# PAVEL VALOV

**Data Analyst / Data Scientist**

Email: valov.pm@gmail.com

Phone: +79111178585 ◇ Telegram: t.me/valovpm

GitHub: bit.ly/3KFv5Ha ◇ Google Scholar: bit.ly/3zp5oDx

## ABOUT ME

Data Analyst with experience in development, research, and analytics both in industry and academy. Strong knowledge of SQL, Python, and data analysis. Honest, responsible, driven, fluent Russian and English proficiency, honed by years of teaching and research.

## SKILLS

| | |
|---|---|
| **Programming Languages** | Python, SQL, R, C# |
| **Technologies & Libraries** | Python (numpy, pandas, scikit-learn, matplotlib, jupyter, etc.)<br>R (dplyr, ggplot2, kernlab, reshape2, rpart, sqldf, tidyr, etc.)<br>.NET (ADO.NET, Entity Framework, LINQ, WinForms, WCF, etc.) |
| **Data Analysis** | Probability theory, Statistics, A/B testing |
| **Database Systems** | YT (Yandex Table), ClickHouse, PostgreSQL, MS SQL Server |
| **Business Intelligence** | Yandex Nirvana, Yandex DataLens, Apache Airflow, Apache Superset |
| **Operating Systems** | Linux (Arch, Manjaro, Ubuntu, Bash, CLI), Windows |
| **Languages** | Russian (native), English (C2, fluent, academic writing) |

## EXPERIENCE

**Analyst** June 2023 - August 2024
*Yandex* *Saint-Petersburg, Russia*

- Proposed, developed, and implemented a methodology for calculating CPT Administration and CPT Recruitment for the Employee Lifecycle Management Center, based on data from the Strategy and Finance Group.
- Developed and implemented processes for setting Administration PPU based on CPT data.
- Enhanced Crowd processes and developed dashboards with Nirvana, Groovy, YT, and DataLens.

**Research & Teaching Assistant** September 2014 - February 2023
*University of Waterloo* *Waterloo, ON, Canada*

- Analyzed and visualized performance data of configurable software systems in heterogeneous hardware and cloud environments, trained prediction models based on performance data, using Python (pandas, scikit-learn, matplotlib), R (tidyr, dplyr, reshape2, ggplot2), Microsoft Azure.
- Conducted lectures, practices, marked code, developed scripts for testing code, for courses: 'Algorithms and complexity', 'Introduction to Computer Science', 'Object-oriented programming', 'Software requirements', 'Functional programming', using: C, Java, Python, Racket (Scheme).
- Developed and published 4 methods for predicting performance of configurable software systems across heterogeneous hardware using machine learning models, and presented the results at 3 international conferences.
- Dedicated about 4000 hours to educational activities, including lectures, practices, seminars.

**Research Assistant** September 2015 - August 2016
*Pratt & Whitney Canada* *Montreal, QC, Canada*

- Analyzed and visualized the architecture of a software system using design patterns, UML, SysML.
- Converted the classical architecture of the analyzed software system to Software Product Line architecture.

**Software Development Engineer** September 2011 - August 2014
*Aller Petfood LLC* *Saint-Petersburg, Russia*

- Analyzed and visualized manufacturing data, developed a comprehensive manufacturing process reporting system using: C#, .NET (ADO.NET, WCF, WinForms), Microsoft SQL Server (Core, Reporting Services).
- Implemented a distributed production documentation workflow system, eliminating manual document writing.
- Automated data collection from equipment and integrated it with the production documentation system.

## EDUCATION

**PhD Thesis** [bit.ly/3P7ja3L], **Master Thesis** [bit.ly/3zsyXUG] — September 2014 - October 2023
*University of Waterloo* — *Waterloo, ON, Canada*

**Bachelor of Applied Science (BASc)** — September 2010 - June 2014
*ITMO University* — *Saint-Petersburg, Russia*

## REPOSITORIES, PUBLICATIONS, ADDITIONAL LINKS

**AB-testing, bootcamp** — September 2024 - current time
*Karpov.Courses* — *www.karpov.courses*

- Analyzed parallel A/B tests, estimated sample size, empirical errors, confidence intervals, and performed error correction using Holm's method [bit.ly/45W1LFs, bit.ly/3VvUE1u]
- Analyzed how CUPED reduces variance in data and how it affects the p-value [bit.ly/4fWVCgI]
- Analyzed how removing different percentages of outliers affects statistical power [bit.ly/3kTMnGb]
- Analyzed how removing different percentages of outliers affects sensitivity [bit.ly/3HYWJQA]
- Analyzed different methods of calculating confidence intervals [bit.ly/4p2wHwn, bit.ly/4mB6PWV]
- Compared methods of introducing effects in synthetic A/B tests [bit.ly/4g7PRNC, bit.ly/4g1RLPH]
- Calculated experimental group sizes and MDE for A/B tests [bit.ly/47VUl7D, bit.ly/4g0hsAd]

**Data Analyst, bootcamp** — March 2023 - May 2023
*Karpov.Courses* — *www.karpov.courses*

- Analysis of the novel newsfeed recommendation algorithm, designed to improve the key metric (CTR):
- A/B testing to demonstrate CTR deterioration with a new recommendation algorithm using: transformations of the initial data (Laplace smoothing, Poisson bootstrap, bucket transformation), normality criterias (Shapiro-Wilk, D'Agostino), distribution difference criterias (Student's T-test, Mann-Whitney U-test), SQL, ClickHouse, Python (pandas, matplotlib) [bit.ly/3kTMnGb]
- Demonstration of increasing of a key metric sensitivity using the linearization method [bit.ly/3IWRXPQ]
- A/A testing to check CTR consistency across different datasets [bit.ly/3L0bmCr]
- ETL-pipelines for sending reports to ClickHouse and Telegram using Apache Airflow, Python, SQL:
- Pipeline for monitoring and sending a report in case of an anomaly in the metrics [bit.ly/3ZJUEeD]
- Pipeline of a report to Telegram on key metrics of two products in different time slices [bit.ly/3ZJhd31]
- Pipeline of a report to Telegram on basic product metrics (DAU, views, likes, CTR) [bit.ly/3mvMCI3]
- Pipeline for sending a report to ClickHouse about the basic product metrics in different slices [bit.ly/3mrzX8N]
- Dashboards for visualization and analysis of key metrics using Apache Superset, ClickHouse, SQL:
- Dashboard for analyzing the abnormal drop in the active audience of the newsfeed [bit.ly/413UzTW]
- Dashboard for analyzing differences in the behavior of 'organic' and 'advertising' users [bit.ly/400X4VR]
- Dashboard for analyzing basic product metrics of the newsfeed (likes, view, CTR, etc.) [bit.ly/4138Tfo]
- Dashboard for analyzing audience metrics of several products (DAU, MAU, WAU, etc.) [bit.ly/3obs2NF]

**Research Assistant** — September 2014 - February 2023
*University of Waterloo* — *Waterloo, Ontario, Canada*

- Designed and published a machine learning approach for *approximating and transferring Pareto frontiers* of systems' properties across different cloud environments, using Python ecosystem (pandas, scikit-learn, matplotlib). Repository [bit.ly/3nuIs3p] Paper [bit.ly/3oPyUxk] Video [bit.ly/3bqLP5W] Slides [bit.ly/3BFIF9f]
- Designed and published a machine learning approach for *generalizing performance prediction models* of configurable systems across different hardware platforms, extensively using R (tidyr, dplyr, reshape2, ggplot2, etc). Repository [bit.ly/3K8o3sA] Paper [bit.ly/3d0FVsF]
- Designed and published a machine learning study on *comparison of various performance prediction methods*, while extensively using R ecosystem (tidyr, dplyr, reshape2, ggplot2, etc).
Repository [bit.ly/43dJL7x] Paper [bit.ly/3d3T30d]