Appendix C. Raw Classification Data

Here we provide additional details about the raw classification data provided in the online supplementary data file[8]. It is written in the binary HDF5 format, in the variant produced by the *pandas* library (supported by the PyTables library[9]).

The general structure is as follows: Each classification submission by an individual volunteer creates a *classification_id*. All objects created by this volunteer receives the same *classification_id*, with the marking data for each object being one entry in the classification database. Each data row also has a *marking* column that identifies if this data is for a fan, a blotch, an interesting feature that will have the string value "interesting" in the *marking* column, or "none", when the volunteer did not create any marking object. Below we describe the columns available in this database:

Column name	Example value	Description
classification_id	50ecaaf760d4050d21000414	Unique ID for each classification by a Planet Four volunteer
created_at	2013-01-08 23:25:43	time of submission
tile_id	APF0000p9t	Planet Four tile identifier

[8] `2018-02-11_planet_four_classifications_queryable_cleaned_seasons2and3.h5`
[9] http://pandas.pydata.org/pandas-docs/stable/io.html#hdf5-pytables

image_name	ESP_021491_0950	HiRISE observation identifier
tile_url	http://www.planetfour.org/subjects/standard/50e741555e2ed211dc002346.jpg	URL to image data for this Planet Four tile
user_name	abc	Originally, the Zooniverse username or non-logged-in session ID. For privacy concerns, we have converted these to anonymous IDs.
marking	blotch	identifier for what data in row is for: blotch, fan, interesting, none
x_tile	1	x coordinate of tile inside larger HiRISE image frame. Starts at 1 in upper left of the HiRISE image, increases to the right.
y_tile	2	y coordinate of tile inside larger HiRISE image frame. Starts at 1 in upper left of the HiRISE image and increase downwards.
acquisition_date	2011-01-01 00:00:00	date only for HiRISE observation time (ignore hours)
local_mars_time	5:43 PM	local mars time for given acquisition date
x	553.65	x pixel coordinate of object in Planet Four tile. Starts at 0 in upper left, increases to the right.
y	355.817	y pixel coordinate of object in Planet Four tile. Starts at 0 in upper left, increases downwards.
image_x	2033.65	x pixel coordinate of object in original HiRISE image. Starts at 0 in upper left, increases to the right.
image_y	37071.8	y pixel coordinate of object in original HiRISE image. Starts at 0 in upper left, increasing downwards.
radius_1	295.195	Semi-major axis of blotch object in pixels. NAN if not applicable (N/A)
radius_2	294.715	Semi-minor axis of blotch object in pixels. NAN if N/A
distance	NaN	Length of fan object in pixels. NAN if data row is for blotch or interesting

angle	27.4331	Orientation of marking object with respect to tile image x-axis in degrees. Positiv clock-wise, zero to image right (same definition as HiRISE)
spread	NaN	Opening angle of fan objects in degrees. NAN if N/A
version	NaN	version of tool used to create fan. NAN if N/A
x_angle	0.887549	cartesian x coordinate of *angle* column on unit circle
y_angle	0.460713	cartesian y coordinate of *angle* column on unit circle

The Planet Four classification interface recorded a different angle than the intended spread angle from the fan marking tool. This was identified and subsequently fixed in the software. The correct spread angle is recoverable from the values stored in the database. We denote those markings generated before the patch with version flag set to 1.0 and those after with the version flag set to 2.0. We provide the corrected spread angle for the fans affected, but leave that version flag in the final catalog, for reference. To gather statistics on the understanding of the tutorial, the Planet Four classification database contains all the tutorial markings, indicated by a HiRISE image name of 'tutorial'. For the delivered raw classification database, the fan angles range has been converted from -180–180 to 0–360, while the range of the blotch angles have been converted to 0–180, due to their rotational symmetry.

Appendix D. Pipeline outputs

The intermediate stages of the pipeline, as output by our clustering and combination pipeline are identified with different level identifiers 1A, 1B, and 1C, indicating different stages of the processing pipeline, where the processing is done on a per-tile-id level. After this is done, the final step of combines all the data from the ten-thousands of tile_id folders into a set of summarizing CSV files.

Appendix D.1. Directory file structure

The directory file structure of the pipeline products are as follows (examples in parentheses):

- HiRISE observation ID (`ESP_011350_0945`)
  - Planet Four tile ID (`APF0000any`)
    - Level 1A (`L1A/APF0000any_L1A_fans.csv`)
    - Level 1B (`L1B/APF0000any_L1B_fnotches.csv`)
    - Level 1C with cut value 0.5 in directory name (`L1C_cut_0.5/APF0000any_L1C_cut_0.5_blotches.csv`)

with the list of HiRISE observation IDs identifying the HiRISE observations that went into Planet Four for this database.

Appendix D.2. Pipeline stage levels

Appendix D.2.1. Level 1A

Level 1A is the data that is directly output from clustering and averaging the cluster members into average markings, as described in Section 4.2. Here, the biggest reduction in terms of numbers of objects in the system occurs, as all the different volunteers data are being combined into one object when the clustering process has determined the markings to be part of one cluster. All newly created average fans and blotches are summarized into one fan and blotch summary file respectively, which each line representing the mean object from averaging all cluster members. As an example, the content of `APF0000p3q_L1A_fans.csv` is shown below. When the column name matches those given in Appendix Appendix C, they have the same meaning. The two new columns are *n_votes*, which records how many members the cluster had that was used to produce this averaged object, and *marking_id*, which have been created at this stage of the pipeline and serve as a tracer throughout the different pipeline outputs:

	x_tile	y_tile	x	y	image_x	image_y	radius_1	radius_2
0	2.0	26.0	123.611111	455.666667	863.611111	14155.666667	NaN	NaN
1	2.0	26.0	157.000000	391.800000	897.000000	14091.800000	NaN	NaN

	distance	angle	spread	version	x_angle	y_angle	n_votes	image_id
0	81.884266	223.712817	71.559689	1.0	-0.691035	-0.660663	9	APF0000any
1	57.742472	248.754137	52.521798	1.0	-0.360802	-0.927999	10	APF0000any

	image_name	marking_id
0	ESP_011350_0945	F006de3
1	ESP_011350_0945	F006de4

Additionally, each L1A folder contains a text file called `clustering_setttings.yaml` that summarizes the clustering settings used for these data for reference. *epsilon* values are static and all the same, but the *min_samples* value is dynamically calculated, see Section 4.2.1 for details.

Appendix D.2.2. Level 1B

At level 1B, the combination pipeline has determined with objects are so close to each other that they should be considered for merging (see Section 4.3). The outputs are between one and three files this time. One only, in case all fans and blotches found were so close that they need to be evaluated by their classification votes. Usually, though, there are two to three files, where one files stores the objects that need voting, and the other file(s) store the objects that don't have any close neighbors and will simply be copied over to the final level later. The fans and blotches in these latter files will receive the 'vote_ratio' value of 1.0, indicating that they had a "perfect" probability for being a fan, or blotch, respectively. The third file that keeps the close objects for the later thresholding contains these temporary meta-objects in sets of 2 rows, one fan and one blotch, and has the term "fnotch" in its filename (fnotches: FaN–blOTCH). This file contains all the clustering statistics data from L1A required to make a cut decision for L1C, with the data for

each meta-object being sorted in alternating rows. Here are the first four rows of the fnotch file `APF0000any_L1B_fnotches.csv`:

	angle	distance	image_id	image_name	image_x	image_y
fan	223.712817	81.884266	APF0000any	ESP_011350_0945	863.611111	14155.666667
blotch	67.261720	NaN	APF0000any	ESP_011350_0945	838.395834	14123.875000
fan	247.146845	58.742330	APF0000any	ESP_011350_0945	832.000000	14306.400000
blotch	70.684606	NaN	APF0000any	ESP_011350_0945	821.666667	14281.428571

	marking_id	n_votes	radius_1	radius_2	spread	version	x	x_angle
fan	F006de3	9	NaN	NaN	71.559689	1.0	123.611111	-0.691035
blotch	B0071f2	8	49.309277	36.981958	NaN	NaN	98.395834	0.379131
fan	F006de5	5	NaN	NaN	81.171448	1.0	92.000000	-0.387419
blotch	B0071ed	7	35.324591	26.493443	NaN	NaN	81.666667	0.217508

	x_tile	y	y_angle	y_tile	vote_ratio
fan	2.0	455.666667	-0.660663	26.0	0.539412
blotch	2.0	423.875000	0.907431	26.0	0.460588
fan	2.0	606.400000	-0.919245	26.0	0.426667
blotch	2.0	581.428571	0.852341	26.0	0.573333

This data stage L1B is what can be used to create a different significance threshold cut for the final data , by filtering on the data column *vote_ratio* in the fnotch file for the required threshold value. For example, if a higher threshold on the probability for a fan is wanted, e.g. 0.8, one would filter out all rows that start with "fan" with a *vote_ratio* value below 0.8. One then needs to decide if one wants to use this threshold as a general "certainty" filter and simply don't take any object with a *vote_ratio* < 0.8, or if one wants the blotch to appear instead of a fan.

Appendix D.2.3. Level 1C

This level contains the data of the final catalog files, but split-up into each Planet Four tiles. At the end of the thresholding stage (Section 4.3), appending the data for the rows that pass the threshold filters into the respective blotch and fan files and copying these completed files into the L1C directory completes that thresholding step and fills up the L1C folders. A final tool walks through each folder and collects all the fan and blotch data into one summary file each, followed by merge operations with meta-data that is useful for future analysis. These files are described in the next section, Appendix E.

Appendix E. Planet Four Catalog files description

Our catalog product files consist of one CSV result file per fan and blotch markings, a Planet Four tile meta-data file, and a HiRISE observation meta-data file. Below, each subsection describes the data columns for these files.

For convenience we provide both the planeto-centric and planeto-graphic latitudes for each fan's base and blotch's center point. Longitudes are measured 0–360, increasing positive to the East. Note that, because the HiRISE images were not co-registered, the conversion of pixel to geographical coordinates can be offset by up to 100 HiRISE pixels between data from different HiRISE images.

Appendix E.1. Fan catalog

Column name	**Example value**	**Description**
marking_id	F00004ab	Consistent identifier for marking after clustering. Fxxx=Fan, Bxxx=Blotch
angle	185.4	Alignment angle of marking measured from 3 o'clock direction, clockwise
distance	179.6	Length of fan in pixels
tile_id	`APF0000cia`	tile identifier in the Planet Four system
image_x	3391.2	Base X coordinate [px] in original HiRISE image
image_y	5640.6	Base Y coordinate [px] in original HiRISE image
n_votes	15	# of markings that went into this average object.
obsid	`ESP_012079_0945`	HiRISE image observation id
spread	21.346	Spreading angle of Fans
version	1	Version number of Fan model used in Planet Four (see Appendix Appendix C)
vote_ratio	1.0	Ratio of votes from a potential combination step. Value of 1.0 means only fan votes occurred.
x	431.206	Base X pixel coordinate in the Planet Four tile
y	160.6	Base Y pixel coordinate in the Planet Four tile
x_angle	-0.995088	Polar X coordinate of alignment angle
y_angle	-0.0938355	Polar Y coordinate of alignment angle
l_s	214.785	Solar longitude of HiRISE observation
map_scale	0.25	Factor for scaling distances to correct for HiRISE binning mode
north_azimuth	126.857	Direction of North in the original unprojected HiRISE input image
BodyFixedCoordinateX	-67.2071	Base X coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateY	257.05	Base Y coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateZ	-3370.63	Base Z coord. [km] in Mars-fixed ref. frame

PlanetoCentricLatitude	-85.493	Latitude of catalog object (-centric)
PlanetoGraphicLatitude	-85.5457	Latitude of catalog object (-graphic)
Longitude	104.652	Longitude of catalog object

Appendix E.2. Blotch catalog

Column name	**Example value**	**Description**
marking_id	B00004ab	Consistent identifier for marking after clustering. Fxxx=Fan, Bxxx=Blotch
angle	185.4	Alignment angle of marking measured from 3 o'clock direction, clockwise
tile_id	APF0000cia	tile identifier in the Planet Four system
image_x	3391.2	Center X pixel coordinate in the original HiRISE image
image_y	5640.6	Center Y pixel coordinate in the original HiRISE image
n_votes	15	Number of markings used for the average object
obsid	ESP_012079_0945	HiRISE image observation id
radius_1	10.4	Semi-major axis of Blotch
radius_2	15.2	Semi-minor axis of Blotch
vote_ratio	0.0	Ratio of votes from a potential combination step. Value of 0.0 means only blotch votes occurred.
x	431.206	Center X pixel coordinate in the Planet Four tile
y	160.6	Center Y pixel coordinate in the Planet Four tile
x_angle	-0.995088	Polar X coordinate of alignment angle
y_angle	-0.0938355	Polar Y coordinate of alignment angle
l_s	214.785	Solar longitude of HiRISE observation
map_scale	0.25	Factor for scaling distances to correct for HiRISE binning mode
north_azimuth	126.857	Direction of North in the original unprojected HiRISE input image
BodyFixedCoordinateX	-67.2071	Center X coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateY	257.05	Center Y coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateZ	-3370.63	Center Z coord. [km] in Mars-fixed ref. frame
PlanetocentricLatitude	-85.493	Latitude of catalog object (-centric)
PlanetographicLatitude	-85.5457	Latitude of catalog object (-graphic)
Longitude	104.652	Longitude of catalog object (Positive East 360)

Appendix E.3. Planet Four tile catalog

Here we provide the data required to position the Planet Four tiles both back into HiRISE images, if so required, or directly onto the Martian surface, by using the provided latitude/longitude values or their map-value equivalents in the BodyFixed-Mars frame in a rectangular coordinate system, measuring kilometers from the south pole. The coordinate values come directly from the ISIS campt utility, while the x_tile and y_tile position indices of tiles inside the HiRISE image are the result of the splitting up routine that was developed by the Zooniverse team at the beginning of the project. All coordinates were calculated at the tile center pixel coordinate of (420, 324). The decimal digits precision was set to 7, guided by the Latitude/Longitude significant bits for a HiRISE pixel diameter on the ground for a 1x1 binning observation.

Column name	**Example value**	**Description**
BodyFixedCoordinateX	-67.2071	Center X coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateY	257.05	Center Y coord. [km] in Mars-fixed ref. frame
BodyFixedCoordinateZ	-3370.63	Center Z coord. [km] in Mars-fixed ref. frame
PlanetocentricLatitude	-85.493	Latitude of catalog object (-centric)
PlanetographicLatitude	-85.5457	Latitude of catalog object (-graphic)
Longitude	104.652	Longitude of catalog object (Positive East 360)
tile_id	`APF0000cia`	tile identifier in the Planet Four system
obsid	`PSP_003092_0985`	HiRISE observation ID of the source image for this tile
x_hirise	840	X pixel coordinate of the tile center in the HiRISE image
x_tile	5	X index of the Planet Four tile inside the HiRISE image (1-based)
y_hirise	648	Y pixel coordinate of the tile center in the HiRISE image
y_tile	11	Y index of the Planet Four tile inside the HiRISE image (1-based)

Appendix E.4. HiRISE observations catalog

This catalog provides the user with a list of HiRISE images and their meta-data that were used to create the Planet Four results presented here. The columns with capital letters were directly taken from the published cumulative EDR index[10]. The decimal digits precision was set to 7, guided by the Latitude/Longitude significant bits for a HiRISE pixel diameter on the ground for a 1x1 binning observation.

Column name	**Example value**	**Description**

[10] https://hirise-pds.lpl.arizona.edu/PDS/INDEX/EDRCUMINDEX.TAB

OBSERVATION_ID	ESP_011296_0975	HiRISE observation identifier
IMAGE_CENTER_LATITUDE	-82.1965000	Planetographic latitude of the HiRISE image center
IMAGE_CENTER_LONGITUDE	225.2530000	Longitude of HiRISE image center (positive west 360)
SOLAR_LONGITUDE	178.8330000	Solar longitude of HiRISE image. Equivalent to column *l_s* in the fan and blotch catalogs.
START_TIME	2008-12-23 16:15:26	UTC time of observation start
map_scale	1.0000000	Units: pixel/m. Calculated from EDRCUMINDEX by 0.25*BINNING
north_azimuth	110.6001067	The median north azimuth value for the HiRISE image, recalculated with ISIS' campt, due to known errors in HiRISE EDR index file.
# of tiles	91	the number of created Planet Four tiles per HiRISE observation. Depends on original image size.