Centre for
Data Analytics

Insight

Data Analytics on VLE Access Data How much can we mine from a mouseclick ?

Alan F. Smeaton

Sinéad Smyth
Owen Corrigan
Aly Egan and John Brennan

Why am I here ?

- Computer Scientist, content-based Information Retrieval
- Machine learning and analytics are important to me
- I did some work on capturing lectures on video (1996) but nothing since
- Looked at analytics for education .. So much is descriptive analytics, analysing and explaining the past, looking for reasons to explain why
- Predictive analytics, data that can predict the future, much more useful, and is a **data-driven** approach

Outline

Motivation and goals
Ethics and approval
A primer on machine learning
Selecting the modules
Building the system
The interventions
- What the Student sees
- What the Lecturer sees

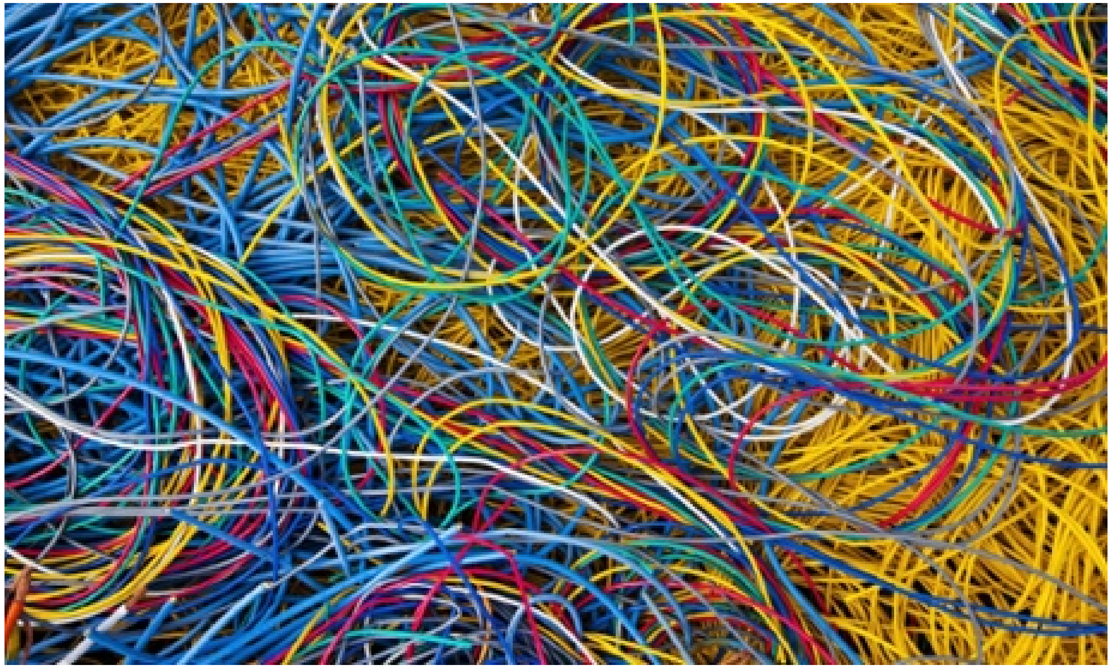
Roll-out
Future plans

Motivation

Guardian Newspaper Article, 2013

University data can be a force for good

Data analytics shouldn't be seen as a dark art but a tool to aid student retention and enhance experience, says **Ruth Drysdale**



If managed well, data analytics can aid student retention and enhance experience. Photograph: Google

So much student data available …

Demographics

Age, home/term address, commuting distance, socio-economic status, family composition, school attended, census information, home property value, sibling activities, …

Academic Performance

CAO and Leaving cert, University exams, course preferences, performance relative to peers in school

On-Campus Activities

Library access, sports centre, clubs and societies, eduroam access yielding co-location with others and peer groupings, lecture/lab attendance,

Online Behaviour

Mood and emotional analysis of Facebook, Twitter, Instagram activities, friends and their actual social network, access to VLE (Moodle)

So much student data we could use …

Demographics

Age, home/term address, commuting distance, socio-economic status, family composition, school attended, census information, home property value, sibling activities, …

Academic Performance

CAO and Leaving cert, University exams, course preferences, performance relative to peers in school

On-Campus Activities

Library access, sports centre, clubs and societies, eduroam access yielding co-location with others and peer groupings, lecture/lab attendance,

Online Behaviour

Mood and emotional analysis of Facebook, Twitter, Instagram activities, friends and their actual social network, access to VLE (Moodle)

… which might be just a bit more palatable

What do we do ?

We use this student data on a weekly basis to predict likelihood of pass-fail in a given module, for each of c.1,600 students

But before the details, lets talk ethics

Importance of Ethics

- Ethics are important to ensure safety of participants and researchers
- Educational Data Analytics is a new area of research
  - Not much previous research to highlight possible ethical issues
  - Requires extensive ethical consideration
- Analytics in business are reployed in an ethically embivalent way
- We have spent a lot of time obtaining institutional approval, converting to the tools and norms that institutions know
- We are following the 8 Principles set out by the Open University who are at EXACTLY the same stage as us

Centre for Data Analytics

Open University Principles

Insight

Open University Principles	This study
Learning analytics is a moral practise which should align with core organisational principles	This study was carried out following DCU's core principles
The purpose and boundaries regarding the use of learning analytics should be well defined and visible	Outlined in PLS
Students should be engaged as active agents in the implementation of learning analytics	Intervention is interactive
The organisation should aim to be transparent regarding data collection and provide students with the opportunity to update their own data and consent agreements at regular intervals	Students change their data by engaging with Moodle
Modelling and interventions based on analysis of data should be free from bias and aligned with appropriate theoretical and pedagogical frameworks wherever possible	Predictions are calculated using the same algorithm therefore not biased
Students are not wholly defined by their visible data or our interpretation of that data	Data used is from Moodle usage alone and therefore does not in any way define an individual
Adoption of learning analytics within the organisation requires broad acceptance of the values and benefits (organisational culture) and the development of appropriate skills	All researchers have the appropriate skills required to handle the data
The organisation has a responsibility to all stakeholders to use and extract meaning from student data for the benefit of students where feasible	Results of the study will be used to better understand how to increase student engagement

We have approval but questioned at many steps … everybody is very nervous about this

Using data analytics like this is minor in light of our everyday exposure to data analytics … from the length of prison sentences in the US, to our retail purchases, it is a fact of life

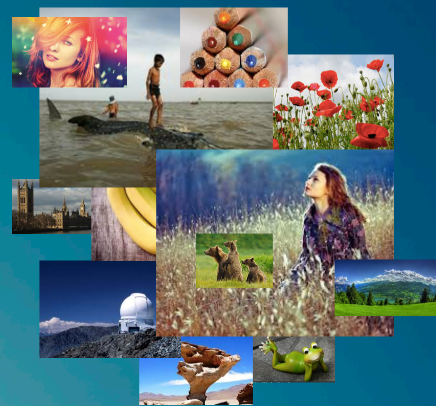
Big data applications are light a lightning rod, the June 2014 Facebook A/B testing debacle showed this

Tools of research ethics committees include an information session, plain language statement, opt-in option, withdraw at any time, no penalties for not participating, anonymity (except among peers) … and we do all these

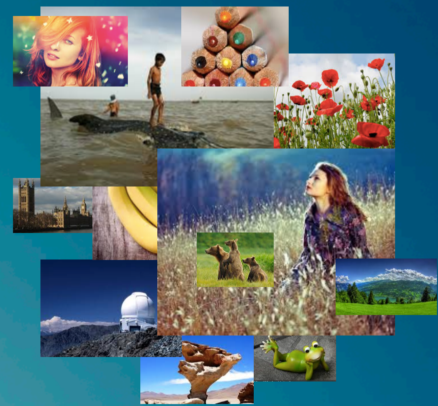
Our right to privacy goes back to Louis Brandeis' article from 1890 but *"the world has quickly become data driven, its time ethics caught up"* (Techcrunch, 2014). By highlighting, we draw attention, potentially spoiling what we're doing

How Does Machine Learning Work ?

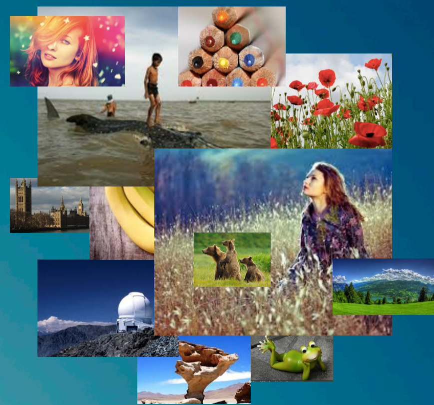
Suppose you want to build a classifier for ‘boat’, you need training data, + and – examples of boat images



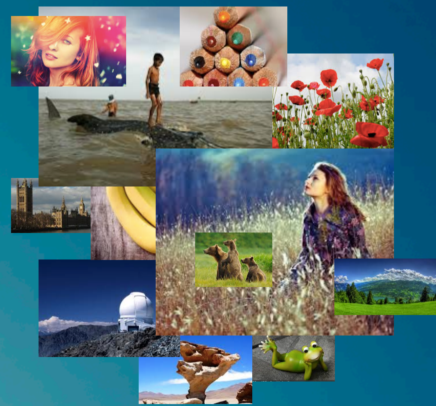
What makes a boat a boat, and a "not boat", not a boat ?
We extract low level features from each boat/non-boat
and try to "learn" the differences



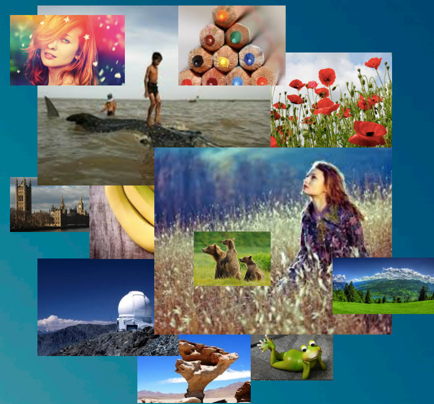
What kind of features … colours, textures, shapes, lines, across all the picture or in regions, calculated at pixel level

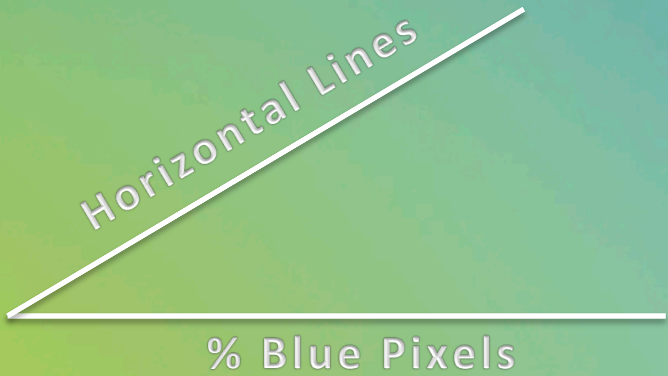


In practice, there are hundreds of such features, but lets look at just two



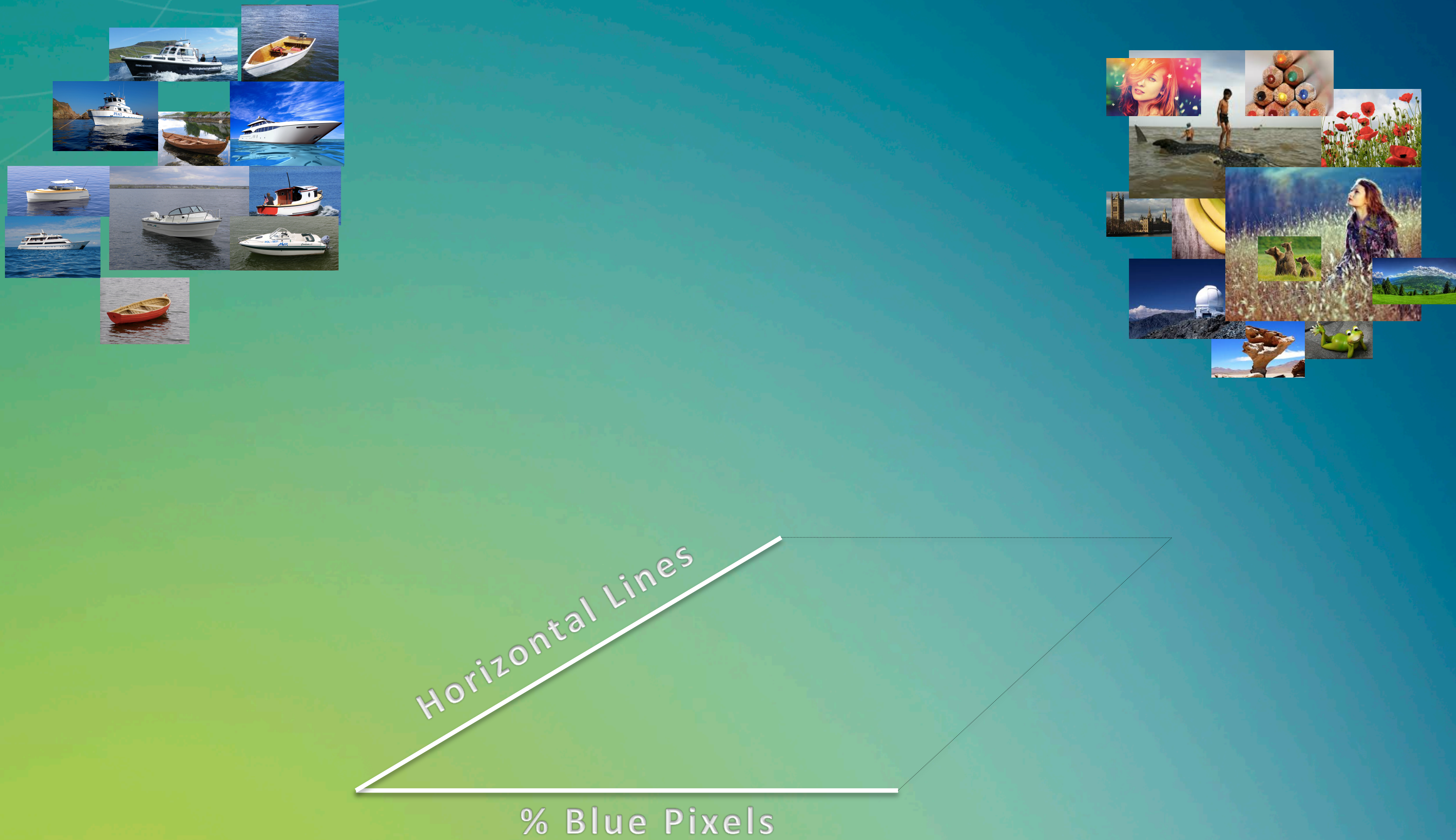
In practice, there are hundreds of such features, but lets look at just two



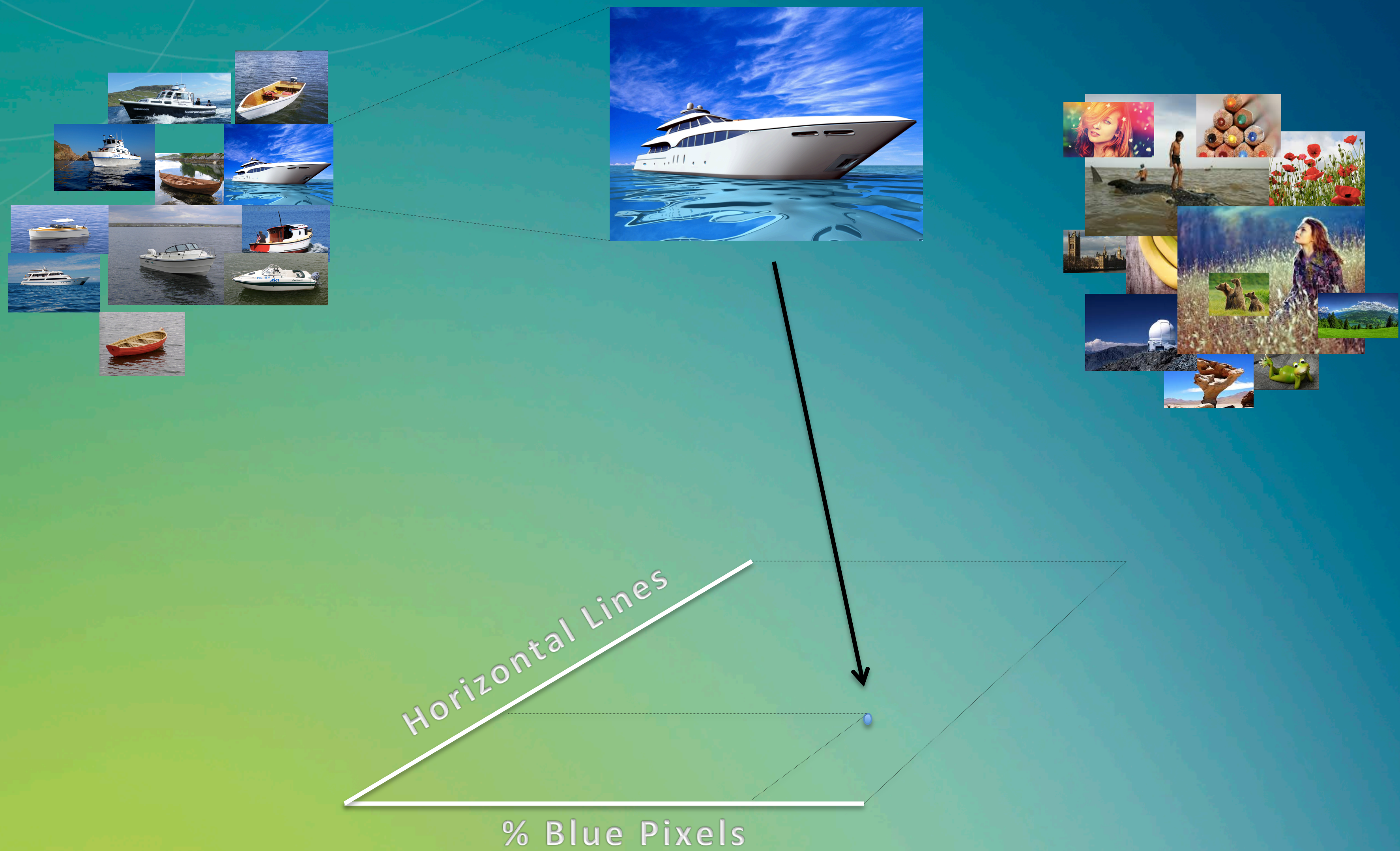


In practice, there are hundreds of such features, but lets look at just two

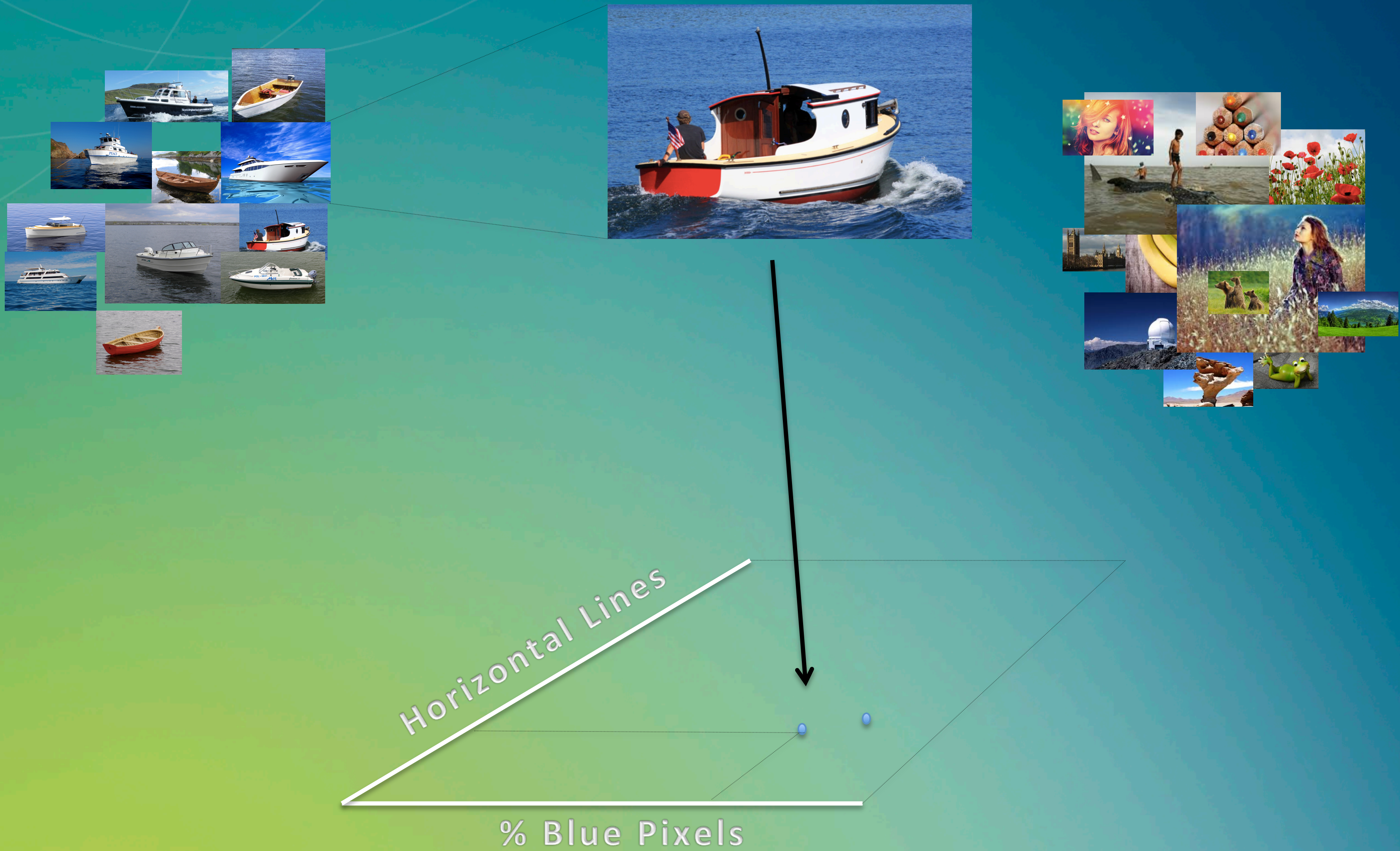
We can then take each image and "plot" it in this 2D "space"



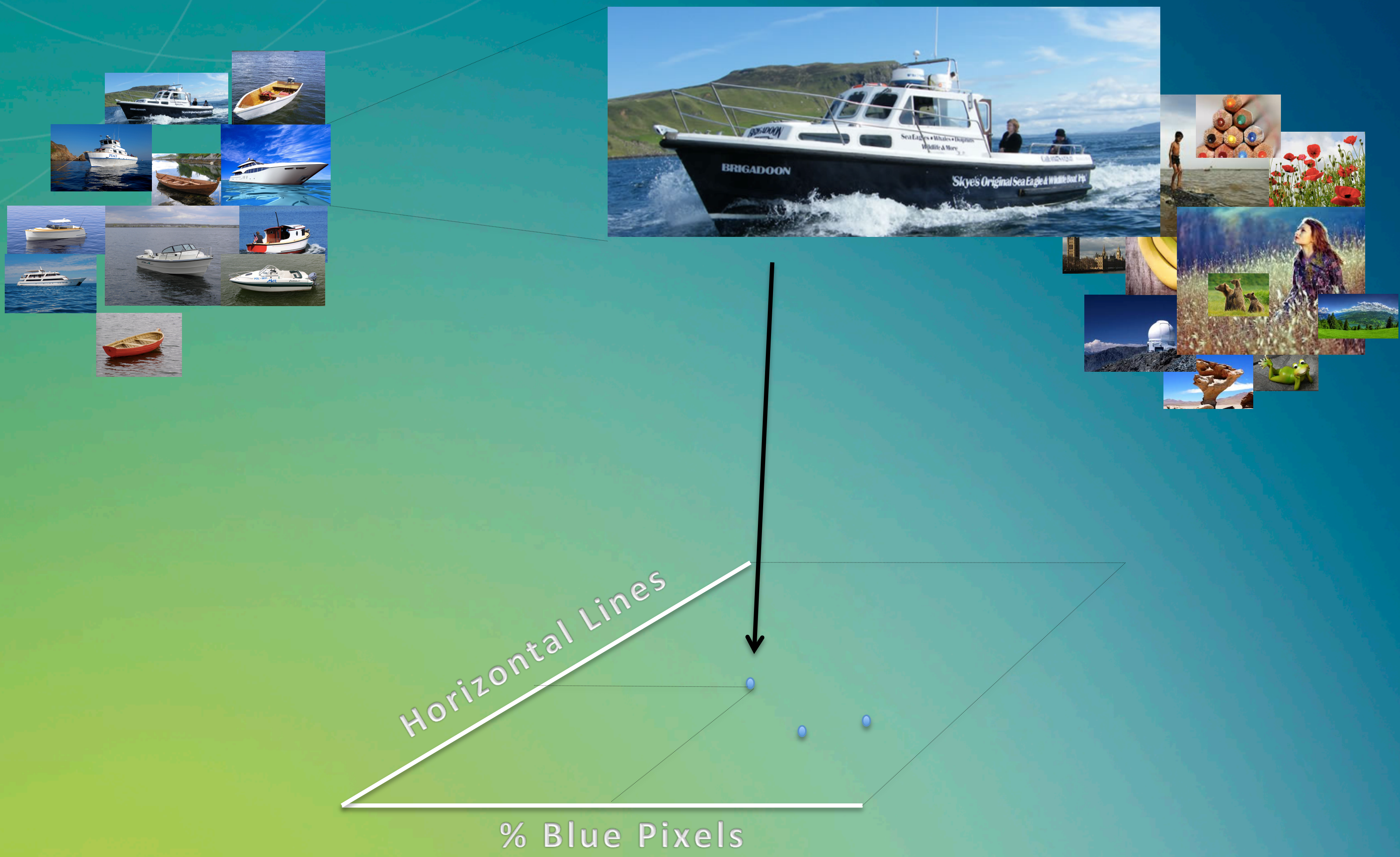
For boats …



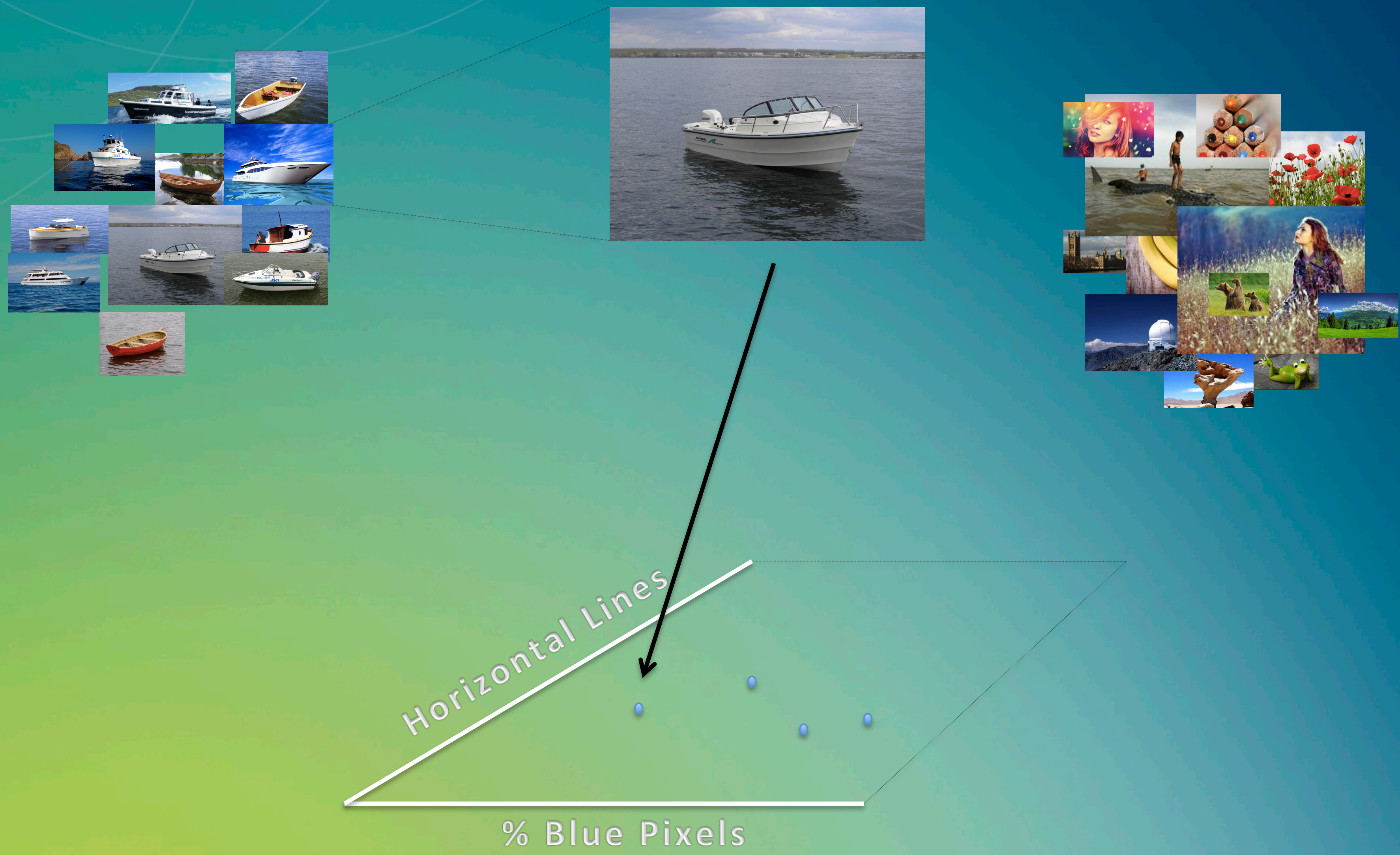
For boats …



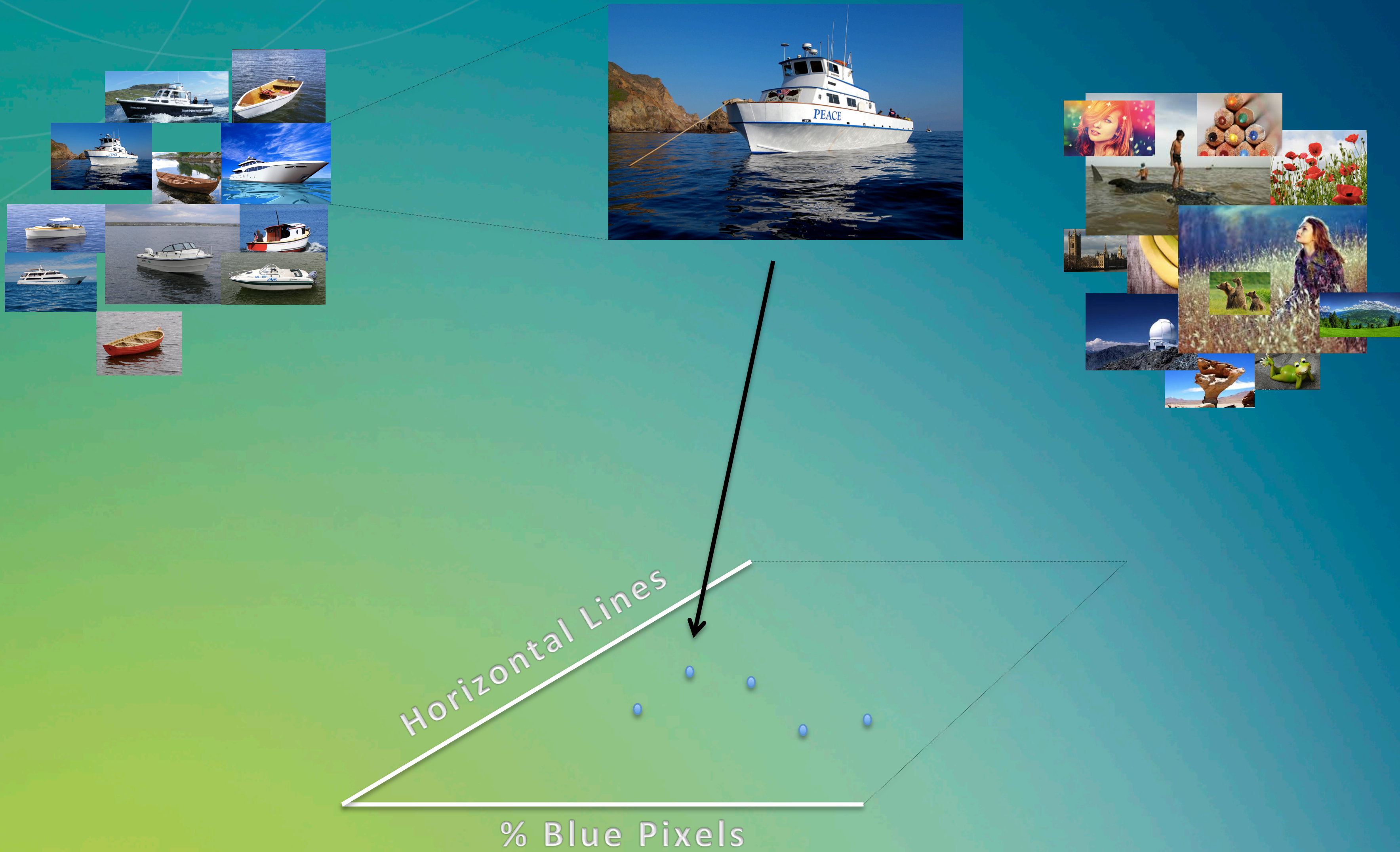
For boats …



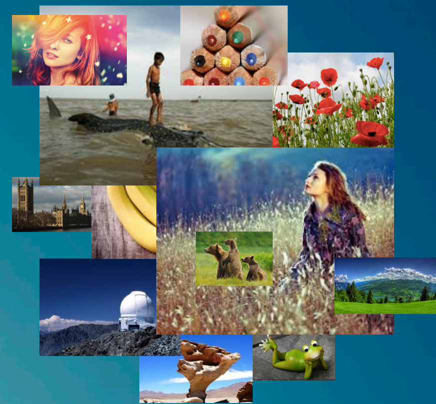
For boats …

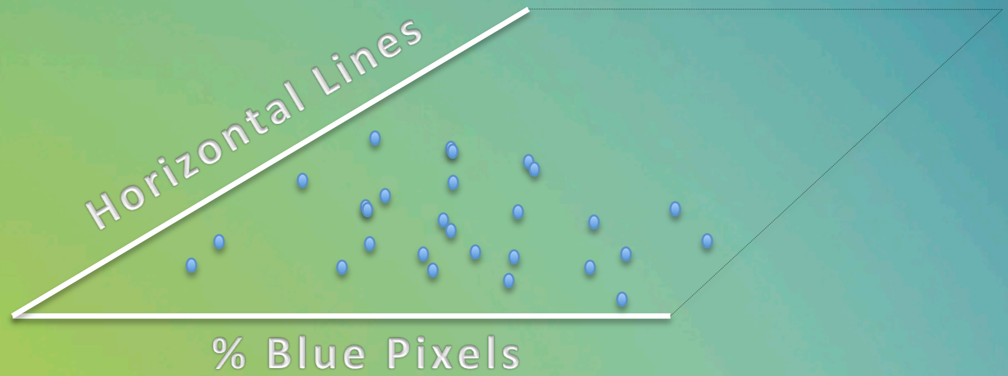


For boats …

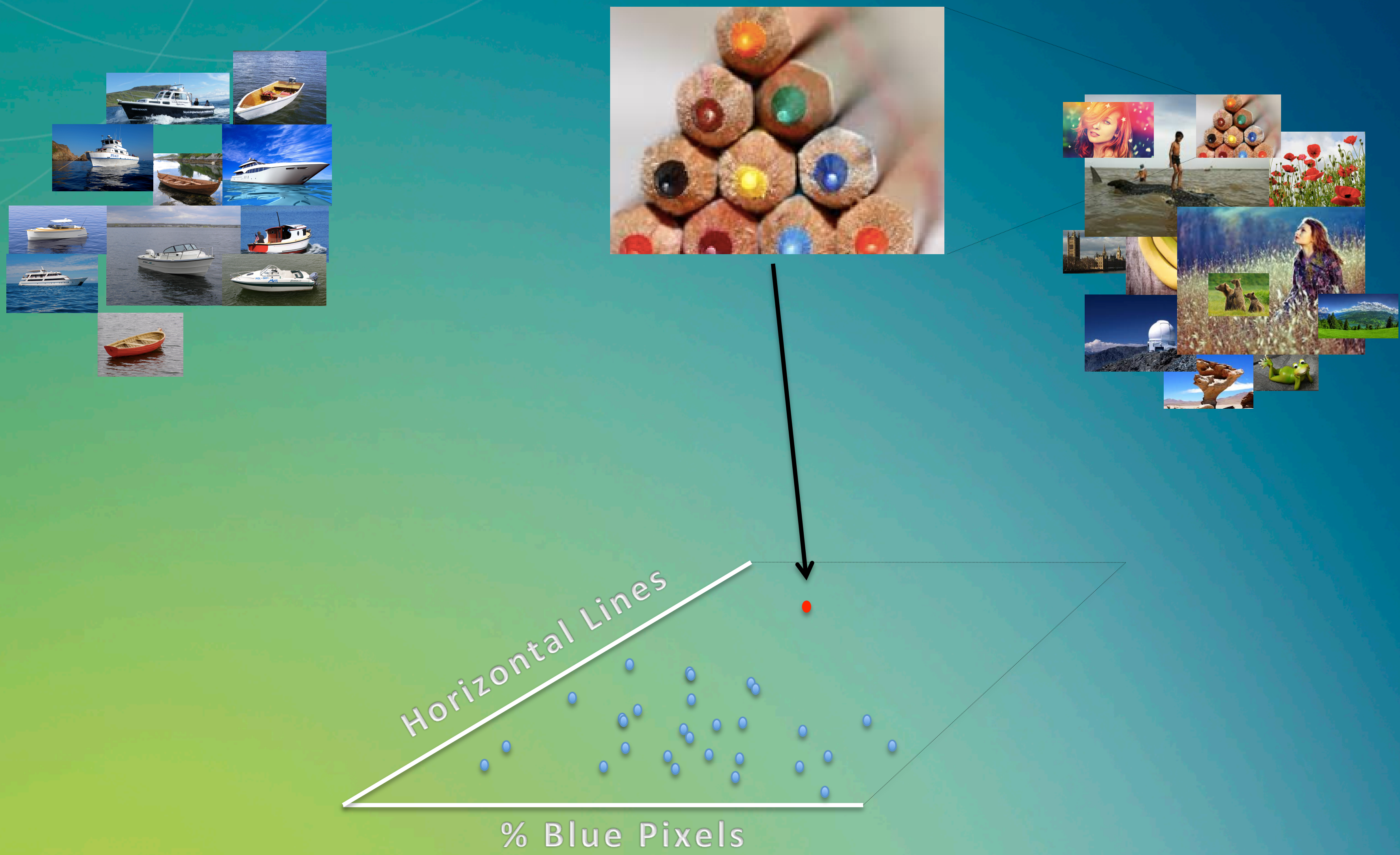


Until there are many of them

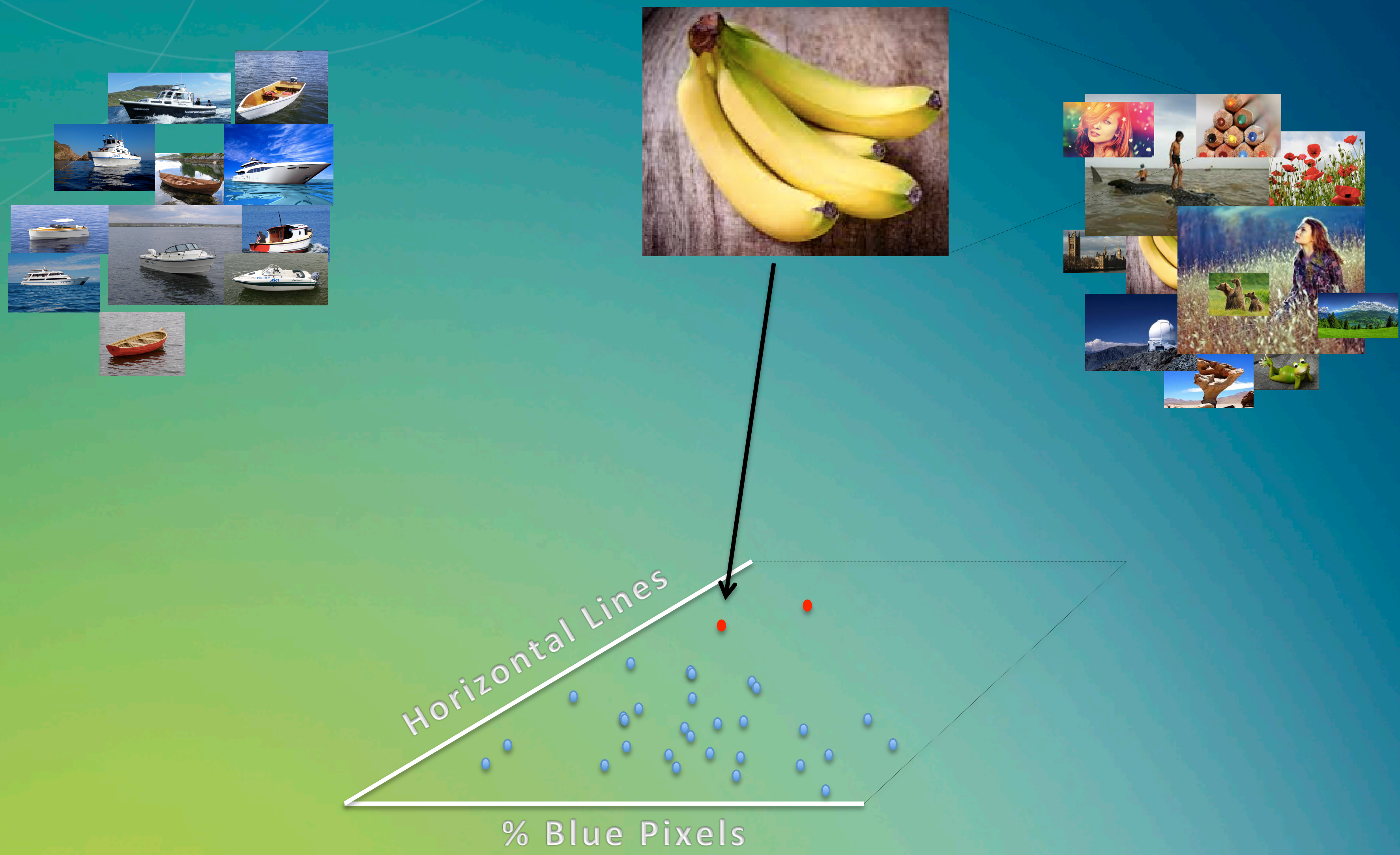


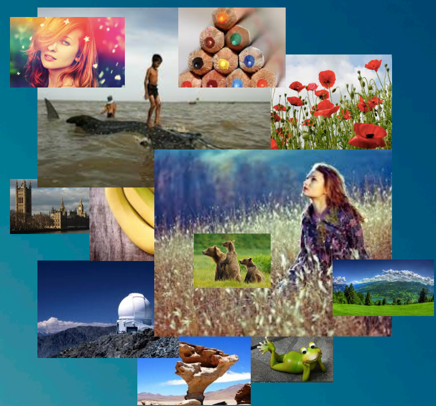


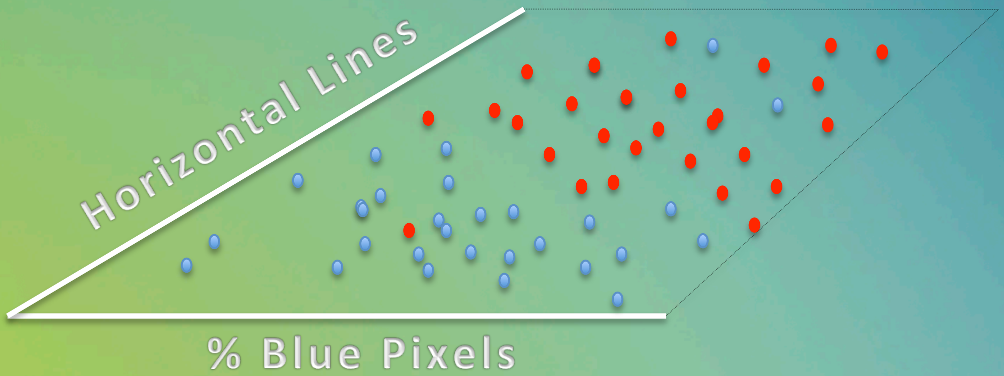
And then for non-boats …



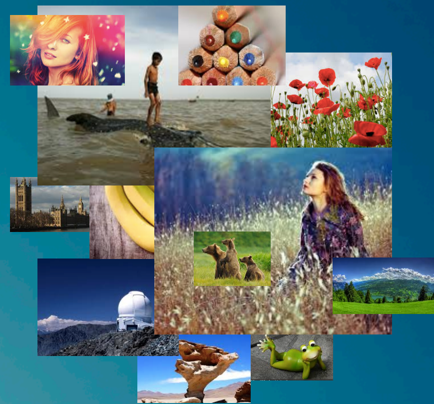
And then for non-boats …

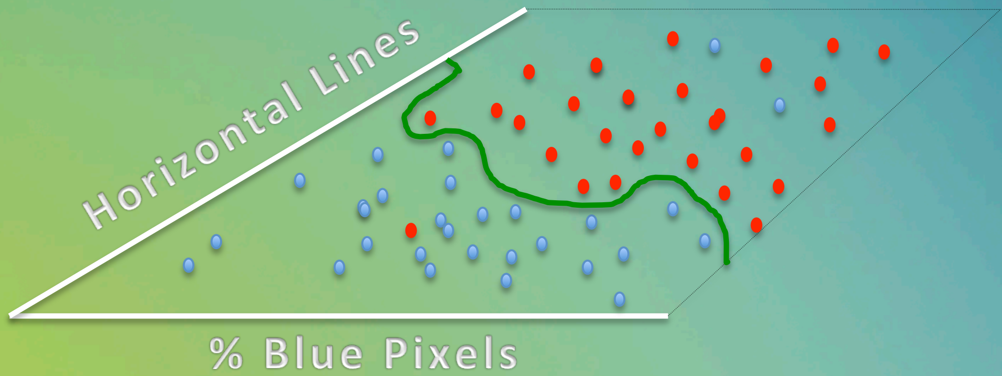




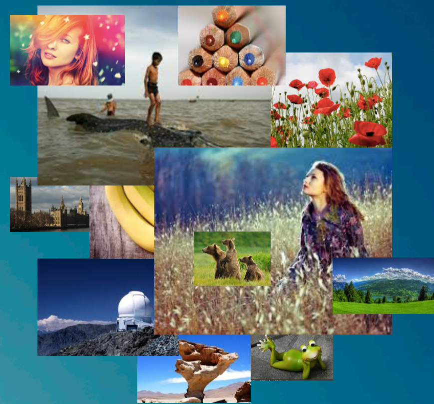

Horizontal Lines
% Blue Pixels

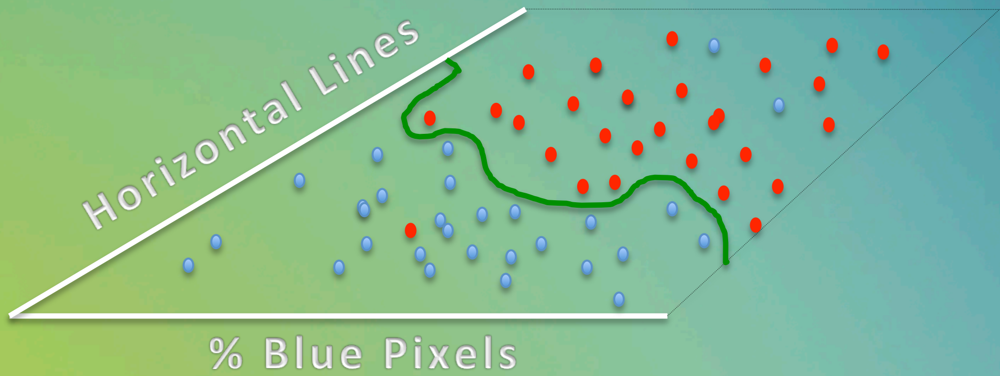
We then “learn” the differences between a boat and a non-boat, in terms of %Blue pixels/Horizontal Lines



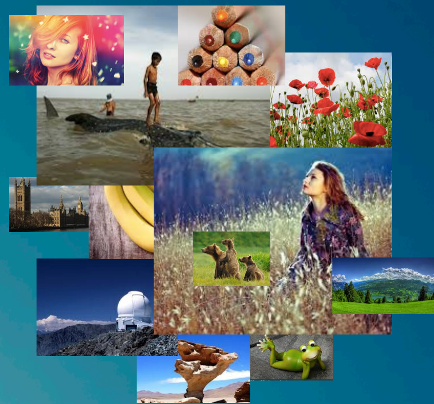


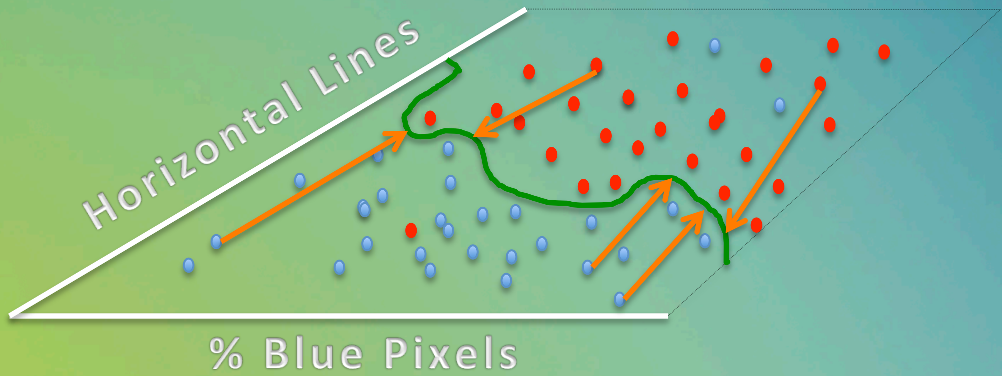
There are outliers, but mostly its correct



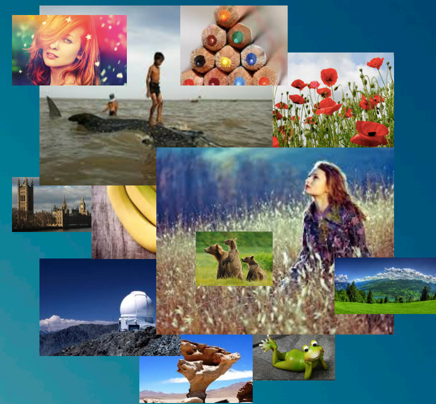


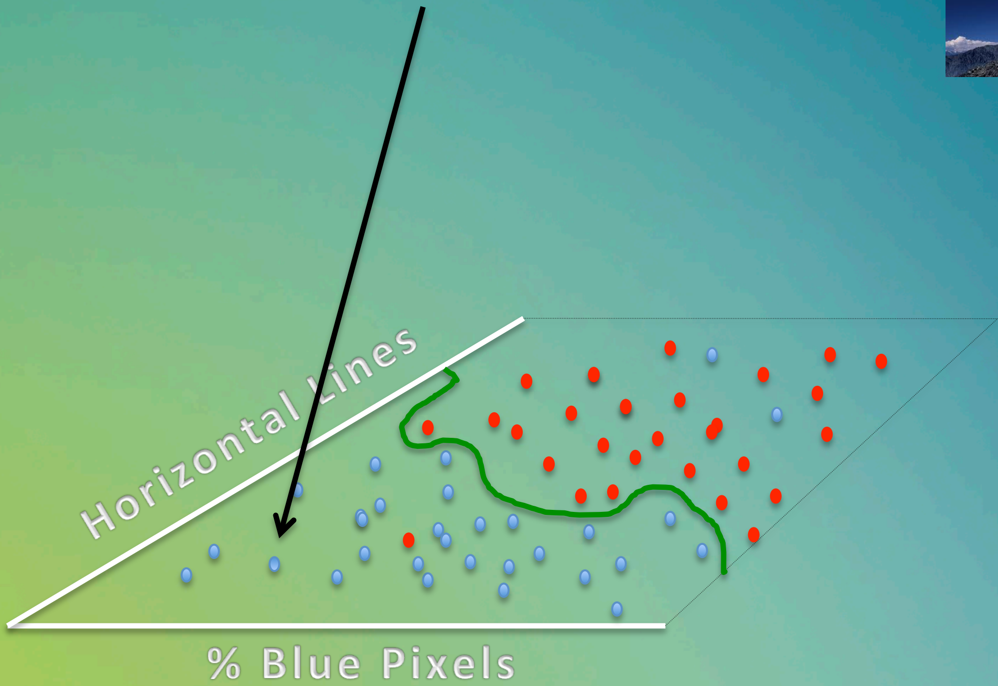
The “distance” from this “hyperplane” is a measure of confidence in boat/non-boat



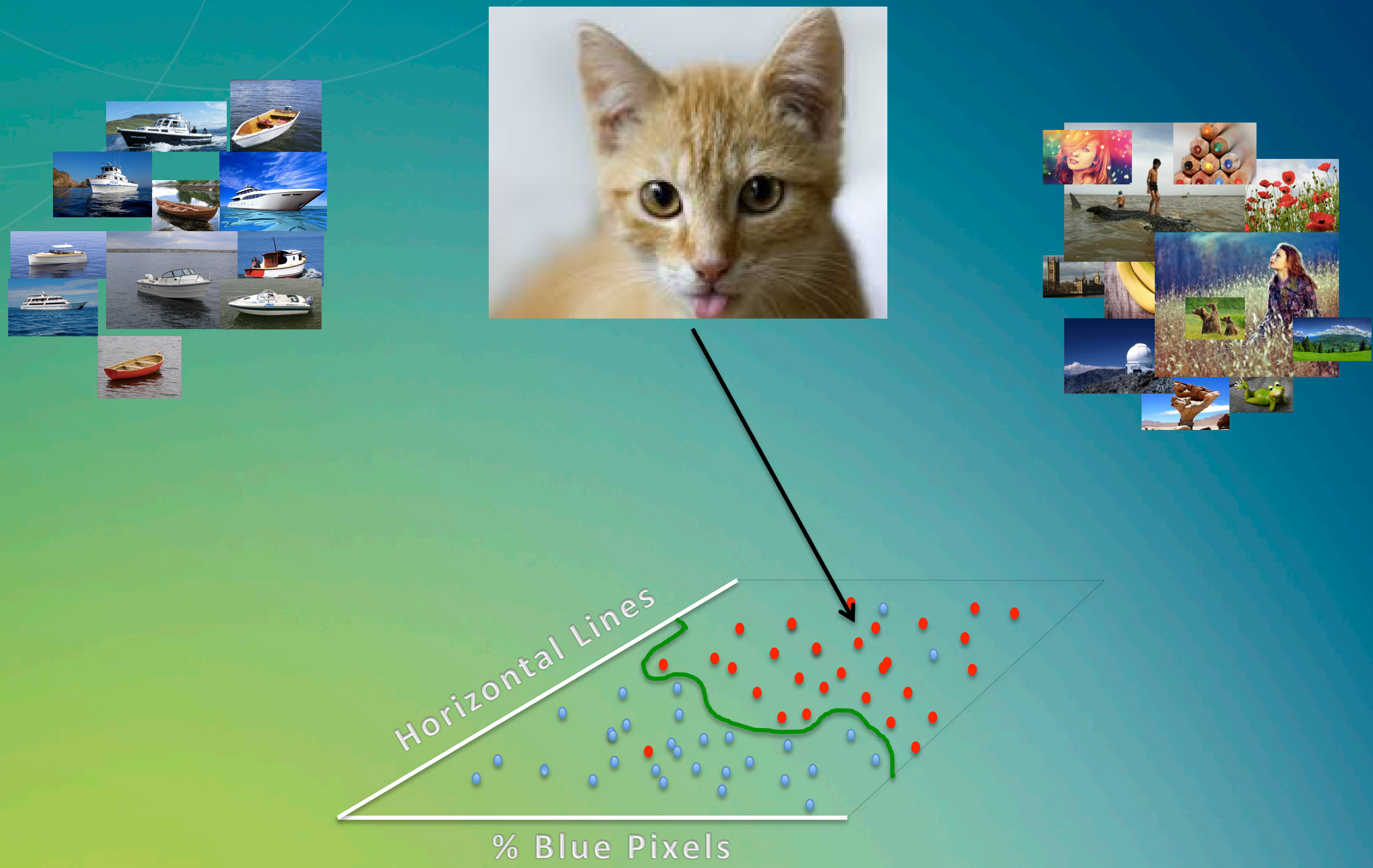


Then take new (untrained) images …

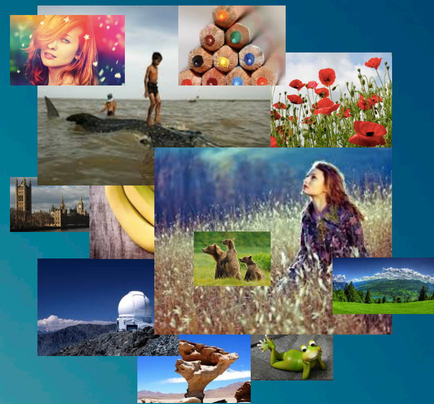


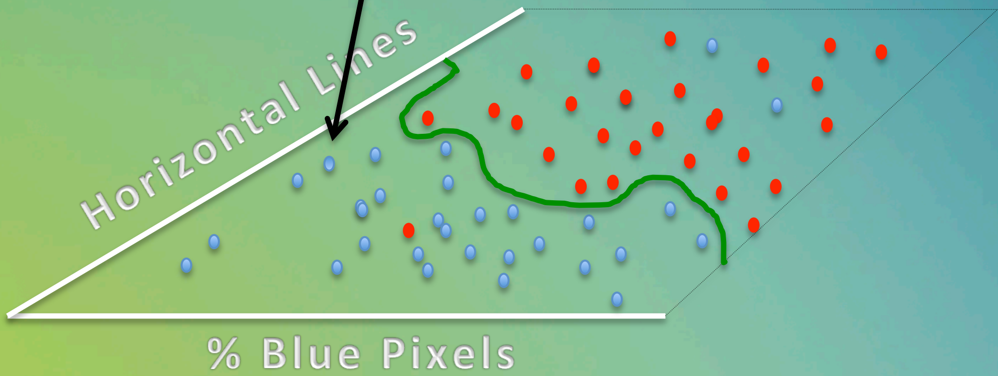


Then take new (untrained) images …

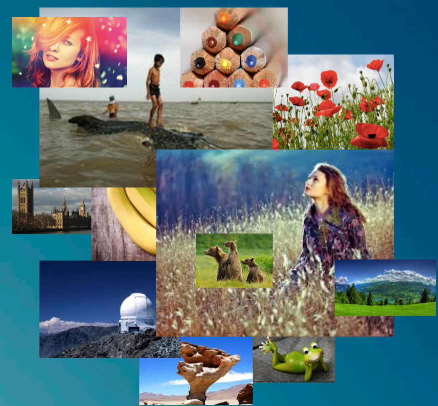


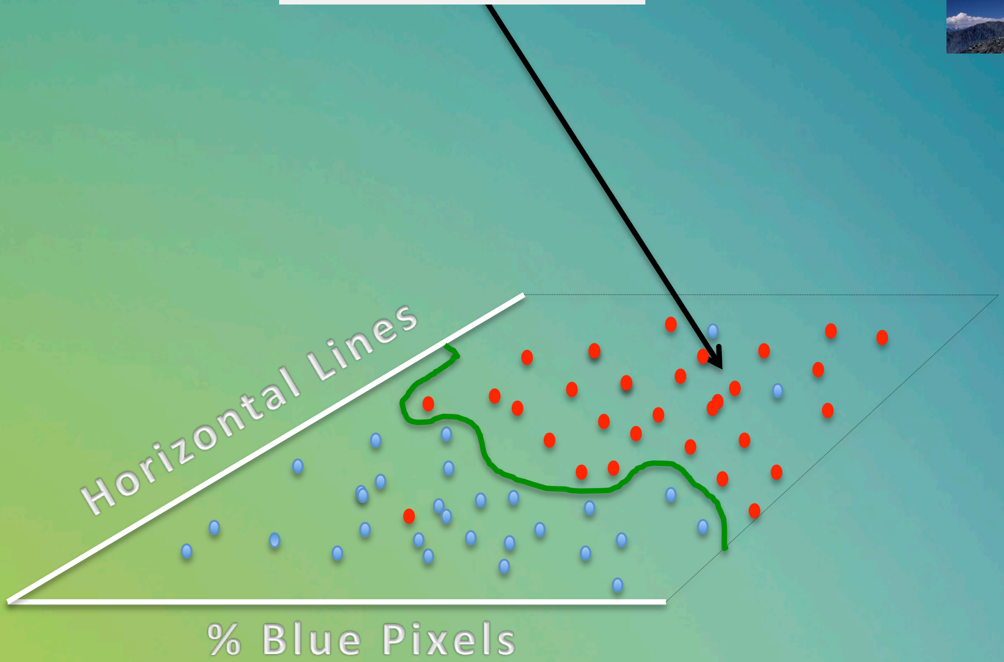
Then take new (untrained) images …





Then take new (untrained) images …





So that's machine learning … building a classifier

Training set (positive and negative examples)
Balanced numbers of each
Features for each
Lots of computing to extract features and learn the classifier

Very fast to run new examples through the classifier

Which learning functions, which kernal, which features … all that is the black art !

So many questions …

We set out to apply machine learning to predict pass-fail on modules, using demographics and past behaviour (from previous years)

Which modules ?

Which students ?

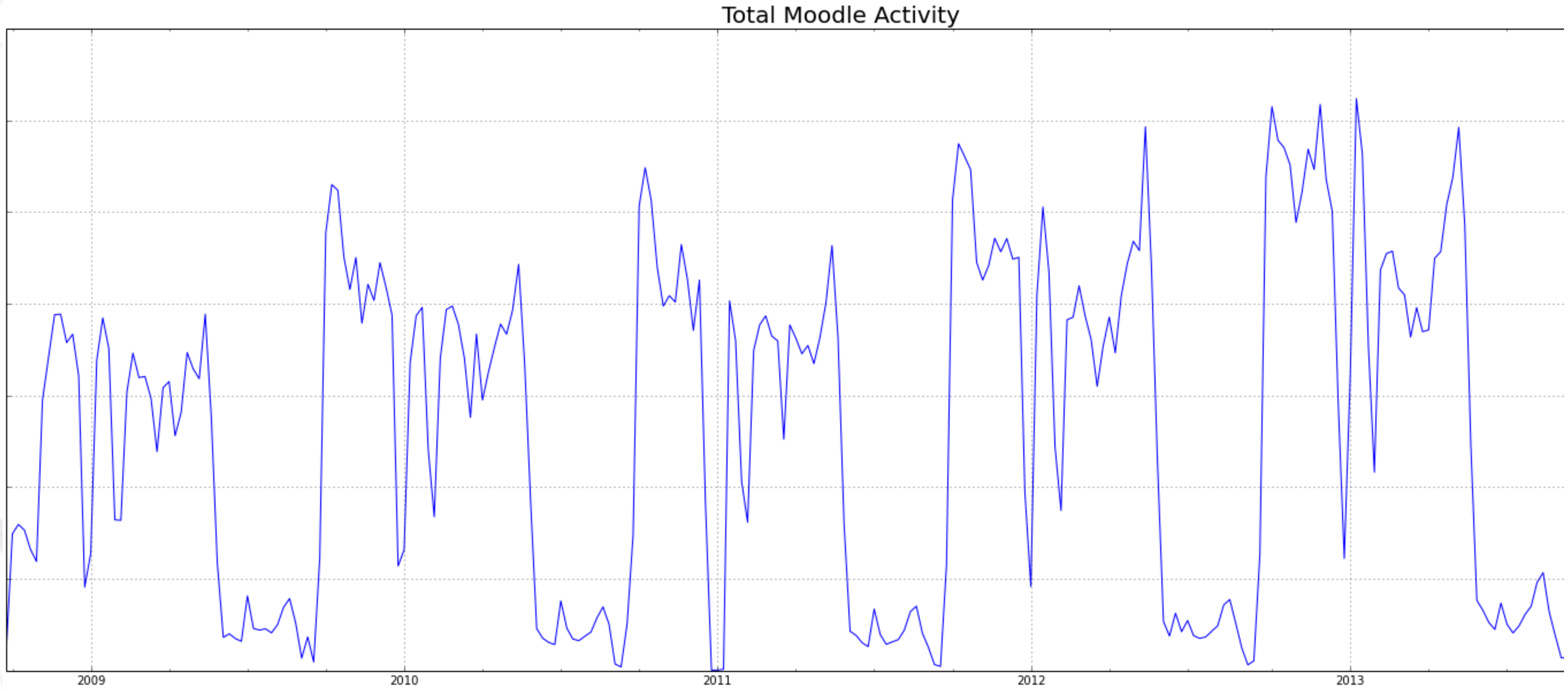
When to calculate prediction ?

How to feed back to students, to lecturers ?

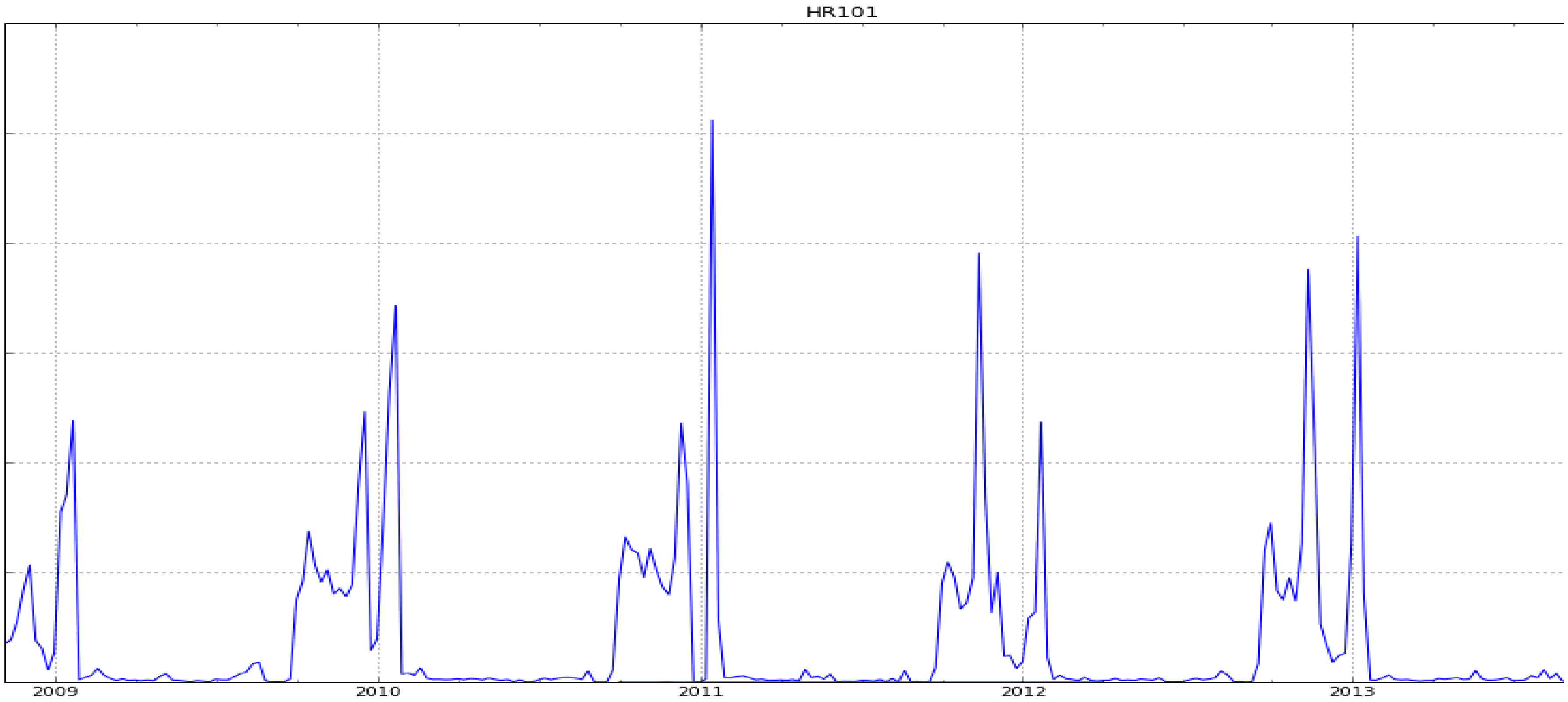
Modules which work well …

- Have periodicity (repeatability) in Moodle access
- Confidence of predictor increases over time
- Don't have high pass rates (< 95%)
- Have large number of students, early-stage

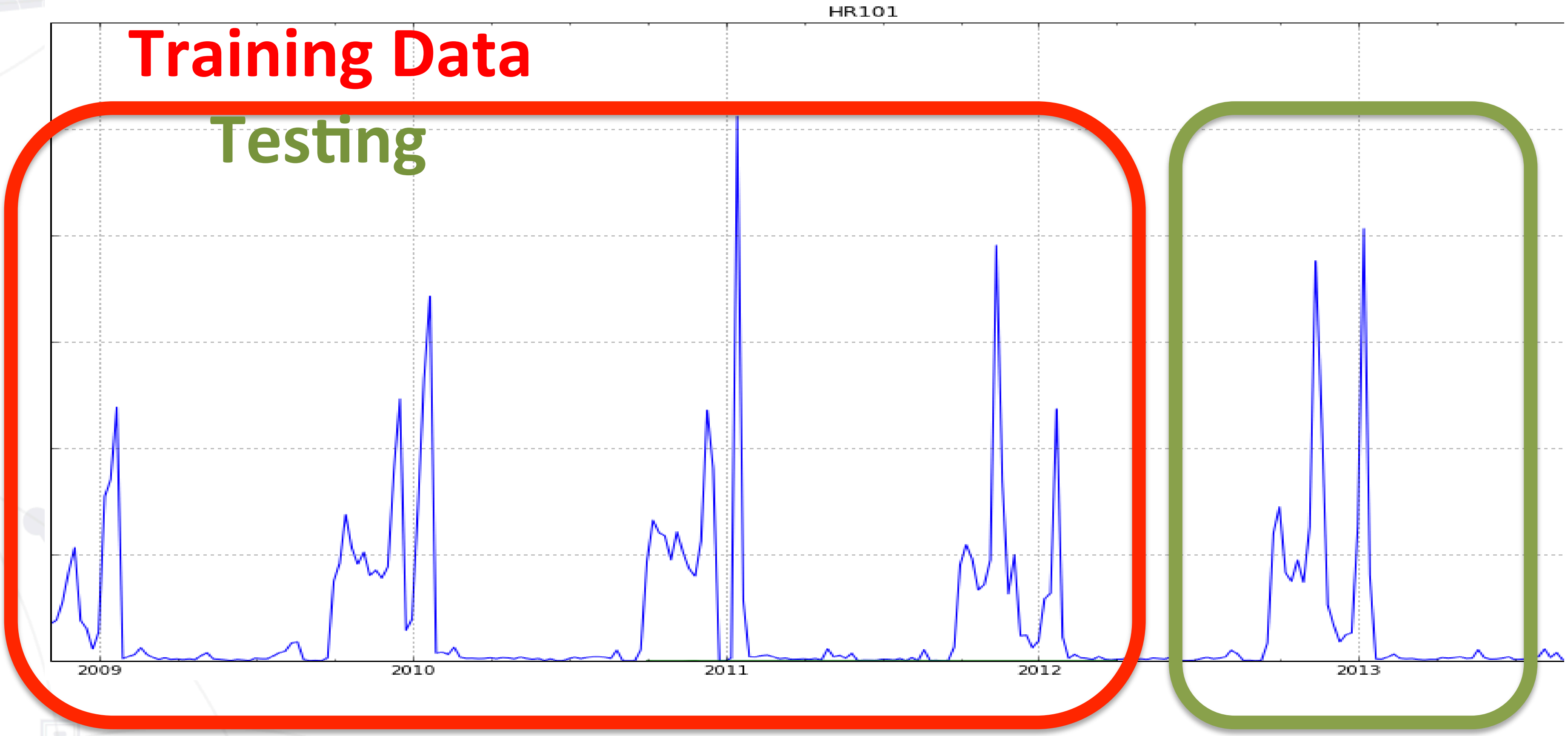
Total DCU Moodle Activity – notice the periodicity



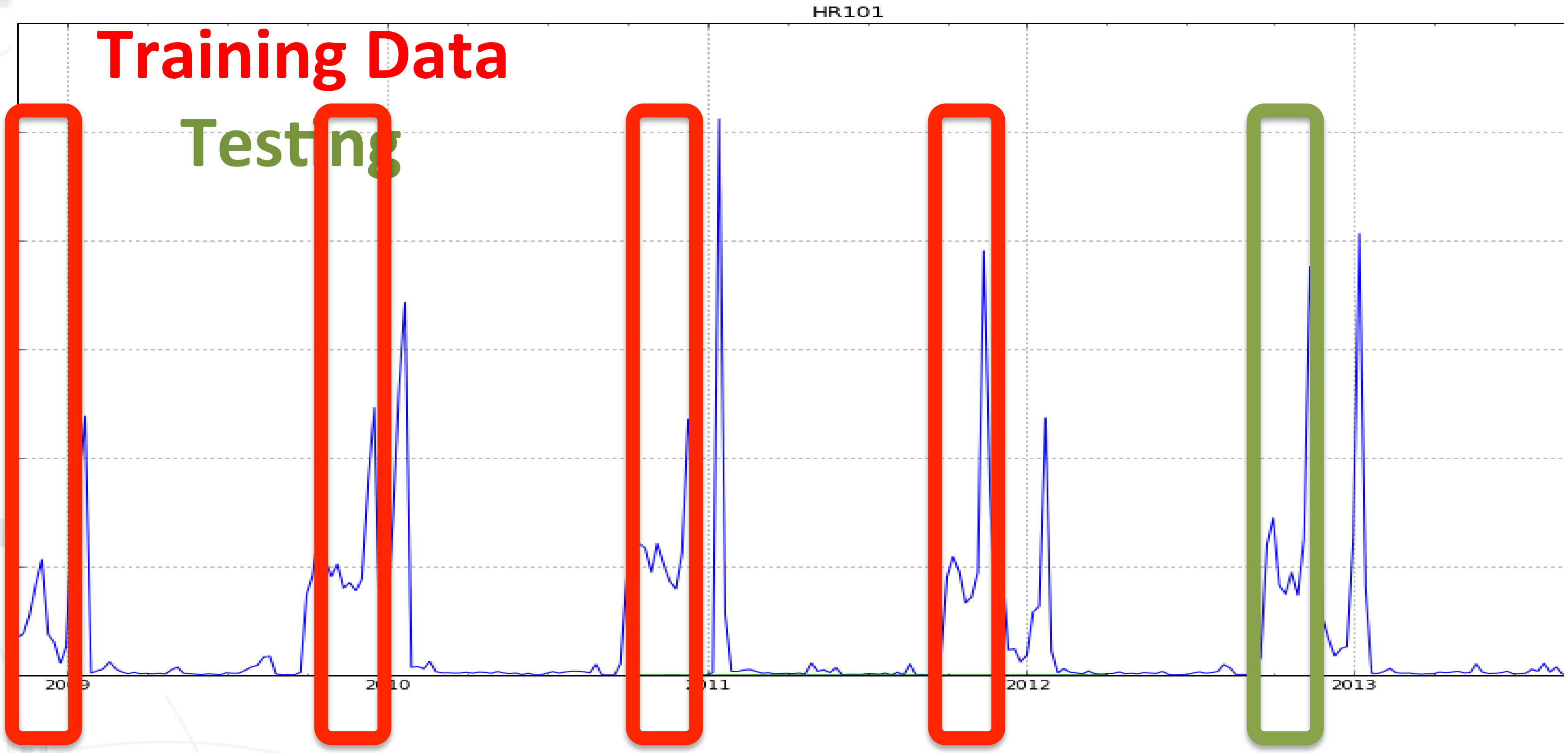
One example module, HR101 – ideal !



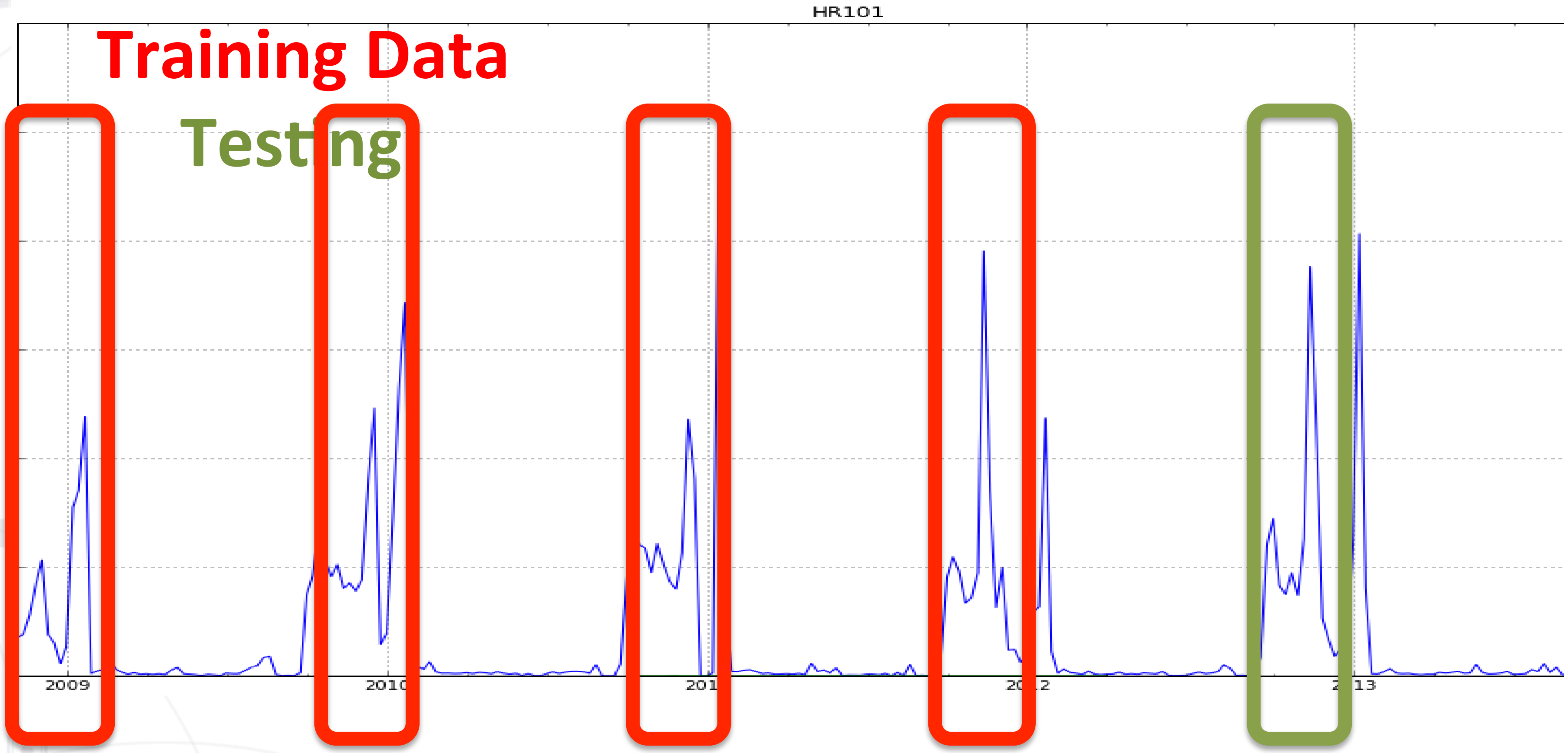
Building classifiers for each week/each module



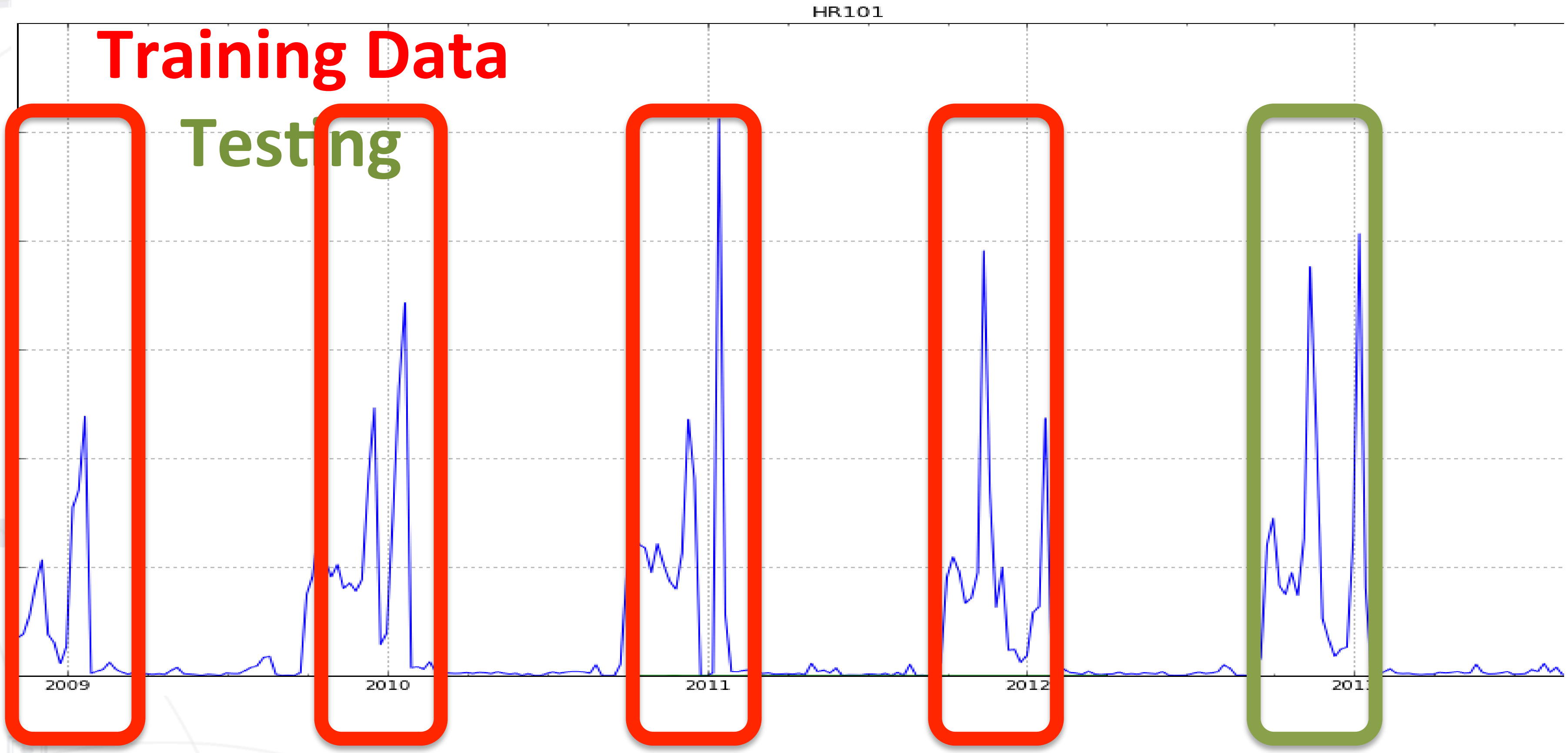
Week 3



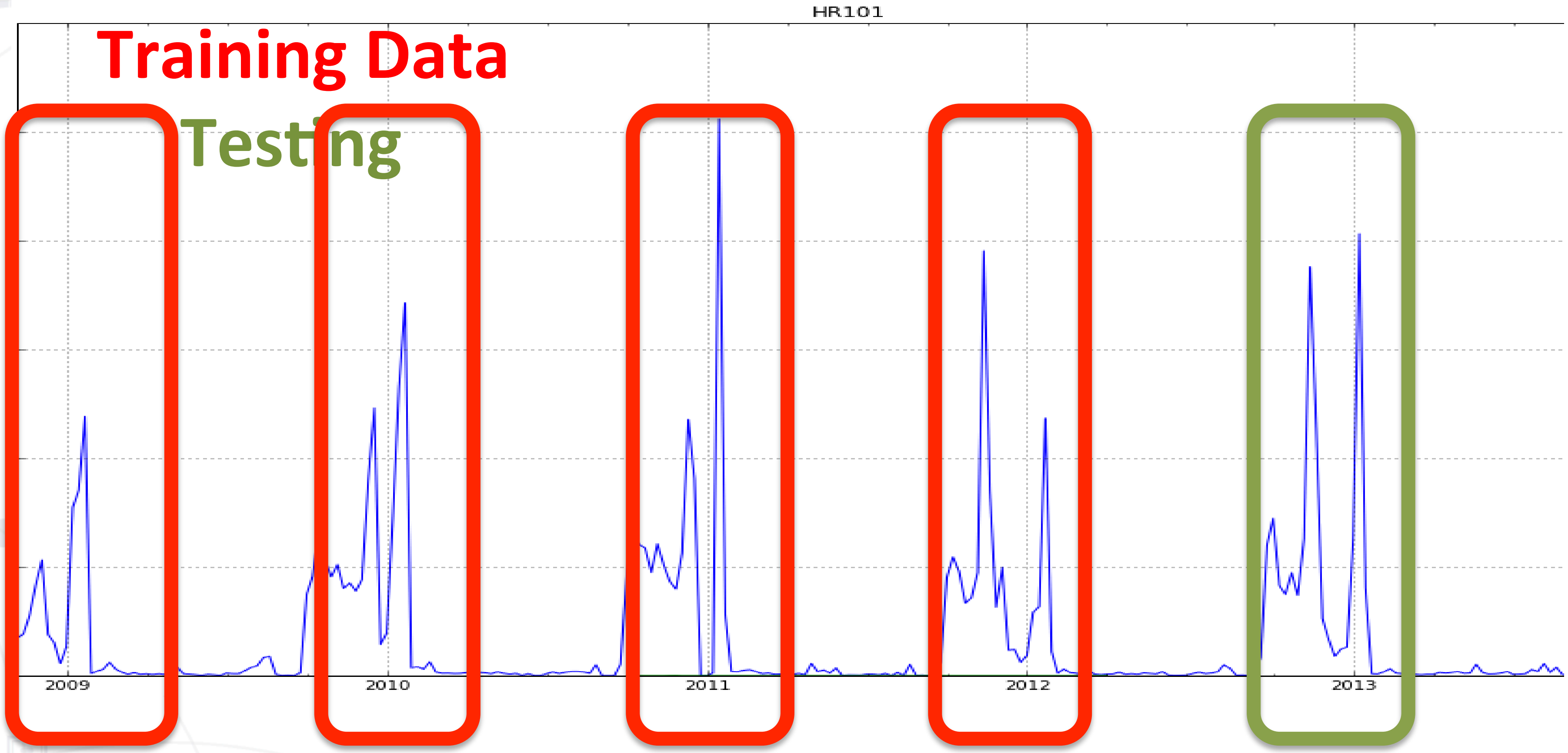
Week 4



Week 5



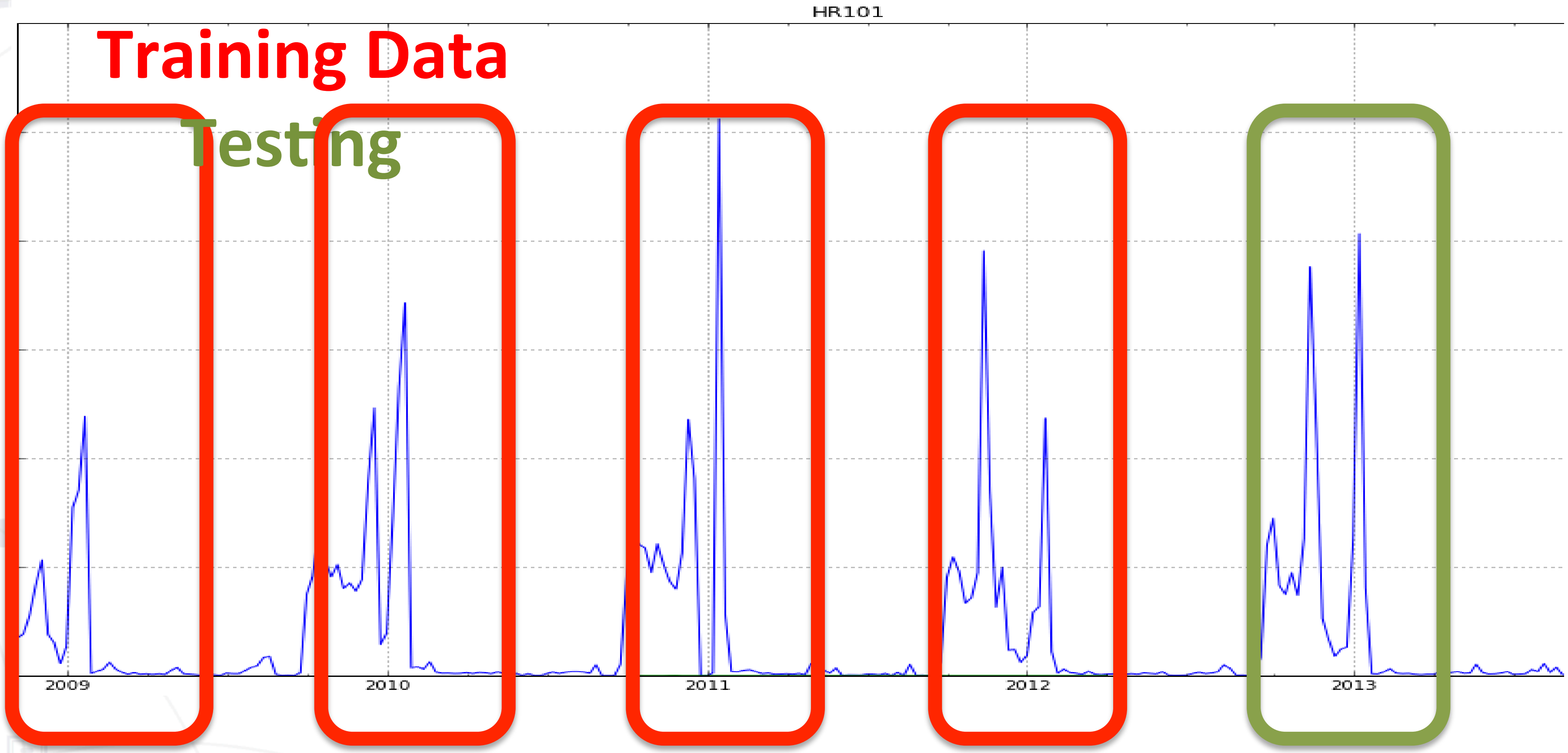
Week 6



Centre for
Data Analytics

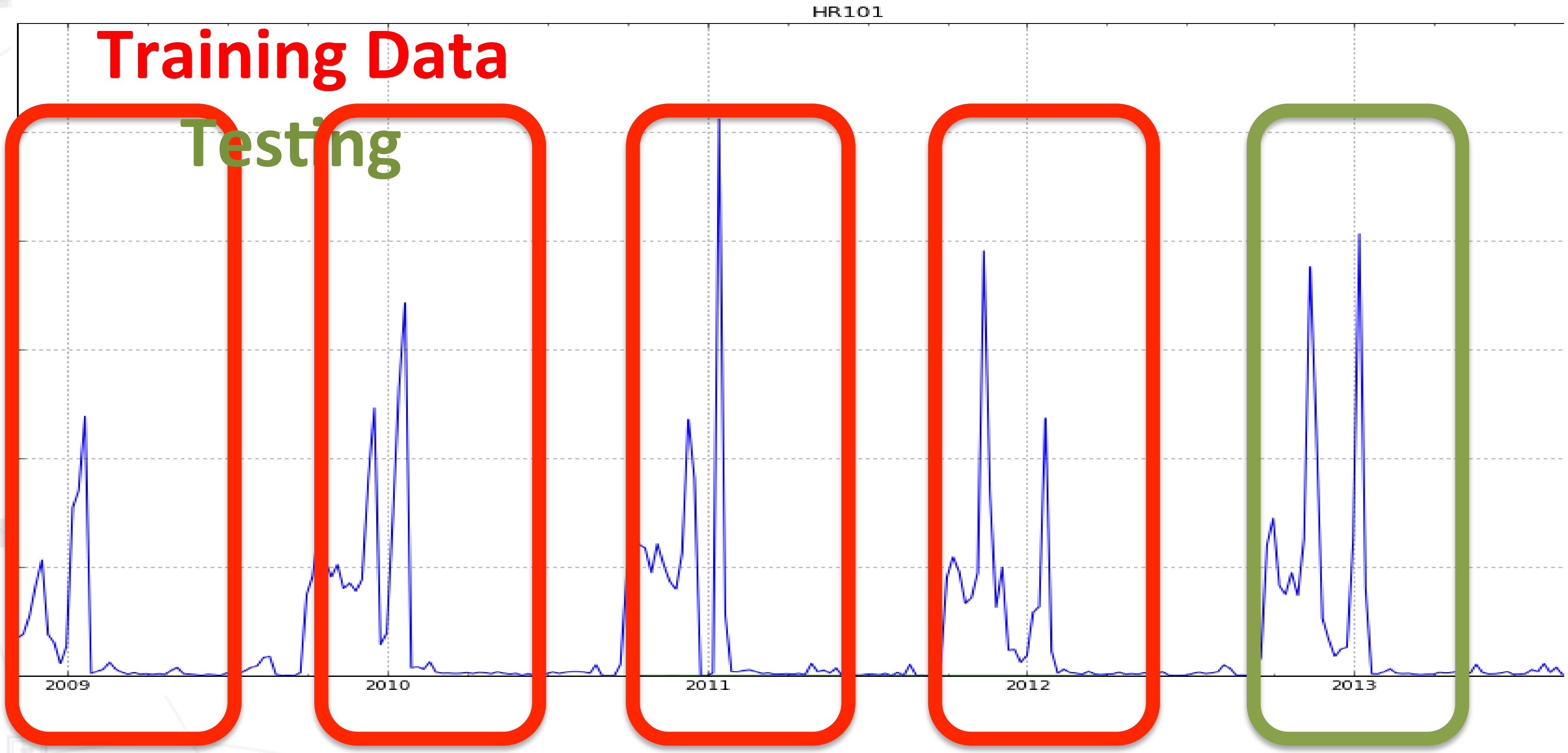
Insight

Week 7

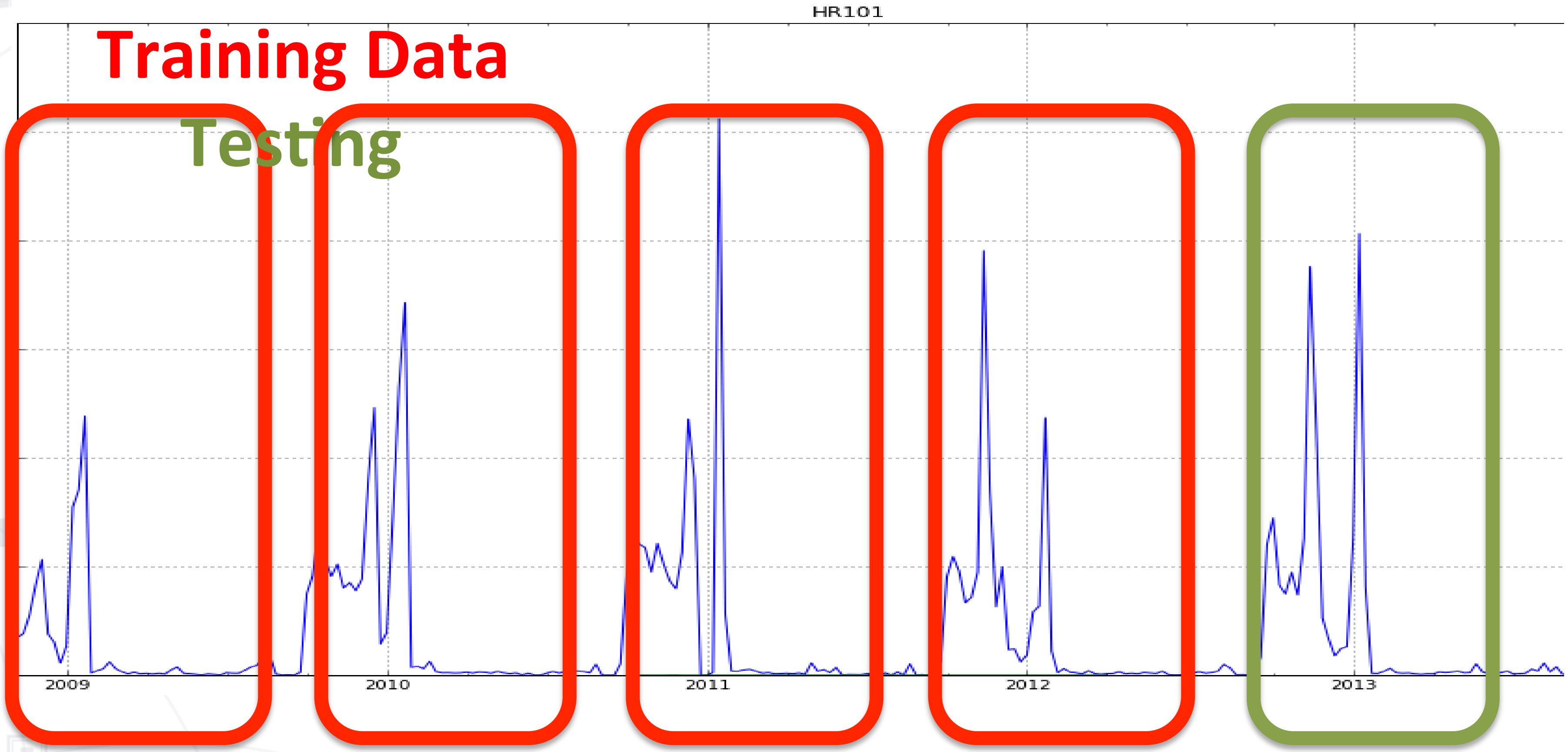


creating a data-driven society

Week 8



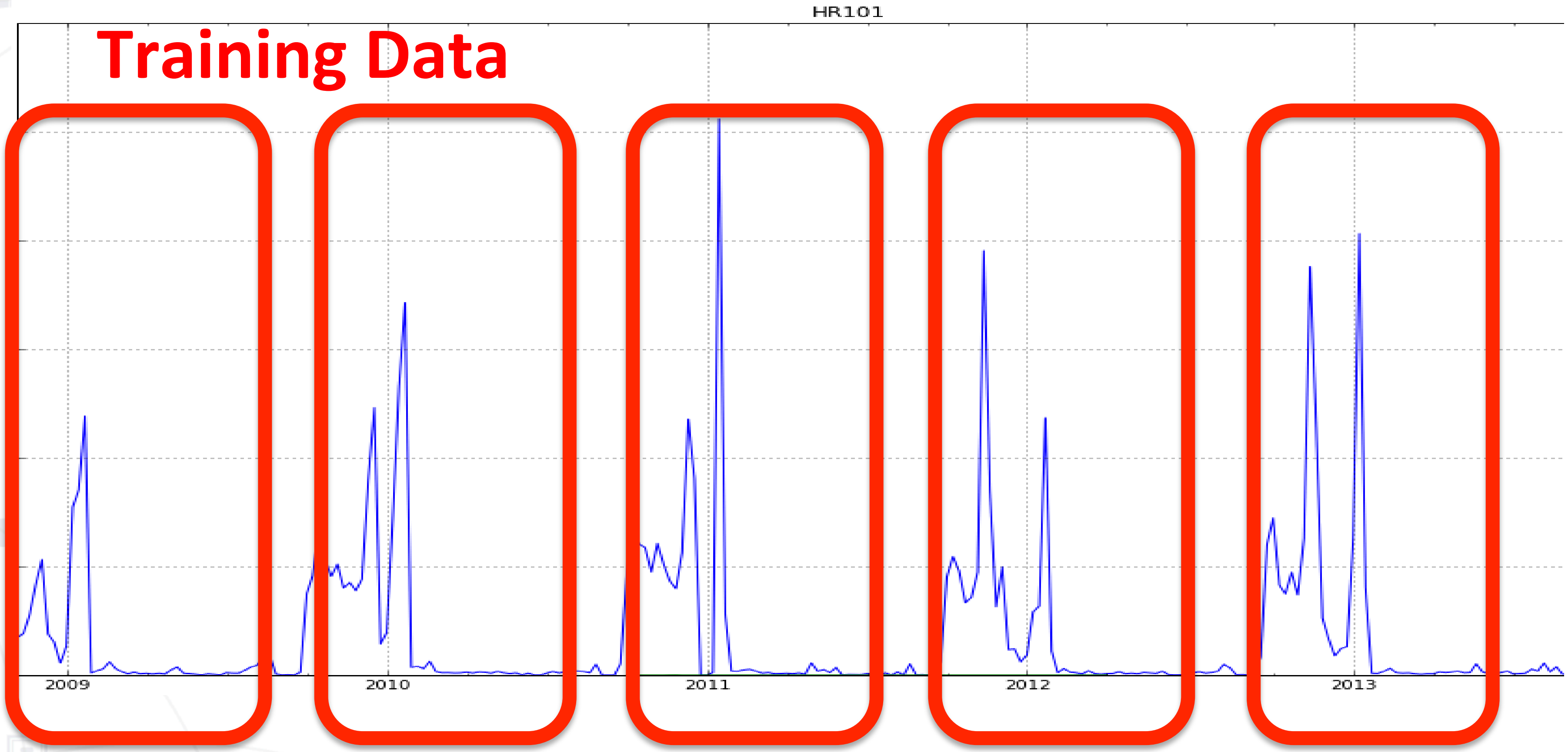
Week 9 … all the time for HR101



So for each module, we have a "tipping point" a week offset at which the predictor will work to an acceptable accuracy, and this varies across modules

When we know this, we then re-train on ALL data we have

For 2014 intake, we then re-trained each combination of module/week on all data – we're now ready for 2014 !

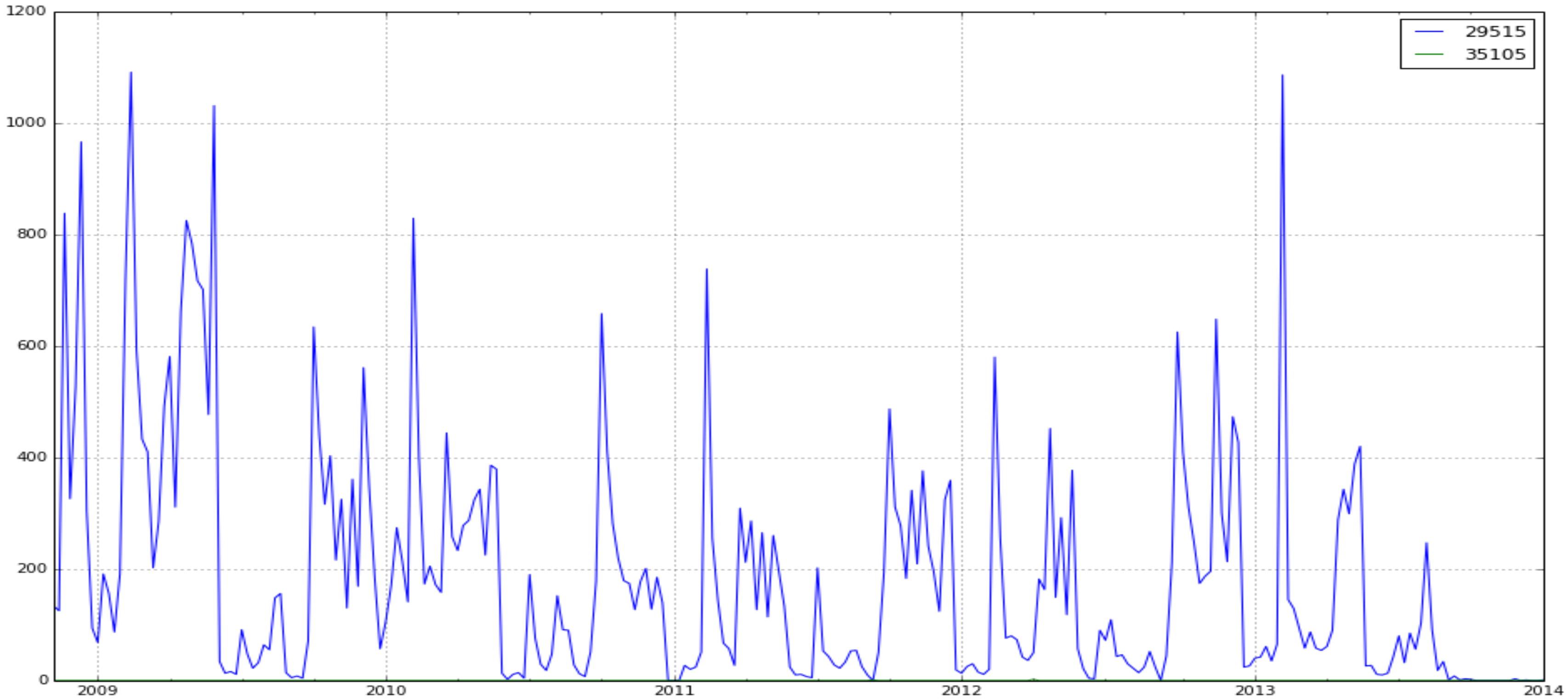


Model confidence

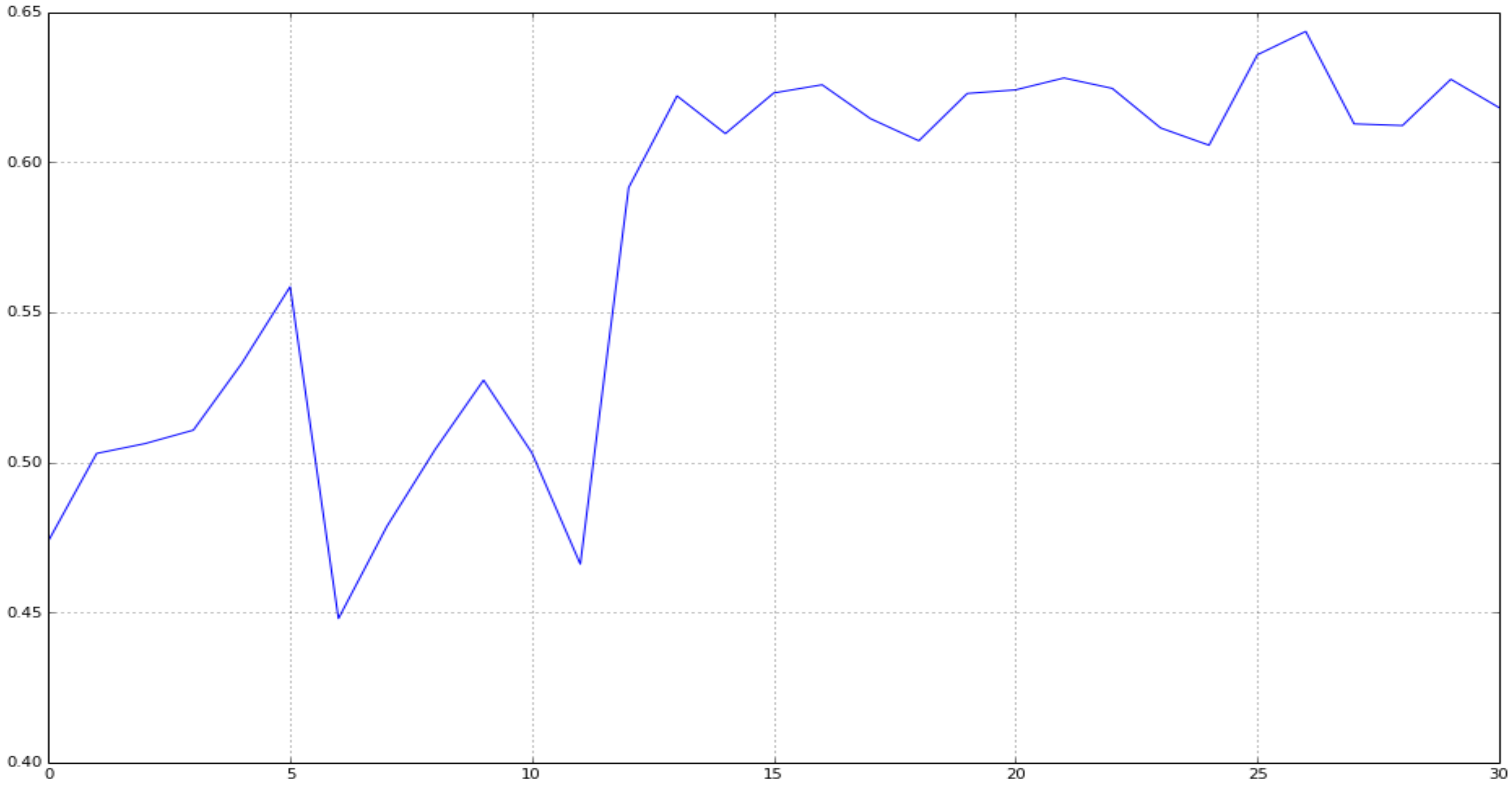
- Y axis is confidence in AUC ROC (not probability)
- X axis is time in weeks
- 0.5 or below is a poor result
- Most Modules start at 0.5 when we don't have much information
- 0.6 is acceptable, 0.7 is really good (for this task)
- The model should increase in confidence over time

LG116: Introduction to Politics

Students / year = ~110
Pass rate = 0.78



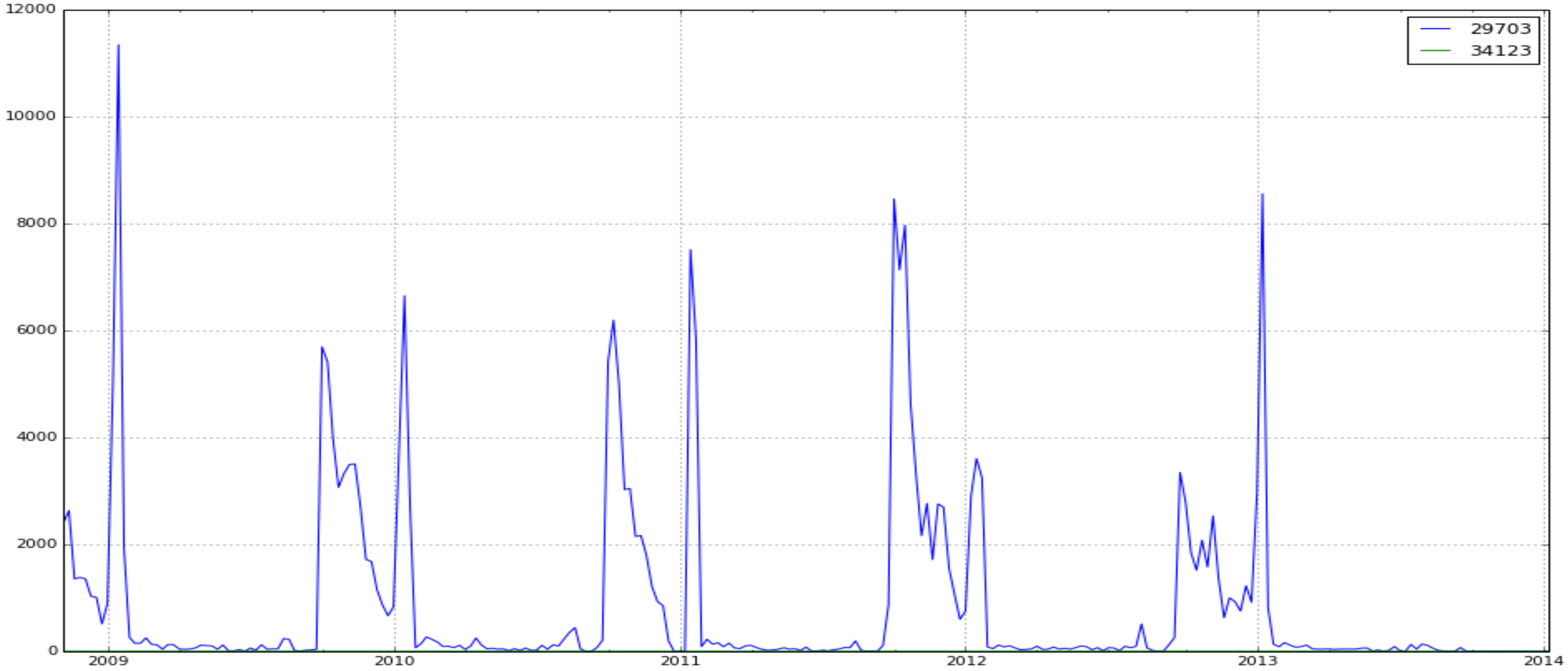
LG116 – Predictor confidence (ROC AUC)



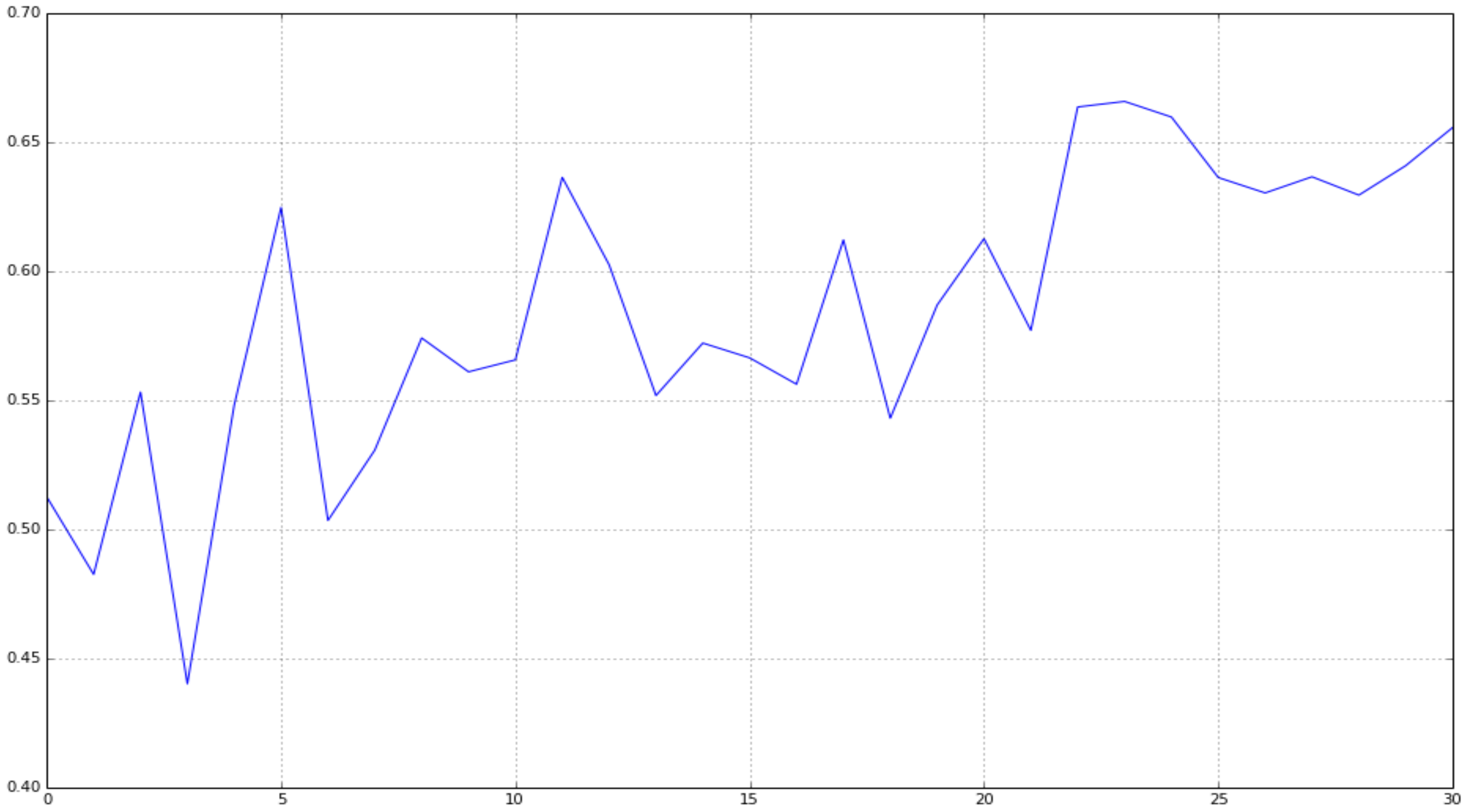
BE101: Intro to Cell Biology

Results / year = ~300

Pass rate = 0.86

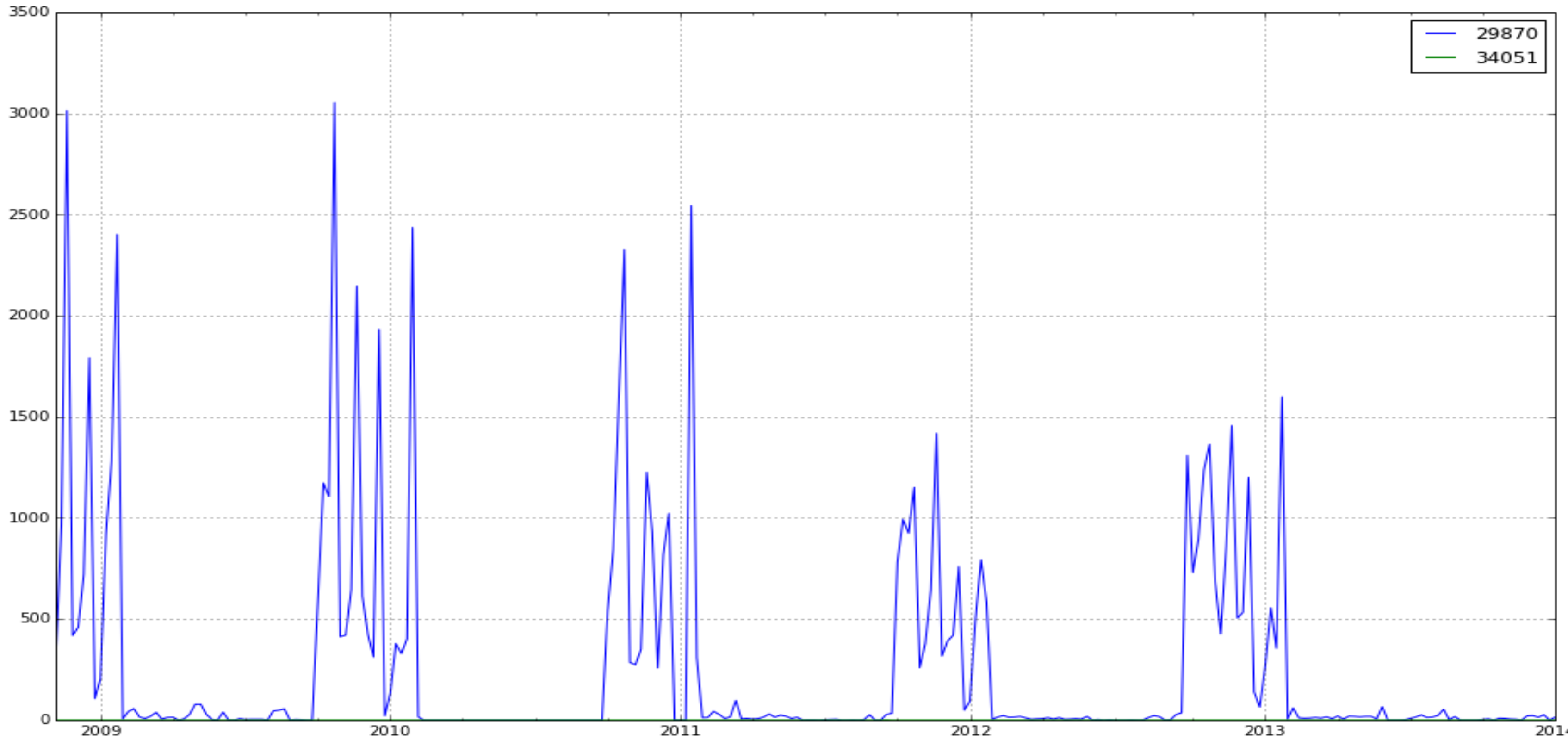


BE101

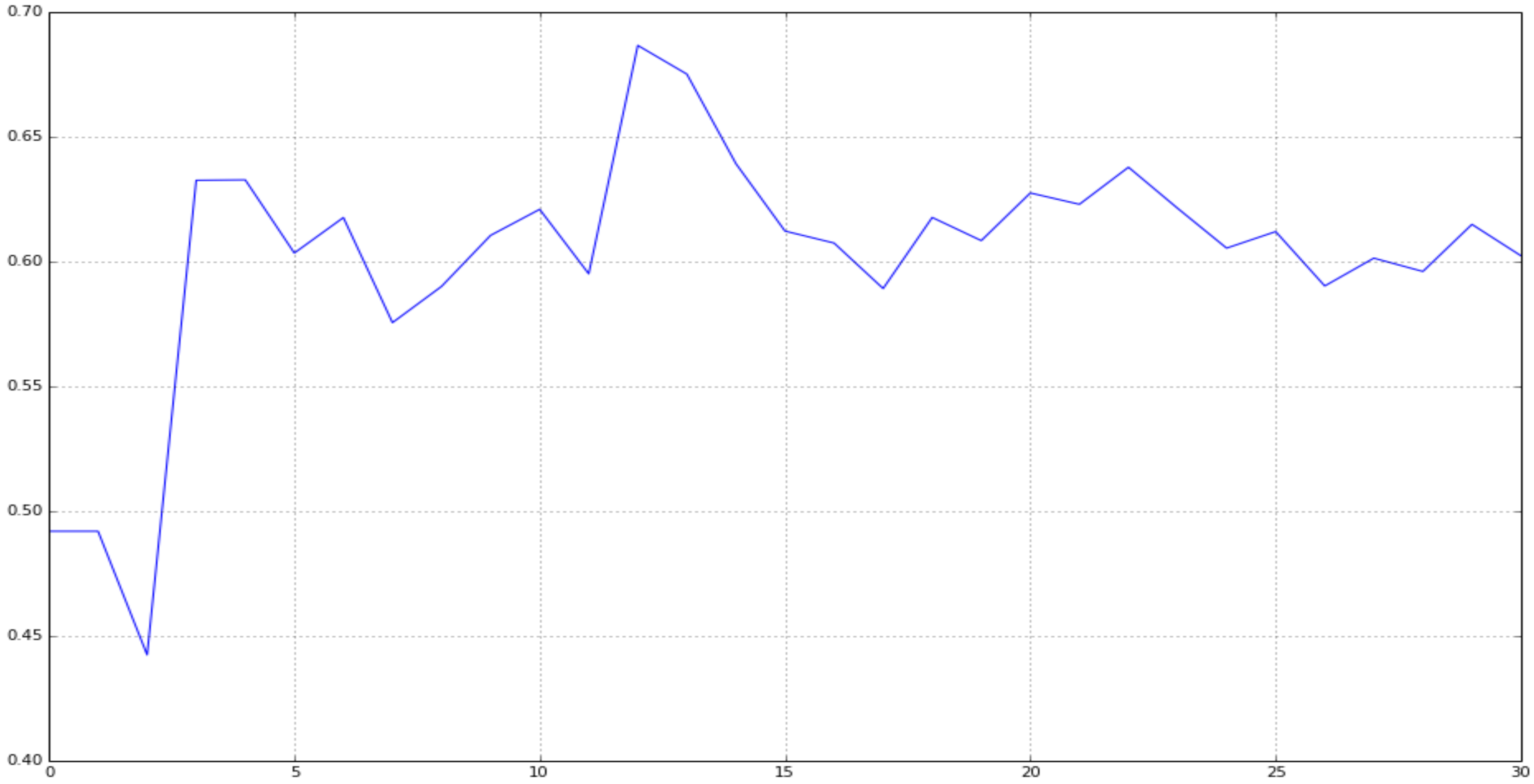


SS103: Physiology for Health Sciences

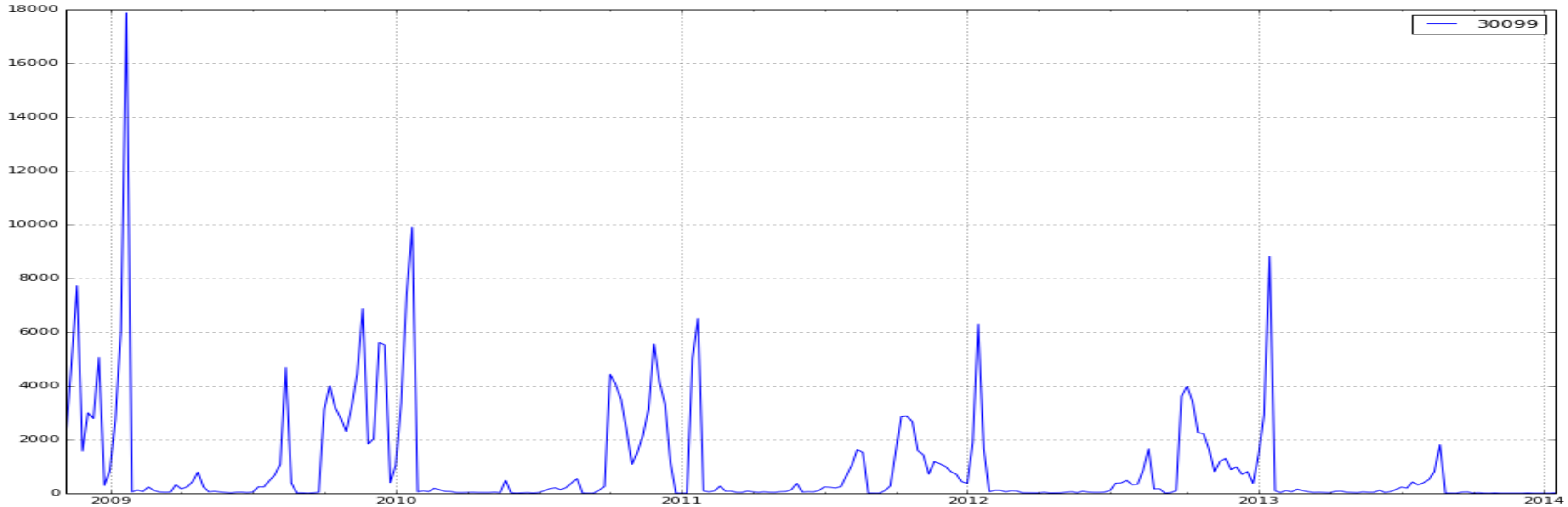
Results / Year = ~150
Pass rate = 0.92



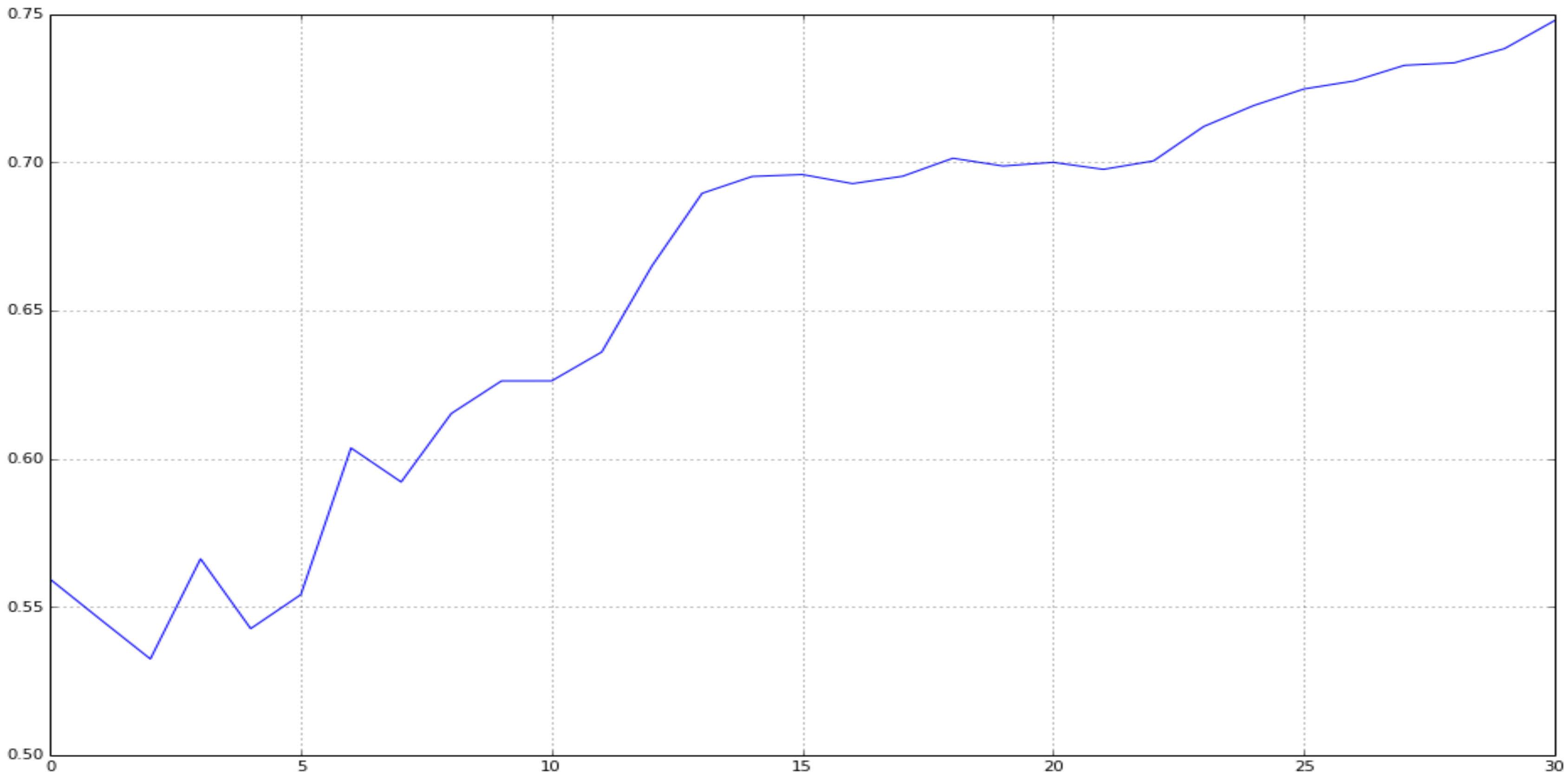
SS103



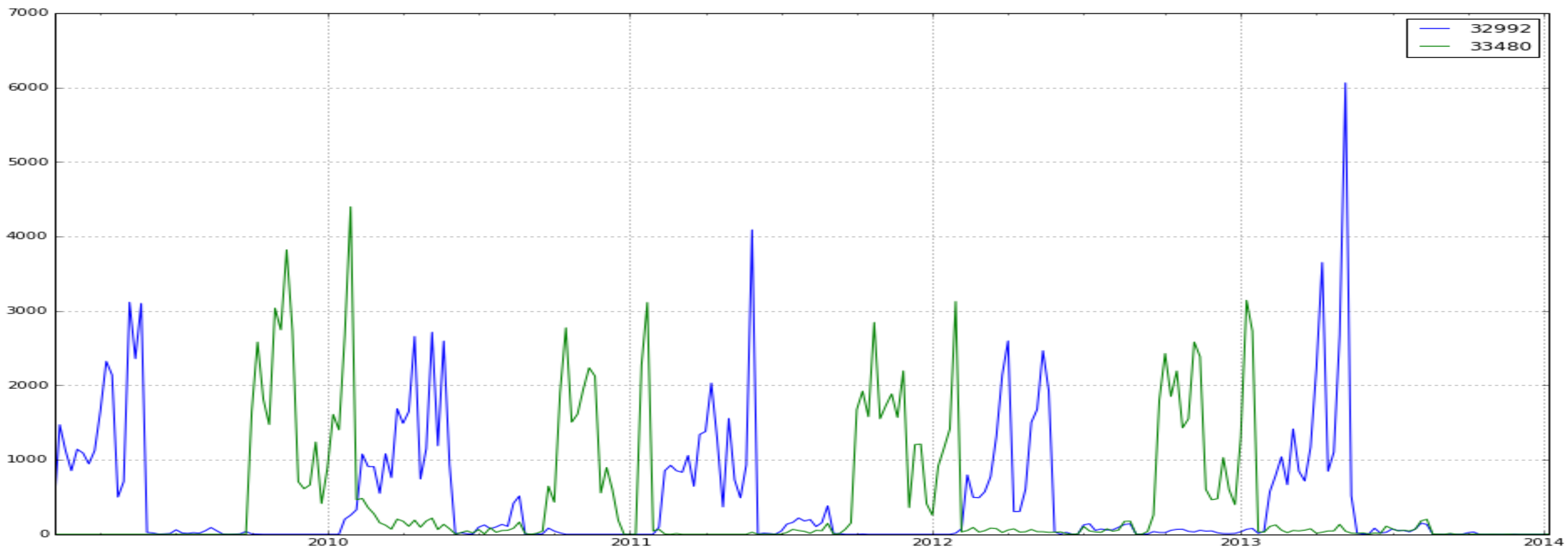
Results / year = ~250
Pass rate = 0.68



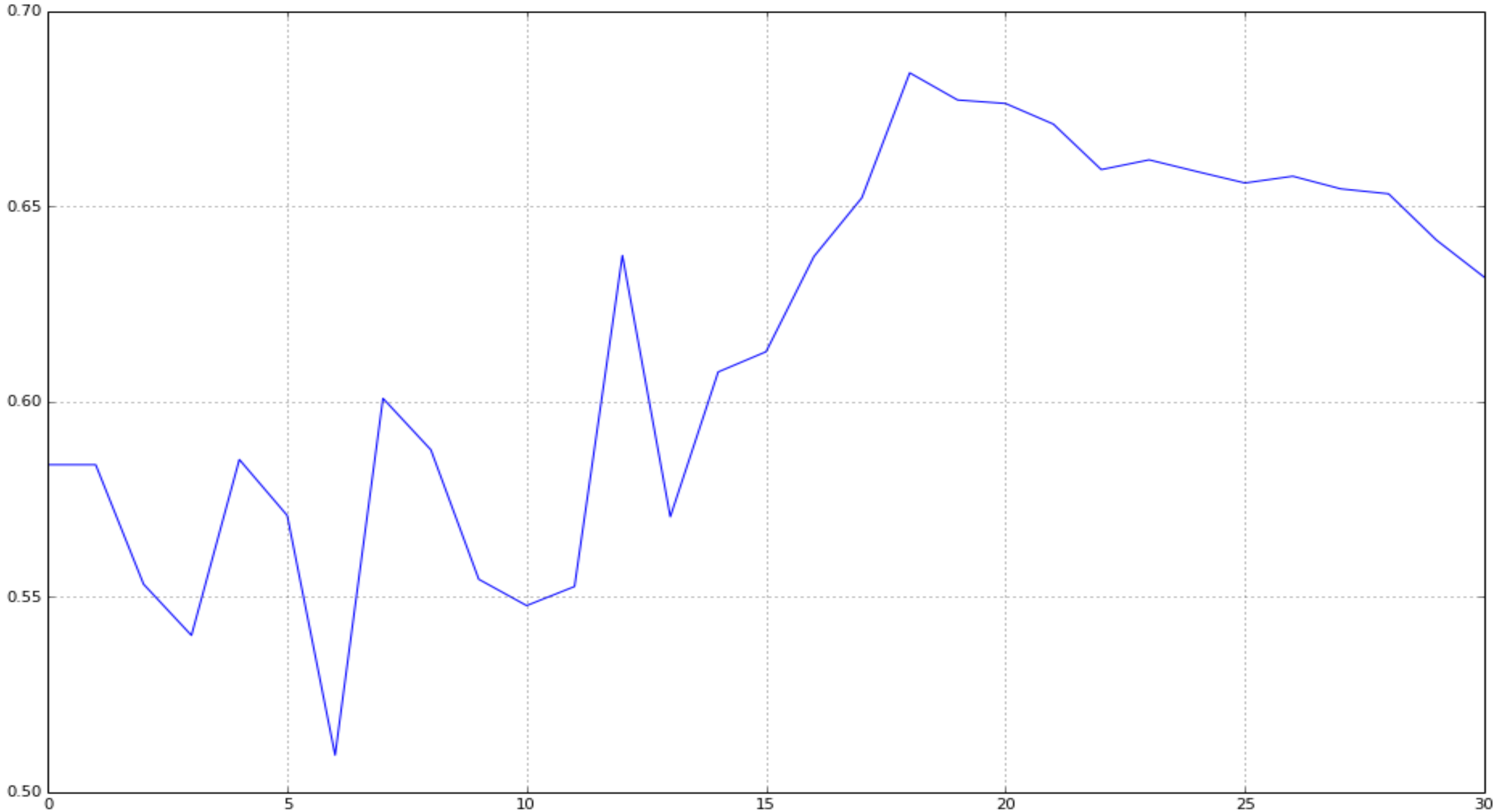
MS136



Results / Year = ~270 per semester
Pass rate = 0.87

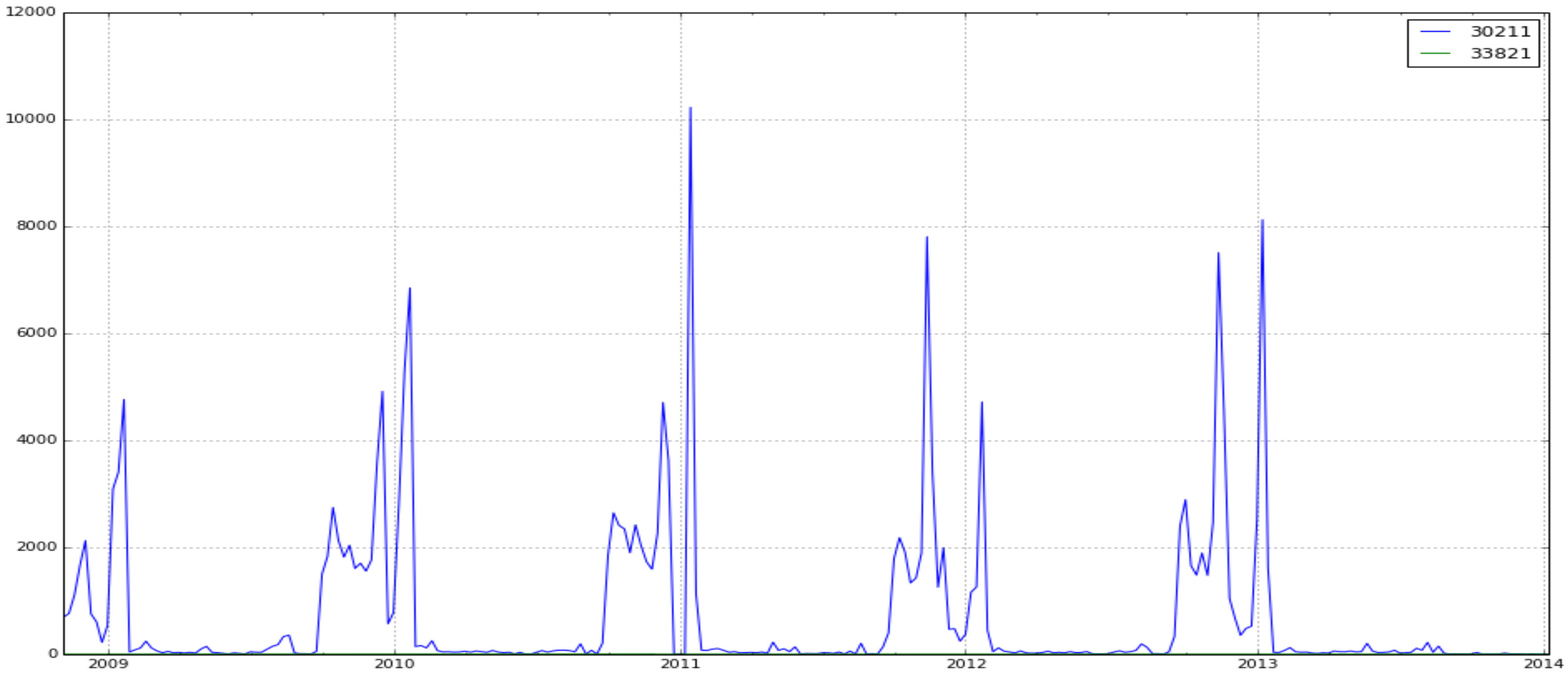


LG101

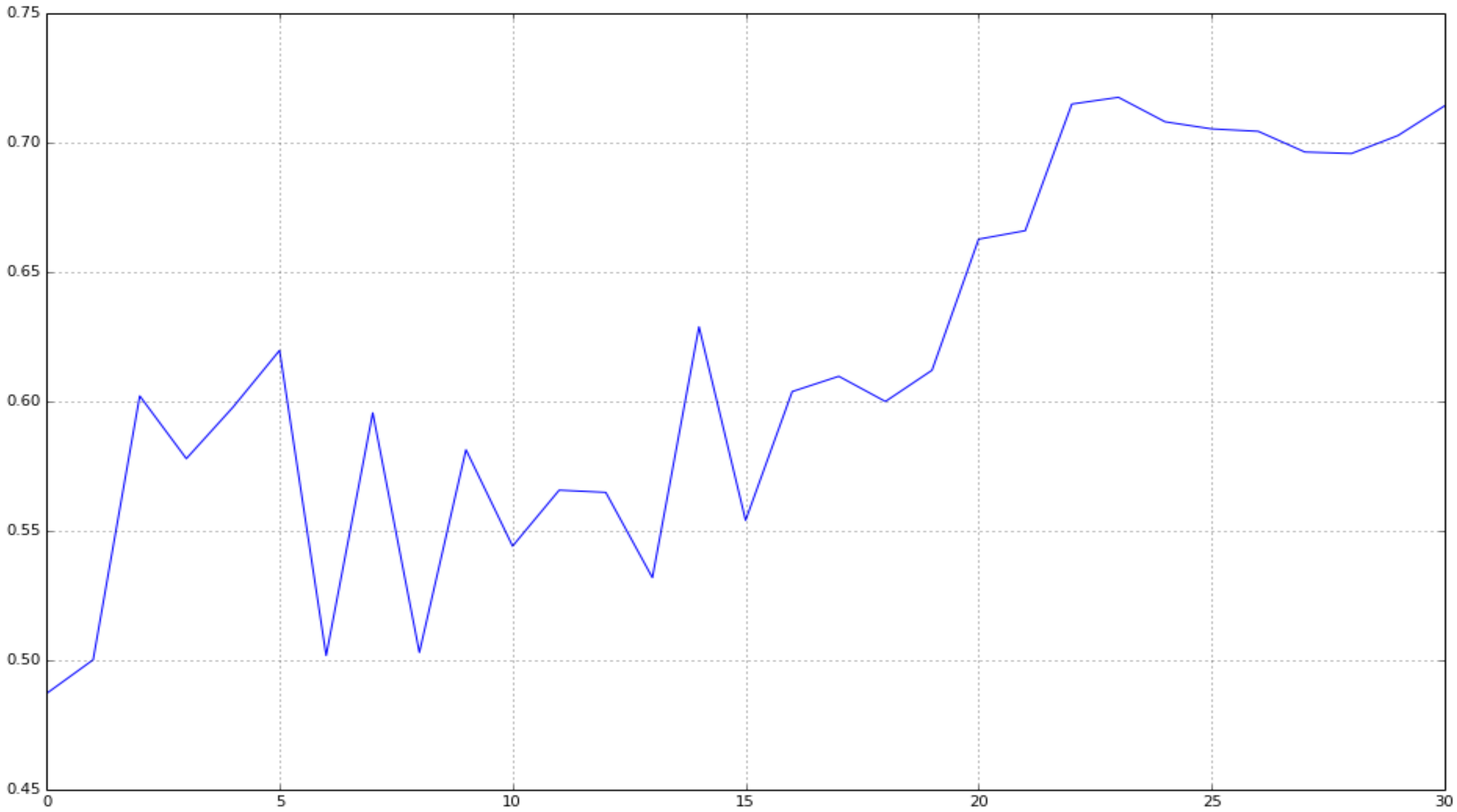


HR101: Psychology in Organisations

Results / Year – ~450
Pass Rate – 0.89

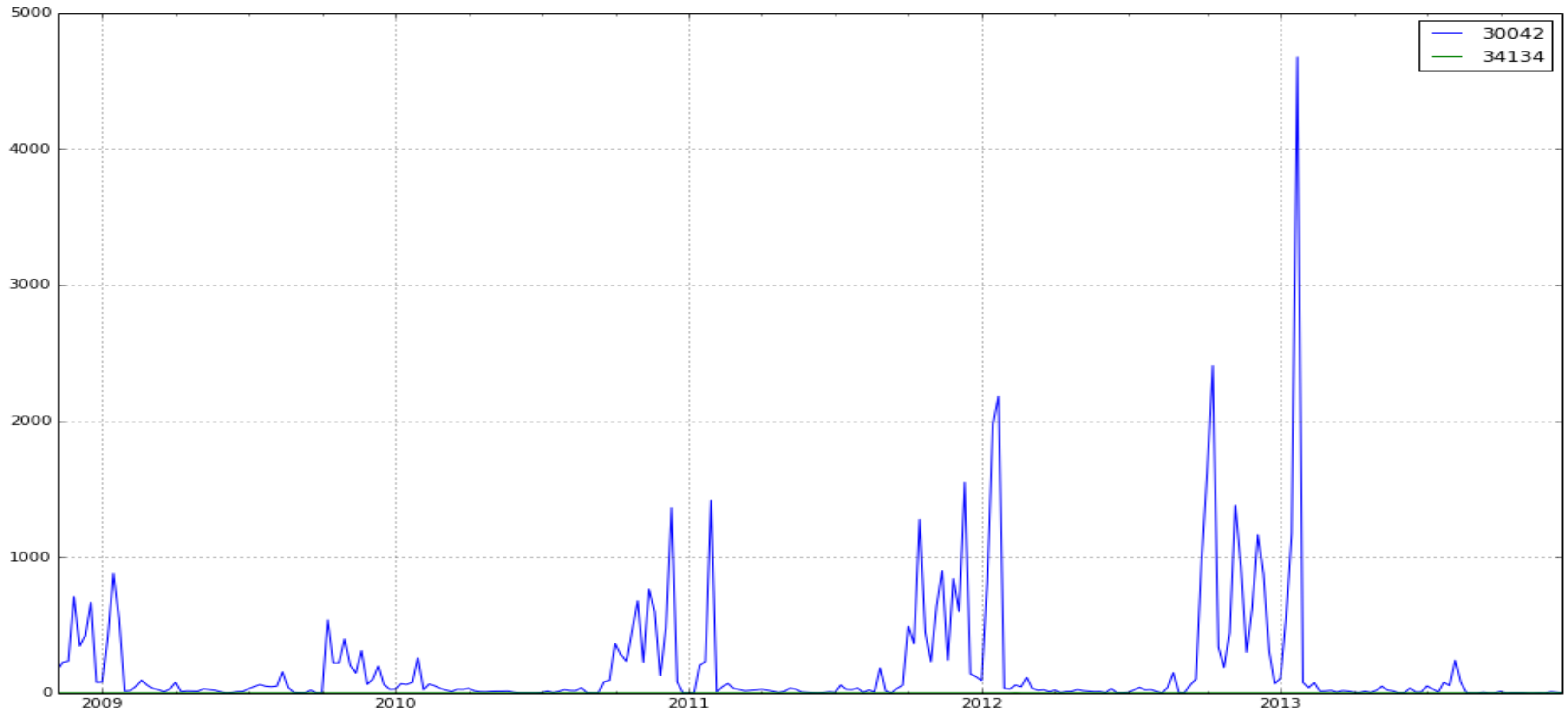


HR101

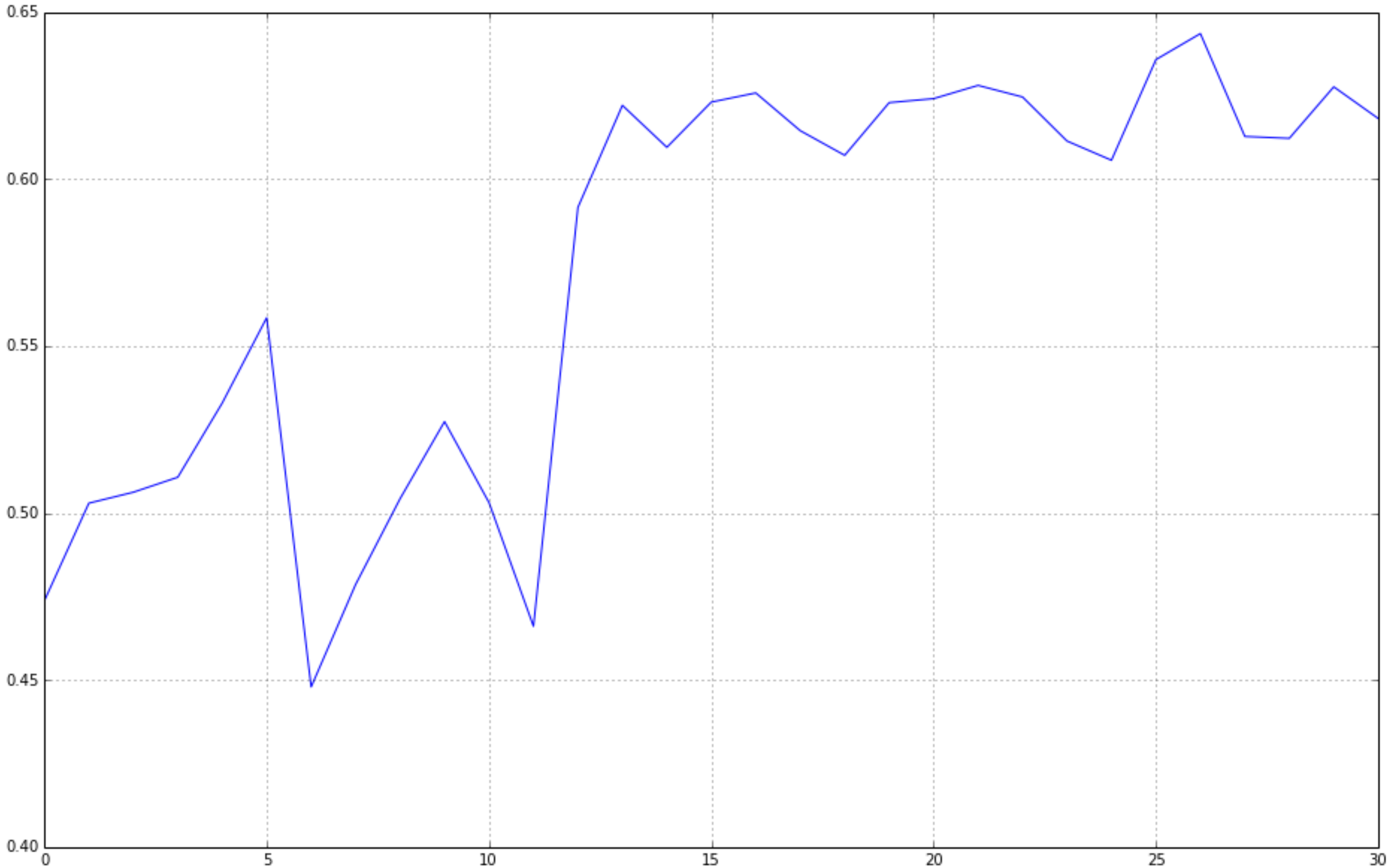


CA103: Computer Systems

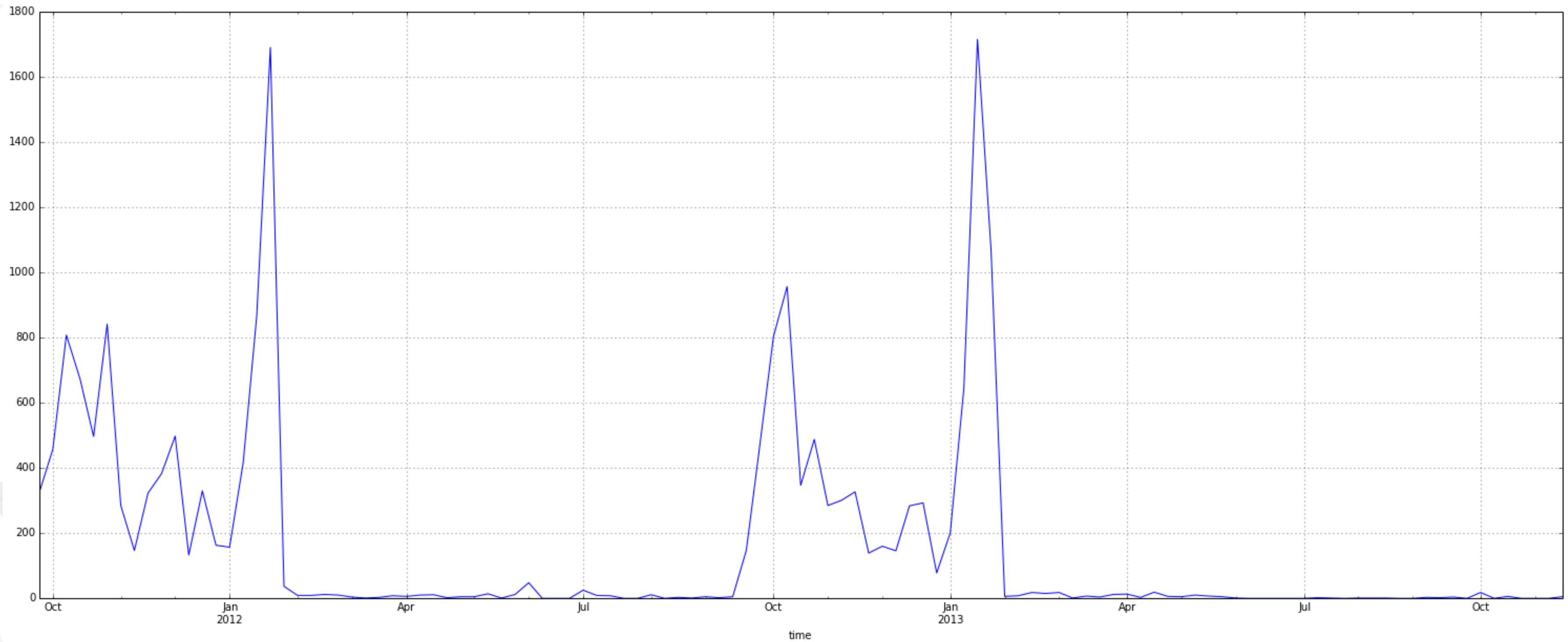
Results / year - ~140
Pass rate – 0.76



CA103

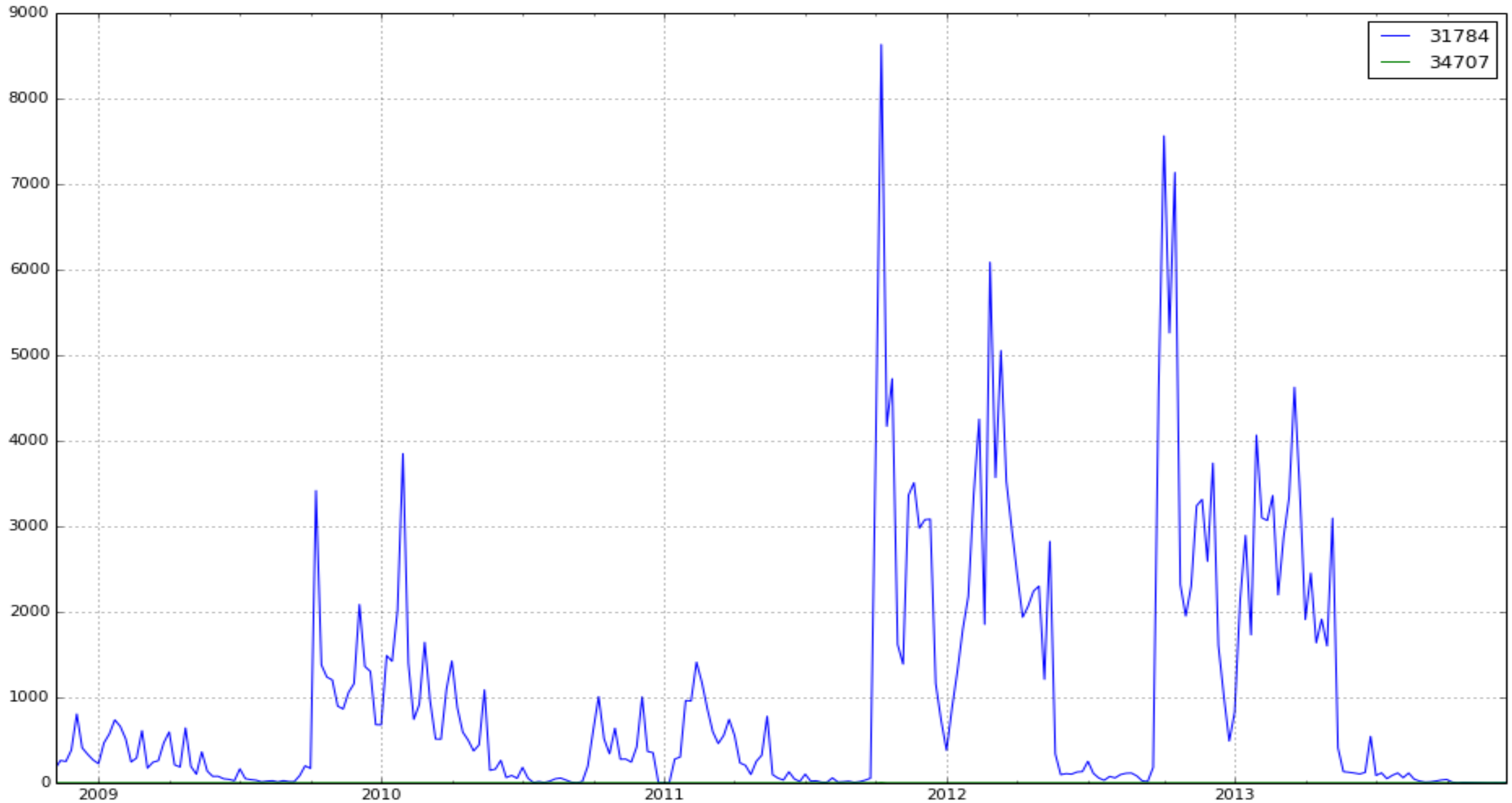


Results / year ~100
Pass rate 0.98



Some unusable modules:

Courses whose pattern changes over time – e.g.



Some unusable Modules

Modules with low numbers of students – below 100 per year

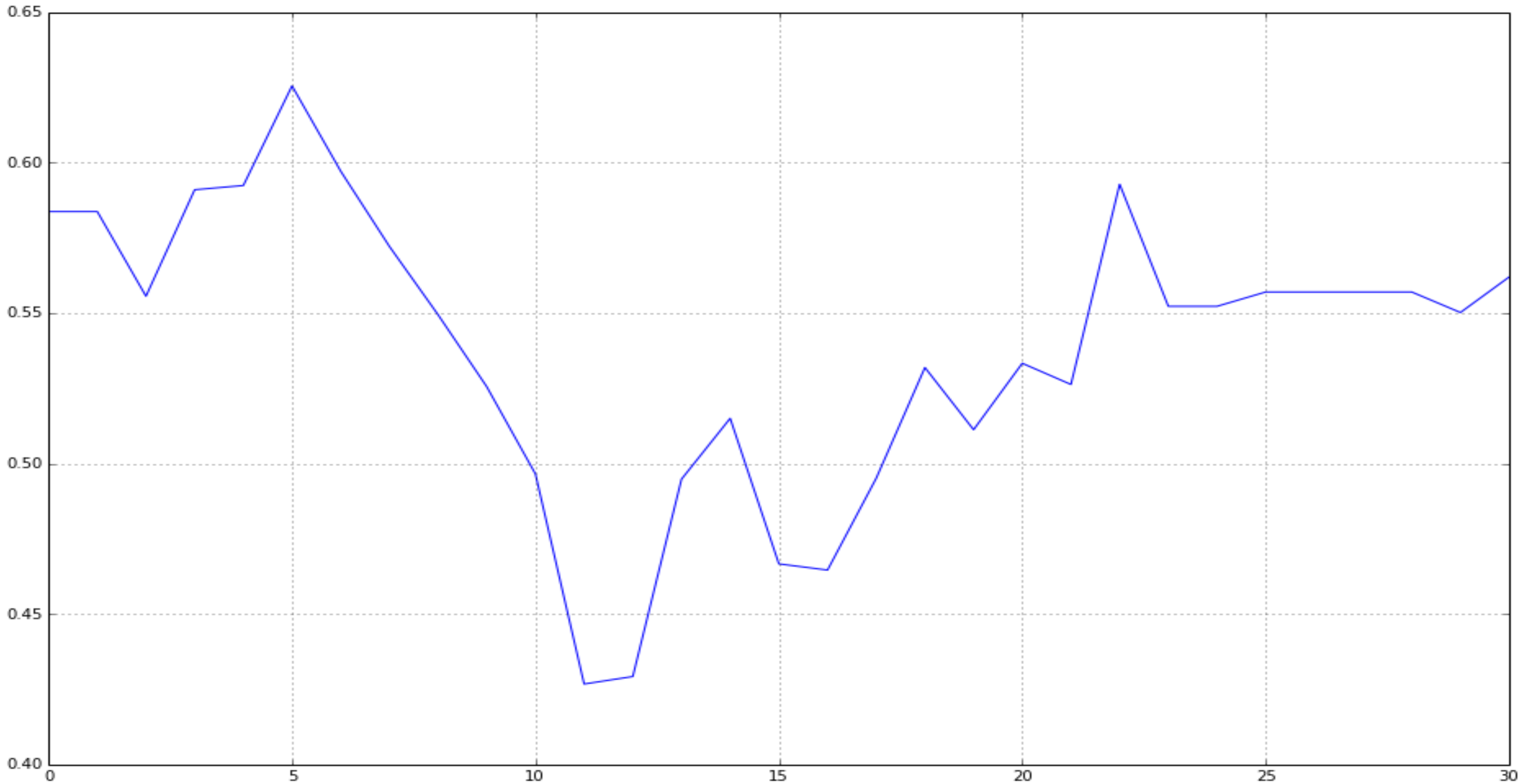
Modules with high pass rates – above 95%

Centre for
Data Analytics

Some unusable modules

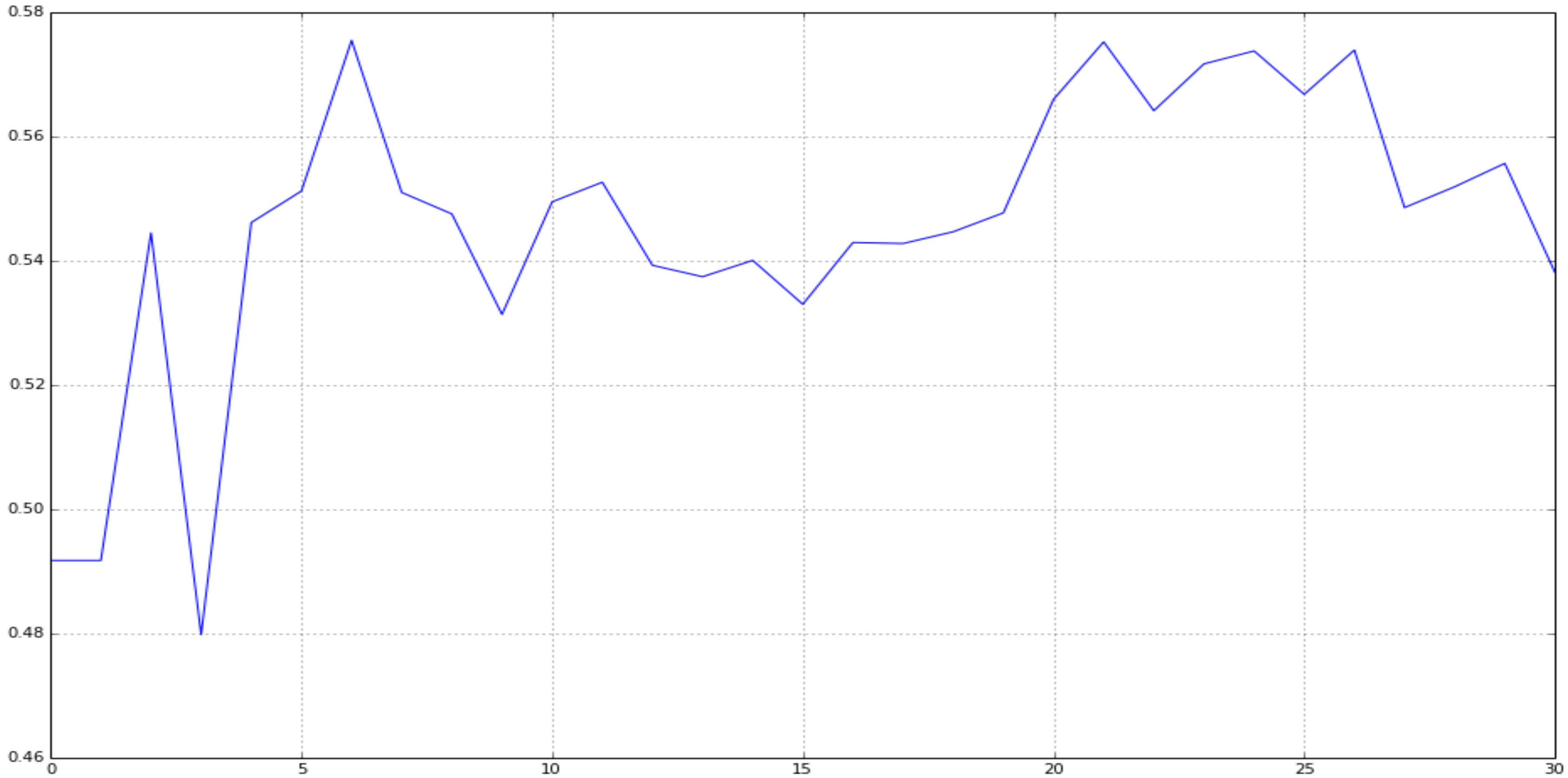
Insight

Modules where the prediction confidence stays the same or decreases – e.g. SS411



Some unusable modules

Modules where the ROC AUC increases slowly (e.g stays below 0.6) e.g. PS122



Building the System

We have 10 modules in semester 1, with impact on c.1670 students for 2014 intake

The Interventions – What Students Experience

Student Interventions: Personal Feedback

- Students in half of modules get a weekly email about personal performance relative to the target.

Dear _____,

This week our records show that your level of Moodle engagement is nearly at the target. If you try a little harder this week you will easily succeed.

Please use this information to help you to increase your engagement with Moodle. We will continue monitoring your Moodle activity for the module XX and will let you know how well you are doing again next week.

Kind Regards,
The Research Team
PredictED

If you feel affected by this and would like to speak to someone, please contact student support services (studentsupport@dcu.ie)

If you would like more information on this project please contact one of the research team members:

Alan Smeaton: alan.smeaton@dcu.ie
Sinead Smyth: sinead.smyth@dcu.ie

Centre for
Data Analytics

Insight

Student Interventions: Public Posting

- Students in other modules are emailed a weekly link to a league table depicting non-anonymised performances of all class members.
- The email is a link to a use-once (c.f. Snapchat) interactive graphic, which is time-limited and disappears forever after 3 minutes
- The evidence is that public postings give a social motivation to perform better
- Its easy for a student to move up the league table … use Moodle … can it be gamed ?

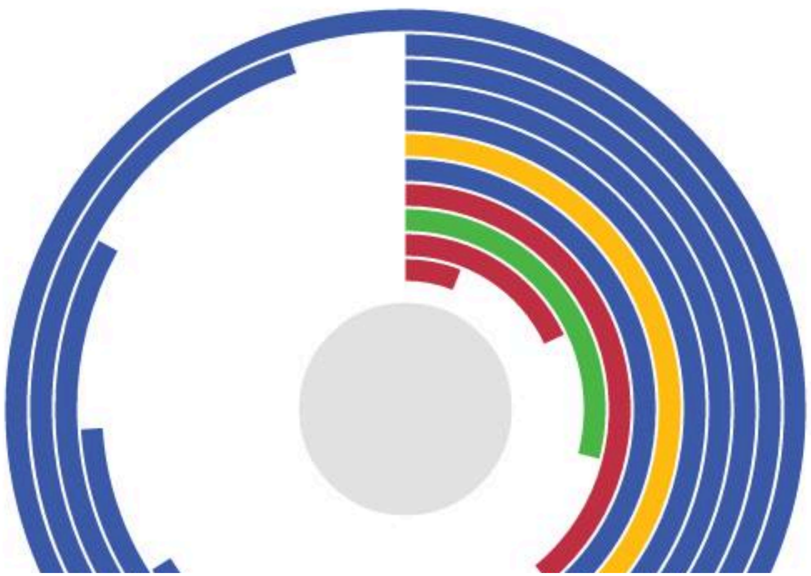
ost:5000

Overview of CA169 for John Brennan

Legend	
Red	Hasn't reached target
Blue	Has reached target
Yellow	This is you
Green	Item currently highlighted
Grey circle	Click to go back to the top level

Rankings

Section 3: hasn't reached target.



What is this?

Welcome to the interactive leaderboard for module CA169. This displays the progress of each student in terms of Moodle engagement in ranked order. Your progress can be seen in the yellow circle. If you lie below the target, this means that our programme predicts that you are at risk of failing this module based on your Moodle activity this week.

To increase your score to meet the target you are advised to engage more with the Moodle platform. Moodle is a learning tool to help you with your studies and has been found to

The Interventions – Lecturers' Experience

- Lecturers get a colour-coded dashboard showing …

  students x weeks x predictions

Centre for
Data Analytics

Insight

» home / engineering and computing / electronic engineering / EE417

Records | Information | Curriculum | Examination | Analysis | Prediction

Pass/Fail Prediction for EE417

Web Application Development : Pass/Fail Prediction

Columns — Page 1 of 1 — 100 — View 1 - 42 of 42

	Firstname	Surname	Std No	OT Desc	QualCo	Period	Exempt	Wk3	Wk4	Wk5	Wk6	Wk7	Wk8	Wk9	Wk10	Wk11	Wk12
1	Christopher	Mercado(*)	12210797	FULL-TIME	MEN	C	Y	1	2	1	2	3	4	3	2	1	1
2	George	Bruton(*)	12210878	FULL-TIME	MEN	C	Y	3	2	3	4	3	2	3	2	1	1
3	Richard	Murphy(*)	12210602	FULL-TIME	MEN	C	Y	4	5	4	5	4	5	6	7	8	7
4	Helen	Liu(*)	13119028	FULL-TIME	ECSA	X	N	6	7	8	9	9	8	7	8	9	9
5	Susan	Zhou(*)	13105124	FULL-TIME	ECSA	X	N	1	1	1	1	2	1	1	1	1	2
6	Brian	Holohan(*)	13119001	FULL-TIME	ECSA	X	N	7	6	5	6	7	6	5	4	5	4
7	Timothy	Wang(*)	14210408	PART-TIME	MTC	C	N	9	8	7	6	7	8	7	6	5	6
8	Brian	Chowdury(*)	13119036	FULL-TIME	ECSA	X	N	7	8	9	8	7	8	9	8	7	6
9	Sandra	Liu(*)	12210474	PART-TIME	MTC	C	N	9	9	8	7	6	7	6	7	8	7
10	Gary	Flynn(*)	13119605	FULL-TIME	MTC	C	N	3	2	3	2	3	4	3	4	5	6
11	Donna	Syed(*)	13212729	PART-TIME	MTC	C	N	2	3	4	5	6	5	6	7	8	7
12	Joseph	Mercado(*)	13211047	FULL-TIME	MTC	C	N	5	4	3	4	3	4	3	4	5	4
13	Michael	Breslin(*)	10319307	FULL-TIME	DME	4	N	6	5	6	7	8	9	9	9	9	8
14	George	Breslin(*)	59536582	FULL-TIME	ICE	4	N	7	8	7	8	7	6	5	4	5	4
15	Barbara	Gilbert(*)	10320107	FULL-TIME	DME	4	N	4	3	4	5	6	5	4	3	2	3
16	Dorothy	Ali(*)	12212354	PART-TIME	MEN	C	N	8	9	9	9	9	8	9	8	9	8
17	Deborah	Chowdury(*)	13210385	FULL-TIME	SMPEC	C	N	3	2	1	1	1	1	2	1	2	3
18	Kimberly	O Brien(*)	12212125	PART-TIME	MEN	C	N	9	9	9	9	9	8	9	8	7	8
19	Mary	Flynn(*)	12210644	FULL-TIME	MTC	C	Y	4	5	4	5	6	5	4	3	2	1
20	Laura	Uddin(*)	13211951	FULL-TIME	MEN	C	N	5	6	7	8	7	6	5	4	3	4
21	Nancy	Flynn(*)	59365249	FULL-TIME	DME	4	N	5	4	3	2	1	1	2	1	1	1
22	Karen	Brennan(*)	13212618	FULL-TIME	MTC	C	N	3	2	1	2	1	1	1	2	3	4
23	Brian	O Reilly(*)	58670617	FULL-TIME	DME	4	N	5	4	5	6	7	6	7	8	9	9

Timescale for Rollout

It starts in 7 days, and we're logging everything

If it works (how to measure ?) then we have to re-select modules each year, and re-train because course may change, lecturer may change, Moodle's offerings will change (tests) so transferability is not trivial

What next with student data ?

Knowing many things about your students can be used for more than predicting their pass/fail

Our roadmap is to improve prediction accuracy by using more information about our students … eduroam, library, sports, lifestyle … FaceBook ?

Not the best application because relationship between engagement & pass-fail is correlation/causation

The bigger prize is adapting course content to personalised models of individuals

US company Knewton does this, partnering with Arizona State University, but using class test results, not behaviour

KNEWTON

Google-fication of education ?

Given the ease with which you can find out anything you want to know, any where, any time, any device … would that be a bad thing ?

Predictive Analytics – change the future

Predicting the future changes it, especially if you want people to change as a result of knowing their future

This was the storyline in the "Back to the Future" movie trilogy, so maybe we'll never know

Thank you !