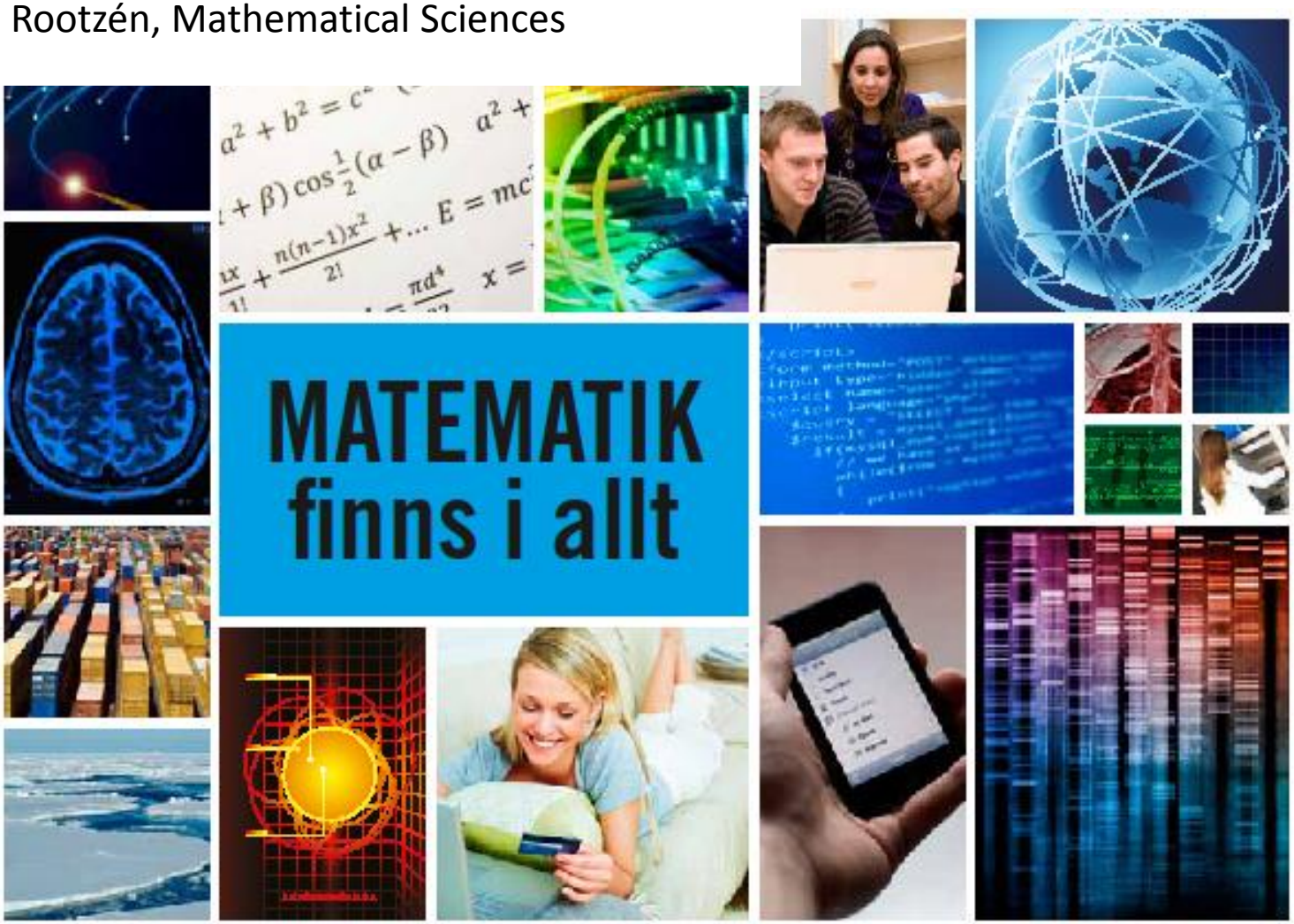
Understanding Big data & Big systems

Holger Rootzén, Mathematical Sciences



We all know that it is happening – but it still is hard to fathom how fast it will revolutionize our world:

Library of congress contains more than 96 miljon books and 100 terabyte ($100x10^{12}$) data. In 2014 a miljon times more will be created, and in 2023 100 times more – and it will be only a mouseclick away

The internet has more than 3 billion users; the worlds telphone networks connect 6 billion mobile phones; physics experiments become bigger and bigger; systems for making astronomical measurements are fantastically sophisticated and complicated.

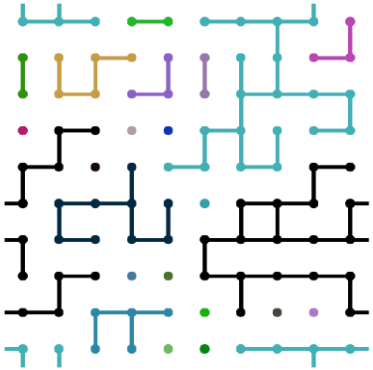
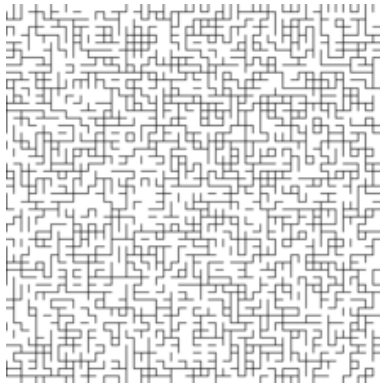
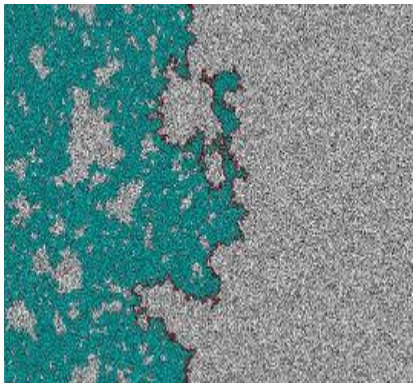
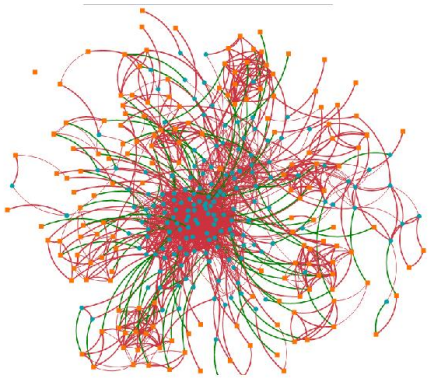
Natural systems – as we see them in larger and larger resolution – exhibit more and more details and complexity

need

more abstraction
more generality

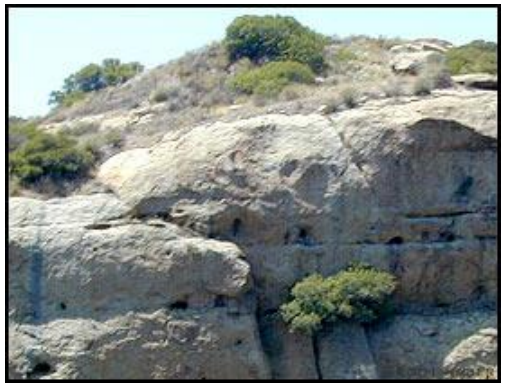
more mathematics

percolation

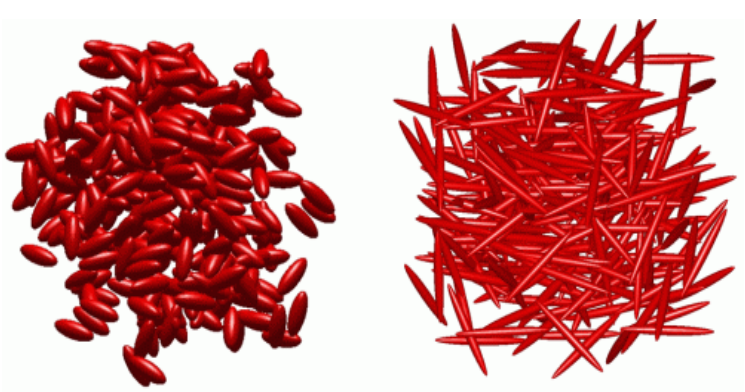
Simple local rules lead to complicated system behavior
System behavior informs about local rules

Helps understanding of



water flow through porous rocks

structure of social networks



drug delivery through gels

Big data and Big systems affects the lives of all of us (and will affect our children)

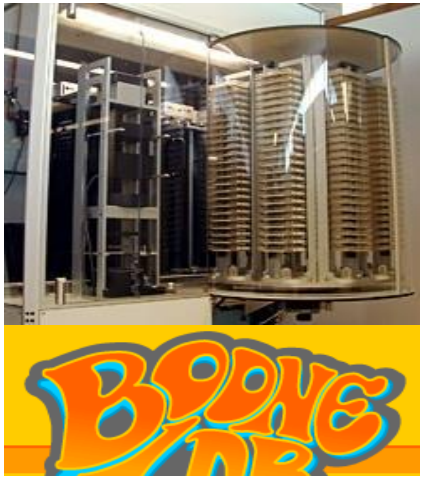
	past	future
see advertizements	in newspapers, TV, billboards	on web, choosen just for you

cost of insurance	accident rates of ”your group” + your own accidents	this + your internet trail: Facebook, Instagram, Twitter, blogs, ….

get credit	credit behaviour of ”your group” + your own credit history	this + your internet trail: Facebook, Instagram, Twitter, blogs, ….

Nothing is ever forgotten

good or bad??



Some things we do:

understand 27 millon p-values from a study of interaction between pairs of genes in yeast, coming from 27 million experiments in 4 replications

understand how huge datasets from weather satellites, buoys, weather stations, and sensor systems can improve safety and efficiencey of marine transports

understand traffic accidents from a naturalistic driving study where 2000 cars equipped with web cams, radars, gps-s, sensors are driven in normal traffic for 3 years by drivers like you and me, and all data is saved

Traffic accidents

- 1.3 miljon deaths/year worldwide, 20-50 million severely injured
- Large economic losses
- Less than 1 death/day in Sweden now. Down from 3 deaths/day a few decades ago – at a time with much less traffic
- First simple measures: seatbelts,helmets, follow traffic rules, drunk driving laws, …, then more sophisticated ones: rebuild roads, better tires, improve driver education, airbags, …, then next level of sophistication: more driver training and retraining, ABS, ESP, … ??

Naturalistic driving studies: cars with drivers like you and me are instrumented with video cameras, radar, GPS, sensors for steering wheel movement, gas- and brake pedal movement, …

Crash Acc.mpg

Generates extremely large data sets – and give completely new opportunities for preventing traffic accidents

Accidents are extreme events – same methods as those for financial risks and for natural catastrophes (later in talk!) can be used

New and exciting area for statistics

Selection bias/errors
Risk estimation

Active safety systems for next generation cars. Important for competition with other car makers and for safety (?)

Driver training, traffic laws, …

Visual behavior/censoring

How much do you look off road while driving?

- ❑ 5% of the time
- ❑ 10% of the time
- ☒ 15% of the time
- ❑ 20% of the time

There is a 1 in 1000 chance that the lengths of an off road glances is longer than

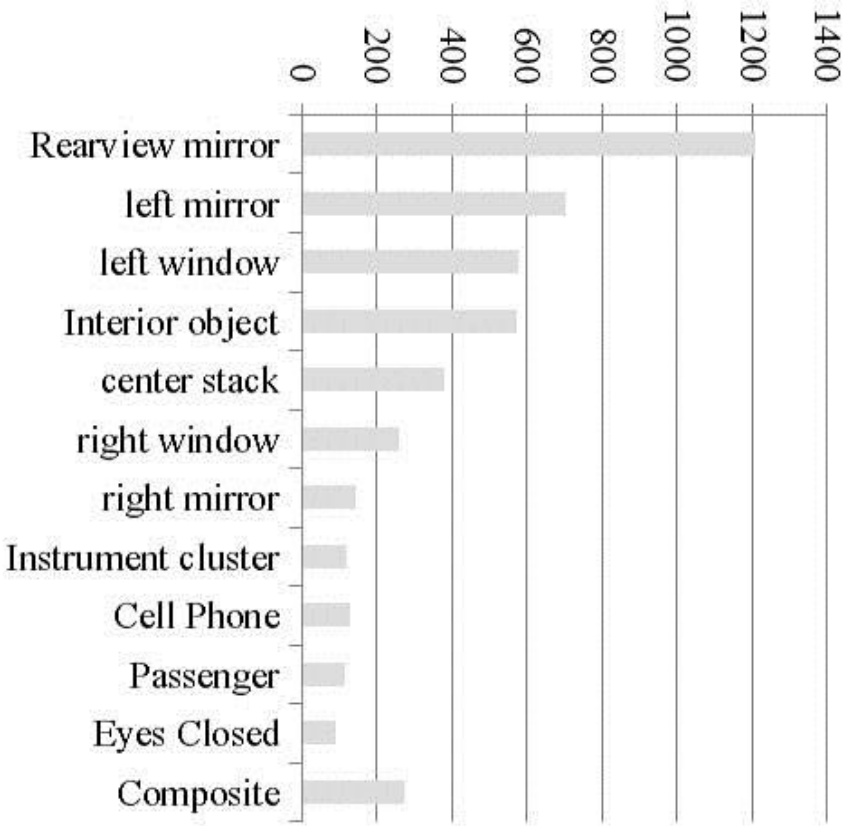
- ❑ 1 second
- ❑ 2 seconds
- ❑ 3 seconds
- ❑ 4 seconds
- ☒ 5 seconds
- ❑ 10 seconds

Is glance behavior different in different circumstances?

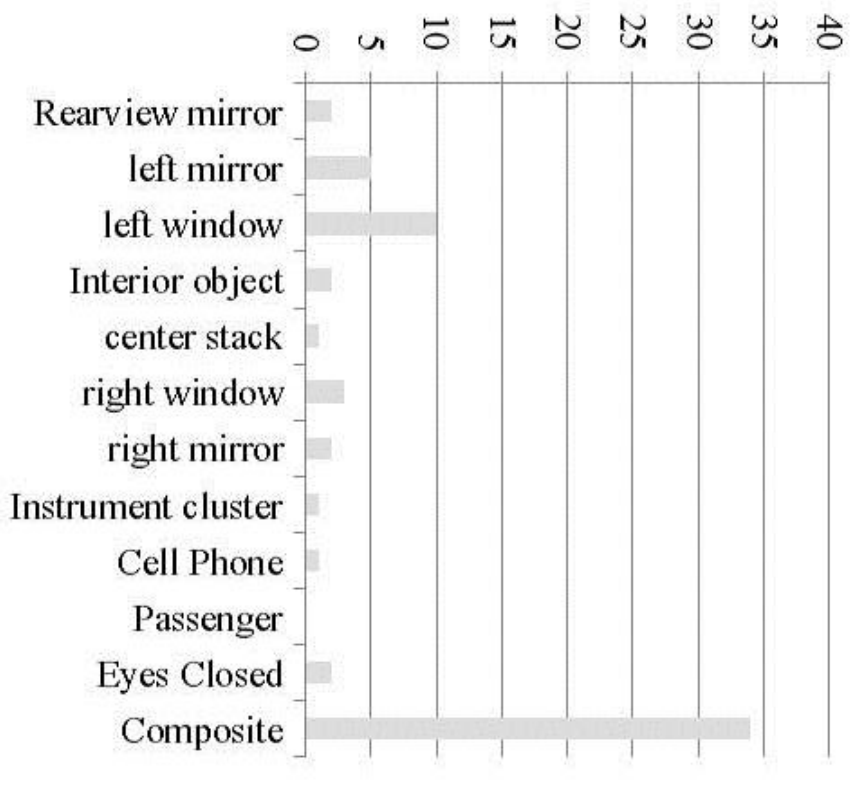
Not well understood

Glance behavior in the 100-car study

Raw data: 19,616 annotated 6-second intervals from 100-car study: 4582 with 1 or more off road glances



Glances shorter than 3 seconds



Glances longer than 3 seconds

One conclusion

different kinds of nearcrashes and crashes; naturalistics driving studies; vehicles; drivers, all lead to different kinds of

- Selection bias
- Crash proximity measures
- Driver behavior – and ”covariates”

All require separate careful analysis

No omnibus answer to ”is there selection bias in choice of near-crashes or the estimation of driver characteristics”

SHRP 2

- 2000 cars
- 3 years
- Much better instrumentation (?)
- Started a two years ago

Some more things we do

Catastrophes

(Risk handling)

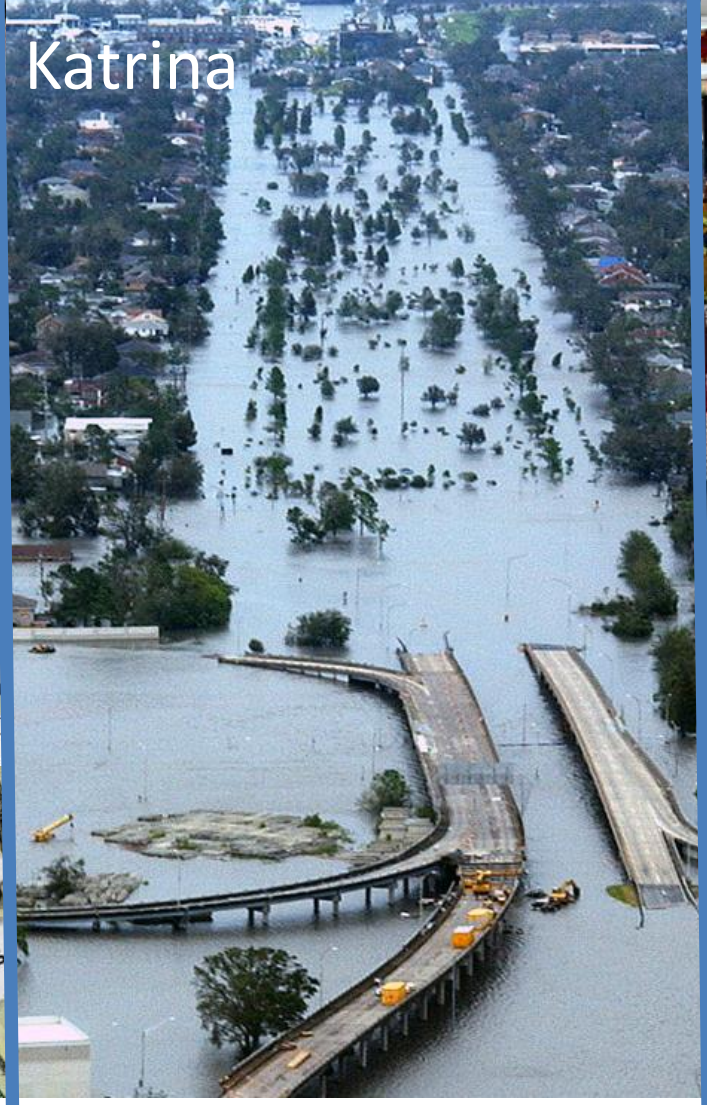
Not this talk

- Normal statistics: uses typical values, helps make life longer, feed the world

This talk

- Extreme value statistics: only uses extreme observations, helps avoid floods, bridges collapsing, increase safety of cars, ships, airplanes, understanding of climate
- The wonderful and useful generality of mathematics and statistics

September 15, 2008
New York



August 29, 2005
New Orleans

All the time
everywhere
1,3 milj dead/year

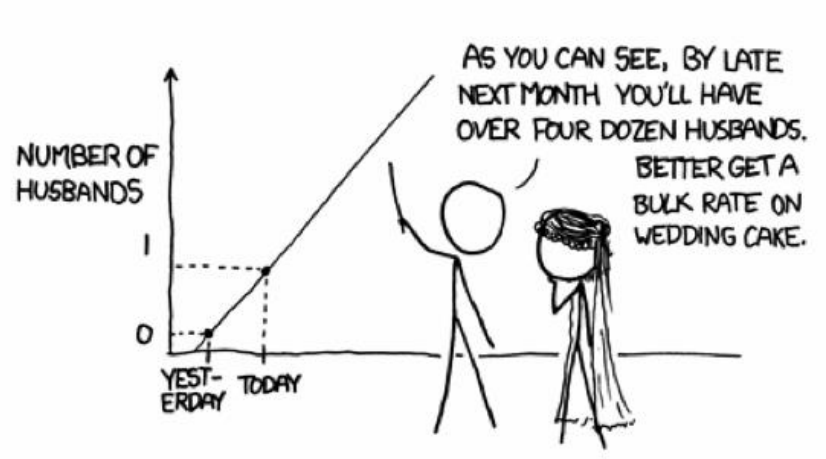
Nassim Taleb: Black swans

Cygnus atratus

- Rare: Have not occured before
- Extreme consequences
- Unpredicatable before, easy to predict after they have happened
- Shape history

Taleb (approximately): History is shaped by black swans – so not worth your while to try to manage risks

Taming grey swans with statistics

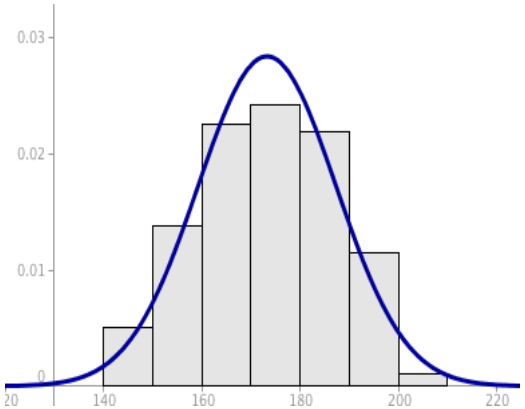


Not (good) statistics

Statistics: to find stable patterns which have happened many times, and which one believes (sometimes because of lack of better knowledge) will continue to repeat themselves

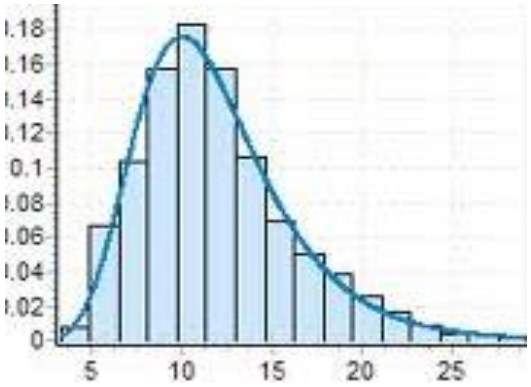
Stable patterns → statistical distributions

Matematics → shapes of possible distributions
Statistics → choice of specific distribution (with uncertainty)
Distribution → quantifies **risk**



Normal distribution

Error in the measured distance = sum of many small measurement errors



Extreme value distribution

Highest water level during year = maximum of daily water levels

Statistics can tame swans ***if***

- One has realized they exist
- One has enough data
- There are sufficiently stable patterns in data – and the patterns continue to repeat themselves in the future

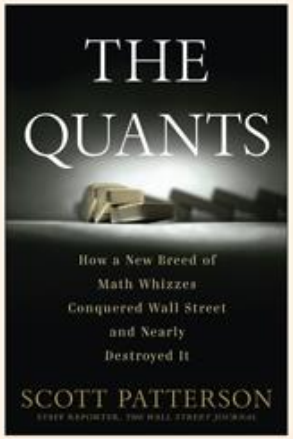
Then statistics can help judging: ”How big is the risk?” -- and what types of catastrophes are most likely (how to handle the risk then depends on the situation)

If not then statistics cannot help

Scott Patterson

The Quants: How a New Breed of Maths Whizzes Conquered Wall Street and Nearly Destroyed It

Crown Business, 2010



$$\Pr[T_A<1, T_B<1] = \Phi_2(\Phi^{-1}(F_A(1)), \Phi^{-1}(F_B(1)), \gamma)$$

The formula that broke Wall street
(David Li's formula)

Complicated credit insurance contracts (CDO-s) were valued by Li's formula
Contributed to starting financial crisis

- **Opinion 1**: Matematics contributed to starting the financial crisis. Ergo – forbid financial mathematics!
- **Opinion 2**: Matematics contributed to starting the financial crisis. Ergo – develop better financial mathematics!

The world has chosen Opinion 2 and implements new mathematics based regulation of banks and insurance companies (e.g. Basel III, Solvency 2)

What do you think?

Megastort skadestånd för Weavering-blåsning

Dagens industri 2011-09-02 08:23

Weavering Capital Founder charged with fraud

The Guardian 2012-12-14

”Weavering Capital-härvan: Magnus Petersons, den svenske grundarens, styvfar och bror döms att solidariskt betala ett skadestånd på 700 miljoner kronor. "Det är fruktansvärt mycket pengar", säger brodern Stefan Peterson. Men det är inte Stefan Petersons inblandning i Weaverings svenska del som nu straffar sig.

Det var i mars 2009 som skandalen kring Weavering Capital, och den svenske grundaren till det brittiska fondbolaget Magnus Peterson, briserade. … Trots den turbulens som följde efter Lehman Brothers-kraschen hösten 2008 såg Weaverings fond ut att klara sig oförskämt bra. Men fondens värde misstänks ha blåsts upp av internaffärer och med hjälp av ränteswappar. Det mesta var dock luft. Sammanlagt uppges Weavering ha förvaltat omkring 5 miljarder kronor, men en stor del av de pengarna misstänks vara försvunna. … I en domstol på Caymanöarna föll i fredags en dom mot Magnus Petersons 85-årige styvfar och bror. Den senare var chef för det svenska dotterbolaget Weavering Capital, ett av dotterbolagen till det brittiska bolaget.”

”**Hedge fund boss Magnus Petersen charged with six offences** after three-and-a-half year investigation by Serious Fraud Office. … ” investigation first closed and then reopend.

Court demands 10 years of accounts from Weavering Capital boss

Joe McGrath
22 Jul 2013

Magnus Peterson – the $600 million hedge fund leader facing a criminal prosecution for alleged fraud – has 10 weeks to provide a personal asset trail to forensic investigators at liquidators Duff & Phelps.

At an unrelated hearing at the High Court last week, the founder of Weavering Capital UK was instructed by deputy chancery master Roger Bartlett to provide a list of all transactions over £2,000 for him and his wife, Amanda, dating to 2003.

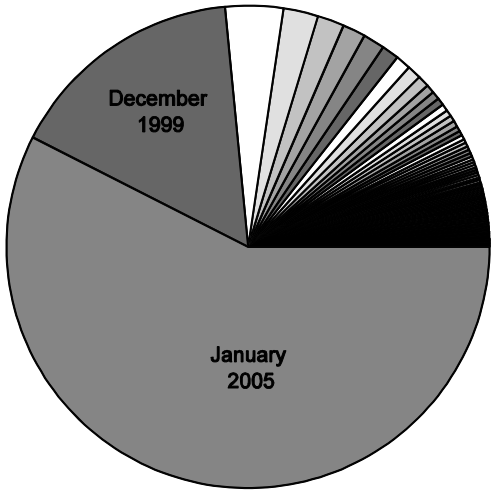
In a trial scheduled for October 2014, Peterson faces criminal charges of forgery, fraudulent trading and false accounting, brought by the Serious Fraud Office. He has not yet made a plea and has not issued any comment through a lawyer as he is representing himself.

Peterson was declared bankrupt in 2012 after the High Court handed down a $450 million judgment against him and three other senior executives at Weavering Capital UK, which was the UK-based manager for the Cayman Islands Weavering Macro Fixed Income hedge fund.

The Petersons are due to leave the family home they share with their four children at the end of next month in order to settle debts from an earlier High Court case, Amanda Peterson said in court. This is unconnected with next year's fraud case.

Financial News, Nov. 26, 2013

Before Gudrun After Gudrun



SEK 2,8 billion loss (LF)
55% of total loss 1982-2005

Windstorm Gudrun, January 2005

Analysis based on data 1982-1993: 1% chance biggest damage next 15-year period will be more than SEK 2,5 billion

???

Analysis basered on data 1982-2005: 10% chance biggest damage next 15-year period will be more than SEK 7 billion (this time we didn't miss damage to forrest – but did we miss something else?)

Pandemics: the SARS Timeline

Cristl Donnelly, Imperial College

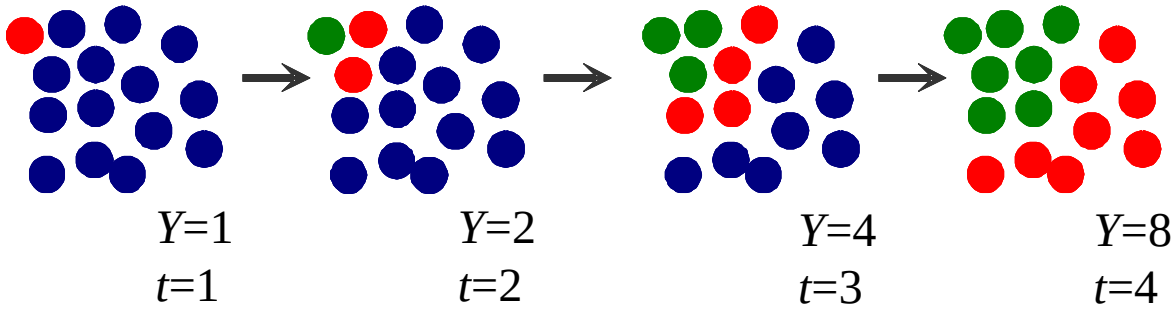
(slides adapted from one of her presentations)

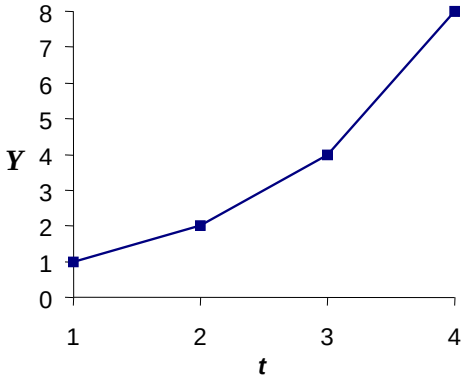
- 16 Nov 02 – a case of atypical pneumonia, Guangdong.
- 26 Feb 03 – cases of unusual pneumonia, Hanoi
- 10 Mar 03 – Dr Carlo Urbani reports an unusual outbreak of the illness he calls sudden acute respiratory syndrome (SARS) to WHO.
- 11 Mar 03 – outbreak of mysterious respiratory disease in Hong Kong.
- 12 Mar 03 – WHO issues a global alert about SARS.
- 15 Mar 03 – WHO issues a heightened global health alert after cases in Singapore and Canada.

Act now → close down (parts of) countries, stop air travel

Don't act → cause millions of deaths ("only" 775 for SARS)

- Pandemics spread through contact between individuals
- ‘Chain reaction’ gives (initial) exponential growth





R = reproduction number = # of secondary infections caused by one primary case at the start of an epidemic.

is R big, in particular >1?

The key is international coordination: better decision systems, better reporting, better (and faster!) data processing, better understanding – and better communication with the public

Total Quality Management
Continuous Improvement

(Also for the finance sector?)

Ford, model T, 1908

Toyota Prius, concept

FMEA – Failure Modes and Effects Analysis

Systematic approach to identifying and diminishing risks

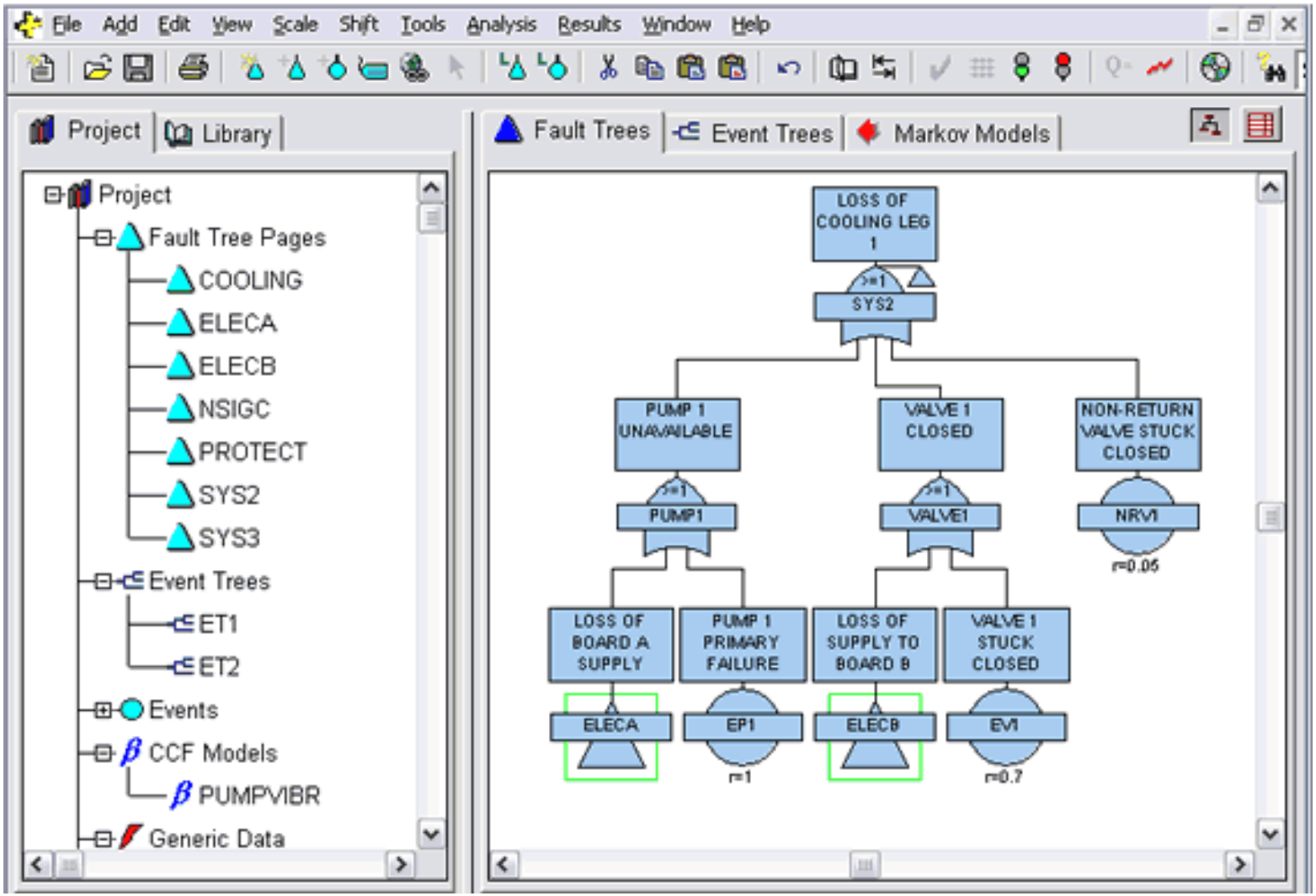
Subcomponent/ Subsystem	Component ID	Component Fail Rate	Failure Mode	Alpha	Failure Effects		Failure Detection Method	Mitigation	Severity Class	Maintenance Action
					Local	End				
Hoist Motor and Gearbox	MFPH- Hoist Motor 3 HP & Gearbox	5.5E-06	Motor fails to start/run	1.00	Cannot raise or lower Tube Plug by motor	Shutdown for repairs ~24 hours	Status light indicator	Use of handwheel for manual application	III	Replace
	BFPH - Tube Plug Hoist Brake	1.15E-05	Fails	1.00	Potential for tube plug drop (see note)	Shutdown for repairs ~24 hours. If tube plug drops then shutdown n > 1 week.	None	None	III (II if plug drops	Replace
	PHOL - Tube Plug Hoist Brake Overload Switch EB33056-16	4.12E-06	Fails Open	.98	Motor will not run	Shutdown for repairs ~8 hours	Status light indicator	Manual handwind	IV	Replace

working group with broad and diverse background – leader – data – training – plan work – brainstorm – take notes – everyone responsible – systematic judgment of likelihoods; consequences; countermeasures – someone responsible for implementing results

Systematic  *Creative*

Seriously and well thought through for many different kinds of FMEA-s

FTA – Fault Tree Analysis
PSA – Probabilistic Safety analysis



Used for

- Estimating probabilities ← ***Unreliable!***
- Identification of important risks

FMEA: Bottom up, simple systems

FTA: Top down, redundant system

Tools for continuous improvement

(but unstructured thinking about risks is also needed)

Holger Rootzén and Claudia Klüppelberg (Ambio, 1999)

A Single Number Can't Hedge Against Economic Catastrophes

Mathematics and statistics have transformed day-to-day trading in the world's financial markets. This has led to new ways to reduce (or "hedge") risks, which provide an important service to society, but also a temptation to very big gambles, with a potential for extreme losses. This paper discusses some of the ways statistics and mathematics can be used to understand and protect against very large, "catastrophic" financial risks. We argue that means don't mean anything for catastrophic risk, that separate large financial risks often are better handled by separate companies, and that the mathematical aspects of risk can't be summarized into one number. We also believe that there is a large potential for improved risk management in financial institutions, where extreme value theory, a speciality of the present authors, may be a useful tool. Improvements, however, will not come for free, but require long and hard work, where mathematics is only one part of the total effort.

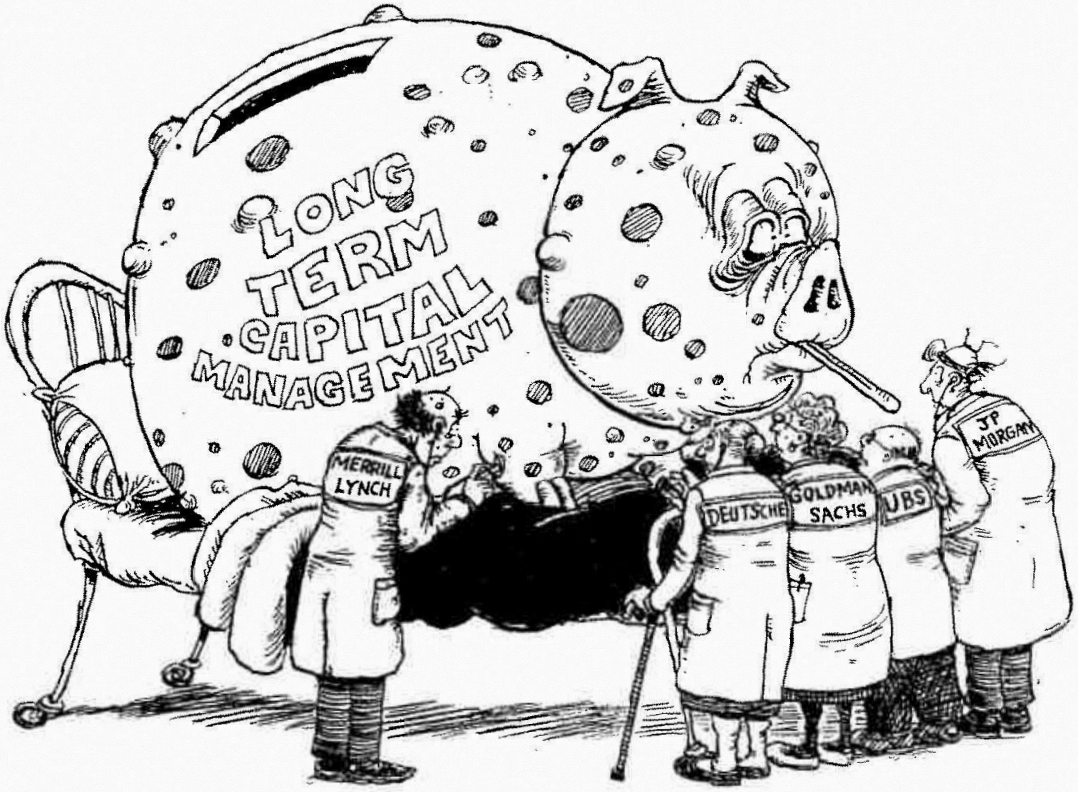
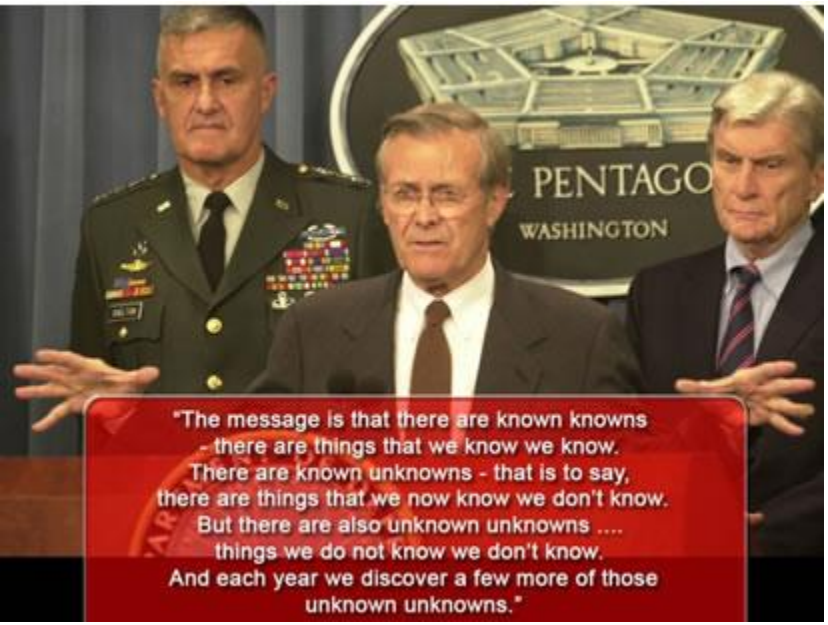


Figure 1. Long-term sickness? Reprinted from the *Economist* (3), with permission of the artist, D. Simmonds.

My own point of view

All risks must be handled as well as can be done – and if you don't handle the small risks, then the big risks become uniteresting! If possible, risk handling should be quantitative, otherwise qualitative (and, don't spend so much effort on "known unkowns" that you forget "unknown unknowns")



Donald Rumsfeld

Black swans

↓ Struktured thinking

Grey swans

↓ Statistics

White swans

Handle known risk

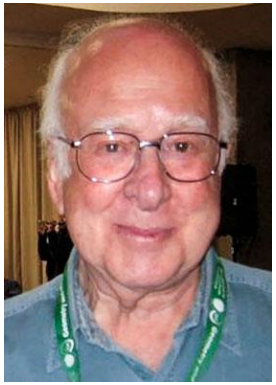
Many small changes together lead to dramatic results

Rome10.000 years ago

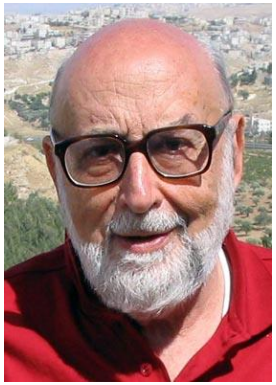
Rom now

Mankind’s total transformation of our earth has occured in very many small steps

Nobel prize in Physics 2013



Peter Higgs



Francois Englert

Which is most Important?

CERN: thousands of researchers

Climate change

slow down, stop mitigate

Extreme climate

Make dyke 1.5 m higher

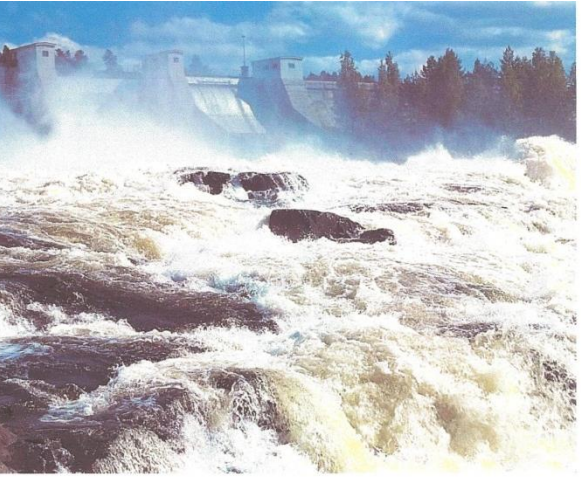
→ costs billions of Euros, popular protests

Keep dyke as is

→ (perhaps) thousands of deaths

My research this week:

Devlop new mathematics and statistics which help us to make such choices as well as we can



Riktlinjer för bestämning av dimensionerande flöden för dammanläggningar

Nyutgåva 2007

”Dammanläggningar som dimensioneras enligt flödesdimensioneringsklass II ska vid dämningsgränsen kunna framsläppa ett tillrinnande flöde med en återkomsttid av minst 100 år.”

Climate change: the standard method above doesn’t work any more, one has to specify

- Design life period (e. g. 2015-2115)
- Risk (e.g. 5% probability of a flood during this period)

→ **Design life level**

First convince engineering and political community that this has to be done, ***then*** the rest is climate modelling and (extreme value) statistics

(***But*** science and statistics is only one part, one has to develop new and much more flexible and adaptive approaches to design)

Design Life	Prob.	Design Life Level	Return Level *(2015 climate)*	EWT	EWT *(trend stopped)*
2015-2064	0.05	11.5	10.9	251	788
2015-2064	0.01	15.2	14.4	431	3839
2065-2114	0.05	12.6	10.9	262	1008
2065-2114	0.01	16.6	14.4	453	5002

Return levels are for $T = 975$ and $T = 4975$, respectively. EWT is expected waiting time

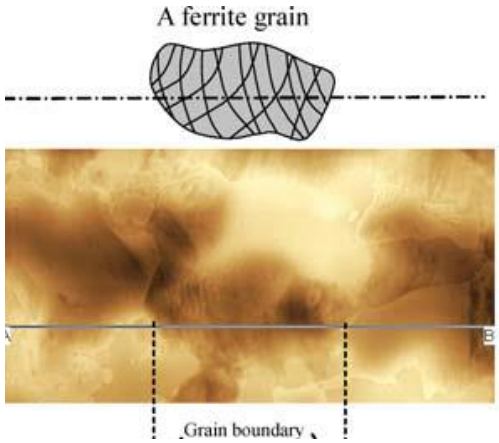
Estimation of the distribution of maximal inclusion sizes in clean steels, from measurements on planar sections

Fatigue: metal specimens subjected to varying stresses develop cracks and eventually break --- the major cause ofmechanical failure

Fatigue limit: if stresses are lower than this limit, cracks stop growing and the specimen doesn't break

Murakami's theory: fatigue limit is determined by the square root of area of the largest inclusion, measured on the projection perpendicular to the direction of stress

measurements: on polished planar sections, size of square root of area of sections of cut inclusions,



- Peaks over Thresholds measurements
- Area Maxima measurements,

choice depends mainly on experimental convenience and habit

model: inclusions are spheres, centers 3-d Poisson process in the specimen, 2-d sectional sizes (=diameters) exponential (PoT) or Gumbel (AM)

aim: distribution of maximal diameters of 3-d spheres, in bigger volumes, streamlined methods for engineering use