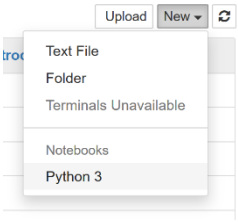
CMS Open Data: Päällekkäisten histogrammien piirtäminen

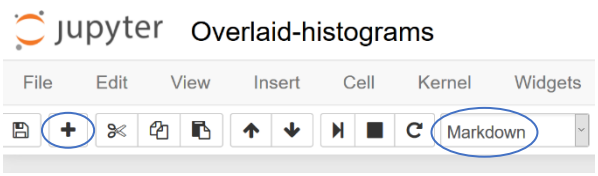
Tässä harjoituksessa analysoidaan CERNin julkaisemaa avointa dataa Jupyter Notebook -sovelluksen avulla. Sovellus voidaan ladata ilmaiseksi asentamalla Anaconda tietokoneeseen osoitteesta https://www.continuum.io/downloads.

Valitse analysoitava datatiedosto osoitteesta http://opendata.cern.ch/record/545. Tässä esimerkissä käytetään tiedostoa *Dimuon_DoubleMu.cvs*, mutta myös muita tiedostoja voidaan käyttää, kunhan niiden tiedostotyyppi on *comma separated values (.csv)*. Lataa valitsemasi tiedosto ja paina mieleen tiedoston tarkka sijainti sekä nimi. Mikäli tiedostonimi sisältää välilyöntejä, korvaa ne alaviivoilla.

Avaa Anaconda Navigator ja käynnistä Jupyter painamalla “Launch”. Avaa tyhjä notebook valitsemalla oikean yläkulman valikosta ”New” ja edelleen ”Pyhton 3”.



Tekstisolut, jotka sisältävät ohjeita tai selityksiä mutta eivät ohjelmointikoodia, voidaan luoda painamalla yläpalkista ”+”-painiketta ja valitsemalla solutyypiksi ”Markdown”. Kirjoittaaksesi koodia valitse solutyypiksi ”Code”. Soluihin voidaan kirjoittaa, kopioida, liittää ja muokata sisältöä eri lähteistä.



Tuo ensimmäiseksi paketit *pandas* ja *matplotlib.pyplot*, jotta voit lukea tiedostoja ja piirtää kuvaajia.

In [1]:
```
import pandas
import matplotlib.pyplot as plt
%matplotlib inline
```

Luotuasi koodisolun paina *Ctrl+Enter* ajaaksesi koodin. Hakasulkujen *In[ ]* sisään ilmestyvä tähti on merkki siitä, että koodia suoritetaan. Etene vasta sitten, kun tähden tilalle on tullut numero eli kun koodi on suoritettu. Numero kertoo ajojen lukumäärän. Mahdolliset virheilmoitukset tulostuvat punaisella ja antavat tietoa virheestä.

Voidaksesi käyttää lataamaasi datatiedostoa, sen tiedot täytyy tallentaa muuttujaan. Kirjoita tiedostosijainti ja -nimi siten, kuin se esiintyy tietokoneellasi. Tässä esimerkissä tiedosto (*Dimuon_DoubleMu.csv*) sijaitsee yhtä hakemistoa ylempänä (*../*) olevassa kansiossa nimeltään *Data* (*Data/*). Tiedoston tuomiseen tarvittava polku on siten *../Data/Dimuon_DoubleMu.csv.*

Tallenna data muuttujaan *datasetti* ja tarkista ensimmäisten viiden rivin sisältö.

In [2]:
```
datasetti = pandas.read_csv('../Data/Dimuon_DoubleMu.csv')

datasetti.head()
```

Mikäli komennot on kirjoitettu oikein, osa taulukosta pitäisi ilmestyä näkyviin ajettuasi koodisolun.

Tässä harjoituksessa olemme kiinnostuneita hiukkasten invariantista massasta. Tallenna datasetin invarianttia massaa vastaava sarake muuttujaan *invariantti_massa* viittaamalla sarakkeen otsakkeeseen siten, kuin se esiintyy taulukossa (esim. *M*). Jos datasetistäsi puuttuu invariantin massan sarake, laske arvot ensin itse. Piirrä histogrammi kertomalla ohjelmalle piirrettävä muuttuja, pylväiden (bins) lukumäärä sekä haluttu piirtoväli (range). Alla olevassa esimerkissä piirretään invariantin massan histogrammi välillä 0-200 GeV ja 50 pylväällä.

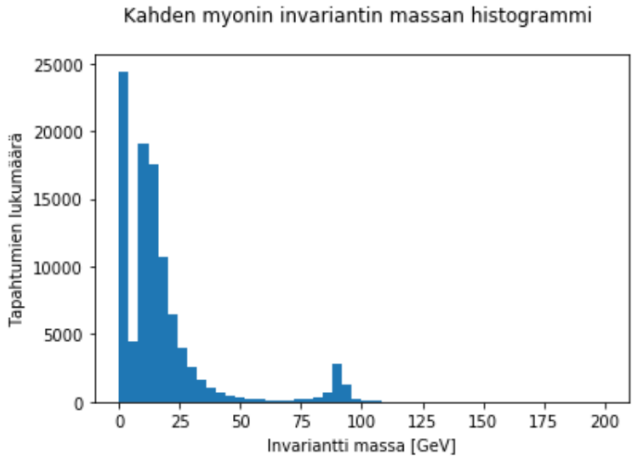
In [3]:
```
invariantti_massa = datasetti['M']

plt.hist(invariantti_massa, bins=50, range=(0,200))
plt.show()
```

Piirtoväliä ja pylväiden lukumäärää voidaan muuttaa tarkemman analyysin tekemiseksi. Piirretään seuraavaksi sama kuvaaja uudestaan siten, että akselit on nimetty ja kuvaajalla on otsikko.

In [4]:
```
plt.xlabel('Invariantti massa [GeV]')
plt.ylabel('Tapahtumien lukumäärä')
plt.title('Kahden myonin invariantin massan histogrammi \n')

plt.hist(invariantti_massa, bins=50, range=(0,200))
plt.show()
```



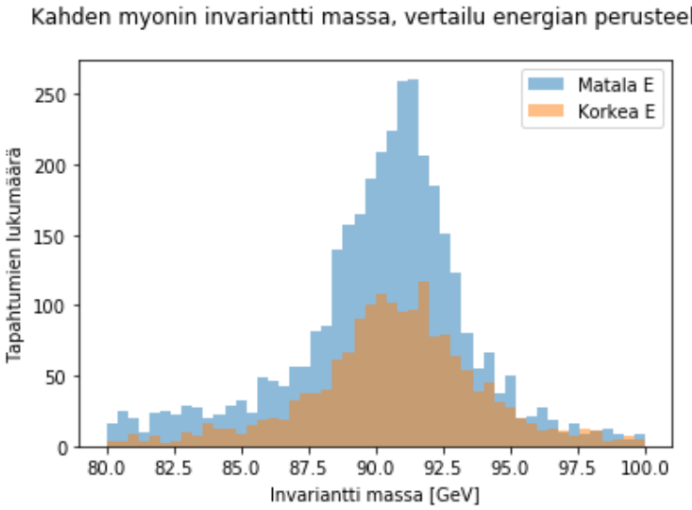
Yllä olevasta kuvaajasta voidaan erottaa piikki noin 90 GeV:n kohdalla. Likimäärin 50 GeV:a matalammat invariantin massan arvot havainnollistavat taustatapahtumia.

Dataa voidaan käsitellä matemaattisilla operaattoreilla, kuten yhteen- ja vähennyslaskulla. Datan lajittelu onnistuu luomalla uusi muuttuja, johon talletetaan vain tietyn ehdon täyttävät arvot. Seuraavassa esimerkissä alkuperäisen datan tapahtumat jaetaan kahteen uuteen datasettiin hiukkasten energian perusteella. Datasetit nimetään ja lajitellaan korkeaan energiaan (> 150 GeV) ja matalaan energiaan (< 150 GeV).

In [5]:
```
uusiSettiKorkeaE = datasetti[datasetti.E1+datasetti.E2>150]
uusiSettiMatalaE = datasetti[datasetti.E1+datasetti.E2<150]
```

Uudet datasetit voidaan piirtää erikseen, kuten tehtiin aiemmin, tai sovittaa yhteen kuvaajaan. Kaksi histogrammia voidaan piirtää päällekkäin säätämällä niiden läpinäkyvyyttä komennolla *alpha*. Histogrammien selitteet tulostetaan kuvaajan oikeaan yläkulmaan selvyyden vuoksi. Lisäksi kuvaajan piirtoväli (range) voidaan rajata analyysin kannalta kiinnostavalle alueelle (piikki 90 GeV:n tuntumassa).

In [6]:
```
plt.xlabel('Invariantti massa [GeV]')
plt.ylabel('Tapahtumien lukumäärä')
plt.title('Kahden myonin invariantti massa, vertailu energian perusteella \n')
plt.hist(uusiSettiMatalaE ['M'], bins=50, range=(80,100),alpha=0.5, label='Matala E')
plt.hist(uusiSettiKorkeaE ['M'], bins=50, range=(80,100),alpha=0.5, label='Korkea E')
plt.legend (loc='upper right')
plt.show()
```



Mitä tapahtuu, jos muutat energiarajan arvoa? Kokeile energiarajan muuttamista luodessasi muuttujat *uusiSettiKorkeaE* ja *uusiSettiMatalaE*. Millä muulla ehdolla voisit lajitella datan?