#DataScienceMilan

Opening Talk

07-06-2017
@Assolombarda
-Milan-

About me

Background in Telematics and Distributed Systems
    Software Engineering

Data Science experience:
    Predictive Marketing
    Web Cyber Security
    Retail & Business Banking
    Automotive

Co-author of Python Deep Learning book

Founder of the Data Science Milan community

We are about 500+ professionals and major data science hub for startups, organizations and academia based in the greater Milan area.

Innovate, collaborate, share...

We are an independent group with the only goal of promoting and pioneering knowledge and innovation of the data-driven revolution in the Italian peninsula and beyond. We encourage international collaboration, sharing and open source tools. The official language of our events, talks and communication is English. Everyone who is involved in data science projects or want to undertake this career is invited to join.

41 **DATA SCIENCE**
Data Scientists, Statisticians, AI, Machine Learning Specialists, Quant Analysts, BI, Analytics

32 **ENGINEERING**
Big Data, Data Engineers, Software Developers, System Design, ETL, NoSQL, Data Warehousing, DevOps

19 **ACADEMIA**
Researchers, PhD, Lecturers, Professors, Students

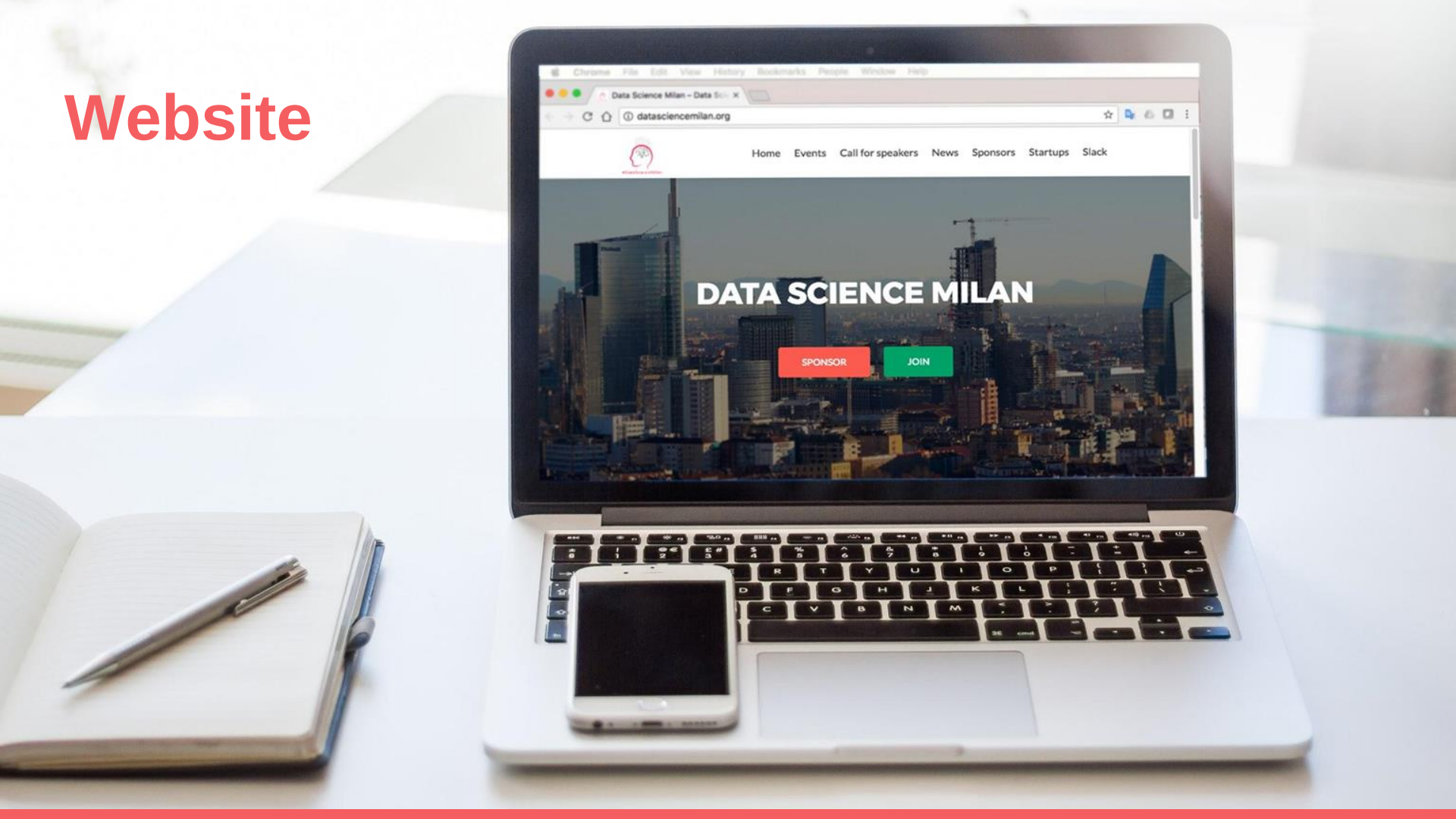
8 **OTHER**

TECH AND KNOWLEDGE SHARING

CONNECTING PEOPLE

STARTUP NETWORK

Website


datasciencemilan.org
Home
Events
Call for speakers
News
Sponsors
Startups
Slack
DATA SCIENCE MILAN
SPONSOR
JOIN

Demystifying Data Science

Gianmario Spacagna

"The Role of the Data Scientist in the Industry 4.0" @ Assolombarda, Milan

2017/06/07

Misconceptions about DS

DS is not just Statistics

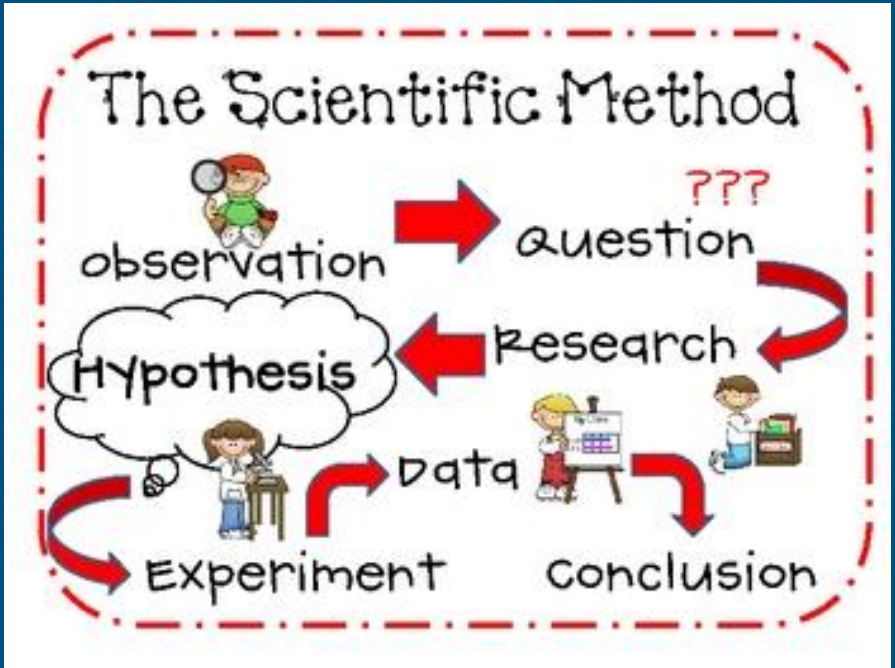
DS is not just Machine Learning / AI

DS is not just Big Data

DS is not just Business Analytics

DS is not magic and Data Scientists are not wizards

DS is Science [+ Technology]



A more exhaustive stack

Explorative analysis

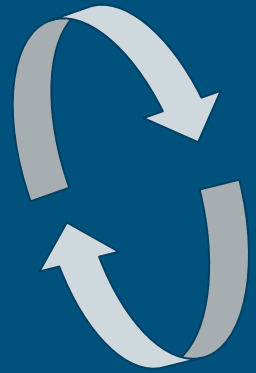
Data cleansing / feature representation

Hypothesis-driven development

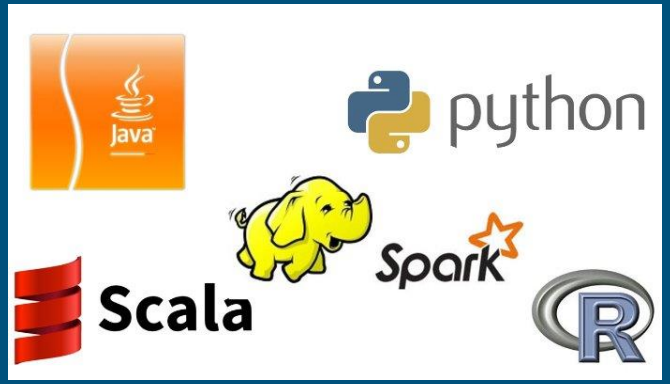
Modeling: Statistics and Machine Learning +
Domain knowledge

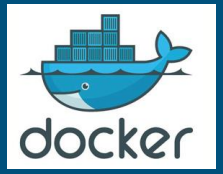
Automation and Engineering

Effective visualization / actionable insights



Programming languages & tools





Regardless of the programming language or tool, the thoughtful methodology makes you a Data Scientist

Focusing on science rather than just data

Big Data Vs. Smart Data

Design of experiments and data collection

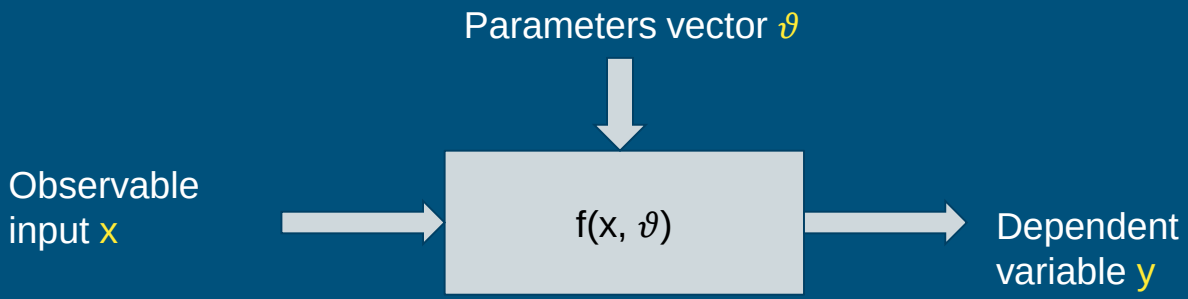
Understanding correlations

Answering the right question

Parametric Model

All models are wrong, but some are useful!

"*A scientific model is a conceptual, mathematical or physical representation of a real-world phenomenon. A model is generally constructed for an object or process when it is at least partially understood, but difficult to observe directly.*"



Learning algorithm

Model parameters define the probability of observing a set of outcomes (data) given the model characterized by those parameters (likelihood):
$\mathcal{L}(\theta|\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{x}|\theta)$

Training/fitting: Minimizing a given loss function (e.g. using Gradient Descent) in order to maximize the likelihood.

A model is fitted when we have found the unique set of parameters θ that best describes the training data.

Testing procedures

Model validation: Validating the hypothesis that the model describe data used for training but also unseen data (generalization property).

Hypothesis space: whole set of parameters representing the algorithm pipeline not just the analytical model (includes pre and post processing, implementation configurations, thresholds, etc...).

Model selection/tuning: process of selecting among a set of candidates, the model/pipeline-configuration that scores the highest in the validation.

KPIs and e2e evaluation

Model validation metrics + business KPIs quantifying the added value, e.g:

- Perceived user experience (impact on the final consumer)
- Deployment and real-time usage simulation (A/B testing ideally)
- Relevance indicators (accuracy depends on the context)
- Reaction time (how many data points need the model to adapt?)
- Failure scenarios (performance in case of wrong data collection)
- Error cost (accuracy meets economics)

From PoC to Production

A data-driven app is not enough, any website is powered by data.

A Data Product must be able to derive value from raw data - not just consuming it as it is - and generate knowledge (in the form of other data or insights) utilized to solve a specific problem.

ML, Statistics and Data Analysis techniques are not new. Innovation comes from integrating streams of information generated from data products able to automate, drive and/or trigger actions.

In the AI context, the data product should also be able to take unbiased decisions on the human's behalf.

Agile Data Product Development

Time-boxed research spikes along with clearly defined feature stories.

Focus on MVP meeting the given acceptance criteria (upon KPIs).

Notebooks good for analysys, entry points and results presentation.
Code developed in modules and functions in a proper IDE.

Git branching system, reproducible analysis, avoid non programmatic operations.

Model versioning: both code, priors and evaluation results.

Test-driven development: replace “assertEqual” with uncertainty ranges.

My experience with data products

Connesso: processing tyre sensors data to estimate latent variables and predicting evolution over time

Demand Forecasting: predicting short and long-term sales of prestige products

First Time Buyer: predicting propensity to buy a first home and gaining advantage in the mortgage competition

SmartBusiness Insights: portal showing revenue and spending at a glance and comparing your performance to competitors

Talos: detecting 0-day web hosts serving threats and malware

Blacklist Feeds Evaluation: quantifying the quality and optimize renewal of third-party data providers

360 Customer Profiles: automatic discovery and segmentation of user clusters

The DS Team Unicorn

The DS team shall be cross-functional over different areas and be able to deliver as end-to-end as possible solutions.

Software Development

Math & Statistics

Subject Matter Expertise

+ a bunch of related skills:
    + Data Engineering
    + System Design and Infrastructures (cloud?)

DS & Analytics (DSA)

Not everything should be rocket science or advanced predictive models, analytics also matter!

Data Science can provide you with business strategy, Data Analytics provide you with day-to-day operational insights.

Many calculations could be simply done with sums and divisions.

Analytics is a good place to start for delivering quick value and gain trustiness and understanding of the business.

Go PRO!

The Professional Data Science Manifesto

Become a signatory at
www.datasciencemanifesto.org

Latest technology advancements have made data processing accessible, cheap and fast for everyone. We believe combining engineering practices with the scientific method will extract the most utility from these advancements. So this manifesto proposes a principled methodology for unifying science and technology by valuing:

- **Minimal Viable Products** over prototypes
- **APIs** over databases
- **Clever use of computation** over convenient assumptions
- **Dashboards** over reports
- **Validation**, **scrutiny** and **repeatability** over convention and ad verecundiam

That is, while there is value in the items on the right, we value the items on the left more.

Principles

Aim to completely remove manual intervention in numerical processing.

Data science is about solving problems, not models or algorithms.

All validation of data, hypotheses and performance should be tracked, reviewed and automated.

Prior to building a model, construct an evaluation framework with end-to-end business focused acceptance criteria.

A product needs a pool of measures to evaluate its quality. A single number cannot capture the complexity of reality.

Even research can be broken down into clearly defined tasks. The smallest of iterations should be preferred in acquiring, integrating and correcting knowledge.

Don't neglect assumptions in models. Make them explicit then aim to have them either verified or removed.