Doing bad data science well

Ben Collier, University of Edinburgh

Data Science – how it should look

Huge amounts of novel social data

Variety, velocity, volume!

ML tools – automate analysis

Death of theory

But – in practice…

Usually not VVV.

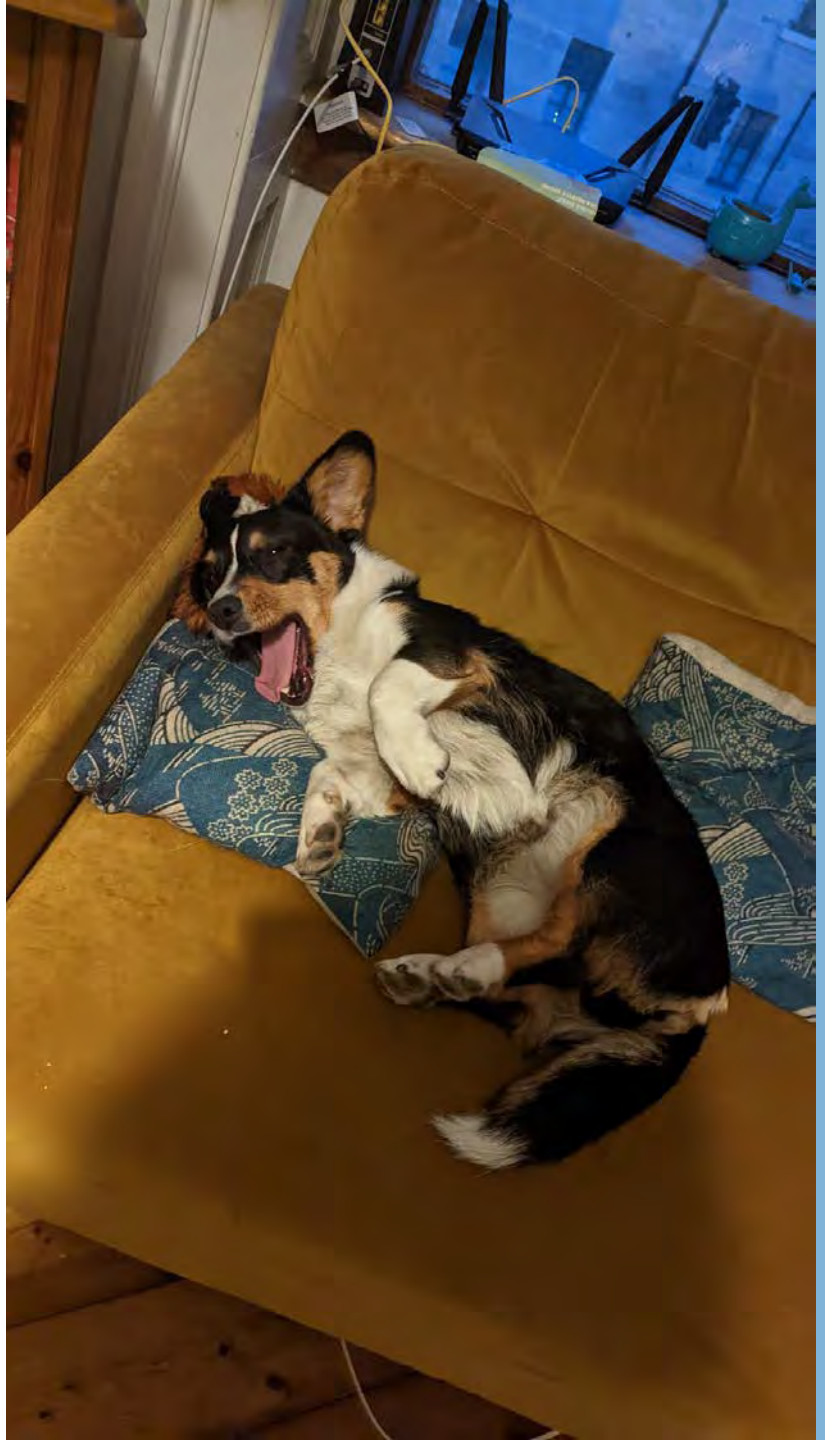
These methods really work best for operational development and process modelling

Insights – less so

A lot of the point of data science is feature reduction – the errors cancel out

Theory's back, and it's good again

Much more rubbish, interesting, and complicated than we thought



Cambridge Cybercrime Centre

Set up in Cambridge University Computer Lab

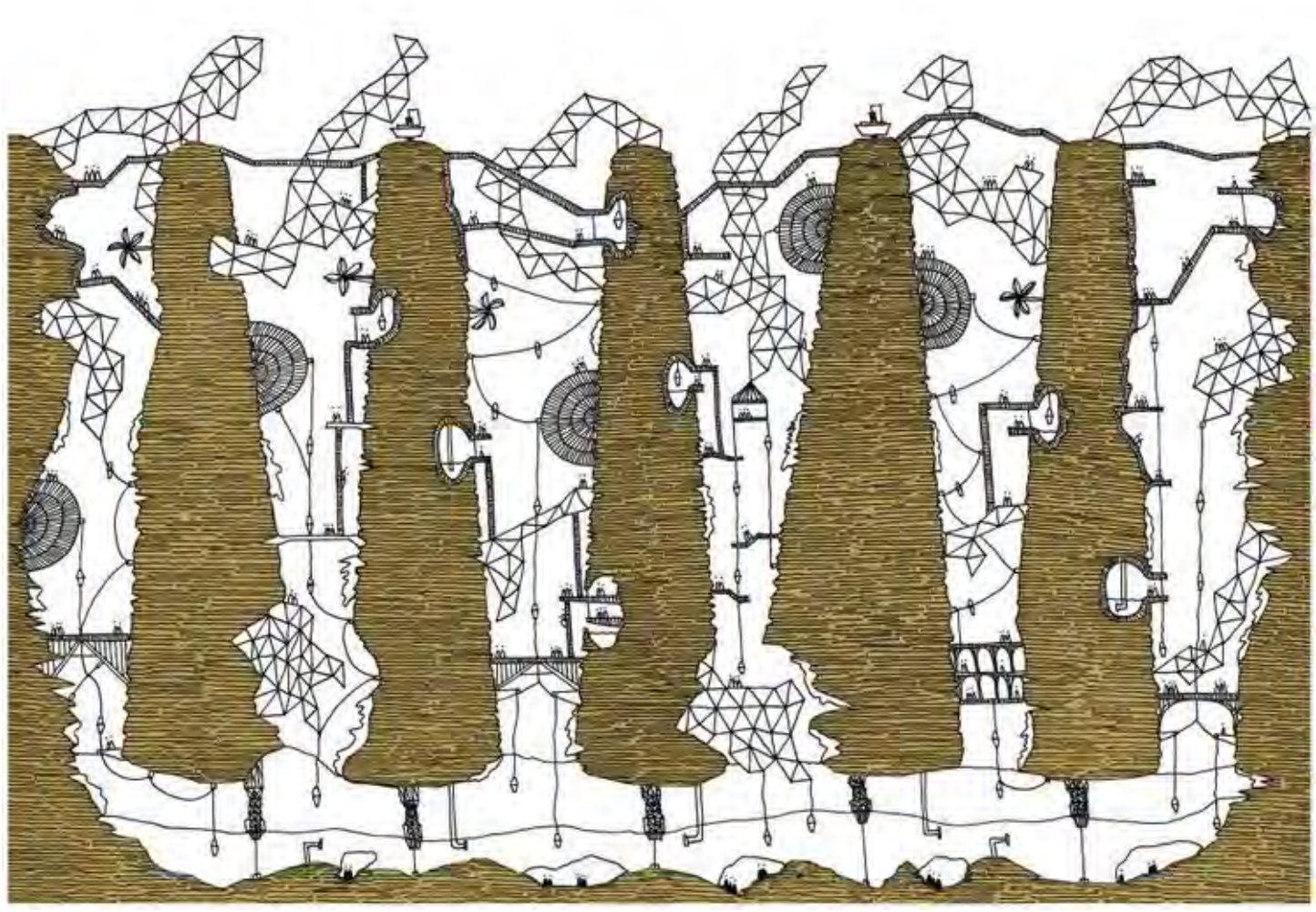
Problem – lots of scrappy data sources

Collect huge amounts of ‘authoritative’ data and license to researchers

Do our own research

Try to implement the Cathy O’Neill approach (old fashioned now!)

My attempts to do bad data science well



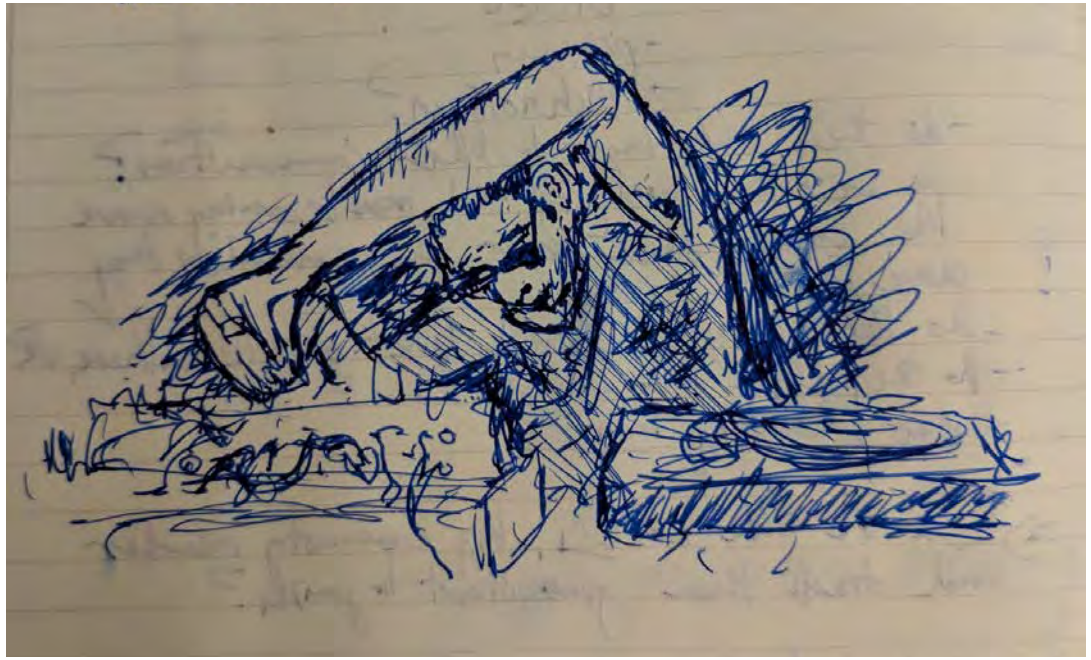
What are we going to do with all ~~this~~ ~~these~~ our data?

Look at the cybercrime-as-a-service ecosystem

Huge transformation about 15 years ago

Cybercrime now about small businesses running services

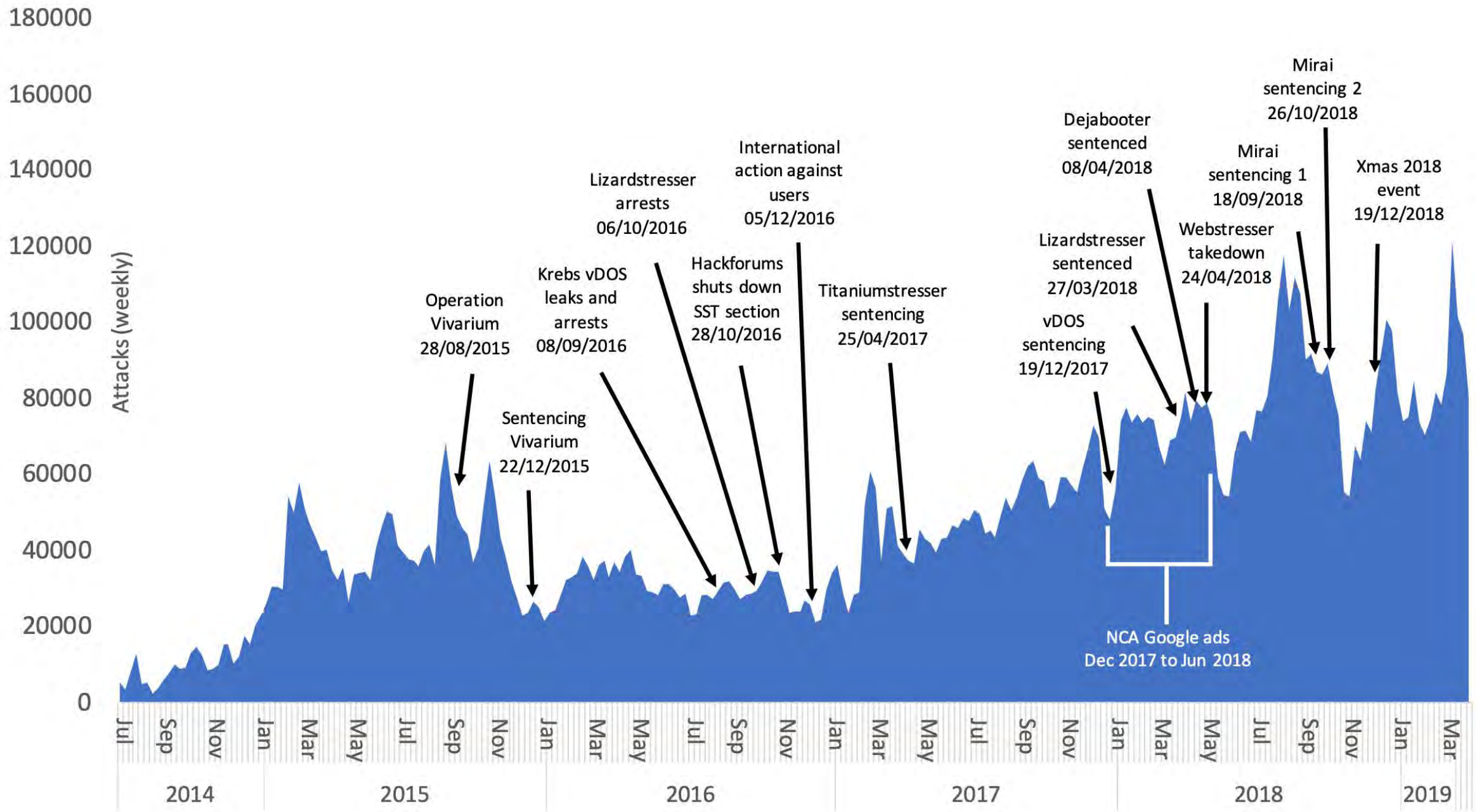
What effects do law enforcement interventions have and why?

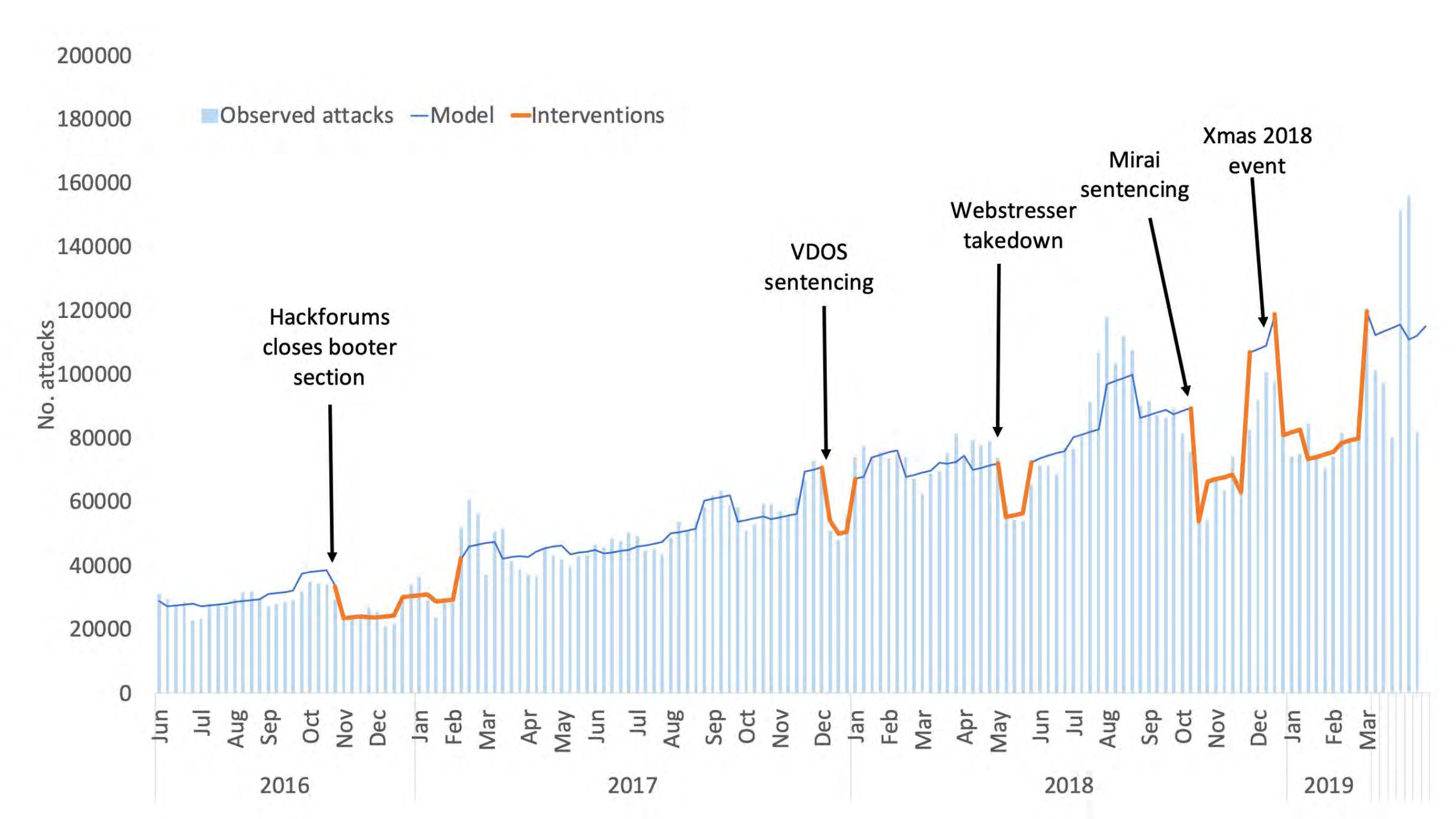


Honeypots – measuring crime directly

Set up servers which pretend to be crime infrastructure

Linear time series – count of attacks per country


Attacks (weekly)
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Operation Vivarium 28/08/2015
Sentencing Vivarium 22/12/2015
Krebs vDOS leaks and arrests 08/09/2016
Lizardstresser arrests 06/10/2016
Hackforums shuts down SST section 28/10/2016
International action against users 05/12/2016
Titaniumstresser sentencing 25/04/2017
vDOS sentencing 19/12/2017
Lizardstresser sentenced 27/03/2018
Dejabooter sentenced 08/04/2018
Webstresser takedown 24/04/2018
Mirai sentencing 1 18/09/2018
Mirai sentencing 2 26/10/2018
Xmas 2018 event 19/12/2018
NCA Google ads Dec 2017 to Jun 2018
Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar
2014
2015
2016
2017
2018
2019


200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
No. attacks
Observed attacks
Model
Interventions
Hackforums closes booter section
VDOS sentencing
Webstresser takedown
Mirai sentencing
Xmas 2018 event
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
2016
2017
2018
2019

Intervention		UK	US	RU	FR	DE	PL	NL	Overall
Xmas2018 Intervention 19/12/2018	Mean	-27%	-49%	-33%	-1%	-28%	-23%	-16%	-32%
	L95/U95	-43/-28%	-55/-42%	-43/-22%	-13/11%	-36/-20%	-37/-5%	-27/-3%	-37/-27%
	Duration	9 weeks	9 weeks	9 weeks	N/A	8 weeks	3 weeks	8 weeks	10 weeks
	Signif.	0.000**	0.000**	0.000**	0.828	0.000**	0.014*	0.018*	0.000**
Mirai sentencing and other actions 24/10/2018	Mean	-27%	-31%	-5%	-9%	-32%	-47%	-19%	-40%
	L95/U95	-42/-9%	-41-20%	-16/7%	-31/21%	-40/-23%	-56/-36%	-35/0%	-46/-34%
	Duration	2 weeks	7 weeks	2 weeks	N/A	6 weeks	2 weeks	6 weeks	8 weeks
	Signif.	0.006**	0.000**	0.41	0.533	0.000**	0.000**	0.053	0.000**
Webstresser takedown 24/04/2018	Mean	-10%	-24%	-16%	-22%	-29%	-29%	146%	-21%
	L95/U95	-21%/3%	-40/-4%	-33/6%	-35/-7%	-36/-22%	-42/-14%	94/211%	-30/-12%
	Duration	N/A	4 weeks	2 weeks	4 weeks	9 weeks	6 weeks	4 weeks	3 weeks
	Signif.	0.120	0.022*	0.14	0.006*	0.000**	0.001**	0.000**	0.000**
vDOS sentencing 16/12/2017	Mean	-20%	-4%	-37%	-30%	-4%	16%	-24%	-24%
	L95/U95	-33/-5%	-18/12%	-47/-24%	-37/-23%	-17/10%	-17/62%	-33/-13%	-32/-25%
	Duration	3 weeks	3 weeks	2 weeks	2 weeks	N/A	N/A	3 weeks	3 weeks
	Signif.	0.011*	0.563	0.000**	0.000**	0.532	0.373	0.000*	0.000**
HackForums 28/10/2016	Mean	-48%	-30%	-13%	-52%	-32%	2%	-35%	-30%
	L95/U95	-53/-42%	-37/-21%	-23/-3%	-59/-43%	-41/-23%	-19/28%	-42/-27%	-33/-25%
	Duration	15 weeks	7 weeks	14 weeks	15 weeks	7 weeks	N/A	15 weeks	13 weeks
	Signif.	0.000**	0.000**	0.02*	0.000**	0.000*	0.86	0.000*	0.000**

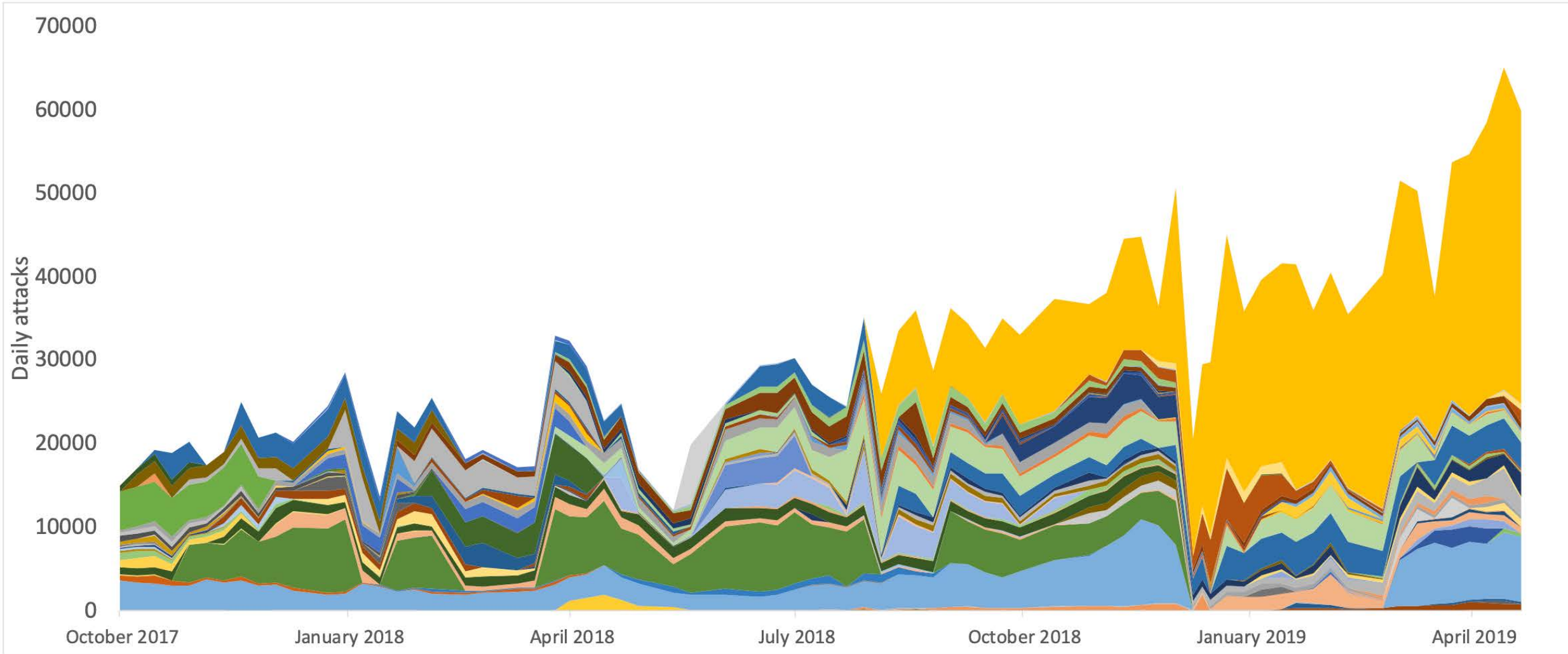
Secret bonus analysis

We had also been collecting what the illegal services *said* they were doing – they advertise a weekly number of attacks

Taking notes on a criminal conspiracy...

Journal hated it, but we used some fun stats magic to ‘prove’ that it was ‘reliable’ and they relented

Normally distributed and heteroskedastic


70000
60000
50000
40000
30000
20000
10000
0
Daily attacks
October 2017
January 2018
April 2018
July 2018
October 2018
January 2019
April 2019

Begin initial analysis – problem generation

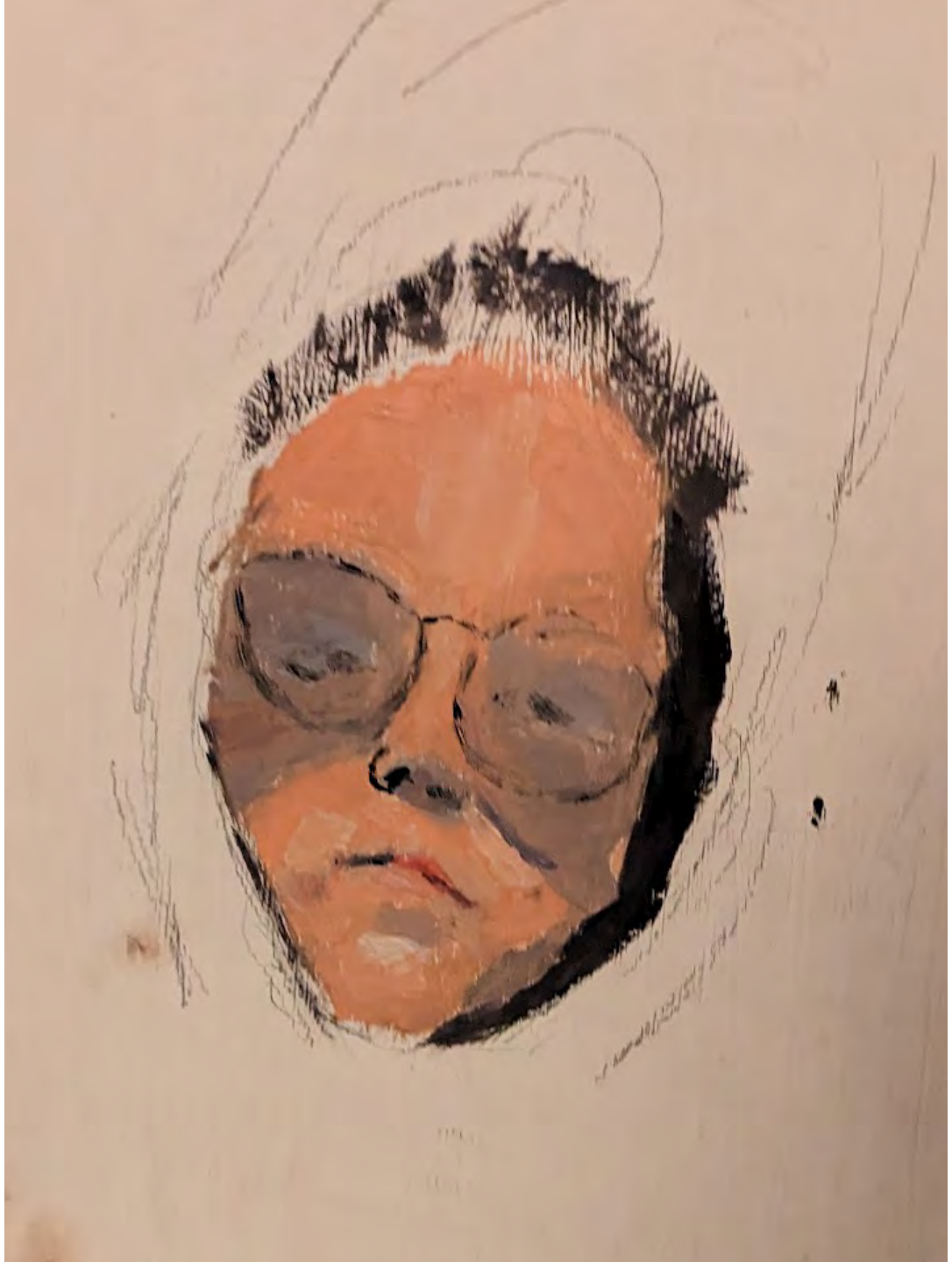
Started by trying to expain our quants findings

Why do some of these strategies work and some don’t?

Hugely overdetermined – we can see what works (ish) but no idea why

Could write a rubbish quants paper at this point but decide not to

Carry out more research!



Interviews

Interview people in the communities running these attack platforms and users Easier than you think – but you get lied to a lot

CrimeBB – data science

Flagship collection of cybercrime forums

Home-made scrapers

Millions of posts across 10 years and 30 forums

CrimeBB – enrichment

Sentiment
Post-type
Crime type
Topics
Group based trajectory modelling
Data tools

In practice…

Research often just doesn't work

QUANTS BAD

Our quants collections were pretty good, but incredibly overdetermined.

QUAL BAD

Our qual collection was based on quite a small number of interviews...

DATA SCIENCE WORSE

And our data science modelling was a nightmare. NLP fell over just looking at the complexity and size of our data, and our GBTM models first refused to converge, then worked but wrecked my hard drive, and finally gave results that look like every other GBTM in the world...

Social Data Science to the rescue!

Take what was good about our collections – the range of different types of knowledge

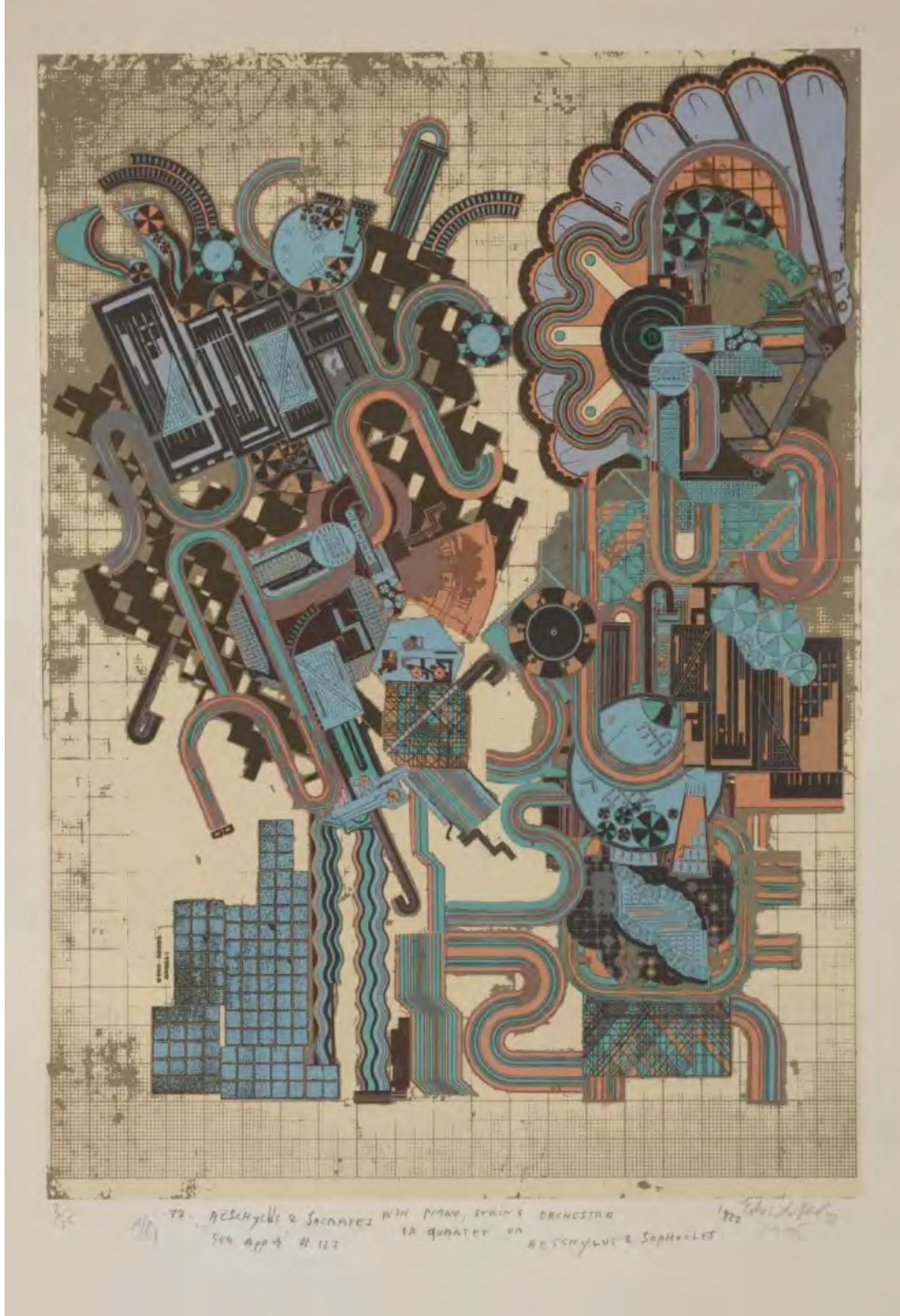
Iterate back and forward,

First – we immerse

- Read the forums manually every day for 20 minutes and a good scan of the new chats
- Try to mimic a lurker
- Tool design round 1 has been done
- But need to build tools to probe the dataset
- Think – what do I need? What can’t I get just from looking?

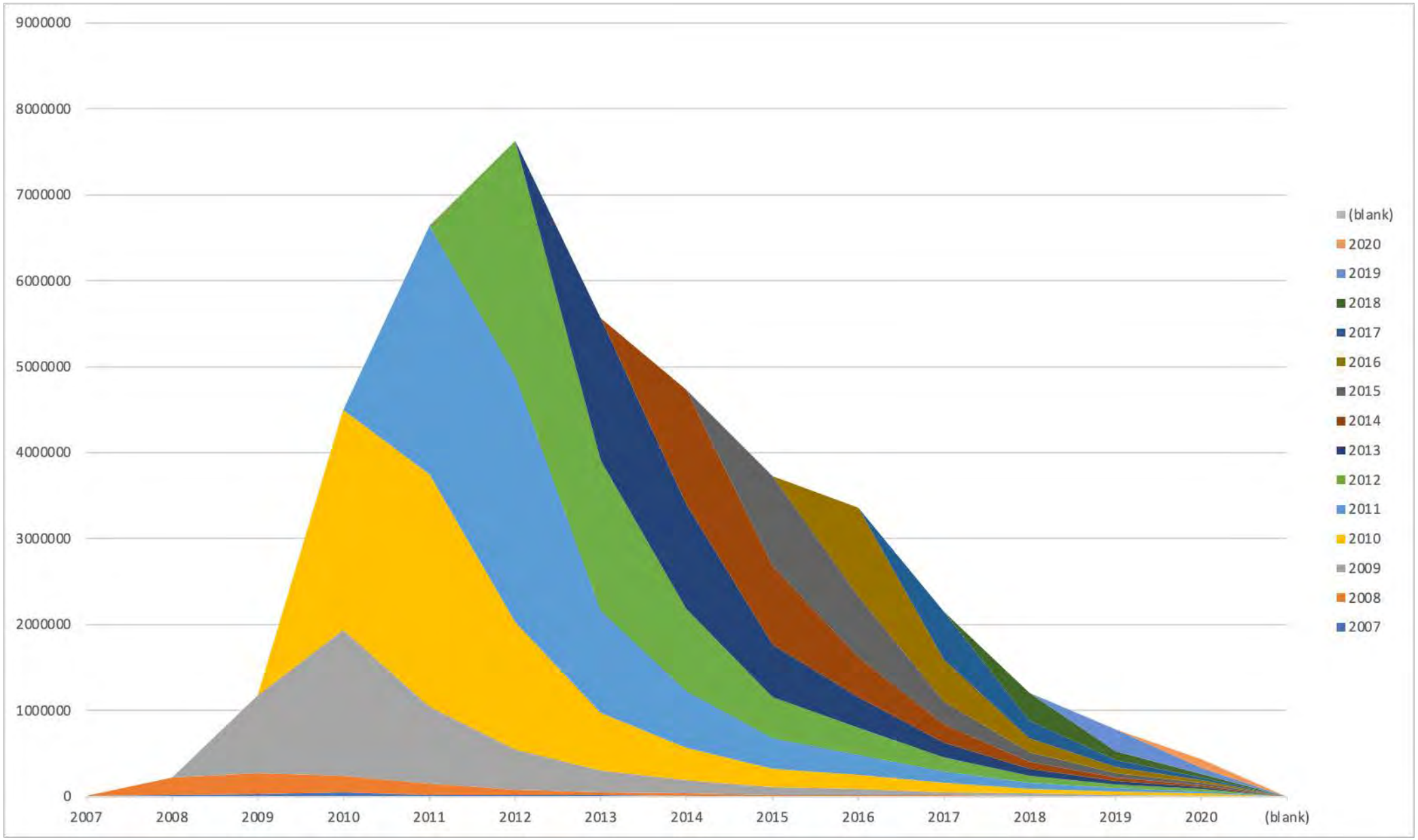
Second – we iterate

- Working with both kinds of data as a team
- Weekly meetings (‘stand ups’) and over-the-shoulder chats
- Draw out the evolving ‘models’ of the codes, concepts, and situations on shared whiteboards
- Bring interim findings and codes back and forward to one another
- Eventually these stabilize – each model should have a list of assertions/findings and the linked evidence
- Plan two papers (one technical and one sociological) from the start - sieve findings into buckets then group buckets within papers



Second – we iterate

- Even active people don't stay in the community for more than a year except in exceptional cases – and permanence getting lower as time goes on
- Huge range of topics – hacker culture, lessons, marketplace etc. Lots of cybercrime services
- Drill down into a particular subforum, then subset again for keywords like 'work' (or which reference particular kinds of work)
- Scan through and sample posts, find language used, then run queries and see if it looks like what you thought it would
- Get a subset of 'meaty' (relevant) discussions
- Apply to wider dataset manually (or topic model within your subset, then apply the topic classifier to the broader dataset)
- Group based trajectory modelling (more or less a time-series LCA)


9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
(blank)
(blank)
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007

Third: initial interviews

Map practices, people, actors, technologies, ideas

Idea of boredom comes up a lot

Hacker ideas still really important – skill etc.

But actual work is very simple

Stage 4: deep-dive

- Expand the map
- Take the concepts from the interviews and search them
- Theory-driven (Star, labour alienation)
- How do key words change over time?
- How is involvement changing over time?
- Can we search for posts using the language we're seeing in the interviews?
- Sort topic models then sample within topics, then model topics within topics etc.
- Can we model posts like this versus other sorts of posts?
- Establishing our rough key themes
  - Very few posts talk about genuinely technical stuff – mostly about low-level, boring, or facile work
  - Lots of evidence of boredom
  - Work is mostly maintenance and administration, or marketing and customer service
  - Technical talk getting less over time
  - Evidence of deskilling and gradual breakdown

Stage 5: exploration

Go back to interviews

Pull findings from the data – established some key themes now

Explore qualitative depth, new findings from data

Findings

People are bored – directly contradicts the core literature, which views hacking as exciting

Mostly administrative work

No respect

Exit through burnout (you can even look at people's last posts)

None of our sources was authoritative on its own – but we could evidence the wider applicability of our interview findings using the quantitative and data science data

Why do some police actions work and not others? Making the work even more boring! Alienation of labour.

Why does cybercrime increase during the pandemic, but not really change its nature? Well-developed service market

Immersive data ethnography

AKA ‘social data science”

Focus on multiple types of data

Immersion – how was data produced

Patch up the rubbish by attacking your data from all directions

READ THE MICRODATA

A lot of this is about how you write it up

THANK YOU

@johnnyhistone