Bogumił Kamiński

Knowledge

- Economy
- Management
- Finance

Methods

- Optimization
- Simulation
- Data Analysis

Domains

- Economy
- Management
- Finance

Skills

Communications

- Presentation
- Reports
- Implementation

Tools

- GNU R, Python, Julia
- SQL, NoSQL
- Excel
Types of decision situations

- Individual
  - Deterministic
  - Uncertainty
  - Risk

- Strategic interactions
Evidence based management

„Man won’t fly for a thousand years.”

*Wilbur Wright, to brother Orville after a disappointing flying experiment, 1901*

„The war in Vietnam is going well and will succeed.”

*Robert McNamara, U.S. Secretary of Defence, 1963*

„Everything that can be invented has been invented.”

*C. H. Duell, Patent Office Director, urging President McKinley to abolish the office, 1899*

„I cannot conceive of anything more ridiculous, more absurd, and more affrontive to sober judgment that the cry that we are profiting by the acquisition of New Mexico and California.”

*U.S. Senator Daniel Webster, 1848*

„I cannot imagine any condition which could cause this ship to founder. I cannot conceive of any vital disaster happening to the vessel. Modern shipbuilding has gone beyond that.”

*E.J. Smith, captain of the Titanic, 1912*

„I think there is a world market for about five computers.”

*Thomas J. Watson, IBM, 1958*
Evidence based management

As you can see, by late next month you’ll have over four dozen husbands. Better get a bulk rate on wedding cake.

NUMBER OF HUSBANDS

0

YESTERDAY

TODAY
Evidence based management

„Man won’t fly for a thousand years.”

Wilbur Wright, to brother Orville after a disappointing flying experiment, 1901

„The war in Vietnam is going well and will succeed.”

Robert McNamara, U.S. Secretary of Defence, 1963

„Everything that can be invented has been invented.”

C. H. Duell, Patent Office Director, urging President McKinley to abolish the office, 1899

„I cannot conceive of anything more ridiculous, more absurd, and more affrontive to sober judgment that the cry that we are profiting by the acquisition of New Mexico and California.”

Daniel Webster, 1848

„I cannot imagine any condition which could cause this ship to founder. I cannot conceive of any vital disaster happening to the vessel. Modern shipbuilding has gone beyond that.”

E.J. Smith, captain of the Titanic, 1912

„I think there is a world market for about five computers.”

Thomas J. Watson, IBM, 1958

Assumption: near future is similar to recent past
CRISP-DM Process
Uncertainty of $R^2$
Uncertainty of $R^2$ (code, R)

```r
sizes <- seq(from = 10, to = 200, by = 10); reps <- 10000
sim.r.squared <- function(n) {
  x <- rnorm(n); y <- 1 + x + rnorm(n); model <- lm(y ~ x)
  return(summary(model)$r.squared)
}
r.squared.q95 <- r.squared.q5 <- r.squared.mean <- numeric(length(sizes))
system.time(for (i in 1:length(sizes)) {
  result <- replicate(reps, sim.r.squared(sizes[i]))
  r.squared.mean[i] <- mean(result)
  r.squared.q5[i] <- quantile(result, 0.05)
  r.squared.q95[i] <- quantile(result, 0.95)
}) # ~2 minutes
plot(sizes, r.squared.mean,
     ylim=c(min(r.squared.q5), max(r.squared.q95)),
     xlab="sample size", ylab=expression(R^2))
lines(sizes, r.squared.q5); lines(sizes, r.squared.q95)
```
using GLM, PyPlot, DataFrames, Statistics
sizes, reps = 10:10:200, 10000
function sim_r²(n)
    x = randn(n);  y = [1 + randn() + v for v in x]
    r²(lm(@formula(y ~ x), DataFrame(x=x, y=y)))
end
r²_q95, r²_q5, r²_mean = Float64[], Float64[], Float64[]
@time for s in sizes
    result = [sim_r²(s) for _ in 1:reps]
    push!(r²_mean, mean(result))
    push!(r²_q5, quantile(result, 0.05))
    push!(r²_q95, quantile(result, 0.95))
end # time: ~20 seconds
scatter(sizes, r²_mean)
plot(sizes, r²_q5, color="black"); plot(sizes, r²_q95, color="black")
xlabel("sample size"); ylabel("R²")
Uncertainty of $R^2$ (code, Julia 2)

```julia
using Random

function sim_r²_k(n, k, reps)
    x, y, result = ones(n, k+1), zeros(n), Float64[]
    for _ in 1:reps
        randn!(view(x, :, 1:k)); sum!(y, x)
        for i in 1:n
            y[i] += randn()
        end
        push!(result, r²(lm(x, y)))
    end
    result
end

r²_q95, r²_q5, r²_mean = Float64[], Float64[], Float64[]

@time for s in sizes
    result = sim_r²_k(s, 1, reps)
    push!(r²_mean, mean(result))
    push!(r²_q5, quantile(result, 0.05)); push!(r²_q95, quantile(result, 0.95))
end # time: ~2 seconds
```
Uncertainty of $R^2$ – nine variables
An example of forecasting model

- **Features:**
  - X1 do X5, random
- **Target (formula unknown for the analyst):**
  - Y=X1+X2+1+ε

- **Simulated process:**
  - Historical data collection
  - Model building
  - Application of the model for new data
  - Collection of target variable for new data and quality assessment
An example of forecasting model

• Features:
  – X1 do X5, random

• Target (formula unknown for the analyst)
  – Y=X1+X2+1+ε

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE historical</th>
<th>MSE forecasted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables X1-X5</td>
<td>1.39</td>
<td>1.45</td>
</tr>
<tr>
<td>Variables X1-X2</td>
<td>1.41</td>
<td>1.42</td>
</tr>
</tbody>
</table>
Example examination questions

1. What are the conditions of application of statistical models for forecasting?
2. Describe stages of CRISP-DM process.
3. Explain why there is a feedback loop between “business understanding” and “data understanding” stages in CRISP-DM process.
4. When comparing two models simpler and more complex one. Which one will have a bigger error on historical data?
5. Does for a given model forecast error grow in comparison to training error and why?
6. Why avoiding of memory allocation speeds up execution of software?
7. Why are compiled languages typically faster than interpreted languages?
8. What is UNICODE standard?