AnacondaCon18

Anacondacon er en konferanse for data analyse og “data science” laget av selskapet Anaconda. Selskapet driver salg av plattformen Anaconda til bedrifter som enten driver analyser eller har en analyse-seksjon internt. Sentralt i Anaconda-pakken – og vårt hovedfokus for konferansen – er Python. Python er et programmeringsspråk med en lang open source-historie samt et stort globalt nettverk som driver utvikling av open source-bibliotek til Python. Dette er også mye av grunnen til at Python er foretrukket som analyseverktøy. Som en konsekvens er Anaconda en stor bidragsyter til mange open source-pakker for å vedlikeholde kvaliteten på mange av bibliotekene til Python.

Konferansen for 2018 gikk av stabelen i Austin, Texas 8-11 april. Søndag 8 april bestod av opplæring i bruk av noen av de mest populære bibliotekene innenfor data science: TensorFlow, Scikit-Learn, Pandas osv. Vi deltok ikke på opplæringsdagen da vi allerede har noe kjennskap og erfaring med å bruke disse pakkene. Programmet de tre etterfølgende dagene besto av forelesninger á 50 minutter som gikk i tre parallelle løp, fordelt etter løst definerte tematiske stier; «Anaconda», «Real World» og «Open Source». Vi var innom alle de forskjellige stiene, som hadde hver sin foredragssal på konferansehotellet. Utenom disse foredragsrommene fantes det et mingleområde hvor en kunne finne demoer og “stands”.

«Anaconda»-stien handlet om innholdet i Anaconda-pakken og om tilleggs bibliotek som enkelt kan installeres for så å fungere sømløst med det en har inneholdt i et “conda environment”. Tittelen på to av foredragene vi gikk på i denne stien var «Scalable Machine Learning with Dask» og «PyViz: Dashboards for Visualizing 1 Billion Data Point in 20 Lines of Python in Your Browser on Your Laptop».

Forelesningene i Real World-stien ble holdt av gjester fra firmaer som benytter seg av Anacondas løsninger på en rekke forskjellige problemstillinger. Eksempler på forelesninger vi deltok på er «Accelerating Scientific Workloads with Numba» og «Causal Inference in Tech».

Open-Source-stien handlet, som navnet tilsier, om open-source-prosjekter. Foreleserne kunne gi gode eksempler på hvordan gratis programvare kan løse vanskelige problemer. Vi likte spesielt godt forelesningene «DeepFashion: Building a REST API to Detect Clothing Styles» og «Accelerating Deep Learning with GPUs».

I tillegg til det faglige materialet var lunchen på konferansehotellet JW Marriot fullstendig utsøkt, og Anaconda inviterte til en sosial samling tirsdag kveld – «AnacondaCon Carne». Det var en god anledning til å bli kjent med de andre konferansedeltakerne. Vi – som ganske unge – syntes det var interessant å høre om alt det ulike arbeidet innen databehandling som foregår i Amerikansk næringsliv. Dette var en slags forsikring om at våre evner også er anvendbare utenfor akademiske kretser.

Vårt fokus

Vi er studenter ved masterprogrammet i Computational Physics ved Fysisk Institutt ved Universitetet i Oslo. I teoretisk fysikk støter en nokså raskt på problemer som er vanskelige eller umulige å løse analytisk (med penn og papir) og en må ofte ty til numeriske metoder som et supplement eller en erstatning. Med numeriske metoder mener vi stort sett, dog ikke begrenset til, approksimative beregninger gjort på en datamaskin.

Ofte må en gjøre implementasjonen av slike numeriske algoritmer i et språk som ligger nærme nok maskinkode for å oppnå tilstrekkelig hastighet på beregningene. Typisk har språk som C/C++ og Fortran vært foretrukket. I senere tid har Python erstattet disse språkene i mange anvendelsesområder. Dette kommer i stor grad av at Python er enklere og lettere å skrive enn de gamle bautaene som C/C++ og Fortran. Dette senker utviklingstiden betydelig! De vitenskapelige tungregnebibliotekene som finnes til Python er også fremragende. Dette gjør det lettere for oss å fokusere på vitenskapen i større grad enn programmeringen. Målet for å dra på denne konferansen er derfor i stor grad å få svar på spørsmålet (drømmen): Kan Python erstatte C++ fullstendig?

Faglig utbytte

Det følger av avsnittet over at AnacondaCons fokus for oss ikke var vårt fag direkte, men heller metodene vi benytter i faget – teoretisk beregning og datanalyse. Vi gjorde definitivt noen nyttige oppdagelser i form av biblioteker til programmeringsspråket Python.

Dask (dask.pydata.org)

Dask er et bibliotek til Python som tilbyr nærmest automatisk parallellisering av ens programkode. Dette er uhyre viktig for oss som fysikere. Det hender ofte at vi må gjøre tunge beregninger på en regneklynge og alt som gjør parallellisering av et program enklere ønsker vi velkommen med åpne armer. Dask er laget for å passe godt sammen med Numerical Python (Numpy) og Pandas, som vi bruker svært ofte. Dask muliggjør også behandling av datastrukturer som er større enn datamaskinens minnekapasitet (RAM) på en smart måte. Pakken kommer med et innebygd visualiseringsverktøy som gir en god oversikt over hvordan det parallelliserte programmet arbeider.

Mellom de to foredragene om Dask som vi deltok på fikk vi god tid til å prate med Tom Augspurger (https://github.com/TomAugspurger). Her fikk vi mulighet til å diskutere styrker og svakheter ved Dask samt få en demonstrasjon av hvordan Dask lar seg visualisere og hvordan det virker mot en regneklynge.

Numba (numba.pydata.org)

Numba er et bibliotek som kompilerer valgte Python-funksjoner til maskinkode ved å bruke «Just-in-time compilation» (JIT). Dette kan potensielt øke beregningshastigheten voldsomt mye. Biblioteket muliggjør en utførelse av beregninger på hastighetsnivå som er sammenlignbart med C/C++ og Fortran uten at en trenger å bytte programmeringsspråk. Numba er bygget for å fungere både på GPU'er og CPU'er og er integrert med mange av de vitenskapelige bibliotekene som er tilgjengelige for bruk med Python.

En av hovedutviklerne – Siu Kwan Lam (https://github.com/sklam) – var veldig interessert i å diskutere våre problemstillinger og hvordan Numba kunne være til hjelp. Han var godt kjent med flere av problemstillingene vi møter innen mange-partikkel kvantemekanikk og hadde flere ideer om hva som kunne være interessant å prøve. Spesielt ved hjelp av Numba.

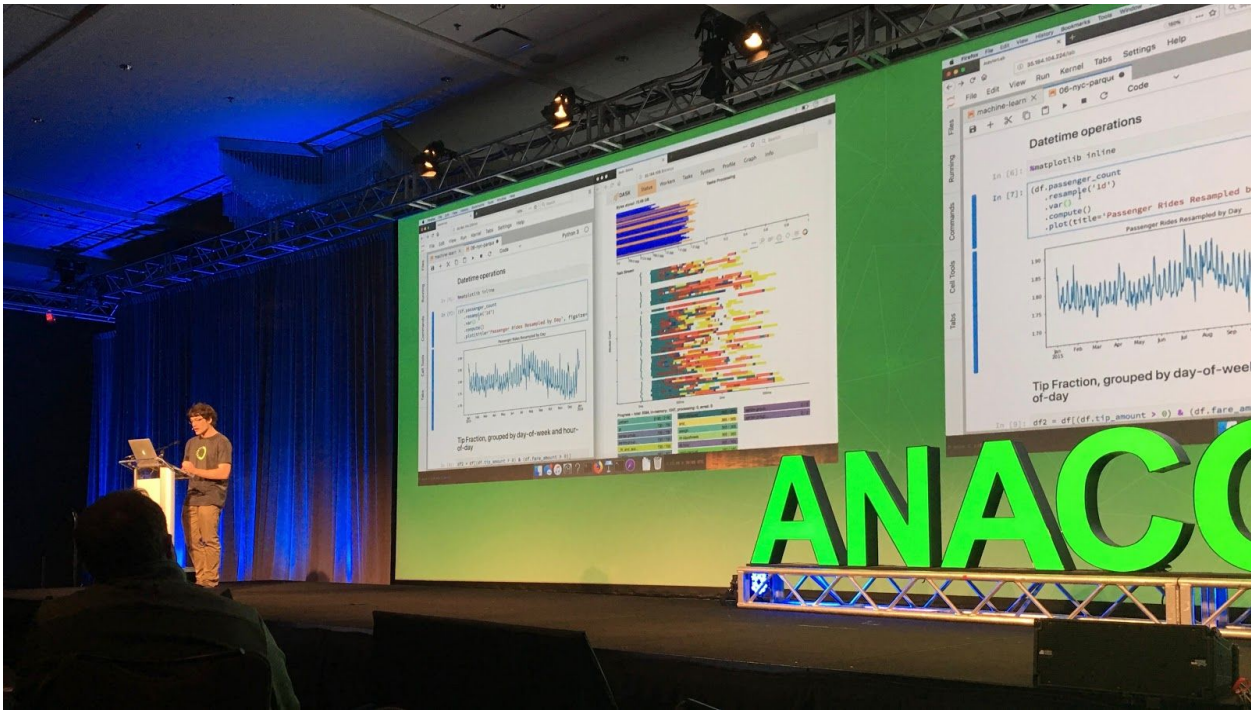
PyViz (pyviz.org)

PyViz er ikke ett enkelt bibliotek, men en samling av pakker for datavisualisering som fungerer godt sammen. PyViz-pakken egner seg godt til å lage interaktive plott og figurer, som kan oppdateres automatisk ettersom ny data er tilgjengelig. Dette er svært god funksjonalitet å ha i problemer knyttet til datainnsamling og kontroll. Som alle andre bibliotek fungerer alt innholdet i PyViz sømløst med Pythons eksisterende vitenskapelige funksjonalitet.

Konklusjon

De som først og fremst vil ha et utbytte av å dra på en konferanse som AnacondaCon er de som driver med en eller annen form for dataanalyse. Som vordende fysikere var dette godt egnet for oss, da mange av de metodene er anvendbare innenfor vårt fagfelt. Vårt mål for å dra på konferansen var først for å lære noe nytt og for å oppdage nye verktøy. Dette målet ble definitivt oppfylt. I tillegg mener vi at foredragene holdt et svært høyt nivå, både når det gjelder det faglige og det pedagogiske.

Mye av det vi lærte på AnacondaCon har vi allerede tatt i bruk i prosjekter og arbeid mot masteroppgaven. Vi skal begge jobbe som gruppelærerer til høsten og kommer til å ha stor glede av kunnskapen vi har fått frå AnacondaCon.



Presentasjon av Dask v/ Tom Augspurger

Mingleområde. Her får vi en grundigere gjennomgang av DASK.