

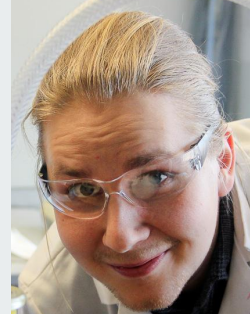


Databearbetning

Steget innan datavetenskap

Lektion 8 - Visualisering recap, Förhör och Inlämningsuppg 3

Dennis Biström
bistromd@arcada.fi



Upplägg - Vi e på slutrakan

Lektion 1 - Kursinfo, verktyg & resurser, Intro till Databearbetning. My first python app

Läxa 1 ut

Lektion 2 - Python Moduler och Klasser, My second and third app. Läxa 1 hjälp?

Förhör 1 ut, ~~Läxa 2 ut~~

Lektion 3 - Python Datastrukturer, Numpy & Matplotlib, Uppg 1 start

Förhör 1 in

Lektion 4 - Pandas, Uppg 1 forts

Läxa 2 ut, Uppg 1 ut

Lektion 5 - Visualisering, Webscraping & BeautifulSoup, Pandas, Uppg 2 start

Uppg 1 in, Uppg 2 ut, Uppg 3 ut

Lektion 6 - Visualisering forts. Matplotlib med textfiler, Ljud och Bilder som data

Uppg 2 in 21.10 kl 16.59

Lektion 7 - Övning med Ljud & kodande på inlämningsuppg 3

Uppg 3 in 28.10 kl 16.59

Lektion 8 - Övning med bild & kodande på inlämningsuppg 4

Uppg 4 in 4.11 kl 16.59

Feedback & Glögg på cornern? ~~4.11 kl 8.11~~ **8.11**

Visualisering

Andrew Gelman - Why tables are better than charts (Skriven första april)

“Graphs are a way of implying results that are often not statistically significant”

En välstrukturerad tabell är ärlig.

Vi är vana med tabulär data och har inga problem att uppfatta förhållanden. Använd för presentation

Alla grafer uppmuntrar eller insinuerar till slutsatser. Använd för att övertyga

Figure 11: GNH index by gender

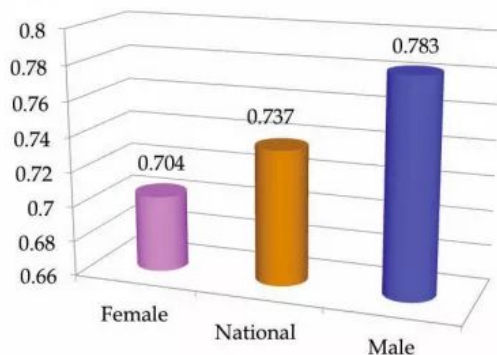


Figure 11

When we decompose the GNH Index by gender we see that men are happier than women.

49% of men are happy, while only one-third of women are happy.

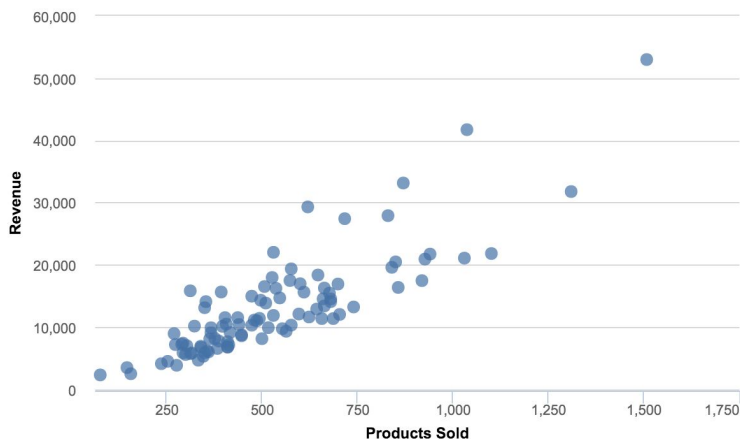
Gender	Happiness
Men	49 %
Women	33 %

Korrelation

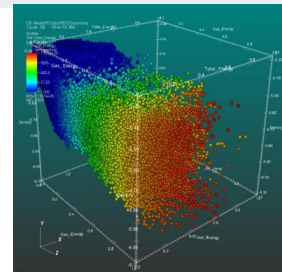
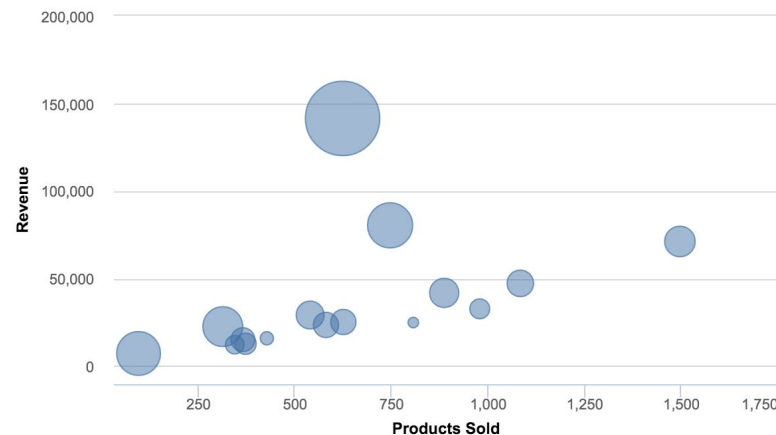
Scatter aka punktdiagram, även spridningsdiagram eller sambandsdiagram

För korrelation och distribution och även snabbt sätt att få insikt i din data

Illustrerar även outliers eller gruppering

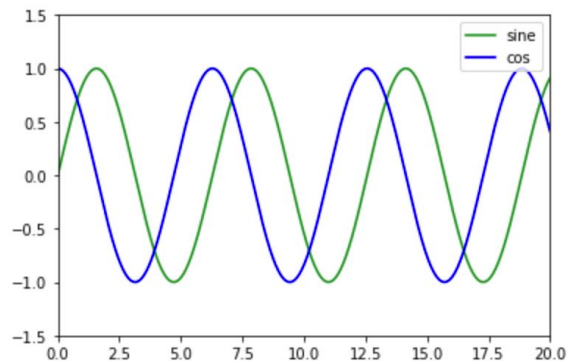


Få en insikt i samband mellan upp till fyra variabler

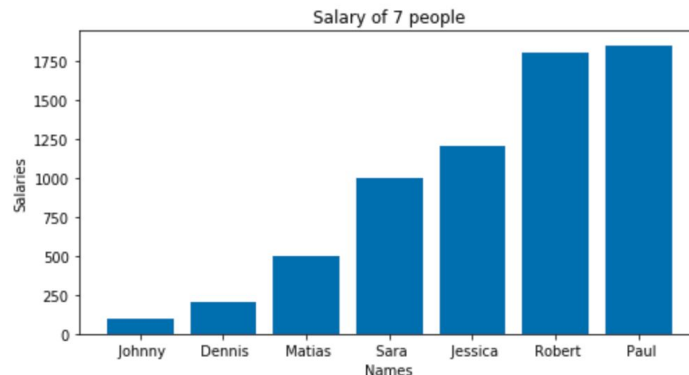


Visualiseringsövning - Matplotlib och Numpy

```
# Sin och Cos med legend & limits|
plt.plot(x , y1, "-g", label="sine")
plt.plot(x , y2, "-b", label="cos")
plt.legend(loc="upper right")
plt.ylim(-1.5, 1.5)
plt.xlim(0, 20)
plt.show()
```



```
# Vi använder indexing för att få bort stuff
plt.figure(figsize=(8, 4))
plt.bar(x[1:-2], salary[1:-2])
plt.xticks(x[1:-2], names[1:-2])
plt.ylabel("Salaries")
plt.xlabel("Names")
plt.title("Salary of 7 people")
plt.show()
```



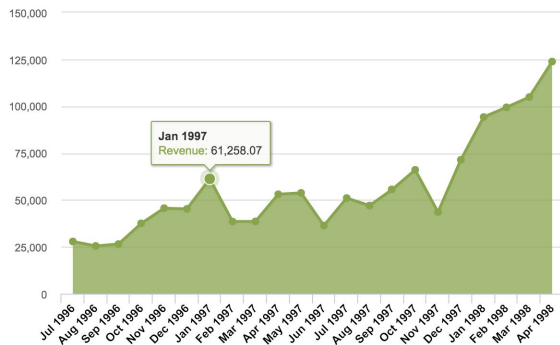
```
# Märk hur median vs mean ändrar
print('\nMedel: ', np.average(salary[1:-2]), "\nMedian: ", np.median(salary[1:-2]))
```

Kontinuerlig data

Kontinuerliga numeriska dataserier?

Ex temperatur

Linje eller Area - Integralen kan ge mervärde
Arean under en hastighet-tidsgraf är plats.



Diskret data

Diskreta numeriska serier?

Försiktigt med sampel och medeltal

Barchart eller Candlestick - Fördelen med candlestick är att man kan illustrera variation



Sampling - Riktiga världen och datorn

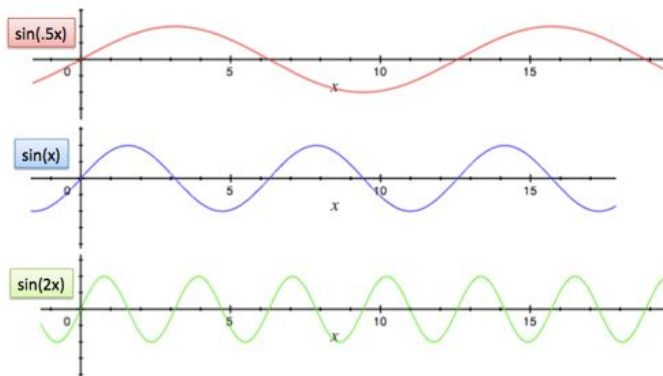
Vågformer - Ljus och ljud är vågformer*

Alla signaler och all data går att uttrycka med sinusvågor.

Det är inte helt omöjligt att spara sinusvågor på en dator, men nästan.

Därför SAMPLAR vi data och sparar diskreta värden.

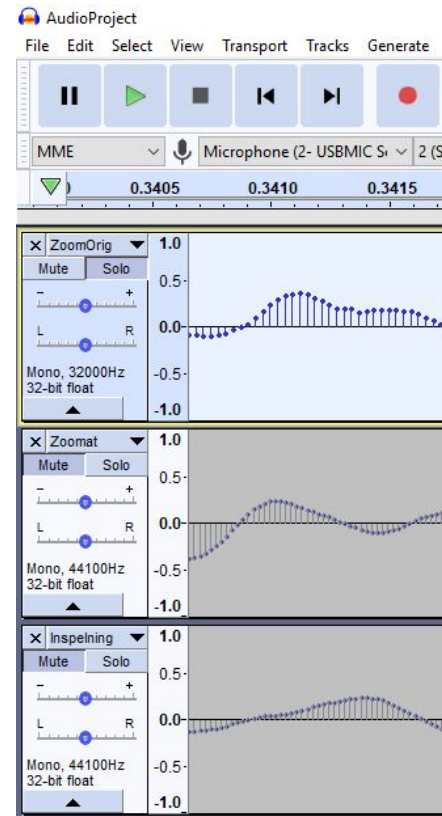
Exempel - Octave och Dataset



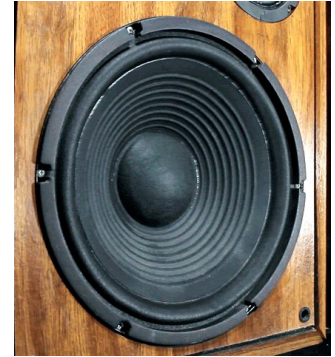
Mono
32kHz
32-bit float?

Varför 44100Hz
aka "CD-Quality"

Nyquist frekvens



Ljud - Vi är påväg mot bilder som data



Vågformer - Ljus och ljud är vågformer*

Många signaler behöver ett medium att färdas i ljud är skillnad i lufttryck

En våglängd är avståndet mellan toppar/dalar. Våglängd betecknas lambda λ

I en kontinuerlig signal bestämmer våglängden frekvensen f

Vi mäter ljudfrekvens i enheten Hz som betyder 1/s "hur många svängar per s"

Människan hör 20-20kHz ljud, men vad ser vi för våglängder? 430-770 THz

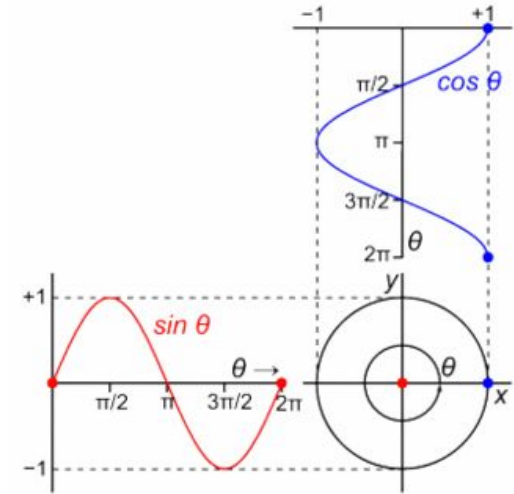
Kan man se ljud? Refraktion Kan man höra hastighet? Dopplereffekt

Amplitud - Hur stora värden sparar vi?

Period - Vad är frekvensen eller våglängden?

Fas - Börjar vi från noll?

Hör människan fasförskjutning? Nej

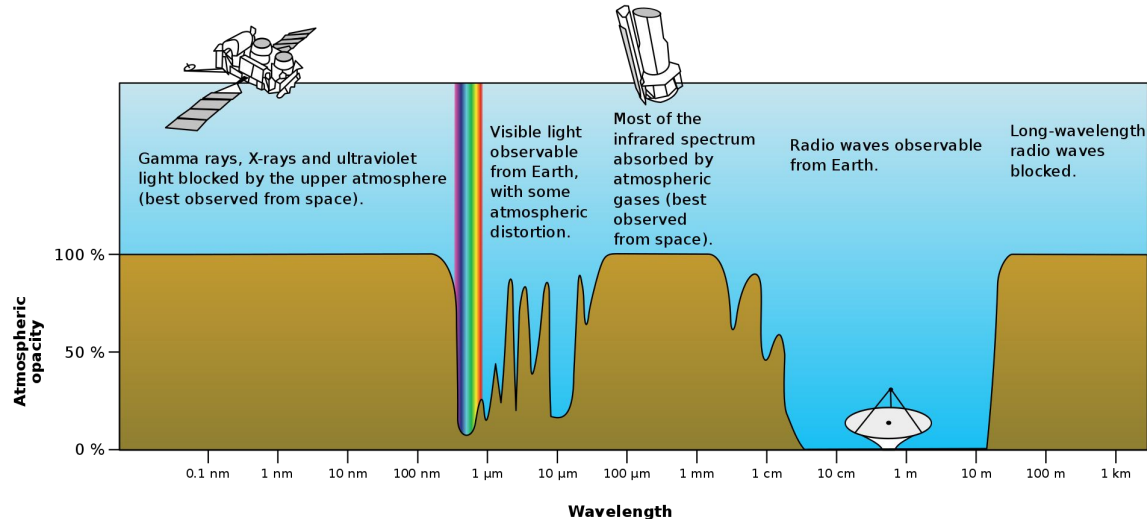
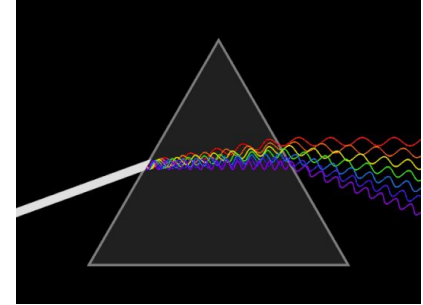


Ljus - EM strålning

Vågformer - Ljus och ljud är vågformer*

Synligt ljus ligger vid 400 - 700 nm ,dvs frekvenser vid 430–770 THz

Vi har valt att spara färg digitalt i rutor (pixlar) med tre färgkomponenter RGB





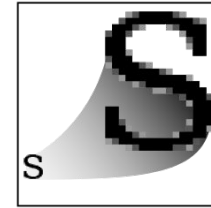
Demonstration of additive color mixing - [Zátonyi Sándor](#)

Bilder - Färger och lagring

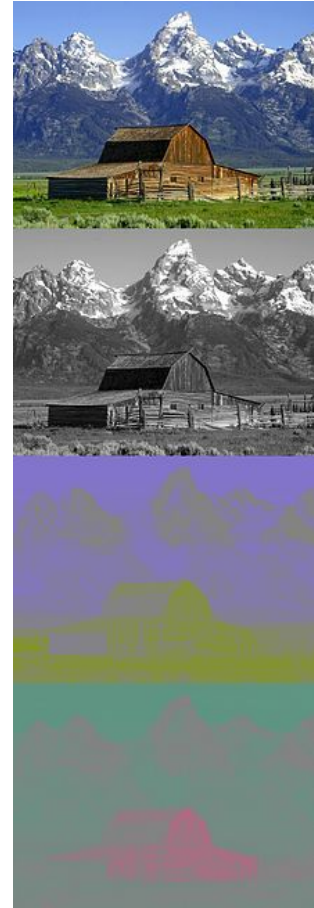
Vågformer - Ljus och ljud är vågformer*

24-bitars färgrymden reserverar 8 bitar per kanal RGB

Det här ger oss $2^{24} = 16,777,216$ färger



Raster
.jpeg .gif .png



8								8								8								8							
Red								Green								Blue								Alpha							
31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

På nätet uttrycker vi färger med hexadecimal, dvs tecken från 00-FF istället för 0-9

 rgb(255,0,255) blir #FF00FF

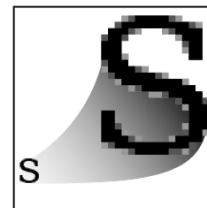
Blu-Ray och DVD standarden använder YCbCr eftersom det möjliggör bättre kompression av färgkanalerna i jämförelse med RGB. (Lossless conv. till RGB)

Bilder - Leka med datan

Pixlar på vår skärm sparas i RGB(A) format.

Varje pixel har tre 8-bits värden för färg, och ett 8-bits värde för transparens

Exempel: Histogrammet från kursen i D3



Raster
.jpeg .gif .png

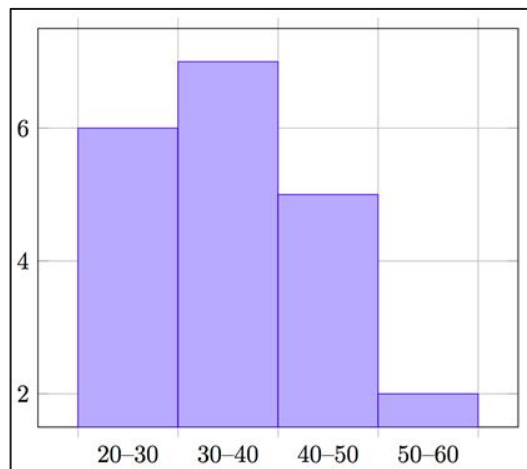


Histogram

Stapeldiagram med kategorier

Åskådliggör distribution

[Matplotlib.pyplot.hist](https://matplotlib.pyplot.hist)

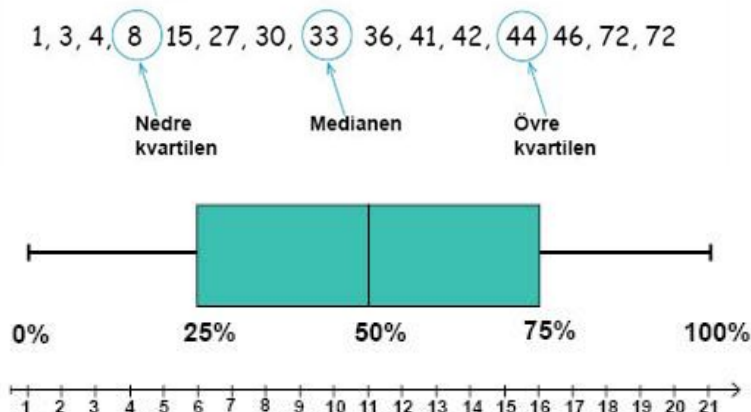


Lådagram

Distribution med

Spridning, Kvartiler & Percentiler

[Pyplot.subplots, axis.boxplot](https://matplotlib.pyplot.boxplot)





[Lena Söderberg](#) shot by photographer [Dwight Hooker](#) från November 1972
The image has produced controversy because *Playboy* is seen as being degrading to women",^[9]
The Lenna photo has been pointed to as an example of sexism in the sciences, reinforcing gender stereotypes.



Fabio Lanzoni, unknown photographer

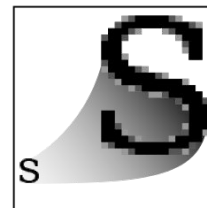
Kompression - Bits och Bytes

Pixlar på vår skärm sparas i RGBA format.

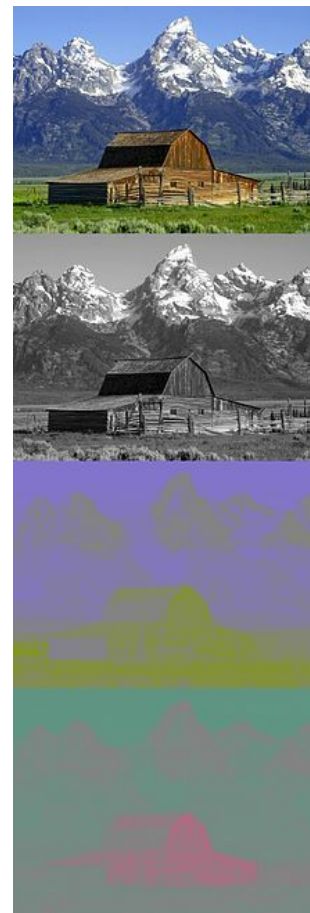
Varje pixel har tre 8-bits värden för färg, och ett 8-bits värde för transparens

Övning med bilder i matplotlib!

Öppna Jupyter



Raster
.jpeg .gif .png



Kompression - Bits och Bytes

Pixlar på vår skärm sparas i RGBA format.

Varje pixel har tre 8-bits värden för färg, och ett 8-bits värde för transparens

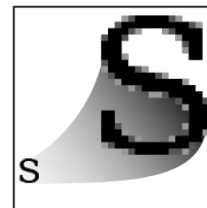
Övning med bilder i matplotlib!

Öppna Jupyter

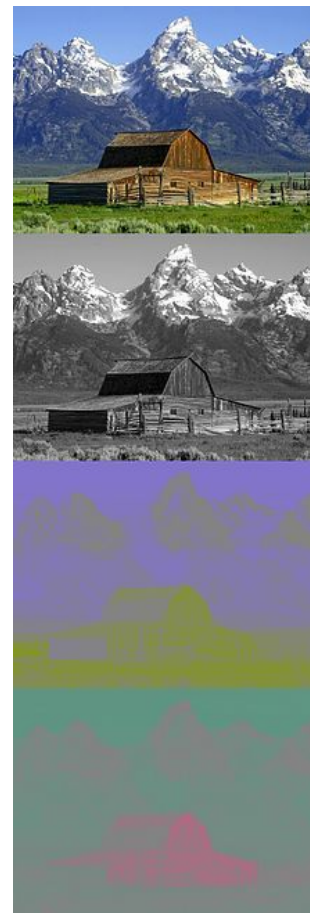
Hur mycket space tar bilder? Exempel Lena:

$64^2 = 4096$ st pixlar (RGB och kanske A) \Rightarrow 4 st 8 bitars värden.

Ett 8-bitars värde: 10000000 vad är det här för färg?



Raster
.jpeg .gif .png





A picture of Mohammed Alim Khan (1880-1944), Emir of Bukhara, taken in 1911

Inlämningsuppg 4 - Delmoment 2 och 3



Bilder som data

- 1) Ladda ner fabio från IL, och göm en hälsning i alphakanalen av bilden (t.ex texten "HEJ!").
Gör en jupyter notebook där du visar hur man får fram informationen:
Använd dig av ett histogram för att illustrera var den gömda datan ligger
Gör en before/after figur med imshow och pseudofärg där meddelande syns
- 2) Skapa en färgbild genom att kombinera luminanskanalerna av kabuto (elr valfri poke)
Ladda in de tre kanalerna som skilda arrays och kombinera dem till en "rgb array"
som du sedan visualiserar.

Inlämningsuppg 3 - Gårdagens hjälp

Inlämningsuppg 3

Moore's lag säger att antalet transistorer i en mikroprocessor fördubblas ungefär varje 2 år.

[Moore's Law over 120 Years](#)

Stämmer det?

Ta in data från https://en.wikipedia.org/wiki/Transistor_count om år och transistorantal

(OBS! scraping, tvättning, sorting av data behövs som vanligt), och rita tre grafer av transistorantalens utveckling.

1. Stapeldiagram CPU:s/Decade
2. Linjediagram nm:s/Decade
3. Scatter transistors/year*

*Använd logaritmisk skala på y-axeln, och sätt in en linje som visar vad ökningen borde vara enligt Moores lag.

Märk ut några valda punkter med proessorns namn.


Jag har redan fallit av kälken! - Brush up ur skills



1. Kolla [Socratica tutorialen](#) videorna 1-17
2. [Chapter 1-3 - Python for data science \(DataCamp\)](#)
3. [Chapter 1 - Writing python functions \(DataCamp\)](#)
4. [Chapter 4 - Numpy \(DataCamp\)](#)
5. [Chapter 1 - Matplotlib \(DataCamp\)](#)
6. [Chapter 1-4 - Pandas for data science \(Lynda: Kelly\)](#)

Sök hjälp bland resurserna om du kör fast.

Lynda och resurser - Kolla även itslearning!

- 
- Cheat sheet: [Anaconda Cheat Sheet - Getting Started - PDF](#)
[Pandas Cheat Sheet - PDF](#)
- Manual/Docs [Conda package manger - Docs](#)
[Pandas - QuickStart & Cookbook](#)
- Tutorials (text) [Anaconda Getting Started - User Guide](#)
[Python - Intro till avancerat - Övningar och förklaringar](#)
[Pandas tutorial - PythonSpot](#)
[Intro to data science Numpy, MatPlot & Panda](#)
[\(Pandas - How do pivot tables work - ExcelCampus\)](#)
- Tutorials (video) [Socratica python tutorial - Youtube](#)
[Derek Banas - "Learn Python in one video"](#)

Lynda och resurser2 - Kolla även itslearning!



Interaktiva:

[Intro to **python** for data science - Gratiskurs - DataCamp](#)

[Intro to python for data science - Ch4 - **Numpy** \(DataCamp\)](#)

[Intermediate python for data science - Ch1 - **MatPlotLib** \(DataCamp\)](#)

Lynda:

[6h nybörjarkurs **Python** för datavetenskap med Lillian Pierson - Lynda](#)

[2h intermediate - **Numpy** Data Science Essentials - Charles Kelly](#)

[Intermediate - Ch3: **Numpy**, Ch4: **Pandas**, Ch 9: **matplotlib** - Miki Tebaka](#)

[2h intermediate kurs i **Pandas** med Jonathan Fernandes - Lynda](#)

[2h intermediate **Pandas** för Datavetenskap med Charles Kelly - Lynda](#)

[Big Data Analysis in python using **Numpy** and **Pandas** - Michele Vallisneri](#)

Lynda och resurser3 - Kolla även itslearning!



Interaktiva roligheter

[The Python Challenge](#)

[Roliga övningar i logisk ordning - Practice Python](#)

[How to think like a Computer Scientist](#)

[CodeSignal - Interaktiva utmaningar, badges, points etc.](#)

[Reddit daily programmer challenges](#)

Vill du vinna 1 miljon \$

[7h gratis tutorial på Kaggle - Känner ni till kaggle?](#)

Interaktiva, bra helheter

Python 2 vs Python 3 trubbel, kolla [här](#)