ABOUT DATA FROM A MACHINE LEARNING PERSPECTIVE

Oriol Pujol

UNIVERSITAT DE BARCELONA

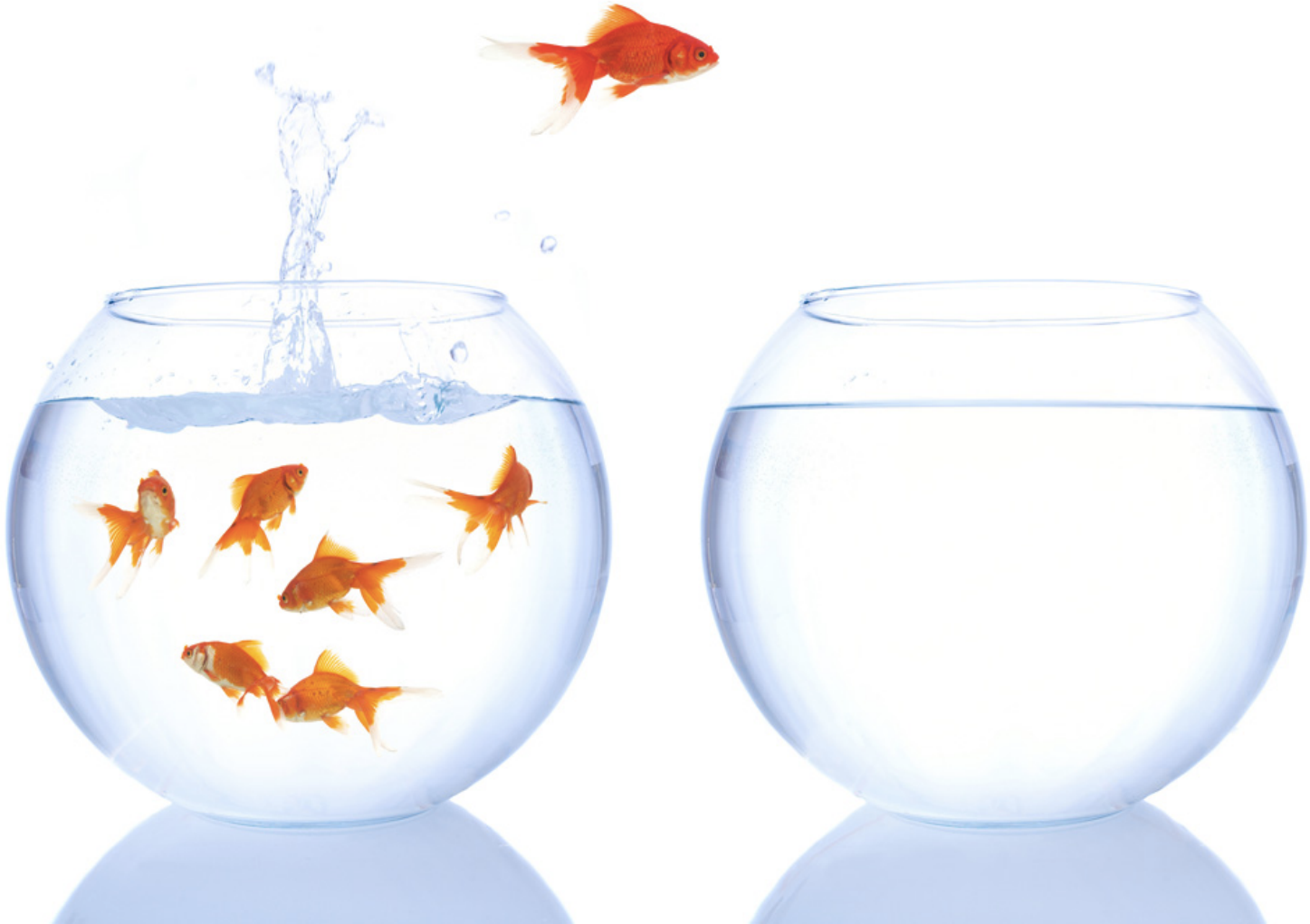
learn

TWO FLAVOURS AND ONE PRESENTATION

Share experiences in openness from the machine learning community

To propose potential solutions to the problem of sharing sensible data

A COMMUNITY SHIFT



THE ROLE OF DATA IN MACHINE LEARNING

- Machine learning is sometimes called learning from data. Thus we are talking about building/coding algorithms that learn from data.

- Data is the key component in most research, but it is obviously crucial in machine learning.

EXAMPLES OF TASKS IN MACHINE LEARNING

Automatic captioning

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."

CURRENT MACHINE LEARNING RESEARCH FOCUS

- Improving results in terms of generalization, robustness, speed, and interpretability.
- Comparison with other techniques is fundamental.
- Many different data sets are needed to assess each property we are tackling in the research, e.g. we need large datasets for speed and big data.
- Sometimes, their application largely depends on the domain, e.g., financial products forecasting, targeted gene identification.

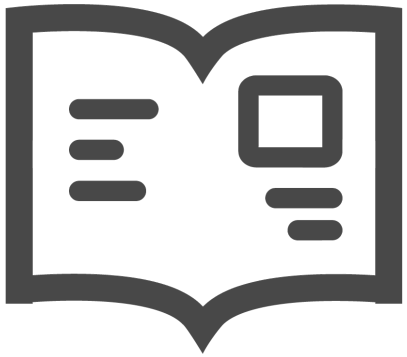
AS A RESEARCHER

We pursue to contribute to advance science: reproducibility of experimentation that helps to eliminate individual biases.

DISEMINATION: REPORTING PROCESS, CLAIMING AUTHORSHIP

QUALITY: COMMUNITY VALIDITY ASSESSMENT, PEER REVIEW
(KPI USED BY ORGANISATIONS FOR RESEARCH QUALITY)

FOR THIS REASON WE USE PUBLISHING



DISSEMINATION
QUALITY ASSESSMENT

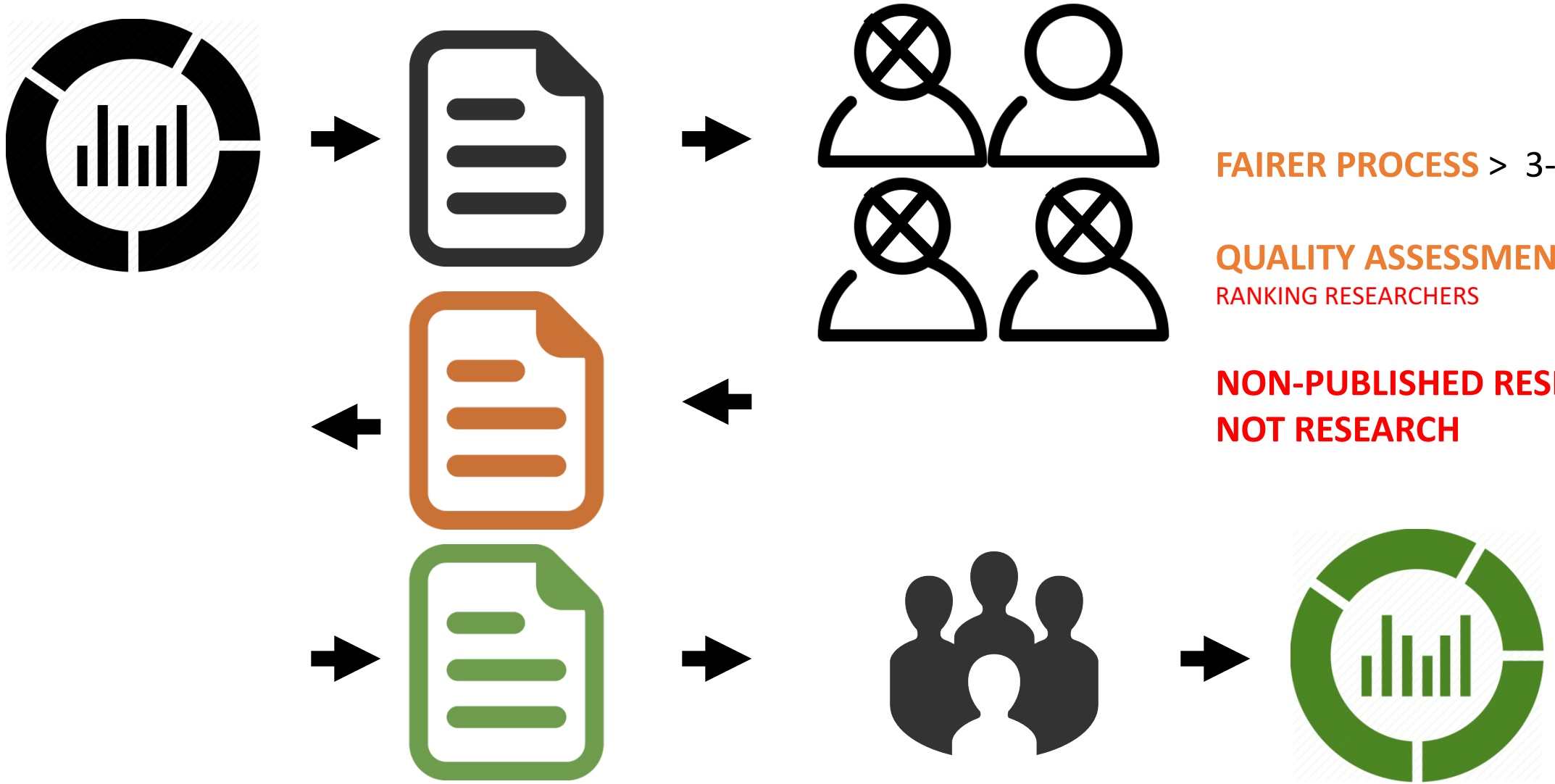
FOR THIS REASON WE USE PUBLISHING

LONG PROCESS > 1 YEAR
NOT A GOOD DISEMINATION SOURCE, OR ATTRIBUTION PROC.

PEER QUALITY ASSESSMENT
CRITICAL FEEDBACK
GUARD AGAINST ERRORS
MAJOR SOURCE FOR RANKING RESEARCHERS

NON-PUBLISHED RESEARCH IS NOT RESEARCH

FOR THIS REASON WE USE PUBLISHING



FAIRER PROCESS > 3-6 month

QUALITY ASSESSMENT
RANKING RESEARCHERS

NON-PUBLISHED RESEARCH IS NOT RESEARCH

THE NIPS CASE: WHEN CONFERENCE PAPERS ARE OLD NEWS

IN MACHINE LEARNING 3-6 MONTHS IS TOO LONG

DECEMBER 5TH – 10TH , 2016

NIPS Proceedings$^{\beta}$ | Books | 2016

Conditional Image Generation with PixelCNN Decoders

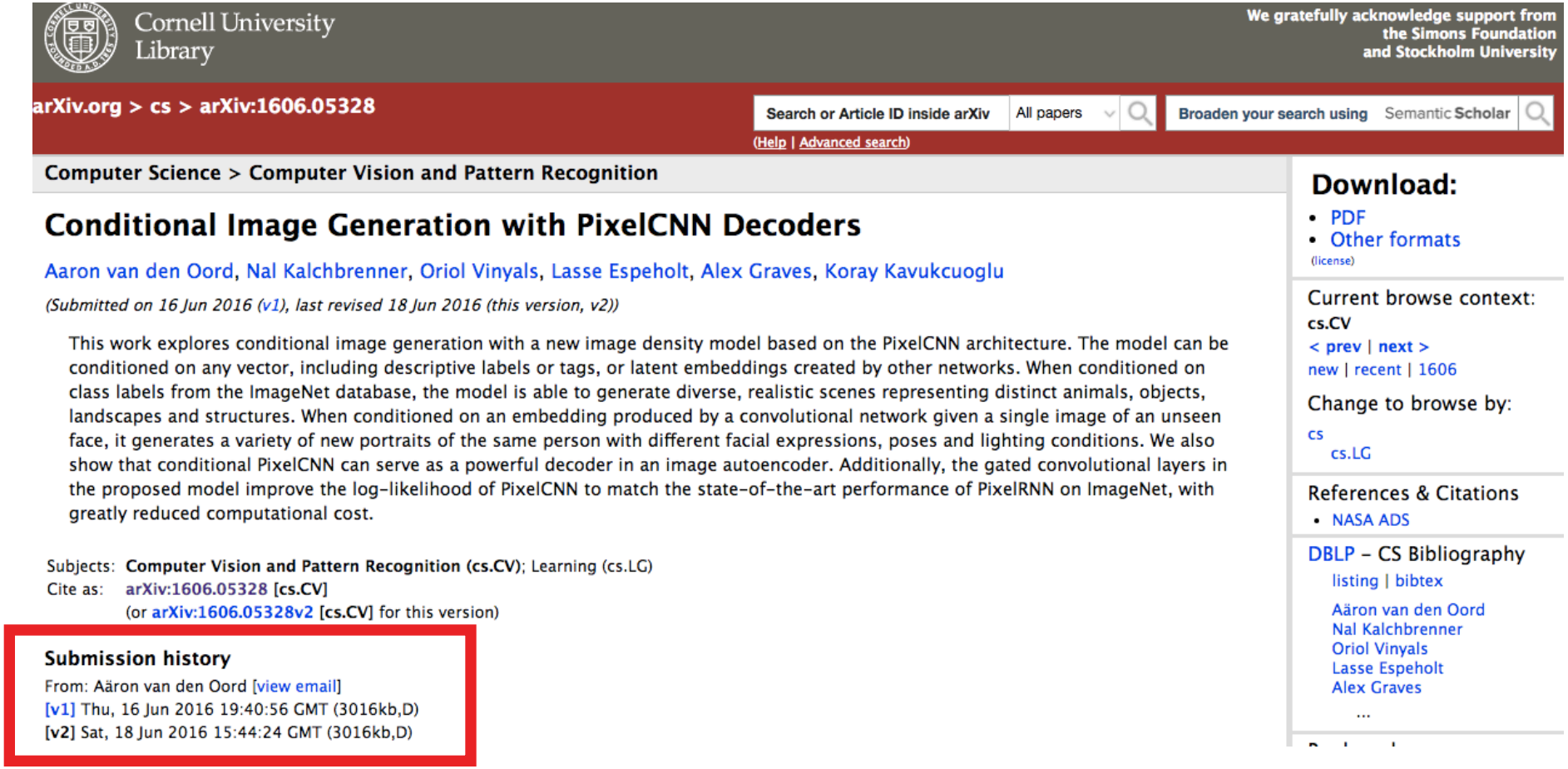
Part of: Advances in Neural Information Processing Systems 29 (NIPS 2016) pre-proceedings

[PDF] [BibTeX] [Supplemental] [Reviews]

Authors

- Aaron van den Oord
- Nal Kalchbrenner
- Lasse Espeholt
- koray kavukcuoglu
- Oriol Vinyals
- Alex Graves

THE NIPS CASE: WHEN CONFERENCE PAPERS ARE OLD NEWS



Cornell University Library

We gratefully acknowledge support from the Simons Foundation and Stockholm University

arXiv.org > cs > arXiv:1606.05328

Search or Article ID inside arXiv | All papers | Broaden your search using Semantic Scholar

(Help | Advanced search)

Computer Science > Computer Vision and Pattern Recognition

Conditional Image Generation with PixelCNN Decoders

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, Koray Kavukcuoglu

(Submitted on 16 Jun 2016 (v1), last revised 18 Jun 2016 (this version, v2))

This work explores conditional image generation with a new image density model based on the PixelCNN architecture. The model can be conditioned on any vector, including descriptive labels or tags, or latent embeddings created by other networks. When conditioned on class labels from the ImageNet database, the model is able to generate diverse, realistic scenes representing distinct animals, objects, landscapes and structures. When conditioned on an embedding produced by a convolutional network given a single image of an unseen face, it generates a variety of new portraits of the same person with different facial expressions, poses and lighting conditions. We also show that conditional PixelCNN can serve as a powerful decoder in an image autoencoder. Additionally, the gated convolutional layers in the proposed model improve the log-likelihood of PixelCNN to match the state-of-the-art performance of PixelRNN on ImageNet, with greatly reduced computational cost.

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Learning (cs.LG)
Cite as: arXiv:1606.05328 [cs.CV]
(or arXiv:1606.05328v2 [cs.CV] for this version)

Submission history

From: Aäron van den Oord [view email]
[v1] Thu, 16 Jun 2016 19:40:56 GMT (3016kb,D)
[v2] Sat, 18 Jun 2016 15:44:24 GMT (3016kb,D)

Download:

- PDF
- Other formats

(license)

Current browse context: cs.CV
< prev | next >
new | recent | 1606

Change to browse by:
cs
cs.LG

References & Citations
- NASA ADS

DBLP - CS Bibliography
listing | bibtex
Aäron van den Oord
Nal Kalchbrenner
Oriol Vinyals
Lasse Espeholt
Alex Graves
...

24/7 PUBLISHING CYCLE

IMMEDIACY, 24/7 PUBLISHING CYCLE

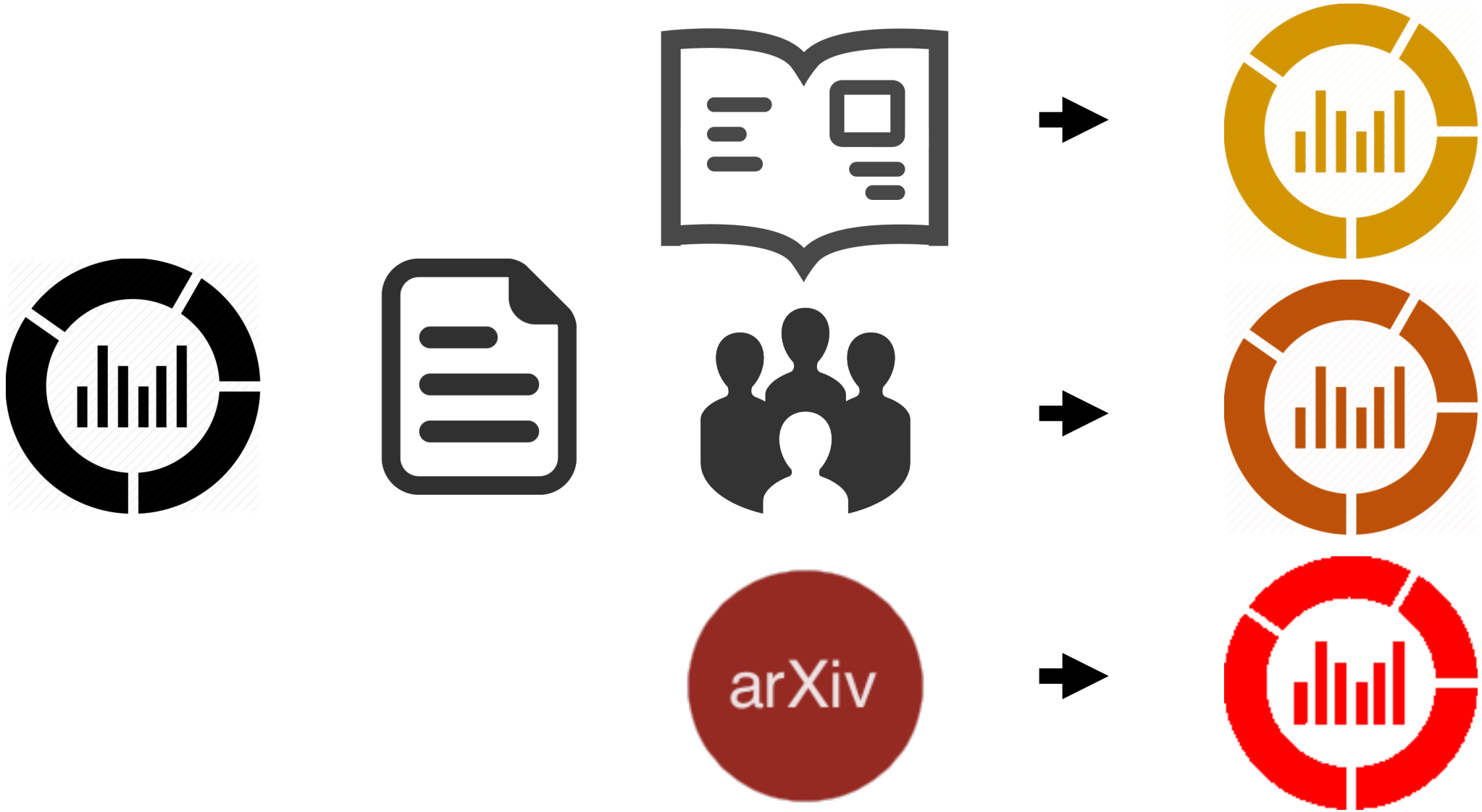
IMMEDIATE ATTRIBUTION

NO QUALITY ASSESSMENT

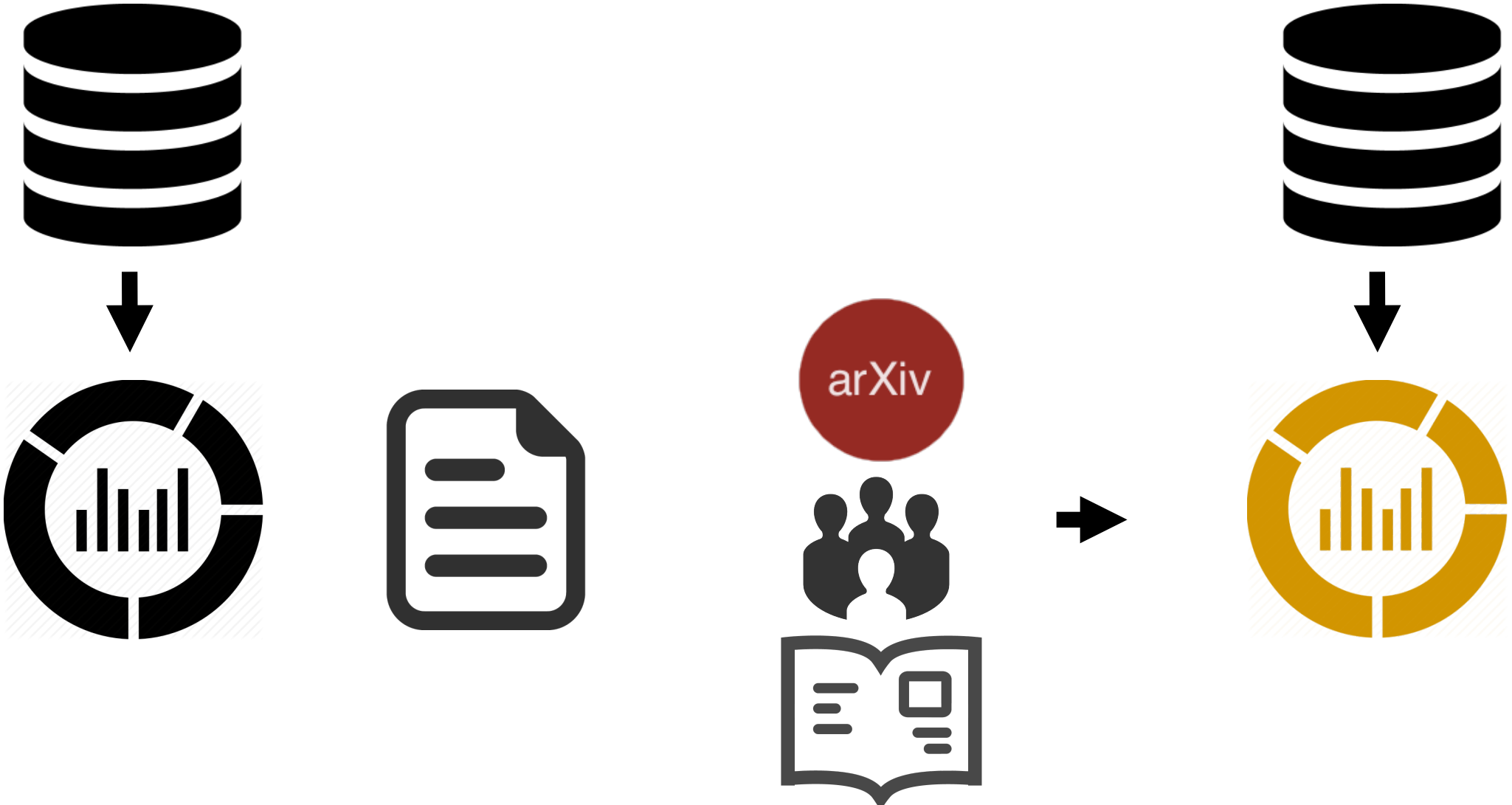
RANKING RESEARCHERS

HOW TO DISTINGUISH GOOD RESEARCH FROM FRAUD

HOWEVER, IN REALITY



TOWARDS REPRODUCIBILITY: WE NEED DATA



AN EXAMPLE: PLOS ONE

FOCUS SHIFT:

METHODOLOGICAL CORRECTNESS

DATA

MACHINE LEARNING DATASET REPOSITORIES

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Loading

View ALL Data Sets

Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 360 data sets as a service to the machine learning community. You may **view all data sets** through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.

Supported By: In Collaboration With: Rexa.info

Latest News:

2013-04-04: Welcome to the new Repository admins Kevin Bache and Moshe Lichman!

2010-03-01: Note from donor regarding Netflix data

2009-10-16: Two new data sets have been added.

2009-09-14: Several data sets have been added.

2008-07-23: Repository mirror has been set up.

2008-03-24: New data sets have been added!

2007-06-25: Two new data sets have been added: UJI Pen Characters, MAGIC Gamma Telescope

Newest Data Sets:

2016-11-23: NIPS Conference Papers 1987-2015

2016-11-16: Amazon book reviews

2016-08-14: Dota2 Games Results

2016-08-05: Facebook metrics

Most Popular Data Sets (hits since 2007):

1189468:  Iris

816286:  Adult

621858:  Wine

534408:  Car Evaluation

MACHINE LEARNING DATASET REPOSITORIES

Facebook metrics Data Set

Download: Data Folder, Data Set Description

Abstract: Facebook performance metrics of a renowned cosmetic's brand Facebook page.

Data Set Characteristics:	Multivariate	**Number of Instances:**	500	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	19	**Date Donated**	2016-08-05
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	7110

Source:

Created by: SÃ©rgio Moro, Paulo Rita and Bernardo Vala (ISCTE-IUL) @ 2016

Data Set Information:

The data is related to posts' published during the year of 2014 on the Facebook's page of a renowned cosmetics brand.
This dataset contains 500 of the 790 rows and part of the features analyzed by Moro et al. (2016). The remaining were omitted due to confidentiality issues.

Attribute Information:

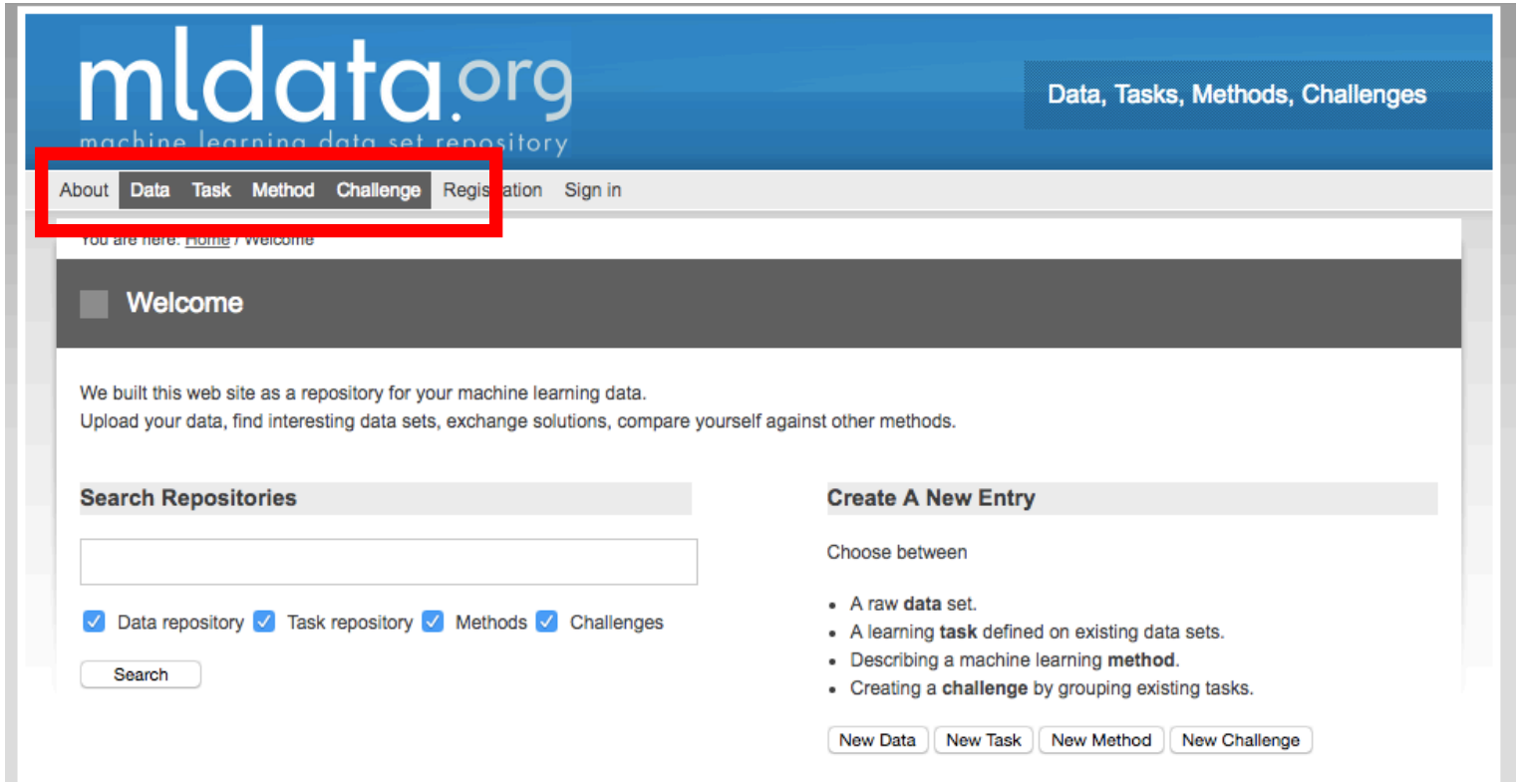
It includes 7 features known prior to post publication and 12 features for evaluating post impact (see Tables 2 and 3 from Moro et al., 2016 - complete reference in the 'Citation Request')

Relevant Papers:

(Moro et al., 2016) Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.

MACHINE LEARNING DATASET REPOSITORIES

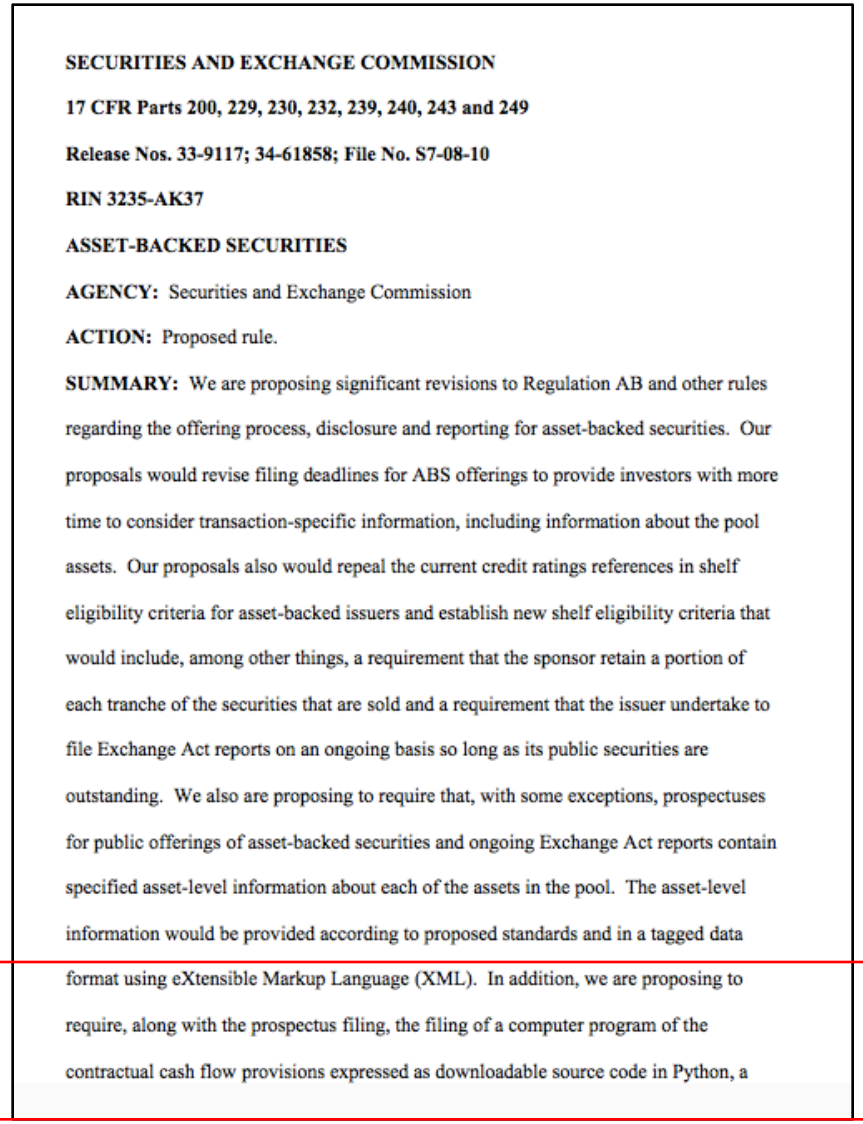
Data Sets - Raw data as a collection of similarily structured objects.
Material and Methods - Descriptions of the computational pipeline.
Learning Tasks - Learning tasks defined on raw data.
Challenges - Collections of tasks which have a particular theme.



WHEN DESCRIPTION AND DATA IS NOT ENOUGH

U.S. SECURITIES AND EXCHANGE COMMISSION

- *The U.S. Securities and Exchange Commission (SEC) 2010 proposal, financial products must include Python code.*



SECURITIES AND EXCHANGE COMMISSION

17 CFR Parts 200, 229, 230, 232, 239, 240, 243 and 249

Release Nos. 33-9117; 34-61858; File No. S7-08-10

RIN 3235-AK37

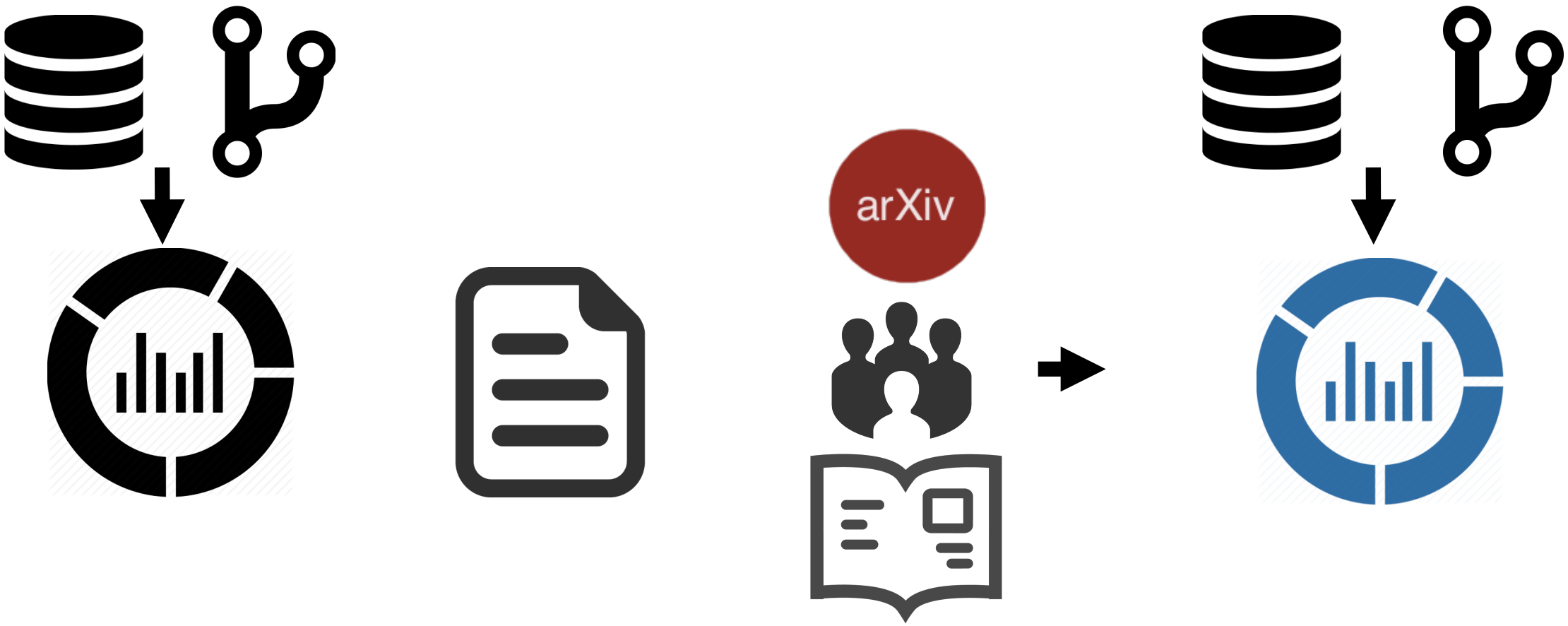
ASSET-BACKED SECURITIES

AGENCY: Securities and Exchange Commission

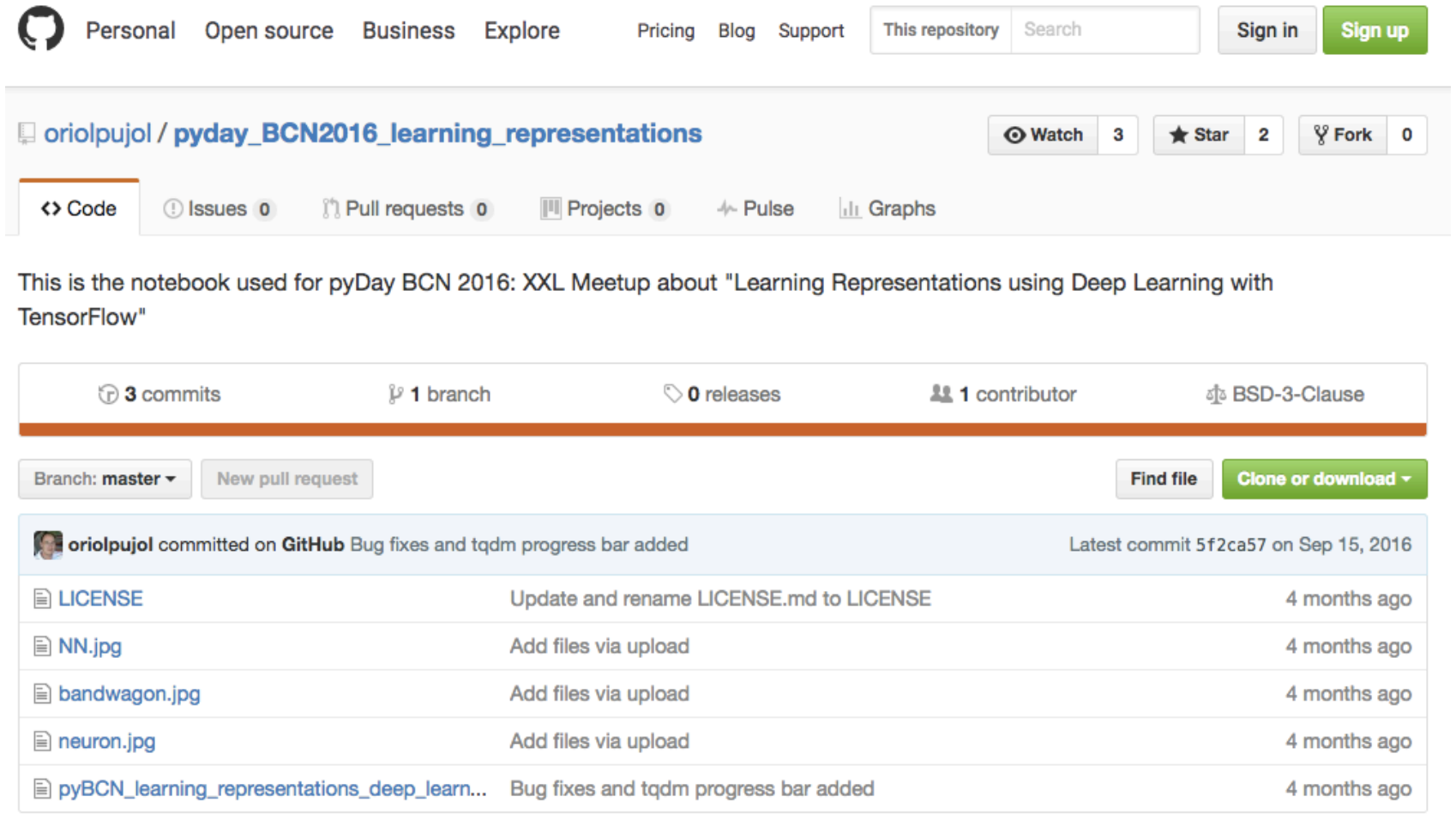
ACTION: Proposed rule.

SUMMARY: We are proposing significant revisions to Regulation AB and other rules regarding the offering process, disclosure and reporting for asset-backed securities. Our proposals would revise filing deadlines for ABS offerings to provide investors with more time to consider transaction-specific information, including information about the pool assets. Our proposals also would repeal the current credit ratings references in shelf eligibility criteria for asset-backed issuers and establish new shelf eligibility criteria that would include, among other things, a requirement that the sponsor retain a portion of each tranche of the securities that are sold and a requirement that the issuer undertake to file Exchange Act reports on an ongoing basis so long as its public securities are outstanding. We also are proposing to require that, with some exceptions, prospectuses for public offerings of asset-backed securities and ongoing Exchange Act reports contain specified asset-level information about each of the assets in the pool. The asset-level information would be provided according to proposed standards and in a tagged data format using eXtensible Markup Language (XML). In addition, we are proposing to require, along with the prospectus filing, the filing of a computer program of the contractual cash flow provisions expressed as downloadable source code in Python, a

WHEN DESCRIPTION AND DATA IS NOT ENOUGH

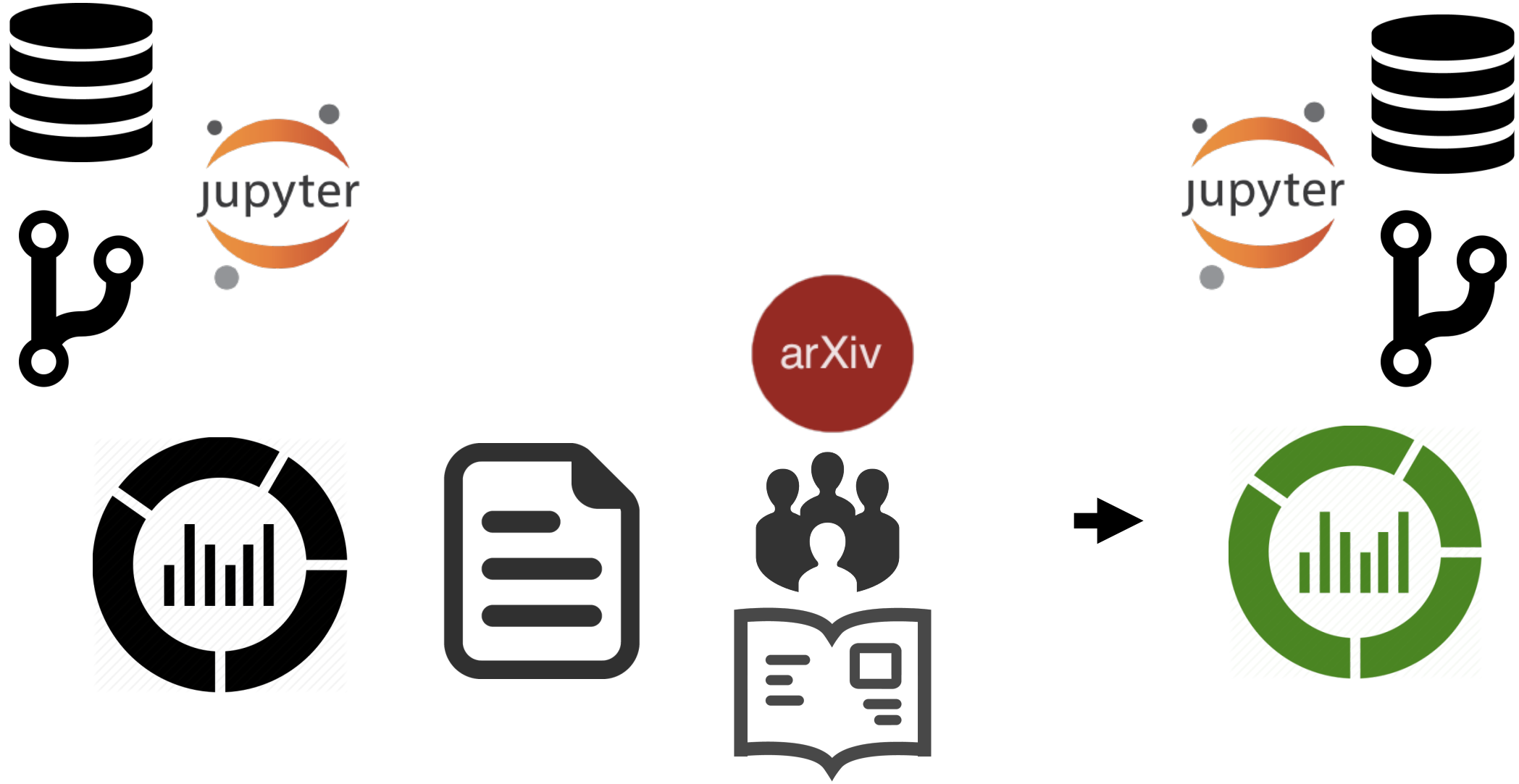


GITHUB PROJECTS



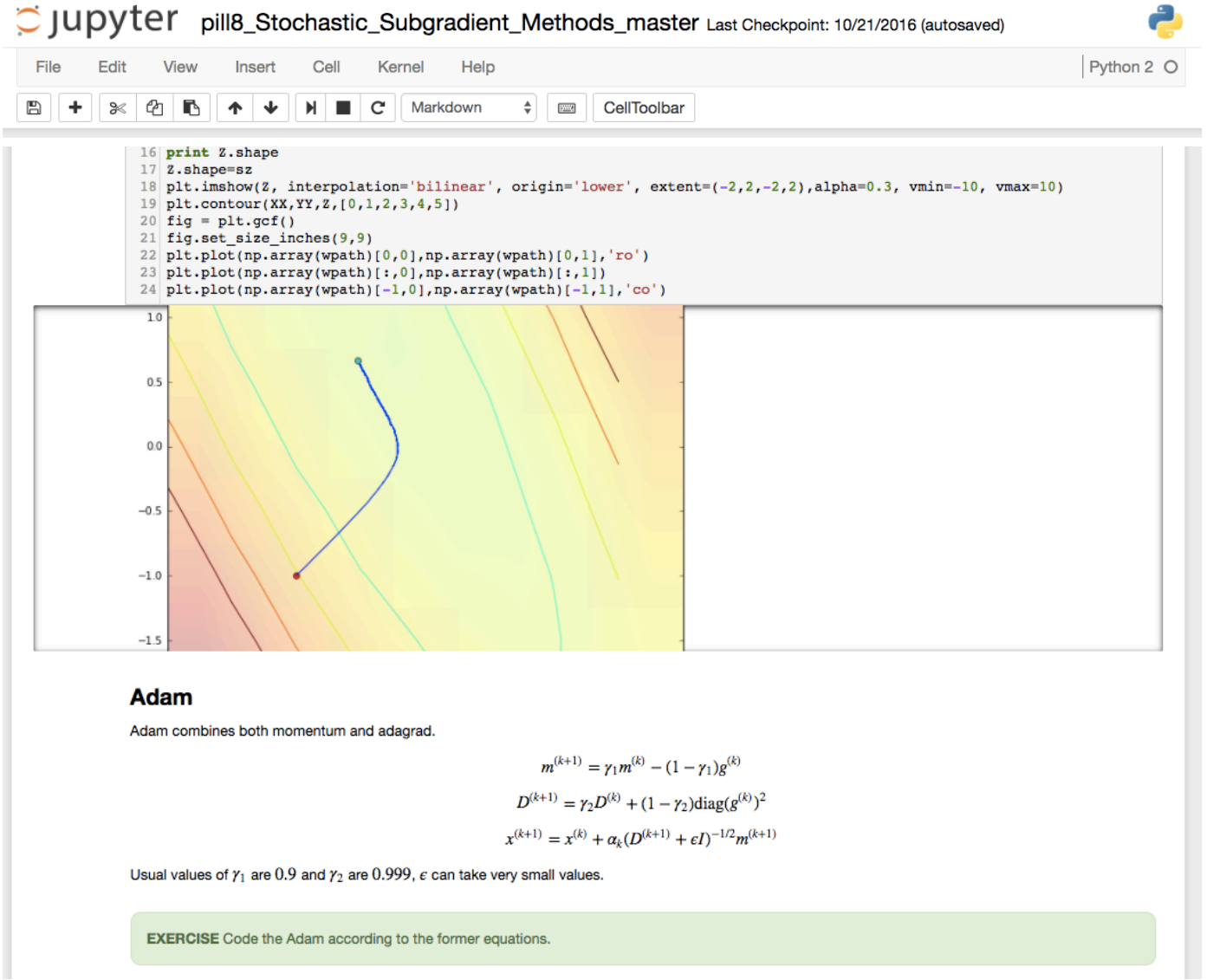
BUT … DEVIL IS IN THE DETAILS

DOCUMENTING THE DETAILS



JUPYTER NOTEBOOKS

EXECUTABLE CODE
LATEX
MARKDOWN
HTML+CSS
HYPERLINKS
EMBEDDED FRAMES
… and more.



PUTTING ALL TOGETHER

Competitions About Categories ▾ Search

GitXiv

Register Sign In **Post**

Receive the best of GitXiv right in your inbox. Your Email **Get Newsletter**

Collaborative Open Computer Science

View : Top New Best Daily

1

Deep Probabilistic Programming

Edward, a Turing-complete probabilistic programming language

BAYESIAN DEEP LEARNING (DL)

Christos Iraklis Tsatsoulis 1 point 5 hours ago **0 Comments**

2

OpenNMT: Open-Source Toolkit for Neural Machine Translation

Industrial-strength, open-source neural machine translation system

NATURAL LANGUAGE PROCESSING (NLP) RECURRENT NEURAL NETWORKS (RNN)

Christos Iraklis Tsatsoulis 1 point 6 days ago **0 Comments**

3

Improved Texture Networks: Quality & Diversity in Feed-forward Stylization & Texture Synthesis

Introducing Instance Normalization for better quality, entropy loss for diversity.

COMPUTER VISION CONVOLUTIONAL NEURAL NETWORKS (CNN) GENERATIVE

Dmitry Ulyanov 2 points 6 days ago **0 Comments**

SalGAN: Visual Saliency Prediction with Generative Adversarial Networks

Adversarial training improves a binary cross-entropy loss in most saliency metrics

ADVERSARIAL NETWORKS COMPUTER VISION CONVOLUTIONAL NEURAL NETWORKS (CNN) DEEP LEARNING (DL)

Xavier Giró-i-Nieto 2 points 13 days ago 0 Comments

This paper introduces the usage of generative adversarial networks (GANs) for visual saliency prediction. In this context, training is driven by two agents. First, the Generator that creates a synthetic sample matching the data distribution modelled by a training data set; second, the Discriminator, that distinguishes between a real sample drawn directly from the training data set and one created by the generator.

arXiv

We introduce SalGAN, a deep convolutional neural network for visual saliency prediction trained with adversarial examples. The first stage of the network consists of a generator model whose weights are learned by back-propagation computed from a binary cross entropy (BCE) loss over downsampled versions of the saliency maps. The resulting prediction is processed by a

GitHub

Our paper presents two convolutional neural networks, one corresponds to the Generator (Saliency Prediction Network) and the another is the Discriminator for the adversarial training. To compute saliency maps only the Generator is needed. SalGAN is implemented in Lasagne, which at its time is developed over Theano.

Links

- https://imatge-upc.github.io/saliency-salgan-2017/.
- http://www.slideshare.net/xavigiro/salgan-visual-saliency-prediction-with-generative-adversarial-networks.

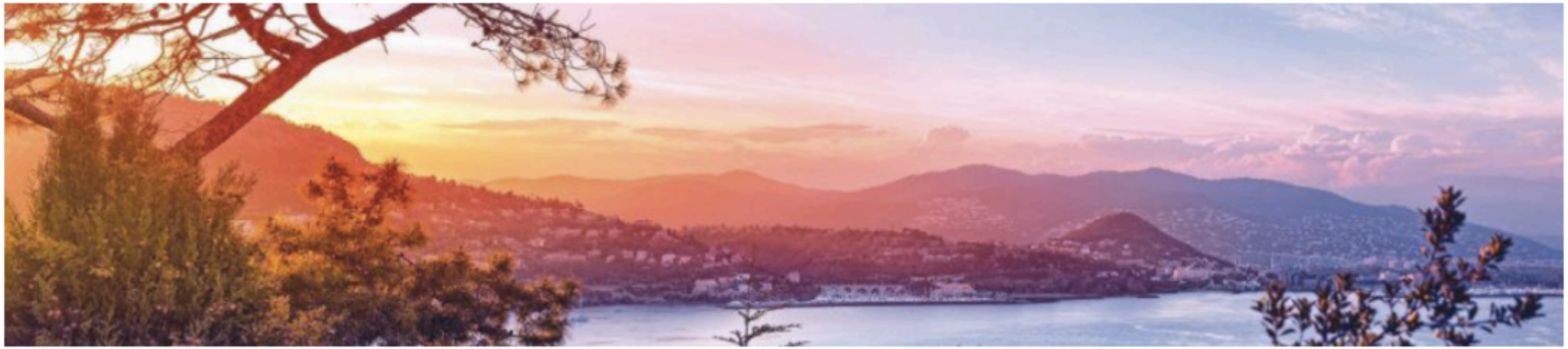
BOTTOM LINE

1. SCIENCE IS MORE THAN A PUBLICATION
2. AT ITS CORE LIES REPRODUCIBILITY…THIS INVOLVES DATA
3. BUT DATA IS NOT EVEN ENOUGH… CODE AND PRECISE EXPERIMENT REPRODUCIBILITY
4. DATA AND PROGRAMMING LITERACY IS NEEDED

BUT … SOMETHING ELSE ABOUT QUALITY ASSESSMENT: OPEN REVIEW CASE

ICLR 2017

MAIN | CONFERENCE CFP | WORKSHOP CFP | COMMITTEE | FAQ | OUR SPONSORS | OTHER YEARS



5th International Conference on Learning Representations

Overview

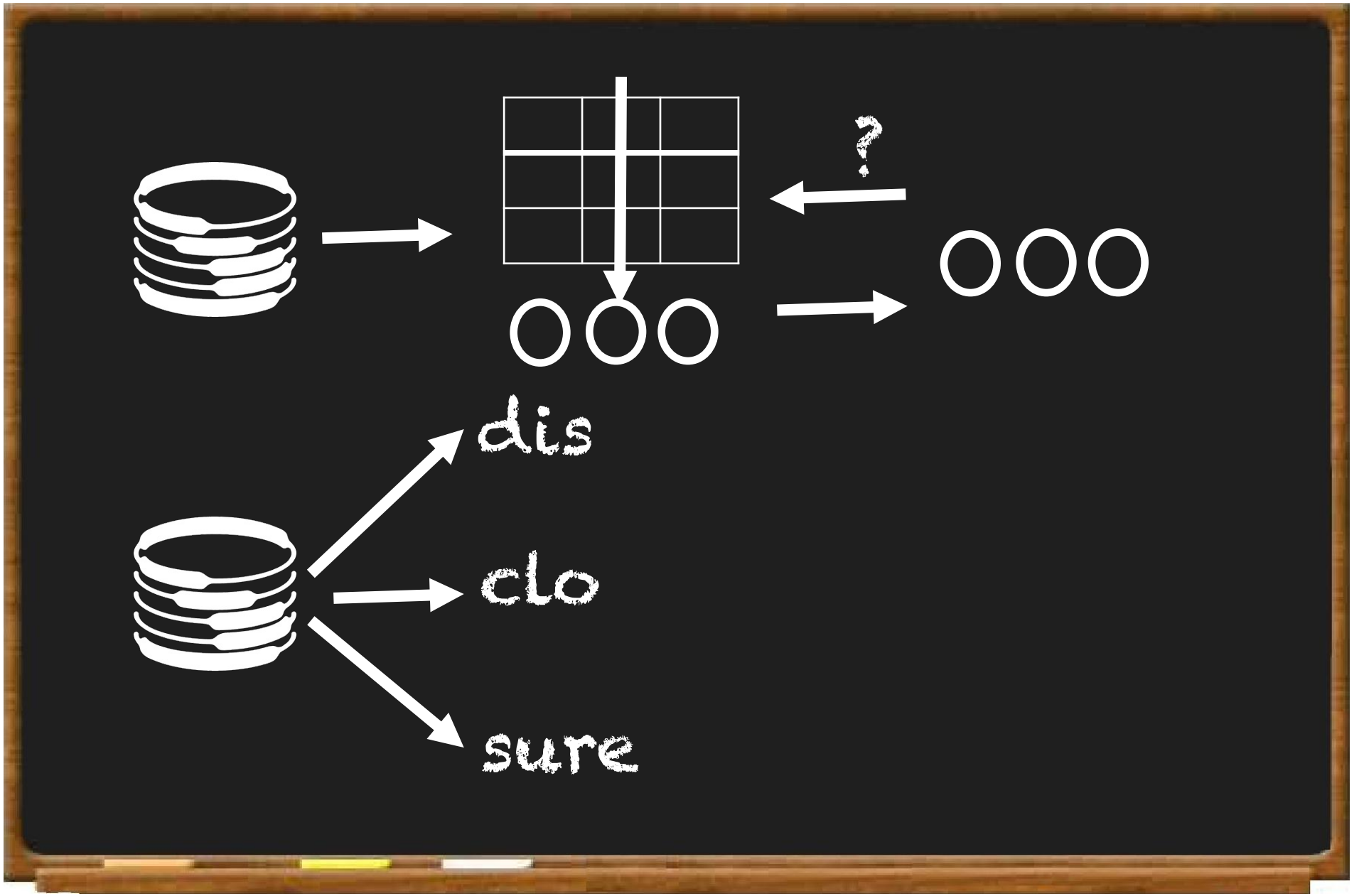
THE RULES

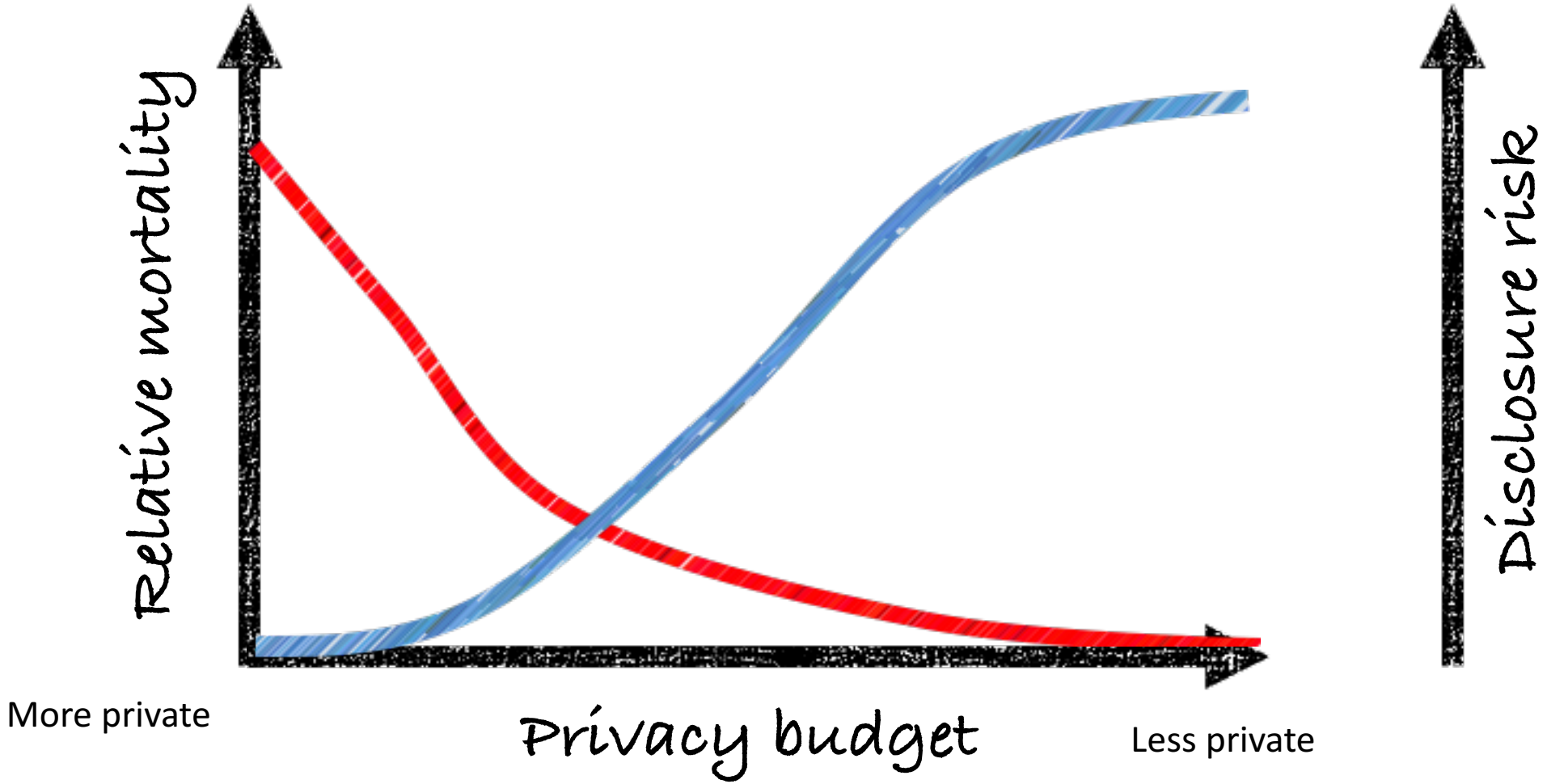
- Submissions posted on arXiv before being submitted to the conference
- Program committee assigns blind reviewers (as always) marked as designated reviewers
- Anyone can ask the program chairs for permission to become an anonymous designated reviewer (open bidding).
- Open commenters will have to use their real names, linked with their Google Scholar profiles.
- Papers that are presented in the workshop track or are not accepted will be considered non-archival, and may be submitted elsewhere (modified or not), although the ICLR site will maintain the reviews, the comments, and the links to the arXiv versions

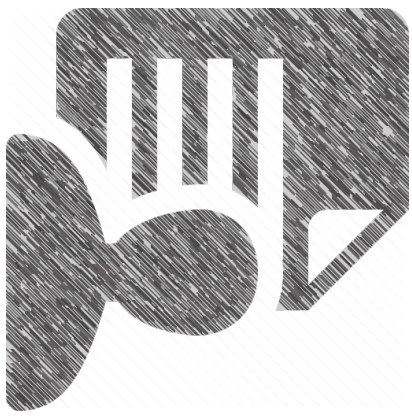
SOME POINTS ABOUT OPEN REVIEW

- Reading the answers from others is valuable
- Rejected papers have influence
- May become a popularity contest

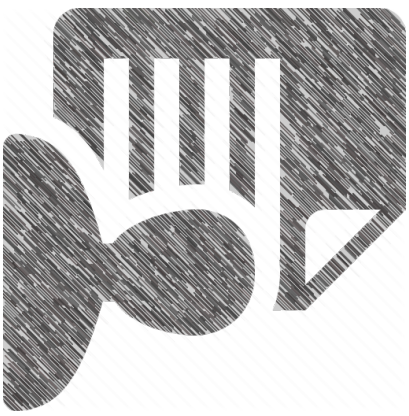
SHARING DATA IN THE AGE OF DEEP LEARNING


?
dis
clo
sure


Relative mortality
Disclosure risk
More private
Privacy budget
Less private

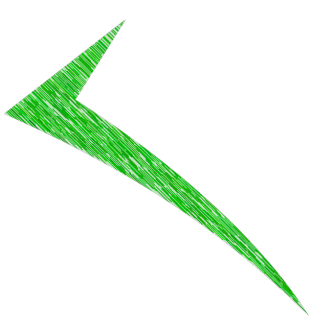


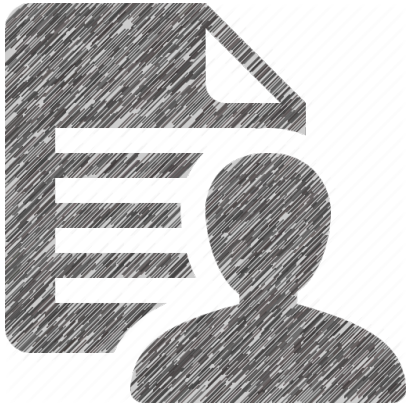
CLASSIFIER

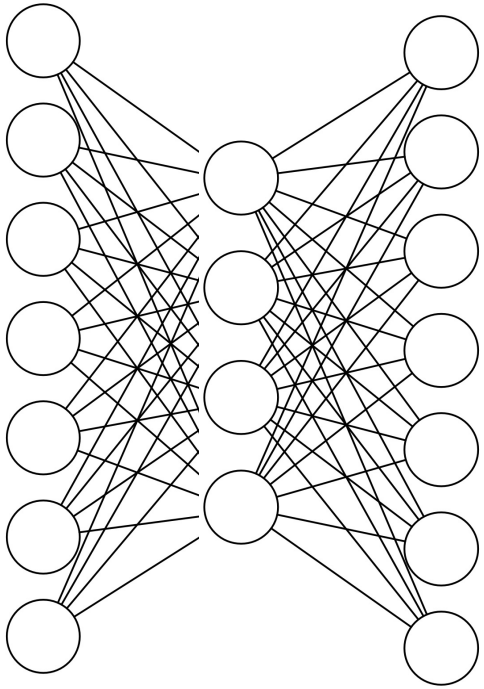


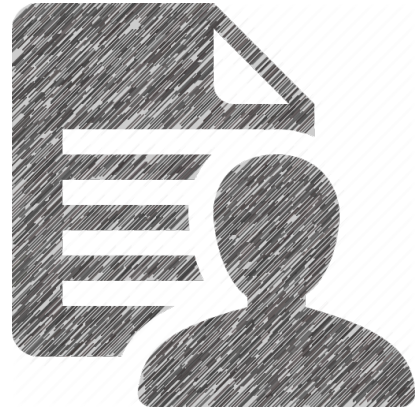
LEARNER

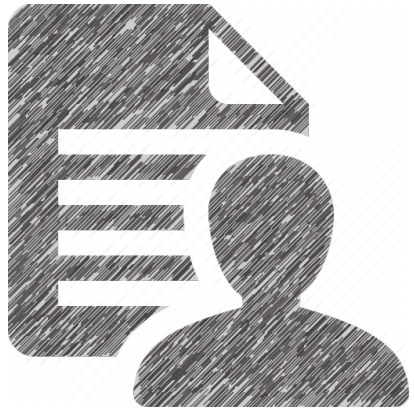
CLASSIFIER

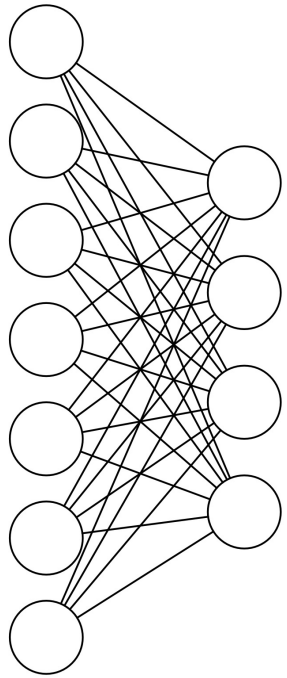


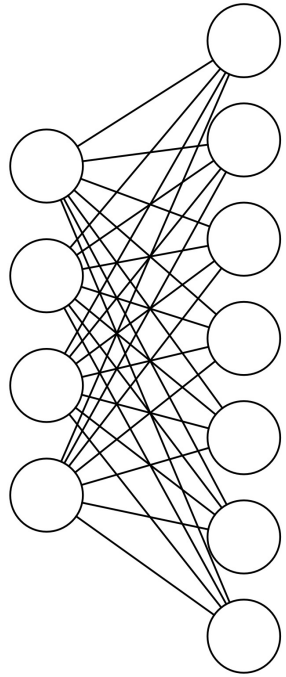


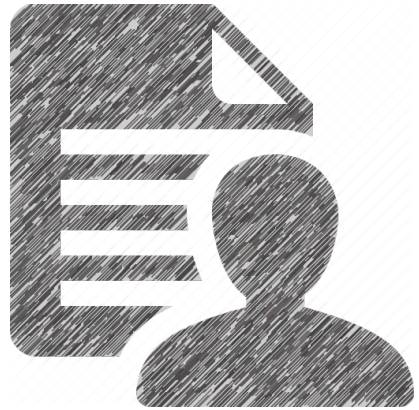


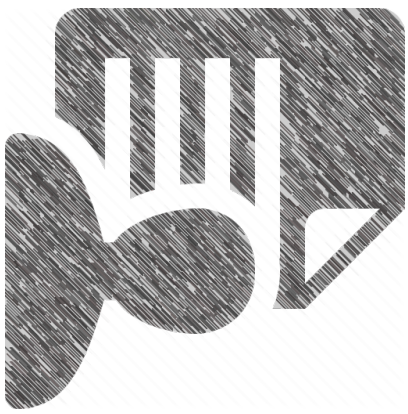


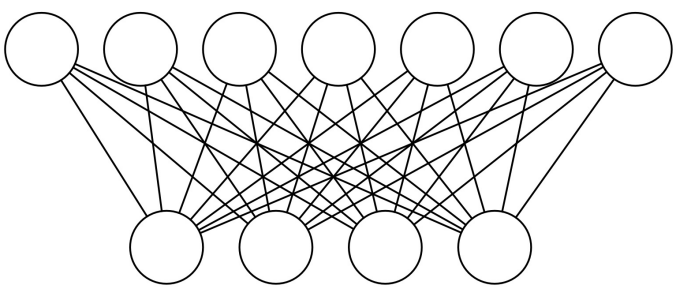




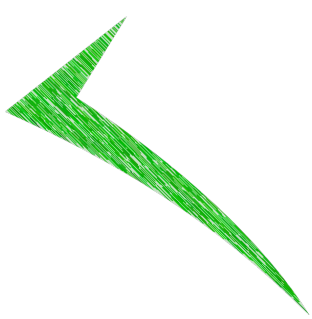


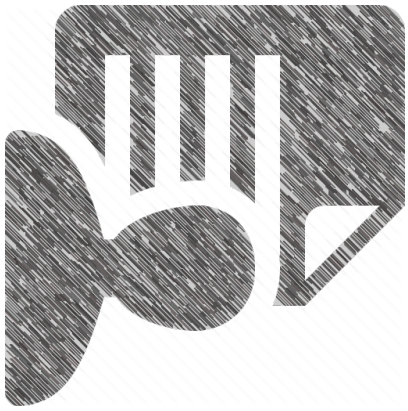


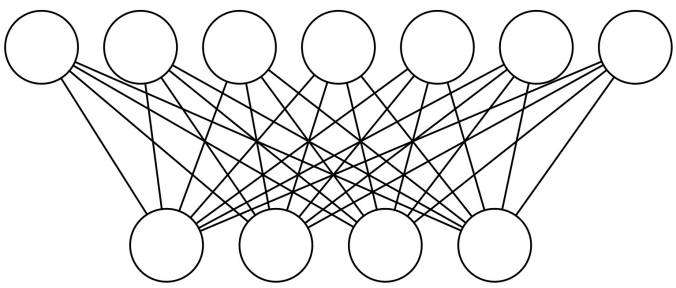




CLASSIFIER







CONCLUSIONS

- Research is based on reproducibility: documentation, data, code, and details.

- Machine learning methods can help in the challenge for democratizing data.

Thank you

oriol_pujol@ub.edu

@oriolpujolvila

UNIVERSITAT DE BARCELONA

learn