H2020 Programme

Rules of Contest for Prize

Big Data Technologies

Version 1.0
18 December 2017

RULES OF CONTEST

1. THEME: Big Data Technologies

1.1. Objective pursued

The objective of this inducement prize is to improve the performance of software for the forecasting of geospatio-temporal data (collections of time-stamped records that are linked to a geospatial location). The prize will reward a solution which improves existing methods in terms of scalability, accuracy, speed and use of computational resources.

1.2. Expected results

Many domains of societal or industrial significance, from epidemiology, to climate change, to transportation to energy production and transmission benefit from our ability to examine historical records and predict how the system under study will evolve.
In all these cases, it is not sufficient for predictions be accurate: they also need to be delivered fast enough for corrective action to be applied on the system observed.
The solution selected will demonstrate the ability to analyse extremely large scale collections of structured geospatial temporal data in a way that is sensitive to the trade-off between the consumption of computational resources and the practical value of the predictions obtained. Datasets will consist of time-series recording weather conditions and parameters of energy grid operations.
This will not only result in the more efficient management of those domains in which spatio-temporal predictions are already used, but also in the applications of such predictive methods where today they are not, due to current limitations of speed, scalability, accuracy and resource efficiency. Possible domains of application include but are not limited to logistics, manufacturing, telecommunications.
This inducement prize also complements the activities of the Big Data cPPP which aims to develop Europe's data driven economy and the prospects offered by Big Data technologies (as outlined in the Communication COM(2017) 9 final and the accompanying Staff Working Paper, adopted on 10 January 20171).

1.3. Glossary of concepts

Working software: the implementation of the prediction algorithm that the participant submits to the contest.

Contest platform: the computing platform on which the working software submissions are run against a test data. The contest platform measures the performance (accuracy, speed, resource consumption) of each working software submission. The contest platform will be used in two occasions: in an interactive mode which allows participants to test their software, and in a back-office mode where the pre-selection panel and the jury will score the working software submitted to the contest.

Starting kit: simulator of the contest platform that allows a participant to get familiar with the contest running environment and allows for the testing of the working software against sample datasets, representative to the actual test dataset

Data sets: After registration, participants will be granted access to **training and validation data** that they will be allowed to use to train and validate their working software on their own IT equipment, using the starting kit. At the contest platform, a different set of **testing data**, not accessible to the participants, will be used for testing, and to score and pre-rank the participants' working software. Yet another data set, **verification data** (data from the same process, but at a different time period) will be used for the verification runs and final ranking of the pre-selected applications by the jury (see section 7.2)

Persistence model, latest observed value (section 7.2): the prediction task that the working software is invited to tackle consists of observed values (consecutive measurements) from which the working software shall infer the future values. **Persistence model** is a trivial prediction method that assumes that the future value is the same as the given observed value. As the persistence model does not require any computing, it is used as an evaluation threshold: any solution that produces a prediction "worse" than the persistence model, is considered below threshold and will thus be eliminated.

2. PRIZE AMOUNT: 2 million EUR

1st place	*1,200,000 EUR*
2nd place	*600,000 EUR*
3rd place	*200,000 EUR*

[1] https://ec.europa.eu/digital-single-market/en/building-european-data-economy

3. DEADLINES & ADMISSIBILITY:

Deadlines	
Opening date for registration	19 December 2017
Closing date for registration	30 January 2018 at 17:00:00 CET
Opening date for application submission:	23 February 2018
Closing date for application submission:	9 April 2018 at 17:00:00 CET[2]

Joint applications by a group of participants are admitted. In this case, the participants must appoint a 'lead participant' to represent them towards the Commission. The participants will be jointly responsible and must all fulfil and respect the conditions set out in these Rules of Contest.

Applications must be submitted by the (lead) participant via the Participant Portal Submission Service.

Only participants that have completed a registration by the registration closing date are allowed to submit an application (see section 7 for details).

Applications must be readable, accessible and printable. Incomplete applications may be considered inadmissible if essential elements are missing (see General Annex B to the Main Work Programme).

A complete submission will consist of successful registration, and a submission of sections 1 and 2 (including the working software) of part B in the participant portal.

The size limit of Section 1 of Part B of the application is **10 pages** (point 5 of General Annex B to the main Work Programme). Content on any excess pages will be ignored. If essential information for assessing admissibility is contained on the excess pages, this may lead to inadmissibility of the submission.

Admissibility conditions related to the working software

The working software must

1. Not infringe the intellectual property rights of, nor be plagiarized from third parties and be written with the purpose of being intelligible to an expert software engineer, i.e. it must not be submitted in an obfuscated form i.e. organizing the source code in a way that makes it unnecessarily difficult to understand its structure and intent..

2. Follow a number of technical measures that will be specified in Section 9 (for example the seeding of random number generators so as to allow repeatability of results).

3. Be written in one of the following programming languages: **C, C++, Python 2, Python 3, Octave, Julia, R**. Script(s) must be written in the Linux scripting languages **sh** or **bash**, compatible with the Ubuntu 16.04 operating system.

4. Be fully of the participant's own work, or, if third-party components/libraries are used, these must be clearly identified, documented (in Section 1 of Part B) and separated from the participant's own work. The licensing status of any third-party components must also be described.

5. Not exceed 1 Megabyte (1 000 000 bytes) (size of the Section 2 of Part B which is the zip file containing the source code of the working software, see section 7).

Participants will be asked in the application forms to declare that they fulfil those admissibility requirements.

The Commission is not responsible for applications or working software not received for any reason, nor for working software received but not functioning as intended on the platform. If the submitted working software does not load, compile or run on the contest platform, or ends in an error or fails to produce required output within the required time, the submission is considered inadmissible. This also applies to any of the final verification runs of the best-ranking working software at the evaluation stage.

Sample application forms will be available on the Participant Portal Reference documents page.

[2] Central European Time = Brussels local time.

4. ELIGIBILITY

4.1. Eligibility criteria

The contest is open to any legal entity (including natural persons) or group of legal entities established in an EU Member State or in a country associated to Horizon 2020.

Please note, however, that special rules may apply for entities from certain countries (see General Annex C to the Main Work Programme).

Please also be aware that participants that have already received an EU or Euratom prize cannot receive a second prize for the same activities.

4.2. Exclusion criteria

Participants will be excluded if they (or one of them):

- are subject to an administrative sanction (i.e. exclusion)[3]
- are in one of the following situations[4]:
  - bankrupt, being wound up, having their affairs administered by the courts, entered into an arrangement with creditors, suspended business activities or subject to any other similar proceedings or procedures under national law (including persons with unlimited liability for the participant's debts)
  - declared in breach of social security or tax obligations by a final judgment or decision (including persons with unlimited liability for the participant's debts)
  - found guilty of grave professional misconduct[5] by a final judgment or decision (including persons having powers of representation, decision-making or control)
  - convicted of fraud, corruption, involvement in a criminal organisation, money laundering, terrorism-related crimes (including terrorism financing), child labour or human trafficking (including persons having powers of representation, decision-making or control)
  - shown significant deficiencies in complying with main obligations under a procurement contract, grant agreement or grant decision financed by the EU or Euratom budget (including persons having powers of representation, decision-making or control)
  - found guilty of irregularities within the meaning of Article 1(2) of Regulation No 2988/95 (including persons having powers of representation, decision-making or control)
- have misrepresented information required for participating in the contest or fail to submit such information
- were involved in the preparation of the prize documents and this entails a distortion of competition.

5. AWARD CRITERIA

The prize will be awarded to the application that best addresses the following cumulative criteria:

1. Accuracy of the prediction, expressed as "root-mean square error" (RMSE) rank of the software, as described in section 7.

2. Speed of delivery of the prediction, low use of computing resources: both criteria are expressed as the "overall elapsed execution time" (OEET) rank of the software, as described in section 7.

The method of calculating RMSE and OEET are described in Section 9.
The method of combining the two scores to determine relative ranking of applications is described in section 7.

[3] See Articles 131(4) and 106(1) Financial Regulation.

[4] See Articles 138(2) and 106(1), 107 of the Regulation (EU, Euratom) No 966/2012 of the European Parliament and of the Council of 25 October 2012 on the financial rules applicable to the general budget of the Union and repealing Council Regulation (EC, Euratom) No 1605/2002 (OJ L 218, 26.10.2012, p.1).

[5] Professional misconduct includes: violation of ethical standards of the profession, wrongful conduct with impact on professional credibility, false declarations/misrepresentation of information, participation in a cartel or other agreement distorting competition, violation of IPR, attempting to influence decision-making processes or obtain confidential information from public authorities to gain an advantage.

6. DOCUMENTS

The mandatory supporting documents are set out in the application form.

Participants may be asked at a later stage for further documents (for legal entity validation, bank account validation, ethics review, declaration of honor on exclusion grounds, etc.).

7. PROCEDURE

7.1. Submission

1. *Registration and trialing*: individuals and organisations wishing to participate in the contest are required to register After registration, participants will be able to download a starting kit allowing them to try out their working software in a small-scale system (running at the participant's own IT equipment) simulating the real contest system and allowing participants to get familiar with the requirements and scoring principles.

2. *Testing of the software on the contest platform*: Registered participants will receive instructions on how to activate their user account and upload their working software to the contest (cloud) platform and run it against the test data. The platform will issue a PDF file with the identification information of the participant, and the score describing the performance of the solution. These results are purely indicative, for the use of the participant only, and will not affect the evaluation. Each participant has up to three attempts[6] to run their working software on the contest platform and to ensure it will conform with the requirements.

3. Submission: registered participants submit their working software using the participant portal.

Section 1 of part B is a description of the working software and the method. Section 2 of part B is a zip file containing the working software (in source code). Part B will be submitted through the Participant Portal.

7.2. Evaluation

Each working software submitted to the participant portal is run on the contest platform and separately pre-ranked according to accuracy (using "root mean square error" (RMSE) as defined in Section 9) and speed (using the overall elapsed execute time, OEET, as defined in Section 9). Two ranking lists are produced, one based on lowest RMSE, the other one on lowest OEET.

The following three thresholds are then applied (in the following order):

1. Applications whose working software fails to complete all forecast steps within 6 hours OEET will be considered below threshold.

2. Applications having a rank greater than 10 in the RMSE rank list[7] *and* greater than 0.5 x (number of all ranked applications) in the same ranking are considered "below threshold".[8]

3. Applications whose RMSE is greater than the RMSE of a model assuming that the predicted variable remains constant at the latest observed value ("persistence model", see Glossary for definition) are considered "below threshold".[9]

The combined pre-ranking of the above-threshold applications is determined by a "combined rank" value which is the RMSE rank + 0.5 x OEET rank. The applications are then pre-ranked by increasing combined rank, with the lowest combined rank appearing at the top. This constitutes the pre-ranking list.

If there are more than 10 applications above threshold, there will be a pre-selection phase to select the best 10 applications to pass to jury review. Otherwise, all applications will pass directly to jury review.

---

[6] The contest platform will be equipped with a reasonable amount of computing resources. However, it cannot be guaranteed that the platform will provide immediate execution slot for all users at all times, especially in heavy usage situations. Heavy use is likely to arise near the closing date and time for submission. Participants are therefore encouraged to test their software early and not in the last day before the closing date for submission.

[7] In case of tied ranks (i.e. two or more applications having identical RMSE values) above and below the cutting line (e.g. 10th and 11th on the RMSE rank), all such tied applications are considered above threshold for this criterion.

[8] Example 1: if there are 30 applications on the RMSE rank list, the lowest-ranking (worst) 15 will be considered below threshold. Example 2: if there are 12 applications on the RMSE rank list, the lowest-ranking (worst) 2 will be considered below threshold.

[9] Note: the reference value of the "persistence model" RMSE is based on the data set to be used at the time of the evaluation, and is therefore specific to the data set used for the evaluation. However, the same data set (and hence, the same "persistence model" RMSE) will be used for all applicants that are being evaluated.

The pre-selection panel and jury usually have a different composition, but jury members may participate in the pre-selection panel.

The **pre-selection panel** will (1) run each submitted working software once on the contest platform, (2) will rank the applications applying the above ranking criteria and thresholds, (3) pre-select ten[10] highest-ranking applications.

The jury evaluation is planned to take place between March 2018 and June 2018.

The **jury** will establish the final ranking of pre-selected applications by running each working software three times and taking the arithmetic averages of the RMSE and OEET values as a basis of a new ranking which is determined using the same method and same thresholds (except threshold 2 which is not applied at this stage) as those used for pre-ranking, but using a different data set (verification data). If any of the three runs exceeds 6 hours OEET, the application is below threshold.

On the basis of the evaluation by the jury, the Commission will decide on the award of the prizes.

If two applications tie for the first rank, the combined prize amount of the first and second prize will be equally divided and awarded to both, and no second prize will be awarded. If three or more applications tie for the first rank, the combined prize amount of the first, second and third prize will be equally divided and awarded to all those applications, and no second and no third prize will be awarded. If two or more applications tie for the second rank, the combined prize amount of the second and third prize will be equally divided and awarded to all those applications, and no third prize will be awarded. If two or more applications tie for the third rank, the combined prize amount of the third position will be equally divided and awarded to all those applications.

All participants will be informed during the third quarter of 2018 on the outcome of their applications.

8. OTHER CONDITIONS

8.1. Payment arrangements

The prize money (EUR 1,200,000 for the first winner, EUR, 600,000 EUR for the second winner, EUR 200,000 for the third winner) will be paid to the (lead) participant(s) in one instalment after the award ceremony by bank transfer, provided all the requested documents have been submitted.

8.2. Publicity — Promoting the prize — Visibility of EU funding

8.2.1. Publicity by the winner(s)

The winners must promote the prize and its results, by providing targeted information to multiple audiences (including the media and the public) in a strategic and effective manner.

Unless the Commission requests or agrees otherwise or unless it is impossible, any communication activity related to the action (including in electronic form, via social media, etc.) must:

(a) display the EU emblem and

(b) include the following text:

> "This participant was one of the winners of the Inducement Prize: Big Data technologies from the European Union's Horizon 2020 research and innovation programme".

When displayed together with another logo, the EU emblem must have appropriate prominence.

For the purposes of their obligations, the winners may use the EU emblem without first obtaining approval from the Commission

This does not, however, give it the right to exclusive use.

Moreover, they may not appropriate the EU emblem or any similar trademark or logo, either by registration or by any other means.

8.2.2. Publicity by the Commission

[10] If there are less than 10 pre-ranked applications, they all proceed to evaluation by the jury.

The Commission may use, for its communication and publicising activities, information relating to the action, documents notably summaries for publication as well as any other material, such as pictures or audio-visual material that it receives from the participants (including in electronic form).

The Commission will publish the name of the winners, their origin, the amount of the prize and its nature and purpose — unless they have requested to waive this publication (because disclosure risks threatening its security and safety or harm its commercial interest).

The winners of the contest will be publicly announced in a public award ceremony.

Photos and videos taken by the Commission either in preparation of or during the award ceremony are the sole property of the Commission.

8.3. Dissemination and exploitation of results

The winners must comply with the obligations set out in Title III of the Horizon 2020 Rules for Participation Regulation No 1290/2013[11]

For more information and best practice, see Articles 23a-31 of the H2020 AGA — Annotated grant agreement.

8.4. Processing of personal data

8.4.1. Processing of personal data by the Commission

Any personal data will be processed by the Commission under Regulation No 45/2001[12] and in accordance with the Participant Portal privacy notice(s).

All winners consent that the Commission publishes the following information:

- name
- Member State of origin (address or NUTS 2 region)
- their activities in relation to the award of the prize (via the summary for publication they provided)
- prize amount

in whatever form and medium.

8.4.2. Processing of personal data by the participants

The participants must process personal data in compliance with applicable EU and national law on data protection (including authorisations or notification requirements, if any).

8.5. Ethics

The activities must be carried out in compliance with:

(a) ethical principles (including the highest standards of research integrity) and

(b) applicable international, EU and national law.

No prize will be awarded for activities carried out outside the EU, if they are prohibited in all Member States.

The participants must ensure that the activities have an exclusive focus on civil applications.

The participants must ensure that the activities do not:

(a) aim at human cloning for reproductive purposes

[11] Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 December 2013 laying down the rules for participation and dissemination in "Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020)" (OJ L 347, 20.12.2013 p.81).

[12] Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data (OJ L 8, 12.01.2001, p. 1).

(b) intend to modify the genetic heritage of human beings which could make such changes heritable (with the exception of research relating to cancer treatment of the gonads) or

(c) intend to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer.

Research activities involving human embryonic stem cells (hESC) are moreover subject to the conditions set out in the Statement of the Commission related to research activities involving human embryonic stem cells.

The participants must respect the highest standards of research integrity — as set out, for instance, in the European Code of Conduct for Research Integrity[13].

For more information and best practice, see the Participant Portal Online Manual, the Guidance — How to complete your ethics self assessment and the Guidance note — Research focusing exclusively on civil applications.

8.6. Security

The activities must be carried out in compliance with Commission Decision 2015/444, i.e. security-sensitive information must be **EU-classified**, if its unauthorised disclosure could adversely impact the interests of the EU or of one (or more) of its Member States. Applications that are too security-sensitive cannot be awarded a prize.

For more information and best practice, see the Guidance — Guidelines for the classification of information in research projects, the Guidance — Guidelines for the handling of classified information in EU research projects, the Guidance note — Potential misuse of research results and the Guidance note — Research involving dual use items.

8.7. Conflict of interests

The participants must take all measures to prevent any situation where the impartial and objective award of the prize is compromised for reasons involving economic interest, political or national affinity, family or emotional ties or any other shared interest ('conflict of interests').

They must inform the Commission without delay of any situation constituting or likely to lead to a conflict of interests and immediately take all the necessary steps to rectify this situation.

The Commission may verify that the measures taken are appropriate and may require additional measures to be taken by a specified deadline.

8.8. Liability for damages

The Commission cannot be held liable for any damage caused to the participants or to third parties as a consequence of the prize, including for gross negligence.

The Commission cannot be held liable for any damage caused by any of the participants in the context of the prize.

8.9. Checks, audits and investigations

The Commission, the European Anti-Fraud Office (OLAF) and the European Court of Auditors may carry out checks, audits and investigations in relation to the prize.

8.10. Withdrawal of the prize — Recovery of undue amounts

The Commission may withdraw the prize after its award and recover all payments made, if it finds out that:

(a) false information, fraud or corruption was used to obtain it

(b) a winner was not eligible or should have been excluded

(c) a winner is in serious breach of its obligations under these Rules of Contest.

8.11. Administrative sanctions

If a participant has committed irregularities or fraud or has made false declarations, the Commission may also:

[13] European Code of Conduct for Research Integrity of ALLEA (All European Academies) and ESF (European Science Foundation) of March 2017 http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf

(a) exclude the participant from all future contracts, grants and contests financed from the EU or Euratom budget for a maximum of five years (or 10 years in case of repetition) and/or

(b) impose a financial penalty between 2% and 10% of the value of the prize (or between 4% and 20% in case of repetition).

8.12. Cancellation of the contest

The Commission may cancel the contest or decide not to award the prize — without any obligation to compensate participants —, if:

(a) no admissible applications are received

(b) the jury does not find a winner

(c) the winner is not eligible or must be excluded or

(d) the objective of the contest has already been achieved.

8.13. Complaints

Complaints against decisions negatively affecting the rights of a participant or winner can be brought before the General Court — or, on appeal, the Court of Justice of the European Union — under Article 263 of the Treaty on the Functioning of the EU (TFEU).

CONTACT

For more information, please see the prize website ec.europa.eu/research/horizonprize/bigdata/.

In case of questions, please contact *CNECT-BIG-DATA-PRIZE@ec.europa.eu*.

9. Technical description of evaluation metrics

This section describes the objective metrics of performance evaluation of submissions, both in terms of accuracy and computational resource use.

Glossary of terms used

Term	Meaning	Range
Data-referenced time	Time as expressed in time stamps in the data itself, relative to the process that generated the data, and recorded as a historical time series	
Prediction generation time	*Data-referenced time* at which the contest platform requests the software to generate a prediction	
Lead time	future *data-referenced time* (positive, a scalar in time units), relative to the *prediction generation time*	
m	*m*-th prediction step,i.e: the m-th generation of a prediction in a given sequence, the *m*-th prediction generation time	$1..M$
M	Total number of prediction steps, the entire sequence of which constitutes the task which the software is expected to perform.	
h	*h*-lead prediction component (e.g. 1st lead: 5 minutes ahead, 2nd lead: 10 minutes ahead, 3rd lead: 15 minutes ahead... up to *H*)	$1..H$
H	Prediction horizon (the number of points in time in the future at which a prediction component is requested for a single generation)	
n	*n*-th component of the target vector (e.g.. the traffic on the *n*-th power line of the power grid)	$1..N$
N	Dimensionality of time-varying target vector (e.g. total number of grid power lines considered in the prediction task). A prediction component has N values, a prediction step generates N*H values, the overall task of the software is to produce M*N*H scalar values	
$\tilde{y}$	Target vector: forecast provided by the participant's software	
y	Target vector: ground truth (the "correct" future values, unknown to the software at the time of the prediction step, but made available to it after the prediction step)	
ET(x,z)	Execution time (between events x and z), relative to a fixed machine	

1. Evaluation/Execution of .zip formatted submissions (summary).

The software provided by participants must be written in a legible way and converted to a zip file. The software will be run and evaluated on a contest platform with a specially built Ubuntu 16.04 Docker container which already has support for multiple programming languages (such as Python, R and Octave), and includes GPU computing support. At each prediction step m=1..M, the following sequence occurs:

A) The participant's software will be presented historical time-series data in a given folder, simulating data that would be available at a given time that a prediction generation is requested (i.e. simulates a real-time forecasting scenario). The available data includes both 1) historical values of the data whose future values are to be predicted, and 2) other spatio-temporal data that may or may not correlate with the data whose values are to be predicted. The time steps in the historical time series are not necessarily spaced at equal intervals. The format of this data includes, but is not limited to, the hdf5 file format. The historical values of the data whose future values are to be predicted will be provided in the hdf5 format.

B) An executable, named *predict.sh*, provided by participants as part of their .zip file submission, is called by the contest platform. This executable may in turn call any of the other code present in the submitted .zip file, or any of the libraries present on the container. It must place the predictions (target vectors for *H successive* future lead times requested) in an output hdf5-formatted file at a pre-specified location, and then call a special time-stamp executable script called *makeready.sh*, provided by organizers. Once this function is successfully called, the output file may no longer be modified and the organizers' software proceeds to C) below.

C) The organizers' software then returns to step A) unless m=M, at which the prediction task is considered complete, or the OEET up to the current time-step has exceeded its upper limit (OEET threshold, specified in the Rules of Contest), in which case the execution of the prediction task is considered unsuccessfully completed.

After every prediction step (except the last one), the ground-truth vectors of the completed prediction step are made available to the software.

2. Performance evaluation: Forecast Accuracy (RMSE)

The Root Mean Square Error (RMSE) is calculated over all the predictions made, with equal weight for all of the scalar components of the predictions generated in the task. In other words**:**

$$RMSE = \sqrt{\frac{1}{MHN}\sum_{m=1}^{M}\sum_{h=1}^{H}\sum_{n=1}^{N}\left(\tilde{y}_{mhn} - y_{mhn}\right)^2}$$

See 'Glossary of terms used' above for definitions of terms.

3. Performance evaluation: Overall Elapsed Execution Time (OEET)

The Overall Elapsed Execution Time (OEET) for any given prediction step m is the elapsed time between the start of event B) and the end of event B) in section 1 above, and summed for all forecast steps m=1..M.

$$OEET = \sum_{m=1}^{M} ET\left(B_{start,m}, B_{end,m}\right)$$

The reference machine is the Amazon Web Services 'p2.8xlarge' machine. It is up to the participants to utilize as much of the hardware (GPUs, CPUs, RAM) available on the reference machine that they see fit, within the overall time limit for OEET.

Note

It is expected that software will not produce the same exact result both in terms of RMSE and OEET when run on identical data. It is up to participants to control the non-deterministic nature of their code execution to minimize its effect (e.g. special attention should be given to seeding of random number generators).