

This is source{d}.

Vadim Markovtsev

source{d}

Machine Learning for
Large Scale Code Analysis

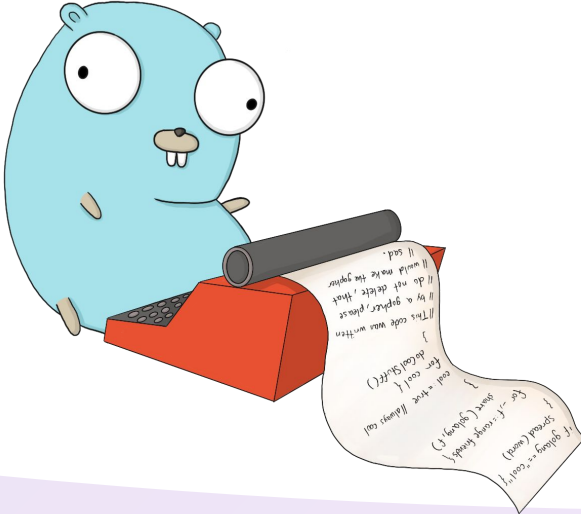
Who I am

Vadim Markovtsev

source{d}

- Machine Learning team leader
- Joined source{d} in mid-2016
- Worked in 5 software companies, e.g. Samsung Research, Mail.Ru
- Spoke 30+ times on IT conferences, from meetups to AAA
- Master in Applied Mathematics (Moscow University of Physics and Technology, 2012)

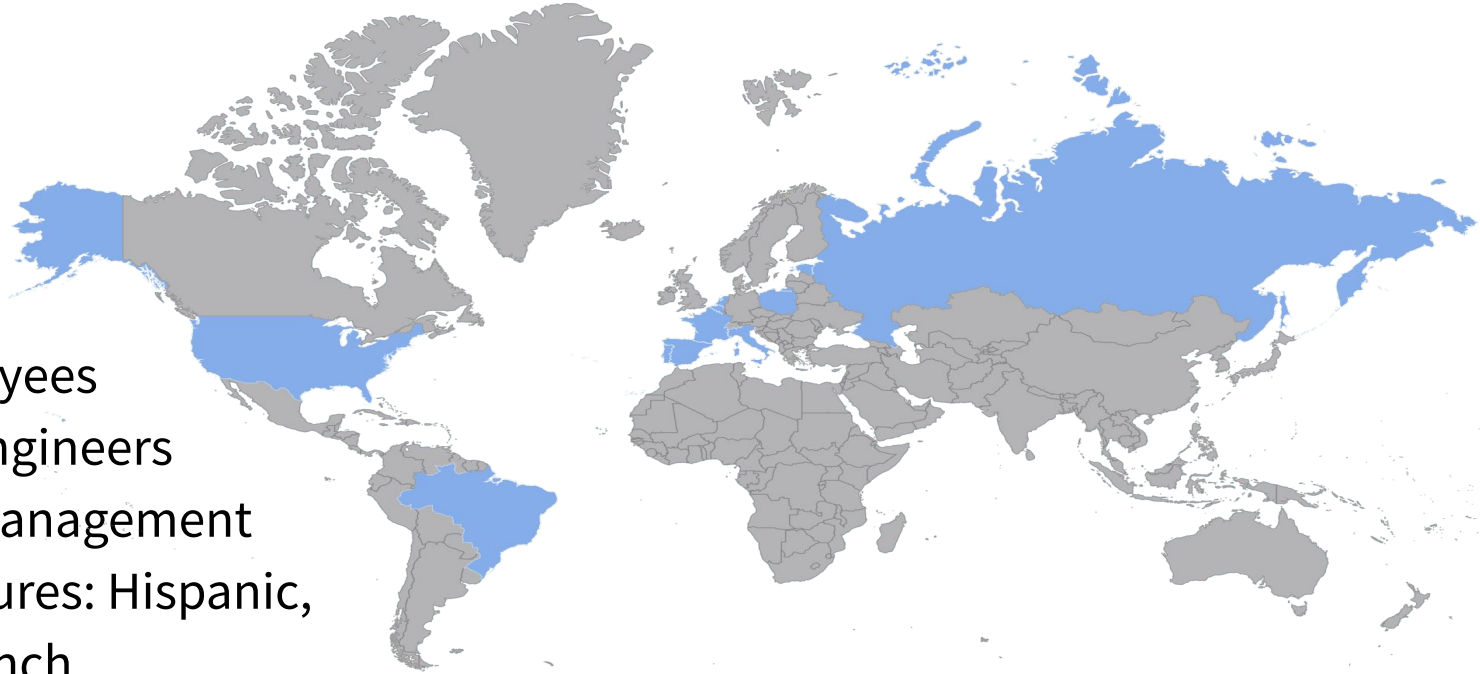
How we work



Who we are

source{d}

- 35+ employees
- 70% are engineers
- 15% are management
- Major cultures: Hispanic, Slavic, French



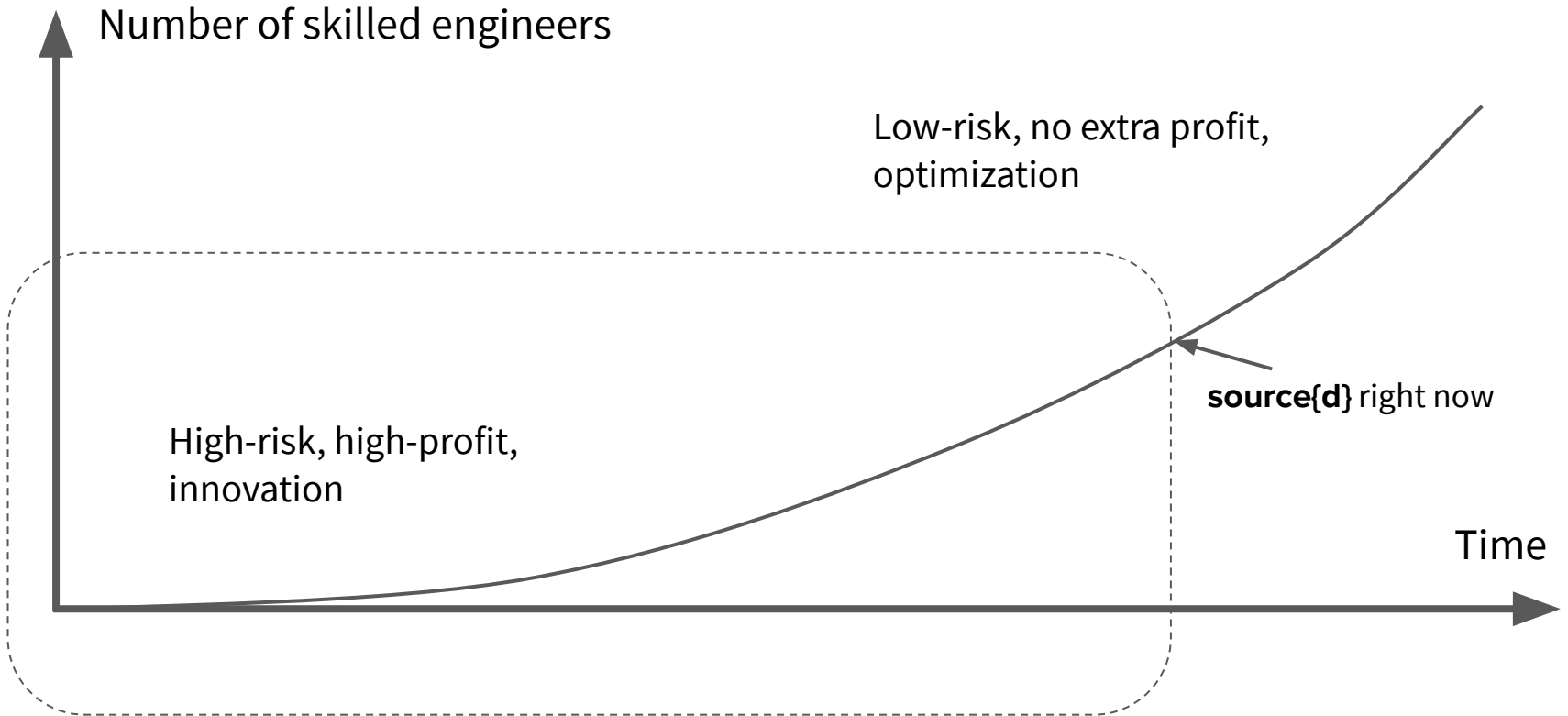
Timeline

source{d}

- Early 2015: appear as Tyba, seed investments
- Late 2015: pivot, become source{d} (new CTO's side project)
- Early 2016: run out of money, some employees leave
- Mid 2016: raise €5mm, salvation
- Late 2016: pivot, stop making money and dismiss 50%
- 2017: let's clone all Git repositories, parse them, and do MLonCode
- Late 2017: remote first
- 2018: assisted code review with MLonCode
- Early 2019: engineering observability... with MLonCode
- Mid 2019: first revenue and raise €Xmm

Human resources

source{d}



Offices

source{d}

- HQ: Madrid and San Francisco
- 60% are remote
- Flexible everything

Perks

source{d}

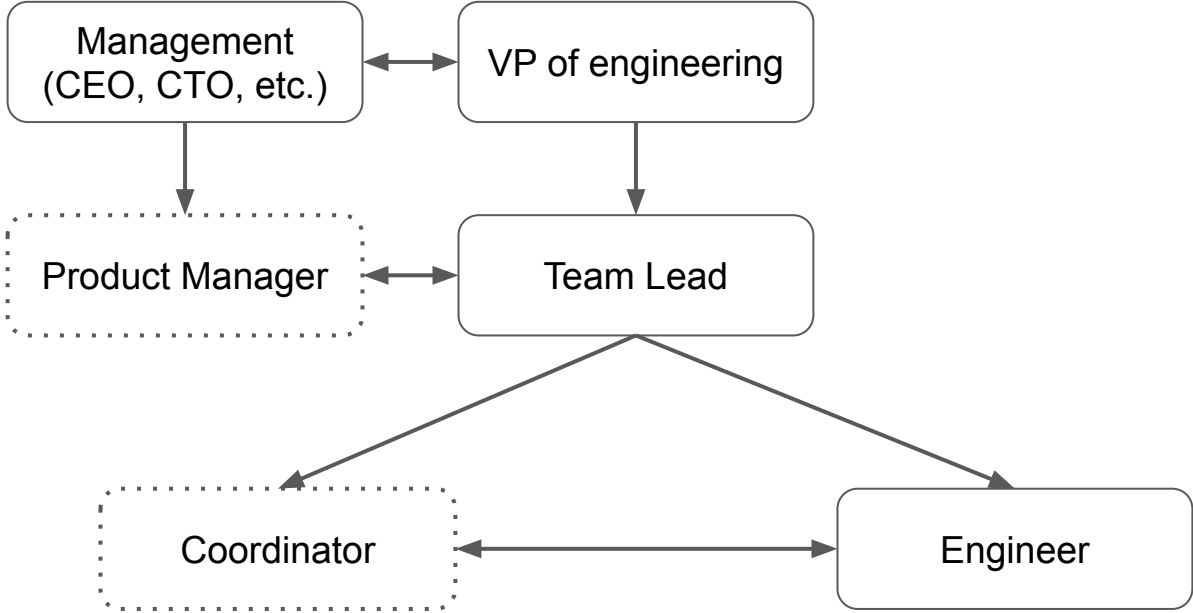
- 23 days of vacations
- Open Source Day
- Research Day
- Conferences
- Papers

Teams

source{d}

- Applications
- Solutions
- Machine Learning, applied
- Machine Learning, research
- Data Processing
- Data Retrieval
- Language Analysis
- Developer Operations
- Developer Relations
- Quality Assurance
- Product management
- Business Intelligence
- Management

Structure



Hiring

source{d}

- Remote coding challenge
- Remote Machine Learning challenge*
- Personal with CTO or VP of Engineering
- Personal with CEO (sometimes)
- Design interview
- Machine Learning interview*
- Q&A interview
- Logical Thinking interview

Open discussion with veto-ing and overriding by the team lead

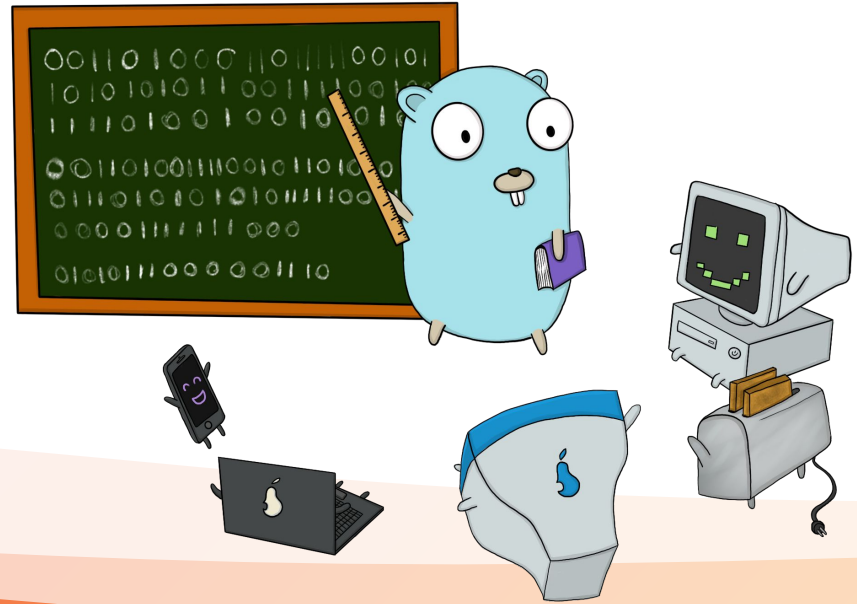
github.com/src-d/guide

Culture

source{d}

- 4 full-time engineers left since mid-2016
- 1 joined back
 - He became our VP of engineering

Technologies



Programming languages

source{d}

1. Go
2. Python 3
3. Scala
4. Javascript, Typescript
5. C/C++
6. CUDA

Our engineers play with

source{d}

- Rust
- Elm
- D
- Haskell

Tools

source{d}

- Git, GitHub
- Linux, macOS
- Visual Studio Code
- PyCharm, GoLand, CLion
- vim
- Ghost
- Gimp, Inkscape
- shwr.me
- Slack
- Google Docs
- Zoom, appear.in
- Octobox

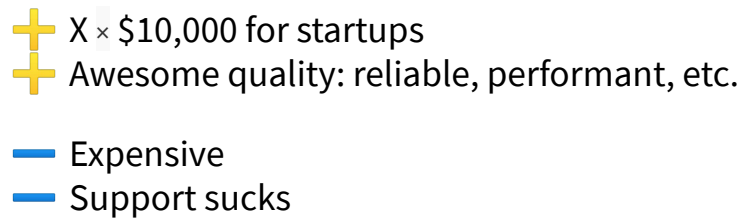
If we go deeper

source{d}

- Indent with spaces
- Switched to Go modules, no vendoring
- go-git
- Python scientific stack
- Tensorflow and Pytorch
- Apache Spark
- React
- Kubernetes, Docker

Clouds

- Amazon
- Azure (Microsoft)
- Google
- etc.



- + X × \$10,000 for startups
- + Awesome quality: reliable, performant, etc.
- Expensive
- Support sucks

Google

GitHub

Computer science stuff we did use

source{d}

Force yourself to study the theory, this is your competitive advantage

- Graph traversal
- Connected components, shortest path, etc.
- Linear Programming: bipartite matching, network flow, and many others
- Convex optimization
- Grammar parsers
- Complexity theory
- String algorithms, e.g. LCS with suffix arrays
- Compression theory
- Disjoint sets
- Merkle tree
- Dynamic Programming
- Max-SAT
- ...

Machine Learning

Natural Language Processing

source{d}

- Working with word distributions
- Pipelines
- word2vec; Swivel
- Transformers

Classics

source{d}

- Linear Regression
- Random Tree Forest
- Production Rules
- GBDT: xgboost, catboost
- Hyperparameter optimization

Clustering

source{d}

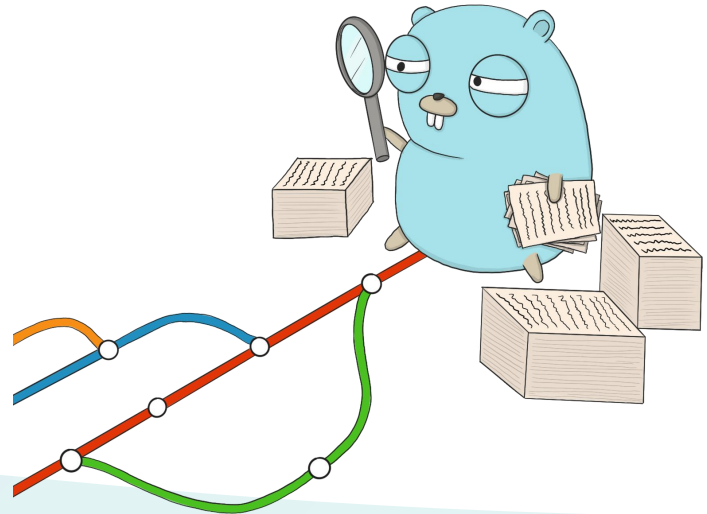
- K-Means
- t-SNE
- UMAP
- k-NN, e.g. KD Tree, VP Tree
- aNN, e.g. hnsw
- MinHash, Weighted MinHash

Deep Learning

source{d}

- Char-level CNN
- Transformers, Inception
- LSTM, GRU
- Gated Graph Neural Networks

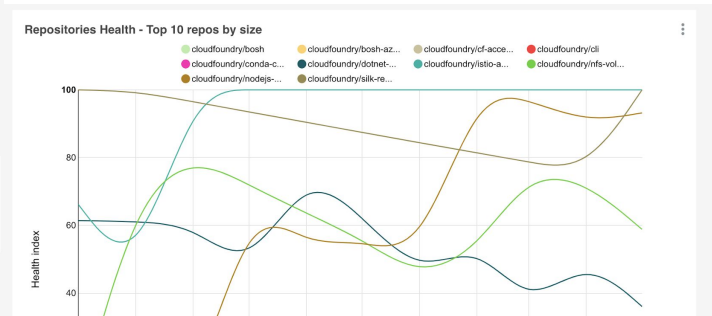
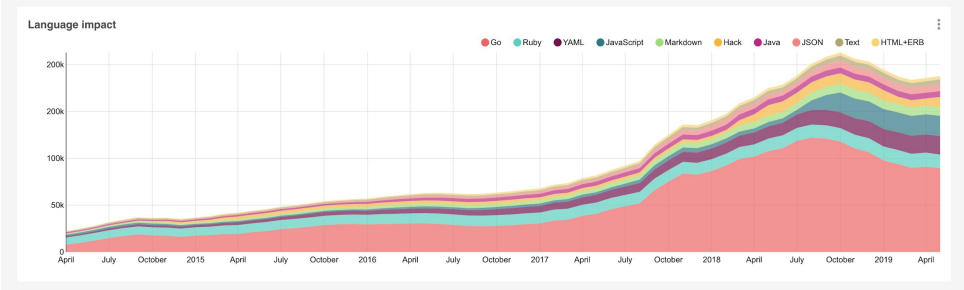
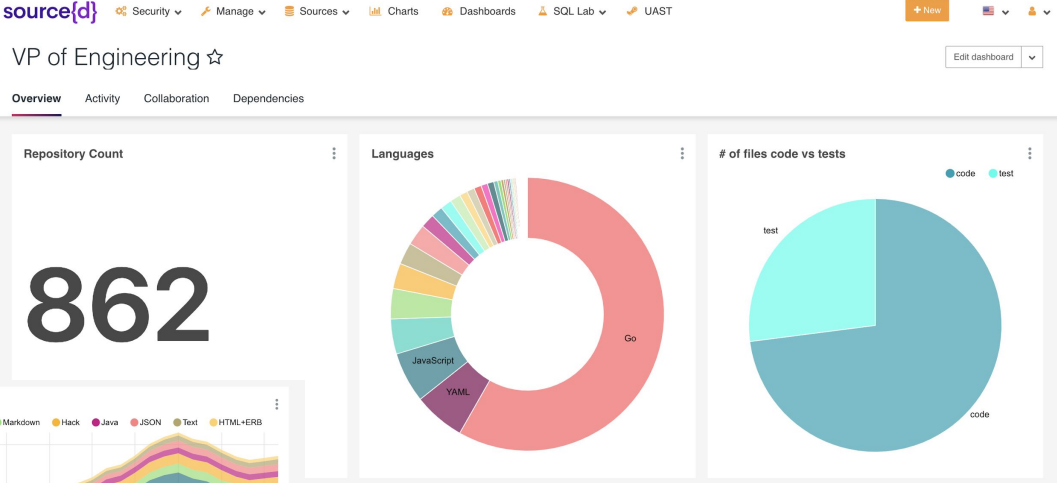
What we do



Code as Data

Engineering Observability

- "Code Lake"
- Dashboards
- Advanced Insights



Machine Learning on Source Code

source{d}

Assisted Code Review

- Automatic Program Repair
- Code Naturalness



lookout-staging bot just now

format: style mismatch:

Suggested change Beta ⓘ [Give us feedback](#)

```
24 - const {Color} = require('./color');  
24 + const { Color } = require('./color');
```

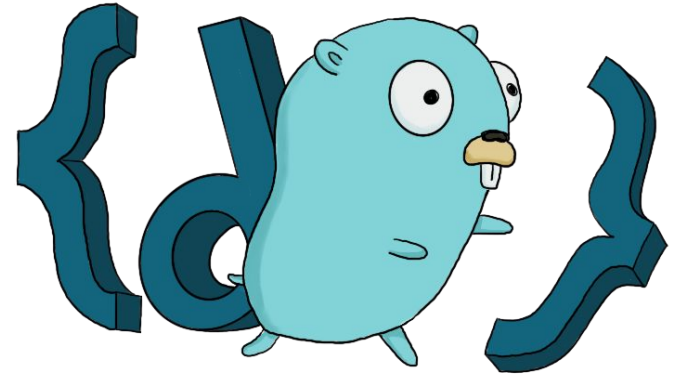
[Commit suggestion](#) ▼ [Add suggestion to batch](#)

05381b86 □ at column 6 should be removed.

f4b2f444 □ should be inserted at column 8.

3688cf1e □ should be inserted at column 13.

Challenges



Remote communication

Talent

Transparency

Pioneering

Academia

Sales

Scaling

source{d}

Machine Learning for Large Scale Code Analysis

sourced.tech · github.com/src-d · [@sourcedtech](https://twitter.com/sourcedtech) · blog.sourced.tech