
ο1-Εισαγωγή στη τέχνη της Ανάλυσης

Δρ. Δημήτρης Καβρουδάκης
Αν. Καθηγητής Γεωγραφικής Ανάλυσης
dimitrisk@aegean.gr
www.dimitrisk.gr

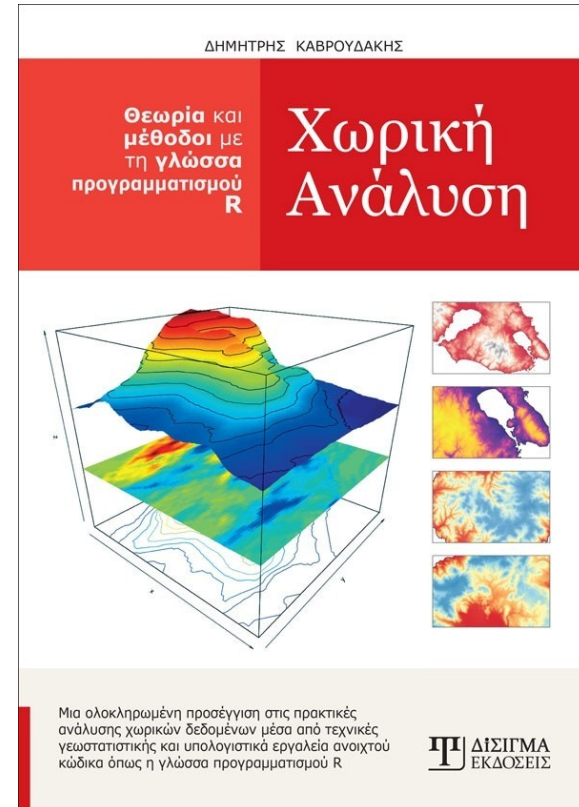
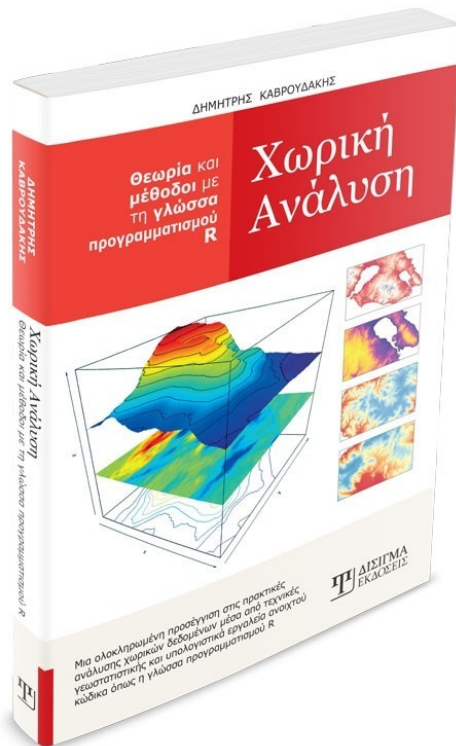


Πανεπιστήμιο Αιγαίου
Τμήμα Γεωγραφίας

SAG*i*SRS
ΧΑ-ΣΓΠ-ΤΗ

Spatial Analysis, Geographical Information Systems and Remote Sensing
Ερευνητική ομάδα Χωρικής Ανάλυσης, Συστημάτων Γεωγραφικών Πληροφοριών και Τηλεπισκόπησης

Δημήτρης Καβρουδάκης (2020) Χωρική Ανάλυση Εκδόσεις ΔΙΣΙΓΜΑ, ISBN13: 978-618-202-021-0



- Ανα-λύω
- Η διαδικασία τμηματοποίησης σύνθετων αντικειμένων σε μικρότερα μέρη για την
 - Κατανόηση, Ερμηνεία, Επίλυση
- Συστατικά
 - Θεωρητικές προσεγγίσεις
 - Ποσοτικές μέθοδοι
- Στάδια
 - **Διατύπωση** του αντικειμένου
 - και των ερευνητικών ερωτημάτων
 - **Κατακερματισμός** των επιμέρους κομματιών
 - **Κατανόηση** των αντικειμένων
 - Σαφή και πλήρη αντίληψη
 - **Ερμηνεία** του προβλήματος
 - Εξήγηση των αγνώστων και του προβλήματος
 - **Επίλυση** του προβλήματος
 - Εύρεση του ζητούμενου



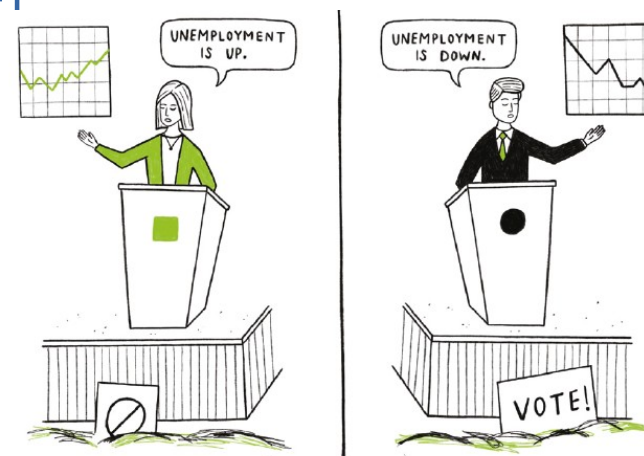


Λάθη στην Ανάλυση

- Δεδομένα απαραίτητα για την απόκτηση νέων πληροφοριών, τη λήψη αποφάσεων
 - υπάρχουν τόσα δεδομένα εκεί έξω, που η ποσότητα τους διπλασιάζεται κάθε δύο χρόνια - και έως το 2020, θα υπάρχουν 45.000 exabytes δεδομένων
- Άνευ προηγούμενου νούμερο και είναι δύσκολο να τεθεί σε προοπτική
 - 1 exabyte ισούται με 1.000.000.000 GB δεδομένων
 - 5 exabyte περίπου ίσο με *“όλες τις λέξεις που έχει μιλήσει ποτέ η ανθρωπότητα”*
- Αναζήτηση και ανάλυση των δεδομένων μπορεί να είναι η πρόκληση
- Τα δεδομένα διατίθενται σε πολλές διαφορετικές μορφές και δεν είναι εύκολο να αναλυθούν όλα
- Δελεαστικό να παίρνουμε συντομεύσεις με δεδομένα ή να προσπαθούμε να προσαρμόζουμε τα δεδομένα στις προκαθορισμένες αντιλήψεις μας για το πώς θα έπρεπε να είναι τα πράγματα
-
- Θα μελετήσουμε τα κοινά λάθη στην ανάλυση δεδομένων:
 - *πηγή εικόνων [1]*

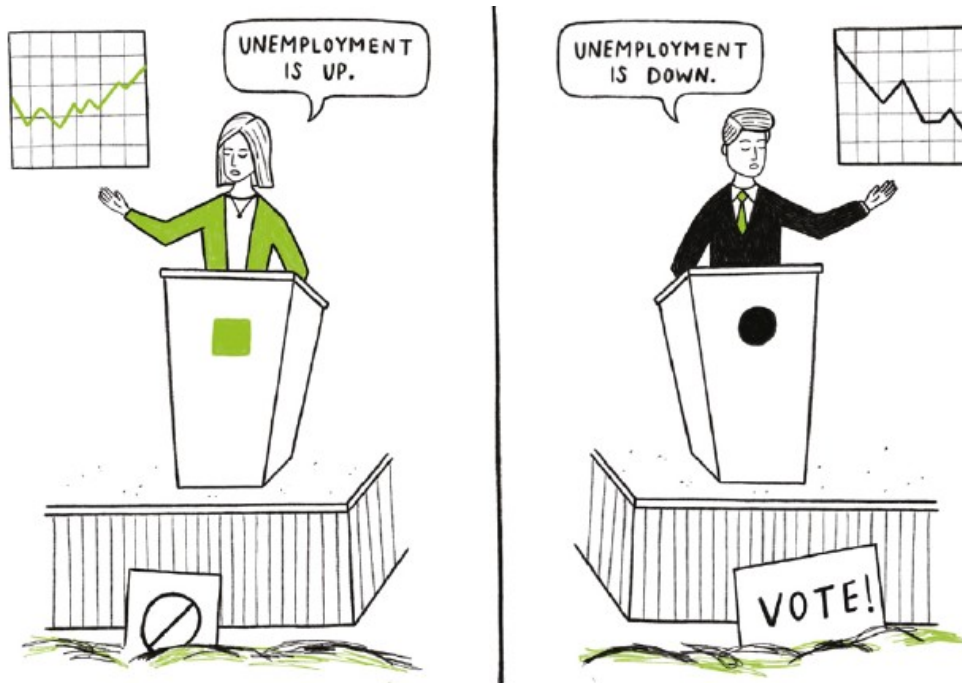
1) Cherry Picking

- απόκρυψη στοιχείων ή πλάνη ελλιπών αποδεικτικών στοιχείων
 - είναι η πράξη κατάδειξης μεμονωμένων υποθέσεων ή δεδομένων
 - που φαίνεται να επιβεβαιώνουν μια συγκεκριμένη θέση,
 - ενώ αγνοείται ένα σημαντικό μέρος σχετικών και παρόμοιων υποθέσεων ή δεδομένων
 - που μπορεί να έρχονται σε αντίθεση με αυτήν τη θέση
- ηθελημένα ή ακούσια



CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.



CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.





- βυθοκόρηση δεδομένων
 - ή αλιεία δεδομένων, καταδίωξη σημασίας, αναζήτηση σημασίας, επιλεκτική εξαγωγή συμπερασμάτων και παραβίαση
- είναι η κακή χρήση της ανάλυσης δεδομένων για την εύρεση προτύπων σε δεδομένα
 - που μπορούν να παρουσιαστούν ως στατιστικά σημαντικά
- αυξάνοντας και υποτιμώντας δραματικά τον κίνδυνο ψευδώς θετικών αποτελεσμάτων
 - Αυτό γίνεται με την εκτέλεση πολλών στατιστικών δοκιμών στα δεδομένα
 - και την αναφορά μόνο εκείνων που επανέρχονται με σημαντικά αποτελέσματα
- 1) κακή χρήση των τεχνικών για την ψευδή: πρόταση τάσεων και συσχετισμών στο υποσύνολο, που δεν εμφανίζονται στο πλήρες σύνολο δεδομένων
 - ή
- 2) εφαρμογή πολλαπλών δοκιμών στατιστικής σημασίας (δηλ. t-test) σε έναν αριθμό μεταβλητών μέσα σε ένα σύνολο δεδομένων για τον καθορισμό συσχετίσεων που μπορεί να προκύψουν στην πραγματικότητα καθαρά τυχαία

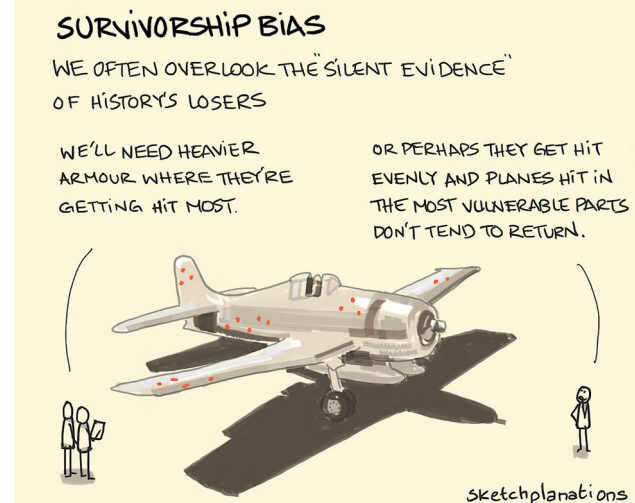


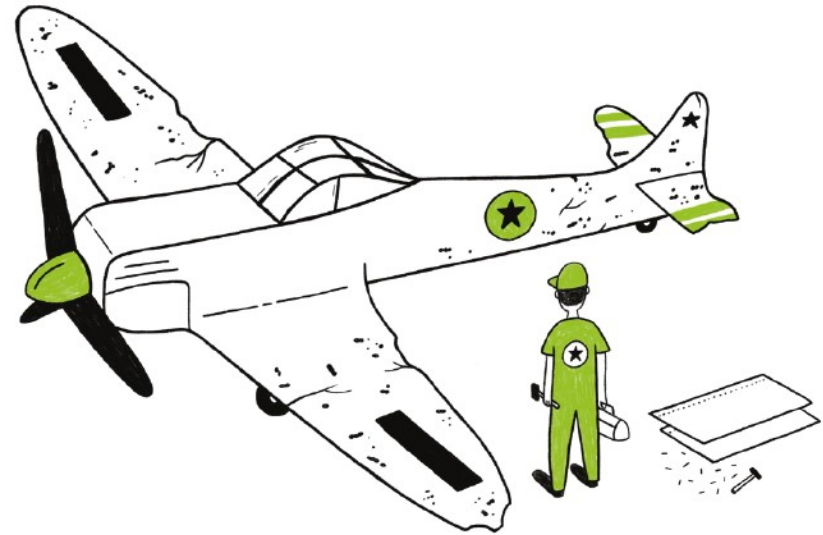
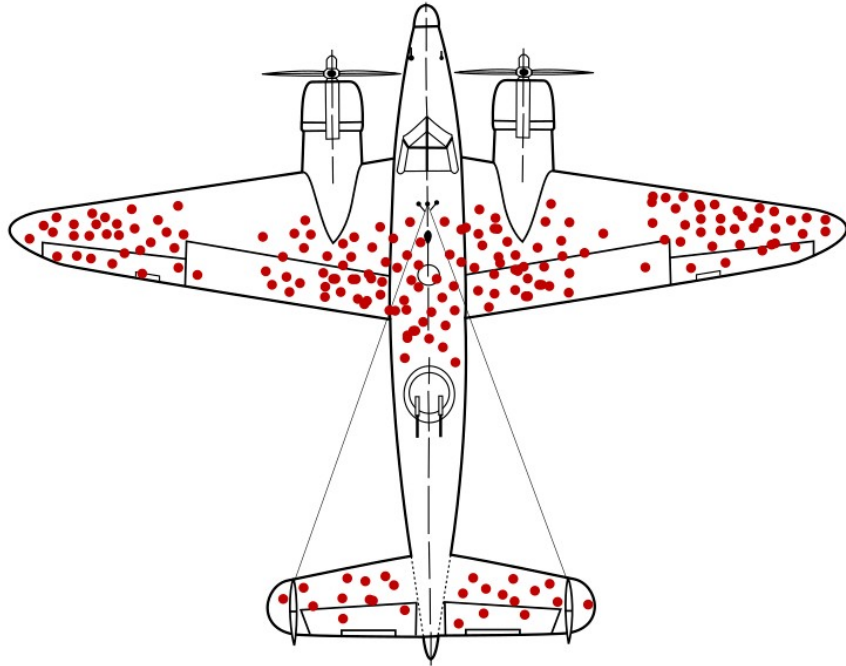
DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

3) Survivor Bias

- Προκατάληψη επιβίωσης ή μεροληψία επιβίωσης
- είναι το λογικό λάθος της συγκέντρωσης στους ανθρώπους ή τα πράγματα που πέτυχαν τη διαδικασία επιλογής
 - και την παράβλεψη εκείνων που δεν το έκαναν,
 - συνήθως λόγω της έλλειψης ορατότητάς τους
- Αυτό μπορεί να οδηγήσει σε ψευδή συμπεράσματα με πολλούς διαφορετικούς τρόπους
 - είναι μια μορφή προκατάληψης επιλογής (selection bias)
- Αυτό μπορεί να οδηγήσει σε υπερβολικά αισιόδοξες πεποιθήσεις
 - επειδή οι αποτυχίες αγνοούνται ή δεν φαίνονται
- εμφανίζεται όταν λαμβάνονται υπόψη μόνο οι νικητές
 - ενώ οι χαμένοι που έχουν εξαφανιστεί
 - δεν λαμβάνονται υπόψη.
- παραμορφώνει τα μεσαία αποτελέσματα προς τα πάνω
 - με αποτέλεσμα να φαίνεται ότι έχουν καλύτερη απόδοση
 - δεδομένου ότι τα αποτελέσματα χαμηλών επιδόσεων
 - έχουν παραβλεφθεί





SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

- Στρεβλό κίνητρο
- Ένα στρεβλό κίνητρο είναι ένα κίνητρο που έχει ένα ακούσιο και ανεπιθύμητο αποτέλεσμα που είναι αντίθετο με τις προθέσεις των σχεδιαστών του
- Το κίνητρο ανταμείβει ακούσια τους ανθρώπους που επιδεινώνουν το ζήτημα
- Η εσφαλμένη διέγερση στην οικονομία/πολιτική μπορεί να προκαλέσει ανεπιθύμητες συνέπειες
 - είναι ένα παράδειγμα της παροιμίας «ο δρόμος προς την κόλαση είναι στρωμένος με καλές προθέσεις»
- Η βρετανική κυβέρνηση, ανήσυχη για τον αριθμό των δηλητηριωδών κόμπρων στο Δελχί, πρόσφερε αμοιβή για κάθε νεκρή κόμπρα.
 - αρχικά, αυτή ήταν μια επιτυχημένη στρατηγική
 - **μεγάλος αριθμός φιδιών σκοτώθηκαν για την ανταμοιβή**
 - τελικά, ωστόσο, οι επιχειρηματίες άρχισαν να εκτρέφουν κόμπρες για το εισόδημα.
 - όταν η κυβέρνηση το αντιλήφθηκε, το πρόγραμμα επιβράβευσης καταργήθηκε.
 - όταν οι εκτροφείς κόμπρας απελευθέρωσαν τα άχρηστα πλέον φίδια τους, ο πληθυσμός της άγριας κόμπρας αυξήθηκε περαιτέρω [2]

THE COBRA EFFECT

A WELL-INTENTIONED MEASURE CAN OFTEN BACKFIRE
AND HAVE THE OPPOSITE EFFECT TO INTENDED



INTENTION
REDUCE COBRA
POPULATION



ACTION
A BOUNTY FOR
DEAD COBRAS!



EFFECT
PEOPLE START
COBRA FARMING

sketchplanations

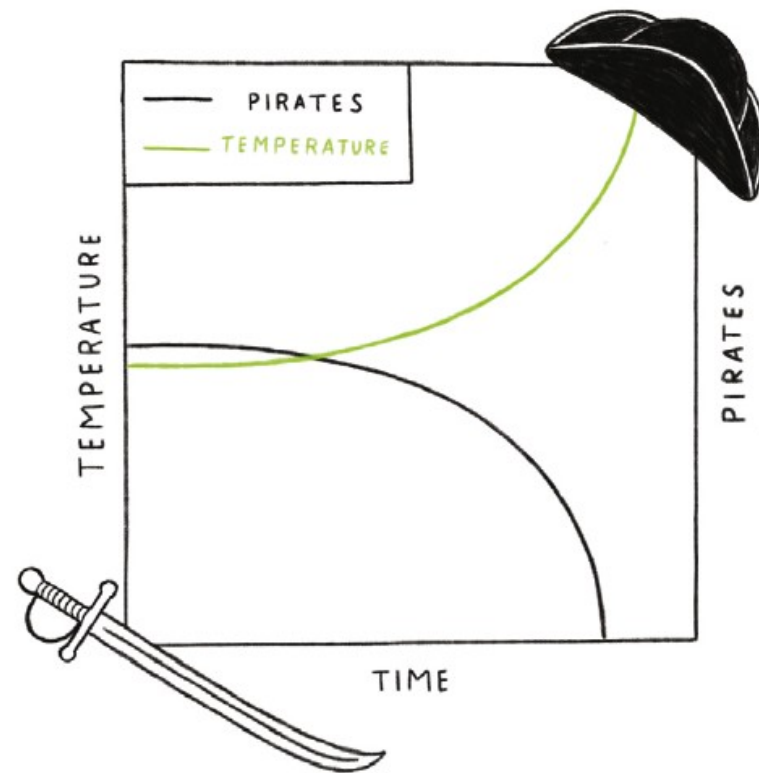
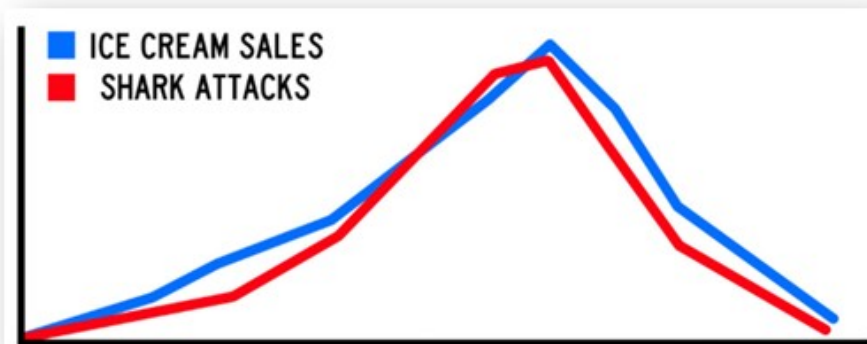


COBRA EFFECT

Setting an incentive that accidentally produces the
opposite result to the one intended. Also known as a
Perverse Incentive.

- Λανθασμένη αιτιότητα, “συσχέτιση δεν συνεπάγεται αιτιότητα”
- Αδυναμία να συνταχθεί η σχέση αιτίου-αποτελέσματος μεταξύ δύο γεγονότων ή μεταβλητών
 - αποκλειστικά με βάση μια παρατηρούμενη συσχέτιση μεταξύ τους
- δύο γεγονότα που συμβαίνουν μαζί
 - θεωρούνται χωρίς βάσιμες αποδείξεις ότι έχουν μια σχέση αιτίου-αποτελέσματος
- Ενώ ίσως μπορεί να υπάρχει μια αιτιώδης σχέση
 - δεν υπάρχουν αρκετές πληροφορίες για να επεξηγήσει την αλήθεια πίσω από τη σύνδεση
- Οδηγεί σε ένα συμπέρασμα που υπονοεί μια αιτιώδη σχέση χωρίς αποδεικτικά στοιχεία
- Αυτή η πλάνη είναι επίσης γνωστή από τη λατινική φράση “cum hoc ergo propter hoc” (“με αυτό, επομένως εξαιτίας αυτού”)
 - Αυτό διαφέρει από την πλάνη που είναι γνωστή ως post hoc ergo propter hoc (“μετά από αυτό, επομένως λόγω αυτού”),
 - στην οποία ένα γεγονός που ακολουθεί ένα άλλο θεωρείται απαραίτητη συνέπεια του προηγούμενου γεγονότος και από τη συγχώνευση, την εσφαλμένη συγχώνευση δύο συμβάντα, ιδέες, βάσεις δεδομένων κ.λπ., σε ένα.



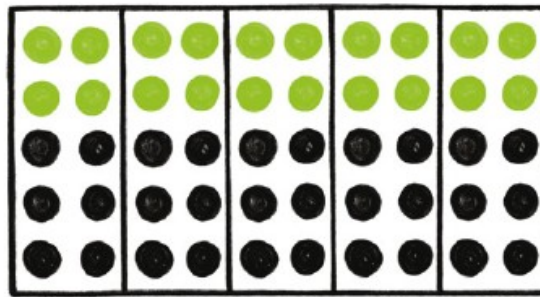


FALSE CAUSALITY

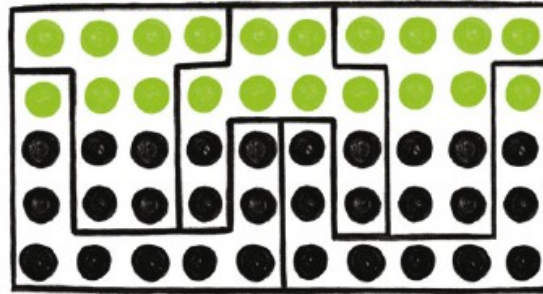
Falsely assuming when two events appear related that one must have caused the other.

- Οι παγκόσμιες θερμοκρασίες έχουν αυξηθεί σταθερά τα τελευταία 150 χρόνια και ο αριθμός των πειρατών έχει μειωθεί με συγκρίσιμο ρυθμό.
-
- Κανείς δεν θα ισχυριζόταν εύλογα ότι η μείωση των πειρατών προκάλεσε την υπερθέρμανση του πλανήτη ή ότι περισσότεροι πειρατές θα την αντιστρέψουν. Αλλά συνήθως δεν είναι τόσο ξεκάθαρο.
-
- Συχνά οι συσχετίσεις μεταξύ δύο πραγμάτων μας βάζουν στον πειρασμό να πιστέψουμε ότι το ένα προκάλεσε το άλλο. Ωστόσο, συχνά πρόκειται για σύμπτωση ή υπάρχει ένας τρίτος παράγοντας που προκαλεί και τα δύο αποτελέσματα που βλέπετε. Στο παράδειγμά μας για τους πειρατές και την υπερθέρμανση του πλανήτη, η αιτία και των δύο είναι η εκβιομηχάνιση.
-
- Ποτέ μην υποθέτετε αιτιώδη συνάφεια μόνο λόγω της συσχέτισης - συλλέγετε πάντα περισσότερα στοιχεία.





BLACK WINS



GREEN WINS

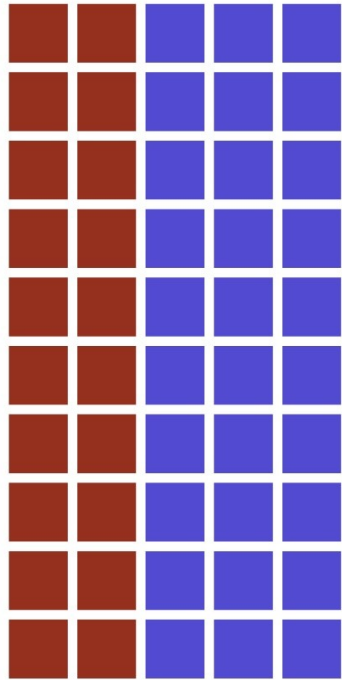
GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.

- Σε πολλά πολιτικά συστήματα, είναι δυνατό να χειραγωγηθεί η πιθανότητα εκλογής ενός κόμματος έναντι ενός άλλου, επαναπροσδιορίζοντας τις πολιτικές περιφέρειες - συμπεριλαμβάνοντας περισσότερες αγροτικές περιοχές σε μια περιφέρεια για να μειώσετε το κόμμα που είναι πιο δημοφιλές στις πόλεις κ.λπ.
- Ένα παρόμοιο φαινόμενο, γνωστό ως Πρόβλημα Τροποποιήσιμων Επιφανειακών Μονάδων (MAUP), μπορεί να συμβεί κατά την ανάλυση δεδομένων.
- Ο τρόπος με τον οποίο ορίζετε τις περιοχές για να συγκεντρώσετε τα δεδομένα σας - π.χ. τι ορίζετε ως «Βόρειες κομητείες» - μπορεί να αλλάξει το αποτέλεσμα.
- Η κλίμακα που χρησιμοποιείται για την ομαδοποίηση δεδομένων μπορεί επίσης να έχει μεγάλο αντίκτυπο.
- Τα αποτελέσματα μπορεί να ποικίλλουν σημαντικά είτε χρησιμοποιούνται ταχυδρομικοί κώδικες, κομητείες είτε πολιτείες.

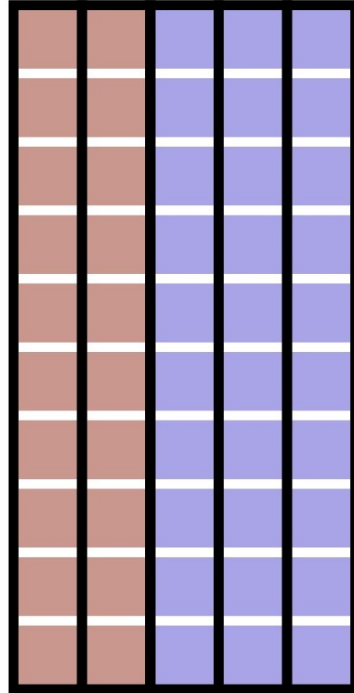
Three different ways to divide 50 people into five districts

50
people



**60% blue,
40% red**

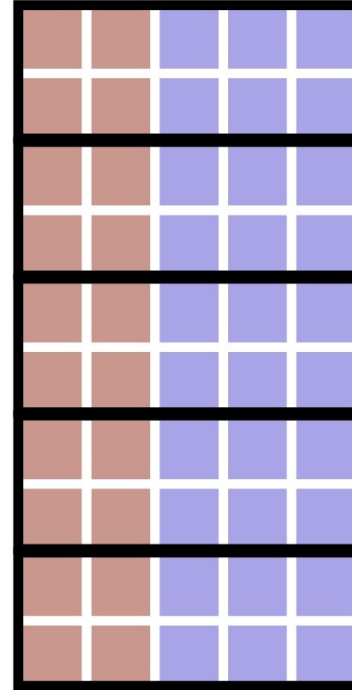
1. Perfect
representation



**3 blue districts,
2 red districts**

BLUE WINS

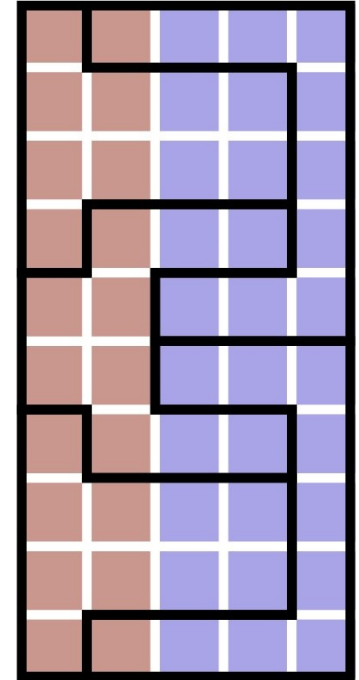
2. Compact,
but unfair



**5 blue districts,
0 red districts**

BLUE WINS

3. Neither compact
nor fair



**2 blue districts,
3 red districts**

RED WINS





SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

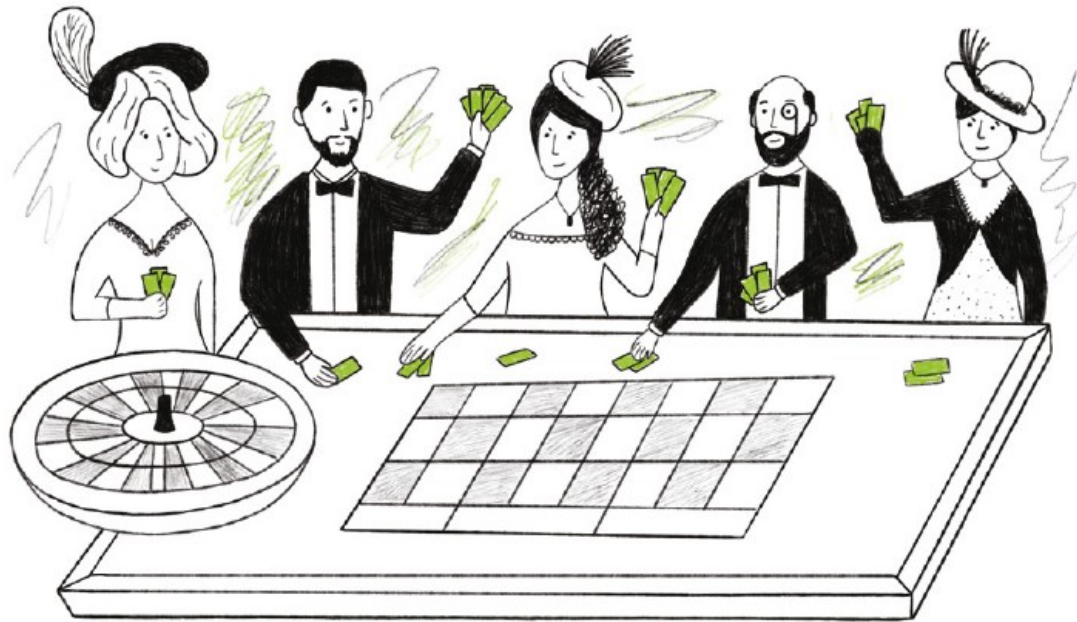
- Ένα κλασικό πρόβλημα στις δημοσκοπήσεις για τις εκλογές, όπου τα άτομα που συμμετέχουν σε μια δημοσκόπηση δεν είναι αντιπροσωπευτικά του συνολικού πληθυσμού, είτε λόγω αυτοεπιλογής είτε λόγω προκατάληψης από τους αναλυτές.
- Ένα διάσημο παράδειγμα συνέβη το 1948, όταν η εφημερίδα The Chicago Tribune προέβλεψε λανθασμένα, βασιζόμενη σε μια τηλεφωνική έρευνα, ότι ο Thomas E. Dewey θα γινόταν ο επόμενος πρόεδρος των ΗΠΑ.
- Δεν είχαν λάβει υπόψη τους ότι μόνο ένα συγκεκριμένο δημογραφικό στοιχείο θα μπορούσε να αντέξει οικονομικά τα τηλέφωνα, αποκλείοντας ολόκληρα τμήματα του πληθυσμού από την έρευνά τους.
- Βεβαιωθείτε ότι έχετε εξετάσει εάν οι συμμετέχοντες στην έρευνά σας είναι πραγματικά αντιπροσωπευτικοί και δεν υπόκεινται σε κάποια προκατάληψη δειγματοληψίας.

- Η προκατάληψη μπορεί να προκύψει, κυρίως, με τρεις τρόπους:
 - **Μεροληψία επιλογής:** Έχει να κάνει με την επιλογή των ομάδων. Τις περισσότερες φορές, η απόφαση δεν έχει ληφθεί βάσει αντικειμενικών μεθόδων δειγματοληψίας
 - **Μεροληψία πληροφοριών:** Αντιμετωπίζουμε προκατάληψη λόγω ανεπαρκών πληροφοριών. Επομένως, δεν μπορούμε να συγκρίνουμε τις ομάδες επειδή έχουμε πολύ περιορισμένες πληροφορίες σχετικά με αυτές.
 - **Μεροληψία σύγχυσης:** Υπάρχει μια λεγόμενη συγχώνευση μεταβλητή, η οποία προκαλεί την προκατάληψη. Συχνά είναι δύσκολο να βρεθεί πού βρίσκεται το πρόβλημα.
-

- Self-selection
 - Συγκεκριμένα άτομα είναι πιο επιρρεπή να λάβουν μέρος σε μια έρευνα
 - People who are more thrill-seeking are likely to take part in pain research studies. This may skew the data.
- Non-response
 - People who refuse to participate or drop out from a study systematically differ from those who take part.
 - Αντικοινωνικά άτομα, είναι λιγότερο πιθανό να λάβουν μέρος σε μια έρευνα σχετικά με την αντικοινωνικότητα
- Υπο-εκπροσώπηση (Undercoverage)
 - Μειωμένη αντιπροσώπευση ομάδων στο δείγμα
 - Διαδικτυακές έρευνες, χάνουν απαντήσεις από άτομα χωρίς πρόσβαση στο διαδίκτυο

- Survivorship
 - Επιτυχημένες μελέτες είναι πιο πιθανό να επιλεγθούν για παρουσίαση
- Pre-screening or advertising
 - Ο τρόπος με τον οποίο οι συμμετέχοντες προελέγχονται ή το πού διαφημίζεται μια μελέτη μπορεί να επηρεάσει ένα δείγμα.
 - Όταν αναζητάτε εθελοντές για να δοκιμάσετε μια νέα παρέμβαση ύπνου, μπορεί να καταλήξετε σε ένα δείγμα που έχει μεγαλύτερο κίνητρο να βελτιώσει τις συνήθειες ύπνου του από τον υπόλοιπο πληθυσμό. Ως αποτέλεσμα, μπορεί να ήταν πιθανό να βελτιώσουν τις συνήθειες ύπνου τους ανεξάρτητα από τις επιπτώσεις της παρέμβασής σας.
- Healthy user
 - Οι εθελοντές για προληπτικές παρεμβάσεις είναι πιο πιθανό να επιδιώξουν συμπεριφορές και δραστηριότητες που ενισχύουν την υγεία σε σχέση με άλλα μέλη του πληθυσμού.
 - Ένα δείγμα σε μια προληπτική παρέμβαση έχει καλύτερη διατροφή, υψηλότερα επίπεδα σωματικής δραστηριότητας, απέχει από το αλκοόλ και αποφεύγει το κάπνισμα περισσότερο από το μεγαλύτερο μέρος του πληθυσμού. Τα πειραματικά ευρήματα μπορεί να είναι αποτέλεσμα της αλληλεπίδρασης της θεραπείας με αυτά τα χαρακτηριστικά του δείγματος, και όχι μόνο της ίδιας της θεραπείας.



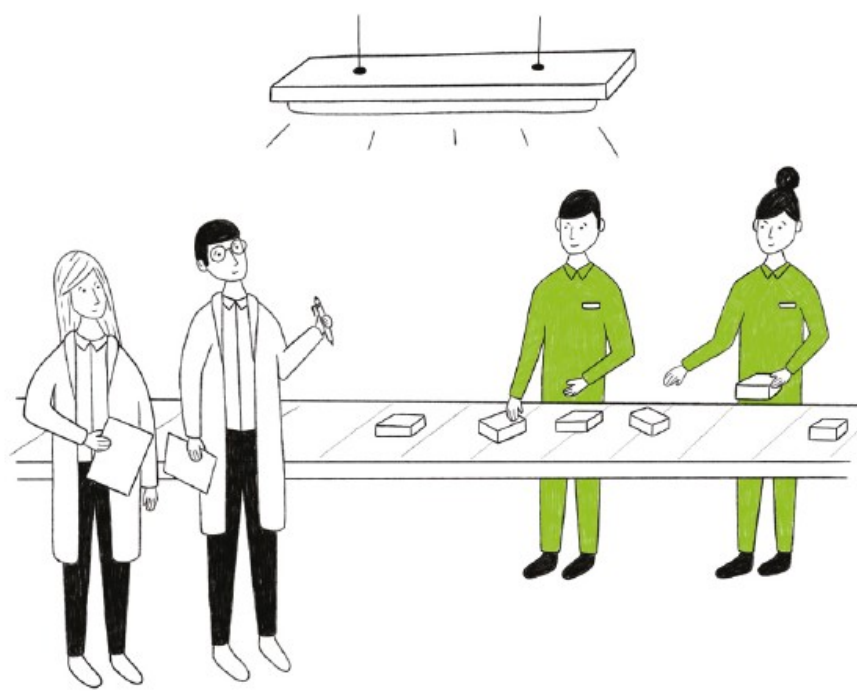


GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).

- Αυτό είναι επίσης γνωστό ως το Μόντε Κάρλο Πλάνη λόγω ενός διαβόητου παραδείγματος που συνέβη σε ένα τραπέζι ρουλέτας εκεί το 1913.
- Η μπίλια έπεσε μαύρη 26 φορές στη σειρά και οι παίκτες έχασαν εκατομμύρια στοιχηματίζοντας εναντίον του μαύρου, υποθέτοντας ότι το σερί έπρεπε να τελειώσει.
- Ωστόσο, η πιθανότητα να είναι μαύρο είναι πάντα η ίδια με το κόκκινο, ανεξάρτητα από το τι έχει συμβεί στο παρελθόν, επειδή η υποκείμενη πιθανότητα παραμένει αμετάβλητη.
- Ένα τραπέζι ρουλέτας δεν έχει μνήμη.
- Όταν μπείτε στον πειρασμό από αυτό το πλάνη, υπενθυμίστε στον εαυτό σας ότι δεν υπάρχει καμία διορθωτική δύναμη στο σύμπαν που να ενεργεί για να «ισορροπήσει τα πράγματα»!





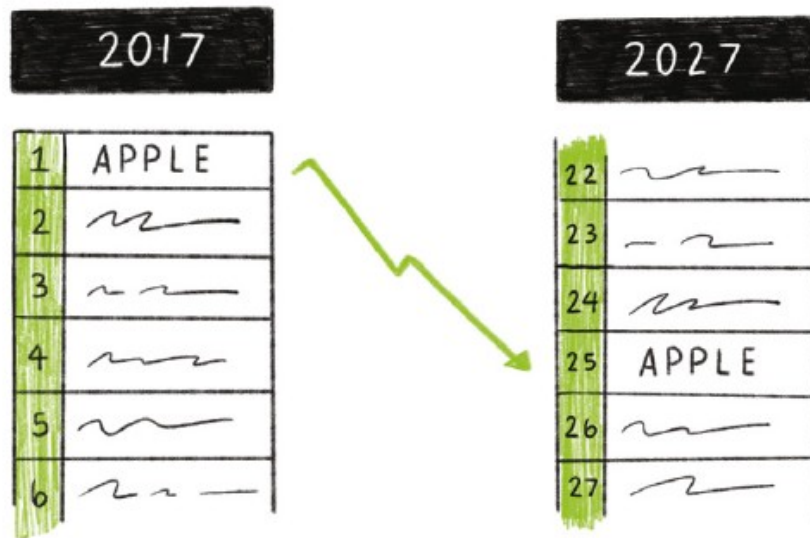
HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.

- Όταν η πράξη της παρακολούθησης κάποιου μπορεί να επηρεάσει τη συμπεριφορά αυτού του ατόμου. Γνωστό και ως Φαινόμενο του Παρατηρητή.
- Τη δεκαετία του 1920 στο Hawthorne Works, ένα εργοστάσιο στο Ιλινόις, ένα πείραμα κοινωνικών επιστημών υπέθεσε ότι οι εργαζόμενοι θα γίνονταν πιο παραγωγικοί μετά από διάφορες αλλαγές στο περιβάλλον τους, όπως οι ώρες εργασίας, τα επίπεδα φωτισμού και οι ώρες διαλειμμάτων.
- Ωστόσο, αποδείχθηκε ότι αυτό που πραγματικά παρακίνησε την παραγωγικότητα των εργαζομένων ήταν κάποιος που έδειχνε ενδιαφέρον για αυτούς. Όταν χρησιμοποιούνται ανθρώπινα ερευνητικά υποκείμενα, είναι σημαντικό να αναλύονται τα δεδομένα που προκύπτουν λαμβάνοντας υπόψη το Φαινόμενο Hawthorne.



TOP COMPANIES



REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.

- Όταν επαναλαμβάνεται ένα εξαιρετικό συμβάν, σιγά-σιγά θα επανέλθει προς τον μέσο όρο του
- Όταν το αποτέλεσμα υπόκεινται σε τυχαιότητα, αργά ή γρήγορα θα επανέλθει προς τον μέσο όρο

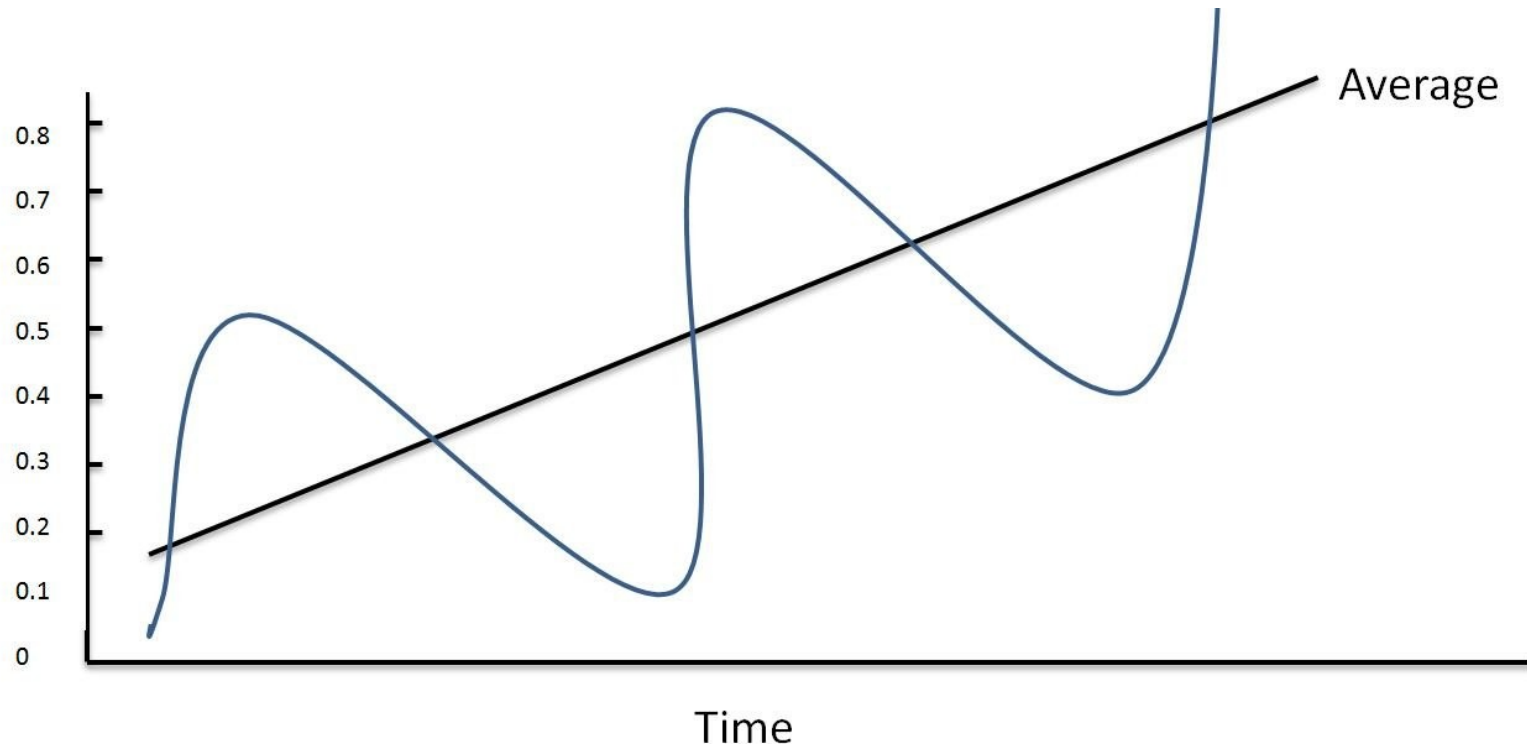


FIG. 7.

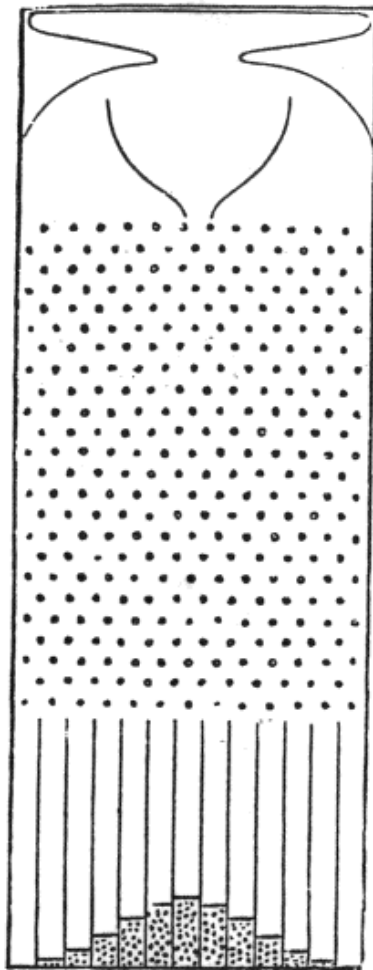


FIG. 8.

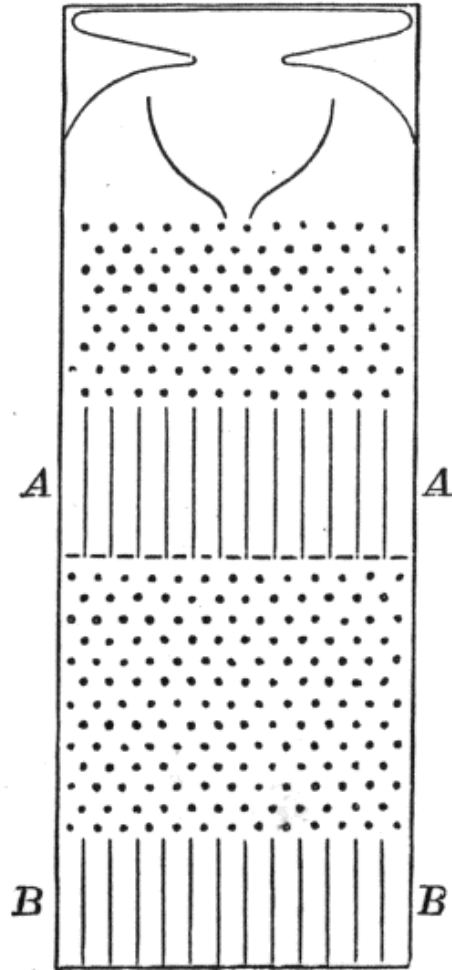
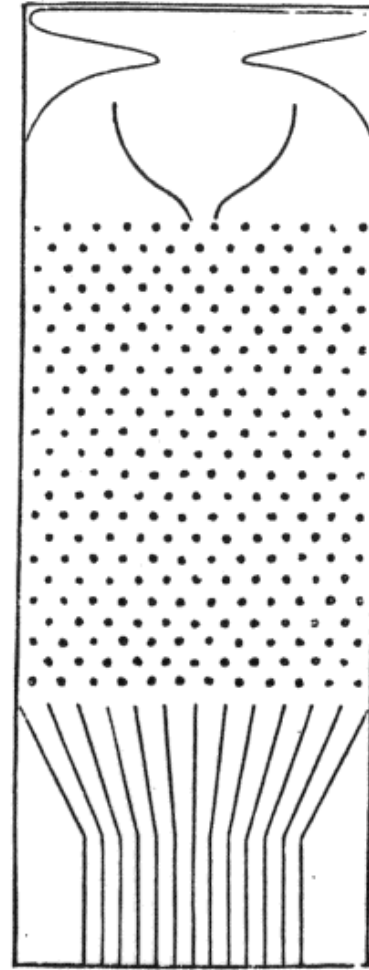


FIG. 9.







APPLICATION SUCCESS RATE

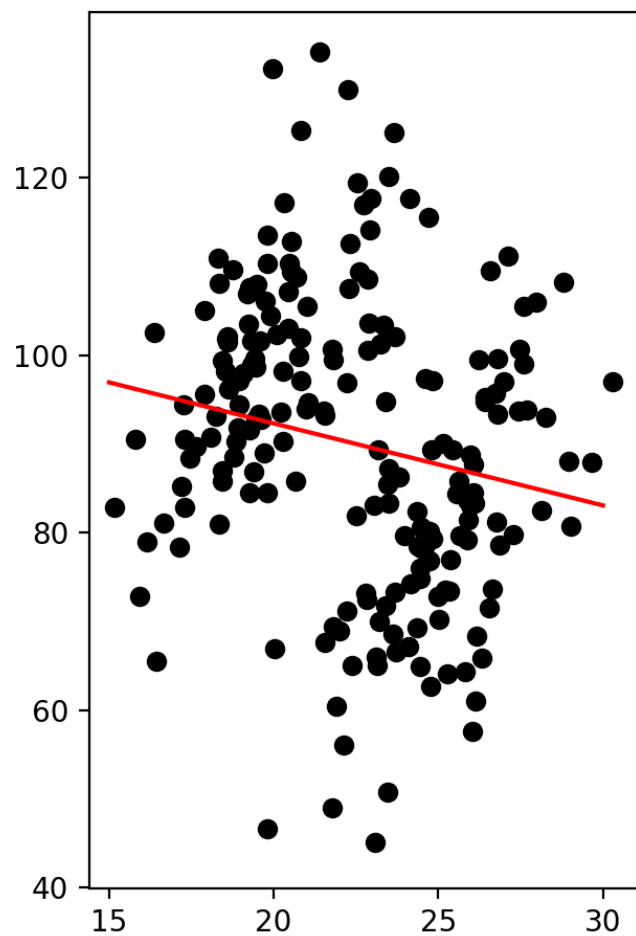
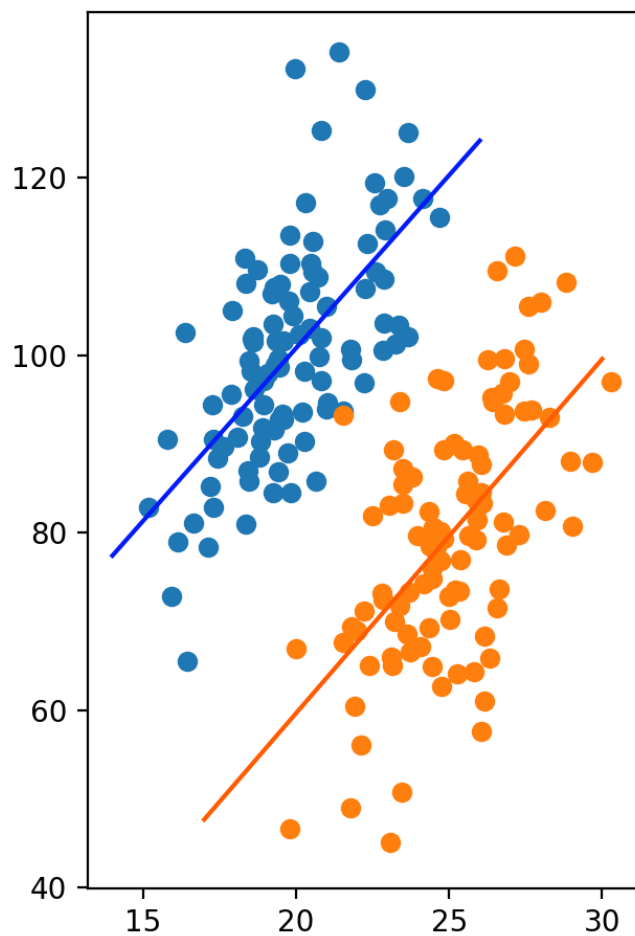
	MALE	FEMALE
SUBJECT 1	14 % (168 of 1200)	15 % (270 of 1800)
SUBJECT 2	50 % (400 of 800)	51 % (102 of 200)
TOTAL	28 % (568 of 2000)	19 % (372 of 2000)

??

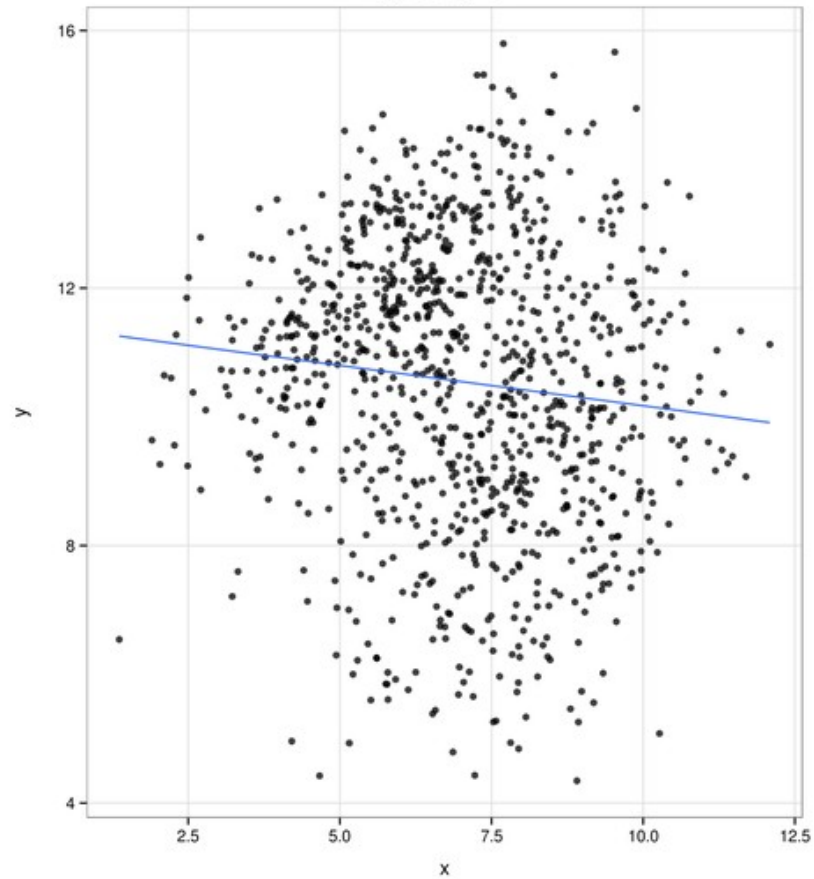
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.

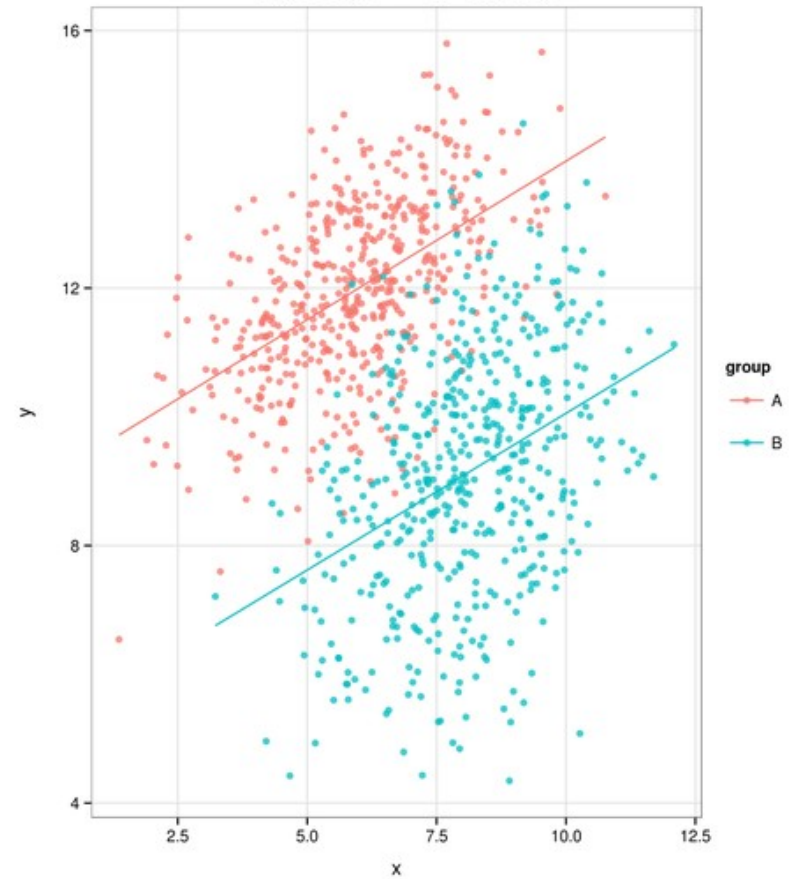
- το 1970s το Berkeley University κατηγορήθηκε για μεροληψία υπέρ των αντρών
 - αιτήσεις από γυναίκες είχαν λιγότερες πιθανότητες
- Ωστόσο προσπαθώντας να βρεθεί η πηγή του προβλήματος
 - βρέθηκε ακριβώς το αντίθετο
- Το παράδοξο προκλήθηκε από τη διαφορά στα μαθήματα για τα οποία υπέβαλαν αιτήσεις άνδρες και γυναίκες.
-
- Ένα μεγαλύτερο ποσοστό γυναικών υποψηφίων υπέβαλε αιτήσεις σε μαθήματα με υψηλό ανταγωνισμό, όπου τα ποσοστά αποδοχής ήταν πολύ χαμηλότερα και για τα δύο φύλα.



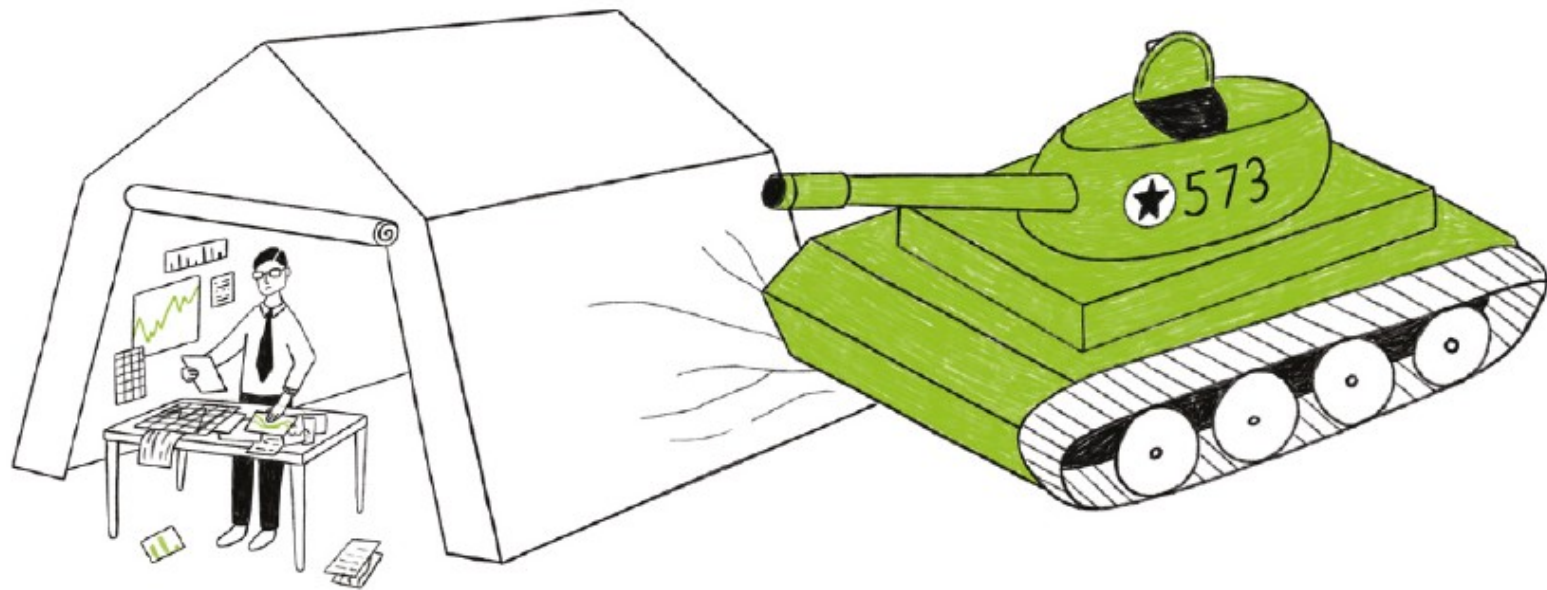
Regress y on x



Regress y on x and group







MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.

- Robert McNamara, τον Υπουργό Άμυνας των ΗΠΑ (1961-1968)
 - πίστευε ότι η αλήθεια μπορούσε να βρεθεί μόνο στα δεδομένα και τις στατιστική τεχνικές αυστηρότητα.
- Η πλάνη αναφέρεται στην προσέγγισή (εμμονή) του να λάβει ως μέτρο επιτυχίας ΜΟΝΟ
 - τον αριθμό των εχθρικών απωλειών στον πόλεμο του Βιετνάμ.
- άλλες σχετικές γνώσεις αγνοήθηκαν σε μεγάλο βαθμό:
 - η μεταβαλλόμενη διάθεση του κοινού των ΗΠΑ
 - τα συναισθήματα του Βιετναμέζικου λαού
- Όταν αναλύουμε περίπλοκα φαινόμενα
 - χρησιμοποιούμε μετρήσεις για κατανόηση/ποσοτικοποίηση της επιτυχίας
 - Ωστόσο, η δογματική βελτιστοποίηση των μετρήσεων
 - και η παράβλεψη **όλων** των άλλων πληροφοριών
 - είναι επικίνδυνη

GOODHART'S LAW

WHEN A MEASURE BECOMES A TARGET,
IT CEASES TO BE A GOOD MEASURE

IF YOU
MEASURE
PEOPLE ON...

NUMBER OF
NAILS MADE

WEIGHT OF
NAILS MADE

THEN YOU
MIGHT GET

1000'S OF
TINY NAILS

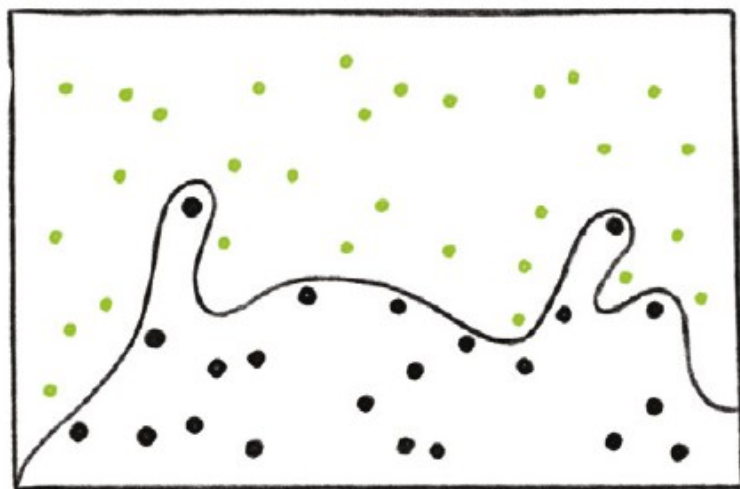
A FEW GIANT,
HEAVY NAILS



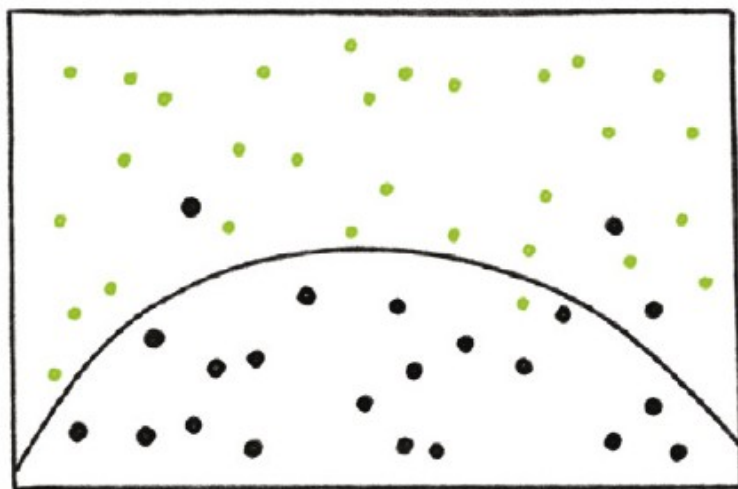
sketchplanations



OVERFITTING



JUST RIGHT

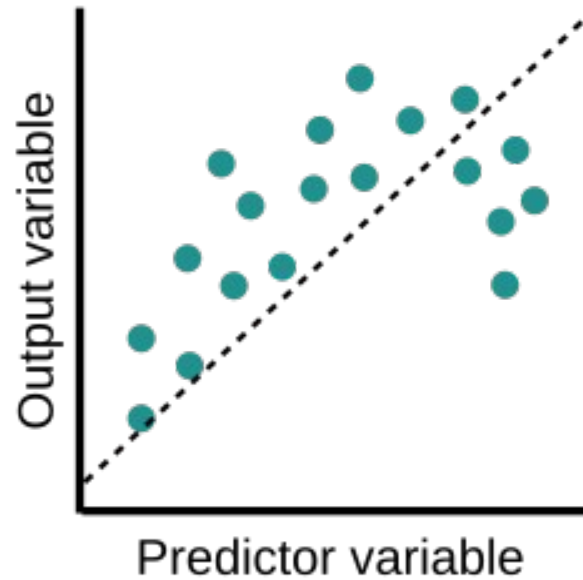


OVERFITTING

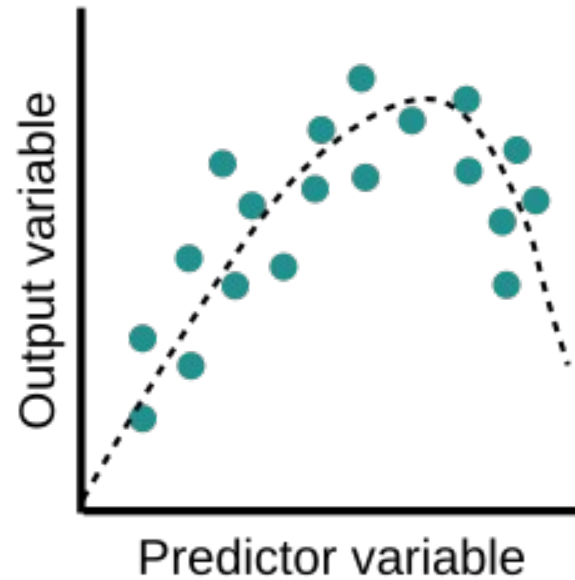
Creating a model that's overly tailored to the data you have and not representative of the general trend.

- Μια πιο σύνθετη εξήγηση συχνά περιγράφει τα δεδομένα σας καλύτερα από μια απλή.
- Ωστόσο, μια απλούστερη εξήγηση είναι συνήθως πιο αντιπροσωπευτική της υποκείμενης σχέσης.
- Όταν εξετάζετε δεδομένα, θα πρέπει να κατανοήσετε ποιες είναι οι υποκείμενες σχέσεις. Για να το κάνετε αυτό, δημιουργείτε ένα μοντέλο που τα περιγράφει μαθηματικά.
- Το πρόβλημα είναι ότι ένα πιο σύνθετο μοντέλο θα ταιριάζει καλύτερα στα αρχικά σας δεδομένα από ένα απλό.
- Ωστόσο, τείνουν να είναι πολύ εύθραυστα: Λειτουργούν καλά για τα δεδομένα που έχετε ήδη, αλλά προσπαθούν πολύ σκληρά να εξηγήσουν τυχαίες παραλλαγές.
- Επομένως, μόλις προσθέσετε περισσότερα δεδομένα, αυτά καταρρέουν. Τα απλούστερα μοντέλα είναι συνήθως πιο ισχυρά και καλύτερα στην πρόβλεψη μελλοντικών τάσεων.

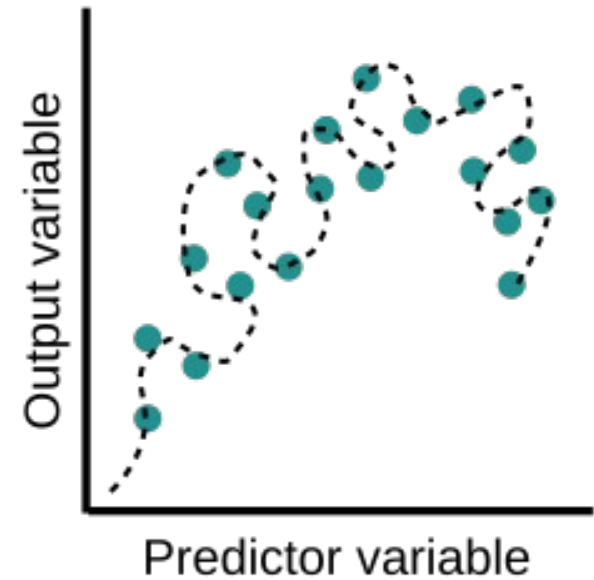
Underfit

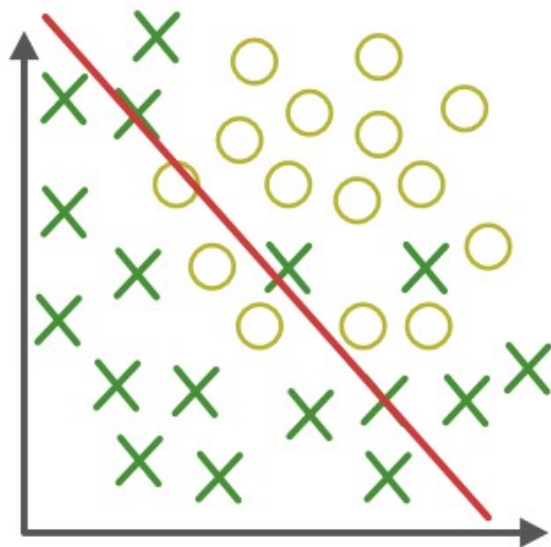


Optimal

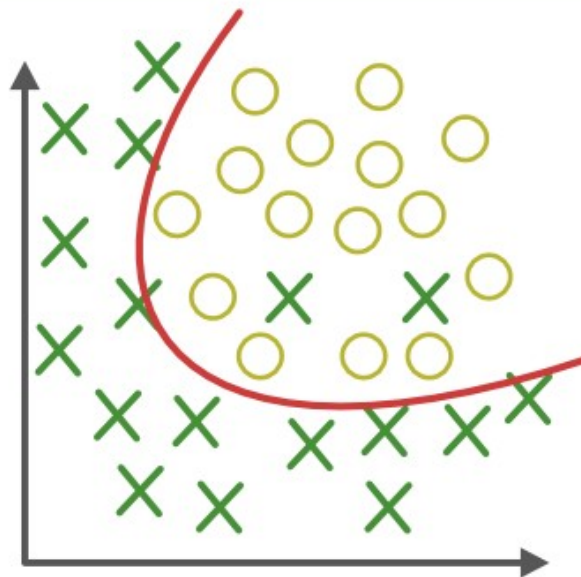


Overfit

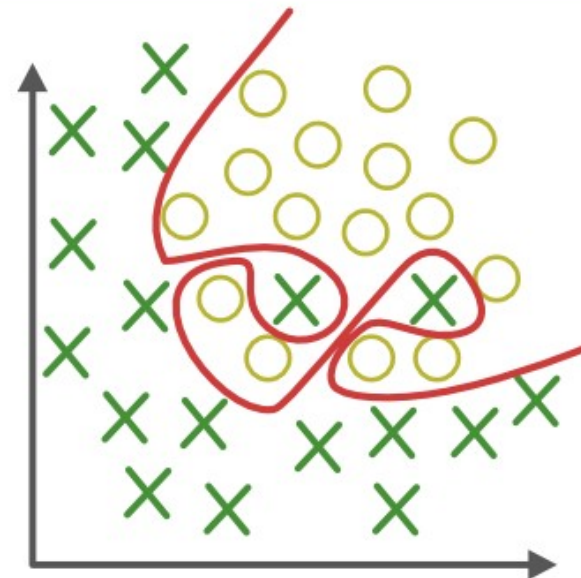




Under-fitting
(too simple to
explain the variance)



Appropriate-fitting



Over-fitting
(forcefitting--too
good to be true)



THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG
PILE OF LINEAR ALGEBRA, THEN COLLECT
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL
THEY START LOOKING RIGHT.





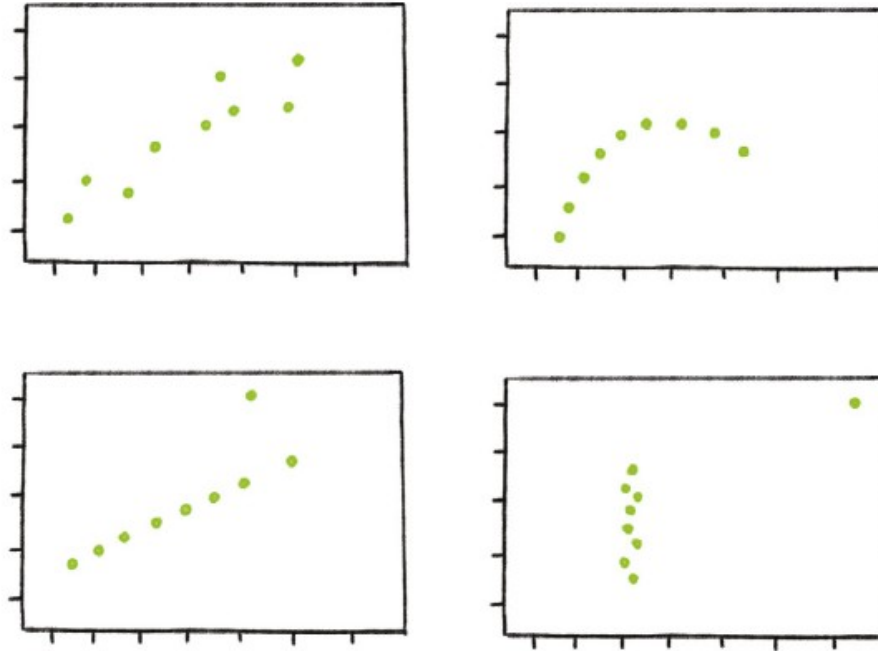


PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.

- Το πόσο ενδιαφέρον είναι ένα ερευνητικό εύρημα επηρεάζει την πιθανότητα δημοσίευσής του, διαστρεβλώνοντας την εντύπωσή μας για την πραγματικότητα.
-
- Για κάθε μελέτη που δείχνει στατιστικά σημαντικά αποτελέσματα, μπορεί να έχουν υπάρξει πολλές παρόμοιες δοκιμές που δεν ήταν οριστικές.
-
- Ωστόσο, τα σημαντικά αποτελέσματα είναι πιο ενδιαφέροντα να τα διαβάσει κανείς και επομένως είναι πιο πιθανό να δημοσιευτούν.
- Το να μην γνωρίζουμε πόσες «βαρετές» μελέτες κατατέθηκαν επηρεάζει την ικανότητά μας να κρίνουμε την εγκυρότητα των αποτελεσμάτων που διαβάζουμε.
-
- Όταν μια εταιρεία ισχυρίζεται ότι μια συγκεκριμένη δραστηριότητα είχε σημαντικό θετικό αντίκτυπο στην ανάπτυξη, άλλες εταιρείες μπορεί να έχουν δοκιμάσει το ίδιο πράγμα χωρίς επιτυχία, επομένως δεν μιλούν γι' αυτό.





DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.

- Για να καταδείξει το αποτέλεσμα, ο στατιστικολόγος Francis Anscombe συγκέντρωσε τέσσερα παραδείγματα συνόλων δεδομένων τη δεκαετία του 1970. Γνωστά ως το Κουαρτέτο του Anscombe, κάθε σύνολο δεδομένων έχει τον ίδιο μέσο όρο, διακύμανση και συσχέτιση.
- Ωστόσο, όταν απεικονίζονται γραφικά, είναι σαφές ότι κάθε ένα από τα σύνολα δεδομένων είναι εντελώς διαφορετικό. Το σημείο που ήθελε να τονίσει ο Anscombe είναι ότι το σχήμα των δεδομένων είναι εξίσου σημαντικό με τις συνοπτικές μετρήσεις και δεν μπορεί να αγνοηθεί στην ανάλυση.

10 κανόνες

Οι 10 κανόνες στην ανάλυση δεδομένων



- 1. Mixing up correlation and causation
 - Η ανάμειξη της συσχέτισης και της αιτιώδους συνάφειας είναι ιδιαίτερα επικίνδυνα κατά την εξερεύνηση ιστορικών δεδομένων ή όταν δεν υπάρχει σαφής υπόθεση προς διαψευσιν. Είναι καλύτερο να αντιμετωπίζετε τυχόν κοινά μοτίβα που συνέβησαν στο παρελθόν ως υποδηλωτικά αιτιώδους συνάφειας και όχι ως αιτία, μέχρι να αποδειχθεί το αντίθετο.
- 2. Αναμεινοντας τα δεδομένα να δώσουν απαντήσεις σε ερωτήσεις που δεν μπορείς να εκφράσεις (formulate)
 - Πάρα πολλές εταιρείες πιστεύουν ότι μπορούν απλώς να συλλέγουν δεδομένα, να χρησιμοποιούν τις πιο σύγχρονες τεχνολογίες και να προσλαμβάνουν ακριβούς επιστήμονες/αναλυτές δεδομένων/MBA για να κατανοήσουν την επιχείρησή τους. Η πραγματικότητα είναι ότι η ποιότητα της επιχειρηματικής σας ευφυΐας είναι άμεσα ανάλογη με το πόσο καλά μπορεί ο οργανισμός σας να διατυπώσει τα ερωτήματα που πρέπει να απαντήσει. Περισσότερα δεδομένα και ταλαντούχοι αναλυτές δεδομένων μπορούν να ενισχύσουν έναν οργανισμό που έχει μια σαφή διαδικασία λήψης αποφάσεων και παραγωγής, αλλά ο διπλασιασμός των μεγάλων δεδομένων δεν θα είναι το θαύμα που θα σώσει έναν οργανισμό που δεν έχει εστίαση.

- 3. Αναζήτηση δεδομένων για υποστήριξη αποφάσεων που ήδη έχουν παρθεί
 - Είναι σύνηθες να συλλέγετε δεδομένα, να τα αναλύετε και να καταλήγετε σε μια απόφαση όταν εσείς (ή άλλοι στην ομάδα) έχετε ήδη πάρει την απόφασή σας. Αντίθετα, θα πρέπει να διατυπώνετε υποθέσεις, να βλέπετε αν μπορείτε να τις διαψεύσετε και να ενημερώνετε την οπτική σας όταν τα δεδομένα παίρνουν την αντίθετη κατεύθυνση.

- 4. Ψαρεύοντας για το θετικό αποτέλεσμα
 - Ένα υποσύνολο αναζήτησης δεδομένων για την υποστήριξη μιας απόφασης: αναζήτηση δεδομένων που να υποστηρίζουν μια ρόδινη εικόνα. Υπάρχει πάντα κάτι που παρουσιάζει ανοδική τάση, ακόμη και σε εταιρείες που βρίσκονται σε τελικό στάδιο. Εάν όλες οι μετρήσεις που θεωρείτε σημαντικές δεν πάνε καλά, αποφύγετε τον πειρασμό να αναζητήσετε μετρήσεις που λένε μια πιο αισιόδοξη ιστορία.
- 5. Αναμενοντας πολύ καθαρά αποτελέσματα
 - Ακόμα κι αν έχετε δει πολλές ταινίες δράσης, όταν παρακολουθείτε έναν πραγματικό αγώνα πυγμαχίας, μπορεί να μην μπορείτε να τον κατανοήσετε. Αν έχετε συνηθίσει να βλέπετε χορογραφημένες σκηνές μάχης, γυρισμένες από τέλειες γωνίες με τέλειο φωτισμό και μοντάζ, το χάος και η ταχύτητα των πραγματικών μαχών μπορεί να σας μπερδέψουν.
 - Το ίδιο ισχύει και για άτομα που έχουν συνηθίσει να παρακολουθούν μαθήματα MBA ή ιδιαίτερα εξιδανικευμένες αναρτήσεις ιστολογίου όταν αντιμετωπίζουν ποσοτική λήψη αποφάσεων στην πραγματική ζωή. Στον πραγματικό κόσμο, τα αποτελέσματα μπορεί να είναι μικρά, ακατάστατα και πολυτροπικά (θα μιλήσουμε για τις παγίδες των μέσων όρων παρακάτω). Θα πρέπει να εργαστείτε με τα δεδομένα που έχετε, όχι με τα τέλεια δεδομένα που φαντάζεστε.

- 6. Expecting to A/B test your way to success

- Ενώ οι προσεκτικά σχεδιασμένες και καλά διεξαγόμενες δοκιμές A/B μπορούν να είναι μετασχηματιστικές για μια εταιρεία, συχνά οδηγούν επίσης στο να κυνηγάς την ουρά κάποιου. Βεβαιωθείτε ότι γνωρίζετε ποιο είναι ένα σημαντικό αποτέλεσμα πριν ξεκινήσετε μια δοκιμή A/B. Μην διακόψετε τη δοκιμή τη στιγμή που μία από τις επιλογές φαίνεται να αποδίδει καλύτερα και συμπεριλάβετε πάντα μια ομάδα ελέγχου. Και όσο μικρότερο είναι το αποτέλεσμα, τόσο μεγαλύτερος είναι ο αριθμός των χρηστών που θα χρειαστείτε. Εάν έχετε μόνο 10.000 μηνιαίους ενεργούς χρήστες, θα ήταν καλύτερα να αναβάλλετε απλώς οποιοδήποτε είδος δοκιμής A/B μέχρι να έχετε περισσότερα άτομα με τα οποία μπορείτε να δοκιμάσετε.
- Επιπλέον, οι δοκιμές A/B δεν θα καθορίσουν τα καλύτερα χαρακτηριστικά του προϊόντος ή το διαφημιστικό κείμενο για εσάς. Τα αποτελέσματα είναι τόσο καλά όσο οι επιλογές που δοκιμάζετε και τα αποτελέσματα είναι πολύ ευαίσθητα στο πόσο καλός είναι ο αρχικός σχεδιασμός. Μην αφήσετε το "θα το δοκιμάσουμε A/B" να γίνει ένα μότο που σταματά τη διαδικασία λήψης απόφασης για το τι είναι στην πραγματικότητα το προϊόν σας. Οι δοκιμές A/B χρησιμοποιούνται καλύτερα για να προσθέσουν αυτό το τελευταίο κομμάτι στιλβώματος.

- 7. Χρηση λανθασμένης χρονικής περιόδου
 - Εάν οι πελάτες σας αγοράζουν σε ένα χρονικό πλαίσιο πολλών μηνών και ο κύκλος παραγωγής του προϊόντος σας κινείται σε σπριντ δύο εβδομάδων, δεν χρειάζεστε αναλύσεις σε πραγματικό χρόνο. Ομοίως, εάν προσπαθείτε να διαγνώσετε σφάλματα σε λειτουργίες δικτύου όπου το κόστος της απώλειας μετριέται σε δεκάδες εκατομμύρια ανά λεπτό, καλύτερα να μην κοιτάτε ωριαία γραφήματα. Είναι σημαντικό να συνδέσετε την χρονική περίοδο αναφοράς με τη φυσική χρονική περίοδο λήψης αποφάσεων. Εάν εξετάζετε τα δεδομένα σας σε μια πολύ λεπτή χρονική περίοδο, θα καταλήξετε να είστε νευρικοί και να παλεύετε μεταξύ των αποφάσεων. Εάν χρησιμοποιείτε μια πολύ μεγάλη χρονική περίοδο, θα βρίσκεστε για πάντα τρεις κινήσεις πίσω.

- 8. Εξέταση μόνο των μέσων τιμών
 - Averages are a great place to hide uncomfortable truths. If you only use blended averages across organic and paid channels, you might end up ignoring the fact that your paid acquisition channels are becoming unsustainably expensive. If you look at average latency across all your web pages, you might not notice that your most important pages are getting slower over time. As a rule, when averages tell you something is getting worse, it's time to worry. When averages tell you things are looking good, it's time to dig deeper.
 - Most average-inspired delusions go away when you break the data out into a histogram. For example, rosy projections regarding the average cost of customer acquisition can disappear when you break the costs out by channel.
- 9. Εστίαση μόνο στα συνολικά (totals) αντι στις αλλαγές (rate of change)
 - Everyone loves charts that go up and to the right. “Total number of signups”, “Cumulative revenue”, and “Total value of goods sold” can make for good press, but for most situations, you should be looking at the rate of change, and possibly even the growth in that rate. If 95% of the information a metric carries relates to events that happened months or years ago, does that help you evaluate how you're doing today? Or how things will look tomorrow?

- 10. Μη-αξιολόγηση των αποτελεσμάτων μιας απόφασης
 - It's common to want to collect lots of information before making a big decision. However, once you make the decision, and the results start to trickle in, it's common to just assume things are going well.
 - Bad decisions are inevitable, but it's the **bad decisions you don't accept and correct** that end up hurting you the most. It's better to understand that you made the wrong call right away than it is find out you've been wrong for months.



- Don't use "data" as a crutch for lazy thinking
- Figure out what you really mean when you talk about data.
-
- It's easy to fall into the trap of using the word "data" and other analytics jargon as ambiguous placeholders for what you really mean, like a dumping ground for miscellaneous information you haven't quite figured out how to articulate.
-
- "Data" means different things to different audiences. Your engineering, legal, and marketing colleagues all have different ideas of what data is, and they're all correct. To cut through this confusion, it's important to be specific when talking about the fields and rows that make up your databases and the information you need to validate decisions. Sticking with straightforward and precise language will help your team form a clearer idea of what information you already have, what questions you're looking to answer, and how answering those questions will help people make progress toward your goals. And the better people understand your new analytics strategy, the more interested they will be, and the more they'll contribute to it.
-
- There's a certain irony to the guidance we're providing — yes, this is all broad advice about how you should be specific when figuring out how analytics should work at your organization, but stay with us here. And no, you probably won't stop using the word "data" entirely, but you can try to avoid ambiguous analytics corporate-speak whenever possible.

- [1]: <https://www.visualcapitalist.com/here-are-15-common-data-fallacies-to-avoid/>
- <https://www.geckoboard.com/best-practice/statistical-fallacies/>
- [2]: Schwarz, Christian A. (1996). NCD Implementation Guide. Carol Stream Church Smart Resources. p. 126. Cited in Brickman, p. 326.
-

- Αποστασιοποίηση από το γεγονός
- Χρήση πολλαπλής οπτικής
- Δεν καθοδηγούμεστε
 - ανέκδοτο με 3 φίλους στο ξενοδοχείο 30€
- Λάθος υποθέσεις, εκφώνηση
 - λεωφορείο επιβάτες, ανεβαίνουν - κατεβαίνουν
- Λανθασμένα δεδομένα
 - αυγό-σύνορα Ελλάδας Βουλγαρίας
-





dimitrisk@aegean.gr

facebook

Dimitris Kavroudakis



Dimitris Kavroudakis



dimitris_k



Instagram

d_kavroudakis



dimitriskavroudakis

Δρ. Δημήτρης Καβρουδάκης
Αν. Καθηγητής Γεωγραφικής
Ανάλυσης

www.dimitrisk.gr

