What did we do last time?

What is pandas?

What are some of the things we can do with pandas?

The autocoder we built last class is called a rule based autocoder

* Last class we used pandas to build a rule based autocoder
* We had this function, foot or not
  + You put some narrative in
  + We gave the computer some logic to determine whether it was a foot injury or not
* For a long time this was how all AI was done, rule based systems
* In BLS this is still how almost all automation is done
  + You have a bunch of people writing programs that tell the computer what to do
* Up in about the early 80’s, things started to change
  + And an approach that people had known for a long time, suddenly started becoming useful
  + That approach is called machine learning
  + The motivation is that people wanted to build computers that learned like humans
  + Humans do not learn by someone opening up their head and writing a little program in there
  + They seem to learn through their experiences and their observations
  + Starting in the mid 80’s, we started being able to create programs that could sort of do this
  + A number of things contributed to this
    - there were important theoretical developments
    - there were new techniques
    - but perhaps most importantly, people started having access to powerful computers that could do very interesting and elaborate computations
  + machine learning has only exploded since then, with invention of internet
  + exponential growth in computer processing and storage
  + and now there are billion dollar companies, basically all they do is machine learning
  + Google is probably best example of this
  + Their whole strategy is if you have enough data, you can teach the computer to all sorts of things
    - Search through webpages
    - Translate between languages
    - Block email spam
    - Recognize spoken words and answer questions
    - Recognize images
    - Teach cars how to drive
    - And the list is just growing, it seems like almost every day
* The rule based approach didn’t go away
* Still used for almost all automation at BLS for example, works great for simple well defined tasks
* Does not work well for poorly defined tasks like classifying written narratives
* That does not stop people from trying to sell rule based crap to you
  + When I first started doing autocoding here at BLS, we had contract with SAS, they wanted to sell us this super expensive rule based system for automatically coding
  + $40,000 dollar rule based system
  + And so I said ok, let’s try this system, but let’s compare it to simple machine learning system I built, and see how they did
  + So they took this team of like 3 engineers, went off and built this elaborate rule based system using their expensive software
  + Worked on it for 3 or 4 months
  + Brought it back and compared it
  + No contest
  + Their elaborate expensive system could not come anywhere close to the simple machine learning system I built in a few hours
  + Today I’m going to show you how to build one of theses machine learning systems
  + In fact we’re going to build several them
* There are many different kinds of machine learning
  + Key to all of them is data
  + Data is what the machine is going to use to learn
  + For our coding tasks – we typically have lots of data that’s been previously coded
  + This is what we’re going to use for our autocoder today
* Earlier I downloaded some MSHA data
* You saw this last time, has these narratives and has these codes associated with them
* This is what we’re going to use to build our autocoder
* Our first step is to divide this into 3 datasets