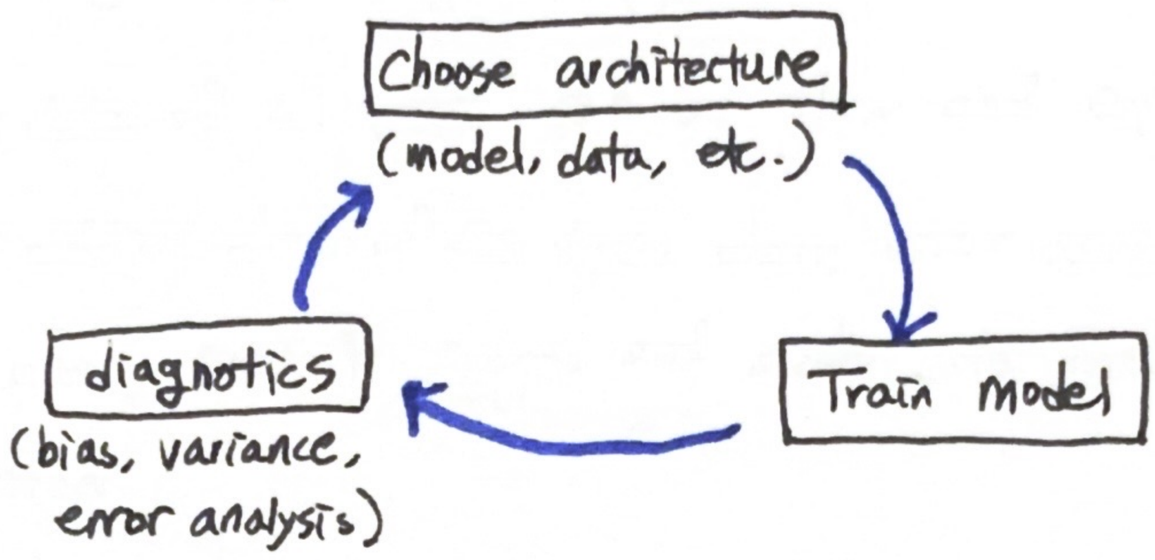
<Machine Learning development process – Iterative loop of ML development>

* Iterative loop of ML development



ex) Spam mail classifier

Supervised learning : $\vec{x}$ = features of email
y = spam(1) or not spam(0)

list the top 10,000 words to compute $x_1, x_2 \cdots x_{10,000}$
then $\vec{x}$ vector is : 추출한 10,000개 단어가 mail에 포함되어있는지 여부를 나타내는 vector

⇓ after initializing a learning algorithm
how to try to reduce spam classifier algorithm's error?

① Collect more data
② Develop sophisticated features based on email routing (from email header)
③ Define sophisticated features from email body (e.g. should "discounting" and "discount" be treated as same word)
④ Design algorithms to detect misspellings (e.g. watches, med1cine, m0rtgage)

<Error Analysis>

- Cross validation set 중 misclassified된 데이터 example들을 수동으로 분석하여 어떤 type의 데이터가 어떤 경향의 error를 발생시키는지 관찰

ex) m_{cv} = 500 examples in CV set.
Algorithm misclassified 100 of them
⇒ Manually examine 100 examples and categorize them based on common traits

ex) - Pharma : 21
- Deliberate misspelling : 3
- Unusual email routing : 17
- Steal password (phishing) : 18
- Spam message in embedded image : 5

→ 통계적으로 misclassfied된 example의 category에 대한 성능을 높이는 방법으로 next step을 결정

- Pharma spam email (collect more data about pharma
come up with new features
(e.g. specific name of drug, pharma product)
- Phishing email (collect more data about phising email
come up with new features
(e.g. specific url in email...)