00:00:00:08 - 00:00:01:03
Hello,

00:00:01:03 - 00:00:01:20
I’m Leigh Smith,

00:00:01:20 - 00:00:02:13
a researcher at the

00:00:02:13 - 00:00:04:23
Institute of Astronomy in Cambridge.

00:00:04:23 - 00:00:06:17
My background is in stellar astronomy,

00:00:06:17 - 00:00:07:18
but a few years ago,

00:00:07:18 - 00:00:09:24
I found myself needing to process more data

00:00:09:24 - 00:00:11:10
than CPU-based tools

00:00:11:10 - 00:00:13:24
could handle in a reasonable time.

00:00:13:24 - 00:00:15:14
So I learned GPU programming

00:00:15:14 - 00:00:17:04
and wrote a GPU-accelerated

00:00:17:04 - 00:00:18:18
exoplanet transit detection algorithm

00:00:18:18 - 00:00:20:24
called CETRA.

00:00:20:24 - 00:00:22:06
I'm not a computer scientist,

00:00:22:06 - 00:00:23:10

and I had little prior

00:00:23:10 - 00:00:25:17
GPU experience before this project,

00:00:25:17 - 00:00:27:18
so if that's where you're starting from too,

00:00:27:18 - 00:00:29:20
hopefully, you'll find this deep dive useful.

00:00:31:01 - 00:00:32:18
The majority of the thousands of exoplanets

00:00:32:18 - 00:00:33:11
we've confirmed to date

00:00:33:11 - 00:00:34:15
were discovered through

00:00:34:15 - 00:00:36:09
a remarkably simple observation.

00:00:36:09 - 00:00:37:17
When a planet passes

00:00:37:17 - 00:00:39:04
in front of its host star,

00:00:39:04 - 00:00:41:06
it blocks some of the star's light.

00:00:41:06 - 00:00:42:14
If you're monitoring the brightness

00:00:42:14 - 00:00:43:23
of the star over time,

00:00:43:23 - 00:00:45:21
what astronomers call a light curve,

00:00:45:21 - 00:00:47:23
you see a brief periodic dip.

00:00:47:23 - 00:00:49:15
That's a transit.

00:00:49:15 - 00:00:51:03
The transit method is powerful

00:00:51:03 - 00:00:52:11
because it's scalable.

00:00:52:11 - 00:00:53:08
You can monitor

00:00:53:08 - 00:00:55:24
hundreds of thousands of stars simultaneously

00:00:55:24 - 00:00:57:22
with a single space telescope

00:00:57:22 - 00:01:00:04
and look for those dips automatically.

00:01:00:04 - 00:01:01:14
That's exactly what missions

00:01:01:14 - 00:01:03:15
like Kepler and TESS have been doing.

00:01:03:15 - 00:01:05:02
And it's what the European Space

00:01:05:02 - 00:01:06:15
Agency's upcoming PLATO

00:01:06:15 - 00:01:07:24
mission is designed to do

00:01:07:24 - 00:01:09:22
at an even greater scale.

00:01:09:22 - 00:01:11:09
The easy planets to find

00:01:11:09 - 00:01:14:10
are the big ones - hot Jupiters, close-in giants.

00:01:14:10 - 00:01:15:20
They produce deep transits

00:01:15:20 - 00:01:17:10
and transit frequently.

00:01:17:10 - 00:01:18:01
But the planets

00:01:18:01 - 00:01:19:00
that motivate the PLATO mission

00:01:19:00 - 00:01:22:06
are the small, rocky, temperate ones.

00:01:22:06 - 00:01:25:05
Earth-like planets around Sun-like stars.

00:01:25:05 - 00:01:27:18
The problem is that an Earth-Sun analogue

00:01:27:18 - 00:01:29:04
produces a transit depth

00:01:29:04 - 00:01:31:04
of around 100 parts per million,

00:01:31:04 - 00:01:32:13
a reduction in the star's

00:01:32:13 - 00:01:35:10
apparent brightness of 0.01%.

00:01:35:10 - 00:01:37:12
That is an extraordinarily small signal

00:01:37:12 - 00:01:39:07
to pull out of real data.

00:01:39:07 - 00:01:42:01

Stars are not perfectly stable light sources.

00:01:42:01 - 00:01:44:03
They have their own intrinsic variability

00:01:44:03 - 00:01:45:07
due to starspots,

00:01:45:07 - 00:01:47:08
granulation, and other activity.

00:01:47:08 - 00:01:48:12
And this variability,

00:01:48:12 - 00:01:50:05
sometimes called red noise,

00:01:50:05 - 00:01:52:22
can easily be larger in amplitude than

00:01:52:22 - 00:01:55:01
the transit signal we're trying to detect.

00:01:55:01 - 00:01:56:01
The standard approach

00:01:56:01 - 00:01:57:15
is to detrend the light curve

00:01:57:15 - 00:01:59:07
before running a transit search.

00:01:59:07 - 00:02:01:21
That is, to model and remove the slow

00:02:01:21 - 00:02:04:00
correlated variations in brightness,

00:02:04:00 - 00:02:04:18
leaving behind

00:02:04:18 - 00:02:06:00
only the transit signal

00:02:06:00 - 00:02:07:19
and uncorrelated noise.

00:02:07:19 - 00:02:09:18
CETRA is designed to work on

00:02:09:18 - 00:02:10:20
detrended light curves,

00:02:10:20 - 00:02:11:13
and we'll take that

00:02:11:13 - 00:02:13:02
as a given throughout.

00:02:13:02 - 00:02:13:22
And then there are

00:02:13:22 - 00:02:15:06
the measurement errors.

00:02:15:06 - 00:02:16:15
Even the best space-based

00:02:16:15 - 00:02:18:04
photometry is not perfect:

00:02:18:04 - 00:02:19:20
detector noise, pointing jitter,

00:02:19:20 - 00:02:21:17
and other instrumental effects

00:02:21:17 - 00:02:23:18
all contribute to uncertainty

00:02:23:18 - 00:02:25:12
in each flux measurement.

00:02:25:12 - 00:02:27:11
PLATO's highest-priority targets -

00:02:27:11 - 00:02:30:03
the P1 sample of around 15,000

00:02:30:03 - 00:02:32:20
bright nearby Sun-like stars -

00:02:32:20 - 00:02:33:24
are specified to achieve

00:02:33:24 - 00:02:34:17
a noise level

00:02:34:17 - 00:02:37:20
better than 50 parts per million per hour.

00:02:37:20 - 00:02:39:06
To detect a transit,

00:02:39:06 - 00:02:40:08
you need its signal

00:02:40:08 - 00:02:42:12
to stand out above the noise.

00:02:42:12 - 00:02:43:11
This signal-to-noise ratio,

00:02:43:11 - 00:02:44:23
or SNR,

00:02:44:23 - 00:02:46:14
is simply the measured transit

00:02:46:14 - 00:02:48:06
depth divided by the uncertainty

00:02:48:06 - 00:02:50:08
on that depth measurement.

00:02:50:08 - 00:02:52:08
The uncertainty depends on the noise

00:02:52:08 - 00:02:53:10

level of the instrument

00:02:53:10 - 00:02:54:24
and how many data points fall

00:02:54:24 - 00:02:56:01
within the transit,

00:02:56:01 - 00:02:58:01
which is why longer transits

00:02:58:01 - 00:02:59:10
higher cadence data,

00:02:59:10 - 00:03:01:09
and multiple observed transits

00:03:01:09 - 00:03:03:17
all improve your SNR.

00:03:03:17 - 00:03:05:08
For an Earth-Sun analogue

00:03:05:08 - 00:03:07:22
observed by PLATO at P1 noise levels,

00:03:07:22 - 00:03:09:20
the SNR reaches around 13

00:03:09:20 - 00:03:11:17
when two transits are combined,

00:03:11:17 - 00:03:14:01
rising to around 15 for three.

00:03:14:01 - 00:03:15:12
To put that in perspective,

00:03:15:12 - 00:03:17:09
you need to observe the same planet

00:03:17:09 - 00:03:19:14
transit its star multiple times

00:03:19:14 - 00:03:21:21
just to be confident the signal is real.

00:03:21:21 - 00:03:24:01
And that's for the best-case targets.

00:03:24:01 - 00:03:25:02
The bulk of PLATO's survey

00:03:25:02 - 00:03:26:23
consists of fainter stars

00:03:26:23 - 00:03:28:11
in the P5 sample -

00:03:28:11 - 00:03:30:05
several hundred thousand stars

00:03:30:05 - 00:03:32:04
extending to fainter magnitudes -

00:03:32:04 - 00:03:33:04
where noise levels

00:03:33:04 - 00:03:34:24
can be several times higher,

00:03:34:24 - 00:03:36:16
making the detection of small planets

00:03:36:16 - 00:03:38:01
significantly harder

00:03:38:01 - 00:03:39:21
and placing even greater demands

00:03:39:21 - 00:03:40:23
on the sensitivity

00:03:40:23 - 00:03:43:04
of the transit search algorithm.

00:03:43:04 - 00:03:45:01
The PLATO mission, planned for launch

00:03:45:01 - 00:03:48:05
in 2027, is specifically designed to find

00:03:48:05 - 00:03:49:23
and characterise terrestrial planets

00:03:49:23 - 00:03:51:23
around Sun-like stars.

00:03:51:23 - 00:03:53:07
It will monitor a few hundred

00:03:53:07 - 00:03:55:08
thousand stars simultaneously

00:03:55:08 - 00:03:57:03
at two different cadences.

00:03:57:03 - 00:03:58:20
Cadence refers to the time

00:03:58:20 - 00:04:01:00
between successive observations.

00:04:01:00 - 00:04:02:21
Long cadence targets are observed

00:04:02:21 - 00:04:04:13
every 600 seconds,

00:04:04:13 - 00:04:05:12
and short-cadence

00:04:05:12 - 00:04:07:12
targets every 25-seconds,

00:04:07:12 - 00:04:08:16
over an epoch baseline

00:04:08:16 - 00:04:10:17

of at least two years.

00:04:10:17 - 00:04:11:24
That combination -

00:04:11:24 - 00:04:13:03
large target numbers,

00:04:13:03 - 00:04:14:04
high cadence,

00:04:14:04 - 00:04:16:04
long baseline - creates a data

00:04:16:04 - 00:04:17:05
processing challenge

00:04:17:05 - 00:04:18:12
that the exoplanet community

00:04:18:12 - 00:04:20:05
hasn't faced before.

00:04:20:05 - 00:04:21:15
Light curves of this length

00:04:21:15 - 00:04:23:06
and cadence are simply larger

00:04:23:06 - 00:04:23:16
than those

00:04:23:16 - 00:04:24:18
from previous missions,

00:04:24:18 - 00:04:25:22
and transit detection

00:04:25:22 - 00:04:27:15
needs to be run on all of them.

00:04:27:15 - 00:04:29:01
To put some numbers on this,

00:04:29:01 - 00:04:30:06
running the current state-of-the-art

00:04:30:06 - 00:04:32:01
transit detection algorithm,

00:04:32:01 - 00:04:34:08
Transit Least Squares,

00:04:34:08 - 00:04:36:07
on a single two-year PLATO light curve

00:04:36:07 - 00:04:40:13
at a 600-second cadence takes around 2 CPU core-hours.

00:04:40:13 - 00:04:42:01
For a four-year light curve,

00:04:42:01 - 00:04:44:07
it's closer to 10 core-hours.

00:04:44:07 - 00:04:46:19
Running at 25 second cadence instead

00:04:46:19 - 00:04:49:01
takes around 50 times longer.

00:04:49:01 - 00:04:51:10
Across hundreds of thousands of targets,

00:04:51:10 - 00:04:53:21
that's a major computational challenge.

00:04:53:21 - 00:04:55:18
Even with a large computing cluster,

00:04:55:18 - 00:04:57:15
this is a significant burden.

00:04:57:15 - 00:04:59:04
This is the problem CETRA

00:04:59:04 - 00:05:00:16
was built to address.

00:05:00:16 - 00:05:02:10
Before getting into how CETRA works,

00:05:02:10 - 00:05:03:11
it's worth understanding

00:05:03:11 - 00:05:04:09
what came before it

00:05:04:09 - 00:05:05:18
and why these approaches

00:05:05:18 - 00:05:07:13
have limitations at this scale.

00:05:07:13 - 00:05:08:16
The Box-fitting Least Squares

00:05:08:16 - 00:05:10:10
algorithm, BLS,

00:05:10:10 - 00:05:12:23
published by Kovács et al. in 2002,

00:05:12:23 - 00:05:14:11
was the standard transit

00:05:14:11 - 00:05:16:17
detection tool for many years.

00:05:16:17 - 00:05:18:18
The core idea is straightforward:

00:05:18:18 - 00:05:21:00
BLS phase-folds the light curve

00:05:21:00 - 00:05:23:03
over a grid of trial periods.

00:05:23:03 - 00:05:24:06

For each period,

00:05:24:06 - 00:05:25:05
it checks a box-shaped

00:05:25:05 - 00:05:26:09
transit model -

00:05:26:09 - 00:05:29:05
flat bottom, zero ingress and egress time -

00:05:29:05 - 00:05:31:10
against the data at a grid of transit

00:05:31:10 - 00:05:33:06
times and durations, and finds

00:05:33:06 - 00:05:34:23
the least-squares best fit.

00:05:34:23 - 00:05:36:21
The result is a periodogram:

00:05:36:21 - 00:05:38:03
a measure of how well

00:05:38:03 - 00:05:39:10
a periodic box-shaped

00:05:39:10 - 00:05:40:17
signal fits the data,

00:05:40:17 - 00:05:42:13
as a function of period.

00:05:42:13 - 00:05:43:05
It works,

00:05:43:05 - 00:05:44:21
and it was a major step forward,

00:05:44:21 - 00:05:46:21
but a box is only an approximation

00:05:46:21 - 00:05:48:12
to a real exoplanet transit,

00:05:48:12 - 00:05:50:16
which has a smooth ingress and egress

00:05:50:16 - 00:05:52:09
shaped by the geometry of the planet

00:05:52:09 - 00:05:53:17
crossing the stellar limb.

00:05:53:17 - 00:05:55:13
At low signal-to-noise ratios -

00:05:55:13 - 00:05:58:02
which is exactly where Earth-like planets live -

00:05:58:02 - 00:06:00:00
this mismatch in model shape costs

00:06:00:00 - 00:06:00:22
you sensitivity.

00:06:00:22 - 00:06:03:07
Transit Least Squares, TLS,

00:06:03:07 - 00:06:06:02
published by Hippke and Heller in 2019,

00:06:06:02 - 00:06:08:08
addressed this by replacing the box model

00:06:08:08 - 00:06:09:21
with a physically motivated

00:06:09:21 - 00:06:12:20
exoplanet transit shape.

00:06:12:20 - 00:06:14:19
TLS also introduced optimised

00:06:14:19 - 00:06:17:05
grids for the period and duration search,

00:06:17:05 - 00:06:18:17
astrophysically motivated

00:06:18:17 - 00:06:19:19
so that you're not wasting

00:06:19:19 - 00:06:21:12
compute time on combinations

00:06:21:12 - 00:06:22:13
that couldn't correspond

00:06:22:13 - 00:06:23:16
to a real planet.

00:06:23:16 - 00:06:25:11
These two changes together

00:06:25:11 - 00:06:26:13
gave meaningful

00:06:26:13 - 00:06:28:23
improvements in detection sensitivity,

00:06:28:23 - 00:06:30:18
particularly for small planets,

00:06:30:18 - 00:06:32:06
without dramatically increasing

00:06:32:06 - 00:06:33:19
the computational cost.

00:06:33:19 - 00:06:35:17
TLS is the current state of the art,

00:06:35:17 - 00:06:37:02
and it's the benchmark against

00:06:37:02 - 00:06:38:19

which CETRA is measured.

00:06:38:19 - 00:06:40:06
But it has some limitations

00:06:40:06 - 00:06:41:13
that become significant

00:06:41:13 - 00:06:43:18
at the scale PLATO demands.

00:06:43:18 - 00:06:45:15
The first is computational cost,

00:06:45:15 - 00:06:46:24
which we've already touched on.

00:06:46:24 - 00:06:48:22
The second is how it handles

00:06:48:22 - 00:06:50:05
gaps in the data.

00:06:50:05 - 00:06:52:07
Real light curves always have gaps -

00:06:52:07 - 00:06:53:15
between observing sectors,

00:06:53:15 - 00:06:54:21
during data downlinks,

00:06:54:21 - 00:06:55:09
or simply

00:06:55:09 - 00:06:56:12
because the target wasn't

00:06:56:12 - 00:06:58:11
observed continuously.

00:06:58:11 - 00:06:59:24
TLS phase-folds

00:06:59:24 - 00:07:01:12
the light curve before searching,

00:07:01:12 - 00:07:02:15
which means that gaps

00:07:02:15 - 00:07:04:21
in the original data become irregular

00:07:04:21 - 00:07:06:12
sampling in phase space.

00:07:06:12 - 00:07:07:21
For short-period planets

00:07:07:21 - 00:07:09:06
with many observed transits

00:07:09:06 - 00:07:10:13
this averages out,

00:07:10:13 - 00:07:11:24
but for long-period planets

00:07:11:24 - 00:07:14:23
where only 2 or 3 transits are observed,

00:07:14:23 - 00:07:16:05
a gap that falls on

00:07:16:05 - 00:07:17:04
one of those transits

00:07:17:04 - 00:07:19:12
can significantly reduce sensitivity.

00:07:19:12 - 00:07:20:24
The third is a subtlety in

00:07:20:24 - 00:07:22:12
how TLS steps through trial

00:07:22:12 - 00:07:23:13
transit times -

00:07:23:13 - 00:07:26:04
that is, the grid of times at which it tests

00:07:26:04 - 00:07:27:24
for the centre of a transit.

00:07:27:24 - 00:07:29:11
For long-period planets,

00:07:29:11 - 00:07:31:08
the effective step size in this grid

00:07:31:08 - 00:07:33:00
can become larger than intended

00:07:33:00 - 00:07:34:20
due to an algorithmic constraint,

00:07:34:20 - 00:07:36:19
again, reducing sensitivity

00:07:36:19 - 00:07:38:00
for exactly the targets

00:07:38:00 - 00:07:39:10
PLATO cares most about.

00:07:39:10 - 00:07:41:07
These aren't fatal flaws in TLS.

00:07:41:07 - 00:07:42:02
It's an excellent

00:07:42:02 - 00:07:43:12
algorithm, and CETRA borrows

00:07:43:12 - 00:07:45:02
heavily from its ideas.

00:07:45:02 - 00:07:46:21

But they're the specific problems

00:07:46:21 - 00:07:48:18
that motivated a new approach,

00:07:48:18 - 00:07:50:04
and they're worth keeping in mind

00:07:50:04 - 00:07:51:24
as we go through the CETRA design.