

# Data-Centric Engineering

## Preparing Your Research Proposal

[qmul.ac.uk/dce](http://qmul.ac.uk/dce)

- What is a Research Proposal?
- What we are looking for
- Examples

Pre-application Workshop 2  
Queen Mary University of London  
13<sup>th</sup> April 2021



Dr Eram Rizvi  
Deputy Dean for Research



## Research Proposal

- Opportunity to express yourself
- Explain what research excites you
- Communicate the project well / succinctly
- Illustrate the potential impact of the project

Research Proposal  
300-700 words

**Show us how you think**  
**Show us how you communicate**

### Why is the research necessary?

Is there a demand?

Has the problem already been solved?

### Who are the beneficiaries of the research?

industry sector

a specific company or organisation

is there a societal benefit?

will this work assist the UK economy?

You will not be bound to this project - need to find matching supervisor  
You can contact potential supervisors now (see links in introduction slides)  
We can assist in finding suitable supervisor  
Projects will be adapted / refined with supervisor if you are successful



## Structure Of A Research Proposal

<https://www.phdassistance.com/blog/how-to-write-your-research-proposal/>



- A catchy title helps grab attention
- Give some background to the problem
- What is the challenge ? Why is this needed?
- What techniques / facilities will be used? Where will the data come from?
- Can you outline a work-plan for the research?
- Include a short bibliography (excluded from word count) not necessarily academic journal papers

# What We are Looking For



Strictly 700 words maximum (excl. bibliography)

Describe 1 or 2 projects (max.)

Not much space to describe your proposal(s)

Below I list some things to consider

we do not expect all of these to be met in your proposal

consider what is most relevant to your topic

Proposal should show

- Clearly structured thought
- Ability to plan out a project - timeline and resources
- Informed background knowledge
- Match to your skills / competencies

Have you thought about the problem from more than one perspective?

Do you have the resources needed, how will you acquire them?

Are there ethical considerations, e.g. using personal data?

Are you framing the right research question?

Can you articulate why it is interesting?

**In-depth academic knowledge of the stat-of-the-art is not necessary  
...but if you have it then tell us about it!**



Examples of Doctoral Research Projects on offer - please email supervisors for more information

Supervisor	School	Project Title
<a href="#">Dr. Chris Jones</a>	Biological & Chemical Sciences	Developing of Novel Cannabinoids with Machine Learning
<a href="#">Prof. Christian Beck</a>	Mathematics	Modelling of power-grid frequency
<a href="#">Dr Akram Alomainy</a>	Electronic	Development of Novel Data-Centric Techniques in emergent wireless
<a href="#">Dr Richard Clewley</a>	Computer Science	Analysis of software
<a href="#">Prof. Adrian Beckett</a>	Physics	Quantum detectors
<a href="#">Prof. David Dunham</a>	Statistics	Bayesian Likelihood Fits in Data-
<a href="#">Dr Lin Wang</a>	Music	Enhancement
<a href="#">Prof. Josh Reiss</a>	Computer Science	Parametric Controls from Data Analytics
<a href="#">Dr Seth Zenz</a>	Physics	Designing novel ultra-thin low mass curved silicon imaging sensors for X-ray diffraction and nuclear security
<a href="#">Dr Yannick Wurm</a>	Biological & Chemical Sciences	A toolkit for pragmatic interrogation exploration & hypothesis testing of disconnected genomic data
<a href="#">Dr Anthony Phillips</a>	Physics	Using machine-learning techniques to identify new perovskite materials with electrical properties that exceeding those of inorganic perovskites
<a href="#">Dr Jens-Dominik Mueller</a>	Engineering	Adaptive multi-fidelity robust design optimisation driven by machine learning

<https://www.qmul.ac.uk/dce/research/>

**Links to Available Doctoral Research Projects in our Schools**

- [School of Electronic Engineering and Computer Science](#)
- [School of Engineering and Materials Science](#)
- [School of Chemical and Biological Sciences](#)
- [School of Physics and Astronomy](#)
- [School of Mathematical Sciences](#)



Some examples given here - looking at titles only for this exercise

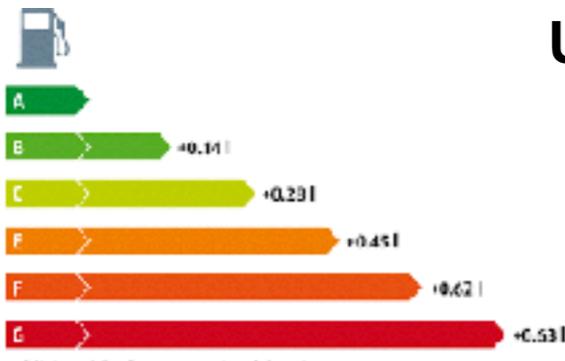
## Research into new black holes



exciting topic - Nobel Prize for physics 2020!  
vague title - what will you investigate?  
not data-centric engineering  
do you have skills in general relativity?  
will industry partners be willing to participate? probably not...  
not applied research

## Building a large-scale distributed database for financial assets

no research element - application of existing technology  
where is the challenge? This is a solved problem



## Using machine learning to increase vehicle fuel efficiency

good topic  
potential interest from many industry organisations  
clearly a data-engineering problem  
you will need a background in vehicle design engineering



## **Using machine-learning to identify new perovskite materials with novel electrical properties**

new application of machine learning

potential benefit to industries using perovskites (e.g. renewable solar energy)

merges chemical / materials engineering with data science approach



## **Designing novel ultra-thin curved silicon sensors for nuclear security**

interesting data-centric engineering topic

requires software modelling skills

methodology is not clear from title

target industry is clear



## **Building a model of urban air pollution**

hot topic currently

potential interest from many industry and governments (incl. internationally)

clearly a data-engineering problem: use input data to generate software model



## **Good exercises to prepare the proposal**

Outline with bullet-point headings

Re-order bullet points for logical flow

Assign rough word counts to each paragraph in order of importance

Draft each paragraph

....

Forget about the draft for 1-2 days

Return to it with fresh eyes and re-order sections if needed

Cut the word-count by 20% or 30% but without losing any content

...

Repeat if you have time



**This process helps you focus on what is important**

**Your final draft will be much 'tighter'**



in the calibration of the electron and HFS energy. This is determined by varying  $\theta_e$  and  $\gamma_h$  by the angular measurement uncertainty.

### 5.3 Neutral Current Measurement Procedure

Inelastic  $ep$  interactions are required to have a well reconstructed interaction vertex to suppress beam induced background events. High  $Q^2$  neutral current events are selected by requiring each event to have a compact and isolated cluster in the electromagnetic part of the LAr calorimeter<sup>4</sup>. The scattered lepton is identified as the cluster of highest transverse momentum. In the central detector region,  $\theta \geq 30^\circ$ , the cluster must be associated to a CTD track. Forward going leptons with  $\theta < 30^\circ$  traverse the region between the FTD and CTD where an increased amount of dead material causes electrons to shower. Since in this kinematic region the scattered lepton has high energy and the contribution from photoproduction background is very small, no tracker information is required to be associated with the lepton for  $\theta < 30^\circ$ .

Energy-momentum conservation requires the variable  $E - P_z$  summed over all final state particles (including the electron) to be approximately equal to twice the initial electron beam energy. Restricting  $E - P_z$  to be greater than 35 GeV considerably reduces the photoproduction background and the radiative processes in which the scattered lepton or bremsstrahlung photons escape undetected in the lepton beam direction.

The photoproduction background increases rapidly with decreasing electron energy, therefore the analysis is separated into two distinct regions: the *nominal* analysis ( $y_e \leq 0.63$  for  $Q_e^2 < 890 \text{ GeV}^2$  and  $y_e < 0.93$  for  $Q_e^2 > 890 \text{ GeV}^2$ ) for which the minimum electron energy is 11 GeV and the *high y* analysis ( $0.63 < y_e < 0.9$  and  $56 < Q_e^2 < 890 \text{ GeV}^2$ ) for which the minimum electron energy is 5 GeV. The techniques employed to contain the background in each analysis are described below.

#### 5.3.1 Nominal Analysis

For the *nominal* analysis the small  $\theta$  background contribution is statistically subtracted using the background simulation. The quality of the background simulation is checked using a sample of data in which a true scattered lepton is observed in the electron tagger which, however, has a small distance.

The comparison of  $e^-p$  data and the simulation is shown in figure 4(a) for the scattered lepton energy spectrum and polar angle, and the distribution of  $E - P_z$ , which are all used in the kinematic reconstruction of  $x$  and  $Q^2$  using the  $e\Sigma$ -method. The corresponding distributions for  $e^+p$  data and simulation are shown in figure 4(b). In the figure the  $R$  and  $L$  data are combined and the simulation is normalised to the luminosity of the data, as is also done for all later performance figures. All distributions are described well by the simulation aside from a small difference in normalisation which is discussed in section 7.2 where the data are compared to the NLO QCD fit.

<sup>4</sup>Small local detector regions are disregarded in the analysis where the cluster of the scattered electron is not fully contained e.g. intermediate space between stacks, or where the trigger is not fully efficient.

For the NC analysis in the region  $y < 0.19$  the noise component has an increasing influence in the transverse momentum balance  $P_{T,h}/P_{T,e}$  through its effect on  $P_{T,h}$ . The event kinematics reconstructed with the  $e\Sigma$ -method in which the HFS is formed from hadronic jets only, limits the noise contribution and substantially improves the  $P_{T,h}/P_{T,e}$  description. The jets are found with the longitudinally invariant  $k_T$  jet algorithm [59, 60] as implemented in FastJet [68, 69] with radius parameter  $R = 1.0$  and are required to have transverse momenta  $P_{T,\text{jet}} > 2 \text{ GeV}$ . In figure 5(a) the quality of the simulation and its description of the  $e^-p$  data for  $y_e < 0.19$  can be seen for the distributions of the  $P_{T,h}/P_{T,e}$ ,  $\gamma_h$ , and  $E - P_z$  where all HFS quantities are obtained using the vector sum of jet four-momenta. Distributions for the  $e^+p$  sample are also shown in figure 5(b). Overall both sets of distributions are well described in shape by the simulation.

At low  $y$ , the forward going hadronic final state particles can undergo interactions with material of the beam pipe. In some cases the products of these secondary interactions are incorrectly assigned as originating from the primary vertex, producing a false determination of the primary interaction vertex position. Such cases are corrected by considering a vertex position calculated using a standard method for the determination of the track associated with the electron cluster [65, 67].

For the nominal analysis the background contribution is low, and this allows the electron candidate track verification on  $\theta \geq 30^\circ$  to be supplemented with an alternative method which increases the efficiency for NC events with no CTD track associated to the electron cluster, the track verification is achieved by searching for hits in the CIP located on the line from the electron cluster to the electron cluster.

The optimised treatment of the vertex determination and verification of the electron cluster using the tracker information improves the reliability of the vertex position determination and increases the efficiency of the procedure to 99.5%.

#### 5.3.2 High y Analysis

In the *high y* region the neutral current analysis is extended to lower energies of the scattered electron,  $E'_e > 5 \text{ GeV}$ . At low energies photoproduction background contributions arise due to  $\pi^0 \rightarrow \gamma\gamma$  decays and charged hadrons being mis-identified as electron candidates. Part of this background is suppressed by requiring a well measured track linked to the calorimeter cluster. The track is furthermore required to have the same charge as the beam lepton. The remaining background in the correctly charged sample is estimated from the number of data events in which the detected lepton has opposite charge to the beam lepton. A charge asymmetry can arise due to the different detector response to particles compared to anti-particles, in particular  $p$  and  $\bar{p}$  [70, 75]. By taking into account the charge asymmetry between negative and positive background, the background estimate is statistically subtracted from the correctly charged sample. The charge asymmetry between fake lepton candidates in the  $e^+p$  and  $e^-p$  data sets is determined by measuring the ratio of wrongly charged fake scattered lepton candidates in  $e^+p$  and  $e^-p$  scattering, taking into account the difference in luminosity. The asymmetry is found to be  $1.03 \pm 0.05$ . This is cross checked using a sample of photoproduction events in which the scattered electron is detected in the electron tagger. Further details are given in [67, 71].

The  $e$ -method using scattered lepton variables alone has the highest precision in this region of phase space and is used to reconstruct the event kinematics.

Example of a paper I wrote - on behalf of my team of 12 researchers



$ep \rightarrow e\gamma$  in the electron and photon "taggers" located at  $z = 35$  and  $z = -103$  cm. A full description of the H1 detector can be found in [20] and [21].

NC events are triggered in the H1 detector mainly using information from the LAr calorimeter. The trigger requires an electromagnetic trigger "tower" to point back to the vertex. For 11 GeV electrons this trigger is more than 99.5% efficient. At lower electron energies the reduced efficiency is supplemented by use of a similar trigger with a lower threshold in co-incidence with an identified track. For 6 GeV electrons the combined efficiency is better than 90% for the high statistics  $e^+p$  data set and 90% for the  $e^-p$  data set.

CC events are triggered on missing transverse momentum  $P_T^{miss}$ . This is determined using the LAr calorimeter vector sum of trigger towers. At low  $P_T^{miss}$  an additional trigger is used requiring hadronic energy pointing to the event vertex with associated track activity measured in the CFC. For a  $P_T^{miss}$  of 12 GeV the efficiency is 60% rising to 90% for  $P_T^{miss}$  of 25 GeV.

### 3.2 Monte Carlo Generation Programs

In order to determine acceptance corrections and background contributions for the DIS cross section measurements, the detector response to events produced by various Monte Carlo (MC) generation programs is simulated in detail using a program based on GEANT [22]. These simulated events are then subjected to the same reconstruction and analysis chain as the real data.

DIS processes are generated using the DJANGO [16] program which is based on LEPTO [15] for the electroweak interaction and on LEPTO [23], using the color dipole model implemented in ARIADNE [24] to generate the QCD dynamics. The hadronization is done using the PYTHIA [25] program for the hadron fragmentation [25]. The simulated events are then analyzed using the next-to-leading order QCD fit performed on and H1 data and is detailed in section 2.2. The fit gives a good description of the data as the "H1-2002 PDF fit" in the following.

The dominant  $ep$  background processes is due to photoproduction of  $e^+e^-$  pairs. These are simulated using the ARIADNE [18] generator with GRV leading order parton distribution functions and photon [19].

### 3.3 Monte Carlo Selection Procedure

High  $Q^2$  NC events are selected by requiring that the event has a compact electromagnetic cluster taken to be the scattered electron, in addition to a vertex position within  $\pm 35$  cm of its nominal position. The cluster is validated by requiring that an extrapolated track have a distance of closest approach to the cluster of less than 12 cm. This loose cluster-track matching requirement is only applied for  $\theta_e \geq 40^\circ$ , where  $\theta_e$  is the polar angle of the scattered electron. In this analysis the polar angle is determined using the position of the electromagnetic cluster. The total  $E - P_z$  summed over all particles is required to be larger than 35 GeV to reduce the photoproduction background, and the influence of QED radiative corrections to the measured cross sections. Fiducial cuts are also made to remove local regions where the electromagnetic energy dep.

A correlated uncertainty of  $1/2/3$  mrad on the determination of the electron polar angle for the region  $\theta_e > 135^\circ / 135^\circ > \theta_e > 120^\circ / \theta_e < 120^\circ$ . The precision of the  $\theta_e$  measurement was checked using a sub-sample of DIS NC events with an accurately measured track associated to the scattered lepton. After alignment of the calorimeter to the tracking chambers, the remaining difference in measurement of  $\theta_e$  from trackers and calorimeter was assigned as a systematic uncertainty. This leads to a typical uncertainty on the NC reduced cross section of less than 1% increasing at high  $Q^2$ .

An uncorrelated 1% uncertainty on the hadronic energy measured within the region  $50 \text{ GeV} > P_{T,h} > 12 \text{ GeV}$ . The uncertainty is increased to 1.7%. In addition, a 1% correlated uncertainty is added in quadrature originating from the calibration of the calorimeter and the uncertainty of the reference scale ( $P_{T,e}$ ). This yields a total uncertainty of 1.7% and 2% in both regions. The resulting influence of correlated systematic uncertainty  $\leq 1\%$  for NC and CC cross sections, but increases at low  $Q^2$ .

An uncertainty on the amount of noise energy subtracted in the LAr calorimeter gives rise to a correlated systematic error at low  $Q^2$ ,  $\approx 10\%$  at  $x = 0.65$  and  $Q^2 \leq 2000 \text{ GeV}^2$  in the NC measurements.

A 7% (3%) uncertainty on the energy of the hadronic final state measured in the SPACAL (tracking system). The influence on the cross section is small compared to the uncorrelated uncertainty of the LAr calorimeter energy, and so the three contributions (LAr, SPACAL, tracks) have been added quadratically, giving rise to the uncorrelated hadronic error which is given in tab. 14 for the NC data and in tab. 15 for the CC data.

The correlated error due to the uncertainty of the efficiency of the anti-photoproduction cut in the CC analysis is estimated by varying the quantity  $V_{\text{cut}}/V_0$  by  $\pm 0.02$ . This leads to a maximum error at low  $P_{T,h}$  of 6%.

In the CC and the NC nominal analysis a 30% uncertainty on the subtracted photoproduction background is estimated from a comparison of data and simulation for a phase space region dominated by photoproduction background. This results in a correlated systematic error of typically 1% for the NC nominal analysis and CC cross sections.

In the NC extended analysis a 10% uncertainty on the charge symmetry of the subtracted photoproduction background is applied. The resulting uncertainty on the measured cross sections is found to be 1% or less.

The following uncertainties, which lead to equivalent uncorrelated systematic errors on the cross sections, have also been taken into account as listed below:

- A 0.5% error originating from the electron identification efficiency in the NC analysis. For  $z_{\text{tag}} > -5$  cm the uncertainty is increased to 2%, where statistics are limited. The precision of this efficiency is estimated using an independent track based electron identification algorithm.

**the editing process was brutal - but helped produce a great final draft!**

Includes ep interaction are used to have a vertex position within  $\pm 35$  cm of its nominal position.

The rest of the cluster is identified with the cluster of highest  $E_{\text{EM}}$ .

Energy from core reg. the scattered  $e^-$  is  $\approx 26$ . Positionally  $E_{\text{EM}} \approx 26$  GeV and the radiative correction is about the size of the beam diameter.

In the NC+CC analysis the  $Q^2$  is estimated from simulation. A 30% uncertainty is determined from the subtracted  $W_{\text{EM}}$ .

In the NC extended analysis the  $Q^2$  is estimated from a charge symmetry of the subtracted photoproduction background.