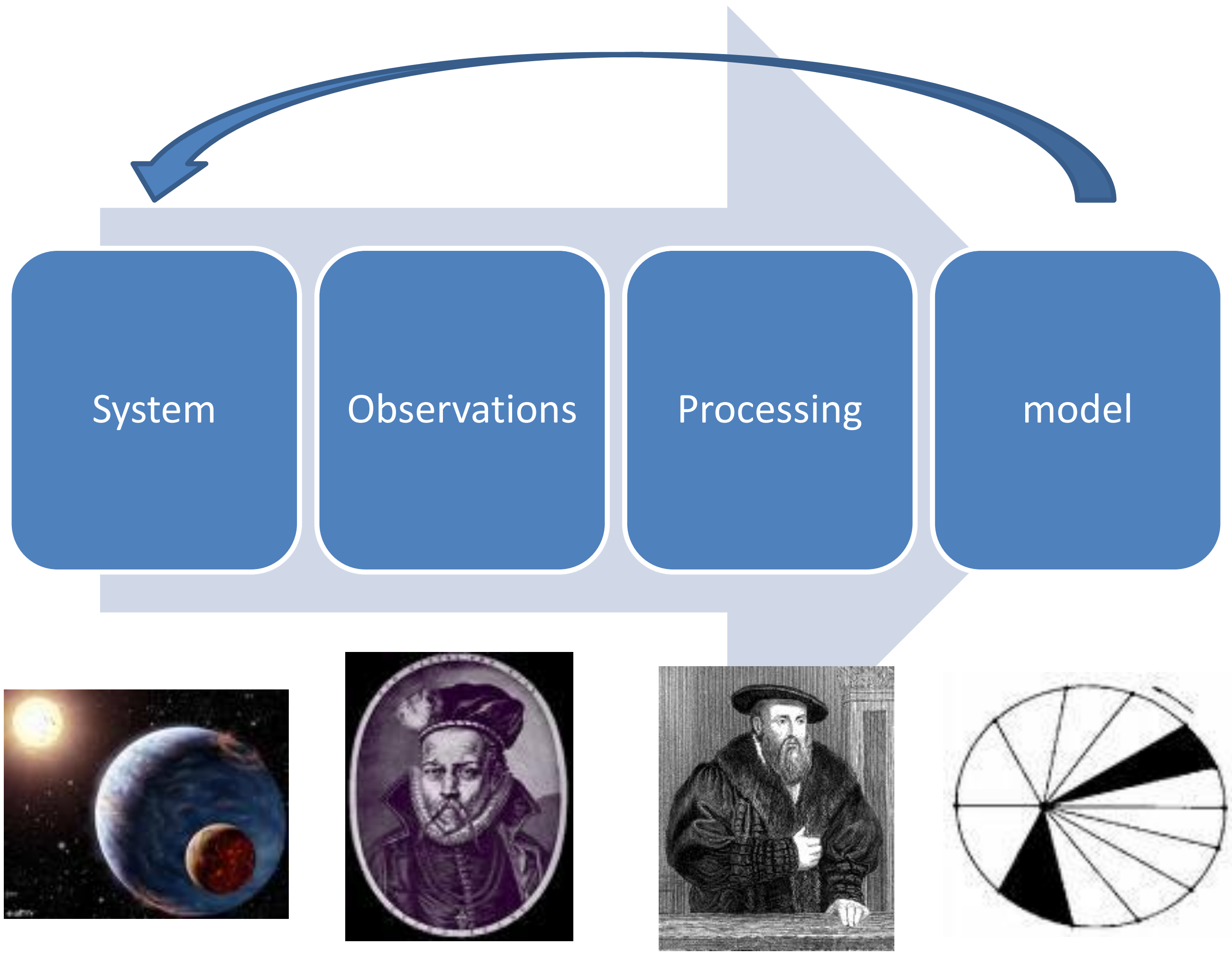
Research Questions

- What is a medically relevant model of an individual human being?
- How complex is such a model?
- How individual is such a model?
- How difficult is it to learn or make such a model?

medically relevant model of an individual human being

- A computational model (or program) that allows us to predict, explain (and if possible, help cure) any disease an individual human being X could have or get.

Research cycle



Research Methodology Proposal I

- Make a list of all possible diseases.
- For every disease study a relevant group of patients empirically and make a model that allows us to make deterministic predictions.

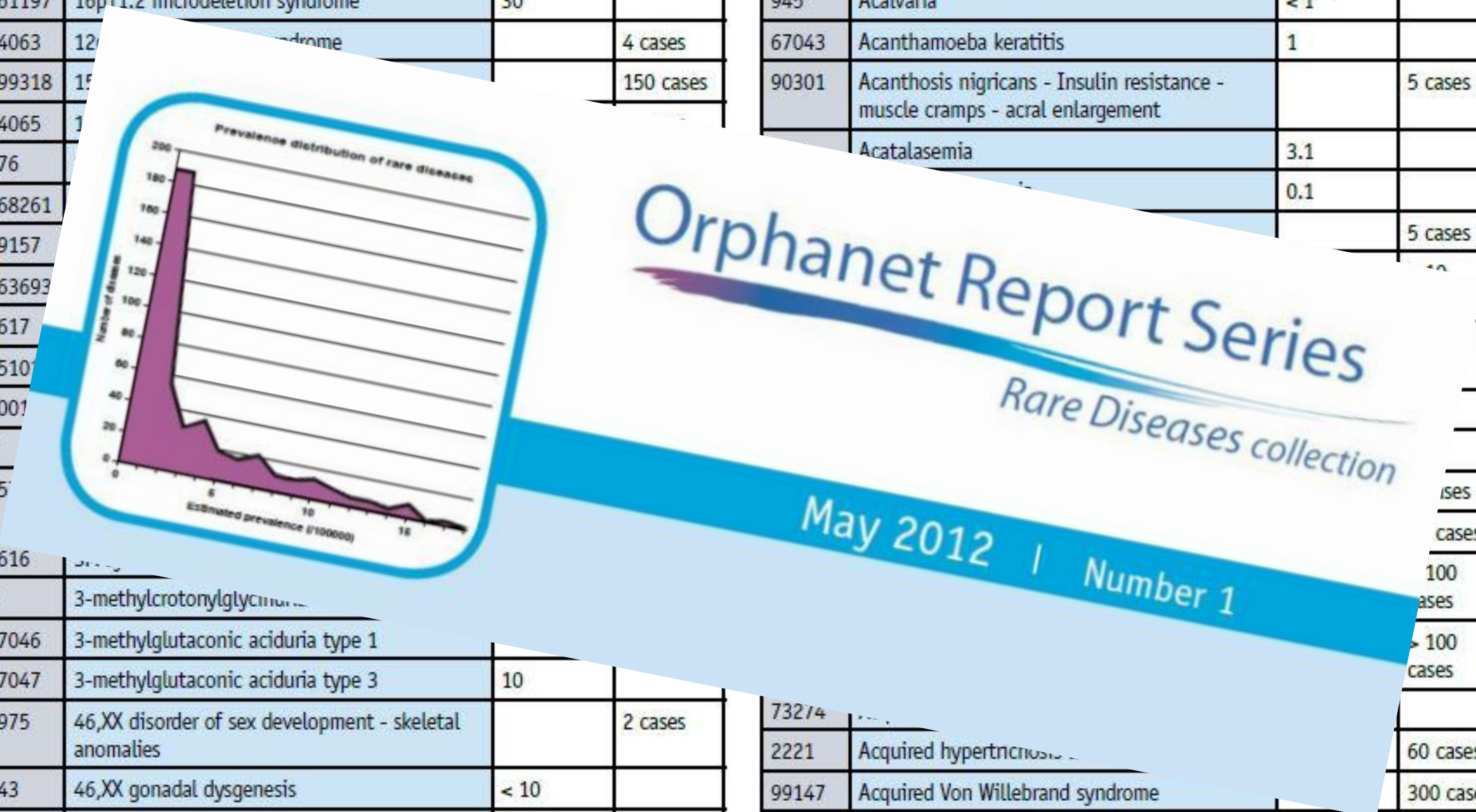
- How much diseases are there?
- 2000?
- 7000?

...w.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf

...e door deze pagina gebruikte plug-ins zijn verouderd. Plug-ins bijwerke...

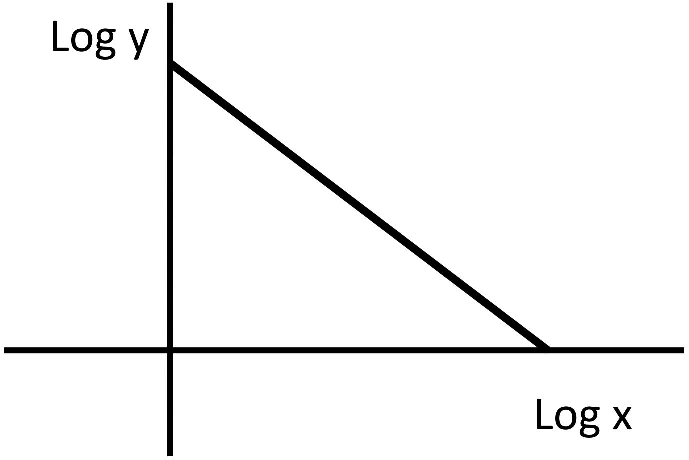
...RPHA ...umber	Disease or group of diseases	Estimated prevalence (/100,000)	Number of published cases or families
...61197	16p11.2 microdeletion syndrome	30	
...4063	12... ...drome		4 cases
...99318	15...		150 cases
...4065	1...		
...76			
...68261			
...9157			
...63693			
...617			
...510...			
...001...			
...616	3-...		
	3-methylcrotonylglycin...		
...7046	3-methylglutaconic aciduria type 1		
...7047	3-methylglutaconic aciduria type 3	10	
...975	46,XX disorder of sex development - skeletal anomalies		2 cases
...43	46,XX gonadal dysgenesis	< 10	
...68558	46,XY disorder of sex development - adrenal insufficiency		9 cases

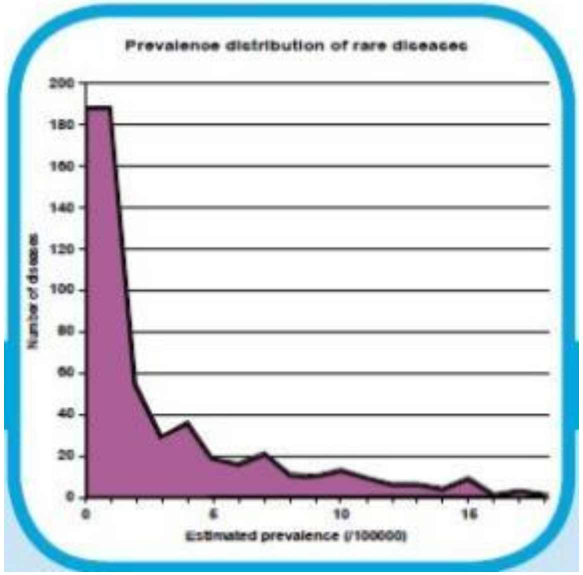
ORPHA Number	Disease or group of diseases	Estimated prevalence (/100,000)	Number of published cases or families
945	Acalvaria	< 1**	
67043	Acanthamoeba keratitis	1	
90301	Acanthosis nigricans - Insulin resistance - muscle cramps - acral enlargement		5 cases
	Acatalasemia	3.1	
		0.1	
			5 cases
			...ses
			... cases
			... 100 ...ases
			> 100 cases
73274			
2221	Acquired hypertrichosis...		60 cases
99147	Acquired Von Willebrand syndrome		300 cases
36	Acrocallosal syndrome		34 cases

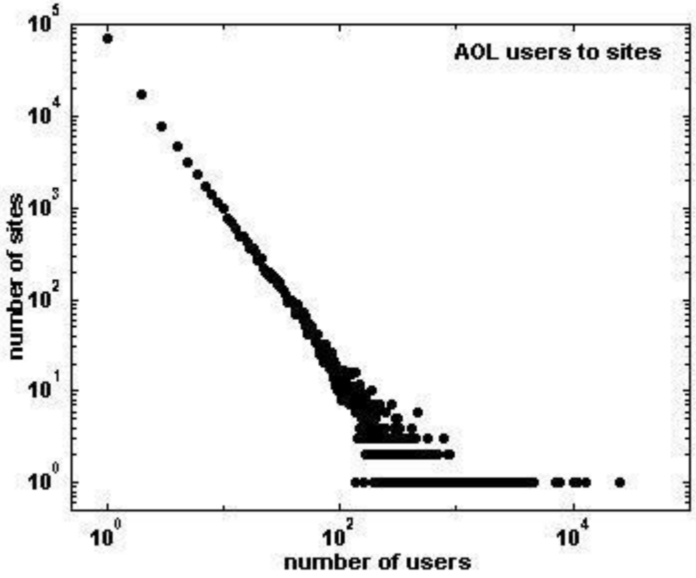


Powerlaws

- $\log_c y = -a \log_c x + b$
- $y = c^b x^{-a}$







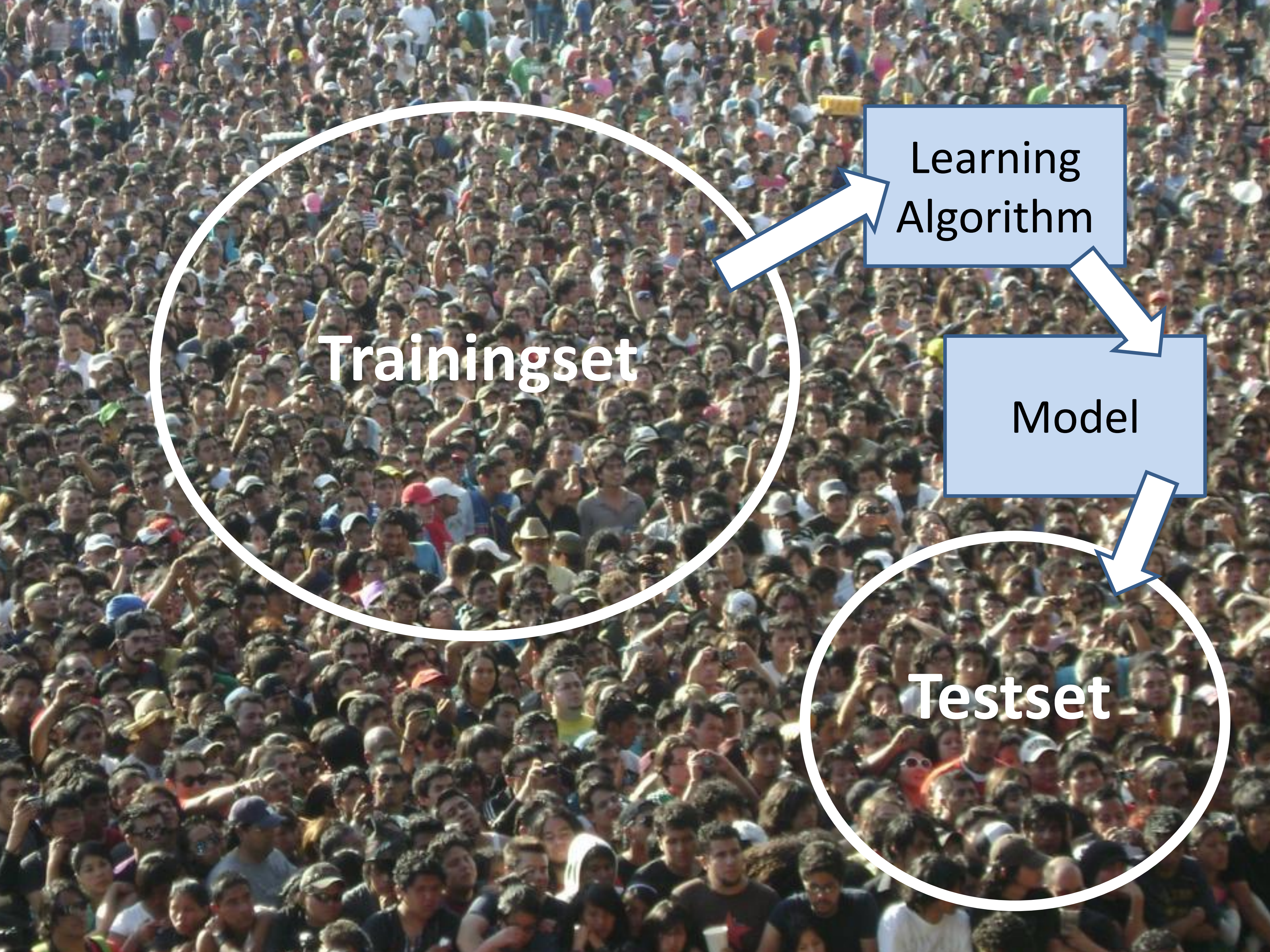
Problems!

- Power laws have no mean (i.e. it is infinite)
- If you double the population, you not only double the number of patients, but also the number of diseases.
- There is a (seemingly) unlimited supply of rare diseases.
- Most possible diseases are so rare that we never see them!

Research Methodology Proposal II

- Forget the list of all possible diseases.
- Forget deterministic prediction.
- Just study a relevant group of patients that is large enough to make statistically relevant risk assessments.

- How big a group of patients?


Learning Algorithm
Trainingset
Model
Testset

VC dimension

The bound on the test error of a classification model:

$$\text{training error} + \sqrt{\frac{h(\log(2N/h)+1) - \log(\eta/4)}{N}}$$

Probability $1 - \eta$

Size of training set *N*
VC dimension: *h*
Condition: *h* << *N*

Problems again!!

Size of training set *N*

- Maximal:
  current world population
  = 0.7 x 10^10

VC dimension: *h*

- Maximal size of the medically relevant model of an individual human being

Condition: *h* << *N*

- 10^8 bits ≈ 12,5 Mb

12,5Mb!!

Maximal size of the medically relevant statistical model of an individual human being.

Information: Some numbers

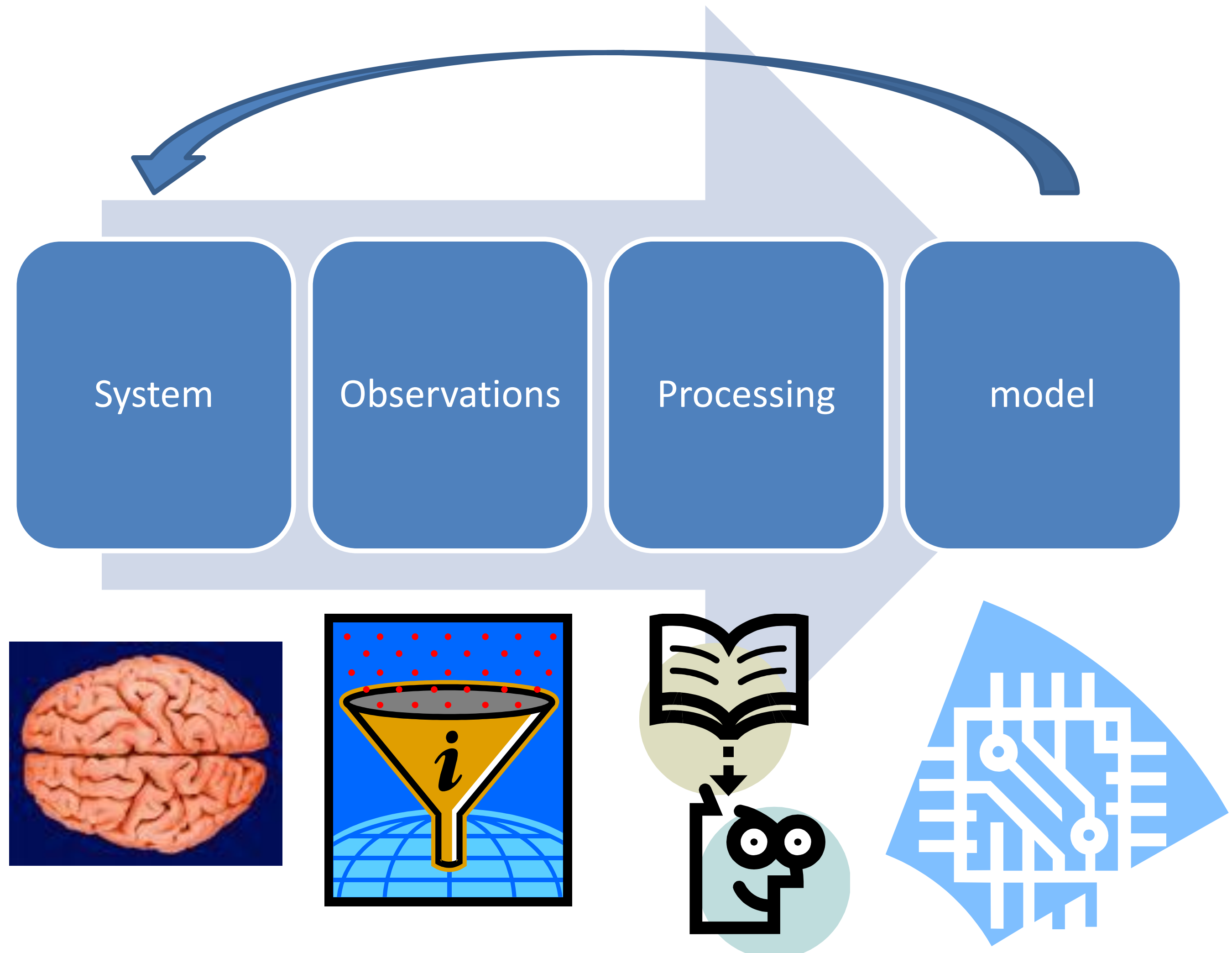
- 10^{10} bits Human genetic code
- 10^{14} bits Human brain
- 10^{32} bits Human body at quantum level (10^{28} electrons)
- 10^{92} bits Total Universe

- 10^{123} Total number of computational steps since Big Bang (Seth Lloyd)

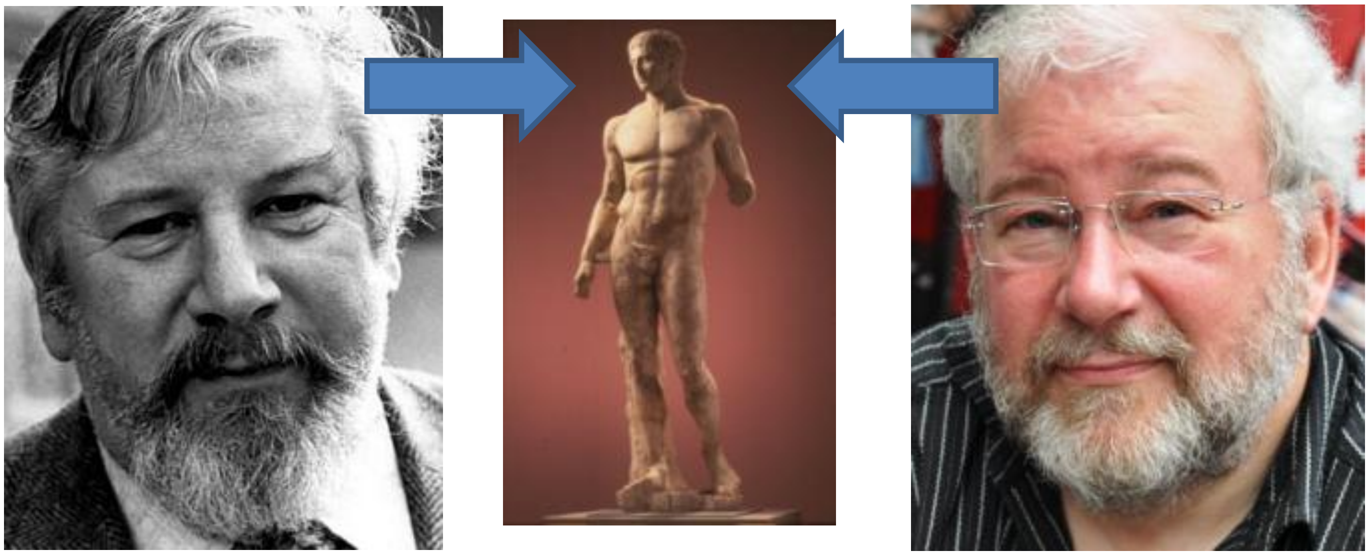
Empirical Incompleteness theory

- Our universe contains classes of entities that have a population size that is much smaller then their relevant model complexity measured in bits.
- No adequate statistical models of such classes of entities based on frequency observations can be constructed.

Research cycle



Mutual model information
The classical view

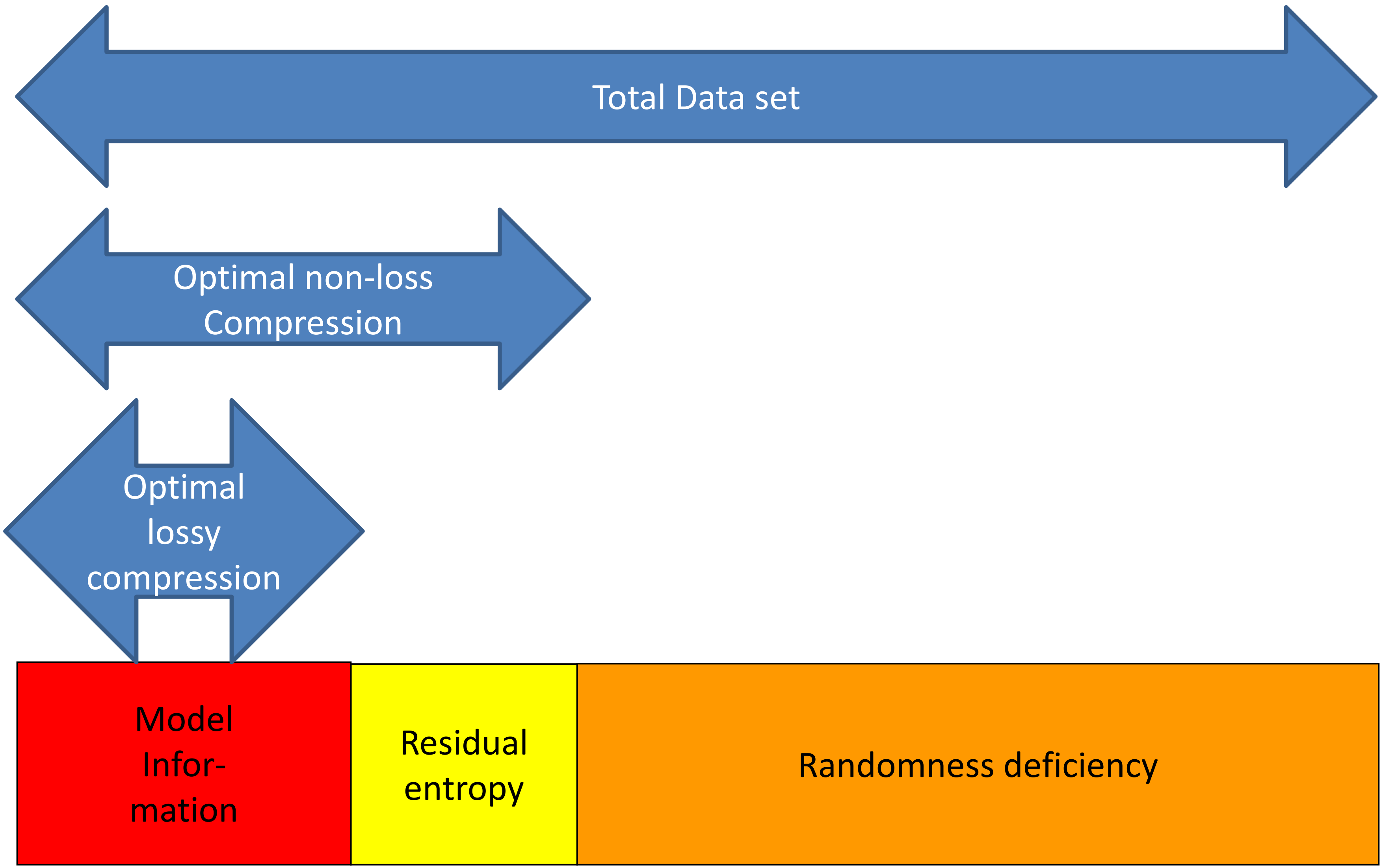


Ideal	Faults

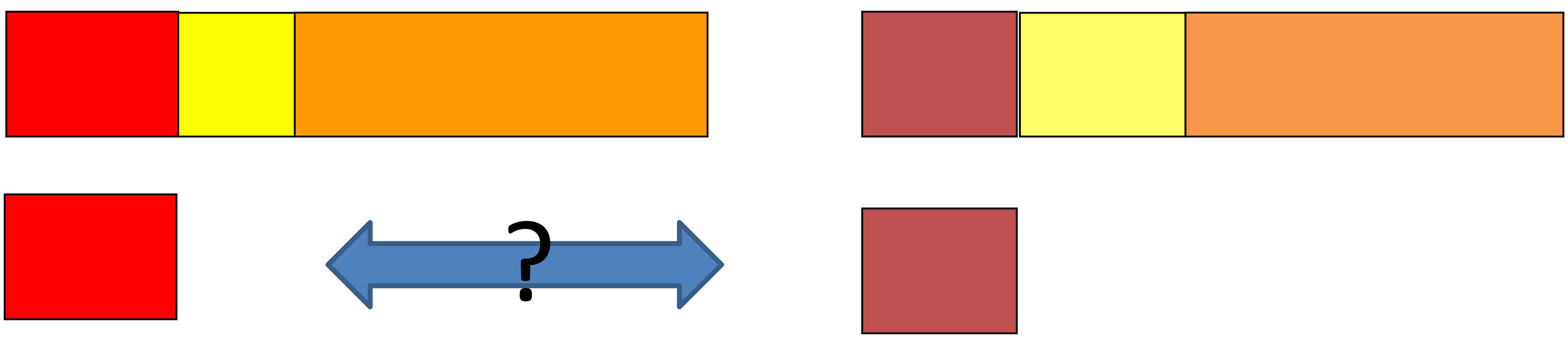
ideal	Faults

Same model

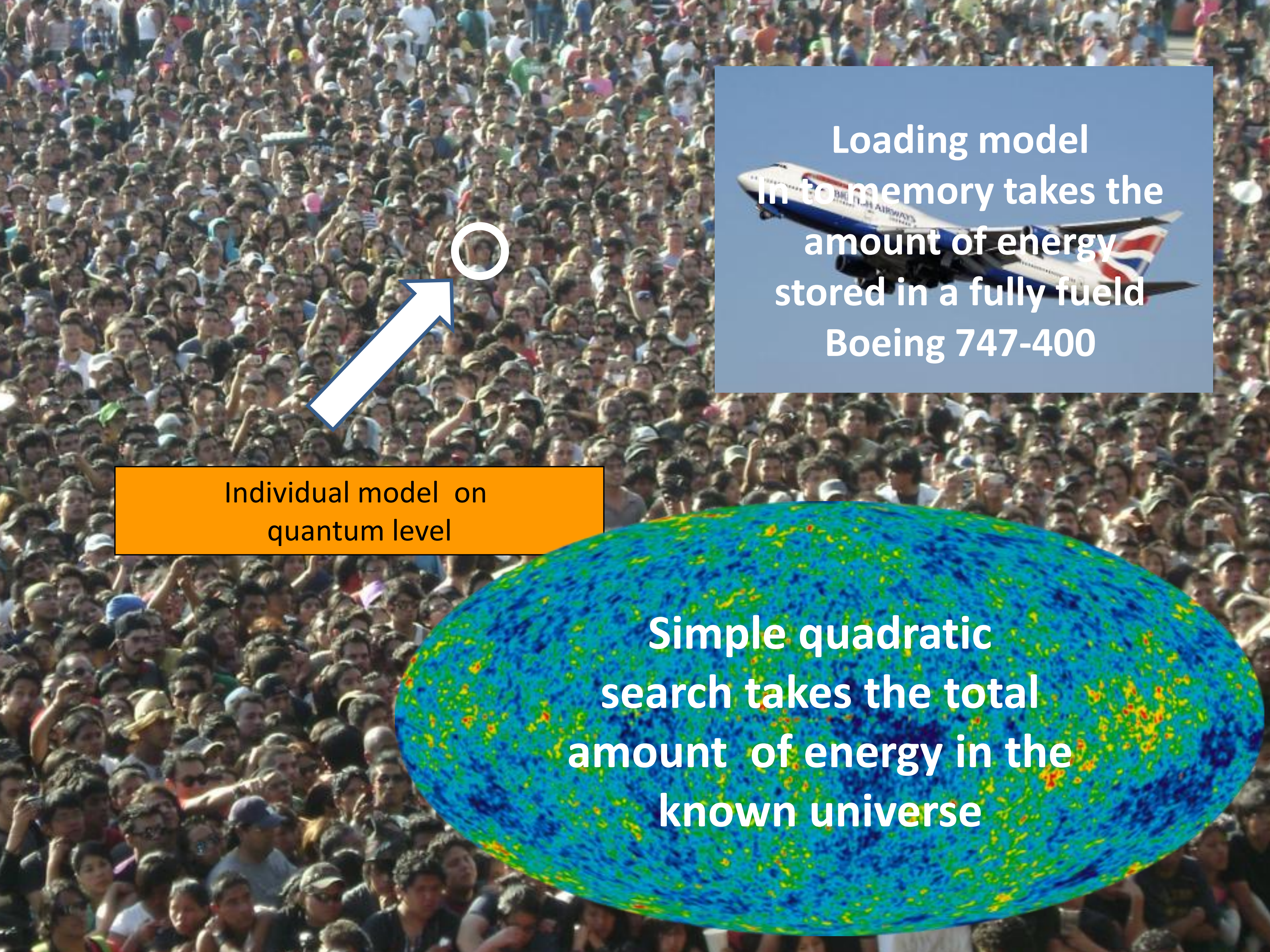
Crash Course Complexity Theory



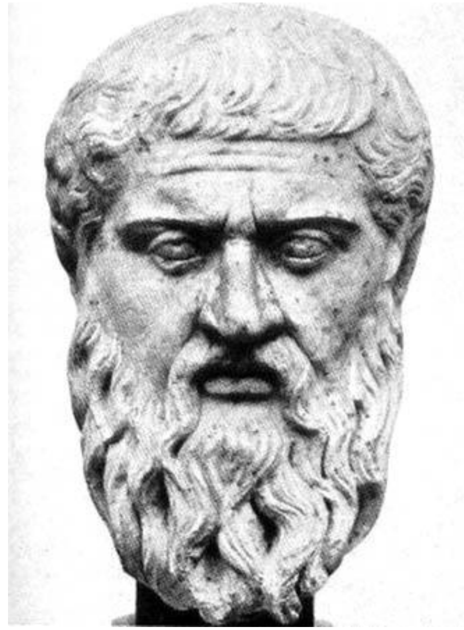
mutual model information?
Computational View

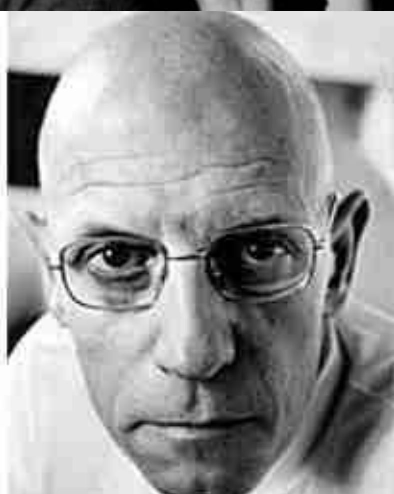


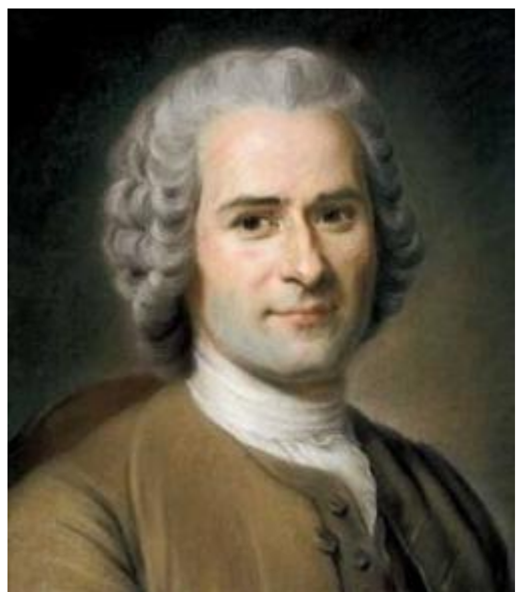
Any two unrelated human beings differ by about 3 million distinct DNA variants.

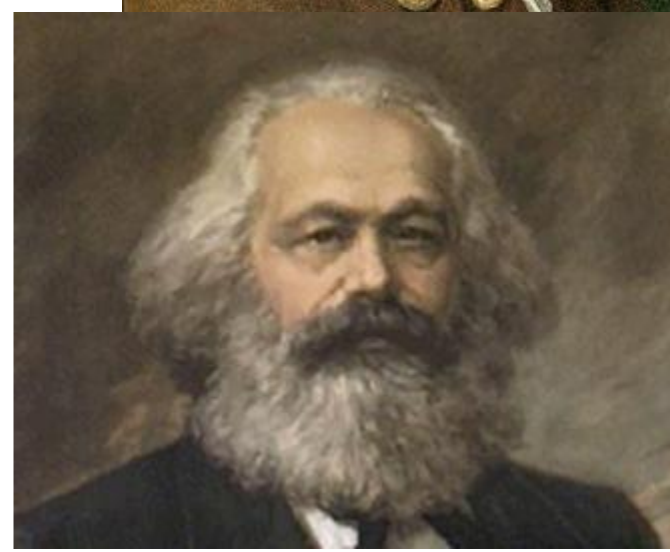

Loading model
In to memory takes the
amount of energy
stored in a fully fueld
Boeing 747-400
Individual model on
quantum level
Simple quadratic
search takes the total
amount of energy in the
known universe

Breaking a long tradition









Conclusions

- Data mining is developing in to the study of Big Data and e-Science
- We need to rethink scientific methodology for the 21st century.
- Our universe contains classes of entities that have a population size that is much smaller then their relevant model complexity measured in bits.
- The individual is the species!
- Consequences for:
  - Political organization of society
  - Ethics and human rights
  - Scientific Methodology
  - Medical policies
    (Caritas versus evidence based medicine)