# Project evolution

## From university to commerce

Agnieszka Suchwałko, QuantUp, Bioavlee

Why R 2018, 2.07.2018, Wrocław

QUANTUP

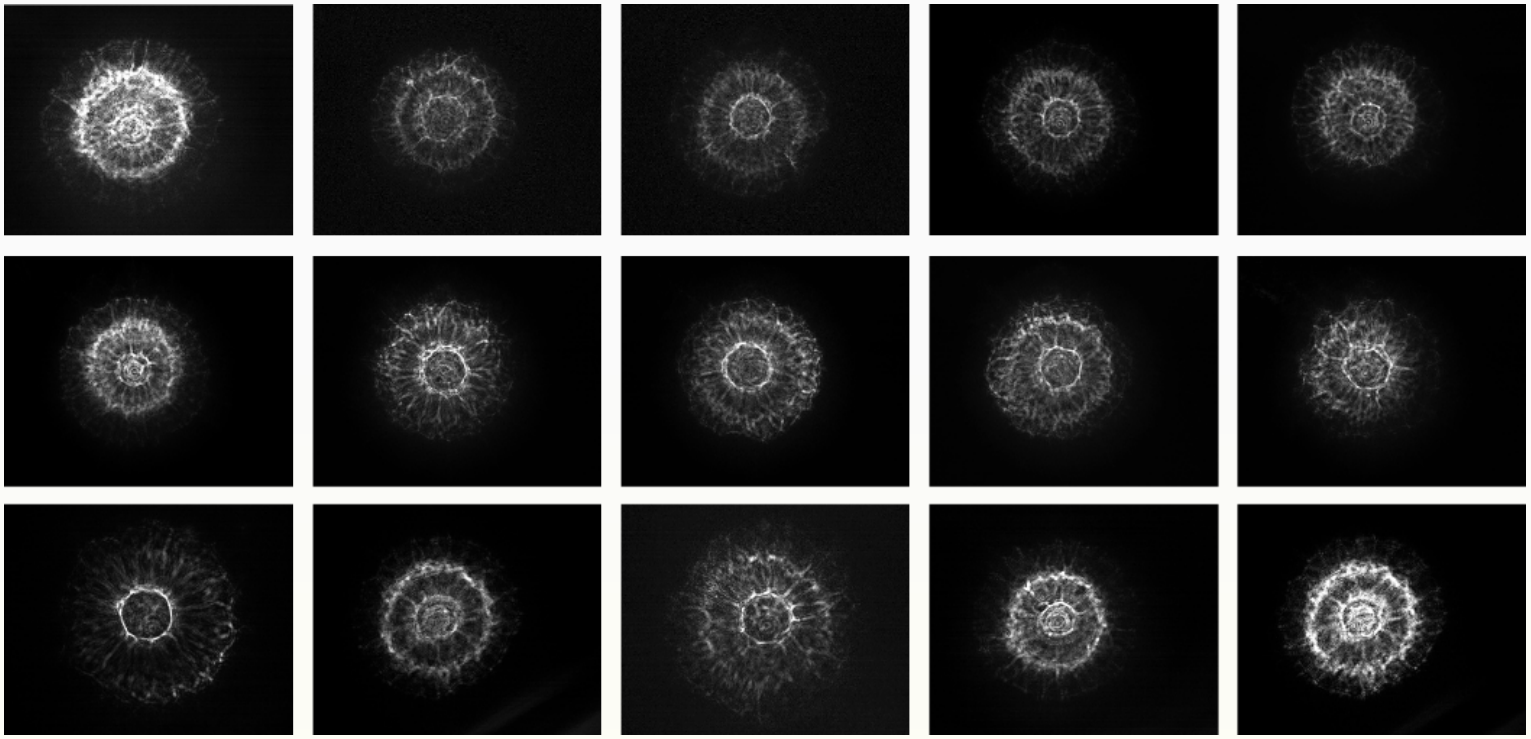

When I was a little girl
I wanted to change the world


When I grew up,
I met people with whom
I could make dreams come true

# Today I will show you my way

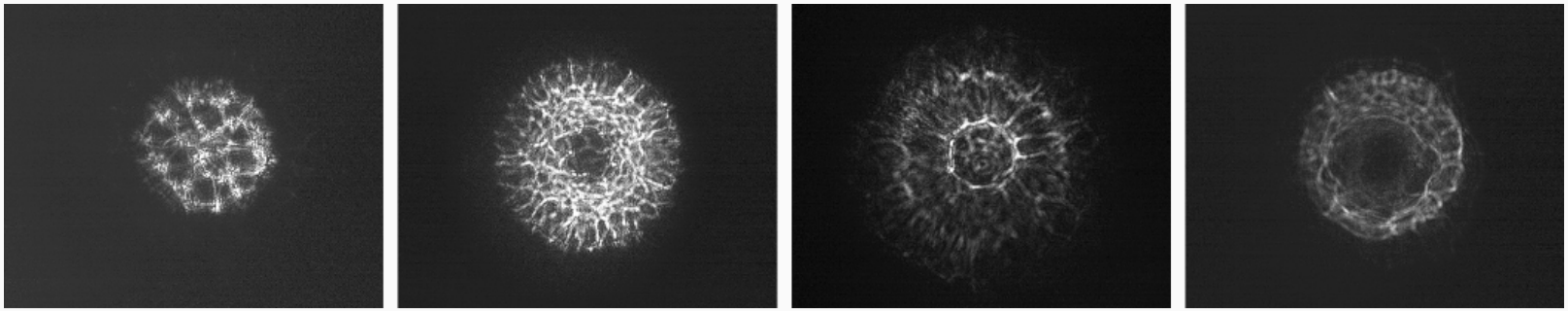
**One day Igor**
**showed me some pictures**

# Diversity within bacteria species

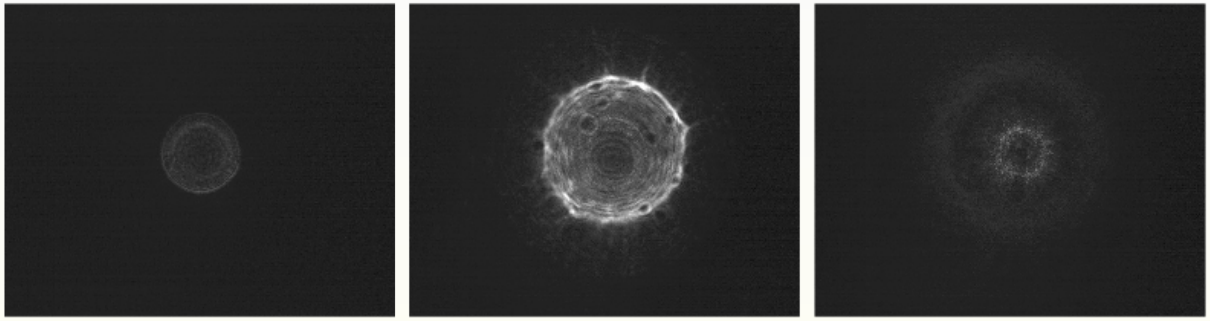


QUANTUP

# Diversity between bacteria species



*Pseudomonas aeruginosa* (ATCC 27853) | *Citrobacter freundii* (PCM 531) | *Escherichia coli* (PCM O119) | *Proteus mirabilis* (PCM 547)



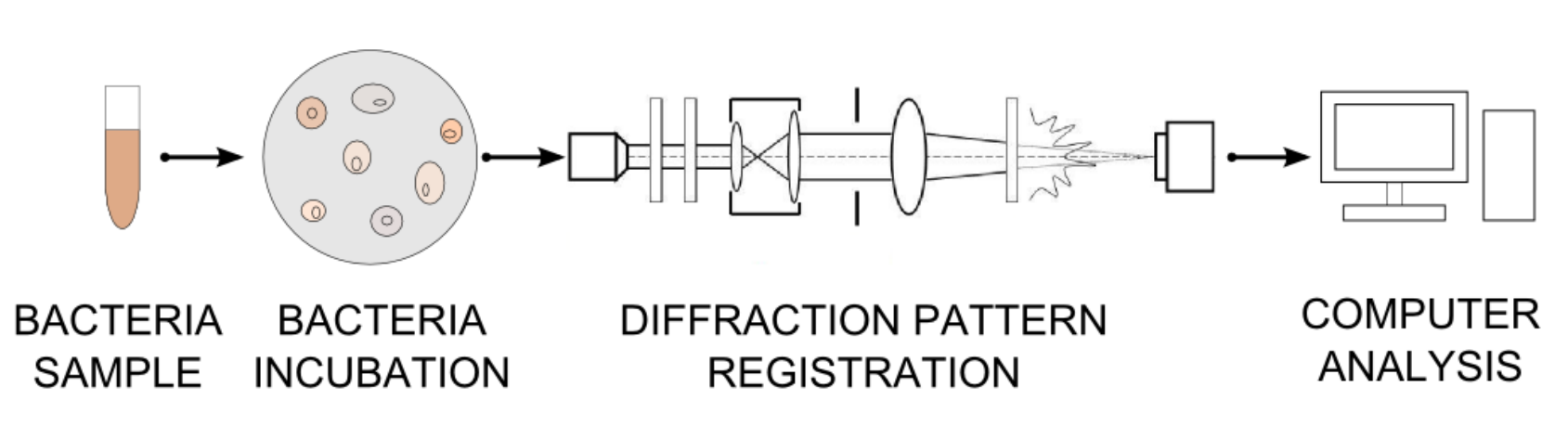
*Staphylococcus aureus* (PCM 2267) | *Salmonella Enteritidis* (ATCC 13076) | *Staphylococcus intermedius* (PCM 2405)

QUANTUP


This was when my adventure began

# What is it all about?!

The experiment workflow



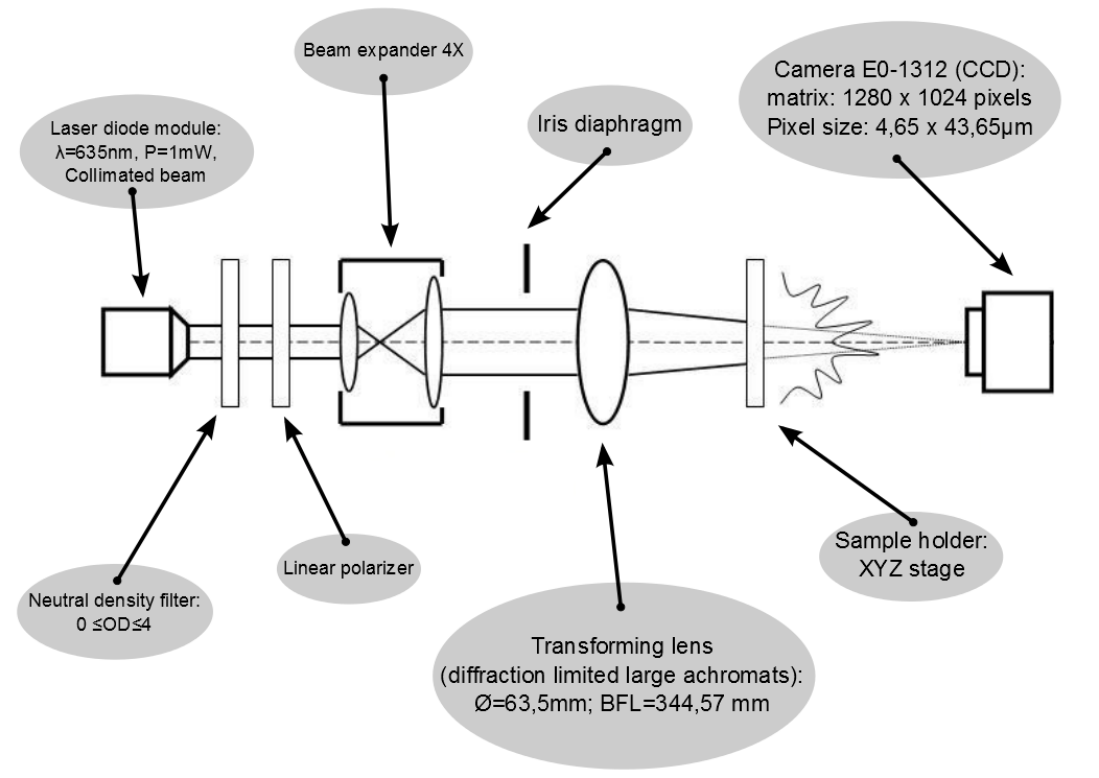

QUANTUP

# Microbiology


QUANTUP

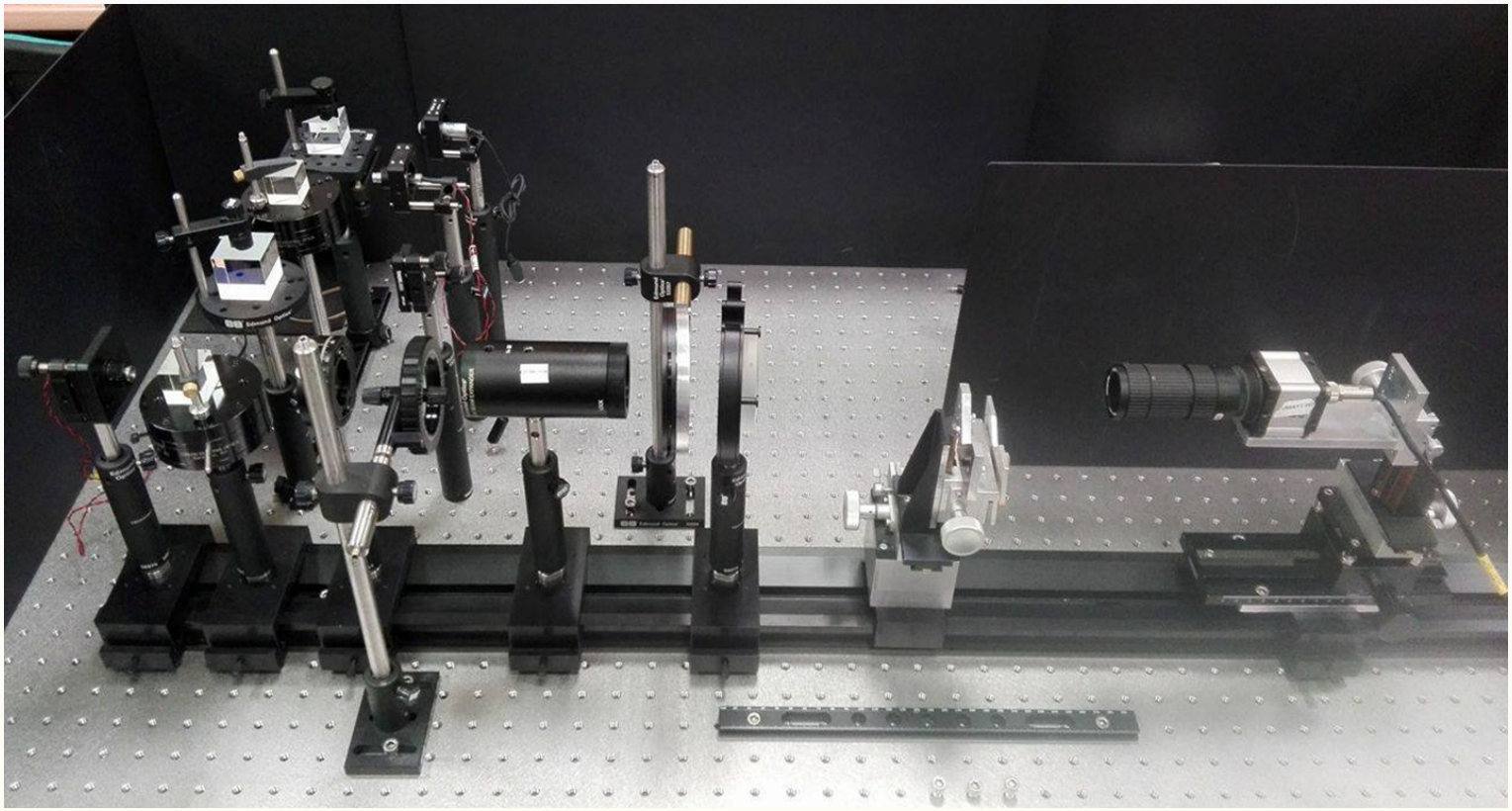
# What was done in optics

Igor invented an optical device for light diffraction on bacterial colonies
(it was patented)



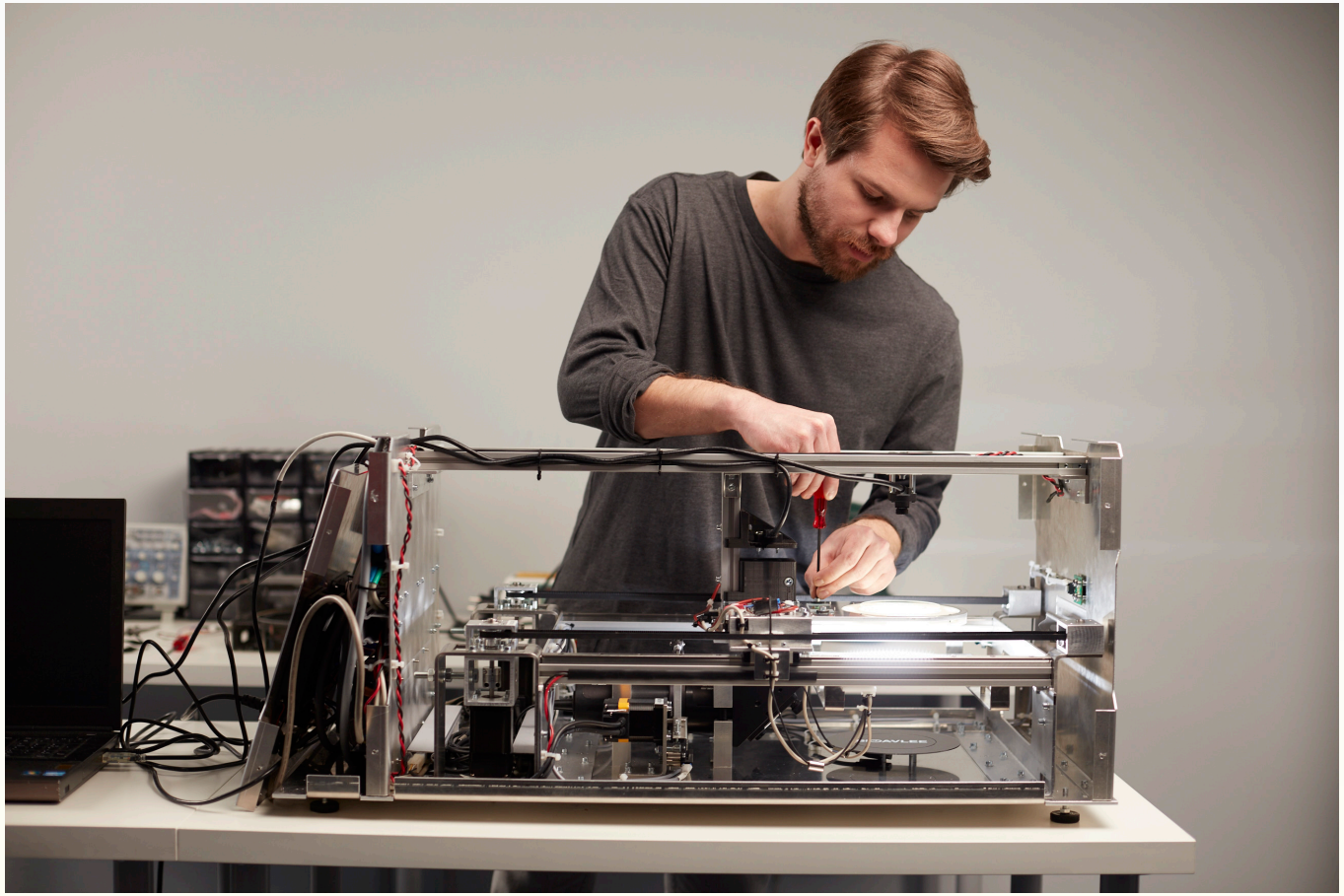

QUANTUP

# Optical system 2008-2012



QUANTUP

# Optical system 2018



QUANTUP

# Analytics

The **analytical part** was mostly about the idea!

But the results needed years of work...

QUANTUP

# What was done in analytics

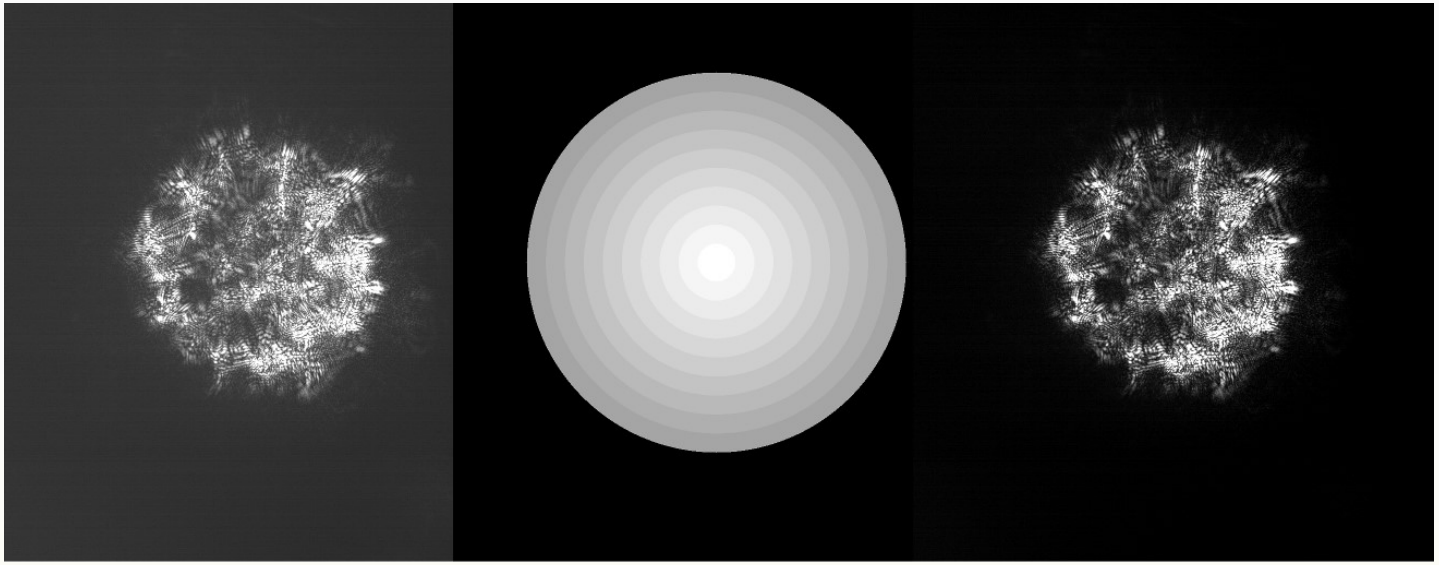
Workflow consisted of modules with packages:

- image processing (`pixmap`, `rtiff`, `imager`)
- feature extraction (`moments`)
- feature selection (`survey`, `pamr`, `ggplot2`, `sets`)
- predictive modeling (`ipred`, `e1071`, `tree`, `klaR`)
- performance assessment (`MASS`, `xtable`, `Multiclasstesting`)

QUANTUP

# Image processing

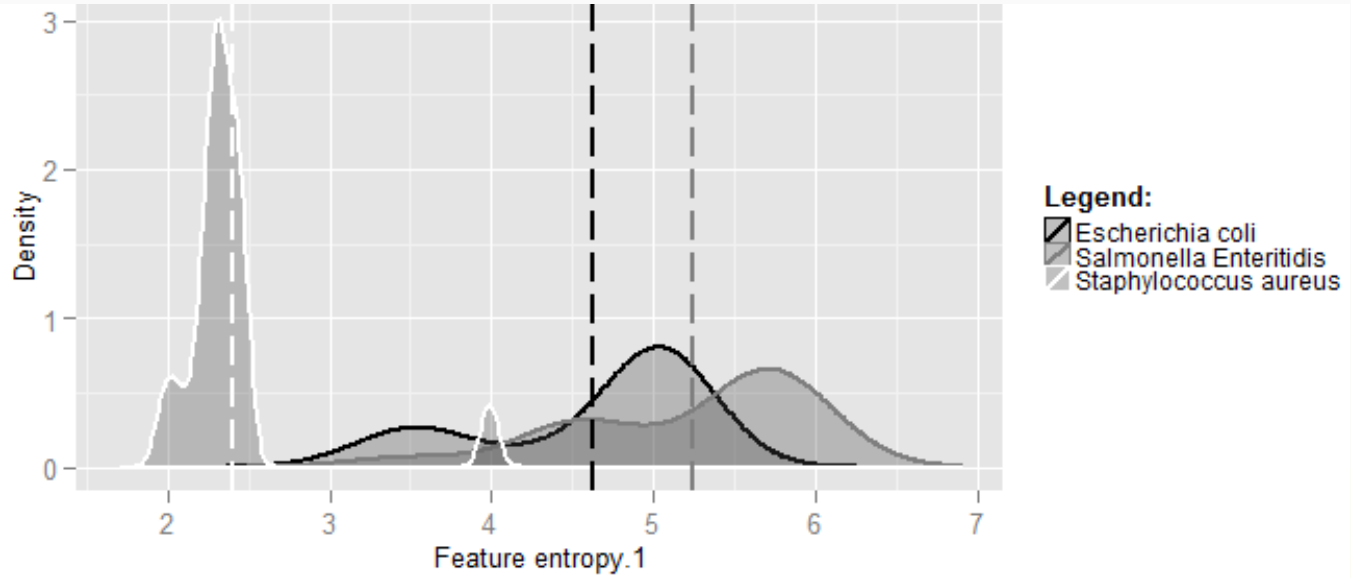
Diffraction patterns must be comparable between each other.



QUANTUP

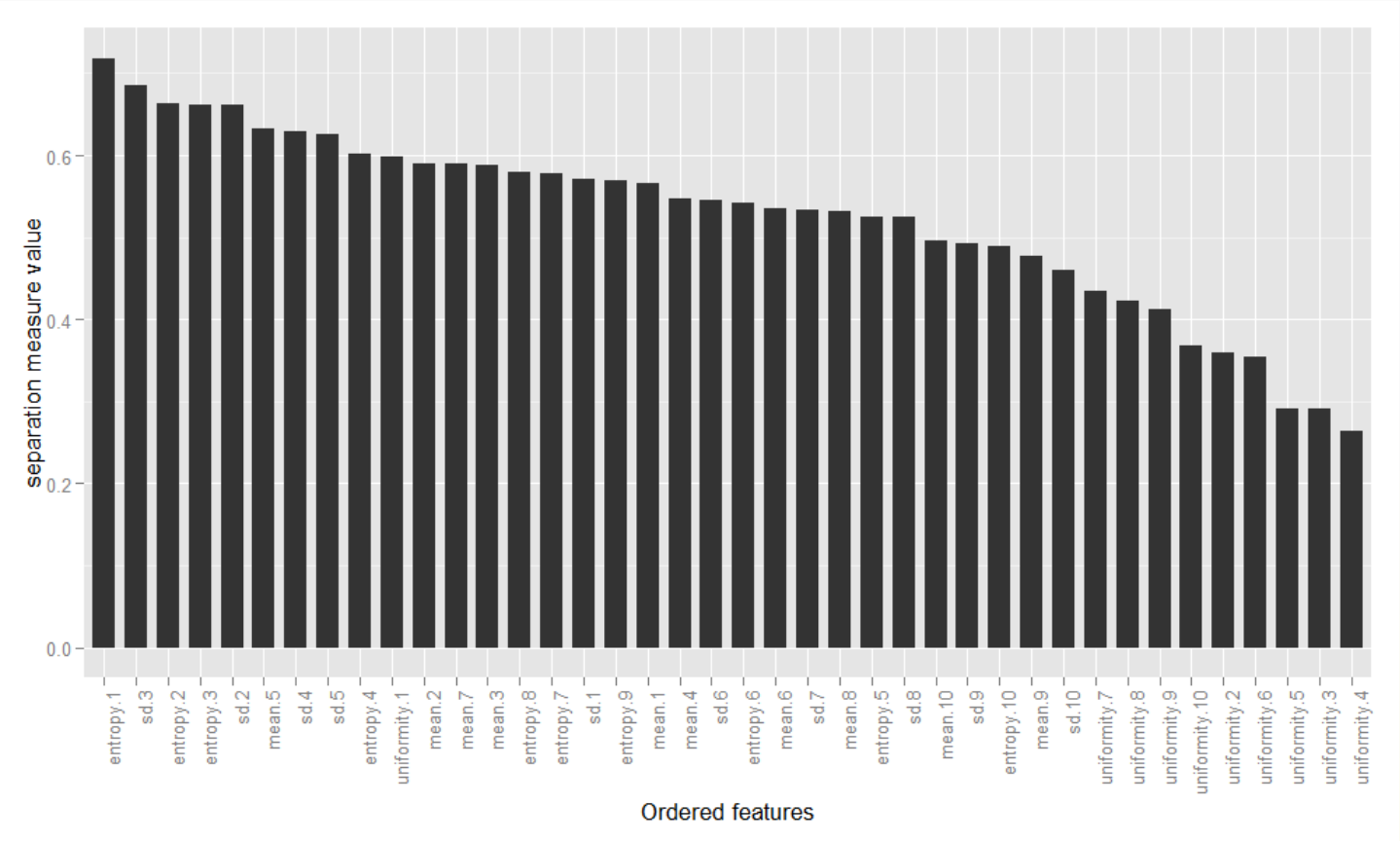
# Feature extraction

Example of empirical density of most discriminant feature.




QUANTUP

# Feature selection




QUANTUP

# Identification of bacteria species (or strains)

Number of methods was tested:

- LDA, QDA (to start with)
- neural networks (numerical features and raw images)
- SVM
- random forest
- other

QUANTUP

# Identification process was…

…far from automated or repeatable!

Human assistance was needed:

- whole laboratory process handling
- whole registration processing
- detection of borders of diffraction patterns
- analysis handling
- summary of results

QUANTUP

# Performance assessment from 2013

CV error estimation with 10 fold stratified sampling for best fitted classification models after ANOVA feature selection for normalized data of 7 laboratory bacteria species.

| | LDA | QDA | SVM |
|---|---|---|---|
| features | sd.4 + sd.3 + mean.5 + mean.4 + sd.2 + sd.5 + sd.1 + mean.1 + mean.7 + mean.3 + mean.2 + mean.6 + mean.10 + mean.8 + mean.9 + sd.9 | sd.4 + sd.3 + mean.5 + mean.4 + sd.2 + sd.5 + sd.1 + mean.1 + mean.7 + mean.3 + mean.2 + mean.6 + mean.10 | sd.4 + sd.3 + mean.5 + mean.4 + sd.2 + sd.5 + sd.1 + mean.1 + mean.7 + mean.3 + mean.2 + mean.6 + mean.10 + mean.8 + mean.9 + sd.9 |
| CV error | 5.11% | 1.42% | 2.27% |

QUANTUP

# Changes over time – analytics and UI

Some novelties came over the years:

- coding standards (`dplyr`, `magrittr`, `lazyeval`, `tidyverse`, `rlang`, `reshape2`)
- UI (`Shiny`, `DT`, `openxlsx`)
- reporting (`Rmarkdown`, `knitr`, `kableExtra`, `flextable`, `wrapr`)
- documentation (`roxygen`)

And then **XGBoost** came and stayed for good :)

QUANTUP

# Changes over time

BIOAVLEE™

- professionals working on every aspect of the method
- a few prototypes (now it’s a pre-production prototype)
- tests on environmental and lab bacterial species
- number of strains for each species
- repeatability tests for devices and microbiological conditions
- automation of registration – elimination of human bias
- dedicated modifications of algorithms

**https://bioavlee.com**

QUANTUP

# Exemplary results 2018 I

XGBoost with SNR feature selection for random division of data into test and teach sets of 4 environmental bacteria species.

Up to date algorithms include detection of *unknown* species and adaptive threshold determination for acceptation of results dependent on percentage match level.

| Real | Identified | Count |
|---|---|---|
| Candida albicans | Candida albicans | 121 |
| Candida albicans | Unknown | 4 |
| Escherichia coli | Escherichia coli | 113 |
| Escherichia coli | Unknown | 12 |
| Pseudomonas aeruginosa | Pseudomonas aeruginosa | 119 |
| Pseudomonas aeruginosa | Unknown | 6 |
| Staphylococcus aureus | Staphylococcus aureus | 123 |
| Staphylococcus aureus | Unknown | 2 |

QUANTUP

# Exemplary results 2018 II

Summary of results obtained on environmental data and over 2000 diffraction patterns from two diffrent devices after applying all the changes.

| Bacteria species | ACC | Sensitivity | Specificity |
|---|---|---|---|
| Staphylococcus aureus | 1.00 | 0.98 | 1.00 |
| Unknown | 0.95 | NaN | 0.95 |
| Pseudomonas aeruginosa | 0.99 | 0.95 | 1.00 |
| Escherichia coli | 0.98 | 0.90 | 1.00 |
| Candida albicans | 0.99 | 0.97 | 1.00 |
| **AVERAGE** | **0.99** | **0.95** | **1.00** |

QUANTUP

# Why R?

Because:

- it is suitable for image processing
- irreplaceable for predictive analysis of numerical data
- perfect for applications with analytical background
- extremely convenient for reporting analytical results

**It is decided to use R and Shiny for MVP.**

QUANTUP

# Has the world changed?

# How I've managed to change the world

- Together with Igor Buzalewicz, I developed a method for identifying bacteria species – faster, cheaper and very accurate
- World acquired a Doctor of Technical Sciences
- A Bioavlee company was created that gives work to about 10 people

**Soon our method will enter the market**
**and the time of next, much bigger changes will come**

QUANTUP


Who was that?

# Agnieszka Suchwałko, Ph.D.

Professionally:

- Associate & Data Scientist at QuantUp
- Data Scientist at Bioavlee
- graduate of Wroclaw University of Technology (M.Sc. and Ph.D.)
- over 10 years of experience
- agnieszka@quantup.pl

Privatelly:

- mother of Two (3,5 year and 6 months)
- dog lover (especially my own Gonzo the beagle)
- wife and partner (come and listen to his talk tommorow)

QUANTUP