Software Engineering Process and Practices for Data Science

Junhua Ding, PhD

Department of Information Science

University of North Texas

Software Crisis

- **The difficulty to deliver useful and efficient software in the required time with planned budget.**
- **Coined at the first NATO Software Engineering Conference in 1968 at Garmisch, Germany.**

- "The major cause of the software crisis is that the machines have become several orders of magnitude more powerful! To put it quite bluntly: as long as there were no machines, programming was no problem at all; when we had a few weak computers, programming became a mild problem, and now we have gigantic computers, programming has become an equally gigantic problem".

— Edsger Dijkstra, The Humble Programmer (EWD340), Communications of the ACM, 72 Turing Award Lecture

What is Software

- A collection of computer instructions and data that tell the computer how to work.
- Software = Algorithms + Data
- Powerful Computer → Powerful Software → Complex Algorithms + Big Data

- Software Engineering has been widely and successfully used for building Algorithms (Functions), but Not for Big Data.

Questions

- **How should we build data intensive software?**

- **How can we integrate software engineering into data science for building data intensive software?**

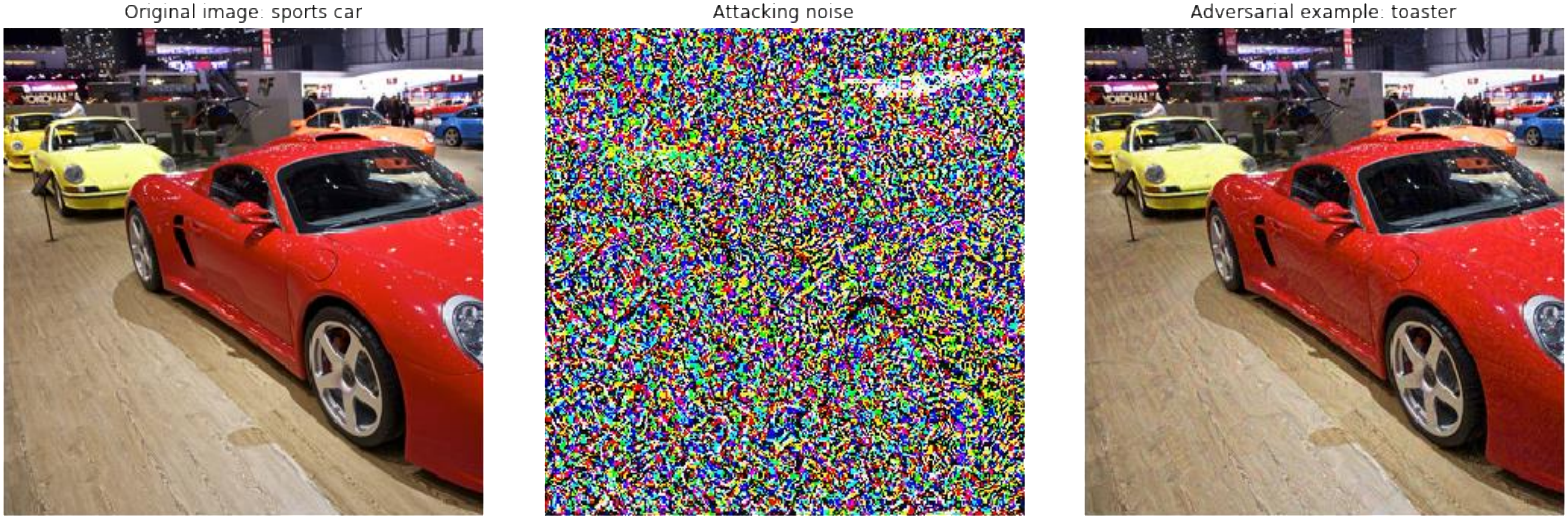
Examples (Why do we need Software Engineering?)

Rajpurkar and et al. introduced a deep learning system (ChexNet) for diagnosing pneumonia diseases based on chest X-ray images. They claimed ``We find that CheXNet exceeds average radiologist performance on pneumonia detection on both sensitivity and specificity''. (ref: https://arxiv.org/abs/1711.05225)

But Oakden-Raynera, a radiologist student and machine learning researcher, questioned the dataset used by ChexNet. He said: ``I believe the ChestXray14 dataset, as it exists now, is not fit for training medical AI systems to do diagnostic work. (1). how accurate are the labels, (2). what do the labels actually mean, medically, and (3). how useful are the labels for image analysis''.

(ref: https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/)

What data should we need, and how to evaluate the them?



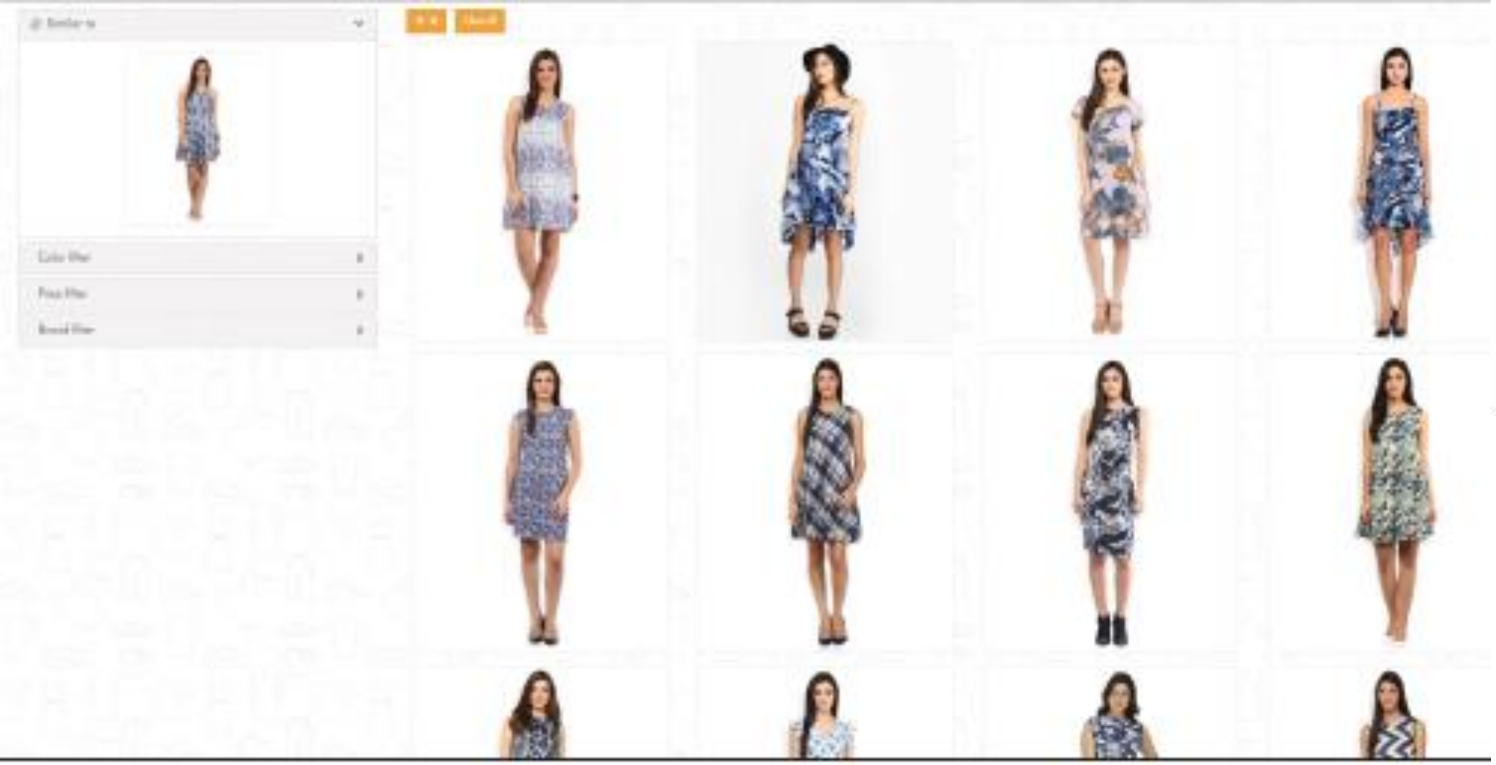
A small number of bad samples added to the training data would diminish learning robustness. Bad samples can be easily generated using GAN.

Ref: Jesus Rodriguez, ``Using Adversarial Attacks to Make Your Deep Learning Model Look Stupid", https://medium.com/@jrodthoughts/using-adversarial-attacks-to-make-your-deep-learning-model-look-stupid-24fb872f06fd

- A. Eklund, T. E. Nichols, and H. Knutsson, ``**Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates**'', PNAS, vol. 112(28), pp. 7900-7905, July 12, 2016.

- ``These results **question the validity of a number of fMRI studies** and may have a large impact on the interpretation of weakly significant neuroimaging results.''

- ``Despite the popularity of fMRI as a tool for studying brain function, the statistical methods used have **rarely been validated using real data**. Validations have instead mainly been performed using simulated data, but it is obviously very hard to simulate the complex spatiotemporal noise that arises from a living human subject in an MR scanner.''



Product search using catalog image as query. The system return similar looking image but the similarity was not very high.

One More

Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, by I. Deborah Raji, and J. Buolamwini, AAAI 2019,

- To analyze gender and skin type performance disparities in commercial facial analysis models.

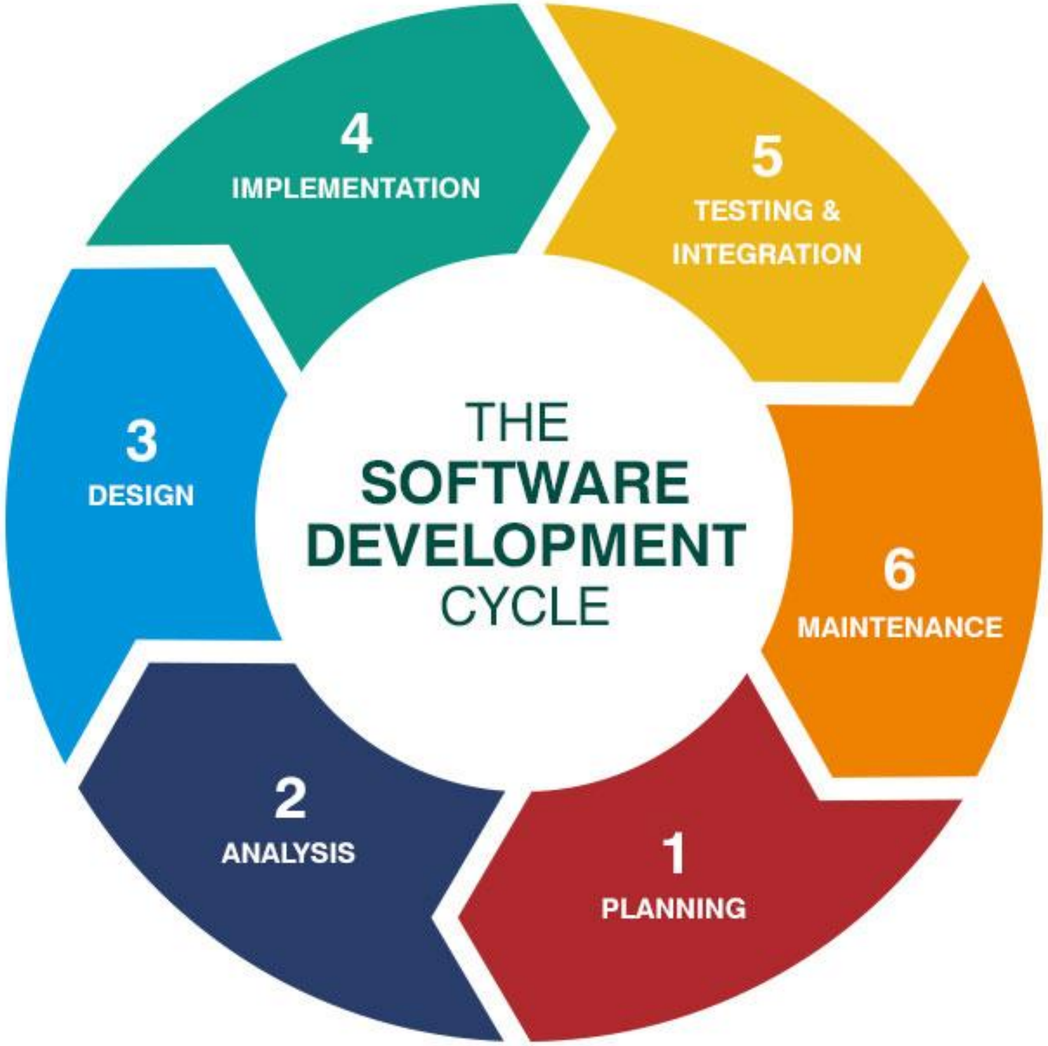
Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)

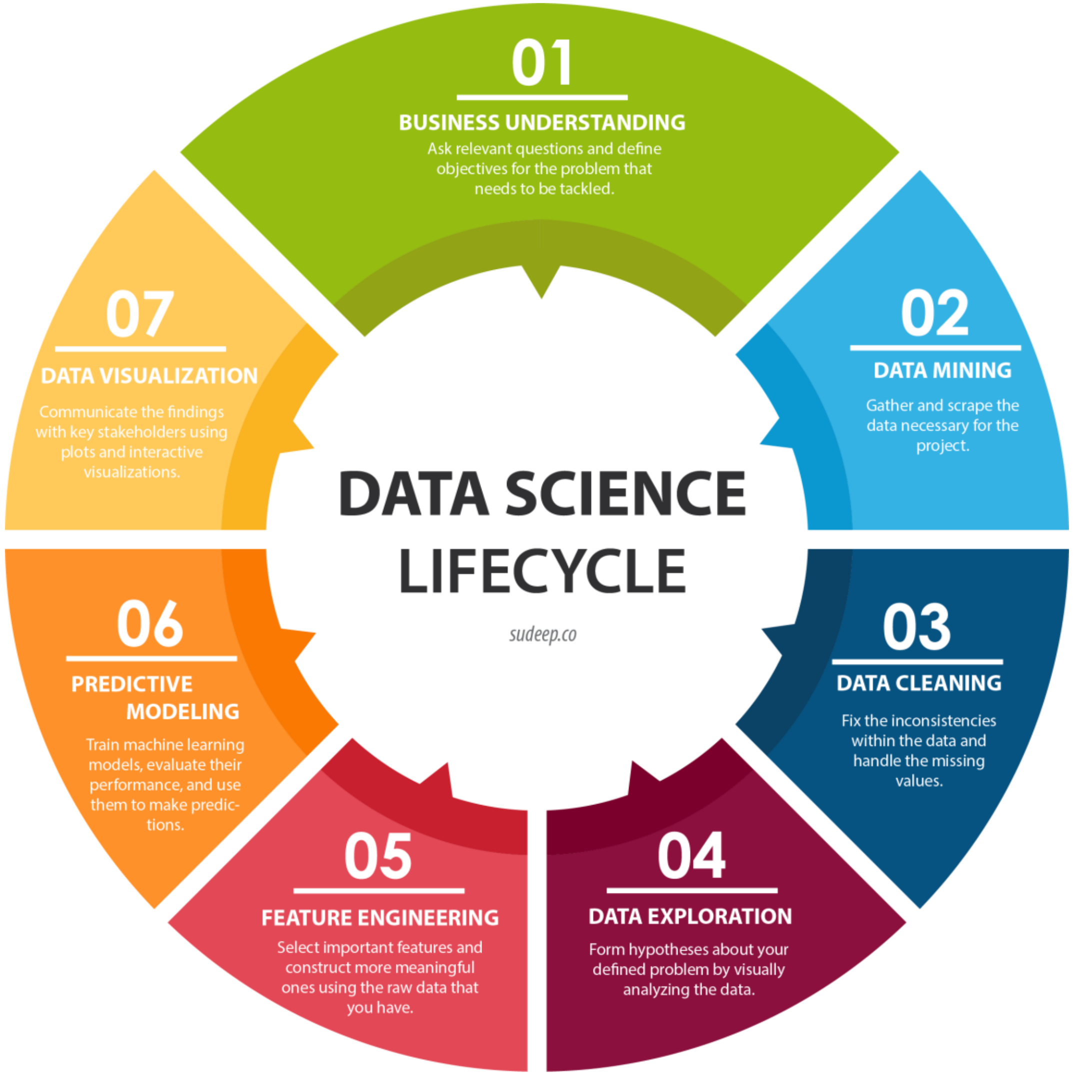
Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Face ++	-8.3	-18.7	0.2	-13.9	-3.9	-30.4	0.6	-8.5	-0.3
MSFT	-5.72	-9.70	-2.45	-12.01	-0.45	-19.28	-5.67	-1.06	0.00
IBM	-7.69	-10.74	-5.17	-14.24	-1.93	-17.73	-11.37	-4.43	-0.04

Table 1: Overall Error on Pilot Parliaments Benchmark, August 2018 (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Target Corporations									
Face ++	1.6	2.5	0.9	2.6	0.7	4.1	1.3	1.0	0.5
MSFT	0.48	0.90	0.15	0.89	0.15	1.52	0.33	0.34	0.00
IBM	4.41	9.36	0.43	8.16	1.17	16.97	0.63	2.37	0.26
Non-Target Corporations									
Amazon	8.66	18.73	0.57	15.11	3.08	31.37	1.26	7.12	0.00
Kairos	6.60	14.10	0.60	11.10	2.80	22.50	1.30	6.40	0.00

Software Life Cycle vs. Data Science Life Cycle

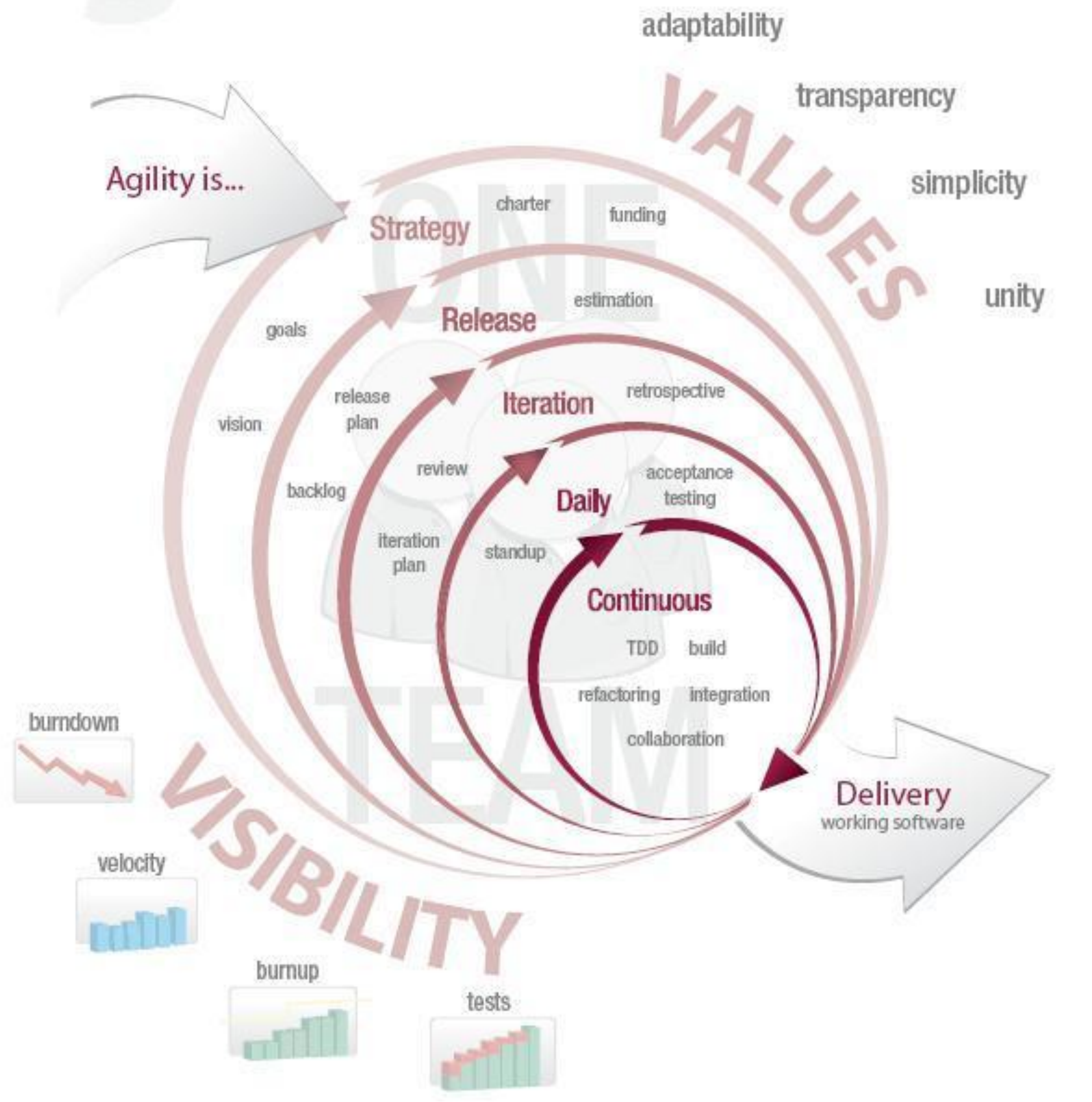




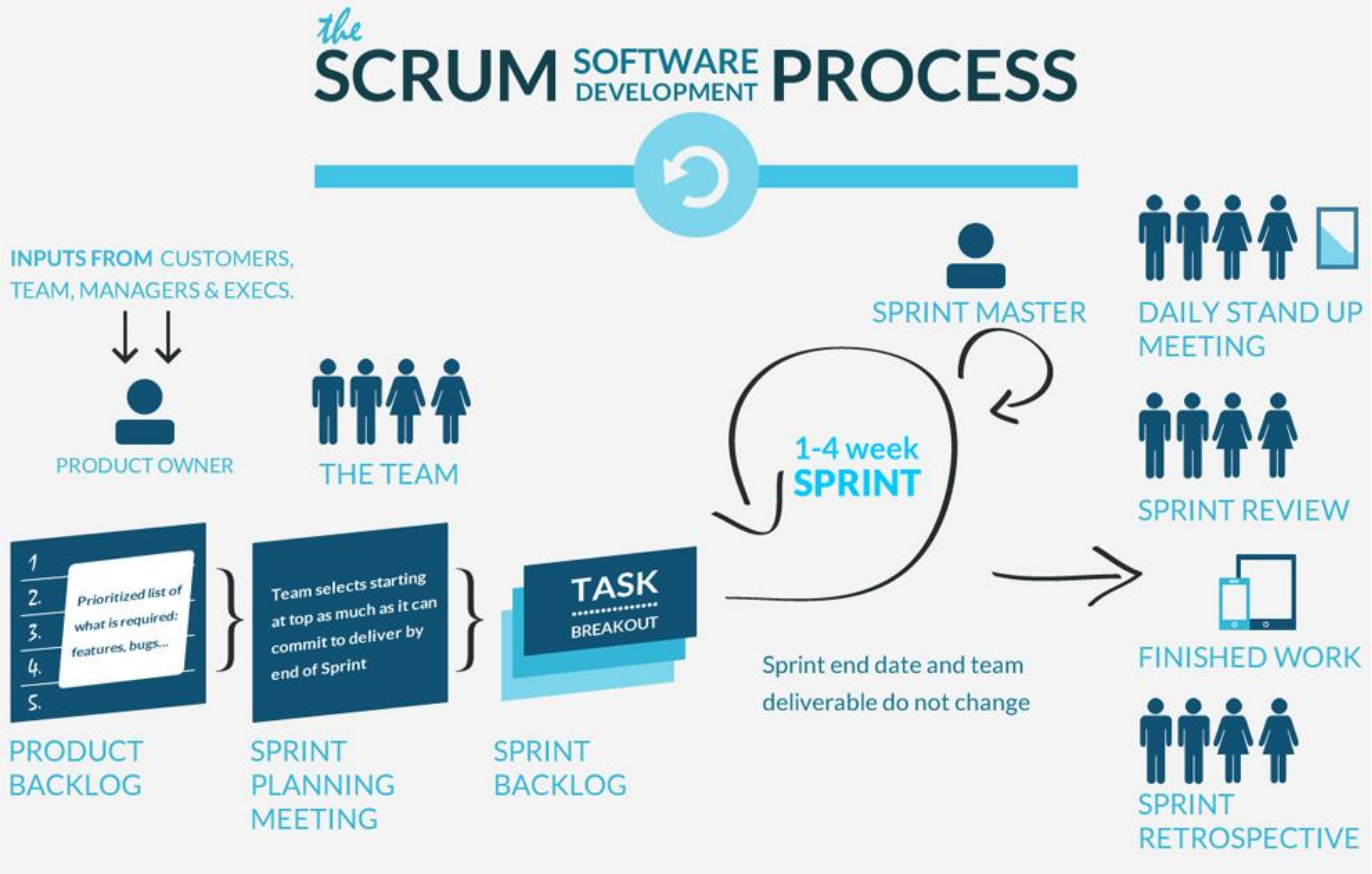
https://medium.com/@jilvanpinheiro/software-development-life-cycle-sdlc-phases-40d46afbe384, by Jilvan Pinheiro

http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/, by Sudeep Agarwal

Agile Development

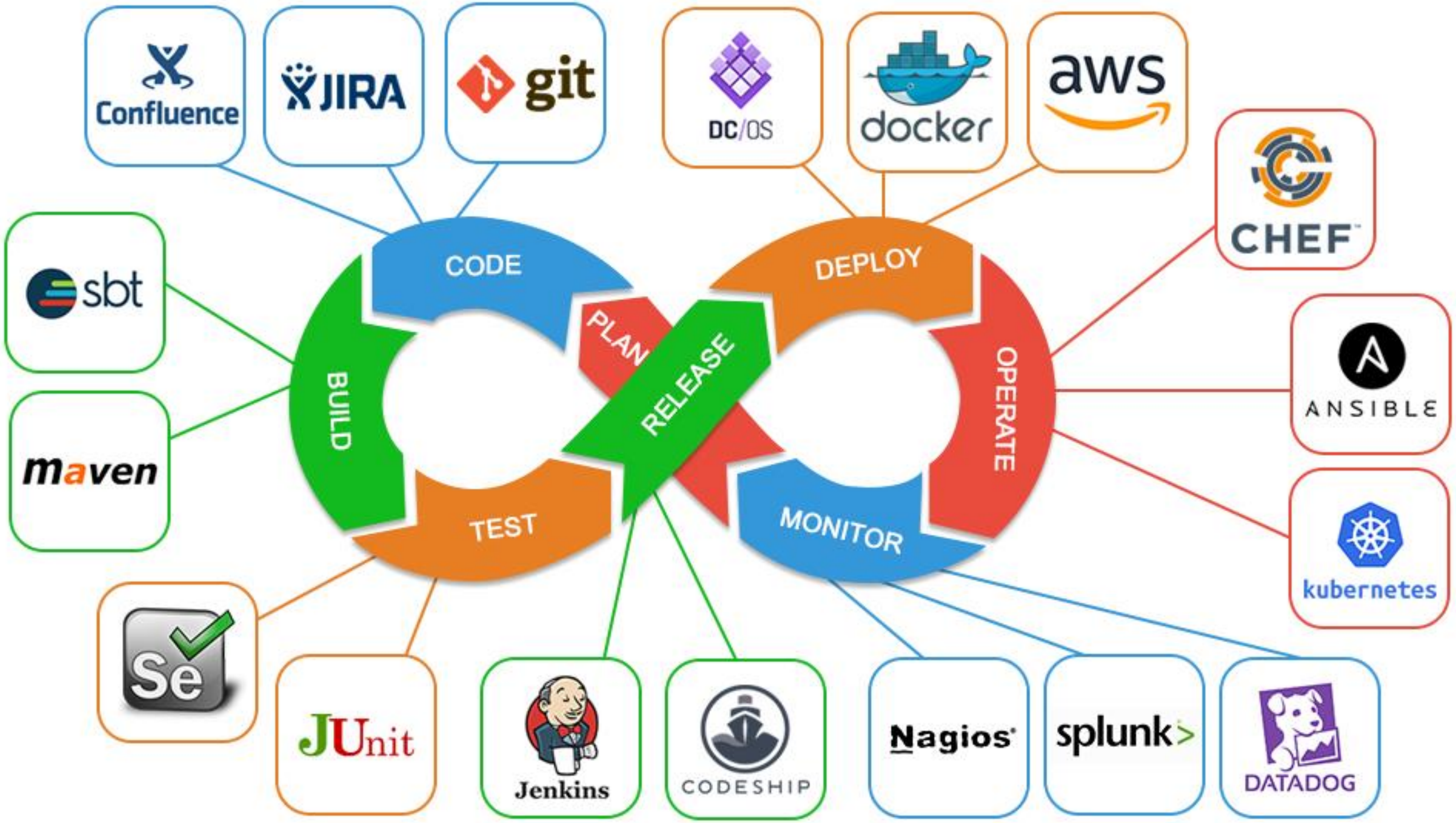


https://blogs.worldbank.org/voices/agile-global-development-using-technology-fight-extreme-poverty, by Chirs Vein



https://www.maxxor.com/software-development-process

DevOps



https://medium.com/devopslinks/devops-without-devops-tools-3f1deb451b1c, by Jagatveer Singh

Tools

- Git, Github, JIRA, Stack
- PSPP/SPSS, Tabular, SAS, etc.
- Apach WeKa
- Google Tensorflow
- Facebook PyTorch
- MongoDB
- Jupyter Notebook, Framework Pandas, TF Learn
- ……

Datasets v.s. Program Libraries

- Kaggle
- ImageNet
- NIST
- Government agencies

Integrate Software Engineering Process and Practices into Data Science Project Development

Evaluate it before use it.

- Fit for purpose: Fidelity, Variety, Veracity?
- Intrinsic: Completeness, Correctness, ...?