ICA 2018 Berlin – Data, decisions and distortions by David J. Hand

Aktuárský seminář 2. listopadu 2018

Miroslav Šimurda

ACTUARIA

Úvod

https://events.crowdcompass.com/ica2018

https://www.actuarial-center.org/data-decisions-and-distortions-decision-making-in-the-modern-world_587a65376.html

ICA CIA BERLIN 2018

Data, decisions, and distortions - decision-making in the modern world

David J. Hand
Imperial College, London

Profesor David J. Hand

About the speaker

ICA CIA BERLIN 2018

LOGO

- **Professor David J. Hand**
- Emeritus Professor of Mathematics, Imperial College, London
- David Hand is Emeritus Professor of Mathematics and Senior Research Investigator at Imperial College, London. He is a Chartered Statistician and Honorary Fellow of the Institute of Actuaries. He serves on the Board of the UK Statistics Authority and on the European Statistical Advisory Committee. He has received many awards for his work, including the Guy Medal of the Royal Statistical Society and the Box Medal from the European Network for Business and Industrial Statistics. His 29 books include *Principles of Data Mining*, *Measurement Theory and Practice*, *The Improbability Principle*, and *The Wellbeing of Nations.*

Rozhodování v moderním světě

Theory-driven example

ICA CIA BERLIN 2018

Model the relationship between the height from which a stone is dropped and the time it takes to hit the ground

→ Data $\left(H_i, t_i\right) \quad i = 1,\ldots,n$

→ Model $t = \sqrt{2H/a} + \varepsilon \quad H = a\left(t+\varepsilon\right)^2/2$

Rozhodování v moderním světě

Data-driven example

ICA CIA BERLIN 2018

Logistic regression model for the probability that someone will purchase a product

based on the values of their characteristics

$x_1, x_2, \ldots, x_n$

$$\log\left(\frac{p}{1-p}\right) = \sum_{j=1}^{d} \beta_j x_j$$

6

ACTUARIA

Zdroje dat

Increasing number of data sources and types
with different properties

- Surveys
- Panel data
- Administrative data
- Transaction data
- Web scraped data
- Social media data
- ….

Experimental/observational
Big data
Open data
etc

Problémy s daty

Some of the challenges

Bad data:

not the data you want, *but a* ***distorted version***

Invisible data:

not just the data you've got, *but also* ***the data you'd like***

Changing data:

not the data you've got, *but* ***the data you'll have***

Alternative data:

not the data you've got, *but* ***the data you would have had***

Misleading data:

not the data you've got, *but* ***the data you think you've got***

ICA CIA BERLIN 2018

ACTUARIA

30 kg nebo 30 g?

BAD DATA

Not the data you want,

but a distorted version

ICA CIA BERLIN 2018

"Britain's largest bat, the greater mouse-eared bat, which was officially declared extinct in the UK 12 years ago, has been rediscovered hibernating in an underground hole in West Sussex. They can weigh up to 30kg and have ears as long as 3 cm."

Source: The Times, December 2002

30 g nebo 300 mg?

Last year:

"Two students suffered 'life threatening reactions' when they were given enough caffeine for 300 cups of coffee.

... spent several days in ICU...dialysis...

Should have been given 0.3g of caffeine. Instead they were given 30g."

The Times, 26 January 2017

ICA CIA BERLIN 2018

The ***Mars Climate Orbiter***

Launched 1998, but communication lost on September 1999 when the spacecraft trajectory brought it too close to Mars

... because one of the software teams forgot to convert Imperial units to SI units

ACTUARIA

Drahé chyby...

ICA CIA BERLIN 2018

"Poor data quality costs the US economy around $3.1 trillion per year"

Source: IBM

- Lidské chyby – 1 akcie za 670 000 jenů nebo 670 000 akcií za jeden jen?
- Moc lékařů narozených 11.11.1911. Proč?
- Podvody...

ACTUARIA

Automaticky generovaná data...

Berry and Linoff (2000) example:

"The data is clean because it is automatically generated – no human ever touches it"

But it turned out that 20% of transactions had

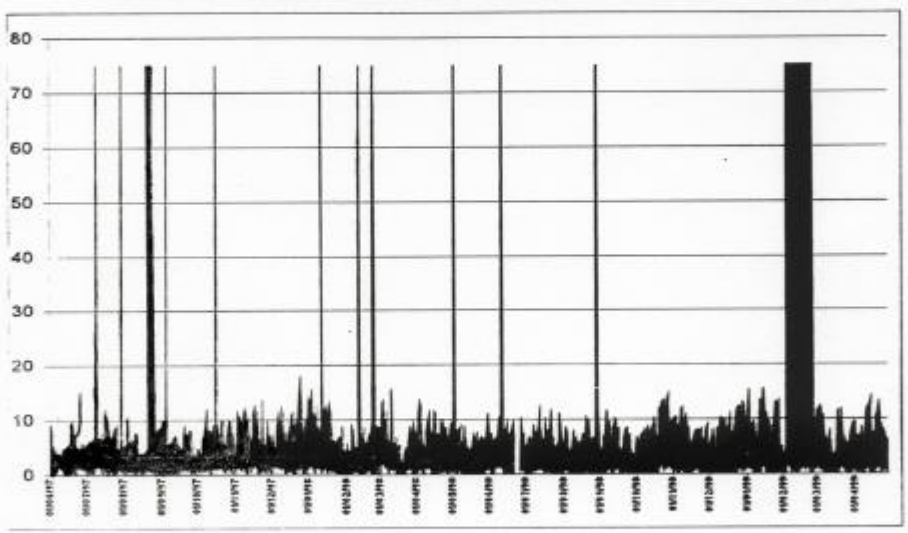
"arrived before they were sent ....
not only did people never touch the data, but they didn't set the clocks on the computers either"

ICA CIA BERLIN 2018

ACTUARIA

...pozor na ně

Not merely human error



Source: Dave Yearling

Smutná pravda

Maintain a healthy scepticism

Twyman's Law:
Any figure that looks interesting or different is usually wrong

Příležitostí k chybám je dost...

Other aspects of bad data:

relevance,
timeliness,
consistency,
coherence,
availability,
and accessibility

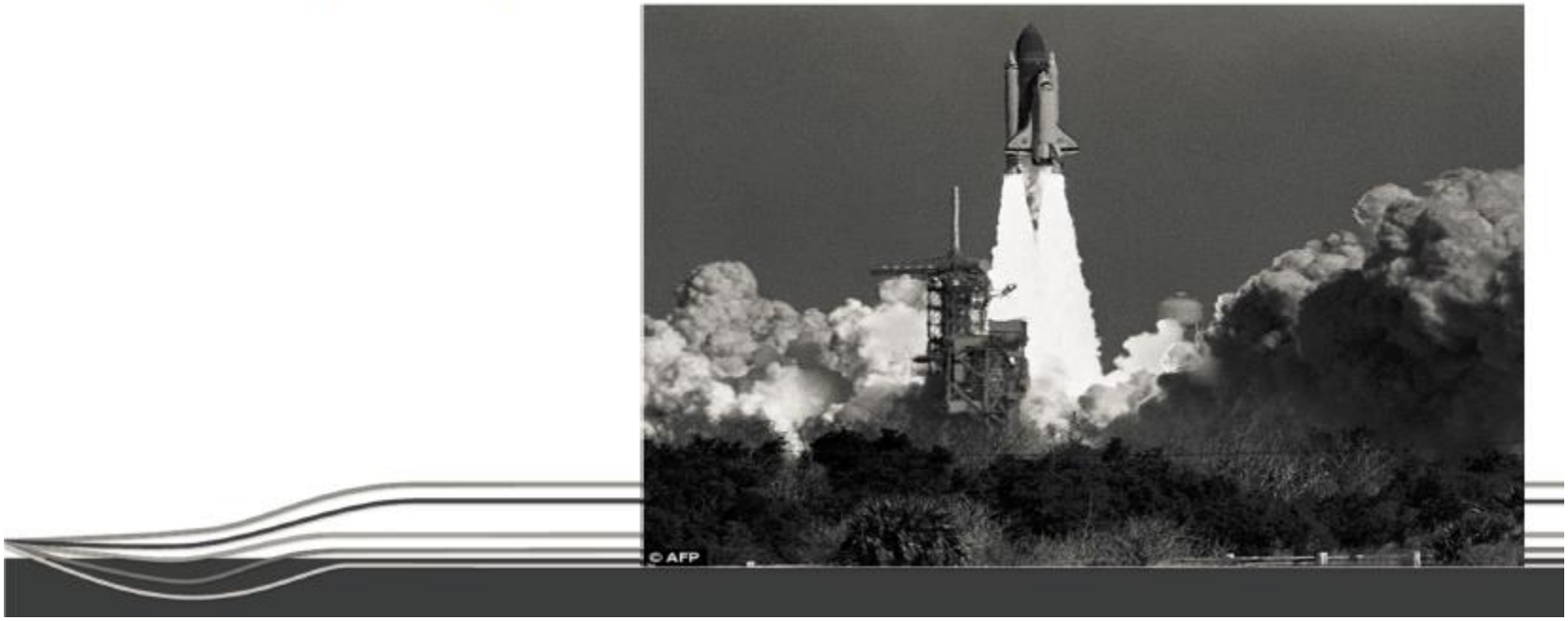
Těžko uvěřitelné...

INVISIBLE DATA

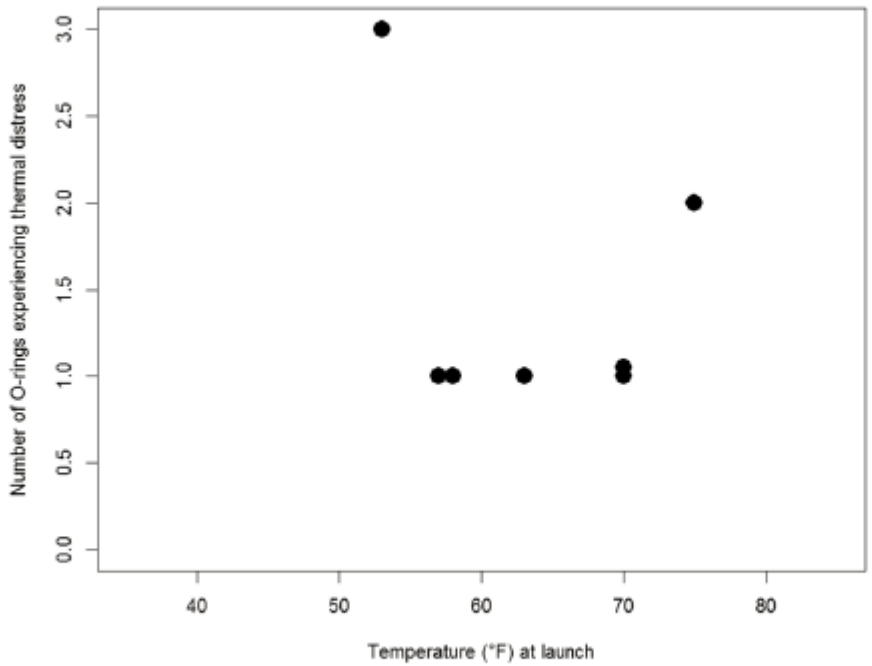
Not just the data you've got,

but also the data you'd like

ICA CIA BERLIN 2018

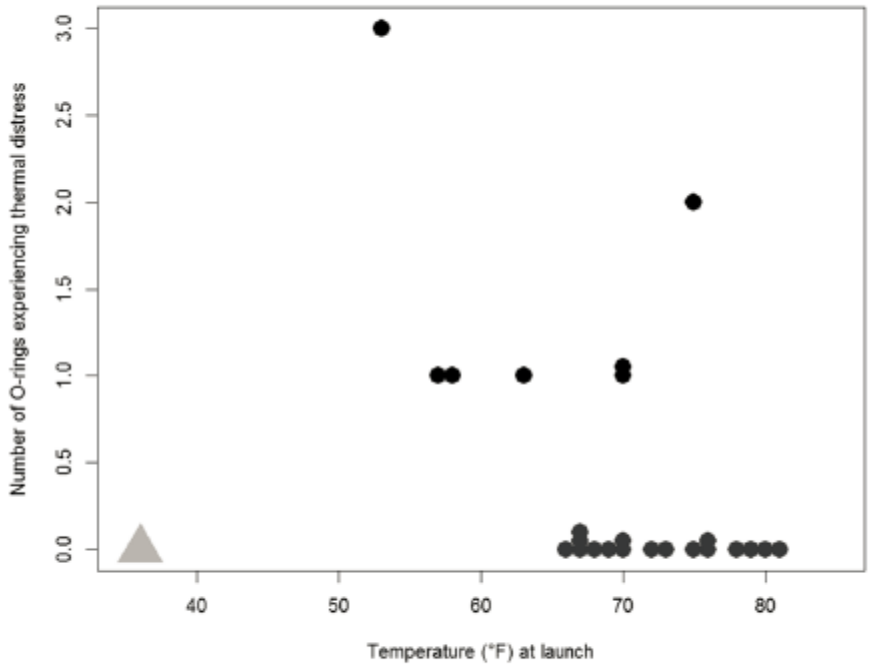


Potíže s těsnícími kroužky – nezávisí na teplotě...



ICA CIA BERLIN 2018

...Opravdu ne?

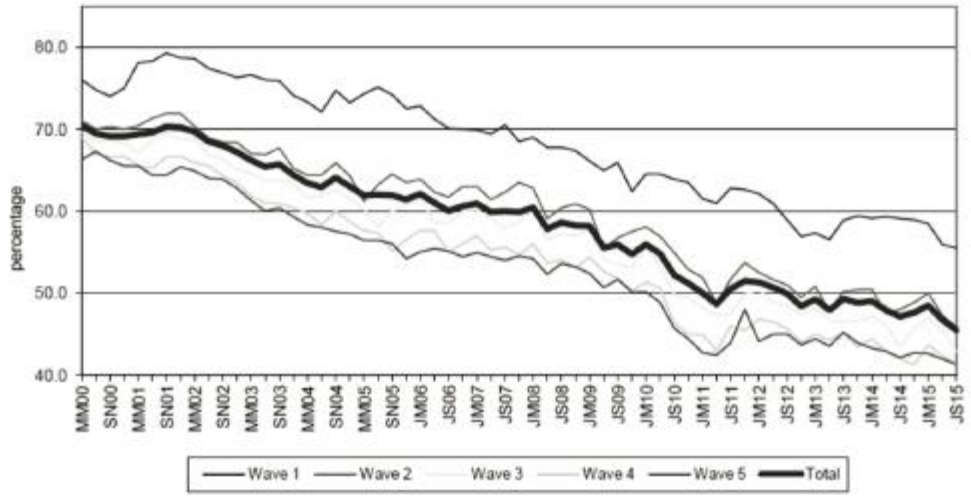


ICA CIA BERLIN 2018

ACTUARIA

Moc průzkumů

Non-response and refusals



LFS quarterly survey wave-specific response rates: March-May 2000 to July-Sept 2015

Source: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-force-survey/index.html

Otázky a co plyne z odpovědí...

The magazine survey which asks one question:

Do you reply to magazine surveys?

And discovers that apparently *all* the readers reply to surveys

The Actuary, July 2006, editorial:

"A couple of months ago I invited all 16,245 of you to participate in our online survey concerning the sex of actuarial offspring."

" ... Well, I'm pleased to say that a number of you (13, in fact) replied to our poll."

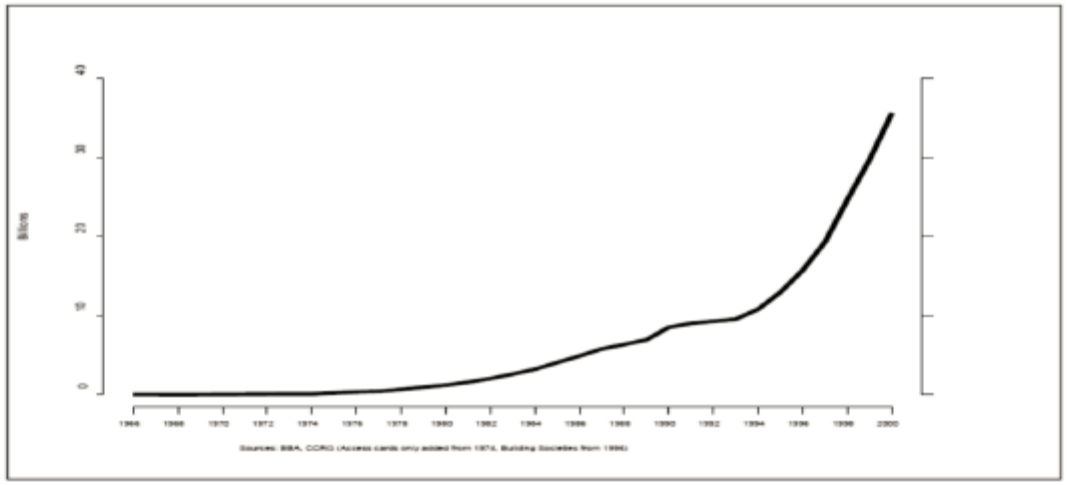
Síla trendů

CHANGING DATA

Not the data you've got,

but the data you'll have

Non-stationarity



ICA CIA BERLIN 2018

Alternativní fakta...

ALTERNATIVE DATA

Not the data you've got,

but the data you would have had ***Counterfactuals***

ICA CIA BERLIN 2018

Credit card transaction fraud detection

- compare existing detector with proposed new one
- stop transaction when ***existing*** detector says its suspicious
- not when proposed (untried, untested) ***new*** one says so

→ data asymmetry

→ comparison artificially favours one method

ACTUARIA

O co doopravdy jde...

MISLEADING DATA

Not the data you collect,

but the data you think you've got

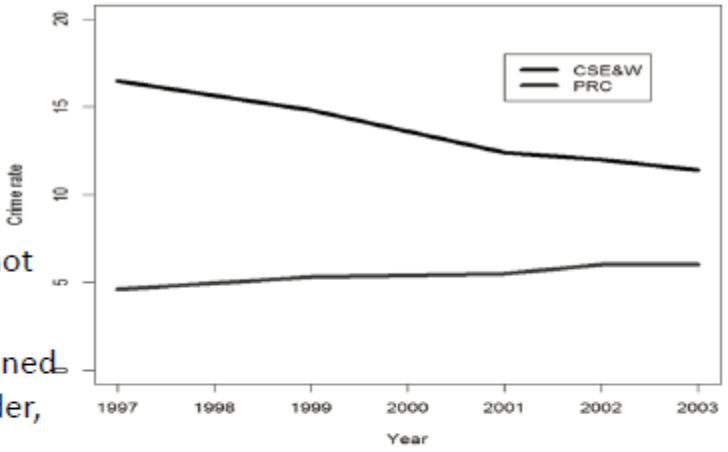
ICA CIA BERLIN 2018

Answering wrong question

Crime rates, 1997-2003

Crime Survey England and Wales

Police Recorded Crime



CSE&W: ...not group residences; not crimes against commercial or public sector bodies; victim-based (not include murder); capping repeat victimisation ...

PRC: reported to and recorded by police; crime defined by "Notifiable Offence List" (incl. murder, public order, ...); incl. residents of institutions and tourists; incl. commercial bodies ...

Co dělat...

So what should we do?

ICA CIA BERLIN 2018

Detection

Prevention

Correction

ACTUARIA

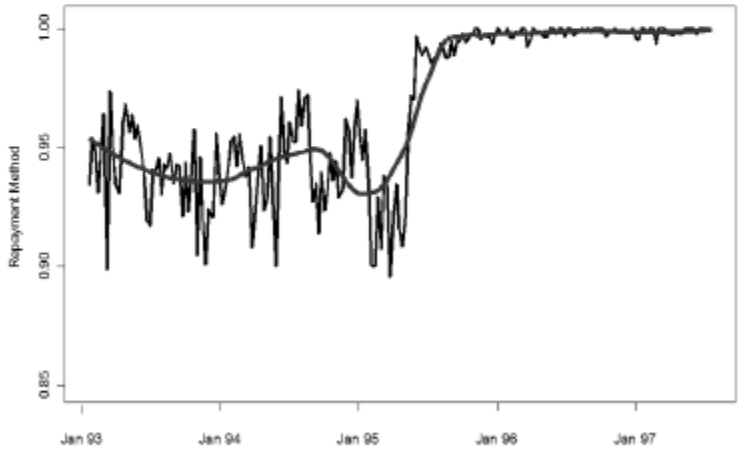
Detekce

Detection

ICA CIA BERLIN 2018

Does the data conform to what you expect?
Are different data sources consistent?

- outliers
- expected distributions
- triangulation
- change points



ACTUARIA

Prevence

Prevention

Consistency checks on data entry
- logical, rule-based

Careful design of data collection systems

Why were so many doctors born on 11th November 1911 ?

Korekce

Correction

Imputation

Sophisticated statistical adjustment

Rule-based consistency

But cannot perform miracles !

Jak dál...

So where does all this leave us?
Some recommendations

ICA CIA BERLIN 2018

1) Consider the origin of the data: don't take at face value

"We don't have any fraud at my bank"

Said to me by a banker at a conference

2) Sense check

The 1970 US Census showed that 289 boys had been both widowed and divorced by the age of 14

3) Statistics and data science MScs need a ***module*** on data quality

Can't teach data quality issues while teaching methods and ideas

Různé pohledy na věc…

"With enough data, the numbers speak for themselves"

Chris Anderson in *Wired* magazine, 2008

"the most reckless and treacherous of all theorists is he who professes to let facts and figures speak for themselves, ..."

Alfred Marshall, Inaugural Lecture to Chair in Political Economy, Cambridge, 1885

...a jejich spojení.

ICA CIA
BERLIN 2018

If the data can speak for themselves

they can also lie for themselves

David Hand

ACTUARIA