WHEN REASONABLE METRICS DESTROY PRODUCTS, PEOPLE AND COMPANIES

Goodhart's law and how to fight it

Ondrej Tulach

Metrics guide our firms and lives

- Our efforts and outcomes measured

- Our goals set

- Yet unforseen consequences loom

- Try to find the common pattern

BROADCOM®

The Czech Railways case

- Czech Railways have a notorious problem with train delays

- So management had their bonuses tied to % of trains adhering to the timetable

The railway paradox. Trains between Prague and Brno will run longer than before the 10 billion crown modernization

15.06.2023 5:00

Journey times between the two biggest cities will increase by up to 11 minutes compared to pre-diversion times

- If the timetable allocates extra time for each station, risk of delay decreases
  - Increased travel time not captured by the metric

BROADCOM

# The USAF case (Government Accountability Office report)

- The number of long-term grounded aircraft as an indicator of the success of their maintenance efforts ...

- Maintenance personnel ... try to have as few aircraft as possible designated in this status"

- „KC-135 unit maintenance personnel we spoke with said they actively rotate parts among their aircraft in an effort to prevent any of them from being designated as a long-term grounded aircraft"

BROADCOM

The Indian snake bounty case (fable)

- During the British rule of India, venomous cobra snakes were a problem
- The british authorities offered snake hunters a bounty on killed cobras turned in
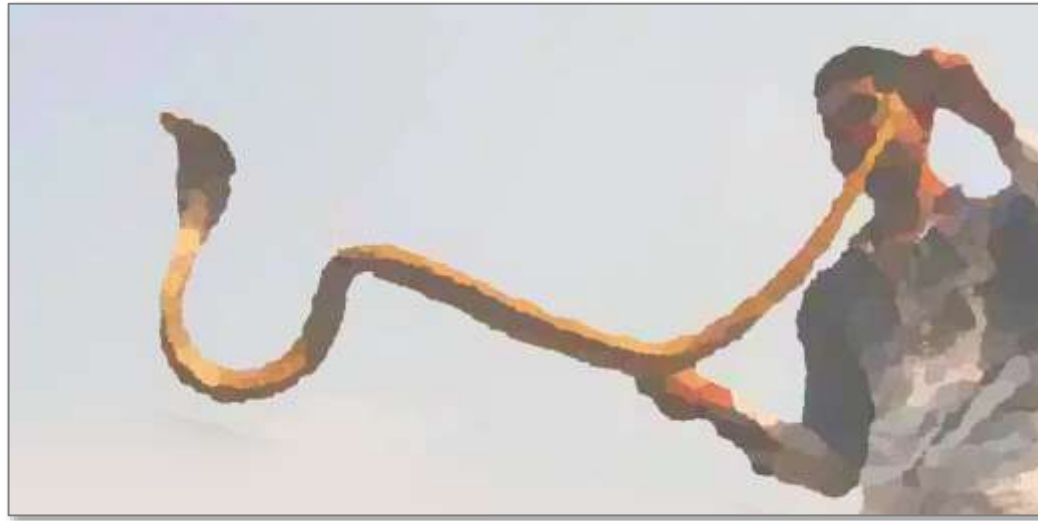
- Supposedly, local hunters wanted to turn this bounty sustainable...
- So they started *farming* cobras to be „hunted" and turned-in for bounty
  - The bounty caused *more* cobras when it should have caused *less*

BROADCOM

The case of USNW Universities ranking system

- 1983: struggling magazine US. News & World report searching for a hit

- Compiled „ranking" of 1800 Us. colleges and universities
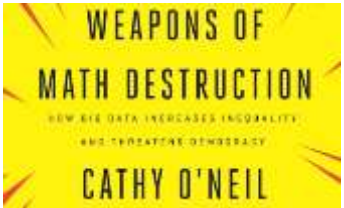
„No direct way to quantify how a 4-year process affects individual students ...
Couldn't measure learning, happiness, usefulness ...
Instead picked proxies which seemed to correlate with success"

- Eventually compiled 15 metrics
  – Based on role-model of the most elite universities
  – Wuth tuition/price NOT being part of the model

- All universities *had* to improve the *same* 15 metrics *at the expense* of all else
  – Rat race: eliminated diversity, „safety colleges" going extinct

- Cost of education rose by 500% (4× faster than inflation)

- Students became *commodity*, a means to reach the true goal (Us. News ranking)
  – Student admission rate, SAT scores, graduation rate, athletic performance etc.: liability or asset

BROADCOM

The Software development metrics incentives case

- Code Coverage vs. Test Class Coverage?
 - Deep, functionally limited coverage Vs. broad, possibly superficial?
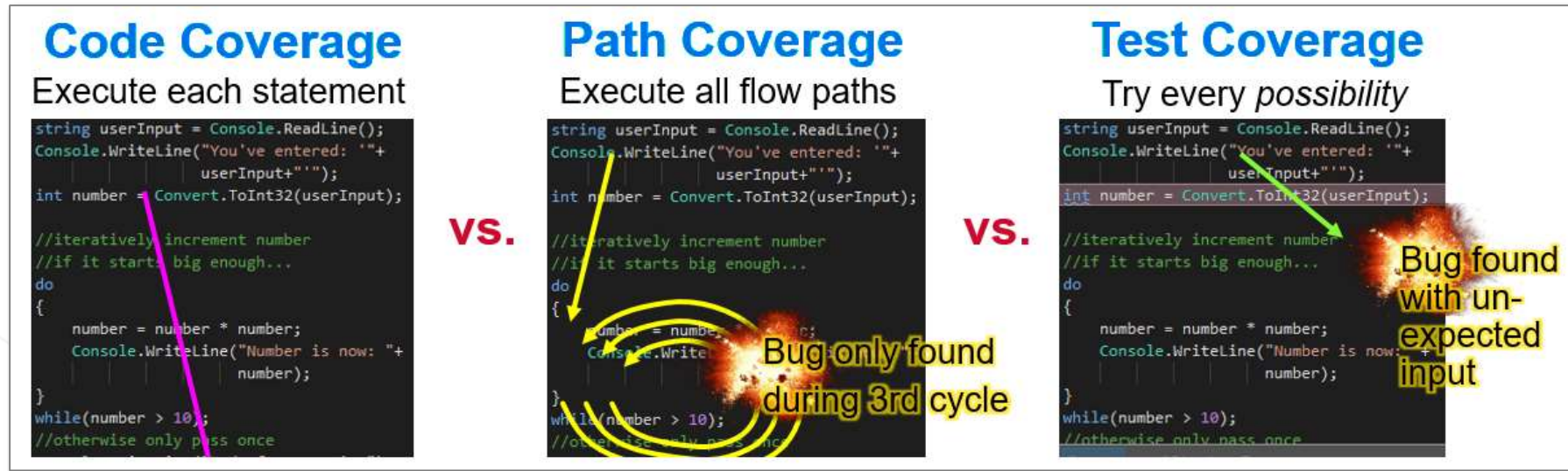
- Test automation vs. Number of tests?
  - The less tests or the more clone-a-like, the more automated
- Number of defects found or Zero-defect policy?

The Boeing 737 MAX case 1/3 – The bad engineering choice forced

- Boeing didn't want to make 737MAX - Jan 2011, Boeing CEO Jim Albaugh:
  „I don't think we will re-engine the 737. Right answer probably a new small plane"

- But Jul 2011: American Airlines will buy *500* airliners all to Airbus...
  - *Unless* Boeing swiftly re-engines 737 <u>as</u> „Common Pilot Rating"
  - Common Pilot Rating with older 737 = <u>*without*</u> difference pilot training, simulators, certification
  - „Training on moving from the old 737 NG to the new 737 MAX consisted of session on an iPad"

Controlling „metric"

- But re-engined 737MAX had different aerodynamics
  - → MCAS system to „hide" the differences, using stabilizer trim

- But MCAS power was found insufficient
  - → undocumented change from hard-limit 0.6° to *incremental* 2.5°
  - → > 100 lbs of force pushing stick & plane nose down

- But CPR requirement – no simulator sessions
  - →All-powerful MCAS not documented, hidden from the pilots

- 6,767 orders on books... 346 killed in Lion Air 610, Ethiopian 302 crashes

BROADCOM
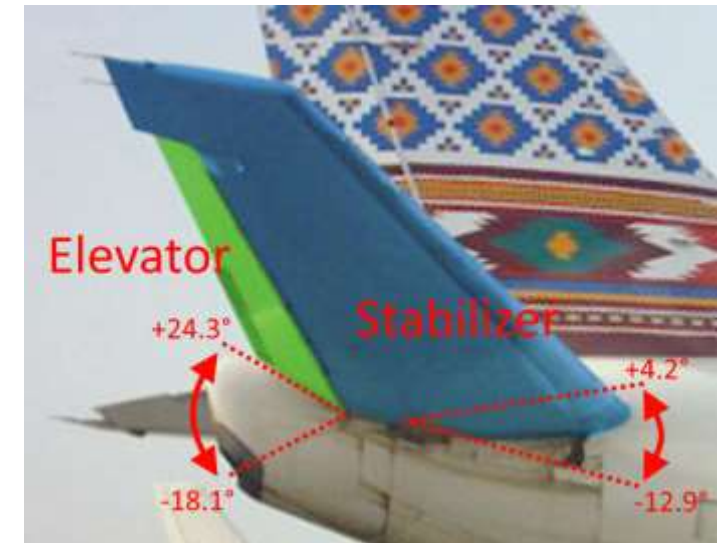
There is something fishy with the metrics

- There seems to be a systemic problem with metrics-driven decisions!

  - Nobody wanted to increase the travel time of trains
  - Nobody wanted to ruin insofar perfectly good airplanes to cannibalize them for broken ones
  - Nobody wanted to kill inadequately-trained pilots with well-hidden all-powerfull MCAS system
  - Nobody wanted to grow *more* venomous snakes
  - Nobody wanted to „game" the software/test/defect metrics
  - Nobody wanted to demote students to commodity, skyrocket costs, homogenize education

- Yet the exact opposite has happened...

- Again and again and again

BROADCOM

The Rule of All Metrics

- Charles Goodhart (statistician):

  "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

- Jón Danielsson (economist):

  "Any statistical relationship will break down when used for policy purposes."

- Marylin Strather (anthropologist)

  "When a measure becomes a target, it ceases to be a good measure."

BROADCOM

The problem with metrics from Systems perspective

- Metrics are 1) a model, and 2) a feedback-control system
  - Attempt to model complex organization, interactions, skills, effort
  - As a „black box" represented by single/few numeric proxies
  - Supposedly tightly correlated with the outcome
  - Driving behavioral change through reward/punishment
  - To achieve single (or few) supposedly correlating priority goals

- Problem with models:

  "All models are wrong. Some are useful." (George Box, statistician)

- Too simple model ➡ important elements of reality ignored

- Metric-as-a-model ➡ by prioritizing one element, all others are neglected

- Reward/Punishment behavioral change ➡ forcing workers to sacrifice the non-measured

  "Any observed statistical regularity will tend to collapse
  once pressure is placed upon it for control purposes."

BROADCOM

It's not about cheating - you get *exactly* what you ask for

- Predominant feeling: selfish, immoral employees want to cheat the metrics

- Reality: some employees know better, but have no autonomy
  - USAF maintainers knew ruining good aircraft is silly, but had no choice
    - If they made the right choice, they would be punished as „low performers"
  - Boeing R&D knew that secret hidden all-powerfull MCAS was bad, but had no choice
    - If they made the right choice, BA would lose the big AA contract and employees would be punished
  - Universities who cared about individual students would lose financing and applications
    - Because they would be punished by poor ranking if they didn't play the Rankings game along

- Reality: some employees only defend against unatainable goals
  - What can Czech Railways management do to fix delays?
    - Delays caused by passengers, rolling stock, international delays, accidents...
  - SW engineers constrained by short release cycles, small teams, sustainment & infra work
    - Never enough resources to do everything properly → something is prioritized and something let go

- Reality: some employees don't understand the true goal (intent)
  - In the India fable, what do the hunters know? Maybe the British are harvesting cobra skins!

BROADCOM

Obvious solution doesn't work

- The usual approach:
  - If a metric failed... Let's find a better metric!
  - If a metric was gamed... Let's hide it from the workers!

- Yet systematically:
  - The „better" metric is just a different piece of complex reality, *still* ignoring the rest
  - The „hidden metric" *still* drives behavioral change (that's the purpose!) to the detriment of *non-measured reality*

- A different approach is needed

BROADCOM

Solution is a 3-step game

1. Never measure an isolated metric – capture the trade-offs and balances
2. Identify true goals – not the misleading proxies
3. Always consult the Subject-matter experts

Balancing trade-offs instead of isolated numbers

- If any single metric
  - Ignores all other important aspects of reality
  - Prioritizes one piece of reality at the expense of all others
  - Incentivizes people to do useless, contraproductive things to satisfy the metric

- Instead, we need to model based on the **engineering trade-offs**
  - Where different aspects of reality/design/processes compete with each other
  - Where a balance is found between the various aspects

- Solution: Spider/Radar-graphs selected to capture the trade-offs balancing

Czech Railways

USAF maintenance

Software quality
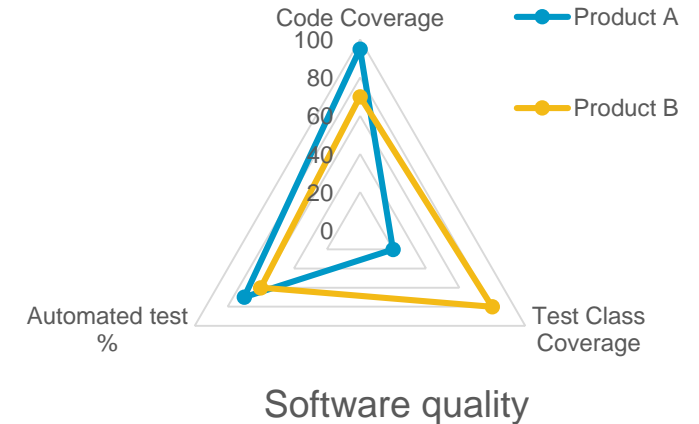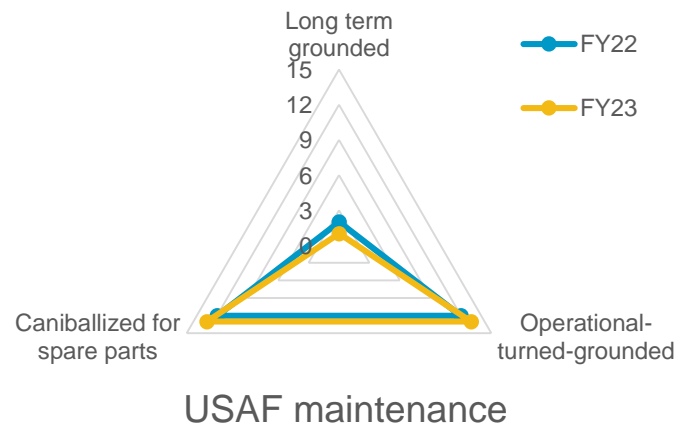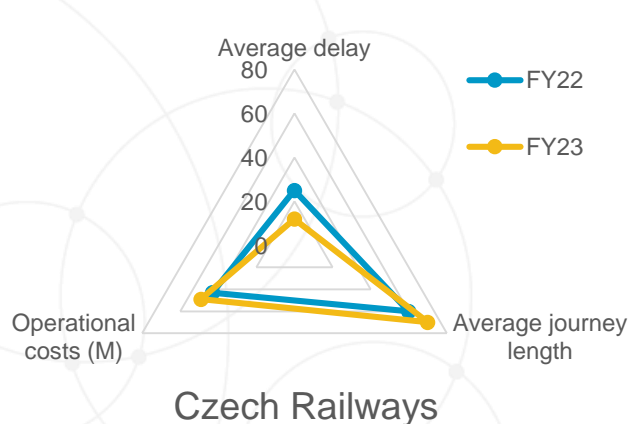
Identify out the *true* goals instead of mid-way proxies

- The real goal was not to „decrease train delays"
  - It was to decrease overall train travel length, and increase predictability of the journey

- The real goal was not to have aircraft broken for less overall time
  - It was to have as many aircraft airworthy as possible, and not to neglect the broken ones

- The real goal was not to hide an all-powerful control system from the pilots
  - It was to avoid the need of official & costly pilot re-training

- The real goal was not to hand over as many dead snakes as possible
  - It was to decrease the overall number of snakes per square mile

- The real goal was not any arbitrary number related to code
  - It was to assure thorough-enough testing of all important code, without coverage gaps

- The real goal was not to capture which school has best athletes and scores
  - It was to find which schools care about the students, help them, give them what they need

BROADCOM

Use the structured Subject Matter Expertise

- If a metric
  - Ignores important elements of reality
  - Depends on factors beyond control of the „metered" individual or organization
  - Doesn't capture competing trade-offs

- Such metric will *always* be the harmful type

- Use the SMEs who *do* understand!
  - The „ground level" SMEs know perfectly what influences the deliverables
  - They know the obstacles, dead ends, backfires, unintended consequences *by experience*
  - They can tell where the Paret Optimum would yield biggest improvement

- Craft the metrics based on expert knowledge of your SMEs
  - Consult competing SME fields to capture the trade-offs and balance
  - Listen carefully to the known limitations and obstacles – no „stop making excuses"

- Never blindly force the buzzword / „industry trend" metrics top-down
  - = an admission of not understanding, ignoring the specifics and reality of your teams

BROADCOM

This is why your incentives fall short

- RSA Animate 2010: **Drive: The surprising truth about what motivates us**

- https://www.youtube.com/watch?v=u6XAPnuFjJc

- But why?

- SME's not allowed to do their job and instead *forced-by-metric* to do poor choices

BROADCOM

True goals + power to SMEs spawn alternative working solutions

- Instead of impossible task to affect foreign countries train delays...
  - Improve delay notification and trip-planning system

- Instead of impossible task to fix more KC-135s than there are spare parts for...
  - Optimize the spares to achieve top realistic sustainable airworthiness

- Instead of the dangerous task to hide MCAS from airline pilots...
  - Find a way to notify the pilots about MCAS *while* avoiding official retraining
  - Implement MCAS in different way – eg. Steering force-feedback, or reliable cross-sensor SW

- Instead of forcing developers to focus on one-size-fits-all metrics...
  - Ask them what are the biggest long-term issues with the product quality
  - Ask them what specific steps could be taken to improve product quality = **true *Kaizen***
  - Allow periodical „free time" for each SW engineer to address things which matter most

- Matter of trust: still incentivizes gaming/cheating for some
  - but *most* employees *are **honest*** and want to improve their product. Mastery & Purpose...

BROADCOM

Summary

- „When a measure becomes a target, it ceases to be a good measure."

- All metrics are a simplified model of a complex reality based on proxies

- "All models are wrong. Some are useful."

- Whichever part of complex reality is not covered by metrics is de-incentivized

- Whichever part of complex reality is covered by metrics gets prioritized

- Metrics always have the Unintended Consequences of deforming human behavior

- People do silly things not to „cheat" and „game" the metrics...

- But because metrics by-design drive behavioral change...

- Not towards the real goal, but towards the measured proxy-value!

- The only defense is to recognize the complex reality, the trade-offs, balances...

- And to dis-incentivize satisfying one metric at the expense of everything else

BROADCOM

Good bye

- Left Broadcom December 31, 2023 to pursue System Safety Engineering
- This was my post-exit parting gift ☺

BROADCOM