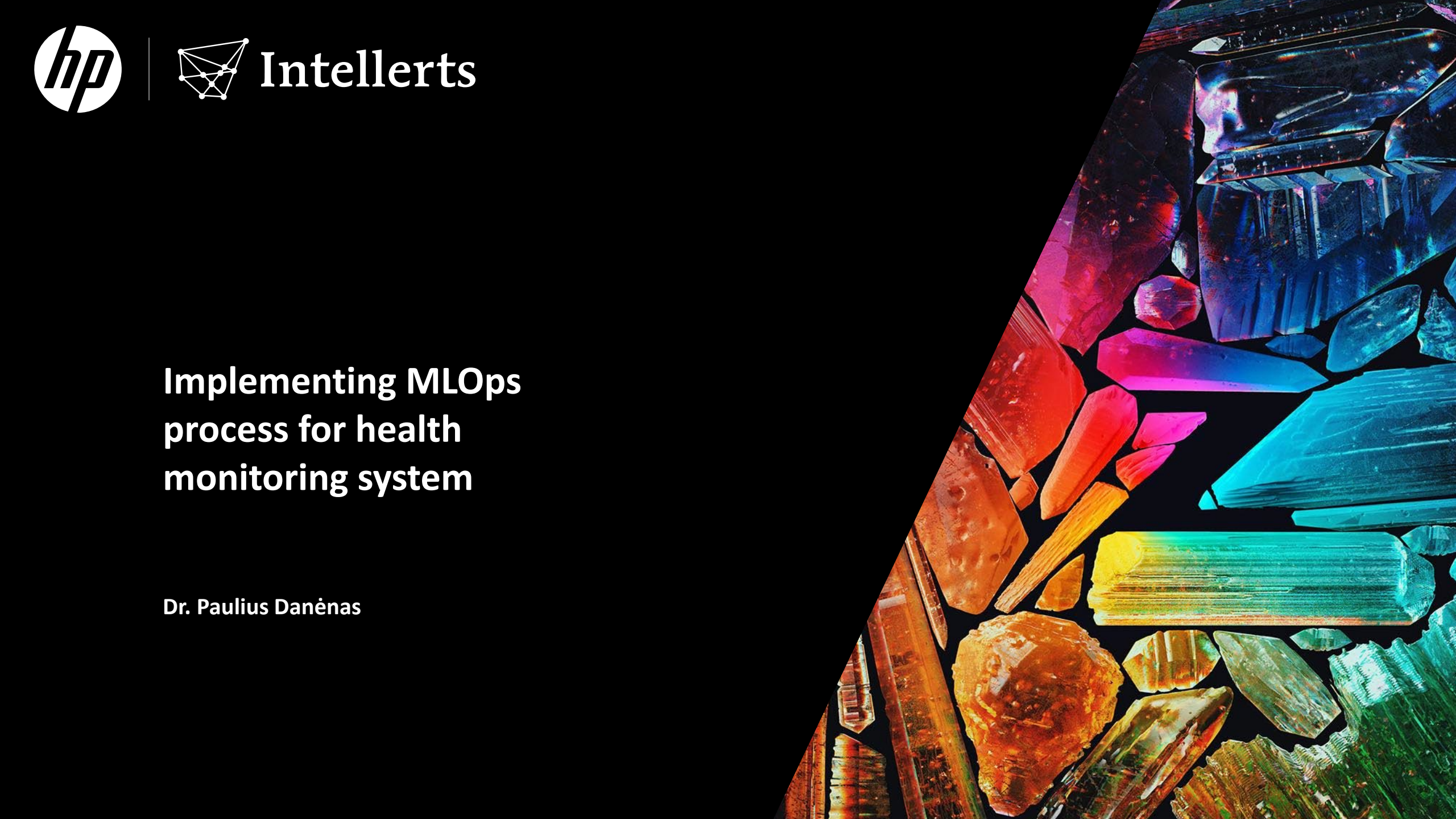
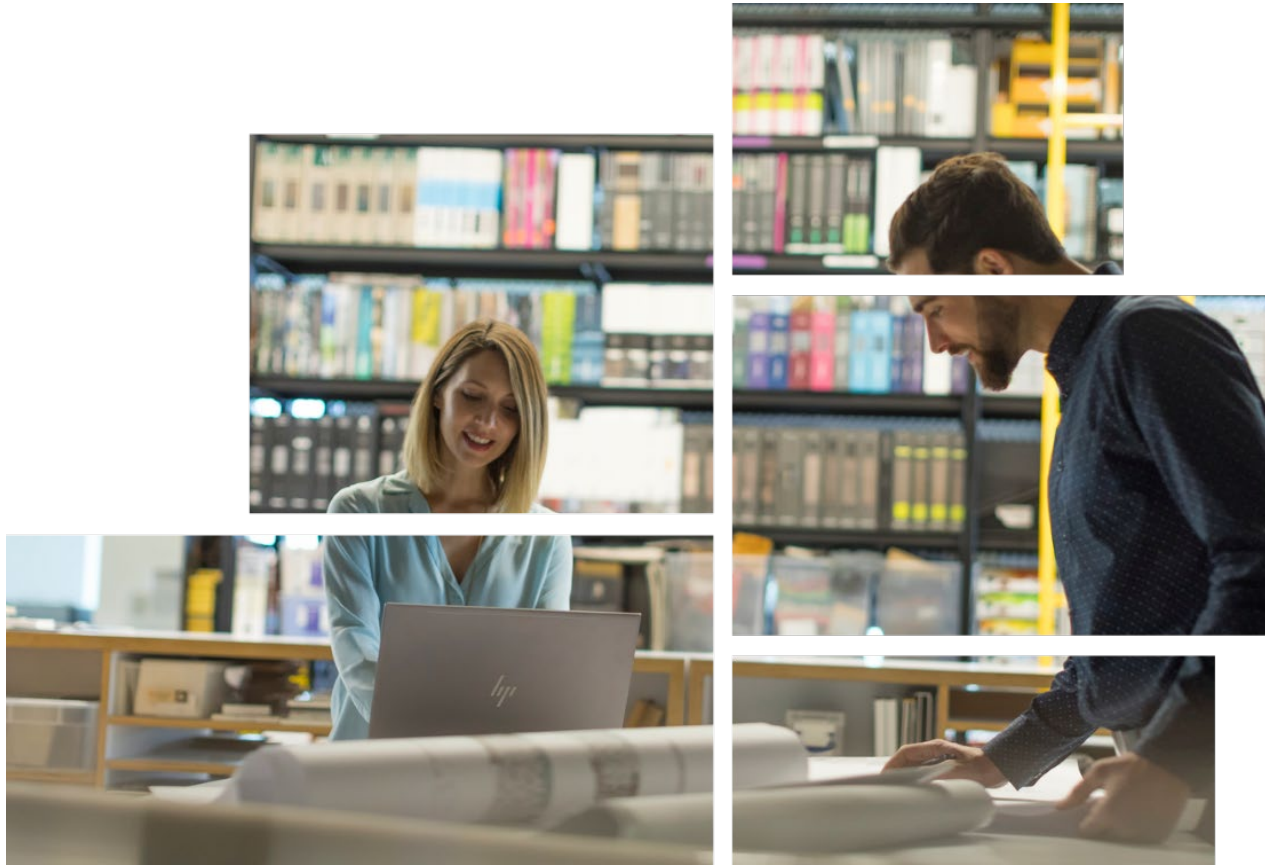


Implementing MLOps process for health monitoring system

Dr. Paulius Danėnas

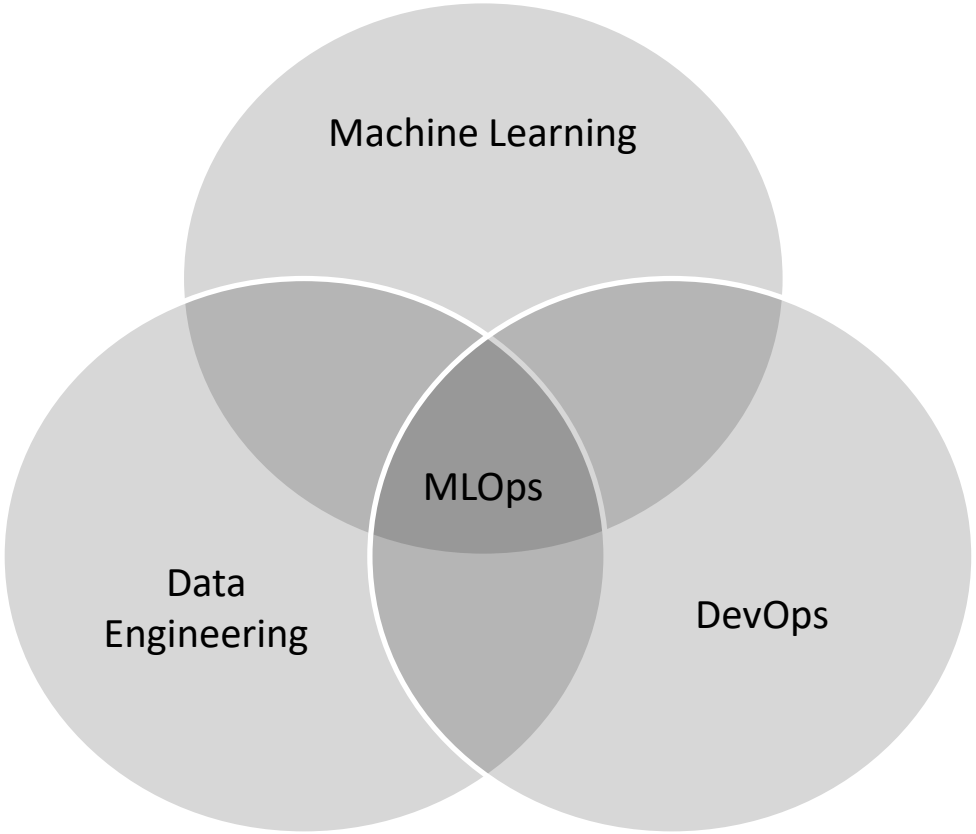
Agenda

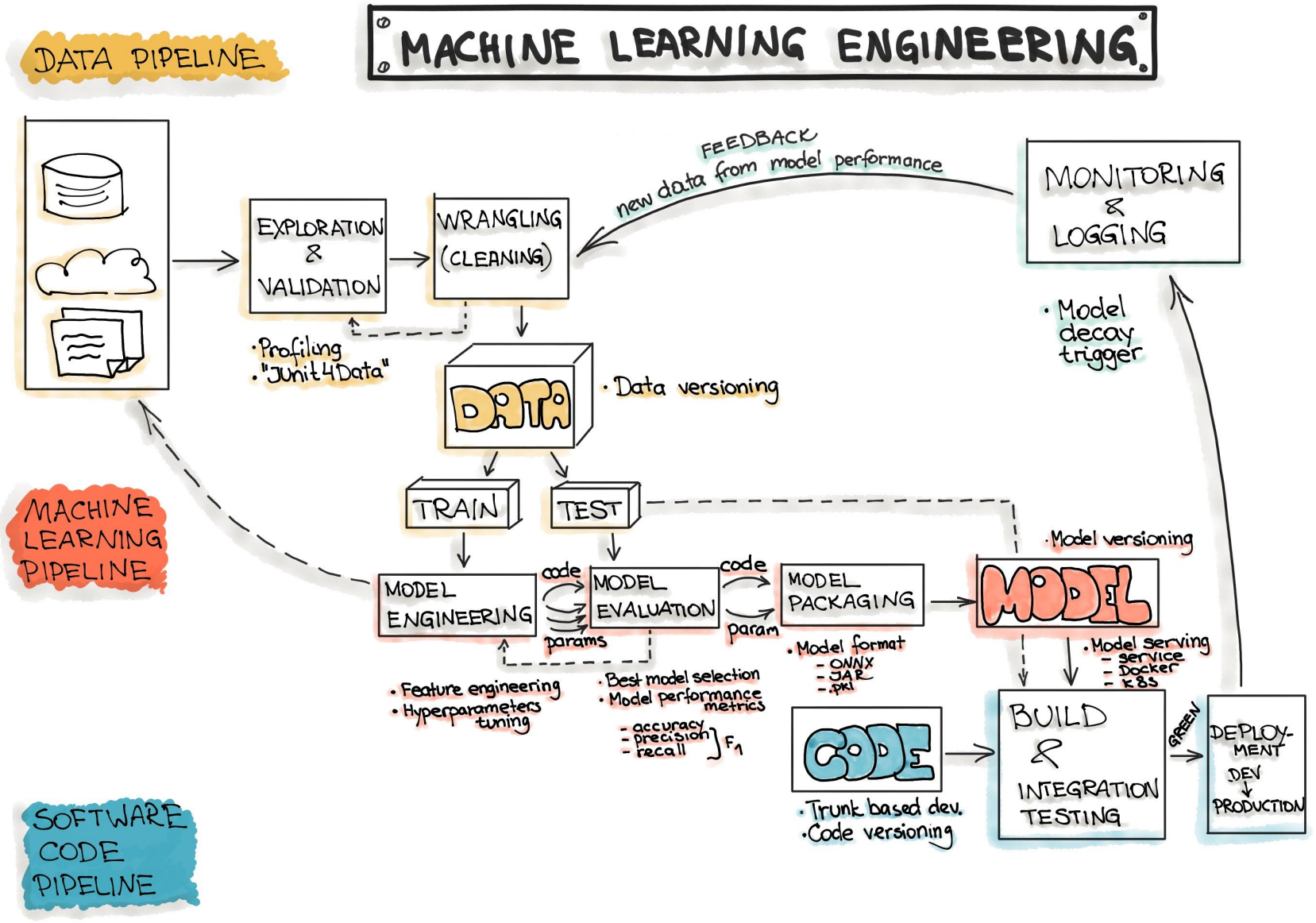
- What is MLOps?
- The key components of MLOps-enabled systems
- Case study
- Infrastructure



MLOps concept

- MLOps = ML + DevOps
- Might include additional xOps, like DataOps, DevSecOps, ModelOps, GitOps…
- Highly emerging area…
- … although both ML and DevOps have existed long before

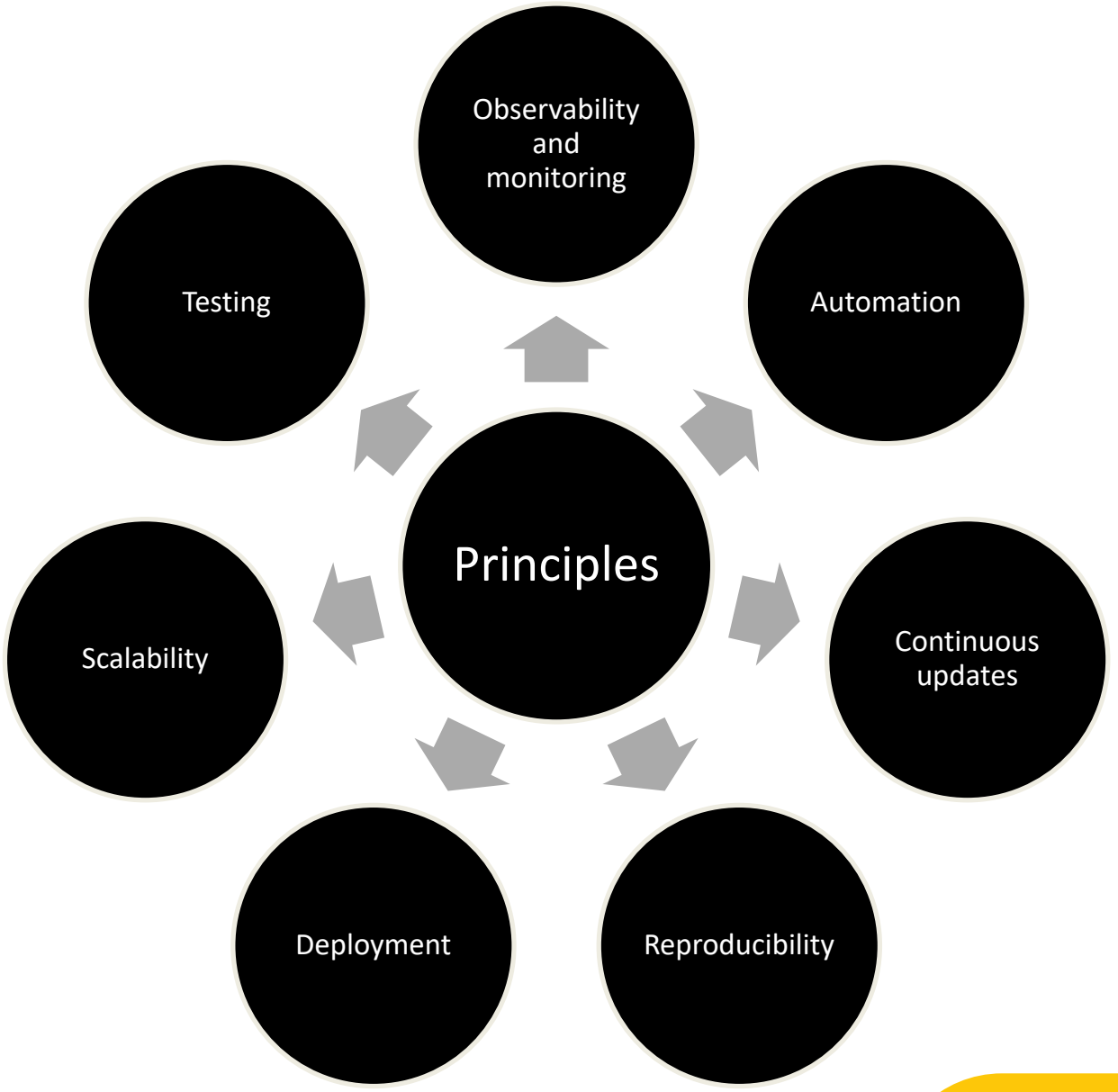




Source: https://ml-ops.org/content/end-to-end-ml-workflow

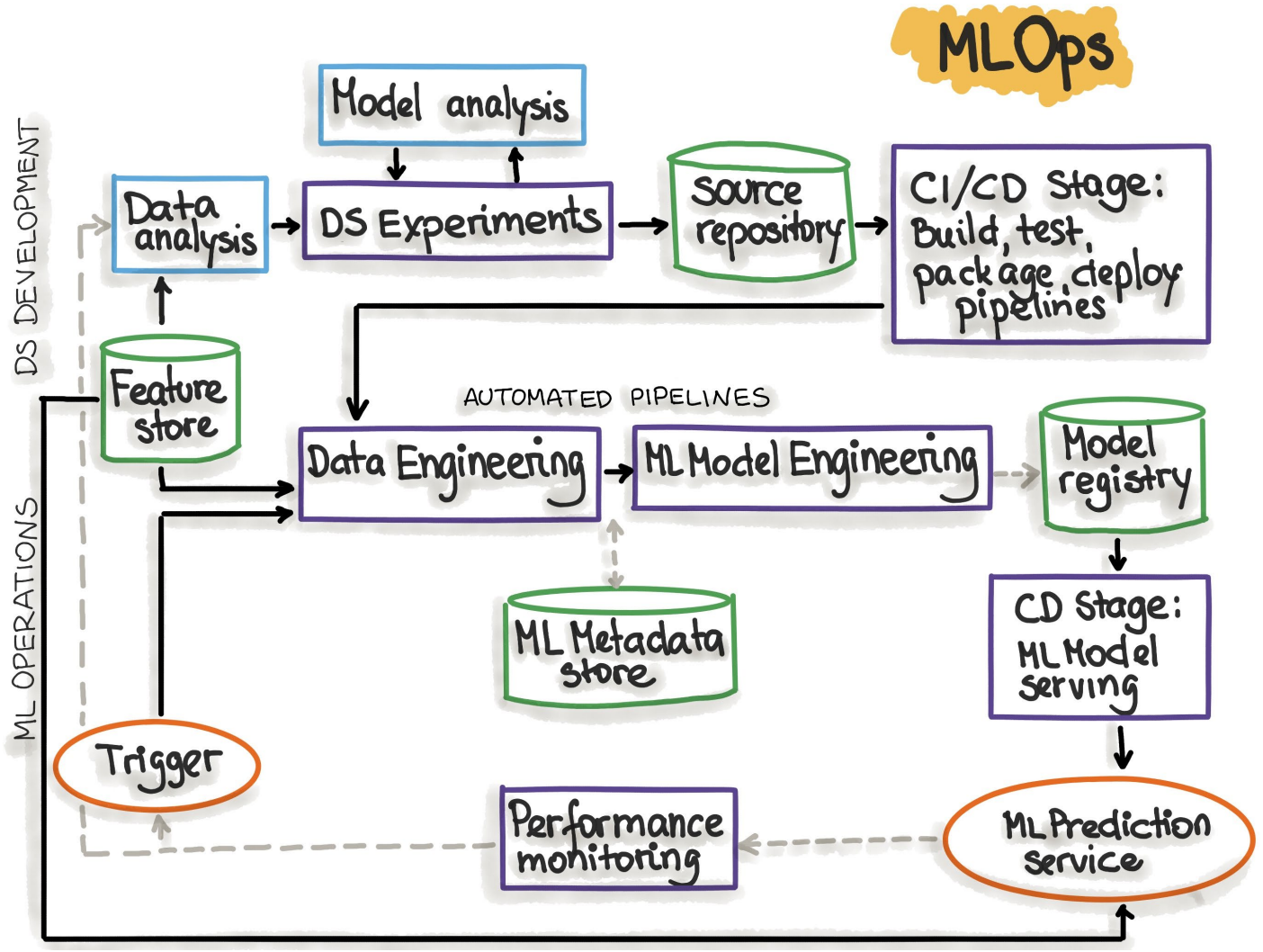
Intellerts

Core principles of MLOps philosophy



Main components

- Feature stores
- Data and code versioning
- Management of experiments and output artifacts
- Monitoring infrastructure
- Deployment and serving



Source: https://ml-ops.org/content/mlops-principles

Intellerts

Monitoring

- Ability to track and identify errors and root causes
- Required at both system and infrastructure levels
- Must be actionable
- Collect – Monitor – Respond – Analyze - Evaluate
- Enabled by tools like Grafana, Prometheus, Telegraf, Icinga...
- Can collect both predefined sets of metrics or custom-defined ones

Intellerts

Automation

- Aims to reduce human error by automating setup, installation, or system run steps
- Relevant for ML-driven systems:
  - Data collection automation (Airflow, Airbytes, ...)
  - Automated data profiling and reporting (Great Expectations)
  - Automated feature extraction and selection
  - Automated experiment runs (Airflow, Prefect, KubeFlow, ...)
  - Automated evaluation
  - Automated deployment (GitHub Actions, Kubernetes, ...)

Continuous updates

- Drift detection
  - Drift in data distribution
  - Drift in labels distribution
  - Drift in data quality
  - Drift in reality
- Model retraining to reflect these changes
  - Data sampling/resampling issues
  - Data window selection
  - Robustness to noise

Intellerts

Reproducibility

- Ability to re-run previous experiments in a deterministic manner
- Produce the same outputs and results as in previous runs
- Several of the issues addressed:
  - Random seeds (are we properly handling randomness?)
  - Code versioning (are we using the exact same code?)
  - Data versioning (are we using exactly the same data?)
  - Parameter settings (are we using the same experimental settings or hyperparameters?)
  - Library versions (are we using exactly the same implementations?)
- Highly important for automated ML (AutoML) settings

Intellerts

Deployment

- Managing which model is deployed in production and tracking its ongoing performance
- Load balancing between multiple instances
- Continuous process reflecting recent changes
- Needs careful evaluation before rolling-out completely (canary deployments)
- May require additional optimizations, like batch inference support to minimize the number of actual calls

Intellerts

dhealthIQ

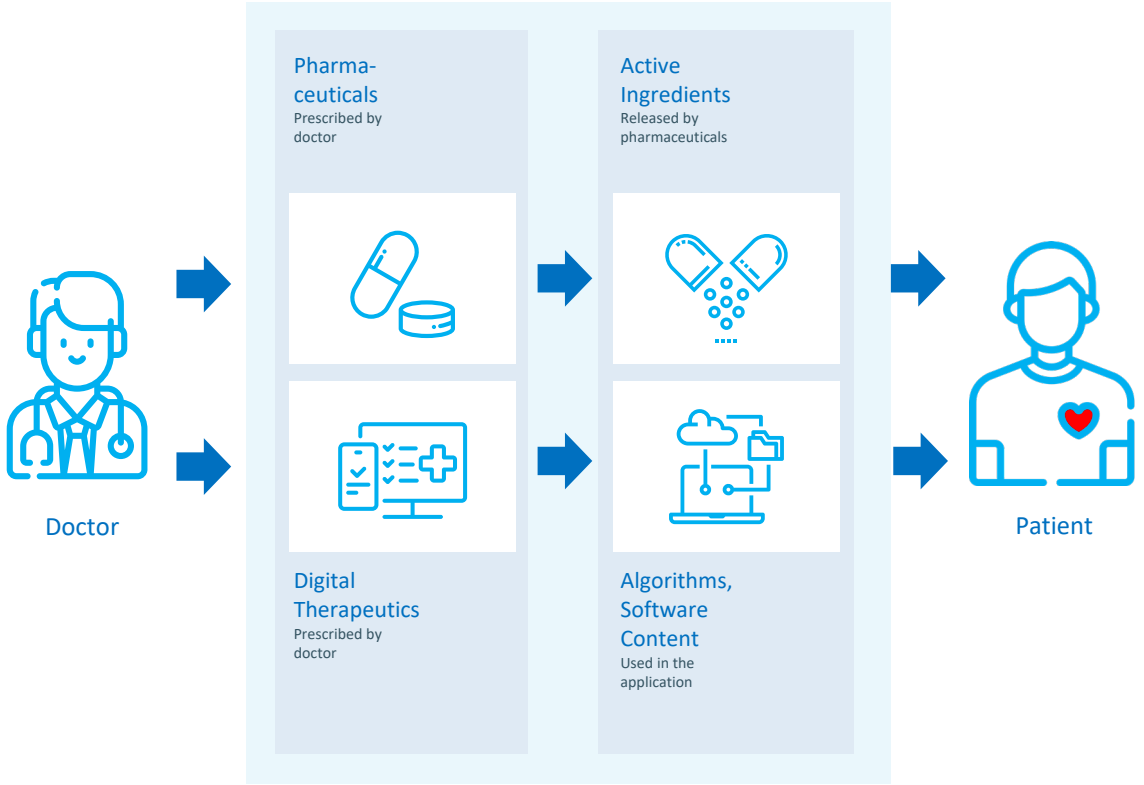
enabling personalized therapeutics

Case study

Intellerts

About dHealthIQ

- Digital Therapeutic company
- Enabling Healthcare providers to treat and manage Diabetes 2, Cardiometabolic, and COVID diseases in an advanced way
- Based on evidence-based and personalized interventions Directly to the patient



dhealthIQ
enabling personalised therapeutics

Intellerts

dHealthIQ

COVID Safe

WORKS WITH fitbit watches

Data Collection

Heart rate

Blood Pressure Accelerator

ECG

Pulse Entropy Respiration

Rate Oximetry

Temperature

Statistical and AI analysis

Health status forecasting

NEWS2 Score for COVID

Advanced correlations

Physical activity intensity/frequency

Heart rate recovery time

Heart rate variability

Digital Alerting Mechanism

Prediction of Health Status

Health Status Change

Deteriorating Patient Condition

Prediction of Required Hospitalization

Actions

Remote Consultation

Review treatment plan

Review medication

Hospitalization

Intellerts

COVID-19 DIGITAL THERAPY HEALTH REPORT

COVID Safe

Duration: 27-04-2022 - 11-05-2022

Start date **27 04 2022** End date **16 05 2022** | Progress **3/3 week**

NR 271 COVID

testing.dhealth.user+271@gmail.com
+37060022555

Gender	Female	Height	190 cm
Birth date	28-04-2000	Weight	70 kg
Age	22	Smoking	

COVID-19 information

Has Covid-19 infection been confirmed?	**Yes**	Is the patient self-isolating?	**Yes**
The date when covid-19 was diagnosed	**27-04-2022**	Date when symptoms were noticed	**26-04-2022**
Does patient experience symptoms of the disease?	**Yes**	Is the patient in the quarantine (returned from abroad)?	**Yes**

Information about a woman

Is the patient currently pregnant?	**Yes**	Is there a postpartum period at the moment?	**Yes**
Have there been any miscarriages?	**Yes**		

Diagnoses

U07.1	COVID-19, virus detected	27-04-2022

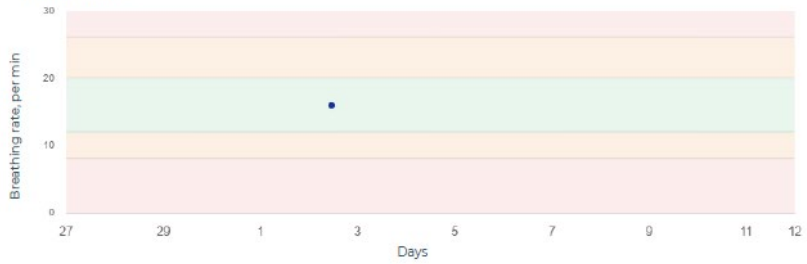
COVID-19 DIGITAL THERAPY HEALTH REPORT

Duration: 27-04-2022 - 11-05-2022

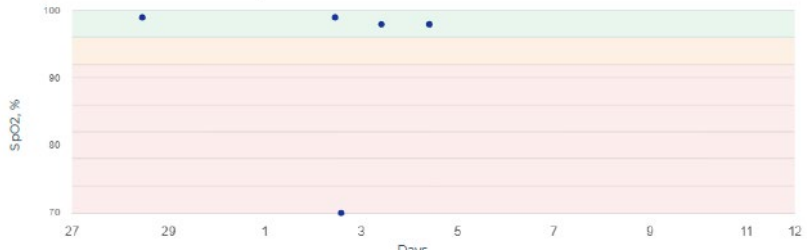
COVID Safe

Health data

Breathing rate



Oxygen saturation of the blood SpO2



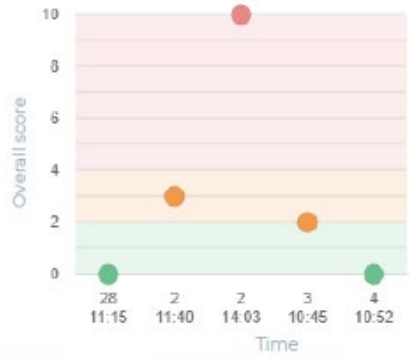
COVID-19 DIGITAL THERAPY HEALTH REPORT

Duration: 27-04-2022 - 11-05-2022

COVID Safe

Patient monitoring

Overall score



Assessed condition	Mild	Moderate	Severe	Moderate	Mild
Assessed condition according to the protocol	Mild	Moderate	Severe	Severe	Mild

Explanation of scores

SYS	120.0	110.0	200.0	128.0	120.0
SPO2	99.0	99.0	70.0	98.0	98.0
Heart rate	73.0	75.0	95.0	78.0	70.0
Temperature	36.5	36.7	38.0		36.4
SYS decrease		-10	90	-72	-8
Fever					
Respiration rate		16	16		
Prolonged fever					

Questionnaire answers

Do you feel chest pain?	No	Yes	Yes	No	No
Do you feel chest pain during moderate exercise?	-	Yes	Yes	-	-
Do you feel chest pain at rest / during minimal physical activity?	-	Yes	Yes	-	-
Do you feel shortness of breath?	No	No	Yes	No	No
Do you feel shortness of breath during light exercise?	-	-	Yes	-	-
Do you feel shortness of breath at rest?	-	-	Yes	-	-
Do you cough up yellowish / greenish / pink sputum?	No	No	Yes	No	No

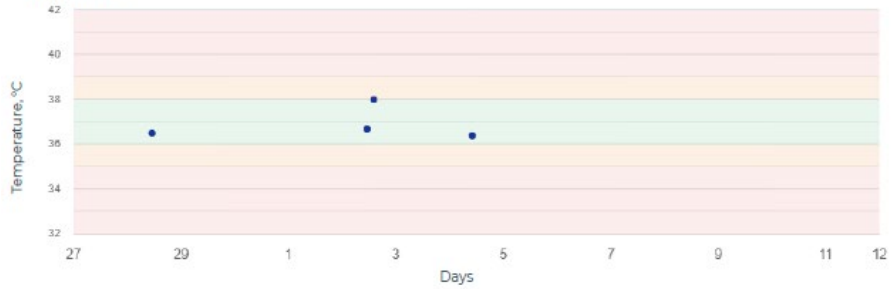
COVID-19 DIGITAL THERAPY HEALTH REPORT

Duration: 27-04-2022 - 11-05-2022

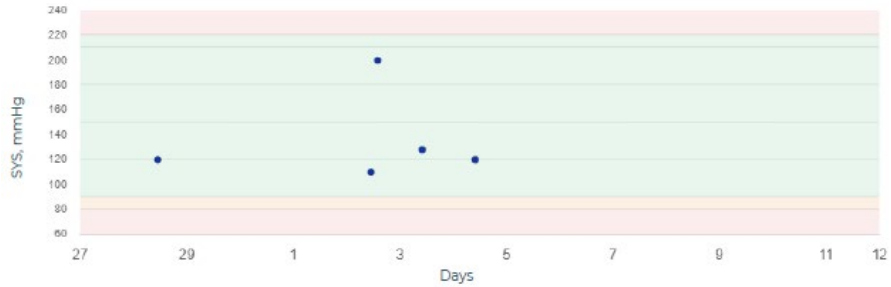
COVID Safe

Health data

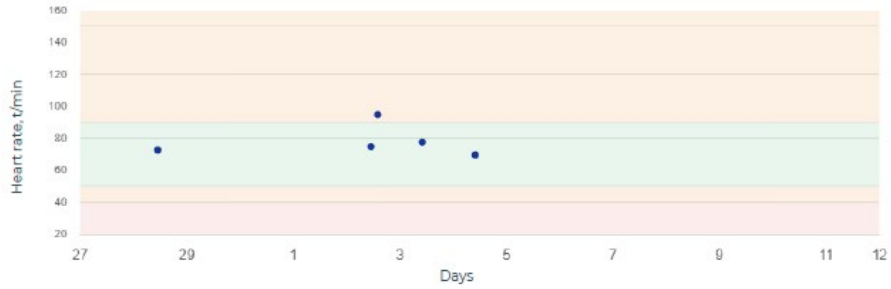
Temperature



Systolic blood pressure



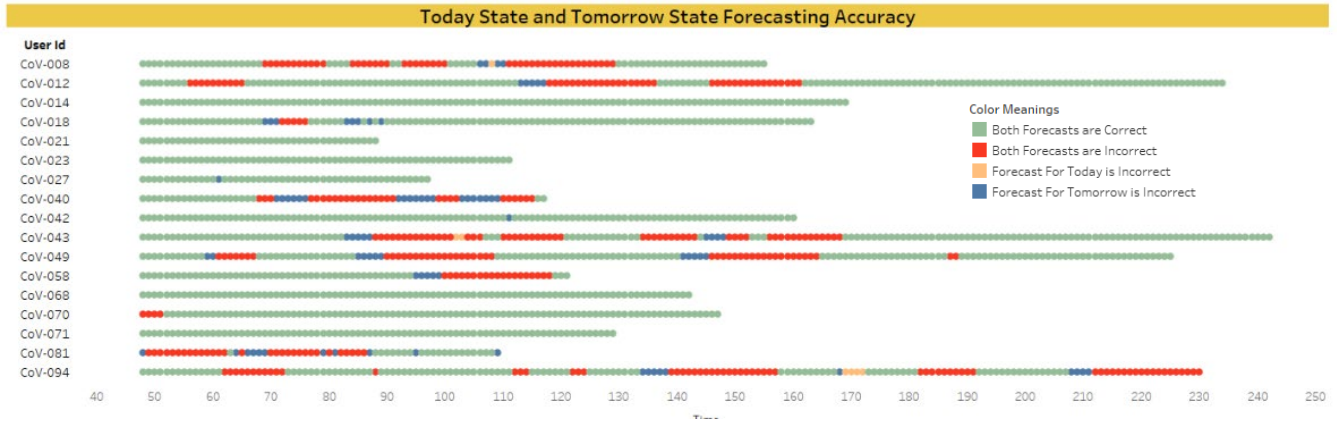
Heart rate

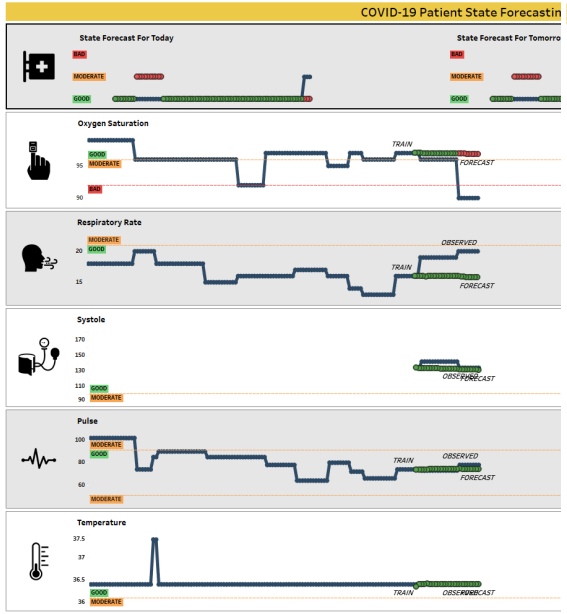


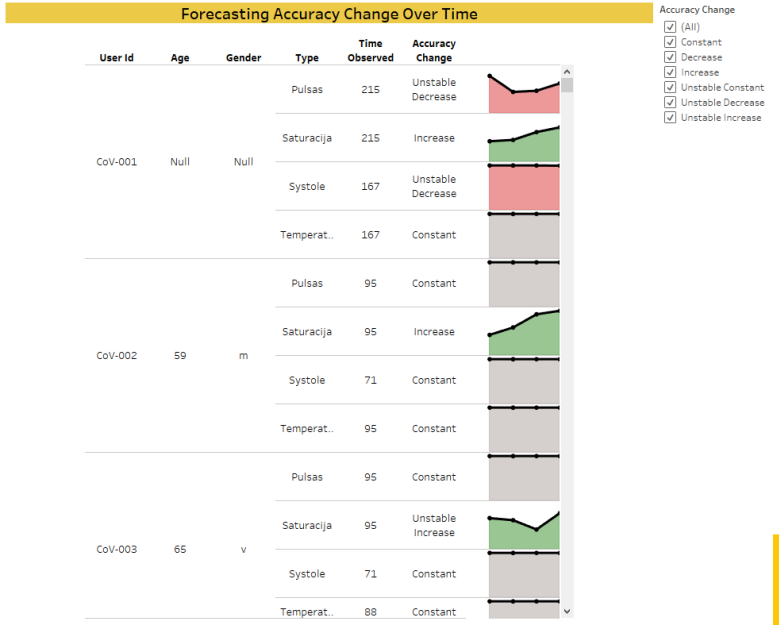
Case study: dHealthIQ

COVID Safe

- Applies time-series analytics and forecasting to measure and forecast the future state of the patient
- Tremendously important during the pandemic times (COVID-19)
- Could be used as a monitoring and decision support system by doctors

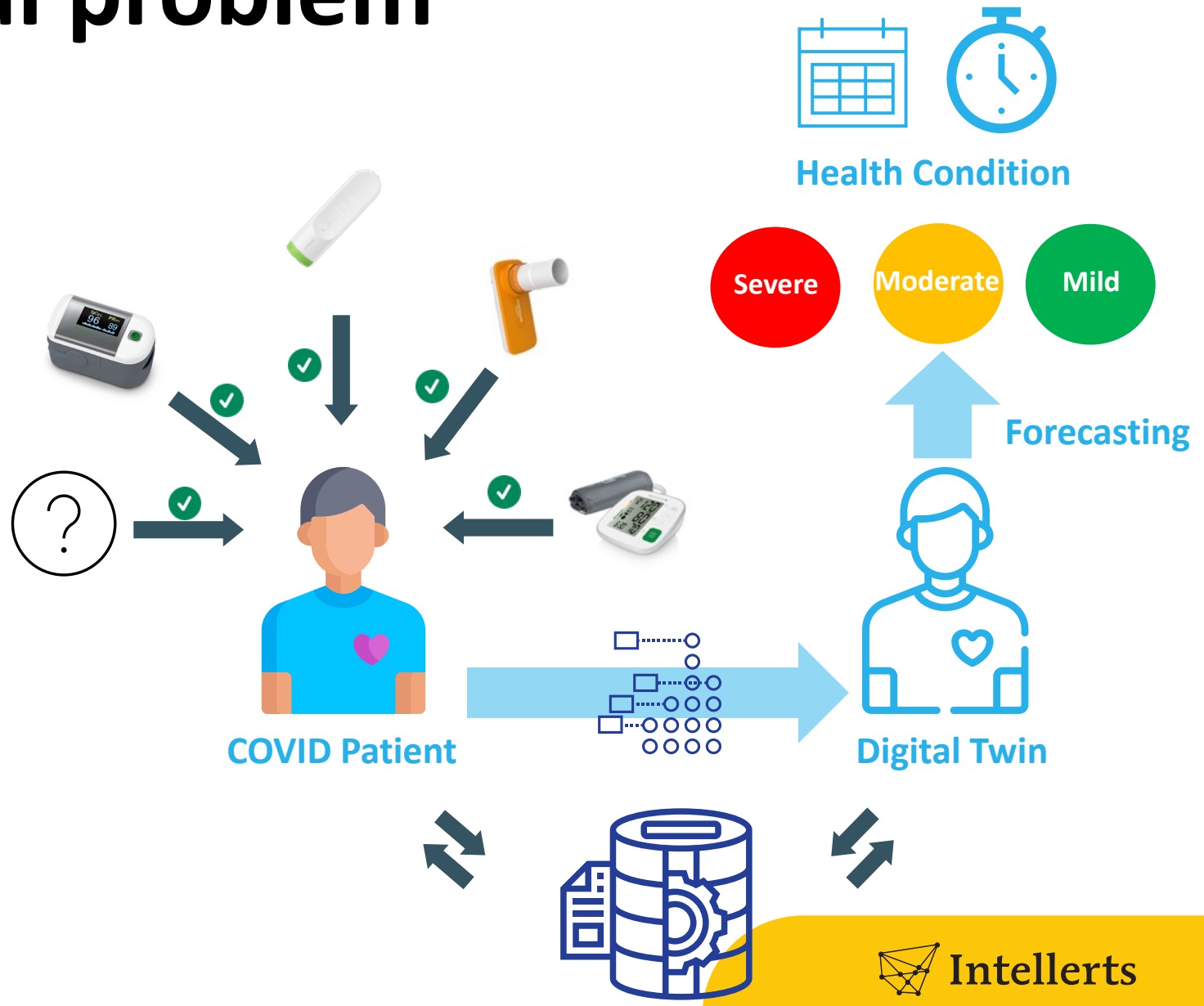






Multidimensional problem

- Multiple time series for input
  - Pulse, respiratory rate, systolic, temperature, $SpO_{2, \ldots}$
- Multiple models: ARIMA, TBATS, STLM, Theta, STL, ...
  - Additional models are also continuously tested
- User dimension
- Time dimension



Monitoring data collection

- Ingest data from multiple external providers
- Provider API / JSON schema changes
- Data quality: near-real-time visual insights (Grafana)
- Drift detection: visual (Grafana, Tableau)
- Collection issues: logging (database, Elastic Search)

- **Missing data:**
  - User does not wear a wearable -> history may be empty for some period
  - The user wears a wearable, but it is not synchronized or connected to the system
  - The system fails to fetch the data from the provider API
  - Discrepancies in data/date formats, etc.

Intellerts

Model training

- Handled using Airflow framework which enables flexibility in managing and orchestrating required calculations, provides required scalability and distribution between multiple nodes
- Continuous retraining (refitting): an ensemble of univariate time series models
- R - stable and fast implementations compared to less established and slower Python ones
- Strict requirements imposed by the medical domain
  - Interpretability
  - Performing calculations on time

Intellerts

Model training

- Multiple models per user on each iteration (*type* x *model* x *user*)
- Research and experimentation must support multiple configurations:
  - Like: training performed in two hours, while predictions are calculated hourly
  - Different window settings
  - Currently: hourly basis retraining and recalculation
  - Depends on the frequency of data collection, how often will the users send the data
- Requirement to exclude some models if they do not perform well or take too long to train (due to hyperparameter selection steps, like ARIMA or TBATS)
  - managed through configuration in Airflow

Intellerts

Storage and monitoring

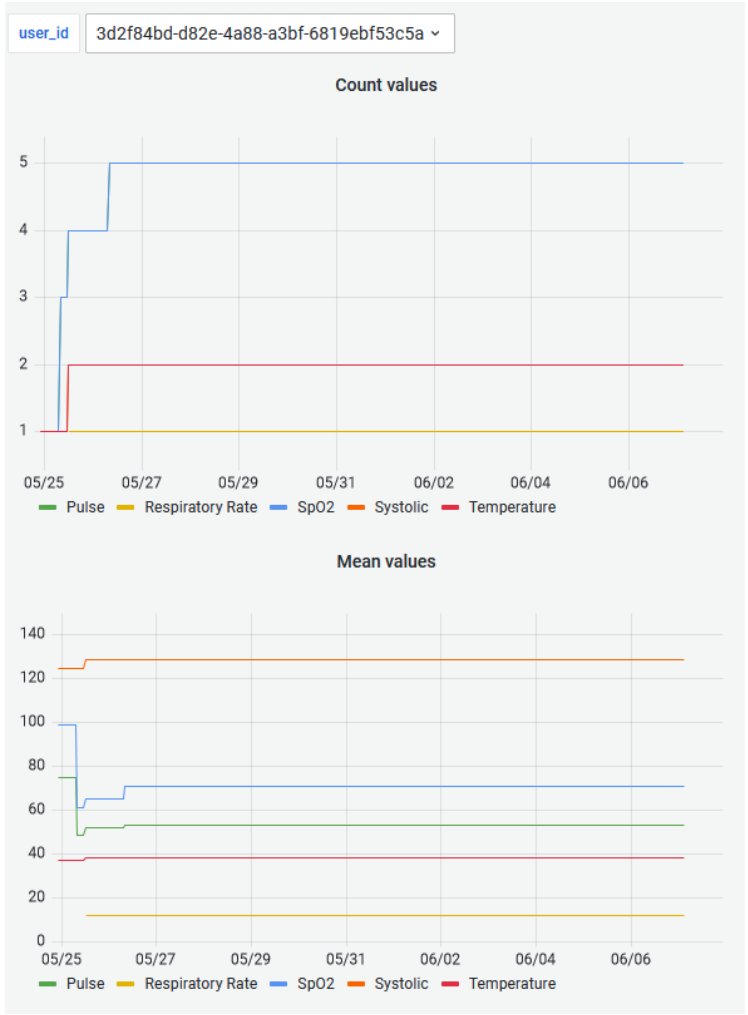
- All the models must be stored for logging or back testing purposes
- Internal Model Registry-like structure
  - Models are serialized into JSON with jsonlite and stored as text fields
  - Can be easily deserialized when required
  - Old models might be dropped after a specific time frame to save storage space - will be adjusted in the future after gathering more data about the system performance
- Data can also be stored as JSON files

Intellerts

Monitoring dashboard example

Logging I

- Log everything that we can:
  - forecasts,
  - performance (actual and back testing) metrics
  - core descriptive stats for each time series :
    - number of entries
    - missing values
    - fact data and forecasting intervals start/end
    - actual min/max dates
    - min/max/mean/std
    - mean time gaps...
- Profiling information
  - Training procedure for some models may take longer than expected
  - Prolonged hyperparameter optimization
- Actual data used
- Training exceptions, errors, inconsistencies



Intellerts

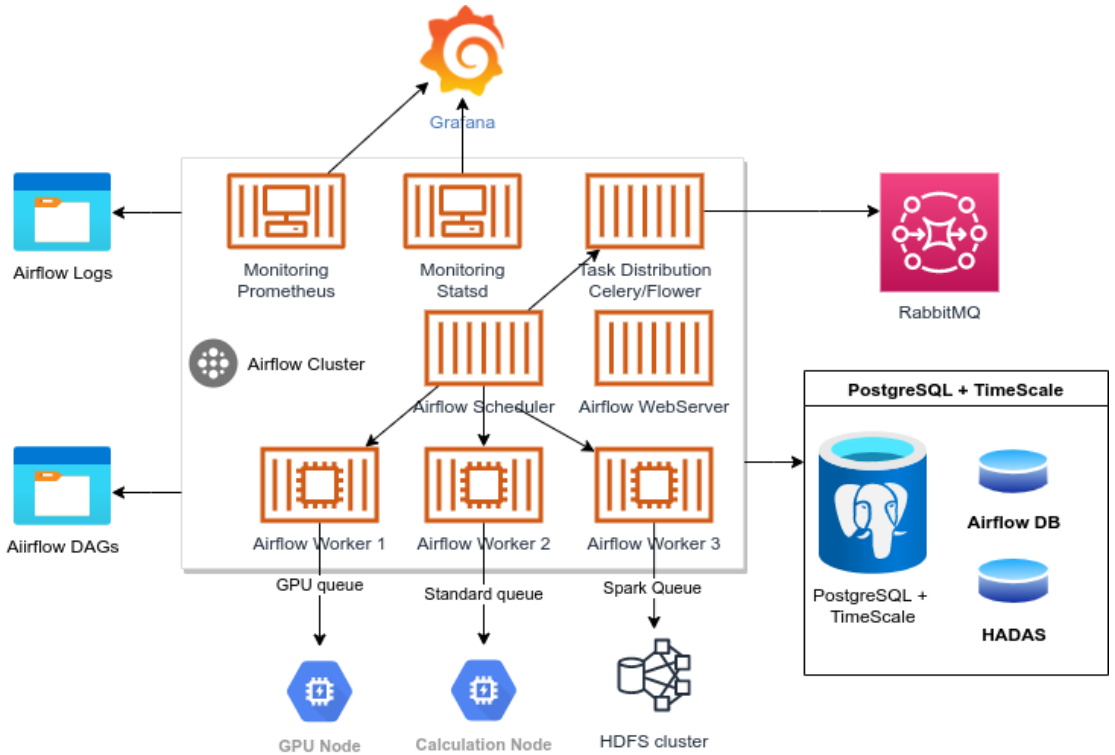
Logging II

- Custom logger implementation based on R logger libraries
- Custom implementations vs. more generic facilities (MLFlow)
  - MLFlow is more suitable for managing a smaller amount of models
- Required to log training exceptions, errors, and inconsistencies for forecasting adjustment
  - Failure to perform fitting procedure is not rare - issues with time series consistency, length, and properties result in failure to perform fitting models successfully
  - Carefully log cases when no fitting could be performed due to missing data
  - Adjust ensemble aggregation to exclude missing members if their fitting procedure failed
  - Alternatively, may use a previous successful instance of the particular model at the same dimension
  - Write to Airflow log and use tools like Elastic Search to extract issues
  - May write exceptions to specialized log as well for easier debugging

Intellerts

CI/CD

- Automated processing pipelines (Airflow)
- System management and updates (Ansible)
- Cloud infrastructure (Terraform, Cloud)
- Virtual environments (packrat, venv)
- Forecasting endpoint (plumber)
  - Deploy the most recent forecasting model versions



Intellerts

R based MLOps ecosystem challenges

- Limitations of the native workflow orchestration/management tools
  - **targets** do not provide runs execution history and are less mature compared to Airflow or Prefect
- Containerization
  - Limited packaging abilities for MLFlow model store – limited to single wrapper framework (**crates**), might require additional efforts and metaprogramming skills
  - System dependency management (Linux native libraries)
- No native libraries for interfacing with other MLOps ecosystem tools, like Grafana, Feast, ClearML
- Performance issues:
  - while some modeling libraries are fast, R language is relatively slow and less suitable for production-ready systems

Intellerts

Infrastructure requirements

- Processing a large number of users is costly

- This is particularly important for the R&D phase
  - Testing multiple techniques, implementations
  - Testing different settings (windows, frequency of calculations)
  - This is difficult and very costly to run in the cloud

- Production level system includes only final developments
  - Hence, the cloud is an option in this case

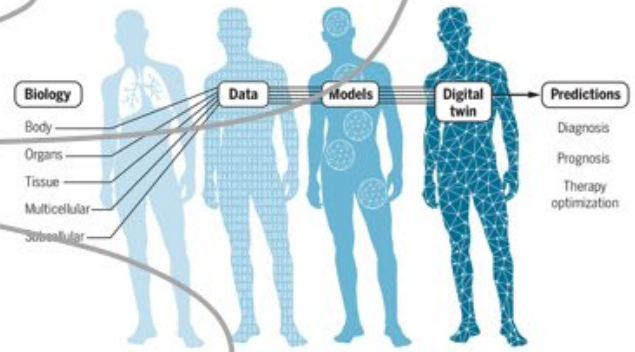
Intellerts

Enabling Personalized Therapeutics with HP Z Data Science Solutions

dhealthIQ
enabling personalized therapeutics

COVID Safe

Edge AI Lab | Data Fabric | MLOps Factory | Data Share Lab | Data Science Lab



Defining a Personal Baseline of Health

High-Definition Prevention

High Precision Treatment

Billions of High-Resolution People

Intellerts

Infrastructure requirements

- It is beneficial to invest in powerful HP Workstations for experimentation
  - HP Z Workstation contains 80 cores and 2 Nvidia RTX 8000 GPUs
  - Is applicable for both large-scale and GPU-based processing experimental workloads
  - Therefore, it also enables comparative testing of SOTA deep-learning based time series analysis and forecasting approaches, e.g., based on RNN or transformers

Intellerts

HP Z8

Max Out on Performance with our Most Powerful Workstation

QuickSpecs

Datasheet

HP Z8 G4 Workstation Desktop PC

When you work with the most complex simulations, video editing, or high-end VFX, only our flagship Z workstation and most powerful PC will do.

Unthrottled, Real-Time Performance

Forget clicking and waiting. With up to a staggering 56 cores, 1.5 TB high-speed memory, and certification for hardcore apps, you get optimal performance from your Z8 workstation PC every time.

Up to 56 cores with 2x Intel® Xeon® Scalable CPUs

Run demanding professional apps and push your multi-threaded applications to the limit for fast iteration cycles with up to 56 cores. Avoid system crashes in the middle of your workflow with Error Correction Code (ECC) memory that detects and corrects soft errors in the memory system on the fly.

ISV Certified

Work with confidence knowing your desktop is certified with leading software applications to ensure peak performance even with complex projects

Up to 56 TB

Say goodbye to bottlenecks and get optimal performance in storage-bound applications.

Up to 2x NVIDIA RTX™ A6000 or 1x AMD Radeon™ Pro W6800 GPUs

Get serious about performance and reach peak productivity with your choice of professional graphics.

Up to 1.5 TB DDR4-2933 ECC SDRAM

Get the ultimate productivity boost when working with massive and complex data sets with Intel® Optane™ DC Persistent Memory.

Up to 1700W

Multiple PSU options, including the 1700W PSU, enable increasing need for GPU, even with the highest power CPUs and rich configurations.

Conclusions

- MLOps is a very prominent and emerging area
- Takes on issues that are well known for software engineers, but less familiar to data scientists (monitoring, automation, deployment, scalability, testability…)
- Yet, it also considers reproducibility and continuous model improvements
- Our case study emphasizes the necessity of monitoring, logging, continuous experimenting, and R&D activities to properly implement a modern data-driven analytics system for health monitoring
- **While R&D-related MLOps tasks might be very resource-intensive, obtaining on-premises powerful HP workstation for experimentation is a paying-off decision**

Intellerts

Thank you!

Questions?