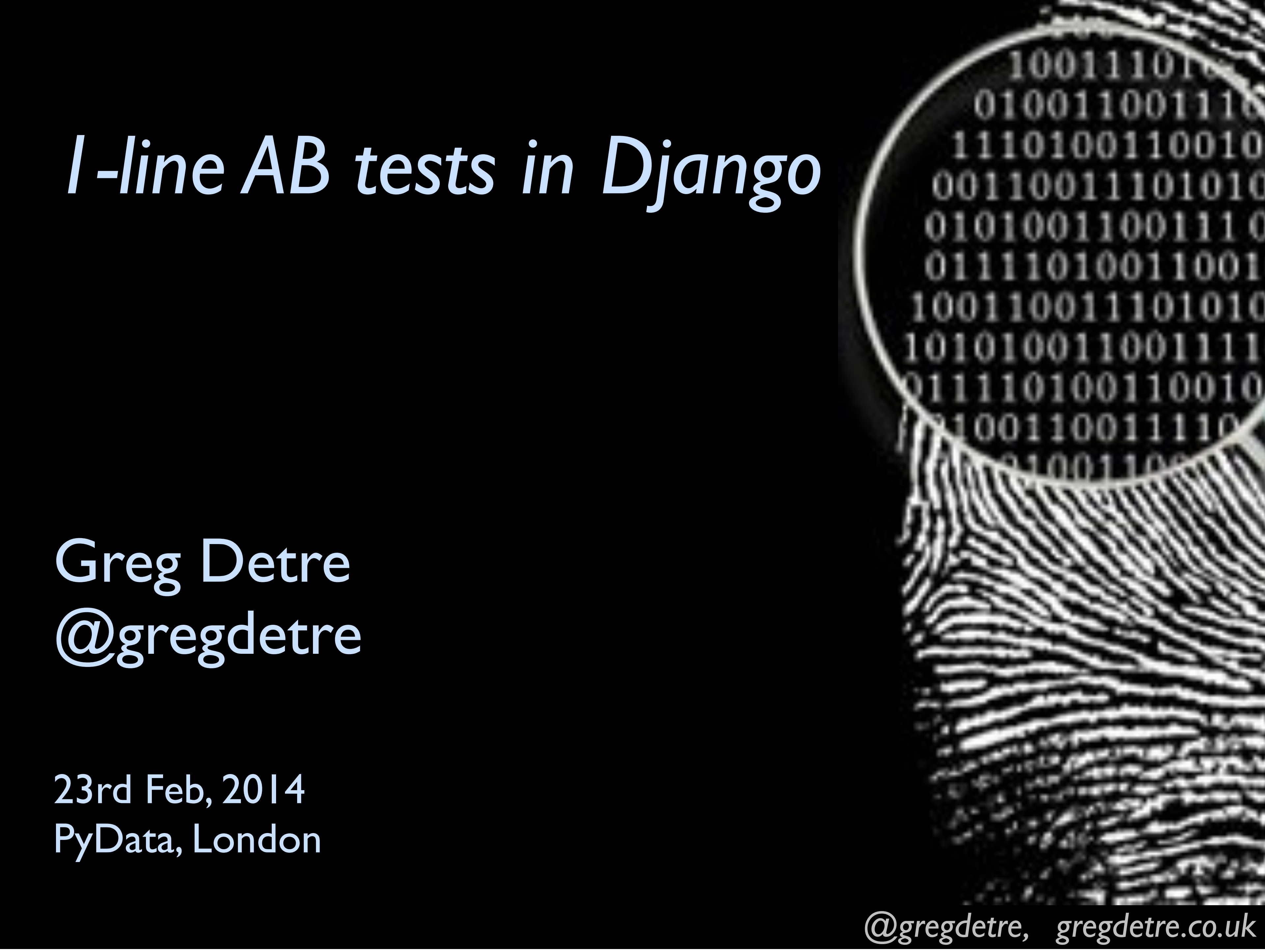


1-line AB tests in Django

Greg Detre
@gregdetre

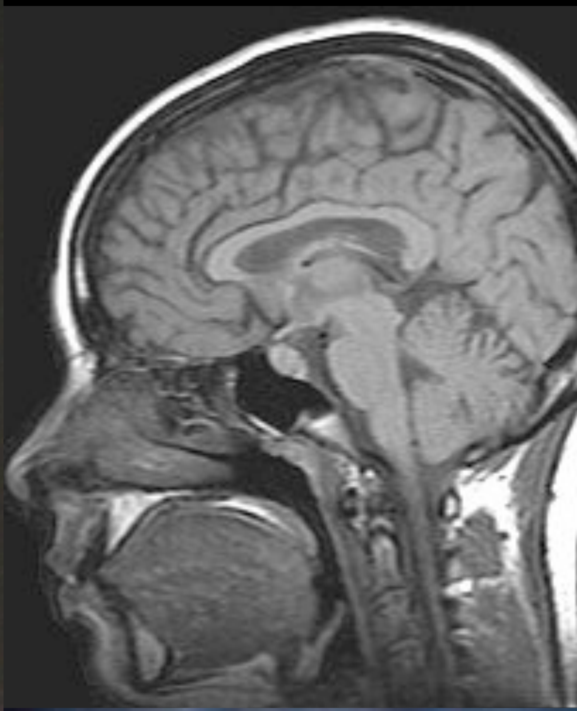
23rd Feb, 2014
PyData, London

Sunday, 23 February 2014

i will show you how to write a 1-line AB test in Django. but it’s only 1 line if you start sufficiently far to the left

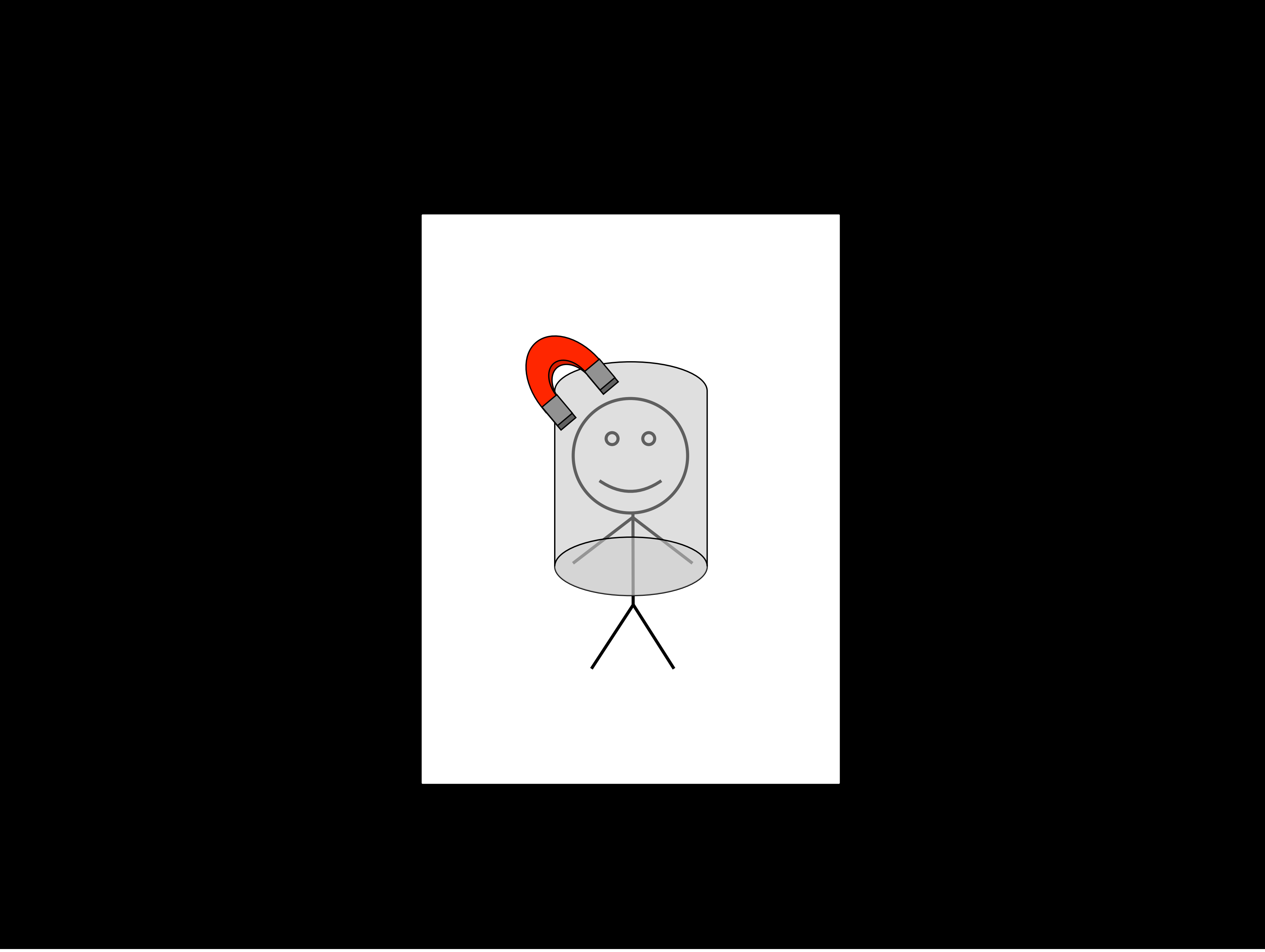
INTRO

Greg Detre



I'm Greg Detre

my PhD was on human memory & forgetting



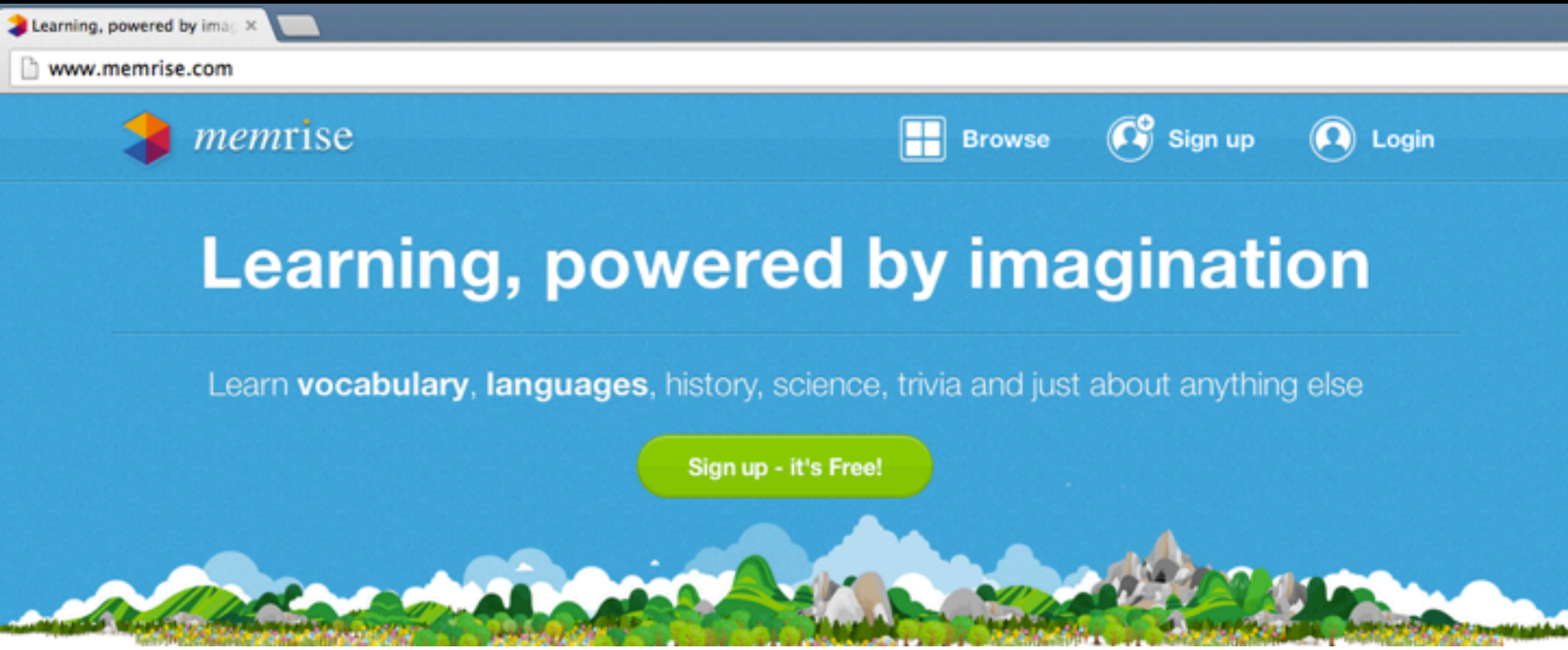
Sunday, 23 February 2014

i spent my days scanning people’s brains

including my own

it turned out to be smaller than I’d hoped

memrise



Learning, powered by imagination

Learn **vocabulary**, **languages**, history, science, trivia and just about anything else

Sunday, 23 February 2014

founded with Ed Cooke, grandmaster of memory, can remember a deck of cards in a minute flat
set out to combine the art, and the science, of memory, to help people learn 10 times faster
venture capital dance, millions of users
did a lot of AB testing, built our own internal framework

Sunday, 23 February 2014

helped build up their data science team
distil AB testing best practices for them

YOU

Hands up if...

you’ve run an AB test

Hands up if...

you’ve used Django

WHAT IS AN AB TEST?

When you release a change, you need to know whether you've made a big step forward...
Or taken two steps back.

The idea behind AB testing is very simple:
- when you change something
- show some people the old version
- show some people the new version
- look at which group are happiest

i.e. it's a scientific experiment on your product

Sunday, 23 February 2014

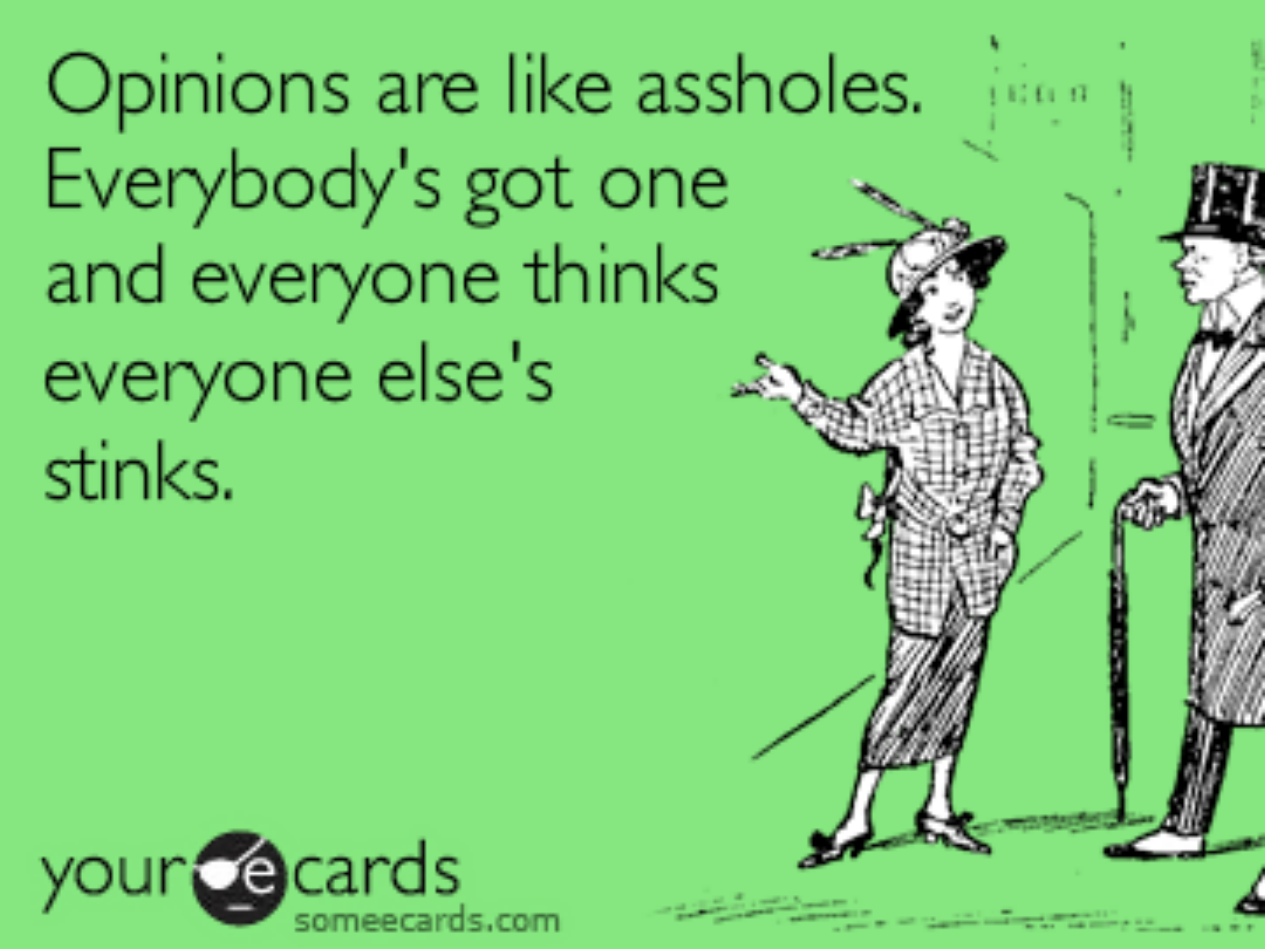
When you release a change, you need to know whether you've made a big step forward...
Or taken two steps back.

The idea behind AB testing is very simple:

- when you change something
- show some people the old version
- show some people the new version
- look at which group are happiest

i.e. it's a scientific experiment on your product

WHY RUN AB TESTS?


Opinions are like assholes. Everybody's got one and everyone thinks everyone else's stinks.
your ecards
someecards.com

Sunday, 23 February 2014

AB testing for making decisions

If opinions are like assholes, I work with a lot of opinions.

Sunday, 23 February 2014

this has nothing to do with the talk

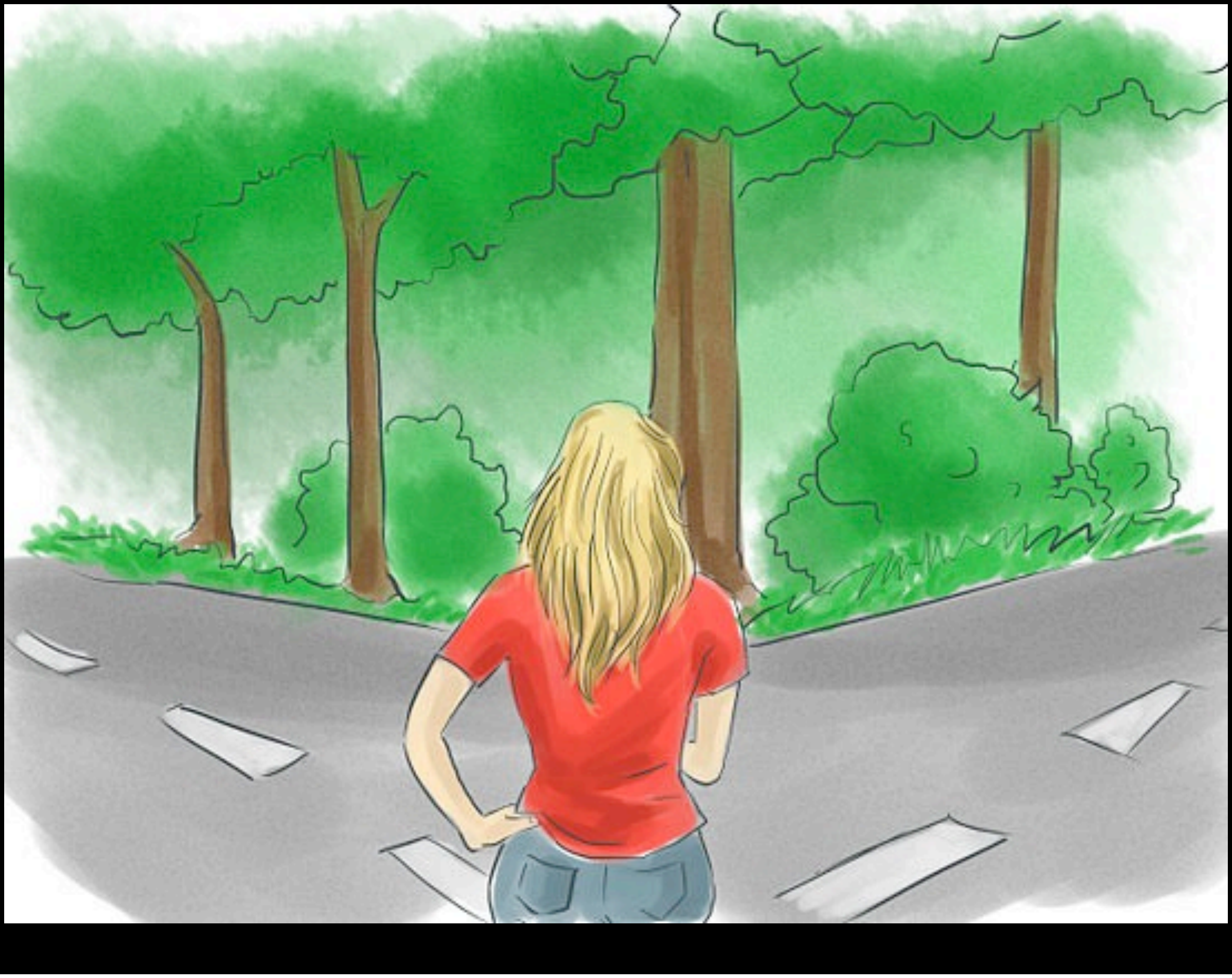
control for external factors

Sunday, 23 February 2014

If I’m a designer at The Guardian, and I change the font today.
Tomorrow, traffic increases by 50%.
Should I get a pay-rise?
Not if the paper just published the NSA leaks this afternoon.

By running old vs new simultaneously, you control for that surge in traffic. Both groups will show the boost, but you’re just looking at the difference between them.

improve your intuitions



Sunday, 23 February 2014

feedback loops, error-driven learning

PREFACE

Sunday, 23 February 2014

Sunday, 23 February 2014

yes, there are gotchas to AB testing

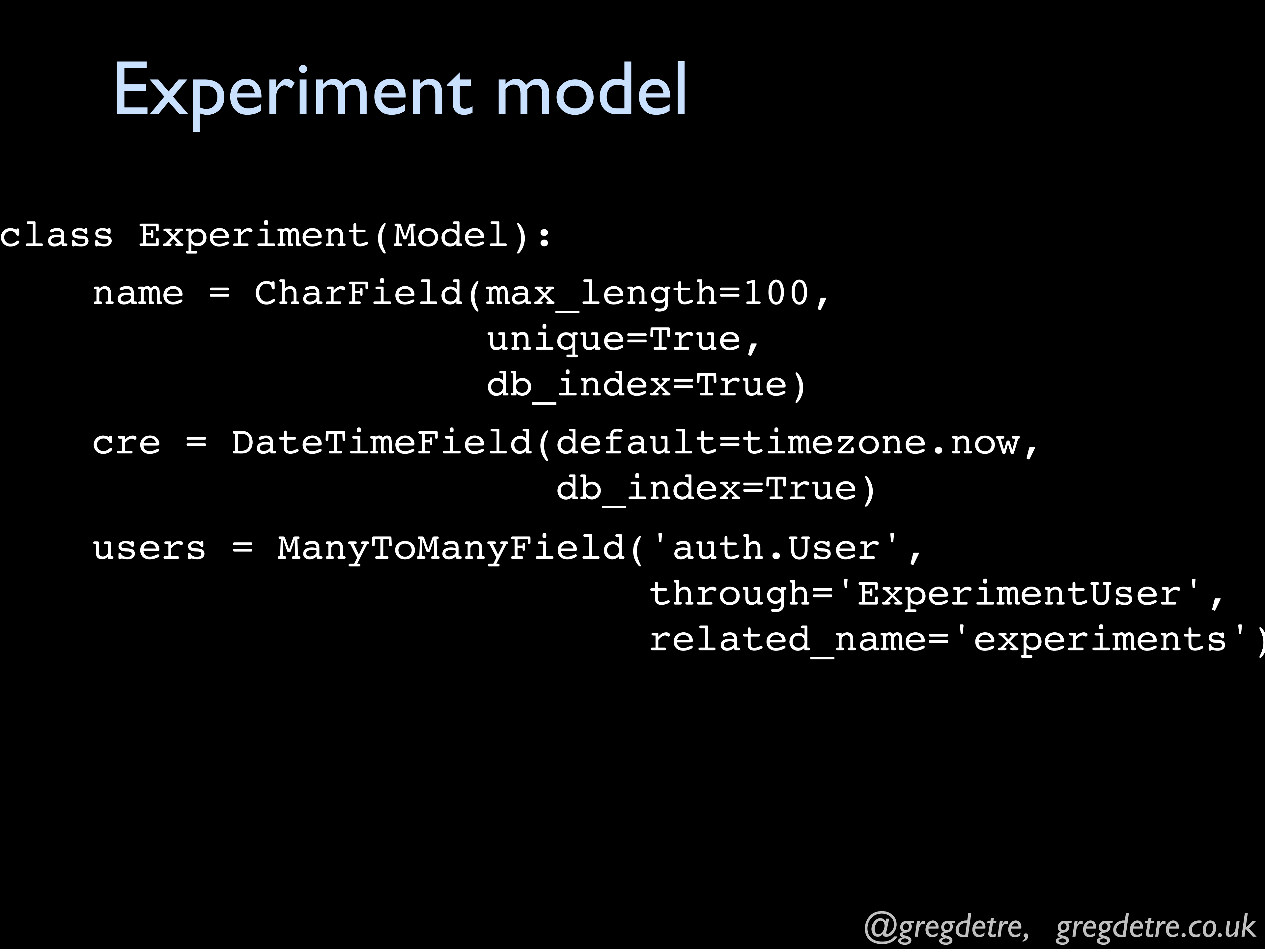
but the main problem in AB testing is that people don’t AB test often enough

CODE

I want to be able to do this

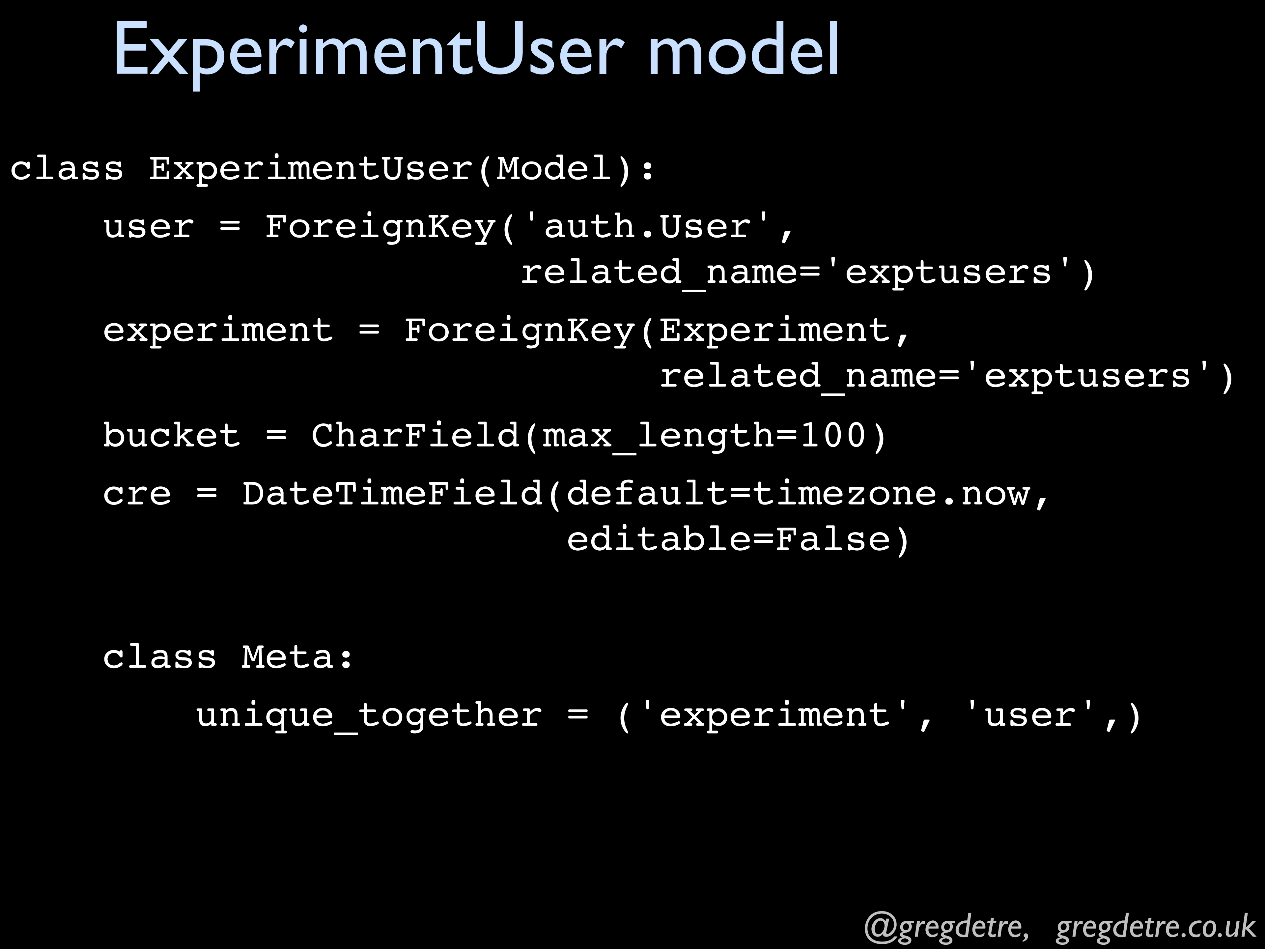
```
bucket = ab(user,
            ‘Expt 37 - red vs green buy button’,
            [‘red’, ‘green’])

if bucket == ‘red’:
    # show a red button
elif bucket == ‘green’:
    # show a green button
else:
    raise Exception(...)
```



Experiment model

```
class Experiment(Model):
    name = CharField(max_length=100,
                     unique=True,
                     db_index=True)
    cre = DateTimeField(default=timezone.now,
                        db_index=True)
    users = ManyToManyField('auth.User',
                            through='ExperimentUser',
                            related_name='experiments')
```



ExperimentUser model

```
class ExperimentUser(Model):
    user = ForeignKey('auth.User',
                      related_name='exptusers')
    experiment = ForeignKey(Experiment,
                            related_name='exptusers')
    bucket = CharField(max_length=100)
    cre = DateTimeField(default=timezone.now,
                        editable=False)

    class Meta:
        unique_together = ('experiment', 'user',)
```

Sunday, 23 February 2014

minimize FKs and indexes on ExperimentUser

Putting a user in a bucket

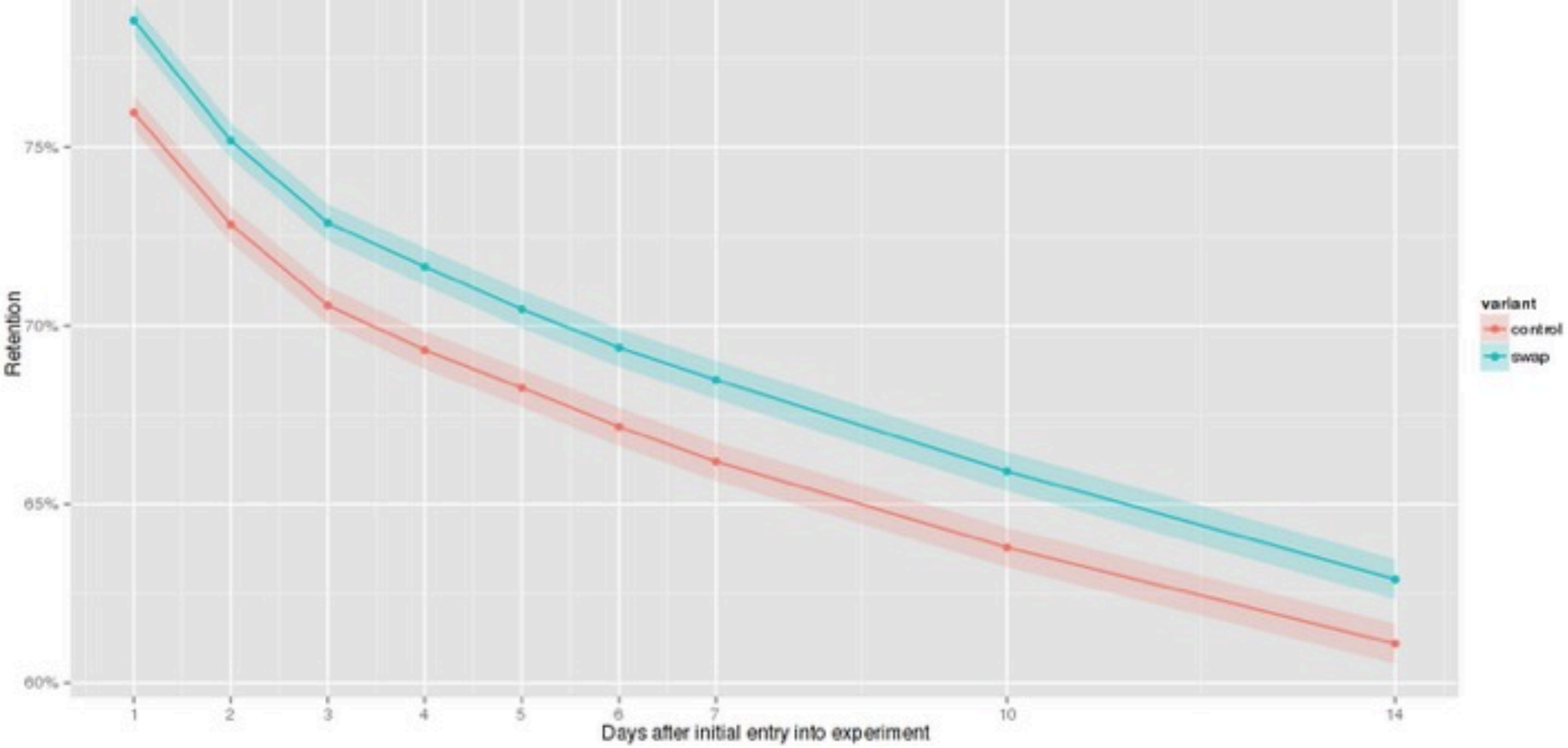
```
def ab(user, name, buckets):
    expt = Experiment.objects.get_or_create(name=name)[0]
    exptuser, cre = ExperimentUser.objects.get_or_create(
        experiment=expt, user=user)
    if created:
        exptuser.bucket = random.choice(buckets)
        exptuser.save()
    return exptuser.bucket
```

probably should be using default= in ExperimentUser get_or_create

actually, why not ExperimentUser.objects.get_or_create(experiment__name=name)???

SQL for calculating retention

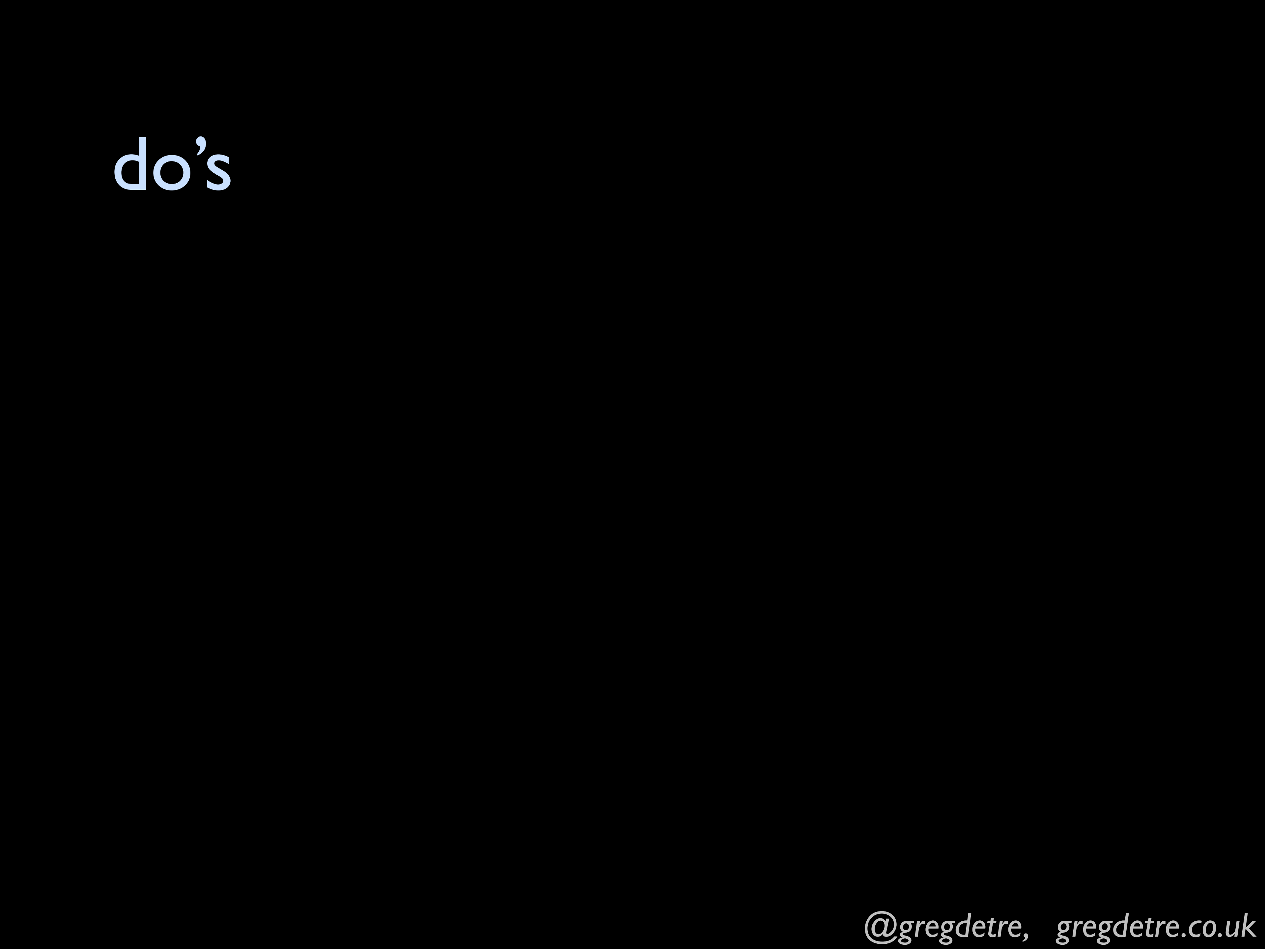
```
select
        d0.user,
        d0.dt as activity_date,
        'd01'::text as retention_type,
        case when dXX.dt is not NULL then true else false end
        as user_returned
from
        user_activity_per_day as d0
left join
        user_activity_per_day as dXX
on
        d0.user = dXX.user
        and
        d0.dt + 1 = dXX.dt
```


Retention
75%
70%
65%
60%
1
2
3
4
5
6
7
10
14
Days after initial entry into experiment
variant
control
swap

username	visited
greg	20 Feb 2014
ed	20 Feb 2014
greg	21 Feb 2014
greg	22 Feb 2014

github.com/gregdetre/abracadjabra

PRO TIPS



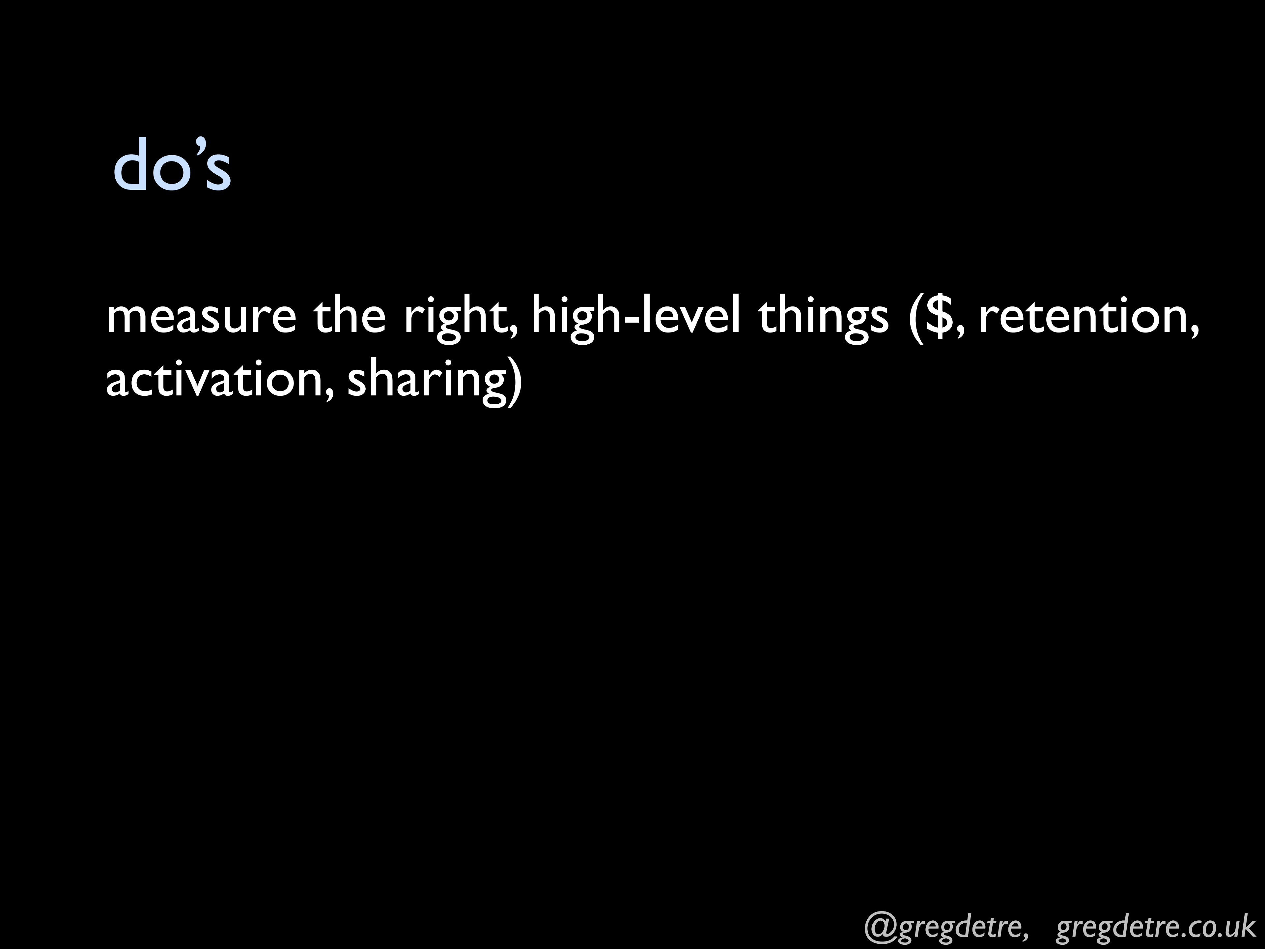
do’s

@gregdetre, gregdetre.co.uk

Sunday, 23 February 2014

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars



do’s

measure the right, high-level things ($, retention, activation, sharing)

@gregdetre, gregdetre.co.uk

Sunday, 23 February 2014

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars

do’s

measure the right, high-level things ($, retention, activation, sharing)

run on a subset

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars

do’s

measure the right, high-level things ($, retention, activation, sharing)

run on a subset

focus the analysis on relevant users

Sunday, 23 February 2014

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars

do’s

measure the right, high-level things ($, retention, activation, sharing)

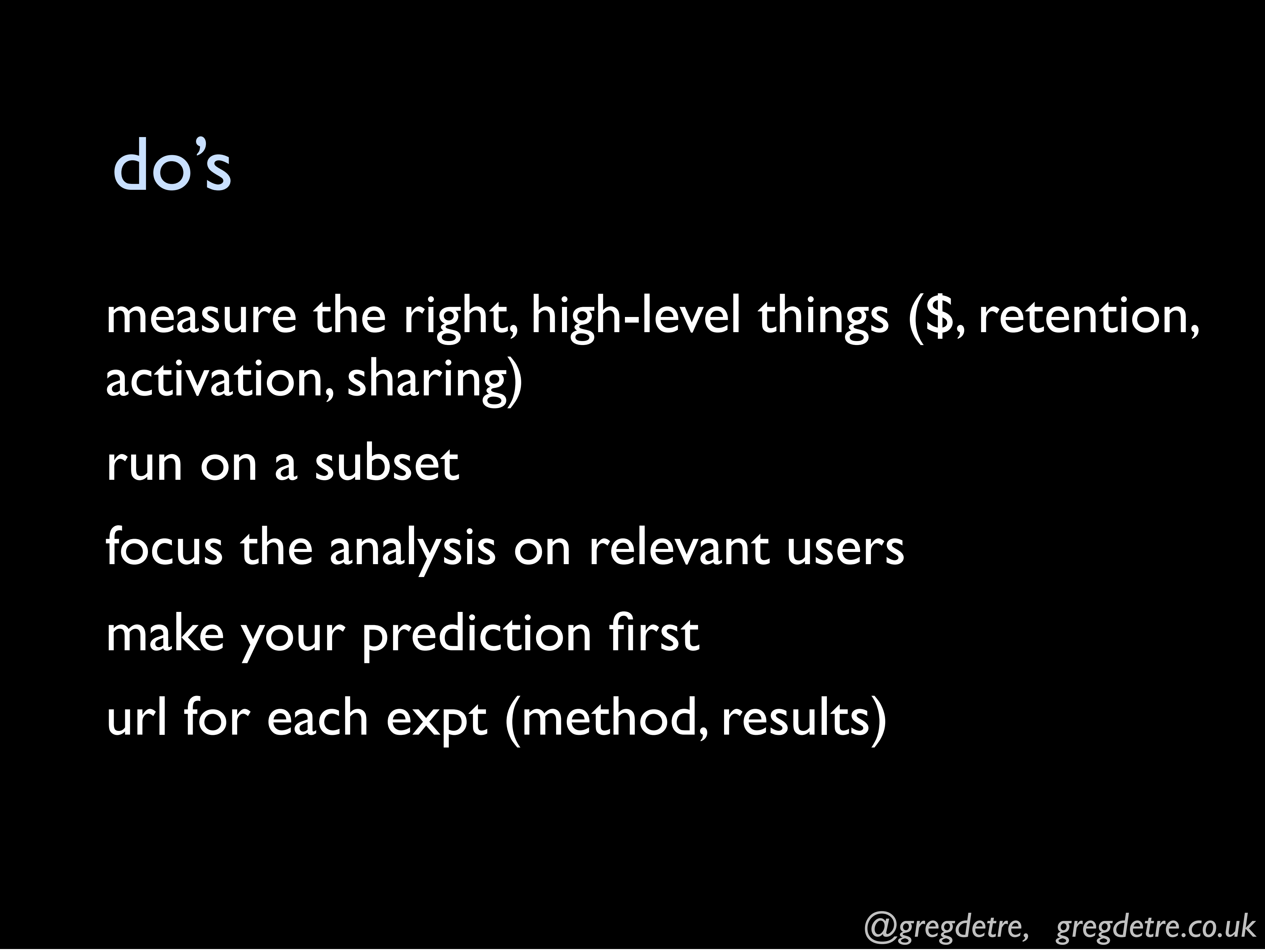
run on a subset

focus the analysis on relevant users

make your prediction first

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars



do’s

measure the right, high-level things ($, retention, activation, sharing)

run on a subset

focus the analysis on relevant users

make your prediction first

url for each expt (method, results)

Sunday, 23 February 2014

measure the right/high-level thing, so you can see if you're making things worse elsewhere/down the line

e.g. eBay hurt their sale of books, but increased sale of cars

don’ts

don’ts

don’t get lost in the weeds

don’ts

don’t get lost in the weeds

don’t expect your AB tests to succeed very often

don’ts

don’t get lost in the weeds

don’t expect your AB tests to succeed very often

don’t keep checking the results

don’ts

don’t get lost in the weeds

don’t expect your AB tests to succeed very often

don’t keep checking the results

sanity checks

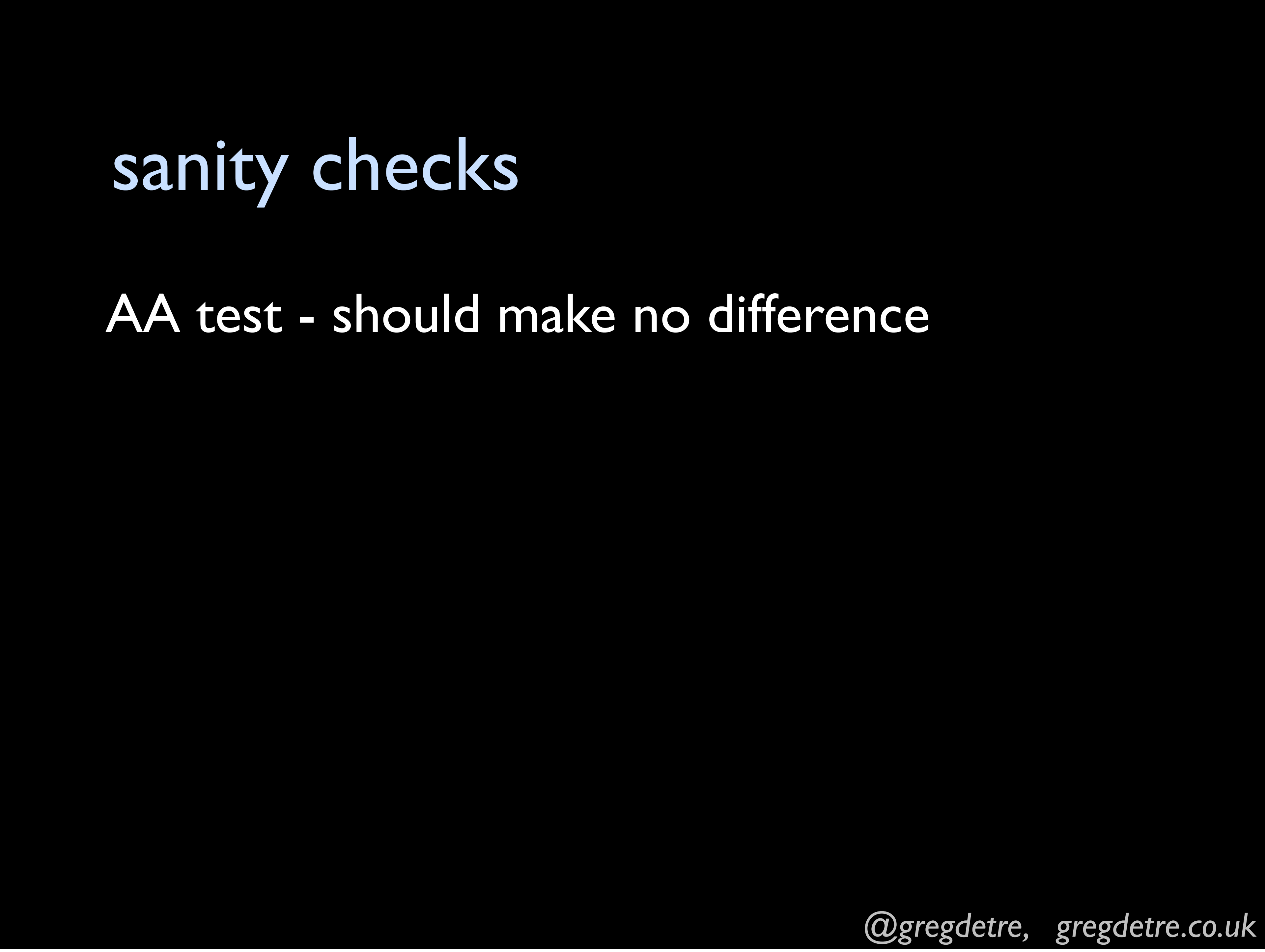
Sunday, 23 February 2014

e.g. if you make the site slower, how much does that hurt you? prioritise dev efforts. or what if you get rid of components? or get rid of ads?

sanity checks

AA test - should make no difference

e.g. if you make the site slower, how much does that hurt you? prioritise dev efforts. or what if you get rid of components? or get rid of ads?



sanity checks

AA test - should make no difference

@gregdetre, gregdetre.co.uk

Sunday, 23 February 2014

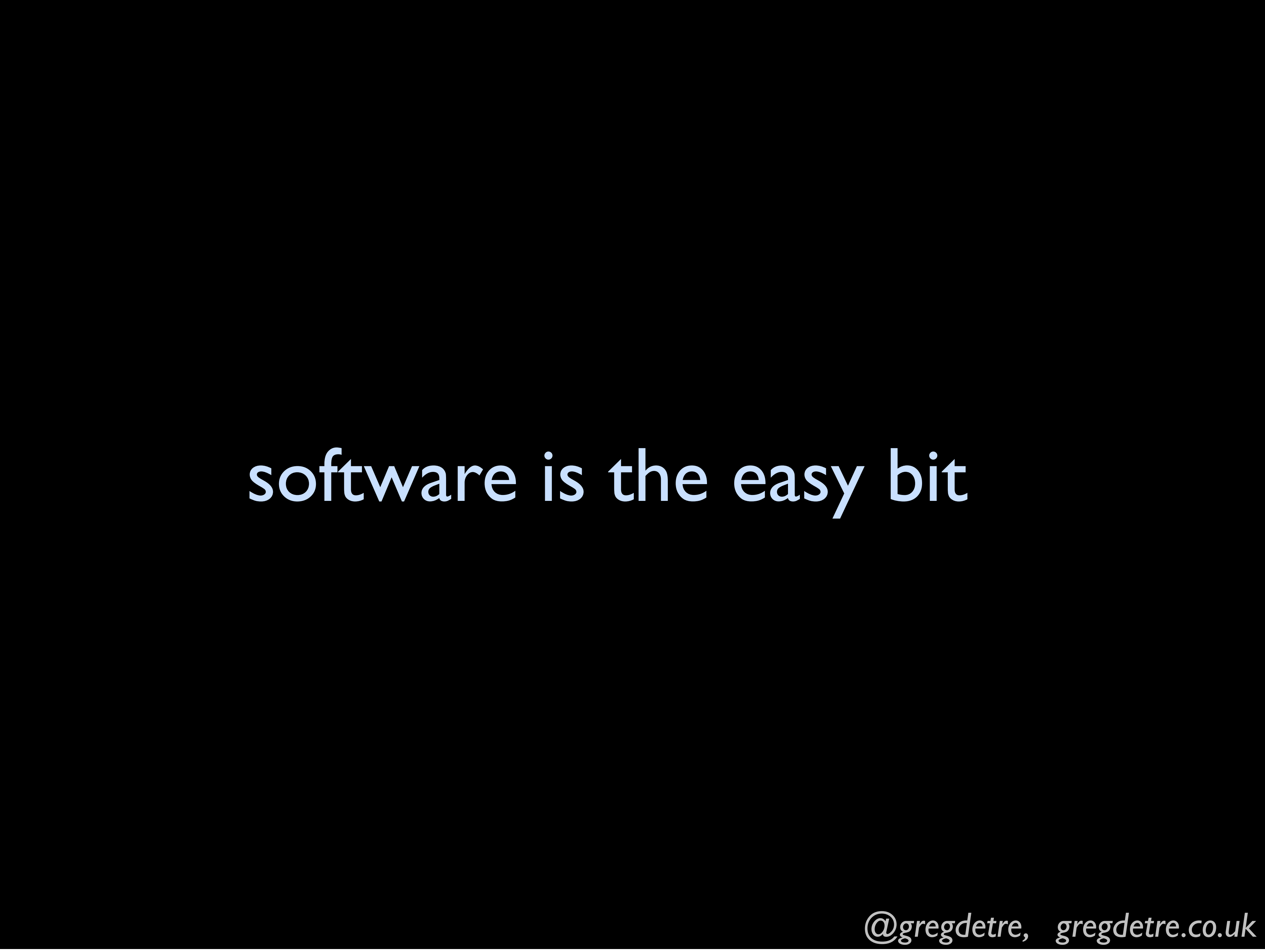
e.g. if you make the site slower, how much does that hurt you? prioritise dev efforts. or what if you get rid of components? or get rid of ads?

sanity checks

AA test - should make no difference

does making things worse make things worse?

e.g. if you make the site slower, how much does that hurt you? prioritise dev efforts. or what if you get rid of components? or get rid of ads?



software is the easy bit

@gregdetre, gregdetre.co.uk

Sunday, 23 February 2014

culture
human intuition to generate hypotheses vs being receptive to the results
most AB tests are null results
storing & sharing conclusions
the big changes are the most important to test, but the hardest

WORKING TOGETHER

software
science
startups

gregdetre.co.uk
@gregdetre

Sunday, 23 February 2014

i’m moving back to London
happy to help if you drop me a line. or you can hire me

THE END

@gregdetre, gregdetre.co.uk

Sunday, 23 February 2014

link to this presentation

resources

Eric Ries, The one line split-test, or how to A/B all the time

http://www.startuplessonslearned.com/2008/09/one-line-split-test-or-how-to-ab-all.html

Kohavi et al (2007), Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO

http://exp-platform.com/Documents/GuideControlledExperiments.pdf

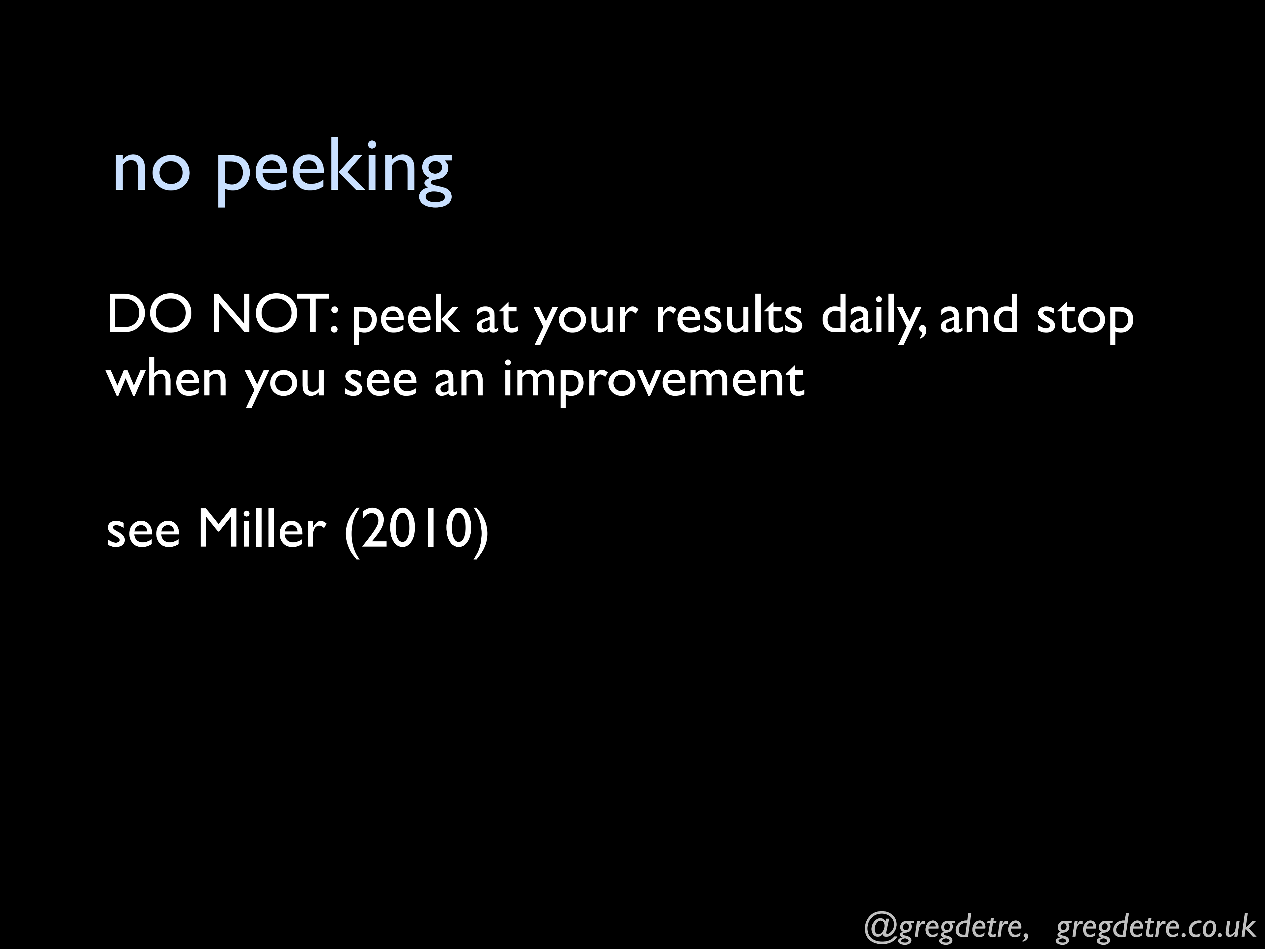
Kohavi et al (2013), Online Controlled Experiments at Large Scale, KDD.

http://www.exp-platform.com/Documents/2013%20controlledExperimentsAtScale.pdf

Miller (2010), How not to run an AB test

http://www.evanmiller.org/how-not-to-run-an-ab-test.html

APPENDIX



no peeking

DOT NOT: peek at your results daily, and stop when you see an improvement

see Miller (2010)

- say you start with a 50% conversion rate
- 2 buckets
- and you decide to stop when 5% significance or after 150 observations
- 26% chance of a false positive!

this is the worst case scenario (running a significance test after every observation)
but peeking to see if there’s a difference and stopping when there is inflates the chances of you seeing a spurious difference