Text-Processing

Text-Processing

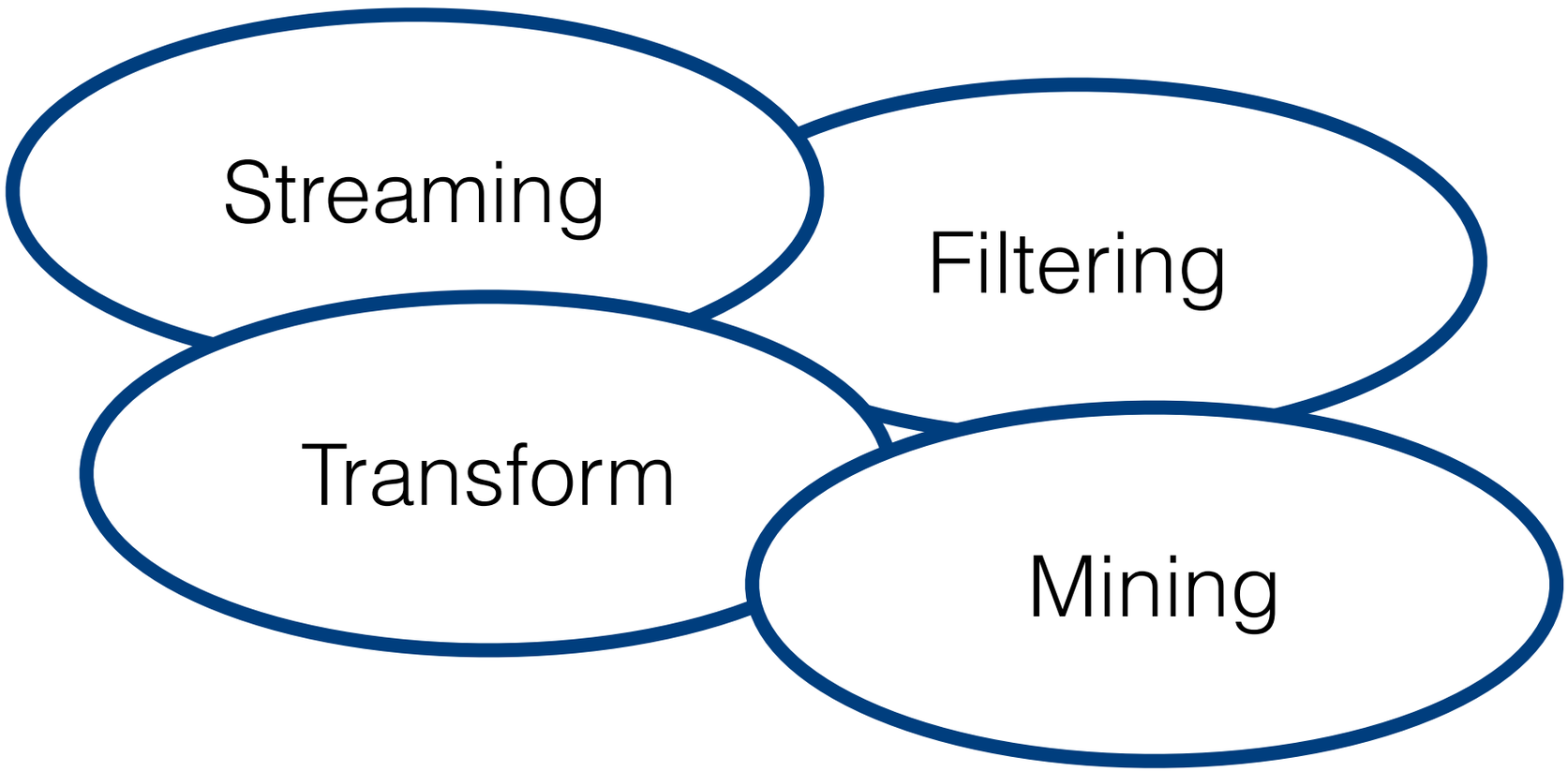
from the
Linux Command-Line

Agenda

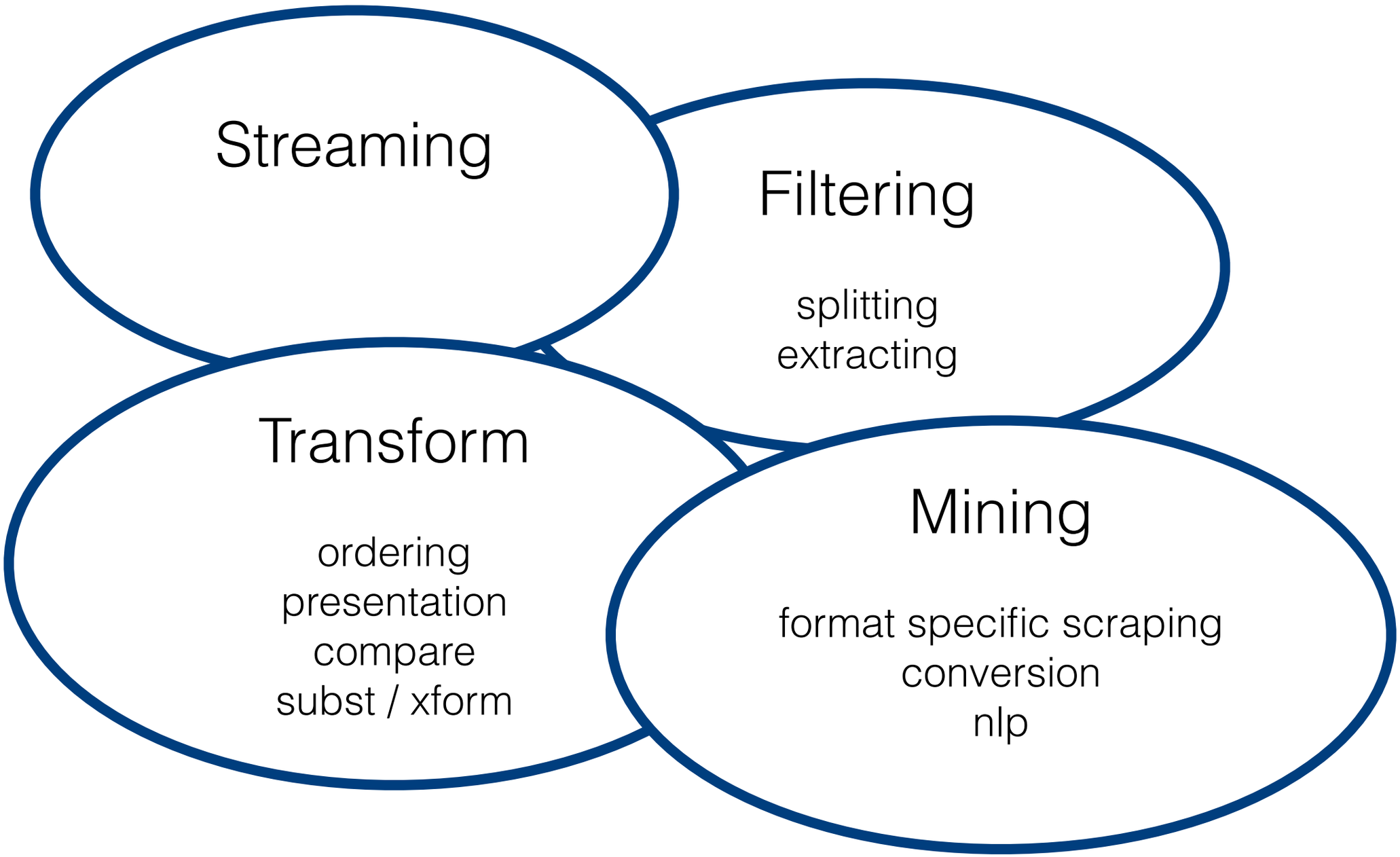
Examine Some Common Tools

- awk
- cat
- csplit
- cut
- echo
- grep
- head
- join
- paste
- perl
- pdftotext
- python
- sed
- sort
- tail
- tee
- tr
- uniq
- xls2csv / xlsx2csv
- …more…

Standard Tools



Standard Tools



Streaming

- **echo** - output strings
- **cat** - output or concatenate files.
- **tee** - stdin to file and stdout

Filtering

Transform

Mining

echo

echo "Penguins are great!\nLinux too."

Penguins are great!\nLinux too. \n

---

echo -n "Penguins are great!\nLinux too."

Penguins are great!\nLinux too.

---

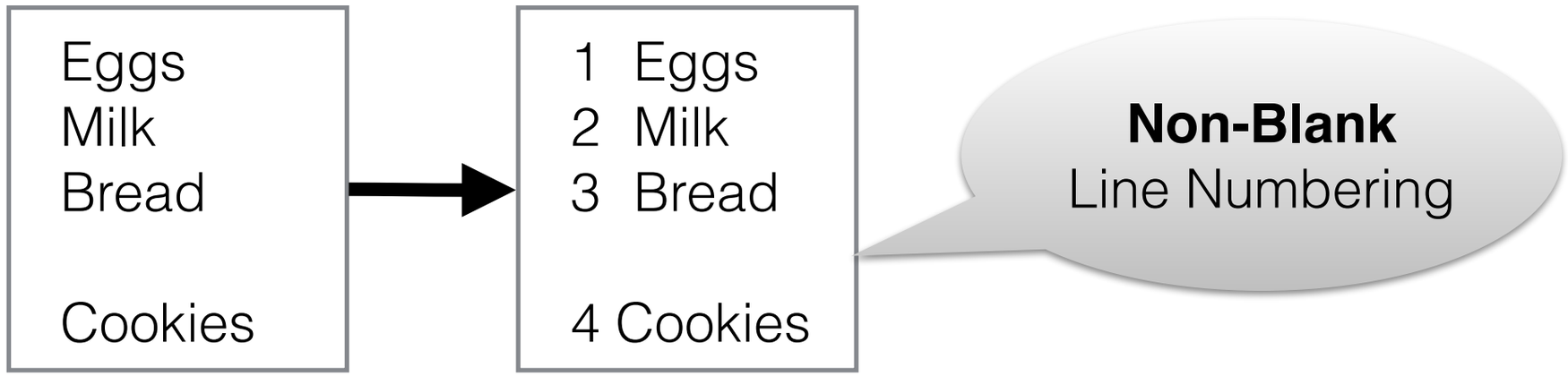
echo -e "Penguins are great!\nLinux too."

Penguins are great! \n
Linux too. \n

cat

Write contents of file to stdout

- cat -n GroceryList.txt



cat

Write contents of file to stdout

- cat -v BabyTux.png

  Replace control characters with ^x (caret + letter)
  Replace high order bytes with M-x.

Display non-printable

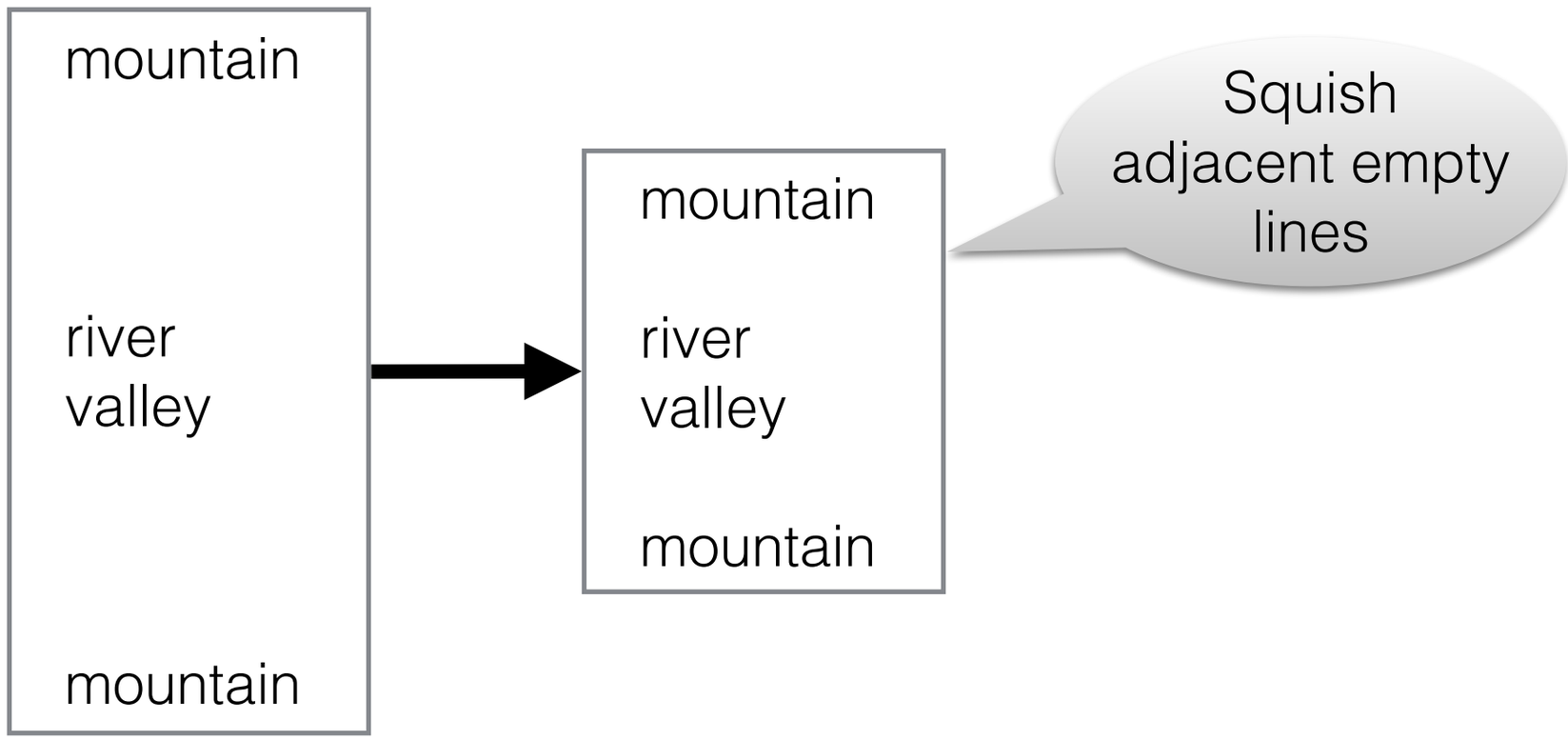
```
M-^IPNG^M
^Z
^@^@^@^MIHDR^@^@^C ^@^@^C ^H^B^@^@^@T^RM-^Q?^@^@^@      pHYs^
M-^?^@M-^?M- M-=M-'M-^S^@^B>^BIDATxM-ZM-lM-]yM-^LM-^UUM-^ZM-G
-vM-^HbM-bM-^VLM-[FM-QM-FmtM-|GM\M-^Z?$.M-]M-ZM-.M-#&80tM-7^T
Alp!M-JbM-1M-UM-^YM-^WM-:M-wM->M-oM-s<M-gM-^\M-wM-^r^EM-|~^R^
E^@^@@M-@^B^@^@ `^A^@^@M-^@M-^@^E^@^@@M-@^B^@^@ `^A^@^@M-^@M-
@^@^@@M-@^B^@^@ `^A^@^@^PM-0^@^@^@@M-@^B^@^@ `^A^@^@^PM-0^@^@
@^@^@ `^A^@^@^PM-0^@^@^@^HX^@^@^@D,^@^@^@^PM-0^@^@^@^HX^@^@^
,^@^@^@^B^V^@^@^@^HX^@^@^@^D,^@^@^@^B^V^@^@^@^HX^@^@^@^D,^@^@
^V^@^@^@^A^K^@^@^@^D,^@^@^@^B^V^@^@^@^A^K^@^@^@^D,^@^@^@^B^V^
E^@^@^@^B^V^@^@^@^A^K^@^@M-^@M-^@^E^@^@^@^B^V^@^@^@^A^K^@^@M-
^@^A^K^@^@M-^@M-^@^E^@^@@M-@^B^@^@^@^A^K^@^@M-^@M-^@^E^@^@@M-
--\M-9rM-mM-^Z5M-^_ww^?M-zM-iM-'M-kM-WM-/M-oM-WM-OoM-\M-8M-)M
-?~M-E^GM-^KM-?^O^[6M-,M-xM-KM-PM-!CM-yM-v^B `^AM-X1uwwM-/ZM-
M-|M-_o}QM-jM-=PM-cuIM-|=M-4>M-d^[M-/ZM-!M-uYM-WM-z{M-oWxM-_M
-nM-9M-SN;M-qM-#^A@M-@^BM-0^](M-*P^_|M-pM-AM-^R%KM-^}M-wM-]"Q
RM-.M-zM- kM-F$M-qM-^YfM-@*_M-%M-duM-(^WM-2V"sM-fJCM-s"M-rM-j
-xM-cM-^O^W-Z4^?M-~M-[^?M-yM-KM-;sM-gM-NYM-=zuM-q^QM-^QM-^MM-
M-;M-l<vM-^_M-1M-^G^\zM-hM-8M--^N^Z?M-~^GM-E:M-#M-W7^A^@^D,^@
LjM-e%M-_M-<M-^RM-rM-^UM-),wM-^U^_M-O][T^[M-+"M-^W-z^E^QM-BM-
M-oM-^?M-^zM-k^?M-gM-LM-^YM-3`M-AM-^BM-nM-nM-O{kRM->M-^Y_M-"J
^_^?M-TQGMM-^Z4M-iM-pM-C^O^_1b^DM-?^X^@^HX^@M-zM-&M-XM-G7^?M-
-J[M-a*^]M-`z3^'M-^QM-+*e%M-^[M-7M-JM-;dM-.M--\M-7M-^T^OM-#M-
```

* Thanks to Dan for mentioning this option

cat

Write contents of file to stdout

- cat -s Spacy.txt

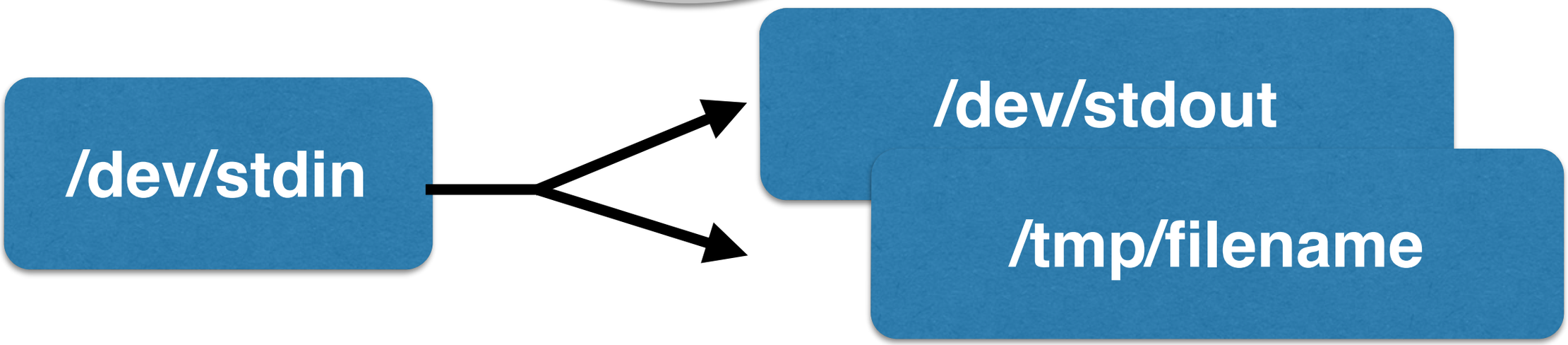


tee

tee /tmp/filename
tee -a /tmp/filename

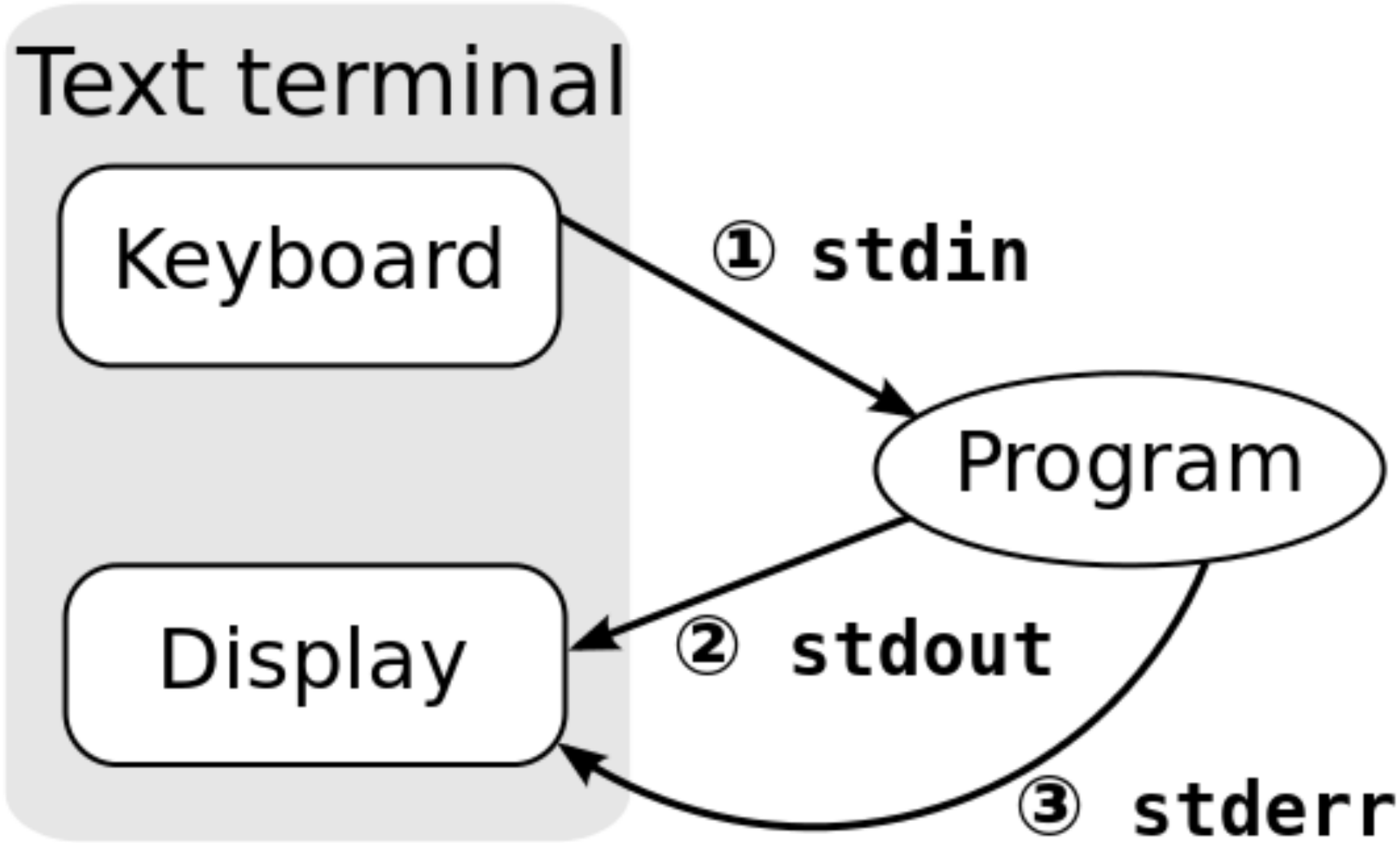
Beware
truncates without
warning

append



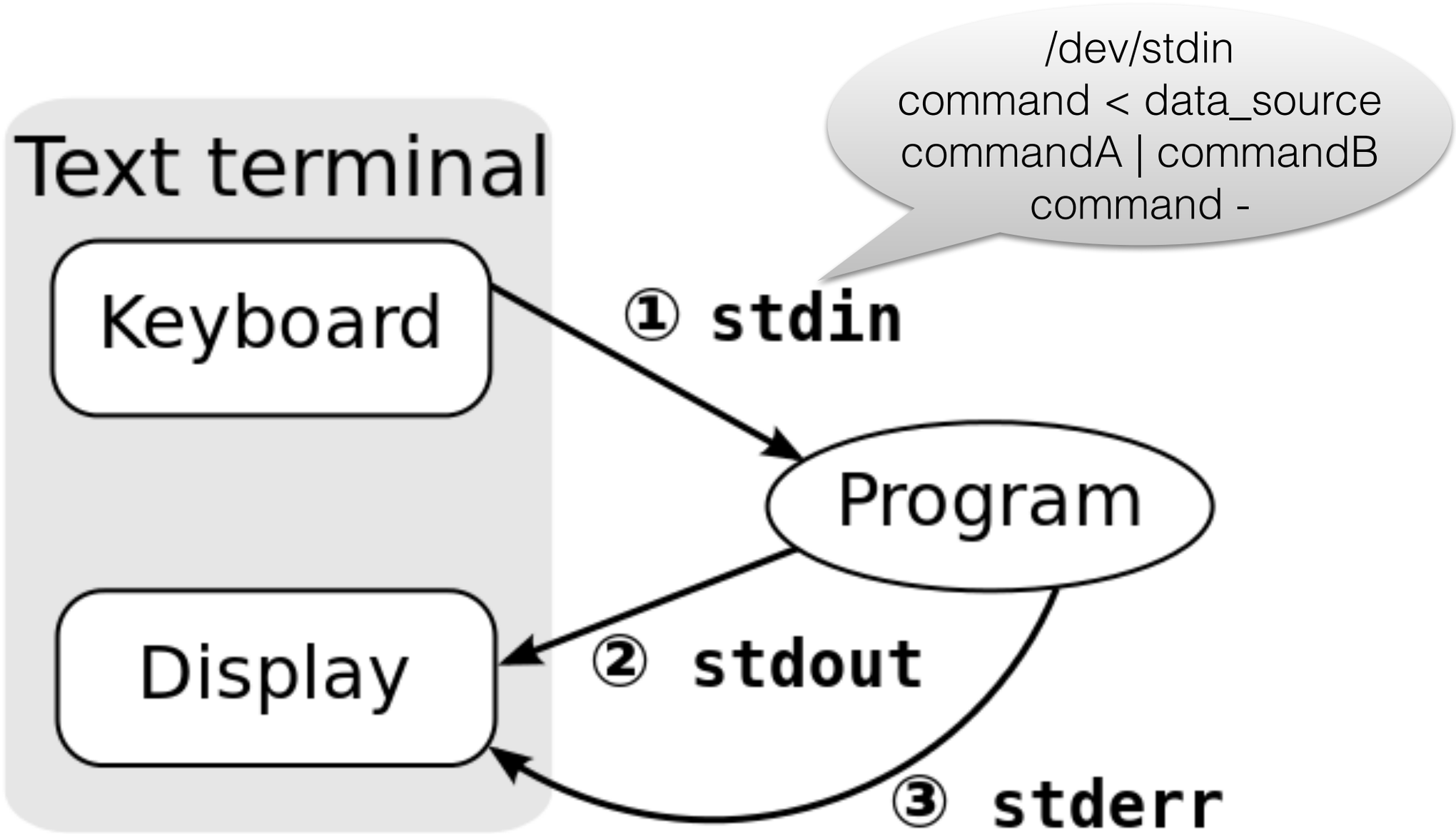

```
while (read STDIN) {
    write to STDOUT;
    write to FILE;
}
```

ASIDE: Standard Streams



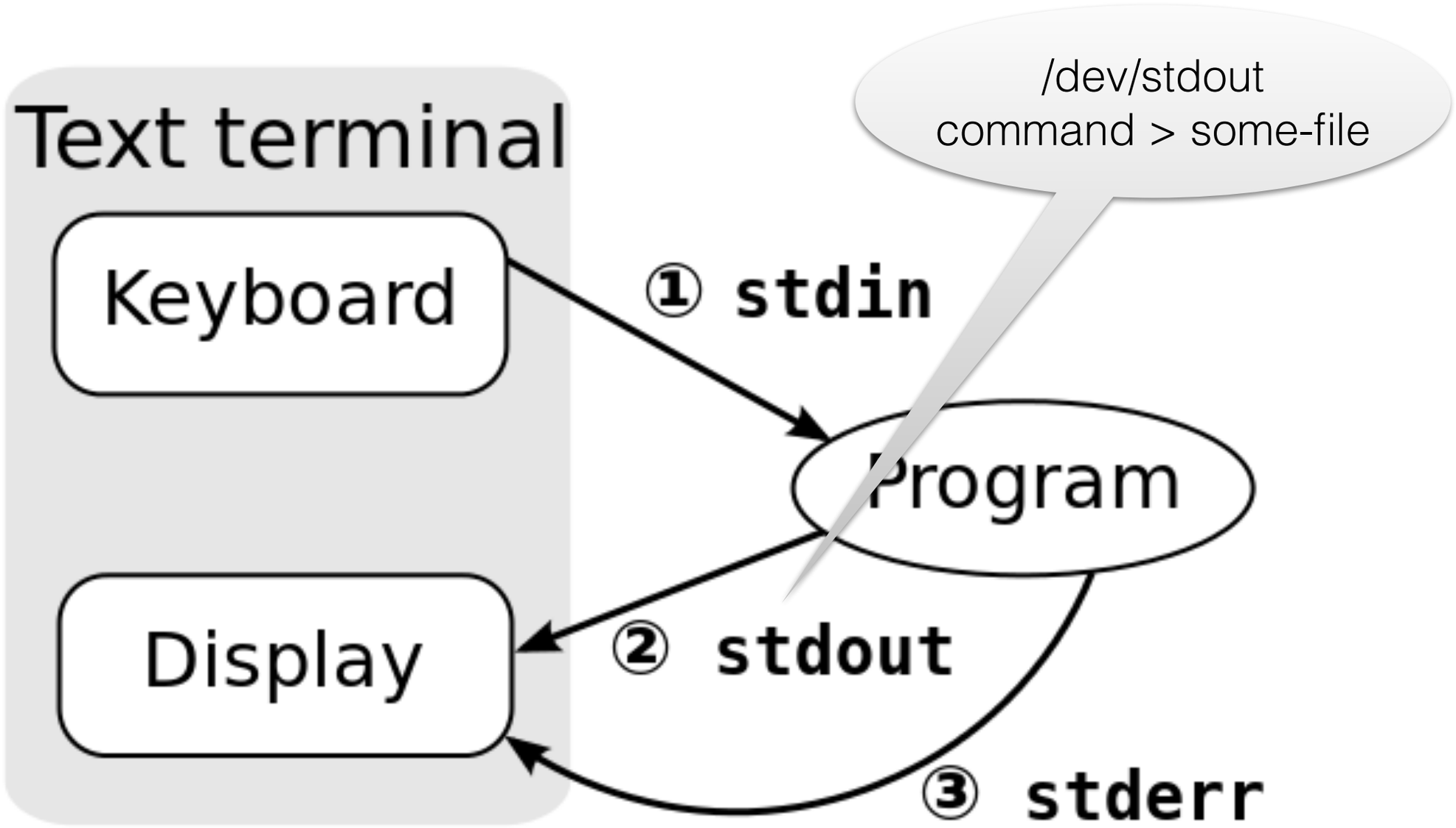
Source:
https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/Stdstreams-notitle.svg/535px-Stdstreams-notitle.svg.png

ASIDE: Standard Streams



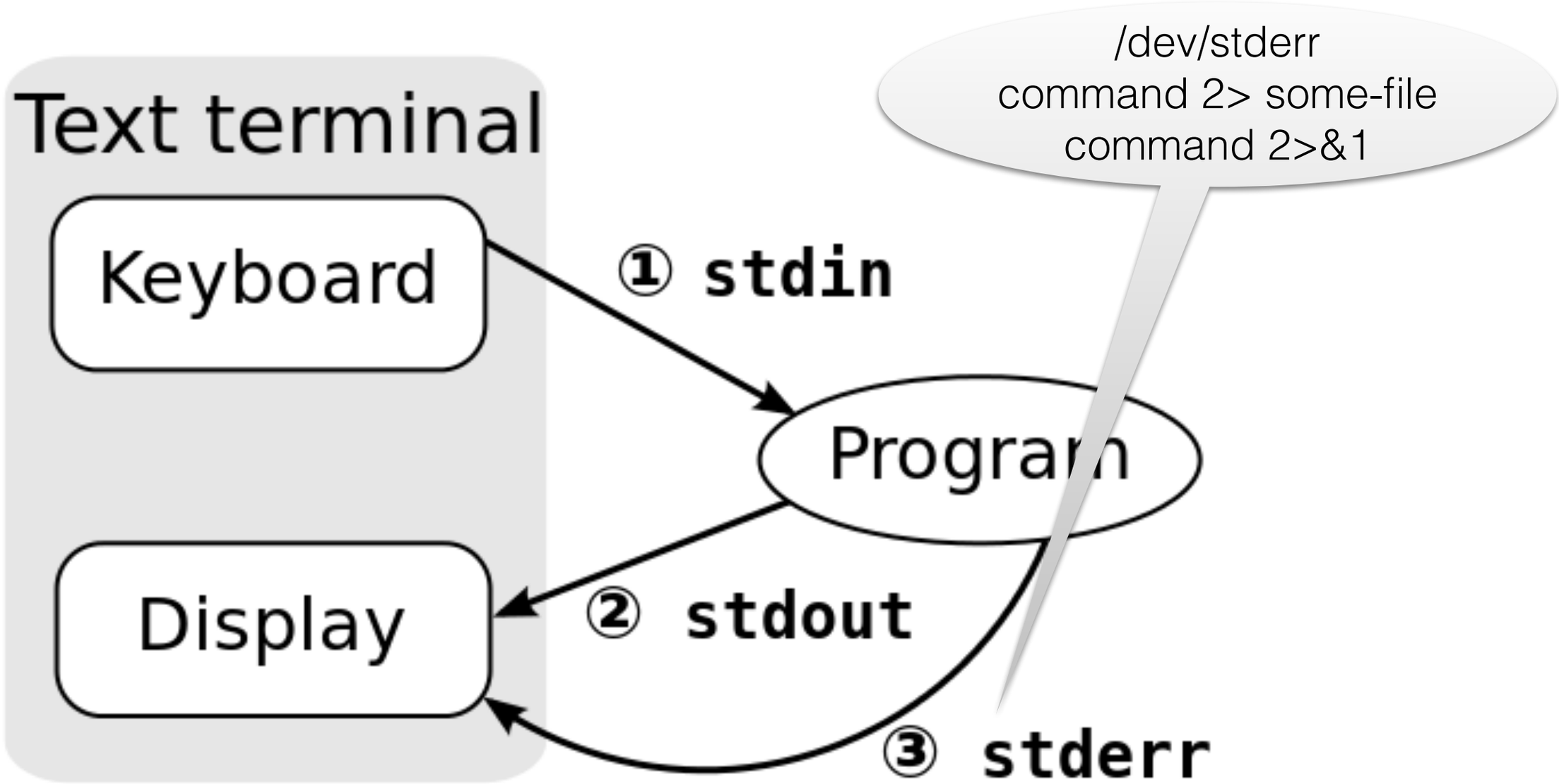
Source:
https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/Stdstreams-notitle.svg/535px-Stdstreams-notitle.svg.png

ASIDE: Standard Streams



Source:
https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/Stdstreams-notitle.svg/535px-Stdstreams-notitle.svg.png

ASIDE: Standard Streams



Source:
https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/Stdstreams-notitle.svg/535px-Stdstreams-notitle.svg.png

splitting
extracting

Filtering

- **head** - start of stream.
- **tail** - end of stream.
- **split** - split files or stdin every n-lines/records/bytes
- **csplit** - split files or stdin at regex pattern

Streaming

Transform

Mining

head

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon Schildenfeld
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Corluka
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

head

head -n 10

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon Schildenfeld
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Corluka
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

10 lines

head worldcup2014-demo.tsv | tr '\t' ','| cut -d, -f1-9| column -s, -t

head

head -n -5

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon S
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Ba
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

omit
5 lines

head -n -5 worldcup2014-demo.tsv | tr '\t' ','| cut -d, -f1-9| column -s, -t

tail

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon Schildenfeld
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Corluka
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

tail

tail -n 10

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Ze
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Co
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

10 lines

tail worldcup2014-demo.tsv | tr '\t' ','| cut -d, -f1-9| column -s, -t

tail

tail -n +5

tab separated — World Cup 2014 Players

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon Schildenfeld
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Corluka
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

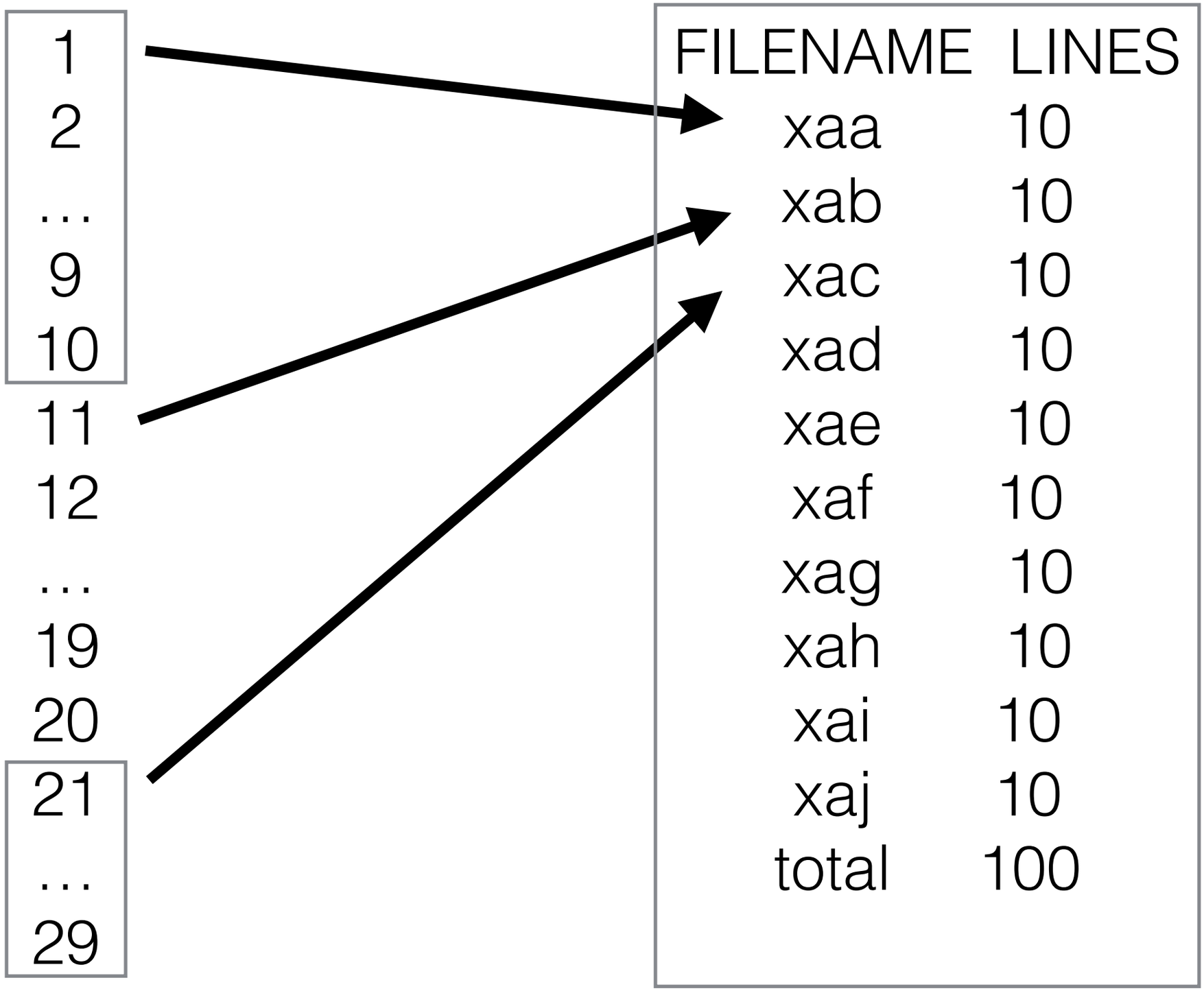
omit
5 lines

tail -n +5 worldcup2014-demo.tsv | tr '\t' ','| cut -d, -f1-9| column -s, -t

split

Create a new file every n-records

seq 100 | split -l 10



split

Create a new file every n-records

ex:100 digits 1 per line

ex: 10 lines per file

```
seq 100 | split -l 10
```

- record separator (default newline -t "\n")
- record count (default 1000 records -l 1000)
- optionally n-bytes instead of records
- optionally n-bytes without breaking records
- optionally round-robin record allocation

FILENAME	LINES
xaa	10
xab	10
xac	10
xad	10
xae	10
xaf	10
xag	10
xah	10
xai	10
xaj	10
total	100

Output summarized with (echo "FILENAME LINES";wc x*|awk '{print $4, $1}') | column -t

Create a new file based on regex match

csplit

```
(seq 10; seq 15; seq 20) | csplit - '/^1$/' '{*}'
```

csplit

Create a new file based on regex match

```
(seq 10; seq 15; seq 20) | csplit - '/^1$/' '{*}'
```

1..10 1..15 1..20

stdin

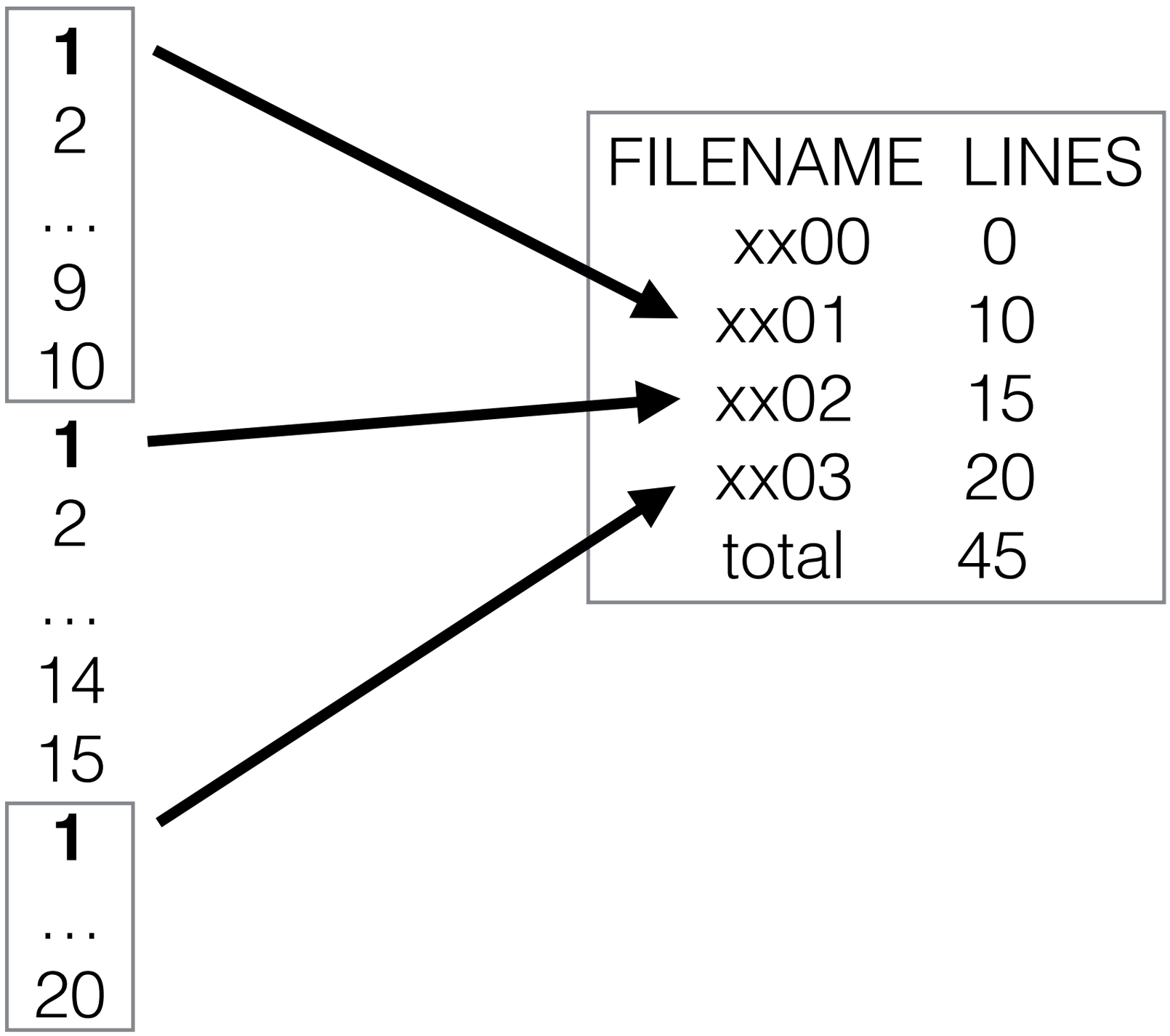
every occurrence

‘1’ occurs alone

csplit

Create a new file based on regex match

```
(seq 10; seq 15; seq 20) | csplit - '/^1$/' '{*}'
```



csplit

Create a new file based on regex match

(seq 10; seq 15; seq 20) | csplit - '/^1$/' '{*}'

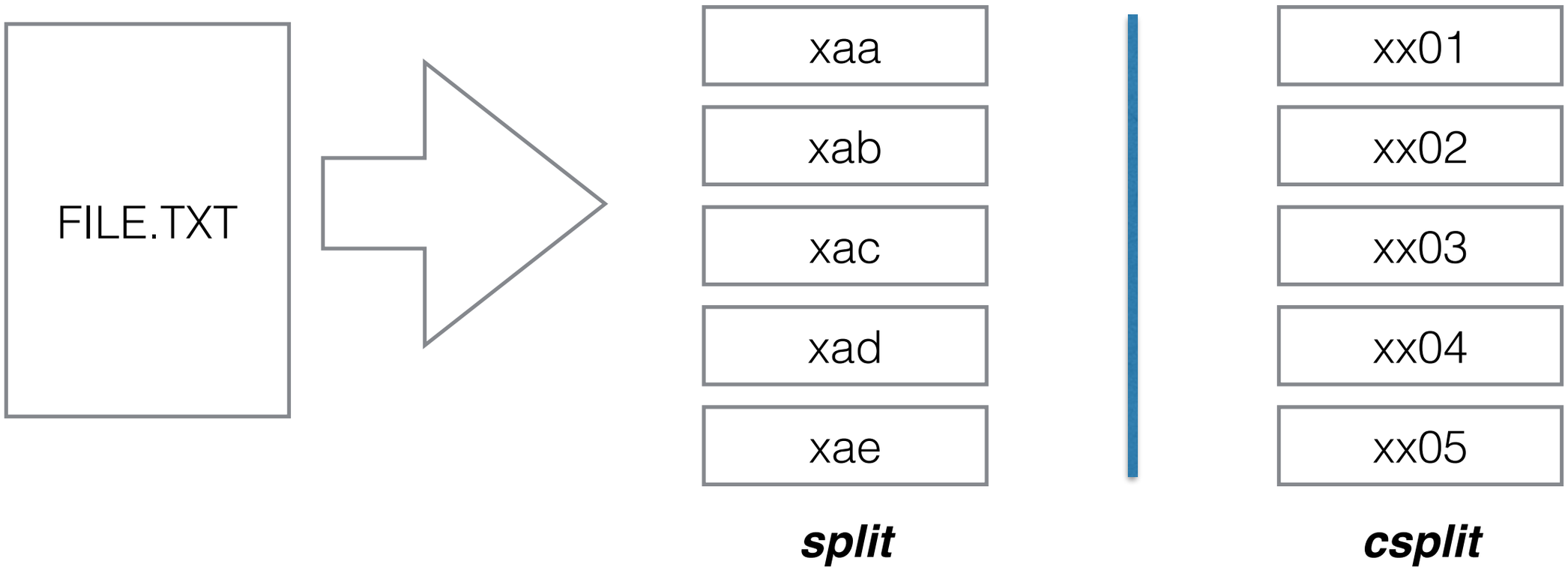
FILENAME	LINES
xx00	**0**
xx01	10
xx02	15
xx03	20
total	45

- %pattern% - skip until matched
- /pattern/ - output until match
- {*} - repeat last pattern
- **--elide-empty-files**
- --keep-files - do not remove output files on errors. An error occurs if pattern not found, for instance.

Output summarized with (echo "FILENAME LINES";wc xx*|awk '{print $4, $1}’) | column -t

split / csplit

Default file naming



Name changing options vary between split and csplit.
See man or help if needed.

splitting
extracting

Filtering

- **cut** - extract sections from each line of output.
- **uniq** - collapses adjacent lines, and makes the output unique.
- **grep** - search for lines matching regular expressions.

Streaming

Transform

Mining

cut

tab separated — World Cup 2014 Players

http://indeedeng.github.io/imhotep/files/worldcupplayerinfo_20140701.tsv

Group	Country	Rank	Jersey	Position	Age	Selections	Club	Player
A	Brazil	3	1	Goalie	31	9	Botafogo	Jefferson
A	Brazil	3	2	Defender	31	75	Barcelona	Dani Alves
A	Brazil	3	6	Defender	26	31	Real Madrid	Marcelo
A	Brazil	3	9	Forward	30	33	Fluminense	Fred
A	Brazil	3	21	Forward	27	17	Atletico Mineiro	Jô
A	Brazil	3	10	Forward	22	49	Barcelona	Neymar
A	Cameroon	56	23	Goalie	24	3	Fethiyespor	Sammy N'Djock
A	Cameroon	56	22	Defender	26	10	Granada	Allan Nyom
A	Cameroon	56	11	Midfielder	31	66	Rennes	Jean Makoun
A	Cameroon	56	17	Midfielder	28	49	Sevilla	Stephane Mbia
A	Cameroon	56	8	Forward	25	17	Nancy	Benjamin Moukandjo
A	Cameroon	56	10	Forward	22	24	Lorient	Vincent Aboubakar
A	Croatia	18	23	Goalie	29	6	Monaco	Danijel Subasic
A	Croatia	18	12	Goalie	21	0	Lokomotiva	Oliver Zelenika
A	Croatia	18	13	Defender	29	21	Panathinaikos	Gordon Schildenfeld
A	Croatia	18	5	Defender	28	72	Lokomotiv Moscow	Vedran Corluka
A	Croatia	18	15	Midfielder	25	9	Hamburger SV	Milan Badelj
A	Croatia	18	10	Midfielder	28	75	Real Madrid	Luka Modric

cut

tab separated — World Cup 2014 Players

http://indeedeng.github.io/imhotep/files/worldcupplayerinfo_20140701.tsv

Group	Country	Rank	**Jersey**	Position	Age	Selections	Club	**Player**
A	Brazil	3	**1**	Goalie	31	9	Botafogo	**Jefferson**
A	Brazil	3	**2**	Defender	31	75	Barcelona	**Dani Alves**
A	Brazil	3	**6**	Defender	26	31	Real Madrid	**Marcelo**
A	Brazil	3	**9**	Forward	30	33	Fluminense	**Fred**
A	Brazil	3	**21**	Forward	27	17	Atletico Mineiro	**Jô**
A	Brazil	3	**10**	Forward	22	49	Barcelona	**Neymar**
A	Cameroon	56	**23**	Goalie	24	3	Fethiyespor	**Sammy N'Djock**
A	Cameroon	56	**22**	Defender	26	10	Granada	**Allan Nyom**
A	Cameroon	56	**11**	Midfielder	31	66	Rennes	**Jean Makoun**
A	Cameroon	56	**17**	Midfielder	28	49	Sevilla	**Stephane Mbia**
A	Cameroon	56	**8**	Forward	25	17	Nancy	**Benjamin Moukandjo**
A	Cameroon	56	**10**	Forward	22	24	Lorient	**Vincent Aboubakar**
A	Croatia	18	**23**	Goalie	29	6	Monaco	**Danijel Subasic**
A	Croatia	18	**12**	Goalie	21	0	Lokomotiva	**Oliver Zelenika**
A	Croatia	18	**13**	Defender	29	21	Panathinaikos	**Gordon Schildenfeld**
A	Croatia	18	**5**	Defender	28	72	Lokomotiv Moscow	**Vedran Corluka**
A	Croatia	18	**15**	Midfielder	25	9	Hamburger SV	**Milan Badelj**
A	Croatia	18	**10**	Midfielder	28	75	Real Madrid	**Luka Modric**

cut -f 4,9

cut

tab separated

cut -f 4,9

Jersey	Player
1	Jefferson
2	Dani Alves
6	Marcelo
9	Fred
21	Jô
10	Neymar
23	Sammy N'Djock
22	Allan Nyom
11	Jean Makoun
17	Stephane Mbia
8	Benjamin Moukandjo
10	Vincent Aboubakar
23	Danijel Subasic
12	Oliver Zelenika
13	Gordon Schildenfeld
5	Vedran Corluka
15	Milan Badelj

Group	Country	Rank	**Jersey**	Position	Age	Selections	Club	**Player**
A	Brazil	3	**1**	Goalie	31	9	Botafogo	**Jefferson**
A	Brazil	3	**2**	Defender	31	75	Barcelona	**Dani Alves**
A	Brazil	3	**6**	Defender	26	31	Real Madrid	**Marcelo**
A	Brazil	3	**9**	Forward	30	33	Fluminense	**Fred**
A	Brazil	3	**21**	Forward	27	17	Atletico Mineiro	**Jô**
A	Brazil	3	**10**	Forward	22	49	Barcelona	**Neymar**
A	Cameroon	56	**23**	Goalie	24	3	Fethiyespor	**Sammy N'Djock**
A	Cameroon	56	**22**	Defender	26	10	Granada	**Allan Nyom**
A	Cameroon	56	**11**	Midfielder	31	66	Rennes	**Jean Makoun**
A	Cameroon	56	**17**	Midfielder	28	49	Sevilla	**Stephane Mbia**
A	Cameroon	56	**8**	Forward	25	17	Nancy	**Benjamin Moukandjo**
A	Cameroon	56	**10**	Forward	22	24	Lorient	**Vincent Aboubakar**
A	Croatia	18	**23**	Goalie	29	6	Monaco	**Danijel Subasic**
A	Croatia	18	**12**	Goalie	21	0	Lokomotiva	**Oliver Zelenika**
A	Croatia	18	**13**	Defender	29	21	Panathinaikos	**Gordon Schildenfeld**
A	Croatia	18	**5**	Defender	28	72	Lokomotiv Moscow	**Vedran Corluka**
A	Croatia	18	**15**	Midfielder	25	9	Hamburger SV	**Milan Badelj**
A	Croatia	18	**10**	Midfielder	28	75	Real Madrid	**Luka Modric**

cut

colon separated

/etc/passwd

nobody:*:-2:-2:Unprivileged User:/var/empty:**/usr/bin/false**
root:*:0:0:System Administrator:/var/root:**/bin/sh**
daemon:*:1:1:System Services:/var/root:**/usr/bin/false**
…

cut

colon separated

/etc/passwd

nobody:*:-2:-2:Unprivileged User:/var/empty:**/usr/bin/false**
root:*:0:0:System Administrator:/var/root:**/bin/sh**
daemon:*:1:1:System Services:/var/root:**/usr/bin/false**
…

cut -d: -f 1,7 /etc/passwd

```
nobody:/usr/bin/false
root:/bin/sh
daemon:/usr/bin/false
```

cut

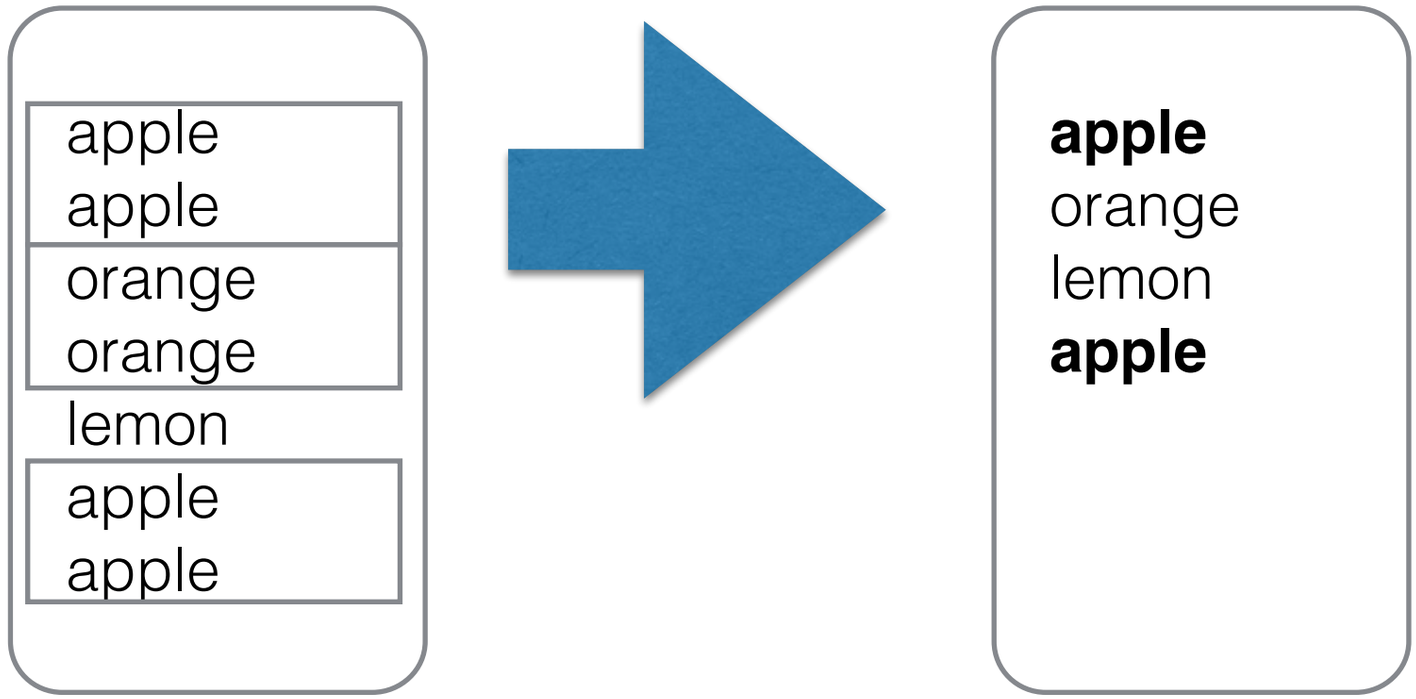
fixed field sizes

echo 1234567890**112233 | cut -c 8-10**

890

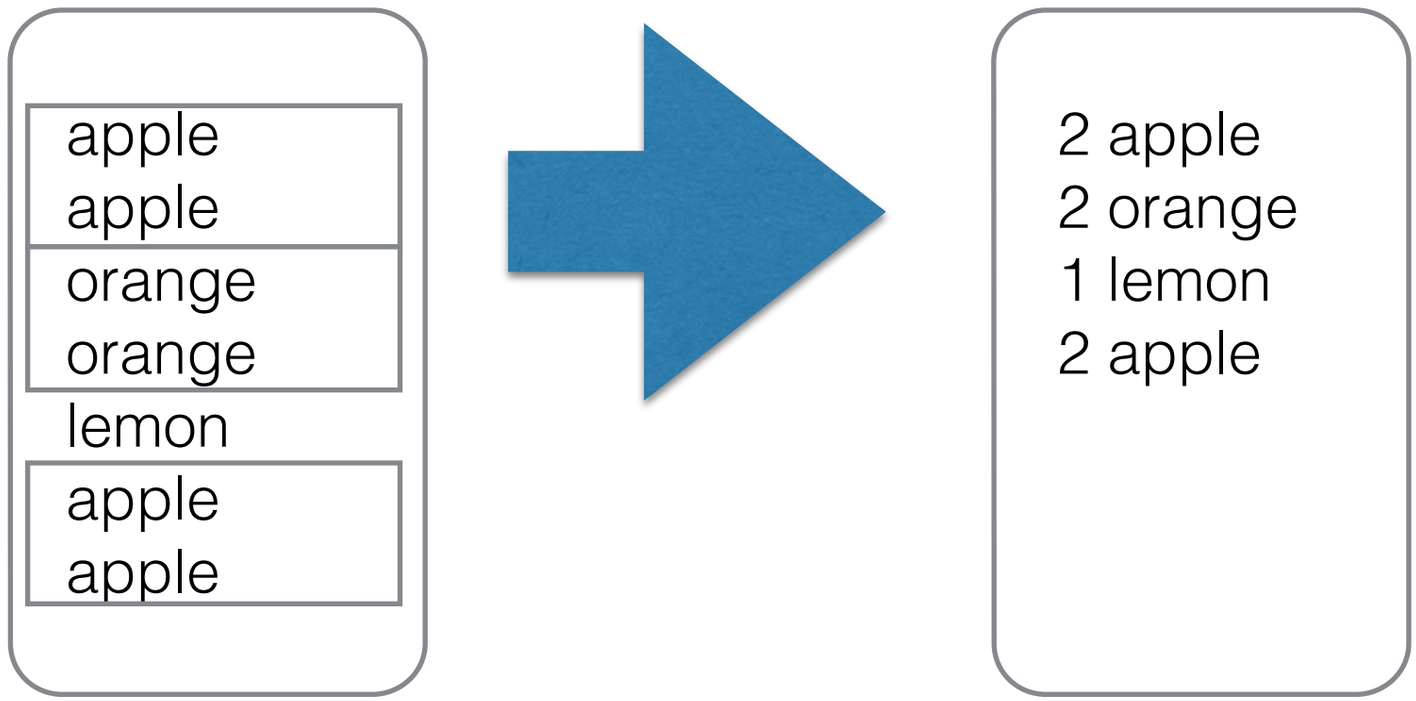
uniq

Remove repeated adjacent lines



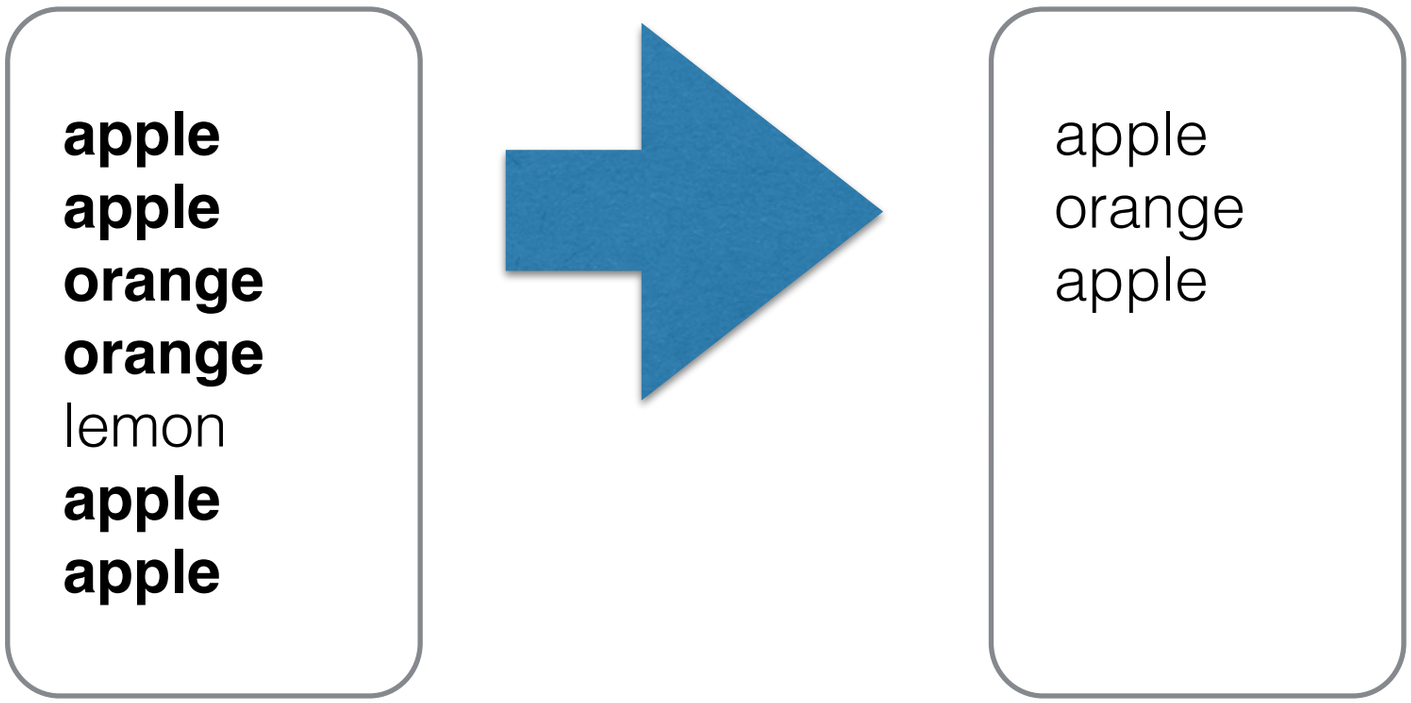
uniq

Count repeated adjacent lines
uniq -c



uniq

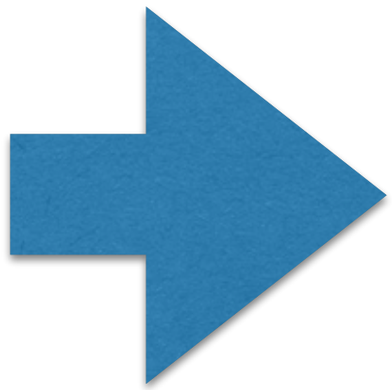
Only show repeating lines
uniq -d



uniq

Only show non-repeating lines
uniq -u

apple
apple
orange
orange
lemon
apple
apple
lemon

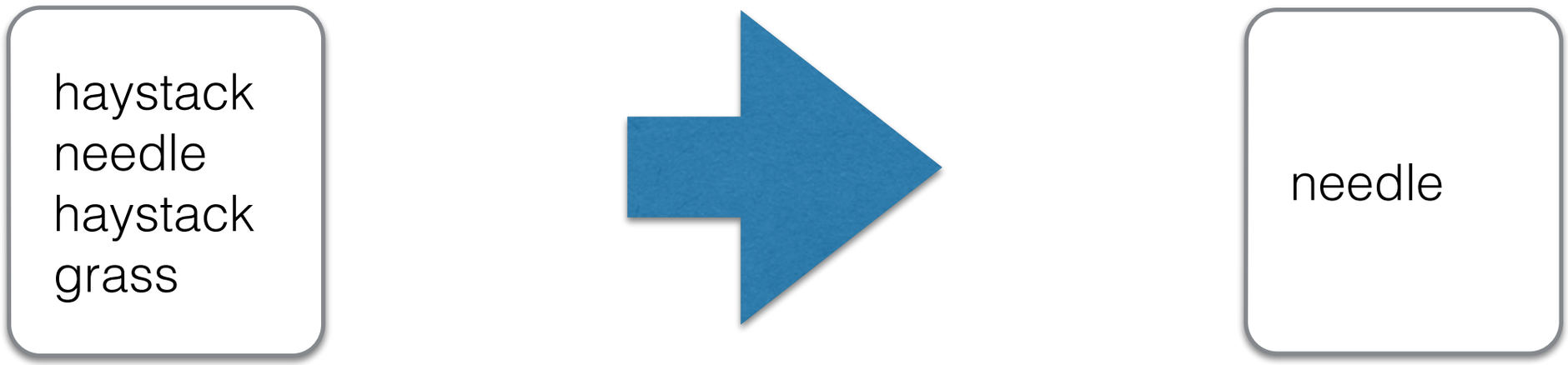


lemon
lemon

grep

- Select lines that match

grep needle

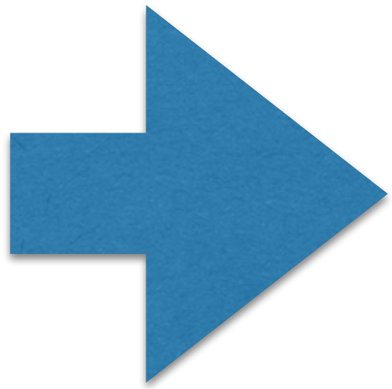


grep

- Select lines that match

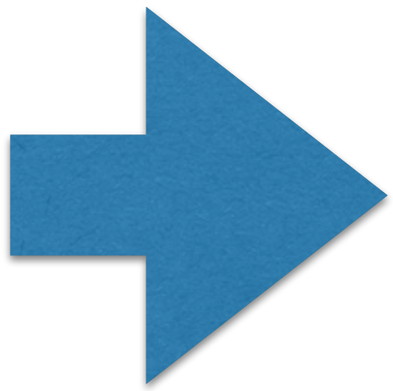
grep needle

haystack
needle
haystack
grass



needle

haystack
haystack needle haystack
haystack
river



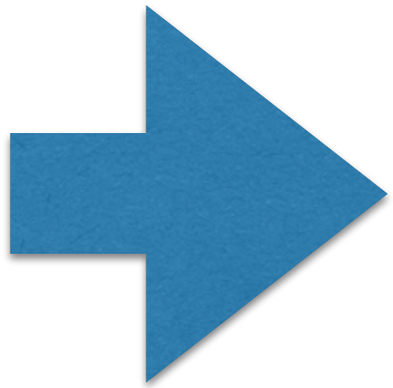
haystack needle haystack

grep

- Extract the matching portion from within line

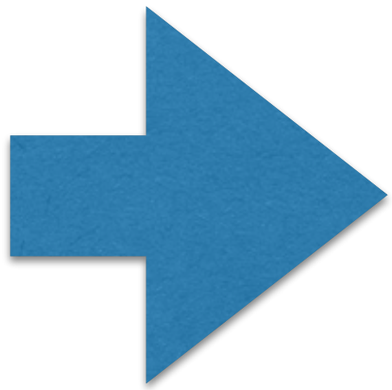
grep -o n…..e

haystack
needle
haystack
grass



needle

haystack
haystack needle haystack
haystack
river



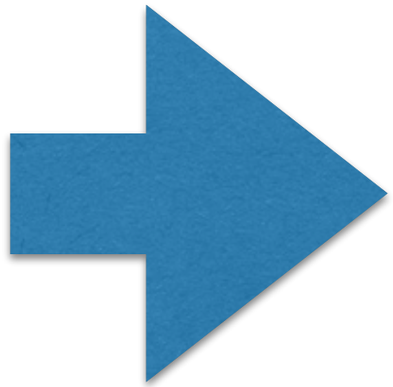
needle

grep

- Select lines NOT matching

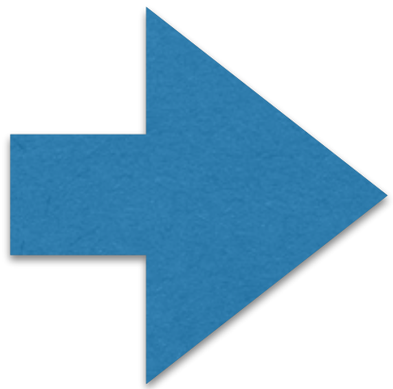
grep -v needle

haystack
needle
haystack
grass



haystack
haystack
grass

haystack
haystack needle haystack
haystack
river

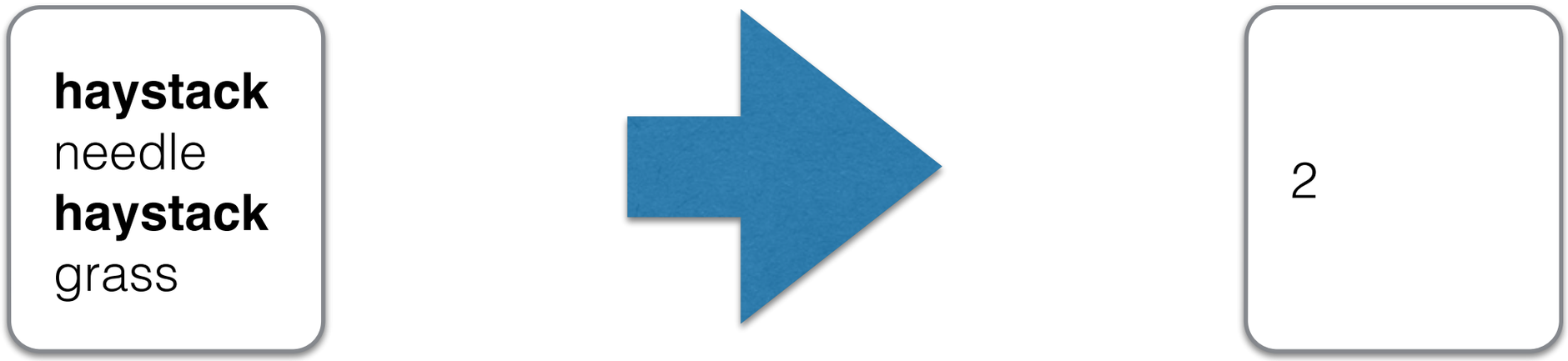


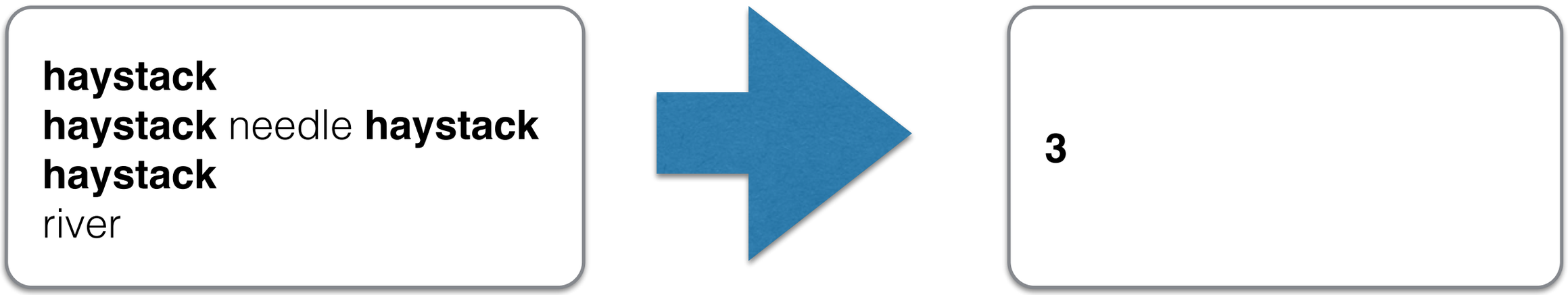
haystack
haystack
river

grep

- Count haystacks (number of lines)

grep -c haystack

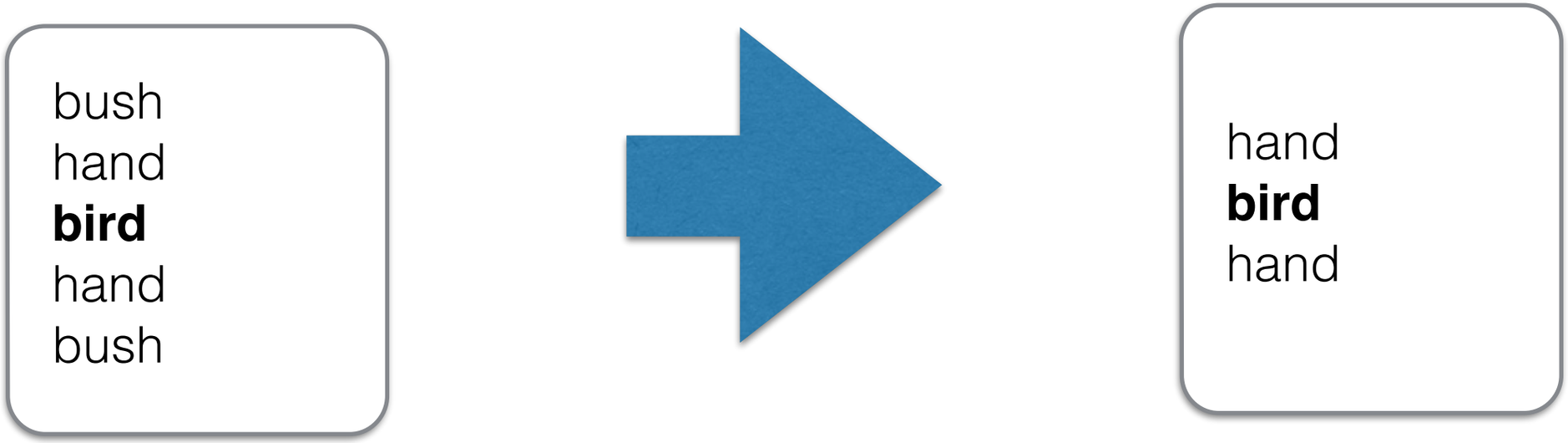




grep

- Print lines just before / after a match (context)

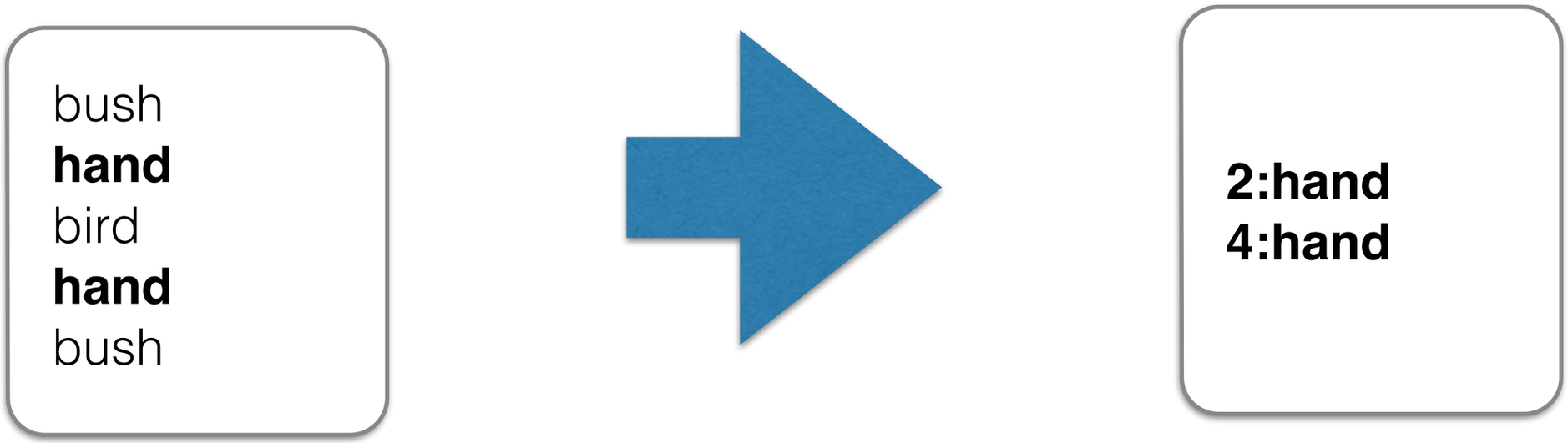
grep -C1 bird



grep

- Print line numbers

grep -n hand

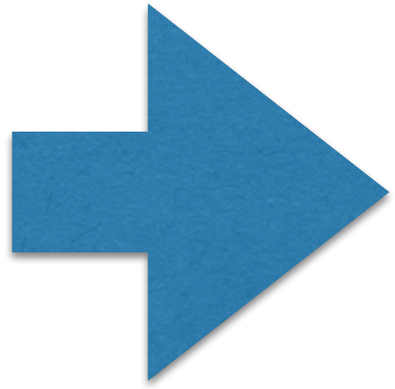


grep

- List files containing match

Files **grep -l crumb ***

pantry.txt
refrigerator.txt
sink.txt

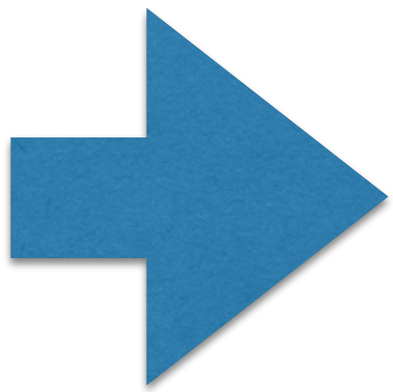


pantry.txt

- List files NOT containing match

Files **grep -L crumb ***

pantry.txt
refrigerator.txt
sink.txt



refrigerator.txt
sink.txt

grep

- Most common regex

	Operator	Example	Matches
Alternative words	**|**	cat|dog	con**cat**s **dogs**
Alternative characters	**[ ]**	[$¢£¥€]	**€**10 **£**20 5**¢**
Line starts with	**^**…	^Rise	**Rise** and shine
Line ends with	…**$**	to be$	to be or not **to be**
Any single char	**.**	p..ch	**peach**
>=1 of previous	**.+**	ya+y	y**a**y y**aaaa**y
0 or more of previous	**.***	yea*h	yeh ye**aaaaa**h

Non bolded . represents character or expression

- Basic grep treats ?, +, {, |, (, and) as literals unless escaped unlike -E extended, -P perl regex

ordering
presentation
compare
subst / xform

Streaming

Filtering

Transform

Mining

- **paste** - join multiple files horizontally.
- **join** - merges lines of two files based on commonalities.
- **sort** - sorts input. (also -u)

paste

Interleave streams into rows of tab separated values

colors.txt

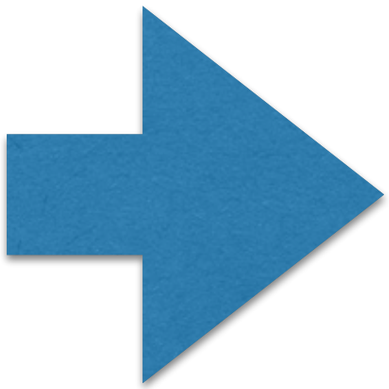
blue
red
yellow
black

nouns.txt

berries
apples
flowers
penguin

verbs.txt

grow
shine
bloom



paste colors.txt nouns.txt verbs.txt

blue	**berries**	**grow**	\n
red	*apples*	*shine*	\n
yellow	flowers	shine	\n
black	***penguin***		\n

paste

Round-robin interleaving from a single stdin stream

```
ls
```

```
file1.txt
file2.txt
file3.txt
file4.txt
```

```
ls | paste - - -
```

```
file1.txt       file2.txt       files3.txt
file4.txt
```

paste

Round-robin interleaving from a single stdin stream

```
ls
```

```
file1.txt
file2.txt
file3.txt
file4.txt
```

```
ls | paste - - -
```

```
file1.txt       file2.txt       files3.txt
file4.txt
```

Specifying delimiters

```
ls | paste -d,: - - -
```

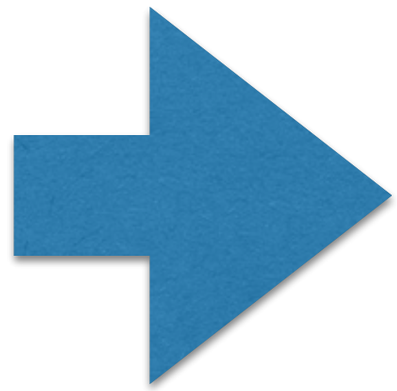
```
file1.txt,file2.txt:files3.txt
file4.txt,:
```

join

Merge rows based on a common field

1 abc
2 lmn
3 pqr

1 ABC
3 LMN
9 OPQ

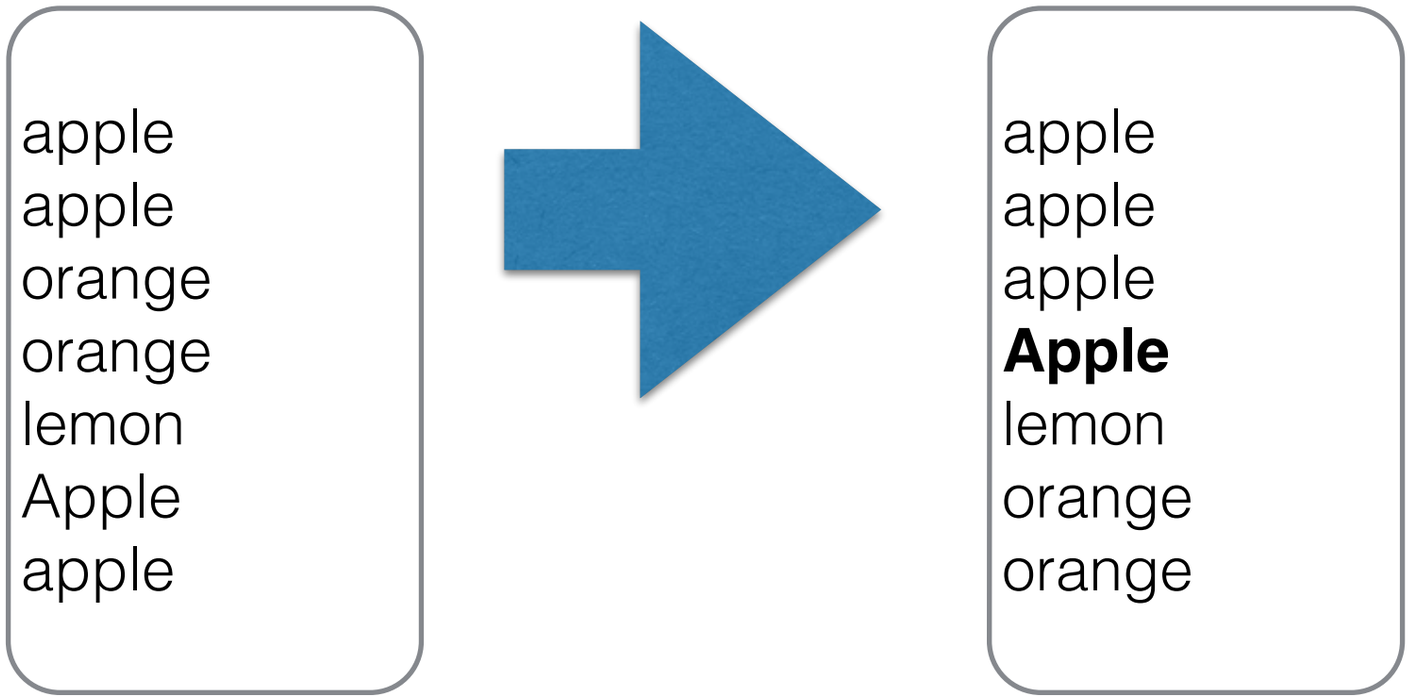


`join file1 file2`

1 abc ABC
3 pqr LMN

- Key column and field delimiter can be specified
- Stream input must be sorted by key column

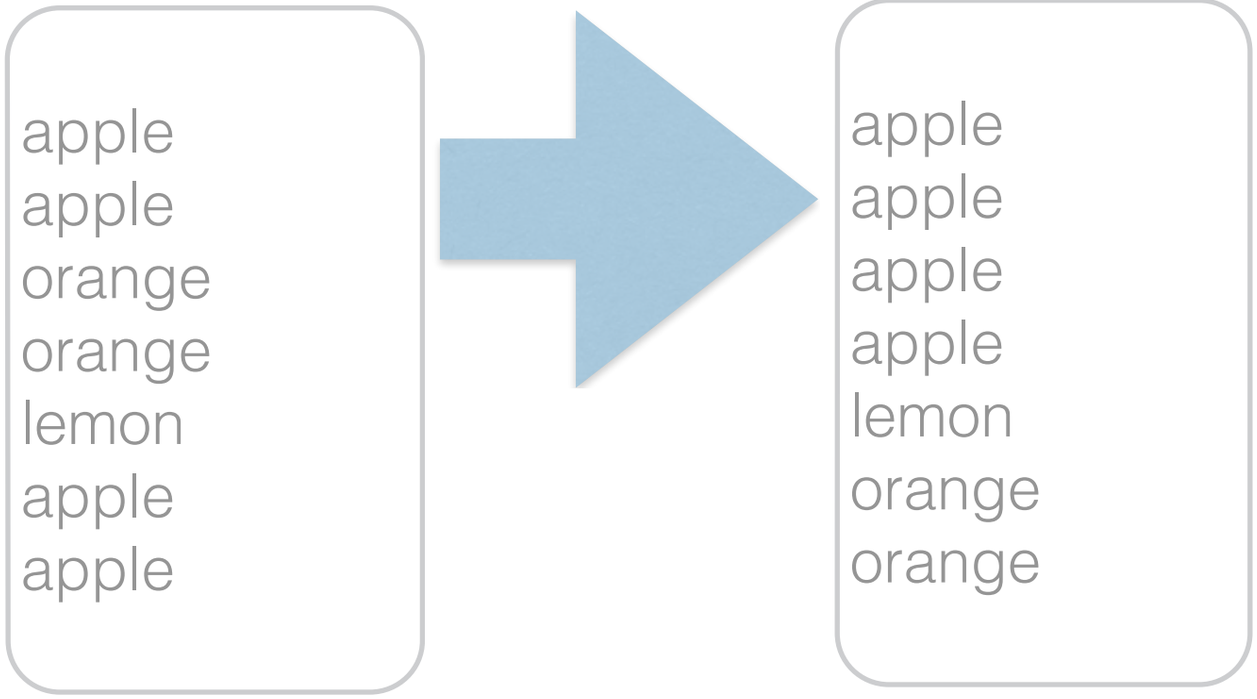
sort



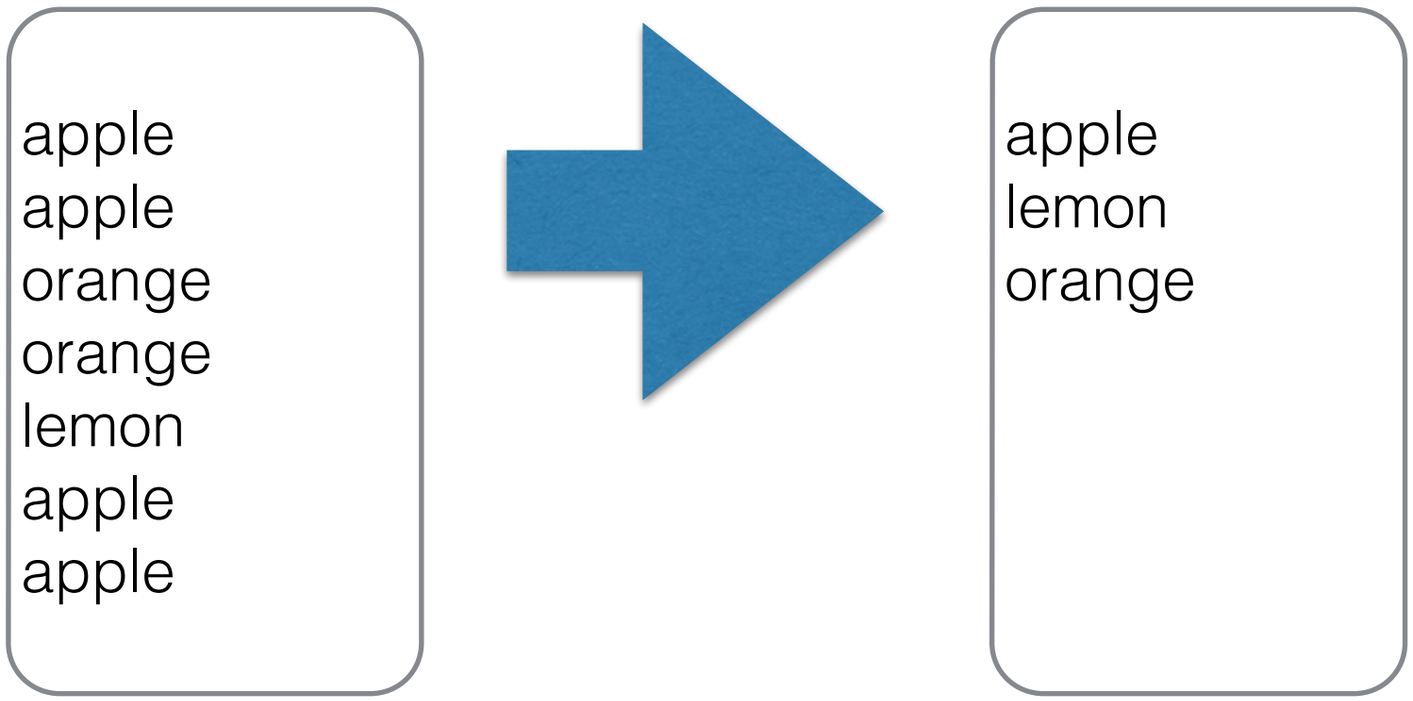
- Note case insensitive using locale en_US.UTF-8
- Numeric ordering by using LC_COLLATE=C

* Reference http://teaching.idallen.org/net2003/06w/notes/character_sets.txt

sort



sort | uniq is mostly equivalent* to **sort -u**



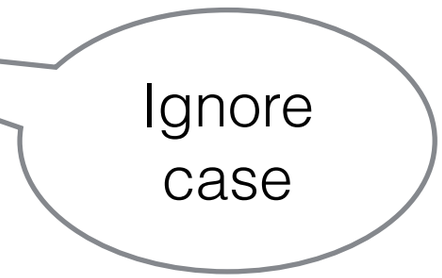
* sort and uniq handle locale differently
https://unix.stackexchange.com/a/76095

sort

Observations on case sensitive sort*

```
1 - sort (no options)
2 - sort --ignore-case
3 - LC_COLLATE=C sort
4 - LC_COLLATE=C sort --ignore-case
5 - sort -u
6 - sort -u --ignore-case
7 - sort | uniq
8 - sort | uniq -i
```

1	2	3	4	5	6	7	8
a	a	B	a	a	a	a	a
b	b	D	B	b	b	b	b
B	B	a	b	B	c	B	c
c	c	b	c	c	d	c	d
d	d	c	D	d		d	
D	D	d	d	D		D	



* Results shown using LANG="en_US.UTF-8"
Example based on https://superuser.com/a/178214

Key field sort

sort -k 8

B	Australia	62	18	Goalie	32	8	**Adelaide United**	Eugene Galekovic	0
G	Ghana	37	1	Goalie	24	7	**Aduana Stars**	Steven Adams	0
D	Costa Rica	28	5	Midfielder	26	63	**AIK**	Celso Borges	0
A	Mexico	20	13	Goalie	28	59	**Ajaccio**	Guillermo Ochoa	0
H	Algeria	22	22	Midfielder	30	22	**Ajaccio**	Mehdi Mostefa	0
B	Netherlands	15	5	Defender	24	12	**Ajax**	Daley Blind	0
B	Netherlands	15	1	Goalie	25	8	**Ajax**	Jasper Cillessen	0
B	Netherlands	15	13	Defender	22	2	**Ajax**	Joel Veltman	0
G	Ghana	37	3	Forward	28	78	**Al-Ain**	Asamoah Gyan	1
E	Honduras	33	9	Forward	32	23	**Alajuelense**	Jerry Palacios	0

```
sort -k 8 -t "   " worldcupplayerinfo_20140701.tsv | tail -n +10 | head | column -s '   ' -t
```

sort

```
du --max-depth=1 | sort
```

```
10032    ./.config
11192    ./perl5
129580   ./.virtualenvs
12       ./.compiz
12       ./.dbus
2884     ./text-tools
364820   .
36       ./.gnupg
36       ./.pki
48       ./.byobu
48       ./.ipython
49328    ./nltk_data
4        ./Desktop
4        ./Documents
4        ./Downloads
4        ./.gconf
4        ./.marks
4        ./Music
4        ./Pictures
```

Alphanumeric sort by default

sort

```
du --max-depth=1 | sort -n
```

```
4       ./Desktop
4       ./Documents
4       ./Downloads
4       ./.gconf
4       ./.marks
4       ./Music
4       ./Pictures
4       ./Public
4       ./Templates
4       ./Videos
8       ./.vim
12      ./.compiz
12      ./.dbus
36      ./.gnupg
36      ./.pki
48      ./.byobu
48      ./.ipython
76      ./.java
2884    ./text-tools
```

Numeric sort

-n

sort

```
du -h --max-depth=1 | sort -n
```

```
4.0K  ./Desktop
4.0K  ./Documents
4.0K  ./Downloads
4.0K  ./.gconf
4.0K  ./Music
4.0K  ./Pictures
4.0K  ./Public
4.0K  ./Templates
4.0K  ./Videos
8.0K  ./.gnupg
 13M  ./.mozilla
 21M  ./.cache
 76K  ./.java
224K  ./.config
416M  ./text-tools
450M  .
944K  ./.local
```

Numeric sort insufficient for quantified values

sort

```
du -h --max-depth=1 | sort -h
```

```
4.0K    ./Desktop
4.0K    ./Documents
4.0K    ./Downloads
4.0K    ./.gconf
4.0K    ./Music
4.0K    ./Pictures
4.0K    ./Public
4.0K    ./Templates
4.0K    ./Videos
8.0K    ./.gnupg
76K     ./.java
224K    ./.config
944K    ./.local
13M     ./.mozilla
21M     ./.cache
416M    ./text-tools
450M    .
```

Human readable quantified values

-h

ordering
presentation 1
compare
subst / xform

Streaming

Filtering

Transform

Mining

- **nl** - add line numbers, many options. (Compare cat/awk)
- **column** - column align fields
- **expand** - tab to spaces.
- **unexpand** - spaces to tabs.

ordering
presentation 2
compare
subst / xform

Streaming

Filtering

Transform

Programming Language and Data Format Reformatting

Mining

C
Java

HTML
XML

binary

javascri

JSON

Perl

C / C++ / C# / ObjC / Java

```
apt install astyle
astyle unformatted.c
```

```
void CTrampDlg::LoadConfigFile()
{
int handle,ptr;
unsigned long size;
char cfgfile[255];

        GetModuleFileName(NULL,configpath,25

        configpath[ptr]=0;
    sprintf(cfgdata.listpath,"%s",configpath
        sprintf(cfgdata.skinpath,"%s\\skins\
        sprintf(cfgdata.songpath,"%s",config

    sprintf(cfgdata.currentskin,"tramp.skn")
    cfgdata.currentsong[0]=0;
        cfgdata.currentlist[0]=0;
        sprintf(cfgfile,"%s\\trampcfg.dat",c

    cfgdata.x=100; cfgdata.y=100;
        cfgdata.w=500; cfgdata.h=175;

    for(ptr=0; ptr<11; ptr++) cfgdata.eqpos[
        cfgdata.shufflecode=cfgdata.loopcode
        cfgdata.eqonoff=true;
        cfgdata.volume=(unsigned long) 0x3ff
    for(ptr=0; ptr<50; ptr++) cfgdata.contro

        handle=open(cfgfile,O_RDWR|O_BINARY)

    if(handle==-1) {SaveConfigFile(); return
```

```
void CTrampDlg::LoadConfigFile()
{
    int handle,ptr;
    unsigned long size;
    char cfgfile[255];

    GetModuleFileName(NULL,configpath,254);

    configpath[ptr]=0;
    sprintf(cfgdata.listpath,"%s",configpath);
    sprintf(cfgdata.skinpath,"%s\\skins\\tramp",configpath);
    sprintf(cfgdata.songpath,"%s",configpath);

    sprintf(cfgdata.currentskin,"tramp.skn");
    cfgdata.currentsong[0]=0;
    cfgdata.currentlist[0]=0;
    sprintf(cfgfile,"%s\\trampcfg.dat",configpath);

    cfgdata.x=100;
    cfgdata.y=100;
    cfgdata.w=500;
    cfgdata.h=175;

    for(ptr=0; ptr<11; ptr++) cfgdata.eqpos[ptr]=50;
    cfgdata.shufflecode=cfgdata.loopcode=false;
    cfgdata.eqonoff=true;
    cfgdata.volume=(unsigned long) 0x3fff3fffL;
    for(ptr=0; ptr<50; ptr++) cfgdata.controls[ptr]=0;
```

Unformatted sample from http://www.polystyle.com/features/c-unformatted.jsp

HTML / XML

```
apt install libxml2-utils
xmllint --format --html index.html
```

Also consider
apt install tidy

```
<div class="langSelect inlineblock">^M
    <div id="langSelect" class="inlineblock pointer" onmouseover="$('#editionContainer').fa
middle inlineblock"></span></div>^M
    <div id="TopFlagsContainer" class="tooltipPopup countrySel
        <div class="content">^M
            <ul><li><a href="http://uk.investing.com"><span
>Русский</a></li><li><a href="http://in.investing.com"><span
<li><a href="http://ca.investing.com"><span class="ceFlags Ca
li><li><a href="http://au.investing.com"><span class="ceFlags
 href="http://za.investing.com"><span class="ceFlags South_A
ef="http://de.investing.com"><span class="ceFlags Germany"><
.com"><span class="ceFlags Spain"></span>Español (España)</a>
"ceFlags Mexico"></span>Español (México)</a></li><li><a href=
</span>Français</a></li><li><a href="http://kr.investing.com"
/li><li><a href="http://cn.investing.com"><span class="ceFlag
://hk.investing.com"><span class="ceFlags Hong_Kong"></span>
an class="ceFlags Indonesia"></span>Indonesia</a></li><li><a
lass="ceFlags Malaysia"></span>Malaysia</a></li><li><a href="
gs Thailand"></span>Thailand</a></li></ul>
    </div>^M
</div>^M
<!-- GEO POPUP -->^M
<div class="topBarTableBox chooseEdition tooltipPopup bigArr
ss="popupInnerWrapper">
 class="header"><span><i class="ceFlags "></i></span><span c
 class="content">
        </div>
 class="topPortfolioFooter">
<a href="javascript:void(0);" onClick="postponeEdition(14)"
    </div>
```

```
                <div id="TopFlagsContainer" class="tooltipPopup countrySelPopup bigArrowTopbar noHeader displa
                    <div class="content">^M
                        <ul>
<li><a href="http://uk.investing.com"><span class="ceFlags United_Kingdom"></span>English (UK)</a></li>
<li><a href="http://ru.investing.com"><span class="ceFlags Russian_Federation"></span>&#1056;&#1091;&#1089;&#10
<li><a href="http://in.investing.com"><span class="ceFlags India"></span>English (India)</a></li>
<li><a href="http://tr.investing.com"><span class="ceFlags Turkey"></span>T&uuml;rk&ccedil;e</a></li>
<li><a href="http://ca.investing.com"><span class="ceFlags Canada"></span>English (Canada)</a></li>
<li><a href="http://sa.investing.com"><span class="ceFlags Saudi_Arabia"></span>&rlm;&#1575;&#1604;&#1593;&#158
<li><a href="http://au.investing.com"><span class="ceFlags Australia"></span>English (Australia)</a></li>
<li><a href="http://gr.investing.com"><span class="ceFlags Greece"></span>&Epsilon;&lambda;&lambda;&eta;&nu;&ic
<li><a href="http://za.investing.com"><span class="ceFlags South_Africa"></span>English (South Africa)</a></li
<li><a href="http://se.investing.com"><span class="ceFlags Sweden"></span>Svenska</a></li>
<li><a href="http://de.investing.com"><span class="ceFlags Germany"></span>Deutsch</a></li>
<li><a href="http://fi.investing.com"><span class="ceFlags Finland"></span>Suomi</a></li>
<li><a href="http://es.investing.com"><span class="ceFlags Spain"></span>Espa&ntilde;ol (Espa&ntilde;a)</a></li
<li><a href="http://il.investing.com"><span class="ceFlags Israel"></span>&#1506;&#1489;&#1512;&#1497;&#1514;</
<li><a href="http://mx.investing.com"><span class="ceFlags Mexico"></span>Espa&ntilde;ol (M&eacute;xico)</a></l
<li><a href="http://jp.investing.com"><span class="ceFlags Japan"></span>&#26085;&#26412;&#35486;</a></li>
<li><a href="http://fr.investing.com"><span class="ceFlags France"></span>Fran&ccedil;ais</a></li>
<li><a href="http://kr.investing.com"><span class="ceFlags South_Korea"></span>&#54620;&#44397;&#50612;</a></li
<li><a href="http://it.investing.com"><span class="ceFlags Italy"></span>Italiano</a></li>
<li><a href="http://cn.investing.com"><span class="ceFlags China"></span>&#20013;&#25991;</a></li>
<li><a href="http://nl.investing.com"><span class="ceFlags Netherlands"></span>Nederlands</a></li>
<li><a href="http://hk.investing.com"><span class="ceFlags Hong_Kong"></span>&#39321;&#28207;</a></li>
<li><a href="http://pl.investing.com"><span class="ceFlags Poland"></span>Polski</a></li>
<li><a href="http://id.investing.com"><span class="ceFlags Indonesia"></span>Indonesia</a></li>
<li><a href="http://pt.investing.com"><span class="ceFlags Portugal"></span>Portugu&ecirc;s (Portugal)</a></li
<li><a href="http://ms.investing.com"><span class="ceFlags Malaysia"></span>Malaysia</a></li>
<li><a href="http://br.investing.com"><span class="ceFlags Brazil"></span>Portugu&ecirc;s (Brasil)</a></li>
<li><a href="http://th.investing.com"><span class="ceFlags Thailand"></span>Thailand</a></li>
</ul>                    </div>^M
                </div>^M
            </div>^M
            <!-- GEO POPUP -->^M
            <div class="topBarTableBox chooseEdition tooltipPopup bigArrowTopbar " id="topAlertBarContainer" st
    <div class="popupInnerWrapper">
        <div class="header">
pan><i class="ceFlags "></i></span><span class="chooseEditionHeaderText"></span><a href="javascript:void(0)
```

Javascript

python -m json.tool


```
{"id":"0001","type":"donut","name":"Cake","image":{"url":"images/0001.jpg","widt
h":200,"height":200},"thumbnail":{"url":"images/thumbnails/0001.jpg","width":32,
"height":32}}
```



```
{
    "id": "0001",
    "image": {
        "height": 200,
        "url": "images/0001.jpg",
        "width": 200
    },
    "name": "Cake",
    "thumbnail": {
        "height": 32,
        "url": "images/thumbnails/0001.jpg",
        "width": 32
    },
    "type": "donut"
```

Perl

```
perl -MCPAN -e 'Perl::Tidy'
perltidy -sil=0 unformatted.pl
```

```
                                                #
                                              sub j(\$){($
          P,$V)=                             @_;while($$P=~s:^
        ([()])::x){                         $V+=('('eq$1)?-32:31
  }$V+=ord(  substr(                       $$P,0,1,""))-74} sub a{
 my($I,$K,$  J,$L)=@_                     ;$I=int($I*$M/$Z);$K=int(
$K*$M/$Z);$J=int($J*$M                    /$Z);$L=int($L*$M/$Z); $G=$
J-$I;$F=$L-$K;$E=(abs($                  G)>=abs($F))?$G:$F;($E<0) and($
 I,$K)=($J,$L);$E||=.01                 ;for($i=0;$i<=abs$E;$i++ ){ $D->{$K
        +int($i*$F/$E)                 }->{$I+int($i*$G/$E)}=1}}sub p{$D={};$
       Z=$z||.01;map{                $H=$_;$I=$N=j$H;$K=$O=j$H;while($H){$q=ord
      substr($H,0,1,"" );if(42==$q){$J=j$H;$L=j$H}else{$q-=43;$L =$
    %9;$J=($q-$L)/9;$L=$q-9*$J-4;$J-=4}$J+=$I;$L+=$K;a($I,$K,$J,$ L
    ($I,$K)=($J,$L)}a($I,$K,$N,$O)}@_;my$T;map{$y=$_;map{ $T.=$D->{
    ->{$_}?$\:' '}(-59..59);$T.="\n"}(-23..23);print"\e[H$T"}$w= ev
    require Win32::Console::ANSI};$b=$w?'1;7;':"";($j,$u,$s,$t,$a,$
    ,$h,$c,$k,$p,$e,$r,$l,$C)=split/}/,'Tw*JSK8IAg*PJ[*J@wR}*JR]*QJ
     'BA*JQK8I*JC}KUz]BAIJT]*QJ[R?-R[e]\RI'.'}Tn*JQ]wRAI*JDnR8QAU}w
     ']n*JEI*EJR*QJ]*JR*DJ@IQ[}*JSe*JD[n]*JPe*'.'JBI/KI}T8@?PcdnfgV
      '?ABKV]]}*JWe*JD[n]*JPe*JC?8B*JE};Vq*OJQ/IP['.'wQ}*JWeOe{n*EE
        'J*JC}/U*OJd[OI@*BJ*JXn*J>w]U}CWq*OJc8KJ?O[e]U/T*QJP?}*JSe*
        'QIAKJR}*JV]wRAI*J?}T]*RJcJI[\]3;U]Uq*PM[wV]W]WCT*DM*SJ'.
           'PZa[\]UKVgogK9K*QJ[\]n[RI@*EH@IddR[Q[]T]T]T3o[dk*JE'.
            '{T]*JPKTKK]*OJ[QIO[PIQIO[[gUKU\k*JE+J+J5R5AI*EJ00'.
                'DMKKJIR[Q+*EJ0*EK';sub h{$\ = qw(% & @ x)[int
                4];map{printf  "\e[$b;%dm",int(rand 6)+101-60*
                 ||0);system(  "cls")if$w ;($A,$S)=    ($_[1],
                  _[0]);($M,   @,)= split  '}';for(     $z=256
                  ;$z>0; $z   -=$S){$S*=   $A;p @,}     sleep$
                  [2];while   ($_[3]&&($    z+=$ S)      <=256
                  p@,}}("".   "32}7D$j"     ."}AG".      "$u}O
                  ."$s}WG"    ."$t","       ."24}("       ."IJ
                  ."}1G$n"    ."}CO$o"      ."}GG$t"      ."}Q
                   ."$h}"      ."^G$e"     ."})IG"        ."$r
                   "32}?"      ."H$p}FG$e}QG$r".          "}ZC
                   ."$l",        "28}(LC" ."              ."".
                   "$h}:"          ."J$a}EG".             "$c"
                   ."}M"           ."C$k}ZG".             "$e"
                   ."}"            ."dG$r","18"          ."}("
                  ."D;"            ."$C"  )}{h(16        ,1,1,
                 );h(8,          .98,0,0   );h(16        ,1,1,
```

```
sub j(\$) {
    ( $P, $V ) = @_;
    while (
        $$P =~ s:^
                ([()])::x
      )
    {
        $V += ( '(' eq $1 ) ? -32 : 31;
    }
    $V += ord( substr( $$P, 0, 1, "" ) ) - 74;
}

sub a {
    my ( $I, $K, $J, $L ) = @_;
    $I = int( $I * $M / $Z );
    $K = int( $K * $M / $Z );
    $J = int( $J * $M / $Z );
    $L = int( $L * $M / $Z );
    $G = $J- $I;
    $F = $L - $K;
    $E = ( abs($G) >= abs($F) ) ? $G : $F;
    ( $E < 0 ) and ( $I, $K ) = ( $J, $L );
    $E ||= .01;

    for ( $i = 0 ; $i <= abs $E ; $i++ ) {
        $D->{ $K + int( $i * $F / $E ) }->{ $I + int( $i * $G / $E ) }
    }
}

sub p {
    $D = {};
    $Z = $z || .01;
    map {
        $H = $_;
        $I = $N = j $H;
        $K = $O = j $H;
        while ($H) {
            $q = ord substr( $H, 0, 1, "" );
            if ( 42 == $q ) { $J = j $H; $L = j $H}
            else {
```

Code sample from https://en.wikipedia.org/wiki/Just_another_Perl_hacker

Binary

`hexdump -C BabyTux.png`

vim-common installed?
xxd converts to AND from

```
00000000  89 50 4e 47 0d 0a 1a 0a  00 00 00 0d 49 48 44 52  |.PNG........IHDR|
00000010  00 00 03 20 00 00 03 20  08 02 00 00 00 54 12 91  |... ... .....T..|
00000020  3f 00 00 00 09 70 48 59  73 00 00 0b 13 00 00 0b  |?....pHYs.......|
00000030  13 01 00 9a 9c 18 00 00  00 07 74 49 4d 45 07 d5  |..........tIME..|
00000040  01 11 08 19 21 82 98 d6  88 00 00 00 06 62 4b 47  |....!........bKG|
00000050  44 00 ff 00 ff 00 ff a0  bd a7 93 00 02 3e 02 49  |D............>.I|
00000060  44 41 54 78 da ec dd 79  8c 95 55 9a c7 f1 73 58  |DATx...y..U...sX|
00000070  05 97 10 08 8a 22 62 6b  b4 85 1e 57 68 95 c4 ad  |......"bk...Wh...|
00000080  c7 d1 f4 88 31 8e 5b ec  49 94 9e 8c b6 33 3d ae  |....1.[.I....3=.|
00000090  a3 f6 88 62 e2 96 4c db  46 d1 c6 6d 74 fc 47 4d  |...b..L.F..mt.GM|
000000a0  5c 9a 3f 24 2e dd da ae  a3 26 38 30 74 b7 14 34  |\.?$.....&80t..4|
000000b0  41 d2 0a 41 6c 70 21 ca  62 b1 d5 99 97 ba f7 be  |A..Alp!.b.......|
000000c0  ef f3 3c e7 9c f7 de 72  05 fc 7e 12 05 aa ea 6e  |..<....r..~....n|
000000d0  55 75 ef fd e5 39 cf 79  8e 0f 21 38 00 00 00 7c  |Uu...9.y..!8...|
000000e0  75 3c 01 0b 00 00 80 80  05 00 00 40 c0 02 00 00  |u<.........@....|
000000f0  20 60 01 00 00 80 80 05  00 00 40 c0 02 00 00 20  | `........@.... |
00000100  60 01 00 00 80 80 05 00  00 40 c0 02 00 00 20 60  |`........@.... `|
00000110  01 00 00 80 80 05 00 00  40 c0 02 00 00 20 60 01  |........@.... `.|
00000120  00 00 10 b0 00 00 00 40  c0 02 00 00 20 60 01 00  |.......@.... `..|
00000130  00 10 b0 00 00 00 40 c0  02 00 00 20 60 01 00 00  |......@.... `...|
00000140  10 b0 00 00 00 40 c0 02  00 00 20 60 01 00 00 10  |.....@.... `....|
00000150  b0 00 00 00 40 c0 02 00  00 20 60 01 00 00 10 b0  |....@.... `.....|
00000160  00 00 00 08 58 00 00 00  20 60 01 00 00 10 b0 00  |....X... `......|
00000170  00 00 08 58 00 00 00 20  60 01 00 00 10 b0 00 00  |...X... `.......|
00000180  00 08 58 00 00 00 20 60  01 00 00 10 b0 00 00 00  |..X... `........|
00000190  08 58 00 00 00 20 60 01  00 00 10 b0 00 00 00 08  |.X... `.........|
000001a0  58 00 00 00 04 2c 00 00  00 10 b0 00 00 00 08 58  |X....,.........X|
000001b0  00 00 00 04 2c 00 00 00  10 b0 00 00 00 08 58 00  |....,.........X.|
000001c0  00 00 04 2c 00 00 00 10  b0 00 00 00 08 58 00 00  |...,.........X..|
```

https://commons.wikimedia.org/wiki/File:BabyTuxAlpha.png

ordering
presentation
compare
subst / xform

Streaming

Filtering

Transform

Mining

- **diff** - line by line comparison
- **wdiff** - word based diff
- Format intelligent diffs also available

diff / wdiff

Sonnet 18

"Shall I compare thee to a **summer's** day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds **of May**,
And **summer's** lease hath all too short a date".

Sonnet 18

"Shall I compare thee to a **Winter's** day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds **of May**,
And **Summer's** lease hath all too short a date".

diff

3c3
< "Shall I compare thee to a summer's day?

> "Shall I compare thee to a Winter's day?
5,6c5,6
< Rough winds do shake the darling buds of May,
< And summer's lease hath all too short a date".

> Rough winds do shake the darling buds of May,
> And Summer's lease hath all too short a date".

-i Ignore case
Omits line 6

wdiff

Sonnet 18

"Shall I compare thee to a **[-summer's-] {+Winter's+}** day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And **[-summer's-] {+Summer's+}** lease hath all too short a date".

wdiff -3

```
==========================================
 [-summer's-] {+Winter's+}
==========================================
 [-summer's-] {+Summer's+}
==========================================
```

-w Ignore whitespace
Omits line 5

jsondiff

apt install python-jsonpatch

jsondiff orig.json new.json | python -m json.tool

orig.json

```
{
  "k1": "abc",
  "a1": [
     "a21",
     "a22",
     "a23",
     "a24"
  ],
  "k2" : { "k2.1": "v2.1" }
}
```

new.json

```
{
  "k2" : { "k2.1": true,
  "k2.2": "v2" },
  "k1": "abc",
  "a1": [ "a21", "a21a", "a22", "a24" ]
}
```

```
[
    {
        "op": "add",
        "path": "/a1/1",
        "value": "a21a"
    },
    {
        "op": "remove",
        "path": "/a1/3"
    },
    {
        "op": "replace",
        "path": "/k2/k2.1",
        "value": true
    },
    {
        "op": "add",
        "path": "/k2/k2.2",
        "value": "v2"
    }
]
```

ordering
presentation
compare
subst / xform

Streaming

Filtering

Mining

Transform

- **tr** - character replacement or removal.
- **sed** - single pass stream processing
- **AWK** - general purpose text extraction and xform
- **Perl** - general purpose
- **Python** - general purpose

tr

	A Small Sample
Substitute character for character	cat -> Dog **tr 'act' 'oDg'** a -> o c -> D t -> g
Remove repeated occurrences of specified letters	-s squash aapplle -> apple **tr -s 'ale'**
Remove non-printable characters	-c !string1 -d delete **tr -cd "[:print:]"**
Parse out character strings, one per line	**tr -cs "[:alpha:]" "\n"**

tr

	Special Values
Octal (base 8) characters	TAB (ASCII dec 9) = \9, \09, **\009** A (ASCII dec 65) = \101
Hex (base 16) characters	TAB (ASCII dec 9) = \x9, **x09** A (ASCII dec 65) = \x42
Special escape codes	\f (formfeed) \n (newline) \r (carriage return) \t (tab)
Range	m-q = mnopq 2-6 = 23456
Character classes*	[:alnum:] = [a-zA-Z0-9] [:alpha:] = [a-zA-Z] [:blank:] = [\t] [=C=] = locale equivalence class for char C eg. French [=e=] is e, é, è and ê

* Reference http://www.regular-expressions.info/posixbrackets.html

AWK

	A Small Sample
Execute on matching line	/*pattern*/ {*action*}
Print 3rd and 5th whitespace delimited column	**{ print $3, $5 }**
Count fields across all rows	**{total=total+NF};END{print total}**

* For more: https://www.shortcutfoo.com/app/dojos/awk/cheatsheet

Perl

Command line Option	Equivalent
-e “EXPR”	Evaluate inline EXPR rather then open named file
-n -e “EXPR”	**while (< >) {** EXPR **}** looping over < > reads successive lines of stdin to $_
-p -e “EXPR”	**while (< >) {** PROG**; print $_ }**
-a -n (similarly -a -p)	**while (< >) { @F=split;** PROG**; }**
-a -n -F/,/	**while (< >) { @F=split /,/, $_;** PROG**; }**

Perl

	A Small Sample
Transform if matched	**perl -pe 's/apple/banana/ if /red/‘** perl -e "while (<>) { s/apple/banana/ if /red/; print}”
Print 4th and 6th column	**perl -aF/\\s+/ -ne 'print "$F[3], $F[5]\n”'** perl -e 'while(<>) { @F=split /\s+/; print "$F[3}, $F[5]\n" }'
Sort by line length	**perl -e 'print sort {length $a <=> length $b} <>'**
Translate lower to uppercase and newline to comma	cat\ndog -> CAT,DOG **perl -pe 'tr/a-z\n/A-Z,/'**

format specific scraping
conversion
nlp

Streaming

Filtering

Transform

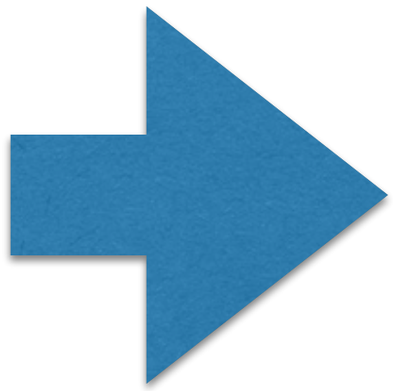
Mining

- **file** - identify file header
- **pandoc** - convert between common document formats
- **pdftotext** - extract text from PDF
- **xls2csv** / **xlsx2csv** - convert Excel format to CSV
- **strings** - Extract printable characters from a file
- **tesseract** - OCR

Pandoc

From

CommonMark
DocBook
EPUB
Emacs Org mode
GitHub-Flavored Markdown
HTML
Haddock markup
LaTeX
Markdown
MediaWiki markup
ODT
OPML
PHP Markdown Extra
TWiki markup
Textile
Word docx
reStructuredText
txt2tags



To

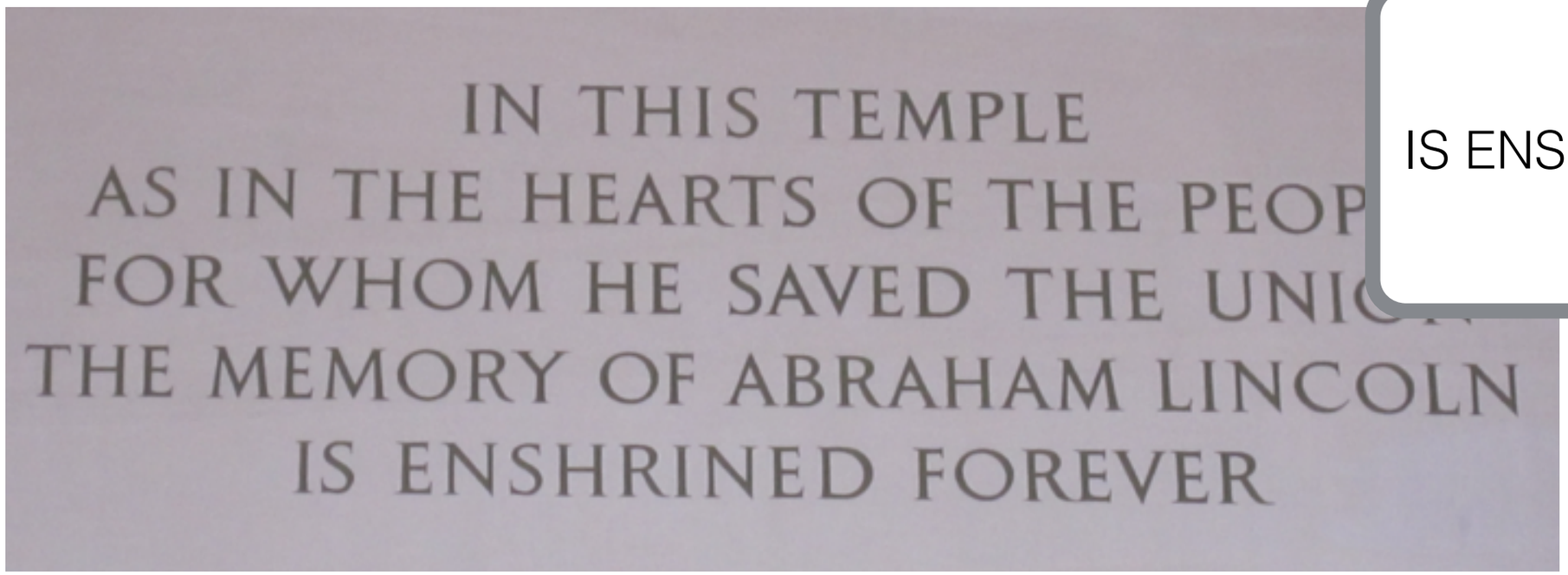
AsciiDoc
CommonMark
ConTeXt
DZSlides
DocBook
DokuWiki markup
EPUB (v2 or v3)
Emacs Org mode
FictionBook2
GNU Texinfo
GitHub-Flavored Markdown
HTML5
Haddock markup
InDesign ICML
LaTeX
Mark-down
MediaWiki markup
ODT
OPML
OpenDocument
PDF
PHP Markdown Extra
RTF

OCR

apt install tesseract-oct

tesseract lincoln.jpg stdout

jpg



text

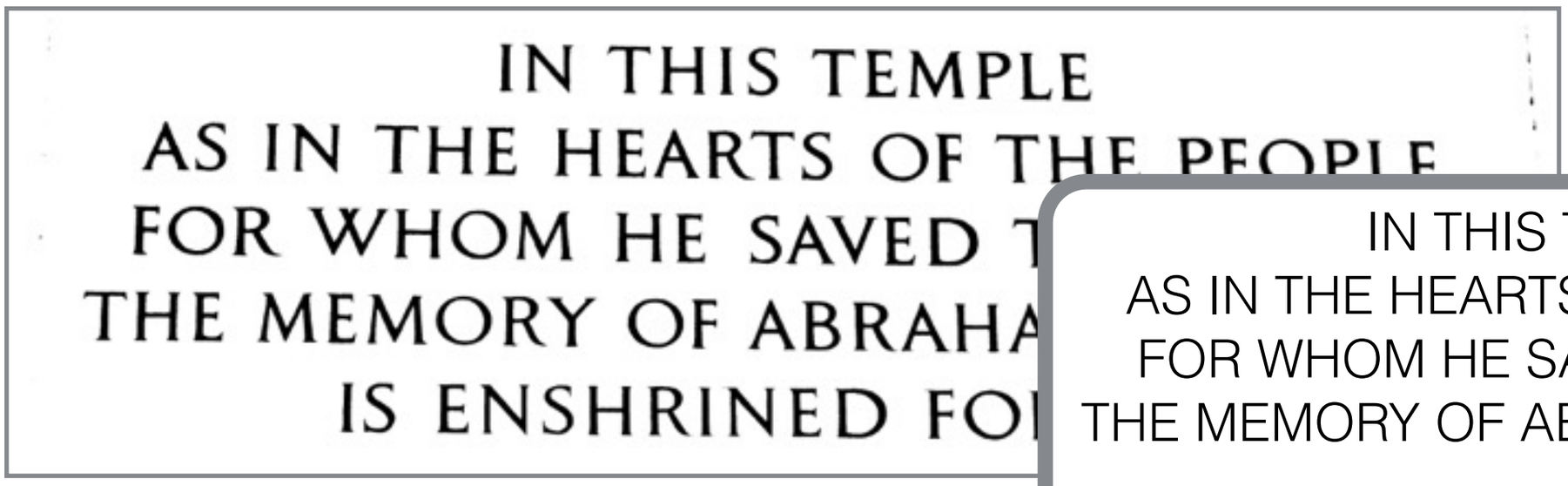
IS ENSHRINED FOREVER

Only one line extracted

Not perfect yet. Perhaps more prefiltering?

Processed with imagemagick 2colorthresh:

jpg



text

IN THIS TEMPLE
AS IN THE HEARTS OF THE PEOPLE
FOR WHOM HE SAVED THE UNION
THE MEMORY OF ABRAHAM LINCOLN ‘

IS EN S H RIN E D "FOREVER

sumy

pip install sumy

sumy lsa --url='http://www.examenglish.com/PTE/pte_academic_writing_1.htm' | perl -pe 'print "\n"'

Public figures include politicians and other public officials such as judges and civil servants as well as celebrities such as film stars, musicians and sports stars. The very nature of these roles opens these people to scrutiny by the press. The extent to which the media are legally free to investigate and publish details of public figures' private lives varies from country to cou
France are much stricter on protecting personal privacy than
People have a right to know about those in power. Whether
case of politicians and civil servants, or by revenue generat
sporting appearances or concerts in the case of celebrities
is dependant on the general public.
People have the right to make informed judgements about t
have. Attempts to restrict what may be reported about publ
could easily become a conspiracy to keep voters in the dar
them. All elections are to a greater or lesser extent about th
leading politicians involved. Unless the voters are allowed i
lives they will lack the information needed to make a fair de
booth. For example, some people believe that a politician w
an affair was equally capable of breaking his promises and
Exposing corruption and dishonesty on the part of public of
a critical part of the function of a free press, and it is essent
free-market economy. If investigative journalists are prevent
private lives of public figures, then corruption and crime wil
Public figures know that with fame comes a price and that p
In fact, many celebrities actively seek media exposure in or
careers, revealing to the media many aspects of their perso
has been bought in such a fashion it is then somewhat hyp
"press intrusion" into those few aspects the star would pr

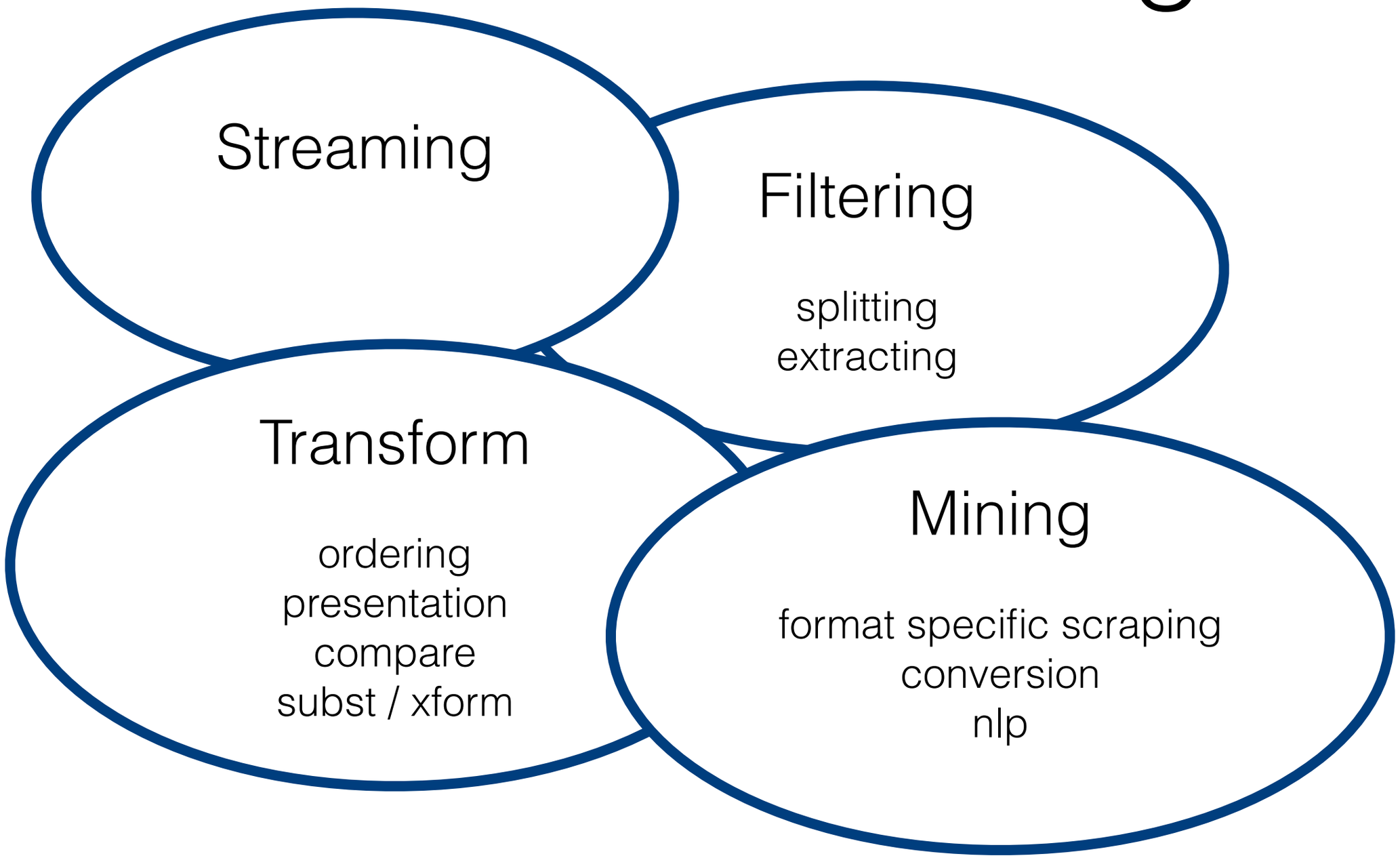
Attempts to restrict what may be reported about public figures in the press could easily become a conspiracy to keep voters in the dark and to manipulate them.

Unless the voters are allowed insights into their private lives they will lack the information needed to make a fair decision at the polling booth.

For example, some people believe that a politician who betrayed his wife in an affair was equally capable of breaking his promises and lying to his country.

If investigative journalists are prevented from scrutinising the private lives of public figures, then corruption and crime will be much easier to hide.

Text Processing Tools



Examples…
Questions and Discussion