

LINKSFOUNDATION.COM

Applied Data Science Project

L3 - Model & data-centric data science projects

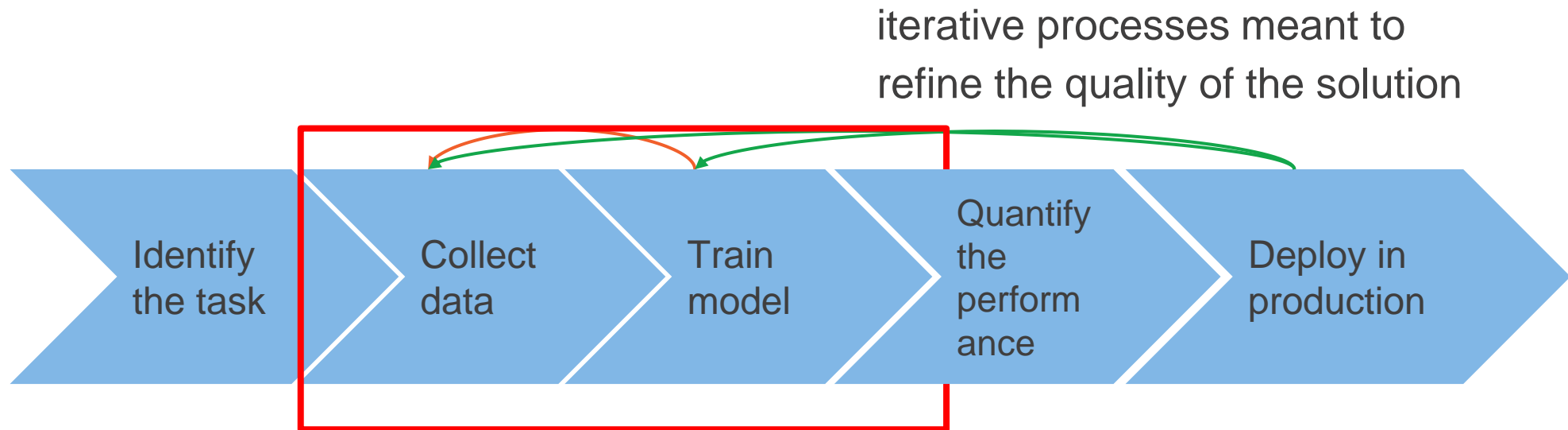


**Politecnico
di Torino**

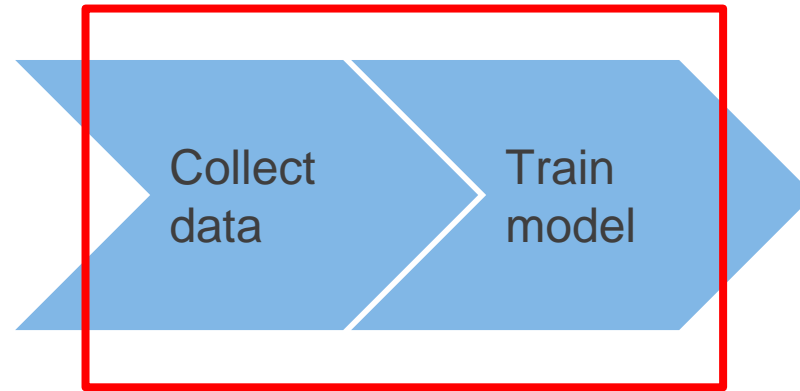


e l i s
European Laboratory for Learning and Intelligent Systems

Artificial intelligence

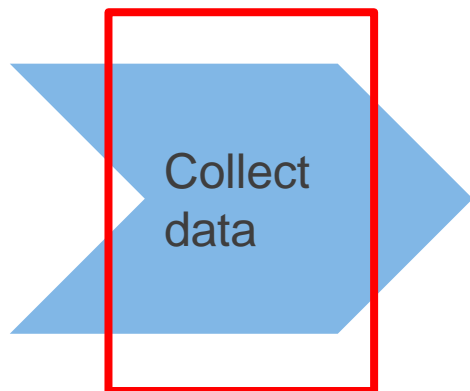


Data + Model

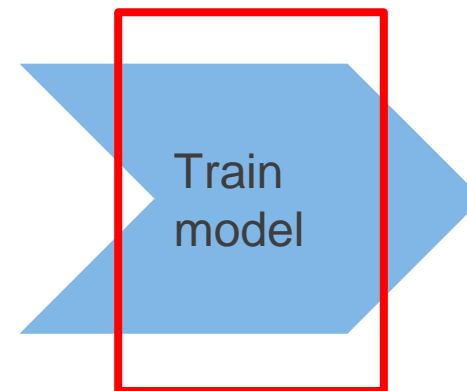


artificial intelligence = data + model (software + algorithm)

Data-centric vs Model-centric



Data-centric: the focus is on acquiring further examples or cleaning the collected ones to retrain the algorithm and generate a new model. The output of this activity is extending the dataset that is used for training



Model-centric: the focus is on modifying the algorithm by extending the neural architecture (for instance having more layers, new residual connections) and then train it with the data at disposal

Data



Collect
data

data is vital for creating any sort of artificial intelligence

improving data has a big impact to artificial intelligence even
more than model optimization

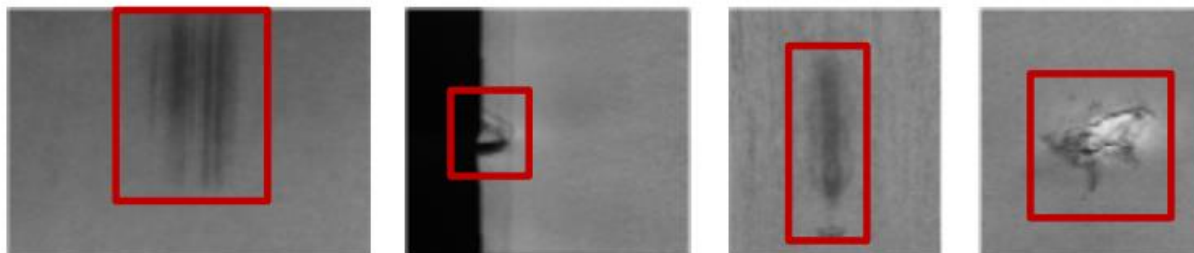
unless of radical changes in the code thus not optimization



Inspecting steel sheets for defects



Examples
of defects



Baseline system: 76.2% accuracy
Target: 90.0% accuracy

Andrew Ng

Improve code vs improve data

	Steel defect detection
Baseline	76.2%
Model-centric	+0% (76.2%)
Data-centric	+16.9% (93.1%)

Other examples

	Solar panel	Surface inspection
Baseline	75.68%	85.05%
Model-centric	+0.04% (75.72%)	+0.00% (85.05%)
Data-centric	+3.06% (78.74%)	+0.4% (85.45%)

Easier step

Improving data turns out to be key for a better artificial intelligence solution

Note: Improving a code is different than designing a new, breakthrough, code however the effort for the latter is way higher than improving data and the return of the effort (may) be very high

Take home message: we consider the data improvement as an easier and necessary step when developing a machine intelligence before starting a new venture

Data improvement

Strategies for data improvement:

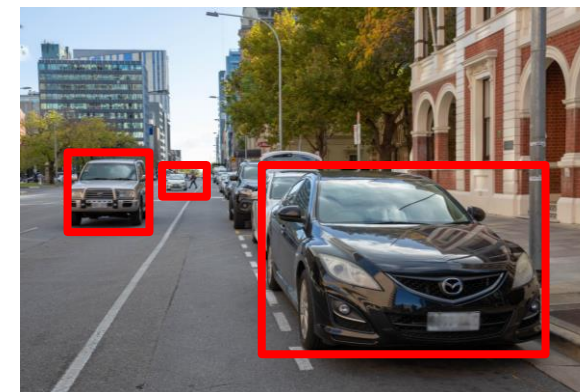
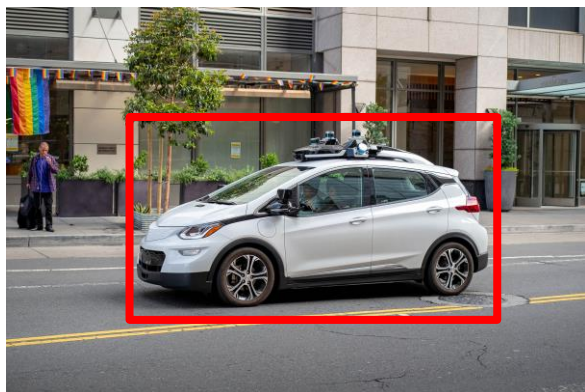
- more examples

augmentation

- completeness
- consistency

cleaning

Augumentation



Take or generate new examples

Consistency

Task: Label cars



Consistency

Task: Label cars

Annotator 1



Consistency

Task: Label cars

Annotator 2



Consistency

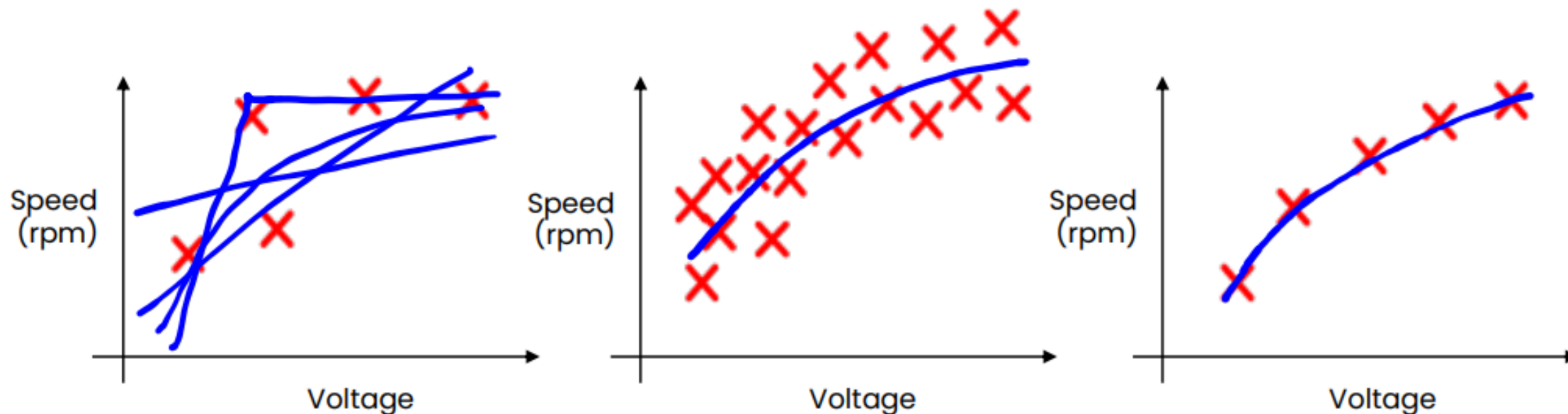
Consistency in annotation turns out to be crucial for minimizing the potential error of the intelligence

However, ensuring a consistent dataset is a not obvious task

It involves:

- how the task has been conceived
- how the intervention of the human has been designed
- how did human annotators perform their task
- how the dataset has been packaged

Small Data and Label Consistency



- Small data
- Noisy labels

- Big data
- Noisy labels

- Small data
- Clean (consistent) labels

Andrew Ng

Completeness

Task: Label cars



Completeness

Task: Label cars

Annotator 1



Completeness

Task: Label cars

Annotator 2



Completeness

Completeness in annotation turns out to be crucial for improving coverage to the intelligence

However, ensuring a complete dataset is a not obvious task

It involves:

- how the task has been conceived
- how the intervention of the human has been designed
- how did human annotators perform their task
- how the dataset has been packaged

Good data

~~Big data vs small data~~
good data

Good data is:

- Defined consistently (definition of labels y is unambiguous)
- Cover of important cases (good coverage of inputs x)
- Has timely feedback from production data (distribution covers data drift and concept drift)
- Sized appropriately

We also refer to good data with the concept of clean data

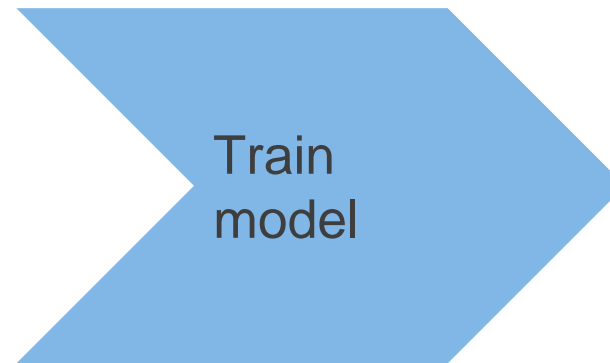
Example: Clean vs. noisy data



Note: Big data problems where there's a long tail of rare events in the input (web search, self-driving cars, recommender systems) are also small data problems.

Andrew Ng

Model



model encapsulates the intelligence in an executable environment that embeds both training data and algorithm



Model

Improving a model is a hard task because it inherits the challenges related to optimize both data and algorithm



Model improvement

Strategies for model optimization

- Any change in the data, if statistically relevant, is propagated to the final output of the model. This links to the previous topic
- Change in the algorithm, for instance the addition of a new layer in a neural architecture, or eventually, a brand new architecture
- Change in the hyperparameter set, for instance `n_layers`, or learning rate. This change modifies the parameter weights



Thank you for your attention.

Questions?

CONTACTS

Giuseppe Rizzo

Program Manager (LINKS Foundation) and
Adjunct Professor (Politecnico di Torino)

giuseppe.rizzo@polito.it



FONDAZIONE LINKS
Via Pier Carlo Boggio 61 | 10138 Torino
P. +39 011 22 76 150
LINKSFOUNDATION.COM