Building A Scalable Data Annotation Strategy

September 6, 2022

SIGMA·AI DATA DONE RIGHT

Outline

- Sigma.AI Brief Introduction
- Why is a Data Annotation Strategy Essential?
- How Much Training Data Do You Need to Build Out an Effective Strategy?
- What Are the Biggest Challenges for Data-Centric AI Project Teams?
- How Do You Know It’s Time to Outsource Data Annotation?
- Building the Proper Data Annotation Strategy
- Positioning your Project for Success

Sigma.AI is part of the Sigma Group, a consortium of AI service organizations covering data services to AI technology development and end-end AI solutions

Who is Sigma.AI :
Sigma is a global training data collection, preparation, and annotation services company.

What Sigma does:
Sigma solves one of the biggest problems in AI; high quality, accurate training data.

How Sigma does it:
Sigma helps you build the future of AI applications by providing the highest quality training data, the way you want it, when you need it, at the right cost.

Why:
Sigma wants to contribute to a better world by maximizing the performance of AI algorithms with large volumes of high quality training data.

Sigma is a **global data services company founded in 2008 with offices in US, Spain and UK.**

99.99%

Up to 99.99% accuracy available

98% accuracy guaranteed

4 out of 5

Of the global technology titans are clients

Trusted

12

12 Point Code of Ethics

Trustworthy

760k+

Audio hours annotated to date

1M+

Words analyzed per day

200M+

Video frames analyzed per day

250+

Languages and dialects

We excel at NLP

25,000+

Vetted, trained, linguists, annotators, and domain specialists, in 5 continents

We can Scale quickly

900,000+

Pool of annotation & collection professionals

SIGMA·AI DATA DONE RIGHT

Sigma **core is data services** where we execute projects in data generation, collection, preparation and annotation **across all media types (e.g. audio, text, image, video)**

GENERATION

An example…
Synthetically generate product opinions in order to train classification models

COLLECTION

An example…
Record conversations of people speaking about cars in Swedish to train or evaluate voice recognition models

PREPARATION

An example…
Help preselect images to be annotated in order to train a robot (enhance variability / avoid bias)

ANNOTATION

An example…
Transcribe audios tagging speakers, topics, noises, and sentiments for decision making in customer service

Sigma's motto is "Data done Right" but that is only a part of it: Sigma seeks to do the right thing not only in data treatment, but also towards our customers, partners and employees

QUALITY

No crowdsourcing

Senior team +30 years experience in training data collection and annotation and solving complex challenges

Adaptation of annotation guidelines to address different needs

Combine automation and human annotation to optimize workflows in order to achieve speed while ensuring the highest quality

SECURITY & PRIVACY

Possibility to use of **highly secure facilities** to avoid information leaks

100% **compliance with GDPR and ISO27001 certified**

Large EU presence – no personal data needs to leave the Union

Strict **confidentiality of clients and their data**

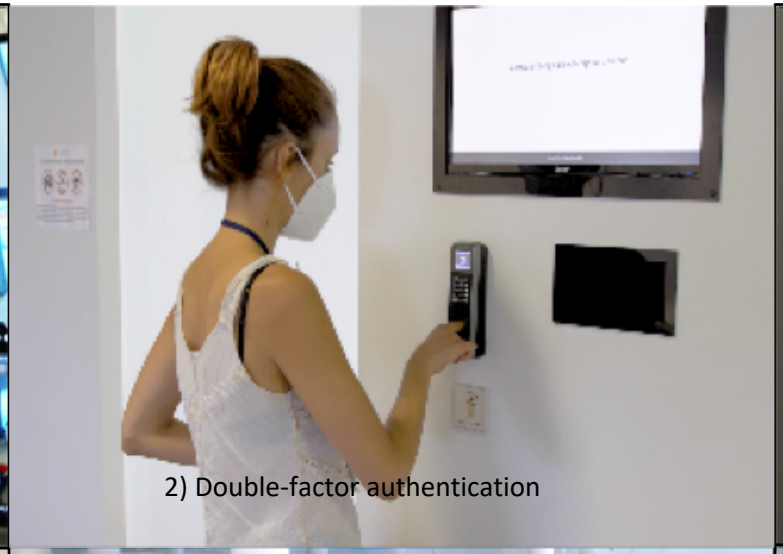
FLEXIBILITY

Sigma **tailors processes and tools for each customer or project** and work hand in hand to understand what success looks like

Sigma **accompanies customers in designing and implementing AI solutions** through Sigma's data enhancement services and consulting practice

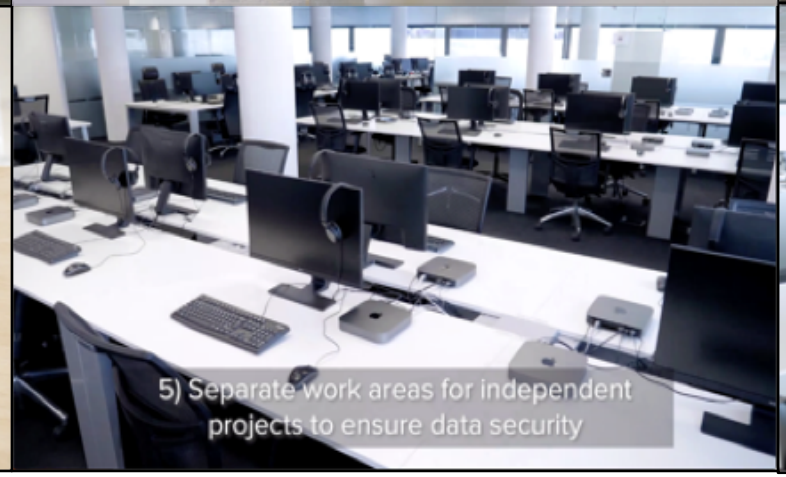
DATA SERVICES - Secure Facilities

1) Video surveillance cameras.


2) Double-factor authentication

3) Security Signs

4) Polarized monitor filters.


5) Separate work areas for independent projects to ensure data security

6) Training rooms.

Why is a Data Annotation Strategy Essential?

SIGMA·AI DATA DONE RIGHT

Two AI approaches to reach the desired performance level

AI = Code + Data

20%**

Model-Centric AI

How can you change the code to improve performance

80%**

Data-Centric AI*

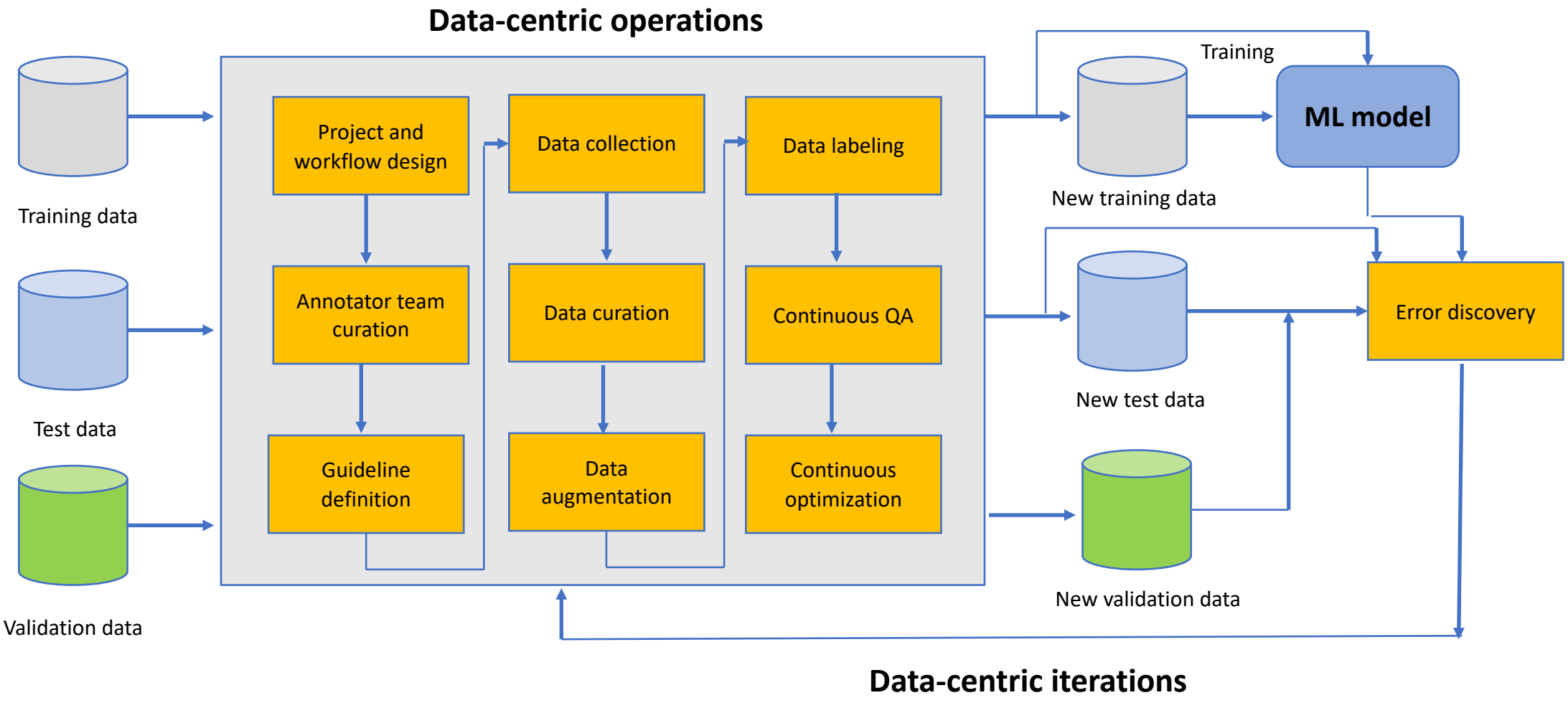
How can you systematically change your data to improve performance

**"Data Engineering, Preparation, and Labelling for AI" – Cognilitica, Jan. 31, 2019.

* Pioneered by Andrew Ng

SIGMA·AI DATA DONE RIGHT

Data-centric technologies are the AI future



SIGMA·AI DATA DONE RIGHT

A data strategy should focus on quality and scalability. It must address:

- Data Consistency
- Security & Privacy Requirements
- Domain Coverage
- Storage
- Bias Mitigation
- Data Annotation Resources

According to McKinsey Global Institute, 75% of AI and ML projects demand learning datasets to be refreshed once per month, and 24% of AI and ML models require refreshed data daily.

SIGMA·AI DATA DONE RIGHT

To be successful a data strategy should include:

- Data Preparation ands Annotation Tools
- Quality Assessments
- Communication
- Timeline
- Budget

SIGMA·AI DATA DONE RIGHT

How Much Training Data Do You Need to Build Out An Effective Strategy?

SIGMA·AI DATA DONE RIGHT

It depends... There is a balance between data volume and managing your project practically. The perfect balance takes into account the following factors:

- Defining performance objectives to guide all decisions
- Complexity
- Variable environmental conditions
- Intra-data variability
- Accessibility of the data and how to augment it with synthetic data

SIGMA·AI DATA DONE RIGHT

What Are The Biggest Challenges For Data-Centric AI Project Teams?

SIGMA·AI DATA DONE RIGHT

Challenges include:

- Data quality and accuracy at scale
- Human resources
- Speed
- Agility

How Do You Know It’s Time to Outsource Data Annotation?

SIGMA·AI DATA DONE RIGHT

Indications that outsourcing data annotation makes the most sense include:

- Lack of state-of-the-art tools
- Lack of in-house resources with the right skill set
- Inefficiency
- Challenges to finding data
- Breakdowns in collaboration
- Tight timeline
- Low Risk Tolerance

Options include:

- Crowdsourcing
- Using a data platform and contractor annotators
- Partnering with a data annotation service provider

Building The Proper Data Annotation Strategy

SIGMA·AI DATA DONE RIGHT

A Proper and Scalable Data Annotation Strategy include:

- Project analysis
- Guidelines and requirements
- Tools and procedures setup
- Comprehensive test
- Client feedback
- Annotation/Collection at scale
- Quality assessment

Positioning Your Project For Success

SIGMA·AI DATA DONE RIGHT

You need to ask these pivotal questions

- Do you need more resources to accomplish data annotation effectively and generate the required volume of annotated data?
- Would your project benefit from advanced tools with features that enable the highest quality of annotation?
- How can you mitigate risk from undesirable outcomes in training and reworking data annotation?
- Would your project benefit from the advice of an experience data annotation provider?

Thank you!

Contact: Jean-Claude Junqua

jean-claude@sigma-ai.com

SIGMA·AI DATA DONE RIGHT