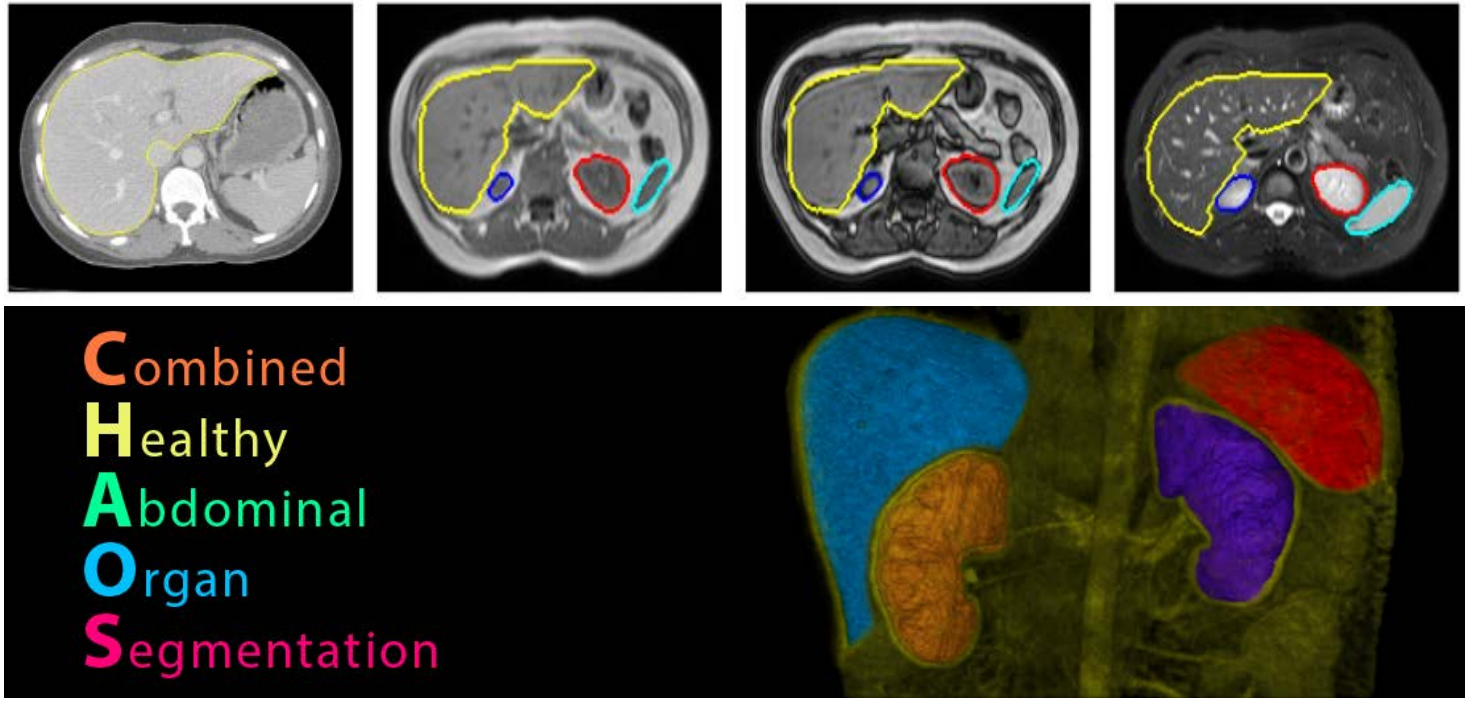
A Story of a Challenge and How To Keep It Impartial

11/03/2020



Ali Emre KAVUR

Graduate school of Natural and Applied Sciences, Dokuz Eylul University Izmir ,Turkey

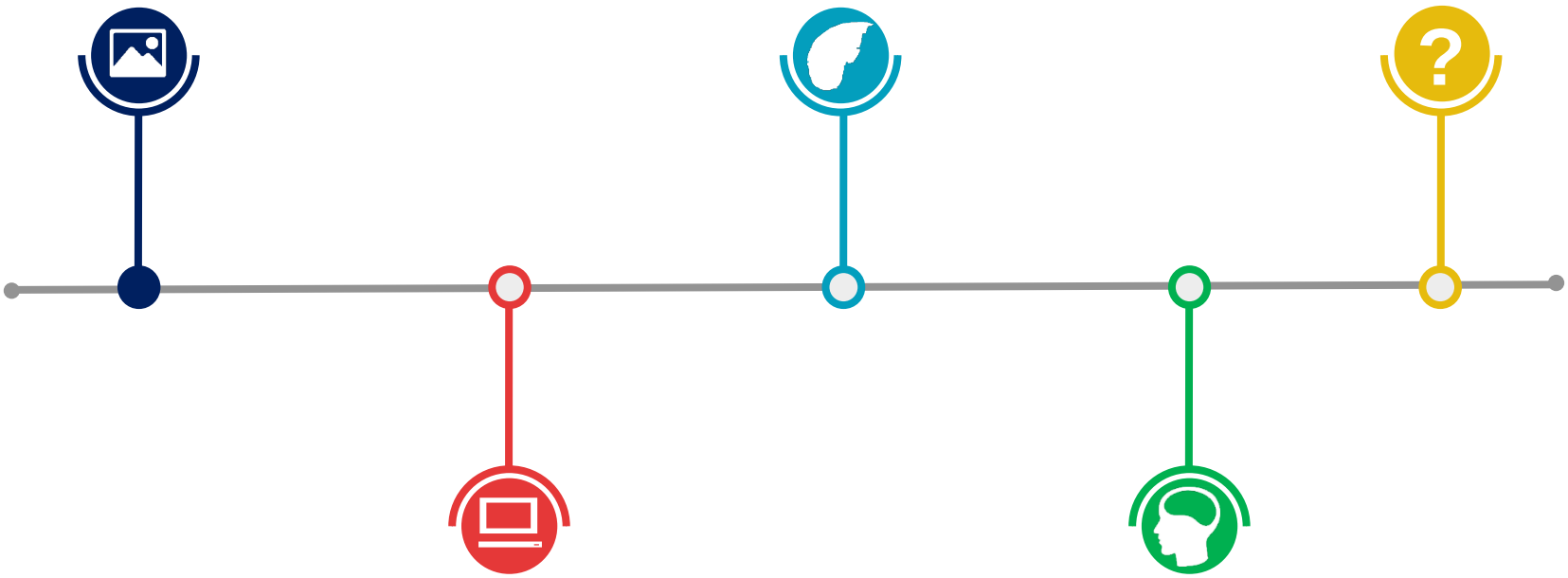
ABOUT CHALLENGES

- Why are they getting so important?

CHAOS CHALLENGE

- What are the aims?
- Why are there many tasks?
- Criticisms

QUESTIONS



FLAWS OF CHALLENGES

- Annotation Quality
- Evaluation Metrics
- “Peeking”

EXPERIENCE DEDUCTIONS ADVICES

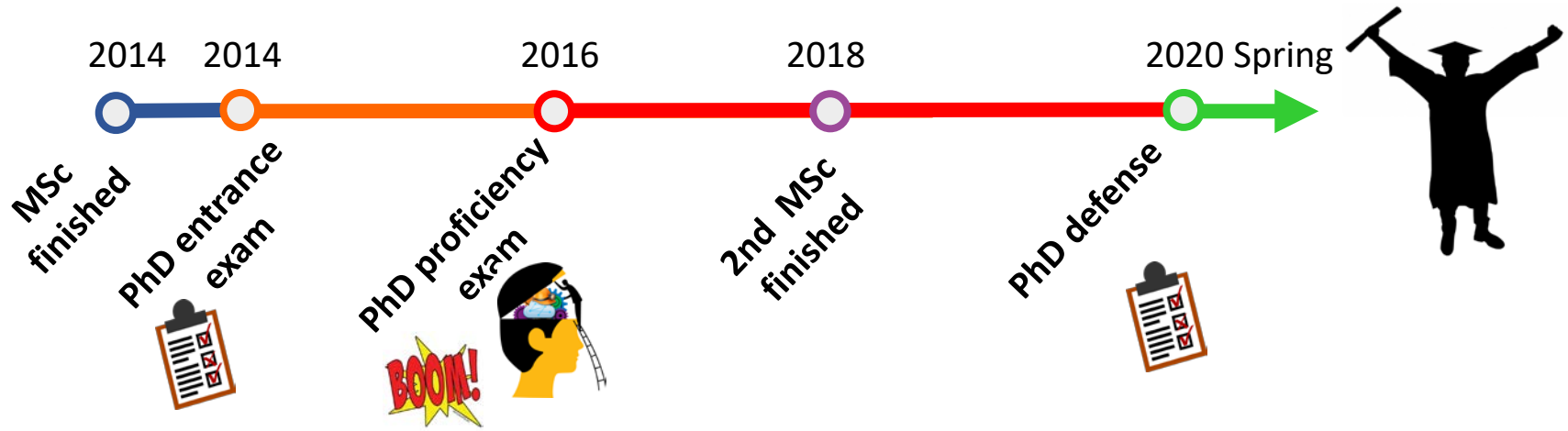
BIO SKETCH

- İzmir is 3rd biggest city in Turkey
- Population: 4+ Million

BIO SKETCH

- BSc: Izmir Institute of Technology (IZTECH) (Electronics and Communication Engineering)
- 1st MSc: Dokuz Eylul University, (DEU) (Electrical and Electronic Engineering)
- 2nd MSc: Izmir Katip Celebi University, (IKCU) (Biomedical technologies)
- PhD: Dokuz Eylul University (DEU), Izmir (Electrical and Electronic Engineering)

  Thesis: "Machine Learning Based Fusion of Different Segmentation Techniques for Liver Visualization for Enhanced Accuracy And Sensitivity"
- Academic Visitor: School of Computer Science and Electronic Engineering, Bangor University, Wales, United Kingdom (Supervisor: Prof. Ludmilla Kuncheva)
- Research of Interests: Medical imaging systems, Medical image processing, Image Segmentation



BIO SKETCH

Medical Image Processing at DEU

- Faculty of Engineering
  - Electrical Electronic Engineering
    - Medical Image Processing

- Faculty of Medicine
  - Radiology

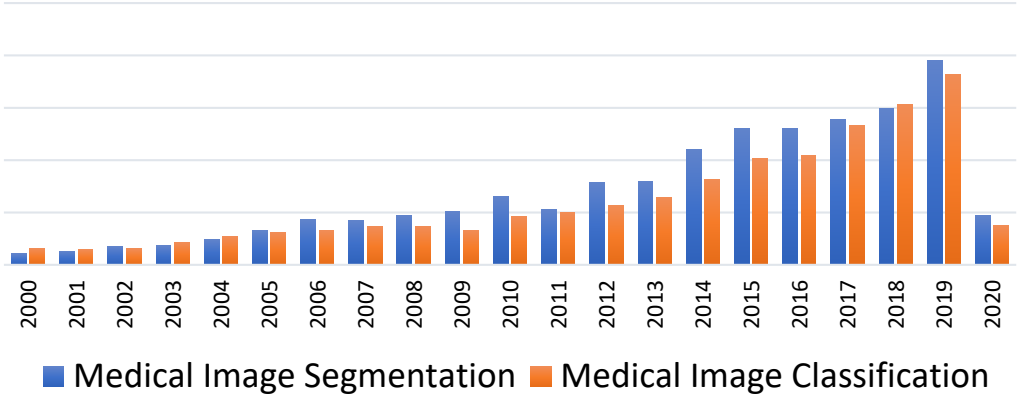
CHALLENGES IN BIOMEDICAL IMAGE ANALYSIS

- As the human population grows, demand for medical imaging solutions in clinics is increasing.
- Segmentation, detection and classification are popular problems in medical imaging field.
- New algorithms are being developed continuously.
- To compare proposed solutions with previous ones, they must be tested under the same conditions.

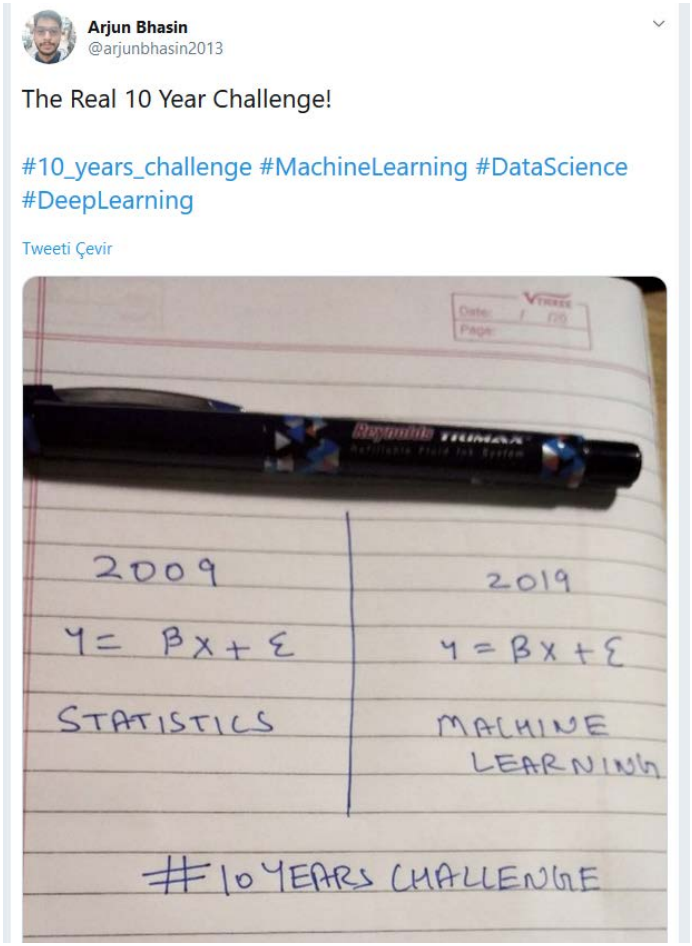
CHALLENGES IN BIOMEDICAL IMAGE ANALYSIS

Why challenges are very popular now?

- Number of new algorithms dramatically increased due to huge interest in Machine Learning studies
- Since competition is getting bigger, challenges are very important benchmark platforms then ever before.



Source: Pubmed



CHALLENGES IN BIOMEDICAL IMAGE ANALYSIS

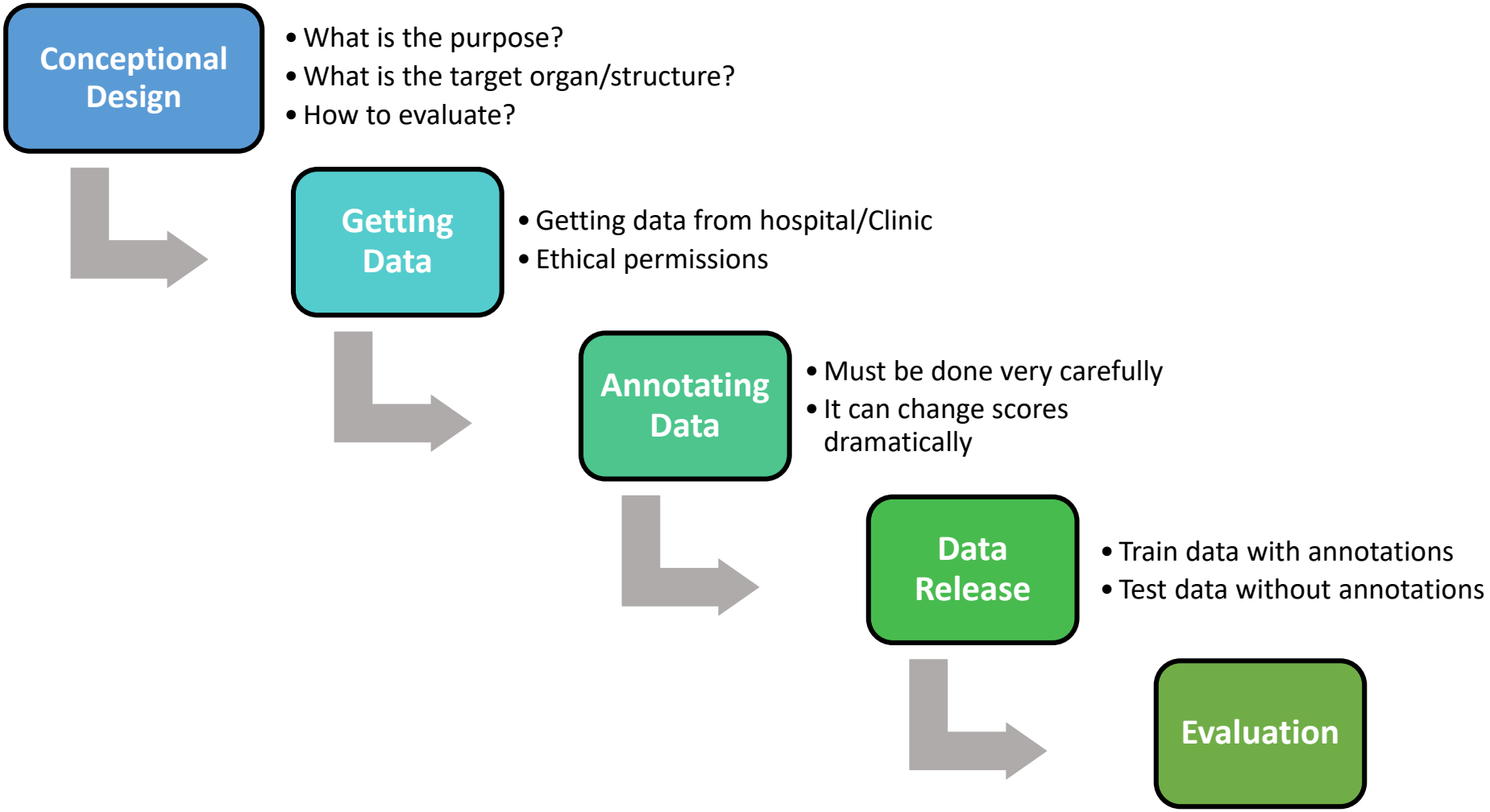
Why challenges are very popular now?

- In the past, benchmarking an algorithm on a single and private data was enough for a publication.

- Now, a proposed algorithm should be performed on multiple and open benchmark platforms in order to classify it as “successful”.

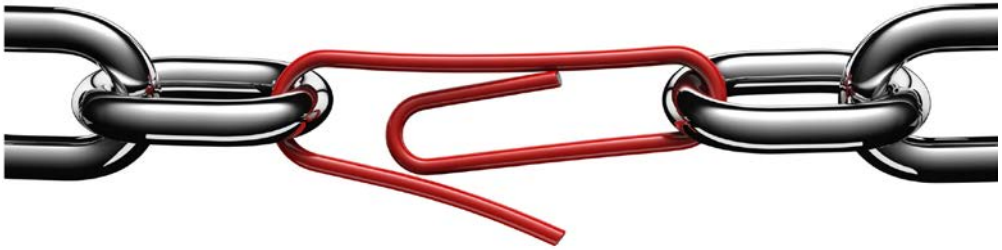
- Performance on different platforms (challenges) play important role in publications, thesis, reports … etc.

MAJOR STAGES OF A CHALLENGE



FLAWS OF CHALLENGES

- A challenge has multiple complicated steps. Any mistake will cause important problem(s).



- These steps might be very time consuming.

- Therefore, it is very difficult to organize a completely perfect challenge.

FLAWS OF CHALLENGES

- Lack of time and/or human sources, inattention will cause problems in challenge designs.

- Some major flaws of challenges are:

  I. Quality of Annotations
  II. Metrics
  III. “Peeking”

1. Quality of Annotations

FLAWS OF CHALLENGES

Quality of Annotations

- To evaluate performance of algorithms, there is a need of ground truths (references).
- The data of challenge is annotated in order to create the references.
- The quality of annotations has direct role on both training and testing of algorithms.

FLAWS OF CHALLENGES

Quality of Annotations

- Annotation is one of the most time consuming steps in challenge design.
- The modality, target organ(s)/structure(s) dramatically affect the time.
- E.g., a single abdomen CT scan may include more than **200 slices**.

FLAWS OF CHALLENGES

Quality of Annotations

- Annotations can be handled via:
  I. Manually (slice-by-slice)
  II. Semi-automatically (with help of a segmentation tool)
  III. Crowdsourcing (from a service such as *Amazon Mechanical Turk*)

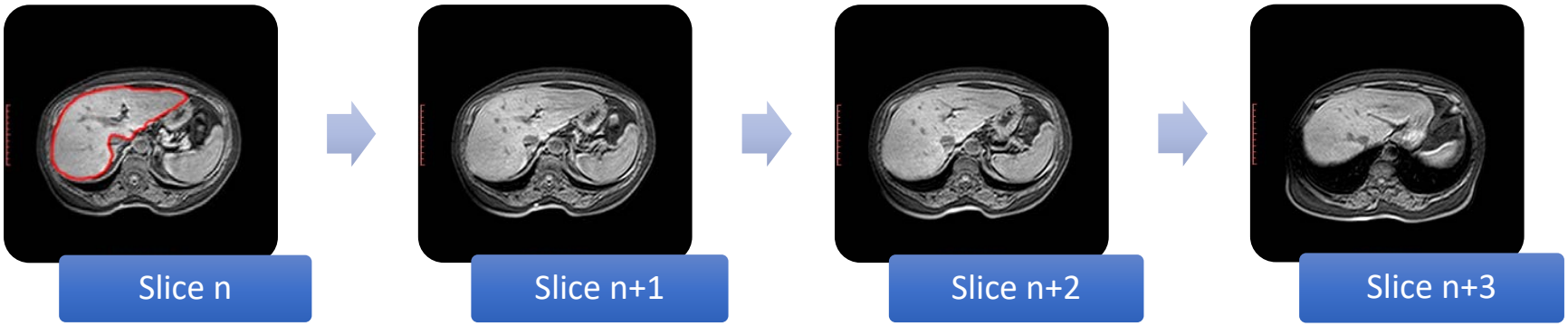
- What is the best way? *(Is there any?)*

FLAWS OF CHALLENGES

Quality of Annotations

I. Manual annotation:

- Each image file is annotated individually (usually by experts).
- It is extremely time consuming.
- It is the safest way.



FLAWS OF CHALLENGES

Quality of Annotations

II. Semi-automatic annotation:

- Annotations are handled via a segmentation tool.
- It may still require interaction (corrections, post-processing).
- It takes less time than manual way but this varies due to many parameters (modality, target structure).
- The output of the software(s) must be subjected to quality control.

FLAWS OF CHALLENGES

Quality of Annotations

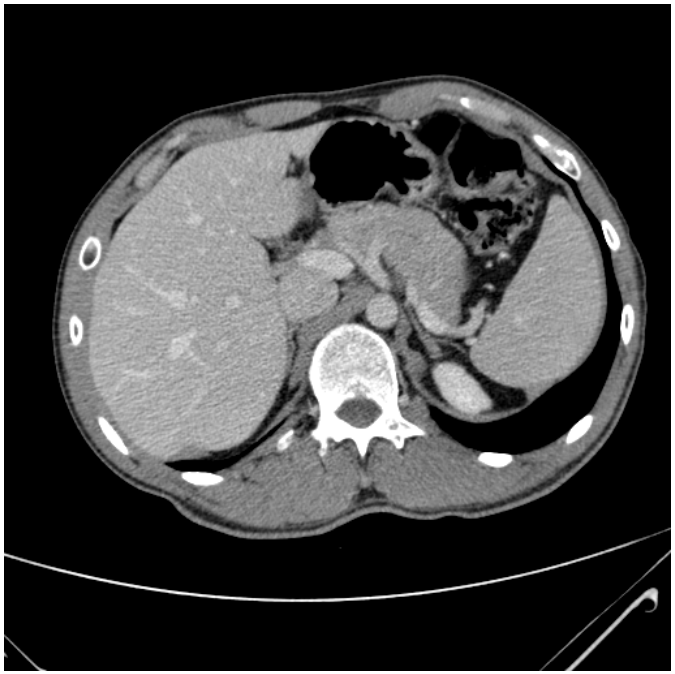
III. Crowdsourcing:

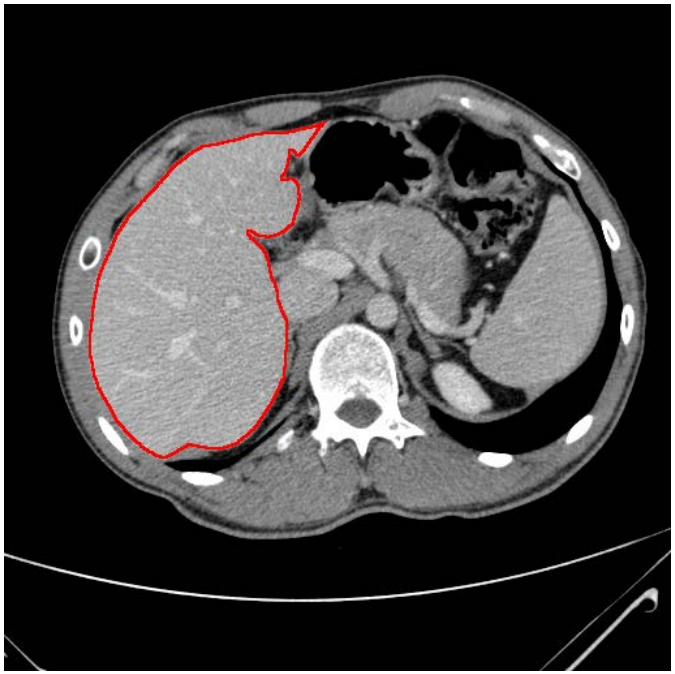
- Annotations are handled by multiple anonymous workers.
- It may be a smart solution to the problem of finding time and human sources.
- There is a need of “fund” for this step.
- Correctness (or quality) of annotations is not guaranteed.
- Additional quality-check might be necessary.

FLAWS OF CHALLENGES

Quality of Annotations – Gray zone

- In many cases, it is not so difficult to define what is “true” or “false” for the target structure/organ:

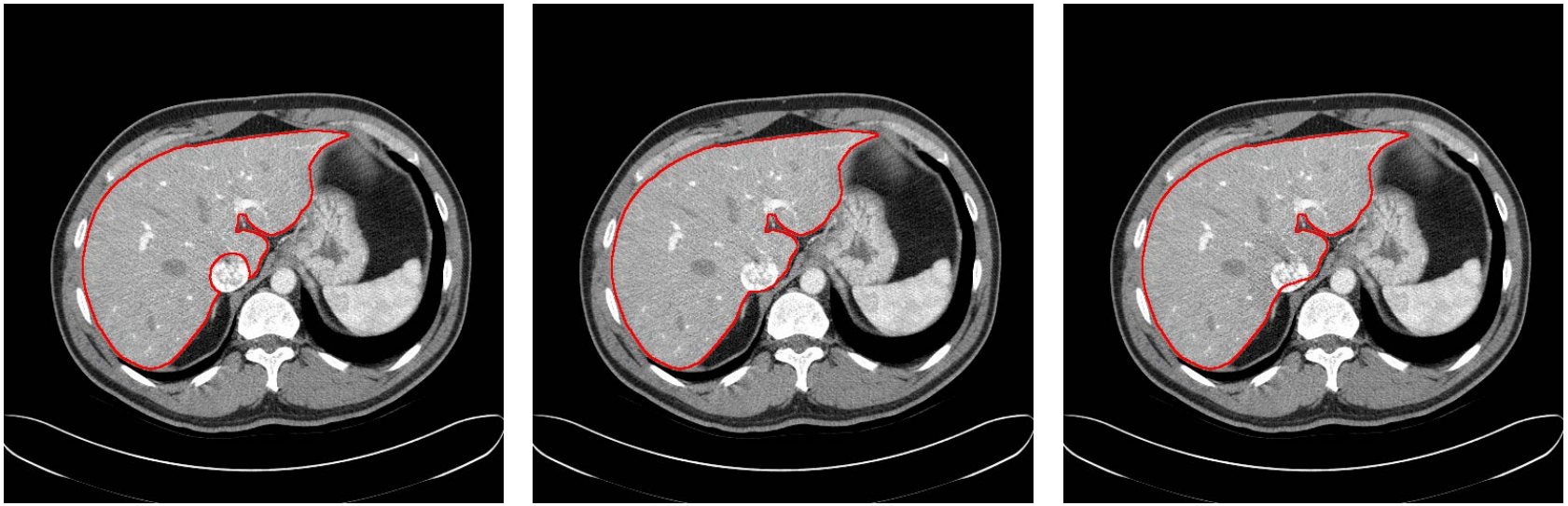




FLAWS OF CHALLENGES

Quality of Annotations – Gray zone

- But sometimes even experts cannot have a consensus:



Ground truths according to 3 different experienced radiologists.

FLAWS OF CHALLENGES

Quality of Annotations – Gray zone

- The solution of unspecific cases can be:
    1. Majority voting of multiple annotations
    2. Reaching a consensus over all cases. In this way, consistency can be achieved.
    3. All of the above
- In any case, multiple annotations are more preferred.
- That is why annotation is one of the most time consuming step of a challenge.

FLAWS OF CHALLENGES

Quality of Annotations – Questions without certain answers

- On the other hand…
  1. Is there a single and absolute ground truth?
  2. Do minor differences between two annotation ways mean anything?
  3. How can we define “minor difference”?

- That is why annotation is very difficult and confusing stage.

2. Evaluation Metrics

FLAWS OF CHALLENGES

Evaluation Metrics

- Evaluation strategy has direct impact on the performance of a proposed algorithm.
- Preference of metric(s) and ranking methods can make critical changes on leaderboard.

FLAWS OF CHALLENGES

Evaluation Metrics

- Evaluation strategy has direct impact on the performance of a proposed algorithm.
- Preference of metric(s) and ranking methods can make critical changes on leaderboard.

Why rankings of biomedical image analysis competitions should be interpreted with care

Lena Maier-Hein, Matthias Eisenmann, [...] Annette Kopp-Schneider

Nature Communications **9**, Article number: 5217 (2018) | Cite this article

5464 Accesses | **17** Citations | **10** Altmetric | Metrics

https://www.nature.com/articles/s41467-018-07619-7

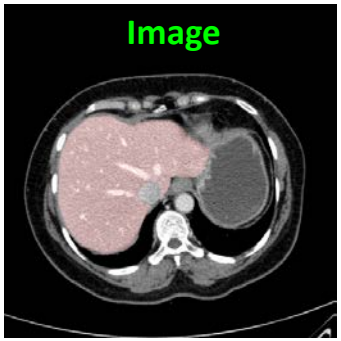
FLAWS OF CHALLENGES

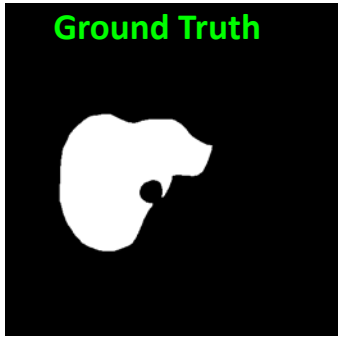
Evaluation Metrics – Single or Multiple?

- Evaluations can be handled by a single metric or aggregation of multiple metrics.
- There is no proven standard of this choice.
- Single metrics may not be adequate to perform a complete evaluation.
- Combination of multiple diverse metrics may eliminate each other’s drawbacks.

FLAWS OF CHALLENGES

Metrics comparison



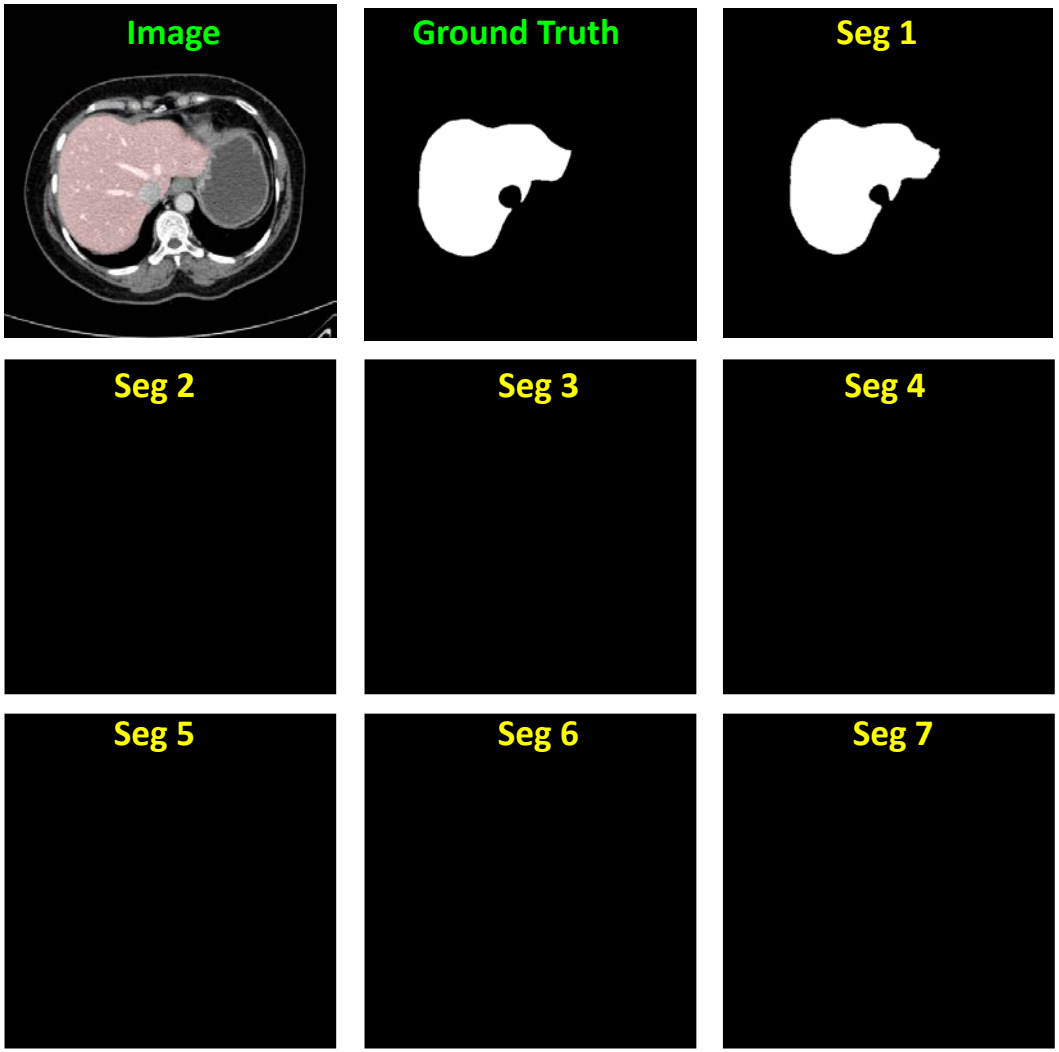


	DICE
Seg 1	0.987

*DICE: Sørensen–Dice coefficient [higher is better]

FLAWS OF CHALLENGES

Metrics comparison

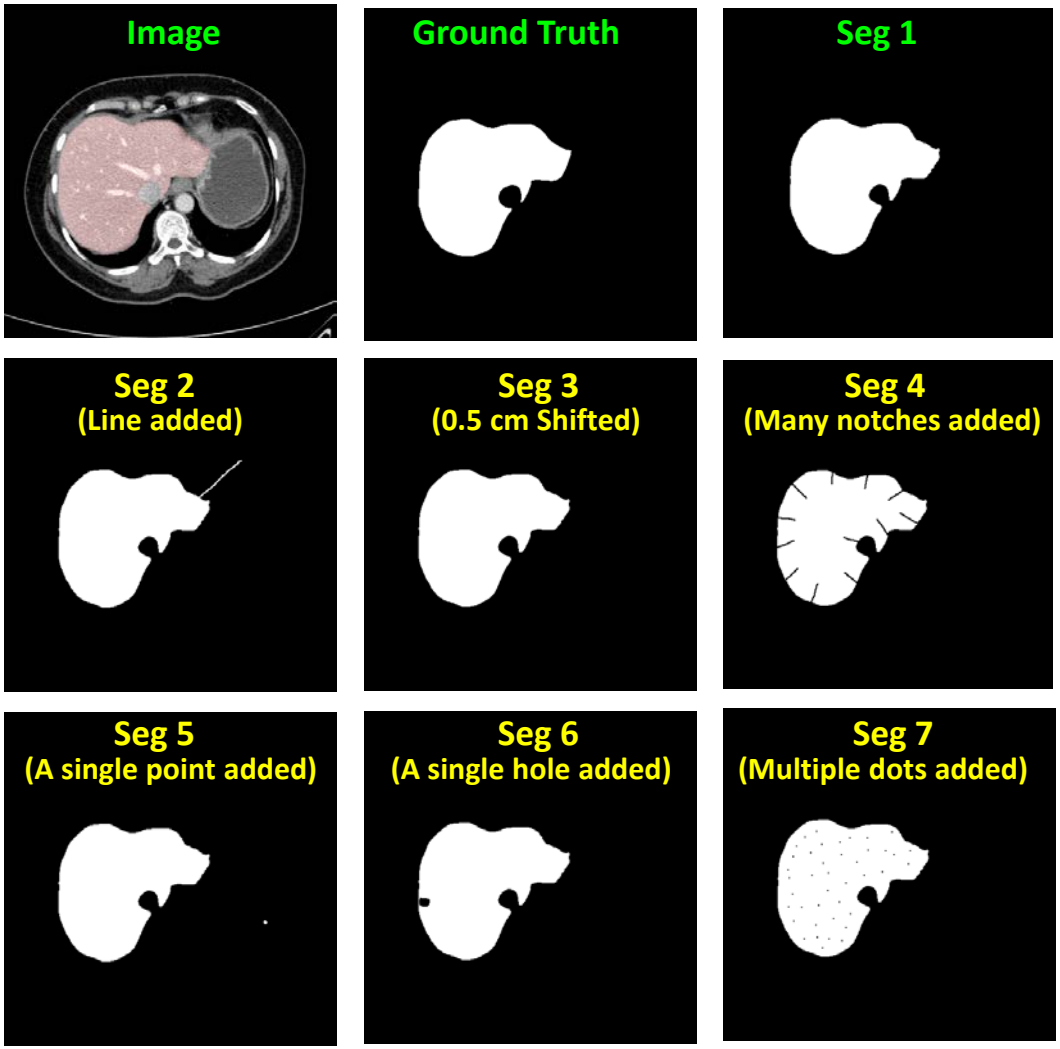


	DICE
Seg 1	0.987
Seg 2	0.984
Seg 3	0.858
Seg 4	0.975
Seg 5	0.986
Seg 6	0.984
Seg 7	0.984

*DICE: Sørensen–Dice coefficient [higher is better]

FLAWS OF CHALLENGES

Metrics comparison

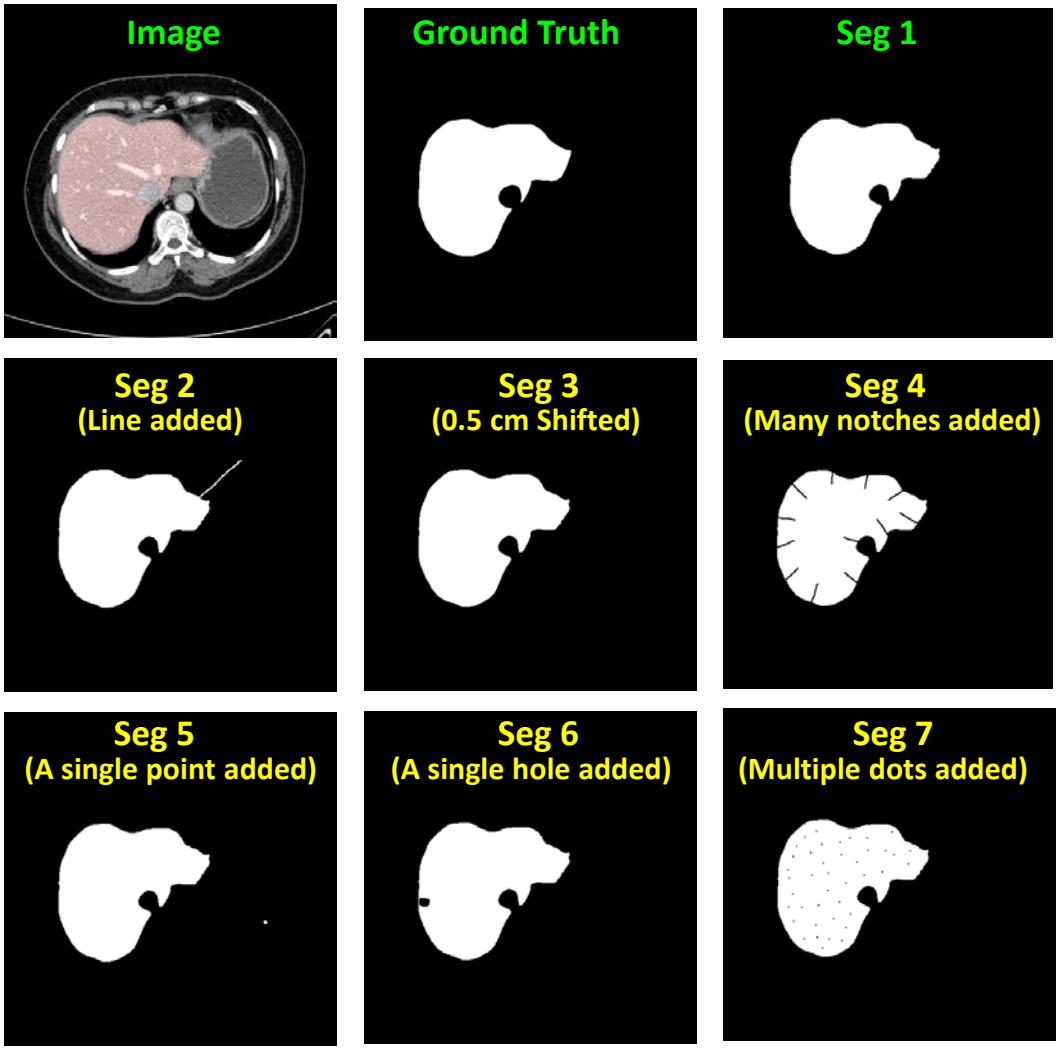


	DICE
Seg 1	0.987
Seg 2	0.984
Seg 3	0.858
Seg 4	0.975
Seg 5	0.986
Seg 6	0.984
Seg 7	0.984

*DICE: Sørensen–Dice coefficient [higher is better]

FLAWS OF CHALLENGES

Metrics comparison



	DICE	RAVD	ASSD	MSSD
Seg 1	0.987	0.805	0.696	3.681
Seg 2	0.984	0.159	2.99	53.731
Seg 3	0.858	0.805	7.406	13.825
Seg 4	0.975	3.089	3.342	22.013
Seg 5	0.986	0.722	1.398	81.531
Seg 6	0.984	1.406	0.864	11.621
Seg 7	0.984	1.327	4.06	44.328

*DICE: Sørensen–Dice coefficient [higher is better]

*RAVD: Relative absolute volume difference [lower is better]

*ASSD: Average symmetric surface distance (mm) [lower is better]

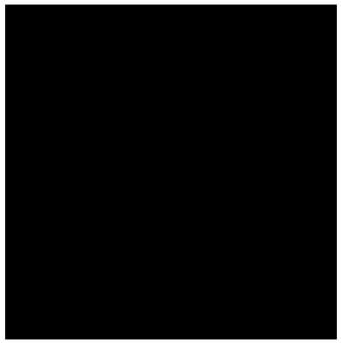
*MSSD: Maximum symmetric surface distance (mm) [lower is better]

FLAWS OF CHALLENGES

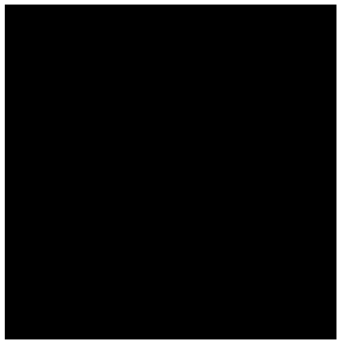
Metrics comparison

Seg 1

Seg 2

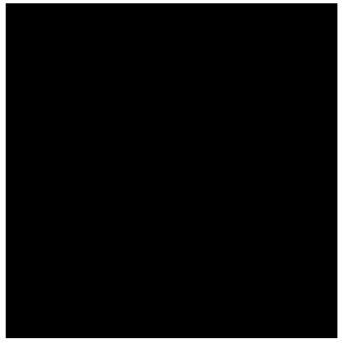


Seg 3

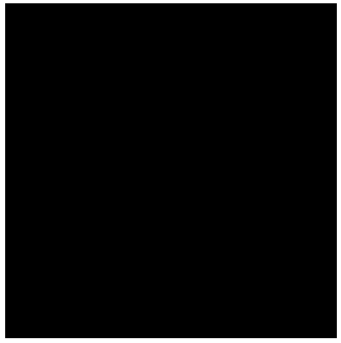


	MSSD
Seg 1	53.731
Seg 2	53.731
Seg 3	53.731
Seg 4	53.731
Seg 5	53.731
Seg 6	53.731

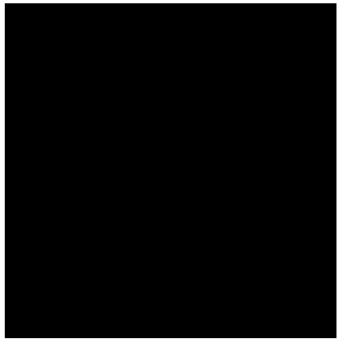
Seg 4



Seg 5



Seg 6

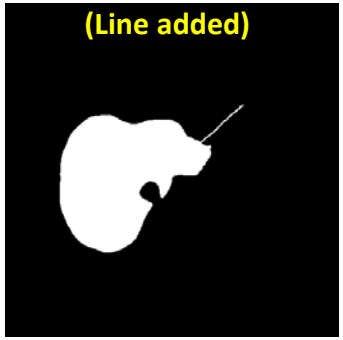


*MSSD: Maximum symmetric surface distance (mm) [lower is better]

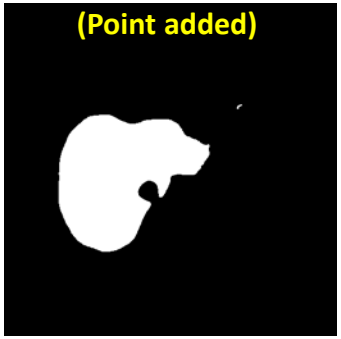
FLAWS OF CHALLENGES

Metrics comparison

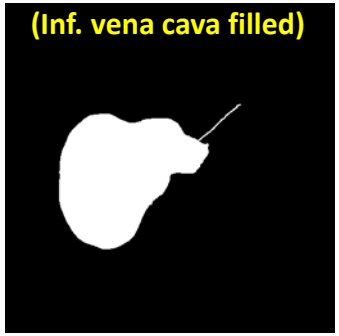
Seg 1
(Line added)



Seg 2
(Point added)

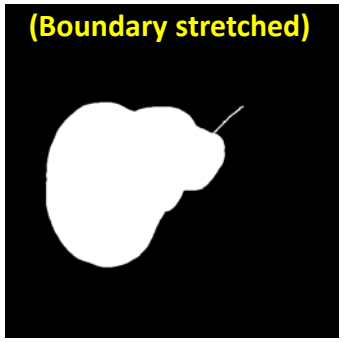


Seg 3
(Inf. vena cava filled)

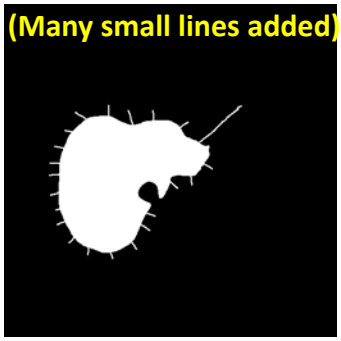


	DICE	RAVD	ASSD	MSSD
Seg 1	0.985	0.159	2.990	53.731
Seg 2	0.986	0.729	1.313	53.731
Seg 3	0.972	2.412	3.731	53.731
Seg 4	0.782	55.701	16.128	53.731
Seg 5	0.974	2.023	3.580	53.731
Seg 6	0.955	6.449	3.632	53.731

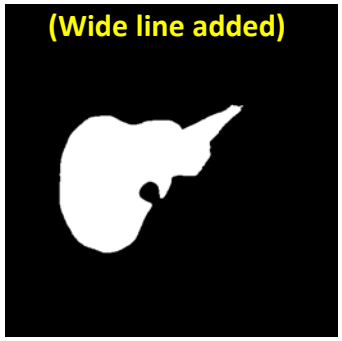
Seg 4
(Boundary stretched)



Seg 5
(Many small lines added)



Seg 6
(Wide line added)



*DICE: Sørensen–Dice coefficient [higher is better]

*RAVD: Relative absolute volume difference [lower is better]

*ASSD: Average symmetric surface distance (mm) [lower is better]

*MSSD: Maximum symmetric surface distance (mm) [lower is better]

FLAWS OF CHALLENGES

Evaluation Metrics

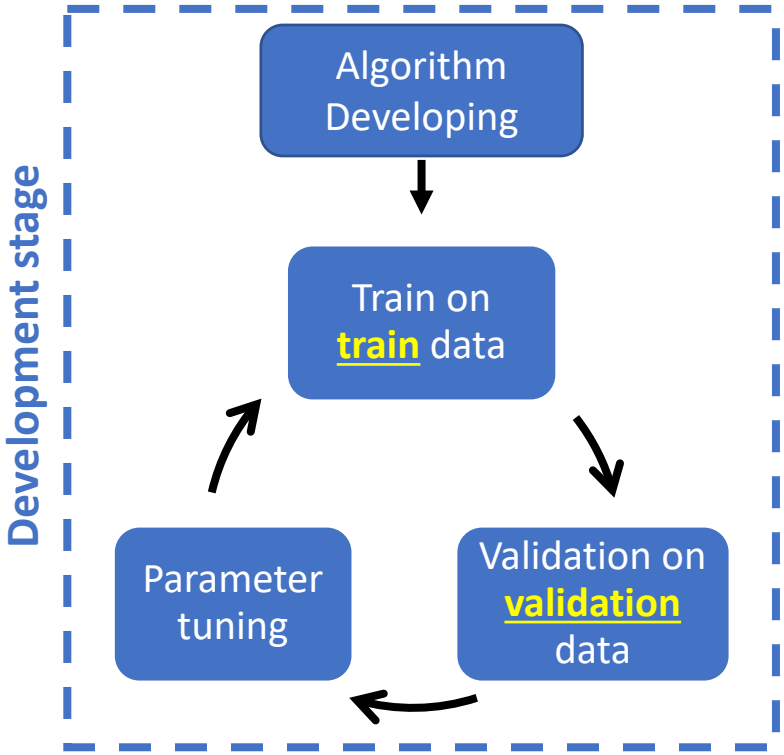
- Each metric has advantages and disadvantages.
- Combination of multiple metrics that uses different specifications may guarantee a reliable comparison of segmentations.
- However, using multiple metrics may come with other problems such as how to aggregate them.
- If metrics have different distributions, additional transformations are necessary to aggregate them.

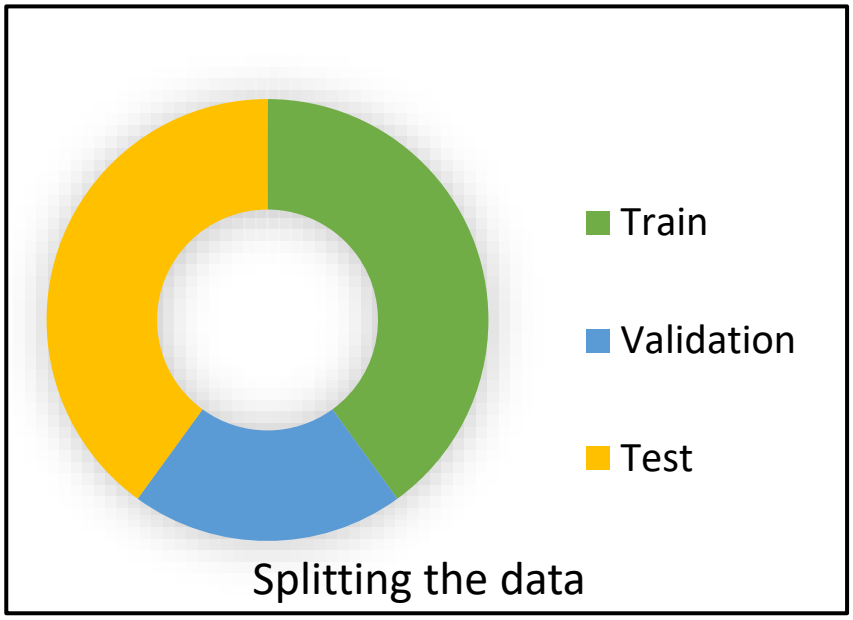
3. Peeking

FLAWS OF CHALLENGES

A fair study

Steps of a fair and proper study:

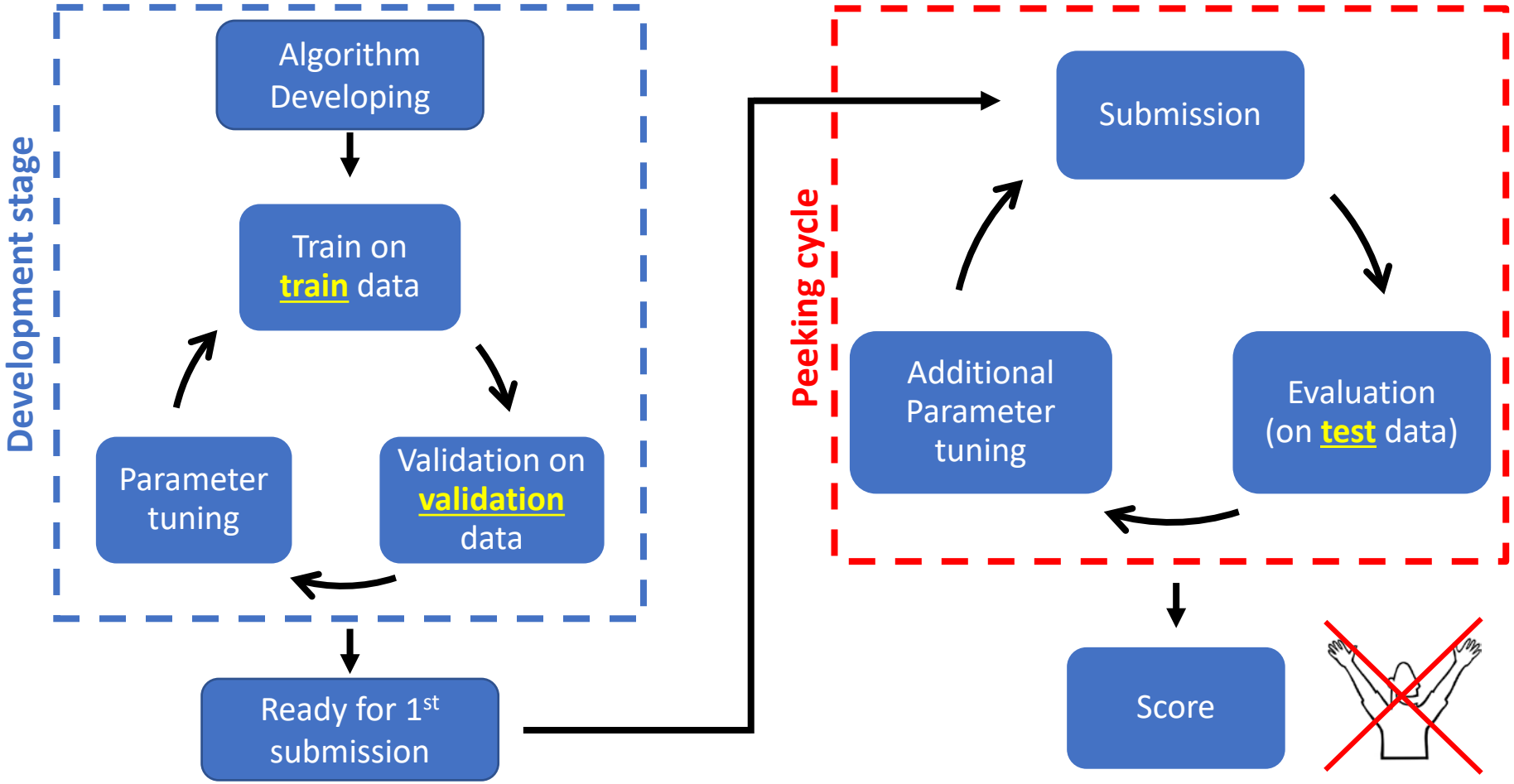




Ready for submission → Evaluation (on test data) → Score

FLAWS OF CHALLENGES

Peeking means tuning parameter(s) on the testing data with lots of iterative submissions.



FLAWS OF CHALLENGES

Peeking

- There is no need for direct access of ground truths in testing data for peeking.
- Just metric outputs of several submissions can be used for parameter tuning.
- Indirectly, peeking makes possible to use testing data for development process even without the ground truths.

Further reading:
“Combining Pattern Classifiers: Methods and Algorithms”, 2nd Edition (Page 17)
Ludmila I. Kuncheva
ISBN: 978-1-118-31523-1

FLAWS OF CHALLENGES

Peeking

- Peeking is an underestimated problem.
- Peeking causes overtraining of an algorithm on a specific data.
- Even indirect usage of testing data for development makes it impossible to observe generalization abilities of the proposed algorithms.
- It leads to deviation from the **main purpose of engineering**.
- Getting high scores with peeking on one specific data does not mean that the method is valid!
- Testing must be done on previously **unseen** data with reasonable number of submissions.

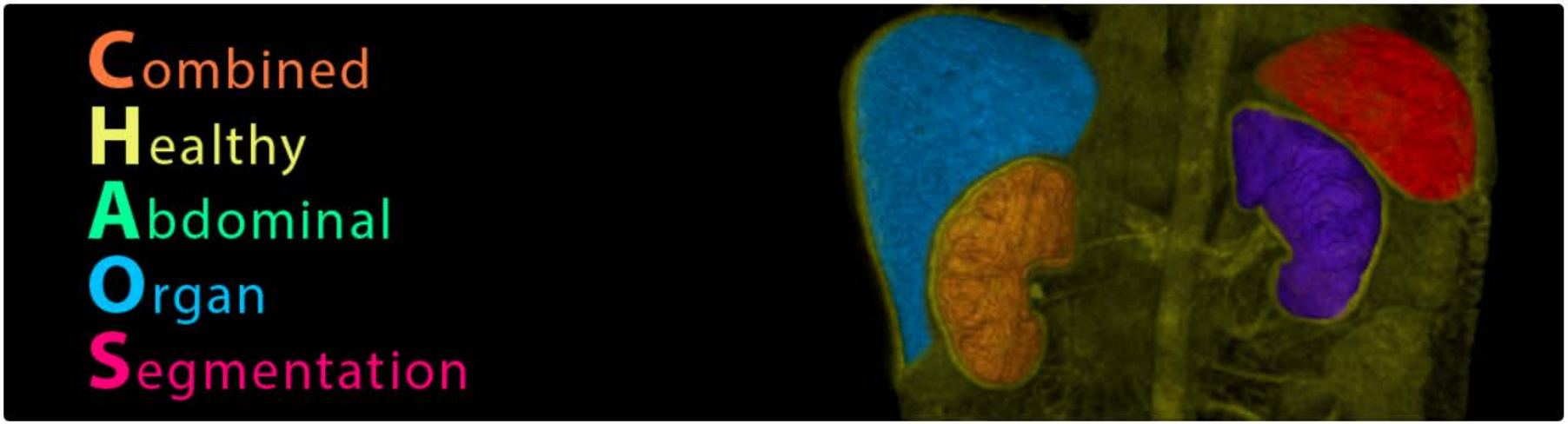
FLAWS OF CHALLENGES

Peeking – Questions?

There are important questions regarding the peeking problem:

- How many submissions can be accepted as “reasonable”?
- How is it possible to prevent peeking in online submitted challenges?
- **How can we understand whether a submission from the same team comes from a brand new algorithm or old one with tuned-parameters?**

CHAOS CHALLENGE



Description

Data Info

Rules

Evaluation

Download

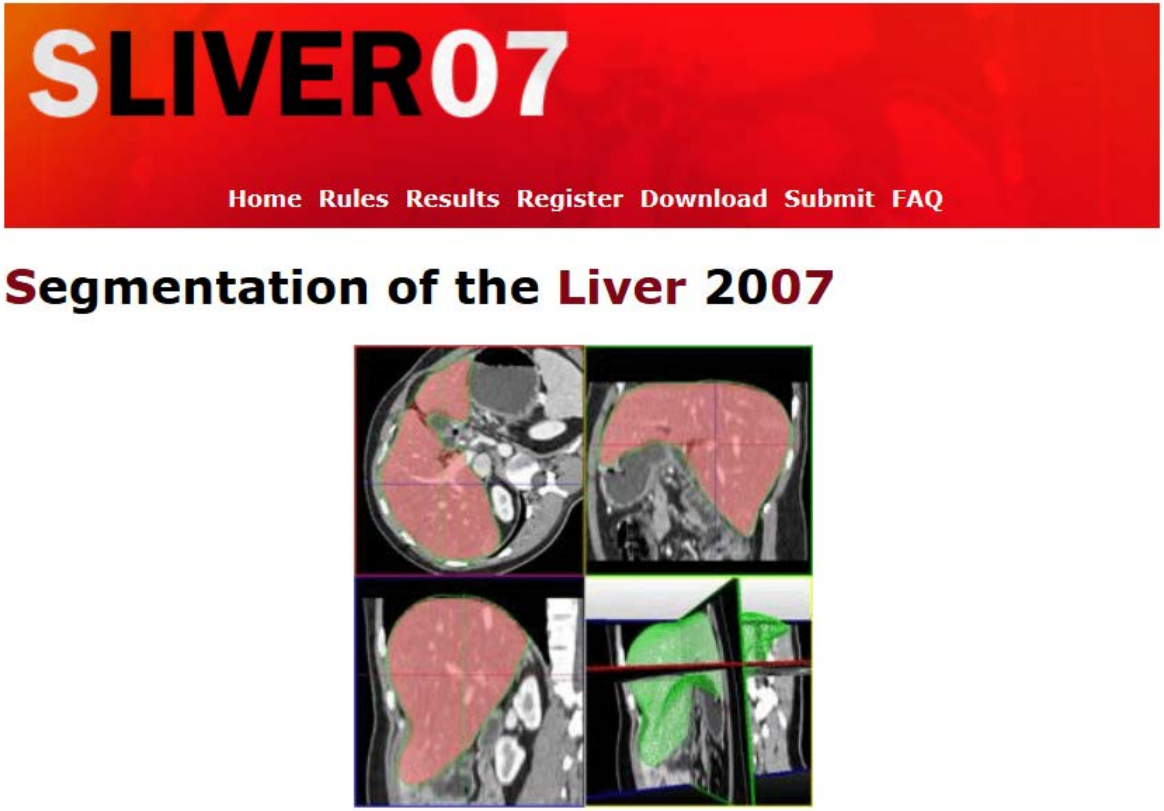
CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation

CHAOS challenge aims the segmentation of abdominal organs (liver, kidneys and spleen) from CT and MRI data. CHAOS was held in The IEEE International Symposium on Biomedical Imaging (ISBI) on April 11, 2019, Venice, ITALY. The results of the challenge were published: https://chaos.grand-challenge.org/Results_CHAOS/

CHAOS CHALLENGE

SLiver07

- SLiver 07 is the first grand challenge about liver segmentation.
- It was organized within the 10th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) at 29. October 2007, Brisbane, Australia.
- There are 20 train (image sets + ground truths) and 20 test (only image sets) of abdomen CT.



http://www.sliver07.org
https://sliver07.grand-challenge.org

3D Segmentation in the Clinic: A Grand Challenge, *Bram van Ginneken, Tobias Heimann, Martin Styner*
http://mbi.dkfz-heidelberg.de/grand-challenge2007/web/p7.pdf

CHAOS CHALLENGE

Before CHAOS…

- Statistics from challenge outputs become very important in our projects.
- We decided to open our private data for a new challenge.
- In May 2018, we organized the first nation-wide challenge in Turkey.
- Aim of the challenge was same as SLiver07.
- 20 abdomen CT scans (10 Train + 10 Testing) were used.
- Same metrics with SLiver07 were used.

CHAOS CHALLENGE

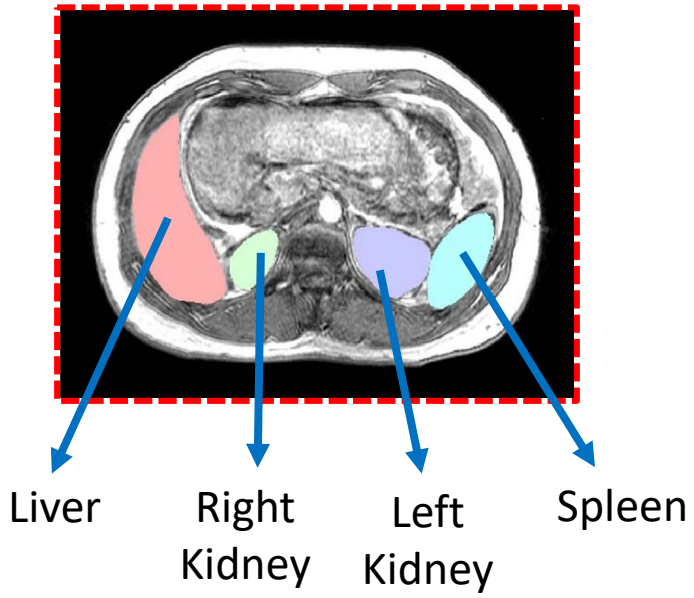
Before CHAOS…

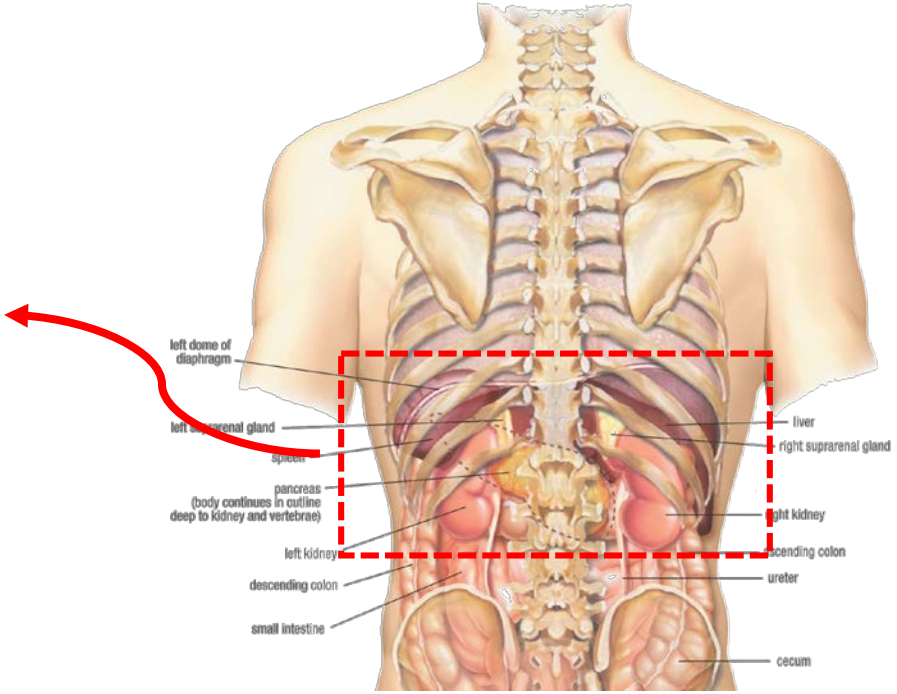
- The challenge was one time event and held only on-site.
- 11 teams were participated to the challenge.
- The challenge lasted six hours. 7 of 11 teams delivered results.
- After seeing the potential of a new challenge, we decided to organize CHAOS challenge.

CHAOS CHALLENGE

Description and Aims

- CHAOS challenge aims the segmentation of abdominal organs from CT (liver) and MRI (liver, kidneys and spleen) data.





CHAOS CHALLENGE

Description and Aims

- Goals of CHAOS are:
  1. Segmentation of target organ(s) from a single modality
  2. Segmentation of target organ(s) from cross modalities (with single model)
- Our aim is to observe capabilities of deep learning based algorithms on more challenging tasks.

CHAOS CHALLENGE

Tasks

1. Liver Segmentation [CT & MRI]
2. Liver Segmentation [CT only]
3. Liver Segmentation [MRI only]
4. Segmentation of abdominal organs [CT(Liver) & MRI(liver, kidneys, spleen)]
5. Segmentation of abdominal organs [MRI only(liver, kidneys, spleen)]

- Different combinations were preferred to push algorithms to their limits.

Combined
Healthy
Abdominal
Organ
Segmentation

News And Faq
Download
Organization Team
Results (Isbi19)
Online Leaderboard
Publications
Terms Of Use
Contact
Join
Admin

Assoc. Prof. Dr. M. Alper SELVER
Dokuz Eylul University (DEU), Electrical and Electronics Engineering Dept., Izmir, TURKEY

Prof. Dr. Gözde ÜNAL
Istanbul Technical University (ITU), Computer Engineering Dept., Istanbul, TURKEY

Prof. Dr. Oğuz DİCLE (MD)
DEU Faculty of Medicine, Radiology Dept., Izmir, TURKEY
Chair of Turkish Association of Medical Informatics (TURKMIA), Ankara, TURKEY

Assoc Prof. Dr. N. Sinem GEZER (MD)
DEU, Faculty of Medicine, Radiology Dept., Izmir, TURKEY.

Dr. Mustafa BARIŞ (MD)
Columbia University (CU) Faculty of Medicine, Radiology Dept., New York, NY, USA (Visiting Fellow till August 2019)

Dr. Sinem ASLAN
Ca' Foscari University of Venice, European Centre for Living Technology (ECLT), Venice, ITALY (Visiting postdoctoral researcher.

Dr. Cemre CANDEMİR
Ege University · Institute of International Computing, Izmir, TURKEY.

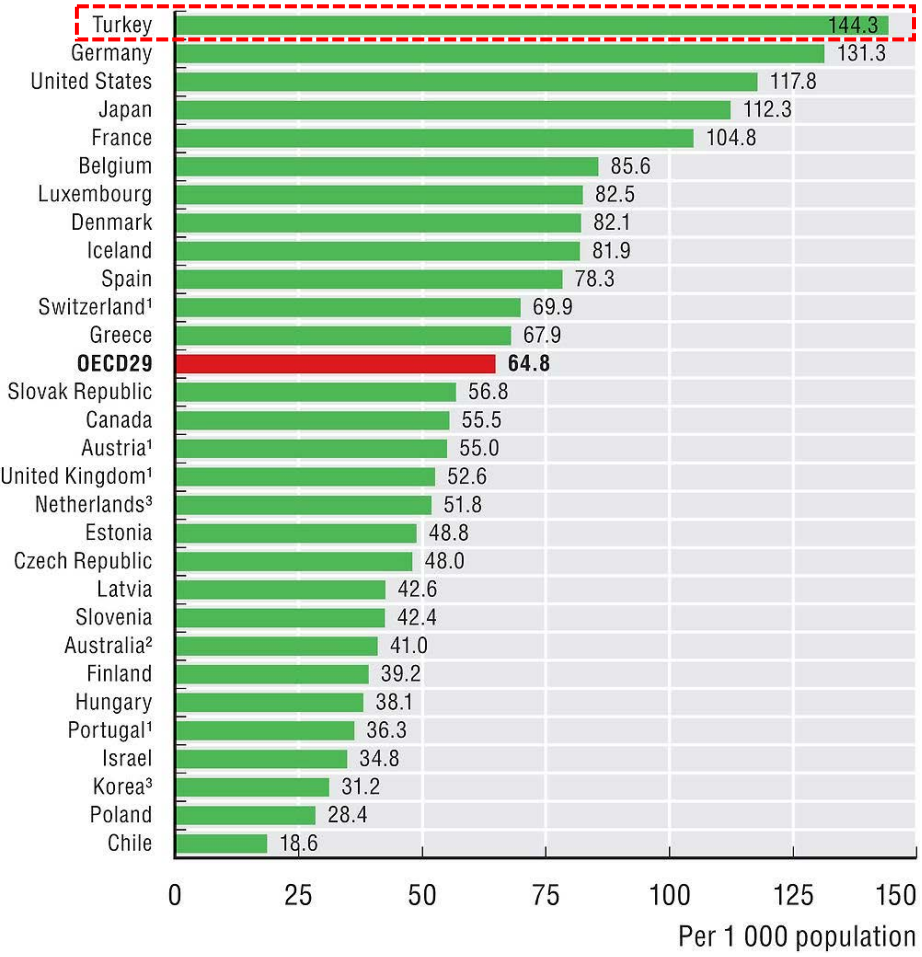
Ali Emre KAVUR (PhD Student)
Dokuz Eylul University (DEU), Institute of Natural and Applied Sciences, Izmir, TURKEY

Esranur Kazaz (B.Sc. Student)
Dokuz Eylul University (DEU), Electrical and Electronics Engineering Dept., Izmir, TURKEY

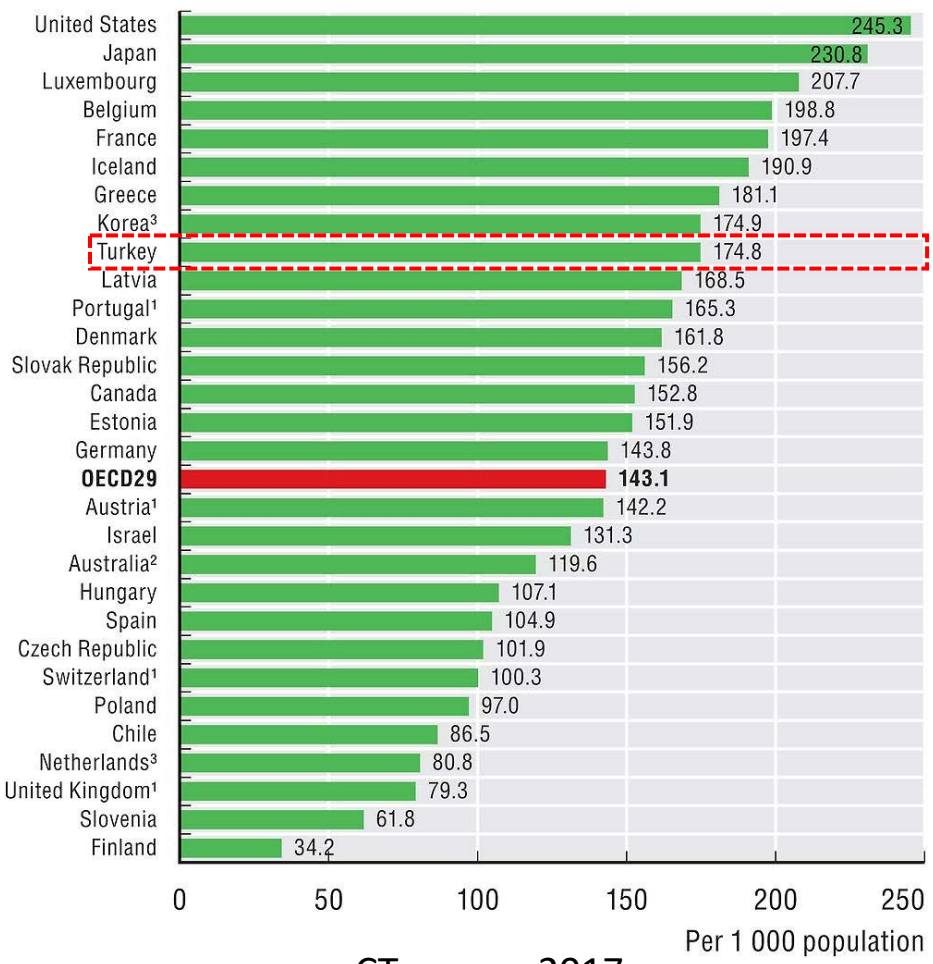
- CHAOS is organized by a relatively small team from different disciplines.

CHAOS CHALLENGE

Data: Medical Imaging Facts in Turkey



MRI exams, 2017
(Total: 11 million MRI scans)



CT exams, 2017
(Total: 14.5 million CT scans)

Source: https://www.oecd-ilibrary.org/sites/health_glance-2017-61-en/index.html?itemId=/content/component/health_glance-2017-61-en

CHAOS CHALLENGE

Data

- Due to huge amount of medical image scans availability, it is relatively easy to find medical data in Turkey.
- After getting ethical approval from university hospital committee (approval from all chair of departments in the board and dean of medicine), there is no need additional permission if the data is anonymized and for only non-commertial scientific work.
- Getting approval from university board can be challenging.

CHAOS CHALLENGE

Data

- There are two medical image databases in CHAOS:

1. **The first database** contains CT images of **40** different patients.

2. **The second database** includes **120** data sets of different MRI sequences from **40** different patients:

   I. T1-DUAL In-phase
   II. T1-DUAL Opposed-phase
   III. T2 SPIR

   ➢ *T1-DUAL sequences are registered so there are single ground truths for both of them. T1 and T2 series are independent (not registered).*

CHAOS CHALLENGE

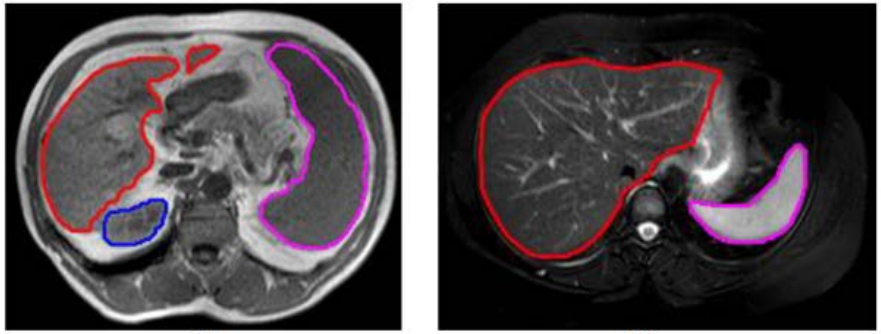
Data

- Both dataset were divided 50-50 portion as train and testing data. (Participants are free to divide train data for train and validation)
- Private patient info were cleared.
- Train data was published with DICOM images and ground truths (annotations).
- Testing data (only DICOM images) was published online after on-site challenge in ISBI 2019 was completed.
- Data is open to usage for scientific work: https://zenodo.org/record/3431873

CHAOS CHALLENGE

Annotation

- All DICOM files in the data were manually (slice-by-slice) annotated via 3D Slicer* software (no segmentation tool was used).



- Reached consensus on uncertain areas (such as excluding *inferior vena cava* outside of the liver).
- Multiple annotations by different radiologists were collected.
- Ground truths were generated by majority voting of them.

*https://www.slicer.org/

CHAOS CHALLENGE

Evaluation

- Aggregation of four metrics is preferred.
- A similar strategy of SLiver07 used.
- The metrics in CHAOS:

1. Sørensen–Dice coefficient (DICE): Provides information about the overlapping parts of segmented and reference volumes (takes value 1 for a perfect segmentation).
2. Relative absolute volume difference (RAVD): Provides information about the differences between volumes of segmented and reference organs (0% for a perfect segmentation).
3. Average symmetric surface distance (ASSD): Determines the average difference between the surface of the segmented object and the reference in 3D (0 mm for a perfect segmentation).
4. Maximum symmetric surface distance (MSSD) or Hausdorff distance: Determines the maximum difference between the surface of the segmented object and the reference in 3D (0 mm for a perfect segmentation).

CHAOS CHALLENGE

Evaluation

- Output of four metrics have different distribution.
- Therefore, outputs were individually transformed to 0-100 scale with pre-defined thresholds [Thresholds calculated by inter-difference of multiple annotations coming from different experts for the same data (Similar with SLiver07)].
- After transformation, mean of four scores defined the score of a case.
- Final score is calculated by mean score of all cases.

CHAOS CHALLENGE

CHAOS at ISBI 2019

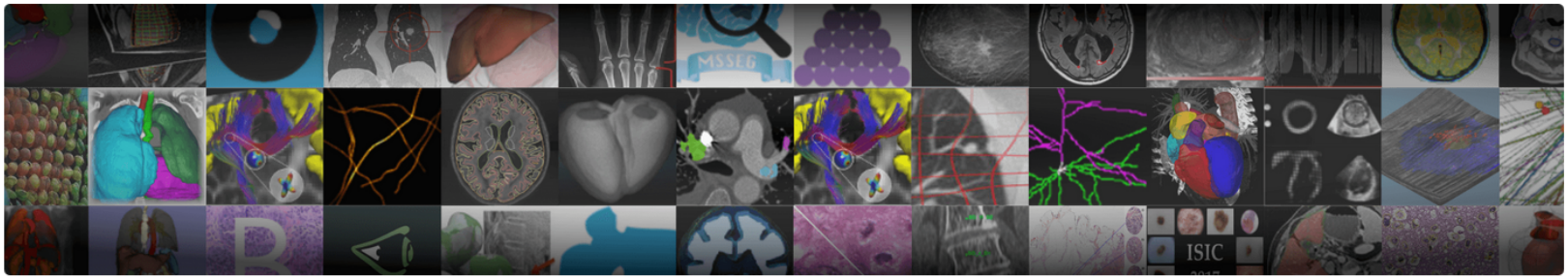
- CHAOS was started in The IEEE International Symposium on Biomedical Imaging (ISBI) on April 11, 2019, Venice, ITALY.

- Online submissions were accepted after first on-site challenge was completed.

- CHAOS has two separate leaderboards for on-site and online submissions.

CHAOS CHALLENGE

Online Submission



Grand Challenges in Biomedical Image Analysis

Every year, thousands of papers are published that describe new algorithms to be applied to medical and biomedical images, and various new products appear on the market based on such algorithms. But few papers and products provide a fair and direct comparison of the newly proposed solution with the state-of-the-art. We believe that such comparisons can help the research community and industry to develop better algorithms. We support the organization of these comparative studies and the dissemination of their results.

Organizing and participating in challenges is not the only way to facilitate better comparisons between new and existing solutions. If it were easy to publish and share your data, and the code you used to evaluate your algorithm's performance on that data, and possibly the algorithm itself, others could directly compare their approach to yours, using the same test data and the same evaluation metrics. With this site we provide tools to make it as easy as possible for you to publish your data and your evaluation for any paper you've written.

Why Challenges? describes the rationale for organizing grand challenges, provides advice for those who want to organize such events, and discusses where we hope the field will move to next.

All Challenges provides an overview of all previous, ongoing and upcoming challenges in biomedical image analysis that we are aware of. Drop us a note if you want your event listed on this overview.

https://grand-challenge.org

CHAOS CHALLENGE

Statistics

Number of participants (on-site)	12
Number of participants (online)	1526
Number of submissions (online)	380
People in organization team	9
Number of DICOM files in the data	10000+

https://chaos.grand-challenge.org/

CHAOS CHALLENGE

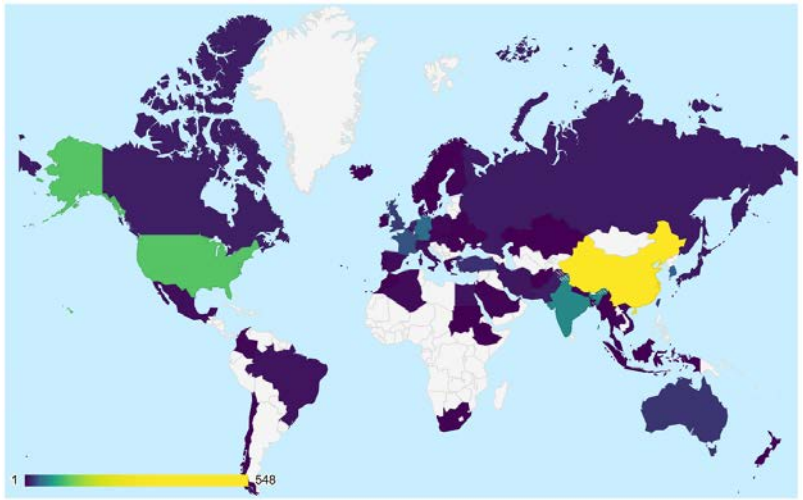
Statistics

Number of participants (on-site)	12
Number of participants (online)	1526
Number of submissions (online)	380
People in organization team	9
Number of DICOM files in the data	10000+
Budget ☺	**0₺, 0$, 0€, £0, 0CHF**

CHAOS CHALLENGE

Positive Facts

- **CHAOS is the first challenge in the literature, that brings cross modality tasks in abdomen imaging**.
- CHAOS showed that new automatic segmentation algorithms (deep learning-based) has a potential of success for abdomen MRI scans.
- Currently, CHAOS has participants from **38 countries**.



CHAOS CHALLENGE

Negative Facts

- Tasks and data of CHAOS can be described as “chaotic”.
- Aims of tasks are complicated.
- Data was not distributed in a compact form.
- The novel part of the CHAOS, cross-modality tasks has not drawn enough attention so far.

	Submissions
Task 1: Liver Segmentation [CT & MRI]	24
Task 2: Liver Segmentation [CT]	298
Task 3: Liver Segmentation [MRI]	47
Task 4: Segmentation of abdominal organs [CT(Liver) & MRI(liver, kidneys, spleen)]	19
Task 5: Segmentation of abdominal organs [MRI only(liver, kidneys, spleen)]	92

CHAOS CHALLENGE

Critics

- Evaluation system was found complicated. Many questions were in regard with why we did not use only DICE score as in other studies and publications.
- Some participants claimed our evaluation system to be ill-designed, because they have higher DICE score while getting lower scores from other three metrics.
- Some participants were angry to the fact that online submissions started after on-site challenge.

CHAOS CHALLENGE

FAQs

1. Why are there annotations for only liver in CT data while MRI data have annotations for four abdomen organs?

CHAOS CHALLENGE

FAQs

1. Why are there annotations for only liver in CT data while MRI data have annotations for four abdomen organs?
    - CT sets have more slices than MRI sets. Due to lack of time and human sources, we could not annotate other organs in CT scans.

CHAOS CHALLENGE

FAQs

1. Why are there annotations for only liver in CT data while MRI data have annotations for four abdomen organs?
    - CT sets have more slices than MRI sets. Due to lack of time and human sources, we could not annotate other organs in CT scans.

2. Why did not we release test data before ISBI19 conference?

CHAOS CHALLENGE

FAQs

1. Why are there annotations for only liver in CT data while MRI data have annotations for four abdomen organs?
    - CT sets have more slices than MRI sets. Due to lack of time and human sources, we could not annotate other organs in CT scans.

2. Why did not we release test data before ISBI19 conference?
    - We wanted to see the scores of on-site results first. (Reason will be explained on further slides.)

EXPERIENCE & DEDUCTIONS

Peeking Problem

- Preventing peeking is too hard and time consuming on online challenges.
- What we tried in CHAOS:
    1. Online submission via e-mail,
    2. Online submission via grand-challenge.org with only university e-mail address,
    3. Online submission via grand-challenge.org with a scientific manuscript which explains the method.

EXPERIENCE & DEDUCTIONS

Peeking Problem: Online submission via e-mail

- Online submission via e-mail is time consuming for the organizers and participants.

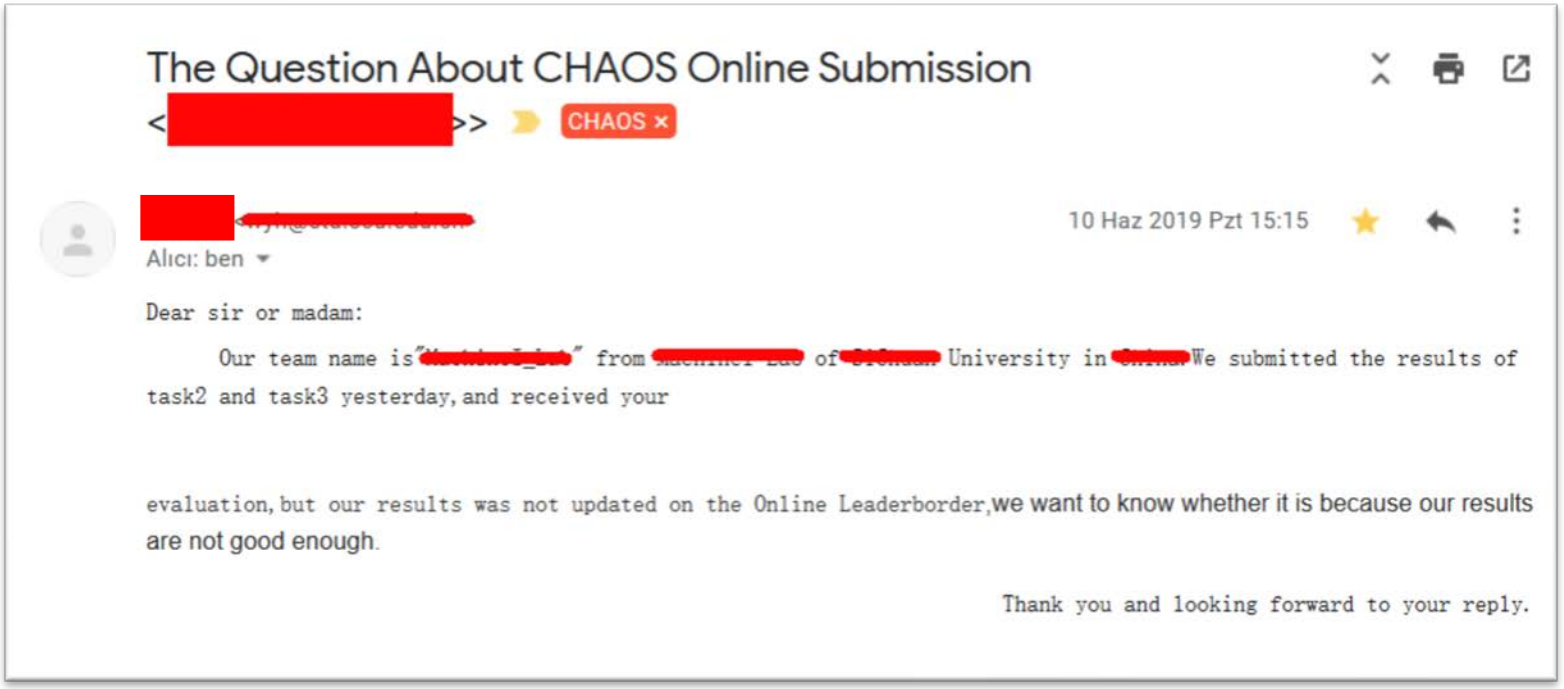
- Each submission must be downloaded and saved to a local computer. Then, evaluation code is run. Finally, the scoreboard on the website is updated manually.

    - This slows down the peeking attempts but not completely. Some teams started using different nicknames.

EXPERIENCE & DEDUCTIONS

Peeking Problem: Online submission via e-mail

- If you don’t publish the results in 24 hours, participants may get angry:



The Question About CHAOS Online Submission

<[redacted]>> CHAOS ×

[redacted] 10 Haz 2019 Pzt 15:15

Alıcı: ben

Dear sir or madam:

Our team name is "[redacted]" from [redacted] of [redacted] University in [redacted] We submitted the results of task2 and task3 yesterday, and received your

evaluation, but our results was not updated on the Online Leaderborder, we want to know whether it is because our results are not good enough.

Thank you and looking forward to your reply.

EXPERIENCE & DEDUCTIONS

Peeking Problem: Online submission via grand-challenge.org

- Online submission via grand-challenge.org with only university mail address is another way to stop peeking.
- Registration of a challenge is controlled by organizers.
- After approval of registration, participants upload their submissions to the site.
- Evaluation is handled by grand-challenge.org servers.
- Disadvantage: it is not hard to find another e-mail address (e.g. from a friend) and re-register to the challenge.

EXPERIENCE & DEDUCTIONS

Peeking Problem: Online submission via grand-challenge.org

- We added another rule for online submission via grand-challenge.org. A scientific manuscript explaining the method was requested during submissions.
- Participants must show the difference of their method if they have a previous submission.
- Only one submission per day is allowed.
- Disadvantage 1: it is possible to provide miss-information in these one page manuscripts.
- Disadvantage 2: cheating the one submission per day limit is easy with multiple registrations.

EXPERIENCE & DEDUCTIONS

Peeking Problem

- It is impossible to stop peeking on online submitted challenges unless examining the source code of the algorithm.
- We can make peeking harder but we cannot stop it completely.
- That is why, "*I think*" on-site results of a challenge represent more realistic and fair results. Only these results should be added to challenge publication.

EXPERIENCE & DEDUCTIONS

A More Serious Problem

- After evaluation of a submission, the final score is published on the leader board on grand-challenge.org.
- It is also possible to see individual scores and metrics of all cases.
- Participants may try to fix just the cases which they have lower scores.
- Unfortunately this “fix” might be handled by manual touches to the segmentation results.

EXPERIENCE & DEDUCTIONS

A More Serious Problem

- It is not hard to "*suspect*" manual corrections on lower scored cases.
- However it is not possible to intervene them without 100% proof.
- There is no way to stop this cheating.

EXPERIENCE & DEDUCTIONS

Problems with Annotation

- Despite of a hard work, it is a high possibility to have mistakes in data and annotations.

- The data and annotations should be checked by another person in organization team who has never seen them before.

ADVICES

My advices for potential challenge organizers

- Teamwork is crucial. A solid team with a solid plan is very important in challenge organization.
- I strongly recommend alternative plans in case of changes in the organization team.
- If the challenge is not the only work in your schedule, it is better to know that it will take too much time than you expected.

EXPERIENCE & DEDUCTIONS

If you will organize a challenge, …

- …annotation stage will take more time than planned.
- …some participants will not read the explanations in your website and documents.
- …there will be someone who requests the ground truth of test data persistently.
- …despite all the precautions, there will be teams those push so hard to exploit the challenge.

ADVICES

If I organized a new challenge, I would…

- … dedicate more time before organizing the challenge
- … push hard to find sponsor(s)
- … keep the tasks as simple as possible
- … prefer NIfTI file format instead of thousands of DICOM files
- … get ready for many e-mails all around the world (some of them will include questions with answers on the challenge website)
- … get ready for some rude comments and e-mails (especially from unsuccessful participants)

ADVICES

Despite of all difficulties, I am very happy…

- … to gain tremendous experience before and after the challenge,
- … to meet and contact many scientist in my field,

ADVICES

Despite of all difficulties, I am very happy…

- … to gain tremendous experience before and after the challenge,
- … to meet and contact many scientist in my field,
- … to find an opportunity to give this speech here. ☺

THANK YOU...

Contact Info

- emrekavur@gmail.com
- https://orcid.org/0000-0002-9328-8140
- https://www.researchgate.net/profile/Ali_Kavur