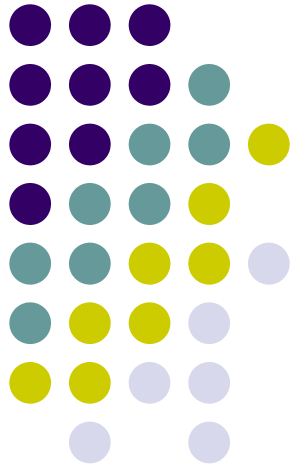
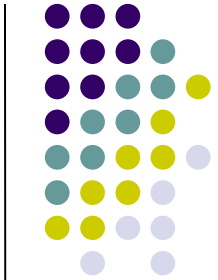
Obrada informacija
Pregled tema

Marko Subašić

http://www.fer.hr/predmet/obrinf_a

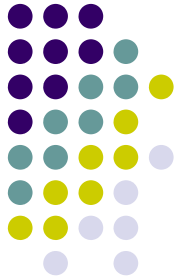


Obrada informacija



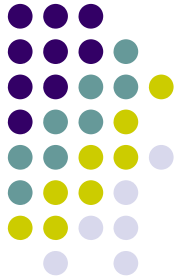
- Vrlo širok pojam
- Postoji puno tipova informacija
- Postoji puno tehnika obrade informacija
- Cjeline su organizirane po uzlaznoj kompleksnosti informacija
  - Kompleksnost “mjerimo” brojem dimenzija

Obrada informacija



- Krećemo prvo sa 1D signalima
  - Vremenska mjerenja fizikalnih veličina
  - Genske sekvence
  - Govor
  - Financijki signali (višedimenzionalni signali)
- 2D signali – slike
  - Analiza slika
  - Rekonstrukcija 3D prostora iz video snimke

Analiza mjerenja fizikalnih veličina

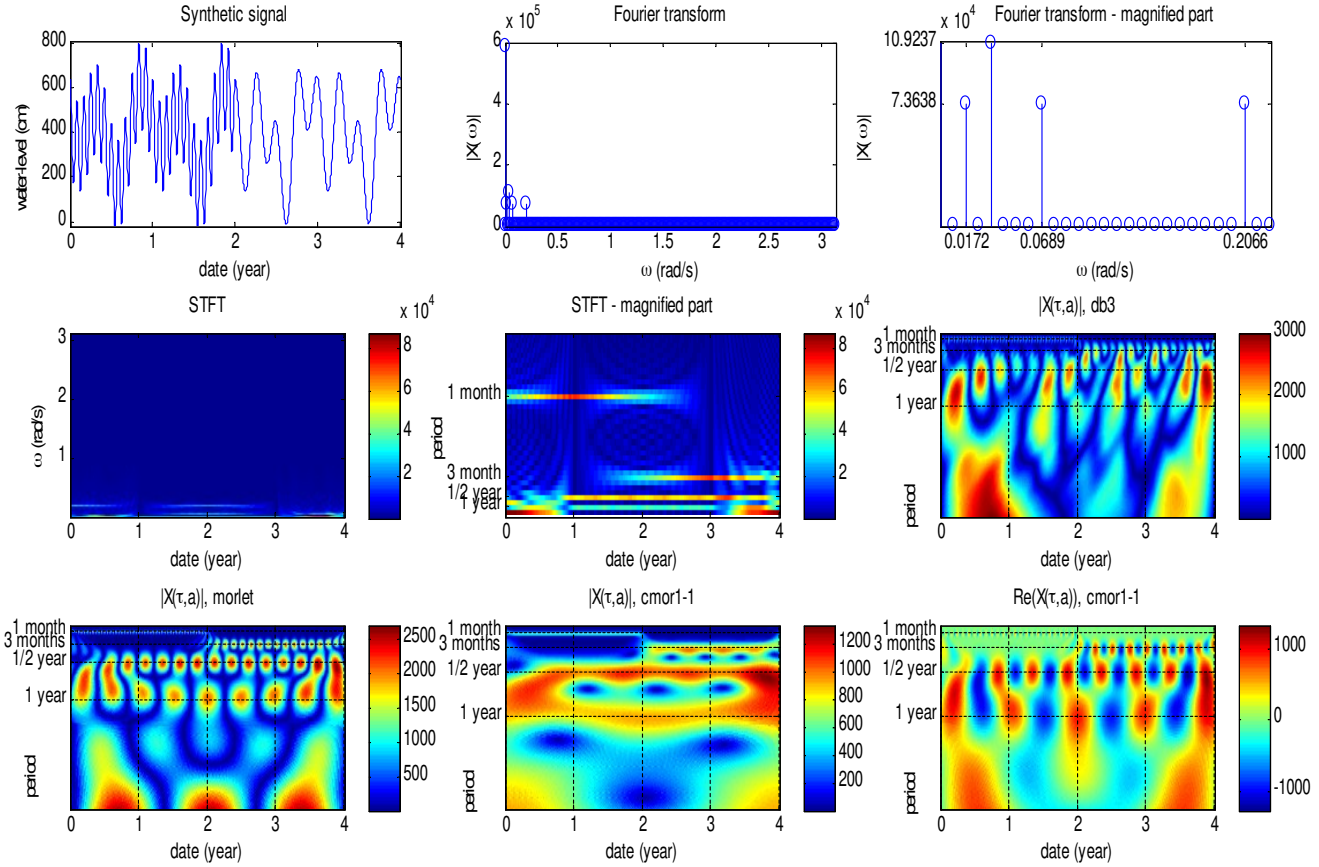


- Vremenski signali – 1D signali
- Možemo li predvidjeti vodostaj Save?
- Koja je frekvencija pojave poplava?
- Kao pomoć u otkrivanju prirodnih pojava može poslužiti:
  - Fourierova transformacija
  - Fourierova transformacija na vremenskom otvoru
  - Valićna transformacija

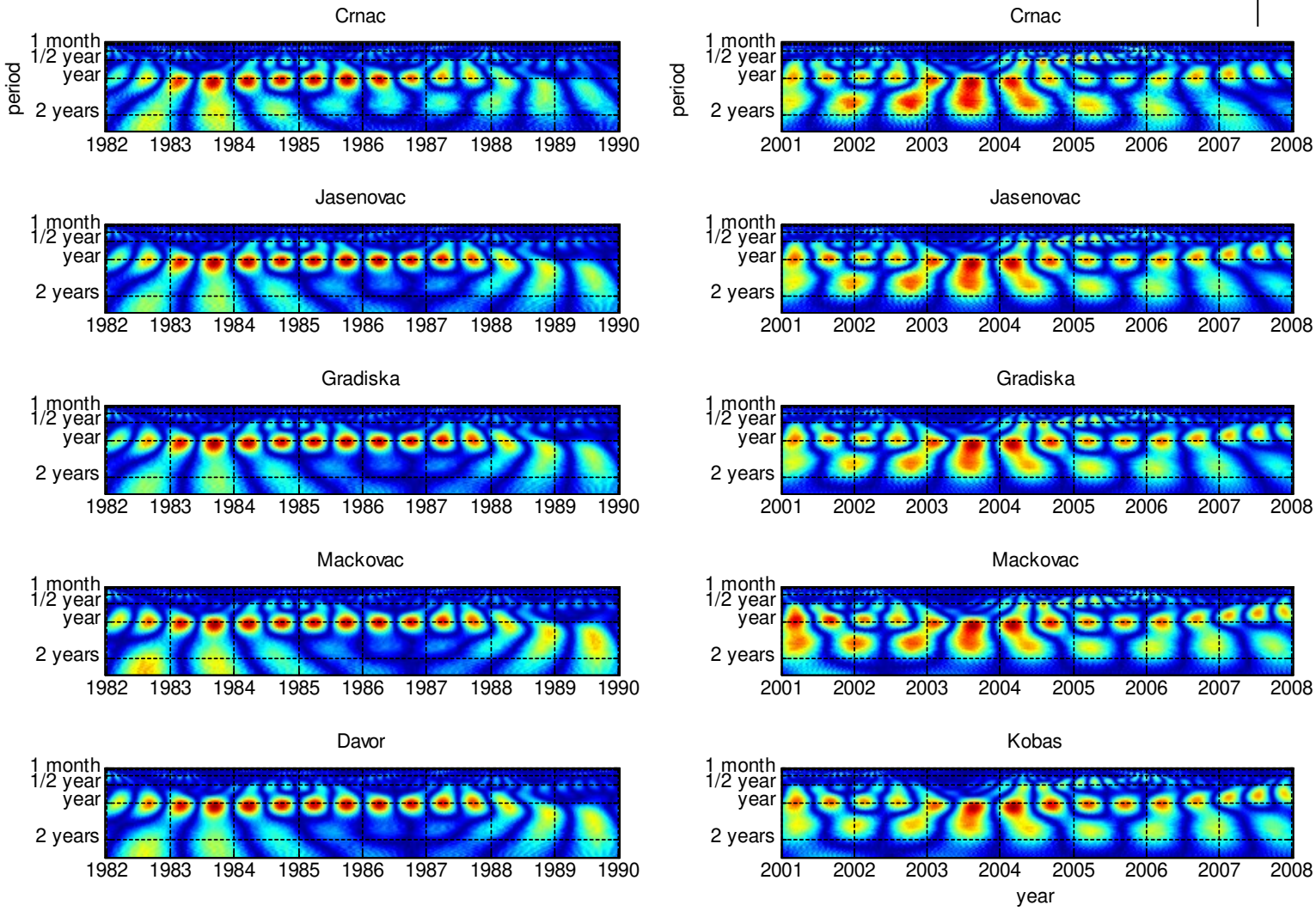
Analiza mjerenja fizikalnih veličina

$$x_1(t) = 400 + 100 \cdot \cos(2\pi \cdot 1/T \cdot t) + 150 \cdot \cos(2\pi \cdot 2/T \cdot t - T/6) + 200 \cdot \cos(2\pi \cdot 12/T \cdot t), \text{for } t \in \langle 0, 2T]$$

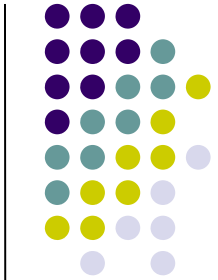
$$x_2(t) = 400 + 100 \cdot \cos(2\pi \cdot 1/T \cdot t) + 150 \cdot \cos(2\pi \cdot 2/T \cdot t - T/6) + 200 \cdot \cos(2\pi \cdot 4/T \cdot t), \text{for } t \in \langle 2T, 4T]$$



Analiza mjerenja fizikalnih veličina



Bioinformatika



- Korištenje računala za analizu bioloških podataka, u užem smislu za alanizu DNK, RNK i proteina.
- Merriam Webster dictionary:
  - „the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics”
- Osnovni zadaci
  - Sastavljanje genoma
  - Određivanje varijacija u genomima
  - Ekspresija gena
  - Određivanje kompozicije uzoraka
- Primjena u medicini, farmaciji, biologiji, poljoprivredi ...

Bioinformatika

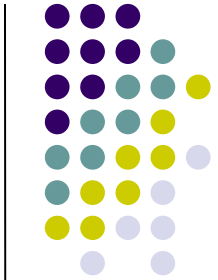
Cijele DNK molekule, sastoje se od ACTG

Postupak sekvenciranja (engl. sequencing)

Očitanja (engl. sequencing reads)

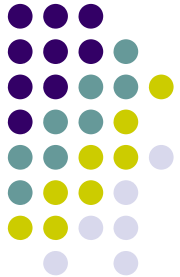
- Postupak sekvenciranja pokušava pročitati DNK molekulu. Očitanja i cijele DNK molekule pohranjuju se kao nizovi znakova **ACTG**.
- Današnje tehnologije ne mogu pročitati cijele DNK molekule već čitaju manje dijelove (očitanja) i to s određenom greškom. Duljina očitanja te učestalost i vrsta pogreške ovisi o korištenoj tehnologiji.

Bioinformatika



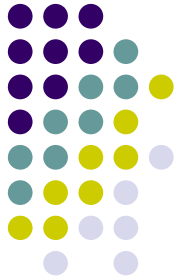
- Osnovni postupak pri svakoj bioinformatičkoj analizi: **Poravnanje očitanja na referencu ili na druga očitanja**
- Poravnanje je postupak kojim se određuju najsličniji djelovi dvaju nizova (npr. dio referentnog genoma koji je najsličniji pojedinom očitanju).
- Načini određivanja poravnanja:
  - Egzaktno pomoću dinamičkog programiranja
  - Približno pomoću abecedno minimalnih podnizova (engl. minimizer) i LCS algoritma (longest common subsequence)
  - **Približno koristeći MAFFT algoritam – poravanje više od dvije sekvence pomoću brze Fourierove transformacije** – tema laboratorijskih vježbi

Govorni signali



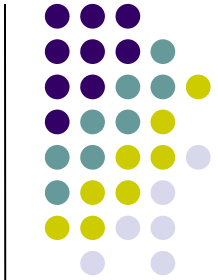
- Svakodnevna primjena obrade informacija - u automatiziranim sustavima za dijalog
  - Automatsko prepoznavanje govora (engl. Automatic Speech Recognition, ASR)
  - Automatska sinteza govora iz teksta (Text to Speech, TTS)
  - Obrada prirodnog jezika (engl. Natural Language Processing, NLP)
  - Osobni asistenti, Amazon Alexa, Apple Siri, Google Assistant, … i mnogi drugi

Principi rada sustava za obradu govora



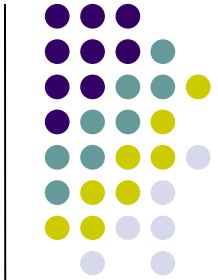
- Intenzivno korištenje najmodernijih postupaka strojnog učenja i umjetne inteligencije
- Najsloženiji i računski najzahtjevniji dio obrade se provodi „u oblaku“.
- Danas su najpopularniji postupci temeljeni na dubokom učenju i konvolucijskim neuronskim mrežama.
- Sve su to postupci kojima se opisuju vremenski dinamička statistička svojstva nosioca informacije, koji je u ovom primjeru ljudski glas.
- Korištenjem takvih statističkih modela moguće je provoditi klasifikaciju, tj. prepoznavanje glasova, riječi, rečenica, čime govor pretvaramo u tekst
- Moguće je male razlike između značajki glasa iskoristiti i za biometrijske primjere u svrhu automatskog prepoznavanja govornika
- Moguće je automatski prepoznavati jezik ili narječje, temeljem svojstava govora
- Može se tvrditi da su postupci digitalne obrade govornog signala bili jedan od prvih i glavnih pokretača razvoja podatkovne znanosti, još u osamdesetim godinama prošlog stoljeća

Govorni signali



- HMM – Skriveni Markovljevi modeli
  - Jedan od vrlo jednostavnih statističkih modela za opis ponašanja vremenskih serija s promjenjivim statističkim svojstvima su Skriveni Markovljevi modeli (engl. Hidden Markov Models, HMM).
  - Pokazali su se iznimno pogodnim za modeliranje govora na više razina: od pojedinačnih glasova, slogova, dijelova ili cijelih riječi i konačno cijelih rečenica.
  - Ovi modeli se grade automatiziranim nadziranim ili nenadziranim postupcima učenja iz uzoraka stvarnog govora, što je konceptualno jednako postupcima koji se koriste kod učenja neuronskih (ili dubokih) mreža.
  - HMM modeli će biti detaljnije predstavljeni u sklopu predmeta, te će se ilustrirati postupci njihove primjene za modeliranje raznovrsnih procesa.Opis signala pomoću vektora značajki

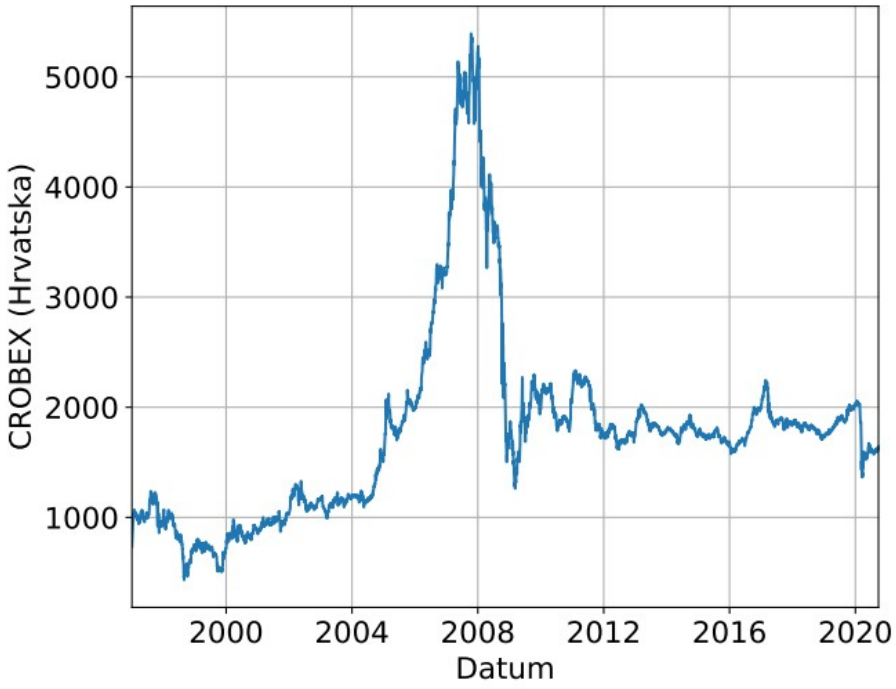
Financijski signali ili financijski vremenski nizovi

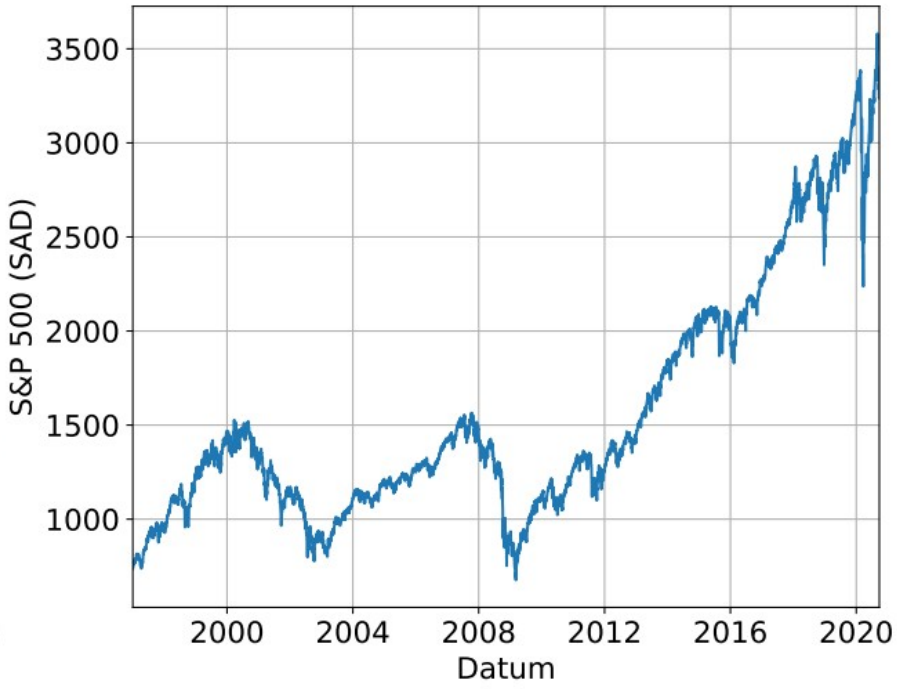


- Cijene financijskih instrumenata
  - cijene dionica (tržišta kapitala)
  - cijene obveznica (tržišta obveznica)
  - cijene opcija (tržišta izvedenica)
  - Itd.
- Makroekonomske varijable
  - kamatne stope
  - tečajne liste (valute)
  - bruto domaći proizvod (BDP)
  - itd.
- Fundamentalni podatci o kompanijama
  - zarada po dionici
  - knjigovodstvena vrijednost po dionici
  - dug po dionici

Primjer: CROBEX i SP&500

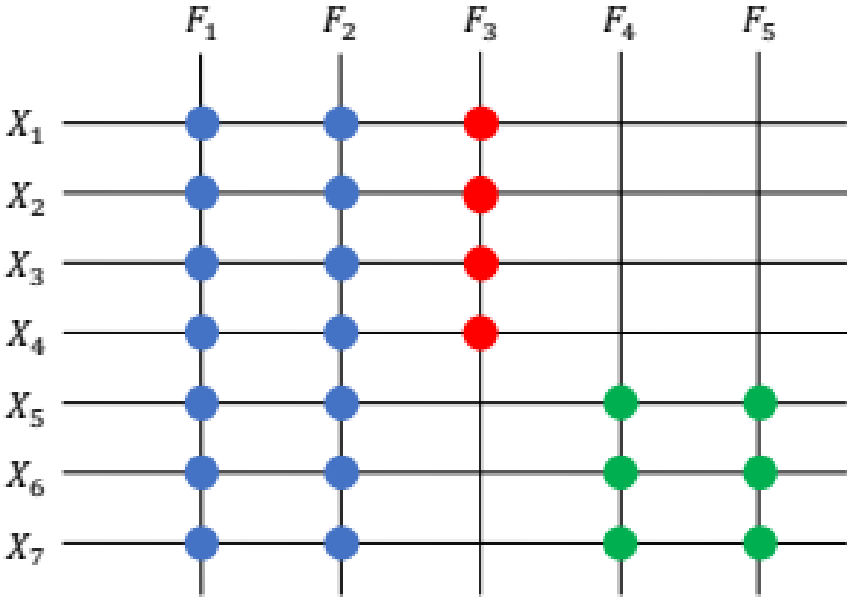
- Dionički indeks - prosjek vrijednosti kompanija kojima se trguje na nekom tržištu





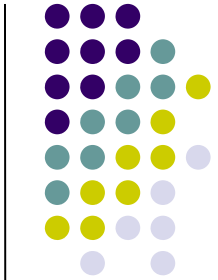
Faktorska struktura

- Na sve dionice na zagrebačkoj burzi utječu kamatne stope, BDP, itd, ali na turistički sektor utječe npr. broj gostiju, dok na građevinski sektor utječu demografski podatci



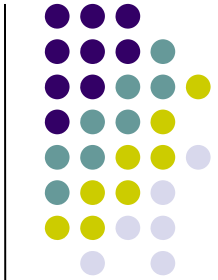
- Za model: **X** = **FB**' + **e**, problem je:
- odrediti **B**, ako su poznati **X** i **F**
- odrediti **F** i **B**, ako je poznat samo **X**

Analiza slika



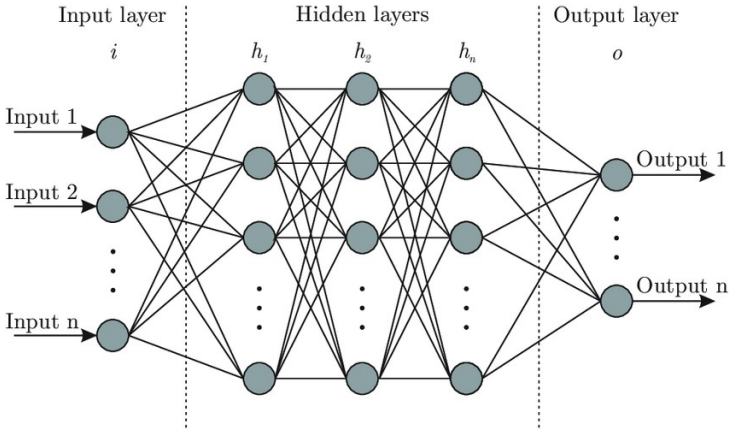
- Pronalazak bitne informacije u slici
- Nužno je eliminirati ili ignorirati nebitne informacije
  - Uklanjanje šuma iz slike
- Razni problemi
  - Klasifikacija slika
  - **Detekcija objekata na slikama**
  - Raspoznavanje objekata na slikama
  - ...

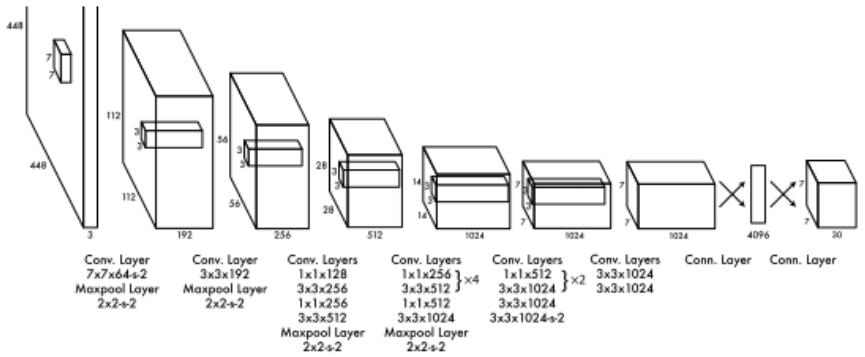
Analiza slika

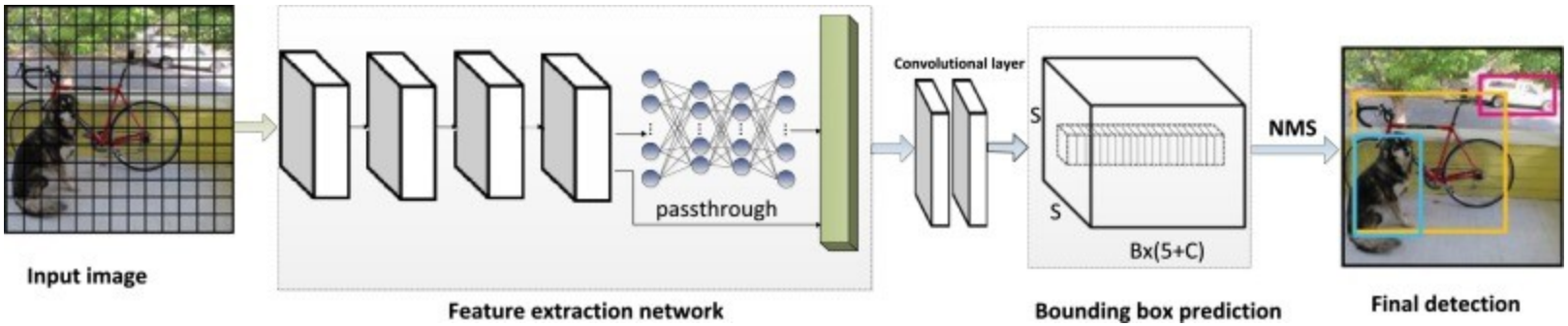


- Složenost zadatka može biti različita i ovisi o konkretnom problemu
- Složeni problemi analize slike uspješno se rješavaju dubokim neuronskim mrežama
- Neuronske mreže dolaze iz područja umjetne inteligencije, odnosno strojnog učenja
  - Duboke neuronske mreže predstavljaju nedavno ostvareni napredak u području

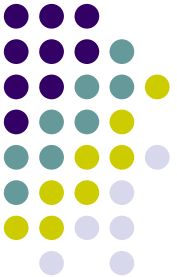
Neuronske mreže







Obrada video podataka



- Video signali kombiniraju vremenski i prostornu kodomenu
- Često je potrebna obrada u stvarnom vremenu
- Primjer analize snimke s projiciranim uzorkom strukturiranog svjetla za 3D rekonstrukciju

Obrada nestrukturiranih podataka

- Oblaci točaka - dobiveni 3D skeniranjem
- Svaka točka uz koordinate (x,y,z) može sadržavati i dodatne podatke kao boja i normala
- Problem spajanja više oblaka točaka u jedan (ICP algoritam)