LINKSFOUNDATION.COM

FONDAZIONE links
PASSION FOR INNOVATION

Applied Data Science Project

Politecnico di Torino

Lecture 2

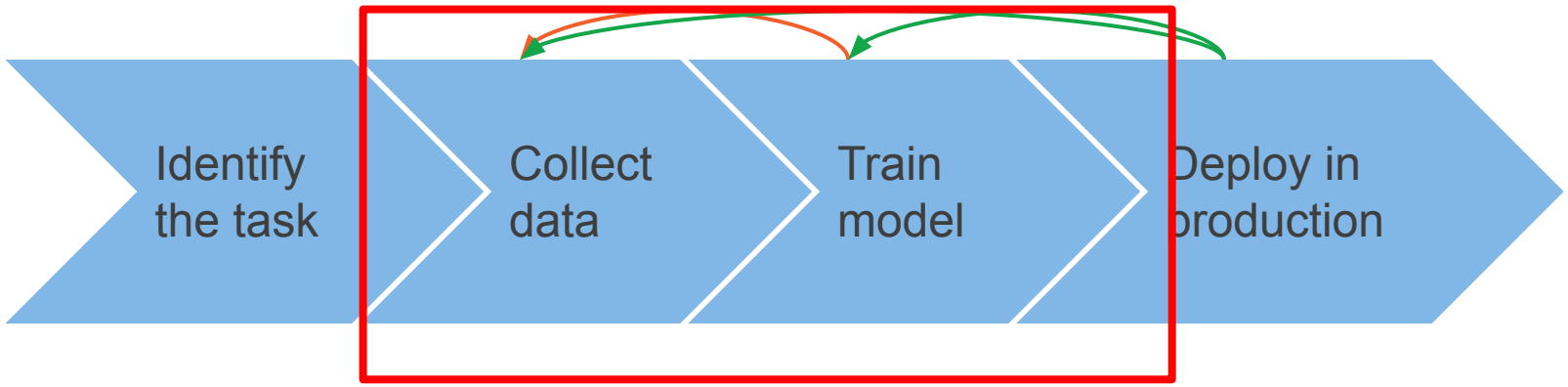
ellis
European Laboratory for Learning and Intelligent Systems

Turin

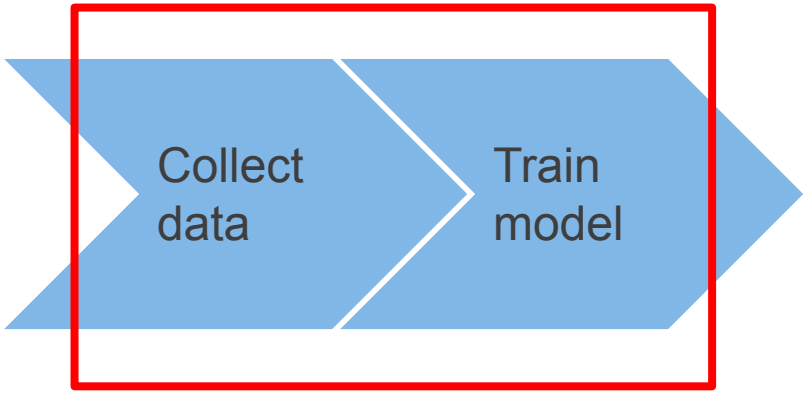
September 28, 2021

Machine intelligence

iterative processes meant to refine the quality of the solution

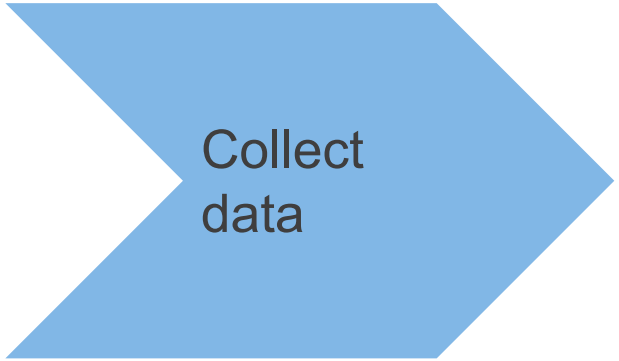


Data + Model



machine intelligence = data + model (software + algorithm)

Data



data is vital for creating any sort of machine intelligence

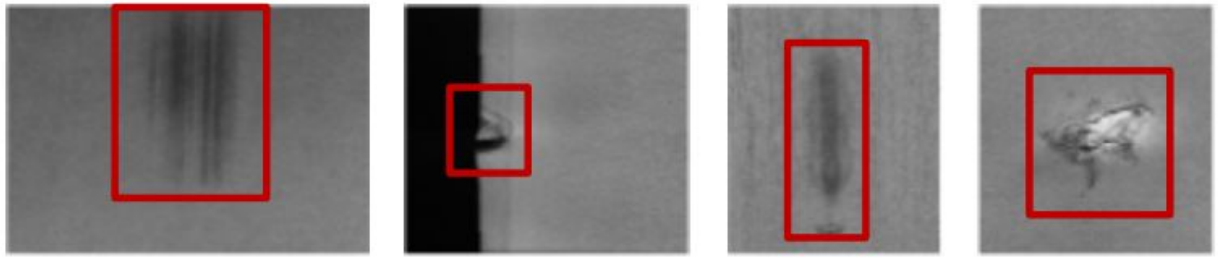
Data

improving data has a big impact to machine 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 machine intelligence

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

Two strategies for data improvement:

- consistency
- completeness

Consistency

Task: Label cars

Consistency

Task: Label cars

Annotator 1

Consistency

Task: Label cars

Annotator 2

Consistency

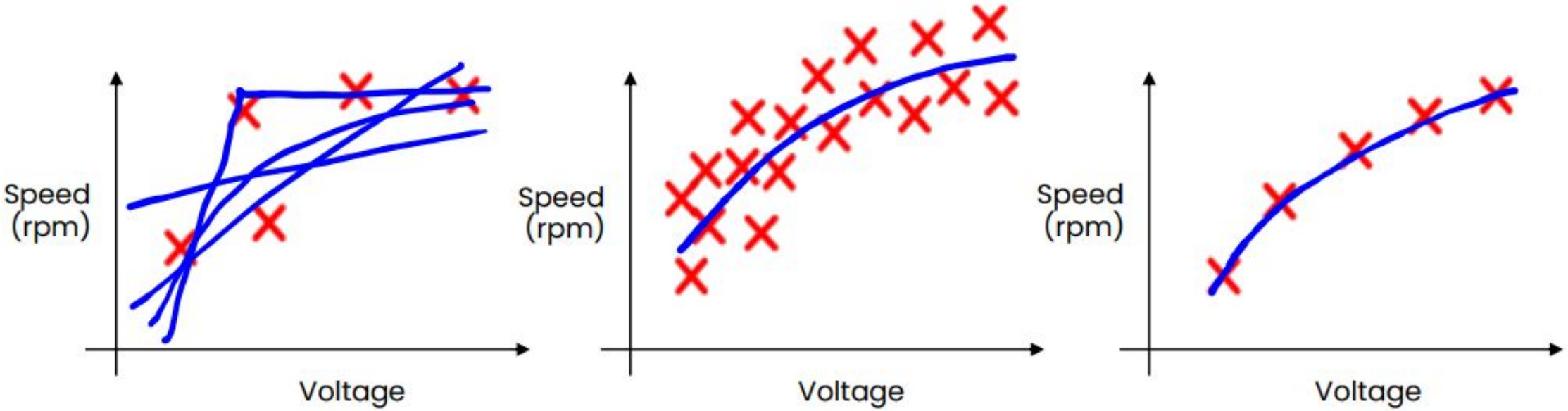
Consistency in annotation turns out to be crucial for the minimizing the potential error of the intelligence

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~~

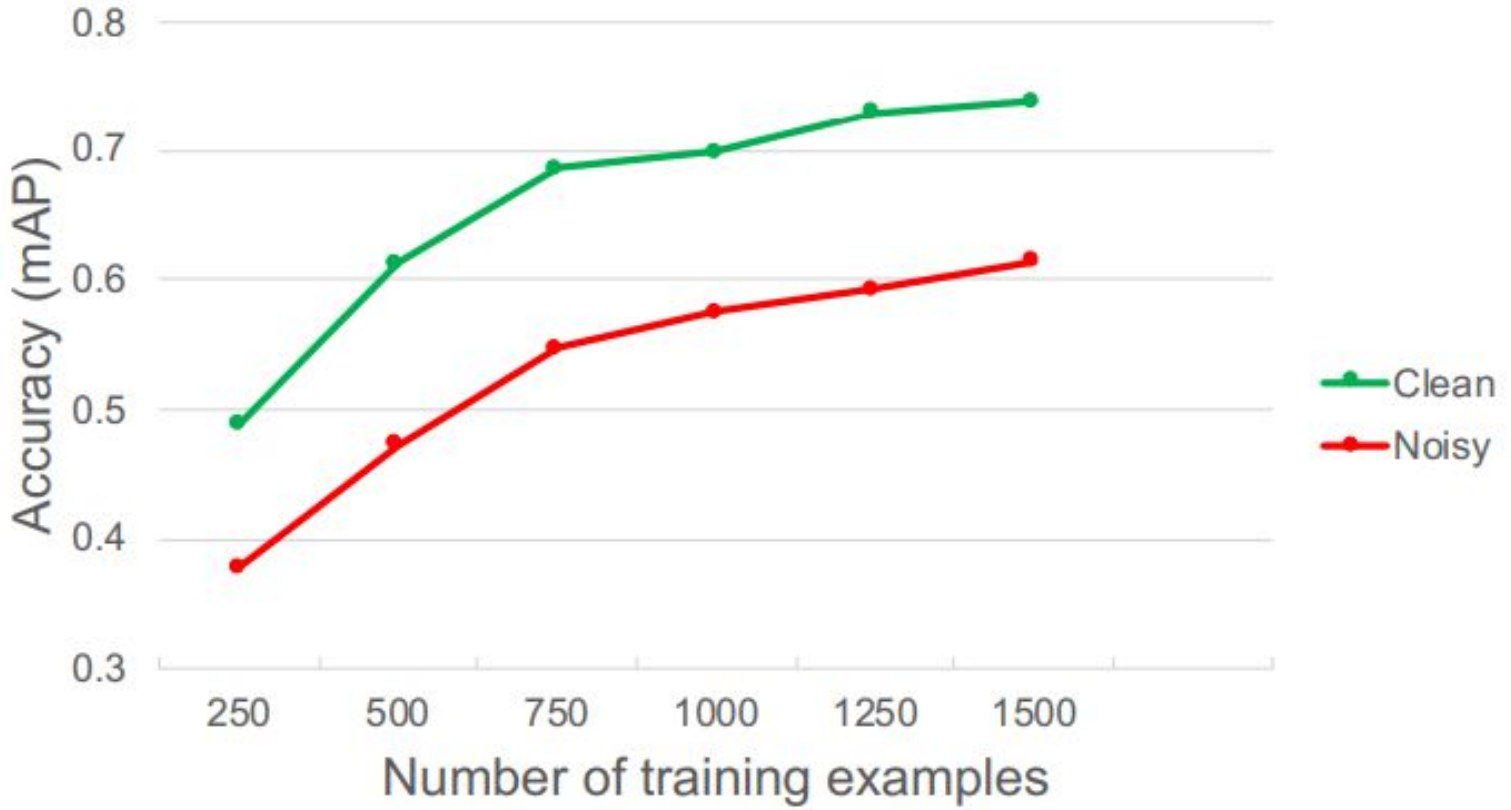
good data

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

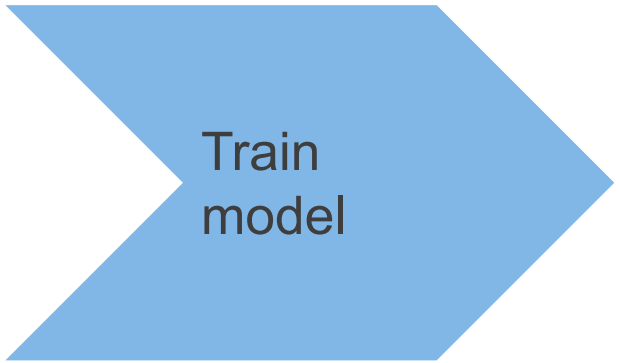
Example: Clean vs. noisy data



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

Model

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

The traditional approach is to:

- collect data as much as possible, then standardize it with preprocessing
- optimize the model to be enough robust to cover noise iteratively by minimizing the error

FONDAZIONE links

PASSION FOR INNOVATION

Thank you for your attention.

Questions?

FONDAZIONE links

PASSION FOR INNOVATION

CONTACTS

Giuseppe Rizzo

Team Leader

p. +39 011 2276244

e. giuseppe.rizzo@linksfoundation.com

FONDAZIONE LINKS

Via Pier Carlo Boggio 61 | 10138 Torino

P. +39 011 22 76 150

LINKSFOUNDATION.COM