Everybody’s got to cheat sometimes...

Christophe Bontemps
Toulouse School of Economics, INRA

@Xtophe_Bontemps

TSE

INRA
SCIENCE & IMPACT

My Job

Above all, show the data

In his seminal book, Edward Tufte proposed some rules for good visualisation practices.

- Truthfulness

Above all, show the data

In his seminal book, Edward Tufte proposed some rules for good visualisation practices.

- Truthfulness
- Data transparency (*”see the data”*)

Above all, show the data

In his seminal book, Edward Tufte proposed some rules for good visualisation practices.

- Truthfulness
- Data transparency (*"see the data"*)
- Minimize the *"lie factor"*

Above all, show the data

In his seminal book, Edward Tufte proposed some rules for good visualisation practices.

- Truthfulness
- Data transparency (*"see the data"*)
- Minimize the *"lie factor*"
- Maximize the *“data-ink ratio”*

Above all, show the data

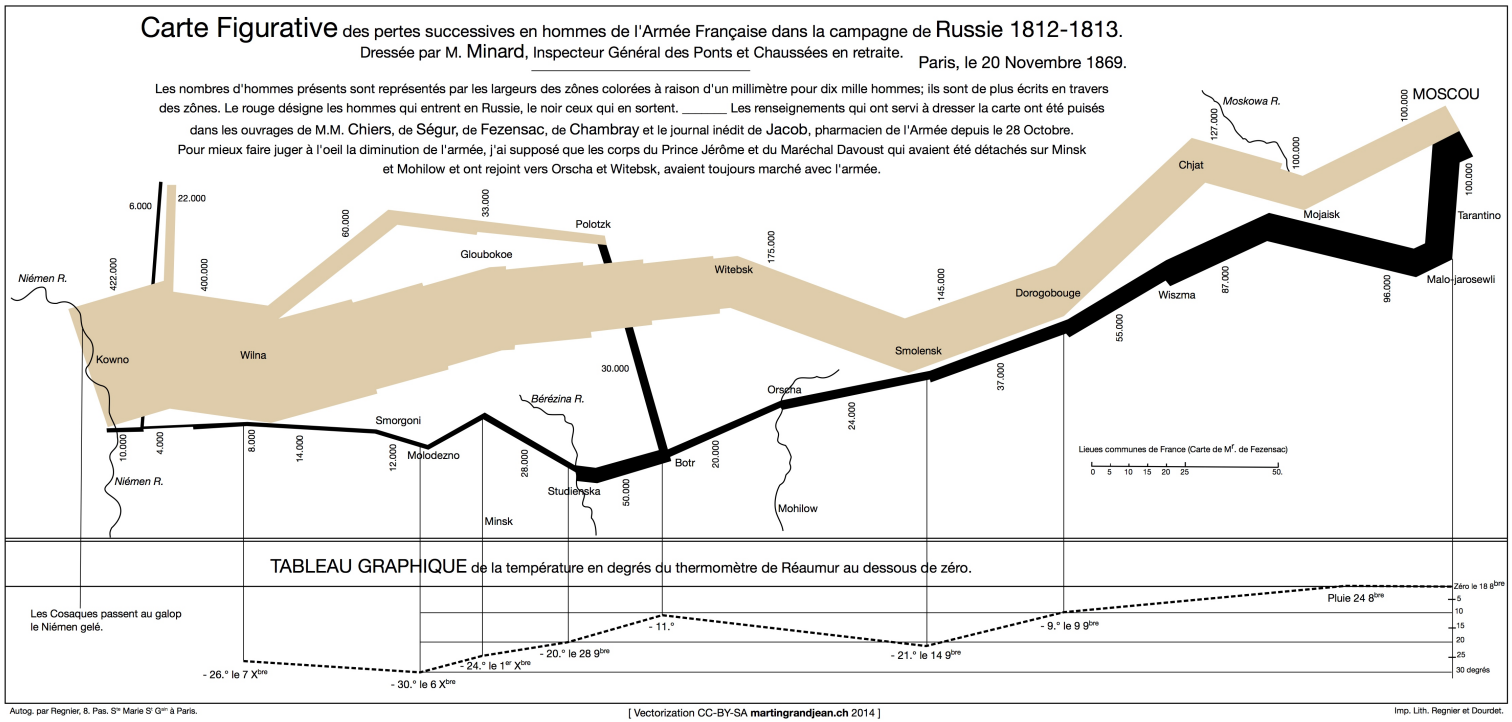
In his seminal book, Edward Tufte proposed some rules for good visualisation practices.

- Truthfulness
- Data transparency (*"see the data"*)
- Minimize the *"lie factor*"
- Maximize the *“data-ink ratio”*
- ...

EARLY "CHEATERS"

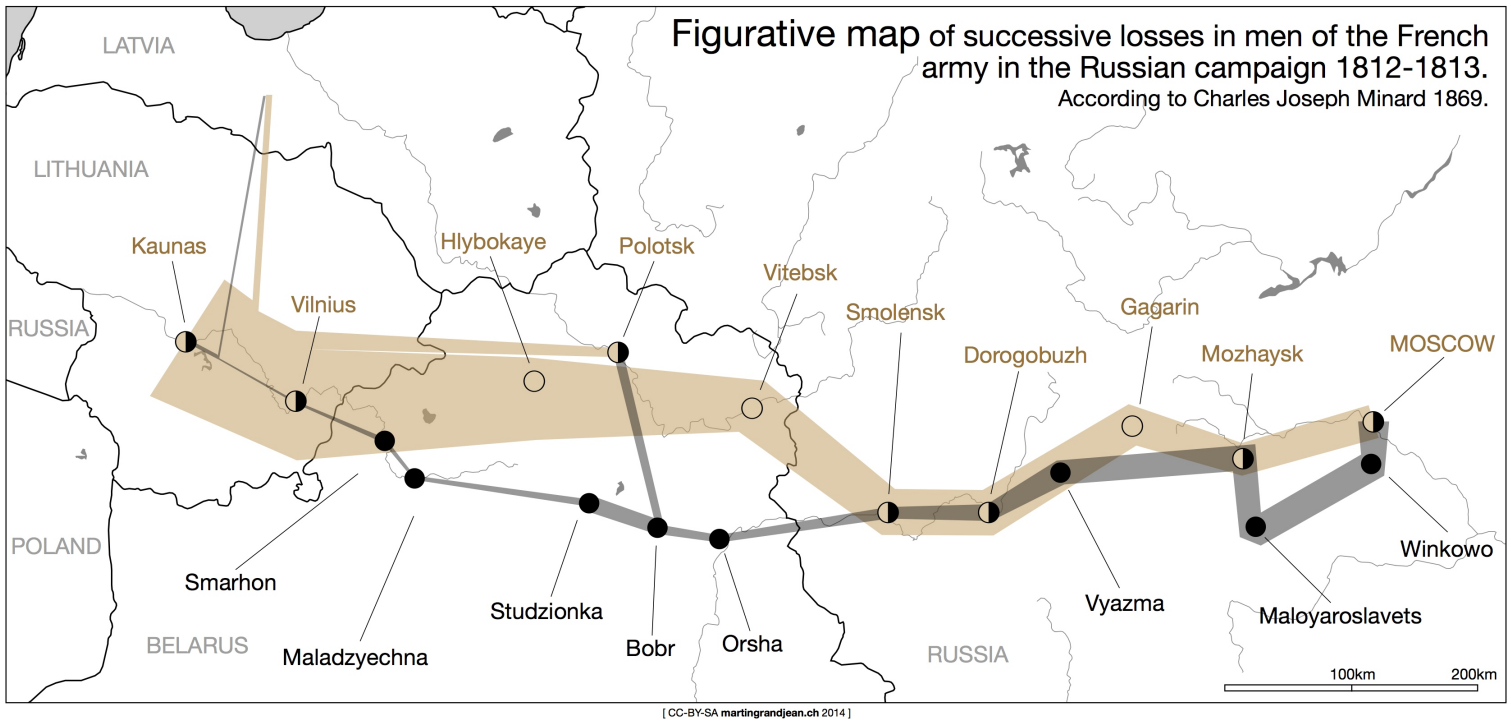
Charles Minard (1869) map graph was quoted by E. Tufte as

"the best statistical graphic ever drawn".



Early "cheaters"

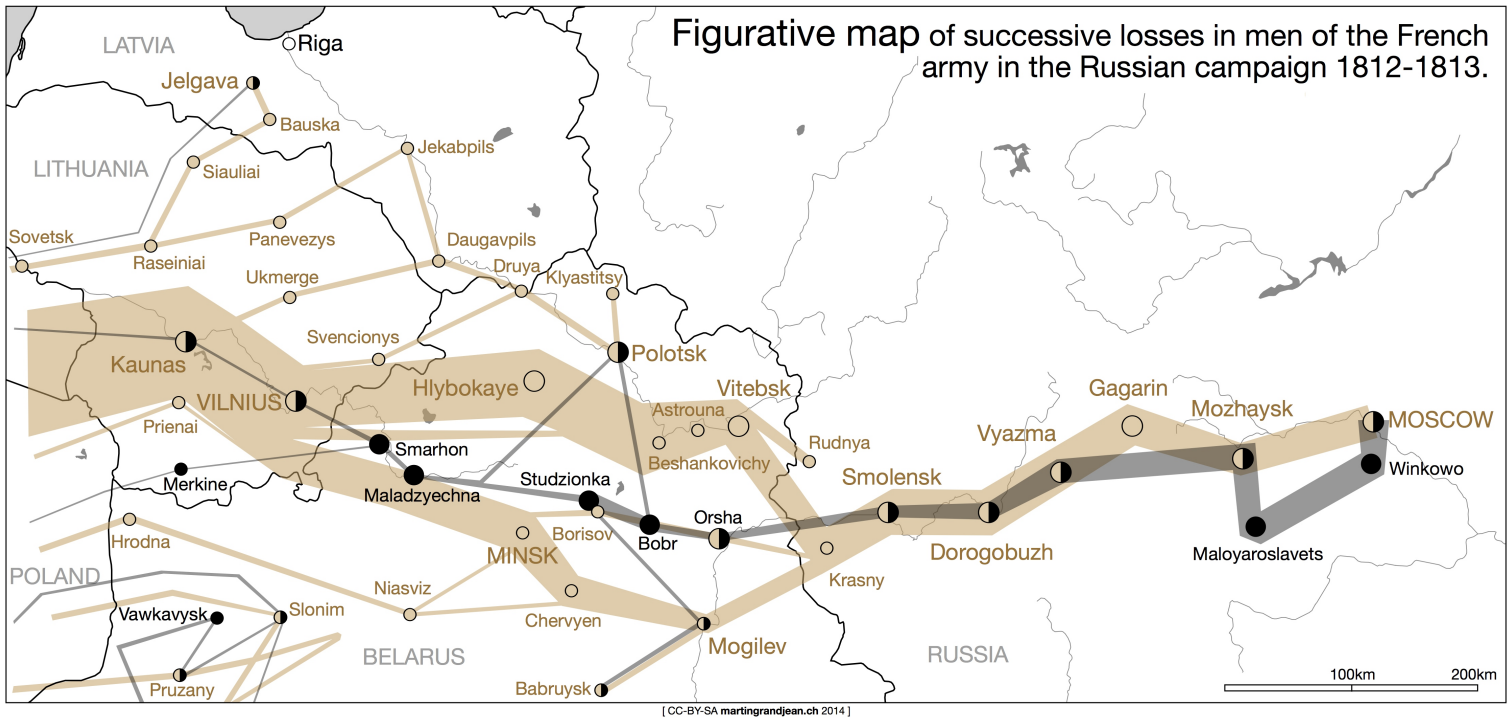
But Charles Minard has *cheated* a little bit :
First on geography



Source Martin Grandjean

EARLY "CHEATERS"

But Charles Minard has *cheated* a little bit :
Second on the historical data and army path



Source Martin Grandjean

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the “great army”

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the “great army”

→ A single stream going and returning

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the “great army”

$\rightarrow$ A single stream going and returning

- Temperatures plotted at irregular intervals.

Early “cheaters”

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the “great army”

$\rightarrow$ A single stream going and returning

- Temperatures plotted at irregular intervals.

...

EARLY "CHEATERS"

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the "great army"

$\rightarrow$ A single stream going and returning

- Temperatures plotted at irregular intervals.

...

In reality, Charles Minard simplified the data representation on purpose (see Martin Grandjean for details).

EARLY "CHEATERS"

As data suggest (Martin Grandjean), the story was a little bit more complex and we can find approximations in the map :

- Lack of projection reference (as quoted by Michael Friendly)
- Towns and places not mentioned in the map
- Arrangements in the aggregation of flows leaving and then rejoining the "great army"

→ A single stream going and returning

- Temperatures plotted at irregular intervals.

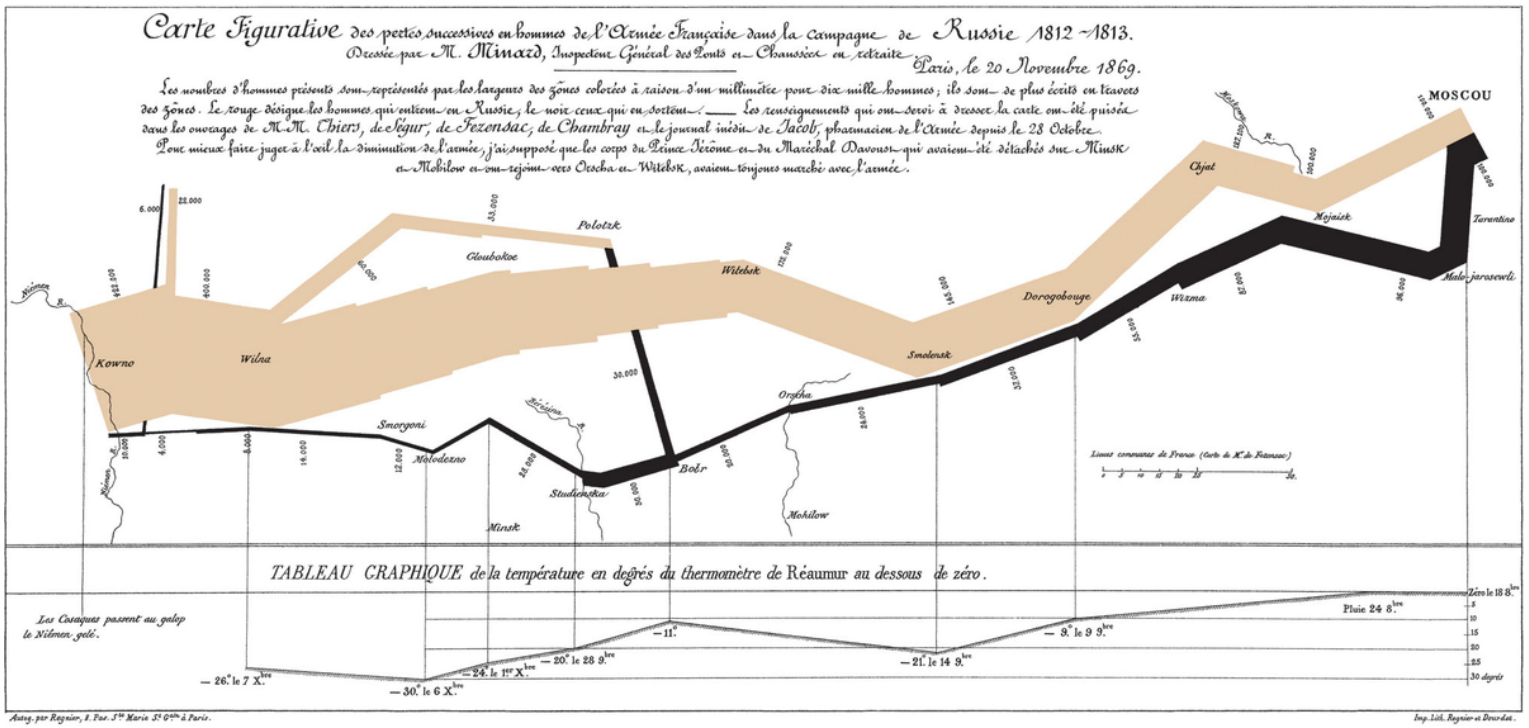
...

In reality, Charles Minard simplified the data representation on purpose (see Martin Grandjean for details).

The result is a clear storytelling map (a brilliant one !)

EARLY "CHEATERS"

Charles Minard (1869) original map :



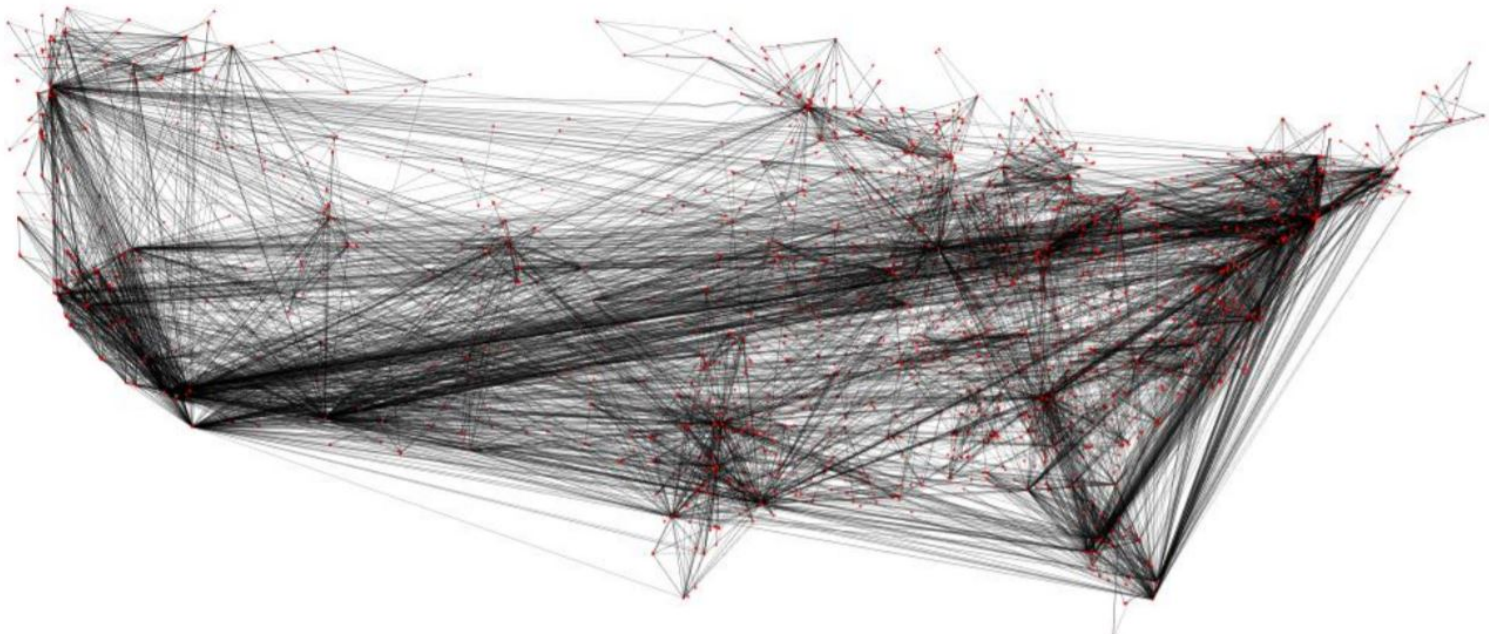
21ST CENTURY "CHEATERS"

Let us take true data and visualize US airlines map (or graph)

Source Christophe Hurter

21ST CENTURY "CHEATERS"

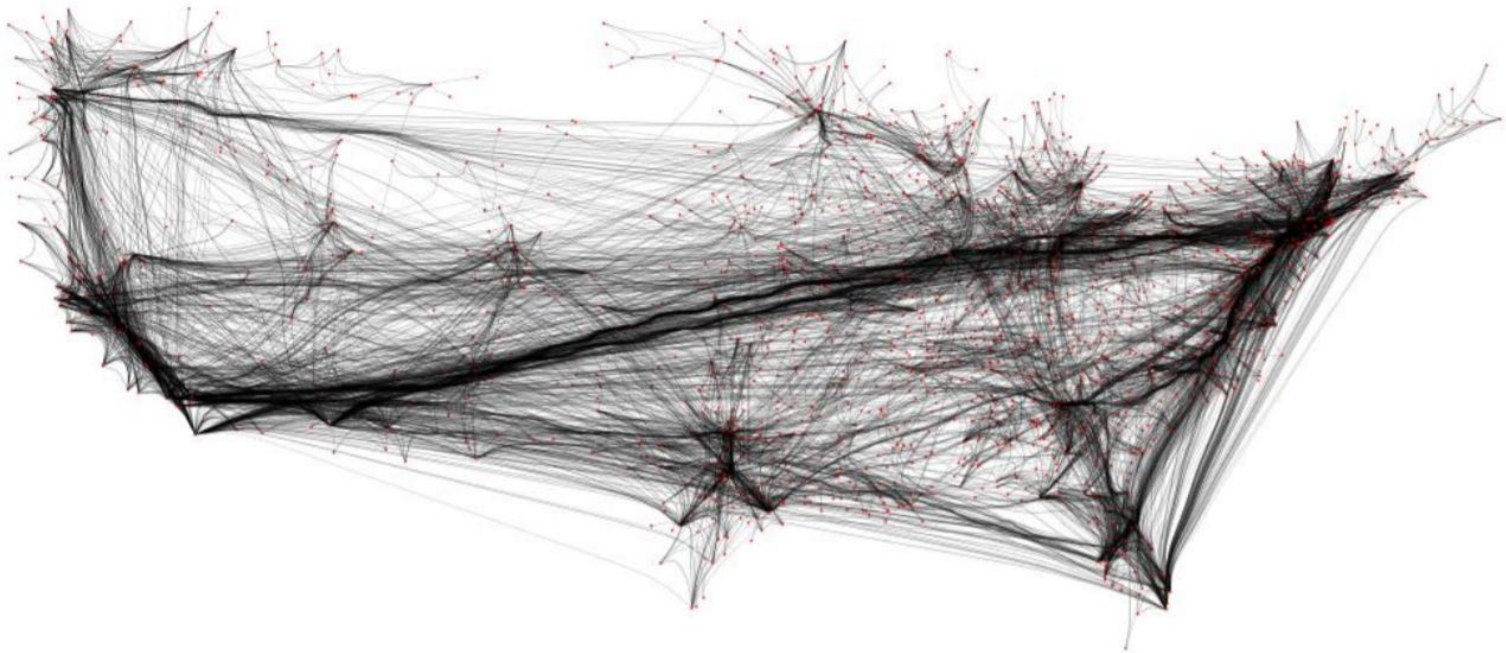
Let us take true data and visualize US airlines map (or graph)



Source Christophe Hurter

21ST CENTURY "CHEATERS"

To simplify the understanding, researcher propose bundling techniques. [1]

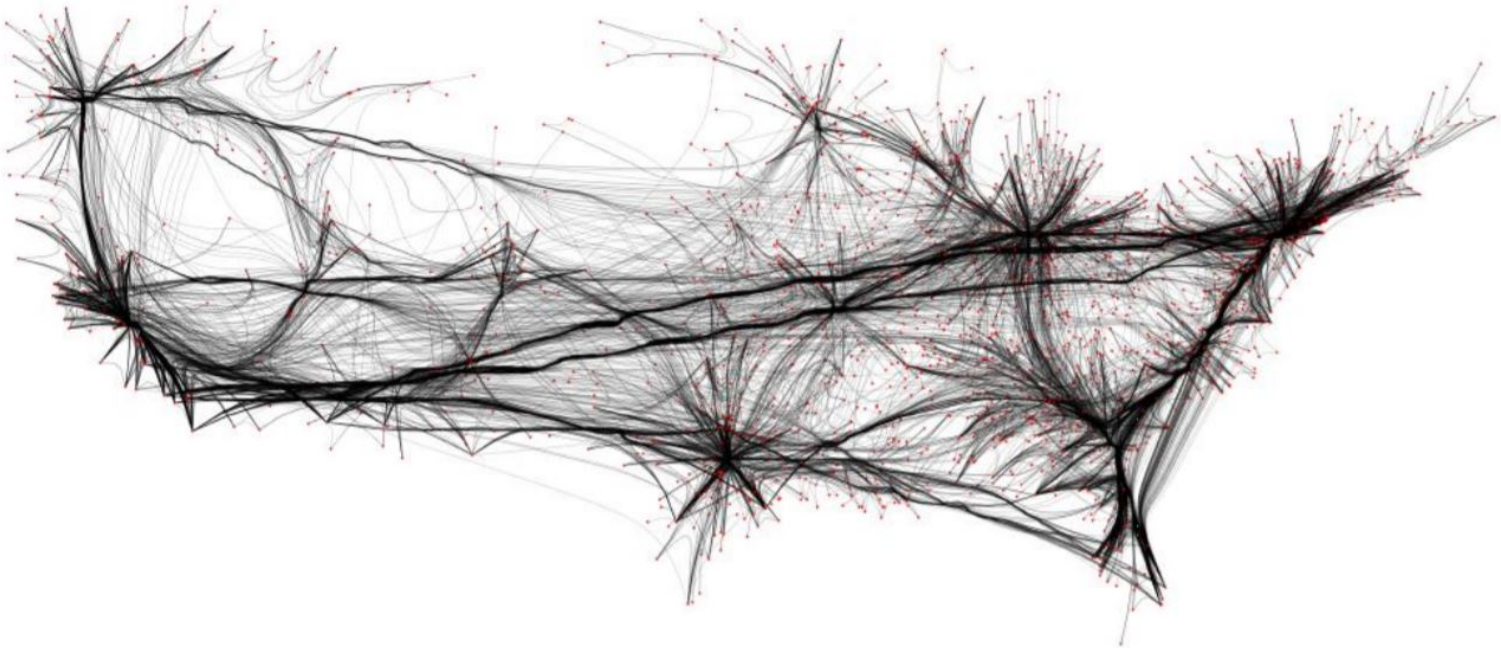


Source Christophe Hurter

1. To bundle : to tie or gather things together

21st century "cheaters"

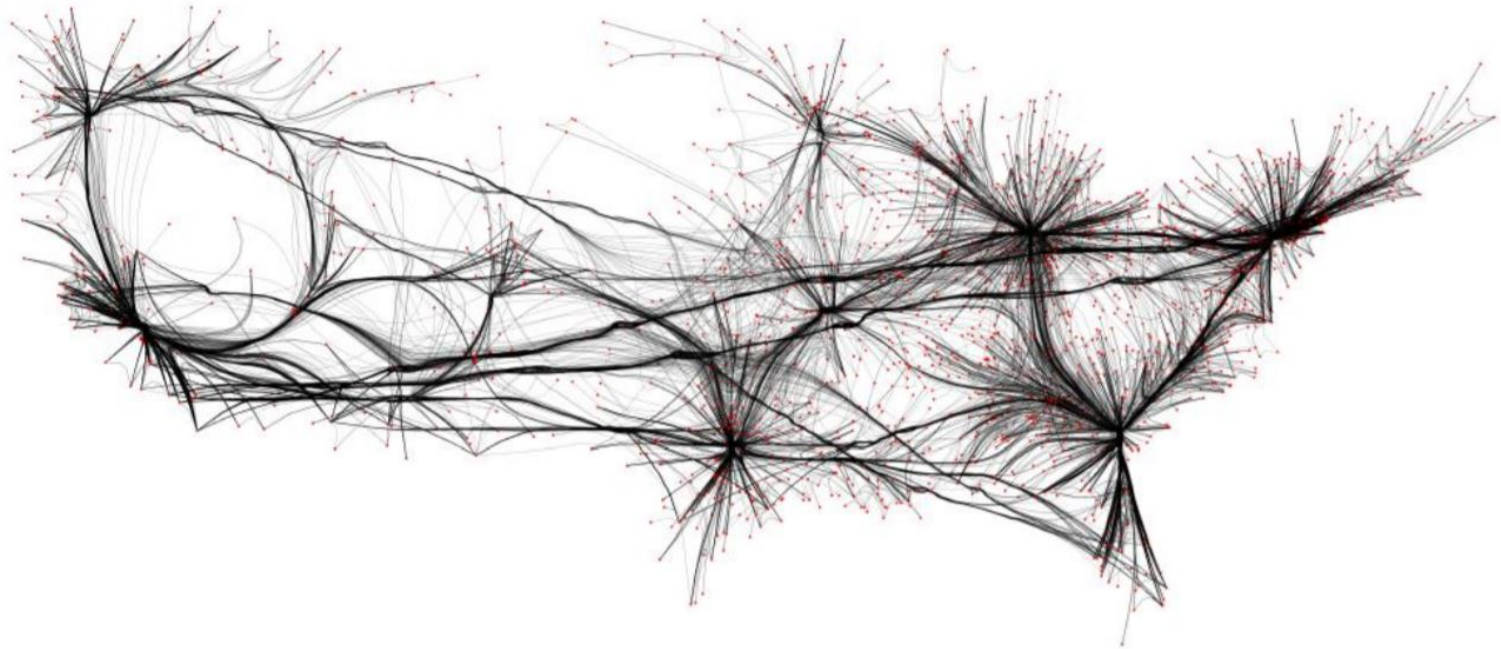
Edges get closer and density gets sharper :



Source Christophe Hurter

21ST CENTURY "CHEATERS"

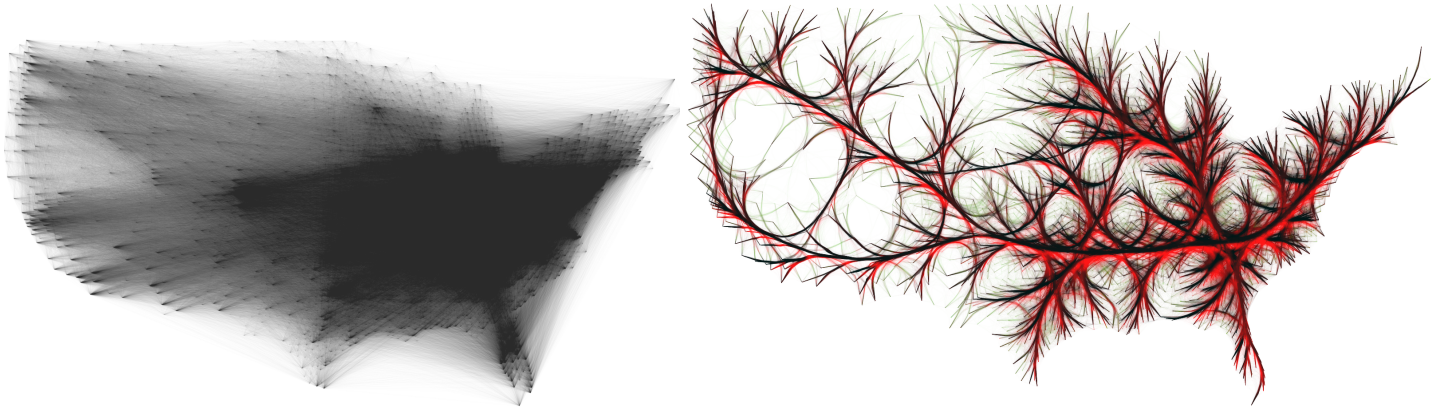
Edges get even closer and density gets even sharper :



Source Christophe Hurter

21ST CENTURY "CHEATERS"

At the end of the day...one can see through darkness !



Source Christophe Hurter

21ST CENTURY "CHEATERS"

The process is more data-driven here (kernel smoothing), but is based on the same leading idea. To simply, to see patterns, structure and to tell a story, even if :

- None of these routes correspond to real traffic routes

21ST CENTURY "CHEATERS"

The process is more data-driven here (kernel smoothing), but is based on the same leading idea. To simply, to see patterns, structure and to tell a story, even if :

- None of these routes correspond to real traffic routes
- Edges (airports) may have moved (geography affected)

21ST CENTURY "CHEATERS"

The process is more data-driven here (kernel smoothing), but is based on the same leading idea. To simply, to see patterns, structure and to tell a story, even if :

- None of these routes correspond to real traffic routes
- Edges (airports) may have moved (geography affected)
- Colours artificial

21ST CENTURY "CHEATERS"

The process is more data-driven here (kernel smoothing), but is based on the same leading idea. To simply, to see patterns, structure and to tell a story, even if :

- None of these routes correspond to real traffic routes
- Edges (airports) may have moved (geography affected)
- Colours artificial

But

21ST CENTURY "CHEATERS"

The process is more data-driven here (kernel smoothing), but is based on the same leading idea. To simply, to see patterns, structure and to tell a story, even if :

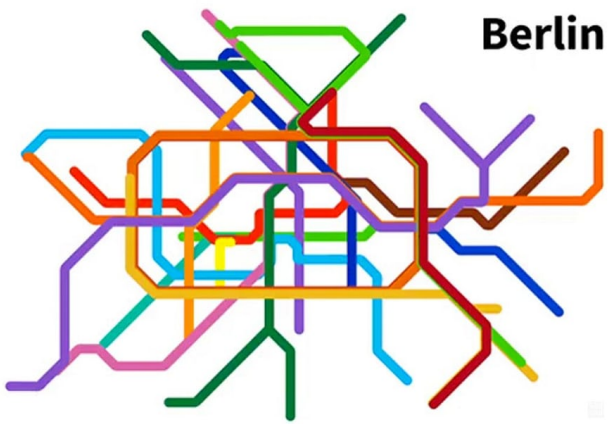
- None of these routes correspond to real traffic routes
- Edges (airports) may have moved (geography affected)
- Colours artificial

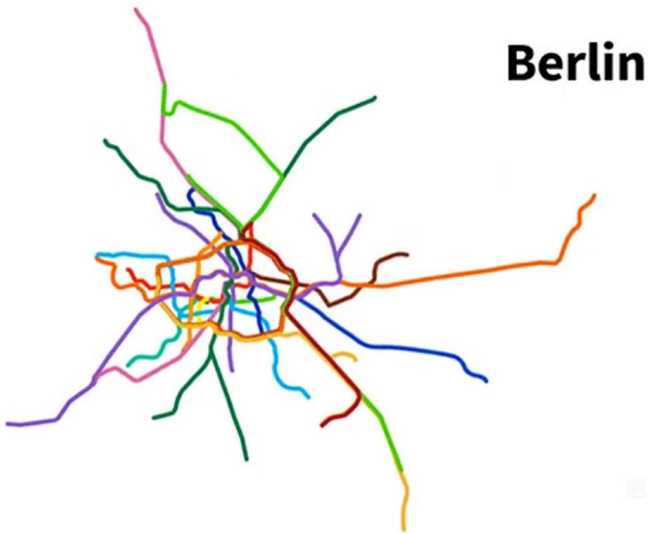
But

- The result is a clear storytelling map (a brilliant one !)

USUAL “CHEATERS”

There are many visualisation that transform the data for clarity : Subway maps for example





Source The Guardian

USUAL "CHEATERS"

Subway maps that match the physical reality are quite rare

Source Benjamin Schmidt

USUAL "CHEATERS"

Ski resort maps are also neither maps, nor pictures, nor projections but paintings !

Source Pierre Novat

We are here in the land of an artist representation of a printed landscape showing both south, east and north-oriented slopes....

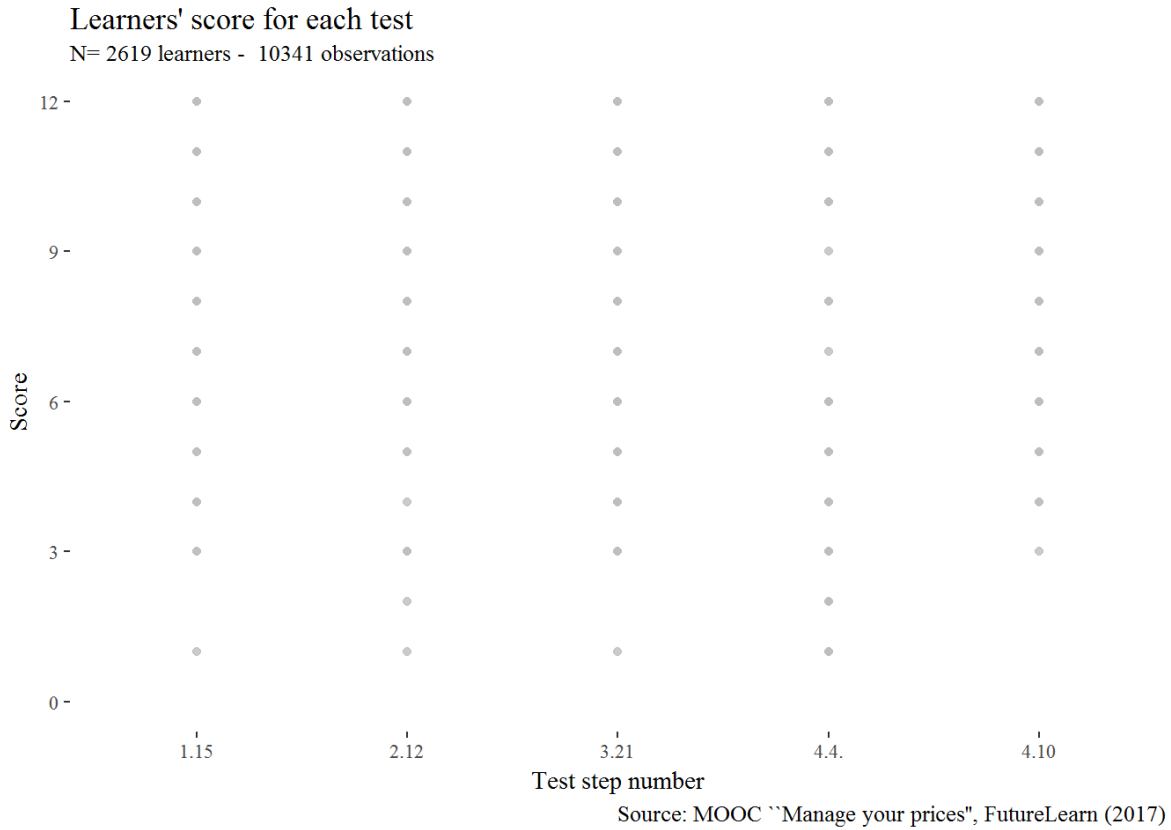
WHY “CHEATING” IS SOMETIMES USEFUL IN DATA SCIENCE

We have a file with 2619 learners, following a MOOC with some "steps" and passing 5 tests :

learner id	step	test score
29e3b4d1-f030-46b7-937e-1f70a3609921	1.15	11
4af89f00-d7d1-41a0-84c4-2e3826363c0d	1.15	9
4af89f00-d7d1-41a0-84c4-2e3826363c0d	3.21	10
53ecc918-1f24-43f7-b3d3-c781d814538e	1.15	7
8059bc2c-fc5a-4392-9db0-f35763183c5f	1.15	11
3db74705-e8c3-42ea-80a2-35ec48f24d83	2.12	12
c55f1f9f-fb94-4e83-84ce-b194b956d4b6	4.10	8
c93990be-c458-42df-b870-f389345380cf	1.15	9
e324839b-c897-467a-b3f6-30c3742afeab	2.12	12
0a35c3a3-60f1-4d13-b40a-84627e96101b	1.15	11
...		

How to detect some patterns, if any ?

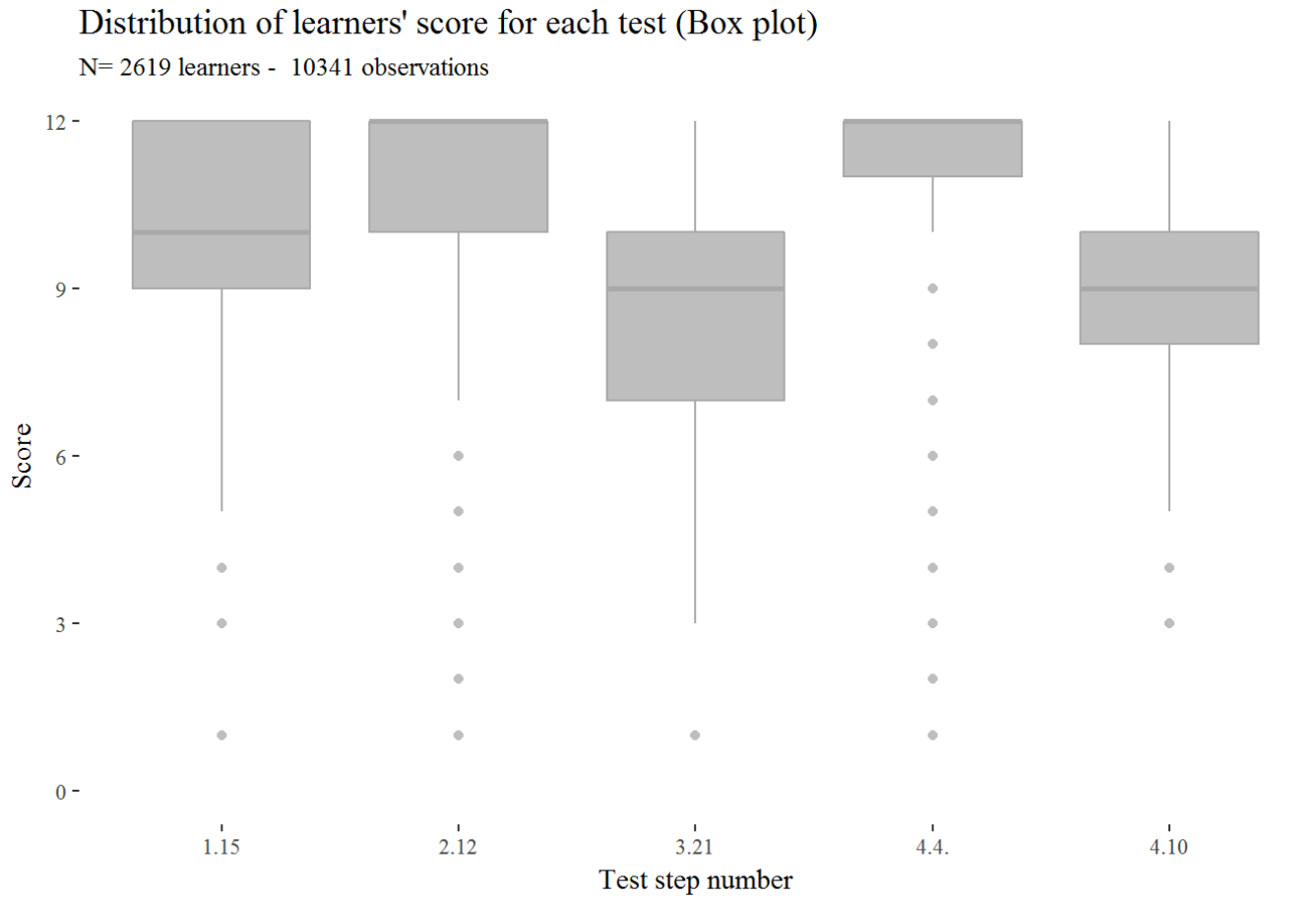
Are there some visible patterns ? Are learners with good results for one test still good at another ? So my first reflex was a plot with all the learners’results over the 5 steps :



BEGINNERS MISTAKE !

With 10341 observation (score x learners) we have a lot of overplotting !

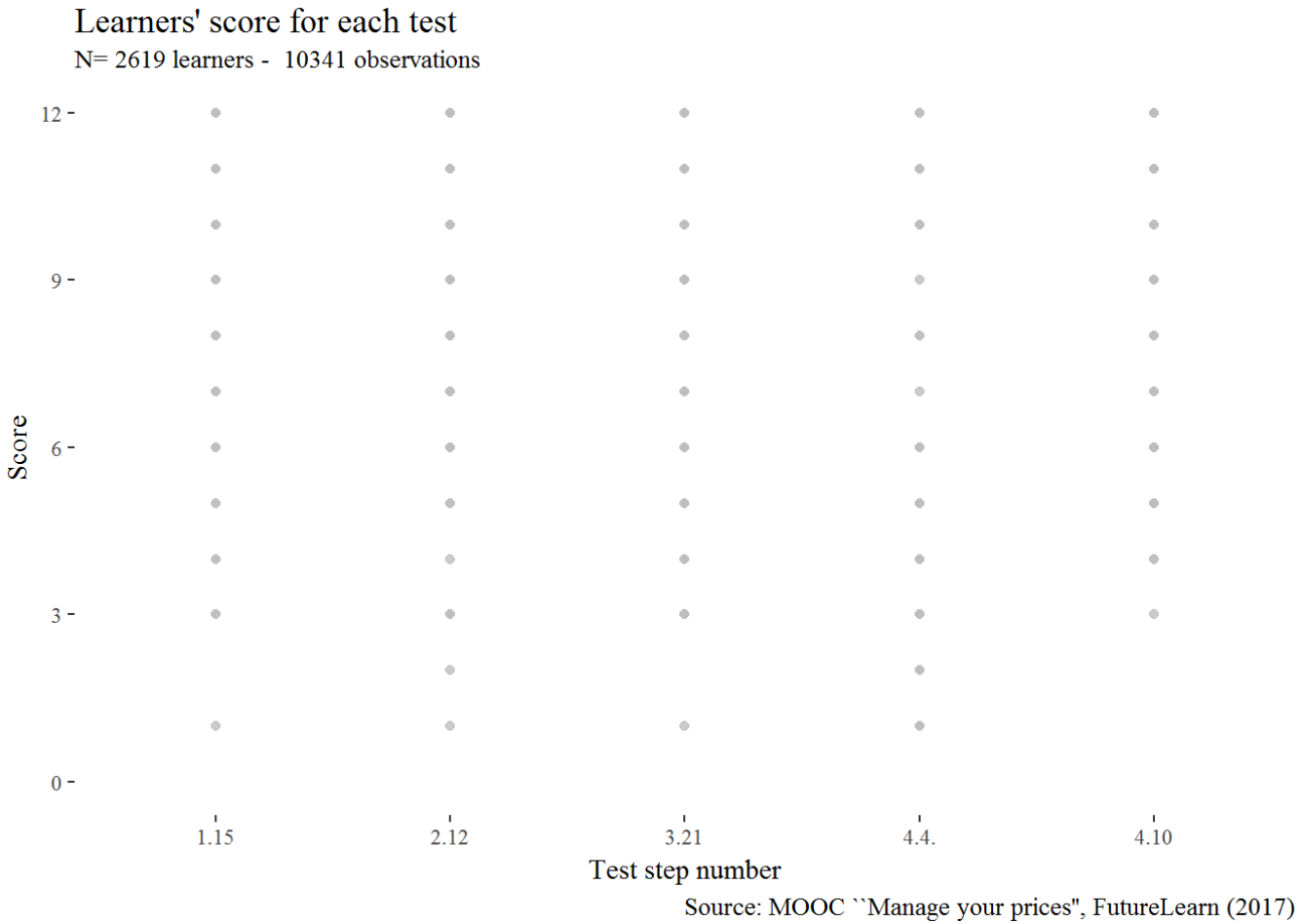
Let us use the good old box-plot



Source: MOOC ``Manage your prices", FutureLearn (2017)

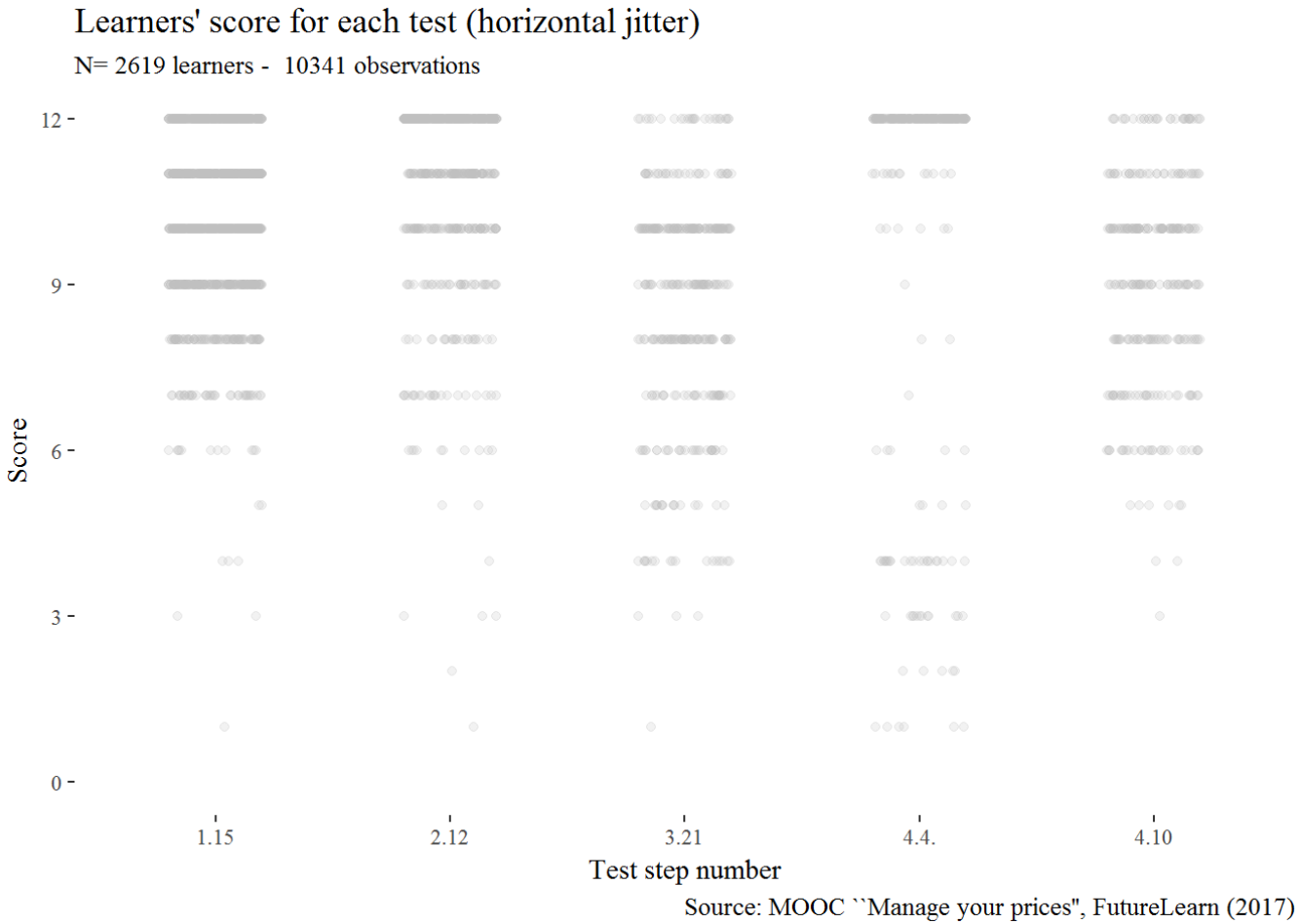
LET’S CHEAT NOW !

Let us add some randomness in the data to avoid overplotting



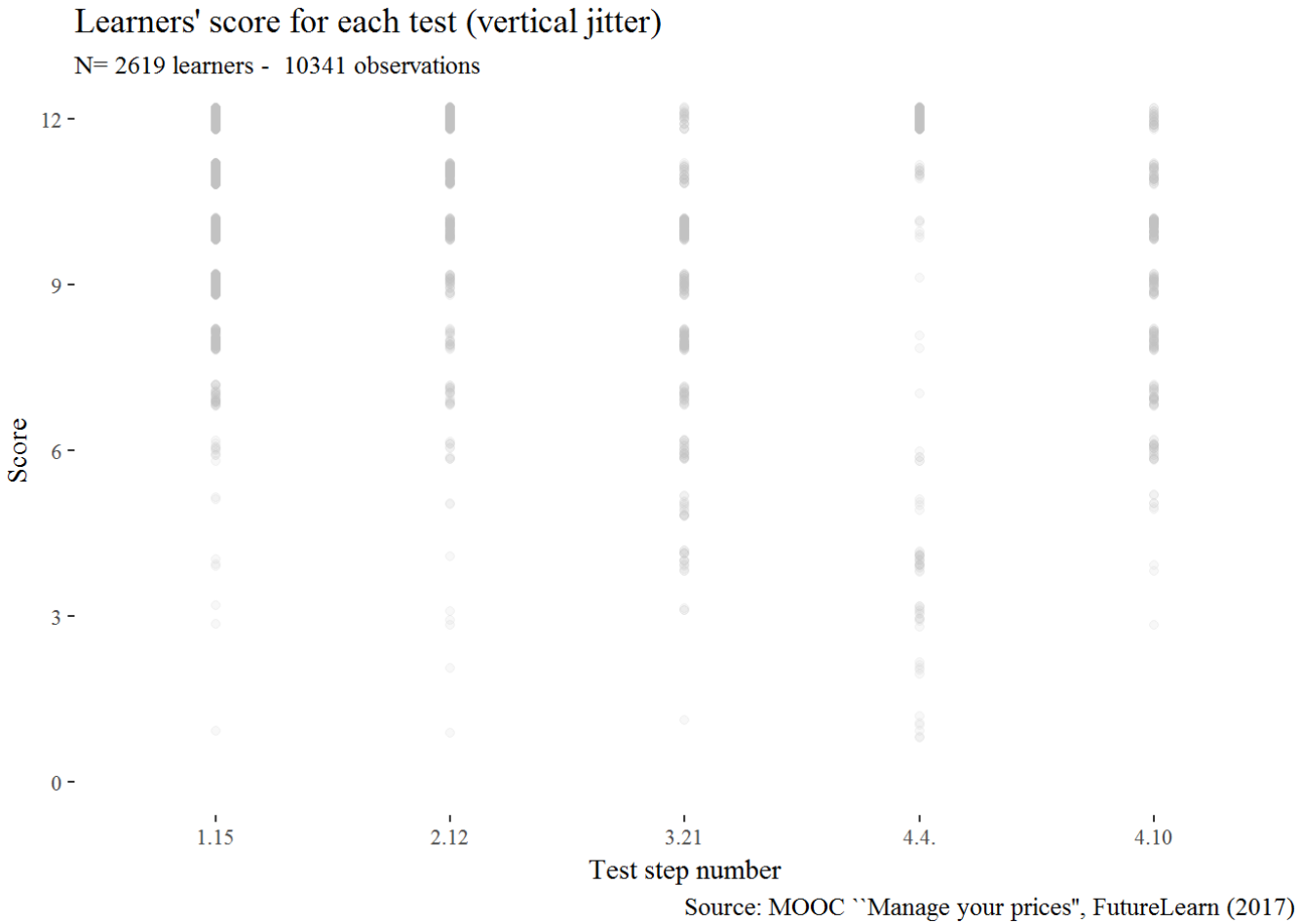
LET’S CHEAT NOW !

Let us add some randomness in the data to avoid overplotting



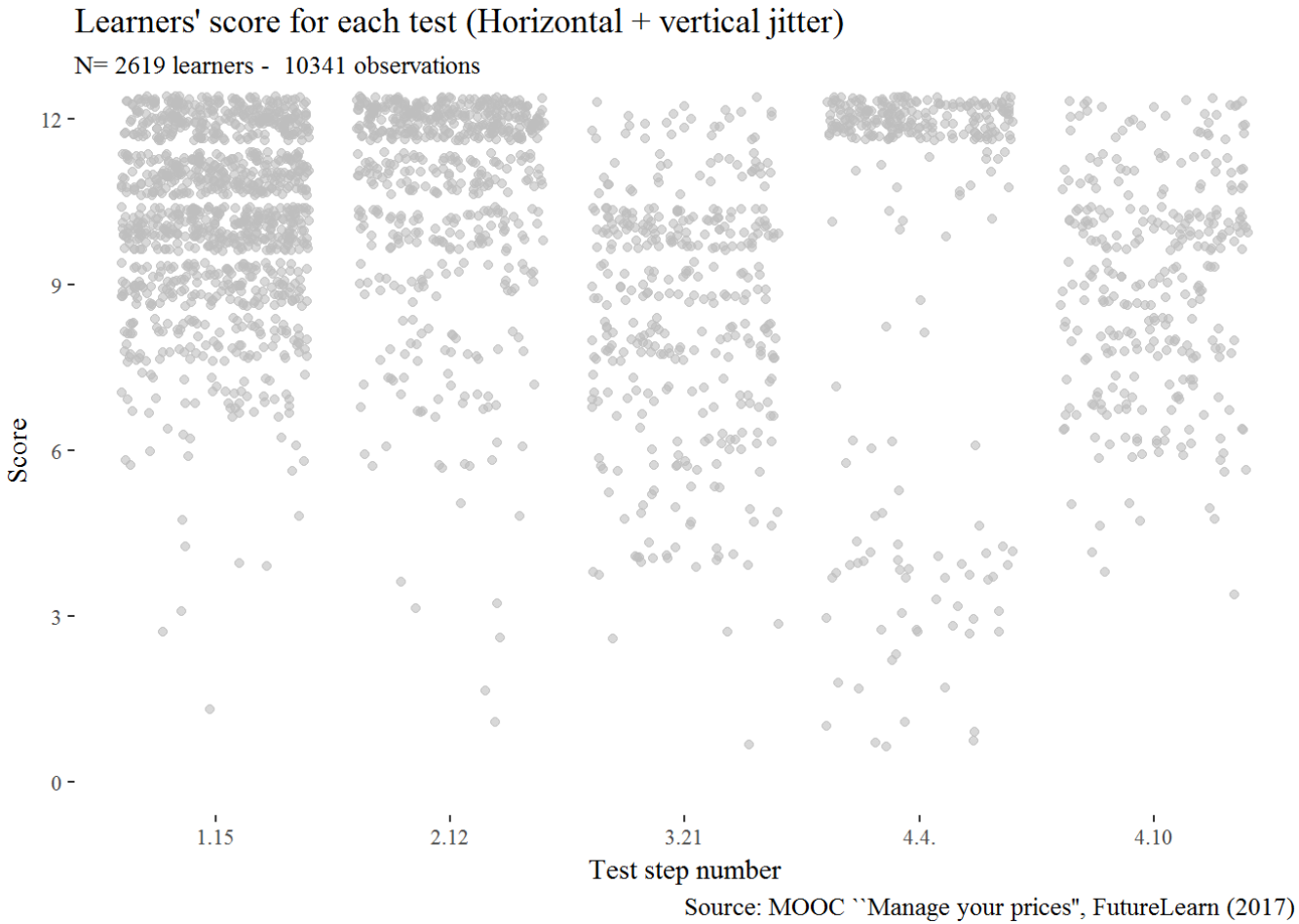
LET’S CHEAT NOW !

Let us add some randomness in the data to avoid overplotting



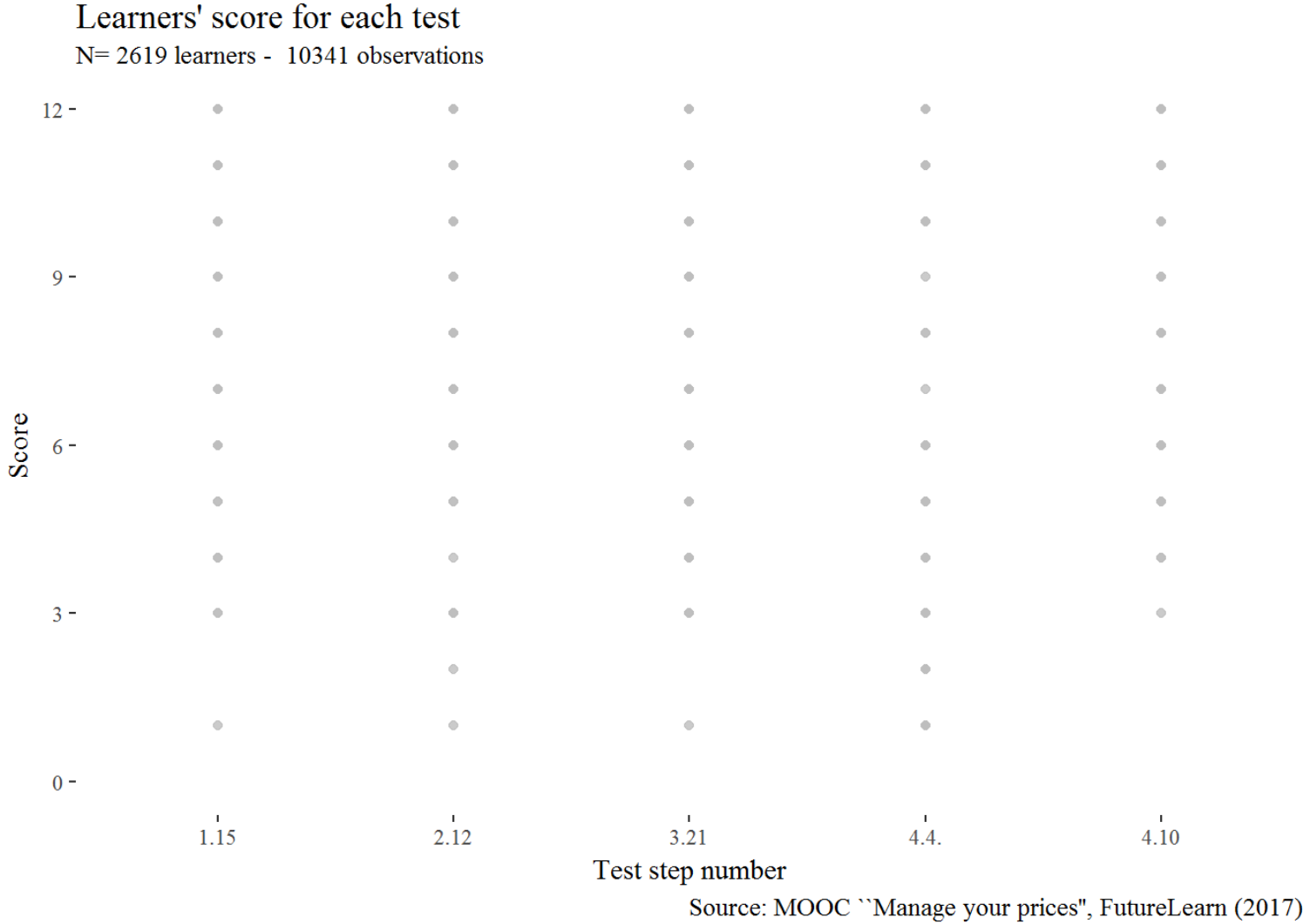
LET’S CHEAT NOW !

Let us add some randomness in the data to avoid overplotting



How to detect some patterns, if any ?

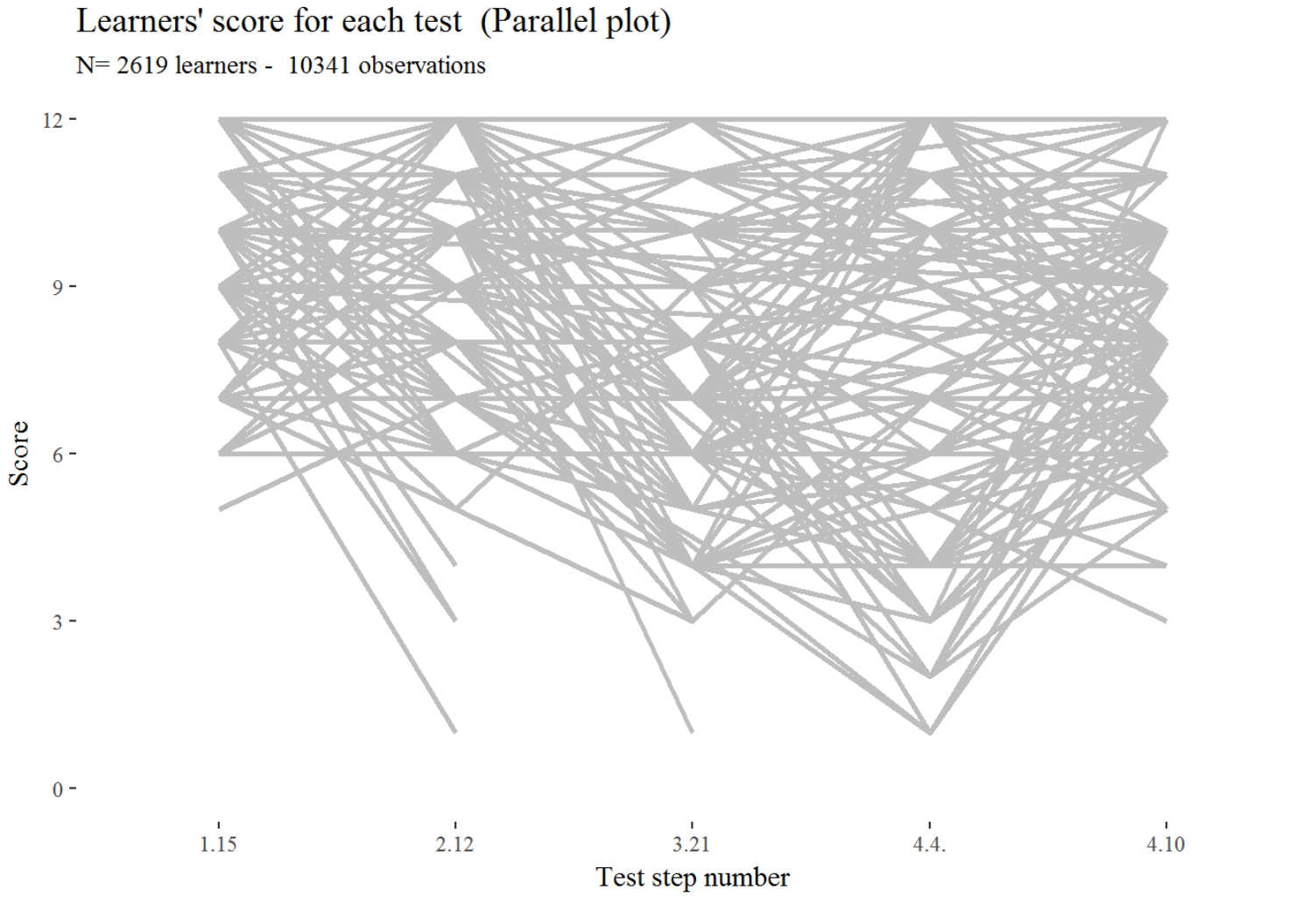
Now for more cheating with parallel plots !



Source: MOOC ``Manage your prices", FutureLearn (2017)

How to detect some patterns, if any ?

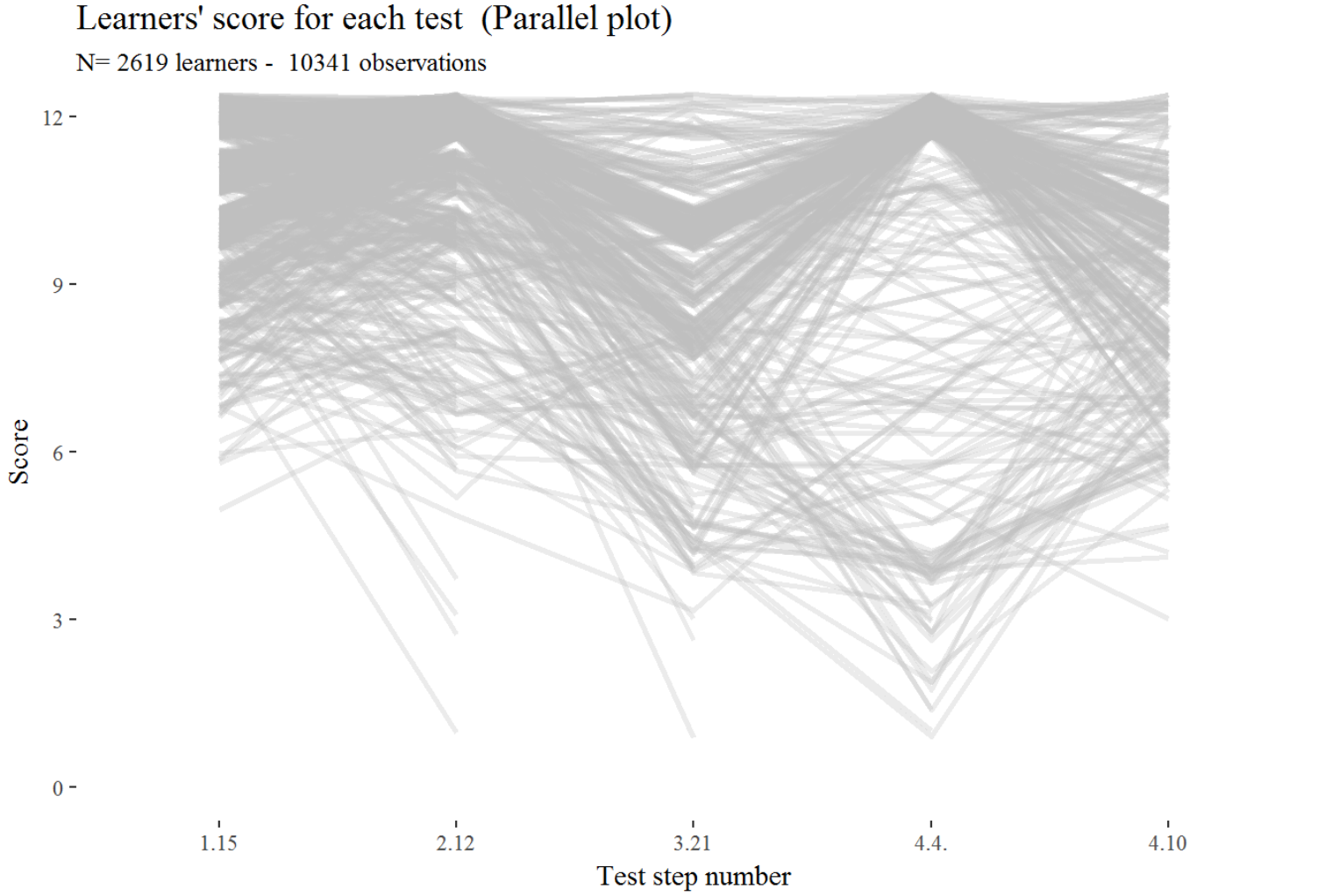
Now for more cheating with parallel plots !



Source: MOOC ``Manage your prices", FutureLearn (2017)

How to detect some patterns, if any ?

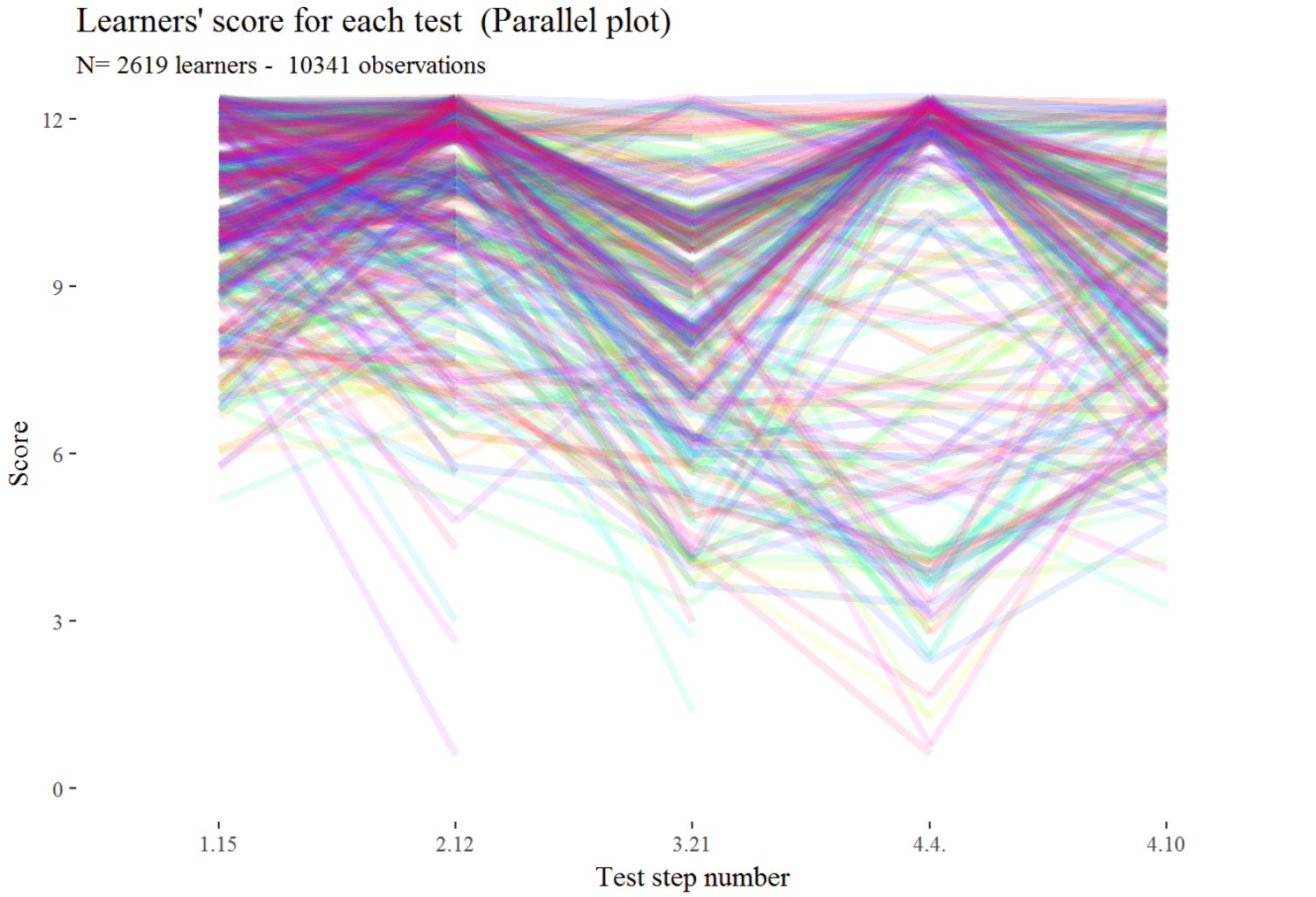
Now for more cheating with parallel plots !



Source: MOOC ``Manage your prices", FutureLearn (2017)

How to detect some patterns, if any ?

Now for more cheating with parallel plots !



Source: MOOC ``Manage your prices", FutureLearn (2017)

TO *cheat* OR NOT TO *cheat* ?

- We have slightly modified the data using "jitter"" [2]

2. To Jitter= to make quick, small movements

TO *cheat* OR NOT TO *cheat* ?

- We have slightly modified the data using "jitter"" [2]
- This solves the *overplotting* problem (together with some α-transparency)

2. To Jitter= to make quick, small movements

To *cheat* or not to *cheat* ?

- We have slightly modified the data using "jitter"" [2]
- This solves the *overplotting* problem (together with some α-transparency)
- None of the above trajectories exists really

2. To Jitter= to make quick, small movements

TO *cheat* OR NOT TO *cheat* ?

- We have slightly modified the data using "jitter"" [2]
- This solves the *overplotting* problem (together with some α-transparency)
- None of the above trajectories exists really
- The reader should be aware of these little adjustments.

2. To Jitter= to make quick, small movements

TO *cheat* OR NOT TO *cheat* ?

- We have slightly modified the data using "jitter"" [2]
- This solves the *overplotting* problem (together with some α-transparency)
- None of the above trajectories exists really
- The reader should be aware of these little adjustments.

$\rightarrow$ Everybody should be allowed to *cheat* sometimes !

2. To Jitter= to make quick, small movements