Should this have been the first session?

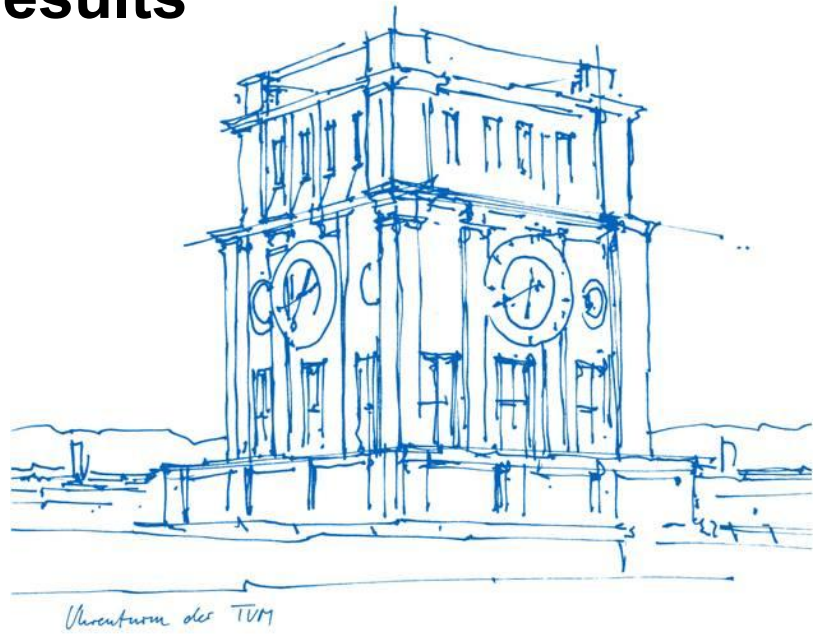
Imperial College London

Model development part 3: evaluation and visualization of results

Martin Menten (martin.menten@tum.de)

AFRICAI/ MICCAI summer school

14th September 2023



Hypothesis-driven research

- Research question:

~~"My method is very good at solving a challenging task."~~

- Testable hypothesis:

"My method is better than the existing state-of-the-art at solving a challenging task."

Hypothesis-driven research

- Deciding on your research hypothesis shapes your entire paper

Methods sections - evaluation metrics

Related works

”My method is better than the existing state-of-the-art at solving a challenging task.’

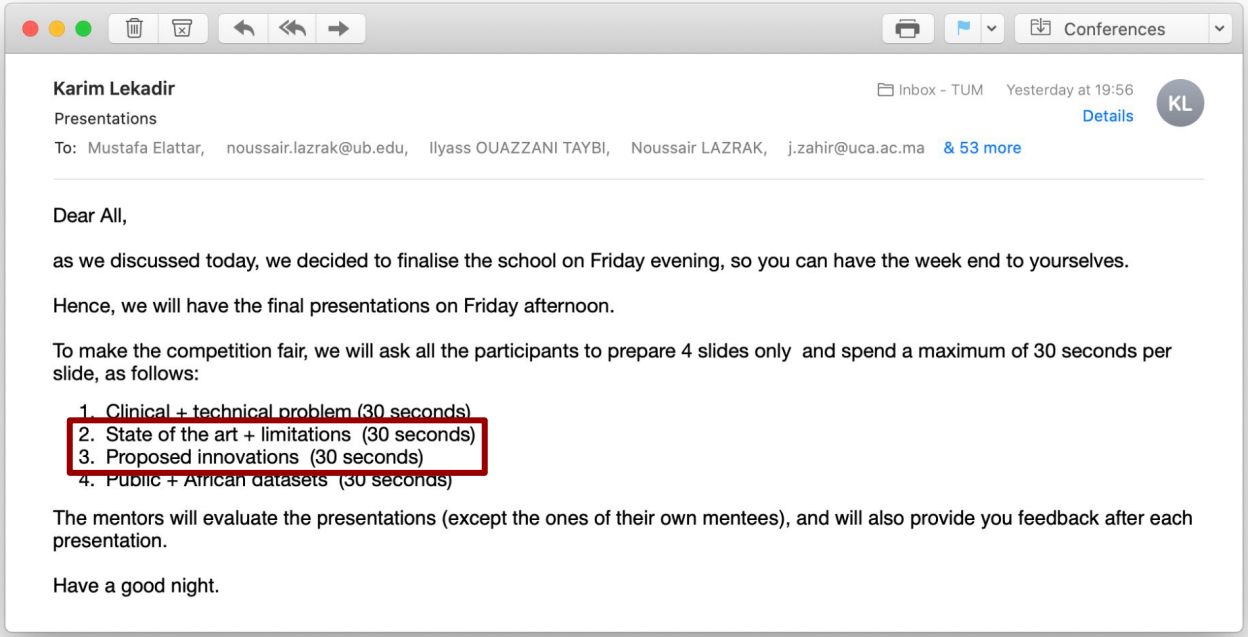
Methods section - baselines

Results

Introduction

Hypothesis-driven research

- Extra motivation by Karim



Conferences

Karim Lekadir

Inbox - TUM Yesterday at 19:56

Details

KL

Presentations

To: Mustafa Elattar, noussair.lazrak@ub.edu, Ilyass OUAZZANI TAYBI, Noussair LAZRAK, j.zahir@uca.ac.ma & 53 more

Dear All,

as we discussed today, we decided to finalise the school on Friday evening, so you can have the week end to yourselves.

Hence, we will have the final presentations on Friday afternoon.

To make the competition fair, we will ask all the participants to prepare 4 slides only and spend a maximum of 30 seconds per slide, as follows:

1. Clinical + technical problem (30 seconds)
2. State of the art + limitations (30 seconds)
3. Proposed innovations (30 seconds)
4. Public + African datasets (30 seconds)

The mentors will evaluate the presentations (except the ones of their own mentees), and will also provide you feedback after each presentation.

Have a good night.

Implement your evaluation pipeline early

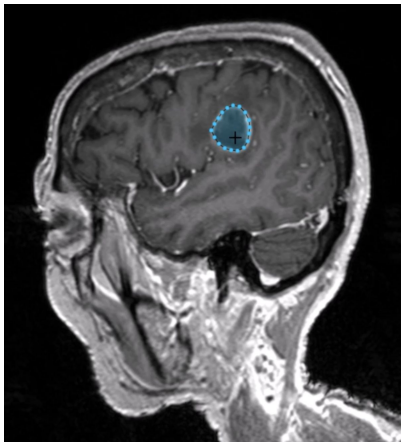
- Provides a quantitative measure where you are with your research
- Helps you understand the strengths and limitations of the baselines
- Culls unpromising research ideas
- Focuses your writing

Selecting evaluation metrics - starting point

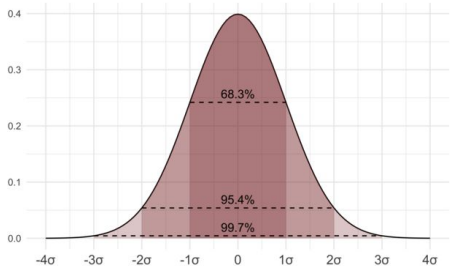
- Check how others have been evaluating their methods
- Decent chance that they know what they are doing
- Intrinsic comparability with other methods

Selecting evaluation metrics - clinical relevance

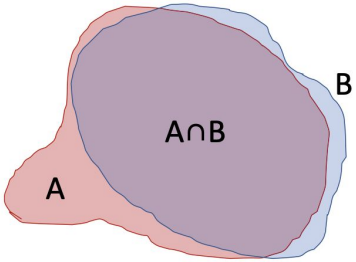
- Envisioned clinical use case may inform focus of evaluation



Brain tumor segmentation

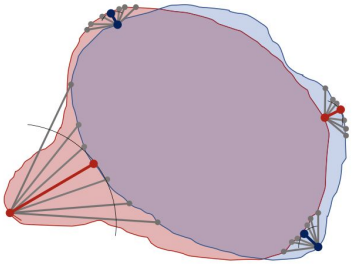


Volume quantification



Dice similarity coefficient

Surgical guidance



Hausdorff distance

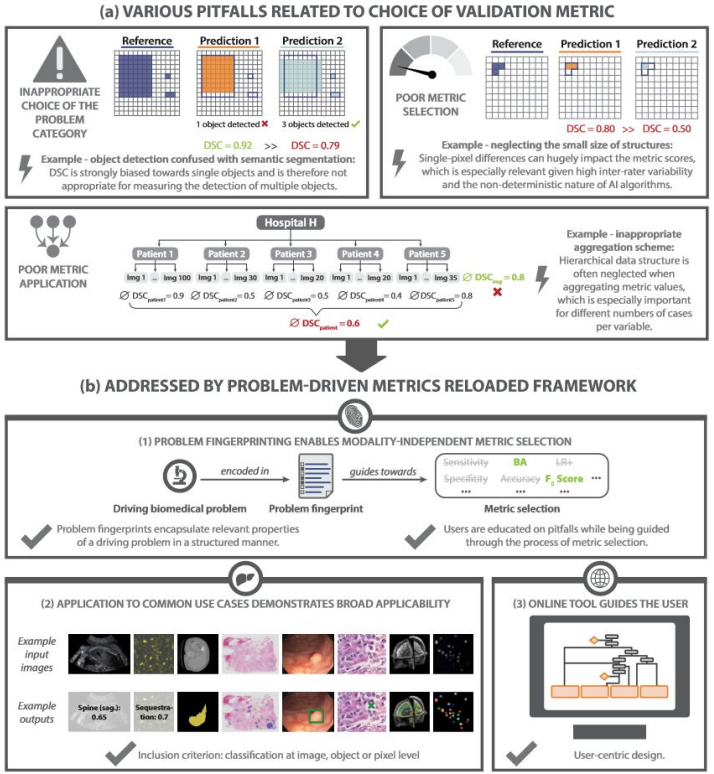
Left images taken from www.wikimedia.org
Right figure taken from Kiser *et al.*, CTRO, 2019

Selecting evaluation metrics - expert consensus

- Metrics reloaded

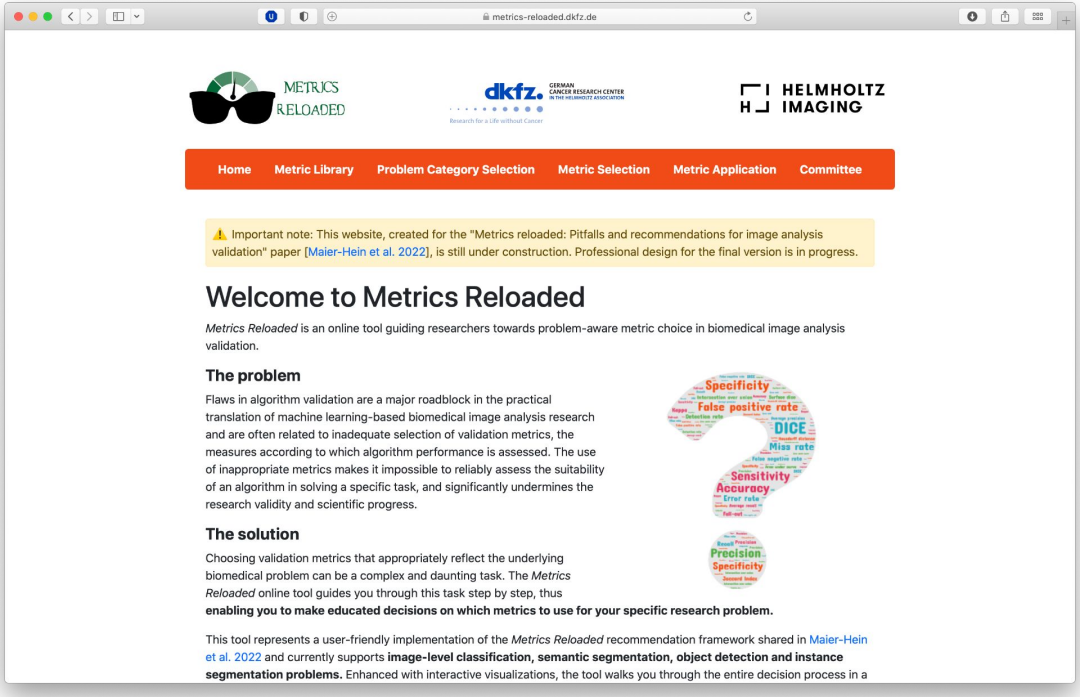
Metrics Reloaded: Recommendations for image analysis validation

LENA MAIER-HEIN[+†], German Cancer Research Center (DKFZ), Germany, Heidelberg University, Germany, and National Center for Tumor Diseases (NCT), Germany
ANNIKA REINKE[†], German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
PATRICK GODAU, German Cancer Research Center (DKFZ), Germany, Heidelberg University, Germany, and National Center for Tumor Diseases (NCT), Germany
MINU D. TIZABI, German Cancer Research Center (DKFZ), Germany and National Center for Tumor Diseases (NCT), Germany
FLORIAN BÜTTNER, German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Germany, Goethe University Frankfurt, Germany, and Frankfurt Cancer Institute, Germany
EVANGELIA CHRISTODOULOU, German Cancer Research Center (DKFZ), Germany
BEN GLOCKER, Imperial College London, UK
FABIAN ISENSEE, German Cancer Research Center (DKFZ), Germany
JENS KLEESIEK, University Medicine Essen, Germany
MICHAL KOZUBEK, Masaryk University, Czech Republic
MAURICIO REYES, University of Bern, Switzerland
MICHAEL A. RIEGLER, Simula Metropolitan Center for Digital Engineering, Norway and UiT The Arctic University of Norway, Norway
MANUEL WIESENFARTH, German Cancer Research Center (DKFZ), Germany
A. EMRE KAVUR, German Cancer Research Center (DKFZ), Germany
CAROLE H. SUDRE, University College London, UK and King's College London, UK
MICHAEL BAUMGARTNER, German Cancer Research Center (DKFZ), Germany
MATTHIAS EISENMANN, German Cancer Research Center (DKFZ), Germany
DOREEN HECKMANN-NÖTZEL, German Cancer Research Center (DKFZ), Germany and National Center for Tumor Diseases (NCT), Germany
TIM RÄDSCH, German Cancer Research Center (DKFZ), Germany
LAURA ACION, CONICET – Universidad de Buenos Aires, Argentina
MICHELA ANTONELLI, King's College London, UK and University College London, UK
TAL ARBEL, McGill University, Canada
SPYRIDON BAKAS, University of Pennsylvania, USA and Perelman School of Medicine at the University of Pennsylvania, USA
ARRIEL BENIS, Holon Institute of Technology, Israel and European Federation for Medical Informatics, Switzerland
MATTHEW BLASCHKO, KU Leuven, Belgium
M. JORGE CARDOSO, King's College London, UK
VERONIKA CHEPLYGINA, IT University of Copenhagen, Denmark
BETH A. CIMINI, Broad Institute of MIT and Harvard, USA
GARY S. COLLINS, University of Oxford, UK
KEYVAN FARAHANI, National Cancer Institute, USA
LUCIANA FERRER, CONICET-UBA, Argentina
ADRIAN GALDRAN, Universitat Pompeu Fabra, Spain and University of Adelaide, Australia
BRAM VAN GINNEKEN, Fraunhofer MEVIS, Germany and Radboud University Medical Center, The Netherlands
ROBERT HAASE, DFG Cluster of Excellence "Physics of Life", Germany and Center for Systems Biology, Germany
DANIEL A. HASHIMOTO, University of Pennsylvania, USA, and Perelman School of Medicine at University of Pennsylvania, USA
MICHAEL M. HOFFMAN, University Health Network, Canada, University of Toronto, Canada, and Vector Institute, Canada
MEREL HUISMAN, Radboud University Medical Center, The Netherlands
PIERRE JANNIN, Université de Rennes 1, France and INSERM, France
CHARLES E. KAHN, University of Pennsylvania, USA



Selecting evaluation metrics - expert consensus

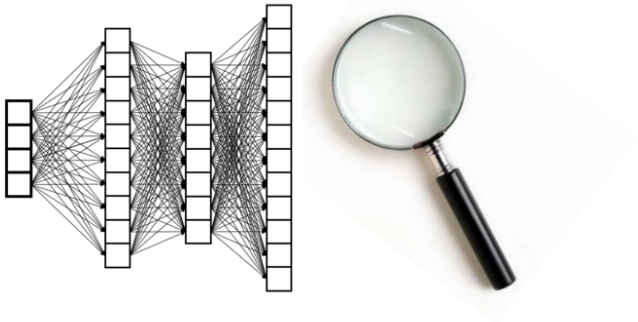
- Metrics reloaded



Beyond correctness

- A method may be beneficial in other aspect than correctness:

Required human effort



Explainability

Fairness

Cost

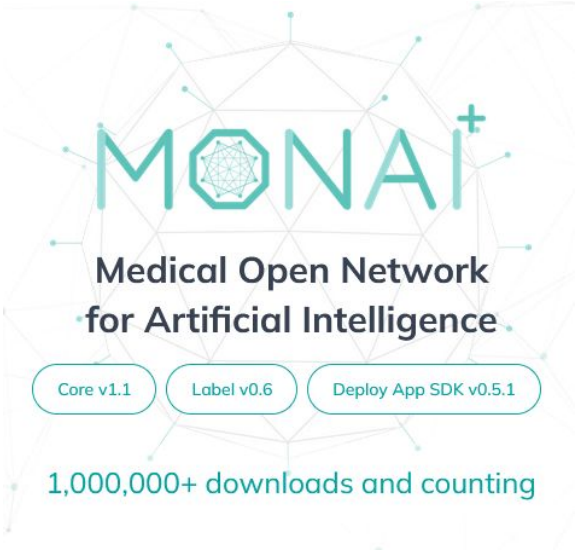
Run time

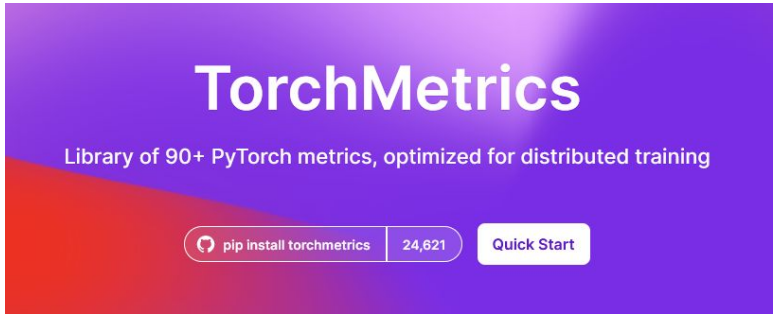
Carbon footprint

Images taken from www.wikimedia.org

Tips when implementing metrics

- Beware of implementation errors
- Save results in “raw format”
  in case you want to add other metrics later
- Use established libraries





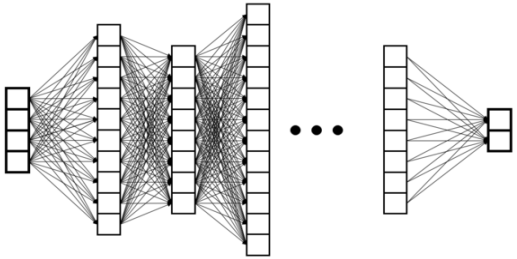
Picking baselines

Best performing methods

nnu-net: Self-adapting framework for u-net-based medical image segmentation
F Isensee, J Petersen, A Klein, D Zimmerer… - arXiv preprint arXiv …, 2018 - arxiv.org
… For each task the **nnU-Net** automatically runs a five-fold cross-validation for three different
… In the context of the Medical Segmentation Decathlon we demonstrate that the **nnU-Net** …
☆ Speichern 99 Zitieren Zitiert von: 708 Ähnliche Artikel Alle 6 Versionen »

Widely established algorithms

Methods with good public code



Ablated versions of proposed method

OpenCV

Pre-deep learning algorithms

All baselines have to be tuned!

Images taken from www.wikimedia.org

Overfitting - case study 1

- Task: classification of skin lesion based on smart phone photographs
- Approach: train a convolutional neural network on all data and report performance

Problem: classical overfitting of network parameters

Images taken from www.wikimedia.org

Overfitting

- Large enough models are able to memorize the entire dataset
- Model does not need to learn general patterns in the data

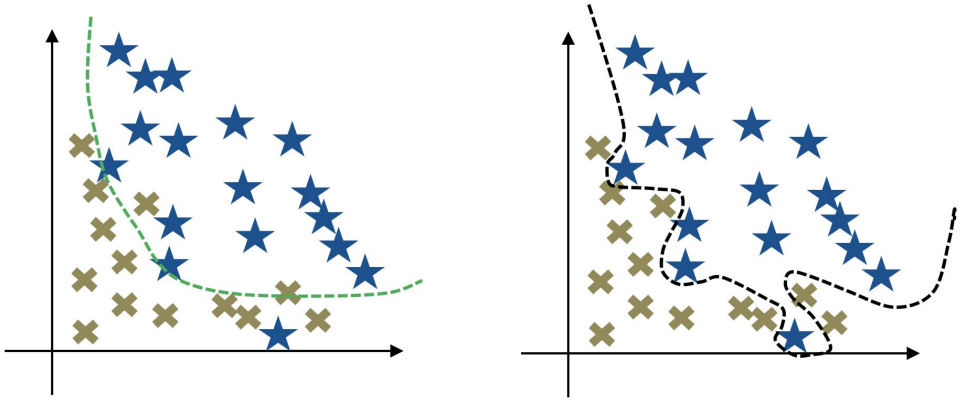
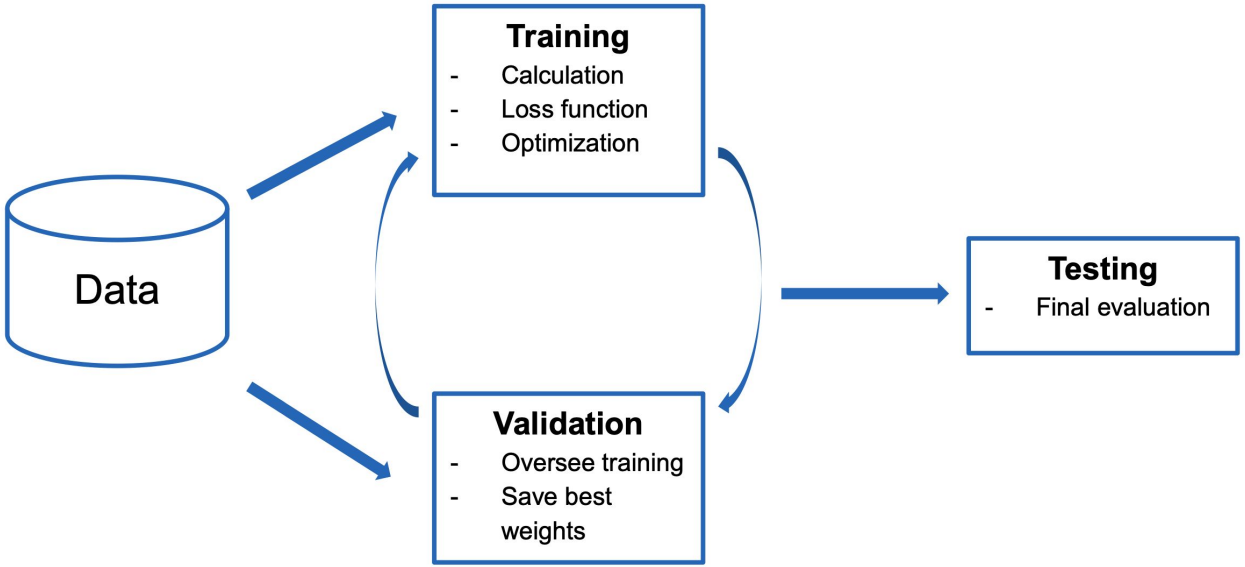


Figure taken from Buduma - fundamentals of deep learning

Overfitting

- Splitting data into training, validation and testing dataset



Overfitting - case study 2

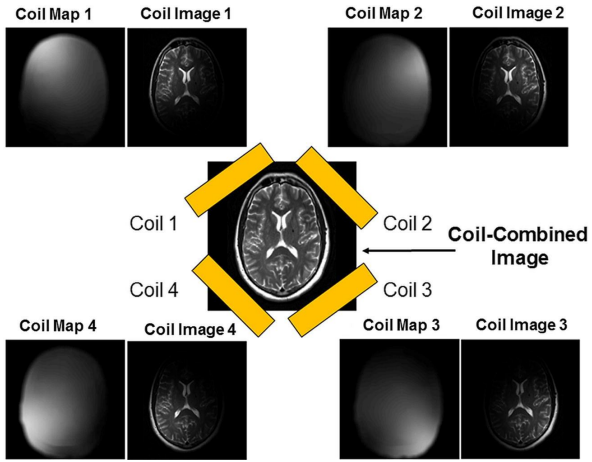
- Task: classification of skin lesion based on smart phone photographs
- Approach: train a convolutional neural network on test and validation data
- After poor test performance, add two additional convolutional blocks to your network

Problem: hyperparameter tweaking based on test performance

Image taken from www.wikimedia.org

Overfitting - case study 3

- Task: reconstruction of MR images from parallel imaging data
- Approach: train reconstruction neural network on large dataset of healthy volunteers
- Reconstruction fails when deployed on clinical data

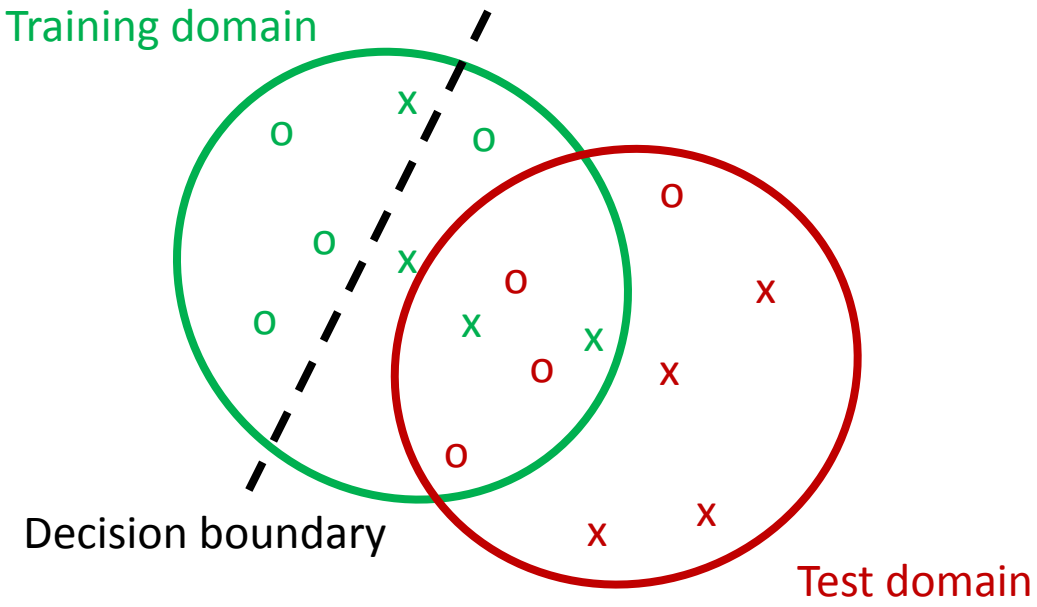


Problem: scans in training dataset feature less artifacts and may be anatomically different

Figure taken from Hamilton *et al.*, Prog Nucl Magn Reson Spectrosc, 2017

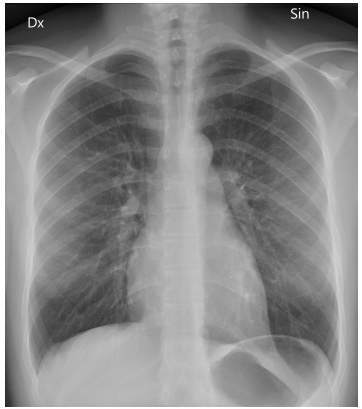
Domain shifts

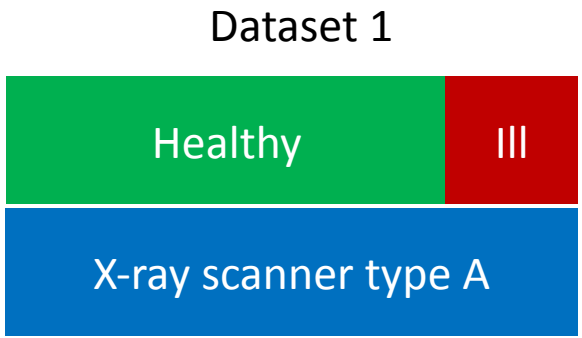
- A domain shift occurs when a machine learning algorithm is trained and deployed on datasets with different characteristics

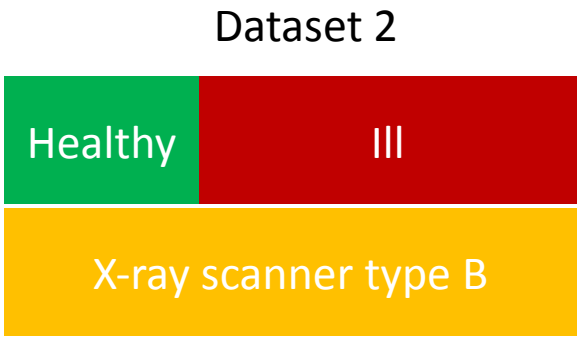


Overfitting - case study 4

- Task: forecast severity of Covid-19 infection from chest radiographs
- Approach: In order to overcome small dataset size,
  scans from different clinics are combined for model training







Problem: the algorithm learns to predict scanner type instead of disease markers

Image taken from www.wikimedia.org

Overfitting - case study 5

- Task: predicting mortality of intensive care unit patients
- Approach: utilize rich dataset consisting of demographic data, vital sign measurements, laboratory tests, medication, caregiver notes, …
- Algorithm fails to generalize well

Problem: the presence of a test or intervention, rather than its outcome already conveys information

Image taken from www.wikimedia.org

Statistical tests – advantages and disadvantages

Pros

- Good scientific practice
- Makes your presented results look more impressive
- Most statistical tests used in MICCAI research can be calculated rather quickly

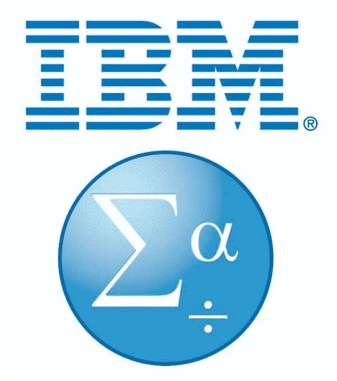
Cons

- Requires expertise
- Potential source for errors
- Limited sensitivity when using a small number of samples

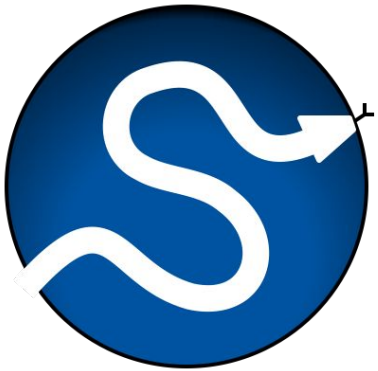
Statistical tests – software and libraries

- Dedicated statistics software is often more advanced than Python toolboxes
- Still, scipy.stats covers most tests needed for MICCAI research

R programming language



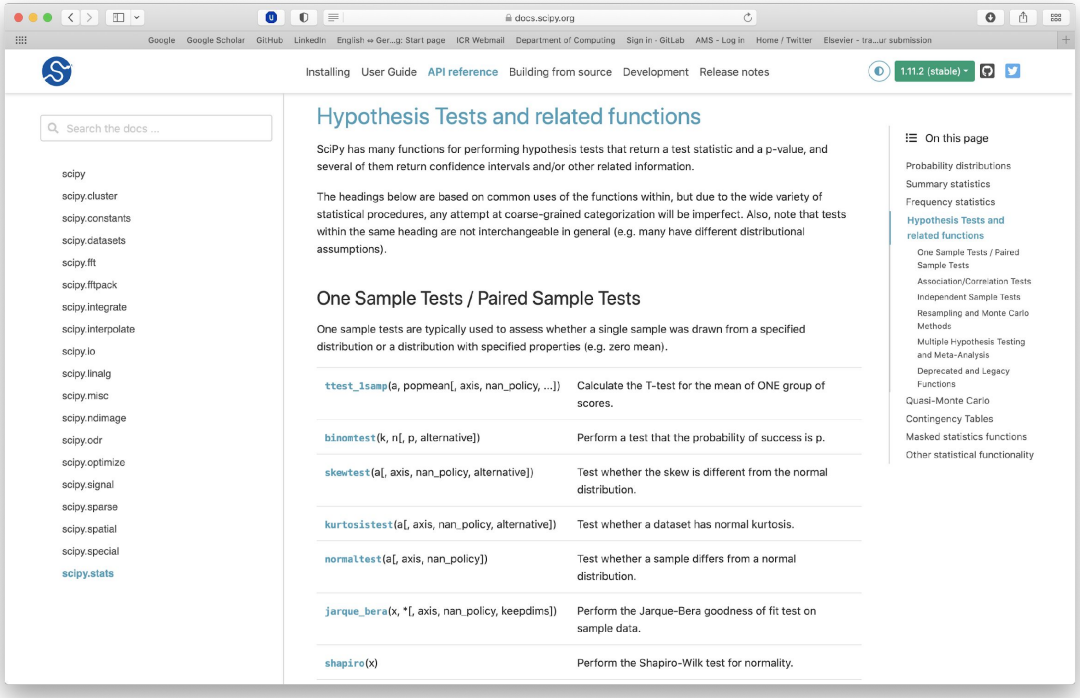
SPSS



SciPy

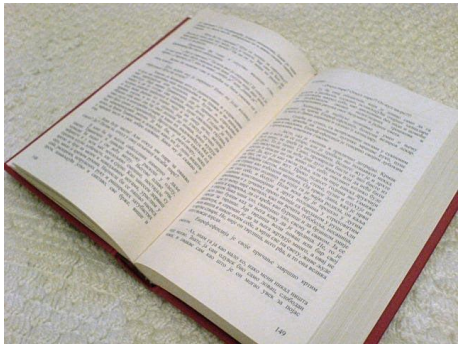
Statistical tests – software and libraries

- Dedicated statistics software is often more advanced than Python toolboxes
- Still, scipy.stats covers most tests needed for MICCAI research



Statistical tests – how to get started

Enroll in a university course



Read a statistics book

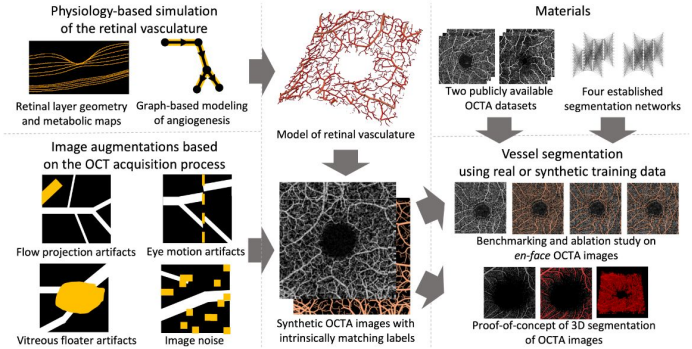
Check what others have done

Ask a friend or colleague

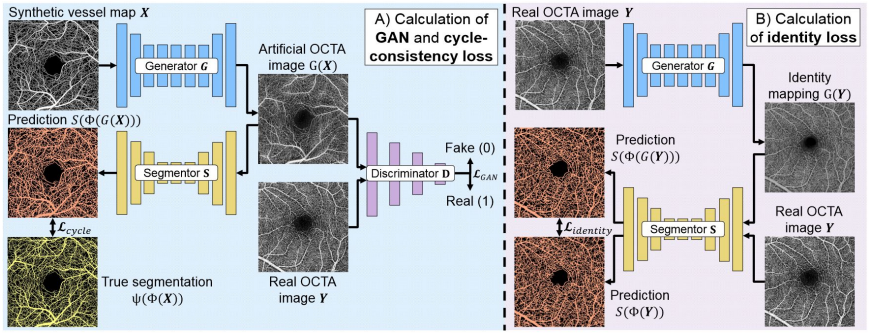
Browse the internet

Images taken from www.wikimedia.org

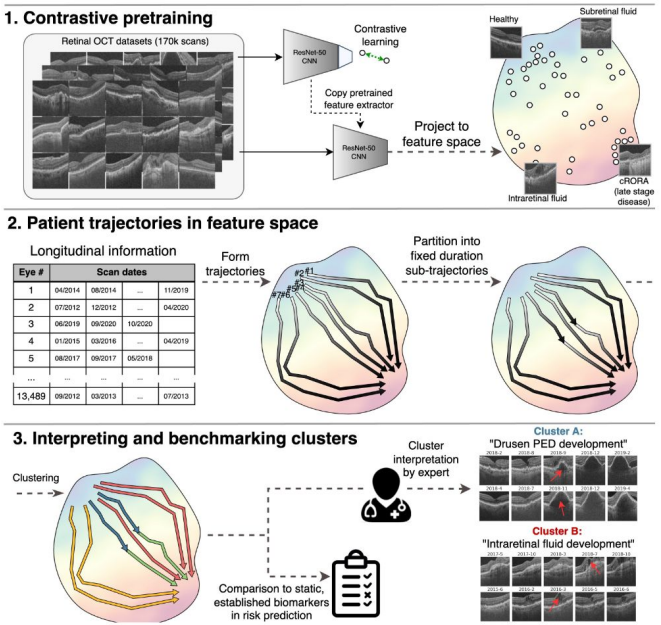
Types of scientific figures - introductory and methods figures



Graphical abstract



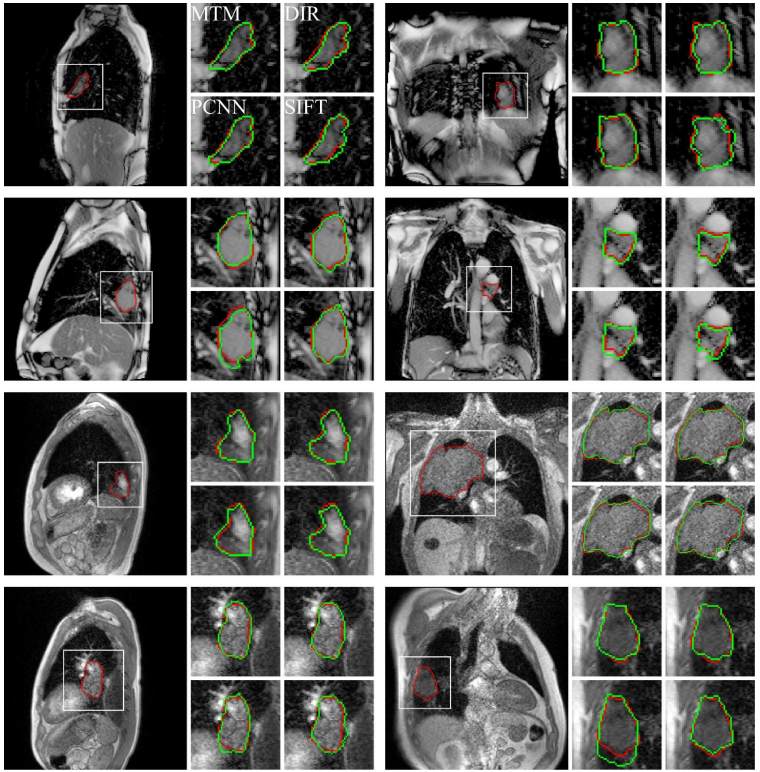
Methods figure



Study flowchart

Types of scientific figures - qualitative results

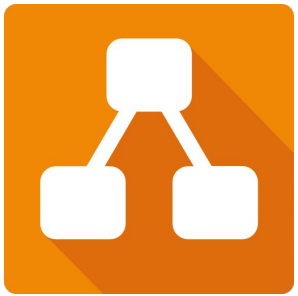
- Provides intuition about methods' strengths and weaknesses
- Consistently crop and set contrast
- Use high-enough image resolution
- Mention whether you are presenting best, worst, representative or randomly selected samples



WYSIWYG tools

- What-you-see-is-what-you get makes it easier to prototype
- Rapid iteration of figures may be needed

PowerPoint

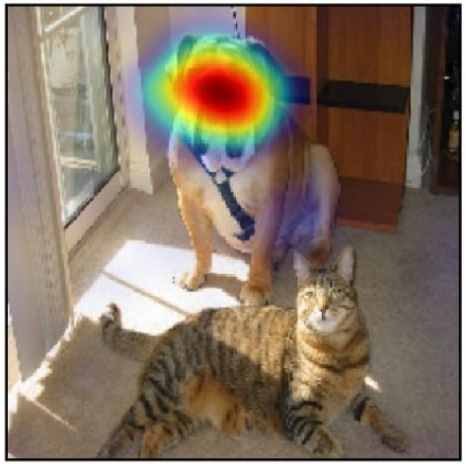


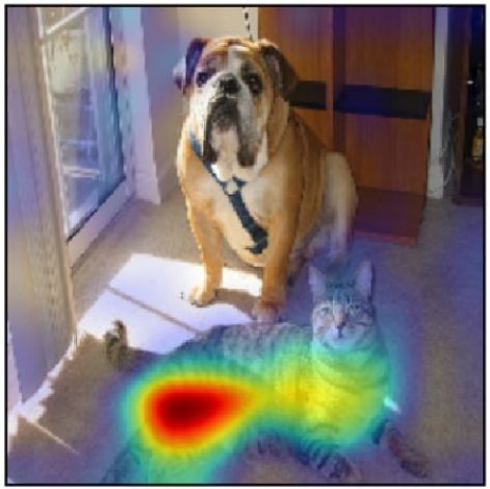
Draw.io

Gimp

Types of scientific figures - saliency maps

- Visualization of image features used by deep neural networks
- Most established on is GradCAM





Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

Abstract We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent and explainable.

Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (*e.g.* VGG), (2) CNNs used for structured outputs (*e.g.* captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g.* visual question answering) or reinforcement learning, all *without architectural changes or re-training*. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures.

In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are robust to adversarial perturbations, (d) are more faithful to the underlying model, and (e) help achieve model generalization by identifying dataset bias.

For image captioning and VQA, our visualizations show that even non-attention based models learn to localize discriminative regions of input image.

We devise a way to identify important neurons through Grad-CAM and combine it with neuron names [4] to provide textual explanations for model decisions. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions. Our code is available at https://github.com/ramprs/grad-cam/, along with a demo on CloudCV [2][1], and a video at youtu.be/COjUB9Izk6E.

Ramprasaath R. Selvaraju
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: ramprs@gatech.edu

Michael Cogswell
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: cogswell@gatech.edu

Abhishek Das
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: abhshkdz@gatech.edu

Ramakrishna Vedantam
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: vrama@gatech.edu

Devi Parikh
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: parikh@gatech.edu

Dhruv Batra
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: dbatra@gatech.edu

1 Introduction

Deep neural models based on Convolutional Neural Networks (CNNs) have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification [33,24], object detection [21], semantic segmentation [37] to image captioning [55,7,18,29], visual question answering [3,20,42,46] and more recently, visual dialog [11, 13, 12] and embodied question answering [10,23]. While

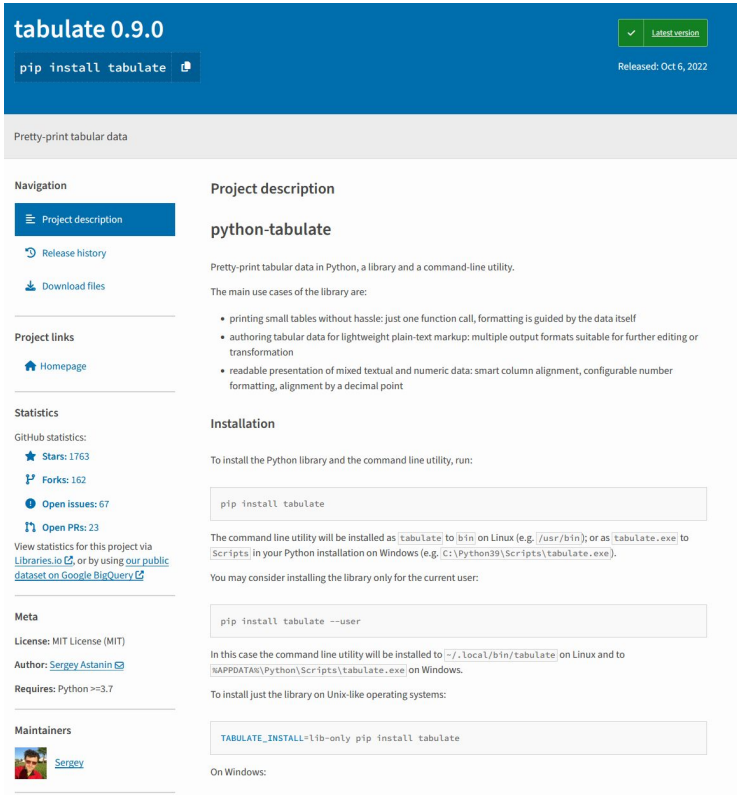
[1] http://gradcam.cloudcv.org

Types of scientific figures - the big results table

OCTA-500	Model	Trained on	Augmentation	DSC	clDice	AUC	ACC	Recall	Precision
Supervised	U-Net	Real	Random	**0.912±0.001**	**0.940±0.002**	**0.950±0.001**	**0.984±0.000**	**0.910±0.003**	**0.916±0.003**
Traditional	Frangi	-	-	0.807±0.003	0.848±0.004	0.895±0.002	**0.975±0.000**	0.802±0.005	0.820±0.003
	OOF			0.734±0.004	0.785±0.003	0.851±0.004	0.966±0.001	0.719±0.009	0.760±0.006
Menten *et al.*	U-Net	Synthetic	Random	0.594±0.011	0.536±0.014	**0.936±0.002**	0.913±0.004	**0.963±0.001**	0.434±0.011
Ours	U-Net	Synthetic	-	0.734±0.006	0.731±0.009	0.827±0.005	0.968±0.001	0.663±0.010	0.827±0.003
			Random	**0.849±0.001**	**0.892±0.002**	0.926±0.002	0.971±0.000	0.871±0.005	0.831±0.004
			Adversarial	0.840±0.003	0.881±0.002	0.915±0.003	0.970±0.000	0.848±0.006	0.837±0.004
			GAN	0.840±0.005	0.876±0.006	0.907±0.002	0.971±0.001	0.829±0.005	**0.855±0.010**
ROSE-1	**Model**	**Trained on**	**Augmentation**	**DSC**	**clDice**	**AUC**	**ACC**	**Recall**	**Precision**
Supervised	U-Net	Real	Random	**0.717±0.004**	**0.700±0.006**	**0.799±0.003**	0.843±0.003	**0.678±0.006**	**0.774±0.005**
Traditional	Frangi	-	-	**0.670±0.003**	0.601±0.007	**0.789±0.003**	**0.875±0.002**	**0.644±0.009**	0.704±0.012
	OOF			0.594±0.003	0.559±0.003	0.752±0.004	0.831±0.006	0.623±0.016	0.578±0.018
Menten *et al.*	U-Net	Synthetic	Random	0.606±0.004	0.534±0.005	0.746±0.003	0.861±0.001	0.556±0.008	0.673±0.004
Ours	U-Net	Synthetic	-	0.613±0.004	0.546±0.004	0.743±0.002	**0.870±0.002**	0.532±0.003	0.731±0.006
			Random	0.649±0.001	0.653±0.002	0.754±0.001	0.806±0.001	0.617±0.002	0.699±0.002
			Adversarial	0.665±0.004	**0.658±0.005**	0.756±0.003	0.816±0.003	0.601±0.006	**0.734±0.002**
			GAN	0.637±0.005	0.639±0.007	0.745±0.003	0.804±0.004	0.591±0.005	0.705±0.005
Giarratano *et al.*	**Model**	**Trained on**	**Augmentation**	**DSC**	**clDice**	**AUC**	**ACC**	**Recall**	**Precision**
Supervised	U-Net	Real	Random	**0.907±0.002**	**0.954±0.003**	**0.887±0.004**	**0.895±0.002**	**0.925±0.003**	**0.981±0.006**
Traditional	Frangi	-	-	0.769±0.009	0.833±0.007	0.815±0.002	0.797±0.003	**0.895±0.010**	0.683±0.012
	OOF			0.812±0.005	0.848±0.006	**0.851±0.003**	**0.854±0.002**	0.808±0.017	0.827±0.019
Menten *et al.*	U-Net	Synthetic	Random	0.471±0.014	0.496±0.012	0.651±0.017	0.710±0.011	0.318±0.014	**0.937±0.004**
Ours	U-Net	Synthetic	-	0.781±0.006	0.808±0.008	0.827±0.009	0.812±0.003	0.887±0.003	0.710±0.012
			Random	**0.850±0.005**	**0.933±0.002**	0.846±0.001	0.839±0.001	0.829±0.003	0.879±0.003
			Adversarial	0.842±0.005	0.931±0.008	0.840±0.004	0.833±0.006	0.814±0.004	0.881±0.010
			GAN	0.834±0.002	0.902±0.004	0.829±0.002	0.822±0.002	0.828±0.004	0.853±0.006

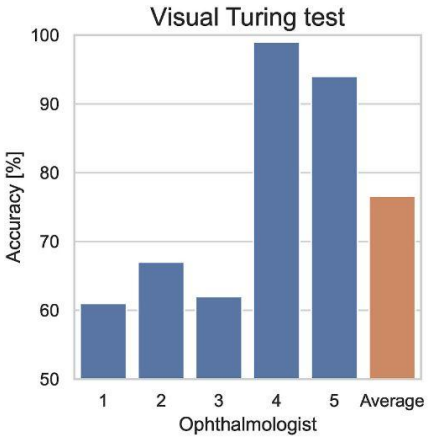
Types of scientific figures - the big results table

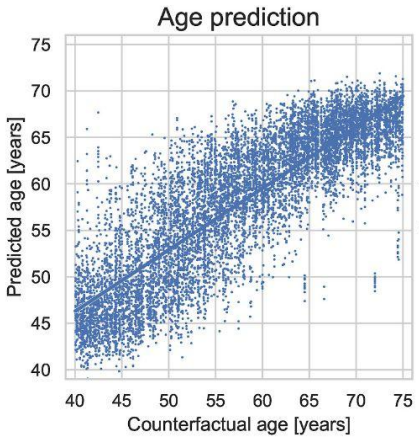
- Usually done in LaTeX
- Sparingly use separating lines
- Add hierarchy with multicolumn and/ or multirow if needed
- Consistent number of digits
- Align numbers
- Remember to update table with new results

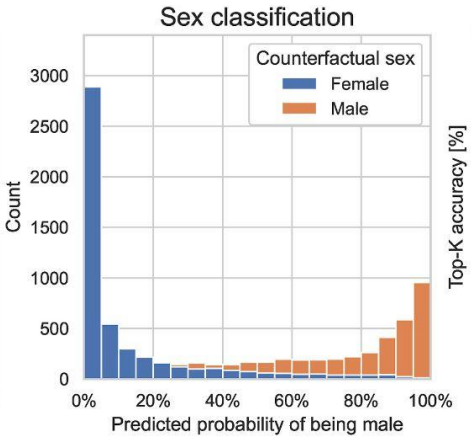


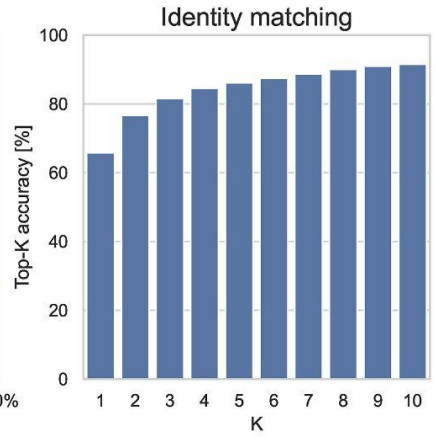
Types of scientific figures - quantitative results figures

- Conveys high-dimensional data in an intuitive way
- Type of plot can be used to emphasize different aspects of the findings









Python plotting libraries

- Dedicated libraries for scientific figure plotting
- Can be directly integrated with the output of your algorithm

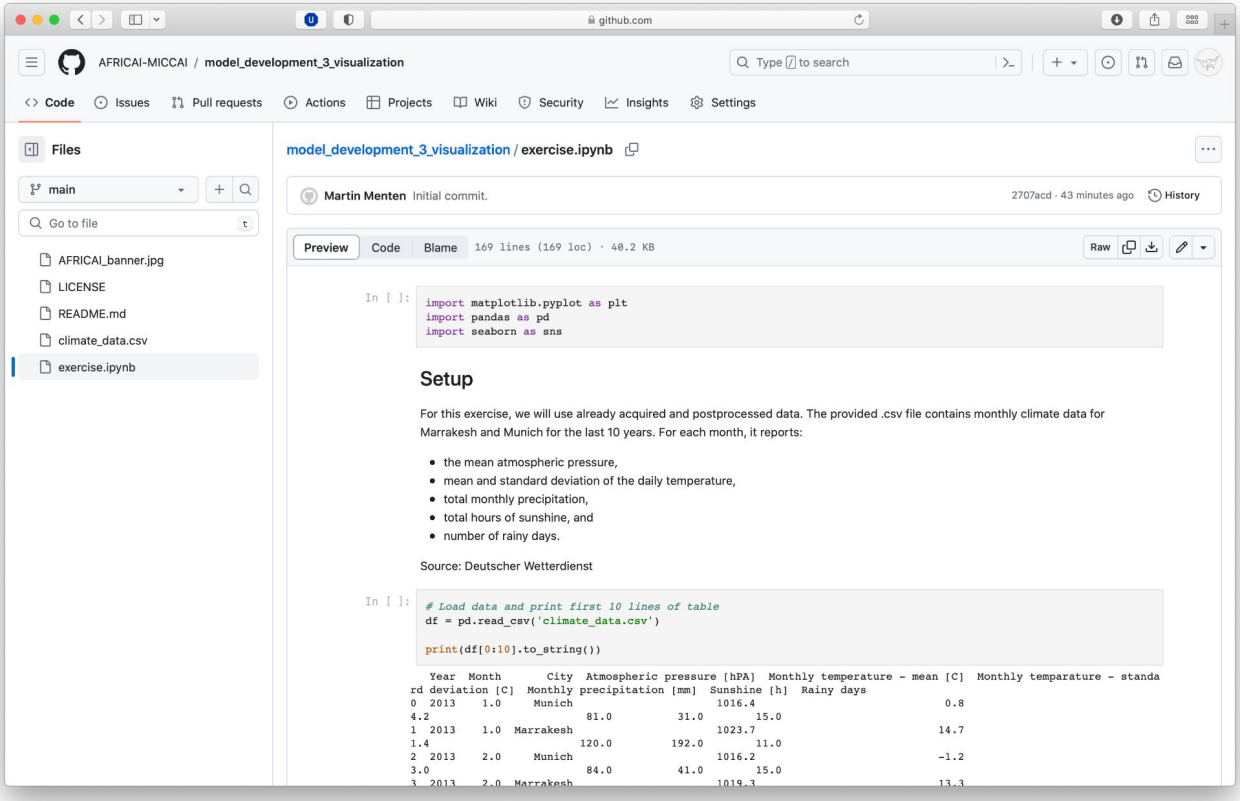
Matplotlib

seaborn

Common mistakes when making results figures

- No clear purpose
- Too easy to understand
- Too difficult to understand
- Too small font
- Too large font
- Too many colors
- Too few colors
- Lack of consistency
- Typos

Interactive: make your own results figures



Five takeaways

- Formulate and think about your research question as testable hypothesis
- Set up your evaluation pipeline including metrics and baselines early
- Beware of overfitting, which comes in many forms and shapes
- Include statistical analysis if you can
- Carefully design your figures as they carry the results section

List of resources

- https://metrics-reloaded.dkfz.de/
- https://monai.io/
- https://lightning.ai/torchmetrics
- https://docs.scipy.org/doc/scipy/reference/stats.html
- https://stats.stackexchange.com/
- https://www.drawio.com/
- https://www.gimp.org/
- https://pypi.org/project/tabulate/
- https://matplotlib.org/
- https://seaborn.pydata.org/