evidation

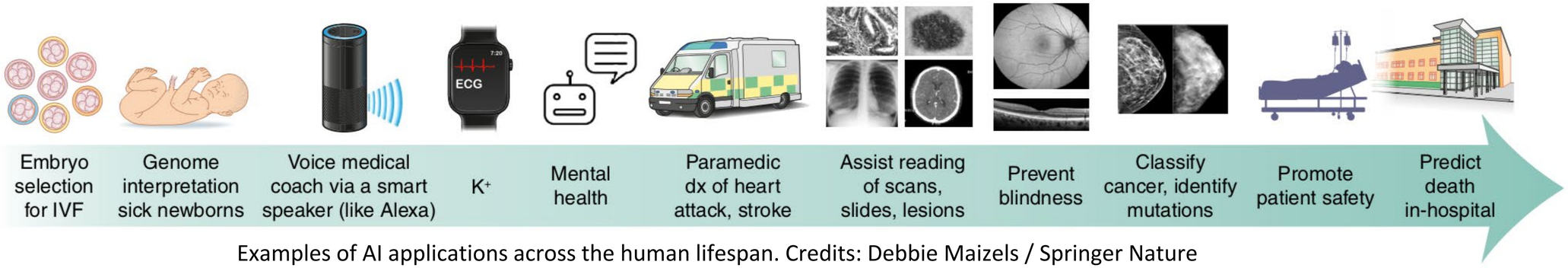
Trustworthiness of Person-generated Health Data

JULY 2019

Luca Foschini, PhD

Co-founder & Chief Data Scientist

luca@evidation.com | @calimagna



Examples of AI applications across the human lifespan. Credits: Debbie Maizels / Springer Nature

Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

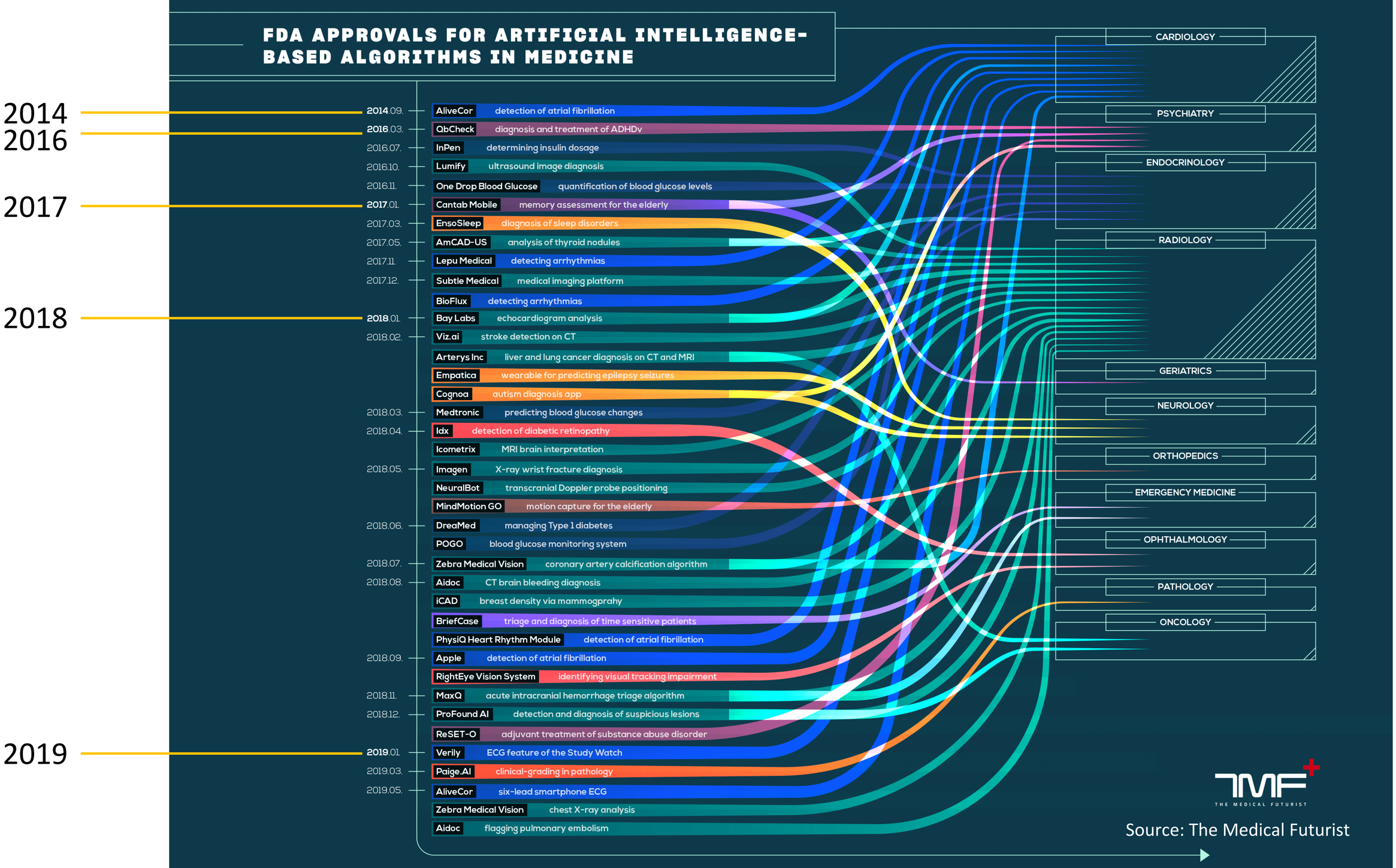
Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93*0.75+0.85#	Rajkomar et al.[96]
All-cause 3–12 month mortality	221,284	0.93^	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	0.85@	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82++	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

LOS, length of stay; *n*, number of patients (training+validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ^, all patients; @, structured+unstructured data; ++, for University of Michigan site.

Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

© EVIDATION HEALTH, INC. 2019

FDA APPROVALS FOR ARTIFICIAL INTELLIGENCE-BASED ALGORITHMS IN MEDICINE



Source: The Medical Futurist

To trust an MI system requires to trust all steps of its workflow

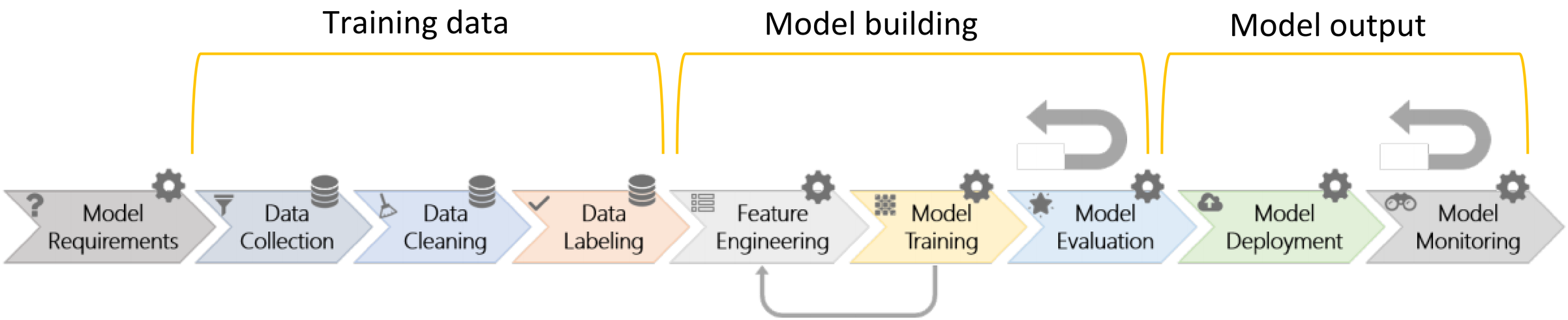


Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: **Software Engineering for Machine Learning: A Case Study** Amershi et al., International Conference on Software Engineering (ICSE) 2019

© EVIDATION HEALTH, INC. 2019

To trust an MI system requires to trust all steps of its workflow

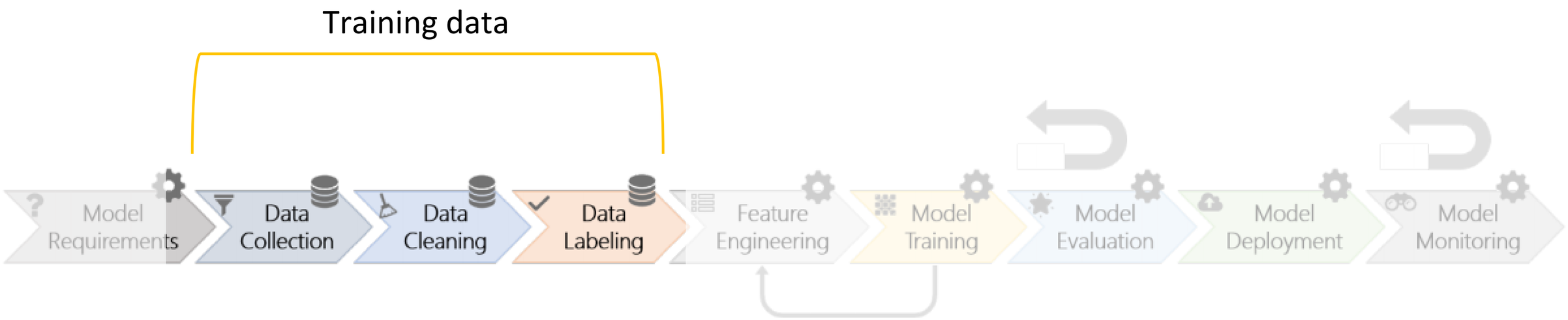
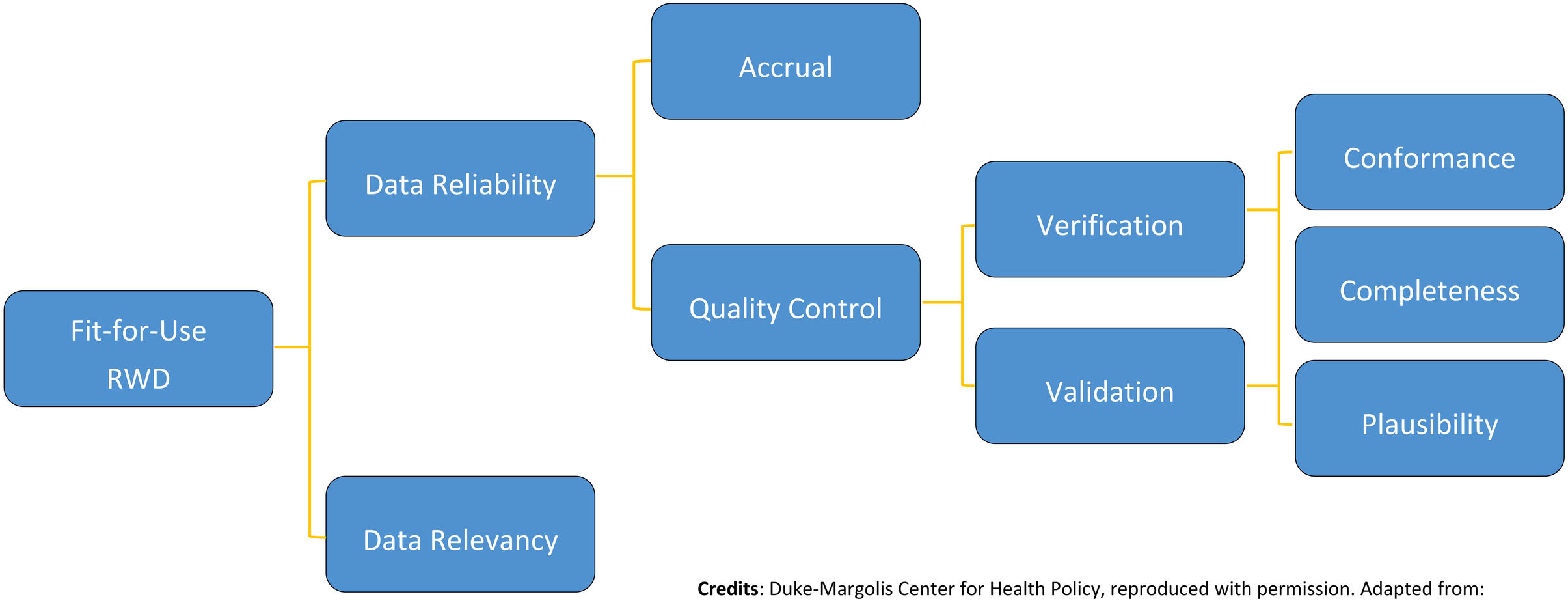


Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: **Software Engineering for Machine Learning: A Case Study** Amershi et al., International Conference on Software Engineering (ICSE) 2019

© EVIDATION HEALTH, INC. 2019

Define ”bad data”: a Real-World Evidence (RWE) perspective



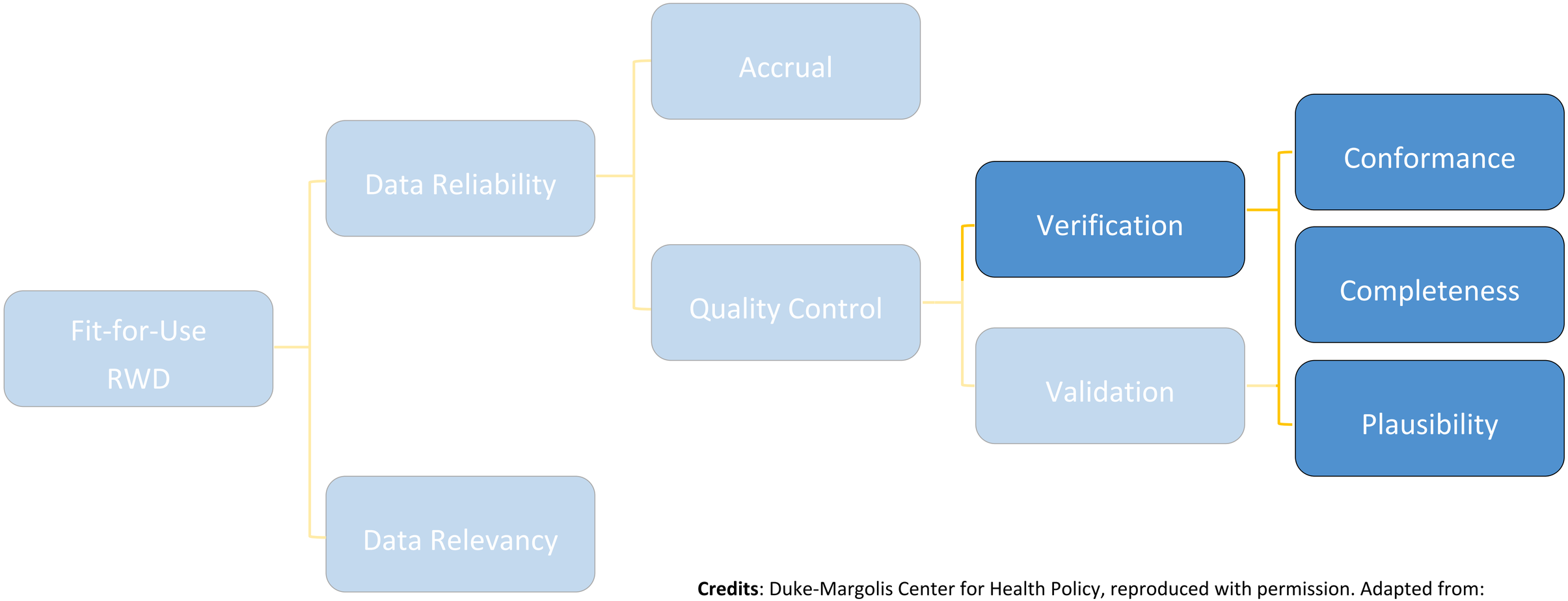
Credits: Duke-Margolis Center for Health Policy, reproduced with permission. Adapted from:

- **A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data**. Kahn, et al. EGEMS 2016
- **Framework for FDA’s Real World Evidence Program**. Food and Drug Administration, 2018

Part of a forthcoming Duke-Margolis RWE Collaborative White Paper on RWD Quality.

© EVIDATION HEALTH, INC. 2019

Define ”bad data”: a Real-World Evidence (RWE) perspective



Credits: Duke-Margolis Center for Health Policy, reproduced with permission. Adapted from:

- **A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data**. Kahn, et al. EGEMS 2016
- **Framework for FDA’s Real World Evidence Program**. Food and Drug Administration, 2018

Part of a forthcoming Duke-Margolis RWE Collaborative White Paper on RWD Quality.

© EVIDATION HEALTH, INC. 2019

Evidation Study Platform ingests and analyze high-frequency Person-generated Health Data (PGHD) from a variety of sensors, services, and applications

Sensors

Services

Examples of PGHD

- Biometric data
- Symptoms
- Self-reported medical history
- Activity tracking
- Medication adherence/effects
- Sleep tracking

© EVIDATION HEALTH, INC. 2019

Conformance*:* Does the data fit our expected schema?

- Ingesting sleep data from Fitbit
- *Two different data formats depending on device and data availability*

```
"data": [
    {
        "datetime": "2017-04-01T23:58:30.000",
        "level": "wake",
        "seconds": <value>
    },
    {
        "datetime": "2017-04-02T00:16:30.000",
        "level": "rem",
        "seconds": <value>
    },
    <...>
],
"shortData": [
    {
        "datetime": "2017-04-02T05:58:30.000",
        "level": "wake",
        "seconds": <value>
    },
```

timestamp	level_deep_seconds	level_light_seconds	level_rem_seconds	level_wake_seconds
2017-04-01T23:58:30.000	NaN	NaN	NaN	1080
2017-04-02T00:16:30.000	NaN	NaN	360	NaN

```
"data": [
    {
        "dateTime": "2017-04-02T12:06:00.000",
        "level": "asleep",
        "seconds": <value>
    },
    {
        "dateTime": "2017-04-02T12:13:00.000",
        "level": "restless",
        "seconds": <value>
    },
    {
        "dateTime": "2017-04-02T12:14:00.000",
        "level": "awake",
        "seconds": <value>
    },
```

timestamp	level_deep_seconds	level_light_seconds	level_rem_seconds	level_wake_seconds
2017-04-01T23:58:30.000	NaN	NaN	NaN	NaN
2017-04-02T00:16:30.000	NaN	NaN	NaN	NaN

© EVIDATION HEALTH, INC. 2019

Completeness*:* How much data do we observe, and do our observations meet our expectations?

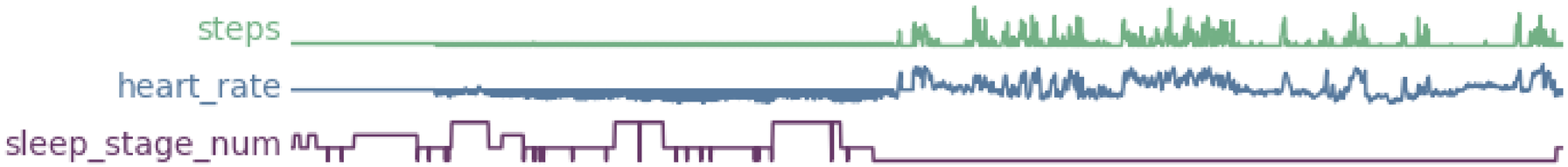
- Data availability relies on both the systems that capture/store **and** the patients who generate it
  - Participant Missing: stopped wearing the device; does not sync
  - Device Missing: battery died and stopped recording
  - System Missing: server/API errors

© EVIDATION HEALTH, INC. 2019

Completeness*:* How much data do we observe, and do our observations meet our expectations?

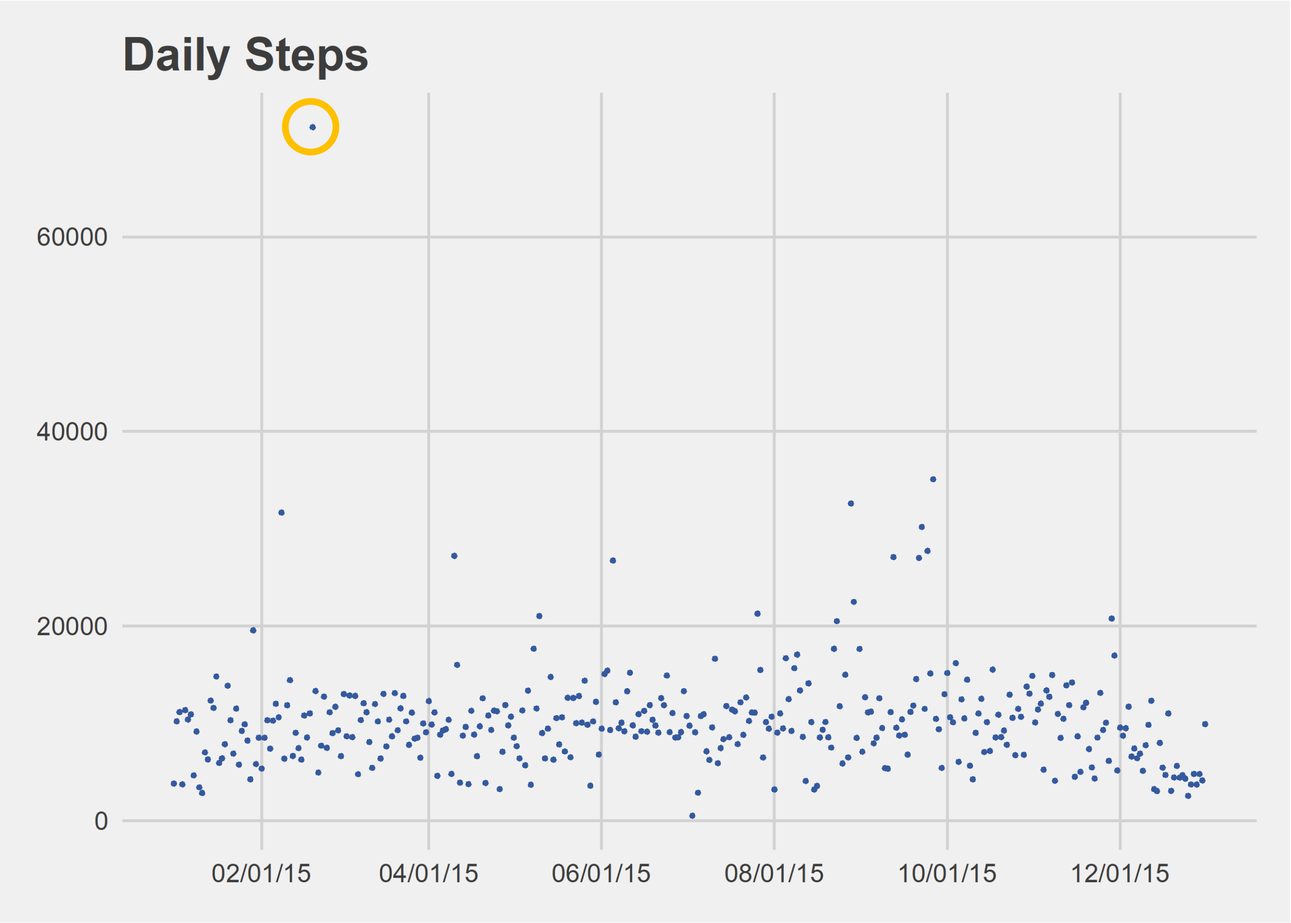
- Data availability relies on both the systems that capture/store **and** the patients who generate it
  - Participant Missing: stopped wearing the device; does not sync
  - Device Missing: battery died and stopped recording
  - System Missing: server/API errors

- Multi-sensor systems allow for data completeness checks that take into account multiple data streams
  - Allows you to differentiate true missing from null values



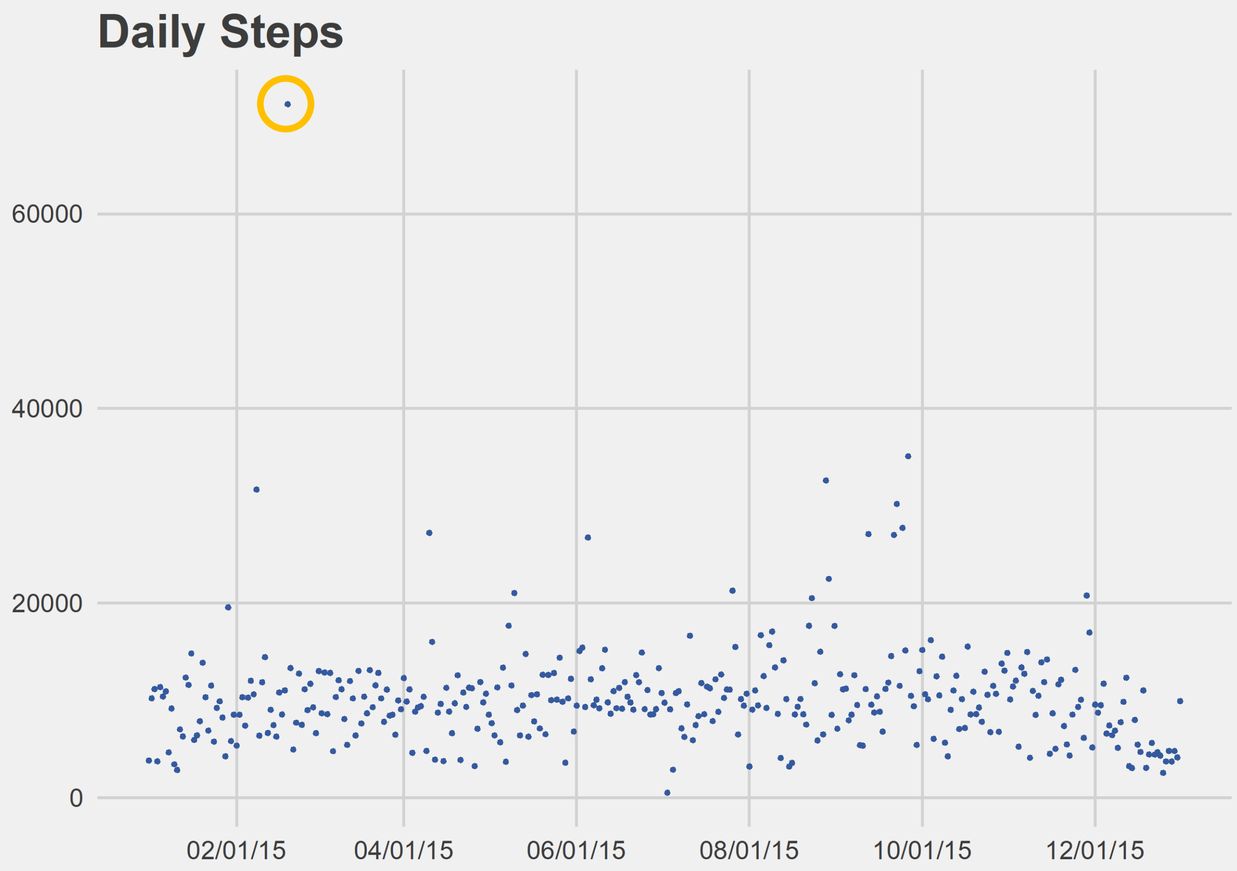
© EVIDATION HEALTH, INC. 2019

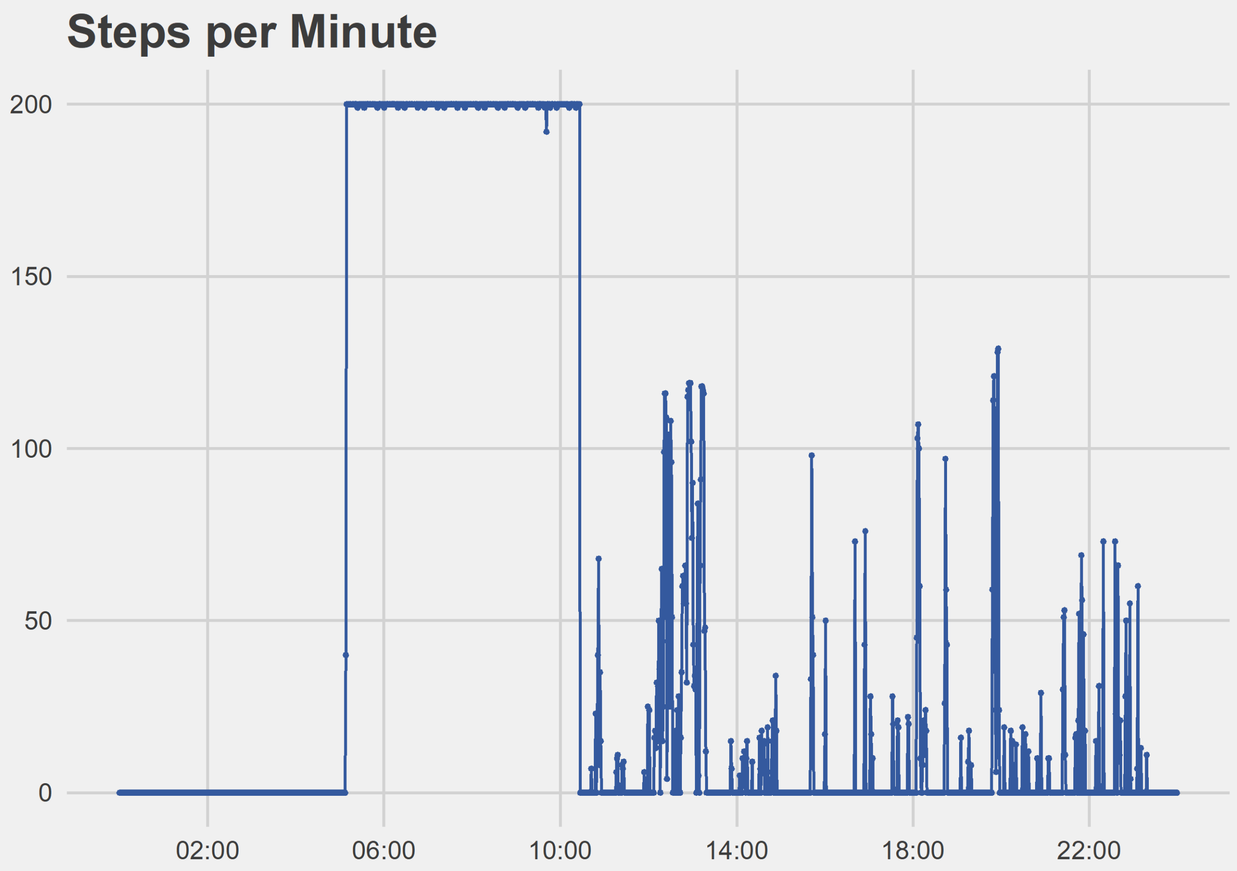
Plausibility: Can we trust the values we observe in the data?



© EVIDATION HEALTH, INC. 2019

Plausibility: Can we trust the values we observe in the data?





© EVIDATION HEALTH, INC. 2019

Need: Data documentation frameworks for medical MI applications

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[3]. The

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The following steps were taken to process the data:

1. **Gathering raw images:** First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.
2. **Running the Viola-Jones face detector**[4] The OpenCV ver-

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Papers using this dataset and the specified training/evaluation protocols are listed under "Methods" section of http://vis-www.cs.umass.edu/lfw/results.html

What (other) tasks could the dataset be used for?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[5]

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from http://vis-www.cs.umass.edu/lfw/index.html#download. The images can be downloaded as a gzipped tar file.

When will the dataset be distributed?

The dataset was released in October, 2007.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted at the University of Massachusetts.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All questions and comments can be sent to Gary Huang: gbhuang@cs.umass.edu.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu. Errata are listed under the "Errata" section of http://vis-www.cs.umass.edu/lfw/index.html

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes to the dataset will be announced through the LFW mailing list.

Source: **Datasheets for Datasets** Gebru et al. Workshop on Fairness, Accountability, and Transparency in Machine Learning, PMLR80, (2018)

Other efforts (not an exhaustive list):

- **Data statements for NLP: Toward mitigating system bias and enabling better science**. Bender et al. Transactions of the Association for Computational Linguistics (2018).
- **The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards**. Holland, et al. arXiv preprint arXiv:1805.03677 (2018).

© EVIDATION HEALTH, INC. 2019

Need: Data documentation frameworks for medical MI applications

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[3]. The

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The following steps were taken to process the data:

1. **Gathering raw images:** First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.
2. **Running the Viola-Jones face detector**[4] The OpenCV ver-

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Papers using this dataset and the specified training/evaluation protocols are listed under "Methods" section of http://vis-www.cs.umass.edu/lfw/results.html

What (other) tasks could the dataset be used for?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[5]

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from http://vis-www.cs.umass.edu/lfw/index.html#download. The images can be downloaded as a gzipped tar file.

When will the dataset be distributed?

The dataset was released in October, 2007.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted at the University of Massachusetts.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All questions and comments can be sent to Gary Huang: gb-huang@cs.umass.edu.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu. Errata are listed under the "Errata" section of http://vis-www.cs.umass.edu/lfw/index.html

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes to the dataset will be announced through the LFW mailing list.

Source: **Datasheets for Datasets** Gebru et al. Workshop on Fairness, Accountability, and Transparency in Machine Learning, PMLR80, (2018)

Other efforts (not an exhaustive list):

- **Data statements for NLP: Toward mitigating system bias and enabling better science**. Bender et al. Transactions of the Association for Computational Linguistics (2018).
- **The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards**. Holland, et al. arXiv preprint arXiv:1805.03677 (2018).

Need to extend data documentation frameworks to medical/health datasets:

- **Integrate RWD framework**: *score* datasets on relevance, reliability, accrual, verification, validation
- **Specifically address conformance, plausibility, and completeness** which have a more crucial role in high-stakes medical/health applications

© EVIDATION HEALTH, INC. 2019

To trust an MI system requires to trust all steps of its workflow

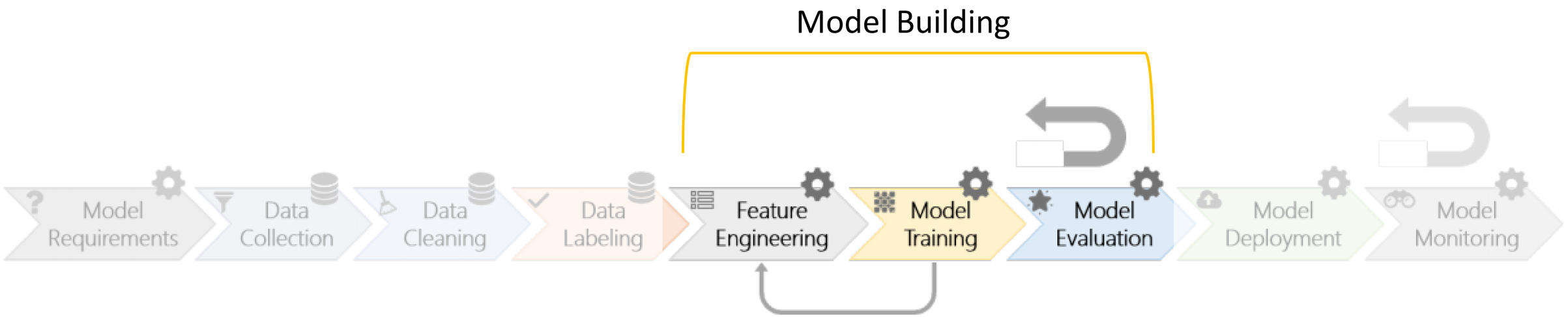
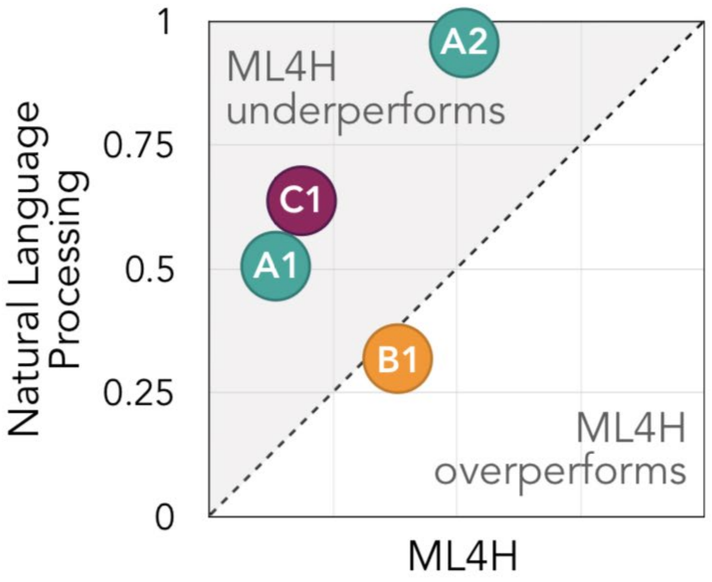


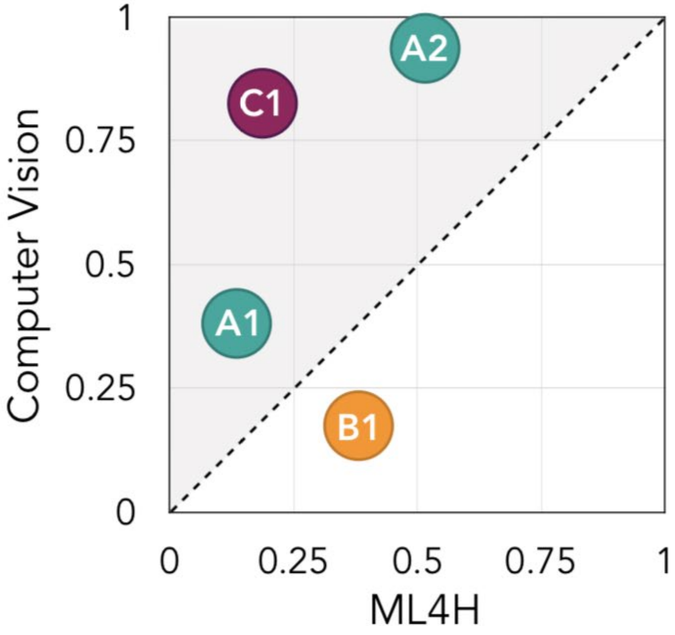
Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

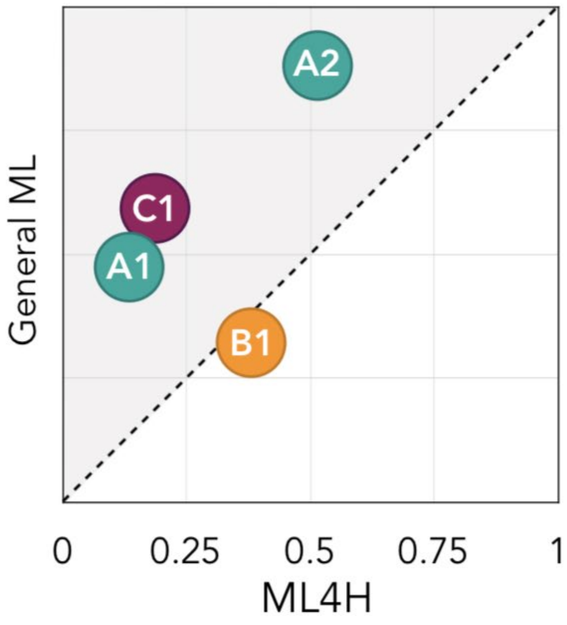
Source: **Software Engineering for Machine Learning: A Case Study** Amershi et al., International Conference on Software Engineering (ICSE) 2019

© EVIDATION HEALTH, INC. 2019

Trusting model building: Reproducibility





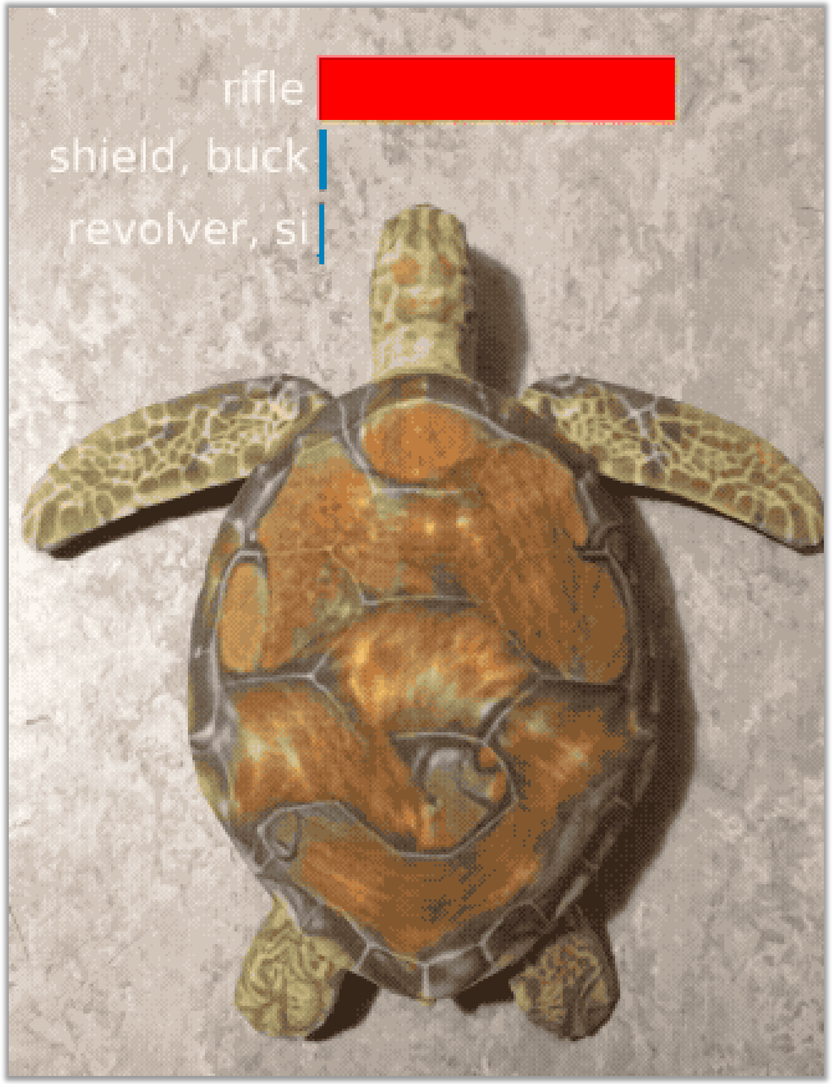


Source: **Reproducibility in Machine Learning for Health** McDermott et al., Workshop in Reproducibility in Machine Learning (ICLR) 2019

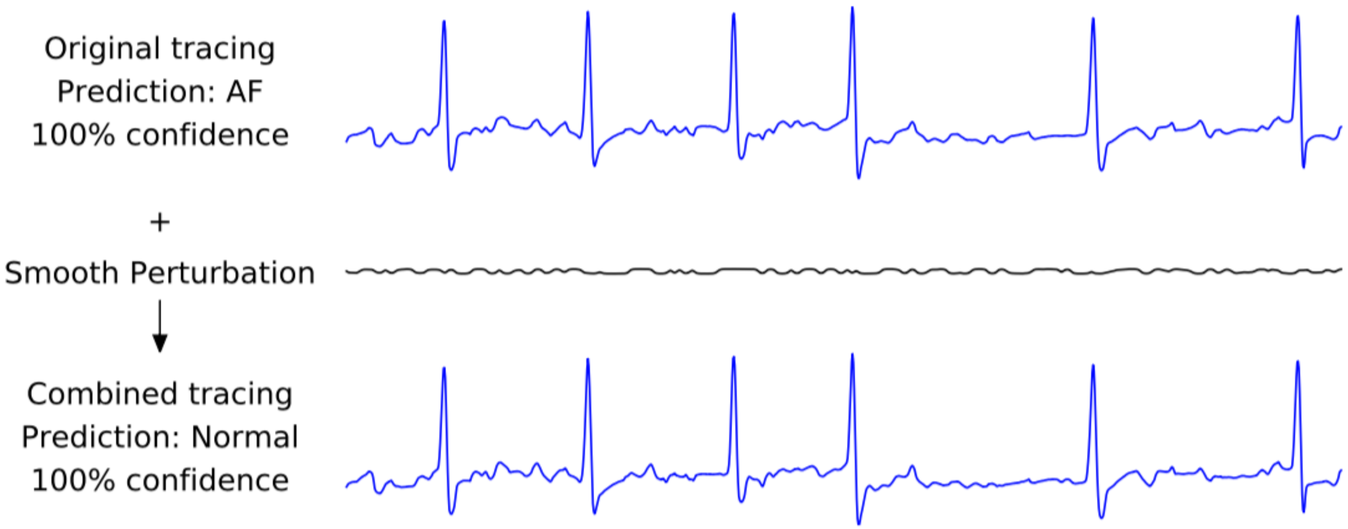
https://arxiv.org/abs/1907.01463

© EVIDATION HEALTH, INC. 2019

Trusting model building: Adversarial examples



Source: **Synthesizing Robust Adversarial Examples** Athalye et al., (ICML) 2018



Source: **Adversarial Examples for Electrocardiograms** Han et al., Workshop on debugging Machine Learning systems (ICLR) 2019 https://arxiv.org/abs/1905.05163

See also: **Adversarial attacks on medical machine learning** Finlayson et al. Science (2019)

© EVIDATION HEALTH, INC. 2019

To trust an MI system requires to trust all steps of its workflow

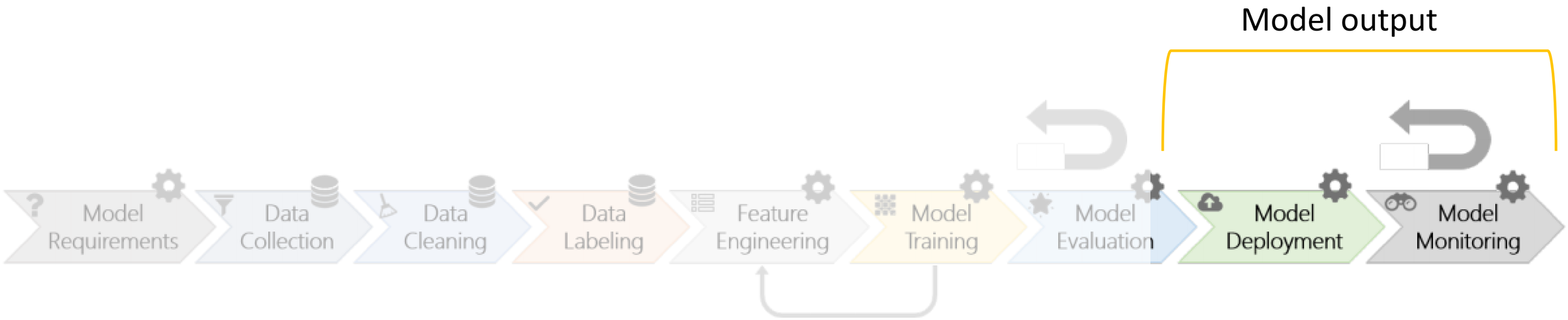


Fig. 1. The nine stages of the machine learning workflow. Some stages are data-oriented (e.g., collection, cleaning, and labeling) and others are model-oriented (e.g., model requirements, feature engineering, training, evaluation, deployment, and monitoring). There are many feedback loops in the workflow. The larger feedback arrows denote that model evaluation and monitoring may loop back to any of the previous stages. The smaller feedback arrow illustrates that model training may loop back to feature engineering (e.g., in representation learning).

Source: **Software Engineering for Machine Learning: A Case Study** Amershi et al., International Conference on Software Engineering (ICSE) 2019

© EVIDATION HEALTH, INC. 2019

Trusting model outputs: test-time monitoring

- **General idea**: An additional system (can be human) evaluate trustworthiness of model output.
- **High level mechanism:** monitoring system determines if test data is ”anomalous” as compared to the data the model has been trained on

- **Active area of research in ML community:**

- **A neural network architecture that computes its own reliability.** Leonard et al. Computers & chemical engineering, 1992.
- **To trust or not to trust a classifier**. Jiang et al. In Advances in Neural Information Processing Systems 2018
- **Can You Trust This Prediction? Auditing Pointwise Reliability After Learning**. Schulam et al. Artificial Intelligence and Statistics (AISTATS), 2019.
- **Deep Weighted Averaging Classifiers**. Card et al. Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 2019.
- **Scalable joint models for reliable uncertainty-aware event prediction**. Soleimani et al. IEEE transactions on pattern analysis and machine intelligence. 2018
- **Learning with rejection**. Cortes et al. In International Conference on Algorithmic Learning Theory 2016.

- **Can be used to detect distribution shifts post-deployment**

For an in depth-discussion: Saria, S. and Subbaswamy, A. (2019). **Tutorial: Safe and Reliable Machine Learning**. ACM Conference on Fairness, Accountability, and Transparency

© EVIDATION HEALTH, INC. 2019

Machine Intelligence systems are Complex Systems. System Engineering principles can be applied to ensure end-to-end reliability

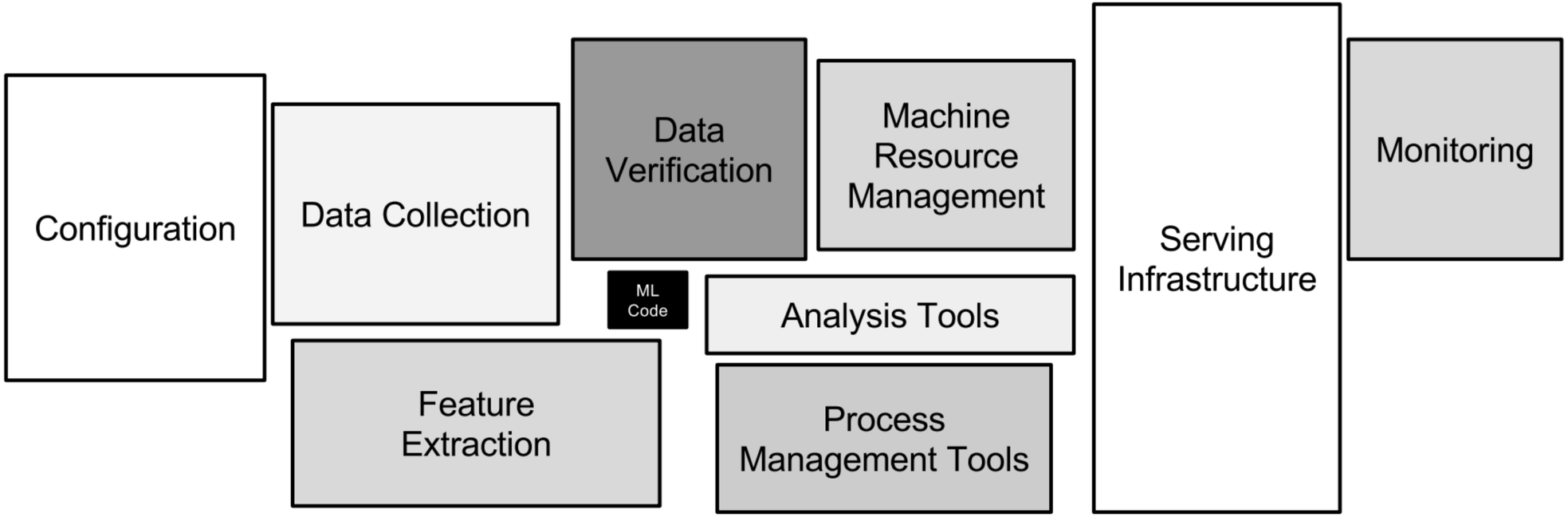


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: **Hidden technical debt in machine learning systems.** Sculley et al., Advances in neural information processing systems (NeurIPS) 2015

For an in depth-discussion: Saria, S. and Subbaswamy, A. (2019). **Tutorial: Safe and Reliable Machine Learning**. ACM Conference on Fairness, Accountability, and Transparency

© EVIDATION HEALTH, INC. 2019

evidation

Thank you

Luca Foschini, PhD

Co-founder & Chief Data Scientist

luca@evidation.com | @calimagna