



DUOPATH.AI

A collaboration platform to bridge the gap between
Data Scientists and Domain Experts

Problem

Increasing demand for data-driven decision-making leads to more collaboration between **Data Scientists** and **Domain Experts** in all industries.

However, there is **no effective collaboration workspace** to help them truly focus on productively mining analytical insights.



Collaboration Conundrum

Knowledge Gap

- **Bi-directional education sessions**
 - Time-consuming, unsystematic, ineffective
- **Lack of transparency** in modelling
 - Hard to comprehend, resulting in misunderstanding and distrust. ¹

Communication Inefficiency

- **Fragmented information** from multiple channels
 - Repeated, chaotic Q&As coming from all tools (Emails/Slack, Drive/Office, GitHub/Ipynb)
 - Requires info tracking and synthetization

Business Needs

- Non-trivial work for data scientists to **translate analytics results into slides**. ²

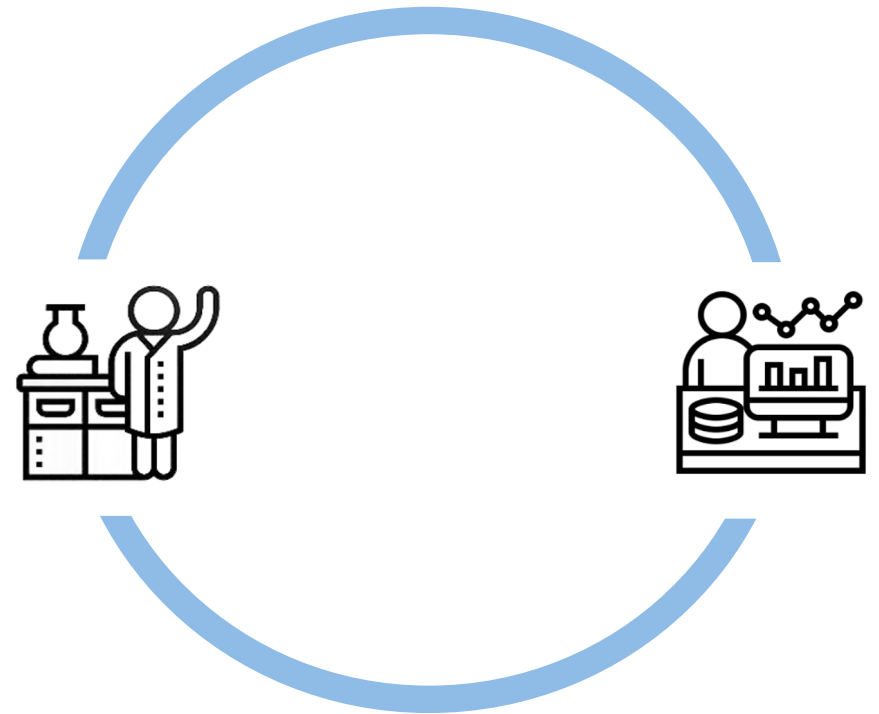
1. Yaoli Mao, Dakuo Wang, et al. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 237 (December 2019), 23 pages. <https://doi.org/10.1145/3361118>

2. David Piorkowski, Soya Park, et al. 2021. How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 131 (April 2021), 25 pages. <https://doi.org/10.1145/3449205>



Solutions

DuoPath.ai is a data science collaboration platform that helps data practitioners and domain experts work together smoothly and seamlessly, leading to **more productive business/scientific understanding and outcomes.**



Features

CORE



Free flow of feedback

- Comment & Tag anywhere on a code piece & slides
- Q&As publicly available via forum



Notebook → PPT

- Auto-transition from templates of code pipeline into slides
- Simultaneous update of plots & charts



Knowledge Warehouse

- Built-in concept-chips with explanation on DS with use cases
- A community for domain knowledge sharing

TECH



Modularization

- Modularized flow of code in ETL, visualization and modelling
- Reusable code chips



AutoML + Cloud

- Free DS from repeated setup work on training & tuning
- Dashboard visualization



Interactive mini apps of models

- Easy for DS to demo models with UI
- Easy for Domain experts to trial-&-error



DUOPATH.AI

MVP

Project

OSIC Pulmonary Fibrosis Progression

Database

Files

Brainstroming space

Forum

Notebook

Slides

Module dashboard

Notebook

OSIC.ipynb

File Edit Insert Tools

Create slides Share

Comments

Code chips

plot

scatter plot

line plot

pair plot

residual plot

Drag to notebook

Import data

Select from database

EDA

```
plt.figure(figsize=(15,10))
sns.lineplot(x='Weeks',y='FVC',data=week,hue='SmokingStatus',style='SmokingStatus',markers=['*','o','^'])
plt.grid()
plt.title('Weekly Distribution OF FVC')
plt.show()
```

Edwin

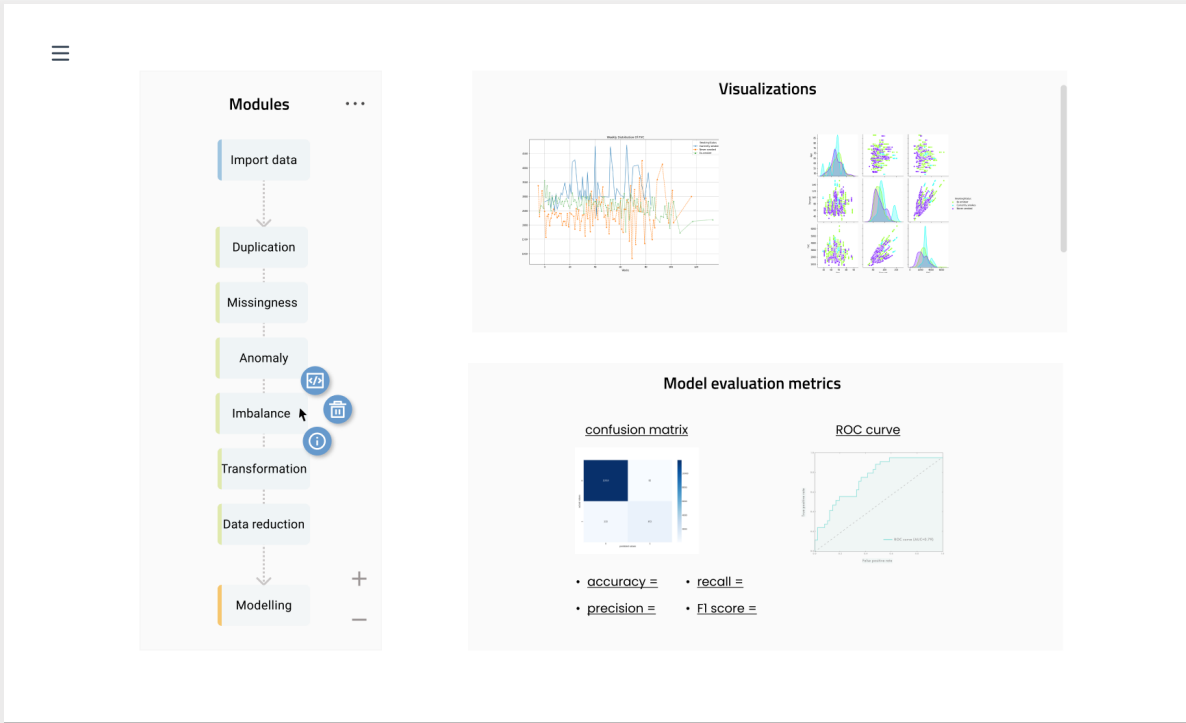
Modelling

AutoML

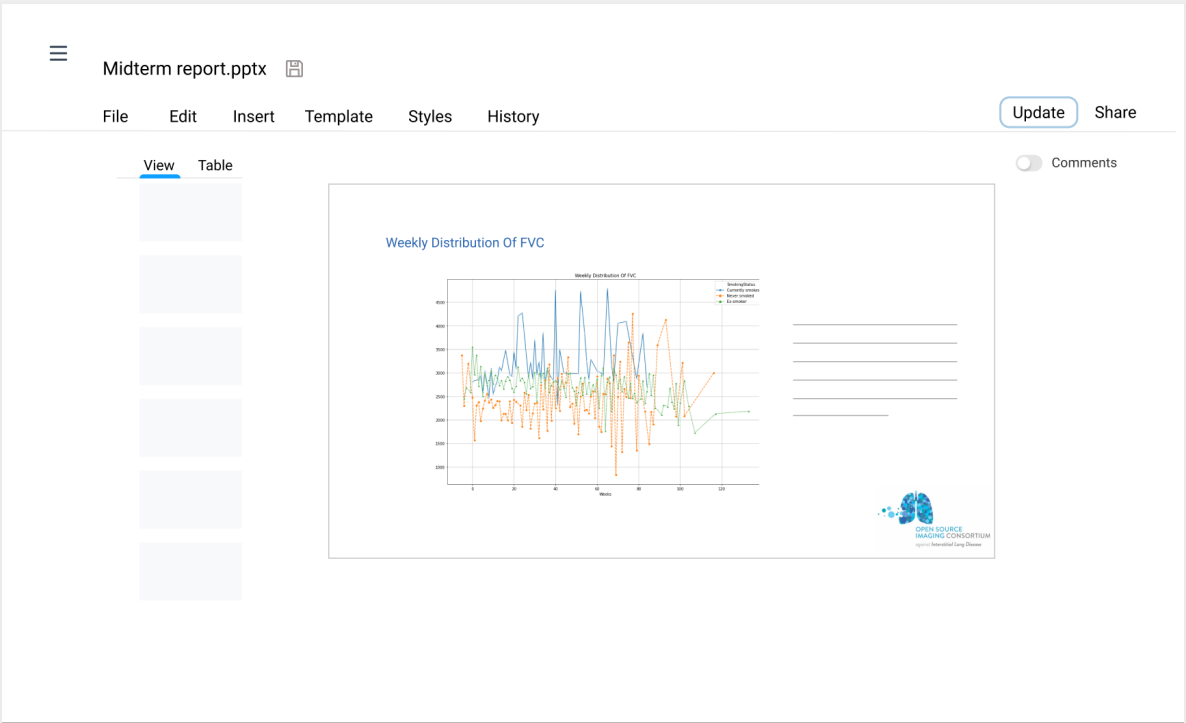
Evaluation metrics

MVP

Module dashboard



Slides



Market

Global revenue from Big data & Business Analytics:
\$168.8 bn in 2018 to
\$215.7 bn estimated in 2021

36.8 million
results for
searching 'data' on
LinkedIn, **14.8**
million in the US

Corporate Data Functions

High demand in digital transformation for all industry

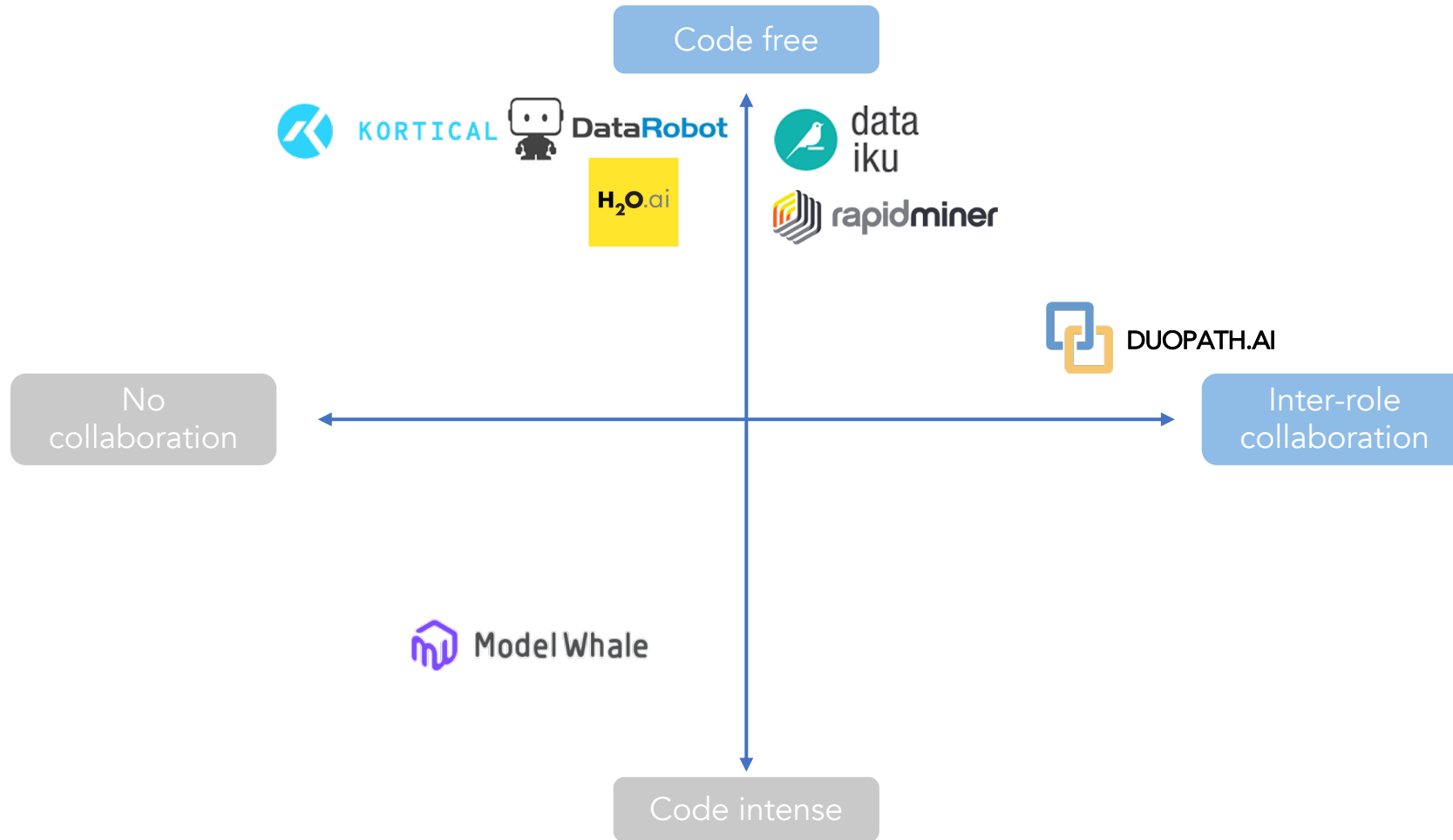
Data-driven Scientific Research

Bio-signal processing, Medical Imaging,
Logistics / Network Analysis, Robotics

Datathons / Student Projects

University-Industry collaboration
in capstone projects / datathons

Competitors



Business Model

Freemium Version

- Free data wrangling and modelling services with ads and up to **5GB of RAM** and **15GB of Storage** of data and plots.
- Extra will be charged by the timely price of cloud resource usage

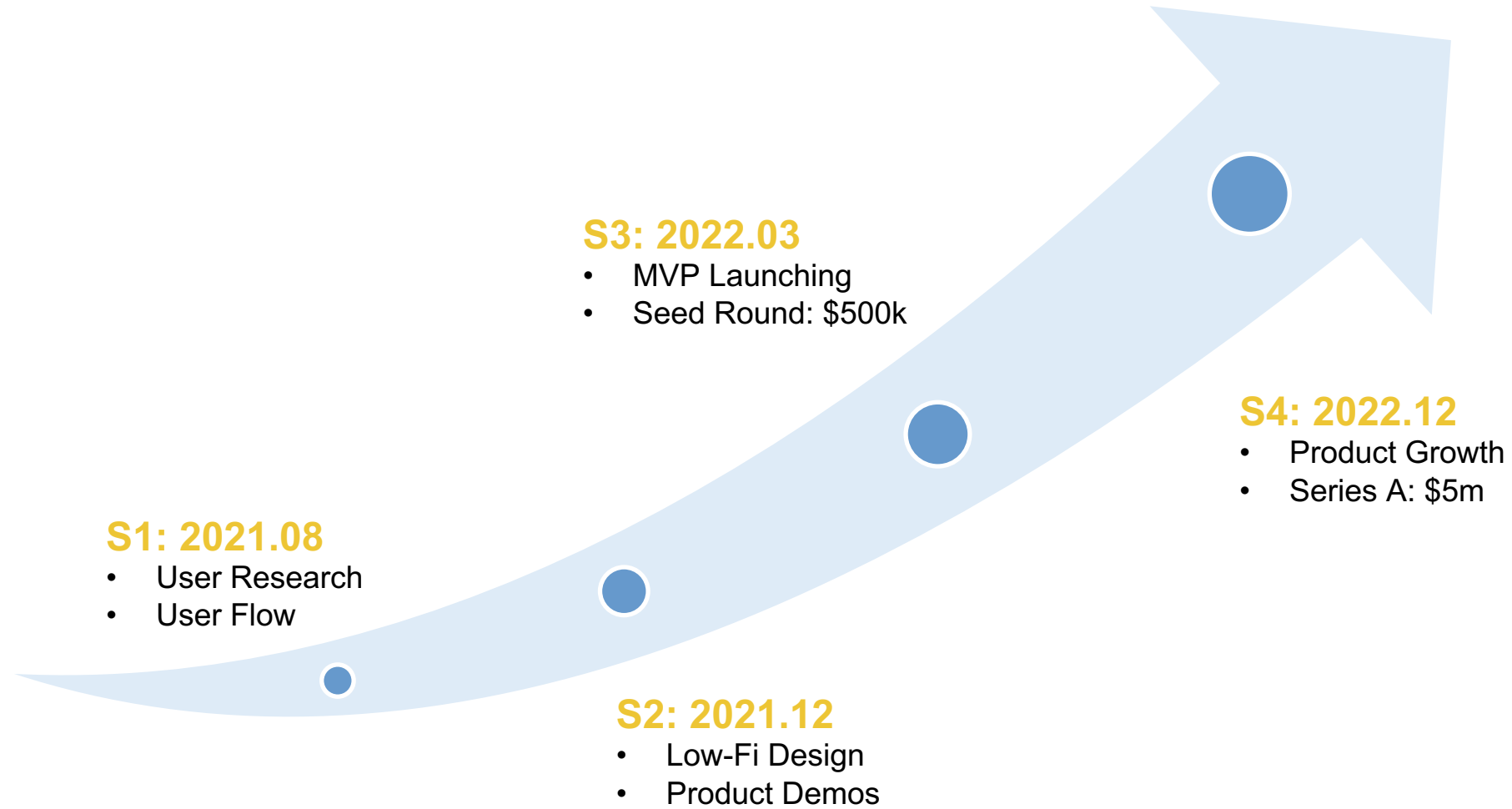
Subscription Version

- **\$4.99 ads-free monthly plan** with **20 GB of RAM** and **50GB of storage**.
- Extra will be charged by the timely price of cloud resource usage

Premium Version

- **\$89 annual fee** on research/industry topic-specific tools and dashboards on the following fields:
 - Healthcare Image Processing
 - Free text editing & embedding
 - Financial time series

Timeline



Team



陈申 Sean

Product Management, Data, Business

- Product Analytics @Google
DS @Microsoft, CVS Health
- MIT, Operations Research & Sloan|
UCL, Statistical Science



李尔晴 Anthea

Operations, Market Research

- Product & DS @ApplySquare,
Series-A EdTech startup
- UCL, Statistics, Economics, Finance



俞润隆 Edwin

ML & AI, MLOps Development

- Full-Stack & ML @Tesla; RA @The Alan
Turing Institute, Oxford Protein
Informatics Groups;
- Oxford, Statistics | UCL, Econ & Stats



唐夕尧 Xiyao

UI/UX Design, Product Development

- UX Specialist @ UHN, Canada;
Digital Marketing VP @ MIT CEO
- U of T, Info Management | Northeastern,
Digital Media | ArtCenter, Graphic Design



Research Reference

1. Yaoli Mao, Dakuo Wang, Michael Muller, Kush R. Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How Data Scientists Work Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 237 (December 2019), 23 pages. <https://doi.org/10.1145/3361118>
2. David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 131 (April 2021), 25 pages. <https://doi.org/10.1145/3449205>
3. Dakuo Wang, Josh Andres, Justin Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan*. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3411764.3445526>
4. Langer, T.; Meisen, T. System Design to Utilize Domain Expertise for Visual Exploratory Data Analysis. *Information* 2021, 12, 140. <https://doi.org/10.3390/info12040140>
5. Rahman P, Nandi A, Hebert C. Amplifying Domain Expertise in Clinical Data Pipelines. *JMIR Med Inform* 2020;8(11):e19612. URL: <https://medinform.jmir.org/2020/11/e19612>. doi: 10.2196/19612. PMID: 33151150