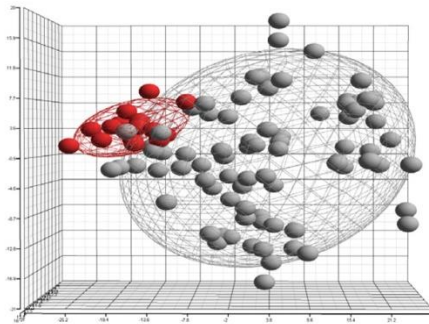
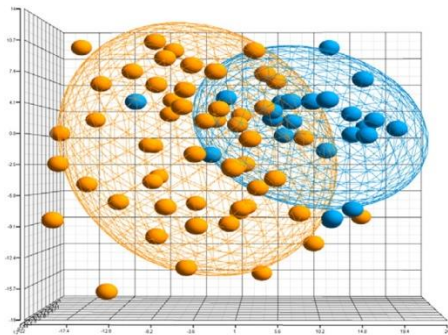
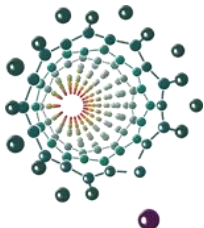




Анализа главних компонената

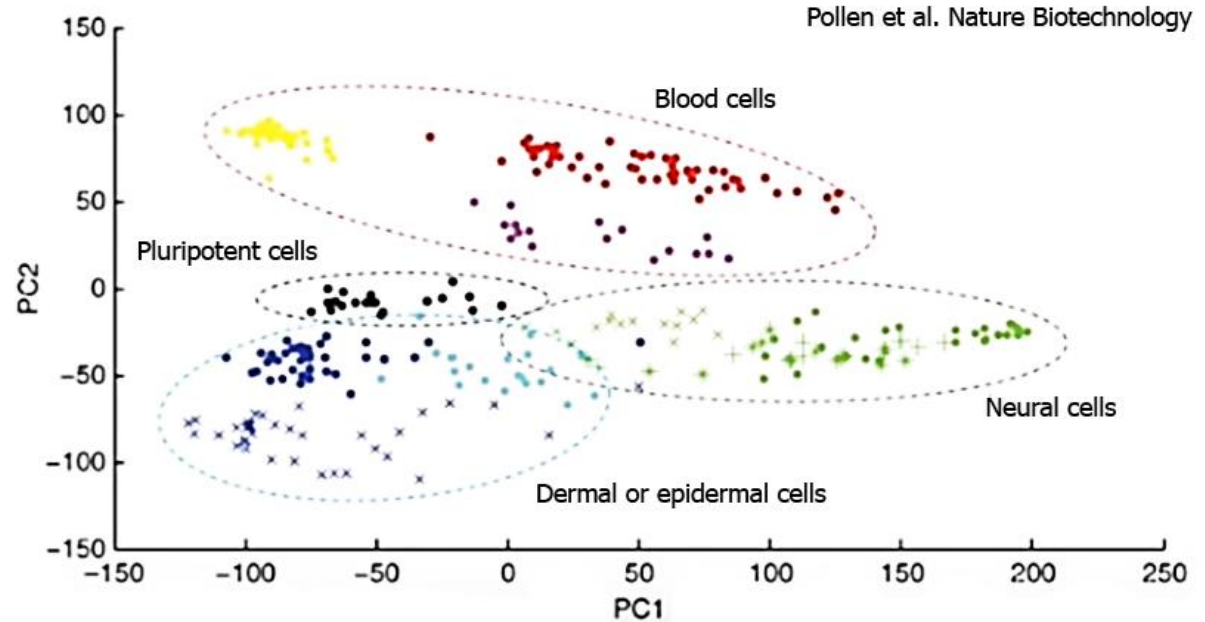
PCA

(Principal Component Analysis)

Милош Мојовић. в. проф.

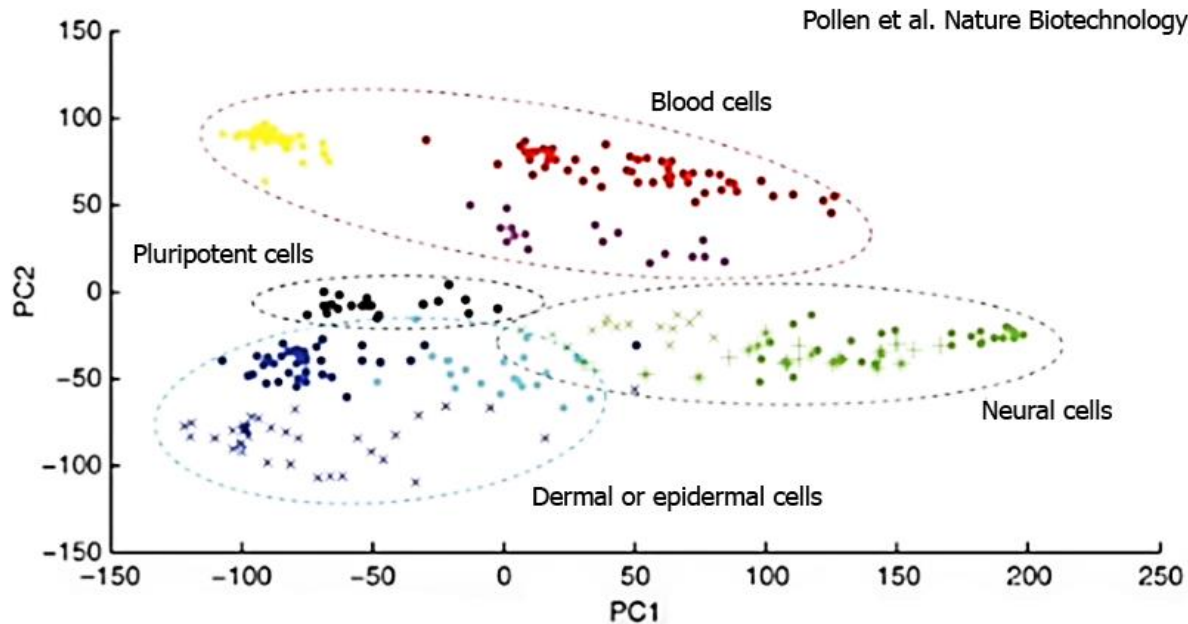
Шта је *PCA*?

- Свака тачка на овом графику представља једну ћелију.
- Ћелије су груписане према сличности експримованих гена
- У свакој ћелији се налази око 10000 гена.
- Како смо успели да извршимо класификацију из оволико великог броја података?
- Како су све информације о експресији транскрипције 10000 гена компримоване у 1 тачку на графику?



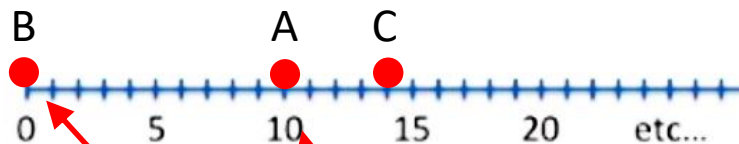
Шта је *PCA*?

- *PCA* је метод који може да из великог броја података извуче само оне који представљају саму суштину добијене информације.
- Шта представљају осе на овом графику (*PC1* и *PC2*)?
- *PC1* и *PC2* су тзв. главне компоненте.
- То су две информације о свакој ћелији на које су сведени сви остали подаци.
- Како смо то успели да урадимо?



Прича о димензијама

- **1 – димензиони простор**
- То је права 😊.
- Представимо на тој правој експресију различитих гена за 1 ћелију.
- Али, уколико наставимо (с обзиром на велики број гена) добићемо један од ова два случаја:

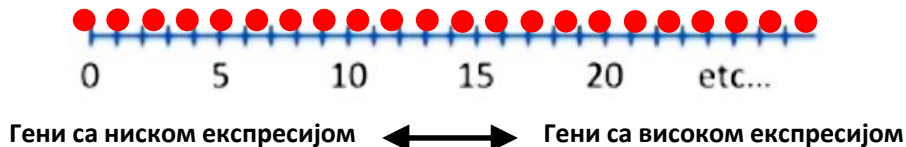


Ген:	Број читавања:
a	10
b	0
c	14
...	...

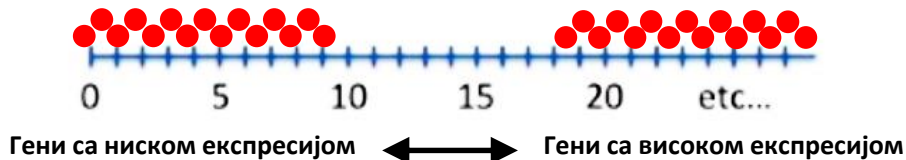
Прича о димензијама

- **1 – димензиони простор**
- То је права 😊
- Представимо на тој правој експресију различитих гена за 1 ћелију.
- Али, уколико наставимо (с обзиром на велики број гена) добићемо један од ова два случаја:
 - Униформна расподела
 - Не-униформна расподела

Униформна расподела експресије гена

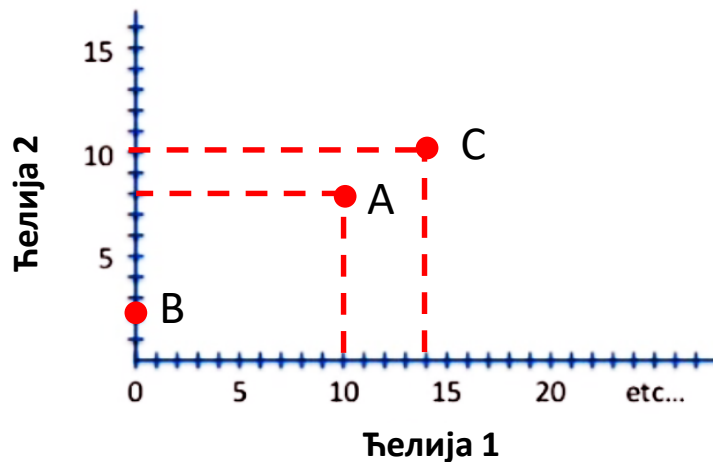


Не-униформна расподела експресије гена



Прича о димензијама

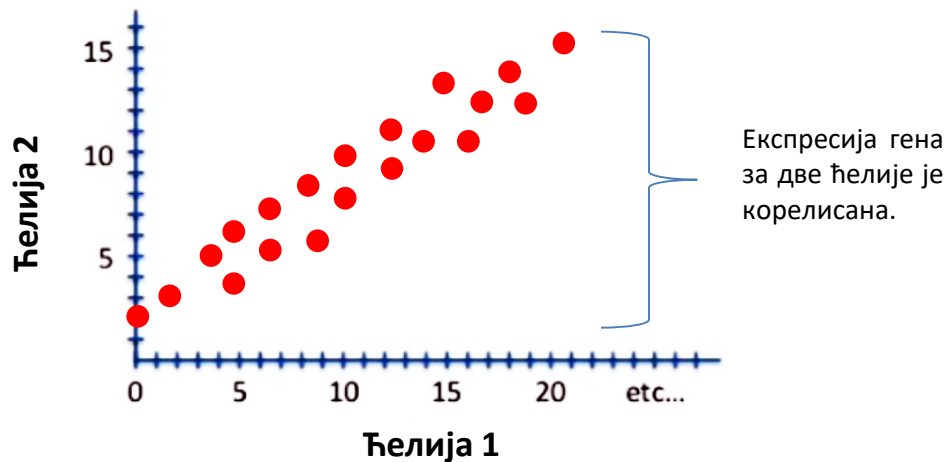
- **2 – димензиони простор**
- То је график са 2 осе (X-Y).
- Представимо сада у 2Д простору експресију различитих гена за 2 ћелије.
- Свака ћелија је има своју осу на којој приказујемо експерисију њених гена.



Ген	Број читавања гена за ћелију 1	Број читавања гена за ћелију 2
a	10	8
b	0	2
c	14	10
...	...	...

Прича о димензијама

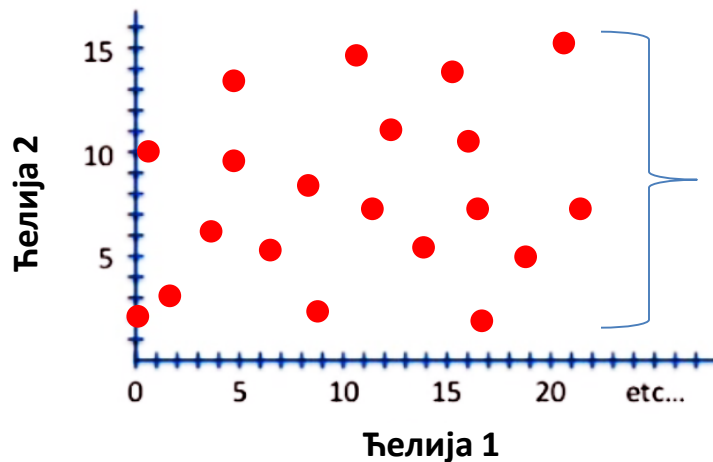
- Ако наставимо да уцртавамо на истом графику све гене за 2 ћелије, добићемо један од ова два случаја:
 - Експресија гена за две ћелије је корелисана.



Ген	Број читавања гена за ћелију 1	Број читавања гена за ћелију 2
a	10	8
b	0	2
c	14	10
...	...	...

Прича о димензијама

- Ако наставимо да уцртавамо на истом графику све гене за 2 ћелије, добићемо један од ова два случаја:
 - Експресија гена за две ћелије је корелисана.
 - Експресија гена за две ћелије није корелисана.

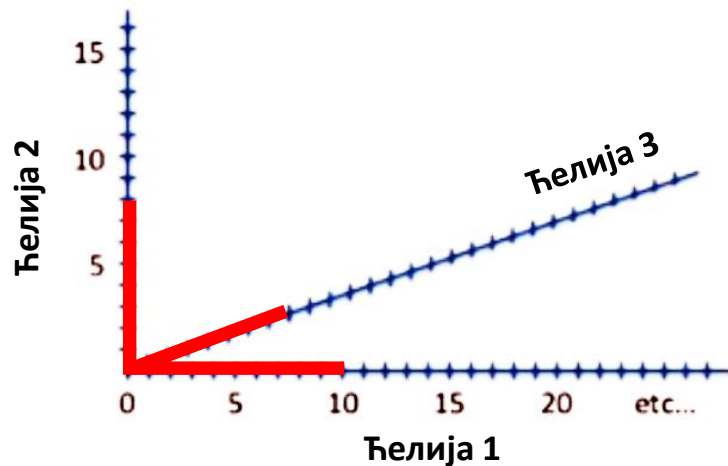


Експресија гена
за две ћелије
није корелисана.

Ген	Број читавања гена за ћелију 1	Број читавања гена за ћелију 2
a	10	8
b	0	2
c	14	10
...	...	...

Прича о димензијама

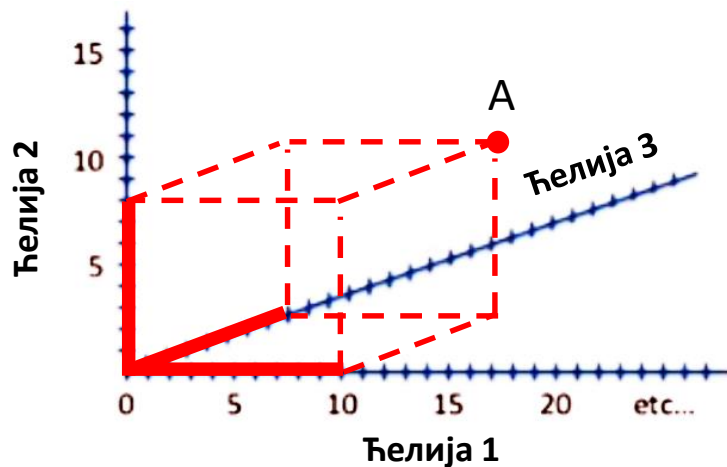
- **3 – димензиони простор**
- То је график са 3 осе (X-Y-Z).
- Представимо сада у 3Д простору експресију различитих гена за 3 ћелије.
- Свака ћелија је има своју осу на којој приказујемо експресију њених гена.



Ген	Број читавања гена за ћелију 1	Број читавања гена за ћелију 2	Број читавања гена за ћелију 3
a	10	8	8
b	0	2	4
c	14	10	12
...	...	...	...

Прича о димензијама

- **3 – димензиони простор**
- То је график са 3 осе (X-Y-Z).
- Представимо сада у 3Д простору експресију различитих гена за 3 ћелије.
- Свака ћелија је има своју осу на којој приказујемо експресију њених гена.
- Тако ћемо за експресију гена А код све три ћелије добити тачку ...
- Исто тако можемо урадити за све остале гене ...



Ген	Број читавања гена за ћелију 1	Број читавања гена за ћелију 2	Број читавања гена за ћелију 3
a	10	8	8
b	0	2	4
c	14	10	12
...	...	...	...

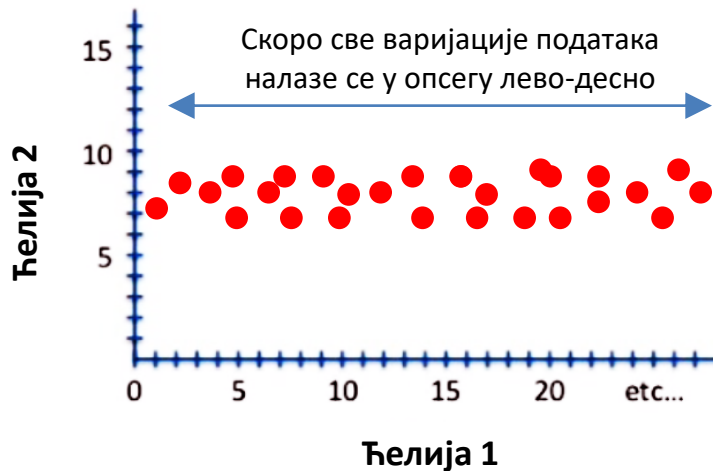
Шта смо установили до сада?

- 1 ћелија = 1Д график (координатна оса)
- 2 ћелије = 2Д график (обичан X-Y график)
- 3 ћелије = 3Д график (познат X-Y-Z график)
- 4 ћелије = 4Д график (овде сада имамо проблем јер нисмо у стању да га нацртамо)
- 200 ћелија = 200Д график (ово дефинитвно нисмо у стању ни да замислимо, а камо ли да нацртамо)

Питање: Да ли су све ове димензије подједнако важне или су неке можда „важније“ од других?

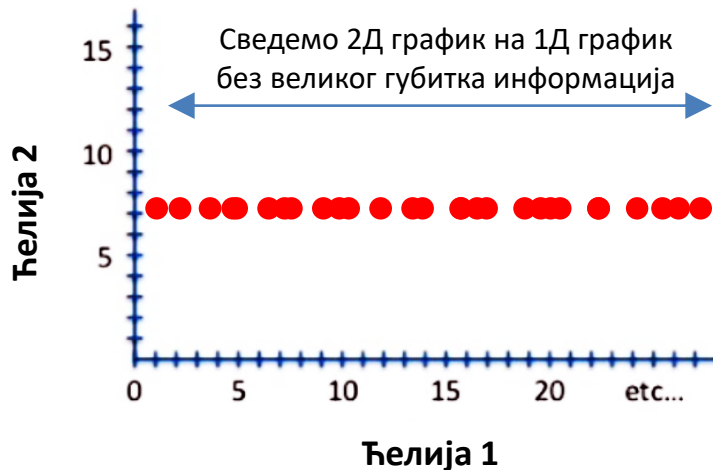
Вратимо се на 2 ћелије и 2Д график

- Претпоставимо да 2Д график који је добијен за 2 ћелије изгледа овако.
- Видимо да варијације у експресији гена за те 2 ћелије постоје само дуж осе која представља ћелију 1.
- У ствари, за ћелију 2, вредности експресије свих гена су веома сличне.
- Зато, нећемо направити велику грешку, уколико експресију гена представимо 1Д графиком (уклонимо варијације горе-доле).
- Овако нисмо изгубили пуно информација о систему, али смо успели да „изгубимо“ једну димензију.



Вратимо се на 2 ћелије и 2Д график

- Претпоставимо да 2Д график који је добијен за 2 ћелије изгледа овако.
- Видимо да варијације у експресији гена за те 2 ћелије постоје само дуж осе која представља ћелију 1.
- У ствари, за ћелију 2, вредности експресије свих гена су веома сличне.
- Зато, нећемо направити велику грешку, уколико експресију гена представимо 1Д графиком (уклонимо варијације горе-доле).
- Овако нисмо изгубили пуно информација о систему, али смо успели да „изгубимо“ једну димензију.



**Значи, ВАЖНИ подаци су били у
димензији лево-десно**

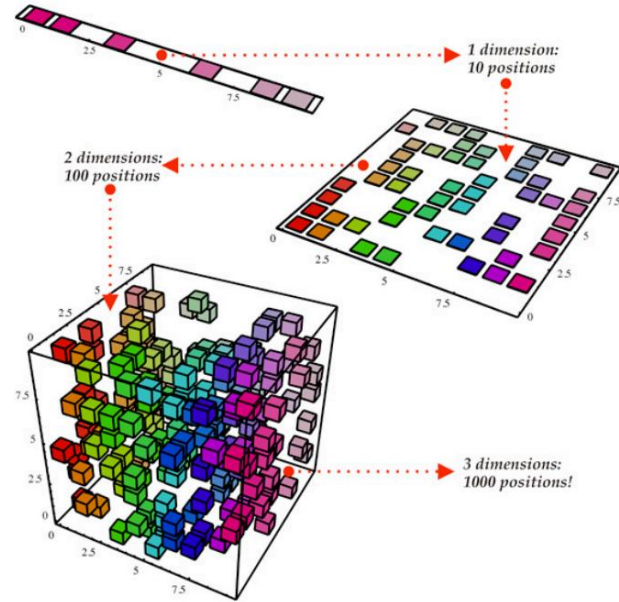
Где често виђамо овакву апроксимацију?

- Када гледамо филм, то је 2Д приказ 3Д објеката (на страну 3Д наочаре).
- Ми немамо никаквих проблема да разликујемо 3Д објекте у 2Д простору филмског платна.
- Изгледа да та трећа димензија често не даје претерано значајних информација.
- Филмска камера узима 3Д информацију и „равна је“ у 2Д информацију без претераних губитака.



Какве везе има ово са **PCA**?

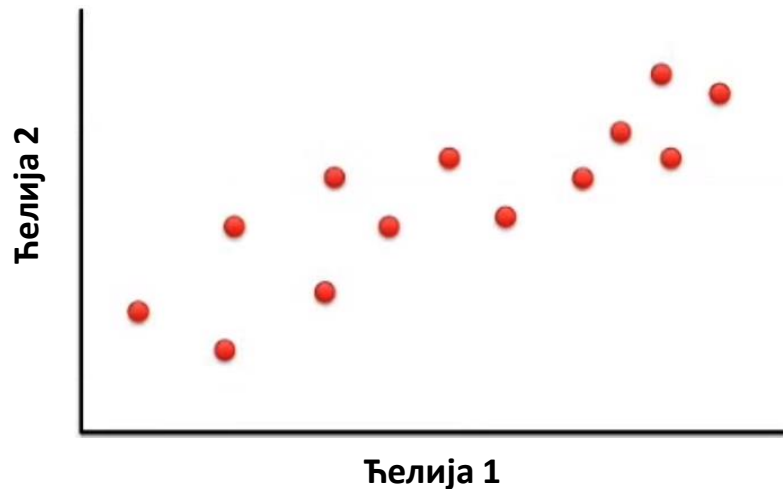
- PCA узима податке са пуно димензија (тј. са великим бројем ћелија) и представља их у 2Д или 3Д простору (да би ми могли да их видимо и схватимо).
- PCA то ради тако што се фокусира само на оне особине по којима се ћелије међусобно разликују.



Пример рада *РСА*

- Узмимо податке за две ћелије, тј. вредности експресије гена за сваку од њих:

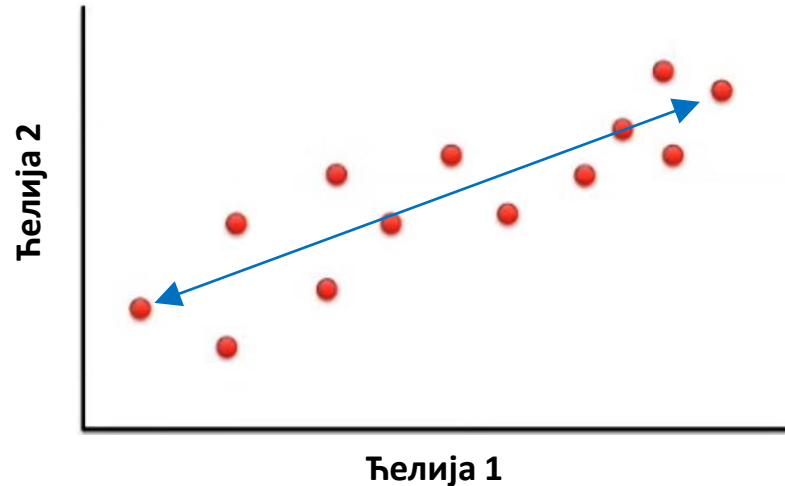
Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...



Пример рада *РСА*

- Узмимо податке за две ћелије, тј. вредности експресије гена за сваку од њих:

Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...

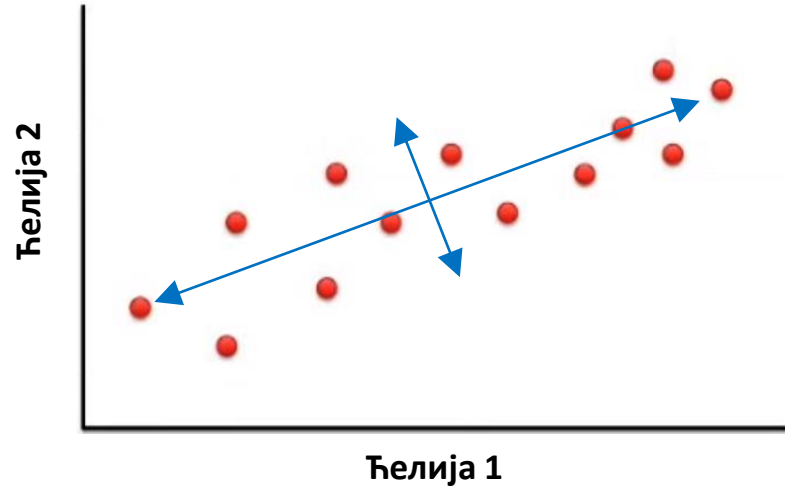


- Можемо приметити да је, генерално, распоред тачака (гена) дуж дијагонале.
- Значи, максимална разлика између вредности података за ове 2 ћелије налази се између два краја ове линије.

Пример рада *РСА*

- Узмимо податке за две ћелије, тј. вредности експресије гена за сваку од њих:

Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...

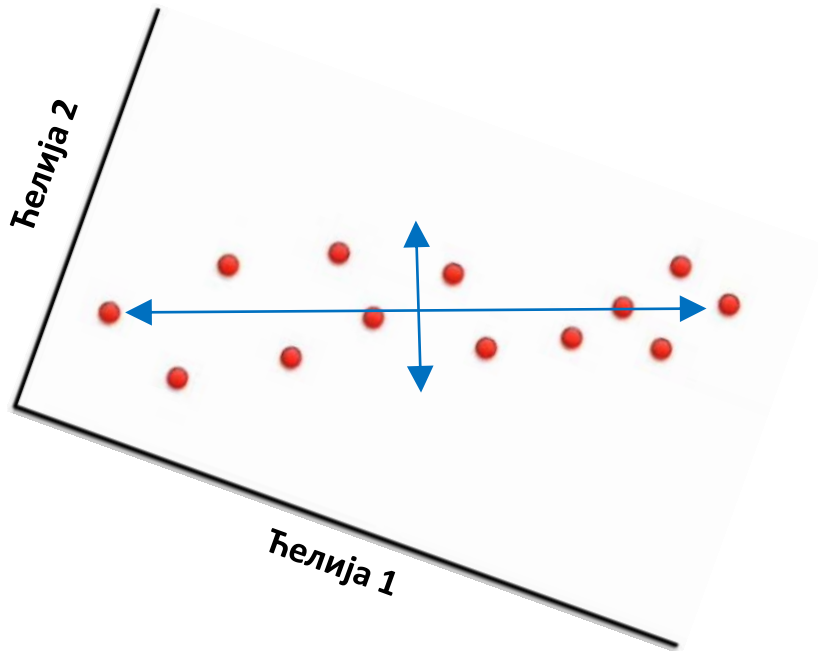


- Са друге стране, расподела тачака (тј. разлика између њих) постоји и дуж друге димензије, тј. изнад и испод прве линије.
- Значи, максимална разлика између вредности података за ове 2 ћелије такође се налази и између два краја нове линије.

Пример рада *PCA*

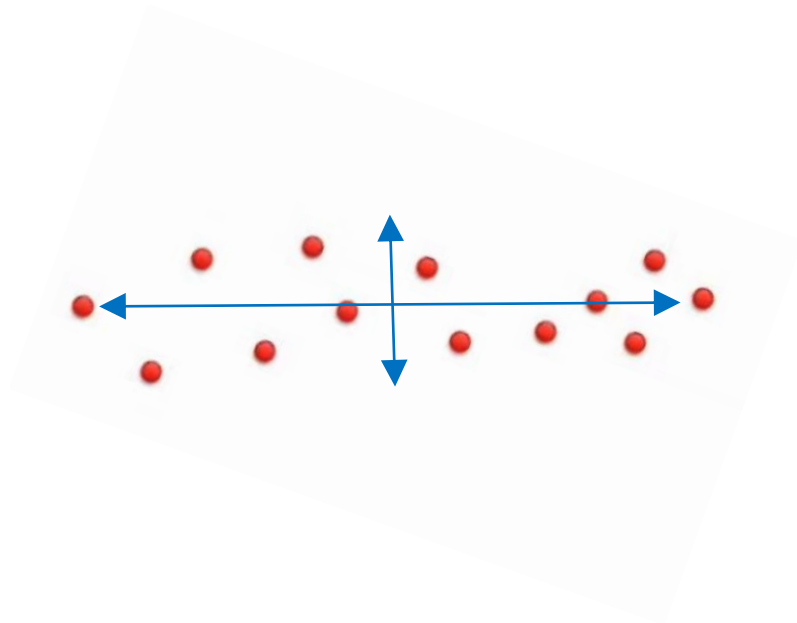
- Узмимо податке за две ћелије, тј. вредности експресије гена за сваку од њих:

Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...

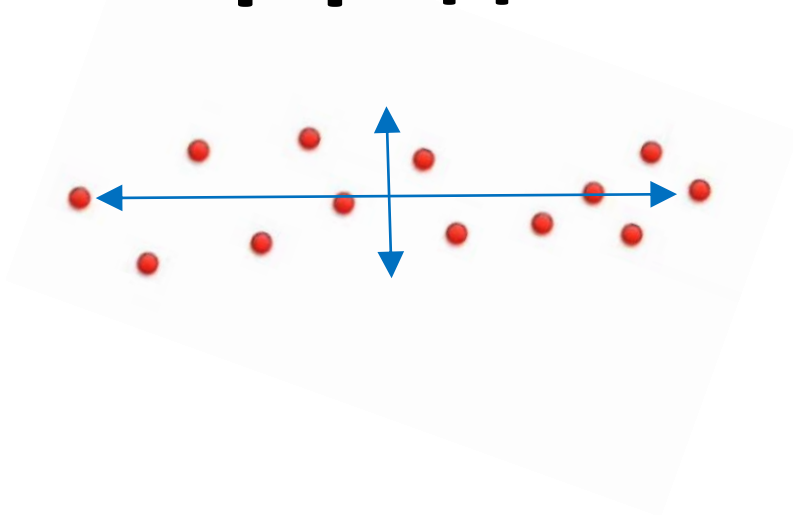


- Ако заротирамо график, ове две линије (дуж којих је разлика између тачака највећа) постаће нове X и Y осе.
- На овај начин, лакше ћемо нумерички изразити ове разлике.

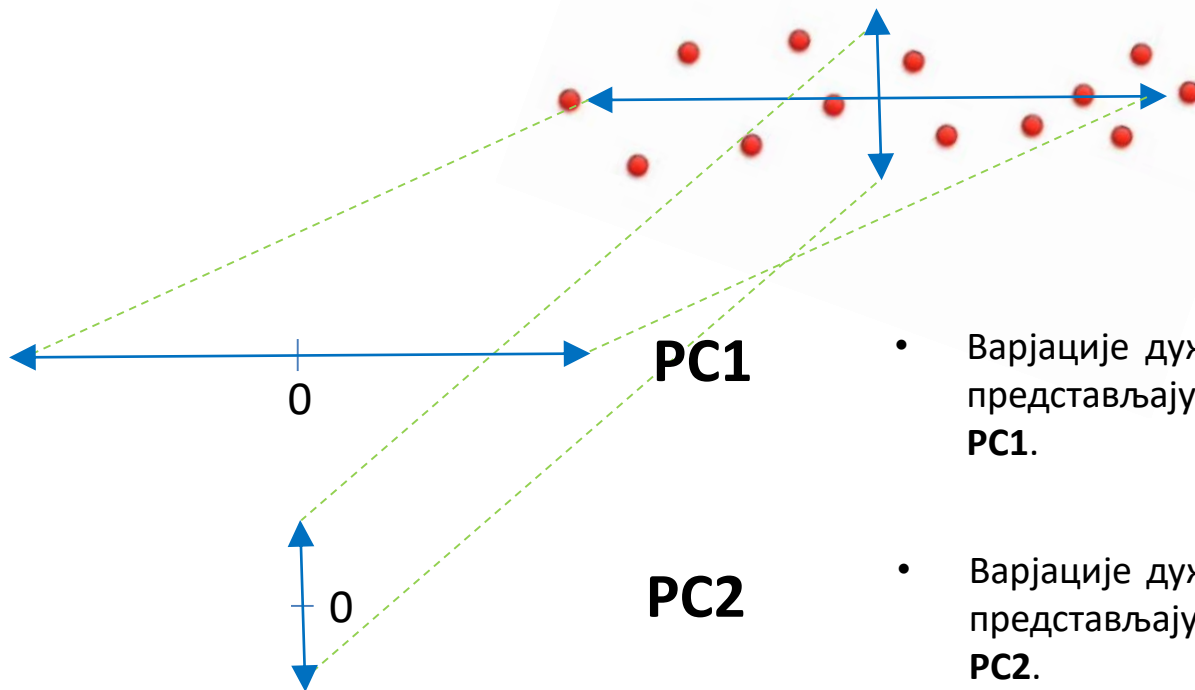
Пример ряда $РСА$



Пример ряда $РСА$

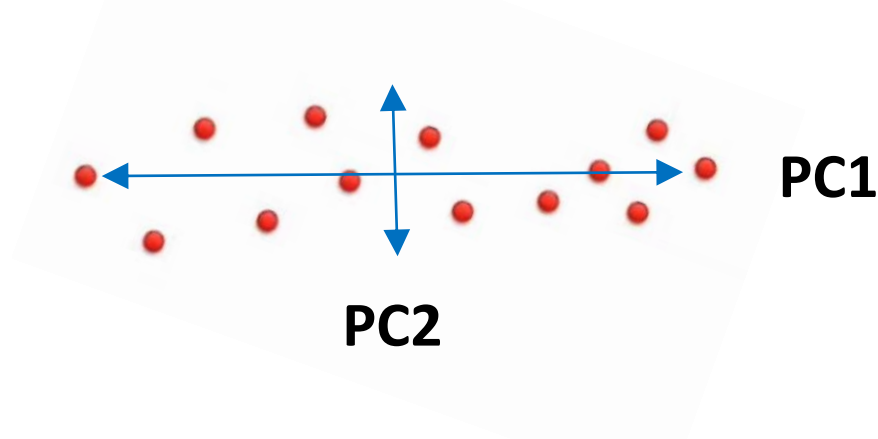


Пример рада *РСА*



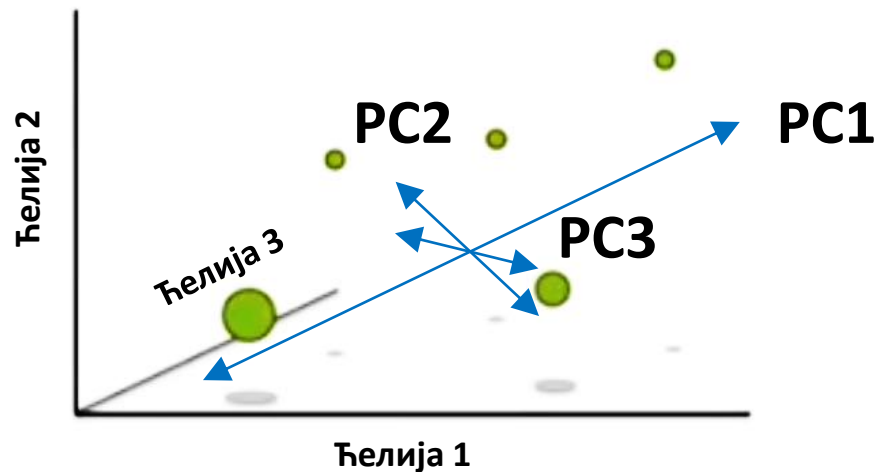
- Варјације дуж ове нове "X" осе у ствари представљају прву "главну" компоненту **PC1**.
- Варјације дуж ове нове "Y" осе у ствари представљају другу "главну" компоненту **PC2**.

Пример рада *РСА*



- **PC1** (првна главна компонента) очигледно представља најзначајнију варијацију међу тачкама.
- **PC2** (друга главна компонента) представља другу по значају варијацију међу тачкама.

Шта би било да имамо 3 ћелије?

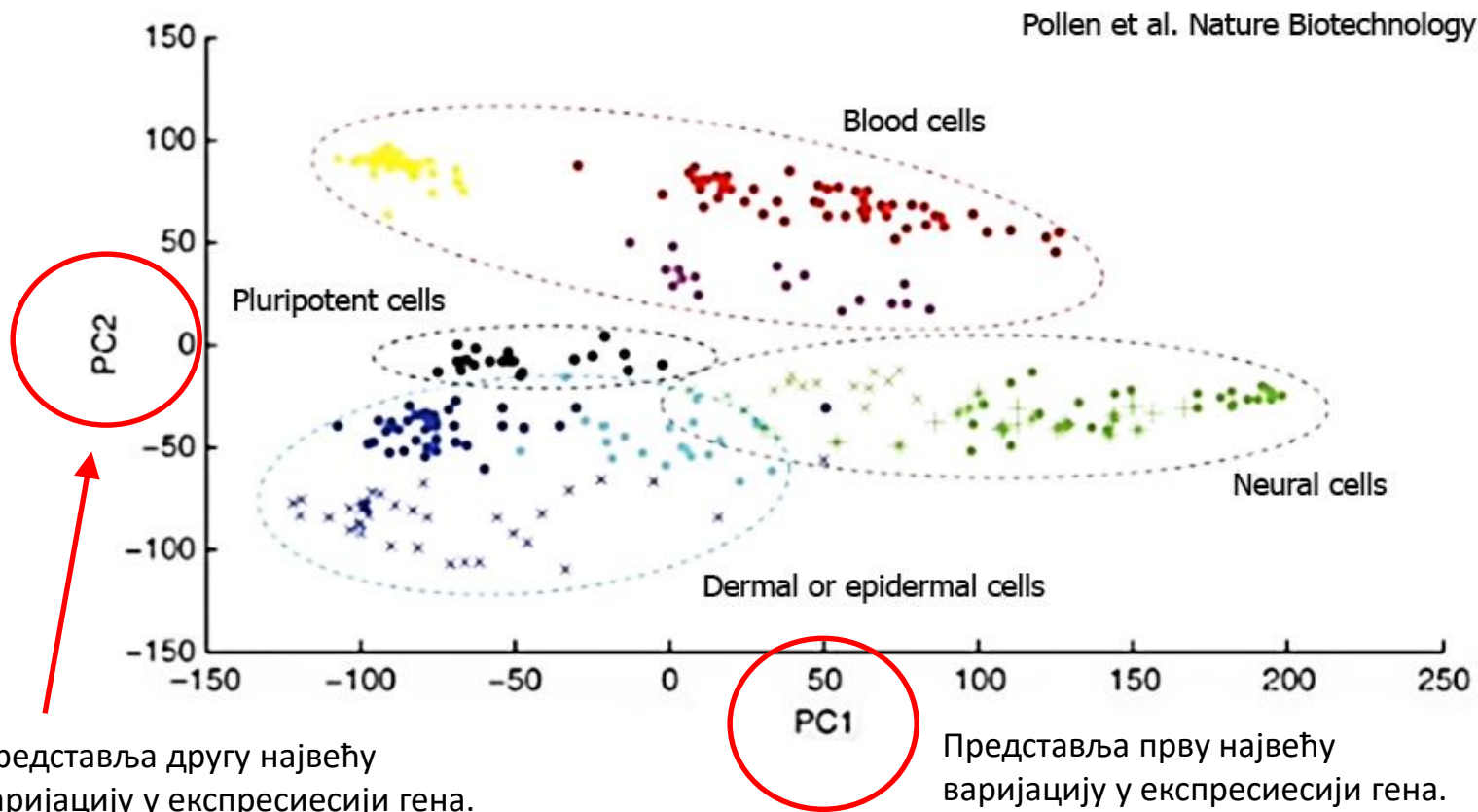


- **PC1** нам показује смер где је најзначајнија варијација.
- **PC2** нам показује смер где је варијација друга по значају.
- **PC3** нам показује смер где је варијација трећа по значају.

Шта би било да имамо 4 ћелије?

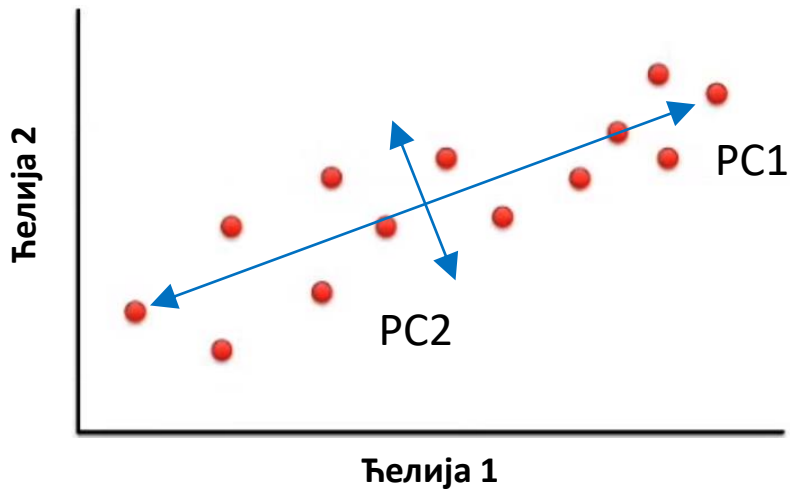
- **PC1** нам показује смер где је најзначајнија варијација.
 - **PC2** нам показује смер где је варијација друга по значају.
 - **PC3** нам показује смер где је варијација трећа по значају.
 - **PC4** нам показује смер где је варијација четврта по значају.
-
- Значи, постоји по једна главна компонента за сваку димензију (тј. ћелију).
 - Да имамо 200 ћелија, имали би 200 главних компонената.

Вратимо се на први график:



Али ово је слика ћелија а не гена!

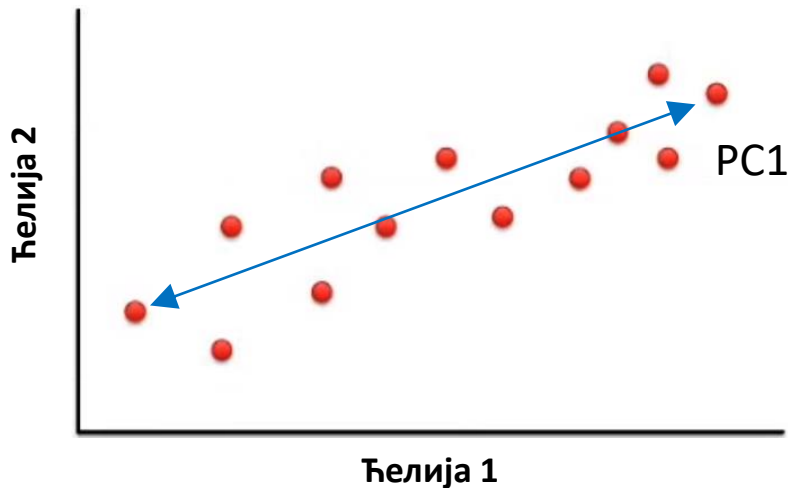
- Како помоћу мање или веће експресије гена можемо да цртамо график приказа ћелија? Зар све време не говоримо о експресији појединих гена?



- Вратимо се на случај са 2 ћелије.
- Задржимо се само на PC1.

Али ово је слика ћелија а не гена!

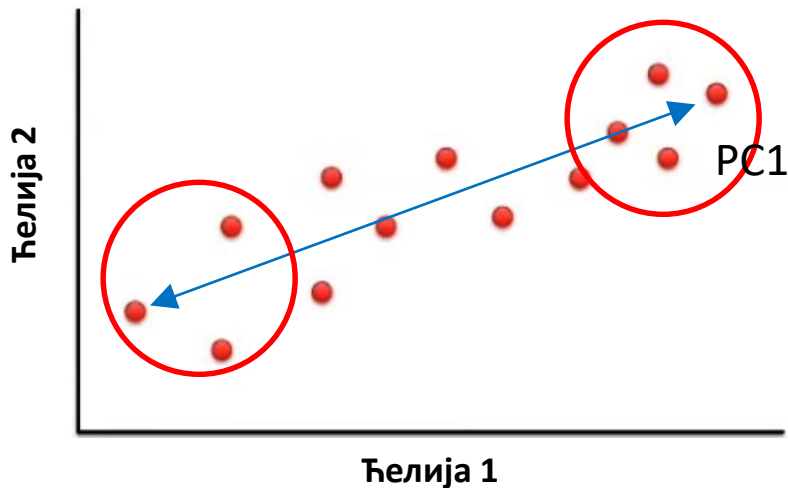
- Како помоћу мање или веће експресије гена можемо да цртамо график приказа ћелија? Зар све време не говоримо о експресији појединих гена?



- Вратимо се на случај са 2 ћелије.
- Задржимо се само на PC1.

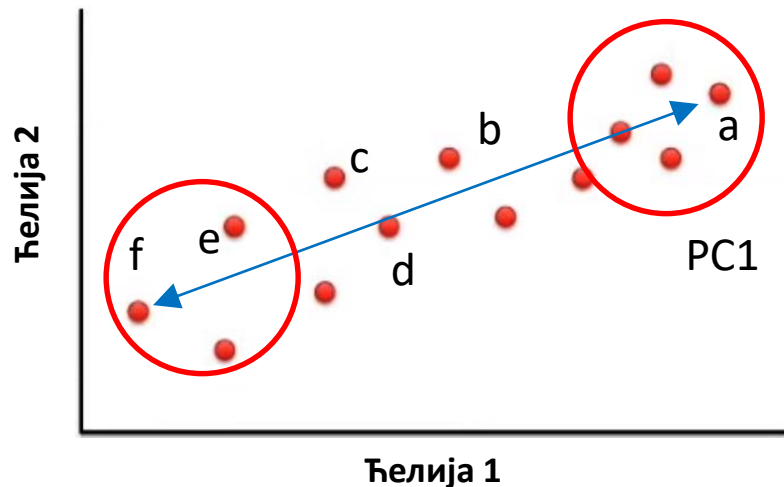
Али ово је слика ћелија а не гена!

- Како помоћу мање или веће експресије гена можемо да цртамо график приказа ћелија? Зар све време не говоримо о експресији појединих гена?



- Вратимо се на случај са 2 ћелије.
- Задржимо се само на PC1.
- Највећи утицај на разлике у PC1 имаће гени који се налазе у оквиру кругова.

Селекција редоследа гена по важности

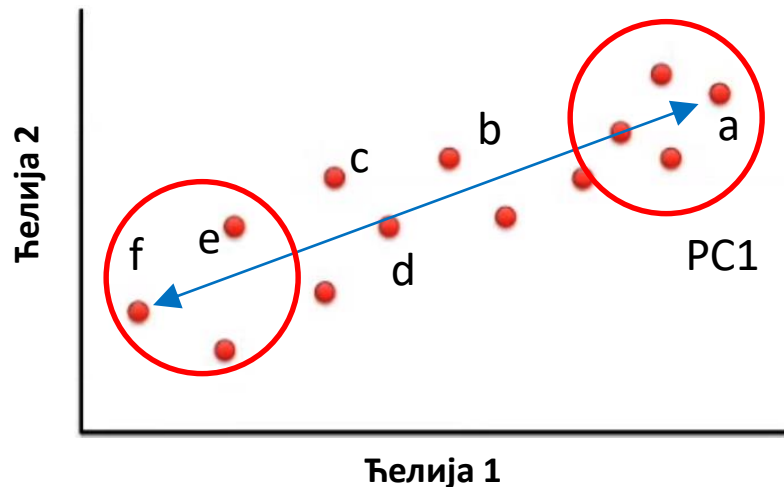


- Направимо селекцију редоследа гена по важности, тј. како утичу на разлике компоненте PC1.

Назив гена	Утицај на PC1
a	Велики
b	Мали
c	Мали
d	Мали
e	Велики
f	Велики

- Значи, гени ближе крајевима линије имају већи утицај од оних близу средини линије.

Селекција редоследа гена по важности

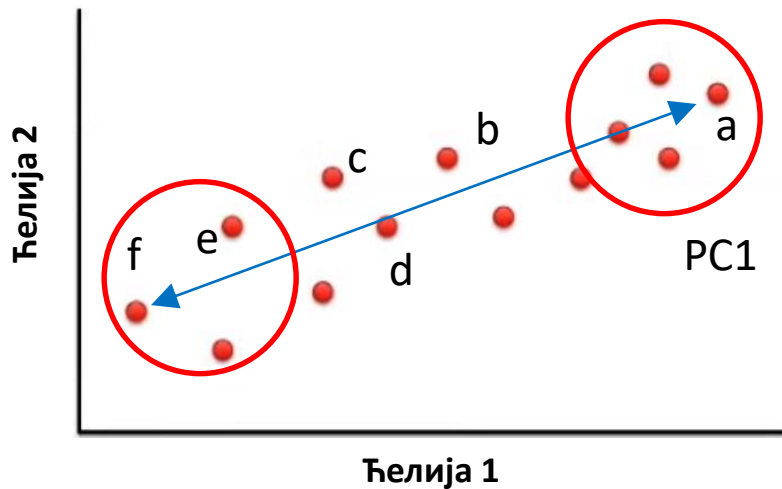


- Значи, гени ближе крајевима линије имају већи утицај од оних близу средини линије.
- То можемо приказати и нумерички.

- Направимо селекцију редоследа гена по важности, тј. како утичу на разлике компоненте PC1.

Назив гена	Утицај на PC1	Нумерички
a	Велики	10
b	Мали	0.5
c	Мали	0.2
d	Мали	-0.2
e	Велики	13
f	Велики	-14

Селекција редоследа гена по важности

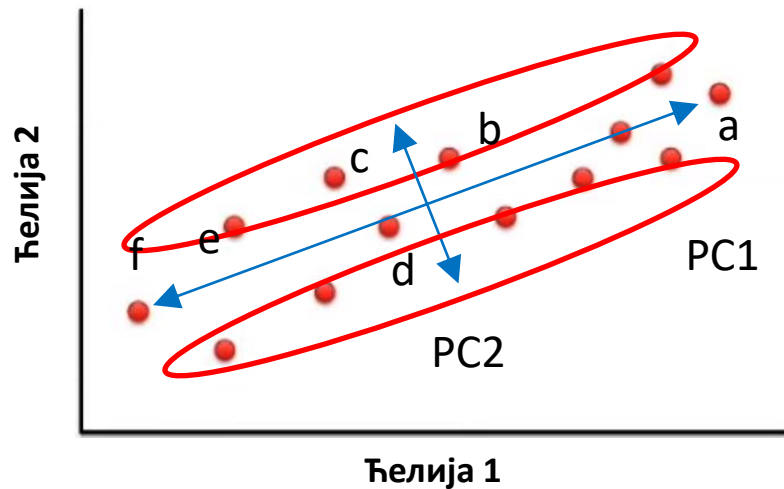


- Разлика да ли је нумеричка вредност плус или минус зависи од тога са које стране линије се гени налазе.

- Видећемо да ће гени који немају пуно утицаја на PC1 имати нумеричку вредност блиску 0, а они који имају утицај ће бити значајно различити од 0.

Назив гена	Утицај на PC1	Нумерички
a	Велики	10
b	Мали	0.5
c	Мали	0.2
d	Мали	-0.2
e	Велики	13
f	Велики	-14

Селекција редоследа гена по важности



- Сличан случај имамо и за PC2.

Назив гена	Утицај на PC1	Нумерички
a	Велики	10
b	Мали	0.5
c	Мали	0.2
d	Мали	-0.2
e	Велики	13
f	Велики	-14

Нумеричко поређење PC1 и PC2

Назив гена	Утицај на PC1	Нумерички	Назив гена	Утицај на PC2	Нумерички
a	Велики	10	a	Средњи	3
b	Мали	0.5	b	Велики	10
c	Мали	0.2	c	Велики	8
d	Мали	-0.2	d	Велики	-12
e	Велики	13	e	Мали	0.2
f	Велики	-14	f	Мали	-0.1

Сада, пошто имамо ове вредности „значајности“ сваког гена за сваку ћелију, то можемо искористити за добијање јединственог броја који ће представљати сваку ћелију 2Д координатном систему. То се ради овако:

Нумеричко поређење PC1 и PC2

Назив гена	Утицај на PC1	Нумерички	Назив гена	Утицај на PC2	Нумерички
a	Велики	10	a	Средњи	3
b	Мали	0.5	b	Велики	10
c	Мали	0.2	c	Велики	8
d	Мали	-0.2	d	Велики	-12
e	Велики	13	e	Мали	0.2
f	Велики	-14	f	Мали	-0.1

Сада, пошто имамо ове вредности „значајности“ сваког гена за сваку ћелију, то можемо искористити за добијање јединственог броја који ће представљати сваку ћелију 2Д координатном систему. То се ради овако:

Комбинујемо бројевну вредност експресије сваког гена са „утицајношћу“ тих гена на PC1 и PC2.

Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...

Назив гена	Утицај на PC1	Нумерички
a	Велики	10
b	Мали	0.5
c	Мали	0.2
d	Мали	-0.2
e	Велики	13
f	Велики	-14

Назив гена	Утицај на PC2	Нумерички
a	Средњи	3
b	Велики	10
c	Велики	8
d	Велики	-12
e	Мали	0.2
f	Мали	-0.1

PC1 скор за ћелију 1 = (експресија гена “a” * утицајност гена “a”) +
(експресија гена “b” * утицајност гена “b”) + ... + и тако за све гене

Комбинујемо бројевну вредност експресије сваког гена са „утицајношћу“ тих гена на PC1 и PC2 (утицајност = тежина).

Ген	Ћелија 1	Ћелија 2
a	10	8
b	0	2
c	14	10
d	33	45
e	50	42
f	80	72
g	95	90
h	44	50
i	60	50
...	...	...

Назив гена	Утицај на PC1	Нумерички
a	Велики	10
b	Мали	0.5
c	Мали	0.2
d	Мали	-0.2
e	Велики	13
f	Велики	-14

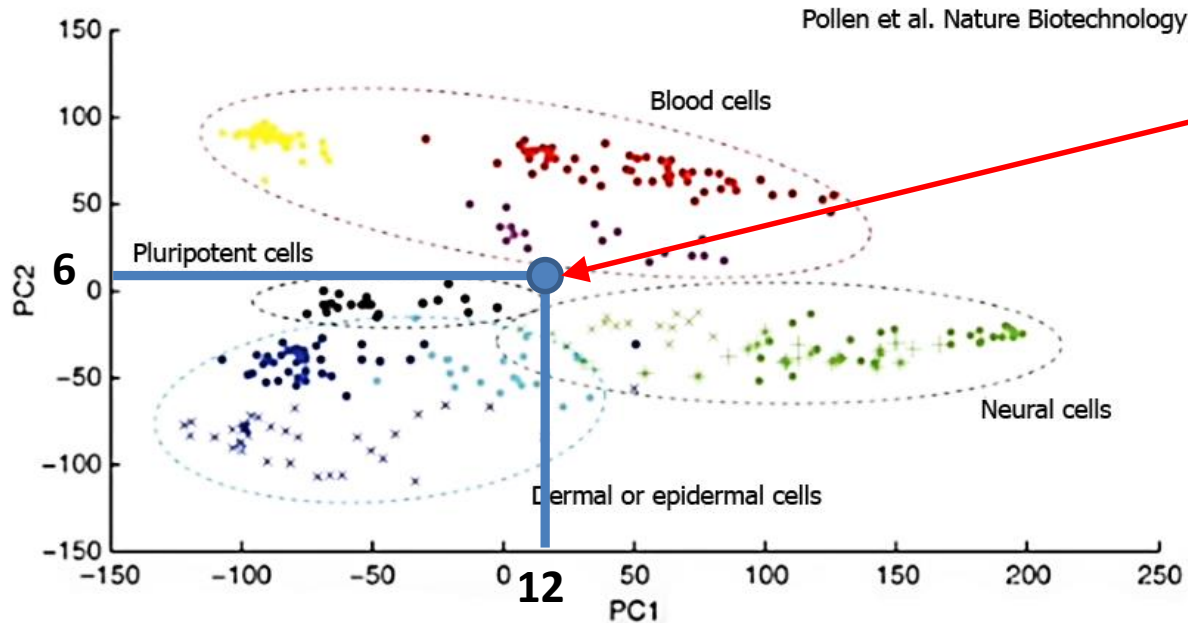
Назив гена	Утицај на PC2	Нумерички
a	Средњи	3
b	Велики	10
c	Велики	8
d	Велики	-12
e	Мали	0.2
f	Мали	-0.1

PC2 скор за ћелију 1 = (експресија гена “a” * утицајност гена “a”) +
(експресија гена “b” * утицајност гена “b”) + ... + и тако за све гене

Рецимо, да смо добили да су вредности PC1 и PC2 компоненете за ћелију 1:

PC1 скор за ћелију 1 = $(10 \cdot 10) + (0 \cdot 0.5) + \dots$ итд. = 12

PC2 скор за ћелију 1 = $(10 \cdot 3) + (0 \cdot 10) + \dots$ итд. = 6

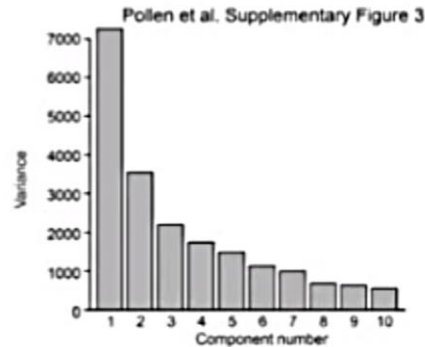


Ћелија 1

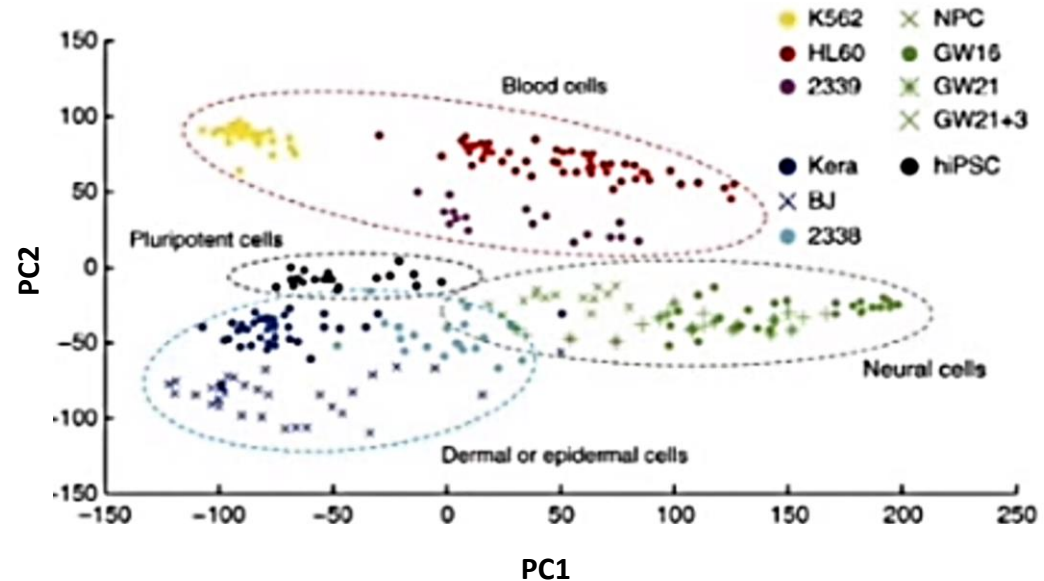
... и тако за за
све ћелије ...

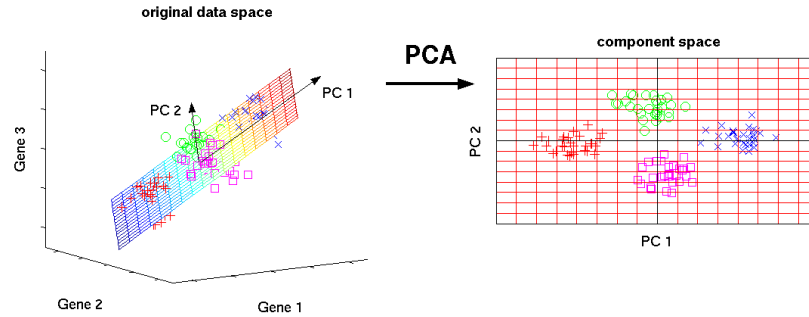
Ћелије са сличним транскрипционом факторима ће се груписати заједно !!!

- Овај график сада може да се искористи да би идентификовали кључни гени!
- Можемо приметити де се ћелије групишу лево-десно и горе-доле.
- Ако хоћемо да сазнамо који гени су одговорни за то што су епидермалне ћелије груписане са леве стране а нервне ћелије са десне стране, само треба да погледамо скорове „утицајности“ појединих гена за PC1, тј. који гени имају највећи PC1 скор.
- Ако хоћемо да видимо који гени су одговорни за то што се дермалне ћелије разликују од крвних ћелија, само треба да погледамо скорове утицајности одређених гена за PC2 тј. који гени имају највећи PC2 скор.



Можемо видети да највећи допринос у добијању разлике између појединих група добијамо помоћу 2 главне компоненте: PC1 и PC2.





PCA и MATLAB

Како можемо користити MATLAB да
би радили PCA анализу

Анализа експресије гена

- MATLAB се лако може користити за PCA анализу.
- Најбоље је то показати на конкретном примеру који се налази на:

<http://www.mathworks.com/help/bioinfo/examples/gene-expression-profile-analysis.html>

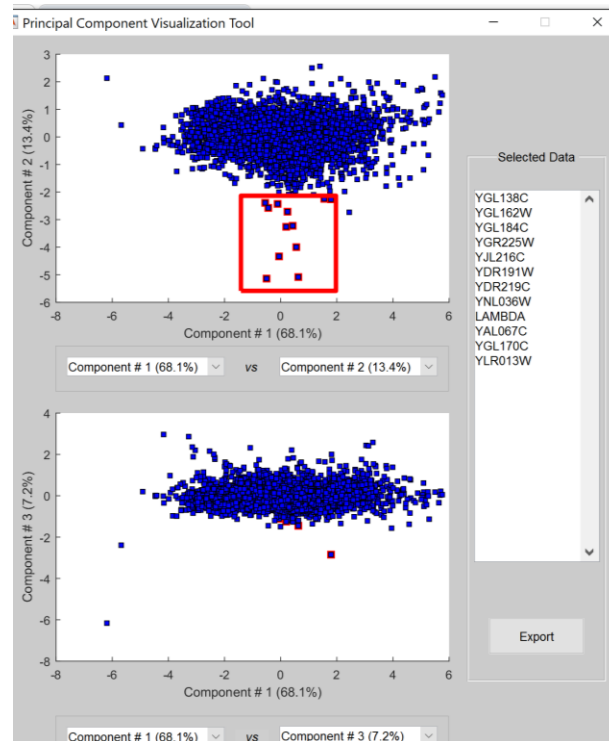
- У командном прозору уписати:
>> load yeastdata.mat

Анализа експресије гена

- У овом примеру имамо вредности експресије 6400 гена у 7 временских интервала, као и називе тих гена.
- Да би применили PCA анализу у командном прозору само треба уписати:

```
>> mapcaplot(yeastvalues,genes)
```

- MATLAB ће урадити аутоматску кластеризацију тј. Груписање по жељеним главним компонентама.
- На нама је само да одаберемо (селекујемо) одређене групе, а MATLAB ће нам у десном прозору аутоматски приказати који гени су најодговорнији за такву кластеризацију.
- За цртање главних компоненти можемо корисити и „scatter“ функцију.



Коришћење „scatter“ функције за PCA

```
>> [pc, zscores, pcvars] = pca(yeastvalues); %radi pca i razbija promenljivu „yeastvalues“ na više njih
>> figure
>> scatter(zscores(:,1),zscores(:,2));
>> xlabel('First Principal Component');
>> ylabel('Second Principal Component');
>> title('Principal Component Scatter Plot');
```

- Или можемо обојити различите кластере користити статистичке алате као што је „gscatter“.

```
>> figure
pcclusters = clusterdata(zscores(:,1:2),'maxclust',8,'linkage','av');
gscatter(zscores(:,1),zscores(:,2),pcclusters)
xlabel('First Principal Component');
ylabel('Second Principal Component');
title('Principal Component Scatter Plot with Colored Clusters');
```

- Уколико желимо идентификацију гена, можемо користити команду „gname“.

```
>> gname(genes)
```